A KULLBACK-LEIBLER DISTANCE APPROACH TO SYSTEM IDENTIFICATION

Rudolf Kulhavý

Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, P.O. Box 18, 182 08 Prague, Czech Republic

Abstract: The use of probability in system identification is shown to be equivalent to measuring Kullback-Leibler distance between the actual (empirical) and model distributions of data. When data are not known completely (being compressed, quantized, aggregated, missing etc.), the minimum distance approach can be seen as an asymptotic approximation of probabilistic inference. A class of problems is pointed out where inference via Kullback-Leibler distance brings an attractive, computationally less demanding alternative to maximum likelihood or Bayesian estimation.

Keywords: System identification, parameter estimation, statistical inference, algorithms, asymptotic approximation, large deviations, model approximation, adaptation.

1. INTRODUCTION

The use of abstract mathematical models to describe the behaviour of real systems has a long history. It has become common in fields as different as physics, statistics, engineering, econometrics, or biology. In general, to build a faithful model of a given system requires a deep understanding of physical, chemical, biological, etc. processes underlying the system operation. A complete analytical solution is thus rather "expensive", measured by time spent and human qualification required.

For this reason, much effort has been made to automate the process of building a model based on *data* observed on the system. Fitting a parametric model to data is an attractive solution, especially as the relative cost of measurements and computations is steadily decreasing. This does not mean, of course, that having enough data and computational power, we do not need to care about the system. There are several features that make system identification a difficult and nontrivial task.

1. Data usually behave in a stochastic (unpredictable) manner. This is a trivial observation that calls for the use of statistical methods of inference. The stochastic behaviour of data is usually the main source of uncertainty in system identification.

2. Information contained in data is often considerably reduced before used for estimation. In practice, data available for identification are rarely complete. In recursive estimation, data are compressed so that only the value of a certain data statistic is available. In economic or social applications, data are often aggregated so that only sums or averages over a certain time period are recorded. Frequently some data items are completely missing. Data may also be corrupted by noise or systematic errors.

3. The model class typically does not include the actual system. As usual in science, to make predictions about the performance of system identification, a number of simplifying assumptions have to be postulated. The assumption that the true system belongs to a certain model class is commonly accepted. Strictly speaking, this is never true in reality.

4. The computer resources are always limited. Only a finite and limited amount of computer memory and time is available for implementation of any identification algorithm. When the algorithm is too complex to

be implemented in its theoretically optimal form, it has to be approximated in some way. The use of approximation increases the uncertainty of system identification.

None of known paradigms of inference seems to address all the points. The probability-based inference often fails because of the extreme dimensionality of related computations. The paper suggests and elaborates the use of Kullback-Leibler distance as a promising alternative to the use of probability.

2. PARAMETER ESTIMATION AND PROBABILITY

Throughout the paper, *data* are supposed to be 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 $\mathcal{X}$. As usual, distributions on finite sets are identified with their probability mass functions.

Suppose that $X_1, X_2, \ldots, X_k$ are independent and identically 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 family $\{S_\theta : \theta \in \mathcal{T}\}$. We shall consider two cases — θ being an integer ranging over a finite set or a real parameter ranging over an open interval. If not stated explicitly, the former — discrete case is considered. The objective of parameter estimation is to guess on a proper value of the parameter θ.

2.1 Joint distribution of sample

For a sequence of observations $\mathbf{x} = (x_1, \ldots, x_k)$, the joint probability $S_\theta^k(\mathbf{x})$ can be expressed in the form

$$\begin{aligned} S_\theta^k(\mathbf{x}) &= \prod_{i=1}^{k} S_\theta(x_i) \\ &= \exp\Big\{ \sum_{i=1}^{k} \log S_\theta(x_i) \Big\} \\ &= \exp\Big\{ \sum_{a \in \mathcal{X}} N_{\mathbf{x}}(a)\ \log S_\theta(a) \Big\} \\ &= \exp\Big\{ k \sum_{a \in \mathcal{X}} \frac{N_{\mathbf{x}}(a)}{k}\ \log S_\theta(a) \Big\} \end{aligned}$$

where $N_{\mathbf{x}}(a)$ counts the number of occurrences of $a \in \mathcal{X}$ in the sequence $\mathbf{x}$.

Introducing an *empirical* probability mass function $R_{\mathbf{x}}$ through the relative frequencies

$$R_{\mathbf{x}}(a) = \frac{N_{\mathbf{x}}(a)}{k},\ a \in \mathcal{X} ,$$

we can rewrite $S_\theta^k(\mathbf{x})$ as follows

$$\begin{aligned} S_\theta^k(\mathbf{x}) &= \exp\Big\{ k \sum_{a \in \mathcal{X}} R_{\mathbf{x}}(a)\ \log S_\theta(a) \Big\} \\ &= \exp\Big\{ -k \Big(-\sum_{a \in \mathcal{X}} R_{\mathbf{x}}(a)\ \log R_{\mathbf{x}}(a) \Big) \Big\} \\ &\quad \times \exp\Big\{ -k \Big(\sum_{a \in \mathcal{X}} R_{\mathbf{x}}(a)\ \log \frac{R_{\mathbf{x}}(a)}{S_\theta(a)} \Big) \Big\} . \end{aligned}$$

The expressions in parentheses are well-known quantities in information theory. *Shannon entropy* of a discrete random variable with probability distribution R is given by

$$H(R) = -\sum_{x \in \mathcal{X}} R(x) \log R(x) .$$

Kullback-Leibler distance between two probability distributions R and S is defined by

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

By continuity, we set in the above definitions $\log \frac{0}{q} = 0$, $p \log \frac{p}{0} = \infty$, $0 \log 0 = 0$.

Using the above notions, we can express the probability $S_\theta^k(\mathbf{x})$ as

$$S_\theta^k(\mathbf{x}) = \exp\big\{ -k \big[H(R_{\mathbf{x}}) + D(R_{\mathbf{x}}\|S_\theta) \big] \big\} . \quad (1)$$

2.2 Posterior distribution of parameter

When the unknown parameter θ is interpreted as a random variable Θ, it is natural to describe its uncertainty by means of the *posterior* distribution of Θ *conditional* on $\mathbf{x}$. Given a prior distribution P on $\mathcal{T}$, the posterior distribution $P_{\mathbf{x}}$ is determined by Bayes rule

$$P_{\mathbf{x}}(\theta) \propto P(\theta)\, S_\theta^k(\mathbf{x})$$

where $\propto$ stands for proportionality, i.e., equality up to a normalizing constant.

Substituting (1) for $S_\theta^k(\mathbf{x})$ in Bayes rule gives

$$P_{\mathbf{x}}(\theta) \propto P(\theta)\ \exp\big\{ -k\, D(R_{\mathbf{x}}\|S_\theta) \big\} \quad (2)$$

since the entropy $H(R_{\mathbf{x}})$ does not depend on θ.

3. KULLBACK-LEIBLER DISTANCE

Kullback-Leibler distance was introduced by Kullback and Leibler (1951). It is also known as relative entropy, cross entropy, informational divergence, or information for discrimination. It has been studied in detail by many authors, including Sanov (1957), Csiszár (1967), Čencov (1972) and Amari (1985).

This section lists some of remarkable properties of Kullback-Leibler distance. Much of the material now belongs to folklore, with many results going back to Kullback and Leibler (1951) and Kullback (1959). From modern literature, an excellent textbook by Cover and Thomas (1991) is recommended to the interested reader.

3.1 Nonnegativity

Kullback-Leibler distance of any two probability mass functions R and S is nonnegative

$$D(R\|S) \geq 0 \tag{3}$$

with equality if and only if $R(x) = S(x)$ for all $x \in \mathcal{X}$. For proof see Kullback and Leibler (1951, Lemma 3.1) or Cover and Thomas (1991, Theorem 2.6.3).

This fundamental property makes it possible to think of $D(R\|S)$ as a "distance" between the distributions R and S. It is not a true distance, however, since it is not symmetric and does not satisfy the triangle inequality.

3.2 Convexity

The Kullback-Leibler distance $D(R\|S)$ is convex in the pair (R, S), i.e., if (R_1, S_1) and (R_2, S_2) are two pairs of probability mass functions, then

$$\begin{aligned} &D(\lambda R_1 + (1-\lambda)R_2 \,\|\, \lambda S_1 + (1-\lambda)S_2) \\ &\leq \lambda D(R_1\|S_1) + (1-\lambda)D(R_2\|S_2)\,. \end{aligned} \tag{4}$$

For proof see e.g. Cover and Thomas (1991, Theorem 2.7.2).

3.3 Locally Euclidean behaviour

Let θ be a real parameter ranging over an open interval. Suppose that the probabilities $S_\theta(x)$ are continuously differentiable functions of θ. Then Kullback-Leibler distance of two neighbouring distributions S_θ and $S_{\theta'}$ for θ' near θ can be approximated — neglecting higher than second-order terms — by a quadratic form

$$D(S_\theta\|S_{\theta'}) \approx \frac{1}{2}\, I(\theta)\,(\theta - \theta')^2\,, \tag{5}$$

with the kernel

$$I(\theta) = \sum_{a\in\mathcal{X}} S_\theta(a)\,\Big(\frac{\partial}{\partial\theta}\log S_\theta(a)\Big)^2 \tag{6}$$

being Fisher information. Note that $I(\theta)$ was introduced by Fisher (1925) as a measure of information contained in one observation from S_θ for estimating θ. For derivation of (5) see e.g. Kullback (1959, Section 2.6).

3.4 Asymptotically Euclidean behaviour

An "empirical" version of (5) holds as well. Suppose $\mathbf{x} = (x_1, \ldots, x_k)$ is a sequence of observations drawn from a distribution S_θ, $\theta \in \mathcal{T}$. Let $\hat\theta$ minimize Kullback-Leibler distance $D(R_\mathbf{x}\|S_\theta)$. Then for k large enough and θ close to $\hat\theta$ it holds approximately

$$D(R_\mathbf{x}\|S_\theta) \approx D(R_\mathbf{x}\|S_{\hat\theta}) + \frac{1}{2}\, I(\hat\theta)\,(\theta - \hat\theta)^2 \tag{7}$$

where $I(\hat\theta)$ is Fisher information (6) at $\hat\theta$. This is just another form of the well-known result on the asymptotic normality of likelihood and its derivation goes along the same line — see e.g. DeGroot (1970, Section 10.9).

4. PARAMETER ESTIMATION AND KULLBACK-LEIBLER DISTANCE

The purpose of this section is to demonstrate that the use of Kullback-Leibler distance for inference is consistent with common statistical approaches.

4.1 Maximum likelihood estimation

When making inference about the unknown θ, we think of the probability mass function $S_\theta^k(\mathbf{x})$ as a function of θ for given $\mathbf{x}$. This function is known as *likelihood*

$$L_\mathbf{x}(\theta) = S_\theta^k(\mathbf{x})\,.$$

A classical statistical method of estimating θ chooses the value of θ that maximizes the likelihood

$$\max_\theta\, L_\mathbf{x}(\theta)\,.$$

Equation (1) exhibits a close relationship between Kullback-Leibler distance and likelihood, namely,

$$D(R_\mathbf{x}\|S_\theta) = -H(R_\mathbf{x}) - \frac{1}{k}\log L_\mathbf{x}(\theta)\,.$$

Kullback-Leibler distance is thus equal, up to an additive constant, to a negative normalized log-likelihood. Since the entropy term does not depend on θ and the function $-\frac{1}{k}\log(\cdot)$ is strictly monotonous, maximum likelihood estimation is equivalent to *minimum Kullback-Leibler distance* estimation

$$\min_\theta\, D(R_\mathbf{x}\|S_\theta)\,. \tag{8}$$

The minimum distance formulation (8) is intuitively appealing. If the empirical probability mass function $R_\mathbf{x}$ converges (entry-wise) to a sampling probability mass function S_{θ_0} for a certain $\theta_0 \in \mathcal{T}$, then also

$$D(R_\mathbf{x}\|S_\theta) \to D(S_{\theta_0}\|S_\theta)$$

for each $\theta \in \mathcal{T}$. By information inequality (3), the minimum distance $\min_\theta D(S_{\theta_0}\|S_\theta)$ is achieved at the point $S_\theta = S_{\theta_0}$.

Example 1. Let X take on just three different values — 1, 2, 3. Probability mass functions on $\mathcal{X} = \{1, 2, 3\}$ can be identified with the probability vectors

$$\big[\,\Pr\{X=1\}, \Pr\{X=2\}, \Pr\{X=3\}\,\big]\,.$$

These vectors can be envisaged as points of a probability simplex.

This view allows us to think of maximum likelihood estimation as searching for the probability vector S_θ, $\theta \in \mathcal{T}$ that is nearest, in terms of Kullback-Leibler distance, to the probability vector $R_\mathbf{x}$ (cf. Fig. 1).

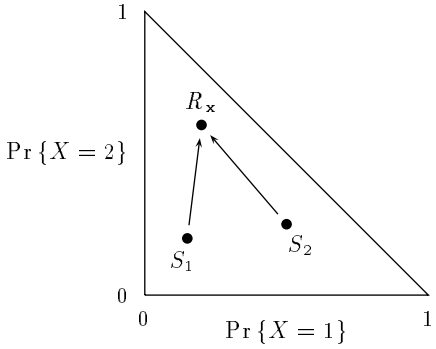

Fig. 1. Sampling and empirical distributions of data can be regarded as points of the same probability simplex.

Example 2. A particular case of the previous example occurs when the sampling distribution of X is *binomial* with the probability mass function

$$S_\theta(x) = \binom{2}{x} \theta^x (1-\theta)^{2-x}, \quad x = 0, 1, 2 .$$

Figure 2 illustrates on simulation data that the empirical distribution $R_{\mathbf{x}}$ gets close, for a sufficiently long sample $\mathbf{x}$, to the true distribution S_{θ_0}.

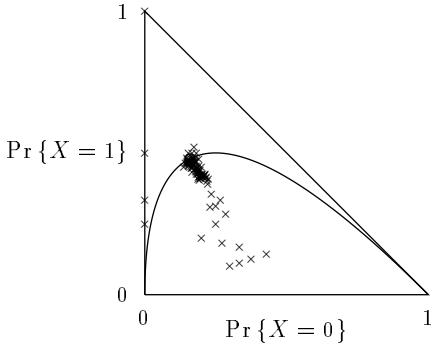

Fig. 2. Empirical distributions $R_{\mathbf{x}}$ (marked by crosses) for samples $\mathbf{x} = (x_1, \ldots, x_k)$ drawn from binomial distribution approach the family of sampling distributions (solid curve) near the true point $\theta_0 = 0.6$ (circle) as k becomes large enough.

4.2 Maximum a posteriori probability estimate

A natural Bayesian counterpart of the maximum likelihood estimate is the estimate maximizing the posterior probability mass (or density) function

$$\max_\theta \; P_{\mathbf{x}}(\theta) .$$

Because of (2), maximum a posteriori probability estimation is equivalent to

$$\min_\theta \left[D(R_{\mathbf{x}} \| S_\theta) - \frac{1}{k} \log P(\theta) \right] . \tag{9}$$

Compared with the form (8) of maximum likelihood estimation, Kullback-Leibler distance is here modified by a normalized log-prior. The effect of this term is important for short samples. When k gets large enough, the term becomes typically negligible.

4.3 Conjugate prior

It is convenient if the prior distribution on $\mathcal{T}$ is chosen from a *conjugate* family, i.e., a family closed under conditioning on observed data (see Robert (1989, Section 3.2)). The expression (2) makes it possible to give a general form of conjugate priors

$$P(\theta) \propto \exp\{-k_0 D(R_0 \| S_\theta)\} \tag{10}$$

where R_0 stands for a "prior" distribution of X and k_0 counts the number of (actual or fictitious) observations R_0 is built on. Note that k_0 is nonnegative but not necessarily integer. By choosing a particular value k_0 we put more or less weight (prior belief) on R_0.

It is easy to verify that given a conjugate prior (10) the posterior distribution is of the same form

$$P_{\mathbf{x}}(\theta) \propto \exp\{-(k_0 + k) D(R_k \| S_\theta)\} \tag{11}$$

where

$$R_k = \frac{k_0}{k_0 + k} R_0 + \frac{k}{k_0 + k} R_{\mathbf{x}} \tag{12}$$

is now a *mixture* (convex combination) of the prior distribution R_0 and the empirical distribution $R_{\mathbf{x}}$. The weight on R_0 tends to zero as $k \to \infty$.

Therefore, when the conjugate prior (10) is chosen, maximum a posteriori probability estimation (9) admits a more compact expression

$$\min_\theta \; D(R_k \| S_\theta) . \tag{13}$$

4.4 Posterior distribution

To sum up, maximum likelihood or maximum a posteriori probability estimation is looking for a point (not necessarily unique) $S_{\hat{\theta}}$ that minimizes Kullback-Leibler distance from the empirical distribution (possibly modified by prior information). Compared with the point estimation, Bayesian estimation is by far more ambitious. It aims at evaluating Kullback-Leibler distances between *all* sampling distributions $S_\theta, \theta \in \mathcal{T}$ and the empirical distribution

$$D(R_{\mathbf{x}} \| S_\theta), \theta \in \mathcal{T} \tag{14}$$

or the modified empirical distribution (12)

$$D(R_k \| S_\theta), \theta \in \mathcal{T} \tag{15}$$

provided the conjugate prior (10) is used. The posterior distribution $P_{\mathbf{x}}$ is related to the above functions through simple transformations (2) and (11), respectively.

4.5 Hypothesis testing

By Neyman-Pearson lemma (Cover and Thomas 1991, Theorem 12.7.1), the optimum test to decide between two hypotheses — sampling distributions S_1 and S_2 has the form

$$\frac{S_1^k(\mathbf{x})}{S_2^k(\mathbf{x})} > \kappa \tag{16}$$

where κ is a certain threshold value. That means that for a sequence of observations $\mathbf{x}$ that satisfy (16) the hypothesis S_1 is accepted, otherwise S_2 is accepted.

Substituting (1) for S_1^k and S_2^k in (16), one easily finds that (16) is equivalent to

$$D(R_{\mathbf{x}}\|S_2) - D(R_{\mathbf{x}}\|S_1) > \frac{1}{k}\,\log\kappa\,. \qquad (17)$$

In other words, the likelihood ratio test (16) is equivalent to the comparison of Kullback-Leibler distances of sampling distributions S_1 and S_2 from the empirical distribution $R_{\mathbf{x}}$.

A Bayesian counterpart based on the posterior ratio test

$$\frac{P_{\mathbf{x}}(1)}{P_{\mathbf{x}}(2)} > \kappa \qquad (18)$$

results in an analogous form of test

$$D(R_k\|S_2) - D(R_k\|S_1) > \frac{1}{k}\,\log\kappa \qquad (19)$$

with the empirical distribution $R_{\mathbf{x}}$ being substituted by the mixture R_k provided the prior is chosen in the conjugate form (10).

4.6 Minimum distance view

The use of Kullback-Leibler distance can be seen as a kind of the minimum distance method. But while for instance the least squares method induces Euclidean distance directly in the data space

$$\min_{\theta} \sum_{i=1}^{k} \big(x_i - \theta\big)^2,$$

the minimum Kullback-Leibler distance method

$$\min_{\theta} D(R_{\mathbf{x}}\|S_\theta)$$

works on *probability distributions* of observed data. The latter view is considerably more general (see Wolfowitz (1957) and Vajda (1989) for more background). It is worth emphasizing that its generality is in the freedom to choose quite freely the model distribution of data. In contrast to this freedom, a single definition of "distance" is used for all models. Remember that Kullback-Leibler distance has not been chosen *ad hoc* but has appeared quite naturally from analysis of the joint distribution of sample.

Example 3. A sequence of 400 samples were simulated according to the model $x_k = \theta + e_k$ with $\theta = 1$ and e_k being a discrete Cauchy-like distribution (see Fig. 3).

The empirical distribution of the sample is compared with Cauchy-like and Gauss-like sampling distributions in Fig. 4 and 5. Notice the heavy tails of the empirical distribution caused by the large amount of "outliers" in data.

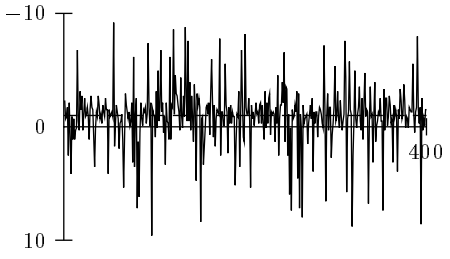

Fig. 3. A sample of 400 data with Cauchy-like distribution around $\theta = 1$.

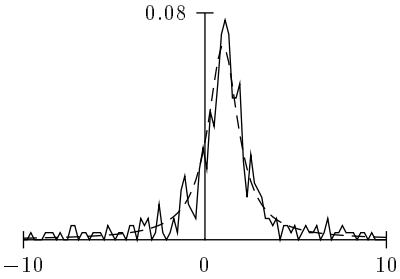

Fig. 4. Empirical distribution of the sample plotted in Fig. 3 against a discrete Cauchy-like sampling distribution for $\theta = 1$. A good fit is achieved.

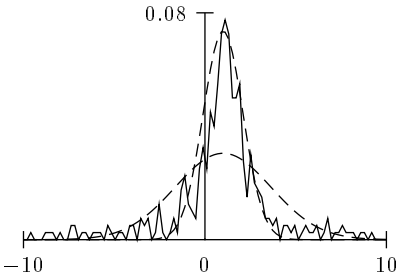

Fig. 5. Empirical distribution of the sample plotted in Fig. 3 against discrete Gauss-like sampling distributions for $\theta = 1$ but different variances. It is difficult to achieve a good agreement for all data values.

5. ASYMPTOTIC APPROXIMATION VIA LARGE DEVIATIONS

Often the empirical distribution of sample is not at disposal. Instead a set $\mathcal{R}$ of distributions R on $\mathcal{X}$ that contains the true empirical distribution $R_{\mathbf{x}}$ is known. Some typical cases are considered in Section 6. The present section gives a general idea of dealing with the lack of information about observed data.

5.1 True inference with partial information

The probability of observing a sample the empirical distribution $R_{\mathbf{x}}$ of which belongs to $\mathcal{R}$ is

$$S_\theta^k(\{\mathbf{x} : R_{\mathbf{x}} \in \mathcal{R}\}) = \sum_{R\in\mathcal{R}} S_\theta^k(\{\mathbf{x} : R_{\mathbf{x}} = R\})$$

where

$$S_\theta^k(\{\mathbf{x} : R_{\mathbf{x}} = R\}) = \sum_{\mathbf{x}\,:\,R_{\mathbf{x}}=R} S_\theta^k(\mathbf{x})\,.$$

It follows from (1) that $S_\theta^k(\mathbf{x})$ is *constant* for all samples with the same empirical distribution. Thus, we

can write

$$S_\theta^k(\{\mathbf{x} : R_\mathbf{x} = R\}) = \big|\{\mathbf{x} : R_\mathbf{x} = R\}\big| \times \exp\big\{-k\big[H(R) + D(R\|S_\theta)\big]\big\}$$

where $|\{\mathbf{x} : R_\mathbf{x} = R\}|$ denotes the number of all sequences $\mathbf{x} \in \mathcal{X}^k$ the empirical distribution $R_\mathbf{x}$ of which is equal to a given distribution R.

It comes as no surprise that the exact evaluation of the probability $S_\theta^k(\{\mathbf{x} : R_\mathbf{x} \in \mathcal{R}\})$ is typically infeasible.

5.2 Probability of large deviations

The probability $S_\theta^k(\{\mathbf{x} : R_\mathbf{x} \in \mathcal{R}\})$ can be approximated in the following asymptotic sense. Two sequences f_k, g_k are said *equal to the first order in the exponent*, $f_k \doteq g_k$, if (Cover and Thomas 1991, Section 3.3)

$$\lim_{k\to\infty} \frac{1}{k} \log \frac{f_k}{g_k} = 0 .$$

By combinatorial arguments, the number of sequences with the same empirical distribution is found (Csiszár and Körner 1981, Lemma 2.3)

$$\big|\{\mathbf{x} : R_\mathbf{x} = R\}\big| \doteq \exp\big\{k\, H(R)\big\}$$

and, consequently, the corresponding probability is (Csiszár and Körner 1981, Lemma 2.6)

$$S_\theta^k(\{\mathbf{x} : R_\mathbf{x} = R\}) \doteq \exp\big\{-k\, D(R\|S_\theta)\big\} .$$

The probability $S_\theta^k(\{\mathbf{x} : R_\mathbf{x} \in \mathcal{R}\})$ is thus

$$S_\theta^k(\{\mathbf{x} : R_\mathbf{x} \in \mathcal{R}\}) \doteq \sum_{R\in\mathcal{R}} \exp\big\{-k\, D(R\|S_\theta)\big\} .$$

Note that the number of different empirical distributions of sequences $\mathbf{x} \in \mathcal{X}^k$ is less than $(k+1)^{|\mathcal{X}|}$ since for every $a \in \mathcal{X}$, $N_\mathbf{x}(a)$ can take (at most) on values $0, 1, \ldots, k$ (Csiszár and Körner 1981, Lemma 2.2). The number of terms in the above sum is thus "only" polynomial in the length of the sequences. Therefore, the sum is (under additional topological assumptions on the set $\mathcal{R}$) equal to the first order in the exponent to its largest term

$$\begin{aligned} S_\theta^k(\{\mathbf{x} : R_\mathbf{x} \in \mathcal{R}\}) &\doteq \max_{R\in\mathcal{R}} \exp\big\{-k\, D(R\|S_\theta)\big\} \\ &\doteq \exp\big\{-k \min_{R\in\mathcal{R}} D(R\|S_\theta)\big\} . \end{aligned}$$

More precisely, if $\mathcal{R}$ is the closure of its interior, then

$$\lim_{k\to\infty} \frac{1}{k} \log S_\theta^k \{\mathbf{x} : R_\mathbf{x} \in \mathcal{R}\} = -D(\mathcal{R}\|S_\theta) \quad (20)$$

with $D(\mathcal{R}\|S_\theta)$ defined as follows

$$D(\mathcal{R}\|S_\theta) = \min_{R\in\mathcal{R}} D(R\|S) . \quad (21)$$

The result is known as Sanov's *large deviation theorem* (Sanov 1957, Theorem 2). The derivation outlined informally above is due to (Csiszár and Körner 1981), cf. (Csiszár *et al.* 1987, Theorem 1) and (Cover and Thomas 1991, Theorem 12.4.1).

The probability that the empirical distribution belongs to a set that does not contain the true sampling distribution is known (by the law of large numbers) to converge to zero. The large deviation theorem refines this statement, showing that the probability converges to zero *exponentially fast*, with the *rate* given by Kullback-Leibler distance between the sampling distribution and a given set.

5.3 Minimum Kullback-Leibler distance

When the set $\mathcal{R}$ is more specific, more can be said about the minimum Kullback-Leibler distance in (21). Often $\mathcal{R}$ is supposed to be bounded by hyperplanes

$$\mathcal{R}_\xi = \left\{ R : \sum_{a\in\mathcal{X}} R(a)\, h_j(a) \geq \xi_j ,\ j = 1, \ldots, n \right\}$$

where h_1,...,h_n are given real functions on $\mathcal{X}$.

Then the minimum Kullback-Leibler distance is

$$D(\mathcal{R}_\xi\|S_\theta) = \max_{\lambda\geq 0} \left[\sum_{j=1}^{n} \lambda_j\, \xi_j - \log N(\lambda) \right] \quad (22)$$

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

$$v_x(\lambda) = S(x)\, \exp\left\{ \sum_{j=1}^{n} \lambda_j\, h_j(x) \right\} . \quad (23)$$

The maximization (22) is basically a dual problem to the convex programming problem (21). For proof see Kullback (1959, Theorem 2.1) and Csiszár (1984, Theorems 2, 3).

Example 4. Suppose a ternary signal with $\mathcal{X} = \{1, 2, 3\}$ again. Let the only information we have about the empirical distribution be that

$$\sum_{a=1}^{3} R_\mathbf{x}(a)\, a \leq 1.5 .$$

To put it other way, $R_\mathbf{x} \in \mathcal{R}_\xi$ with $h(x) = -x$ and $\xi = 1.5$. Further, consider an exponential family $S_\theta(x) \propto S_1(x)^\theta\, S_2(x)^{1-\theta}$ with distributions S_1, S_2 given and $\theta \in (-5, 5)$ (see Fig. 6).

The minimum Kullback-Leibler distance $D(\mathcal{R}_\xi\|S_\theta)$ is compared in Fig. 7 with $D(R\|S_\theta)$ for one point R in $\mathcal{R}_\xi$. Note that by definition (21), it holds

$$D(\mathcal{R}\|S_\theta) \leq D(R\|S_\theta)$$

for every θ and every $R \in \mathcal{R}_\xi$.

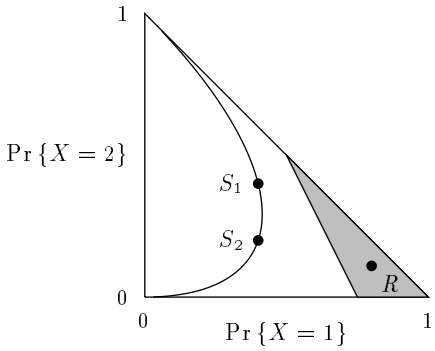

Fig. 6. An example of estimation with partial information: $\{S_\theta\}$ is an exponential family (solid curve), $\mathcal{R}_\xi$ is composed of distributions R with bounded mean (shaded area).

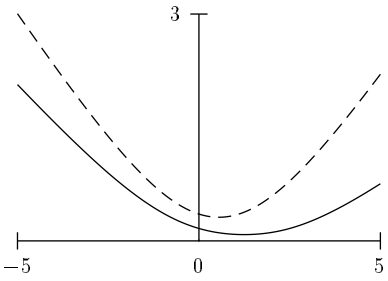

Fig. 7. Minimum Kullback-Leibler distance $D(\mathcal{R}_\xi \| S_\theta)$ (solid line) compared with Kullback-Leibler distance $D(R\|S_\theta)$ (dashed line) for $\{S_\theta\}$, $\mathcal{R}_\xi$ and R shown in Fig. 6.

5.4 Minimum Kullback-Leibler distance distribution

One can be interested in the distribution $R^* \in \mathcal{R}$ minimizing Kullback-Leibler distance from S_θ. The distribution has the form

$$R^*(x) \propto S_\theta(x)\, \exp\left\{\sum_{j=1}^{n} \lambda_j^*\, h_j(x)\right\} \qquad (24)$$

where the coefficients $\lambda_1^*, \ldots, \lambda_n^*$ are chosen so to satisfy

$$\sum_{a\in\mathcal{X}} R^*(a)\, h_j(a) \geq \xi_j\,,\ j = 1, \ldots, n\,. \qquad (25)$$

The form (24) follows by using Lagrange multipliers to solve (21), see Kullback (1959, Theorem 2.1) and Cover and Thomas (1991, Section 12.5).

Example 5. Consider a variant of Example 4 with a single sampling distribution S given. The point $R^* \in \mathcal{R}_\xi$ that minimizes Kullback-Leibler distance from S lies in the intersection of the exponential family

$$S_\lambda(x) \propto S(x)\, \exp\{\lambda x\},\ \lambda \in \mathbb{R}$$

with the region $\mathcal{R}_\xi$ (see Fig. 8).

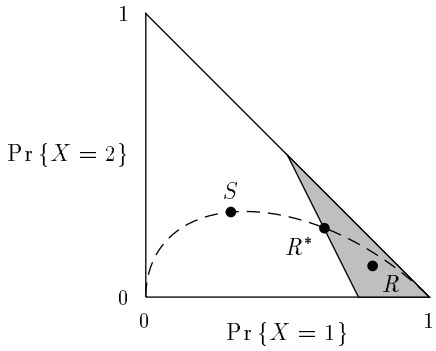

Fig. 8. The minimum Kullback-Leibler distance projection of S onto $\mathcal{R}_\xi$

6. COPING WITH "BAD" DATA

This section shows how the large deviation theorem (20) can be used to handle various cases of incomplete data.

6.1 Compressed data

Suppose that data are compressed by means of a statistic that takes the form of a sample average

$$T(\mathbf{x}) = \frac{1}{k}\sum_{i=1}^{k} h(x_i)$$

where h is a given vector function on $\mathcal{X}$. It is easy to see that the last expression can be rewritten with the help of the empirical distribution $R_\mathbf{x}$ as an empirical mean of the random vector $h(X)$

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

Let the only information about the sample $\mathbf{x}$ be that $T(\mathbf{x}) \geq \xi$. Then by the large deviation theorem (20) the probability of observing such a sample is approximately

$$S_\theta^k(\{\mathbf{x} : T(\mathbf{x}) \geq \xi\}) \doteq \exp\{-k\, D(\mathcal{R}_\xi \| S_\theta)\}\,. \quad (26)$$

The posterior distribution conditional on the statistic value can thus be approximated by

$$\hat{P}_\xi(\theta) \propto P(\theta)\, \exp\{-k D(\mathcal{R}_\xi \| S_\theta)\}\,. \qquad (27)$$

For more discussion on the choice of T see Kulhavý (1993*a*) and Kulhavý (1993*b*).

6.2 Building of prior

Suppose that prior information about the distribution R of X is available in the form

$$T(R) = \sum_{a\in\mathcal{X}} R(a)\, h(a)$$

where h is a given vector function on $\mathcal{X}$. If the distribution R is given the meaning of an empirical distribution $R_\mathbf{x}$ for a sample $\mathbf{x}$ of k_0 data, then the probability

of observing a sample that satisfies $T(R_{\mathbf{x}}) \geq \xi$ is approximately

$$S_\theta^{k_0}(\{\mathbf{x} : T(R_{\mathbf{x}}) \geq \xi\}) \doteq \exp\{-k_0\, D(\mathcal{R}_\xi \| S_\theta)\,.$$

The posterior distribution conditional on this knowledge can be taken as a prior distribution for further estimation

$$P(\theta) = \hat{P}_\xi(\theta) \propto P_0(\theta)\, \exp\{-k_0\, D(\mathcal{R}_\xi \| S_\theta)\,. \quad (28)$$

Here P_0 stands for the initial prior.

Note that the prior (28) depends on the model family $\{S_\theta\}$. When the model changes, the prior changes too. It is because the prior information of the form $T(R_{\mathbf{x}}) \geq \xi$ speaks of data, not the parameter.

Example 6. Suppose data are generated according to the model $x_k = \theta + e_k$ with $\theta = 1$ and e_k being a discrete Cauchy-like distribution. Let prior information about X be that its first two moments are 2 and 12, respectively. Then the prior distribution (28) with $k_0 = 1$, $h_1(x) = x$, $h_2(x) = x^2$ and $P_0(\theta) \propto 1$ gets the shape shown in Fig. 9.

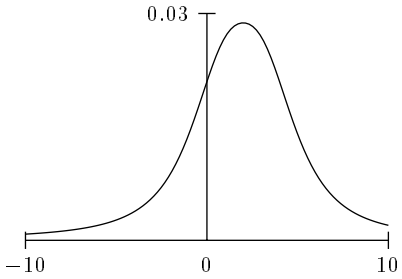

Fig. 9. Prior distribution $P(\theta)$ of the location parameter of a Cauchy-like distribution given information that the mean and variance of data are 2 and 8, respectively.

6.3 Quantized data

Sometimes observed data are known only with accuracy up to a subset of $\mathcal{X}$. More precisely, when $\mathcal{X}$ is partitioned into a system of disjoint subsets $\mathcal{X}_j$, $j = 1, \ldots, n$, observations are known only to belong to a certain subset $\mathcal{X}_j$.

The problem can be regarded as a special case of estimation with compressed data. In fact, when the functions $h_j(x)$ are chosen as follows

$$h_j(x) = \begin{cases} 1, & \text{if } x \in \mathcal{X}_j\,, \\ 0, & \text{if } x \notin \mathcal{X}_j\,, \end{cases}$$

the statistic $T(\mathbf{x})$ gives the relative frequency of observations in particular subsets $\mathcal{X}_j$, $j = 1, \ldots, n$

$$T_j(\mathbf{X}) = \frac{1}{k}\sum_{i=1}^{k} h_j(x_i) = \frac{1}{k}\big|\{x_1, \ldots, x_k\} \cap \mathcal{X}_j\big|\,.$$

With the above choice of h, the approximations (26) and (27) apply to this case without change.

6.4 Aggregated data

In some applications, "microscopic" data are aggregated over longer time periods (hours, days, weeks) so that only the resulting "macroscopic" data become available for identification. Aggregation compresses data neighbouring in time while quantization gets together data close in level.

The sample $\mathbf{x}$ can thus be seen as a juxtaposition of m subsamples $\mathbf{x}_\alpha$, $\alpha = 1, \ldots, m$. Because particular observations are independent, the probability of the whole sample is

$$S_\theta^k(\mathbf{x}) = \prod_{\alpha=1}^{m} S_\theta^{k_\alpha}(\mathbf{x}_\alpha)$$

where $k_1, \ldots, k_m$ denotes the lengths of subsamples. Clearly, $k_1 + \cdots + k_m = k$.

Let information available about subsamples be in the form of statistics

$$T(\mathbf{x}_\alpha) = \frac{1}{k_\alpha}\sum_{i=1}^{k_\alpha} h(x_{\alpha,i}),\ \alpha = 1, \ldots, m\,.$$

A simple example of aggregation is the use of a sample average of data through $h(x) = x$.

Provided all we know about the sample $\mathbf{x}$ is that $T(\mathbf{x}_\alpha) \geq \xi_\alpha$ for $\alpha = 1, \ldots, m$, the probability of observing such a sample is approximately

$$S_\theta^k(\{\mathbf{x} : T(\mathbf{x}_\alpha) \geq \xi_\alpha\,, \alpha = 1, \ldots, m\}) \doteq \exp\Big\{-k \sum_{\alpha=1}^{m} \frac{k_\alpha}{k}\, D(\mathcal{R}_{\xi_\alpha} \| S_\theta)\Big\}$$

and the posterior distribution conditional on $\xi = (\xi_1, \ldots, \xi_m)$ is approximated by

$$\hat{P}_\xi(\theta) \propto P(\theta)\, \exp\Big\{-k \sum_{\alpha=1}^{m} \frac{k_\alpha}{k}\, D(\mathcal{R}_{\xi_\alpha} \| S_\theta)\Big\}\,.$$

6.5 Missing data

Consider another common situation when some observations are completely missing. Since particular observations are independent, the sample $\mathbf{x}$ can be regarded as a juxtaposition of the observed and unobserved subsamples $\mathbf{x}_a$ and $\mathbf{x}_b$ of length k_a and $k_b = k - k_a$, respectively. Let the empirical distribution $R_{\mathbf{x}_a}$ of the observed sample be the only information available.

The probability of the observed part is approximately

$$S_\theta^{k_a}(\{\tilde{\mathbf{x}}_a : R_{\tilde{\mathbf{x}}_a} = R_{\mathbf{x}_a}\}) \doteq \exp\{-k_a\, D(R_{\mathbf{x}_a} \| S_\theta)\}$$

while the probability of the unobserved part is

$$S_\theta^{k_b}(\{\mathbf{x}_b\}) \doteq \exp\{-k_b \min_{R_{\mathbf{x}_b}} D(R_{\mathbf{x}_b} \| S_\theta)\} = 1\,.$$

Since the probability of the whole sample is

$$S_\theta^k(\mathbf{x}) = S_\theta^{k_a}(\mathbf{x}_a)\, S_\theta^{k_b}(\mathbf{x}_b)\,,$$

we have as a result

$$S_\theta^k(\{\tilde{\mathbf{x}} : R_{\tilde{\mathbf{x}}_a} = R_{\mathbf{x}_a}\}) \doteq \exp\{-k_a\, D(R_{\mathbf{x}_a} \| S_\theta)\}$$

and

$$\hat{P}_{R_{\mathbf{x}_a}}(\theta) \propto P(\theta)\, \exp\{-k_a D(R_{\mathbf{x}_a} \| S_\theta)\}$$

which agrees with the optimal solution that is to evaluate the marginal distribution of $S_\theta^{k_a}(\mathbf{x}_a)$.

6.6 Hypothesis testing

It was shown in Section 4.5 that the optimum test between two hypothesis — sampling distributions S_1 and S_2 — depends on the data sequence $\mathbf{x}$ only through the empirical distribution $R_{\mathbf{x}}$. Thus, design of optimum test can be regarded as splitting the set of all possible distributions on $\mathcal{X}$ into two disjoint subsets, $\mathcal{R}_1$ and $\mathcal{R}_2$, with the following meaning

$$\begin{aligned} &\text{if } R_{\mathbf{x}} \in \mathcal{R}_1, \quad \text{then } S_1 \quad \text{is accepted}, \\ &\text{if } R_{\mathbf{x}} \in \mathcal{R}_2, \quad \text{then } S_2 \quad \text{is accepted}. \end{aligned}$$

The probabilities of errors of the first and second kind are then approximately (cf. Cover and Thomas (1991, Section 12.7))

$$\begin{aligned} S_1^k(\{\mathbf{x} : R_{\mathbf{x}} \in \mathcal{R}_2\}) &\doteq \exp\{-k\, D(\mathcal{R}_2 \| S_1)\}, \\ S_2^k(\{\mathbf{x} : R_{\mathbf{x}} \in \mathcal{R}_1\}) &\doteq \exp\{-k\, D(\mathcal{R}_1 \| S_2)\}. \end{aligned}$$

7. COPING WITH "BAD" MODEL

The assumption that the model family contains the "true" model of data is unbearable in practice. Any model is only an approximate description of the true behaviour of data. When the empirical distribution $R_{\mathbf{x}}$ does not converge to any sampling distribution from a model family

$$R_{\mathbf{x}} \to R_\infty \notin \{S_\theta : \theta \in \mathcal{T}\},$$

the minimum Kullback-Leibler distance

$$\min_\theta D(R_{\mathbf{x}} \| S_\theta) \tag{29}$$

does not approach zero even for very large samples. The quantity (29) thus provides a natural measure of "distance" between the model and data.

Example 7. Let $\{S_\theta : \theta = 1, \ldots, N\}$ and $\{S'_\theta : \theta = 1, \ldots, N\}$ be two model families such that

$$D(R_{\mathbf{x}} \| S'_\theta) = D(R_{\mathbf{x}} \| S_\theta) + \Delta,\ \theta = 1, \ldots, N$$

where $\Delta > 0$ is a constant independent of θ. Note that the posterior distributions for such families coincide

$$\begin{aligned} P'_{\mathbf{x}}(\theta) &\propto P(\theta)\, \exp\{-k\, D(R_{\mathbf{x}} \| S'_\theta)\} \\ &\propto P(\theta)\, \exp\{-k\, D(R_{\mathbf{x}} \| S_\theta)\} \propto P_{\mathbf{x}}(\theta). \end{aligned}$$

It is because the posterior distribution does not carry any information about how well the model family, as a whole, fits the actual distribution of data. The posterior distribution gives only a relative comparison of "goodness-of-fit" of particular sampling distributions *within* a given family (cf. Fig. 10).

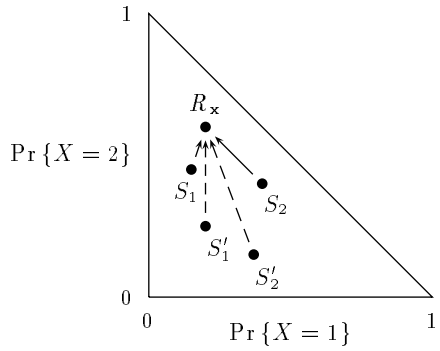

Fig. 10. Kullback-Leibler distance gives an absolute measure of goodness of fit between the model and data. Thus it makes it possible to compare not only the sampling distributions S_1 against S_2 and S'_1 against S'_2, but also the model class $\{S_\theta\}$ against $\{S'_\theta\}$.

A natural extension of (29) to the case of incomplete data is the minimum Kullback-Leibler distance from the corresponding set of empirical distributions $\mathcal{R}$

$$\min_\theta D(\mathcal{R} \| S_\theta). \tag{30}$$

Note that with less information accumulated, it is more difficult to discriminate between model and data. The bigger the set $\mathcal{R}$ is, the smaller (closer to zero) the minimum Kullback-Leibler distance (30) typically is.

Example 8. A sample of 400 data was simulated according to the model $x_k = \theta + e_k$ with $\theta = 1$ and e_k being a discrete Student-like distribution with 3 degrees of freedom. The model family $\{S_\theta\}$ was deliberately considered different — discrete Cauchy-like. The statistic T was defined through the following functions (for motivation see Kulhavý (1993*b*))

$$h_j(x) = \log S_{\theta^*_{j+1}}(x) - \log S_{\theta^*_j}(x),\ j = 1, 2$$

with $\theta^*_1 = -3$, $\theta^*_2 = 0$, $\theta^*_3 = 3$. Both the empirical values

$$T(\mathbf{x}) = \frac{1}{k} \sum_{i=1}^{k} h(x_i),\ k = 1, 2, \ldots, 400$$

and the theoretically expected asymptotic values

$$\hat{T}_\theta(\mathbf{x}) = \sum_{a \in \mathcal{X}} S_\theta(a)\, h(a),\ \theta \in \mathcal{T}$$

were computed. Fig. 11 illustrates the effect of mismodelling — even for large k the empirical value of the statistic T is far from its expected value.

Note that the empirical and model distributions of data appear in Kullback-Leibler distances $D(R_{\mathbf{x}} \| S_\theta)$ and $D(\mathcal{R} \| S_\theta)$ separately. Therefore, it is possible to change the model family $\{S_\theta\}$ in the course of identification without losing information contained in the past data. "Adaptive" identification is driven by the objective to make (29) or (30) small enough.

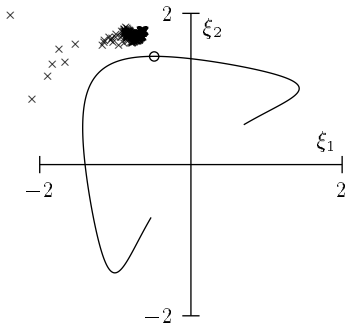

Fig. 11. A sequence of values of a data statistic (crosses) for Student-distributed data against the theoretically expected values of the same statistic for Cauchy distribution around different values of θ (solid curve, circle marks $\theta = 1$).

8. MARKOV CHAINS

To give an example how the above approach can be extended to dependent data, the *Markov chain* model is analysed in detail. Much of the following can be extended straightforwardly to the case of general regression with external input.

Suppose that $X_1, X_2, \ldots, X_{k+1}$ form a Markov chain of first order with a conditional probability mass function

$$S(y|z) = \Pr\{X_{k+1} = y | X_k = z\} .$$

The transition probability distribution is known only partially — it is assumed to belong to a family $\{S_\theta : \theta \in \mathcal{T}\}$ with θ ranging over a finite set or an open interval. The objective is to estimate the parameter θ.

8.1 Joint distribution of sample

Given a sequence of observations $\mathbf{x} = (x_1, \ldots, x_{k+1})$, the joint probability $S_\theta^k(\mathbf{x}|x_1)$ conditional on the initial value x_1 is

$$S_\theta^k(\mathbf{x}|x_1) = \prod_{i=1}^{k} S_\theta(x_{i+1}|x_i) .$$

Let $R_\mathbf{x}$ be an empirical distribution of second order defined by

$$R_\mathbf{x}(a,b) = \frac{N_\mathbf{x}(a,b)}{k},\ (a,b) \in \mathcal{X}^2$$

where $N_\mathbf{x}(a,b)$ counts the number of occurrences of the pair $(a,b) \in \mathcal{X}^2$ in the sequence $\mathbf{x}$. Then, proceeding analogously as in Section 2.1, we can put $S_\theta^k(\mathbf{x}|x_1)$ into the form

$$S_\theta^k(\mathbf{x}|x_1) = \exp\{-k\,[\bar{H}(R_\mathbf{x}) + \bar{D}(R_\mathbf{x}\|S_\theta)]\} \quad (31)$$

where

$$\bar{H}(R) = -\sum_{(y,z)\in\mathcal{X}^2} R(y,z)\log R(y,z) + \sum_{z\in\mathcal{X}} R(z)\log R(z)$$

is *conditional Shannon entropy* of a random variable Y given another variable Z described jointly by probability distribution R, and

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

is *conditional Kullback-Leibler distance* of joint probability distribution R and conditional distribution S.

8.2 Posterior distribution of parameter

Adopting the Bayesian viewpoint, the *posterior* distribution of the unknown parameter *conditional* on $\mathbf{x}$ is

$$P_\mathbf{x}(\theta) \propto P(\theta)\, S_\theta^k(\mathbf{x}|x_1) .$$

Substituting (31) for $S_\theta^k(\mathbf{x}|x_1)$ gives

$$P_\mathbf{x}(\theta) \propto P(\theta)\, \exp\{-k\,\bar{D}(R_\mathbf{x}\|S_\theta)\} \quad (32)$$

since the conditional entropy $\bar{H}(R_\mathbf{x})$ does not depend on θ.

Example 9. Any of conditional distributions $S_\theta(y|z)$, $\theta \in \mathcal{T}$ can be envisaged as a set of distributions $S_\theta^z(y) = S_\theta(y|z)$, $z \in \mathcal{X}$. Similarly, the conditional empirical distribution $R_\mathbf{x}(y|z)$ can be regarded as a set of points $R_\mathbf{x}^z(y) = R_\mathbf{x}(y|z)$, $z \in \mathcal{X}$. The conditional Kullback-Leibler distance is then an empirical expectation of Kullback-Leibler distance between the conditional distributions (cf. Fig.12)

$$\bar{D}(R_\mathbf{x}\|S_\theta) = \sum_{z\in\mathcal{X}} R_\mathbf{x}(z)\, D(R_\mathbf{x}^z\|S_\theta^z) .$$

Note that the conditional empirical distribution $R_\mathbf{x}(y|z)$ is not uniquely determined for such $z \in \mathcal{X}$ that have not been observed. Because $R_\mathbf{x}(z) = 0$ then, the ambiguity does not affect the resulting value.

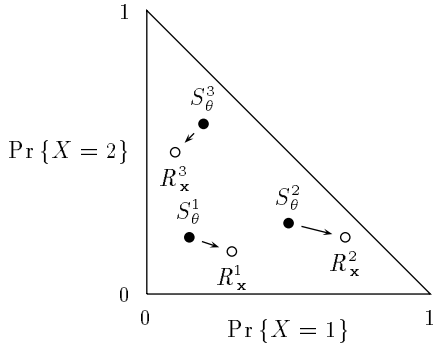

Fig. 12. Conditional Kullback-Leibler distance $\bar{D}(R_\mathbf{x}\|S_\theta)$ is an empirical expectation of Kullback-Leibler distance $D(R_\mathbf{x}^z\|S_\theta^z)$ for z fixed.

8.3 Probability of large deviations

Suppose that the empirical distribution $R_\mathbf{x}$ is known only to belong to a certain set $\mathcal{R}$ of distributions on $\mathcal{X}^2$. Then it makes sense to ask what is the probability

$$S_\theta^k(\{\mathbf{x} : R_\mathbf{x} \in \mathcal{R}\}|x_1) .$$

A Markov chain version of the *large deviation theorem* is the following. Suppose that $S(y|z) > 0$ for all $(y,z) \in \mathcal{X}^2$. Let $\mathcal{C}$ be the set of all distributions on $\mathcal{X}^2$ such that their both marginals coincide, i.e.,

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

Thus, any distribution from $\mathcal{C}$ determines a stationary Markov chain. Suppose $\mathcal{R}$ is the closure of its interior and denote $\bar{\mathcal{R}} = \mathcal{R} \cap \mathcal{C}$. Then it holds

$$\lim_{k \to \infty} \frac{1}{k} \log S_\theta^k(\{\mathbf{x} : R_{\mathbf{x}} \in \mathcal{R}\} | x_1) = -\bar{D}(\bar{\mathcal{R}} \| S_\theta) \tag{33}$$

where

$$\bar{D}(\bar{\mathcal{R}} \| S_\theta) = \min_{R \in \bar{\mathcal{R}}} \bar{D}(R \| S_\theta) . \tag{34}$$

For proof see Boza (1971, Theorem 3.1) and Natarajan (1985, Theorem 1). Csiszár *et al.* (1987, Lemma 2) refrained (under additional regularity assumptions) from the strict positivity of S_θ. This is essential for generalization to higher-order Markov chains.

8.4 Minimum Kullback-Leibler distance

Suppose that the set $\mathcal{R}$ is bounded by hyperplanes

$$\mathcal{R}_\xi = \left\{ R : \sum_{(a,b) \in \mathcal{X}^2} R(a,b)\, h_j(a,b) \geq \xi_j ,\ j = 1, \ldots, n \right\}$$

where h_1,...,h_n are given real functions on $\mathcal{X}^2$.

Then

$$D(\bar{\mathcal{R}}_\xi \| S_\theta) = \max_{\lambda \geq 0} \left[\sum_{j=1}^{n} \lambda_j\, \xi_j - \log N(\lambda) \right] \tag{35}$$

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

$$M_{y,z}(\lambda) = S_\theta(y|z)\, \exp\left\{ \sum_{j=1}^{n} \lambda_j\, h_j(y,z) \right\} . \tag{36}$$

A simpler result for $S(y|z) = \text{const.}$ was proved in Spitzer (1972, Theorem), Justesen and Høholdt (1984, Theorem 1). For the above result see Csiszár *et al.* (1987, Equation 28).

8.5 Minimum Kullback-Leibler distance distribution

The distribution R^* minimizing conditional Kullback-Leibler distance is given by

$$R^*(y,z) = \frac{u_y(\lambda)\, M_{y,z}(\lambda)\, v_z(\lambda)}{N(\lambda)} \tag{37}$$

where $N(\lambda)$ is the largest eigenvalue and $u(\lambda)$ and $v(\lambda)$ are the corresponding left and right eigenvectors, normalized to have inner product 1, of the matrix $M(\lambda)$ (36) .

An analogous result for $S(y|z) = \text{const.}$ was proved in Spitzer (1972, Theorem) and Justesen and Høholdt (1984, Theorem 1). The above form is due to Csiszár *et al.* (1987, Equation 29).

8.6 Coping with compressed data

The large deviation results can be used to cope with incomplete data analogously as in the case of independent data. Consider data compression as an illustrative example.

Suppose that data are compressed using a second-order data statistic of the sample average form

$$T(\mathbf{x}) = \frac{1}{k} \sum_{i=1}^{k} h(x_i, x_{i+1}) \tag{38}$$

where h is a given vector function on $\mathcal{X}^2$. In terms of the second-order empirical distribution, the data statistic T can be regarded as an empirical mean of $h(Y,Z)$

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

Provided the only information one has about the sample $\mathbf{x}$ is that $T(\mathbf{x}) \geq \xi$, the probability of observing such a sample is approximately

$$S_\theta^k(\{\mathbf{x} : R_{\mathbf{x}} \in \mathcal{R}_\xi\} | x_1) \doteq \exp\{-k\, \bar{D}(\bar{\mathcal{R}}_\xi \| S_\theta)\} \tag{39}$$

and the posterior distribution conditional on the statistic value can be approximated as follows

$$\hat{P}_\xi(\theta) \propto P(\theta)\, \exp\{-k \bar{D}(\bar{\mathcal{R}}_\xi \| S_\theta)\} . \tag{40}$$

Note that the approximate distribution (40) has formally the same structure as the approximation (27) for independent data.

9. CONCLUDING REMARKS

There are cases when it is convenient to view inference as measuring of Kullback-Leibler distance between the empirical and model distributions of observed data. This is the main message of the paper. The use of Kullback-Leibler distance turns out conceptually and computationally simpler than the use of probability. Both views of uncertainty are, however, firmly related.

From probability to distance. The view of parameter estimation through the Kullback-Leibler distance optics has some noteworthy features. First of all, information contained in data and its explanation through model play a symmetric role in this view. As a result, both data and model can be modified rather freely during estimation. Thus, for instance, model can be built adaptively — dependent on previous data, or various kinds of reduced or corrupted information can be handled.

One can ask whether this is a unique feature of the Kullback-Leibler distance approach. Certainly not.

The same could be achieved in terms of probability as long as we kept data and model separate. The point is that it is not done so in typical cases, rather the likelihood or posterior distribution are evaluated and propagated directly. The Kullback-Leibler distance view insists on separating data and model by principle.

Intuitive appeal. The author's teaching and lecturing experience is that newcomers to the area, including students, usually take quite a long time to get used to a fully probabilistic view of inference. To master methods such as least squares that measure a distance between data and model *directly in the data space* is a lot easier. With the latter approach, however, one quickly loses insight when dealing with nonstandard problems such as handling outliers in data.

The Kullback-Leibler distance approach is a natural generalization of the least-squares view. One can think of a "distance" between the actual (empirical) and model distributions of data without necessarily introducing the conceptually more demanding concepts of likelihood or posterior. The natural question "Why should I use just Kullback-Leibler distance?" has a good answer "Because you are consistent then with what probability is doing (even if you don't realize it)".

Curse of dimensionality. Probability can be regarded as a natural, consistent, unifying concept in all problems where uncertainty plays an important role. This position seems to be unshakable. Yet, the correct use of probability is often infeasible in practical problems. The computational complexity of the fundamental operations of probability theory — conditioning and marginalization — quickly grows with the dimension of the underlying spaces. This is why it makes sense to look for alternatives that would be computationally less demanding, yet close to probability-based inference. The use of Kullback-Leibler distance is believed to offer such an alternative. The main "trick" (explained in detail in Section 5) is that summing of probability over data of specific form can be approximated by minimizing of Kullback-Leibler distance. This shift is accompanied with massive reduction of computational complexity. How far the potential of this approach extends is an open and challenging question.

ACKNOWLEDGMENT

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

REFERENCES

Amari, S. (1985). *Differential-Geometrical Methods in Statistics*. Vol. 28 of *Lecture Notes in Statistics*. Springer-Verlag, Berlin.

Boza, L. B. (1971). Asymptotically optimal tests for finite Markov chains. *Ann. Math. Statist.* **42**, 1992–2007.

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

Cover, T. M. and J. A. Thomas (1991). *Elements of Information Theory*. Wiley, New York.

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.* **2**, 299–318.

Csiszár, I. (1984). Sanov property, generalized I-projection and a conditional limit theorem. *Ann. Probab.* **12**, 768–793.

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

Csiszár, I., T. M. Cover and B.-S. Choi (1987). Conditional limit theorem under Markov conditioning. *IEEE Trans. Inform. Theory* **33**, 788–801.

DeGroot, M. H. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.

Fisher, R. A. (1925). Theory of statistical estimation. *Proc. Camb. Phil. Soc.* **22**, 700–725.

Justesen, J. and T. Høholdt (1984). Maxentropic Markov chains. *IEEE Trans. Inform. Theory* **30**, 665–667.

Kulhavý, R. (1993*a*). 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.

Kulhavý, R. (1993*b*). On design of approximate finite-dimensional estimators: the Bayesian view. In: *Mutual Impact of Computing Power and Control Theory* (K. Warwick and M. Kárný, Eds.). pp. 13–39. Plenum Press. New York.

Kullback, S. (1959). *Information Theory and Statistics*. Wiley, New York.

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

Natarajan, S. (1985). Large deviations, hypotheses testing, and source coding for finite Markov chains. *IEEE Trans. Inform. Theory* **31**, 360–365.

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

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

Spitzer, F. (1972). A variational characterization of finite Markov chains. *Ann. Math. Statist.* **43**, 580–583.

Vajda, I. (1989). *Theory of Statistical Inference and Information*. Kluwer, Dordrecht.

Wolfowitz, J. (1957). The minimum distance method. *Ann. Math. Statist.* **28**, 75–88.