On extension of information geometry of parameter estimation to state estimation

Rudolf Kulhavý
Honeywell Technology Center Europe and
Institute of Information Theory and Automation, AS CR
Pod vodárenskou věží 4, 182 08 Prague, Czech Republic
kulhavy@htc.honeywell.cz

Keywords: parameter estimation, state estimation, Bayesian methods, information measures, approximation

Abstract

The paper indicates a possible way of extending the information-geometric picture of parameter estimation to state estimation (filtering). The three model situations—independent observations, general regression, and state-space model—are considered alongside to stress the similarities and differences.

1 Introduction

The probabilistic methods of parameter estimation implicitly entail approximation of the empirical distribution of observed data with a model-based parametric distribution where the goodness-of-fit criterion coincides with the 'Kerridge inaccuracy' information measure [1]. The specific information geometry born by this picture has an interesting methodological consequence—estimation is approached as an approximation rather than 'inverse' problem (determination of the true distribution generating the data). Even more importantly, the approximation view of estimation opens a route towards systematic computational approximation of recursive estimation in cases when the model considered has no finite-dimensional statistic. A specific Pythagorean geometry of information measures that links Kerridge inaccuracy and relative entropies suggests a finite-dimensional approximation of the inaccuracy function [1].

The above picture holds for independent and identically distributed observations, and can be extended even to controlled dynamic systems modelled through general (non-linear and non-Gaussian) ARX regression. In both cases, the information geometry of (empirical and model-based) distributions of data has a dual counterpart in the information geometry of the posterior distributions of model parameters [2]. The latter view shows that even with a finite-dimensional (but properly chosen) statistic, one can propagate exact knowledge of the set where the true posterior distribution lies.

The paper indicates a possible way of extending the above picture to state estimation. Because of the limited space, we focus on the initial step—translation of the probabilistic formulations of classical estimation scenarios into information-based ones.

2 Independent Observations

The case of independent identically distributed observation has been studied in statistics thoroughly. It is supposed that one observes outcomes of a certain random experiment that is repeated in perfectly identical conditions many (possibly infinitely many) times. The outcomes are taken as samples from a fixed but unknown probability distribution. The objective is to find the probability distribution.

Model Assumptions

A sample of data is a sequence of continuous random variables

$$Y^N = (Y_1, \ldots, Y_N)$$

with values in a subset $\mathcal{Y}$ of $\mathbb{R}^{\dim y}$. The observed sample is formed by the sequence of observed values

$$y^N = (y_1, \ldots, y_N).$$

We make the following assumptions:

1. The variable Y_k is independent of the past data Y^{k-1}. In terms of density functions, it holds

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

   for $k = 2, \ldots, N$.

2. The distribution of Y_k is identical for all k, i.e., $s_k(y) = s(y)$ for $k = 1, \ldots, N$.

3. The unknown density $s(y)$ comes from a known family

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

   parameterized by the parameter θ taking values in a subset $\mathcal{T}$ of $\mathbb{R}^{\dim \theta}$.

4. The density s_θ is strictly positive, $s_\theta(y) > 0$, for all $y \in \mathcal{Y}$ and $\theta \in \mathcal{T}$.

The objective of parameter estimation is to infer the parameter θ from the observed sample y^N.

Bayesian Estimation

We interpret the unknown parameter θ as a random variable Θ.

Under the above model assumptions, the joint density of Y^N takes the form

$$\pi(y_1, \ldots, y_N | \theta) = \prod_{k=1}^{N} s_\theta(y_k). \tag{1}$$

The uncertainty of Θ is described by the *posterior density* $p_N(\theta) = p(\theta | y^N)$ conditional on the observed sample y^N. Given a prior density $p_0(\theta) = p(\theta)$, the posterior density follows from (1) by Bayes's theorem

$$p_N(\theta) \propto p_0(\theta) \prod_{k=1}^{N} s_\theta(y_k) \tag{2}$$

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

Information-Based View

Given the observed sample y^N, the *empirical* density of Y is defined as

$$r_N(y) = \frac{1}{N} \sum_{k=1}^{N} \delta(y - y_k)$$

where $\delta : \mathcal{Y} \to \mathbb{R}$ is a Dirac function satisfying $\delta(y) = 0$ for $y \neq 0$ and

$$\int \delta(y)\, \mathrm{d}y = 1.$$

The *inaccuracy* [3] of $r_N(y)$ relative to $s_\theta(y)$ is defined as

$$K(r_N : s_\theta) = \int r_N(y) \log \frac{1}{s_\theta(y)}\, \mathrm{d}y.$$

Using the above notions, the joint density (1) can be rewritten as follows

$$\begin{aligned} &\pi(y_1, \ldots, y_N | \theta) \\ &\quad = \exp\left(-N \frac{1}{N} \sum_{k=1}^{N} \log \frac{1}{s_\theta(y_k)} \right) \\ &\quad = \exp\left(-N \int r_N(y) \log \frac{1}{s_\theta(y)}\, \mathrm{d}y \right) \\ &\quad = \exp\bigl(-N\, K(r_N : s_\theta) \bigr). \end{aligned}$$

The *posterior* density of Θ conditional on $Y^N = y^N$ then takes the form

$$p_N(\theta) \propto p_0(\theta) \exp\bigl(-N\, K(r_N : s_\theta) \bigr). \tag{3}$$

3 General Regression

The basic problem of system identification is to find a proper model of a dynamic (and possibly controlled) system. The model describes the dependence of the system output on its past values and perhaps on some external inputs as well. The conditional character of the model and the existence of external inputs, measured but typically unmodelled, makes the problem of parameter estimation more difficult compared with the independent observations.

Model Assumptions

The sample of data is formed by two sequences of continuous random variables

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

which take values in subsets $\mathcal{Y}$ and $\mathcal{U}$ of $\mathbb{R}^{\dim \mathrm{y}}$ and $\mathbb{R}^{\dim \mathrm{u}}$, respectively. U_k is a directly manipulated input to the system at time k. Y_k is the system output, i.e., 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 output and input values form the observed sample

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

We make the following assumptions:

1. The output Y_k is conditionally independent of the past data Y^{k-1}, U^k given $Z_k = z_k$ where $Z_k = z(U_{k-m}^k, Y_{k-m}^{k-1})$ is a known vector function taking values in a subset $\mathcal{Z}$ of $\mathbb{R}^{\dim \mathrm{z}}$. In terms of density functions, it holds

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

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

2. The conditional distribution of Y_k given $Z_k = z_k$ is identical for all k, i.e., $s_k(y|z) = s(y|z)$ for $k = m+1, \ldots, N+m$.

3. The density $s(y|z)$ belongs to a known family

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

   parameterized by the parameter θ taking values in a subset $\mathcal{T}$ of $\mathbb{R}^{\dim \theta}$.

4. The density s_θ is strictly positive, $s_\theta(y|z) > 0$, for all $(y, z) \in \mathcal{Y} \times \mathcal{Z}$ and all $\theta \in \mathcal{T}$.

5. The input U_k and the parameter Θ, interpreted as a random variable, are conditionally independent given the past data Y^{k-1}, U^{k-1}. In terms of density functions, it holds

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

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

The objective of parameter estimation is to infer the parameter θ from the observed sample y^{N+m}, u^{N+m}.

Bayesian Estimation

Under the above model assumptions, the joint density of Y_{m+1}^{N+m}, U_{m+1}^{N+m} conditional on the initial values of $Y^m = y^m$, $U^m = u^m$ takes the form

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

The uncertainty of Θ is described by the *posterior density* $p_N(\theta) = p(\theta | y^{N+m}, u^{N+m})$ conditional on the observed sample y^{N+m}, u^{N+m}. Given a prior density $p_0(\theta) = p(\theta | y^m, u^m)$ conditional on available prior information and possibly the m initial values y^m, u^m, the posterior density follows from (4) by conditioning

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

Information-Based View

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

$$r_N(y, z) = \frac{1}{N} \sum_{k=m+1}^{N+m} \delta(y - y_k, z - z_k) \quad (6)$$

where $\delta : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$ is a Dirac function satisfying $\delta(y, z) = 0$ for $y \neq 0$ or $z \neq 0$ and

$$\int \delta(y, z)\, \mathrm{d}y\, \mathrm{d}z = 1.$$

The *conditional inaccuracy* of $r_N(y, z)$ relative to $s_\theta(y | z)$ is defined as

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

Using the above notions, the θ-dependent part of the joint density (4) can be expressed as

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

Hence, the *posterior* density of Θ conditional on the observed sample y^{N+m}, u^{N+m} can be written as

$$p_N(\theta) \propto p_0(\theta)\, \exp\big(-N\, K(r_N : s_\theta)\big) \quad (7)$$

4 State-Space Model

Often it is natural to explain the observed behaviour of a system through its internal state. The state cannot be observed directly, but we can infer it from the observed data using the system model.

Model Assumptions

The sample of data is formed by the input and output sequences

$$Y^N = (Y_1, \ldots, Y_N),\ U^N = (U_1, \ldots, U_N),$$

taking values in subsets $\mathcal{Y}$ and $\mathcal{U}$ of $\mathbb{R}^{\dim \mathrm{y}}$ and $\mathbb{R}^{\dim \mathrm{u}}$, respectively.

The response of the system at time $k = 1, \ldots, N$ is determined by its internal state X_k taking values in a subset $\mathcal{X}$ of $\mathbb{R}^{\dim \mathrm{x}}$

We make the following assumptions:

1. The state X_{k+1} is conditionally independent of the past data Y^k, U^{k-1} given $X_k = x_k$ and $U_k = u_k$. In terms of density functions, it holds

   $$q_k(x_{k+1} | x_k, y^k, u^k) = q_k(x_{k+1} | x_k, u_k)$$

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

2. The output Y_k is conditionally independent of the past data Y^{k-1}, U^{k-1} given $X_k = x_k$ and $U_k = u_k$. That is, we have

   $$s_k(y_k | x_k, y^{k-1}, u^k) = s_k(y_k | x_k, u_k)$$

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

3. The densities $q_k(\xi | x, u)$ and $s_k(y | x, u)$ are identical for all k, i.e., $q_k(\xi | x, u) = q(\xi | x, u)$ and $s_k(y | x, u) = s(y | x, u)$ for $k = 1, \ldots, N$.

4. The densities $q(\xi | x, u)$ and $s(y | x, u)$ are known.

5. The densities q and s are strictly positive, $q(\xi | x, u) > 0$ and $s(y | x, u) > 0$, for all $(\xi, x, y, u) \in \mathcal{X} \times \mathcal{X} \times \mathcal{Y} \times \mathcal{U}$.

6. The input U_k and the state X_k are conditionally independent given the past data Y^{k-1}, U^{k-1}. In terms of density functions, it holds

   $$g_k(u_k | x_k, y^{k-1}, u^{k-1}) = g_k(u_k | y^{k-1}, u^{k-1})$$

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

The objective of parameter estimation is to infer the state values x_k, $k = 1, \ldots, N + 1$ from the observed sample y^N, u^N.

Bayesian Estimation

Under the above model assumptions, the joint density of X_2^{N+1}, Y^N, U^N conditional on the initial state $X_1 = x_1$ takes the form

$$\pi(x_2^{N+1}, y^N, u^N | x_1) = \prod_{k=1}^{N} q(x_{k+1}|x_k, u_k) \cdot s(y_k|x_k, u_k)\, g(u_k|y^{k-1}, u^{k-1}). \quad (8)$$

The uncertainty of X^{N+1} is described by the *posterior* density conditional on the observed sample y^N, u^N

$$p_N(x^{N+1}) = p(x^{N+1}|y^N, u^N).$$

Given the prior density of the initial state $p_0(x_1) = p(x_1)$, the posterior density follows from (8) by conditioning

$$p_N(x^{N+1}) \propto p_0(x_1) \prod_{k=1}^{N} q(x_{k+1}|x_k, u_k)\, s_\theta(y_k|x_k, u_k). \quad (9)$$

Information-Based View

We introduce the notation $\Xi_k = X_{k+1}$. Given particular sequences x^{N+1}, y^N, u^N, the *joint empirical density* of (Ξ, X, Y, U) is defined as

$$r_N(\xi, x, y, u) = \frac{1}{N} \sum_{k=1}^{N} \delta(\xi - \xi_k, x - x_k, y - y_k, u - u_k) \quad (10)$$

where $\delta : \mathcal{X} \times \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$ is a Dirac function satisfying $\delta(\xi, x, y, u) = 0$ for $\xi \neq 0$ or $x \neq 0$ or $y \neq 0$ or $u \neq 0$ and

$$\int \delta(\xi, x, y, u)\, \mathrm{d}\xi\, \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}u = 1.$$

The *conditional inaccuracy* of $r_N(\xi, x, y, u)$ relative to the product density $q(\xi|x, u)\, s(y|x, u)$ is defined by

$$K(r_N : qs) = \int r_N(\xi, x, y, u) \log \frac{1}{q(\xi|x, u)\, s(y|x, u)}\, \mathrm{d}\xi\, \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}u.$$

The x-dependent part of the joint density (8) can be written as

$$\begin{aligned} &\prod_{k=1}^{N} q(x_{k+1}|x_k, u_k)\, s(y_k|x_k, u_k) \\ &= \exp\left(-N \frac{1}{N} \sum_{k=1}^{N} \log \frac{1}{q(x_{k+1}|x_k, u_k)\, s(y_k|x_k, u_k)}\right) \\ &= \exp\Big(-N \int r_N(\xi, x, y, u) \log \frac{1}{q(\xi|x, u)\, s(y|x, u)} \cdot \mathrm{d}\xi\, \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}u\Big) \\ &= \exp\big(-N\, K(r_N : qs)\big). \end{aligned}$$

Hence, the *posterior* density of X^{N+1} conditional on the observed sample y^N, u^N can be written as

$$p_N(x^{N+1}) \propto p_0(x_1) \exp\big(-N\, K(r_N : s_\theta)\big) \quad (11)$$

5 Objective of Approximation

The density r_N is rarely known completely. The data sample is either compressed to limit the amount of stored information or some of the information required is missing or uninteresting. The following are typical scenarios met in practice.

1. The empirical density $r_N(y)$ is replaced by the set $\mathcal{R}_N$ of densities $r(y)$ such that

$$\int r(y)\, h(y)\, \mathrm{d}y = \int r_N(y)\, h(y)\, \mathrm{d}y = \frac{1}{N} \sum_{k=1}^{N} h(y_k)$$

   where $h : \mathcal{Y} \to \mathbb{R}^n$ is an appropriate vector statistic.

2. The empirical density $r_N(y, z)$ is replaced by the set $\mathcal{R}_N$ of densities $r(y, z)$ such that

$$\int r(y, z)\, h(y, z)\, \mathrm{d}y\, \mathrm{d}z = \int r_N(y, z)\, h(y, z)\, \mathrm{d}y\, \mathrm{d}z = \frac{1}{N} \sum_{k=m+1}^{N+m} h(y_k, z_k)$$

   where $h : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}^n$ is an appropriate vector statistic.

3. The empirical density $r_N(\xi, x, y, u)$ is replaced by the set $\mathcal{R}_N$ of empirical densities of all sequences (ξ_k, x_k, y_k, u_k), $k = 1, \ldots, N$ starting at x_1 and ending at x_{N+1} (with an arbitrary trajectory in between), with known sequences (or statistics) of outputs $(y_1, \ldots, y_N)$ and inputs $(u_1, \ldots, u_N)$.

The bottom line of approximation is a Pythagorean-like decomposition of inaccuracy in the vein of [4]. As a result, the inaccuracy $K(r_N : s_\theta)$ or $K(r_N : qs)$ is replaced with minimum relative entropy $D(\mathcal{R}_N \| s_\theta)$ or $D(\mathcal{R}_N \| qs)$, respectively. The estimation case is treated in detail in [1], the filtering case is still a topic of ongoing research.

References

[1] R. Kulhavý. *Recursive Nonlinear Estimation: A Geometric Approach.* London: Springer-Verlag, 1996.

[2] R. Kulhavý. Can we preserve the structure of recursive Bayesian estimation in a limited-dimensional implementation?. In *Systems and Networks: Mathematical Theory and Applications* (U. Helmke, R. Mennicken, and J. Saurer, eds.), vol. I, pp. 251–272, Berlin: Akademie Verlag, 1994.

[3] D. F. Kerridge. Inaccuracy and inference. *J. Roy. Statist. Soc. Ser. B*, 23:284–294, 1961.

[4] S. Amari. *Differential-Geometrical Methods in Statistics.* Berlin: Springer-Verlag, 1985.