RECURSIVE NONLINEAR ESTIMATION THROUGH GLOBAL APPROXIMATION OF MODEL*

R. Kulhavý
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
P. O. Box 18, 182 08 Prague, Czech Republic
kulhavy@utia.cas.cz

Keywords: Parameter estimation, statistical inference, recursive algorithms, model approximation, exponential family.

Abstract

A natural way of beating the complexity of statistical parameter estimation is to approximate directly the intractable model. Although this approach is often used in practice, it is mostly disliked or suspected by theorists because of the danger of accumulation of approximation errors in recursive estimation. The paper shows that a globally valid approximation of model can be built by projecting sampling distributions onto a suitably chosen exponential family. The case of independent and identically distributed data is considered.

1 Introduction

There has been much interest, practical as well as theoretical, in applying statistical methods of parameter estimation to 'complex' models such as non-Gaussian, non-linear, or high-dimensional ones. The crucial obstacle to direct application is the extreme computational complexity of major estimation schemes. When the posterior density or likelihood function cannot be computed in a closed form, some approximation of the infeasible solution is inevitable.

The problem has been treated in various ways. One natural approach is a *direct approximation of model*. The trick is that a simpler, easily identifiable model is substituted for the original, intractable one. This approach was applied with some success e.g. to solve the discrete-time nonlinear filtering problem. Simpler models result mostly from *local* approximation of a given model. Approximation techniques include local linearization, second-order expansion of nonlinearities, Gram-Charlier and Edgeworth expansions of non-Gaussian distributions (for more information see e.g. [1]).

Unfortunately, the use of just a locally valid approximation of the original model can suffer from accumulation of approximation errors in recursive estimation. It may happen that the results of parameter estimation for the true and approximate model differ significantly. This fact together with the little theoretical insight seem to diminish the potential value of the model approximation approach.

A challenging question in this context is whether a *global* approximation of model can be designed so that the results of estimation for the true and approximate models are related to each other in a way that admits an *a priori* analysis. This would enable the user to draw conclusions about the accuracy of approximation or the increase of estimation uncertainty.

The paper demonstrates on the case of Bayesian estimation for independent and identically distributed data that such an approximation can be constructed. The starting point in design is to understand how the choice of model affects the relative entropy between particular sampling distributions and the empirical distributions consistent with the value of the used data statistic. It is so because the relative entropy, taken as a function of the unknown parameter, essentially determines the likelihood function and the posterior distribution. The paper makes use of recent results in this direction [2, 3].

2 Bayesian Estimation

Consider a sequence of random variables (data) $\mathbf{X} = (X_1, \ldots, X_k)$ taking values in a finite set $\mathcal{X} = \{1, 2, \ldots, N\}$. Suppose that $X_1, X_2, \ldots$ are independent and identically distributed according to a common probability distribution S_θ parameterized by a parameter $\theta \in \mathcal{T}$. For simplicity, θ is supposed to be an integer ranging over a finite set although extension of what follows to the case of a real parameter is straightforward. Note that the numbers of elements of $\mathcal{X}$ and $\mathcal{T}$ are very large in typical cases. The problem is to estimate the unknown parameter θ given a sequence of observations $\mathbf{x} = (x_1, \ldots, x_k)$.

The Bayesian solution of the problem is given by the probability distribution $P_{\mathbf{x}}$ of the unknown parameter Θ (regarded as a random variable) conditional on $\mathbf{X} = \mathbf{x}$. Given a prior distribution P of Θ, the posterior probability mass function $P_{\mathbf{x}}(\theta),\ \theta \in \mathcal{T}$ is determined by the well-known Bayes formula. We present it in a form that is not quite common but useful for our purpose. Two notions need to be introduced first.

*Work supported in part by grant 102/94/0314 of the Grant Agency of the CR, grant 275109 of the Academy of Sciences of the CR and the EC 'Copernicus' project CT94–0237.

The *empirical distribution* of $\mathbf{x} = (x_1, \ldots, x_k)$ is a probability distribution defined by the relative frequencies

$$R_{\mathbf{x}}(a) = \frac{1}{k} N_{\mathbf{x}}(a), \quad a \in \mathcal{X} \tag{1}$$

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

The *relative entropy* (Kullback-Leibler distance, informational divergence) of two distributions R and S on a finite set $\mathcal{X}$ is defined by the quantity

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

Theorem 1: Provided $S_\theta(x) > 0$ for all $x \in \mathcal{X}$ and $\theta \in \mathcal{T}$, the conditional distribution of Θ given a sample $\mathbf{X} = \mathbf{x}$ of size k is given by the formula

$$P_{\mathbf{x}}(\theta) \propto P(\theta)\, \exp\bigl(-k\, D(R_{\mathbf{x}}\|S_\theta)\bigr) \tag{3}$$

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

Proof: Elementary rules of probability calculus give

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

The argument of the exponential function can be rewritten as follows

$$\begin{aligned} \sum_{i=1}^{k} \log S_\theta(x_i) \;&=\; k \sum_{a\in\mathcal{X}} R_{\mathbf{x}}(a) \log S_\theta(a) \\ &=\; -k\, D(R_{\mathbf{x}}\|S_\theta) + \text{const.} \end{aligned}$$

where the constant does not depend on θ. ∎

What Theorem 1 says is that the posterior distribution of Θ is a simple transform of the relative entropy between the empirical distribution $R_{\mathbf{x}}$ and particular sampling distributions S_θ, $\theta \in \mathcal{T}$. The transform depends on the prior distribution P and the sample size k. One could say that Bayesian inference implicitly measures a "distance" between the actual and model distributions of observed data.

3 Data Compression

Recursive estimation requires compression of observed data $\mathbf{x} = (x_1, \ldots, x_k)$ using a suitable data statistic $T_k : \mathcal{X}^k \to \mathbb{R}^n$. We restrict our attention to statistics of the following form

$$T_k(\mathbf{x}) = \frac{1}{k} \sum_{j=1}^{k} h(x_j) = \sum_{a\in\mathcal{X}} R_{\mathbf{x}}(a)\, h(a) \tag{4}$$

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

Suppose that the only information saved about the observed sample $\mathbf{x}$ of size k is $T_k(\mathbf{x}) = \bar{h}$. All we can say then about the empirical distribution $R_{\mathbf{x}}$ is that it belongs to a certain subset of the set $\mathcal{R}$ (probability simplex) of all distributions on $\mathcal{X}$

$$\mathcal{R}_{\bar{h}} = \biggl\{ R \in \mathcal{R} : \sum_{x\in\mathcal{X}} R(x)\, h(x) = \bar{h} \biggr\} \,. \tag{5}$$

The optimal Bayesian solution to estimation with compressed data is given by the conditional distribution

$$P_{\bar{h}}(\theta) \propto P(\theta)\, S_\theta^k(\{\mathbf{x} : R_{\mathbf{x}} \in \mathcal{R}_{\bar{h}}\}) \tag{6}$$

where S_θ^k denotes the joint distribution of a sample $\mathbf{X} = (X_1, \ldots, X_k)$, i.e., $S_\theta^k(x) = \prod_{i=1}^k S_\theta(x_i)$. In typical cases, the computation of $S_\theta^k(\{\mathbf{x} : R_{\mathbf{x}} \in \mathcal{R}_{\bar{h}}\})$ is not feasible.

4 Global Approximation of Model

We suggest to approximate the ideal but infeasible solution (6) in the following way.

4.1 Choice of an Approximating Family

A parametric approximating family $\mathcal{W} = \{W_{O,\lambda} : \lambda \in \mathbb{R}^n\}$ is defined as an exponential family generated by the vector function $h : \mathcal{X} \to \mathbb{R}^n$

$$W_{O,\lambda}(x) = O(x)\, \exp\bigl(\lambda'\, h(x) - \psi(\lambda)\bigr) \tag{7}$$

where $\lambda' = (\lambda_1, \ldots, \lambda_n)$ and

$$\psi(\lambda) = \log \sum_{x\in\mathcal{X}} O(x)\, \exp\bigl(\lambda'\, h(x)\bigr) \,. \tag{8}$$

Remember that the function $h(x)$ comes from the statistic definition (4). A proper choice of the origin O of the exponential family $\mathcal{W}$ is essential in approximation and is discussed later. A simpler notation W_λ is used below when the origin O need not be stressed.

4.2 Approximation of the Original Family

The original sampling distributions S_θ, $\theta \in \mathcal{T}$ are projected onto the exponential family $\mathcal{W}$ so to meet the vector equality

$$\sum_{x\in\mathcal{X}} S_\theta(x)\, h(x) = \sum_{x\in\mathcal{X}} W_\lambda(x)\, h(x) \,. \tag{9}$$

The projection defines a mapping $\theta \mapsto \lambda$ that assigns to each $\theta \in \mathcal{T}$ the corresponding value of $\lambda \in \mathbb{R}^n$. A short-hand notation $\hat{S}_\theta = W_{\lambda(\theta)}$ is used in the sequel for the approximate sampling distributions.

4.3 Estimation for the Approximate Family

The parameter θ of the approximate sampling family $\hat{\mathcal{S}} = \{\hat{S}_\theta : \theta \in \mathcal{T}\}$ is estimated. By Theorem 1, the posterior distribution conditional on $\mathbf{x}$ can be given the form

$$\hat{P}_{\mathbf{x}}(\theta) \propto P(\theta)\, \exp\bigl(-k\, D(R_{\mathbf{x}}\|\hat{S}_\theta)\bigr) \,. \tag{10}$$

Because of data compression, $R_{\mathbf{x}}$ is unknown to us. But, owing to the definition of the enveloping exponential family $\mathcal{W} \supset \hat{\mathcal{S}}$, the following Pythagorean relationship holds [4, Theorem 22.1], [5, Theorem 2.2], [6, Theorem 3.6]

$$D(R_{\mathbf{x}}\|W_\lambda) = D(\mathcal{R}_{\bar{h}}\|W_\lambda) + \text{const.}, \quad \lambda \in \mathbb{R}^n \tag{11}$$

where $D(\mathcal{R}_{\bar{h}}\|W_\lambda) = \min_{R\in\mathcal{R}_{\bar{h}}} D(R\|W_\lambda)$ and the constant does not depend on λ. Therefore, the posterior distribution (10) conditional on complete data coincides with the posterior distribution conditional on just the statistic value, i.e., the statistic (4) is sufficient for estimation of the parameter λ

$$\hat{P}_{\bar{h}}(\theta) \propto P(\theta)\, \exp\bigl(-k\, D(\mathcal{R}_{\bar{h}}\|\hat{S}_\theta)\bigr)\,. \tag{12}$$

4.4 Approximate Relative Entropy

Substituting (7) for W_λ in $D(R_{\mathbf{x}}\|W_\lambda)$ and taking (11) into account, we find that

$$D(\mathcal{R}_{\bar{h}}\|\hat{S}_\theta) = -\lambda'(\theta)\,\bar{h} + \psi(\lambda(\theta)) + \text{const.} \tag{13}$$

where the constant does not depend on θ. Thus, the relative entropy as a function of θ is a linear combination of fixed basis functions $\lambda_i(\theta)$, $i = 1,\dots,n$ plus the term $\psi(\lambda)$ due to the normalization of W_λ. Let us emphasize that the functions $\lambda_i(\theta)$ and $\psi(\lambda)$ can be precomputed before estimation.

Given a vector function h, the only free parameter of the approximation scheme is the distribution O. The choice of the origin O is required to ensure that

$$D(\mathcal{R}_{\bar{h}}\|\hat{S}_\theta) \le D(\mathcal{R}_{\bar{h}}\|S_\theta) \tag{14}$$

for each $\theta\in\mathcal{T}$ and every possible $\bar{h}$. The condition (14) gives the approximate estimation of θ attractive properties.

5 Key Properties of Approximation

5.1 Monotonicity

It is easy to verify that (14) implies the following chain of inequalities

$$0 \le D(\mathcal{R}_{\bar{h}}\|\hat{S}_\theta) \le D(\mathcal{R}_{\bar{h}}\|S_\theta) \le D(R_{\mathbf{x}}\|S_\theta)\,. \tag{15}$$

The rightmost inequality follows trivially from the fact that $R_{\mathbf{x}}\in\mathcal{R}_{\bar{h}}$ while the leftmost inequality is a consequence of the following fundamental property of relative entropy.

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

Proof: See e.g. [7, Theorem 2.6.3]. ∎

Note that (15), (11) and Lemma 1 imply the following implications:

- $\mathcal{W}\supset\mathcal{S} \;\Rightarrow\; D(R_{\mathbf{x}}\|S_\theta) = D(\mathcal{R}_{\bar{h}}\|\hat{S}_\theta) + \text{const.},\ \theta\in\mathcal{T} \;\Rightarrow\; \hat{P}_{\bar{h}}(\theta) = P_{\mathbf{x}}(\theta)$. Thus, if the exponential family $\mathcal{W}$ is rich enough to include the sampling distributions S_θ for all $\theta\in\mathcal{T}$, the approximation $\hat{P}_{\bar{h}}$ coincides with the true posterior $P_{\mathbf{x}}$.
- $\mathcal{W}=\{O\} \;\Rightarrow\; D(\mathcal{R}_{\bar{h}}\|\hat{S}_\theta) = \text{const.},\ \theta\in\mathcal{T} \;\Rightarrow\; \hat{P}_{\bar{h}}(\theta) = P(\theta)$. In other words, if the approximation family $\mathcal{W}$ is so poor that it reduces to a single point, then the data statistic brings no information about data at all and the approximation $\hat{P}_{\bar{h}}$ returns the prior distribution P.

Loosely speaking, the approximate distribution $\hat{P}_{\bar{h}}$ always lies "somewhere between" the prior P and the true posterior $P_{\mathbf{x}}$. The "richer" is the family $\mathcal{W}$, the "closer" $\hat{P}_{\bar{h}}$ is to $P_{\mathbf{x}}$.

5.2 Asymptotics

Suppose $R_{\mathbf{x}}\to S_{\theta_0}$ (pointwise) for a certain value $\theta_0\in\mathcal{T}$. Then, by Lemma 1, $D(R_{\mathbf{x}}\|S_{\theta_0})\to 0$. The inequality (15) forces $D(\mathcal{R}_{\bar{h}}\|\hat{S}_{\theta_0})\to 0$. If $P(\theta_0)>0$, then also $\lim_{k\to\infty}\hat{P}_{\bar{h}}(\theta_0)>0$. Thus, the asymptotic behaviour of the approximate estimation is consistent with the ideal estimation.

More than that can be said. The following classical result of large deviation theory describes the asymptotic behaviour of the probability $S_\theta^k(\{\mathbf{x}: R_{\mathbf{x}}\in\mathcal{R}_{\bar{h}}\})$ in (6).

Lemma 2: Let $\mathcal{C}\subset\mathcal{R}$ be such that the closure of the interior of $\mathcal{C}$ equals $\mathcal{C}$. Then for k independent drawings from a distribution S_θ with an arbitrary fixed $\theta\in\mathcal{T}$, the probability of the sample with an empirical distribution belonging to $\mathcal{C}$ has the asymptotics

$$\lim_{k\to\infty}\frac{1}{k}\log S_\theta^k(\{\mathbf{x}: R_{\mathbf{x}}\in\mathcal{C}\}) = -D(\mathcal{C}\|S_\theta)\,. \tag{16}$$

Proof: See [8, Theorem 1] or [7, Theorem 12.4.1]. ∎

Lemma 2 says that the probability $S_\theta^k(\{\mathbf{x}: R_{\mathbf{x}}\in\mathcal{C}\})$ converges for $S_\theta\notin\mathcal{C}$ to zero exponentially fast, with the rate given by the relative entropy $D(\mathcal{C}\|S_\theta)$. What our approximation suggests to do through (14) is simply to limit the rate of convergence $D(\mathcal{R}_{\bar{h}}\|\hat{S}_\theta)$ by the ideal expression $D(\mathcal{R}_{\bar{h}}\|S_\theta)$ (see [3] for more details).

6 Construction of Approximating Family

6.1 Dual Parameterization of Exponential Family

The only information about the sampling distributions that is used in the projection (9) is given by the expectations

$$\hat{h}(\theta) = \sum_{x\in\mathcal{X}} S_\theta(x)\,h(x), \qquad \hat{h}(\lambda) = \sum_{x\in\mathcal{X}} W_\lambda(x)\,h(x)\,.$$

As for the exponential family $\mathcal{W}$, the expectation $\hat{h}\in\mathbb{R}^n$ can be regarded as another way of its parameterization, dual in a sense to the vector $\lambda\in\mathbb{R}^n$. In fact, when the functions $h_i(x)$ are linearly independent, there is a one-to-one correspondence between $\lambda\in\mathbb{R}^n$ and $\hat{h}$ on the set of its possible values [4, Theorem 19.1]. This justifies a dual notation $W_{O,\hat{h}}$ for an exponential distribution (7) that goes through a point O and satisfies

$$\sum_{x\in\mathcal{X}} W_{O,\lambda}(x)\,h(x) = \hat{h}\,.$$

6.2 Covariance Condition

It is a well-known fact (see e.g. [4, Theorem 22.3]) that the relative entropy $D(R\|W_{O,\hat{h}})$ for $R\in\mathcal{R}_{\bar{h}}$ achieves minimum at the projection (9) of $R_{\mathbf{x}}$ onto $\mathcal{W}$

$$D(\mathcal{R}_{\bar{h}}\|W_{O,\hat{h}}) = D(W_{O,\bar{h}}\|W_{O,\hat{h}})\,. \tag{17}$$

Since the approximate family $\hat{\mathcal{S}}$ results from the projection (9) of $\mathcal{S}$ onto $\mathcal{W}$, the condition (14) is clearly guaranteed if

$$D(W_{O,\bar{h}}\|W_{O,\hat{h}}) \le D(W_{S_\theta,\bar{h}}\|W_{S_\theta,\hat{h}})\,. \tag{18}$$

holds for every $\theta \in \mathcal{T}$ and every possible $\bar{h}$ and $\hat{h}$.

Lemma 3: The inequality (18) is satisfied if

$$(\hat{h}-\bar{h})' \nabla_{\hat{h}} D(W_{O,\bar{h}} \| W_{O,\hat{h}}) \le (\hat{h}-\bar{h})' \nabla_{\hat{h}} D(W_{S_\theta,\bar{h}} \| W_{S_\theta,\hat{h}})$$

holds for every $\theta \in \mathcal{T}$ and every possible $\bar{h}$ and $\hat{h}$.

Proof: Let us denote $\delta_{O,\bar{h}}(\hat{h}) = D(W_{O,\bar{h}} \| W_{O,\hat{h}})$. Since the function $\delta_{O,\bar{h}}(\hat{h})$ is convex and differentiable on the set of all possible values of $\hat{h} \in \mathbb{R}^n$, we have (see e.g. [9, Theorem 3.3.3])

$$\delta_{O,\bar{h}}(\hat{h}) - \delta_{O,\bar{h}}(\bar{h}) \le (\hat{h} - \bar{h})' \nabla \delta_{O,\bar{h}}(\hat{h})$$

for every possible $\hat{h}$. By Lemma 1, $\delta_{O,\bar{h}}(\hat{h})$ is zero if and only if $\hat{h} = \bar{h}$ and positive if $\hat{h} \neq \bar{h}$. Thus, when $\hat{h} \neq \bar{h}$, the above inequality holds even for ratios

$$\frac{\delta_{O,\bar{h}}(\hat{h})}{\delta_{S_\theta,\bar{h}}(\hat{h})} \le \frac{(\hat{h} - \bar{h})' \nabla \delta_{O,\bar{h}}(\hat{h})}{(\hat{h} - \bar{h})' \nabla \delta_{S_\theta,\bar{h}}(\hat{h})}$$

which directly implies the proposition. When $\hat{h} = \bar{h}$, the proposition holds trivially. ∎

Lemma 4: Given an exponential family $\mathcal{W} = \{W_{O,\hat{h}}\}$,

$$(\hat{h} - \bar{h})' \nabla_{\hat{h}} D(W_{O,\bar{h}} \| W_{O,\hat{h}}) = (\hat{h} - \bar{h})' G^{-1}(\hat{h}) (\hat{h} - \bar{h})$$

where $G(\hat{h}) = \mathrm{Cov}_{O,\hat{h}}(h)$ denotes the covariance matrix of $h(X)$, i.e., the Fisher information matrix of λ.

Proof: After substituting (7) for $W_{O,\hat{h}}$ we get

$$D(W_{O,\bar{h}} \| W_{O,\hat{h}}) = -\lambda'(\hat{h})\, \bar{h} + \psi(\lambda(\hat{h}))\,.$$

Partial differentiation with respect to $\hat{h}_i$ gives

$$\frac{\partial}{\partial \hat{h}_i} D(W_{O,\bar{h}} \| W_{O,\hat{h}}) = \sum_{j=1}^{n} \frac{\partial \lambda_j}{\partial \hat{h}_i} (\hat{h}_j - \bar{h}_j)\,.$$

Since $\sum_{j=1}^{n} \frac{\partial \hat{h}_i}{\partial \lambda_j} \frac{\partial \lambda_j}{\partial \hat{h}_k} = \delta_{ik}$, the matrix $(\frac{\partial \hat{h}_i}{\partial \lambda_j})$ is inverse to $(\frac{\partial \lambda_j}{\partial \hat{h}_k})$. A simple computation gives

$$\frac{\partial \hat{h}_i}{\partial \lambda_j} = \mathrm{E}_{O,\bar{h}}(h_i h_j) - \mathrm{E}_{O,\bar{h}}(h_i)\, \mathrm{E}_{O,\bar{h}}(h_j)\,.$$

which is the covariance of $h_i(X)$ and $h_j(X)$. ∎

Theorem 2: If the origin O of the exponential family $\mathcal{W}$ is chosen so that

$$\mathrm{Cov}_{O,\hat{h}}(h) \ge \mathrm{Cov}_{S_\theta,\hat{h}}(h) \tag{19}$$

for every $\theta \in \mathcal{T}$ and every possible $\hat{h}$, then the inequality (14) is satisfied for every $\theta \in \mathcal{T}$ and every possible $\bar{h}$.

The matrix inequality $C \ge C'$ here means that $C - C'$ is positive semidefinite.

Proof: The proposition follows immediately from Lemmas 3 and 4. ∎

6.3 Location of the Origin

Note that all we need to check about the exponential family $\mathcal{W}$ are the first two moments of $h(X)$, i.e., $\mathrm{E}_{O,\hat{h}}(h)$ and $\mathrm{Cov}_{O,\hat{h}}(h)$ for all possible $\hat{h}$. But these moments represent a dual parameterization of an augmented exponential family

$$W_{\mu,\lambda}(x) = \exp\big(\lambda'\, h(x)\big)\, \exp\big(h'(x)\, \mu\, h(x)\big) \tag{20}$$

where λ is an n-vector and μ is a symmetric $(n \times n)$-matrix. Thus, trying to make $\mathrm{Cov}_{O,\hat{h}}(h)$ large enough, we can search for a proper origin O only within the above family. This greatly simplifies design of the approximating family $\mathcal{W}$.

7 Illustrative Examples

Three rather different problems illustrate the proposed construction of the approximating family $\mathcal{W}$.

7.1 Mixture Family in a Low Dimension

Suppose that x takes on just three possible values, i.e., $\mathcal{X} = \{1, 2, 3\}$. Let the sampling distribution belong to a mixture family composed of the probability mass functions

$$S_\theta(x) = (1 - \theta)\, S_0(x) + \theta\, S_1(x) \tag{21}$$

where $S_0 \equiv [0.1, 0.5, 0.4]$, $S_1 \equiv [0.5, 0.1, 0.4]$ and $\theta \in [0, 1]$. Suppose that a data sequence $\mathbf{x}$ is observed with the empirical distribution $R_{\mathbf{x}} \equiv [0.4, 0.5, 0.1]$. Let the only information available about data be a difference between the number of occurrences of 1 and 2 in the sample, i.e., let $h \equiv [1, -1, 0]$ (cf. Fig. 1).

Three approximating exponential families $\mathcal{W} = \{W_{O,\lambda}\}$ for different origins $O_1 = [0.2, 0.2, 0.6]$, $O_2 = [0.3, 0.3, 0.4]$, $O_3 = [0.4, 0.4, 0.2]$ are shown as dashed curves in Fig. 1.

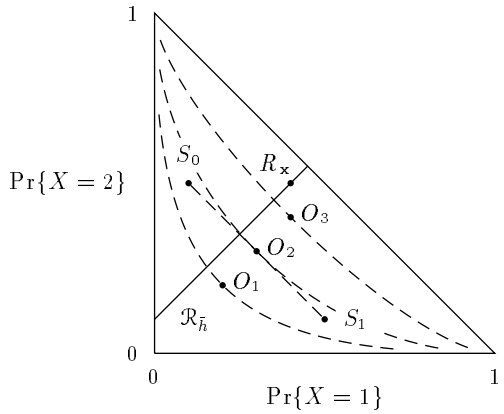

Fig. 1. A probability simplex with the original — mixture family (solid line connecting S_0 and S_1) and three approximating — exponential families (dashed lines). The approximate distributions are produced by projecting points of the mixture family onto the exponential families in the direction parallel to the line $\mathcal{R}_{\bar{h}}$.

The relative entropies for these approximating families are compared in Fig. 2 against the ideal solution. Clearly, it is O_2 that gives the best performance. With O_1 the inequality (14) is violated. O_3 satisfies (14) but results in an unnecessarily large difference from the ideal solution. The optimal approximating family is thus the exponential family tangent to (i.e., touching but not crossing) the original mixture family.

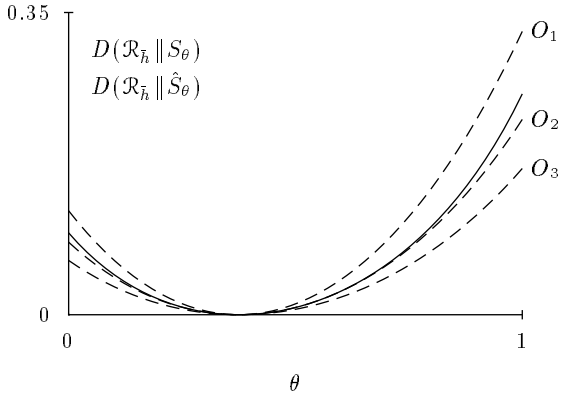

Fig. 2. The relative entropies $D(\mathcal{R}_{\bar{h}} \| \hat{S}_\theta) = D(W_{O,\bar{h}} \| W_{O,\hat{h}})$ for three different origins O (dashed lines) compared with the relative entropy $D(\mathcal{R}_{\bar{h}} \| S_\theta) = D(W_{S_\theta,\bar{h}} \| W_{S_\theta,\hat{h}})$ for the original model (solid line).

This empirical observation can be supported by the formal analysis proposed in Section 6. The optimal origin $O_{\mu,\lambda}$ is taken from the augmented exponential family (20) and chosen so that the function $\mathrm{E}_{O,\lambda}(h) \mapsto \mathrm{Var}_{O,\lambda}(h^2)$ is the minimal upper bound of the functions $\mathrm{E}_{S_\theta,\lambda}(h) \mapsto \mathrm{Var}_{S_\theta,\lambda}(h^2)$ for all $\theta \in \mathcal{T}$ (see Fig. 3). Numerical optimization as well as explicit computation gives $\mu^* = \log(3/4)$. One easily verifies that $O_{\mu^*,\lambda=0}$ coincides with O_2.

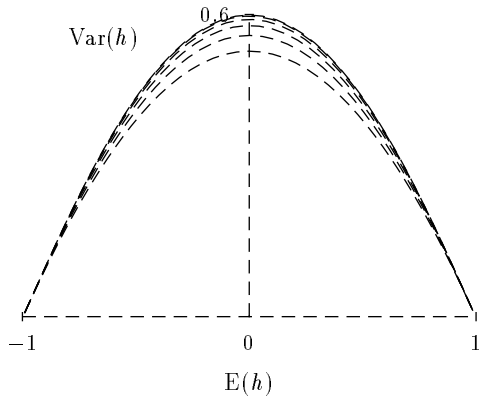

Fig. 3. The variance of $h(X)$ plotted against the mean of $h(X)$ for exponential distributions $W_{S_\theta,\lambda}$ (dashed lines correspond to $\theta = 0, 0.1, \ldots, 0.9, 1$) and $W_{O_2,\lambda}$ (solid line).

7.2 Gaussian Family

Suppose $\mathcal{S} = \{s_\theta\}$ is composed of univariate Gaussian densities with a mean θ and a variance $\sigma^2(\theta)$ *dependent* on the mean. Let the only information available about data be the value of the sample average $\bar{x}$, i.e., let $h(x) = x$.

We apply the proposed construction of the approximating family $\mathcal{W}$ to this case only formally as data are continuous now and $\mathcal{X} = \mathbb{R}$. Alternatively, we could treat the problem as a limit case for a large number of grid points over a real line.

Since $h(x) = x$, the exponential densities going through particular sampling densities s_θ are Gaussian again

$$\begin{aligned} w_\lambda(x) &\propto s_\theta(x)\, \exp(\lambda x) \\ &\sim \mathrm{N}(\theta + \sigma^2(\theta)\lambda, \sigma^2(\theta)) \,. \end{aligned}$$

The augmented exponential family (20) is Gaussian as well

$$\begin{aligned} w_\lambda(x) &\propto \exp\left(\lambda x + \mu x^2\right) \\ &\sim \mathrm{N}(-\tfrac{\lambda}{2\mu}, -\tfrac{1}{2\mu}) \,. \end{aligned}$$

Thus, to satisfy the requirement (19), one needs only to set the variance $-\frac{1}{2\mu}$ of the approximating family $\mathcal{W}$ to the maximum of variances $\sigma^2(\theta)$ over all θ.

The result agrees with our intuition. The price we pay for the model approximation is a possible increase of the variance of X.

It is easy to compute that the (integral) relative entropy between two Gaussian densities of the same variance σ^2 but different means θ_1, θ_2 is

$$D(s_{\theta_1} \| s_{\theta_2}) = \frac{1}{2\sigma^2} (\theta_1 - \theta_2)^2 \,.$$

Thus, increasing the variance σ^2 implies decreasing the relative entropy value. This proves directly that the covariance inequality (19) implies the relative entropy inequality (14).

A multivariate version of the above procedure for a vector x easily follows. To satisfy (14), the covariance matrix C^* of the approximating family $\mathcal{W}$ has to be greater or equal to $C(\theta)$ for all θ.

7.3 Discrete Student-like Family

Let the sampling distribution be a discrete Student-like distribution with 3 degrees of freedom

$$S_\theta(x) \propto \left(1 + \frac{(x-\theta)^2}{3}\right)^2$$

defined on a discrete equidistant grid $\mathcal{X}$ over the interval $(-30, 30)$. The parameter θ is supposed to take values in the interval $(-10, 10)$.

The statistic considered is a function of the minimal sufficient statistic, namely generated by the score function at $\theta = 0$ (see [10] for more details)

$$h(x) = \frac{\mathrm{d}}{\mathrm{d}\theta} \log S_\theta(x)\big|_{\theta=0} = \frac{-4x}{3 + x^2} \,. \qquad (22)$$

The parameter $\mu = 3.8$ of the augmented exponential family (20) is found again from the covariance condition (19) (see

Fig. 4). The relative entropies $D(R_{\mathbf{x}}\|S_\theta)$, $D(\mathcal{R}_{\bar{h}}\|S_\theta)$ and $D(\mathcal{R}_{\bar{h}}\|\hat{S}_\theta)$ are compared in Fig. 5. Because of the extremely low dimension of the used statistic, a good fit of the approximate solution can be expected only close to the global minimum and close to the point $\theta = 0$ where the derivative in (22) is taken. The monotonicity relationship (15) holds, however, globally.

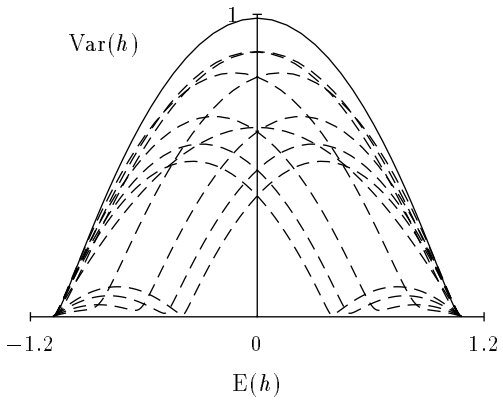

Fig. 4. The variance of $h(X)$ plotted against the mean of $h(X)$ for exponential distributions $W_{S_\theta,\lambda}$ (dashed lines correspond to $\theta = -10, -8, \ldots, 8, 10$) and $W_{O_\mu,\lambda}$ (solid line).

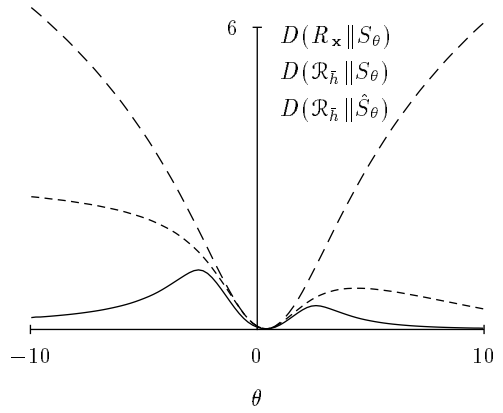

Fig. 5. Comparison of relative entropies for the case of complete data (long-dashed line), incomplete data but true model (short-dashed line), incomplete data and approximate model (solid line).

8 Concluding Remarks

The paper indicates a rather surprising possibility to build a global approximation of a given model. 'Global' here means that the approximation is performed *a priori* — before really starting estimation. This enables the user to analyse and judge the effects of approximation on estimation in advance.

The key point in the proposed approximation scheme is the restriction (14) which formalizes the intuitive requirement to make the uncertainty of estimation depend on the imprecision of its computational implementation. The restriction (14) pushes the relative entropy values closer to zero, making thus the corresponding models more difficult to discriminate.

Much of the presented approximation scheme can be generalized to the case of dependent data (cf. [3]), yet the extension is not straightforward and waits for more investigation.

Acknowledgments

The author thanks the anonymous referees for their helpful and stimulating comments. A discussion with F. Hrnčíř on Lemma 3 is gratefully acknowledged.

References

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

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

[3] R. Kulhavý and F. Hrnčíř, "Approximation and uncertainty in parameter estimation", in *Preprints of the European IEEE Workshop on Computer-intensive Methods in Control and Signal Processing*, Prague, Czech Republic, 1994, pp. 61–70.

[4] N. N. Čencov, *Statistical Decision Rules and Optimal Inference*, vol. 53 of *Transl. of Math. Monographs*, Amer. Math. Soc., 1982.

[5] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems", *Ann. Probab.*, **3** (1975), pp. 146–158.

[6] S. Amari, *Differential-Geometrical Methods in Statistics*, vol. 28 of *Lecture Notes in Statistics*, Springer-Verlag, second edition, 1990.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, 1991.

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

[9] M. S. Bazaraa and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, Wiley, 1979.

[10] R. Kulhavý, "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, 1993.