Initial Description of Multi-Modal Dynamic Models

Miroslav Kárný, Petr Nedoma, Ivan Nagy, Markéta Valečková * [1]

*ÚTIA, AV ČR, Pod vodárenskou věž 4, Prague 8, Czech Republic e-mail: school@utia.cas.cz.

Abstract

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.

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. The proposed technique is also expected to be useful out of probabilistic context and to serve to radial-basis neural networks, multiple regression, cluster analysis etc.

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 [2].

Notation: The following symbols are useded:
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.

Basic scenario: A sequence $d(\mathring{t})$ of data records d_t is observed. Their relationships are modelled by joint pdf conditioned on unknown parameters Θ

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

The considered *mixture model* has the form

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

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 Θ_c, $c \in c^*$, of components. The components are decomposed by the chain rule

$$f(d_t|d(t-1), \Theta_c) = \prod_{i=1}^{\mathring{d}} f(d_{i;t}|\psi_{ic;t}, \Theta_{ic}, i, c) \quad (2)$$

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 $d^*(t-1), d^*_{1;t}, \ldots, d^*_{i-1;t} \to \psi^*_{ic;t}$. The factorisation (2) allows us to deal with parsimonious description of components and to combine entries of logical and continuous nature.

[1] This research was partially supported by EC, grant IST-99-12058, e-mail: school@utia.cas.cz

The adopted Bayesian estimation modifies a chosen prior pdf $f(\Theta)$ and applies Bayes rule [2] in order get posterior pdf $f(\Theta|d(\mathring{t}))$

$$f(\Theta|d(\mathring{t})) \propto f(d(\mathring{t})|\Theta)f(\Theta). \quad (3)$$

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 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.

3 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

$$\mathcal{D}(f||\bar{f}) \equiv \int_{\Theta^*} f \ln\left(\frac{f}{\bar{f}}\right) d\Theta. \quad (4)$$

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

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 it. The omitted proof relies on properties of $\mathcal{D}(\cdot||\cdot)$.

Proposition 3.1 (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

$$\hat{f} \quad \propto \quad f^{\Lambda}\bar{f}^{1-\Lambda},\ \Lambda = \frac{1}{1+q} \in (0,1). \quad (5)$$

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.

$$f_{n+1}(\Theta) \propto \left[\mathcal{L}_n(\Theta)f_n(\Theta)\right]^{\Lambda}\left[\bar{f}(\Theta)\right]^{1-\Lambda},\ f_0 \equiv \bar{f}. \quad (6)$$

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 $\Lambda = 0.5$.

Factor shifting: As outlined in Section 2, 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 3.2 (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*

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

has the form $\quad \hat{f} \propto f \exp(- < \mu, g(\Theta) >). \quad (7)$

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 (7).

Proof: The Lagrangian of this convex optimisation on the convex set $\hat{f}^* \equiv \{\hat{f} \geq 0,\ \int \int f = 1\ \&(7)\}$ 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 $\int \hat{\delta}[\ln(\hat{f}/f) + \nu + 2\rho < \tilde{\mu}, g(\Theta) >]\, d\Theta$, 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 $\mu = \rho\tilde{\mu}$. □

4 Application to normal factors

Normal factor and conjugate prior: The considered normal factors have the form

$$f(d|\psi, \Theta) = \mathcal{N}_d(\theta'\psi, r) = (2\pi r)^{-0.5} \exp\left\{ -\frac{1}{2r}\left([-1, \theta']\Psi\right)^2 \right\} \text{ where} \quad (8)$$

$'$ 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 (8) is Gauss-inverse-Wishart (GiW) pdf [2]

$$f(\Theta) = GiW_{[\theta,r]}(L, D, \kappa) \propto r^{-\frac{\kappa}{2}} \exp\left\{ -\frac{1}{2r}[-1, \theta']L'DL[-1, \theta']' \right\} \text{ where} \quad (9)$$

$\kappa > 0$ 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

$$L \equiv \begin{bmatrix} 1 & 0 \\ L_{d\psi} & L_{\psi} \end{bmatrix}, D = \mathrm{diag}[D_d, D_{\psi}] \quad (10)$$

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.

Shifting of GiW factors: We use the KL distance of a pair of GiW pdfs. Proof is omitted.

Proposition 4.1 (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*

$$\begin{aligned}
\mathcal{D}(\hat{f}||f) &= \ln\left(\frac{\Gamma(0.5\kappa)}{\Gamma(0.5\hat{\kappa})}\right) - 0.5 \ln\left(\prod_{i=2}^{\mathring{\Psi}} \frac{D_{ii}}{\hat{D}_{ii}}\right) + \quad (11)\\
&+ 0.5\kappa \ln\left(\frac{\hat{D}_d}{D_d}\right) + 0.5(\hat{\kappa} - \kappa)\frac{\partial}{\partial(0.5\hat{\kappa})}\ln(\Gamma(0.5\hat{\kappa}))\\
&- 0.5\mathring{\psi} + 0.5\mathrm{tr}[\hat{L}_{\psi}^{-1}\hat{D}_{\psi}^{-1}(\hat{L}'_{\psi})^{-1} L'_{\psi} D_{\psi} L_{\psi}] +\\
&+ 0.5\frac{\hat{\kappa}}{\hat{D}_d}[(\hat{\hat{\theta}} - \hat{\theta})'\hat{L}'_{\psi}\hat{D}_{\psi}\hat{L}_{\psi}(\hat{\hat{\theta}} - \hat{\theta}) - \hat{D}_d + D_d]
\end{aligned}$$

where Γ denotes Euler gamma function.

Proposition 4.2 (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

$$\begin{aligned}
\hat{L} &\equiv \begin{bmatrix} 1 & 0 \\ L_{d\psi} + \rho L_{\psi} x & L_{\psi} \end{bmatrix} \Leftrightarrow \hat{\hat{\theta}} = \hat{\theta} + \rho x\\
\hat{D} &= D,\ \hat{\kappa} = \kappa.. \quad (12)
\end{aligned}$$

The vector x in (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 choice of the optional scalar $\rho \neq 0$ replaces the choice of ω.

Proof: Proposition 3.2 implies functional form (12) of $\hat{f}$. The choice of the function $g(\Theta)$ preserves in GiW class. The form (12) allows no changes in κ, L_{ψ} and D. Thus, $\hat{L}_{d\psi}$ is the only optional quantity influencing just $(\hat{\hat{\theta}} - \hat{\theta})' L'_{\psi} D_{\psi} L_{\psi} (\hat{\hat{\theta}} - \hat{\theta})$.

The restriction (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 meet the considered restriction. □

Overall algorithm: The particular results in the outlined algorithm give *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

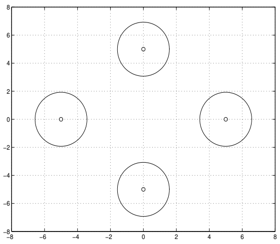
estimation [3] reduces to weighted recursive LS).
$*$ Evaluate the value of the corresponding predictive pdf $f(d(\mathring{t}))$ (a by-product of the step $\bullet$).
$*$ Select the component $\bar{c}$ with the largest sum of $\kappa_{icn;\mathring{t}}$ (explaining the majority of data).
$*$ Set flags 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 4.2.
$*$ 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 3.1, 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 $\bullet$.

5 Illustrative example

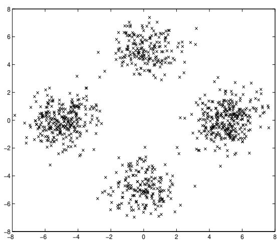
The algorithm has been thoroughly tested and found satisfactory. The example just illustrates its behaviour. In it, $\mathring{t} = 1000$ records of two-dimensional data from a static normal mixture were simulated. Mean values of components together with iso-probability ellipsoids encircling probability 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.

6 Conclusions

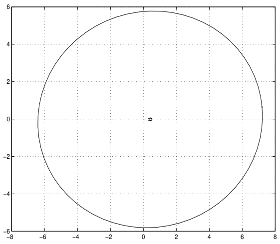
The proposed novel algorithm of initiating mixture estimation was found satisfactory. Its applicability to dynamic mixtures is its exceptional feature. Naturally, it is not able to cope with all cases and should be combined with other techniques containing random search steps. Its applicability to non-normal 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.



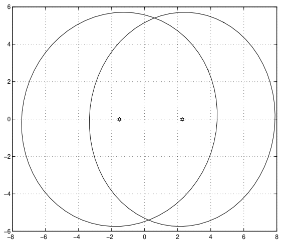
Simulation setup



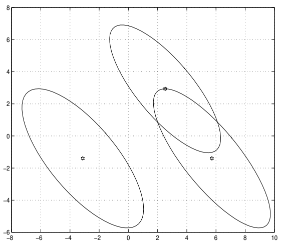
Data sample



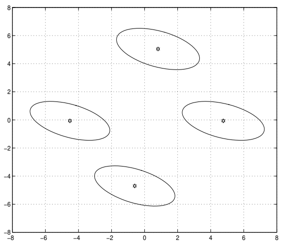
Initialization setup



1th iteration



2nd iteration



3th iteration

References

[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.