Recursive Nonlinear Estimation of Non-linear/Non-Gaussian Dynamic Models

Rudolf Kulhavý

Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
P.O. Box 18, 182 08 Prague, Czech Republic
kulhavy@utia.cas.cz

Abstract

While the general theory of recursive Bayesian estimation of dynamic models is well developed, its practical implementation is restricted to a narrow class of models, typically models with linear dynamics and Gaussian stochastics. The theoretically optimal solution is infeasible for non-linear and/or non-Gaussian models due to its excessive demands on computational memory and time. Parameter estimation of such models requires approximation of the theoretical solution. The paper describes one possible framework for such approximation that is based on measuring of Kullback-Leibler distance between the empirical and theoretical distributions of observed data.

1. Introduction

Suppose we are to control a dynamic system that depends on some unknown parameter θ. We have basically two options how to cope with the uncertainty of θ. Either we satisfy with a point estimate of θ, or we take the uncertainty of θ into account. In the latter case, the expectation in a cost function applies to both the stochastic behaviour of the system and the uncertainty of the parameter θ. This converts the original problem into a hyperproblem [1] that has no more unknowns. Its *information state* is formed by the posterior probability density function of the original state and the parameter θ conditional on the observed data.

The beauty of the latter—*Bayesian* approach is that the resulting solution looks as if all uncertainty vanished. As soon as the prior density is chosen, the state evolves in a definite way, governed by the laws of probability theory. The appeal of the solution is paid, however, by the immense dimension of the information state. Unless the problem has a finite-dimensional statistic, there is no feasible way of updating the full information state.

The limitation is an inherent difficulty of the Bayesian inference that is not bound to just the control problem. The problem is particularly pressing when estimation is to be recursive and adaptive. The recursive character of computations calls for massive compression of data which causes that the posterior density can be restored from compressed data only approximately. In addition, to make estimation adaptive—capable of tracking parameter variations, only a limited amount of past data relevant for the current system behaviour can be taken into account. Both the features produce significant posterior uncertainty of the unknown parameters which cannot be neglected or easily approximated. This makes approximate Bayesian estimation a delicate matter where one must really care how far he is from the theoretically optimal solution.

Approximation of recursive Bayesian estimation was investigated in engineering science intensively in the late 1960s and early 1970s. The increased interest was largely stimulated by the success of Kalman filter for linear problems. A variety of methods have been developed then for non-linear and non-Gaussian models. They used either local simplification (typically linearization) of a given model in estimated parameters, or functional approximation of posterior densities using more tractable functions. A good survey of these techniques can be found in [2], [3].

Roughly speaking, most of the known methods try to bring the problem back, at least locally, to a linear one, or to simplify intermediate results of estimation using classical approximation theory. The sequential character of estimation is given little attention if any. Insufficient theoretical insight needs to be compensated by simulation or practical experience with a particular algorithm.

The paper presents a new approach to systematic design of approximate Bayesian estimation [4] which is based upon the use of information measures and Pythagorean geometry of probability spaces. Section 2 sums up the essentials of probability-based estimation for controlled dynamic systems. The classical solution is put in a new perspective in Section 3 where we introduce the key concept of inaccuracy of the empirical distribution of observed data relative to their theoretical distribution. Section 4 provides us with a major theoretical tool which is Pythagorean-like decomposition of the inaccuracy into sum of two terms—one independent of

observed data given the value of a certain statistic, the other independent of the parameter of an approximating exponential family. The Pythagorean relationship is used in Section 5 to justify a conceptual scheme of approximate estimation. The main properties of the approximation are discussed in Section 6. Section 7 shows how to choose a suitable data statistic that preserves as much information from data as possible. The idea of approximate estimation is illustrated in Section 8 for an autoregressive model with a non-Gaussian noise. The concluding Section 9 indicates possible applications of approximate estimation in system identification.

2. Probability-based Estimation

The basic problem of system identification is to fit a proper model to a dynamic, possibly controlled system. The models used in system identification typically describe the dependence of the system output on its past values and possibly on some external inputs as well.

Sample of Data: Consider a system on which two sequences of continuous random variables are measured

$$\begin{aligned} Y^{N+m} &= (Y_1,\dots,Y_{N+m}),\\ U^{N+m} &= (U_1,\dots,U_{N+m}) \end{aligned}$$

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 output—response of the system at time k to the past history of data represented by the sequences Y^{k-1} and U^k. Both the above sequences form together a *sample* of data.

A sequence of observed (measured) values

$$\begin{aligned} y^{N+m} &= (y_1,\dots,y_{N+m}),\\ u^{N+m} &= (u_1,\dots,u_{N+m}) \end{aligned}$$

is called a *realization* of the sample Y^{N+m}, U^{N+m}, an *observed sample* or a *given sample*.

General Regression: Suppose that the output values Y_k depend on a limited amount of past data, namely Y_{k-m}^{k-1}, U_{k-m}^{k} through a known vector function $Z_k = z(U^k, Y^{k-1})$ taking values in a subset $\mathcal{Z}$ of $\mathbb{R}^{\dim z}$. More precisely, assume that Y^k is conditionally independent [5] of Y^{k-1}, U^k given $Z_k = z_k$

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

for $k = m+1,\dots,N+m$. In terms of density functions, the condition reads

$$s_k(y_k|y^{k-1},u^k) = s_k(y_k|z_k). \tag{2}$$

In addition, we assume that the conditional distribution of Y_k given $Z_k = z_k$ is identical for all k

$$s_k(y|z) = s(y|z).$$

Finally, it is assumed 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

$$(y_N, z_N) = F\big((y_{N-1}, z_{N-1}), (y_N, u_N)\big).$$

Model Family: We assume that the density $s(y|z)$ comes from a given family

$$\mathcal{S} = \{s_\theta(y|z) : \theta \in \mathcal{T}\} \tag{3}$$

parametrized by the parameter θ 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 θ given the observed sample y^{N+m}, u^{N+m}.

Natural Conditions of Control: Let the dependence of the input U_k on the past data Y^{k-1}, U^{k-1} and the parameter θ be expressed through a conditional density $\gamma_k(u_k|y^{k-1},u^{k-1},\theta)$. In many cases of practical interest, we may adopt a simplifying assumption that the only information about θ used for computation of the new input is the information contained in the past data.

More precisely, we assume that U_k and Θ, interpreted as a random variable, are conditionally independent given $Y^{k-1} = y^{k-1}$, $U^{k-1} = u^{k-1}$

$$U_k \perp \Theta \,|\, Y^{k-1}, U^{k-1}, \quad k = m+1,\dots,N+m. \tag{4}$$

which, in terms of density functions, reads

$$\gamma_k(u_k|y^{k-1},u^{k-1},\theta) = \gamma_k(u_k|y^{k-1},u^{k-1}). \tag{5}$$

Note that the condition (5) introduced in [6] is really natural in control of technological processes. The condition is clearly satisfied when the input is produced by an open-loop input generator, a closed-loop fixed controller (pretuned using prior information) or closed-loop adaptive controller (based on prior information *and* observed data).

Joint Density: By chain rule, the joint density q_θ^N of Y_{m+1}^{N+m} and U_{m+1}^{N+m} conditional on m initial values of Y_k and U_k can be rewritten as follows

$$\begin{aligned} &q_\theta^N(y_{m+1}^{N+m}, u_{m+1}^{N+m}|y^m,u^m)\\ &\quad= \prod_{k=m+1}^{N+m} s_\theta(y_k|y^{k-1},u^k)\,\gamma_k(u_k|y^{k-1},u^{k-1},\theta). \end{aligned}$$

Taking into account the conditional independence assumption (2) and the natural conditions of control (5), we obtain

$$\boxed{\begin{aligned} &q_\theta^N(y_{m+1}^{N+m}, u_{m+1}^{N+m}|y^m,u^m)\\ &\quad= \prod_{k=m+1}^{N+m} s_\theta(y_k|z_k)\,\gamma_k(u_k|y^{k-1},u^{k-1}). \end{aligned}} \tag{6}$$

Likelihood Function: 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 it is regarded as a function of θ for given y^{N+m}, u^{N+m}

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

We use the subscript N to indicate N data points available

$$(y_{m+1}, z_{m+1}), \ldots, (y_{N+m}, z_{N+m}).$$

Posterior Density: When the unknown parameter θ is treated as a random variable Θ, its uncertainty can naturally be expressed through the *posterior* density conditional on the observed sample y^{N+m}, u^{N+m}

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

The subscript N indicates again conditioning on N data points

$$(y_{m+1}, z_{m+1}), \ldots, (y_{N+m}, z_{N+m}).$$

Given a prior density conditional on available *a priori* information and possibly m initial values y^m, u^m

$$p_0(\theta) \stackrel{\triangle}{=} p(\theta|y^m, u^m),$$

the posterior density $p_N(\theta)$ follows by Bayes's theorem [5]. Substituting for the joint density $q_\theta^N(y_{m+1}^{N+m}, u_{m+1}^{N+m}|y^m, u^m)$ from (6) and taking the natural conditions of control (5) for granted, we obtain

$$\begin{aligned} & p_N(\theta) \\ & \propto p_0(\theta)\, q_\theta^N(y_{m+1}^{N+m}, u_{m+1}^{N+m}|y^m, u^m) \\ & \propto p_0(\theta)\, l_N(\theta) \\ & \propto p_0(\theta) \prod_{k=m+1}^{N+m} s_\theta(y_k|z_k)\, \gamma_k(u_k|y^{k-1}, u^{k-1}) \end{aligned}$$

where $\propto$ stands for equality up to a normalizing factor. It follows that

$$\boxed{p_N(\theta) \propto p_0(\theta) \prod_{k=m+1}^{N+m} s_\theta(y_k|z_k).} \tag{7}$$

The computation of the posterior density can easily be organized recursively

$$p_k(\theta) \propto p_{k-1}(\theta)\, s_\theta(y_{k+m}|z_{k+m})$$

for $k = 1, \ldots, N$.

Prior Density: The piece of information contained in the initial data y^m, u^m can be used, in principle, to update the prior, unconditional density $p(\theta)$. Bayes's theorem gives the clue

$$p_0(\theta) \propto p(\theta)\, q_\theta^0(y^m, u^m). \tag{8}$$

In practice, however, the piece of information carried by y^m, u^m is usually neglected and $p_0(\theta) = p(\theta)$ [6].

3. Estimation via Inaccuracy

Borrowing the notion of inaccuracy from information theory, we can transpose probability-based estimation into the form of an explicit approximation problem.

Empirical Density: Given the sample y^{N+m}, u^{N+m}, a *joint empirical density* of (Y, Z) is defined as

$$r_N(y,z) = \frac{1}{N} \sum_{k=m+1}^{N+m} \delta(y \perp y_k, z \perp z_k) \tag{9}$$

where $\delta_{y,z}$ is a Dirac function satisfying $\delta(y,z) = 0$ for $y \neq 0$ or $z \neq 0$ and

$$\iint_{\mathcal{Y} \times \mathcal{Z}} \delta(y,z)\, \mathrm{d}y\, \mathrm{d}z = 1.$$

Similarly as with likelihood and posterior density, we use the subscript N to indicate the number of data points

$$(y_{m+1}, z_{m+1}), \ldots, (y_{N+m}, z_{N+m})$$

the empirical density is based on.

The empirical density can be updated recursively according to

$$r_k(y,z) = \frac{k \perp 1}{k}\, r_{k-1}(y,z) + \frac{1}{k}\, \delta(y \perp y_{k+m}, z \perp z_{k+m}) \tag{10}$$

for $k = 1, \ldots, N$.

We shall denote the *marginal empirical density* of Z as

$$\begin{aligned} \tilde{r}_N(z) &= \int r_N(y,z)\, \mathrm{d}y \\ &= \frac{1}{N} \sum_{k=m+1}^{N+m} \delta(z \perp z_k). \end{aligned} \tag{11}$$

Theoretical Density: The empirical density $r_N(y,z)$ represents a raw description of observed data which is not "contaminated" by any model assumption—except the structural assumption about the conditional independence, i.e., the definition of Z. Yet, in most applications we prefer 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 *theoretical* or *model* or *sampling* density.

Note that by using $s_\theta(y|z)$, we drastically reduce the complexity of computations. While the whole sample (y^N, u^N) is basically needed to construct $r_N(y,z)$, the parameter value θ 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 bring a substantial piece of prior information into play. While the empirical density $r_N(y,z)$ describes only the past data, the theoretical density $s_\theta(y|z)$ makes it possible to predict the future behaviour of data as well.

Conditional Inaccuracy: Given the joint empirical density $r_N(y,z)$ and a conditional theoretical density $s_\theta(y|z)$, we define *conditional inaccuracy* as

$$\bar{K}(r_N{:}s_\theta) \triangleq \iint r_N(y,z)\,\log\frac{1}{s_\theta(y|z)}\,\mathrm{d}y\,\mathrm{d}z. \qquad (12)$$

The concept of conditional inaccuracy is generalization of Kerridge's innaccuracy [7] introduced for the case of independent and identically distributed data.

Joint Density of Sample: The joint density of sample (6) can be rewritten as

$$q_\theta^N(y_{m+1}^{N+m}, u_{m+1}^{N+m}|y^m,u^m) = \Gamma_{N+m}\prod_{k=m+1}^{N+m} s_\theta(y_k|z_k)$$

where

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

is a factor independent of θ. Using conditional inaccuracy, we can rewrite the θ-dependent part as follows

$$\begin{aligned}
&\prod_{k=m+1}^{N+m} s_\theta(y_k|z_k)\\
&= \exp\left(N\frac{1}{N}\sum_{k=1}^{N}\log s_\theta(y_k|z_k)\right)\\
&= \exp\left(\perp N\iint r_N(y,z)\,\log\frac{1}{s_\theta(y)}\,\mathrm{d}y\,\mathrm{d}z\right)\\
&= \exp\left(\perp N\bar{K}(r_N{:}s_\theta)\right).
\end{aligned}$$

Note that we made use of the assumption $s_\theta(y_k|z_k) > 0$.

As a result, we have the following expression

$$\boxed{\begin{aligned}&q_\theta^N(y_{m+1}^{N+m}, u_{m+1}^{N+m}|y^m,u^m)\\ &\quad= \Gamma_{N+m}\exp\left(\perp N\bar{K}(r_N{:}s_\theta)\right).\end{aligned}} \qquad (13)$$

Likelihood Function: The *likelihood function* $l_N(\theta)$ for given samples y^{N+m} and u^{N+m}, i.e., the joint density $q_\theta^N(y_{m+1}^{N+m}, u_{m+1}^{N+m}|y^m,u^m)$ taken as a function of the unknown parameter θ takes after substitution from (13) the form

$$\boxed{l_N(\theta) = \Gamma_{N+m}\exp\left(\perp N\bar{K}(r_N{:}s_\theta)\right).} \qquad (14)$$

Posterior Density: Applying Bayes's theorem and substituting for the joint density of sample from (13), we find that the *posterior* density of Θ conditional on the observed sample y^{N+m}, u^{N+m} takes the form

$$\begin{aligned}
p_N(\theta) &\propto p_0(\theta)\,q_\theta^N(y_{m+1}^{N+m}, u_{m+1}^{N+m}|y^m,u^m)\\
&\propto p_0(\theta)\,l_N(\theta)\\
&\propto p_0(\theta)\exp\left(\perp N\bar{K}(r_N{:}s_\theta)\right).
\end{aligned}$$

The resulting expression

$$\boxed{p_N(\theta) \propto p_0(\theta)\exp\left(\perp N\bar{K}(r_N{:}s_\theta)\right)} \qquad (15)$$

separates explicitly the key ingredients of Bayesian estimation—the amount of data, empirical and theoretical densities of observed data and prior density of unknown parameters.

4. Pythagorean Geometry of Estimation

Kullback-Leibler distance: Given two probability density functions $s(y)$ and $s'(y)$, *Kullback-Leibler (K-L) distance* [8] between s and s' is defined as

$$D(s||s') \triangleq \int s(y)\,\log\frac{s(y)}{s'(y)}\,\mathrm{d}y \qquad (16)$$

where the logarithm is understood to the base e.

Analogously, we can define *conditional Kullback-Leibler distance* between the joint density $r(y,z)$ and conditional density $s(y|z)$ as

$$\bar{D}(r||s) \triangleq \iint r(y,z)\,\log\frac{r(y,z)}{s(y|z)\,\tilde{r}(z)}\,\mathrm{d}y\,\mathrm{d}z \qquad (17)$$

where $\tilde{r}(z) = \int r(y,z)\,\mathrm{d}y$ denotes the marginal density of Z. The definition (17) can be formally rewritten as

$$\bar{D}(r||s) = D(r||s\tilde{r}) \qquad (18)$$

using the convention $0\log\frac{0}{0} = 0$ whenever $\tilde{r}(z) = 0$.

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

$$\begin{aligned}
\bar{K}(r_N{:}s_\theta) &= \iint r_N(y,z)\,\log\frac{1}{s_\theta(y|z)}\,\mathrm{d}y\,\mathrm{d}z\\
&= K(r_N{:}s_\theta).
\end{aligned}$$

Joint Exponential Family: Given any density $s_\theta(y|z)$ for a particular value of θ, we can construct a *joint* exponential family [9] $\mathcal{S}_{\theta;h}$ composed of the joint densities

$$s_{\theta,\lambda}(y,z) = s_\theta(y|z)\exp\left(\lambda^T h(y,z) \perp \psi(\theta,\lambda)\right) \qquad (19)$$

where $\lambda \in \mathbb{R}^n$ is a natural or canonical parameter of the family, $h{:}\,\mathcal{Y}\times\mathcal{Z}\mapsto\mathbb{R}^n$ is a given function (canonical statistic) of (y,z) and

$$\psi(\theta,\lambda) = \log\iint s_\theta(y|z)\exp\left(\lambda^T h(y,z)\right)\mathrm{d}y\,\mathrm{d}z \qquad (20)$$

is logarithm of the normalizing divisor.

It is assumed that the functions $h_0(y,z)\equiv 1, h_1(y,z), \ldots, h_n(y,z)$ are linearly independent. Since two densities

$s_{\theta,\lambda}(y,z)$ and $s_{\theta,\lambda'}(y,z)$ are equal if and only if the right-hand side of

$$\log \frac{s_{\theta,\lambda}(y,z)}{s_{\theta,\lambda'}(y,z)} = (\lambda \perp \lambda')^T h(y,z) \perp \psi(\theta,\lambda) + \psi(\theta,\lambda')$$

vanishes, the assumption implies a one-to-one correspondence between the vector parameter λ and the joint density $s_{\theta,\lambda}(y,z)$. The dimension of $\mathcal{S}_{\theta;h}$ then equals n.

Normalizing Divisor: The parameter λ is assumed to run through all values from $\mathbb{R}^n$ for which the normalizing divisor is finite

$$\exp\big(\psi(\theta,\lambda)\big) < \infty.$$

The set of all such values of λ will be denoted by $\mathcal{N}_\theta$. It can be shown by Hölder inequality [5] that the set $\mathcal{N}_\theta$ is convex and $\psi(\theta,\lambda)$ is a convex function of λ on $\mathcal{N}_\theta$. In the sequel, $\mathcal{S}_{\theta;h}$ is understood to be the maximal family of densities that can be expressed as (19) for some $\lambda \in \mathbb{R}^n$.

***h*-Projection:** Suppose a sample y^{N+m}, u^{N+m} is given with the empirical density $r_N(y,z)$. The necessary condition for $\hat{\lambda}$ to minimize the unnormalized joint inaccuracy (and maximize likelihood) is

$$\begin{aligned} 0 &= \nabla_\lambda K(r_N : s_{\theta,\hat{\lambda}}) \\ &= \iint r_N(y,z) \Big(\perp h(y,z) \\ &\quad + \iint s_{\theta,\hat{\lambda}}(y,z)\, h(y,z)\, \mathrm{d}y\,\mathrm{d}z \Big)\, \mathrm{d}y\,\mathrm{d}z \\ &= \perp \iint r_N(y,z)\, h(y,z)\, \mathrm{d}y\,\mathrm{d}z \\ &\quad + \iint s_{\theta,\hat{\lambda}}(y,z)\, h(y,z)\, \mathrm{d}y\,\mathrm{d}z \end{aligned}$$

that is

$$\boxed{\begin{aligned} &\iint s_{\theta,\hat{\lambda}}(y,z)\, h(y,z)\, \mathrm{d}y\,\mathrm{d}z \\ &\quad = \iint r_N(y,z)\, h(y,z)\, \mathrm{d}y\,\mathrm{d}z. \end{aligned}} \qquad (21)$$

The density $s_{\theta,\hat{\lambda}}(y,z)$ that satisfies the condition (21) will be called a *h-projection* of $r_N(y,z)$ onto $\mathcal{S}_{\theta;h}$. Introducing the notation

$$\begin{aligned} \bar{h}_N &\triangleq \iint r_N(y,z)\, h(y,z)\, \mathrm{d}y\,\mathrm{d}z \\ &= \frac{1}{N} \sum_{k=m+1}^{N+m} h(y_k, z_k), \qquad (22) \\ \hat{h}(\theta,\lambda) &\triangleq \iint s_{\theta,\lambda}(y,z)\, h(y,z)\, \mathrm{d}y\,\mathrm{d}z, \qquad (23) \end{aligned}$$

we can write (21) as

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

We denote the set of all densities $r(y,z)$ with the same h-projection as

$$\begin{aligned} \mathcal{R}_N \triangleq \Big\{ r(y,z) : &\iint r(y,z)\, h(y,z)\, \mathrm{d}y\,\mathrm{d}z = \bar{h}_N, \\ &\iint r(y,z)\, \mathrm{d}y\,\mathrm{d}z = 1,\ r(y,z) \geq 0 \Big\}. \qquad (24) \end{aligned}$$

The expectation $\hat{h}(\theta,\lambda)$ can be viewed as an alternative way of parametrizing the joint exponential family $\mathcal{S}_{\theta;h}$ which is dual to the canonical λ-parametrization. The connection between both is exhibited by the fact that $\hat{h}(\theta,\lambda)$ coincides with the gradient of the normalizing divisor $\psi(\theta,\lambda)$ with respect to λ

$$\begin{aligned} &\nabla_\lambda \psi(\theta,\lambda) \\ &= \nabla_\lambda \log \iint s_\theta(y|z)\, \exp\big(\lambda^T h(y,z)\big)\, \mathrm{d}y\,\mathrm{d}z \\ &= \frac{\iint s_\theta(y|z)\, \nabla_\lambda \exp\big(\lambda^T h(y,z)\big)\, \mathrm{d}y\,\mathrm{d}z}{\iint s_\theta(y|z)\, \exp\big(\lambda^T h(y,z)\big)\, \mathrm{d}y\,\mathrm{d}z} \\ &= \iint \frac{s_\theta(y|z)\, \exp\big(\lambda^T h(y,z)\big)}{\iint s_\theta(y|z)\, \exp\big(\lambda^T h(y,z)\big)\, \mathrm{d}y\,\mathrm{d}z}\, h(y,z)\, \mathrm{d}y\,\mathrm{d}z \\ &= \iint s_{\theta,\lambda}(y,z)\, h(y,z)\, \mathrm{d}y\,\mathrm{d}z \\ &= \hat{h}(\theta,\lambda). \end{aligned}$$

Pythagorean Relationship: Let $s_{\theta,\lambda}(y,z)$ be exponential (19) and $\hat{\lambda}$ satisfy (21). Then we can write

$$\begin{aligned} &K(r_N : s_\theta) \perp K(r_N : s_{\theta,\hat{\lambda}}) \\ &= \iint r_N(y,z)\, \log \frac{s_{\theta,\hat{\lambda}}(y,z)}{s_\theta(y|z)}\, \mathrm{d}y\,\mathrm{d}z \\ &= \hat{\lambda}^T \Big(\iint r_N(y,z)\, h(y,z)\, \mathrm{d}y\,\mathrm{d}z \Big) \perp \psi(\theta,\hat{\lambda}) \\ &= \hat{\lambda}^T \Big(\iint s_{\theta,\hat{\lambda}}(y,z)\, h(y,z)\, \mathrm{d}y\,\mathrm{d}z \Big) \perp \psi(\theta,\hat{\lambda}) \\ &= \iint s_{\theta,\hat{\lambda}}(y,z)\, \log \frac{s_{\theta,\hat{\lambda}}(y,z)}{s_\theta(y|z)}\, \mathrm{d}y\,\mathrm{d}z \\ &= D(s_{\theta,\hat{\lambda}} \| s_\theta) \end{aligned}$$

where we used the notation $D(s_{\theta,\hat{\lambda}} \| s_\theta)$ for the *unnormalized* joint K-L distance. We have obtained in this way an analogue of the Pythagorean relationship (cf. Fig. 1)

$$\boxed{K(r_N : s_\theta) = K(r_N : s_{\theta,\hat{\lambda}}) + D(s_{\theta,\hat{\lambda}} \| s_\theta).} \qquad (25)$$

The Pythagorean-like relationship that links together inaccuracies and K-L distance was shown first in [10]. It can be regarded as generalization of the well-known Pythagorean theorem that holds for K-L distances between probability distributions [11], [12], [13].

Minimum Inaccuracy Projection: Assume that the joint inaccuracy $K(r_N : s_{\theta,\hat{\lambda}})$ of $r_N(y,z)$ relative to the h-projection $s_{\theta,\hat{\lambda}}(y,z)$ is finite

$$K(r_N : s_{\theta,\hat{\lambda}}) < \infty. \qquad (26)$$

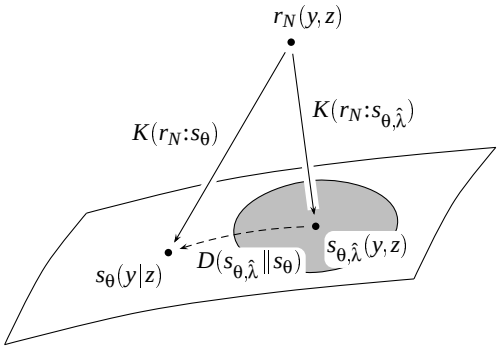

Figure 1: Pythagorean-like decomposition of inaccuracy for dependent observations. The projection "surface" corresponds to the set of all functions $C s_\theta(y|z)\exp(\lambda^T h(y,z))$ with $C>0$. The shaded area indicates a subset of normalized densities with $C=\exp(-\psi(\theta,\lambda))$.

The following Pythagorean relationship holds for every $\lambda\in\mathcal{N}_\theta$

$$\begin{aligned}
&K(r_N{:}s_{\theta,\lambda}) \perp K(r_N{:}s_{\theta,\hat\lambda}) \\
&= \iint r_N(y,z)\log\frac{s_{\theta,\hat\lambda}(y,z)}{s_{\theta,\lambda}(y,z)}\,dy\,dz \\
&= (\hat\lambda\perp\lambda)^T\left(\iint r_N(y,z)\,h(y,z)\,dy\,dz\right) \\
&\quad \perp\psi(\theta,\hat\lambda)+\psi(\theta,\lambda) \\
&= (\hat\lambda\perp\lambda)^T\left(\iint s_{\theta,\hat\lambda}(y,z)\,h(y,z)\,dy\,dz\right) \\
&\quad \perp\psi(\theta,\hat\lambda)+\psi(\theta,\lambda) \\
&= \iint s_{\theta,\hat\lambda}(y,z)\log\frac{s_{\theta,\hat\lambda}(y,z)}{s_{\theta,\lambda}(y,z)}\,dy\,dz \\
&= D(s_{\theta,\hat\lambda}\|s_{\theta,\lambda}).
\end{aligned}$$

Since the joint K-L distance $D(s_{\theta,\hat\lambda}\|s_{\theta,\lambda})$ is nonnegative [8] and the inaccuracy $K(r_N{:}s_{\theta,\hat\lambda})$ was assumed finite in (26), we have

$$K(r_N{:}s_{\theta,\lambda})\geq K(r_N{:}s_{\theta,\hat\lambda})$$

with equality if and only if $s_{\theta,\lambda}(y,z)=s_{\theta,\hat\lambda}(y,z)$ almost everywhere. Thus, under the assumption (26), the h-projection $s_{\theta,\hat\lambda}(y,z)$ is a unique solution to the minimum inaccuracy problem

$$\boxed{K(r_N{:}s_{\theta,\hat\lambda})=\min_{\lambda\in\mathcal{N}_\theta}K(r_N{:}s_{\theta,\lambda}).} \tag{27}$$

Minimum K-L Distance Projection: A dual interpretation of the h-projection is also possible. Assume that the unnormalized K-L distance of the h-projection $s_{\theta,\hat\lambda}(y,z)$ and the conditional model density $s_\theta(y|z)$ is finite

$$D(s_{\theta,\hat\lambda}\|s_\theta)<\infty. \tag{28}$$

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

$$\begin{aligned}
&D(r\|s_\theta)\perp D(r\|s_{\theta,\hat\lambda}) \\
&= \iint r(y,z)\log\frac{s_{\theta,\hat\lambda}(y,z)}{s_\theta(y|z)}\,dy\,dz \\
&= \hat\lambda^T\left(\iint r(y,z)\,h(y,z)\,dy\,dz\right)\perp\psi(\theta,\hat\lambda) \\
&= \hat\lambda^T\left(\iint s_{\theta,\hat\lambda}(y,z)\,h(y,z)\,dy\,dz\right)\perp\psi(\theta,\hat\lambda) \\
&= \iint s_{\theta,\hat\lambda}(y,z)\log\frac{s_{\theta,\hat\lambda}(y,z)}{s_\theta(y|z)}\,dy\,dz \\
&= D(s_{\theta,\hat\lambda}\|s_\theta).
\end{aligned}$$

Since the joint K-L distance $D(r\|s_{\theta,\hat\lambda})$ is nonnegative [8] and we consider only $r(y,z)\in\mathcal{R}_N$ such that $D(r\|s_\theta)<\infty$, we have for every such $r(y,z)$

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

with equality if and only if $r(y,z)=s_{\theta,\hat\lambda}(y,z)$ almost everywhere. Thus, under the assumption (28), the h-projection $s_{\theta,\hat\lambda}(y,z)$ is a unique solution to the minimum K-L distance problem

$$\boxed{D(s_{\theta,\hat\lambda}\|s_\theta)=\min_{r\in\mathcal{R}_N}D(r\|s_\theta).} \tag{29}$$

Minimum K-L Distance: One possible way of calculating the minimum unnormalized K-L distance $D(s_{\theta,\hat\lambda}\|s_\theta)$ is to determine the h-projection $s_{\theta,\hat\lambda}(y,z)$ explicitly and substitute it in $D(s_{\theta,\hat\lambda}\|s_\theta)$. This yields

$$\begin{aligned}
D(s_{\theta,\hat\lambda}\|s_\theta) &= \iint s_{\theta,\hat\lambda}(y,z)\log\frac{s_{\theta,\hat\lambda}(y,z)}{s_\theta(y|z)}\,dy\,dz \\
&= \hat\lambda^T\hat h(\theta,\hat\lambda)\perp\psi(\theta,\hat\lambda) \\
&= \hat\lambda^T\bar h_N\perp\psi(\theta,\hat\lambda).
\end{aligned} \tag{30}$$

Another possibility follows from the identity

$$\begin{aligned}
0 &= \min_\lambda D(s_{\theta,\hat\lambda}\|s_{\theta,\lambda}) \\
&= \min_\lambda\iint s_{\theta,\hat\lambda}(y,z)\log\frac{s_{\theta,\hat\lambda}(y,z)}{s_{\theta,\lambda}(y,z)}\,dy\,dz \\
&= D(s_{\theta,\hat\lambda}\|s_\theta)\perp\max_\lambda\left(\lambda^T\hat h(\theta,\hat\lambda)\perp\psi(\theta,\lambda)\right) \\
&= D(s_{\theta,\hat\lambda}\|s_\theta)\perp\max_\lambda\left(\lambda^T\bar h_N\perp\psi(\theta,\lambda)\right).
\end{aligned}$$

As a result, we have the expression

$$\boxed{D(s_{\theta,\hat\lambda}\|s_\theta)=\max_\lambda\left(\lambda^T\bar h_N\perp\psi(\theta,\lambda)\right).} \tag{31}$$

Thus, the minimum K-L distance follows by maximizing $\lambda^T\bar h_N\perp\psi(\theta,\lambda)$ over λ.

Enveloping Joint Exponential Family: Taking (25) and (31) together, we have

$$K(r_N{:}s_\theta) = K(r_N{:}s_{\theta,\hat\lambda}) + \max_\lambda \big(\lambda^T \bar h_N \perp \psi(\theta,\lambda)\big). \quad (32)$$

The Pythagorean relationship (25) thus enables us to evaluate the inaccuracy $K(r_N{:}s_\theta)$, with precision up to an additive constant, *without* knowledge of $r_N(y,z)$ *provided* $K(r_N{:}s_{\theta,\hat\lambda})$ is independent of θ for every $r_N(y,z)$.

The latter means that for every θ, θ' and every $r_N(y,z)$, it holds

$$K(r_N{:}s_{\theta,\hat\lambda}) = K(r_N{:}s_{\theta',\hat\lambda'})$$

where $\hat\lambda$, $\hat\lambda'$ are such that $\hat h(\theta,\hat\lambda) = \hat h(\theta',\hat\lambda') = \bar h_N$. This condition is to be satisfied for every $r_N(y,z)$, including

$$r_N(y,z) = \delta(y \perp a, z \perp b), \quad a \in \mathcal{Y}, \quad b \in \mathcal{Z}$$

which implies

$$s_{\theta,\hat\lambda}(y,z) = s_{\theta',\hat\lambda'}(y,z).$$

Hence, the h-projections of any $r_N(y,z)$ onto $\mathcal{S}_{\theta;h}$ and $\mathcal{S}_{\theta';h}$ coincide. But this may happen only if the exponential families $\mathcal{S}_{\theta;h}$ and $\mathcal{S}_{\theta';h}$ coincide as a whole. If this is the case for every $\theta \in \mathcal{T}$, the model family $\mathcal{S}$ can be parametrized so that for every θ there exists $\lambda(\theta)$ such that

$$s_\theta(y|z) = C\, s_{\theta_0}(y|z)\, \exp\big(\lambda^T(\theta)\, h(y,z)\big) \quad (33)$$

where $s_{\theta_0}(y|z)$ is a fixed conditional density from the model family $\mathcal{S}$ and C is a constant independent of (y,z).

Thus, the unnormalized joint inaccuracy $K(r_N{:}s_{\theta,\hat\lambda})$ is independent of θ provided the function $s_\theta(y|z)$ for every θ belongs to an unnormalized exponential family of positive functions

$$\Big\{ C\, s_{\theta_0}(y|z)\, \exp\big(\lambda^T h(y,z)\big) : C > 0,\ \lambda \in \mathbb{R}^n \Big\}$$

with a fixed origin $s_{\theta_0}(y|z)$ and the canonical statistic $h(y,z)$. If we choose $h(y,z)$ as a canonical statistic of any unnormalized exponential family *enveloping* the model family $\mathcal{S}$, then (33) is satisfied by definition and $K(r_N{:}s_{\theta,\hat\lambda})$ is independent of θ.

Under the condition (33), it follows from (32) that

$$\boxed{K(r_N{:}s_\theta) = C + \max_\lambda \big(\lambda^T \bar h_N \perp \psi(\theta,\lambda)\big)} \quad (34)$$

where C is a constant independent of θ.

5. Approximate Estimation

The Pythagorean geometry of estimation provides a natural framework for design of suboptimal solutions.

Estimation Problem: Suppose that a certain statistic of (Y,Z) is chosen

$$h{:}\, \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}^n.$$

Let the only information available about the empirical density $r_N(y,z)$ be the empirical expectation of $h(Y,Z)$

$$\bar h_N \triangleq \iint r_N(y,z)\, h(y,z)\, dy\, dz = \frac{1}{N} \sum_{k=m+1}^{N+m} h(y_k,z_k).$$

The empirical density $r_N(y,z)$ is thus only known to lie within the set $\mathcal{R}_N$ defined by (24). The problem is to compute the conditional inaccuracy $\bar K(r_N{:}s_\theta)$ as a function of θ given the above partial information about $r_N(y,z)$.

Approximation of Inaccuracy: Consider the Pythagorean relationship (25)

$$K(r_N{:}s_\theta) = K(r_N{:}s_{\theta,\hat\lambda}) + D(s_{\theta,\hat\lambda}\|s_\theta)$$

and suppose that through a proper choice of $h(y,z)$ we have ensured that

$$K(r_N{:}s_{\theta,\hat\lambda}) \approx C \quad (35)$$

where C is a constant independent of θ.

By (29) and (31), we have

$$\begin{aligned} D(s_{\theta,\hat\lambda}\|s_\theta) &= \min_{r\in\mathcal{R}_N} D(r\|s_\theta) \\ &= \max_\lambda \big(\lambda^T \bar h_N \perp \psi(\theta,\lambda)\big). \end{aligned}$$

Note that $D(r\|s_\theta)$ stands for the unnormalized K-L distance of $r(y,z)$ and $s_\theta(y|z)$.

Introducing the notation

$$D(\mathcal{R}_N\|s_\theta) \triangleq \min_{r\in\mathcal{R}_N} D(r\|s_\theta) \quad (36)$$

and substituting for $\psi(\theta,\lambda)$ from (20), we obtain

$$\boxed{\begin{aligned} D(\mathcal{R}_N\|s_\theta) = \max_\lambda \Big(&\lambda^T \bar h_N \\ &\perp \log \int s_\theta(y|z)\, \exp\big(\lambda^T h(y,z)\big)\Big). \end{aligned}} \quad (37)$$

Hence, under (35) and taking into account that $\bar K(r_N{:}s_\theta) = K(r_N{:}s_\theta)$, we have

$$\boxed{\bar K(r_N{:}s_\theta) \approx C + D(\mathcal{R}_N\|s_\theta).} \quad (38)$$

Posterior Approximation: Substituting from (38) for $\bar K(r_N{:}s_\theta)$ in (14), we obtain the following approximate expression of the likelihood function

$$\boxed{\hat l_N(\theta) = C\, \exp\big(\perp N\, D(\mathcal{R}_N\|s_\theta)\big).} \quad (39)$$

Similarly, substituting from (38) for $\bar K(r_N{:}s_\theta)$ in (15), we obtain the approximate posterior density in the form

$$\boxed{\hat p_N(\theta) \propto p_0(\theta)\, \exp\big(\perp N\, D(\mathcal{R}_N\|s_\theta)\big).} \quad (40)$$

6. Key Properties of Approximation

The approximation (38) is supported by some appealing properties.

Unnormalized Inaccuracy: The (unnormalized) joint K-L distance $D(r\|s_\theta)$ can be decomposed as follows

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

provided all the integrals exist. As a result, we have

$$\boxed{D(r\|s_\theta) = \bar D(r\|s_\theta) \perp H(\tilde r).} \tag{41}$$

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's 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 $\mathcal{R}_N$.

Upper and Lower Bounds: Taking together the identity (41), the definition (36) of the minimum K-L distance $D(\mathcal{R}_N\|s_\theta)$ and the nonnegativity of the conditional K-L distance $\bar D(r\|s_\theta)$ [8], we get the following bounds on $D(\mathcal{R}_N\|s_\theta)$

$$\boxed{\perp \max_{r\in\mathcal{R}_N} H(\tilde r) \le D(\mathcal{R}_N\|s_\theta) \le D(r\|s_\theta)} \tag{42}$$

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

Monotonicity: It follows directly from the definition (36) 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

$$\boxed{\mathcal{R}_N \subseteq \mathcal{R}'_N \quad \text{implies} \quad D(\mathcal{R}_N\|s_\theta) \ge D(\mathcal{R}'_N\|s_\theta).} \tag{43}$$

7. Choice of Statistic

Sufficient Statistic: Consider a family of distributions $\mathcal{S} = \{s_\theta(y|z) : \theta\in\mathcal{T}\}$. A statistic

$$T_N\colon \mathcal{Y}^{N+m}\times\mathcal{U}^{N+m}\to\mathbb{R}^n$$

is called *sufficient* with respect to $\mathcal{S}$ if the sample Y^{N+m}, U^{N+m} and the parameter Θ (taken as a random variable) are conditionally independent given the value of the statistic $T_N(Y_{m+1}^{N+m},U_{m+1}^{N+m}) = T_N(y_{m+1}^{N+m},u_{m+1}^{N+m})$ and the initial values $Y^m = y^m$, $U^m = u^m$

$$Y_{m+1}^{N+m},U_{m+1}^{N+m} \perp \Theta \,|\, T_N(Y^{N+m},U^{N+m}),Y^m,U^m \tag{44}$$

for every prior density $p(\theta|y^m,u^m)$.

Minimal Sufficient Statistic: A statistic $T_N(Y^{N+m},U^{N+m})$ is called *minimal sufficient* if it is a function of any other sufficient statistic $T'_N(Y^{N+m},U^{N+m})$.

Sufficient statistics for practically interesting models have often very large or even infinite dimension. To make estimation for such models feasible, we have to use a statistic of limited dimension—not sufficient for restoration of the true likelihood. The choice of the statistic seriously affects the resulting discrepancy between the true and approximate likelihoods. In the following we present a class of statistics which are the next to try if the sufficient statistics cannot be used.

Necessary Statistic: Consider a family of distributions $\mathcal{S} = \{s_\theta(y|z) : \theta\in\mathcal{T}\}$. A statistic

$$T_N\colon \mathcal{Y}^{N+m}\times\mathcal{U}^{N+m}\to\mathbb{R}^n$$

is said to be *necessary* for $\mathcal{S}$ if $T_N(Y^{N+m},U^{N+m})$ is a function of any sufficient statistic $T_N^*(Y^{N+m},U^{N+m})$. Thus, the necessary statistic is a function of a minimal sufficient statistic.

Under weak regularity assumptions [4], the necessary statistic can be constructed as follows. Consider the linear space $\mathcal{H}$ spanned by constants and the functions

$$\log s_\theta(y|z) \perp \log s_{\theta_0}(y|z)$$

for all $\theta\in\mathcal{T}$ where θ_0 is an arbitrary fixed point in $\mathcal{T}$. Pick up n linearly independent, non-constant functions $h_1(y,z),\dots,h_n(y,z)$ from $\mathcal{H}$. The functions

$$h_0(y,z)\equiv 1, h_1(y,z),\dots,h_n(y,z)$$

span an $(n+1)$-dimensional linear subspace $\mathcal{H}_0$ of $\mathcal{H}$. Given the vector statistic of single observation (y,z)

$$h(y,z) = [h_1(y,z),\dots,h_n(y,z)]^T,$$

we define a vector statistic of the whole sample (y^{N+m},u^{N+m}) as the empirical expectation or sample average

$$T_N(y^{N+m},u^{N+m}) \triangleq E_N\big(h(Y,Z)\big) = \frac{1}{N}\sum_{k=m+1}^{N+m} h(y_k,z_k). \tag{45}$$

As the empirical expectation of any set of basis vectors of the linear space $\mathcal{H}$ is a minimal sufficient statistic, a statistic defined through the empirical expectation of basis vectors of a linear subspace $\mathcal{H}_0$ of $\mathcal{H}$ is clearly necessary.

Construction of Single-Data Statistic: There are many possible constructions of the vector statistic $h(y,z)$ with the above property. The following are perhaps the most typical ones.

Differencing: Pick up $n+1$ points $\theta_1^*, \ldots, \theta_{n+1}^*$ in the parameter space $\mathcal{T}$ and set

$$\boxed{h_i(y,z) = \log s_{\theta_{i+1}^*}(y|z) \perp \log s_{\theta_i^*}(y|z).} \qquad (46)$$

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

$$\boxed{h_i(y,z) = \omega_i^{*T} \nabla_\theta \log s_{\theta_i^*}(y|z).} \qquad (47)$$

Weighted Integration: Pick up n weighting functions $w_1^*(\theta), \ldots, w_n^*(\theta)$ such that

$$\int w_i^*(\theta)\, \mathrm{d}\theta = 0, \quad i = 1, \ldots, n$$

and set

$$\boxed{h_i(y,z) = \int w_i^*(\theta) \log s_\theta(y|z)\, \mathrm{d}\theta.} \qquad (48)$$

General Construction of Single-Data Statistic: Consider a vector space $\mathcal{V}$ that contains functions

$$v(\theta) = \log s_\theta(y|z)$$

for all $(y,z) \in \mathcal{Y} \times \mathcal{Z}$. Let L_i, $i = 1, \ldots, n$ be a set of linear functionals defined on the vector space $\mathcal{V}$. Suppose in addition that the linear functionals are normalized so that

$$L_i(1) = 0$$

for $i = 1, \ldots, n$. Then define

$$\boxed{h_i(y,z) = L_i\big(\log s_\theta(y|z)\big)} \qquad (49)$$

for $i = 1, \ldots, n$.

When the functions $h_0(y,z) \equiv 1$, $h_1(y,z)$, $\ldots$, $h_n(y,z)$ are linearly independent, the vector function $h(y,z) = [h_1(y,z), \ldots, h_n(y,z)]^T$ forms a single-data vector statistic that defines through (45) an n-dimensional statistic *necessary* for the family $\mathcal{S}$.

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

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

the empirical expectation of the single-data statistics (46), (47), (48) yields

$$E_N\left(\log \frac{s_{\theta_{i+1}^*}(Y|Z)}{s_{\theta_i^*}(Y|Z)}\right) = \frac{1}{N} \log \frac{l_N(\theta_{i+1}^*)}{l_N(\theta_i^*)},$$

$$E_N\Big(\omega_i^{*T} \nabla_\theta \log s_{\theta_i^*}(Y|Z)\Big) = \frac{1}{N} \omega_i^{*T} \nabla_\theta \log l_N(\theta_i^*),$$

$$E_N\Big(\int w_i^*(\theta) \log s_\theta(Y|Z)\, \mathrm{d}\theta\Big) = \frac{1}{N} \int w_i^*(\theta) \log l_N(\theta)\, \mathrm{d}\theta,$$

respectively.

In general, we have by (49)

$$E_N\Big(L_i\big(\log s_\theta(Y|Z)\big)\Big) = \frac{1}{N} L_i\Big(\log l_N(\theta)\Big).$$

The necessary statistic thus carries, in a condensed form, partial information about the "shape" of the log-likelihood $\log l_N(\theta)$.

Likelihood Matching: Let us apply the linear functionals $L_i(\cdot)$, $i = 1, \ldots, n$ to both sides of the Pythagorean relationship (25)

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

Substituting from (14) and (39)

$$L_i\big(K(r_N{:}s_\theta)\big) = L_i\Big(\perp \frac{1}{N} \log l_N(\theta)\Big),$$

$$L_i\big(D(s_{\theta,\hat{\lambda}} \| s_\theta)\big) = L_i\Big(\perp \frac{1}{N} \log \hat{l}_N(\theta)\Big),$$

we obtain

$$L_i\Big(\perp \frac{1}{N} \log l_N(\theta)\Big) = L_i\big(K(r_N{:}s_{\theta,\hat{\lambda}})\big) + L_i\Big(\perp \frac{1}{N} \log \hat{l}_N(\theta)\Big).$$

Owing to the definition (49) of h_i, it holds $L_i\big(K(r_N{:}s_{\theta,\hat{\lambda}})\big) = 0$ and

$$\boxed{L_i\big(\log l_N(\theta)\big) = L_i\big(\log \hat{l}_N(\theta)\big)} \qquad (50)$$

for $i = 1, \ldots, n$. Thus, the use of a necessary statistic ensures that the approximate log-likelihood $\log \hat{l}_N(\theta)$ matches partially, in the sense defined by the functionals $L_i(\cdot)$, the true log-likelihood $\log l_N(\theta)$.

8. Illustrative Example

Model: We considered a sequence of observations $Y_1, \ldots, Y_{N+1}$ modelled as

$$Y_k = \theta Y_{k-1} + E_k$$

where $E_1, \ldots, E_{N+1}$ was a sequence of independent, Cauchy-distributed random variables with a common density

$$n(e) = \frac{1}{\pi} \frac{1}{1+e^2}.$$

The theoretical density of Y_k given $Z_k = Y_{k-1}$ was thus

$$s_\theta(y|z) = \frac{1}{\pi} \frac{1}{1 + (y \perp \theta z)^2}.$$

Simulated Data: We simulated a sequence of 101 data $(y_1, \ldots, y_{101})$ shown in Fig. 2. The corresponding joint empirical density $r_N(y,z)$ of Y_k and $Z_k = Y_{k-1}$ is envisaged in Fig. 3 through a scatterplot of (Y,Z).

The problem was to estimate the regression coefficient $\theta = 0.98$ given the observed sample.

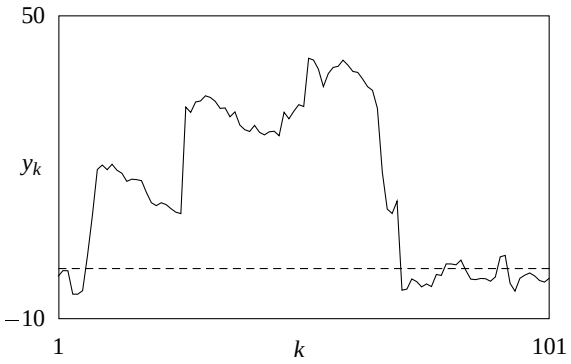

Figure 2: A sequence of 101 samples of $Y_k = 0.98Y_{k-1} + E_k$ with Cauchy-distributed noise $E_k \sim C(0,1)$.

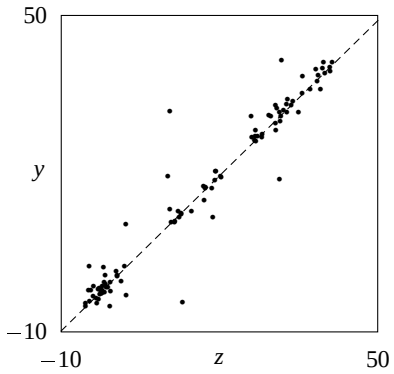

Figure 3: The scatterplot of data shown in Fig. 2 gives a good idea about the joint empirical distribution of (Y,Z) with $Z_k = Y_{k-1}$.

Choice of Statistic: We used a vector statistic $h(y)$ of dimension $n = 5$ composed of score functions, i.e., the first-order derivatives of the log-density $\log s_\theta(y|z)$ with respect to θ

$$h_i(y,z) = \frac{2\,(y \perp \theta_i^*)}{1 + (y \perp \theta_i^*)^2} \tag{51}$$

at the points $\theta_i^* = \perp 1, \perp 0.5, 0, +0.5, +1$.

Approximate Estimation: Given the statistic $h(y,z)$ and the observed data $(y_1,\ldots,y_{101})$, we computed the sample average

$$\bar{h}_N = \frac{1}{N}\sum_{k=m+1}^{N+m} h(y_k, z_k)$$

for $N = 100$ and $m = 1$.

Given the value $\bar{h}_N$, we solved the optimization problem (37) for a set of different values of the unknown parameter θ, namely 121 values evenly spaced within the interval $[\perp 2, 2]$. The resulting K-L distance $D(\mathcal{R}_N \| s_\theta)$ is shown as a solid line in the upper plot in Fig. 4.

To illustrate that K-L distance $D(r\|s_\theta)$ for every $r(y,z) \in \mathcal{R}_N$ is bounded from below by $D(\mathcal{R}_N\|s_\theta)$, we calculated $D(r_i\|s_\theta)$ for six such densities $r_i(y,z)$, constructed as minimum K-L distance projections of the theoretical densities $s_{\theta_i^*}(y|z)$ for $\theta_i^* = \perp 1.5, \perp 0.9, \perp 0.3, +0.3, +0.9, +1.5$ onto $\mathcal{R}_N$.

The lower plot in Fig. 4 compares the normalized likelihoods

$$\begin{aligned} \hat{l}_N(\theta) &= \exp\big(\perp N D(\mathcal{R}_N\|s_\theta)\big), \\ l_N(\theta; r_i) &= \exp\big(\perp N D(r_i\|s_\theta)\big) \end{aligned}$$

for $N = 100$.

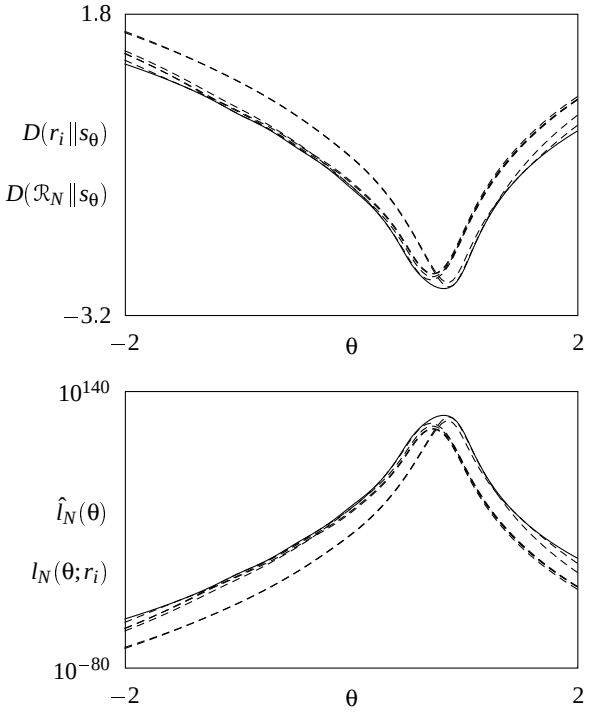

Figure 4: Linear autoregression with Cauchyian noise: results of approximate estimation for $N = 100$. The upper plot compares K-L distances, the lower plot shows the normalized likelihoods.

9. Numerical Implementation

The use of approximation (38) is accompanied with massive drop in computational complexity. First, the approximations use a statistic of finite, limited dimension. Second, the dimension of the optimization problem invoked is given by the dimension of data entering model at one time instant only. Compare it with the ideal solution which in general requires all data to be stored and processed. Yet, to solve the optimization problem (37) for all or a sufficient number of values of the unknown parameter, we may still need a lot of computing power. The following lines give some recommendations as to efficient numerical implementation of the estimation algorithm.

Convex Minimization Problem: Since most optimization packages provide algorithms for finding minimum of a given function, we rewrite (37) explicitly as minimization problem

$$\min_{\lambda} J(\theta,\lambda)$$

where the optimized function is

$$J(\theta,\lambda) = \perp\lambda^T\bar{h}^N + \log\iint s_\theta(y|z)\,\exp\bigl(\lambda^T h(y,z)\bigr)\,\mathrm{d}y\,\mathrm{d}z.$$

The problem we solve is to find for selected values of $\theta \in \mathcal{T}$ the minimum of the above convex function over a convex set (remember the set $\mathcal{N}_\theta$ of all values of λ for which logarithm of the normalizing divisor (20) takes on a finite value is convex).

Gradient and Hessian: The application of gradient and Newton methods is facilitated by the conceptually easy computation of the gradient and Hessian of $J(\theta,\lambda)$ with respect to λ.

In particular, the gradient of the function $J(\theta,\lambda)$ with respect to λ is

$$\nabla_\lambda J(\theta,\lambda) = \perp\bar{h}_N + \hat{h}(\theta,\lambda),$$

i.e., equal to the difference between the theoretical and empirical means of $h(Y,Z)$

$$\boxed{\nabla_\lambda J(\theta,\lambda) = E_{\theta,\lambda}\bigl(h(Y,Z)\bigr) \perp E_N\bigl(h(Y,Z)\bigr).} \qquad (52)$$

The Hessian of $J(\theta,\lambda)$ with respect to λ is

$$\begin{aligned}&\nabla^2_\lambda J(\theta,\lambda)\\ &= \nabla_\lambda \hat{h}^T(\theta,\lambda)\\ &= E_{\theta,\lambda}\Bigl(\bigl(h(Y,Z)\perp\hat{h}(\theta,\lambda)\bigr)\,\bigl(h(y,z)\perp\hat{h}(\theta,\lambda)\bigr)^T\Bigr),\end{aligned}$$

i.e., equal to the theoretical covariance of $h(Y,Z)$

$$\boxed{\nabla^2_\lambda J(\theta,\lambda) = \mathrm{Cov}_{\theta,\lambda}\bigl(h(Y,Z)\bigr).} \qquad (53)$$

Numerical Integration: Integration involved in calculation of the function $\psi(\theta,\lambda)$ and possibly the mean $E_{\theta,\lambda}\bigl(h(Y,Z)\bigr)$ and covariance $\mathrm{Cov}_{\theta,\lambda}\bigl(h(Y,Z)\bigr)$ is taken over the space of all possible values of (Y,Z). Since the dimension of Z for practically interesting problems is usually beyond the margin when quadrature formulae can be used efficiently, Monte Carlo techniques can generally be recommended.

The Monte Carlo computation of the normalizing divisor of the function $f_{\theta,\lambda}(y,z) = s_\theta(y|z)\,\exp\bigl(\lambda^T h(y,z)\bigr)$ is based upon the approximation

$$\begin{aligned}\iint f_{\theta,\lambda}(y)\,\mathrm{d}y\,\mathrm{d}z &= \iint s(y,z)\,\frac{f_{\theta,\lambda}(y,z)}{s(y,z)}\,\mathrm{d}y\,\mathrm{d}z\\ &\approx \frac{1}{M}\sum_{k=1}^{M}\frac{f_{\theta,\lambda}(y_k,z_k)}{s(y_k,z_k)}\end{aligned}$$

where $s(y,z) > 0$ is a density function and (y_1,z_1), …, (y_M,z_M) are independent samples drawn from $s(y,z)$. The density function $s(y,z)$ should be chosen close enough to $s_{\theta,\lambda}(y,z)$. In a similar way, we can approximate the moments $E_{\theta,\lambda}\bigl(h(Y,Z)\bigr)$ and $\mathrm{Cov}_{\theta,\lambda}\bigl(h(Y,Z)\bigr)$.

For more information about the advanced methods of multivariate integration in Bayesian statistics, see [14].

Iterative Optimization: When the dimension of the statistic $h(y,z)$ is too large, it may be more efficient to organize the calculation of the minimum K-L distance $D(\mathcal{R}_N\|s_\theta)$ so that we optimize *one entry* λ_i of the vector λ *at a time*

$$\min_{\lambda_i}\Bigl(\perp\lambda^T\bar{h}^N + \log\iint s_\theta(y|z)\,\exp\bigl(\lambda^T h(y,z)\bigr)\,\mathrm{d}y\,\mathrm{d}z\Bigr)$$

while the other entries λ_j, $j \neq i$ are fixed at their last values. The optimization is done for $i = 1,\ldots,n$ and then the whole loop is repeated—until the minimum is found with a prescribed precision. Owing to the convexity of K-L distance and linearity of the constraints, the whole procedure ultimately converges to the true solution [12].

10. Concluding Remarks

One of the noteworthy features of the above approximation scheme is that compression of data and restoration of inaccuracy (likelihood, posterior) are quite separate steps linked only through the value of $(N,\bar{h}_N)$. This contrasts with 'local' methods of point estimation (such as approximate maximum-likelihood estimation using Newton-Raphson procedure) which permanently optimize the statistic definition. The 'global' character of approximate estimation proposed in the paper makes it possible to solve difficult problems of practical interest.

Just-in-Time Modelling: The idea of just-in-time models is to make use of archived process data for estimation of a local model valid just for the current working point. In [15] a neigbourhood of the current regressor z_N is used to define the "working point". It sounds natural to generalize this approach by taking the statistic value $\bar{h}_N$ as a "wider-angle snapshot" of process data. The methodology described in the paper provides immediately a recipe for on-demand restoration of the posterior density from such a piece of information.

Building of Semiphysical Models: Advanced control starts from advanced models that incorporate enough physical (chemical, biological) insight into the process behaviour. This raises the problem of estimation of "semiphysical" models [16], i.e., models built at least partially from the first principles. Such models are typically non-linear and possibly non-Gaussian as well. The methodology of approximate Bayesian estimation provides a sophisticated tool for exploration of the "error function" represented by Kullback-Leibler distance

between the empirical and model distributions of data. This allows the user to organize the modelling process interactively—trying various model definitions and checking immediately the corresponding "error function" value.

Nearest-neighbour Identification: Recursive computation of the whole posterior density continues to be a challenge whose solution will probably require an explicit approximation of the model itself [17]. One can imagine simpler schemes, however, that are well in reach of the current computer technology. An example is iterative identification that builds a finite number of models in the "neighbourhood" of the current model and computes the corresponding Kullback-Leibler distances (likelihoods, posterior probabilities) for the set of models considered. The whole process can be repeated recursively so that a new set of models is built around the best one from the previous step. Conceptually, computations for particular models can be performed in parallel on separate processors. Note that the data statistic has to be "rich enough" to bring enough information for all potentially considered models.

Acknowledgments

The work was supported in part by the EC grant no. CIPA-CT94-0237 and the of Czech Academy of Sciences grant no. A2075603.

References

[1] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification, and Adaptive Control*. Englewood Cliffs, N.J.: Prentice-Hall, 1986.

[2] H. W. Sorenson, "On the development of practical nonlinear filters," *Inform. Sci.*, vol. 7, pp. 253–270, 1974.

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

[4] R. Kulhavý, *Recursive Nonlinear Estimation: A Geometric Approach*, vol. 216 of *Lecture Notes in Control and Information Sciences*. London: Springer-Verlag, 1996.

[5] G. R. Grimmett and D. R. Stirzaker, *Probability and Random Processes*. Oxford: Oxford University Press, second ed., 1992.

[6] V. Peterka, "Bayesian approach to system identification," in *Trends and Progress in System Identification* (P. Eykhoff, ed.), ch. 8, pp. 239–304, Elmsford, N.Y.: Pergamon, 1981.

[7] D. F. Kerridge, "Inaccuracy and inference," *J. Roy. Statist. Soc. Ser. B*, vol. 23, pp. 284–294, 1961.

[8] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, pp. 79–86, 1951.

[9] L. D. Brown, *Fundamentals of Statistical Exponential Families (with Applications in Statistical Decision Theory)*, vol. 9 of *Lecture Notes — Monograph Series*. Hayward, CA: Inst. Math. Statist., 1987.

[10] R. Kulhavý, "A geometric approach to statistical estimation," in *Proceedings of the 34th IEEE Conference on Decision and Control*, vol. 2, (New Orleans, LA), pp. 1097–1102, 1995.

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

[12] I. Csiszár, "*I*-divergence geometry of probability distributions and minimization problems," *Ann. Probab.*, vol. 3, no. 1, pp. 146–158, 1975.

[13] S. Amari, *Differential-Geometrical Methods in Statistics*, vol. 28 of *Lecture Notes in Statistics*. Berlin: Springer-Verlag, 1985.

[14] A. F. M. Smith, "Bayesian computational methods," *Phil. Trans. R. Soc. Lond. Ser. A*, vol. 337, pp. 369–386, 1991.

[15] A. Stenman, F. Gustafsson, and L. Ljung, "Just in time models for dynamical systems," in *IEEE Control and Decision Conference*, (Kobe, Japan), 1996.

[16] L. Ljung, *System Identification: Theory for the User*. Englewood Cliffs, N.J.: Prentice-Hall, 1987.

[17] R. Kulhavý, "Recursive nonlinear estimation through global approximation of model," in *Proceedings of the 3rd European Control Conference*, vol. 2, (Rome, Italy), pp. 1273–1278, 1995.