Akademie věd české republiky
ÚSTAV TEORIE INFORMACE A AUTOMATIZACE

Academy of Sciences of the Czech Republic
INSTITUTE OF INFORMATION THEORY AND AUTOMATION

RESEARCH REPORT

R. Kulhavý

Bayesian Estimation, Large Deviations, and Incomplete Data

No. 1810 March 1994

ÚTIA AV ČR, P. O. Box 18, 182 08 Prague, Czech Republic
Telex: 122018 atom c, Fax: (42)(2) 6641 4903
Tel: (42)(2) 6605 2313, E-mail: kulhavy@utia.cas.cz

Abstract

Can the Bayesian estimation be modified so to cope with compressed, aggregated or partially missing data? Naturally, the best answer would be to make inference conditional on just information given. But, this is rarely feasible so we have to look for a suitable approximation. The paper shows that the conditional distribution can be approximated so to keep the key features of the optimal solution. First, under weak assumptions, we get the same asymptotic behaviour, only with a lower rate of convergence. Second, the less information we get from data, the more uncertain our inference is. The properties follow essentially from the large deviation theorem and the elementary properties on the informational (Kullback-Leibler) divergence.

1 Introduction

In practice, we often meet the need to estimate a model of a real process from incomplete data. In recursive estimation, for instance, we are forced to compress data to stay within the limits of computer memory. It may happen that the statistic used is not enough to realize the ideal estimation. In econometrics, it is common to have just aggregates of the original data, yet we may be interested in modelling of microeconomic relationships. In engineering and statistics, it is quite frequent that some data are completely missing.

While the theory of parameter estimation is full-grown and able of coping with standard problems, there is no feasible, ready-to-use solution to the above cases. This is true especially for the Bayesian scheme of parameter estimation that is notorious for extreme dimensionality. The theory and practice of approximate Bayesian estimation is currently a domain of numerical analysis [1] and Monte Carlo simulation [2]. During the last two decades a lot of useful methods have been developed, yet there is no commonly accepted approach. While the Bayesian paradigm stresses consistency of our inductive reasoning, in design of approximate estimators we often care only about "staying close enough to the optimum". The connection between the inaccuracy of approximation and the overall uncertainty of parameter estimation is rarely studied.

The paper suggests a new approach that solves, at least partially, the dilemma. By the way, the same scheme has been suggested recently from a differential-geometric perspective [3]. Here we adopt another point of view. Referring to the large deviation theorem [4], we show that the suggested approximation is directly related to the ideal solution which is the distribution of an unknown parameter conditional on incomplete data. To make presentation clear and succinct, we deal with a rather simple example although the basic idea can be applied to more complex model situations as well.

2 Problem Statement

Consider a sequence of random variables ("data") $X_1, \ldots, X_k$ that take values in a finite set Γ. Assume that $X_1, \ldots, X_k$ are mutually independent and identically distributed with a common distribution S on Γ. Suppose further that the "true" distribution S is unknown but it belongs to a finite set $\{S_\theta : \theta \in \Psi\}$ where Ψ is a finite parameter set. For simplicity, assume that $S_\theta(x) > 0$ for each $x \in \Gamma$ and $\theta \in \Psi$. The cardinality of both Γ and Ψ is supposed very large in typical cases.

The problem is to estimate the unknown parameter θ provided the only information about the sample $X = (X_1, \ldots, X_k)$ of size k is the value $\xi = T_k(X)$ of a data statistic $T_k : \Gamma^k \to \mathbb{R}^n$. To be more realistic, we shall suppose that the value of ξ is not known exactly but that $\alpha \leq \xi \leq \beta$ where $\alpha, \beta \in \mathbb{R}^n$ and $\alpha < \beta$ (the inequalities hold entry-wise).

The statistic T_k is usually defined as an empirical average

$$T_k(x) = \frac{1}{k} \sum_{i=1}^{k} T_1(x_i) \tag{1}$$

of a suitable vector function $T_1 : \Gamma \to \mathbb{R}^n$. In the sequel, we shall have in mind only this case although there are other situations that can be handled analogously.

3 Bayesian Estimation (Revisited)

The aim of this section is to remind the Bayesian solution of the problem when the uncertain θ is understood as a random variable. We start by computing the distribution conditional on the sample $X = x$ of size k. The result is well known, of course, but we present it in a form that is not quite common. For that, we need to introduce two notions.

Definition 1 The empirical distribution of a sequence $X_1, \ldots, X_k$ is a (random) distribution defined by the relative frequencies

$$R_X(a) = \frac{1}{k} \sum_{i=1}^{k} \delta_{X_i}(a), \ a \in \Gamma \tag{2}$$

where δ_{X_i} is a point distribution concentrated at X_i.

Clearly, the empirical distribution represents a complete description of observed data. That is, no information is lost by storing only the relative frequencies of observed values (in addition to the sample size k, of course).

Definition 2 The informational divergence (I-divergence) of two distributions R and S on a finite set Γ is defined by the quantity

$$D(R \,||\, S) = \sum_{a \in \Gamma} R(a) \log \frac{R(a)}{S(a)}. \tag{3}$$

where we suppose logarithm to the base e and the standard notational conventions $\log 0 = -\infty$, $\log \frac{a}{0} = \infty$ if $a > 0$, $0 \log 0 = 0 \log \frac{0}{0} = 0$.

Note that the I-divergence $D(R\,||\,S)$ is also known as information for discrimination, Kullback-Leibler information number, information gain, or entropy of R relative to S.

To some extent, we can think of the I-divergence as a nonsymmetric measure of distance between two distributions. The following property is crucial in this context — see e.g. [5, Theorem 3.1].

Lemma 1 *For any two distributions R and S, it holds $D(R\,||\,S) \geq 0$ with the equality if and only if $R = S$.*

With the notions of the empirical distribution and I-divergence defined, we can put the posterior distribution in the following form.

Proposition 1 *The conditional distribution of θ given a sample $X = x$ of size k is given by the formula*

$$P(\theta\,|\,k, x) \propto P(\theta)\ \exp\{-k\, D(R_x\,||\,S_\theta)\} \tag{4}$$

where $\propto$ means equality up to the normalizing factor.

Proof. Applying elementary equations of probability calculus we derive

$$P(\theta\,|\,x) \propto P(\theta)\ \prod_{i=1}^{k} S_\theta(x_i)\,.$$

After simple manipulations we get for the second factor

$$\prod_{i=1}^{k} S_\theta(x_i) = \exp\left\{\sum_{i=1}^{k} \log S_\theta(x_i)\right\} = \exp\left\{k \sum_{a\in\Gamma} R_x(a)\log S_\theta(a)\right\}\,.$$

The proposition then follows by Definition 2. ∎

Proposition 1 shows that the posterior distribution of θ is essentially a function of the I-divergence between the empirical distribution R_x and particular sampling distributions $S_\theta,\ \theta \in \Psi$. Thus, in a sense, the core of the Bayesian estimation is in measuring a "distance" between the actual and model distributions of observed data.

Now, when all we know about a sample X of size k is the value of the data statistic $\xi = T_k(X) \in [\alpha, \beta]$, the empirical distribution R_X becomes also uncertain. Yet, we know it belongs to a certain subset of the set $\mathcal{M}(\Gamma)$ of all distributions on Γ

$$R_X \in \mathcal{R}_\xi = \left\{R \in \mathcal{M}(\Gamma) : E_R[T_1] = \sum_{a\in\Gamma} R(a)\,T_1(a) \in [\alpha, \beta]\right\}\,.$$

Clearly, $\mathcal{M}(\Gamma)$ can be identified with the subset of the Euclidean space that consists of all vectors R whose components are nonnegative, $R(a) \geq 0,\ a \in \Gamma$, and summing up up to 1, $\sum_{a\in\Gamma} R(a) = 1$. In the sequel, the topological terms on $\mathcal{M}(\Gamma)$ refer to the metric topology defined by the Euclidean distance. In particular, the convergence $R_k \to R$ means that $R_k(a) \to R(a)$ for each $a \in \Gamma$.

The optimal Bayesian solution to the incomplete data problem is the conditional distribution

$$P(\theta\,|\,k,\xi) \propto P(\theta)\, S_\theta^k(\{x : R_x \in \mathcal{R}_\xi\}) \tag{5}$$

where S_θ^k denotes the joint distribution of a sample $X = (X_1, \ldots, X_k)$, i.e., $S_\theta^k(x) = \prod_{i=1}^k S_\theta(x_i)$. In most cases, the computation of $S_\theta^k(\{x : R_x \in \mathcal{R}_\xi\})$ is not feasible. The only thing to do then is to look for a suitable approximation.

4 Large Deviation Theorem

Let us remind one of the classical results of large deviation theory.

Theorem 1 *Let $\mathcal{R}_\xi \subset \mathcal{M}(\Gamma)$ be a set of probability distributions on Γ such that the closure of the interior of $\mathcal{R}_\xi$ equals $\mathcal{R}_\xi$. Then for k independent drawings from a distribution S_θ with a fixed $\theta \in \Psi$, the probability of a sample with an empirical distribution belonging to $\mathcal{R}_\xi$ has the asymptotics*

$$\lim_{k\to\infty} \frac{1}{k} \log S_\theta^k(\{x : R_x \in \mathcal{R}_\xi\}) = -D(\mathcal{R}_\xi \,||\, S_\theta) \tag{6}$$

where

$$D(\mathcal{R}_\xi \,||\, S_\theta) = \min_{R\in\mathcal{R}_\xi} D(R \,||\, S_\theta) \tag{7}$$

The original proof of the large deviation theorem was done in [4]. For more background, the reader is referred to [6] or [7].

Note that the sufficient condition stated in Theorem 1 is clearly fulfilled with our formulation.

5 Approximation

Theorem 1 says that the probability $S_\theta^k(\{x : R_x \in \mathcal{R}_\xi\})$ converges to zero (provided $S_\theta \notin \mathcal{R}_\xi$) exponentially fast, with the rate given by the I-divergence between the set $\mathcal{R}_\xi$ of distributions consistent with the value of the used statistic, and the true sampling distribution S_θ. Strictly speaking, it does not say anything about the probability $S_\theta^k(\{x : R_x \in \mathcal{R}_\xi\})$ itself, but it describes its factor that decides about the asymptotic behaviour.

This leads us to the following **approximation** of (5)

$$\hat{P}(\theta \mid k, \xi) \propto P(\theta) \exp\{-k\, D(\mathcal{R}_\xi \,\|\, S_\theta)\}\,. \tag{8}$$

Compare (4) and (8). With just partial information about data, we can measure only a "distance" from the entire set $\mathcal{R}_\xi$ of possible empirical distributions.

Lemma 1 and (7) imply the following *monotonicity property*

$$0 \le D(\mathcal{R}_\xi \mid S_\theta) \le D(R_x \mid S_\theta)\,. \tag{9}$$

Lemma 1 and (9) give rise to the following implications.

1. *If* $\mathcal{R}_\xi = \{R_x\}$, *then* $\hat{P}(\theta \mid k, \xi) = P(\theta \mid k, x)$. In other words, if the statistic makes it possible to reconstruct the true empirical distribution, the approximation gives the true posterior distribution.

2. *If* $\mathcal{R}_\xi = \mathcal{M}(\Gamma)$, *then* $\hat{P}(\theta \mid k, \xi) = P(\theta)$. In other words, if the statistic brings no information about data at all, the approximation returns the prior distribution.

3. *If* $R_x \to S_{\theta_0}$ *for some* θ_0, *then* $D(\mathcal{R}_\xi \,\|\, S_{\theta_0}) \to 0$. *If, moreover,* $P(\theta_0) > 0$, *then also* $\lim_{k\to\infty} \hat{P}(\theta \mid k, \xi) > 0$. In other words, the asymptotic behaviour of approximate estimation is consistent with the ideal estimation.

Figure 1 illustrates some of the properties. The example behind the plots was estimation of a location parameter of an ε-contaminated normal distribution $(1 - \varepsilon)\mathcal{N}(-2, 3) + \varepsilon\,\mathcal{N}(-2, 30)$ with $\varepsilon = 0.1$. To stay with our problem formulation, we supposed just a finite but large number of possible values of x and θ. The dimension of the statistic used was extremely low, $n = 2$.

6 Concluding Remarks

For convenience, the problem solved in the paper has been very simple. But, the idea of approximation works in more general setups too. It can easily be extended to the case of dependent data modelled by a Markov chain. A suitable form of the posterior distribution is suggested again by the appropriate large deviation theorem (see e.g. [8]). The extension to the case of continuous parameters is straightforward. So is the case of continuous data which, however, requires more advanced mathematical tools (cf. [6]).

Acknowledgment

This work was supported in part by the Czech Academy of Sciences Grant 275109.

References

[1] 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.

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

[3] R. Kulhavý, "Can approximate Bayesian estimation be consistent with the ideal solution?," in *Proceedings of the 12th IFAC World Congress*, vol. 4, (Sydney, Australia), pp. 225–228, 1993.

[4] I. N. Sanov, "On the probability of large deviations of random variables (in Russian)," *Mat. Sb. (N.S.)*, vol. 42, pp. 11–44, 1957. English translation in *Sel. Transl. Math. Statist. Probab.* **1** (1961), 213–244.

[5] S. Kullback, *Information Theory and Statistics*. New York: John Wiley & Sons, 1959.

[6] R. S. Ellis, *Entropy, Large Deviations, and Statistical Mechanics*. Berlin: Springer-Verlag, 1985.

[7] J. A. Bucklew, *Large Deviation Techniques in Decision, Simulation, and Estimation*. New York: John Wiley & Sons, 1990.

[8] I. Csiszár, T. M. Cover, and B.-S. Choi, "Conditional limit theorem under Markov conditioning," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 788–801, Nov 1987.

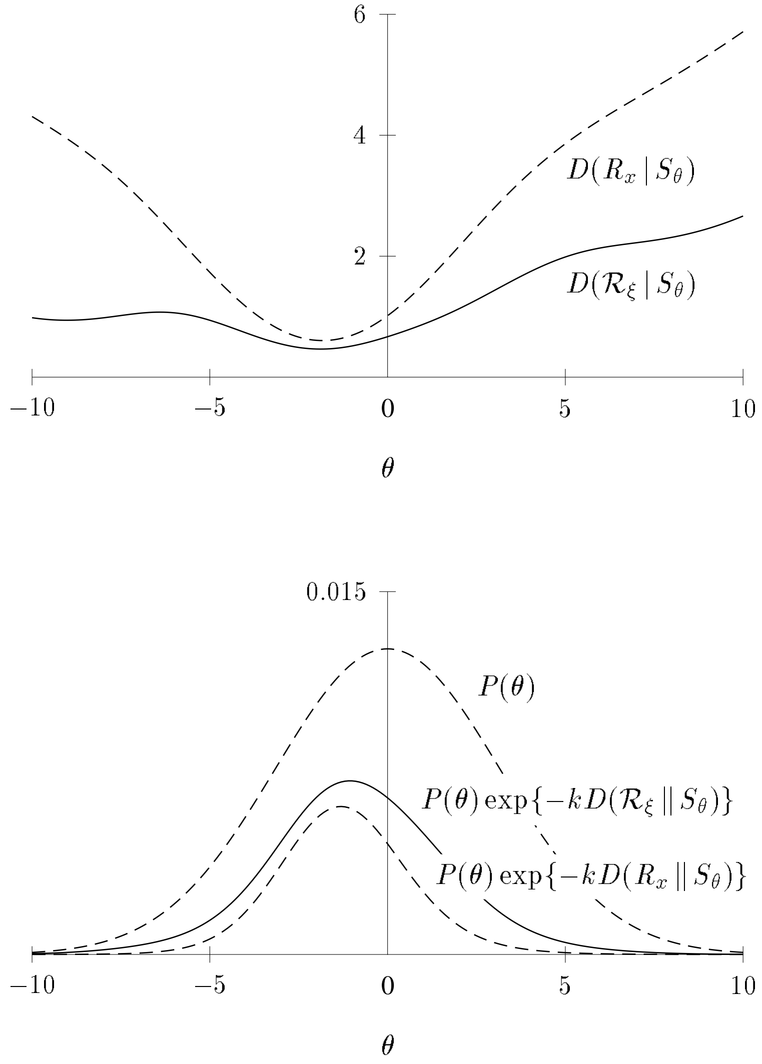

Figure 1: The plots demonstrate that the approximation preserves the key feature of the optimal solution — the uncertainty of parameter estimation increases as a result of data compression. The upper plot compares the I-divergence, taken as a function of the unknown parameter θ, for complete and reduced data. The other plot illustrates that, up to normalization, the approximate distribution lies always "between" the prior and posterior distributions.