ON IDENTIFICATION OF PROBABILISTIC MIXTURE MODELS WITH DYNAMIC WEIGHTS

Josef Andrýsek

Department of Adaptive Systems
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Prague, CZECH REPUBLIC
E-mail: andrysek@utia.cas.cz

Abstract: The probabilistic mixtures with constant weights provide a universal approximation of almost any probabilistic density function and thus can be successfully used in modelling of complex systems and are applicable to real live problems. Nevertheless, there are cases, where the mixtures with constant weights do not provide good results. This paper improves the probabilistic mixture model with introducing data dependent component weights. Parameters of the improved model are estimated with a modification of the PB estimation algorithm.

Keywords: probabilistic mixture, Bayesian estimation, probabilistic modelling

1. INTRODUCTION

The choice of a suitable model is one of the most important tasks when dealing with complex systems. A good system model is a necessary condition for consequent control or decision-making tasks on any application domain. Complexity of real-life problems, however, makes often detailed system modelling unfeasible. This orients us to the use of simplified black-box models, which learn their parameters to match the measured data.

From a wide range of possible probabilistic models, we selected the probabilistic mixtures as they provide a universal approximation of almost any probability density function (Titterington *et al.*, 1985) and their form allows relatively simple use in consequent control or decision-making tasks. The mixture model is a convex combination of simpler models called components, the coefficients of the convex combination are called component weights. If the components model dependency of the samples, we speak about dynamic components, otherwise, we speak about static components. Similarly, if the component weights depends on historical data, they are called dynamic, otherwise, they are called static.

We adopted the Bayesian methodology (Peterka, 1981) as a general framework for the model learning. It provides compact theoretical solution of all tasks related to model learning. Unfortunately, these theoretical results are directly applicable only in limited class of models. The probabilistic mixtures are completely out of this class. Hence, approximations of Bayesian learning must be used (Andrýsek, 2004*b*).

In the currently used dynamic mixture model (Kárný *et al.*, 2005), the individual components are dynamic, but the component weights remain static. Although this model doesn't exactly

describe all dynamic probability distributions, it was proven that it is correct at least asymptotically (Kárný *et al.*, 2005). Practical applications, e.g: (Ettler *et al.*, 2005; Heřmanská *et al.*, 2004) verify this approach, but in difficult dynamic cases a need for improvement arises.

This leads to the need for mixtures with both dynamic components and dynamic weights. We can expect that models with dynamic weights will give better results, but their estimation is much more difficult. This paper deals with algorithms for estimation of parameters of dynamic probabilistic mixture model with data-dependent component weights.

2. NOTIONS AND NOTATIONS

d_t - data record at discrete time t, finite dimensional vector

ϕ_{t-1} - state vector formed from relevant historical values, e.g.: $\phi_{t-1} \equiv (d_{t-1}, d_{t-2})$

Θ - unknown parameter, finite dimensional vector

f, π, ρ, h - letters reserved for probability density functions (pdf)

$f(d_t|\phi_{t-1}, \Theta)$ - parameterized model of the system

$\pi(\Theta|\mathcal{G}_t)$ - approximated posterior pdf determined by finite dimensional statistic $\mathcal{G}_t$

3. PROBLEM FORMULATION

In this section, the mixture model with dynamic weights is defined and the main estimation task based on Projection Based (PB) estimation (Andrýsek, 2004*a*) is formulated.

3.1 Dynamic Probabilistic Mixture

The parameterized mixture model with dynamic weights is defined as follows:

$$
\begin{aligned}
f(d_t|\phi_{t-1}, \Theta) &\equiv \sum_{c=1}^{\mathring{c}} \alpha_c(\phi_{t-1}|\Omega) f_c(d_t|\phi_{t-1}, \Theta_c), \mathring{c} < \infty, \quad \text{where} \qquad (1)\\
\mathring{c} &\equiv \text{number of components,}\\
f_c(d_t|\phi_{t-1}, \Theta_c) &\equiv \text{c-th component given by the component parameters } \Theta_c,\\
\alpha_c(\phi_{t-1}|\Omega) &\equiv \text{c-th component weighting function (cwf) given by the parameter } \Omega,\\
&\quad \alpha_c(\phi_{t-1}|\Omega) \geq 0, \quad \sum_{c=1}^{\mathring{c}} \alpha_c(\phi_{t-1}|\Omega) = 1, \quad \forall \phi_{t-1}, \ \forall c\\
\Theta &\equiv \{\Theta_1, \cdots, \Theta_{\mathring{c}}, \ \Omega\} \text{ is unknown parameter.}
\end{aligned}
$$

Verbally: The dynamic probabilistic mixture is a convex combination of several dynamic pdfs called components. The actual weights depends on the state vector ϕ_{t-1}. Mixture parameter Θ is formed by the component parameters $\{\Theta_1, \cdots, \Theta_{\mathring{c}}\}$ and by the parameter Ω determining the behavior of component weighting functions. The parameter Θ represents our only uncertainty about the system model, i.e. we assume to know the functional form of the components f_c and component weighting functions α_c.

3.2 Class of the Posterior Pdfs

According to the general rules of PB estimation (Andrýsek, 2004*a*), we need to choose well manipulable class of posterior pdfs. This motivates us to select this simple class formed by product of pdfs:

$$\begin{aligned} \pi(\Theta|\mathcal{G}_t) &\equiv \rho(\Omega|\mathcal{H}_t)\prod_{c=1}^{\mathring{c}}\pi_c(\Theta_c|\mathcal{S}_{c;t}), \text{ where} \qquad (2)\\ \rho(\Omega|\mathcal{H}_t) &\text{ is pdf on cwf parameter } \Omega \text{ determined by the finite-dimensional statistic } \mathcal{H}_t,\\ \pi_c(\Theta_c|\mathcal{S}_{c;t}) &\text{ are pdfs on factor parameters } \Theta_{c;t} \text{ determined by the statistics } \mathcal{S}_{c;t},\\ \mathcal{G}_t &\equiv (\mathcal{H}_t, \mathcal{S}_{\bullet;t}). \end{aligned}$$

Verbally: The parameters $\Theta_c,\ c \in \{1,\cdots,\mathring{c}\}$, of the individual parameterized components are considered to be conditionally independent, and also, independent of the parameter Ω of component weighting functions. The posterior statistic $\mathcal{G}_t$ is formed by the statistic $\mathcal{H}_t$ determining the pdf of the parameter of cwfs and by the statistics $\{\mathcal{S}_{c;t}\}_{c=1}^{\mathring{c}}$ determining the pdf of parameters of particular components.

3.3 Addressed Problem

Now, it is time to exactly define the problem addressed. We apply the PB approximation (Andrýsek, 2004*a*) to the introduced mixture model (1) and selected class of approximate posterior pdfs (2) and get the following problem:

Find the statistic $\mathcal{G}_t$, which minimizes KL divergence $\mathcal{D}\left(\hat{\pi}_t(\Theta)\,\middle\|\,\pi(\Theta|\mathcal{G}_t)\right)$, where

$$\begin{aligned} \hat{\pi}_t(\Theta) &\equiv \frac{f(d_t|\phi_{t-1},\Theta)\pi(\Theta|\mathcal{G}_{t-1})}{\int f(d_t|\phi_{t-1},\Theta)\pi(\Theta|\mathcal{G}_{t-1})d\Theta},\\ \pi(\Theta|\mathcal{G}_{t-1}) &\equiv \rho(\Omega|\mathcal{H}_{t-1})\prod_{c=1}^{\mathring{c}}\pi_c(\Theta_c|\mathcal{S}_{c;t-1}),\\ f(d_t|\phi_{t-1},\Theta) &\equiv \sum_{c=1}^{\mathring{c}}\alpha(\phi_{t-1}|\Omega)f_c(d_t|\phi_{t-1},\Theta_c). \end{aligned}$$

In other words, we are looking for $\mathcal{G}_t \equiv (\mathcal{H}_t, \mathcal{S}_{\bullet;t})$, knowing $\mathcal{G}_{t-1} \equiv (\mathcal{H}_{t-1}, \mathcal{S}_{\bullet;t-1})$ and d_t, ϕ_{t-1}.

4. PROBLEM SOLUTION

Because the results for statistics $\mathcal{S}_{c;t}$ determining posterior pdfs on component parameters Θ_c are the same as in the case with static component weights, presented in (Andrýsek, 2004*a*), we can focus on results of optimization of statistics $\mathcal{H}_t$ related to cwf parameters.

4.1 General Minimization

For $\mathcal{H}_t$ solving the addressed problem it holds:

$$\mathcal{H}_t \in \operatorname{Arg}\min_{\mathcal{H}_t}\mathcal{D}\left(h(\Omega)\,\middle\|\,\rho(\Omega|\mathcal{H}_t)\right), \text{ where}$$

$$h(\Omega) \equiv \rho(\Omega|\mathcal{H}_{t-1}) \sum_{c=1}^{\mathring{c}} \frac{w_{c;t}}{\hat{\alpha}_{c;t-1}} \alpha_c(\phi_{t-1}|\Omega), \quad w_{c;t} \equiv \frac{\hat{\alpha}_{c;t-1}\beta_{c;t}}{\sum_{\tilde{c}=1}^{\mathring{c}} \hat{\alpha}_{\tilde{c};t-1}\beta_{\tilde{c};t}},$$
$$\hat{\alpha}_{c;t-1} \equiv \int \alpha_c(\phi_{t-1}|\Omega)\rho(\Omega|\mathcal{H}_{t-1})d\Omega, \qquad \beta_{c;t} \equiv \int f_c(d_t|\phi_{c;t}, \Theta_c)\pi_c(\Theta_c|\mathcal{S}_{c;t-1})d\Theta_c.$$

The presented result describes the condition that must be met by optimal statistic $\mathcal{H}_t$, but it does not provide any rule how the minima can be found. The rest of this section is focused on finding such rules by making some assumptions on the parameter Ω and by selecting suitable class of approximate posterior pdfs on Ω. The constants β_c are evaluated in the same way as in the case with static component weights, hence we will not deal with them here.

4.2 Class of Posterior Pdfs

Let us assume that Ω consists of n conditionally independent vectors $\Omega \equiv (\theta_1, \cdots, \theta_n)$. Then, the class of posterior pdfs on Ω can be selected as a product of simpler pdfs. Here, for simplicity, we assume that the product is formed by Gaussian pdfs only. Results for posterior class in a form of product of Gaussian and Gauss-inverse Wishart pdfs can be found in (Andrýsek, 2005).

$$\rho(\Omega|\mathcal{H}_t) = \prod_{k=1}^{n} \mathcal{N}_{\theta_k}\left(M_{k;t}, R_{k;t}\right), \ \mathcal{H}_t \equiv (M_{1;t}, R_{1;t}, \cdots, M_{n;t}, R_{n;t}) \tag{3}$$

4.3 Optimization Result

For the selected class of posterior pdfs on Ω (3), the solution can be found in terms of moments of marginal pdfs $h(\theta_k)$ of $h(\Omega)$:

$$M_{k;t} \equiv \mathcal{E}\left[\theta_k\right] = \int \theta_k h(\theta_k) d\theta_k = \int \theta_k h(\Omega) d\Omega,$$
$$R_{k;t} \equiv \mathbf{cov}\left[\theta_k\right] = \int \theta_k \theta_k' h(\theta_k) d\theta_k - M_{k;t} M_{k;t}'.$$

These results are very important, because they converted the problem of minimization and divergence evaluation into the evaluation of moments "only". Unfortunately, these moments can be rarely evaluated analytically.

4.4 Approximation

Our ability to obtain feasible algorithms depends on the ability to approximate the integrals

$$\hat{\alpha}_{c;t-1} = \int \alpha_c(\phi_{t-1}|\Omega)\rho(\Omega|\mathcal{H}_{t-1})d\Omega \ \text{ and } \int \theta_k h(\Omega) d\Omega \ \text{ and } \int \theta_k \theta_k' h(\Omega) d\Omega.$$

The simplest and universal approximation of the mentioned integrals is Monte-Carlo integration. Hence it was used on the examined cases. In future research, other approximations of the integrals have to be used.

Let us generate N samples from $\rho(\Omega|\mathcal{H}_{t-1})$ and denote them $(\Omega^1, \cdots \Omega^N)$. Then, the mentioned integrals can be approximated as follows:

$$\hat{\alpha}_{c;t-1} \equiv \int \alpha_c(\phi_{t-1}|\Omega)\rho(\Omega|\mathcal{H}_{t-1})d\Omega \approx \frac{1}{N}\sum_{l=1}^{N} \alpha_c(\phi_{c;t-1}|\Omega^l),$$

$$\int \theta_k h(\Omega) d\Omega \approx \frac{1}{N} \sum_{l=1}^{N} \theta_k^l \left(\sum_{c=1}^{\mathring{c}} \frac{w_{c;t}}{\hat{\alpha}_{c;t}} \alpha_c(\phi_{c;t-1}|\Omega^l) \right),$$

$$\int \theta_k \theta_k' h(\Omega) d\Omega \approx \frac{1}{N} \sum_{l=1}^{N} \theta_k^l \theta_k^{l\,\prime} \left(\sum_{c=1}^{\mathring{c}} \frac{w_{c;t}}{\hat{\alpha}_{c;t}} \alpha_c(\phi_{c;t-1}|\Omega^l) \right).$$

To apply this approximation, we only need to be able to take efficiently samples from $\rho(\Omega|\mathcal{H}_{t-1})$ and to evaluate $\alpha_c(\phi_{t-1}|\Omega)$ for given ϕ_{t-1} and Ω. Because the posterior $\rho(\Omega|\mathcal{H}_{t-1})$ is a product of Gaussian pdfs, the sampling is very easy.

5. EXAMPLE

Simple example of the presented algorithm is displayed here. The description of the example is incomplete, because we focused on the parts corresponding to component weights. Data are scalar valued, mixture has 2 components ($\mathring{c} \equiv 2$), state of the model consists of one historical value ($\phi_{t-1} \equiv (d_{t-1})$). Very simple type of cwfs parameterized with a scalar Ω is considered here:

$$\alpha_1(\phi_{t-1}|\Omega) \equiv \alpha_1(d_{t-1}|\Omega) = \begin{cases} 0 & \text{if } d_{t-1} > \Omega \\ 1 & \text{if } d_{t-1} \leq \Omega \end{cases} \quad \text{(1st cwf)}$$

$$\alpha_2(\phi_{t-1}|\Omega) \equiv \alpha_2(d_{t-1}|\Omega) = \begin{cases} 1 & \phi_{t-1} > \Omega \\ 0 & \phi_{t-1} \leq \Omega \end{cases} \quad \text{(2nd cwf)}$$

According to the assumptions from subsection 4.2, the posterior pdf on Ω is selected in the following form: $\rho(\Omega|\mathcal{H}_t) \equiv \rho(\Omega|M_t, R_t) \equiv \mathcal{N}_\Omega(M_t, R_t)$. Initial values of statistic was set to: $M_0 \equiv -2.000,\ R_0 \equiv 40.000$.

We simulated 500 data records with $\Omega_{true} = -0.108$. Figure 1 shows evolution of statistics M_t and R_t during the estimation. Because M_t is in fact a point estimate of the unknown cwf parameter Ω, we can simply see that the point estimate approaches the true value. Because the statistic R_t is in fact variance of point estimate M_t, the decreasing trend of R_t indicates increasing quality of the point estimate.

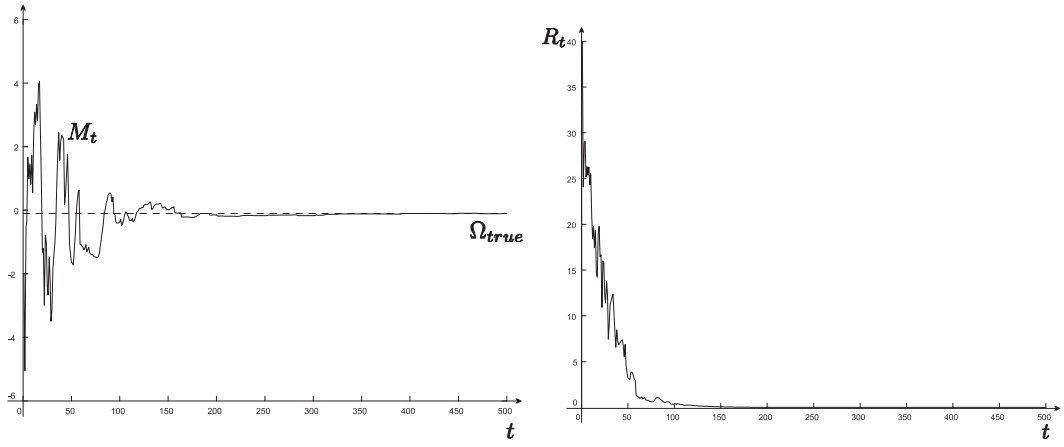

Fig. 1: Evolution of statistics M_t and R_t

ACKNOWLEDGEMENT

This work was supported by GA ČR 102/03/0049, AV ČR S1075351, AV ČR 1ET 100 750 401, MŠMT 8-2006-06 and MŠMT 1M0572.

6. CONCLUSIONS

Dynamic probabilistic mixture model with dynamic weights was defined as a generalization of the current dynamic mixture with static weights. General algorithm for recursive estimation of the generalized model was elaborated. Problem of minimization of KL divergence was converted into a simpler task of evaluation of moments of involved pdfs for special, but important class of posterior pdfs. Monte-Carlo integration was successfully used for evaluating these moments in low-dimensional cases.

Future research will focus on another approximations of the integrals, so that the mixtures with dynamic weights can be estimated also for high-dimensional component weighting functions.

REFERENCES

Andrýsek, J. (2004*a*), Approximate recursive Bayesian estimation of dynamic probabilistic mixtures, *in* J. Andrýsek, M. Kárný and J. Kracík, eds, 'Multiple Participant Decision Making', Advanced Knowledge International, Adelaide, pp. 39–54.

Andrýsek, J. (2004*b*), Projection based algorithms for estimation of complex models, *in* 'Proceedings of the 5th International PhD Workshop on Systems and Control - a Young Generation Viewpoint', Hungarian Academy of Sciences, Budapest, pp. 5–10.

Andrýsek, J. (2005), Estimation of Dynamic Probabilistic Mixtures, Technical Report 2150, ÚTIA AV ČR, Praha.

Ettler, P., Kárný, M. and Guy, T. V. (2005), Bayes for rolling mills: From parameter estimation to decision support, *in* P. Horáček, M. Šimandl and P. Zítek, eds, 'Preprints of the 16th World Congress of the International Federation of Automatic Control', IFAC, Prague, pp. 1–6.

Heřmanská, J., Křížová, H., Gebouský, P., Kárný, M., Wald, M., Adámek, J. and Zimák, J. (2004), Bayesian evaluation of lymphoscintigraphy of secondary lymphedema of upper limbs, *in* F. j. a. f. i. ČVUT v Praze, ed., 'XXVI. Dny radiační ochrany: Sborník rozšířených abstraktů', ČVUT v Praze, Luhačovice, pp. 103–106.

Kárný, M., Böhm, J., Guy, T., Jirsa, L., Nagy, I., Nedoma, P. and Tesař, L. (2005), *Optimized Bayesian Dynamic Advising: Theory and Algorithms*, Springer, London.

Peterka, V. (1981), Bayesian system identification, *in* P. Eykhoff, ed., 'Trends and Progress in System Identification', Pergamon Press, Oxford, pp. 239–304.

Titterington, D., Smith, A. and Makov, U. (1985), *Statistical Analysis of Finite Mixtures*, John Wiley, New York.