Akademie věd České republiky
Ústav teorie informace a automatizace

Academy of Sciences of the Czech Republic
Institute of Information Theory and Automation

RESEARCH REPORT

Miroslav Kárný, Petr Nedoma, Ivan Nagy,
Markéta Valečková

Topics in Mixture Estimation

No. 1994 November 1999

ÚTIA AV ČR, P. O. Box 18, 182 08 Prague, Czech Republic
Fax: (+42)(2)688 4903
E-mail: utia@utia.cas.cz

This report constitutes an unrefereed manuscript which is intended to be submitted for publication. Any opinions and conclusions expressed in this report are those of the author(s) and do not necessarily represent the views of the Institute.

Acknowledgement:

This research was partially supported by EC, grant IST-99-12058
and GAČR 02/99-1564.

Contents

Chapter 1

Introduction

The project ProDaCTool (IST-99-12058) solves has a rather involved task of advising an operator controlling a complex process. On it, many variables, indicating the quality of the production can be measured. The solution idea is based on clustering and approximate Bayesian approach to indentification. In the framework of this project, many sub-problems have to be solved. Two of them, which were submitted to the International Conference on Artificial Neural Networks and Genetic Algorithm (ICANNGA 2001) held at the Lichtenstein Palace, Prague, Czech Republic April 22-25, 2001. Their draft versions are collected in this report.

The first problem, presented in Chapter 2, concerns initiation of a mixture model to be identified from vector-valued data with high dimension.

The second problem, presented in Chapter 3, gives a numerically stable solution of the EM algorithm for identification of large mixture models on basis of data with high dimension.

Chapter 2

Initial Description of Multi-Modal Dynamic Models

Multiple models, neural networks, cluster analysis and probabilistic mixtures are prominent examples of situations when complex multi-modal models [1] are built using vast amount of data. Complexity and non-unicity of modelled situation imply that resulting description depends heavily on the initial phase of search. The safest repetitive purely random search is mostly inhibited by computational complexity of the addressed task. For this reasons, various techniques have been designed. None of them, to our best knowledge, suits to cases when *dynamic* models are constructed. The paper describes a novel technique that fills this gap in a promising way. Essentially, the trial description is gradually split whenever there is possibility that a unimodal sub-model hides more modes. Probabilistic paradigm is used for describing the algorithm but the applicability is much wider.

2.1 Introduction

Modelling of relationships in sequences of multivariate records is addressed in many ways and has a wide range of applications covering majority of human activities. As a rule, the surface in the modelled data space is multi-modal and random. Thus, probabilistic mixtures are its “natural” description. Their estimation is a computationally hard task that has to be approximated using heuristic techniques. Many of them have been developed and successfully used. Predominantly, they assume that processed data records are independent, that the modelled system is static. At the same time, dynamics, i.e. dependence of future on past, is an important feature of majority of practically important cases. The available techniques are able to cope with them at least approximately but the decisive task – the search for initial guess of the model – is almost unsupported. This led us to a development of a novel technique that suits to this case. The paper specialises it to normal mixtures.

It is worth to underline that the proposed technique is also expected to be useful out of probabilistic context and to serve to radial-basis neural networks, multiple local regression, cluster analysis etc.

2.2 Preliminaries

The assumed preliminary knowledge of probabilistic language can be well reduced to the basic algebra of probability density functions (pdf), i.e. their non-negativity, normalisation to unity, chain rule, rule for marginalisation and Bayes rule. A broader engineering insight into the used tools gives [1].

2.2.1 Notation

The following common symbols are adopted:
x^* means the set of all values of x,
$\mathring{x}$ denotes cardinality of x^*,
x_t is x at discrete time $t \in t^* \equiv \{1, \ldots, \mathring{t}\}$,
$x(t) \equiv x_1, \ldots, x_t$,
$x_{i;t}$ is i-th entry of x_t,
$\Theta \in \Theta^*$ denotes unknown model parameters,
$f(\cdot|\cdot)$ is a common symbol for conditional pdfs: versions are distinguished by identifiers in arguments,
$\propto$ means equality up to a normalising factor.

2.2.2 Basic scenario

A sequence $d(\mathring{t})$ of data records d_t is observed and mutual relationships are modelled by the joint pdf

$$f(d(\mathring{t})|\Theta) = \prod_{t \in t^*} f(d_t|d(t-1), \Theta)$$

conditioned on unknown parameters Θ. The considered *mixture model* has the form

$$f(d_t|d(t-1), \Theta) = \sum_{c \in c^*} \alpha_c f(d_t|d(t-1), \Theta_c, c). \tag{2.1}$$

The individual pdfs $f(d_t|d(t-1), \Theta_c, c)$ are called *components*. The unknown parameter Θ is formed by probabilistic weights of components $\alpha \equiv (\alpha_1, \ldots, \alpha_{\mathring{c}}) \in$

$\alpha^* \equiv \{\alpha_c \geq 0, \sum_{c \in c^*} \alpha_c = 1\}$ and by individual parameters $\Theta_c,\ c \in c^*$, of components. The components are decomposed by the chain rule

$$f(d_t|d(t-1), \Theta_c) = \prod_{i=1}^{\mathring{d}} f(d_{i;t}|\psi_{ic;t}, \Theta_{ic}, i, c) \tag{2.2}$$

where $f(d_{i;t}|\psi_{ic;t}, \Theta_{ic}, i, c)$ are called *parameterised factors*. They predict scalar entries $d_{i;t}$ of d_t. They are assumed to depend on regression vectors $\psi_{ic;t}$ that consist of values of known finite-dimensional mappings $d^*(t-1), d^*_{1;t}, \ldots, d^*_{i-1;t} \rightarrow \psi^*_{ic;t}$. The factorisation (2.2) allows us to deal with parsimonious description of components and to combine entries of logical and continuous nature.

The adopted Bayesian estimation modifies a chosen prior pdf $f(\Theta)$ and applies Bayes rule [1] in order get posterior pdf $f(\Theta|d(\mathring{t}))$

$$f(\Theta|d(\mathring{t})) \propto f(d(\mathring{t})|\Theta) f(\Theta). \tag{2.3}$$

This pdf is the most general result of Bayesian estimation and may serve for obtaining point estimates of the unknown Θ, provides information about precision of these estimates and can be directly used for computing predictors that relate arbitrary predicted data item $d_{i;t}$ with an arbitrary regression vector ψ_{ic} within c-th component

$$f(d_i|\psi_{ic}, d(\mathring{t})) = \int_{\theta^*_{ic}} f(d_i|\psi_{ic}, \Theta_{ic}, i, c) f(\Theta_{ic}|d(\mathring{t}))\, d\Theta_{ic}.$$

The posterior pdfs $f(\Theta_{ic}|d(\mathring{t}))$ describing parameterised factors are called *factors*.

2.3 Addressed problem and idea of solution

Application of the Bayes rule requires evaluation of the likelihood function, i.e. the pdf $f(d(\mathring{t})|\Theta)$ taken as a function of parameters. For mixtures, it is computationally hard as it consists of an exponentially blowing number of terms depending on (high-dimensional) Θ. The evaluation can be done approximately, see e.g. [3]. The quality of the approximation depends heavily on the choice of the mixture structure and on the choice of the prior pdf $f(\Theta)$. This choice is addressed here.

The solution is based on a simple iterative search. First, the non-iterative Bayesian evaluation of the posterior pdf is converted into an iterative application of Bayes rule followed by an appropriate flattening. Flattening prevents the iterations to over-fit data. Second, (approximate) Bayes rule is applied for a given number of components. In the component describing majority of data, a most uncertain factor (pdf on $\Theta^*_{..}$!) is selected as being suspicious for hiding more modes. This factor is appropriately split in pair of factors so that a new richer mixture arises. The splitting is performed till predictive abilities of the mixture are increasing.

2.4 Solution details

During evaluations, we need to measure distances between various pdfs involved. Kullback-Leibler (KL) distance [4] is a prominent tool for this. For a pair of pdfs $f, \bar{f}$ posttive on Θ^*, it is defined

$$\mathcal{D}(f||\bar{f}) \equiv \int_{\Theta^*} f \ln\left(\frac{f}{\bar{f}}\right) d\Theta. \tag{2.4}$$

The KL distance is convex in f, non-negative and equals to zero iff $f = \bar{f}$ almost everywhere.

2.4.1 Iterative Bayes learning

Let $f_n(\Theta)$ be n-th guess of the prior pdf and $\mathcal{L}_n(\Theta) \approx f(d(\mathring{t})|\Theta)$ the corresponding approximation of the likelihood function. Then the posterior pdf $f_n(\Theta|d(\mathring{t})) \propto \mathcal{L}_n(\Theta) f_n(\Theta)$ gives a better clue about appropriate parameters and is a candidate for a new guess of the prior pdf. It is, however, over-concentrated. Thus, we select a new guess of the prior pdf $f_{n+1}(\Theta)$ as a compromise between $f_n(\Theta|d(\mathring{t}))$ and a flat prior pdf $\bar{f}(\Theta)$. A construction of such a compromise is described by the proposition below. It also hints how to choose the knob that controls the compromise. The omitted simple proof relies just on properties of the KL distance.

Proposition 2.4.1 (Iterative Bayes learning) *1. Let pdfs $f, \bar{f}$ have a common support Θ^*. Then the pdf $\hat{f}$ minimising $\mathcal{D}(\hat{f}||f) + q\mathcal{D}(\hat{f}||\bar{f}),\ q > 0$ is*

$$\hat{f} \ \propto\ f^{\Lambda}\bar{f}^{1-\Lambda},\ \Lambda = \frac{1}{1+q} \in (0,1). \tag{2.5}$$

2. Let us generate sequence of pdfs $f_n(\Theta)$ by flattening the result of the Bayesian estimation $f \propto \mathcal{L}_n(\Theta) f_n(\Theta)$ to the flat prior pdf $\bar{f}$, i.e.

$$f_{n+1}(\Theta) \propto [\mathcal{L}_n(\Theta) f_n(\Theta)]^{\Lambda} \left[\bar{f}(\Theta)\right]^{1-\Lambda},\ f_0 \equiv \bar{f}. \tag{2.6}$$

If the likelihood function $\mathcal{L}_n(\Theta)$ is evaluated exactly, then $f_{n+1}(\Theta)$ converges for $n \to \infty$ to the exact posterior pdf iff $\Lambda = 0.5$.

2.4.2 Factor shifting

As outlined in Section 2.3, a suitable factor is to be split. The created factors are required to have different modes than the split one. Expected value a suitable function $g(\Theta)$ is used as a mode characterisation. Uncertainty of the factor f is measured by its KL distance $\mathcal{D}(f||\bar{f})$ to a flat factor $\bar{f}$. KL distance $\mathcal{D}(\hat{f}||f)$ measures similarity of the new factors $\hat{f}$ to the split f.

Proposition 2.4.2 (Shifted factor) *Let pdf f and a function g be defined on Θ^*. Then the pdf $\hat{f}$*
- *minimising the KL distance $\mathcal{D}(\hat{f}||f)$,*
- *having a fixed norm $\omega > 0$ of a shift in expectation*

$$\left\langle \int \Theta(\hat{f}(\Theta) - f(\Theta))\, d\Theta, \int g(\Theta)(\hat{f}(\Theta) - f(\Theta))\, d\Theta \right\rangle$$

$$\textit{has the form} \quad \hat{f} \propto f \exp(- < \mu, g(\Theta) >). \tag{2.7}$$

The solution is "indexed" by an array μ that is compatible with a scalar product $< \cdot, \cdot >$. This converts infinite-dimensional optimisation into finite-dimensional search for μ indexing solutions (2.7).

Proof: The Lagrangian of this convex optimisation on the convex set $\hat{f}^* \equiv \{\hat{f} \geq 0,\ \int \int f = 1\ \&(2.7)\}$ is

$$\mathcal{J} \equiv \mathcal{D}(\hat{f}||f) + \nu \int \hat{f}(\Theta)\, d\Theta\ + \rho$$
$$\left\langle \int g(\Theta)(\hat{f}(\Theta) - f(\Theta))\, d\Theta, \int g(\Theta)(\hat{f}(\Theta) - f(\Theta))\, d\Theta \right\rangle.$$

It is parameterised by scalars ν, ρ that have to be chosen so that restrictions are met. Its variation $\delta\mathcal{J}$, that should equal to zero for the optimal $\hat{f}$, reads

$$\delta\mathcal{J} = \int \hat{\delta}[\ln(\hat{f}/f) + \nu + 2\rho < \tilde{\mu}, g(\Theta) >]\, d\Theta, \text{ where}$$

- $\hat{\delta} \equiv \hat{\delta}(\Theta)$ is an arbitrary variation of $\hat{f}(\Theta)$
- $\tilde{\mu} \equiv \int g(\Theta)(\hat{f}(\Theta) - f(\Theta))\, d\Theta$.

This proves the claimed results with $\mu = \rho\tilde{\mu}$. □

2.5 Application to normal factors

2.5.1 Normal factor and conjugate prior

The considered normal factors have the form

$$f(d|\psi, \Theta) = \mathcal{N}_d(\theta'\psi, r) = \tag{2.8}$$
$$(2\pi r)^{-0.5} \exp\left\{ -\frac{1}{2r} \left([-1, \theta']\Psi\right)^2 \right\} \text{ where}$$

$'$ denotes transposition,
$\Theta = [\theta, r] = $ [regression coefficients, noise variance],
$\Psi = [d, \psi']' = $ [regressand, regression vector].

The conjugate prior pdf $f(\Theta)$ that preserves its functional form during Bayes estimation of the model (2.8) is Gauss-inverse-Wishart (GiW) pdf [1]

$$\begin{aligned} f(\Theta) = GiW_{[\theta,r]}(L, D, \kappa) \propto & \\ r^{-\frac{\kappa}{2}} \exp\left\{-\frac{1}{2r}[-1, \theta']L'DL[-1, \theta']'\right\} & \text{ where} \end{aligned} \tag{2.9}$$

$\kappa > 0$ is an effective counter $f(\Theta)$,
$L'DL$ is an extended information matrix in numerically advantageous $L'DL$ decomposition in which
L is lower triangular matrix with a unit diagonal,
D is diagonal matrix with positive entries. The split version of $L'DL$ decomposition

$$L \equiv \begin{bmatrix} 1 & 0 \\ L_{d\psi} & L_{\psi} \end{bmatrix}, D = \text{diag}[D_d, D_\psi] \tag{2.10}$$

can be related to well known least squares (LS) quantities $\hat{\theta} = L_\psi^{-1} L_{d\psi}$ is LS estimate of θ, D_d/κ is LS estimate of r, $D_d/\kappa L_\psi^{-1} D_\psi^{-1} (L'_\psi)'$ is covariance matrix of the LS estimate of θ.

2.5.2 Shifting of GiW factors

We need the formula for the KL distance of a pair of GiW pdfs. We list it without its lengthy proof.

Proposition 2.5.1 (KL distance of GiW pdfs) *Let $\hat{f} \equiv GiW_\Theta(\hat{L}, \hat{D}, \hat{\kappa})$, $f \equiv GiW_\Theta(L, D, \kappa)$ be a pair of GiW pdfs of parameters $\Theta \equiv (\theta, r)$. Their KL distance is given by the formula*

$$\begin{aligned} \mathcal{D}(\hat{f}||f) = & \ln\left(\frac{\Gamma(0.5\kappa)}{\Gamma(0.5\hat{\kappa})}\right) - 0.5\ln\left(\prod_{i=2}^{\mathring{\Psi}} \frac{D_{ii}}{\hat{D}_{ii}}\right) + \\ + & \ 0.5\kappa \ln\left(\frac{\hat{D}_d}{D_d}\right) + 0.5(\hat{\kappa} - \kappa)\frac{\partial}{\partial(0.5\hat{\kappa})}\ln(\Gamma(0.5\hat{\kappa})) \\ - & \ 0.5\mathring{\psi} + 0.5\text{tr}[\hat{L}_\psi^{-1}\hat{D}_\psi^{-1}(\hat{L}'_\psi)^{-1}L'_\psi D_\psi L_\psi] + \\ + & \ 0.5\frac{\hat{\kappa}}{\hat{D}_d}[(\hat{\hat{\theta}} - \hat{\theta})'\hat{L}'_\psi \hat{D}_\psi \hat{L}_\psi(\hat{\hat{\theta}} - \hat{\theta}) - \hat{D}_d + D_d] \end{aligned} \tag{2.11}$$

where Γ denotes Euler gamma function.

Proposition 2.5.2 (Shifted GiW factor)
Let $f = GiW_{\theta,r}(L, D, \kappa)$ *and* $g(\Theta)' = \theta'/r$. *Then the pdf* $\hat{f}$ *minimising the KL distance* $\mathcal{D}(\hat{f}||f)$ *and having a fixed norm* $\omega > 0$ *of the shift*

$$\left\langle \int g(\Theta)(\hat{f}(\Theta) - f(\Theta))\, d\Theta, \int g(\Theta)(\hat{f}(\Theta) - f(\Theta))\, d\Theta \right\rangle$$

has the form $\hat{f} = GiW_{\theta,r}(\hat{L}, \hat{D}, \hat{\kappa})$ *with*

$$\begin{aligned} \hat{L} &\equiv \begin{bmatrix} 1 & 0 \\ L_{d\psi} + \rho L_\psi x & L_\psi \end{bmatrix} \Leftrightarrow \hat{\hat{\theta}} = \hat{\theta} + \rho x \\ \hat{D} &= D,\ \hat{\kappa} = \kappa.. \end{aligned} \tag{2.12}$$

The vector x *in (2.12) is a unit eigenvector of the matrix* $D_d L_\psi^{-1} D_\psi^{-1} (L_\psi')^{-1}$ *corresponding to the maximum eigenvalue. The optional scalar* $\rho \neq 0$ *can be chosen both positive and negative and its choice replaces the choice of* ω.

Proof: Proposition 2.4.2 implies functional form (2.12) of $\hat{f}$. The choice of the function $g(\Theta)$ preserves in GiW class. The form (2.12) allows no changes in κ, L_ψ and D. Thus, $\hat{L}_{d\psi}$ is the only optional quantity. The influenced part of the KL distance $\mathcal{D}(\hat{f}||f)$ is

$$(\hat{\hat{\theta}} - \hat{\theta})' L_\psi' D_\psi L_\psi (\hat{\hat{\theta}} - \hat{\theta}).$$

The restriction (2.12) with $\hat{D} = D$ enforces $(\hat{\hat{\theta}} - \hat{\theta})'(\hat{\hat{\theta}} - \hat{\theta}) > \hat{r}\omega > 0$. Obviously, $\hat{\hat{\theta}} - \hat{\theta} = \gamma x$ with x being eigenvector corresponding to the smallest eigenvalue of $D_d^{-1} L_\psi' D_\psi L_\psi$ (to the largest one of its inversion) is the minimiser searched for. The normalising factor ρ has to be chosen so that the considered restriction is fulfilled. □

2.5.3 Overall algorithm

"Inserting" the particular results into the outlined algorithm we get *initiation of normal mixtures.* Only GiW pdfs are involved used. Their statistics L, D, κ are distinguished by indices $(icn; t)$ pointing to (factor,component,iteration;time). This allows us to use simplified notation

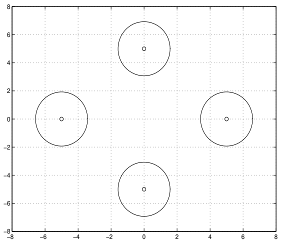
$$GiW(icn; t) \equiv GiW_{\theta_{ic}, r_{ic}}(L_{icn;t}, D_{icn;t}, \kappa_{icn;n})$$

$*$ Set iteration counter $n = 0$ and $\mathring{c} = 1$.
$*$ Select a guess of prior pdf $\hat{f}_0(\Theta) \equiv \prod_{i=1}^{\mathring{d}} GiW(i1n; 0)$ corresponding to the trivial mixture with single component.
• Evaluate approximately $f_n(\Theta|d(\mathring{t})) = \prod_{c \in c^*} \prod_{i=1}^{\mathring{d}} GiW(icn; t)$, (quasi-Bayes

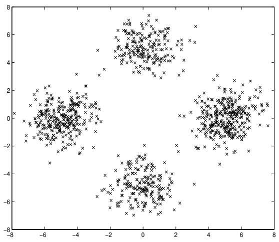
estimation [3] reduces to weighted recursive LS).
$*$ Evaluate the value of the corresponding predictive pdf $f(d(\mathring{t}))$ (a by-product of the step $\bullet$).
$*$ Select the component $\bar{c}$ with the largest sum of $\kappa_{icn;\mathring{t}}$ (explaining the majority of data).
$*$ Set flags to all factors within the component $\bar{c}$ to the value "admissible".
$\bullet\bullet$ Find among all estimated factors with the flag "admissible" the most uncertain factor $GiW(\bar{i}\bar{c}n;\mathring{t})$, i.e. $\mathcal{D}(GiW(\bar{i}\bar{c}n;\mathring{t})||\bar{f}) \leq \mathcal{D}(GiW(i\bar{c}n;\mathring{t})||\bar{f})$ where $\bar{f}$ is a flat (uniform) pdf.
$*$ Generate new from $GiW(\bar{i}\bar{c}n;\mathring{t})$ shifted factors, say $GiW(\bar{i}^{+}\bar{c}n;\mathring{t}),\ GiW(\bar{i}^{-}\bar{c}n;\mathring{t})$, Proposition 2.5.2.
$*$ Create a new trial estimate of the mixture $f_{n+1|n}(\Theta|d(\mathring{t}))$ by replacing the splitted factor $GiW(\bar{i}\bar{c}n;\mathring{t})$ by the pair $GiW(\bar{i}^{+}\bar{c}n;\mathring{t}),\ GiW(\bar{i}^{-}\bar{c}n;\mathring{t})$ of new ones in the selected component with index $\bar{c}$. It splits the component $\bar{c}$ in two and increases $\mathring{c} = \mathring{c} + 1$.
$*$ Apply the flattening operation to $f_{n+1|n}$, see Proposition 2.4.1, to get a trial prior pdf $f_{n+1|n}(\Theta) = \prod_{c \in c^*} \prod_{i=1}^{\mathring{d}} GiW(ic(n+1|n);0)$.
$*$ Make the trial estimation with this prior and evaluate the corresponding value of predictive pdf.
$*$ Set the flag of the split factor to the value not admissible, take all intermediate evaluations as trial ones and go to the step $\bullet\bullet$ if no prediction improvement is recorded and there are admissible factors. Stop if no admissible factor remains.
$*$ Set $f_{n+1} = f_{n+1|n}(\Theta)$ if an increase of of predictive pdf is recorded, set $n = n+1$, go to $\bullet$.

2.6 Illustrative example

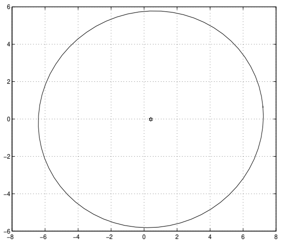
The proposed algorithm has been thoroughly tested and its behaviour was found satisfactory. The following example is intended for illustrating its behaviour. In it, $\mathring{t} = 1000$ records of two-dimensional data from a static normal mixture were simulated. Mean values of components together with isoprobability ellipsoids encircling probability 0.75 are shown on left-hand side of the 1st figure. Data samples are on its right hand side. A bit denser leftist and rightist clusters correspond to the choice $\alpha = [0.3, 0.2, 0.2, 0.3]$. The subsequent figures show centres and the same ellipsoids found during three iterations of the proposed algorithm.



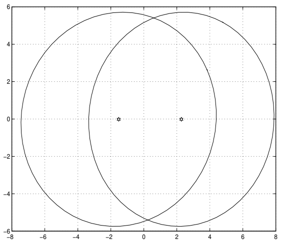
Simulation setup



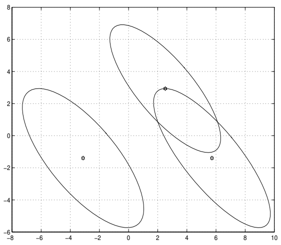
Data sample



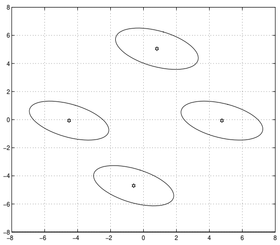
Initialization setup



1th iteration



2nd iteration



3th iteration

2.7 Conclusions

The proposed novel algorithm of initiating mixture estimation was found satisfactory. Its applicability to dynamic mixtures is its exceptional feature. Naturally, it is not able to cope with all cases and should be combined with other techniques containing random search steps. Its applicability to non-normal components is worth noticing. The adopted factorised description of components opens a way for describing mixed discrete-continuous mixtures. Also, the algorithm generates potentially mixtures with factors that are common to several components. This feature is important both for physical modelling and parsimonity of the final description.

Bibliography

[1] R. Murray-Smith and T. Johansen, *Multiple Model Approaches to Modelling and Control*. London: Taylor & Francis, 1997.

[2] V. Peterka, "Bayesian system identification," in *Trends and Progress in System Identification* (P. Eykhoff), Oxford: Pergamon Press, 1981.

[3] M. Kárný, et al, "Quasi-Bayes estimation applied to normal mixture," in *3rd IEEE Workshop CMP* (J. Rojíček et al), ÚTIA, 1998.

[4] S. Kullback and R. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, 1951.

Chapter 3

Factorized EM Algorithm for Mixture Estimation

A classical version of the EM algorithm is considered in the paper. Its numerical properties are improved using factorized algorithms for maximization in M step of the algorithm. The results are illustrated on simulated examples.

3.1 Introduction

Expectation and Maximization (EM) algorithm is well known and widely spread procedure for parameter estimation in cases when some of the modeled data can and some cannot be measured [2]. The parameter estimates are computed iteratively with monotonic convergence guaranteed [3].

Estimation of mixture models [4] is a well known theoretical application of the EM algorithm. Here, the measured data are represented by the output of the system. The internal pointers, indicating the active components of the mixture at respective time moments t, are not known. They correspond to the unmeasured part of the data.

Mixture models have a wide practical use as they have universal approximation property: they can be interpreted as approximators of pdfs or approximators based on radial basis functions. In practical applications, they may serve for describing rather large systems with many output variables and rather high number of components. In such cases, however, numerical instabilities occur while manipulating estimates of covariances that are extensively used in each iterative estimation step. These estimates of covariance matrices can easily loose their theoretical positive definiteness. Then, the subsequent estimation produces unbounded values.

The aim of this paper is to utilize numerical procedures adopted from factorized on-line estimation [5, 1] and to combine them with EM algorithm to improve its numerical properties.

3.2 EM algorithm

Let us consider a system generating at time t data $d_t = [y_t, c_t]$. The part y_t of the data is measured and thus known, the rest c_t cannot be measured and thus it is unknown to us. We suppose to have a model of the system, but this model depends also on the unknown part of the data c_t. Consequently, it cannot be directly used in estimation for constructing likelihood function. Instead, the expectation of the logarithmic likelihood taken with respect to unknown data c_t is used for parameter estimation.

3.2.1 General task

The model of the system is supposed in the form of conditional probability density function (pdf)

$$f(d_t|d_1,\ldots,d_{t-1},\Theta) = f(d_t|\Theta) = f(y_t,c_t|\Theta), \quad \text{where}$$

d_t is the complete data pair $d_t = [y_t, c_t]'$,
y_t are measured data,
c_t are unmeasured data and
Θ is a vector of all unknown parameters of the pdf.

The data y_t are measured at discrete time instances $t = 1, 2, ..., \mathring{t}$. By Y, we denote a vector of all collected measured data $Y = [y_1, y_2, ..., y_{\mathring{t}}]$. Similarly we denote a vector of unmeasured part of data c_t by C and complete data by D. The likelihood is

$$L(\Theta, Y, C) = \prod_{t=1}^{\mathring{t}} f(y_t, c_t|\Theta).$$

As the likelihood contains unknown variables C, it is a random variable. The main trick of the EM algorithm is that it minimizes expectation of its logarithm. The expected value of the log-likelihood, with respect to C, given data Y and current estimate of the parameters Θ, denoted by $\hat{\Theta}$, is

$$\begin{aligned}\mathcal{L}(\Theta, Y) &= E[\log L(\Theta, Y, C)|Y, \hat{\Theta}] = \\ &= \sum_{C \in C^*} \log L(\Theta, Y, C) f(C|Y, \hat{\Theta}),\end{aligned}$$

where the summation is performed over all possible sequences of unmeasured data $C = [c_1, c_2, ..., c_{\mathring{t}}] \in C^*$; C^* denotes the sample space of the random vector

C. The estimation of parameters Θ is based on maximization of this expectation of log-likelihood. The maximum is searched iteratively. In each iteration step, the expectation is computed on the basis of old parameter estimates (E step). Then the resulting function of Θ is maximized giving new parameter estimates (M step).

3.2.2 Normal mixture model

Let us specify the system under consideration in the following way. It consists of several subsystems, on which the set of data y_t, $t = 1, 2, ..., \mathring{t}$ is measured. Each subsystem is modeled by Gaussian pdf

$$f_c(y_t|\theta_c) \sim \mathcal{N}_{y_t}(\mu_c, R_c), \quad c \in c^*, \text{ where}$$

$c \in c^* = \{1, 2, ..., \mathring{c}\}$ is a label of a particular sub-model (mixture component), $\theta_c = [\mu_c, R_c]'$ is a sub-model parameter with μ_c and R_c being the corresponding mean and covariance of y_t, respectively.

We suppose that at each time instant, only one of the models is valid and we take the sequence $c_t \in c^*$ for $t = 1, 2, ..., \mathring{t}$ as pointers to the active subsystem at each time instant t. As we do not know which model is currently active, the individual terms of the sequence are random. They correspond to the unmeasured system output from the general EM algorithm formulation. We suppose them to be independent of $y_\tau, \tau \neq t$ at time t and of the component parameters $\theta_c, c \in c^*$. Their probabilisties are

$$f(c_t|\Theta) = \alpha_{c_t}, \quad c_t \in c^*, \quad t = 1, 2, ..., \mathring{t}, \text{ where}$$

$\Theta = [\theta', \alpha']'$,
$\theta = [\theta'_1, \ldots, \theta'_{\mathring{c}}]'$ are parameters of the components,
$\alpha = [\alpha_1, \ldots, \alpha_{\mathring{c}}]$ are component probabilities.

With this assumption, the system model is equivalent to a mixture with $\mathring{c}$ static Gaussian components, each component having the probability α_c, $c \in c^*$, to be active at time $t = 1, 2, ..., \mathring{t}$. The complete description of such mixture model at time t is given by the joint pdf

$$\begin{aligned} & f(y_t, c_t|d_1, \ldots, d_{t-1}, c_1, \ldots, c_{t-1}, \Theta) = \\ = \; & f(y_t|c_t, \Theta) f(c_t|\Theta) = f_{c_t}(y_t|\theta)\, \alpha_{c_t}. \end{aligned}$$

The pdf $f_{c_t}(y_t|\theta)$ models the c_t-th component of the mixture. We have used its assumed independence on α and independence of pdfs $f(c_t|\Theta)$ on θ.

3.3 EM algorithm for normal mixtures

Our task is to identify the parameters of the mixture μ_c, R_c, α_c for $c \in c^*$. For the solution of the task, we will follow the standard EM algorithm with the only exception that for the numerically sensitive parts of the data covariance estimation we will use factorized algorithms, worked up in connection with the Bayesian approach to identification [1, 6].

3.3.1 Distribution of unmeasured data

To be able to take the expectation of the log-likelihood, we need the probability $f(C|Y,\Theta)$ of all unknown pointers conditioned not only on parameters Θ but also on measured data $y_t,\ t = 1, 2, ...\mathring{t}$. This joint probability is the product

$$f(C|Y,\Theta) = \prod_{t=1}^{\mathring{t}} f(c_t|y_t,\Theta)$$

in which we have used the independency of the pointer c_t of other pointers c_τ and data y_τ for $t \neq \tau$. The particular probability $f(c_t|y_t,\Theta)$ can be viewed as a prediction of the activity of a component at time t based on the measured data. This probability is determined by Bayes rule

$$f(c_t|y_t,\Theta) \propto f(y_t|c_t,\Theta) f(c_t|\Theta) = f_{c_t}(y_t|\theta)\alpha_{c_t}.$$

It uses the models of components and their probabilities. In accordance with EM algorithm, the expectation defined by this pdf is evaluated with the current parameter estimates substituted for the unknown parameter values. Using the measured data Y we can compute values of the needed pdf $f(c_t|y_t, \Theta = \hat{\Theta})$ for all t and at each time t for all possible values $c \in c^*$ of random variable c_t. To stress that the results are just numbers – probabilistic weights for expectation and consequently for weighted data accumulation, we denote them

$$w_{c,t} = f(c_t = c|y_t, \hat{\Theta}),$$

which is the probability, that the component labeled by c is active at time t.

3.3.2 E step

The expectation of log-likelihood $\mathcal{L}(\Theta, Y) =$

$$= \sum_{C \in C^*} \left[\sum_{i=1}^{\mathring{t}} \log[\alpha_{c_i} f_{c_i}(y_i|\theta)] \prod_{j=1}^{\mathring{t}} f(c_j|y_j,\hat{\Theta}) \right] =$$

$$
\begin{aligned}
&= \sum_{c=1}^{\mathring{c}} \sum_{t=1}^{\mathring{t}} \log[\alpha_c f_c(y_t|\theta)] w_{c,t} = \\
&= \underbrace{\sum_{c=1}^{\mathring{c}} \sum_{t=1}^{\mathring{t}} \log[\alpha_c] w_{c,t}}_{\mathcal{L}_1} + \underbrace{\sum_{c=1}^{\mathring{c}} \sum_{t=1}^{\mathring{t}} \log[f_c(y_t|\theta)] w_{c,t}}_{\mathcal{L}_2} .
\end{aligned}
$$

The first equality is implied by the fact, that probabilities sum to unity.

3.3.3 M step

This step consists in maximization of the expectation of log-likelihood $\mathcal{L}(\Theta, Y)$ constructed in the previous step. The maximization can be performed independently in two steps. The first term $\mathcal{L}_1$ gives the estimate of α. For explicit maximization of the second term, we suppose the component models in the form of static and normally distributed pdfs

$$
f_c(y_t|\theta) \propto |R_c|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y_t - \mu_c)' R^{-1} (y_t - \mu_c)\right\}
$$

for $c = 1, 2, \ldots \mathring{c}$, where
μ_c is the expectation of the c^{th} component,
R_c is the covariance matrix of the c^{th} component.

Then the second term $\mathcal{L}_2$ gives estimates of μ_c and R_c for $c = 1, 2, \ldots, \mathring{c}$. The classical results can be derived by the method of Lagrange multipliers. In their simplified description, the symbols without the sign ˆ mean estimation results obtained at previous iteration step. For $c = 1, 2, \ldots, \mathring{c}$, the formulae are

$$
\begin{aligned}
w_{c,t} &= \\
&= \frac{|R_c|^{-0.5} \exp\{-0.5(y_t - \mu_c)' R_c^{-1} (y_t - \mu_c)\}}{\sum_{\tilde{c}} |R_{\tilde{c}}|^{-0.5} \exp\left\{-0.5(y_t - \mu_{\tilde{c}})' R_{\tilde{c}}^{-1} (y_t - \mu_{\tilde{c}})\right\}}
\end{aligned}
$$

$$
\hat{\alpha}_c = \frac{1}{\mathring{t}} \sum_{t=1}^{\mathring{t}} w_{c,t}, \qquad \hat{\mu}_c = \frac{1}{\mathring{t}\hat{\alpha}_c} \sum_{t=1}^{\mathring{t}} y_t w_{c,t},
$$

$$
\hat{R}_c = \frac{1}{\mathring{t}\hat{\alpha}_c} \sum_{t=1}^{\mathring{t}} (y_t - \mu_c)(y_t - \mu_c)' w_{c,t}.
$$

These formulas are simple and clearly interpretable. The explicit form of the weights $w_{c,t}$ shows the source of numerical troubles. If an iteration fails to give numerically a positive definite $\hat{R}_c$ then the next iteration breaks down. It reflects that the maximized function $\mathcal{L}_2$ is not concave any more and its maximizer has infinite-valued entries.

3.4 Factorized estimation

The evaluation of $\hat{R}_c$ whose inversion and determinant are used in weighting is bottle neck of the classical formulas. To avoid difficulties mentioned, we will update not the whole matrix but its LD factorization. The formula for computing weights shows, that the inverse of the matrix $\hat{R}_c$ is needed. That is why we will compute directly the LD factorization of the inverse of $\hat{R}_c$.

The unnormalized version of the estimate

$$\tilde{R}_c = \sum_{t=1}^{\mathring{t}} (y_t - \mu_c)(y_t - \mu_c)' w_{c,t}$$

can be simply updated recursively

$$\tilde{R}_{c,t} = \tilde{R}_{c,t-1} + (y_t - \mu_c)(y_t - \mu_c)' w_{c,t} = \tilde{R}_{c,t-1} + z_{c,t} z'_{c,t}$$

where
$z_{c,t} = (y_t - \mu_c)\sqrt{w_{c,t}}$,
$\tilde{R}_{c,0} = 0$ and $\hat{R}_c = \frac{\tilde{R}_{c,\mathring{t}}}{\mathring{t}\hat{\alpha}_c}$.

As we have stated, we need to update LD factorization of the inverse of $\tilde{R}_c$. We drop the fixed subscript $_c$ and introduce factorizations of inversions of $\tilde{R}_{t-1}$ and $\tilde{R}_t$

$$\tilde{R}_{t-1}^{-1} = L\,D\,L', \;\; \tilde{R}_t^{-1} = \bar{L}\bar{D}\bar{L}', \quad \text{where}$$

$L, \bar{L}$ are lower triangular matrices with unit diagonals and
$D, \bar{D}$ are diagonal matrices with positive diagonals.

All involved square matrices have dimension n.

With these notations, the recursion for $\tilde{R}$ becomes

$$\begin{aligned} \bar{L}\bar{D}\bar{L}' &= \left[(L\,D\,L')^{-1} + z_t z_t'\right]^{-1} = \\ &= L - \left[D\frac{Dff'D}{1 + f'Df}\right] L', \text{ with } f = L'z_t. \end{aligned}$$

The definition of LD decomposition implies that the expression in brackets [ ] can be decomposed analytically into the LD form, say $H\bar{D}H'$, see [5]. The following algorithm provides it

$$\begin{aligned} &\text{for } i = n, n-1, ..., 1 \\ &\sigma^2 = 1 + \sum_{k=i}^{n} D_k f_k^2 \\ &\bar{D}_{i,i} = D_{i,i}\frac{\sigma_{i+1}^2}{\sigma_i^2} \end{aligned}$$

$$
\begin{aligned}
H_{i,i} &= 1 \\
&\text{for } j = i+1, i+2, ..., n \\
&\qquad H_{j,i} = -\frac{f_i f_j}{\sigma_{i+1}^2} D_{j,j} \\
&\text{end for } j \\
&\text{end for } i
\end{aligned}
$$

We attach to this algorithm the multiplications $f = L' z_t$ and $\bar{L} = LH$ and organize carefully manipulations with individual entries. It gives the overall very compact algorithm that replaces L, D by $\bar{L}, \bar{D}$ after including new data vector z_t:

$s = 1$
for $j = N$ to 1
  $h = z_j$
  for $i = j + 1$ to N
    $h = h + L_{ij}$
  end j
  $a = h/s$
  $v_j = D_j * h$
  $r = s + v_j * h$
  $D_j = s/r * D_j$
  $s = r$
  for $i = j + 1$ to N
    $p = L_{ij} - v_i * a$
    $v_i = v_i + L_{ij} * v_j$
    $L_{ij} = p$
  end i
end j

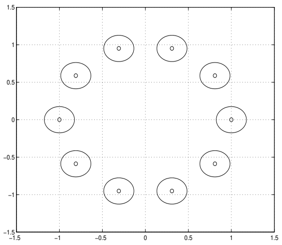
3.5 Illustrative example

The EM algorithm and its factorized version have been thoroughly tested. The following example illustrates advantages of the proposed factorized modification.

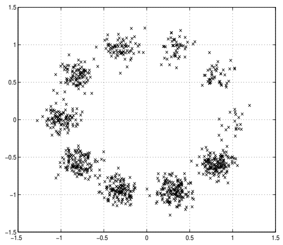
A simulated data sample of the length of 1000 data items was generated from a static normal mixture with 10 components. The mean values of the components together with iso-probability ellipsoids encircling probability 0.75 are shown on the left-hand side of the 1st figure. The data sample is plotted on its right hand side. The weights of the components grow linearly from the first one to the last as it can be seen from the plot.

The estimation setup is on left-hand side of the 2nd figure. The centers of the initial ellipsoid differ from the simulation ones by a random value. Uniform initial estimates of component weights were chosen.

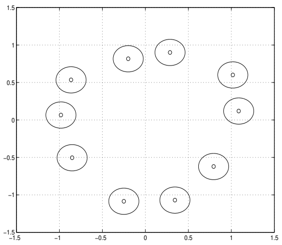
The standard EM algorithm broke down after the 1st iteration step when at least one estimate of R_c lost its definiteness. When running the square root version of EM algorithm, no problems were encountered. The estimated mixture after 20 iterations is shown in the right-hand side of the figure 2.



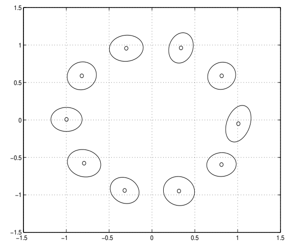
Simulaton setup



Data sample



Estimation setup



Estimation result

3.6 Conclusions

A new numerical solution to the estimation of normal mixture by EM algorithm has been proposed. The algorithm accepts original estimation of component probabilities α and means θ. Instead of the estimate of the noise covariances R, it updates LD decomposition of its inversion. It simplifies evaluation of weights and, first of all, it prevents these estimates from loosing their positive definiteness. The numerical stability of the proposed algorithm in comparison with the "rough" solution has been demonstrated on simulated examples.

The approach can be extended to the case of dynamic components when the expected value of the predicted data is a linear combination of estimated regression coefficients and known regression vector. The overall description is made in probabilistic terms. It opens a way of including a non-trivial prior probability on estimated parameters giving maximum a posteriori probability estimates. The interpretation can be also suppressed and the algorithm used for estimating normal radial basis neural network. These options will be elaborated in detail in foreseen future.

Bibliography

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

[2] A.P.Dempster, N.M.Lair, and D.B.Rubin, "Maximum-likelihood from incomplete data via the em algorithm," *J.Royal Statist. Soc. Ser. B.*, vol. 39, 1977.

[3] C.F.J.Wu, "On the convergence properties of the em algorithm," *The Annals of Statistics*, vol. 11, pp. 95–103, 1983.

[4] R.Redner and H.Walker, "Mixture densities, maximum likelihood and the em algorithm," *SIAM Review*, vol. 26, 1984.

[5] G. Bierman, *Factorization Methods for Discrete Sequential Estimation.* New York: Academic Press, 1977.

[6] M. Kárný, A. Halousková, J. Böhm, R. Kulhavý, and P. Nedoma, "Design of linear quadratic adaptive control: Theory and algorithms for practice," *Kybernetika*, vol. 21, 1985. Supplement to No. 3, 4 ,5, 6.