# System Identification: From Matching Data to Matching Empirical Probabilities

Rudolf Kulhavý*

Honeywell Technology Center Prague and
Institute of Information Theory and Automation,
Academy of Sciences of the Czech Republic
CZ–18208 Prague, Czech Republic
*kulhavy@utia.cas.cz*

**Abstract**

An information view of system identification is presented which is based on projecting the empirical distribution of data onto the model family. Projections onto an exponential family give rise to a specific Pythagorean geometry which serves as a general tool for both analysis of the model accuracy and design of finite memory estimation algorithms.

## 1 Introduction

System identification is usually described as a collection of methods that allow the user to build mathematical models of *dynamic* systems based on measured data. Most of the currently used methods have their roots in statistical techniques. Among them the algorithm of *least squares* (LS) plays quite a special role. One of the reasons for the LS popularity is certainly its simplicity. The possibility to interpret the LS solution as an orthogonal projection of the vector of observed outputs on a subspace spanned by the past data vectors simplifies analysis of the LS properties and provides helpful geometric insight.

However broad is the scope of the LS applications, the LS method does not address many problems of practical interest. It was shown already by Gauss that the LS method implicitly assumes that the deviations of the actual data from the model are normally distributed. In reality, however, non-Gaussian distributions appear quite naturally—as a result of mixing different types of

*This work was supported in part by the grant A2075603 of the Academy of Sciences of the Czech Republic.

disturbances or because of asymmetric bounds on data due to physical constraints. Within the LS framework, there is no straightforward way of taking into account such noise "peculiarities".

Also, the LS method in its pure form makes no use of prior information that may exist about the unknown parameters or the observed data. In cases when data do not carry sufficient information and when prior information is the only way of regularizing estimation, this becomes a serious drawback.

To enlarge the scope of possible applications of the LS algorithm, many *ad hoc* work-arounds have been suggested. For the most part, it was, however, at the cost of losing the original geometric insight.

To cope—in a unified manner—with a variety of "non-standard" cases such as non-Gaussian stochastics, poor system excitation or highly non-linear dynamics, we have to upgrade to a more general statistical framework. The description of uncertainty via *probability* leads us to the realms of frequentist and Bayesian statistics centred around the concepts of likelihood and posterior distribution of unknown quantities, respectively. While these approaches are general enough to deal with the "difficult" cases, their applications are far less frequent than one could expect.

First, compared with the Euclidean distance of data vectors, the likelihood function and posterior probability distribution are considerably more *advanced concepts*. Some training in probability calculus and statistics is typically necessary before one accepts that maximum-likelihood principle or Bayesian inference does the right thing. The classical textbook formulation of statistical estimation sees data as an input to the 'inference engine' rather than the target for direct model-based approximation. Needless to say, most engineers dislike inaccessible 'inference engines', however sophisticated they are.

The other reason why probability-based methods are slow in gaining grounds beyond the LS domain is the notorious *computational complexity* of probabilistic operations. When the limits on computational memory and time are met, either the model or the estimation algorithm has to be approximated. A number of approaches to approximation have been proposed in the past three decades (for surveys see, e.g., Sorenson 1988, Kulhavý 1996), usually motivated by functional-approximation rather than statistical considerations.

We thus face a dilemma to choose between the minimum-distance (LS) framework—simple and intuitive but limited to a specific class of models—and the probability-based frameworks—broadly applicable in principle but difficult to implement for many models of practical interest. This dilemma turns out difficult to resolve within the boundaries of classical paradigms.

The purpose of this work is to show another view of estimation that seem to reconcile some of the above contradictions. While in the LS framework the output vector is projected onto a vector subspace of the past data, here an empirical probability distribution is projected onto an exponential family. In-

stead of matching the data trajectories, we match the frequency distribution of past data. Except for the shift from data to distributions of data, the geometric picture remains much the same. An analogy of Pythagorean geometry and minimum distance properties hold here as well—with Euclidean distance replaced by an appropriate information-theoretic measure.

The view is far from being new. The key ideas were elaborated in the statistical literature long before system identification established as an independent discipline (Kullback and Leibler 1951, Kullback 1959, Wolfowitz 1957). The further theoretical work has concentrated on developing the Pythagorean geometry of information measures and understanding the dual properties of maximum likelihood and maximum entropy projections (Čencov 1972, Csiszár 1975, Amari 1985). On the practical side, information measures have been applied, among others, in system identification (Akaike 1974), fault detection (Basseville and Nikiforov 1993), nonlinear filtering (Brigo *et al.* 1995) and parameter estimation (Kulhavý 1995, Kulhavý 1996).

In most of the literature, information-based estimation is regarded as measuring of an information "distance" between the hypothetical 'true' and model-based distributions of data. Since the 'true' distribution is unknown, all expectations with respect to it are approximated by sample averages. What we do in the following is essentially turning the way of thinking around—the target object for us is the *empirical* distribution of past data which is explicitly approximated by a model-based distribution. Although the views via the 'true' and empirical distributions yield similar results, the latter approach seems to be more natural and also closer to the Bayesian way of thinking.

The rest of the text is organized as follows. Section 2 sums up the basics of probability-based estimation for controlled dynamic systems. Section 3 lists all information measures used in this work. Section 4 presents the view of parameter estimation via conditional inaccuracy. The intuitive appeal of the information-based view is illustrated in Section 5. Section 6 presents a Pythagorean geometry of minimum inaccuracy projections onto conditionally exponential families. The geometry is used in Section 7 to analyse the optimum choice of the regressor structure. Section 8 extends the Pythagorean geometry to jointly exponential families which is used in Section 9 for design of a general finite memory approximation of recursive estimation. The concluding Section 10 gives a summary of major benefits of the information view of system identification.

## 2 Parameter Estimation: Statistical View

The basic problem of system identification is to fit a proper model to a dynamic and possibly controlled system. The model describes the dependence of the

system output on its past values and on possible external inputs.

**Observed Data.** We consider a *sample* of data to be formed by two sequences of continuous random variables

$$Y^{N+m} = (Y_1, \dots, Y_{N+m}), \quad U^{N+m} = (U_1, \dots, U_{N+m})$$

which take values in subsets $\mathcal{Y}$ and $\mathcal{U}$ of $\mathbb{R}^{\dim y}$ and $\mathbb{R}^{\dim u}$, respectively. $U_k$ is defined as a directly manipulated input to the system at time $k$ while $Y_k$ is the system output—response of the system at time $k$ to the past history of data represented by the sequences $Y^{k-1}$ and $U^k$.

The sequences of observed values

$$y^{N+m} = (y_1, \dots, y_{N+m}), \quad u^{N+m} = (u_1, \dots, u_{N+m})$$

form a *realization* of the sample $Y^{N+m}$, $U^{N+m}$.

**General Regression Model.** In the sequel we suppose that the output values $Y_k$ depend on the past data $Y_{k-m}^{k-1}$, $U_{k-m}^{k}$ through a known vector function $Z_k = z(U^k, Y^{k-1})$ that takes values in a subset $\mathcal{Z}$ of $\mathbb{R}^{\dim z}$. Thus $k = m + 1$ is the first instant when $Z_k$ is defined.

More specifically, we assume that $Y^k$ is conditionally independent of $Y^{k-1}$, $U^k$ given $Z_k = z_k$

$$Y^k \perp Y^{k-1}, U^k \mid Z_k$$

for $k = m + 1, \dots, N + m$ which in terms of density functions reads

$$s_k(y_k | y^{k-1}, u^k) = s_k(y_k | z_k). \tag{1}$$

In addition, we assume that the conditional distribution of $Y_k$ given $Z_k = z_k$ is identical for all $k$, $s_k(y|z) = s(y|z)$, and that $(y_N, z_N)$ is recursively computable given its last value $(y_{N-1}, z_{N-1})$ and the latest data $(y_N, u_N)$, i.e., there exists a map $F$ such that $(y_N, z_N) = F\big((y_{N-1}, z_{N-1}), (y_N, u_N)\big)$.

For the purpose of modelling, we suppose that the density $s(y|z)$ comes from the family

$$\mathcal{S} = \{s_\theta(y|z) : \theta \in \mathcal{T}\} \tag{2}$$

parametrized by the parameter $\theta$ taking values in a subset $\mathcal{T}$ of $\mathbb{R}^{\dim \theta}$. We restrict ourselves to the case that $s_\theta(y|z) > 0$ for all $(y, z) \in \mathcal{Y} \times \mathcal{Z}$ and all $\theta \in \mathcal{T}$.

The objective of parameter estimation is to find a proper value of the parameter $\theta$ given the observed sample $y^{N+m}$, $u^{N+m}$.

**Parameter Estimation.** When the system has an external input, the dependence of the input $U_k$ on the past data $Y^{k-1}$, $U^{k-1}$ and the parameter $\theta$ needs to be specified explicitly through a conditional density $\gamma_k(u_k\,|y^{k-1}, u^{k-1}, \theta)$. Luckily, in most cases of practical interest, we may adopt a simplifying assumption that the only information about $\theta$ used for computation of the new input is the information contained in the past data.

More precisely, we can assume that $U_k$ and $\Theta$ (the unknown parameter regarded as a random variable) are conditionally independent given $Y^{k-1} = y^{k-1}$, $U^{k-1} = u^{k-1}$

$$U_k \perp \Theta \,|\, Y^{k-1}, U^{k-1}, \quad k = m+1, \ldots, N+m$$

which in terms of density functions reads

$$\gamma_k(u_k\,|y^{k-1}, u^{k-1}, \theta) = \gamma_k(u_k\,|y^{k-1}, u^{k-1}). \tag{3}$$

The assumption (3), introduced in Peterka (1981) under name 'natural conditions of control', is clearly satisfied when the input is produced by an open-loop input generator, a closed-loop fixed controller (based on prior information) or closed-loop adaptive controller (based on prior information *and* observed data).

Under the modelling assumptions (1), (2) and(3), the joint density $q_\theta^N$ of $Y_{m+1}^{N+m}$ and $U_{m+1}^{N+m}$ conditional on $m$ initial values of $Y_k$ and $U_k$ takes the form

$$q_\theta^N(y_{m+1}^{N+m}, u_{m+1}^{N+m}|y^m, u^m) = \prod_{k=m+1}^{N+m} s_\theta(y_k\,|z_k)\ \gamma_k(u_k\,|y^{k-1}, u^{k-1}). \tag{4}$$

The joint density of observed data $q_\theta^N(y_{m+1}^{N+m}, u_{m+1}^{N+m}|y^m, u^m)$ conditional on the initial values $y^m$ and $u^m$ is called a *likelihood function* when regarded as a function of $\theta$ for given $y^{N+m}$, $u^{N+m}$

$$l_N(\theta) \stackrel{\triangle}{=} q_\theta^N(y_{m+1}^{N+m}, u_{m+1}^{N+m}|y^m, u^m).$$

The subscript $N$ in $l_N(\theta)$ indicates that the likelihood is based on $N$ available data points $(y_{m+1}, z_{m+1}), \ldots, (y_{N+m}, z_{N+m})$.

When the unknown parameter $\theta$ is treated as a random variable $\Theta$, its uncertainty can be described by the *posterior* density conditional on the observed sample $y^{N+m}$, $u^{N+m}$

$$p_N(\theta) \stackrel{\triangle}{=} p(\theta|y^{N+m}, u^{N+m}).$$

Given a prior density $p_0(\theta) = p(\theta|y^m, u^m)$ conditional on available *a priori* information and possibly $m$ initial values $y^m$, $u^m$, the posterior density $p_N(\theta)$ follows by Bayes's theorem. Substituting for the joint density

$q_\theta^N(y_{m+1}^{N+m}, u_{m+1}^{N+m} | y^m, u^m)$ from (4) and taking the natural conditions of control (3) for granted, we obtain the formula

$$p_N(\theta) \propto p_0(\theta) \prod_{k=m+1}^{N+m} s_\theta(y_k | z_k) \tag{5}$$

where $\propto$ stands for equality up to the normalizing factor.

# 3 Preliminaries on Information Measures

For an easy reference, we put here definitions of all information measures used in the following.

**Joint Measures.** Given two joint densities $r(y,z)$ and $s(y,z)$, we define *inaccuracy* (Kerridge 1961) of $r$ relative to $s$ as

$$K(r:s) = \iint r(y,z) \log \frac{1}{s(y,z)} \,\mathrm{d}y \,\mathrm{d}z,$$

*Kullback-Leibler (K-L) distance* (Kullback and Leibler 1951) of $r$ and $s$ as

$$D(r\|s) = \iint r(y,z) \log \frac{r(y,z)}{s(y,z)} \,\mathrm{d}y \,\mathrm{d}z$$

and *Shannon entropy* (Shannon 1948) of $r$ as

$$H(r) = \iint r(y,z) \log \frac{1}{r(y,z)} \,\mathrm{d}y \,\mathrm{d}z$$

The logarithm is always understood to the base $e$.

In Section 8, we use formally the same notation even in the case when $s(y,z)$ is not normalized, i.e., $\iint s(y,z) \,\mathrm{d}y \,\mathrm{d}z \neq 1$.

**Conditional Measures.** Given a joint density $r(y,z)$ and a conditional density $s(y|z)$, we define *conditional inaccuracy* of $r$ relative to $s$ as

$$\bar{K}(r:s) = \iint r(y,z) \log \frac{1}{s(y|z)} \,\mathrm{d}y \,\mathrm{d}z,$$

*conditional Kullback-Leibler distance* of $r$ and $s$ as

$$\bar{D}(r\|s) = \iint r(y,z) \log \frac{r(y,z)}{s(y|z)\,\tilde{r}(z)} \,\mathrm{d}y \,\mathrm{d}z$$

and *conditional Shannon entropy* of $r$ as

$$H(r) = \iint r(y,z)\,\log\frac{\tilde r(z)}{r(y,z)}\,\mathrm{d}y\,\mathrm{d}z$$

where $\tilde r(z) = \int r(y,z)\,\mathrm{d}y$ stands for the marginal density of $Z$.

Alternatively, given conditional densities $s_1(y|z)$ and $s_2(y|z)$ and a marginal density $\tilde r(z)$, we define *conditional inaccuracy* of $s_1$ relative to $s_2$ given $\tilde r$ as

$$\bar K(s_1{:}\,s_2|\tilde r) = \iint s_1(y|z)\,\tilde r(z)\,\log\frac{1}{s_2(y|z)}\,\mathrm{d}y\,\mathrm{d}z,$$

*conditional Kullback-Leibler distance* of $s_1$ and $s_2$ given $\tilde r$ as

$$\bar D(s_1\|s_2|\tilde r) = \iint s_1(y|z)\,\tilde r(z)\,\log\frac{s_1(y|z)}{s_2(y|z)}\,\mathrm{d}y\,\mathrm{d}z$$

and *conditional Shannon entropy* of $s_1$ given $\tilde r$ as

$$H(s_1|\tilde r) = \iint s_1(y|z)\,\tilde r(z)\,\log\frac{1}{s_1(y|z)}\,\mathrm{d}y\,\mathrm{d}z.$$

Note that $\bar D(r\|s) = D(r\|s\tilde r)$ and $\bar D(s_1\|s_2|\tilde r) = D(s_1\tilde r\|s_2\tilde r)$.

**Kullback-Leibler distance.** Even though K-L distance is not a true 'distance' (it is not symmetric, for instance), it is always nonnegative and equal to zero if and only if its arguments coincide almost everywhere. This fundamental property was shown first in Kullback and Leibler (1951).

In particular, the joint K-L distance $D(s_1\|s_2)$ is zero if and only if $s_1(y,z) = s_2(y,z)$. The conditional K-L distance $\bar D(r\|s)$ is zero if and only if $r(y,z) = s(y|z)\,\tilde r(z)$. Similarly, $\bar D(s_1\|s_2|\tilde r)$ is zero if and only if $s_1(y|z)\,\tilde r(z) = s_2(y|z)\,\tilde r(z)$. All equalities are understood almost everywhere on $\mathcal{Y}\times\mathcal{Z}$.

# 4 Parameter Estimation: Information View

The difficulty that a newcomer usually experiences with the formulae (4) and (5) is the lack of any visible mechanism of approximation. It is not immediately apparent why the model maximizing the likelihood (4) or the posterior density (5) should be a good approximation of data. The arguments usually used to convince someone are to show the connection with the LS method for linear normal ARX models, relate the likelihood of a parameter value to the probability of observations or demonstrate how the posterior density concentrates with the increasing number of observations on the 'true' value.

Here we suggest another view of estimation—as a direct approximation of the empirical distribution of data with a model-based distribution.

**Empirical Density.** The empirical distribution is nothing but a limit case of the histogram (frequency function) of data. Given the sample $y^{N+m}$, $u^{N+m}$, the *joint empirical density* of $(Y, Z)$ is defined as a weighted mixture of Dirac functions pointing at particular data points

$$r_N(y,z) \stackrel{\triangle}{=} \frac{1}{N} \sum_{k=m+1}^{N+m} \delta(y \perp y_k, z \perp z_k).$$

Remember the Dirac function $\delta(y,z)$ is defined by the properties $\delta(y,z) = 0$ for $y \neq 0$ or $z \neq 0$ and $\iint \delta(y,z)\,\mathrm{d}y\,\mathrm{d}z = 1$.

The *marginal empirical density* of $Z$ is defined as

$$\tilde{r}_N(z) = \int r_N(y,z)\,\mathrm{d}y = \frac{1}{N} \sum_{k=m+1}^{N+m} \delta(z \perp z_k).$$

**Theoretical Density.** The empirical density $r_N(y,z)$ provides a raw description of the observed data. The only model assumption used in definition of $r_N(y,z)$ concerns the structure of the regressor $Z$. In most applications we prefer, however, to approximate the distribution of $Y$ given $Z = z$ using a density $s_\theta(y|z)$ taken from a suitable parametric family. The density $s_\theta(y|z)$ is called a *theoretical*, *model* or *sampling* density. By using $s_\theta(y|z)$, we accomplish substantial compression of data. While the whole sample $(y^N, u^N)$ is necessary to construct $r_N(y,z)$, the parameter value $\theta$ is sufficient to identify the theoretical density $s_\theta(y|z)$ within a given family $\mathcal{S}$. In addition, through the choice of the parametric family $\mathcal{S}$, we incorporate a substantial piece of prior information. While the empirical density $r_N(y,z)$ describes only the past data behaviour, the theoretical density $s_\theta(y|z)$ allows us to predict the future behaviour of data as well.

**Parameter Estimation.** In terms of the conditional inaccuracy (3), the joint density of sample (4) can be rewritten as

$$q_\theta^N(y_{m+1}^{N+m}, u_{m+1}^{N+m}|y^m, u^m) = \Gamma_N \exp\bigl(\perp N\,\bar{K}(r_N : s_\theta)\bigr) \tag{6}$$

where

$$\Gamma_N = \prod_{k=m+1}^{N+m} \gamma_k(u_k|y^{k-1}, u^{k-1})$$

is a factor independent of $\theta$.

After substitution from (6), the *likelihood function* $l_N(\theta)$ for the observed samples $y^{N+m}$ and $u^{N+m}$ takes the form

$$l_N(\theta) = \Gamma_N \exp\bigl(\perp N\,\bar{K}(r_N : s_\theta)\bigr). \tag{7}$$

Applying Bayes's theorem and substituting for the joint density of sample from (6), we find that the *posterior* density of $\Theta$ conditional on the observed sample $y^{N+m}$, $u^{N+m}$ takes the form

$$p_N(\theta) \propto p_0(\theta)\, \exp\bigl(\perp N\, \bar{K}(r_N : s_\theta)\bigr). \tag{8}$$

Note that the formulae (7) and (8) separate explicitly the key ingredients of estimation—the amount of data $N$, the joint empirical density $r_N(y,z)$, the conditional theoretical density $s_\theta(y\,|\,z)$ and possibly the prior density $p_0(\theta)$.

**How to Interpret Inaccuracy?** To give the reader a better feel for the concept of conditional inaccuracy, we suppose for a while that the observed data take on a finite number of different values only.

Under this assumption, there exists a close connection between the *conditional inaccuracy* $\bar{K}(r_N : s_\theta)$

$$\bar{K}(r_N : s_\theta) = \sum_{(y,z)\in\mathcal{Y}\times\mathcal{Z}} r_N(y,z)\, \log \frac{1}{s_\theta(y\,|\,z)},$$

*conditional Shannon entropy* of the joint empirical mass function $r_N(y,z)$

$$\bar{H}(r_N) = \sum_{(y,z)\in\mathcal{Y}\times\mathcal{Z}} r_N(y,z)\, \log \frac{\tilde{r}_N(z)}{r_N(y,z)}$$

and *conditional Kullback-Leibler (K-L) distance* between the joint empirical and conditional theoretical mass functions $r_N(y,z)$ and $s_\theta(y\,|\,z)$, respectively

$$\bar{D}(r_N \,\|\, s_\theta) = \sum_{(y,z)\in\mathcal{Y}\times\mathcal{Z}} r_N(y,z)\, \log \frac{r_N(y,z)}{s_\theta(y\,|\,z)\,\tilde{r}_N(z)}$$

where $\tilde{r}_N(z) = \sum_{y\in\mathcal{Y}} r(y,z)$ is the marginal empirical mass function. From the definitions it follows immediately that

$$\bar{K}(r_N : s_\theta) = \bar{H}(r_N) + \bar{D}(r_N \,\|\, s_\theta). \tag{9}$$

The conditional inaccuracy can thus be seen as a measure of the average uncertainty of $Y$ given $Z$. The Shannon entropy $\bar{H}(r_N)$ measures the intrinsic uncertainty of $Y$ given $Z$ caused by the random behaviour of data. This component of inaccuracy is "objective"; it cannot be influenced by the choice of $\theta$ but is affected by the definition of $Z$. On the other hand, the K-L distance $\bar{D}(r_N \,\|\, s_\theta)$ quantifies the increase of uncertainty due to the use of the theoretical mass function $s_\theta(y\,|\,z)$ to predict $Y$ given $Z$. To put it different way,

$\bar{D}(r_N \| s_\theta)$ measures the amount of information that speaks against the model; it quantifies the deviation of $r_N(y,z)$ from $s_\theta(y|z)\,\tilde{r}_N(z)$.

Note that the formula (9) does not hold for continuous $(Y,Z)$; the formal evaluation for a discrete $r_N$ and continuous $s_\theta$ gives $\bar{H}(r) = \perp\infty$ and $\bar{D}(r\|s) = \infty$. However, a similar formula can be obtained after replacing the discrete empirical distribution with its smooth approximation (such as the kernel estimate).

# 5 Basic Concepts Revisited

One of the benefits of the information-based view of statistical estimation is in more insight it yields into the fundamental statistical concepts.

## 5.1 Prior

For practical implementation of the Bayesian estimation scheme, it is convenient if the prior density $p_0(\theta)$ is chosen from a *conjugate* family which is closed under conditioning on observed data (cf. Robert 1989). The expression (8) suggests the general form of conjugate priors

$$p_\nu(\theta) \propto \exp\bigl(\perp\nu\,\bar{K}(r_\nu : s_\theta)\bigr). \tag{10}$$

Here $r_\nu(y,z)$ stands for a "prior" density of $(Y,Z)$ built upon prior information and possibly $m$ initial values $y^m$ and $u^m$. The scalar $\nu$ counts the number of actual or fictitious observations $r_\nu(y,z)$ is built on. The density (10) can be interpreted as a "posterior" density given a uniform prior $p_0(\theta) \propto 1$ and $\nu$ data with the empirical density $r_\nu(y,z)$. The scalar $\nu$ is supposed to be nonnegative but not necessarily integer. Its role is to put an appropriate weight (prior belief) on $r_\nu(y,z)$.

Given the conjugate prior (10), the posterior density (8) clearly preserves its form

$$\begin{aligned} p_{\nu+N}(\theta) &\propto p_\nu(\theta)\,\exp\bigl(\perp N\,\bar{K}(r_N : s_\theta)\bigr) \\ &\propto \exp\bigl(\perp\nu\,\bar{K}(r_\nu : s_\theta)\bigr)\,\exp\bigl(\perp N\,\bar{K}(r_N : s_\theta)\bigr) \\ &\propto \exp\bigl(\perp(\nu+N)\,\bar{K}(r_{\nu+N} : s_\theta)\bigr) \end{aligned} \tag{11}$$

where

$$r_{\nu+N}(y,z) = \frac{\nu}{\nu+N}\,r_\nu(y,z) + \frac{N}{\nu+N}\,r_N(y,z)$$

stands for a *mixture* of the prior density $r_\nu(y,z)$ and the empirical density $r_N(y,z)$. In accordance with our intuition, the weight on $r_\nu(y,z)$ tends to zero as $N \to \infty$.

## 5.2 Estimation 'Principles'

It has been observed repeatedly in the literature (cf. Kullback and Leibler 1951, Akaike 1974) that maximizing likelihood coincides asymptotically with minimizing K-L distance between the 'true' and model-based distributions of data. Using the concept of inaccuracy, we get a stronger observation: maximizing likelihood amounts to minimizing inaccuracy between the empirical and model-based distributions of data. Using the conjugate prior, the connection can be extended to maximum a posteriori probability estimation.

**Maximum Likelihood Estimation.** The expression (7) of likelihood makes it possible to characterize the conditional inaccuracy as follows

$$\bar{K}(r_N : s_\theta) = \perp \frac{1}{N} \log l_N(\theta) + \frac{1}{N} \log \Gamma_N.$$

Since logarithm is monotonous and the factor $\Gamma_N$ is independent of $\theta$, the maximum likelihood (ML) estimate

$$\hat{\theta}_{\mathrm{ML}} = \arg \max_{\theta \in \mathcal{T}} l_N(\theta)$$

is equivalent to the minimum *conditional* inaccuracy estimate

$$\hat{\theta}_{\mathrm{ML}} = \arg \min_{\theta \in \mathcal{T}} \bar{K}(r_N : s_\theta) \tag{12}$$

provided the extremum point exists.

**Maximum A Posteriori Estimation.** The maximum a posteriori probability (MAP) estimate defined as

$$\hat{\theta}_{\mathrm{MAP}} \stackrel{\triangle}{=} \arg \max_{\theta \in \mathcal{T}} p_N(\theta)$$

is with respect to the expression (8) of the posterior density equivalent to

$$\hat{\theta}_{\mathrm{MAP}} = \arg \min_{\theta \in \mathcal{T}} \Big( \bar{K}(r_N : s_\theta) \perp \frac{1}{N} \log p_0(\theta) \Big).$$

Compared with the form (12) of the ML estimate, the above is modified by the normalized log-prior. When the conjugate prior (10) is used, the MAP estimate can be put in the following compact form

$$\hat{\theta}_{\mathrm{MAP}} = \arg \min_{\theta \in \mathcal{T}} \bar{K}(r_{\nu+N} : s_\theta). \tag{13}$$

**Bayesian Estimation.** In contrast to point estimation, the objective of Bayesian estimation is to calculate the whole functions

$$\bar{K}(r_N : s_\theta) \quad \text{or} \quad \bar{K}(r_{\nu+N} : s_\theta), \quad \theta \in \mathcal{T},$$

i.e., the conditional inaccuracy of $r_N(y,z)$ or $r_{\nu+N}(y,z)$, respectively, relative to *all* theoretical densities $s_\theta(y|z)$, $\theta \in \mathcal{T}$. The posterior density is constructed from the above functions by means of simple transformations (8) and (11), respectively.

## 5.3 Forgetting

In practice, the parameters $\theta$ of the underlying model are rarely constant, rather they vary slowly in time. The reason is either nonstationarity of the system itself or the fact that the model captures the system behaviour only locally. Whatever the case, the estimation algorithm needs to be modified in order to become capable of tracking parameter changes.

The simplest and most often used solution consists in assigning a lesser weight to the older data. This idea can be expressed quite naturally in terms of the empirical distribution of data. We have seen that for a single structure of the regressor $Z$ a sufficient description of data is given by the pair $(N, r_N)$ where $N > 0$ is the number of observations and $r_N(y,z)$ is the empirical density. Introducing the following operations over the set of all such pairs

$$\begin{aligned}(N_1, r_{N_1}) + (N_2, r_{N_2}) &= \Big(N_1 + N_2, \frac{N_1\, r_{N_1} + N_2\, r_{N_2}}{N_1 + N_2}\Big),\\ \lambda\,(N, r_N) &= (\lambda\, N, r_N),\end{aligned}$$

we can describe the update of $(N, r_N)$ with the new data $(1, \delta_{N+1})$ in the following compact form

$$(N+1, r_{N+1}) = (N, r_N) + (1, \delta_{N+1}). \tag{14}$$

**Plain Forgetting.** Introducing forgetting as a simple discounting of the number of observations, we get a modified recursive algorithm

$$(\nu_{N+1}, r_{N+1}) = \lambda\,(\nu_N, r_N) + (1, \delta_{N+1}) \tag{15}$$

starting from $\nu_0 = 0$. The forgetting factor is chosen as $0 < \lambda < 1$. With $\lambda = 1$, the algorithm (15) coincides with the original one (14). Recursive application of (15) results in

$$\begin{aligned}\nu_{N+1} &= \frac{1 \perp \lambda^{N+1}}{1 \perp \lambda},\\ r_{N+1}(y,z) &= \frac{1}{\nu_{N+1}} \sum_{k=1}^{N+1} \lambda^{N-k+1}\, \delta_k(y,z).\end{aligned}$$

It is easy to see that $\nu_N \to \frac{1}{1-\lambda}$ as $N \to \infty$. Thus, as a result of forgetting, the empirical density becomes a *time-discounted* mixture of Dirac functions pointing at particular data points.

**Regularized Forgetting.** When prior information about the observed data is available as $(\nu_0, r_0)$ with $\nu_0 > 0$ and a prior empirical density $r_0(y, z)$, we can modify the formula (15) so as to preserve the prior information through forgetting

$$(\nu_{0,N+1}, r_{0,N+1}) = \lambda\,(\nu_{0,N}, r_{0,N}) + (1 \perp \lambda)\,(\nu_0, r_0) + (1, \delta_{N+1}). \qquad (16)$$

The recursion starts with $(\nu_{0,0}, r_{0,0}) = (\nu_0, r_0)$. Recursive application of (16) yields

$$\nu_{0,N+1} = \nu_0 + \frac{1 \perp \lambda^{N+1}}{1 \perp \lambda},$$

$$r_{0,N+1}(y, z) = \frac{\nu_0}{\nu_0 + \nu_{N+1}}\, r_0(y, z) + \frac{\nu_{N+1}}{\nu_0 + \nu_{N+1}}\, \frac{1}{\nu_{N+1}} \sum_{k=1}^{N+1} \lambda^{N-k+1}\, \delta_k(y, z).$$

Note that $\nu_{0,N} \to \nu_0 + \frac{1}{1-\lambda}$ as $N \to \infty$. The incorporation of prior information is reflected by the fact that the empirical density $r_{0,N+1}$ is now a *weighted mixture* of the prior density $r_0$ and a time-discounted mixture of Dirac functions pointing at particular data points. The corresponding weights are $\frac{\nu_0}{\nu_0+\nu_{N+1}}$ and $\frac{\nu_{N+1}}{\nu_0+\nu_{N+1}}$, respectively.

# 6 Conditional Pythagorean Geometry

The view of estimation as 'probability matching' centres around the notion of the empirical density. To determine the empirical density, we need, however, have access to the complete sample. This is impractical, especially in recursive estimation when the amount of data grows with time. In cases like that, a condensed description of data is required.

It is well known in statistical theory for the case of independent observations that if the model family is of special—exponential type, estimation can be organized in a recursive form, using only a finite-dimensional statistic of past data. A similar property can be shown to hold even for controlled dynamic systems, but the conditions on the model family are considerably stricter. To stress the difference between the conditionally and jointly exponential families, we split our discussion into two parts—Sections 6 and 8.

**Conditionally Exponential Family.** The model family $\mathcal{S}$ is said to be *conditionally exponential* if it is composed of densities in the form

$$s_\theta(y\,|\,z) = s_0(y\,|\,z)\,\exp\bigl(\theta^T h(y,z) \perp \psi(\theta,z)\bigr) \tag{17}$$

where $s_0(y\,|\,z)$ is any fixed density from $\mathcal{S}$, $\theta \in \mathbb{R}^n$ is a natural (canonical) parameter of the family, $h\colon \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}^n$ is a given vector function (canonical statistic) of single observation $(y,z)$ and $\psi(\theta,z)$ is logarithm of the normalizing divisor

$$\psi(\theta,z) = \log \int s_\theta(y\,|\,z)\,\exp\bigl(\theta^T h(y,z)\bigr)\,\mathrm{d}y.$$

The functions $h_0(y,z) \equiv 1$, $h_1(y,z)$, $\ldots$, $h_n(y,z)$ are supposed linearly independent modulo a function $C(z)$ of $z$. Thus, there is no vector $\theta \neq 0$ such that $\theta^T h(y,z) = C(z)$ where $C(z)$ would be a function of $z$ only. The assumption implies a one-to-one correspondence of $\theta$ and $s_\theta(y\,|\,z)$. Under this assumption, the dimension of the family $\mathcal{S}$ is equal to $n$.

The parameter $\theta$ is assumed to run through all values from $\mathbb{R}^n$ for which the normalizing divisor $\exp\bigl(\psi(\theta,z)\bigr)$ is finite for every $z$.

**Conditional $h$-Projection.** Suppose that a sample $y^{N+m}$, $u^{N+m}$ is given with the empirical density $r_N(y,z)$. The necessary condition for $\hat\theta_N$ to minimize the conditional inaccuracy (to maximize likelihood) $\nabla_\theta\, \bar K(r_N : s_{\hat\theta_N}) = 0$ implies the equality

$$\iint s_{\hat\theta_N}(y\,|\,z)\,\tilde r_N(z)\,h(y,z)\,\mathrm{d}y\,\mathrm{d}z = \iint r_N(y,z)\,h(y,z)\,\mathrm{d}y\,\mathrm{d}z. \tag{18}$$

The density $s_{\hat\theta_N}(y\,|\,z)$ that satisfies the condition (18) will be called a *conditional $h$-projection* of $r_N(y,z)$ onto $\mathcal{S}$. Introducing the notation

$$\bar h_N \stackrel{\triangle}{=} \iint r_N(y,z)\,h(y,z)\,\mathrm{d}y\,\mathrm{d}z = \frac{1}{N}\sum_{k=m+1}^{N+m} h(y_k,z_k),$$

$$\hat h(\theta,z) \stackrel{\triangle}{=} \int s_\theta(y\,|\,z)\,h(y,z)\,\mathrm{d}y,$$

we can rewrite the condition (18) as follows

$$E_N\bigl(\hat h(\hat\theta_N, Z)\bigr) = \bar h_N.$$

The set of all densities $r(y,z)$ with the same conditional $h$-projection as $r_N(y,z)$ has will be denoted as

$$\mathcal{R}_N \stackrel{\triangle}{=} \Bigl\{ r(y,z) : \iint r(y,z)\,h(y,z)\,\mathrm{d}y\,\mathrm{d}z = \bar h_N,\ \iint r(y,z)\,\mathrm{d}y\,\mathrm{d}z = 1,\ r(y,z) \geq 0 \Bigr\}.$$

**Conditional Pythagorean Relationship.** Let $s_\theta(y|z)$ be conditionally exponential (17) and $\hat{\theta}_N$ satisfy (18). By straightforward manipulations we prove that

$$\begin{aligned}
&\bar{K}(r_N{:}\,s_\theta) \perp \bar{K}(r_N{:}\,s_{\hat{\theta}_N}) \\
&\quad = \iint r_N(y,z)\,\log\frac{s_{\hat{\theta}_N}(y|z)}{s_\theta(y|z)}\,\mathrm{d}y\,\mathrm{d}z \\
&\quad = \iint r_N(y,z)\,\log\frac{s_0(y|z)\,\exp\bigl(\hat{\theta}_N^T\,h(y,z) \perp \psi(\hat{\theta}_N,z)\bigr)}{s_0(y|z)\,\exp\bigl(\theta^T h(y,z) \perp \psi(\theta,z)\bigr)}\,\mathrm{d}y\,\mathrm{d}z \\
&\quad = \iint s_{\hat{\theta}_N}(y|z)\,\tilde{r}_N(z)\,\log\frac{s_0(y|z)\,\exp\bigl(\hat{\theta}_N^T\,h(y,z) \perp \psi(\hat{\theta}_N,z)\bigr)}{s_0(y|z)\,\exp\bigl(\theta^T h(y,z) \perp \psi(\theta,z)\bigr)}\,\mathrm{d}y\,\mathrm{d}z \\
&\quad = \iint s_{\hat{\theta}_N}(y|z)\,\tilde{r}_N(z)\;\log\frac{s_{\hat{\theta}_N}(y|z)}{s_\theta(y|z)}\,\mathrm{d}y\,\mathrm{d}z \\
&\quad = \bar{D}(s_{\hat{\theta}_N}\|s_\theta|\tilde{r}_N).
\end{aligned}$$

This implies the following analogue of the Pythagorean relationship (cf. Fig. 1)

$$\bar{K}(r_N{:}\,s_\theta) = \bar{K}(r_N{:}\,s_{\hat{\theta}_N}) + \bar{D}(s_{\hat{\theta}_N}\|s_\theta|\tilde{r}_N). \tag{19}$$

This identity can be regarded as generalization of the well-known Pythagorean theorem that holds for K-L distances between probability distributions of independent observations (Čencov 1972, Csiszár 1975, Amari 1985).

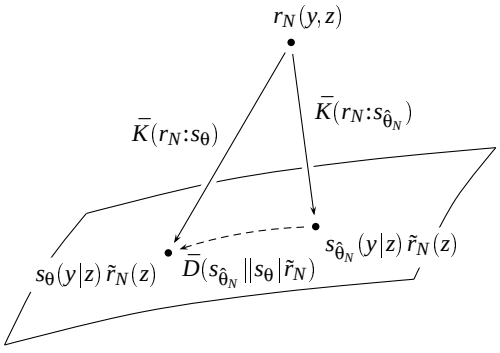


Figure 1: A geometric illustration of the Pythagorean-like decomposition (19) of the conditional inaccuracy.

**Minimum Conditional Inaccuracy Projection.** Suppose that the conditional inaccuracy $\bar{K}(r_N{:}\,s_{\hat{\theta}_N})$ relative to the $h$-projection $s_{\hat{\theta}_N}(y|z)$ is fi-

nite. Then it follows from the Pythagorean relationship (19) and the non-negativeness of K-L distance that for every $\theta$

$$\bar{K}(r_N : s_\theta) \geq \bar{K}(r_N : s_{\hat{\theta}_N})$$

with equality if and only if $s_\theta(y|z)\,\tilde{r}_N(z) = s_{\hat{\theta}_N}(y|z)\,\tilde{r}_N(z)$ almost everywhere. Therefore, the conditional $h$-projection $s_{\hat{\theta}_N}(y|z)$ is a solution to the minimum conditional inaccuracy problem

$$\bar{K}(r_N : s_{\hat{\theta}_N}) = \min_\theta \bar{K}(r_N : s_\theta). \tag{20}$$

It is worth stressing that the uniqueness of $s_{\hat{\theta}_N}(y|z)\,\tilde{r}_N(z)$ does not imply necessarily the uniqueness of $s_{\hat{\theta}_N}(y|z)$ itself.

**Pythagorean View of ARX Model Estimation.** Consider the linear normal ARX model that defines the conditional density

$$s_\theta(y|z) = \frac{1}{\sqrt{2\pi\sigma^2}}\,\exp\Big(\perp\frac{1}{2\sigma^2}(y \perp \theta^T z)^2\Big) \tag{21}$$

where $\theta$ and $z$ are column vectors. The family of all densities $s_\theta(y|z)$ for $\theta \in \mathbb{R}^{\dim z}$ is clearly conditionally exponential

$$s_\theta(y|z) = s_0(y|z)\,\exp\big(\theta^T h(y,z) \perp \psi(\theta,z)\big)$$

with

$$s_0(y|z) = \frac{1}{\sqrt{2\pi\sigma^2}}\,\exp\Big(\perp\frac{1}{2\sigma^2}\,y^2\Big),\;\; h(y,z) = \frac{zy}{\sigma^2},\;\; \psi(\theta,z) = \frac{\theta^T z z^T \theta}{2\sigma^2}.$$

The conditional $h$-projection is given by solving (18)

$$E_N(ZZ^T)\,\hat{\theta}_N = E_N(ZY)$$

where $E_N(\cdot)$ stands for the empirical mean. It is easy to verify that given $\hat{\theta}_N$, the Pythagorean relationship (19) is satisfied for every $\theta$. Indeed, in terms of the following statistics

$$\begin{aligned}
\hat{\theta}_N &= C_N^{-1}\,E_N(ZY),\\
V_N &= E_N(Y^2) \perp E_N(Y\,Z^T)\;C_N^{-1}\,E_N(ZY),\\
C_N &= E_N(ZZ^T),
\end{aligned}$$

where the matrix $C_N$ is supposed positive definite, the particular terms in (19) read

$$\begin{aligned}
\bar{K}(r_N{:}\,s_\theta) &= \frac{1}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}V_N + \frac{1}{2\sigma^2}\big(\theta \perp \hat{\theta}_N\big)^T C_N \big(\theta \perp \hat{\theta}_N\big),\\
\bar{K}(r_N{:}\,s_{\hat{\theta}_N}) &= \frac{1}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}V_N,\\
\bar{D}(s_{\hat{\theta}_N}\|s_\theta\,|\tilde{r}_N) &= \frac{1}{2\sigma^2}\big(\theta \perp \hat{\theta}_N\big)^T C_N \big(\theta \perp \hat{\theta}_N\big)
\end{aligned}$$

which clearly satisfy the Pythagorean relationship.

It follows directly from the above formulae that the minimum conditional inaccuracy over all possible values of $\theta$ is attained at $\theta = \hat{\theta}_N$

$$\min_\theta \bar{K}(r_N{:}\,s_\theta) = \bar{K}(r_N{:}\,s_{\hat{\theta}_N}).$$

We can try to make the conditional inaccuracy even smaller by optimizing the value of $\sigma^2$. Simple computation shows that the minimum conditional inaccuracy over all possible values of $\sigma^2$ is attained at $\sigma^2 = V_N$

$$\min_{\theta,\sigma^2} \bar{K}(r_N{:}\,s_\theta) = \frac{1}{2}\log 2\pi V_N + \frac{1}{2}.$$

Within the class of linear normal ARX models with a fixed structure of $Z$, we cannot do better.

# 7 Choice of Model Class

The last example raises the natural question: Can we achieve a better fit of data using a different model class$\Gamma$ And does the minimum inaccuracy value tell us something about the "distance" from the global minimum over *all possible models*$\Gamma$ The Pythagorean relationship gives us some insight again.

**Limits of Probability Matching.** Suppose that the empirical density $r_N(y,z)$ with the increasing number of samples $N$ "approaches" a density $s^*(y|z)\,\tilde{r}_N(z)$ where $s^*(y|z)$ is a fixed continuous density so that we can set $\bar{K}(r_N{:}\,s_\theta) \approx \bar{K}(s^*{:}\,s_\theta|\tilde{r}_N)$. Let the model family $\mathcal{S}$ be conditionally exponential. Then, analogously as in (19), we can derive the following Pythagorean relationship

$$\bar{K}(s^*{:}\,s_\theta\,|\tilde{r}_N) = \bar{K}(s^*{:}\,s_{\hat{\theta}_N}\,|\tilde{r}_N) + \bar{D}(s_{\hat{\theta}_N}\|s_\theta\,|\tilde{r}_N).$$

Moreover, as $s^*(y|z)$ is now continuous, we can decompose the conditional inaccuracy $\bar{K}(s^*{:}\,s_{\hat{\theta}_N}\,|\tilde{r}_N)$ as follows

$$\bar{K}(s^*{:}\,s_{\hat{\theta}_N}\,|\tilde{r}_N) = \bar{H}(s^*\,|\tilde{r}_N) + \bar{D}(s^*\|s_{\hat{\theta}_N}\,|\tilde{r}_N).$$

As a result, we have

$$\bar{K}(s^*\colon s_\theta\,|\,\tilde{r}_N) = \bar{H}(s^*\,|\,\tilde{r}_N) + \bar{D}(s^*\|s_{\hat{\theta}_N}\,|\,\tilde{r}_N) + \bar{D}(s_{\hat{\theta}_N}\|s_\theta\,|\,\tilde{r}_N). \tag{22}$$

The particular terms in (22) can be given intuitive interpretations.

1. The conditional entropy term $\bar{H}(s^*\,|\,\tilde{r}_N)$ measures the average uncertainty (residual randomness) of the output $Y$ given a specific structure of the regressor structure $Z$. The purpose of modelling is to make—by a proper choice of the regressor structure—the conditional entropy $\bar{H}(s^*\,|\,\tilde{r}_N)$ as small as possible. It is important to realize that the value of the conditional entropy cannot decrease by augmenting the optimum regressor. For instance, if $s^*(y\,|\,z_\mathrm{a}, z_\mathrm{b}) = s^*(y\,|\,z_\mathrm{a})$, i.e., if $Y$ is conditionally independent of $Z_\mathrm{b}$ given $Z_\mathrm{a} = z_\mathrm{a}$, then

$$\begin{aligned}&\iiint s^*(y\,|\,z_\mathrm{a}, z_\mathrm{b})\,\tilde{r}_N(z_\mathrm{a}, z_\mathrm{b})\,\log\frac{1}{s^*(y\,|\,z_\mathrm{a}, z_\mathrm{b})}\,\mathrm{d}y\,\mathrm{d}z_\mathrm{a}\,\mathrm{d}z_\mathrm{b}\\ &= \iint s^*(y\,|\,z_\mathrm{a})\,\tilde{r}_N(z_\mathrm{a})\,\log\frac{1}{s^*(y\,|\,z_\mathrm{a})}\,\mathrm{d}y\,\mathrm{d}z_\mathrm{a}.\end{aligned}$$

2. The K-L distance $\bar{D}(s^*\|s_{\hat{\theta}_N}\,|\,\tilde{r}_N)$ measures the average discrepancy between the actual distribution of data $s^*(y\,|\,z)$ and the minimum inaccuracy (maximum likelihood) estimate of model distribution $s_{\hat{\theta}_N}(y\,|\,z)$ given the marginal empirical density $\tilde{r}_N(z)$. This term is influenced by the choice of the model family and can be made, in principle at least, arbitrarily small by careful fitting of the actual distribution of data.

3. The K-L distance $\bar{D}(s_{\hat{\theta}_N}\|s_\theta\,|\,\tilde{r}_N)$ measures the average discrepancy between the minimum inaccuracy density $s_{\hat{\theta}_N}(y\,|\,z)$ and a given distribution $s_\theta(y\,|\,z)$. This can easily be made zero by setting $\theta = \hat{\theta}_N$ but from the modelling perspective it is more interesting to analyse how the K-L distance changes over the whole model class or at least around the minimum inaccuracy point. The K-L distance behaviour is clearly affected by the choice of the model class.

The above picture can easily be extended so as to take prior information into account. Formally, the only change consists in replacing the "raw" empirical density $r_N(y, z)$ with the mixture $r_{\nu+N}(y, z)$ of prior and empirical density.

**Model Validation.** In off-line system identification, we are usually interested in obtaining a "globally valid" model, i.e., model that fits reasonably well the process behaviour at all operating points of interest and that accounts even for small variations from the 'nominal' behaviour.

The above analysis must be taken with care then. Since we always work with just a finite sample of data, we must count with the fact that the empirical distributions built from different data segments (different "realizations") differ from each other. The variations are the larger, the shorter is the length $N$ of samples. The fact that the behaviour of real data is always changing in time only adds to the empirical density variations.

If we are able to collect $M$ separate samples, we can construct the empirical densities $r_{N_j}(y,z)$, $j=1,\ldots,M$ for each of them and then to solve

$$\min_{\theta} \sum_{j=1}^{M} \pi_j \, \bar{K}(r_{N_j} : s_\theta)$$

where $\pi_j \geq 0$, $\sum_{j=1}^{M} \pi_j = 1$ are relative weights of particular samples in the criterion. Owing to the linearity of inaccuracy in $r_{N_j}$, the optimization amounts to solving

$$\min_{\theta} \bar{K}\left(\sum_{j=1}^{M} \pi_j \, r_{N_j} : s_\theta\right).$$

The mixture density $\sum_{j=1}^{M} \pi_j \, r_{N_j}(y,z)$ lies within the convex hull of densities $r_{N_j}(y,z)$, $j=1,\ldots,M$. When the weights are proportional to the lengths $N_j$ of the samples, the solution to the optimization problem coincides with the minimum inaccuracy estimate for a single sample produced by juxtaposition of given samples. The length of the "mixed" sample is, however, correspondingly shorter, given by $\sum_{j=1}^{M} \pi_j N_j$.

Alternatively, we can formulate estimation as min-max problem

$$\min_{\theta} \max_{j} \bar{K}(r_{N_j} : s_\theta).$$

It can be verified using the Lagrange multiplier method that the solution to the min-max problem coincides with the solution to the previous problem—with a special set of 'least favourable' weights $\pi_j$ following from the optimization.

When separate samples are not available, we can "cheat" by generating subsamples from a given sample, e.g., using a fixed-length window moving over the sample. Alternatively, we can use the time-discounted (exponentially-weighted) empirical densities at particular time instants.

For linear normal ARX models, the above algorithms result in mixing the statistics $E_{N_j}(ZY)$ and $E_{N_j}(ZZ^T)$ with the weights $\pi_j$.

The usual textbook recommendation is "use one part of available data for estimation and the other part for model validation" and then (in terms of inaccuracy) "check that the conditional inaccuracy for the validation data is not significantly larger than that obtained for the estimation data". The above approach goes beyond that simple measure but the results should be treated with extreme care anyway.

**Optimum Model Structure.** Consider several distinct model structures, i.e., different definitions of the regressor $Z_k = z(U^k, Y^{k-1})$. Suppose that all the model families are conditionally exponential. To measure the goodness-of-fit obtained with particular model structures, we have to get rid of dependence of the inaccuracy $\bar{K}(r_N : s_\theta)$ on a particular model density $s_\theta(y|z)$.

Adopting the Bayesian point of view, the likelihood value corresponding to the $i$-th structure is obtained by integration of the joint density of observed data and model parameters over the unknown parameters. From (4), (6), (11) it follows that

$$l_N(i) = \Gamma_N \int p(\theta) \prod_{k=m+1}^{N+m} s_\theta(y_k|z_k)\,\mathrm{d}\theta = \Gamma_N \frac{\int \exp\bigl(\perp(\nu+N)\,\bar{K}(r_{\nu+N} : s_\theta)\bigr)\,\mathrm{d}\theta}{\int \exp\bigl(\perp\nu\,\bar{K}(r_\nu : s_\theta)\bigr)\,\mathrm{d}\theta}$$

where $\Gamma_N$ is a constant independent of $i$. Owing to the Pythagorean relationship (19), the integrals in both the numerator and denominator can be decomposed as follows

$$\int \exp\bigl(\perp N\,\bar{K}(r_N : s_\theta)\bigr)\,\mathrm{d}\theta = \exp\bigl(\perp N\,\bar{K}(r_N : s_{\hat{\theta}_N})\bigr) \int \exp\bigl(\perp N\,\bar{D}(s_{\hat{\theta}_N} : s_\theta\,|\,\tilde{r}_N)\bigr)\,\mathrm{d}\theta.$$

By Taylor series expansion of the log-density $\log s_\theta(y|z)$ around $\hat{\theta}_N$, we obtain a quadratic approximation of the conditional K-L distance

$$\bar{D}(s_{\hat{\theta}_N} : s_\theta\,|\,\tilde{r}_N) \approx \frac{1}{2}\,(\theta \perp \hat{\theta}_N)^T C_N(\hat{\theta}_N)\,(\theta \perp \hat{\theta}_N)$$

where $C_N(\theta)$ stands for the empirical expectation of the conditional Fisher information matrix

$$\begin{aligned} C_N(\theta) &= \int \tilde{r}_N(z) \int s_\theta(y|z)\,\bigl(\nabla_\theta \log s_\theta(y|z)\bigr)\bigl(\nabla_\theta \log s_\theta(y|z)\bigr)^T\,\mathrm{d}y\,\mathrm{d}z \\ &= E_N\,E_\theta\,\bigl(h(Y,Z) \perp \hat{h}(\theta,Z)\bigr)\bigl(h(Y,Z) \perp \hat{h}(\theta,Z)\bigr)^T. \end{aligned}$$

In particular, for the linear normal ARX model, $C_N(\theta) = E_N\bigl(\frac{ZZ^T}{\sigma^2}\bigr)$.

With the above approximation, the integral can be evaluated analytically

$$\begin{aligned} \int \exp\bigl(\perp N\,\bar{D}(s_{\hat{\theta}_N} : s_\theta\,|\,\tilde{r}_N)\bigr)\,\mathrm{d}\theta &\approx \int \exp\Bigl(\perp\frac{N}{2}\,(\theta \perp \hat{\theta}_N)^T C_N(\hat{\theta}_N)\,(\theta \perp \hat{\theta}_N)\Bigr)\,\mathrm{d}\theta \\ &= \left(\frac{2\pi}{N}\right)^{\frac{\dim\theta}{2}} \bigl|C_N(\hat{\theta}_N)\bigr|^{-\frac{1}{2}} = \left(\frac{N\bar{\gamma}}{2\pi}\right)^{-\frac{\dim\theta}{2}} \end{aligned}$$

where $\bar{\gamma}$ denotes the geometric mean of all eigenvalues of the matrix $C_N(\hat{\theta}_N)$.

Putting the above steps together, we get the following approximation

$$\perp \log \int \exp\bigl(\perp N\,\bar{K}(r_N\!:\!s_\theta)\bigr)\,\mathrm{d}\theta \approx N\,\bar{K}(r_N\!:\!s_{\hat{\theta}_N}) + \frac{\dim\theta}{2}\,\log\frac{N\,\bar{\gamma}}{2\pi}.$$

Here the first term (increasing with $N$) measures the minimum inaccuracy attainable over the model family while the second term (increasing with $\log N$) penalizes the model complexity expressed through $\dim\theta$.

As a result, the negative log-likelihood for the $i$-th structure takes the form

$$\perp \log\frac{l_N(i)}{\Gamma_N} \approx (\nu{+}N)\,\bar{K}(r_{\nu+N}\!:\!s_{\hat{\theta}_{\nu+N}}) \perp \nu\,\bar{K}(r_\nu\!:\!s_{\hat{\theta}_\nu}) + \frac{\dim\theta}{2}\,\log\Bigl(\frac{\nu+N}{\nu}\,\frac{\bar{\gamma}_{\nu+N}}{\bar{\gamma}_\nu}\Bigr).$$

Note that the last two formulae hold exactly for the linear normal ARX models.

Alternatively, adopting the frequentist point of view, we can compute expectation of $\bar{K}(r_N\!:\!s_\theta)$ directly. This results in the well-known Akaike's information criterion (Akaike 1974, Matsuoka and Ulrych 1986).

# 8 Joint Pythagorean Geometry

Throughout Section 6 we have supposed that the marginal empirical density $\tilde{r}_N(z)$ is available to us. This is, of course, unrealistic in most cases. In order to get rid of dependence on the whole empirical density, we have to restrict further the model family $\mathcal{S}$. Below we show that if $\mathcal{S}$ is jointly exponential, then it is sufficient to know the minimum conditional inaccuracy projection of the empirical density $r_N(y,z)$ onto $\mathcal{S}$ in order to restore the conditional inaccuracy $\bar{K}(r_N\!:\!s_\theta)$ with precision up to an additive constant.

**Joint Exponential Family.** We call the family $\mathcal{S}$ *jointly exponential* if it is composed of densities in the form

$$s_\theta(y|z) = s_0(y|z)\,\exp\Bigl(\lambda^T(\theta)\,h(y,z) \perp \psi\bigl(\lambda(\theta)\bigr)\Bigr) \tag{23}$$

where $s_0(y|z)$ is any fixed density from $\mathcal{S}$, $\lambda \in \mathbb{R}^n$ is a vector function of $\theta$, $h\colon \mathcal{Y}\times\mathcal{Z} \mapsto \mathbb{R}^n$ is a vector function (canonical statistic) of $(y,z)$ and $\psi(\lambda(\theta))$ is independent of $z$.

It should be emphasized that the definition (23) is much stronger than (17). The conditionally exponential family (17) is jointly exponential (23) if the logarithm of the normalizing divisor $\psi(\theta,z)$ can be factorized as $\psi(\theta,z) = \mu^T g(z) + \psi(\theta)$ for some vector parameter $\mu$ and vector function $g(z)$. The dimension of $h(y,z)$ in (23) may be considerably larger than the dimension of $h(y,z)$ in (17).

We introduce an exponential family $\mathcal{W}$ of joint densities

$$w_\lambda(y,z) = s_0(y|z)\,\exp\bigl(\lambda^T h(y,z) \perp \psi(\lambda)\bigr) \tag{24}$$

where $\lambda \in \mathbb{R}^n$ is a natural (canonical) parameter of the family and

$$\psi(\lambda) = \log \iint s_\theta(y|z)\,\exp\bigl(\lambda^T h(y,z)\bigr)\,\mathrm{d}y\,\mathrm{d}z$$

is logarithm of the normalizing divisor.

The functions $h_0(y,z) \equiv 1$, $h_1(y,z)$, ..., $h_n(y,z)$ are supposed linearly independent which implies a one-to-one correspondence between the vector parameter $\lambda$ and the joint density $w_\lambda(y,z)$. The dimension of $\mathcal{S}$ then equals $n$.

The parameter $\lambda$ is assumed to run through all values from $\mathbb{R}^n$ for which the normalizing divisor $\exp\bigl(\psi(\lambda)\bigr)$ is finite.

**Joint *h*-Projection.** In the sequel we use the fact that the conditional inaccuracy $\bar{K}(r_N\!:\!s_\theta)$ can be formally regarded as the *unnormalized joint inaccuracy* of $r_N(y,z)$ relative to the function $s_\theta(y|z)$, $\bar{K}(r_N\!:\!s_\theta) = K(r_N\!:\!s_\theta)$.

The necessary condition for $\hat{\lambda}_N$ to minimize the unnormalized joint inaccuracy $\nabla_\lambda K(r_N\!:\!s_{\hat{\lambda}_N}) = 0$ implies

$$\iint w_{\hat{\lambda}_N}(y,z)\,h(y,z)\,\mathrm{d}y\,\mathrm{d}z = \iint r_N(y,z)\,h(y,z)\,\mathrm{d}y\,\mathrm{d}z. \tag{25}$$

The density $w_{\hat{\lambda}_N}(y,z)$ that satisfies the condition (25) will be called a *joint h-projection* of $r_N(y,z)$ onto $\mathcal{S}$. Introducing the notation

$$\bar{h}_N \stackrel{\triangle}{=} \iint r_N(y,z)\,h(y,z)\,\mathrm{d}y\,\mathrm{d}z = \frac{1}{N}\sum_{k=m+1}^{N+m} h(y_k,z_k),$$

$$\hat{h}(\lambda) \stackrel{\triangle}{=} \iint w_\lambda(y,z)\,h(y,z)\,\mathrm{d}y\,\mathrm{d}z,$$

we can rewrite (25) as

$$\hat{h}(\hat{\lambda}_N) = \bar{h}_N.$$

The set of all densities $r(y,z)$ with the same $h$-projection as $r_N(y,z)$ has will be denoted as $\mathcal{R}_N$ again.

**Joint Pythagorean Relationship.** Let $w_\lambda(y,z)$ be exponential (24) and $\hat{\lambda}$ satisfy (25). By straightforward manipulations of (25) we prove that

$$K(r_N\!:\!s_\theta) \perp K(r_N\!:\!w_{\hat{\lambda}_N})$$

$$
\begin{aligned}
&= \iint r_N(y,z)\, \log \frac{w_{\hat{\lambda}_N}(y,z)}{s_\theta(y|z)}\, \mathrm{d}y\, \mathrm{d}z \\
&= \iint r_N(y,z)\, \log \frac{s_0(y|z)\, \exp\bigl(\hat{\lambda}_N^T\, h(y,z) \perp \psi(\hat{\lambda}_N)\bigr)}{s_0(y|z)\, \exp\bigl(\lambda^T(\theta)\, h(y,z) \perp \psi(\lambda(\theta))\bigr)}\, \mathrm{d}y\, \mathrm{d}z \\
&= \iint w_{\hat{\lambda}_N}(y,z)\, \log \frac{s_0(y|z)\, \exp\bigl(\hat{\lambda}_N^T\, h(y,z) \perp \psi(\hat{\lambda}_N)\bigr)}{s_0(y|z)\, \exp\bigl(\lambda^T(\theta)\, h(y,z) \perp \psi(\lambda(\theta))\bigr)}\, \mathrm{d}y\, \mathrm{d}z \\
&= \iint w_{\hat{\lambda}_N}(y,z)\, \log \frac{w_{\hat{\lambda}_N}(y,z)}{s_\theta(y|z)}\, \mathrm{d}y\, \mathrm{d}z \\
&= D(w_{\hat{\lambda}_N} \| s_\theta)
\end{aligned}
$$

where we used the notation $D(w_{\hat{\lambda}_N} \| s_\theta)$ for the *unnormalized* joint K-L distance. We have thus obtained another analogue of the Pythagorean relationship (cf. Fig. 2)

$$K(r_N{:}\, s_\theta) = K(r_N{:}\, w_{\hat{\lambda}_N}) + D(w_{\hat{\lambda}_N} \| s_\theta). \tag{26}$$

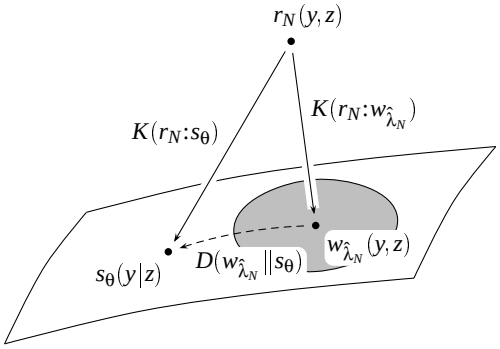


Figure 2: A geometric illustration of the Pythagorean-like decomposition (26) of the conditional inaccuracy. The projection "surface" corresponds to the set of all functions $C\, s_0(y|z)\, \exp(\lambda^T h(y,z))$ with $C > 0$. The shaded area indicates a subset of normalized densities with $C = \exp(\perp\psi(\lambda))$.

**Minimum Inaccuracy Projection.** Assume that the joint inaccuracy $K(r_N{:}\, w_{\hat{\lambda}_N})$ of $r_N(y,z)$ relative to the $h$-projection $w_{\hat{\lambda}_N}(y,z)$ is finite. The following Pythagorean relationship clearly holds for every $\lambda$

$$K(r_N{:}\, w_\lambda) \perp K(r_N{:}\, w_{\hat{\lambda}_N}) = D(w_{\hat{\lambda}_N} \| w_\lambda).$$

Since the joint K-L distance $D(w_{\hat{\lambda}_N} \| w_\lambda)$ is nonnegative, we have

$$K(r_N{:}\, w_\lambda) \geq K(r_N{:}\, w_{\hat{\lambda}_N})$$

with equality if and only if $w_\lambda(y,z) = w_{\hat{\lambda}_N}(y,z)$ almost everywhere. Thus, the joint $h$-projection $w_{\hat{\lambda}_N}(y,z)$ is a unique solution to the minimum inaccuracy problem

$$K(r_N : w_{\hat{\lambda}_N}) = \min_\lambda K(r_N : w_\lambda). \tag{27}$$

**Minimum K-L Distance Projection.** The joint $h$-projection can also be given a dual interpretation. Assume that the unnormalized K-L distance of the $h$-projection $w_{\hat{\lambda}_N}(y,z)$ and the conditional model density $s_\theta(y|z)$ is finite. Then there are $r(y,z) \in \mathcal{R}_N$ such that $D(r\|s_\theta) < \infty$. For every such $r(y,z) \in \mathcal{R}_N$, the following Pythagorean relation holds

$$D(r\|s_\theta) \perp D(r\|w_{\hat{\lambda}_N}) = D(w_{\hat{\lambda}_N}\|s_\theta).$$

Since the joint K-L distance $D(r\|w_{\hat{\lambda}_N})$ is nonnegative, we have for every $r(y,z)$ such that $D(r\|s_\theta) < \infty$,

$$D(r\|s_\theta) \geq D(w_{\hat{\lambda}_N}\|s_\theta)$$

with equality if and only if $r(y,z) = w_{\hat{\lambda}_N}(y,z)$ almost everywhere. Thus, the joint $h$-projection $w_{\hat{\lambda}_N}(y,z)$ is a unique solution to the minimum K-L distance problem

$$D(w_{\hat{\lambda}_N}\|s_\theta) = \min_{r \in \mathcal{R}_N} D(r\|s_\theta). \tag{28}$$

**Data Compression.** From the obvious identity $\min_\lambda D(w_{\hat{\lambda}_N}\|w_\lambda) = 0$ we can derive the dual expression

$$D(w_{\hat{\lambda}_N}\|s_\theta) = \max_\lambda \big(\lambda^T \bar{h}_N \perp \psi(\lambda)\big). \tag{29}$$

Taking together (26) and (29), we have

$$K(r_N : s_\theta) = K(r_N : w_{\hat{\lambda}_N}) + \max_\lambda \big(\lambda^T \bar{h}_N \perp \psi(\lambda)\big). \tag{30}$$

This expression reveals that if the model family is jointly exponential, then $K(r_N : w_{\hat{\lambda}_N})$ is constant for all $\theta$ so that all we need to know about the observed data is the empirical expectation $\bar{h}_N$ of the statistic $h(Y,Z)$.

**Pythagorean View of ARX Model Estimation.** For the linear normal ARX model, the family of all densities $s_\theta(y|z)$ (21) for $\theta \in \mathbb{R}^{\dim z}$ is clearly jointly exponential

$$s_\theta(y|z) = s_0(y|z)\, \exp\Big(\lambda_1(\theta)^T h_1(y,z) + \operatorname{tr}\big(\Lambda_2(\theta)\, h_2(y,z)\big) \perp \psi\big(\lambda(\theta)\big)\Big)$$

with

$$\lambda_1(\theta) = \theta,\ \Lambda_2(\theta) = \theta\theta^T,\ h_1(y,z) = \frac{zy}{\sigma^2},\ h_2(y,z) = \perp\frac{zz^T}{2\sigma^2},\ \psi\big(\lambda(\theta)\big) = 0.$$

The joint $h$-projection is given by solving (25)

$$E_{\hat\lambda_N}(ZZ^T)\,\hat\lambda_{1,N} = E_N(ZY), \quad E_{\hat\lambda_N}(ZZ^T) = E_N(ZZ^T).$$

# 9 Finite Memory Approximation.

Except for the linear normal ARX models, most practically interesting model families cannot be imbedded in a jointly exponential family of a finite (reasonably small) dimension. When looking for a well-justified approximation, we find the joint Pythagorean geometry useful again.

**Approximation of Inaccuracy.** The "trick" in applying the joint Pythagorean geometry to general model families is to define the family $\mathcal{W}_\theta$ separately for each of the model points $s_\theta(y|z)$. In doing so, we set $w_{\theta,\lambda}(y|z) = w_\lambda(y,z)$ for $s_0(y|z) = s_\theta(y|z)$. With this notation, the Pythagorean relation (26) takes the form

$$K(r_N{:}\,s_\theta) = K(r_N{:}\,w_{\theta,\hat\lambda_N}) + D(w_{\theta,\hat\lambda_N}\|s_\theta). \tag{31}$$

The inaccuracy $K(r_N{:}\,w_{\theta,\hat\lambda_N})$ is not independent of $\theta$ any more but we can attempt to make it almost constant through a proper choice of $h(y,z)$. Suppose that $K(r_N{:}\,w_{\hat\lambda_N}) \approx C$ where $C$ is a constant independent of $\theta$. Then the Pythagorean relationship (31) together with (28) and (29) imply

$$\bar K(r_N{:}\,s_\theta) \approx C + D(\mathcal{R}_N\|s_\theta) \tag{32}$$

with

$$\begin{aligned} D(\mathcal{R}_N\|s_\theta) &\stackrel{\triangle}{=} \min_{r\in\mathcal{R}_N} D(r\|s_\theta) \\ &= \max_\lambda \Big(\lambda^T\bar h_N \perp \log\iint s_\theta(y|z)\,\exp\big(\lambda^T h(y,z)\big)\,\mathrm{d}y\,\mathrm{d}z\Big). \end{aligned} \tag{33}$$

**Approximation of Likelihood and Posterior.** Substituting from (32) for $\bar K(r_N{:}\,s_\theta)$ in (7), we obtain the following approximate expression of the likelihood function

$$\hat l_N(\theta) = \Gamma_N\,\exp\big(\perp N\,D(\mathcal{R}_N\|s_\theta)\big). \tag{34}$$

Similarly, applying the same substitution in (8), we obtain the approximate posterior density in the form

$$\hat p_N(\theta) \propto p_0(\theta)\,\exp\big(\perp N\,D(\mathcal{R}_N\|s_\theta)\big). \tag{35}$$

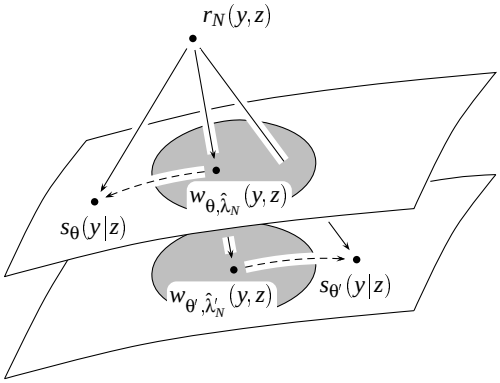


Figure 3: The idea of approximate estimation for dependent data. Instead of computing the unmeasurable inaccuracy $\bar{K}(r_N : s_\theta)$, the minimum K-L distance $D(\mathcal{R}_N \| s_\theta) = D(w_{\theta,\hat{\lambda}_N} \| s_\theta)$ is computed for every $\theta \in \mathcal{T}$ (cf. Fig. 2).

**Key Properties of Approximation.** The unnormalized joint K-L distance $D(r\|s_\theta)$ can be decomposed as follows

$$D(r\|s_\theta) = \iint r(y,z)\, \log \frac{r(y,z)}{s_\theta(y\,|\,z)\,\tilde{r}(z)}\, \mathrm{d}y\, \mathrm{d}z \perp \int \tilde{r}(z)\, \log \frac{1}{\tilde{r}(z)}\, \mathrm{d}z$$

provided all the integrals exist. This implies

$$D(r\|s_\theta) = \bar{D}(r\|s_\theta) \perp H(\tilde{r}). \tag{36}$$

Hence, when minimizing $D(r\|s_\theta)$ over $r \in \mathcal{R}_N$, we seek a compromise between minimizing the conditional K-L distance $\bar{D}(r\|s_\theta)$ and maximizing the marginal Shannon entropy $H(\tilde{r})$. In other words, we look for a trade-off between attaining the best fit of model to data given a particular $\tilde{r}(z)$ and choosing the maximum-entropy $\tilde{r}(z)$ from $\mathcal{R}_N$.

Taking together the identity (36), the definition (33) of the minimum K-L distance $D(\mathcal{R}_N\|s_\theta)$ and the nonnegativity of the conditional K-L distance $\bar{D}(r\|s_\theta)$, we get the following bounds on $D(\mathcal{R}_N\|s_\theta)$

$$\perp \max_{r\in\mathcal{R}_N} H(\tilde{r}) \le D(\mathcal{R}_N\|s_\theta) \le D(r\|s_\theta)$$

for all $r(y,z) \in \mathcal{R}_N$ .

It follows directly from the definition (33) that the minimum K-L distance $D(\mathcal{R}_N\|s_\theta)$ regarded as a function of the set argument $\mathcal{R}_N$ is (anti)monotonous in the sense that

$$\mathcal{R}_N \subseteq \mathcal{R}'_N \quad \text{implies} \quad D(\mathcal{R}_N\|s_\theta) \ge D(\mathcal{R}'_N\|s_\theta).$$

**Choice of Statistic.** As shown above, the choice of the statistic $h(Y, Z)$ is essential in order to make $K(r_N : w_{\theta,\hat{\lambda}_N})$ almost constant. The minimum requirement is to choose $h(y, z)$ so that $E_N\big(h(Y, Z)\big)$ is a *necessary statistic*, i.e., a function of any sufficient statistic (Dynkin 1951). The following definitions of $h(y, z)$ can be shown to result in necessary statistics:

1. Choose $n + 1$ points $\theta_1^*, \ldots, \theta_{n+1}^*$ in the parameter space $\mathcal{T}$ and set

$$h_i(y, z) = \log s_{\theta_{i+1}^*}(y \,|\, z) \perp \log s_{\theta_i^*}(y \,|\, z). \tag{37}$$

2. Provided that $\log s_\theta(y|z)$ is differentiable at every $\theta \in \mathcal{T}$ for all $(y, z) \in \mathcal{Y} \times \mathcal{Z}$, choose $n$ points $\theta_1^*, \ldots, \theta_n^*$ in the parameter space $\mathcal{T}$ and $n$ vectors $\omega_1^*, \ldots, \omega_n^*$ from $\mathbb{R}^{\dim \theta}$ and set

$$h_i(y, z) = \omega_i^{*T} \nabla_\theta \, \log s_{\theta_i^*}(y \,|\, z). \tag{38}$$

3. Choose $n$ weighting functions $w_1^*(\theta), \ldots, w_n^*(\theta)$ such that $\int w_i^*(\theta) \, \mathrm{d}\theta = 0$ for $i = 1, \ldots, n$ and set

$$h_i(y, z) = \int w_i^*(\theta) \, \log s_\theta(y \,|\, z) \, \mathrm{d}\theta. \tag{39}$$

Taking into account the connection between the empirical expectation of the log-density $\log s_\theta(Y \,|\, Z)$ and the log-likelihood $\log l_N(\theta)$

$$E_N\big(\log s_\theta(Y \,|\, Z)\big) = \text{const.} + \frac{1}{N} \log l_N(\theta),$$

the empirical expectation of $h$-statistics (37), (38), (39) yields

$$\begin{aligned}
E_N\left(\log \frac{s_{\theta_{i+1}^*}(Y \,|\, Z)}{s_{\theta_i^*}(Y \,|\, Z)}\right) &= \frac{1}{N} \log \frac{l_N(\theta_{i+1}^*)}{l_N(\theta_i^*)}, \\
E_N\Big(\omega_i^{*T} \nabla_\theta \, \log s_{\theta_i^*}(Y \,|\, Z)\Big) &= \frac{1}{N} \omega_i^{*T} \nabla_\theta \, \log l_N(\theta_i^*), \\
E_N\Big(\int w_i^*(\theta) \, \log s_\theta(Y \,|\, Z) \, \mathrm{d}\theta\Big) &= \frac{1}{N} \int w_i^*(\theta) \, \log l_N(\theta) \, \mathrm{d}\theta,
\end{aligned}$$

respectively.

# 10 Concluding Remarks

The benefits of the information-based view of estimation can be seen in three major directions.

**More Insight.** The information-based view brings a lot of insight into statistical estimation by showing its intuitive minimum "distance" interpretation which is normally hidden behind the concepts of likelihood or posterior (cf. Cover and Thomas 1991, Blahut 1987). When trying to maximize likelihood, we solve essentially an inverse problem—trying to guess on the model that yields a large enough value of the likelihood. When minimizing the conditional inaccuracy, we solve a direct problem—shaping the conditional density so as to match the empirical distribution. Even though the practical implementation of the latter approach may be difficult, it provides a helpful mental framework mainly when dealing with non-linear and non-Gaussian models.

**Extensions of Known Results.** Some of the facts well-known in the LS framework or for linear normal ARX models can be generalized—without losing the intuitive appeal—to more general, namely non-Gaussian models. The Pythagorean relationship naturally generalizes the bias-variance considerations (cf. Barron and Sheu 1991). The expression of conjugate prior through the empirical density offers an alternative expression of prior information—directly in terms of observed data rather than parameters of a particular model (Kulhavý and Tesař 1997). The use of information measures makes it possible to formulate exponential and linear forgetting strategies as decision problems whose solution result in naturally regularized strategies (Kulhavý and Kraus 1996).

**Finite Memory Estimation.** The finite memory approximations (34) and (35) have potentially a great importance for on-line system identification. The efficient numerical implementation of (33) is still a topic of active research. While the optimization problem is relatively easy to solve, the numerical integration is known to be cumbersome in higher dimensions. It is generally agreed that for dimensions greater than $4 \perp 6$ only Monte Carlo techniques can be used effectively. What decides about the computational complexity then is the efficiency of generating a sample from the multivariate probability distribution. The Metropolis or Langevin methods of sampling are currently considered as most powerful.

The optimization problem (33) needs to be solved separately for each $\theta$. Therefore, without further approximation, the algorithm can be used 'as is' only when the number of models considered at the same time is limited. This is the case of *evolutionary modelling*, for instance, when the class of potential models (with possibly different structures of regressor, different nonlinearities and different noise characteristics) is dynamically evolving—new candidates are generated in the neighbourhood of the best model found so far. The algorithm (33) can also be used to get a better idea about the shape of the "cost function" $\bar{K}(r_N : s_\theta)$ at larger distances from $\hat{\theta}_N$ where the asymptotic normality cannot

be used. Another possible application is approximation of the generalized likelihood ratio test for nonlinear or non-Gaussian models (Kulhavý 1997).

For the purpose of on-line identification, the core map $(\bar{h}_N, \theta) \mapsto D(\mathcal{R}_N \| s_\theta)$ needs to be computed and explicitly approximated before estimation—using, e.g., a neural network or nonparametric regression.

# References


Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **19**, 716–723.

Amari, S. (1985). *Differential-Geometrical Methods in Statistics*. Vol. 28 of *Lecture Notes in Statistics*. Springer-Verlag. Berlin.

Barron, A. R. and C. H. Sheu (1991). Approximation of density functions by sequences of exponential families. *Ann. Statist.* **19**(3), 1347–1369.

Basseville, M. and I. V. Nikiforov (1993). *Detection of Abrupt Changes: Theory and Application*. Prentice Hall. Englewood Cliffs, NJ.

Blahut, R. E. (1987). *Principles and Practice of Information Theory*. Addison-Wesley. Reading, MA.

Brigo, D., B. Hanzon and F. Le Gland (1995). A differential-geometric approach to nonlinear filtering: the projection filter. In: *Proceedings of the 34th IEEE Conference on Decision and Control*. Vol. 4. New Orleans, LA. pp. 4006–4011.

Čencov, N. N. (1972). *Statistical Decision Rules and Optimal Inference* (in Russian). Nauka. Moscow. English translation in *Translations of Mathematical Monographs* **53** (1982), Amer. Math. Soc., Providence, RI.

Cover, T. M. and J. A. Thomas (1991). *Elements of Information Theory*. Wiley. New York.

Csiszár, I. (1975). $I$-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* **3**(1), 146–158.

Dynkin, E. B. (1951). Necessary and sufficient statistics for a family of probability distributions (in Russian). *Uspekhi matem. nauk* **VI**(1), 68–90.

Kerridge, D. F. (1961). Inaccuracy and inference. *J. Roy. Statist. Soc. Ser. B* **23**, 284–294.

Kulhavý, R. (1995). A Kullback-Leibler distance approach to system identification. In: *Preprints of the IFAC Symposium on Adaptive Systems in Control and Signal Processing*. Budapest, Hungary. pp. 55–66.

Kulhavý, R. (1996). *Recursive Nonlinear Estimation: A Geometric Approach*. Vol. 216 of *Lecture Notes in Control and Information Sciences*. Springer-Verlag. London.

Kulhavý, R. (1997). Approximate fault detection and isolation using compressed data. In: *IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes*. Hull, UK.

Kulhavý, R. and F. J. Kraus (1996). On duality of regularized exponential and linear forgetting. *Automatica* **32**, 1403–1415.

Kulhavý, R. and L. Tesař (1997). On dual expression of prior information in Bayesian parameter estimation. In: *11th IFAC Symposium on System Identification*. Fukuoka, Japan.

Kullback, S. (1959). *Information Theory and Statistics*. Wiley. New York.

Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *Ann. Math. Statist.* **22**, 79–86.

Matsuoka, T. and T. J. Ulrych (1986). Information theory measures with application to model identification. *IEEE Trans. Acoust., Speech, Signal Processing* **34**, 511–517.

Peterka, V. (1981). Bayesian approach to system identification. In: *Trends and Progress in System Identification* (P. Eykhoff, Ed.). Chap. 8, pp. 239–304. Pergamon. Elmsford, N.Y.

Robert, C. P. (1989). *The Bayesian Choice: A Decision-Theoretic Motivation*. Springer-Verlag. Berlin.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* **26**, 379–423, 623–656.

Sorenson, H. W. (1988). Recursive estimation for nonlinear dynamic systems. In: *Bayesian Analysis of Time Series and Dynamic Models* (J. C. Spall, Ed.). pp. 127–165. Marcel Dekker. New York.

Wolfowitz, J. (1957). The minimum distance method. *Ann. Math. Statist.* **28**, 75–88.