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 and F. Hrnčíř

Approximation and Uncertainty in Parameter Estimation

No. 1811 July 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

Except for simple models, limited computer resources necessarily make parameter estimation more uncertain or less precise compared to the theoretical ideal. However obvious this implication is, there seems to be no formal expression of it in the theory. The paper looks for an estimation scheme close to the Bayesian one that would make the approximation an integral part of inference.

1 Approximate Bayesian Estimation: the State of the Art

Real computer resources are always limited — they offer only a limited space (memory) and time for numerical computations. Under these constraints, the theoretically optimal statistical inference is often impracticable and needs to be approximated in some way. The problem of parameter estimation for non-linear, non-Gaussian, or high-dimensional models has been pursued intensively in the last three decades.

Engineering Science

In the late 1960s and early 1970s, much effort was spent in engineering science to find a recursive solution to nonlinear identification. To some extent, the research was driven by the practical need of the aerospace industry that looked for reliable algorithms of guidance and navigation. On the theoretical side, the interest was largely stimulated by the success of Kalman filter in solving linear problems. A variety of methods have been developed then. Most of them come into one of two groups. The "local" methods try to build a simpler model that would be close, locally at least, to the original model. The "global" methods substitute a more tractable function for the true posterior density while preserving the original model. A good survey of the progress made in this area can be found in [1, 2].

Practical Bayesian Statistics

Research in Bayesian statistics has focused on the non-recursive case when a complete sample of data is available for analysis. The key problem is then effective numerical solution of integrals appearing in the evaluations of the predictive distribution, marginal distributions, or expectations. Several approaches have been pursued: numerical integration using various quadrature formulae [3], analytical approximation using e.g. the Laplace method for integrals [4], and advanced Monte Carlo methods that take advantage of powerful techniques of sampling from highly multivariate distributions, like Gibbs sampler [5] or sampling-resampling [6].

Theory or Art?

The progress made in solving extremely difficult identification tasks is tangible. Yet, there are so many open questions in this area that one tends to speak of the art rather than the theory of approximate estimation. The central issue appears to be the relationship between approximation and inference. One intuitively feels that the uncertainty of estimation should be affected by the imprecision of its computational implementation. The farther the actual estimator is from the optimal one, the more uncertain the inference is likely to be (for finite samples, at least). An adequate expression of this seems to be missing in the theory of parameter estimation. The paper outlines a possible compromising solution of the dilemma.

2 Bayesian Estimation: Textbook Solution

We shall consider two standard model situations, namely the cases of independent data and Markov chain conditioning. In both cases, we suppose a sequence of random variables $X_1, X_2, \ldots$ with values in a *finite* set $\mathcal{X} = \{1, 2, \ldots, |\mathcal{X}|\}$ where $|\mathcal{X}|$ denotes the number of elements of $\mathcal{X}$. For simplicity, distributions of finite sets will be identified with their probability mass functions.

Independent Data

Suppose that $X_1, X_2, \ldots, X_k$ are independent and identically distributed (i.i.d.). Thus, for $k = 1, 2, \ldots$

$$\Pr\{X_k = x_k | X_{k-1} = x_{k-1}, \ldots, X_1 = x_1\} = \Pr\{X_k = x_k\} \tag{1}$$

for all $x_1, x_2, \ldots, x_k \in \mathcal{X}$. Further, each of $X_1, X_2, \ldots, X_k$ is distributed according to a common probability mass function $S(x) = \Pr\{X_k = x\}$. The distribution is not known completely, but it is assumed to belong to a finite set $\{S_\theta : \theta \in \mathcal{T}\}$. Without loss of generality, we can set $\mathcal{T} = \{1, 2, \ldots, |\mathcal{T}|\}$. The objective is to estimate the unknown parameter θ.

Under the Bayesian viewpoint that interprets the unknown parameter as a random variable Θ, the estimation task is solved by evaluating the posterior distribution $P_{\boldsymbol{x}}$ of Θ conditional on the event $\boldsymbol{X} = \boldsymbol{x}$ where $\boldsymbol{X} = (X_1, \ldots, X_k)$ and $\boldsymbol{x} = (x_1, \ldots, x_k)$. A prior distribution P of Θ needs to be chosen beforehand.

Proposition 1: *Under the i.i.d. assumption, the posterior probability mass function of the unknown parameter Θ conditional on $\boldsymbol{X} = \boldsymbol{x}$ takes the form*

$$P_{\boldsymbol{x}}(\theta) \propto P(\theta) \prod_{j=1}^{k} S_\theta(x_j)\,. \tag{2}$$

Throughout the paper, the symbol $\propto$ stands for proportionality, i.e., equality up to a normalizing factor.

Proof: The proposition follows immediately by the i.i.d. assumption and Bayes rule. ■

Markov Chain Conditioning

Suppose that $X_1, X_2, \ldots, X_{k+1}$ form a Markov chain, i.e., for $k = 1, 2, \ldots$

$$\Pr\{X_{k+1} = x_{k+1} | X_k = x_k, \ldots, X_1 = x_1\} = \Pr\{X_{k+1} = x_{k+1} | X_k = x_k\} \tag{3}$$

for all $x_1, x_2, \ldots, x_{k+1} \in \mathcal{X}$. In addition, assume that the conditional probability $S(y|z) = \Pr\{X_{k+1} = y | X_k = z\}$ does not depend on k, i.e., the Markov chain is time-invariant. The transition probability S in known only partially — it is assumed to belong to a finite set $\{S_\theta : \theta \in \mathcal{T}\}$ where $\mathcal{T} = \{1, 2, \ldots, |\mathcal{T}|\}$. The objective is to estimate the parameter θ again.

The Bayesian solution to the problem is given analogously as above by evaluating the posterior distribution $P_{\boldsymbol{x}}$ of Θ conditional on the event $\boldsymbol{X} = \boldsymbol{x}$ where $\boldsymbol{X} = (X_1, \ldots, X_{k+1})$ and $\boldsymbol{x} = (x_1, \ldots, x_{k+1})$.

Proposition 2: *Under the Markov chain conditioning, the posterior probability mass function of the unknown parameter Θ conditional on $\boldsymbol{X} = \boldsymbol{x}$ takes the form*

$$P_{\boldsymbol{x}}(\theta) \propto P(\theta) \prod_{j=1}^{k} S_\theta(x_{j+1} | x_j)\,. \tag{4}$$

Proof: The proposition follows easily by the Markov chain assumption and Bayes rule. ■

Curse of Dimensionality

What prevents us in a general case from evaluating the posterior distributions according to the formulae (2) and (4) is the large cardinality of the sample and parameter sets $\mathcal{X}$ and $\mathcal{T}$.

Note that if $|\mathcal{T}|$ is small enough, one can compute directly the posterior probability vector. For instance, in the i.i.d. case, (2) can be put in the vector form

$$\begin{bmatrix} P_{\boldsymbol{x}}(1) \\ \vdots \\ P_{\boldsymbol{x}}(|\mathcal{T}|) \end{bmatrix} \propto \begin{bmatrix} P(1) \\ \vdots \\ P(|\mathcal{T}|) \end{bmatrix} \prod_{j=1}^{k} \begin{bmatrix} S_1(x_j) \\ \vdots \\ S_{|\mathcal{T}|}(x_j) \end{bmatrix} .$$

When $|\mathcal{T}|$ is large, but $|\mathcal{X}|$ is relatively small, one can write the posterior distribution in a split form, letting the likelihood components, corresponding to particular elements $a \in \mathcal{X}$, separate. For instance, in the i.i.d. case, the posterior probability mass function can be written as

$$P_{\boldsymbol{x}}(\theta) \propto P(\theta) \prod_{a \in \mathcal{X}} S_\theta(a)^{N_{\boldsymbol{x}}(a)}$$

where $N_{\boldsymbol{x}}(a)$ counts the number of occurrences of a in $\boldsymbol{x}$.

When both $|\mathcal{T}|$ and $|\mathcal{X}|$ are too large and $S_\theta(x)$, $\theta \in \mathcal{T}$ are of "general form" (namely, they do not belong to a lower-dimensional exponential family), the formulae (2) and (4) are infeasible and need to be approximated in some way. The choice of a suitable approximation is the topic of the rest of the paper.

3 Approximate Estimation with I.I.D. Data

Bayesian Estimation Revisited

We start with showing that the posterior probability mass function is nothing but transformation of a distance between the actual distribution and particular model distributions of observed data.

Definition 1: *The first-order type of a sequence $\boldsymbol{x} = (x_1, \ldots, x_k)$ of elements of $\mathcal{X}$ is the distribution $R_{\boldsymbol{x}}$ defined by the relative frequencies*

$$R_{\boldsymbol{x}}(a) = \frac{1}{k} \left|\{i \in \{1, \ldots, k\} : x_i = a\}\right| , \ a \in \mathcal{X} .$$

Remember that $|A|$ denotes the number of elements of the set A.

For a given sequence of random variables $X_1, X_2, \ldots$ with values in $\mathcal{X}$, the first-order type of the sample $(X_1, \ldots, X_k)$ is called the first-order empirical distribution.

Definition 2: *The relative entropy or Kullback-Leibler distance between two probability distributions R and S on $\mathcal{X}$ is defined as*

$$D(R\|S) = \sum_{x\in\mathcal{X}} R(x)\,\log\frac{R(x)}{S(x)}\,. \tag{5}$$

Throughout the paper, we consider logarithms to the base e, with the standard notational conventions $\log 0 = -\infty$, $\log\frac{a}{0} = \infty$ if $a > 0$, $0\log 0 = \log\frac{0}{0} = 0$.

Using the notions of the sequence type and relative entropy, the posterior distribution can be put the following appealing form.

Proposition 3: *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with a common distribution S_θ such that $S_\theta(x) > 0$ for all $x \in \mathcal{X}$ and $\theta \in \mathcal{T}$. Then the posterior probability distribution $P_{\boldsymbol{x}}(\theta)$ of Θ conditional on $\boldsymbol{X} = \boldsymbol{x}$ can be written as*

$$P_{\boldsymbol{x}}(\theta) \propto P(\theta)\,\exp\{-k\,D(R_{\boldsymbol{x}}\|S_\theta)\}\,. \tag{6}$$

Proof: The result follows by applying Definitions 1 and 2 to Proposition 1 and using some straightforward manipulations. ∎

It is well-known that the relative entropy can be thought of as a nonsymmetric measure of distance between two distributions (see [7, 8, 9, 10] for details). Specifically, it holds $D(R\|S) \geq 0$ for any two distributions, with equality if and only if $R = S$. Thus, we can say that the posterior probability $P_{\boldsymbol{x}}(\theta)$ is the higher, the smaller the distance of the sampling distribution S_θ from the type $R_{\boldsymbol{x}}$ of a given sequence $\boldsymbol{x}$ is. Note that the size k of the sample accounts for the time dynamics of the posterior distribution while the modifying effect of the prior distribution P typically diminishes with $k \to \infty$.

Data Compression

Suppose now a more general (and more realistic) case that we do not know the whole sequence $\boldsymbol{x} = (x_1, \ldots, x_k)$, but only the value of a certain statistic.

Definition 3: *Given a sample $\boldsymbol{X} = (X_1, \ldots, X_k)$ we define a first-order data statistic as the sample average*

$$T(\boldsymbol{X}) = \frac{1}{k}\sum_{j=1}^{k} h(X_j) \tag{7}$$

of a vector function $h : \mathcal{X} \to \mathbb{R}^n$ of X.

Note that using the concept of the empirical distribution, we can write the data statistic T as the empirical mean of $h(X)$, i.e.,

$$T(\boldsymbol{X}) = \sum_{a \in \mathcal{X}} R_{\boldsymbol{X}}(a)\, h(a)\,.$$

Let the only information we have about the sequence $\boldsymbol{x} = (x_1, \ldots, x_k)$ be that

$$T(\boldsymbol{x}) \geq \xi$$

where $\xi \in \mathbb{R}^n$. Clearly, the type $R_{\boldsymbol{x}}$ of the sequence $\boldsymbol{x}$ is also uncertain then; all we know is that

$$R_{\boldsymbol{x}} \in \mathcal{R}_{\xi} = \left\{ R \in \mathcal{R} : \sum_{a \in \mathcal{X}} R(a)\, h(a) \geq \xi \right\} \tag{8}$$

where $\mathcal{R}$ stands for the set of all probability distributions on $\mathcal{X}$. Topological concepts on $\mathcal{R}$ refer to the topology of pointwise convergence.

The following *large deviation theorem* provides a useful tool for approximate evaluation of the probability $\Pr\{\boldsymbol{x} \in \mathcal{X}^k : R_{\boldsymbol{x}} \in \mathcal{R}_{\xi}\}$.

Theorem 1: *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with a common distribution S such that $S(x) > 0$ for all $x \in \mathcal{X}$. Let $\mathcal{C} \subset \mathcal{R}$ be a nonempty set of probability distributions on $\mathcal{X}$. If $\mathcal{C}$ is the closure of its interior, then*

$$\lim_{k \to \infty} \frac{1}{k} \log \Pr\{\boldsymbol{x} \in \mathcal{X}^k : R_{\boldsymbol{x}} \in \mathcal{C}\} = - \min_{R \in \mathcal{C}} D(R \| S)\,. \tag{9}$$

Proof: The result was proved first in [11, Theorem 2], a more general form was shown in [12, Theorem 1]. Both proofs apply Stirling's formula. The proof using the method of types (developed in [9]) can be found in [13, Theorem 1] and [10, Theorem 12.4.1]. The basic argument is that the number of sequences of any type grows with k exponentially fast while the number of types grows "only" polynomially. ■

The probability that the empirical distribution belongs to a set not containing the true sampling distribution is known (by the law of large numbers) to converge to zero. Note that the large deviation theorem refines this statement, showing that the probability converges to zero exponentially fast, with the rate given by the relative entropy between the sampling distribution and a given set.

Approximate Posterior

By definition (8), $\mathcal{R}_\xi$ satisfies the assumption of Theorem 1 unless its interior is empty. The large deviation theorem suggests then to approximate the posterior distribution as follows

$$\hat{P}_\xi(\theta) \propto P(\theta)\, \exp\Big\{-k\, D(\mathcal{R}_\xi \,\|\, S_\theta)\Big\} \tag{10}$$

where

$$D(\mathcal{R}_\xi \,\|\, S_\theta) = \min_{R \in \mathcal{R}_\xi} D(R \,\|\, S_\theta)\,. \tag{11}$$

For simplicity, the sample size k is omitted in the subscript of $\hat{P}_\xi$.

Compare (6) and (10): the approximation copes with compressed data by measuring the relative entropy between the sampling distributions S_θ and the set $\mathcal{R}_\xi$ *as a whole*.

The properties of the approximation (10) result from the inequality

$$0 \le D(\mathcal{R}_\xi \,\|\, S_\theta) \le D(R_{\boldsymbol{x}} \,\|\, S_\theta) \ \text{ for all } \ \theta \in \mathcal{T} \tag{12}$$

that follows from the nonnegativeness of the relative entropy and the fact that $R_{\boldsymbol{x}} \in \mathcal{R}_\xi$.

Note that the bounds in (12) correspond to two extreme cases. If the statistic T makes is possible to reconstruct the type of $\boldsymbol{x}$, i.e., $\mathcal{R}_\xi = \{R_{\boldsymbol{x}}\}$, the approximation returns the true posterior distribution, $\hat{P}_\xi(\theta) = P_{\boldsymbol{x}}(\theta)$. On the contrary, if the statistic T brings no information about data at all and so $\mathcal{R}_\xi = \mathcal{R}$, the approximation returns the prior distribution, $\hat{P}_\xi(\theta) = P(\theta)$.

In general, the approximation $\hat{P}_\xi(\theta)$ is the closer to the true posterior $P_{\boldsymbol{x}}(\theta)$, the more information the statistic carries. More precisely, the following implication holds

$$\mathcal{R}_\xi \subset \mathcal{R}'_\xi \Rightarrow D(\mathcal{R}'_\xi \,\|\, S_\theta) \le D(\mathcal{R}_\xi \,\|\, S_\theta) \ \text{ for all } \ \theta \in \mathcal{T}\,. \tag{13}$$

In addition, the asymptotic behaviour of the approximate solution is consistent with the ideal estimation. Namely, if $R_{\boldsymbol{x}} \to S_{\theta_0}$ for some $\theta_0 \in \mathcal{T}$, then $D(\mathcal{R}_\xi \,\|\, S_{\theta_0}) \to 0$. If $P(\theta_0) > 0$ *a priori*, then also $\lim_{k\to\infty} \hat{P}_\xi(\theta_0) > 0$. More about the consistency of estimators of this kind can be found in [14, 15].

Minimum Relative Entropy

The minimum relative entropy (11) is determined by the following theorem.

Theorem 2: *Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables with a common distribution S on $\mathcal{X}$. Then*

$$D(\mathcal{R}_\xi \,\|\, S) = \max_{\lambda \geq 0} \left[\sum_{i=1}^{n} \lambda_i \, \xi_i - \log \alpha(\lambda) \right] \tag{14}$$

where $\lambda = (\lambda_1, \dots, \lambda_n)$ and $\alpha(\lambda)$ is the sum of entries of an $|\mathcal{X}|$-vector $v(\lambda)$ whose x entry is

$$v_x(\lambda) = S(x) \, \exp\left\{ \sum_{i=1}^{n} \lambda_i \, \xi_i(x) \right\}. \tag{15}$$

Proof: This is a standard convex programming problem that can be solved either by applying the Kuhn-Tucker conditions, or the saddle-point optimality conditions with the help of the basic properties of the relative entropy. A more general solution can be found in [16, Theorems 2 and 3], its statistical aspects are discussed in [17]. ∎

4 Example: Cauchy-like Distribution

To illustrate the basic properties of the approximation (10)–(11), we simulated a sequence of 100 independent data distributed according to the Cauchy-like (discretized) sampling distribution

$$S(x) = \text{const.} \times \frac{1}{\sigma} \frac{1}{1 + \left(\frac{x-\mu}{\sigma}\right)^2}$$

with $\mu = 1$ and $\sigma = 2$ (see Fig. 1). The data set $\mathcal{X}$ was formed by 50 values equidistantly located within the interval $[-20, 20]$.

We considered two estimation problems. In the first example, the estimated parameter θ was the location parameter μ. The parameter set $\mathcal{T}$ was formed by 50 values equidistantly located within the interval $[-10, 10]$. The prior distribution was uniform for simplicity. The statistic generating functions were chosen as

$$h_i(x) = \log S_{\mu_i}(x) - \log S_{\mu_{i+1}}(x), \ i = 1, 2$$

with $\mu_1 = -3$, $\mu_2 = 0$, $\mu_3 = 3$ (cf. Fig. 2). Motivation for this choice can be found in [18, 19, 20].

The relative entropy functions $D(\mathcal{R}_\xi \,\|\, S_\mu)$, $D(R_{\boldsymbol{x}} \,\|\, S_\mu)$ and the posterior distributions $\hat{P}_\xi(\mu)$, $P_{\boldsymbol{x}}(\mu)$ are compared in Fig. 3 and 4, respectively. Note that $D(\mathcal{R}_\xi \,\|\, S_\mu) \leq D(R_{\boldsymbol{x}} \,\|\, S_\mu)$ for all $\mu \in \mathcal{T}$ indeed. The large difference was because

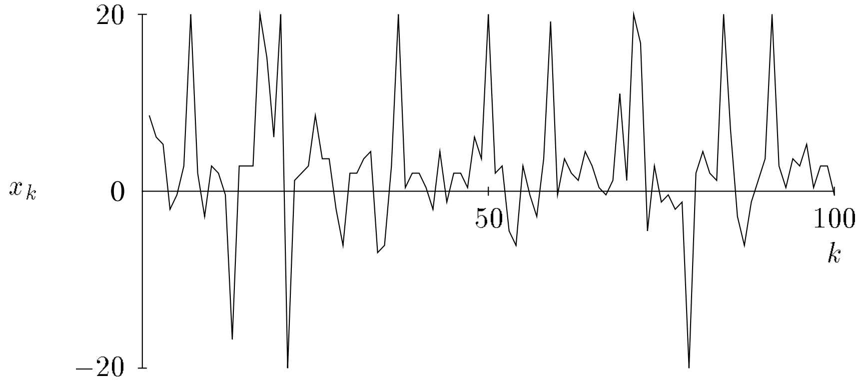

Figure 1: A sequence of 100 i.i.d. data drawn from a Cauchy-like distribution modelling the presence of outliers in observations.

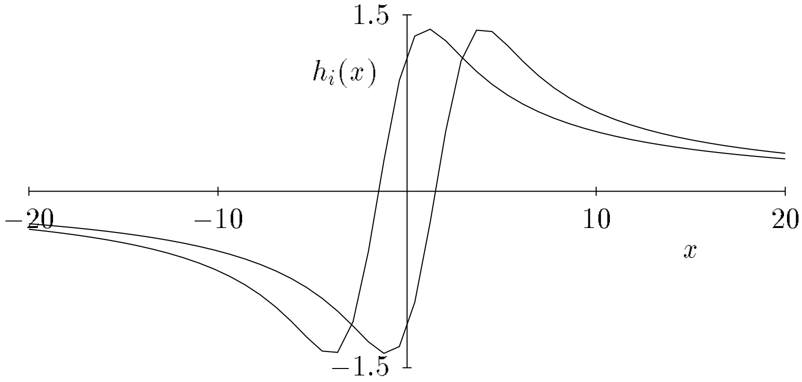

Figure 2: A pair of statistic generating functions $h_i(x),\ i = 1, 2$ used in estimation of the location parameter μ.

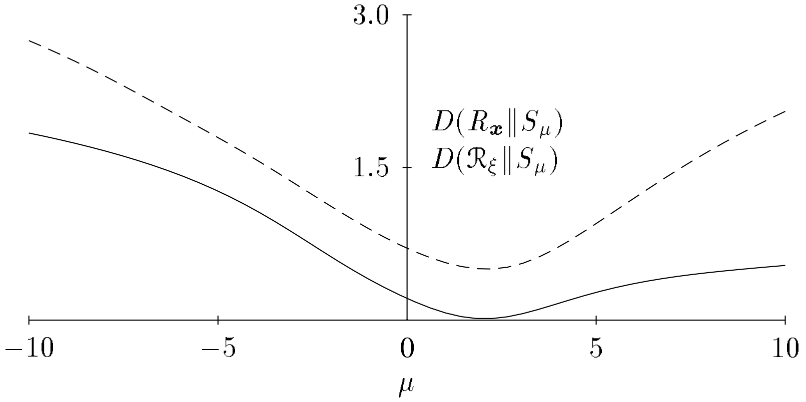

Figure 3: A comparison of the relative entropy, taken as a function of the location parameter μ, for the compressed data (solid curve) and full data (dashed curve).

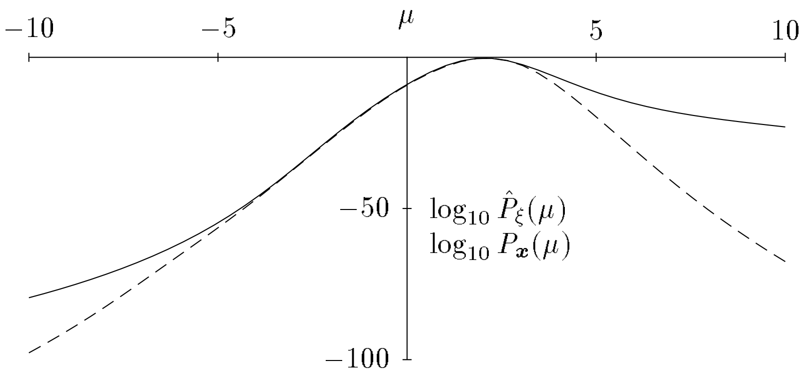

Figure 4: The approximate posterior distribution conditional on the data statistic (solid curve) versus the true posterior distribution conditional on the complete sample (dashed curve) for the location parameter μ.

of the low dimension of the used statistic ($n = 2$). The small amount of information about $\boldsymbol{x}$ resulted also in much heavier tails of $\hat{P}_\xi(\mu)$ compared to $P_{\boldsymbol{x}}(\mu)$.

In the second example the estimated parameter θ was the scale parameter σ. The parameter set $\mathcal{T}$ was formed by 50 values chosen in equal logarithmic distances within the interval $[0.01, 100]$. The prior distribution was uniform again. The statistic generating functions were chosen analogously as above

$$h_i(x) = \log S_{\sigma_i}(x) - \log S_{\sigma_{i+1}}(x),\ i = 1, 2$$

with $\sigma_1 = 0.5$, $\sigma_2 = 5$, $\sigma_3 = 50$ (see Fig. 5).

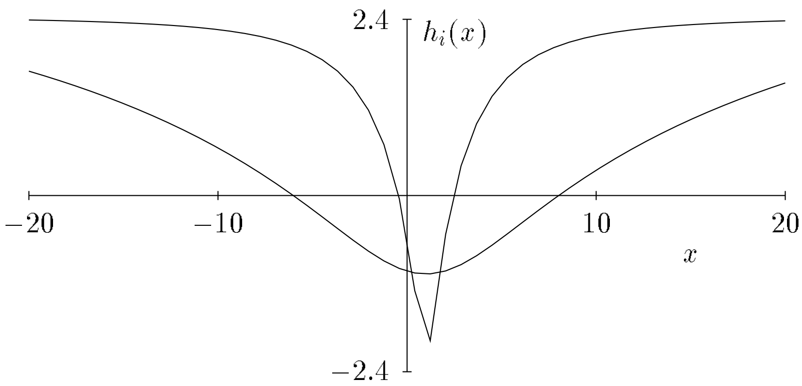

Figure 5: A pair of statistic generating functions $h_i(x),\ i = 1, 2$ used for estimation of the location parameter σ.

The relative entropy functions $D(\mathcal{R}_\xi \| S_\sigma)$, $D(R_{\boldsymbol{x}} \| S_\sigma)$ and the posterior distributions $\hat{P}_\xi(\sigma)$, $P_{\boldsymbol{x}}(\sigma)$ are compared in Fig. 6 and 7, respectively. Note that the results were quite analogous to those achieved for the location parameter.

5 Approximate Estimation under Markov Chain Conditioning

In this section we extend the above scheme of approximate estimation to the case of Markov chain conditioning. The exposition goes along the same lines as for the i.i.d. data.

Bayesian Estimation Revisited

We start again by showing the role of the relative entropy in Bayesian estimation.

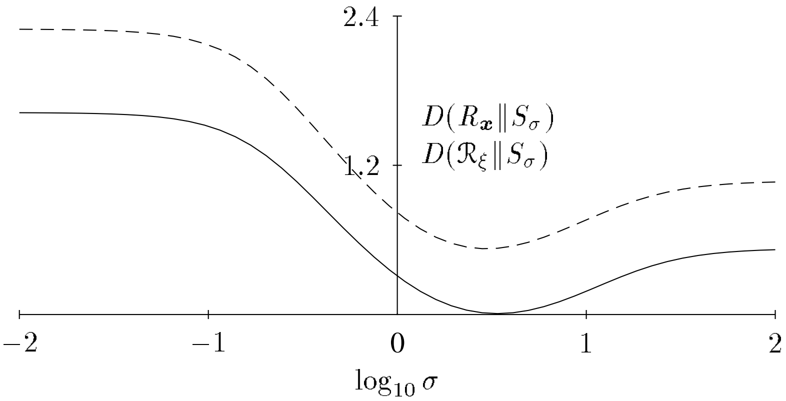

Figure 6: A comparison of the relative entropy, taken as a function of the scale parameter σ, for the compressed data (solid curve) and full data (dashed curve).

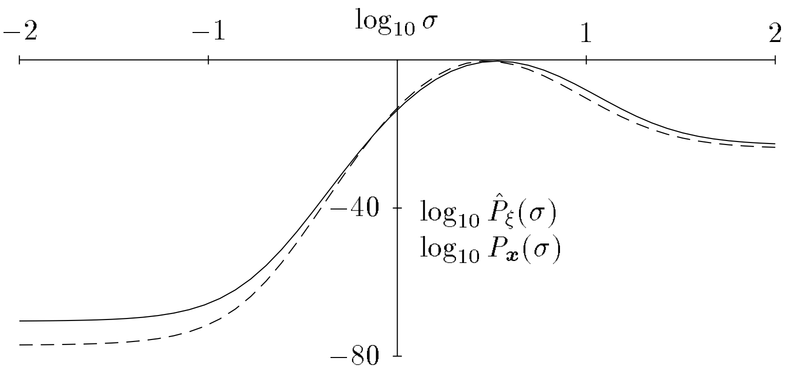

Figure 7: The approximate posterior distribution conditional on the data statistic (solid curve) versus the true posterior distribution conditional on the complete sample (dashed curve) for the scale parameter σ.

Definition 4: *The second-order type of a sequence* $\boldsymbol{x} = (x_1, \ldots, x_{k+1})$ *of elements of* $\mathcal{X}$ *is the distribution* $R_{\boldsymbol{x}}$ *defined by the relative frequencies*

$$R_{\boldsymbol{x}}(\boldsymbol{a}) = \frac{1}{k} \left| \{ i \in \{1, \ldots, k\} : (x_i, x_{i+1}) = \boldsymbol{a} \} \right| , \; \boldsymbol{a} \in \mathcal{X}^2 . \tag{16}$$

For a given sequence of random variables $X_1, X_2, \ldots$ with values in $\mathcal{X}$, the second-order type of the sample $(X_1, \ldots, X_k)$ is called the second-order empirical distribution.

For simplicity, we introduce no special notation for the second-order objects; the meaning should be clear from the context.

Definition 5: *Let* $R(z)$ *be the marginal of the joint distribution* $R(y, z)$*. With some abuse of notation, we define the relative entropy between the probability distributions* $R(y, z)$ *and* $S(y\,|\,z)\, R(z)$ *as*

$$D(R \| S) = \sum_{(y,z) \in \mathcal{X}^2} R(y, z) \, \log \frac{R(y\,|\,z)}{S(y\,|\,z)} . \tag{17}$$

Using the notions of the sequence type and relative entropy, the posterior distribution of the Markov chain parameter can be rearranged as follows.

Proposition 4: *Let* $X_1, X_2, \ldots$ *be a Markov chain with a time-invariant transition probability* $S_\theta(y\,|\,z)$ *such that* $S_\theta(y\,|\,z) > 0$ *for all* $(y, z) \in \mathcal{X}^2$ *and* $\theta \in \mathcal{T}$*. Then the posterior probability distribution* $P_{\boldsymbol{x}}(\theta)$ *of* Θ *conditional on* $\boldsymbol{X} = \boldsymbol{x}$ *can be written as*

$$P_{\boldsymbol{x}}(\theta) \propto P(\theta) \, \exp\{-k \, D(R_{\boldsymbol{x}} \| S_\theta)\} . \tag{18}$$

Proof: The formula follows easily after applying Definitions 4 and 5 to Proposition 2 and straightforward manipulations. ∎

It is not difficult to see that the assumption of Proposition 4 can be weakened by requiring only that the support of S_θ, i.e., the set $\{(y, z) : S_\theta(y\,|\,z) > 0\}$, includes the support of $R_{\boldsymbol{x}}$ for a given $\boldsymbol{x} = (x_1, \ldots, x_{k+1})$.

Data Compression

Suppose that the observed data are compressed using a suitable statistic.

Definition 6: *Given a sample* $\boldsymbol{X} = (X_1, \ldots, X_{k+1})$ *we define a second-order data statistic as the sample average*

$$T(\boldsymbol{X}) = \frac{1}{k} \sum_{j=1}^{k} h(X_j, X_{j+1}) \tag{19}$$

of a vector function $h : \mathcal{X}^2 \to \mathbb{R}^n$ *of* X.

In terms of the second-order empirical distribution, the data statistic T can be regarded as the empirical mean of $h(Y, Z)$, i.e.,

$$T(\boldsymbol{X}) = \sum_{(a,b)\in\mathcal{X}^2} R_{\boldsymbol{X}}(a, b)\, h(a, b)\,.$$

Let all we know about the data sequence $\boldsymbol{x} = (x_1, \ldots, x_{k+1})$ be the vector inequality

$$T(\boldsymbol{x}) \geq \xi$$

where $\xi \in \mathbb{R}^n$. With just partial information about $\boldsymbol{x}$, the type $R_{\boldsymbol{x}}$ of $\boldsymbol{x}$ is also uncertain; one knows only that

$$R_{\boldsymbol{x}} \in \mathcal{R}_\xi = \Big\{ R \in \mathcal{R} : \sum_{(a,b)\in\mathcal{X}^2} R(a, b)\, h(a, b) \geq \xi \Big\} \tag{20}$$

where $\mathcal{R}$ stands now for the set of all probability distributions on $\mathcal{X}^2$.

A Markov chain version of the *large deviation theorem* suggests again a suitable approximation of the probability $\Pr\{\boldsymbol{x} \in \mathcal{X}^{k+1} : R_{\boldsymbol{x}} \in \mathcal{R}_\xi\}$.

Theorem 3: *Let* $X_1, X_2, \ldots$ *form a Markov chain with a time-invariant transition probability* S *such that* $S(y\,|\,z) > 0$ *for all* $(y, z) \in \mathcal{X}^2$. *Let* $\mathcal{C} \subset \mathcal{R}$ *be a nonempty set of probability distributions on* $\mathcal{X}^2$. *Let* $\mathcal{R}_0$ *be the set of all distributions* $R \in \mathcal{R}$ *such that their both marginals coincide, i.e.,*

$$\sum_{z\in\mathcal{X}} R(y, z) = \sum_{z\in\mathcal{X}} R(z, y) \;\; \text{for all} \;\; y \in \mathcal{X}\,.$$

If $\mathcal{C} \cap \mathcal{R}_0$ *is the closure of its interior, then*

$$\lim_{k\to\infty} \frac{1}{k} \log \Pr\{\boldsymbol{x} \in \mathcal{X}^{k+1} : R_{\boldsymbol{x}} \in \mathcal{C}\} = -\min_{R\in\mathcal{C}\cap\mathcal{R}_0} D(R\|S)\,. \tag{21}$$

Proof: The technically simplest way of proving the proposition is to use the counting approach developed by [21, 22] as done e.g. in [23, Theorem 3.1] and [24, Theorem 1]. A more general version of the above proposition was proved in [13, Lemma 2]. ∎

The assumption of strict positivity of S can be removed by adding some regularity assumptions as done in [13, Lemma 2]. This is essential for generalization to higher-order Markov chains.

Approximate Posterior

By definition (20), $\mathcal{R}_\xi$ satisfies the assumption of Theorem 3 unless the interior of $\mathcal{C} \cap \mathcal{R}_0$ is empty. The large deviation theorem leads us then to approximate the posterior distribution as follows

$$\hat{P}_\xi(\theta) \propto P(\theta)\, \exp\Big\{-k\, D(\mathcal{R}_\xi \,\|\, S_\theta)\Big\} \tag{22}$$

where

$$D(\mathcal{R}_\xi \,\|\, S_\theta) = \min_{R \in \mathcal{R}_\xi} D(R \,\|\, S_\theta)\,. \tag{23}$$

Note that the approximate distribution (22) has formally the same structure as the approximation (10) for the i.i.d. data. As a result, quite analogous properties can be shown to hold for (22) too.

Minimum Relative Entropy

The minimum relative entropy (23) is determined by the following theorem.

Theorem 4: *Let $X_1, X_2, \ldots$ form a Markov chain with a time-invariant transition probability S on $\mathcal{X}^2$. Then*

$$D(\mathcal{R}_\xi \,\|\, S) = \max_{\lambda \geq 0} \left[\sum_{i=1}^{n} \lambda_i\, \xi_i - \log \alpha(\lambda) \right] \tag{24}$$

where $\lambda = (\lambda_1, \ldots, \lambda_n)$ and $\alpha(\lambda)$ is the largest eigenvalue of a $|\mathcal{X}| \times |\mathcal{X}|$-matrix $M(\lambda)$ whose (y, z) entry is

$$M_{y,z}(\lambda) = S(y\,|\,z)\, \exp\Big\{ \sum_{i=1}^{n} \lambda_i\, h_i(y, z) \Big\}\,. \tag{25}$$

Proof: This is again a convex programming problem that can be solved using standard tools. An analogous result for $S(y\,|\,z) = \mathrm{const.}$ was proved in [25, 26]. The above general form was shown in [13]. ∎

6 Example: Non-Gaussian Autoregression

To illustrate the basic properties of the approximation (22)–(23), we simulated 100 steps of a Markov chain that appeared as a discretized version of the first-order autoregression with the double-exponential-like distribution

$$S(y\,|\,z) = \mathrm{const.} \times \frac{1}{\sigma}\, \exp\left\{ -\left| \frac{y - \mu z}{\sigma} \right| \right\}$$

with $\mu = 0.99$ and $\sigma = 2$ (see Fig. 8). The data set $\mathcal{X}$ was formed by 50 values equidistantly located within the interval $[-10, 10]$.

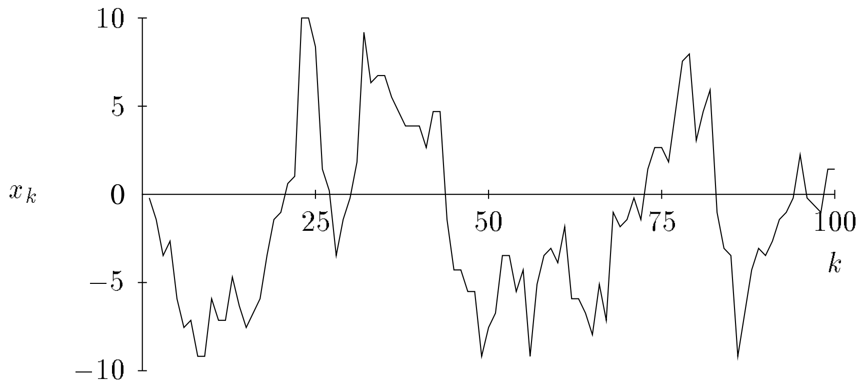

Figure 8: A sequence of 100 values of a Markov chain process modelling a first-order autoregression with a double-exponential noise.

We considered two estimation problems. In the first example, the estimated parameter θ was the regression parameter μ. The parameter set $\mathcal{T}$ was formed by 50 values equidistantly located within the interval $[-5, 5]$. For simplicity, the prior distribution was uniform. The statistic generating functions were chosen as

$$h_i(y, z) = \log S_{\mu_i}(y \,|\, z) - \log S_{\mu_{i+1}}(y \,|\, z), \; i = 1, 2$$

with $\mu_1 = -1.5$, $\mu_2 = 0$, $\mu_3 = 1.5$.

The relative entropy functions $D(\mathcal{R}_\xi \,\|\, S_\mu)$, $D(R_{\boldsymbol{x}} \,\|\, S_\mu)$ and the posterior distributions $\hat{P}_\xi(\mu)$, $P_{\boldsymbol{x}}(\mu)$ are compared in Fig. 9 and 10, respectively.

In the second example the estimated parameter θ was the scale parameter σ. The parameter set $\mathcal{T}$ was formed by 50 values located in equal logarithmic distances within the interval $[0.05, 500]$. The prior distribution was uniform again. The statistic generating functions were defined as

$$h_i(y, z) = \log S_{\sigma_i}(y \,|\, z) - \log S_{\sigma_{i+1}}(y \,|\, z), \; i = 1, 2$$

with $\mu_1 = 0.5$, $\mu_2 = 5$, $\mu_3 = 50$ (see Fig. 5).

The relative entropy functions $D(\mathcal{R}_\xi \,\|\, S_\sigma)$, $D(R_{\boldsymbol{x}} \,\|\, S_\sigma)$ and the posterior distributions $\hat{P}_\xi(\sigma)$, $P_{\boldsymbol{x}}(\sigma)$ are shown in Fig. 11 and 12, respectively. Note that the difference between $\hat{P}_\xi(\sigma)$ and $P_{\boldsymbol{x}}(\sigma)$ was quite negligible in this particular case.

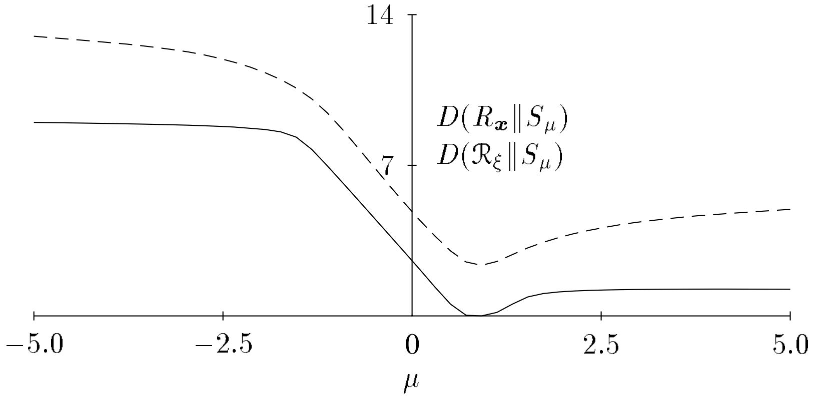

Figure 9: A comparison of the relative entropy, taken as a function of the regression parameter μ, for the compressed data (solid curve) and full data (dashed curve).

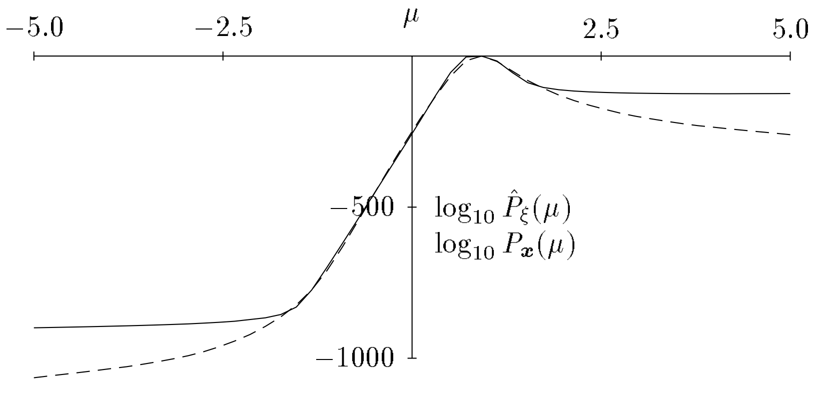

Figure 10: The approximate posterior distribution conditional on the data statistic (solid curve) versus the true posterior distribution conditional on the complete sample (dashed curve) for estimation of the regression parameter μ.

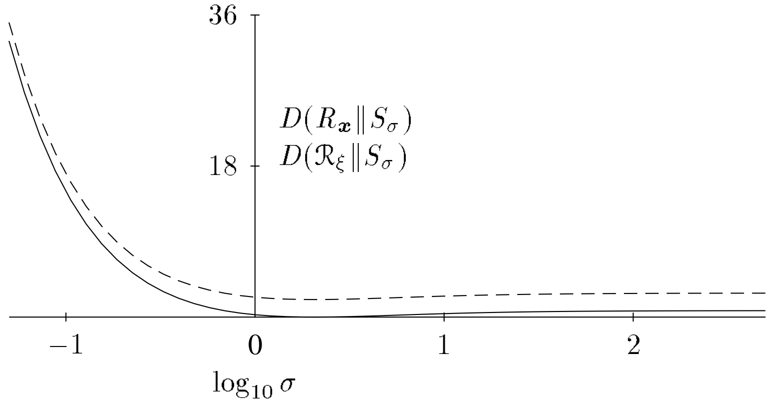

Figure 11: A comparison of the relative entropy, taken as a function of the scale parameter σ, for the compressed data (solid curve) and full data (dashed curve).

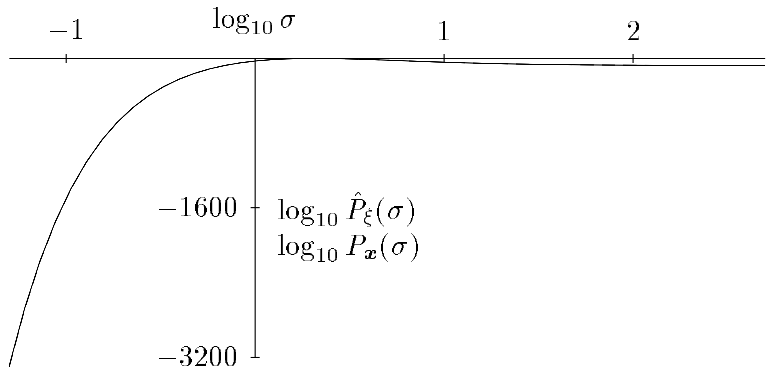

Figure 12: The approximate posterior distribution conditional on the data statistic (solid curve) versus the true posterior distribution conditional on the complete sample (dashed curve) for the scale parameter σ.

7 Concluding Remarks

The paper has outlined a novel scheme of approximate Bayesian estimation with attractive properties supported by the large deviation theory results as well as intuitive geometric arguments (cf. [20]).

Possible Generalizations

It is worth stressing that although the idea of approximation has been developed here only for *finite* data and parameter sets, it works in more general setups too. The extension to the case of continuous parameters is straightforward. So is the case of continuous data which, however, requires more advanced mathematical tools (see e.g. [27, 28]).

Curse of Dimensionality II.

It is only fair to say that the described scheme does not beat the curse of dimensionality completely although it enlarges substantially the class of feasible tasks. The crucial point in the scheme is the evaluation of the relative entropy function $D(\mathcal{R}_\xi \,\|\, S_\theta)$. This may be quite difficult when the cardinality of the sample and parameter sets $\mathcal{X}$ and $\mathcal{T}$ is large.

There are more ways of reducing the complexity of this task that are all consistent with the presented scheme. One is to compute the relative entropy for distributions of suitably transformed data. The relative entropy between such distributions is known to be always smaller or at most equal to the relative entropy between the original distributions (see [29, 30]).

Another possibility is to enlarge suitably the set $\mathcal{R}_\xi$ in the hope to make the evaluation of $D(\mathcal{R}_\xi \,\|\, S_\theta)$ easier. Here the monotonicity relation (13) follows immediately by definition of $D(\mathcal{R}_\xi \,\|\, S_\theta)$ (cf. also the contraction principle in the large deviation theory [27, 28]).

Optimal Inference with Limited Resources

There seems to be no way out of constraints imposed by limited computer resources, without violating the established statistical principles of rational behaviour. This raises some challenging questions. What is the optimal inference under given constraints? Does there exist a well-defined optimum at all?

The scheme outlined in the paper is believed to be "a step in the right direction". Note that it is the relative entropy or Kullback-Leibler distance $D(\mathcal{R}_\xi \,\|\, S_\theta)$, taken as a function of the unknown parameter θ, that becomes the main result of inference, rather than the conditional probability $P_\xi(\theta)$ itself. Yet, owing to the large deviation property, the approximate distribution $\hat{P}_\xi(\theta)$ is known to be asymptotically close to the optimum $P_\xi(\theta)$. The appealing small-sample behaviour of $\hat{P}_\xi(\theta)$ follows from the monotonicity properties (12)–(13) of the relative entropy.

Acknowledgment

The work was supported in part by the grants 102/94/0314 of the Czech Grant Agency and 275109 of the Czech Academy of Sciences.

All simulations reported in the paper were performed using a Lisp-based statistical environment XLISP-STAT [31]. The functions realizing the simulation experiments are available from the authors. The XLISP-STAT itself is available by anonymous *ftp* from `umnstat.stat.umn.edu`.

References

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

[2] 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.

[3] A. F. M. Smith, A. M. Skene, J. E. H. Shaw, J. C. Naylor, and M. Dransfield, "Progress with numerical and graphical methods for practical Bayesian statistics," *The Statistician*, vol. 36, pp. 75–82, 1987.

[4] L. Tierney and J. B. Kadane, "Accurate approximations for posterior moments and marginal densities," *J. Amer. Statist. Assoc.*, vol. 81, no. 393, pp. 82–86, 1986.

[5] A. E. Gelfand and A. F. M. Smith, "Sampling-based approaches to calculating marginal densities," *J. Amer. Statist. Assoc.*, vol. 85, pp. 398–409, 1990.

[6] A. F. M. Smith and A. E. Gelfand, "Bayesian statistics without tears: a sampling-resampling perspective," *Amer. Statist.*, vol. 46, pp. 84–88, 1992.

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

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

[9] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic Press, 1981.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: John Wiley & Sons, 1991.

[11] 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**, 213–244, 1961.

[12] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *Ann. Math. Statist.*, vol. 36, pp. 369–401, 1965.

[13] 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.

[14] L. Györfi, I. Vajda, and E. van der Meulen, "Parameter estimation by projecting on structural statistical models," in *Proceedings of the 5th Prague Symposium on Asymptotic Statistics*, (Prague, Czech Republic), 1994.

[15] I. Vajda and V. Kůs, "Dimensionality reduction by projecting probability distributions on simple families," in *Preprints of the IEEE Workshop on Computer-intensive Methods in Control and Signal Processing*, (Prague, Czech Republic), 1994.

[16] I. Csiszár, "Sanov property, generalized I-projection and a conditional limit theorem," *Ann. Probab.*, vol. 12, no. 3, pp. 768–793, 1984.

[17] I. Csiszár, "An extended maximum entropy principle and a Bayesian justification," in *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, eds.), pp. 83–98, Amsterdam: North Holland-Elsevier Science Publishers, 1985.

[18] R. Kulhavý, "On design of approximate finite-dimensional estimators: the Bayesian view," in *Mutual Impact of Computing Power and Control The-*

ory (K. Warwick and M. Kárný, eds.), pp. 13–39, New York: Plenum Press, 1993.

[19] R. Kulhavý, "Can we preserve the structure of recursive Bayesian estimation in a limited-dimensional implementation?," in *Proceedings of the 10th International Symposium on the Mathematical Theory of Networks and Systems*, (Regensburg, Germany), 1993. To be published by Akademie-Verlag, Berlin.

[20] 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.

[21] P. Whittle, "Some distributions and moment formulae for the Markov chain," *J. Roy. Statist. Soc. Ser. B*, vol. 17, pp. 235–242, 1955.

[22] P. Billingsley, "Statistical methods in Markov chains," *Ann. Math. Statist.*, vol. 32, pp. 12–40, 1961.

[23] L. B. Boza, "Asymptotically optimal tests for finite Markov chains," *Ann. Math. Statist.*, vol. 42, no. 6, pp. 1992–2007, 1971.

[24] S. Natarajan, "Large deviations, hypotheses testing, and source coding for finite Markov chains," *IEEE Trans. Inform. Theory*, vol. IT-31, no. 3, pp. 360–365, 1985.

[25] F. Spitzer, "A variational characterization of finite Markov chains," *Ann. Math. Statist.*, vol. 43, no. 1, pp. 580–583, 1972.

[26] J. Justesen and T. Høholdt, "Maxentropic Markov chains," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 665–667, Jul 1984.

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

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

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

[30] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.

[31] L. Tierney, *Lisp-Stat: An Object-oriented Environment for Statistical Computing and Dynamic Graphics*. New York: John Wiley & Sons, 1990.