Projection Based Algorithms for Estimation of Complex Models

J. Andrýsek

Department of Adaptive Systems, Institute of Information Theory and Automation,
Academy of Sciences of the Czech Republic, Prague, Czech Republic

Abstract: The Bayesian methodology is a consistent tool for dealing with recursive parameter estimation. The general solution of the estimation problem is available, however, it is analytically tractable only for a limited class of models. Various approximations must be used to achieve recursive estimation of complex models. In this paper, we apply the Projection Based (PB) approximation to a simple but practical example of outlier filtration.

Keywords: Bayesian estimation, recursive estimation, approximate estimation, adaptive systems

I. Introduction

The choice of a suitable model is an essential step in control and decision making applications dealing with complex systems. One way to face complexity is the principle of adaptivity, i.e. the use of models which evolve during their use. The demand for adaptivity of the model leads to the recursive estimation of its parameters, i.e. permanent updating of its parameter estimates by new data. In other words, statistics describing estimates are corrected by newly acquired data.

The recursive Bayesian estimation [1] evaluates the posterior distribution of the parameters at time t as an update of the posterior distribution at time $t-1$ using the Bayes rule and the data acquired at time t. The recursion starts at $t=1$ with updating of the prior distribution which must be chosen before the estimation starts. The posterior distribution obtained by the Bayes rule may not be, however, analytically tractable and thus unsuitable for the next estimation step.

In such case, we seek an approximate recursive estimation algorithm. The principle of the Projection Based (PB) method is restriction of the approximate posterior density to a particular (well manipulable) class of probability densities. The optimal posterior pdf is found as the best projection of the intractable correct posterior pdf into this class.

II. Basic notions and notations

$\equiv$ means the equality by definition.

d_t is a data record at the discrete time t, $d(t) \equiv d_1, \ldots, d_t$.

$d(0)$ is an empty sequence and reflects just the prior information.

Θ is the unknown parameter, finite-dimensional vector.

f, π are the letters reserved for probability density functions(pdf).

$\propto$ is the proportion sign, $h \propto g$ means that function h equals to the function g up to the normalization.

$\mathcal{D}(\,||\,)$ means the Kullback-Leibler divergence [2]. This "distance" is familiarly used in Bayesian analysis as the measure how good the second pdf approximates the first pdf. For conciseness, the Kullback-Leibler divergence is referred to as the KL divergence. $\mathcal{D}\left(f \,\middle\|\, g\right) = \int f \log\left(\frac{f}{g}\right)$

III. Problem formulation

The task of recursive parameter estimation is to determine the posterior density $\pi_t(\Theta|d(t))$ based on the knowledge of

- last posterior density $\pi_{t-1}(\Theta|d(t-1))$
- new data record d_t
- model of the system $f(d_t|d(t-1),\Theta)$ parameterized by the unknown parameter Θ

The algorithm starts from prior pdf $\pi_0(\Theta) \equiv \pi_0(\Theta|d(0))$.

The standard Bayesian approach determines $\pi_t(\Theta|d(t))$ as

$$\pi_t(\Theta|d(t)) = \frac{f(d_t|d(t-1),\Theta)\pi_{t-1}(\Theta|d(t-1))}{\int f(d_t|d(t-1),\Theta)\pi_{t-1}(\Theta|d(t-1))d\Theta}. \tag{1}$$

Though the above formula looks quite simple, its repetitive use can lead to very complex pdfs. For example, if the system model is a sum of two pdfs, repetitive multiplication of such models lead to a posterior density in the form of a sum with number of its elements growing exponentially with number of data.

Our task is to design such a methodology, which: (i) yields tractable posterior estimates, and (ii) these estimates are as close to the correct Bayesian ones as possible.

IV. Problem solution

The general PB estimation [3] is defined as follows:

1. *Choose suitable class of probability density functions $\pi(\Theta|\bullet)$.*
2. *Select statistic $\mathcal{G}_0$ so that the pdf $\pi(\Theta|\mathcal{G}_0)$ reflects the prior information.*
3. *In each step of estimation find the statistic $\mathcal{G}_t$ in such a way that $\pi(\Theta|\mathcal{G}_t)$ is the best projection of the "correct Bayesian" pdf to the chosen class of pdfs $\pi(\Theta|\bullet)$.*

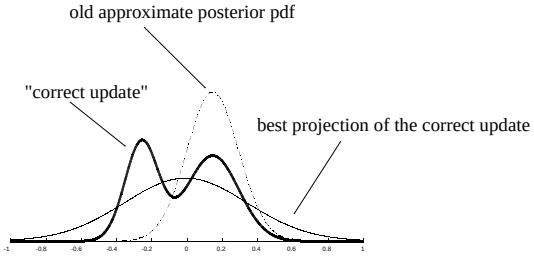

Figure 1: Basis of the approach

It remains to define what we mean by: "correct Bayesian" and "best projection". Let's have the approximate posterior pdf $\pi(\Theta|\mathcal{G}_{t-1})$, which reflects the information obtained from the data records before time t. If we handle this pdf as the true posterior pdf, we can perform one step of the Bayes rule (1), which results into the "correct Bayesian" posterior pdf $\hat{\pi}_t(\Theta)$. The term "best projection" means the closest in the sense of Kullback-Leibler divergence [2]. It means that we want to find $\mathcal{G}_t$ so that

$$\mathcal{D}\left(\hat{\pi}_t(\Theta) \,\middle|\middle|\, \pi(\Theta|\mathcal{G}_t)\right)$$

is minimized.

V. Application

Consider a real system generating many data records per second. The data contains a lot of outliers. The task is to to design a quick on-line filter.

The filtering task can be formulated as estimation of parameter Θ of Cauchy distribution, which describes well systems with outliers. Cauchy pdf doesn't have any moments, the parameter Θ is the median of this pdf. Intuitively, this fact suggests the use of the median filter, which is known to give good results. However, the median filter is not fast enough and yields only point estimates. We seek the Bayesian estimate of Θ.

Let's suppose that the system is modelled by a Cauchy pdf:

$$f(d_t|d(t-1),\Theta) = \frac{1}{\pi(1+(d_t-\Theta)^2)}.$$

The prior information on its parameter is: $\Theta \in (\Theta_{min}, \Theta_{max})$.

The Cauchy pdf doesn't have a conjugate pdf, hence it's parameter can not be estimated analytically using standard Bayesian approach. We will estimate it's parameter using PB estimation.

1. Selecting a suitable class of posterior pdfs

We have chosen to use Normal distribution with parameters μ, σ^2.

$$\pi(\Theta|\mathcal{G}_t) \equiv \pi(\Theta|\mu_t,\sigma_t^2) \equiv \mathcal{N}_\Theta(\mu_t,\sigma_t^2) = \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left(\frac{(\Theta-\mu_t)^2}{2\sigma_t^2}\right)$$

2. Specifying the prior distribution

The prior distribution must also be in the form chosen in the previous step. It remains to choose the parameters μ_0, σ_0^2, so that $\mathcal{N}_\Theta(\mu_0,\sigma_0^2)$ reflects the prior information. We choose $\mu_0 = \frac{\Theta_{min}+\Theta_{max}}{2}$, $\sigma_0^2 = (\frac{\Theta_{max}-\Theta_{min}}{3})^2$, which maps the given prior interval onto 3σ interval of the Normal pdf.

3. Performing one step of estimation

Using the selected class, one step of Bayesian estimation yields:

$$\hat{\pi}_t(\Theta) \propto f(d_t|d(t-1),\Theta)\pi(\Theta|\mathcal{G}_{t-1}) = \frac{1}{\pi(1+(d_t-\Theta)^2)} \frac{1}{\sqrt{2\pi\sigma_{t-1}^2}} \exp\left(-\frac{(\Theta-\mu_{t-1})^2}{2\sigma_{t-1}^2}\right) \equiv L_t(\Theta)$$

We seek minimum of $\mathcal{D}\left(\hat{\pi}_t(\Theta) \,\middle|\middle|\, \mathcal{N}_\Theta(\mu_t,\sigma_t^2)\right)$ with respect to μ_t, σ_t^2. Using Proposition 1 (Section VII.), the minimizing arguments equals to the mean and variance of $\hat{\pi}_t(\Theta)$ respectively:

$$\mu_t = \mathcal{E}\left[\Theta|\hat{\pi}_t(\Theta)\right], \;\; \sigma_t^2 = \mathcal{E}\left[\Theta^2|\hat{\pi}_t(\Theta)\right] - \mathcal{E}\left[\Theta|\hat{\pi}_t(\Theta)\right]^2$$

However, neither the pdf $\hat{\pi}_t(\Theta)$ nor its moments can be evaluated analytically in this case, because we are not even able to evaluate the normalizing integral $\int L_t(\Theta)d\Theta$. Because of the speed requirements on our algorithm, we can not solve the integrals numerically. Hence, we are looking for an approximation of $L_t(\theta)$ giving directly an approximation of $\hat{\pi}_t(\Theta)$.

We will use the Laplace approximation method, which is defined in terms of the first three elements of the Taylor series expansion of the logarithm of $L_t(\Theta)$ in its maximum. Hence, we need to find the maximum of

$$l_t(\Theta) \equiv \log\left(L_t(\Theta)\right) = \log\left(\frac{1}{\pi\sqrt{2\pi\sigma_{t-1}^2}}\right) - \log\left(1+(d_t-\Theta)^2\right) - \frac{(\Theta-\mu_{t-1})^2}{2\sigma_{t-1}^2}.$$

Because $\lim_{\Theta\to\pm\infty} l_t(\Theta) = -\infty$ and $l_t(\Theta)$ is continuous, we know that there exist a local maximum, which is also a global maximum. Hence, we can find the maximum as a point where the first derivative is equal to zero.

$$\begin{aligned} l_t'(\Theta) &= \frac{2(d_t-\Theta)}{1+(d_t-\Theta)^2} - \frac{(\Theta-\mu_{t-1})}{\sigma_{t-1}^2} = \frac{2(d_t-\Theta)\sigma_{t-1}^2 - (\Theta-\mu_{t-1})(1+(d_t-\Theta)^2)}{(1+(d_t-\Theta)^2)\sigma_{t-1}^2} \\ &= \frac{2d_t\sigma_{t-1}^2 - 2\sigma_{t-1}^2\Theta - [\Theta(1+d_t^2-2d_t\Theta+\Theta^2)] + \mu_{t-1}(1+d_t^2-2d_t\Theta+\Theta^2)}{(1+(d_t-\Theta)^2)\sigma_{t-1}^2} \\ &= \frac{-\Theta^3+\Theta^2(2d_t+\mu_{t-1}) + \Theta(-2\sigma_{t-1}^2-1-d_t^2-2d_t\mu_{t-1}) + 2d_t\sigma_{t-1}^2 + \mu_{t-1}(1+d_t^2)}{(1+(d_t-\Theta)^2)\sigma_{t-1}^2} \end{aligned}$$

We seek such Θ_t^M that fulfills $l'(\Theta_t^M) = 0$. It can be found as solution of cubic equation. According to Proposition 2 (Section VII.), this equation has at least one and at most three real solutions. If there are more than one real solution, we have to try, which of them gives greatest value of $l_t(\Theta)$. We denote the argument of maximum as Θ_t^M.

The Laplace approximation $\tilde{L}_t(\Theta)$ of $L_t(\Theta)$ is then found by approximating $l_t(\Theta)$:

$$\begin{aligned} l_t(\Theta) &\doteq l_t(\Theta_t^M) + \underbrace{l_t'(\Theta_t^M)}_{=0}(\Theta-\Theta_t^M) + 0.5 l_t''(\Theta_t^M)(\Theta-\Theta_t^M)^2, \\ l_t''(\Theta) &= \frac{-2(1+(d_t-\Theta)^2)+4(d_t-\Theta)^2}{(1+(d_t-\Theta)^2)^2} - \frac{1}{\sigma_{t-1}^2} = \frac{2(d_t-\Theta)^2-2}{(1+(d_t-\Theta)^2)^2} - \frac{1}{\sigma_{t-1}^2}, \\ L_t(\Theta) &= \exp\left(l_t(\Theta)\right) \doteq \exp\left(-\frac{(\Theta-\Theta_t^M)^2}{2(\frac{-1}{l_t''(\Theta_t^M)})} + l_t(\Theta_t^M)\right) \equiv \tilde{L}_t(\Theta). \end{aligned}$$

The approximation $\tilde{\pi}_t(\Theta)$ of $\hat{\pi}_t(\Theta)$ is then given by

$$\hat{\pi}_t(\Theta) = \frac{L_t(\Theta)}{\int L_t(\Theta)d\Theta} \doteq \frac{\tilde{L}_t(\Theta)}{\int \tilde{L}_t(\Theta)d\Theta} = \mathcal{N}_\Theta(\Theta_t^M, -\frac{1}{l_t''(\Theta_t^M)}) \equiv \tilde{\pi}_t(\Theta).$$

The searched parameters of PB-posterior pdf are simply:

$$\mu_t = \Theta_t^M, \sigma_t^2 = -\frac{1}{l_t''(\Theta_t^M)}.$$

The estimation method is summarized by the following algorithm.

Algorithm 1 (One step of PB estimation) *INPUTS:* $d_t, \mu_{t-1}, \sigma_{t-1}^2$, *OUTPUTS:* μ_t, σ_t^2

1. $a_0 = -2d_t\sigma_{t-1}^2 - \mu_{t-1}(1+d_t^2),\ a_1 = d_t^2 + 2d_t\mu_{t-1} + 1 + 2\sigma_{t-1}^2,\ a_2 = -2d_t - \mu_{t-1}$
2. $Q = \frac{3a_1-a_2^2}{9},\ R = \frac{9a_2a_1-27a_0-2a_2^3}{54},\ D = Q^3+R^2,\ S = \sqrt[3]{R+\sqrt{D}},\ T = \sqrt[3]{R-\sqrt{D}}$
3. $\mu_t = -\frac{1}{3}a_2 + (S+T)$
4. *IF* $D > 0$ *go to 8*
5. $\Theta_t^M = -\frac{1}{3}a_2 - \frac{1}{2}(S+T) + \frac{1}{2}i\sqrt{3}(S-T)$, *IF* $l(\Theta_t^M) > l(\mu_t)$ $\mu_t = \Theta_t^M$
6. *IF* $D = 0$ *go to 8*
7. $\Theta_t^M = -\frac{1}{3}a_2 - \frac{1}{2}(S+T) - \frac{1}{2}i\sqrt{3}(S-T)$, *IF* $l(\Theta_t^M) > l(\mu_t)$ $\mu_t = \Theta_t^M$
8. $\sigma_t^2 = \left(\frac{2(d_t-\mu_t)^2-2}{(1+(d_t-\mu_t)^2)^2} - \frac{1}{\sigma_{t-1}^2}\right)^{-1}$

VI. Simulated example

Consider the Cauchy distribution with $\Theta_{true} = 2,\ \Theta_{min} = -3,\ \Theta_{max} = 3$. The parameter Θ was estimated using the PB algorithm on 30 data records. The KL divergence of $\hat{\pi}_t(\Theta)$ and its approximation $\tilde{\pi}_t(\Theta)$ was numerically evaluated in each estimation step, and it is displayed in Figure 2 (left). Note that the divergence decreases rapidly towards a small positive value. This observation can be interpreted as follows: after some time, the PB estimation behave almost exactly as the Bayesian estimation. In general, however, the convergence of the PB estimation is not guaranteed. In this example, we have studied convergence of the method using numerical evaluation of the full Bayesian posteriors $\pi_t(\Theta)$. KL divergence of $\pi_t(\Theta)$ and $\tilde{\pi}_t(\Theta)$ is displayed in Figure 2 (right). The divergence decreases rapidly towards a small positive value. Hence, the approximate posteriors are almost identical with the full Bayesian posteriors after a few estimation steps.

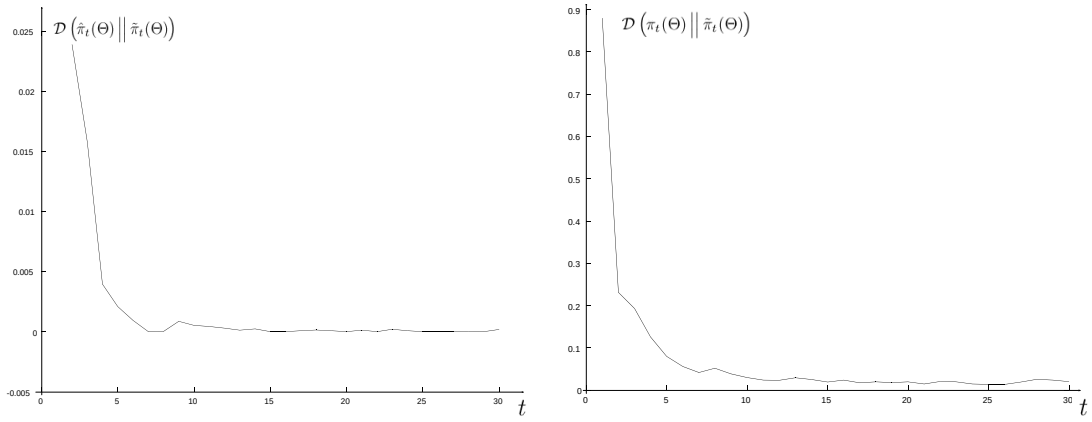

Figure 2: Convergence of the PB estimation

VII. Propositions used

Proposition 1 *Let $\hat{\pi}(\Theta)$ be an arbitrary pdf, which has first two moments defined. Let's denote them $\mathcal{E}_\Theta, \mathcal{E}_{\Theta^2}$. Then*

$$\arg\min_{\mu,\sigma^2} \mathcal{D}\left(\hat{\pi}(\Theta)\,\middle\|\,\mathcal{N}_\Theta(\mu,\sigma^2)\right) = (\mathcal{E}_\Theta, \mathcal{E}_{\Theta^2} - \mathcal{E}_\Theta^2).$$

Proof: We are minimizing

$$\int \hat{\pi}(\Theta) \log\left(\frac{\hat{\pi}(\Theta)}{\mathcal{N}_\Theta(\mu,\sigma^2)}\right) d\Theta.$$

By omitting the terms which doesn't influence the optimization, we can minimize $K(\mu,\sigma^2) =$

$$\begin{aligned} = -\int \hat{\pi}(\Theta) \log\left(\mathcal{N}_\Theta(\mu,\sigma^2)\right) d\Theta &= -\int \hat{\pi}(\Theta)\left(-0.5\log(2\pi) - 0.5\log\left(\sigma^2\right) - \frac{(\Theta-\mu)^2}{2\sigma^2}\right) d\Theta = \\ &= 0.5\log(2\pi) + 0.5\log\left(\sigma^2\right) + \frac{0.5}{\sigma^2}\int (\Theta-\mu)^2 \hat{\pi}(\Theta) d\Theta = \\ &= 0.5\log(2\pi) + 0.5\log\left(\sigma^2\right) + \frac{0.5}{\sigma^2}\left[\mathcal{E}_{\Theta^2} - 2\mu\mathcal{E}_\Theta + \mu^2\right]. \end{aligned}$$

$$\frac{\partial K(\mu,\sigma^2)}{\partial \mu} = \frac{0.5}{\sigma^2}(2\mu - 2\mathcal{E}_\Theta) \tag{2}$$

$$\frac{\partial K(\mu,\sigma^2)}{\partial \sigma^2} = \frac{0.5}{\sigma^2} - \frac{0.5}{(\sigma^2)^2}\left[\mathcal{E}_{\Theta^2} - 2\mu\mathcal{E}_\Theta + \mu^2\right] \tag{3}$$

By equaling the derivatives to zero, we get unique solution

$$\mu = \mathcal{E}_\Theta,\ \sigma^2 = \mathcal{E}_{\Theta^2} - \mathcal{E}_\Theta^2.$$

Now we need only to prove that the found extreme is a minimum. It can be simply proven evaluating the second derivative of $K(\mu, \sigma^2)$.

□

Proposition 2 *The equation* $x^3 + a_2x^2 + a_1x + a_0 = 0$ *has three solutions:*

$$x_0 = -\frac{1}{3}a_2 + (S+T) \tag{4}$$

$$x_1 = -\frac{1}{3}a_2 - \frac{1}{2}(S+T) + \frac{1}{2}i\sqrt{3}(S-T) \tag{5}$$

$$x_2 = -\frac{1}{3}a_2 - \frac{1}{2}(S+T) - \frac{1}{2}i\sqrt{3}(S-T) \tag{6}$$

where
$Q = \frac{3a_1 - a_2^2}{9},\ R = \frac{9a_2a_1 - 27a_0 - 2a_2^3}{54},\ D = Q^3 + R^2,\ S = \sqrt[3]{R + \sqrt{D}},\ T = \sqrt[3]{R - \sqrt{D}}$

Determining which roots are real and which are complex can be accomplished by noting that if the polynomial discriminant $D > 0$*, one root is real and two are complex conjugates; if* $D = 0$*, all roots are real and at least two are equal; and if* $D < 0$*, all roots are real and unequal.*

Proof: See [4]. □

VIII. Conclusions

In this paper, we have presented an algorithm for recursive estimation of parameters of Cauchy distribution using Projection Based (PB) estimation. The intractable posterior distribution was approximated by a Normal distribution. Performance of the algorithm was studied on a simulation study. It was shown, that the PB algorithm yields results very close to the results obtained by numerical aided Bayesian estimation. However, evaluation of the PB algorithm is much faster and therefore more suitable for on-line implementation.

Acknowledgment

This work was supported by AV ČR S1075351 and GA ČR 102/03/0049.

References

[1] V. Peterka, "Bayesian system identification," in *Trends and Progress in System Identification*, P. Eykhoff, Ed., pp. 239–304. Pergamon Press, Oxford, 1981.

[2] S. Kullback and R. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, 22:79–87, 1951.

[3] J. Andrýsek, "Projection Based Estimation of Dynamic Probabilistic Mixtures," Tech. Rep. 2098, ÚTIA AV ČR, Praha, 2004.

[4] E. W. Weisstein, *www.mathworld.com*, chapter Cubic Formula, http://mathworld.wolfram.com/CubicFormula.html.