Bayesian Estimation, Large Deviations, and Incomplete Data

Rudolf Kulhavý[1]

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

The paper suggests an approximation of Bayesian parameter estimation for the case that data are incomplete. Attractive properties of the approximation follow from the large deviation theorem and the elementary properties of the informational divergence.

1. Introduction

In practice, we often meet the need to estimate the 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. In econometrics, it is common to have just aggregates of the original data. In engineering and statistics, some data are often missing.

While the theory of parameter estimation is full-grown and able to cope 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 paper suggests a new approach with appealing both asymptotic and short-sample properties. Recently the same scheme has been suggested from a differential-geometric perspective [1]. Here, referring to the large deviation theorem [2], we show that the suggested approximation is directly related to the ideal solution which is the distribution of the 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 $\mathcal{X}$. Assume that $X_1, \ldots, X_k$ are mutually independent and identically distributed with a common distribution S on $\mathcal{X}$. Suppose further that the "true" distribution S is unknown but it belongs to a finite set $\{S_\theta : \theta \in \mathcal{T}\}$ where $\mathcal{T}$ is a finite parameter set. For simplicity, assume that $S_\theta(x) > 0$ for each $x \in \mathcal{X}$ and $\theta \in \mathcal{T}$. The cardinality of both $\mathcal{X}$ and $\mathcal{T}$ 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 that $T_k(X) \geq \xi$ where $T_k : \mathcal{X}^k \to \mathbb{R}^n$ is a given data statistic and $\xi \in \mathbb{R}^n$ is a fixed vector.

The statistic T_k is usually defined as the empirical average

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

of a suitable vector function $T_1 : \mathcal{X} \to \mathbb{R}^n$. 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 Bayesian solution of the problem is given by the probability distribution of the unknown parameter Θ (regarded as a random variable) conditional on the sample $X = x$ of size k. The result is naturally well known, but we present it in a form that is not so common. Two notions need to be introduced first.

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

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

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

It is easy to verify that the empirical distribution together with the sample size k form a sufficient statistic for the problem.

Definition 2 The informational divergence (I-divergence) of two distributions R and S on a finite set $\mathcal{X}$ is defined by the quantity

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

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

To some extent, the I-divergence can be thought of as a non-symmetric measure of distance between two distributions. The following property is crucial in this context — see e.g. [3, 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 write 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.

[1] This work was supported in part by Grant 102/94/0314 of the Czech Grant Agency and Grant 275109 of the Czech Academy of Sciences.

Proof. Elementary rules of probability calculus give

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

where $\sum_{i=1}^{k} \log S_\theta(x_i) = k \sum_{a \in \mathcal{X}} R_x(a) \log S_\theta(a)$. The proposition follows by Definition 2. ∎

Proposition 1 shows that the posterior distribution of Θ is a function of the I-divergence between the empirical distribution R_x and particular sampling distributions $S_\theta,\ \theta \in \mathcal{T}$. Thus, in Bayesian inference one implicitly measures "distances" between the actual and model distributions of observed data.

Suppose now that all we know about the sample X of size k is that $T_k(X) \geq \xi$. As a result, the empirical distribution R_X becomes uncertain. Owing to (1) and (2) we know, however, that R_X belongs to a certain subset of the set $\mathcal{R}$ (probability simplex) of all distributions on $\mathcal{X}$

$$\mathcal{R}_\xi = \left\{ R \in \mathcal{R} : \textstyle\sum_{a \in \mathcal{X}} R(a)\, T_1(a) \geq \xi \right\}\,. \tag{5}$$

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

$$P(\theta \,|\, k, \xi) \propto P(\theta)\, S_\theta^k(\{x : R_x \in \mathcal{R}_\xi\}) \tag{6}$$

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

The following classical result of large deviation theory suggests a suitable approximation of the probability $S_\theta^k(\{x : R_x \in \mathcal{R}_\xi\})$.

Theorem 1 (Sanov [2]) *Let $\mathcal{R}_\xi$ (5) be 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 an arbitrary fixed $\theta \in \mathcal{T}$, the probability of the sample with an empirical distribution belonging to $\mathcal{R}_\xi$ has the asymptotics*

$$\lim_{k \to \infty} \frac{1}{k} \log S_\theta^k(\{x : R_x \in \mathcal{R}_\xi\}) = -D(\mathcal{R}_\xi \,\|\, S_\theta) \tag{7}$$

where

$$D(\mathcal{R}_\xi \,\|\, S_\theta) = \min_{R \in \mathcal{R}_\xi} D(R \,\|\, S_\theta) \tag{8}$$

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 says nothing 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 (6)

$$\hat{P}(\theta \,|\, k, \xi) \propto P(\theta)\, \exp\{-k\, D(\mathcal{R}_\xi \,\|\, S_\theta)\}\,. \tag{9}$$

Compare (4) and (9). What the approximation suggests to do is to measure "distances" of $S_\theta, \theta \in \mathcal{T}$ from the *entire* set $\mathcal{R}_\xi$ of possible empirical distributions.

Lemma 1 and (8) imply the following *monotonicity property*

$$0 \leq D(\mathcal{R}_\xi \,|\, S_\theta) \leq D(R_x \,|\, S_\theta)\,. \tag{10}$$

Lemma 1 and (10) give rise to the following implications.

1. *If $\mathcal{R}_\xi = \{R_x\}$, then $\hat{P}(\theta \,|\, k, \xi) = P(\theta \,|\, k, x)$.* Thus, if the statistic makes it possible to reconstruct the true empirical distribution, the approximation returns the true posterior distribution.
2. *If $\mathcal{R}_\xi = \mathcal{R}$, then $\hat{P}(\theta \,|\, k, \xi) = P(\theta)$.* 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 $\lim_{k \to \infty} \hat{P}(\theta_0 \,|\, k, \xi) > 0$.* The asymptotic behaviour of approximate estimation is consistent with the ideal estimation.

Figure 1 illustrates the behaviour of the approximation (9). We estimated the location parameter of an ε-contaminated normal distribution $(1-\varepsilon)\,\mathsf{N}(-2,3) + \varepsilon\,\mathsf{N}(-2,30)$ with $\varepsilon = 0.1$. To keep 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, $n = 2$.

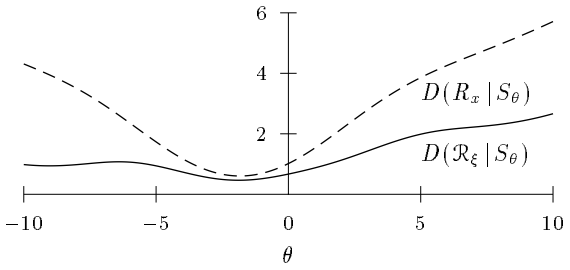

Figure 1: The comparison of the I-divergences for complete and reduced data.

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. [4]). The extension to the case of continuous parameters is straightforward. So is the case of continuous data which, however, requires more advanced mathematical tools.

7. References

[1] 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.

[2] 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.* **I** (1961), 213–244.

[3] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.

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