On-line Nonlinear Estimation

Rudolf Kulhavý

Honeywell Technology Center Europe and
Institute of Information Theory and Automation, AS CR
Pod vodárenskou věží 4, 182 08 Prague, Czech Republic
kulhavy@htc.honeywell.cz

Abstract

The Bayesian identification of complex models is known to require extensive (possibly infinite) computer resources. Practical implementation requires approximation of the theoretically optimal solution. The paper discusses three major approaches to approximate Bayesian estimation—local weighting of data, reduction of model family to "representative" points and minimum relative entropy restoration of the information divergence of the empirical and model distributions of data.

1. Introduction

The lack of an analytical solution to Bayesian identification of non-linear or non-Gaussian models gave birth to a multitude of approximation algorithms. It is often difficult for the user to compare the existing options and make an appropriate choice.

Additional insight can be gained from the fact that estimation itself is an approximation problem. As shown in [1], the posterior density is determined by the information divergence of the empirical and model distributions of observed data (Section 2). The empirical distribution, the family of model distributions and the information divergence between the empirical and model distributions represent three major ingredients of the estimation problem. Any approximation of the Bayesian paradigm necessarily affects at least one of these objects (Section 3).

Following this pattern, we present three approximation methods—based on local weighting of observed data (Section 4), reduction of the model family (Section 5) and direct approximation of the information divergence (Section 6). The approaches are shown to address different application scenarios and to complement in a sense each other (Section 7).

The results presented in the paper are stated without proofs; the reader interested in details is referred to the references.

2. Bayesian Identification

The fundamental problem of system identification is to select a proper model that describes best the behaviour of a *dynamic* and possibly *controlled* system.

2.1. Model Family

We shall consider a controlled system on which two sequences of continuous random variables are measured

$$\begin{aligned} Y^{N+m} &= (Y_1, \ldots, Y_{N+m}), \\ U^{N+m} &= (U_1, \ldots, U_{N+m}), \end{aligned}$$

which take values in subsets $\mathcal{Y}$ and $\mathcal{U}$ of $\mathbb{R}^{\dim y}$ and $\mathbb{R}^{\dim u}$, respectively. U_k is the directly manipulated input to the system at time k. Y_k is the the system output—the response at time k to the past history of data represented by the sequences Y^{k-1} and U^k. The two sequences make up a *sample* of data whereas the sequences of observed or measured values

$$\begin{aligned} y^{N+m} &= (y_1, \ldots, y_{N+m}), \\ u^{N+m} &= (u_1, \ldots, u_{N+m}) \end{aligned}$$

form a *realization* of the sample Y^{N+m}, U^{N+m}.

From now on we suppose that the output values Y_k depend on just a limited amount of past data, namely Y_{k-m}^{k-1}, U_{k-m}^{k}, through a known vector function $Z_k = z(U^k, Y^{k-1})$ taking values in a subset $\mathcal{Z}$ of $\mathbb{R}^{\dim z}$. Hence, we suppose that

$$s_k(y_k \,|\, y^{k-1}, u^k) = s_k(y_k \,|\, z_k) \tag{1}$$

for $k = m+1, \ldots, N+m$. Furthermore, we assume that the conditional distribution of Y_k given $Z_k = z_k$ is identical for all k, i.e.,

$$s_k(y \,|\, z) = s(y \,|\, z).$$

Finally, we assume that (y_k, z_k) is recursively computable given its last value (y_{k-1}, z_{k-1}) and the latest data (y_k, u_k), i.e., there exists a map F such that

$$(y_k, z_k) = F\big((y_{k-1}, z_{k-1}), (y_k, u_k)\big).$$

The density $s(y|z)$ is supposed to come from a model family

$$\mathcal{S} = \{s_\theta(y|z) : \theta \in \mathcal{T}\} \tag{2}$$

parameterized by the parameter θ taking values in a subset $\mathcal{T}$ of $\mathbb{R}^{\dim\theta}$. We restrict ourselves to the case that $s_\theta(y|z) > 0$ for all $(y,z) \in \mathcal{Y} \times \mathcal{Z}$ and all $\theta \in \mathcal{T}$.

The objective of parameter estimation is to find a proper value of the parameter θ given the observed sample y^{N+m}, u^{N+m}.

2.2. Bayes's Theorem

The dependence of the input U_k on the past data Y^{k-1}, U^{k-1} and the parameter θ is most generally described by a conditional density $\gamma_k(u_k|y^{k-1}, u^{k-1}, \theta)$. Mostly, we can adopt a simplifying assumption [2] that the only information about θ used for computation of the new input is the information contained in the past data, i.e.,

$$\gamma_k(u_k|y^{k-1}, u^{k-1}, \theta) = \gamma_k(u_k|y^{k-1}, u^{k-1}) \tag{3}$$

for $k = m+1, \ldots, N+m$.

With the unknown parameter θ interpreted as a random variable Θ, the parameter uncertainty is described by the *posterior* density conditional on the observed sample y^{N+m}, u^{N+m}

$$p_N(\theta) \triangleq p(\theta|y^{N+m}, u^{N+m}).$$

The subscript N indicates conditioning on N data points $(y_{m+1}, z_{m+1}), \ldots, (y_{N+m}, z_{N+m})$.

Given a prior density conditional on available *a priori* information and possibly m initial values y^m, u^m

$$p_0(\theta) \triangleq p(\theta|y^m, u^m),$$

the posterior density $p_N(\theta)$ follows by Bayes's theorem [3]. The model assumptions (1)–(3) yield

$$p_N(\theta) \propto p_0(\theta) \prod_{k=m+1}^{N+m} s_\theta(y_k|z_k)\ \gamma_k(u_k|y^{k-1}, u^{k-1})$$

where $\propto$ stands for equality up to the normalizing factor. After eliminating the parameter-independent part, we obtain the formula

$$p_N(\theta) \propto p_0(\theta) \prod_{k=m+1}^{N+m} s_\theta(y_k|z_k). \tag{4}$$

When convenient, the posterior density can be calculated recursively for $k = 1, \ldots, N$

$$p_k(\theta) \propto p_{k-1}(\theta)\, s_\theta(y_{k+m}|z_{k+m}). \tag{5}$$

2.3. Information View

A *joint empirical density* of (Y, Z) for a given sample y^{N+m}, u^{N+m} is defined as

$$r_N(y,z) = \frac{1}{N} \sum_{k=m+1}^{N+m} \delta(y - y_k, z - z_k) \tag{6}$$

where $\delta(\cdot,\cdot)$ is a Dirac function satisfying $\delta(y,z) = 0$ for $y \neq 0$ or $z \neq 0$ and

$$\iint_{\mathcal{Y}\times\mathcal{Z}} \delta(y,z)\,\mathrm{d}y\,\mathrm{d}z = 1.$$

Similarly as with the posterior density, we use the subscript N to indicate the number of data points $(y_{m+1}, z_{m+1}), \ldots, (y_{N+m}, z_{N+m})$ the empirical density is based on.

A *conditional inaccuracy* of the (joint) empirical density $r_N(y,z)$ relative to the (conditional) model density $s_\theta(y|z)$ is defined as

$$K(r_N{:}\, s_\theta) \triangleq \iint r_N(y,z)\, \log \frac{1}{s_\theta(y|z)}\,\mathrm{d}y\,\mathrm{d}z. \tag{7}$$

The conditional inaccuracy generalizes Kerridge's inaccuracy [4] introduced for independent and identically distributed data.

It follows directly from the empirical density and conditional inaccuracy definitions that the *posterior* density (4) can be rewritten as

$$p_N(\theta) \propto p_0(\theta)\, \exp\bigl(-N\, K(r_N{:}\, s_\theta)\bigr). \tag{8}$$

3. Three Approaches to Approximation

For non-linear or non-Gaussian models $s_\theta(y|z)$, the conditional inaccuracy $K(r_N{:}\, s_\theta)$ can rarely be computed analytically. In the following sections, we outline three major approaches to approximate Bayesian inference.

1. *Replacement of the empirical density with its locally-weighted version* is a natural way of focusing on the local behaviour of data. The advantage is that relatively simple models are sufficient to describe the local behaviour, hence Bayesian estimation can often be performed analytically. But, except for special cases, the local weighting requires that any of the past data is retrievable on-line.

2. *Reduction of the model family to a finite model set* converts Bayesian estimation to calculation of posterior probabilities for the selected models. The models can be chosen beforehand or sampled randomly. The calculation can be performed recursively. The target features of the posterior distribution are explored via the corresponding features of the sample.

3. *Restoration of the information divergence between the empirical and model distributions from compressed data* is a natural option when estimation of complex models needs to be performed recursively. Additional constraints can be imposed on the restored divergence through information inequalities.

4. Local Weighting of Data

The local weighting of data results in replacing the true data statistic (N, r_N) with a locally-weighted statistic (ν_N, ρ_N). The weighted empirical density $\rho_N(y,z)$ puts more weight on the data points close to the time or regressor of interest and suppresses the points far away. The effective number of data ν_N is appropriately decreased, producing thus a more "flattened" posterior density

$$p_N(\theta) \propto p_0(\theta)\, \exp\big(-\nu_N\, K(\rho_N : s_\theta)\big). \qquad (9)$$

The purposeful modification of the data statistic $(N, r_N) \mapsto (\nu_N, \rho_N)$ is certainly a heuristic measure, but it extends significantly the scope of problems that can be addressed within the Bayesian framework.

4.1. Subsampling

A simple way of focusing on the recent data is to base estimation on just the n latest data points (y_k, z_k), $k = N+m-n+1, \ldots, N+m$. The statistic (ν_N, ρ_N) takes in this case the form

$$\begin{aligned} \nu_N &= n, \\ \rho_N(y,z) &= \frac{\sum_{k=N+m-n+1}^{N+m} \delta(y-y_k, z-z_k)}{n}. \end{aligned}$$

In sequential estimation, data come from a moving time window [5].

4.2. Exponential Discounting

Another natural way of focusing on the recent data is to make the weight on the data point (y_k, z_k) exponentially decreasing with its "age" $N+m-k$. In particular, the data point (y_k, z_k) is assigned the weight λ^{N+m-k} where $\lambda \in (0,1)$ acts as a discounting or forgetting factor [6]. With this choice, the statistic (ν_N, ρ_N) modifies as follows,

$$\begin{aligned} \nu_N &= \textstyle\sum_{k=m+1}^{N+m} \lambda^{N+m-k}, \\ \rho_N(y,z) &= \frac{\sum_{k=m+1}^{N+m} \lambda^{N+m-k}\, \delta(y-y_k, z-z_k)}{\sum_{k=m+1}^{N+m} \lambda^{N+m-k}}. \end{aligned}$$

Note that $\nu_N \to \frac{1}{1-\lambda}$ as $N \to \infty$.

A special advantage of exponential discounting is that the statistic (ν_k, ρ_k) can be updated recursively.

4.3. Kernel-Based Discounting

The age discounting of data can be extended to a general weighting profile. Consider a kernel function $K(x)$ that equals to 1 at $x = 0$ and decreases to 0 as x increases. Examples of such functions include the Gaussian kernel $K(x) = \exp(-x^2)$ or the Epanechnikov kernel $K(x) = \max\{1-x^2, 0\}$.

We assign the weight $K\big(\frac{k-k^*}{h}\big)$ to the data point (y_k, z_k) where k^* is the time instant of interest. The scalar h is a smoothing factor that determines how quickly the weight on the data point approaches zero as $|k-k^*|$ increases. With this choice, the statistic becomes

$$\begin{aligned} \nu_N &= \textstyle\sum_{k=m+1}^{N+m} K\big(\frac{k-k^*}{h}\big), \\ \rho_N(y,z) &= \frac{\sum_{k=m+1}^{N+m} K\big(\frac{k-k^*}{h}\big)\, \delta(y-y_k, z-z_k)}{\sum_{k=m+1}^{N+m} K\big(\frac{k-k^*}{h}\big)}. \end{aligned}$$

The kernel-based discounting cannot be implemented recursively but it is far more flexible than the exponential discounting. When k^* is a time instant back in the process history, data from a *two-sided neighborhood* of k^* enter estimation.

4.4. Locally Weighted Smoothing

The prediction $s(y|z^*)$ for a particular value z^* of the regressor vector can often be based on just the data points within a neighbourhood of z^*. This is the bottom line of non-parametric regression [7], [8], which discounts the data according to "similarity" of the current regressor z_k and the regressor of interest z^*.

Consider a kernel function $K(x)$ of a vector argument x. Suppose that $K(0) = 1$ and $K(x)$ decreases to 0 as $\|x\|$ increases. Examples of such functions are the Gaussian kernel $K(x) = \exp(-x^T x)$ or the Epanechnikov kernel $K(x) = \min\{1 - x^T x, 0\}$.

We assign to the data point (y_k, z_k) the weight $K(\|z_k - z^*\|_H)$ dependent on the Euclidean distance

$$\|z_k - z^*\|_H = (z_k - z^*)^T H^{-1} (z_k - z^*).$$

The shape of the kernel is defined by the symmetric, positive definite matrix H.

With the above choice of the weighting profile, the statistic takes the form

$$\begin{aligned} \nu_N &= \textstyle\sum_{k=m+1}^{N+m} K(\|z_k - z^*\|_H), \\ \rho_N(y,z) &= \frac{\sum_{k=m+1}^{N+m} K(\|z_k - z^*\|_H)\, \delta(y-y_k, z-z_k)}{\sum_{k=m+1}^{N+m} K(\|z_k - z^*\|_H)}. \end{aligned}$$

5. Reduction of Model Family

The model family can be reduced to a finite model set in two different ways. First, we can prespecify

a selection of parameter values and compute (possibly recursively) the posterior probabilities for all of them. Second, we can draw a finite sample of parameter values from the posterior density. The former usually suffers from the "curse of dimensionality" in higher dimensions. The Monte Carlo simulation is then the only solution.

5.1. Multiple-Model Estimation

For lower dimensions of the parameter space, evaluation of the posterior density over a finite grid of parameter values is a natural way of getting an idea about the shape of the posterior density. The method is known as a point-mass approximation in engineering science [9]. The parameter grid is either fixed or adapted recursively so as to cover the parameter region of high posterior probability.

5.2. Non-Iterative Monte Carlo Sampling

In non-iterative Monte Carlo simulation, we draw a sample of θ from an approximate distribution $\pi(\theta)$ and then correct the draw so as to make it closer the target posterior distribution $p_N(\theta)$. The algorithm is known as sampling-importance resampling scheme [10] or weighted bootstrap [11].

The algorithm proceeds as follows.

1. Draw a sample $\theta^{(1)}, \ldots, \theta^{(M)}$ from an approximate density $\pi(\theta) \approx p_N(\theta)$.
2. For $j = 1, \ldots, M$ calculate
$$w_j = \frac{p_N(\theta^{(j)})}{\pi(\theta^{(j)})}, \quad q_j = \frac{w_j}{\sum_{l=1}^{M} w_l}.$$
3. Draw θ^* from the discrete distribution that assigns probability q_j at $\theta^{(j)}$ for $j = 1, \ldots, M$.

The algorithm generates θ^* that is distributed approximately according to $p_N(\theta)$ whereas the approximation is improving with increasing M. The posterior density $p_N(\theta)$ and the approximate density $\pi(\theta)$ are sufficient to know with precision up to the normalizing constant.

The non-iterative sampling is ideally suited for recursive Bayesian estimation (5) where the approximate density is set as $\pi(\theta) = p_{k-1}(\theta)$.

The recursive algorithm proceeds as follows.

1. Draw a sample $\theta^{(m,1)}, \ldots, \theta^{(m,M)}$ from the prior density $p_0(\theta) = p(\theta | y^m, u^m)$.
2. For $k = m+1, \ldots, N+m$:
   (a) For $j = 1, \ldots, M$, calculate
$$q_j = \frac{s_{\theta^{(k-1,j)}}(y_k | z_k)}{\sum_{l=1}^{M} s_{\theta^{(k-1,l)}}(y_k | z_k)}.$$
   (b) Draw a sample $\theta^{(k,1)}, \ldots, \theta^{(k,M)}$ from the discrete distribution that places mass q_j at $\theta^{(k-1,j)}$ for $j = 1, \ldots, M$.

Due to the unequal weighting q_j, some points are resampled more often than the others. The number of different values within the sample $\theta^{(k,1)}, \ldots, \theta^{(k,M)}$ thus decreases with the increasing number of observations. In a recursive setting, the sample may quickly degenerate to a couple of points within the region of high posterior probability. A possible remedy is to sample from a kernel-smoothed approximation to the density $p_k(\theta)$ or, equivalently, add a jitter to the samples drawn [12].

5.3. Iterative Monte Carlo Sampling

In iterative Monte Carlo simulation, the samples are drawn sequentially so that the distribution of the next draw depends on the last value drawn. Hence, the draws form a Markov chain. Several algorithms built on this idea are used in practice.

Metropolis Algorithm: The Metropolis algorithm was proposed in the early 1950s [13]. A closely related algorithm was proposed three decades later in global optimization—as simulated annealing [14].

The algorithm proceeds as follows.

1. Draw a starting point $\theta^{(0)}$ from a proper starting distribution concentrated around the mode of the posterior density.
2. For $j = 1, 2, \ldots$:
   (a) Sample a candidate point θ^* from a jumping density $\pi(\theta^* | \theta^{(j-1)})$. The jumping distribution must be symmetric so that $\pi(\theta | \theta') = \pi(\theta' | \theta)$ for all θ, θ'.
   (b) Calculate the density ratio
$$w = \frac{p_N(\theta^*)}{p_N(\theta^{(j-1)})}.$$
   (c) Set
$$\theta^{(j)} = \begin{cases} \theta^* & \text{with probability } \min\{w, 1\}, \\ \theta^{(j-1)} & \text{otherwise}. \end{cases}$$

A simple example of the underlying Markov process is a random walk driven by a zero-mean, normally-distributed, white noise.In order to compute the relative importance ratio w, it is sufficient to know the posterior density $p_N(\theta)$ with precision up to the normalizing constant.

The efficiency of the Metropolis algorithm is determined by the ratio of accepted and all generated

samples. This depends on how well the underlying Markov chain explores the regions of high posterior probability. Both too small and too large variances of the driving noise may result in inefficient sampling.

Metropolis-Hastings Algorithm: Hastings [15] generalized the Metropolis algorithm by allowing also asymmetric jumping distributions $\pi(\theta|\theta')$. To correct for the asymmetry in the jumping rule, the relative importance ratio is replaced by the ratio of importance ratios

$$w = \frac{p_N(\theta^*)/\pi(\theta^*|\theta^{(j-1)})}{p_N(\theta^{(j-1)})/\pi(\theta^{(j-1)}|\theta^*)}.$$

In the Metropolis-Hastings algorithm, the candidate samples can be taken even from a fixed density [16].

1. Draw a starting point $\theta^{(0)}$ from a proper starting distribution.
2. For $j = 1, 2, \ldots$:
   (a) Sample a candidate point θ^* from the approximate density $\pi(\theta)$.
   (b) Calculate the density ratio
   $$w = \frac{p_N(\theta^*)/\pi(\theta^*)}{p_N(\theta^{(j-1)})/\pi(\theta^{(j-1)})}.$$
   (c) Set
   $$\theta^{(j)} = \begin{cases} \theta^* & \text{with probability } \min\{w, 1\}, \\ \theta^{(j-1)} & \text{otherwise}. \end{cases}$$

Clearly, the closer the approximate density $\pi(\theta)$ is to the posterior density $p_N(\theta)$, the closer the number of accepted samples is to the total number of generated samples.

Gibbs Sampler: The concept of alternating conditional sampling was introduced first in the image processing literature [17]. In the early 1990s, it was applied with remarkable success in Bayesian statistics [18].

To illustrate how the Gibbs sampler works, we consider a parameter vector composed of three entries only, $\theta = (\theta_1, \theta_2, \theta_3)$. The algorithm proceeds in this case as follows.

1. Draw a starting point $\theta^{(0)}$ from a proper starting distribution.
2. For $j = 1, 2, \ldots$:
   (a) Draw $\theta_1^{(j)}$ from $p_N(\theta_1|\theta_2^{(j-1)}, \theta_3^{(j-1)})$.
   (b) Draw $\theta_2^{(j)}$ from $p_N(\theta_2|\theta_1^{(j)}, \theta_3^{(j-1)})$.
   (c) Draw $\theta_3^{(j)}$ from $p_N(\theta_3|\theta_1^{(j)}, \theta_2^{(j)})$.

The algorithm extends straightforwardly to higher dimensions. When appropriate, the parameter vector can be subdivided into subvectors rather than scalar entries. At any rate, the complete conditionals must remain lower-dimensional and relatively easy to sample from.

The Gibbs sampler is a natural solution to estimation of hierarchical or highly structured models [19].

Other Monte Carlo Schemes: The Gibbs sampler can be combined with the Metropolis-Hastings algorithm where the latter serves as a general tool for sampling from the univariate full conditionals. The candidate samples are taken from a kernel-smoothed approximation to $p_N(\theta)$ based on its values over a grid of fixed points. The algorithm is known as a *griddy Gibbs sampler* [20].

An idea similar to the Gibbs sampler is used in the *hit-and-run algorithm* where sampling is made in randomly chosen directions rather than entry-by-entry [21].

Yet another possibility of drawing a sample from the posterior density $p_N(\theta)$ is solve the *Langevin stochastic differential equation* [22]

$$\mathrm{d}\theta^{(t)} = \mathrm{d}t \ \mathrm{grad} \log p_N(\theta^{(t)}) + \sqrt{2}\,\mathrm{d}v_t$$

driven by a multivariate Brownian motion v_t.

6. Restoration of Information Divergence

When data are compressed through a finite-dimensional statistic, only a set containing the true empirical density $r_N(y, z)$ is known. The empirical distribution can be restored approximately from this partial knowledge using the following generalization of the maximum entropy principle.

6.1. Data Compression

Consider a certain statistic of (Y, Z)

$$h\colon \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}^n.$$

Let the only information available about the empirical density $r_N(y, z)$ be the empirical expectation of $h(Y, Z)$

$$\bar{h}_N \triangleq \iint r_N(y, z)\, h(y, z)\, \mathrm{d}y\, \mathrm{d}z = \frac{1}{N} \sum_{k=m+1}^{N+m} h(y_k, z_k).$$

The empirical density $r_N(y,z)$ is thus only known to lie within the set

$$\mathcal{R}_N \triangleq \Big\{ r(y,z) : \iint r(y,z)\, h(y,z)\, \mathrm{d}y\, \mathrm{d}z = \bar{h}_N, \iint r(y,z)\, \mathrm{d}y\, \mathrm{d}z = 1,\ r(y,z) \geq 0 \Big\}. \quad (10)$$

The problem is to compute the conditional inaccuracy $K(r_N : s_\theta)$ as a function of θ given the above partial information about $r_N(y,z)$.

6.2. Information Geometry

For each density $s_\theta(y|z)$, $\theta \in \mathcal{T}$, we define an exponential family $\mathcal{S}_{\theta;h}$ composed of the joint densities

$$s_{\theta,\lambda}(y,z) = s_\theta(y|z)\, \exp\big(\lambda^T h(y,z) - \psi(\theta,\lambda)\big) \quad (11)$$

where $h(\cdot,\cdot)$ is a canonical statistic of the family, $\lambda \in \mathbb{R}^n$ is its natural parameter and

$$\psi(\theta,\lambda) = \log \iint s_\theta(y|z)\, \exp\big(\lambda^T h(y,z)\big)\, \mathrm{d}y\, \mathrm{d}z \quad (12)$$

is logarithm of the normalizing divisor.

Consider a sample y^{N+m}, u^{N+m} with the empirical density $r_N(y,z)$. The necessary condition for $\hat{\lambda}$ to minimize the unnormalized joint inaccuracy

$$K(r_N : s_{\theta,\lambda}) = \iint r_N(y,z)\, \log \frac{1}{s_{\theta,\lambda}(y,z)}\, \mathrm{d}y\, \mathrm{d}z$$

is $\nabla_\lambda K(r_N : s_{\theta,\hat{\lambda}}) = 0$. The condition reads

$$\iint s_{\theta,\hat{\lambda}}(y,z)\, h(y,z)\, \mathrm{d}y\, \mathrm{d}z = \iint r_N(y,z)\, h(y,z)\, \mathrm{d}y\, \mathrm{d}z = \bar{h}_N. \quad (13)$$

The density $s_{\theta,\hat{\lambda}}(y,z)$ that satisfies the condition (13) is called a *h-projection* of $r_N(y,z)$ onto $\mathcal{S}_{\theta;h}$.

Let $s_{\theta,\lambda}(y,z)$ be exponential (11) and $\hat{\lambda}$ satisfy (13). The following Pythagorean relationship holds

$$K(r_N : s_\theta) = K(r_N : s_{\theta,\hat{\lambda}}) + D(s_{\theta,\hat{\lambda}} \| s_\theta) \quad (14)$$

where

$$D(s_{\theta,\hat{\lambda}} \| s_\theta) = \iint s_{\theta,\hat{\lambda}}(y,z)\, \log \frac{s_{\theta,\hat{\lambda}}(y,z)}{s_\theta(y|z)}\, \mathrm{d}y\, \mathrm{d}z$$

stands for the *unnormalized* relative entropy [23].

The Pythagorean-like relationship (14) that links together inaccuracies and relative entropy was shown first in [24]. The formula can be regarded as generalization of the Pythagorean theorem shown in [25], [26], [27] to hold for relative entropies between probability distributions.

6.3. Approximation of Inaccuracy

The Pythagorean relationship (14) implies that if $D(s_{\theta,\hat{\lambda}} \| s_\theta)$ is finite, then the h-projection $s_{\theta,\hat{\lambda}}(y,z)$ is a unique solution to the minimum relative entropy problem

$$D(s_{\theta,\hat{\lambda}} \| s_\theta) = D(\mathcal{R}_N \| s_\theta) \triangleq \min_{r \in \mathcal{R}_N} D(r \| s_\theta). \quad (15)$$

The minimum relative entropy is found by solving a dual optimization problem

$$D(s_{\theta,\hat{\lambda}} \| s_\theta) = \max_\lambda \big(\lambda^T \bar{h}_N - \psi(\theta,\lambda) \big). \quad (16)$$

Suppose that through a proper choice of $h(y,z)$ (see [1] for details) we have ensured that $K(r_N : s_{\theta,\hat{\lambda}}) \approx C$ where C is a constant independent of θ. Then the Pythagorean relationship (14) suggests the following approximation of inaccuracy

$$K(r_N : s_\theta) \ \approx\ C + D(\mathcal{R}_N \| s_\theta). \quad (17)$$

Substituting from (17) for $K(r_N : s_\theta)$ in (8), we obtain the approximate posterior density

$$\hat{p}_N(\theta) \propto p_0(\theta)\, \exp\big(-N\, D(\mathcal{R}_N \| s_\theta)\big). \quad (18)$$

6.4. Implementation

The major difficulty in computation of (16) is multivariate integration that is required to evaluate logarithm of the normalizing divisor $\psi(\theta,\lambda)$. When the dimension of (y,z) is small, we can apply numerical integration over a grid of points in the data space $\mathcal{Y} \times \mathcal{Z}$. When the dimension of (y,z) is greater than 3–4, the grid size becomes impracticable. The Monte Carlo simulation is then the only possibility.

Below we outline a possible implementation of the Monte Carlo simulation. For brevity, we use the notation $x = (y,z)$.

1. Generate a sample $x^{(1)}, \ldots, x^{(M)}$ from the density $s_{\theta,\lambda}(x)$ using the Metropolis sampler with random walk chains:
   (a) Generate a candidate value x^* via random walk $x^* = x^{(j-1)} + v$ where v is a sample from the normal distribution $N(0,Q)$ with a suitable covariance Q.
   (b) Accept x^* (and set $x^{(j)} = x^*$) with probability $\min\{w,1\}$ where

$$w = \frac{s_{\theta,\lambda}(x^*)}{s_{\theta,\lambda}(x^{(j-1)})} = \frac{\exp\big(\lambda^T h(x^*)\big)}{\exp\big(\lambda^T h(x^{(j-1)})\big)}.$$

2. Construct a kernel-smoothed estimate $\hat{s}_{\theta,\lambda}(x)$ of the exponential density $s_{\theta,\lambda}(x)$ [28].

3. Approximate logarithm of the normalizing divisor by

$$\hat{\psi}(\theta,\lambda) = \frac{1}{M}\sum_{j=1}^{M} \log \frac{f_{\theta,\lambda}(x^{(j)})}{\hat{s}_{\theta,\lambda}(x^{(j)})} \qquad (19)$$

where $f_{\theta,\lambda}(x) = s_\theta(x)\,\exp\bigl(\lambda^T h(x)\bigr)$ stands for the unnormalized part of $s_{\theta,\lambda}(x)$.

The approximation (19) results from the assumption that the kernel estimate $\hat{s}_{\theta,\lambda}(x)$ is close enough to the true density $s_{\theta,\lambda}(x)$ in the relative entropy sense

$$D(s_{\theta,\lambda}\|\hat{s}_{\theta,\lambda}) = \int s_{\theta,\lambda}(x)\,\log\frac{f_{\theta,\lambda}(x)}{\hat{s}_{\theta,\lambda}(x)}\,\mathrm{d}x - \psi(\theta,\lambda) = \epsilon.$$

With this approximation, the algorithm (16) takes the form

$$D(s_{\theta,\hat{\lambda}}\|s_\theta) \approx \max_\lambda \bigl(\lambda^T \bar{h}_N - \hat{\psi}(\theta,\lambda)\bigr).$$

The optimization over λ must be done with care. The random variations of $\hat{\psi}(\theta,\lambda)$ (being dependent upon a particular sample) may easily cause numerical instability when the Newton-Raphson or similar algorithm is used. A more robust implementation is obtained when the values of $\hat{\psi}(\theta,\lambda)$ are explicitly smoothed (typically, using kernel regression) and the optimum is searched for using an annealing scheme.

7. Which Approximation?

The three approximation approaches outlined in the paper have no clear winner. The decision which approximation to use depends on the nature of a particular application.

Small Samples: Laboratory analysis, clinical tests, preventive maintenance, fault diagnosis, insurance risk estimation, forecasting in transient economies—all work with a limited number of data but sophisticated, often hierarchical or otherwise structured models.

This scenario calls clearly for iterative Monte Carlo sampling. The Gibbs sampler, perhaps using the Metropolis-Hastings algorithm for sampling from the lower-dimensional full conditionals, is the algorithm of choice. Due to the lack of data, the prior information becomes crucially important.

Huge Data Sets: The technology of low-cost collection and storage of digital information has made huge repositories of data today's reality . With the current database management systems, one can retrieve quickly data at any time instant of the process history and any point of the (y,z)-data space. With all the past data available, the data itself can play now the role of model. We do not need to build and update a single compact model, rather we can build *ad hoc* models to help answer specific questions.

A typical application scenario is short-term (hour-, day- and month-ahead) forecasting where the regressor determines the points in the database that can explain the local data behavior. Relatively simple models can often do the job. For models linear in parameters, Bayesian estimation can be performed in a closed form. If a more complex model is necessary, iterative Monte Carlo sampling can be used to process the sample of data retrieved from the database.

Compressed Data: In recursive estimation, data compression is required to keep the dimension of the input information limited. In non-iterative Monte Carlo sampling, the data compression is implicit—all we carry from the previous step is the sample of parameter points. In this way, we propagate the "statistic" and the parameter sample at the same time, which is certainly an attractive feature. A proper smoothing of the resampled discrete distribution built over this sample is, however, necessary to avoid the sample degeneration. This heuristic step produces an additional uncertainty in approximate estimation.

When using an explicit data statistic, we can keep the uncertainty under control. Not only we know the set which the true empirical distribution belongs to, but if the statistic is chosen properly (as a *necessary* statistic, see [1]), then we can determine even the set where the ideal posterior density lies. The restoration of the posterior density from knowledge of the statistic only is, however, considerably more computer-intensive; it requires to solve a (generalized) maximum entropy problem for each (user-defined or sampled) parameter value. The computation involves a multivariate integration that can be effectively computed using iterative Monte Carlo sampling.

Note that the posterior density is restored only when needed—the restoration uses only the statistic value, which is updated recursively.

Acknowledgments

The work was supported in part by the Grant Agency of CR under Grant 102/97/0466 and the Academy of Sciences of CR under Grant A2075603.

References

[1] R. Kulhavý, *Recursive Nonlinear Estimation: A Geometric Approach*, vol. 216 of *Lecture Notes in Control and Information Sciences*. London: Springer-Verlag, 1996.

[2] V. Peterka, "Bayesian approach to system identification," in *Trends and Progress in System Identification* (P. Eykhoff, ed.), ch. 8, pp. 239–304, Elmsford, N.Y.: Pergamon, 1981.

[3] G. R. Grimmett and D. R. Stirzaker, *Probability and Random Processes*. Oxford: Oxford University Press, second ed., 1992.

[4] D. F. Kerridge, "Inaccuracy and inference," *J. Roy. Statist. Soc. Ser. B*, vol. 23, pp. 284–294, 1961.

[5] R. H. Middleton, G. C. Goodwin, D. J. Hill, and D. Q. Mayne, "Design issues in adaptive control," *IEEE Trans. Automat. Control*, vol. 33, no. 1, pp. 50–58, 1988.

[6] R. G. Brown, *Smoothing, Forecasting and Prediction of Discrete Time Series*. Englewood Cliffs, NJ: Prentice-Hall, 1962.

[7] R. Härdle, *Applied Non-parametric Regression*. Cambridge: Cambridge University Press, 1990.

[8] T. J. Hastie and R. J. Tibshirani, *Generalized Additive Models*. London: Chapman & Hall, 1990.

[9] R. S. Bucy and K. D. Senne, "Digital synthesis of non-linear filters," *Automatica — J. IFAC*, vol. 7, pp. 287–298, 1971.

[10] D. B. Rubin, "Using the SIR algorithm to simulate posterior distributions (with discussion)," in *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, eds.), pp. 395–402, Oxford: Clarendon Press, 1988.

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

[12] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, "A novel approach to nonlinear/non-Gaussian Bayesian state estimation," *Proc. IEE-F*, vol. 140, no. 2, pp. 107–113, 1993.

[13] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equations of state calculations by fast computing machines," *J. Chem. Phys.*, vol. 21, pp. 1087–1092, 1953.

[14] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, pp. 671–680, 1983.

[15] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[16] L. Tierney, "Markov chains for exploring posterior distributions," *Ann. Stat.*, vol. 22, pp. 1701–1762, 1994.

[17] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, no. 6, pp. 721–741, 1984.

[18] 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.

[19] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall, 1996.

[20] M. A. Tanner, *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. Berlin: Springer-Verlag, second ed., 1993.

[21] B. Schmeiser and M.-H. Chen, "General hit-and-run Monte Carlo sampling for evaluating multidimensional integrals," technical report, School of Industrial Engineering, Purdue University, 1991.

[22] P. J. Doll, J. D. Doll, and H. L. Friedman, "Brownian dynamics as smart Monte Carlo simulation," *J. Chem. Phys.*, vol. 69, pp. 4628–4633, 1978.

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

[24] R. Kulhavý, "A geometric approach to statistical estimation," in *Proceedings of the 34th IEEE Conference on Decision and Control*, vol. 2, (New Orleans, LA), pp. 1097–1102, 1995.

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

[26] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *Ann. Probab.*, vol. 3, no. 1, pp. 146–158, 1975.

[27] S. Amari, *Differential-Geometrical Methods in Statistics*, vol. 28 of *Lecture Notes in Statistics*. Berlin: Springer-Verlag, 1985.

[28] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall, 1986.