ON DUAL EXPRESSION OF PRIOR INFORMATION IN BAYESIAN PARAMETER ESTIMATION*

R. Kulhavý [1,2] **and L. Tesař** [2]

[1] *Honeywell Technology Center—Prague, 182 08 Prague, Czech Republic*
[2] *Institute of Information Theory and Automation, Acad. Sci. CR, 182 08 Prague, Czech Rep.*

Abstract: In Bayesian parameter estimation, *a priori* information can be used to shape the prior density of unknown parameters of the model. When chosen in a conjugate, self-reproducing form, the prior density of *parameters* is nothing but a model-based transform of a certain "prior" density of observed *data*. This observation suggests two possible ways of expressing *a priori* knowledge—in terms of parameters of a particular model and in terms of data entering the model. The latter way turns out useful when dealing with statistical models whose parameters lack a direct physical interpretation. In practice, the amount of *a priori* information is usually not sufficient for complete specification of the prior density of data. The paper shows an information-based way of converting such incomplete information into the prior density of unknown parameters.

Keywords: Parameter estimation, Bayesian methods, prior information, information measures, approximation.

1. INTRODUCTION

The role of *a priori* information in parameter estimation has become widely acknowledged in the control community. It is well known that information about the system dynamics may significantly affect parameter estimates. Even rudimentary knowledge of the identified system can effectively regularize parameter estimation. This is crucial namely in poorly conditioned cases when data carry little information.

The use of prior information depends on the estimation methodology used. The Bayesian paradigm provides a framework which is particularly well suited for merging of prior and data-based information. *A priori* information is used here to shape the prior density of unknown parameters while Bayes's rule serves as a tool for updating prior knowledge by information contained in observed data.

Indeed, when speaking about *a priori* information, one usually has in mind the choice of the prior density, in other words, the choice of the "initial conditions" for the parameter estimator (Peterka, 1981). More generally, when parameters vary in time, prior information may differ at every time instant. When such information is available, it can be used to shape reference densities in the generalized forgetting schemes (Kulhavý and Kraus, 1996). Finally, there are situations when no data are available and decision is made solely on the basis of prior information. A careful specification of the prior density is of vital importance then.

Provided the model is structured so that its parameters admit a direct physical interpretation, the choice of a particular prior density is a rather intuitive way of expressing available knowledge. The examples of such parameters include rates and orders of chemical reactions, growth rates of biomass or characteristics of valves.

The notorious difficulty with physically parametrized models is their nonlinearity which makes estimation of parameters a non-trivial task. It is why in on-line identification one meets statistically-rooted models more often than first-principle models. A typical example of statistical models is the autoregressive model with external input (ARX). While the parameters of the ARX model are relatively easy to estimate, they usually lack a simple physical interpretation. The regression coefficients describe nothing but a linearized dependence of the system output on the regressor entries (which may be nonlinear functions of past data). The variance

*The work was supported in part by the grants 102/97/0466 and 402/96/0902 of the Grant Agency of the Czech Republic and grant A2075603 of the Academy of Sciences of the Czech Republic.

of the stochastic component is an integral expression of the stochastic behaviour of the system, sensor imprecisions and approximation (mismodelling) errors. As a result, one can rarely specify the prior density of parameters of statistical models on a purely intuitive basis.

It may be easier in such cases to express prior knowledge directly in terms of data observed on the system rather than parameters of a specific model. This alternative approach follows from the observation that the prior density of parameters, when chosen in a self-reproducing form, is a model-based transform of a certain "prior" density of data entering the model. Section 3 shows that this connection holds for any regression-type models. The difference between the two—basically dual—approaches is illustrated on the ARX model.

The amount of information available on-line is typically insufficient to specify the "prior" density of data completely. The major result of Section 4 is a method of building the prior density of model parameters from incomplete information about observed data. The solution suggested here is based upon a natural "information geometry" of probability distributions of observed data (Kulhavý, 1996).

2. BAYESIAN ESTIMATION REVISITED

The basic problem of system identification is to fit a proper model to a dynamic, possibly controlled system. The purpose of this section is to resume the classical view of Bayesian estimation of parameter estimation and to compare it with an alternative viewpoint based upon measuring an information "distance" between the empirical and model-based densities of observed data.

2.1 Probabilistic Model of Dynamic System

The models used in statistical methods of system identification describe—in probabilistic terms—how the system output depends on its past values and possible external inputs.

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

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

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 its 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

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

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

General Regression Model. We shall suppose that the output values Y_k depend on the past data Y_{k-m}^{k-1}, U_{k-m}^{k} only through a known vector function $Z_k = z(U^k, Y^{k-1})$ taking values in a subset $\mathcal{Z}$ of $\mathbb{R}^{\dim z}$. Hence, we assume that

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

for $k = m+1, \ldots, N+m$. Moreover, we suppose that the conditional density of Y_k given $Z_k = z_k$ is identical for all k, i.e., $s_k(y|z) = s(y|z)$. Finally, we assume 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. The density $s(y|z)$ is considered to be a member of a given family

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

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. In general, the dependence of the input U_k on the past data Y^{k-1}, U^{k-1} and the parameter θ is expressed through a conditional density $\gamma_k(u_k|y^{k-1}, u^{k-1}, \theta)$. In most cases of practical interest, we may adopt a simplifying assumption (Peterka, 1981) that the only information about θ used in computation of the input is the information contained in the past data. More precisely, we assume that the equality

$$\gamma_k(u_k|y^{k-1}, u^{k-1}, \theta) = \gamma_k(u_k|y^{k-1}, u^{k-1}) \tag{2}$$

holds at $k = m+1, \ldots, N+m$.

2.2 Bayesian Estimation: The Classical View

Joint Density of Sample. 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

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

Taking into account the model assumption (1) and the natural conditions of control (2), we can write

$$q_\theta^N(y_{m+1}^{N+m}, u_{m+1}^{N+m}|y^m, u^m) = \prod_{k=m+1}^{N+m} s_\theta(y_k|z_k)\, \gamma_k(u_k|y^{k-1}, u^{k-1}). \tag{3}$$

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

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

The subscript N indicates conditioning on N data points (y_{m+1}, z_{m+1}), ..., (y_{N+m}, z_{N+m}).

Given a prior density conditional on *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. When substituting for the joint density $q_\theta^N(y_{m+1}^{N+m}, u_{m+1}^{N+m}|y^m, u^m)$ from (3) and taking the natural conditions of control (2) for granted, we obtain

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

where the symbol $\propto$ stands for equality up to a normalizing factor. As a result, we have the formula

$$\boxed{p_N(\theta) \propto p_0(\theta) \prod_{k=m+1}^{N+m} s_\theta(y_k|z_k).} \qquad (4)$$

2.3 Bayesian Estimation: A View via Inaccuracy

Introducing the notion of conditional inaccuracy, we can reformulate probability-based estimation as an explicit approximation problem.

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

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

where $\delta(\cdot)$ is a Dirac function satisfying $\delta(y,z) = 0$ for $y \neq 0$ or $z \neq 0$ and $\iint_{\mathcal{Y}\times\mathcal{Z}} \delta(y,z)\,\mathrm{d}y\,\mathrm{d}z = 1$. The subscript N is used again to indicate the number of data points (y_{m+1}, z_{m+1}), ..., (y_{N+m}, z_{N+m}) the empirical density is based on.

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

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

The concept of conditional inaccuracy is generalization of Kerridge's inaccuracy (Kerridge, 1961) introduced originally for the case of independent and identically distributed data.

Joint Density of Sample. The joint density of sample (3) 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 Γ_{N+m} is a factor independent of θ. Further, using conditional inaccuracy, we can rewrite the θ-dependent part as follows

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

Putting both together, we obtain the following expression

$$q_\theta^N(y_{m+1}^{N+m}, u_{m+1}^{N+m}|y^m, u^m) = \Gamma_{N+m} \exp\left(-N \bar{K}(r_N{:}s_\theta) \right).$$

Posterior Density. Applying Bayes's theorem and substituting for the joint density of sample from the last formula, we get the *posterior* density of Θ conditional on the observed sample y^{N+m}, u^{N+m} in the form

$$\boxed{p_N(\theta) \propto p_0(\theta) \exp\left(-N \bar{K}(r_N{:}s_\theta) \right)} \qquad (5)$$

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

3. CONJUGATE PRIORS

In practical estimation, it is convenient if the prior density $p_0(\theta)$ is chosen from a *conjugate* family—closed under conditioning on the observed data (Robert, 1989).

General Form of Conjugate Prior. The expression (5) suggests the general form of self-reproducing priors

$$\boxed{p_\nu(\theta) \propto \exp\left(-\nu \bar{K}(r_\nu{:}s_\theta) \right).} \qquad (6)$$

Here $r_\nu(y,z)$ stands for a joint "prior" density of (Y,Z) which is based upon prior information and possibly m initial values y^m and u^m. The scalar ν describes the degree of belief in r_ν; the bigger ν, the more concentrated the prior density $p_\nu(\theta)$ is. The factor ν is supposed to be nonnegative but not necessarily integer.

By formal analogy with (5), ν can be regarded as the number of actual or fictitious observations $r_\nu(y,z)$ is built on. The density (6) then can be seen as a "posterior" density given a uniform prior density $p_0(\theta) \propto 1$

and ν data with the empirical density $r_\nu(y,z)$ (cf. the use of fictitious data in Kárný *et al.*, 1985).

Hence, instead of expressing prior knowledge in terms of probability distribution of parameters of a specific model, we suggest to express it directly through distribution of observed data. The latter way enables us to keep a single description of prior information even when considering different model families. The formula (6) gives a clue how the prior density of (Y,Z) is to be converted, given a particular model $s_\theta(y|z)$, into the prior density of Θ.

Posterior Density. It is easy to verify that given the conjugate prior (6), the posterior density (5) preserves its form

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

Indeed, the posterior density can be written as

$$\boxed{p_{\nu+N}(\theta) \propto \exp\bigl(-(\nu+N)\bar{K}(r_{\nu+N}{:}s_\theta)\bigr)} \qquad (7)$$

where

$$r_{\nu+N}(y,z) = \frac{\nu}{\nu+N}\,r_\nu(y,z) + \frac{N}{\nu+N}\,r_N(y,z) \qquad (8)$$

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 $N\to\infty$.

Exponential Family. An important special case occurs when the model family $\{s_\theta(y|z)\}$ is imbedded in an exponential family, i.e., when for every θ the density $s_\theta(y|z)$ can be expressed as

$$s_\theta(y|z) = s_0(y|z)\exp\Bigl(\lambda^T(\theta)\,h(y,z) - \psi\bigl(\lambda(\theta)\bigr)\Bigr) \qquad (9)$$

where $s_0(y|z)$ is a fixed "origin", $h{:}\,\mathcal{Y}\times\mathcal{Z}\to\mathbb{R}^n$ defines a canonical statistic and $\lambda\in\mathbb{R}^n$ is a natural parameter of the enveloping exponential family.

Note that the factor $\psi\bigl(\lambda(\theta)\bigr)$ does *not* depend on z. This is a crucial assumption which essentially says that the logarithm of $s_\theta(y|z)$ can always be factorized so that the data (y,z) and the parameter θ are separated. To stress that the above definition is stronger than the usual definition of the exponential family of conditional densities, we shall call the enveloping exponential family *jointly exponential* (in both y and z).

Given the model (9), the prior density (6) takes the form

$$p_\nu(\theta) \propto \exp\bigl(-\nu\lambda^T(\theta)\,\bar{h}_\nu\bigr)$$

where

$$\bar{h}_\nu = \mathcal{E}_\nu\bigl(h(Y,Z)\bigr) = \iint r_\nu(y,z)\,h(y,z)\,\mathrm{d}y\,\mathrm{d}z$$

denotes the expectation of the canonical statistic $h(Y,Z)$ of the enveloping exponential family with respect to the prior density $r_\nu(y,z)$.

The posterior density $p_{\nu+N}(\theta)$ preserves the form of $p_\nu(\theta)$

$$p_{\nu+N}(\theta) \propto \exp\bigl(-(\nu+N)\,\lambda^T(\theta)\,\bar{h}_{\nu+N}\bigr)$$

where

$$\bar{h}_{\nu+N} = \mathcal{E}_{\nu+N}\bigl(h(Y,Z)\bigr) = \iint r_{\nu+N}(y,z)\,h(y,z)\,\mathrm{d}y\,\mathrm{d}z$$

denotes the expectation of $h(Y,Z)$ with respect to the prior-modified empirical density $r_{\nu+N}(y,z)$ defined by (8).

ARX Model. Consider the model

$$Y_k = \theta^T Z_k + E_k, \quad E_k \sim N(0,\sigma^2)$$

where θ and Z_k are column vectors and $E_1, E_2, \ldots$ are independent, normally distributed random variables with zero mean and known variance σ^2. The conditional density function of Y_k given $Z_k = z_k$ thus takes the form

$$s_\theta(y|z) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\Bigl(-\frac{1}{2\sigma^2}(y-\theta^T z)^2\Bigr).$$

After straightforward arrangements we find that the prior density (6) can be written as

$$p_\nu(\theta) \propto \exp\Bigl(-\frac{1}{2\sigma^2}\,\nu\,(\theta-\hat{\theta}_\nu)^T P_\nu^{-1}(\theta-\hat{\theta}_\nu)\Bigr) \qquad (10)$$

with

$$\begin{aligned} \hat{\theta}_\nu &= \mathcal{E}_\nu\bigl(ZZ^T\bigr)^{-1}\mathcal{E}_\nu(ZY), \\ P_\nu^{-1} &= \mathcal{E}_\nu(ZZ^T) \end{aligned}$$

where $\mathcal{E}_\nu(\cdot)$ stands for the expectation with respect to $r_\nu(y,z)$. The matrix P_ν^{-1} is supposed positive definite.

The expression (10) indicates that we can provide prior information about the ARX model parameters in two different ways—either via the expectations of model-specific functions of data (Y,Z)

$$\mathcal{E}_\nu(ZY), \quad \mathcal{E}_\nu(ZZ^T) \qquad (11)$$

or via the mean and covariance of the parameter vector θ

$$\mathcal{E}_\nu(\theta) = \hat{\theta}_\nu, \quad \mathcal{E}_\nu(\theta-\hat{\theta}_\nu)(\theta-\hat{\theta}_\nu)^T = \frac{\sigma^2 P_\nu}{\nu}. \qquad (12)$$

Note that the subscript ν in (12) indicates expectation with respect to $p_\nu(\theta)$.

The above raises the natural question what is a better way of expressing prior information. Although (12) is the standard choice in the Bayesian literature, (11) may often be a simpler and more natural way of expressing available knowledge. Prior information about data can often be accumulated quite cheaply, e.g., by processing archived process data or data generated by a simulation model.

4. PRIOR RESTORATION

In practice, we are often unable to specify the density $r_\nu(y,z)$ completely. The results for the ARX model suggest a more realistic problem formulation when, in addition to ν, the expectation of a certain statistic $h(Y,Z)$ with respect to $r_\nu(y,z)$

$$\mathcal{E}_\nu\big(h(Y,Z)\big) = \bar{h}_\nu$$

is available.

Clearly, as $r_\nu(y,z)$ is not known now, the formula (6) cannot be used directly. In the following paragraphs we show how the inaccuracy $\bar{K}(r_\nu{:}s_\theta)$ can be decomposed into sum of two terms one of which can be made (approximately at least) independent of θ while the other depends on r_ν only through $\bar{h}_\nu$.

Joint Exponential Family. Given any model density $s_\theta(y|z)$, we can construct a *joint* exponential family $\mathcal{S}_{\theta;h}$ composed of the joint densities

$$s_{\theta,\lambda}(y,z) = s_\theta(y|z)\,\exp\big(\lambda^T h(y,z) - \psi(\theta,\lambda)\big) \qquad (13)$$

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

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

is logarithm of the normalizing divisor. The functions $h_0(y,z)\equiv 1$, $h_1(y,z)$, …, $h_n(y,z)$ are assumed linearly independent.

We define a *h-projection* $s_{\theta,\hat\lambda}(y,z)$ of the density $r_\nu(y,z)$ onto the exponential family $\mathcal{S}_{\theta;h}$ by the equality

$$\iint s_{\theta,\hat\lambda}(y,z)\,h(y,z)\,\mathrm{d}y\,\mathrm{d}z = \iint r_\nu(y,z)\,h(y,z)\,\mathrm{d}y\,\mathrm{d}z \qquad (14)$$

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

$$\mathcal{R}_\nu \stackrel{\triangle}{=} \Big\{r(y,z) : \iint r(y,z)\,h(y,z)\,\mathrm{d}y\,\mathrm{d}z = \bar{h}_\nu,\ \iint r(y,z)\,\mathrm{d}y\,\mathrm{d}z = 1,\ r(y,z)\ge 0\Big\}.$$

Pythagorean Relationship. The conditional inaccuracy $\bar{K}(r_\nu{:}s_\theta)$ can be regarded as an *unnormalized joint inaccuracy* of $r_\nu(y,z)$ relative to the function $s_\theta(y|z)$

$$\bar{K}(r_\nu{:}s_\theta) = \iint r_\nu(y,z)\,\log\frac{1}{s_\theta(y|z)}\,\mathrm{d}y\,\mathrm{d}z = K(r_\nu{:}s_\theta).$$

Let $s_{\theta,\lambda}(y,z)$ be exponential (13) and $\hat\lambda$ satisfy (14). It holds then

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

where $D(s_{\theta,\hat\lambda}\|s_\theta)$ stands for an *unnormalized* joint Kullback-Leibler distance (Kullback and Leibler, 1951). This implies the following Pythagorean-like relationship that links together the information measures (cf. Fig. 1)

$$\boxed{K(r_\nu{:}s_\theta) = K(r_\nu{:}s_{\theta,\hat\lambda}) + D(s_{\theta,\hat\lambda}\|s_\theta).} \qquad (15)$$

The formula can be seen as analogy of the Pythagorean theorem that holds for Kullback-Leibler distances between probability distributions (Čencov, 1982).

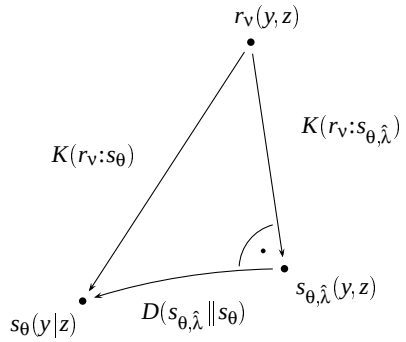

Fig. 1. Pythagorean-like decomposition of conditional inaccuracy.

Approximation of Inaccuracy. It follows directly from the Pythagorean relationship (15) that if $K(r_\nu{:}s_{\theta,\hat\lambda})$ can be made—by a proper choice of $h(y,z)$—"almost independent" of θ

$$K(r_\nu{:}s_{\theta,\hat\lambda}) \approx \text{const.}, \qquad (16)$$

then the inaccuracy $K(r_\nu{:}s_\theta)$ of the unknown $r_\nu(y,z)$ relative to $s_\theta(y|z)$ can be approximated as follows

$$K(r_\nu{:}s_\theta) \approx D(s_{\theta,\hat\lambda}\|s_\theta) + \text{const.} \qquad (17)$$

With this approximation, the conjugate prior density (6) can be approximated as follows

$$\boxed{\hat p_\nu(\theta) \propto \exp\big(-\nu D(s_{\theta,\hat\lambda}\|s_\theta)\big).} \qquad (18)$$

A number of attractive properties of the approximation stem from the use of information measures (for details see Kulhavý, 1996).

The unnormalized Kullback-Leibler distance $D(s_{\theta,\hat\lambda}\|s_\theta)$ can be computed from the following identity

$$0 \;=\; \min_\lambda D(s_{\theta,\hat\lambda}\|s_{\theta,\lambda})$$

$$= \min_{\lambda} \iint s_{\theta,\hat{\lambda}}(y,z) \log \frac{s_{\theta,\hat{\lambda}}(y,z)}{s_{\theta,\lambda}(y,z)} \, \mathrm{d}y \, \mathrm{d}z$$
$$= D(s_{\theta,\hat{\lambda}} \| s_\theta) - \max_{\lambda} \left(\lambda^T \hat{h}(\theta,\hat{\lambda}) - \psi(\theta,\lambda) \right)$$
$$= D(s_{\theta,\hat{\lambda}} \| s_\theta) - \max_{\lambda} \left(\lambda^T \bar{h}_\nu - \psi(\theta,\lambda) \right)$$

which implies

$$\boxed{D(s_{\theta,\hat{\lambda}} \| s_\theta) = \max_{\lambda} \left(\lambda^T \bar{h}_\nu - \psi(\theta,\lambda) \right).} \qquad (19)$$

Choice of h-Statistic. To ensure that (16) holds, we have to choose the statistic $h(y,z)$ in a proper way.

The optimum choice is to set the statistic $h(y,z)$ equal to the canonical statistic of a jointly exponential family (9) enveloping the model family $\{s_\theta(y|z)\}$. In such a case $K(r_\nu : s_{\theta,\hat{\lambda}})$ is independent of θ and (17) holds with equality.

Unfortunately, in many cases of practical interest, the enveloping jointly exponential family has a too large or even infinite dimension which forces us to use a lower-dimensional statistic not sufficient for precise (up to an additive constant) restoration of the inaccuracy $K(r_\nu : s_\theta)$. A general class of h-statistics closest to the above optimum is constructed as follows (cf. Zacks, 1971). Consider a vector space Λ that contains functions $\lambda(\theta) = \log s_\theta(y|z)$ for all $(y,z) \in \mathcal{Y} \times \mathcal{Z}$. Let $L_i(\cdot)$, $i = 1, \ldots, n$ be a set of linear functionals defined on the vector space Λ. Suppose in addition that the linear functionals are normalized so that $L_i(1) = 0$ for $i = 1, \ldots, n$. Then define

$$\boxed{h_i(y,z) = L_i\big(\log s_\theta(y|z)\big)} \qquad (20)$$

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

A simple example of the above definition is the logarithm of density ratio: given $n+1$ points $\theta_1^*, \ldots, \theta_{n+1}^*$ in the parameter space $\mathcal{T}$, we set

$$h_i(y,z) = \log \frac{s_{\theta_{i+1}^*}(y|z)}{s_{\theta_i^*}(y|z)}. \qquad (21)$$

5. CONCLUDING REMARKS

The information-based framework presented in the paper offers an appealing possibility of building the prior density $p_\nu(\theta)$ of unknown parameters via model-based transformation of the prior density $r_\nu(y,z)$ of observed data. The view of prior information via observed data, which can be seen as dual to the standard Bayesian view through model parameters, suggests a number of recipes for quantification of prior knowledge:

(a) Take N archived process data, set $r_\nu(y,z)$ equal to the empirical density $r_N(y,z)$ and choose ν according to the relevance or reliability of the data used.

(b) Do the same with data produced by a faithful (perhaps first-principle) simulation model.

(c) Express your knowledge about (Y,Z) via the prior expectation $\mathcal{E}_\nu\big(h(Y,Z)\big)$. Choose ν so as to reflect your confidence in the expectations provided.

(d) Allow yourself to be more vague by saying only

$$\bar{h}_{\nu,\min} \le \mathcal{E}_\nu\big(h(Y,Z)\big) \le \bar{h}_{\nu,\max}.$$

(e) Play with various definitions of the h-statistic. It may be easier to specify $\mathcal{E}_\nu\big(h(Y,Z)\big)$ for some particular $h(Y,Z)$. Compare the difficulty of guessing *a priori* on the second moments of data

$$\mathcal{E}_\nu(YZ), \quad \mathcal{E}_\nu(ZZ^T)$$

or the log-likelihood ratios

$$\mathcal{E}_\nu\left(\log \frac{s_{\theta_{i+1}^*}(Y|Z)}{s_{\theta_i^*}(Y|Z)} \right), \quad i = 1, \ldots, n.$$

The latter may be easier, especially when having experience with a set of particular models $s_{\theta_i^*}(y|z)$ on archived data or a simulation model. Since the latter is the empirical expectation of (21), it may also bring more information into estimation, especially for nonlinear or non-Gaussian models.

REFERENCES

Čencov, N. N. (1982). *Statistical Decision Rules and Optimal Inference*, Vol. 53 of *Transl. of Math. Monographs*. Amer. Math. Soc., Providence, RI.

Kárný, M., A. Halousková, J. Böhm, R. Kulhavý and P. Nedoma (1985). Design of linear quadratic adaptive control: theory and algorithms for practice. *Kybernetika*, Supplement to No. 3–6.

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

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

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

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

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

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

Zacks, S. (1971). *The Theory of Statistical Inference*. Wiley, New York.