Akademie věd České republiky
Ústav teorie informace a automatizace, v.v.i.

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

RESEARCH REPORT

Kamil Dedecius

Notes on projection based modelling of beta-distributed weights of a two-component mixture

No. 2297 March 25, 2011

ÚTIA AV ČR, P.O.Box 18, 182 08 Prague, Czech Republic
Tel: +420 286892337, Fax: +420 266052068, Url: http://www.utia.cas.cz,
E-mail: dedecius@utia.cas.cz

Abstract

This report contains brief notes on estimation of beta-distributed weight of a Gaussian mixture. The results are directly applied in paper [Kárný, M.: On approximate Bayesian recursive estimation]. First, we develop a method to update the beta distribution of weights by new data (evidences) and show, that a projection is needed to preserve the low modelling complexity. Then, we show how forgetting may be applied to improve adaptivity. The results can be immediately applied to multicomponent mixtures.

Table 1: Notation

Symbol	Meaning
$B(x,y)$	beta function with arguments x, y
$c \geq 0$	value of the Kullback-Leibler divergence
$D(f\|\|g) \geq 0$	Kullback-Leibler divergence of f from g
$f_0(\Theta) \equiv \beta_\alpha(s_1^*, s_2^*)$	alternative (prior) pdf of Θ conditioned on prior knowledge (data)
$f_t(\Theta) \equiv \beta_\alpha(s_1, s_2)$	true pdf of Θ conditioned on prior knowledge (data)
$\hat{f}_t(\Theta) \equiv \beta_\alpha(s_1', s_2')$	approximate posterior pdf of Θ conditioned on piror knowledge (data)
$\tilde{f}_t(\Theta) \equiv \sum_{i=1}^2 w_i \beta_{\alpha,i}(s_1, s_2)$	posterior mixture
$m_t(\Theta)$	model with parameter Θ conditioned on prior knowledge (data)
$w_i \in [0,1]$	weight of a beta mixture component
$\mathcal{N}_y(\mu, r)$	Gaussian distribution of y with nonnegative parameters (statistics) μ, r
$\alpha \in [0,1]$	weight of a Gaussian mixture component
$\beta_\alpha(s_1, s_2)$	beta distribution of α with parameters s_1, s_2
$\lambda \in [0,1]$	forgetting factor
$\psi(\cdot)$	digamma function of $\cdot$
Θ	parameter, possible multivariate

1 Two-component mixture

Let us have a two-component Gaussian mixture model with a multivariate parameter Θ

$$m_t(\Theta) \equiv m_t(y|\alpha, \mu_1, \mu_2, r_1, r_2) = \alpha \mathcal{N}_y(\mu_1, r_1) + (1-\alpha)\mathcal{N}_y(\mu_2, r_2), \tag{1}$$

where μ_1, μ_2, r_1, r_2 are known parameters and the task is to estimate α on base of a data set $y = \{y_1, \ldots, y_n\}$ generated by the mixture (with known α). From this time on, to preserve the consistence with the paper mentioned in the abstract, whenever a function of Θ (being a parameter) occurs, only the parameter α is thought. This is possible since we consider μ_i, r_i to be known and not estimated.

Since $\alpha \sim \beta_\alpha(s_1, s_2)$ and it is modelled on base of the data, the task consists in the search of the mean value of α, formally given by

$$\mathbb{E}[\alpha] = \frac{s_1}{s_1 + s_2}. \tag{2}$$

2 Data update of a conditional beta distribution

The data update reads

$$\tilde{f}_t(\Theta) \propto m_t(\Theta)\hat{f}_{t-1}(\Theta) \tag{3}$$

$$\propto \underbrace{\left[\alpha\mathcal{N}_y(\mu_1, r_1) + (1-\alpha)\mathcal{N}_y(\mu_2, r_2)\right]}_{m_t(\Theta)} \underbrace{\alpha^{s_1-1}(1-\alpha)^{s_2-1}}_{\hat{f}_{t-1}(\Theta)=\beta_\alpha(s_1,s_2)} \tag{4}$$

$$= \frac{\alpha^{s_1+1-1}(1-\alpha)^{s_2-1}}{B(s_1+1, s_2)} \underbrace{\frac{\mathcal{N}_y(\mu_1, r_1)}{\mathcal{N}_y(\mu_1, r_1) + \mathcal{N}_y(\mu_2, r_2)}}_{w_1} \tag{5}$$

$$+ \frac{\alpha^{s_1-1}(1-\alpha)^{s_2+1-1}}{B(s_1, s_2+1)} \underbrace{\frac{\mathcal{N}_y(\mu_1, r_1)}{\mathcal{N}_y(\mu_1, r_1) + \mathcal{N}_y(\mu_2, r_2)}}_{w_2} \tag{6}$$

$$= \beta_\alpha\left(s_1+1, s_2\right) w_1 + \beta_\alpha\left(s_1, s_2+1\right) w_2 \tag{7}$$

$$= \sum_{i=1}^{2} w_i \beta_{\alpha,i}, \tag{8}$$

where $\hat{f}_{t-1}(\Theta)$ is a prior pdf approximating the true but unknown $f_{t-1}(\Theta)$. The need for this approximation arises from the resulting data updated pdf, being a mixture.

The approximation is evaluated w.r.t. the Kullback-Leibler divergence [1] of two β distributions [2], formally

$$D(\beta, \beta') = \ln\frac{B(s_1', s_2')}{B(s_1, s_2)} - (s_1' - s_1)\psi(s_1) - (s_2' - s_2)\psi(s_2) + (s_1' - s_1 + s_2' - s_2)\psi(s_1 + s_2). \tag{9}$$

where, in our case, $\tilde{\beta} \equiv \beta$ and β' is the mixture (1); B and ψ are the beta and the digamma functions, respectively.

Differentiation of (9) w.r.t. s_1', s_2' and setting equal to zero yields

$$\frac{\partial D}{\partial s_1'} = \psi(s_1') - \psi(s_1' + s_2') - \sum_{i=1}^{2} w_i\psi(s_{i,1}) + \sum_{i=1}^{2} w_i\psi(s_{i,1} + s_{i,2}) = 0 \tag{10}$$

$$\frac{\partial D}{\partial s_2'} = \psi(s_2') - \psi(s_1' + s_2') - \sum_{i=1}^{2} w_i\psi(s_{i,2}) + \sum_{i=1}^{2} w_i\psi(s_{i,1} + s_{i,2}) = 0 \tag{11}$$

The resulting system of equations cannot be solved analytically. Therefore, we find an approximate solution numerically. For instance, in Matlab we can use the following m-function:

```
function f = kldivmin(x, w, s1, s2)
f = [psi(x(1)) - psi(x(1) + x(2)) - dot(w, psi(s1)) + dot(w, psi(s1+s2));
     psi(x(2)) - psi(x(1) + x(2)) - dot(w, psi(s2)) + dot(w, psi(s1+s2))]
```

and find the solution by calling

```
>> [x,fval] = fsolve(@kldivmin,x0,options, w, s1, s2)
```

where `x(1), x(2)` stand for s'_1 and s'_2; `w, s1, s2` are arbitrary-shape 2-vectors of weights and statistics and `x0` are any suitable initial values. `options` contains user-defined optimization options, consult `optimset` for more information.

Following the above steps we have obtained

$$\hat{f}_t(\alpha) = \beta_\alpha(s'_1, s'_2). \tag{12}$$

2.1 Approximation error

The approximation is surely not exact. We can use the Kullback-Leibler divergence to calculate the "approximation error". However, since this would be analytically not tractable, we will stick with the upper bound c of $D(\tilde{f}_t||\hat{f}_t)$ found using the Jensen's inequality [3]:

$$\begin{aligned}
D(\tilde{f}_t||\hat{f}_t) &= \int \tilde{f}_t(\alpha) \ln \frac{\tilde{f}_t(\alpha)}{\hat{f}_t(\alpha)} d\alpha && (13)\\
&= \int \sum_{i=1}^{2} w_i \beta_{\alpha,i} \ln \frac{\sum_{i=1}^{2} w_i \beta_{\alpha,i}}{\hat{\beta}_\alpha} d\alpha && (14)\\
&= \int \underbrace{\sum_{i=1}^{2} w_i \beta_{\alpha,i}}_{\mathbb{E}\beta} \underbrace{\ln \sum_{i=1}^{2} w_i \beta_{\alpha,i}}_{\ln \mathbb{E}\beta} d\alpha - \int \sum_{i=1}^{2} w_i \beta_{\alpha,i} \ln \hat{\beta}_\alpha d\alpha && (15)\\
&\leq \int \underbrace{\sum_{i=1}^{2} w_i \beta_{\alpha,i} \ln \beta_{\alpha,i}}_{\mathbb{E}\beta \ln \beta} d\alpha - \int \sum_{i=1}^{2} w_i \beta_{\alpha,i} \ln \hat{\beta}_\alpha d\alpha && (16)
\end{aligned}$$

From which it follows that

$$c = \sum_{i=1}^{2} \int w_i \beta_{\alpha,i} \ln \frac{\beta_{\alpha,i}}{\hat{\beta}_\alpha} d\alpha = \sum_{i=1}^{2} w_i \, D(\beta_{\alpha,i}||\hat{\beta}_\alpha) \tag{17}$$

Practically, c is obtained using (9).

3 Adaptive forgetting

Forgetting, also known as the time update, is often used to improve the adaptivity of estimation. We develop its idea for the beta pdf. Generally, the (stabilized) exponential forgetting method flattens the posterior pdf, which in our case has the following form:

$$\begin{aligned}
f_{t;\lambda_t} &\propto f_0^{1-\lambda_t} \hat{f}_t^{\lambda} && (18)\\
&\propto \left[\alpha^{s_1^*-1}(1-\alpha)^{s_2^*-1}\right]^{1-\lambda} \left[\alpha^{s'_1-1}(1-\alpha)^{s'_2-1}\right]^{\lambda} && (19)\\
&= \frac{\alpha^{\lambda(s'_1-s_1^*)+s_1^*-1}(1-\alpha)^{\lambda(s'_2-s_2^*)+s_2^*-1}}{B\Big(\lambda(s'_1-s_1^*)+s_1^*, \lambda(s'_2-s_2^*)+s_2^*\Big)}. && (20)
\end{aligned}$$

We search for λ_t such that

$$D(f_{t;\lambda_t}||\hat{f}_t) = c. \tag{21}$$

Since the special functions in (9) prevent analytical solution, we have to find the value of λ_t numerically. A very slight modification of the approach presented in Section 2 can be used.

4 Example

The following simple example briefly demonstrates the effects of forgetting used in Bayesian parameter estimation. The task is to estimate the parameter α given the set of 200 data randomly generated from the Gaussian mixture

$$\alpha\mathcal{N}_y(\mu_1, r_1) + (1-\alpha)\mathcal{N}_y(\mu_2, r_2) \tag{22}$$

where $\mu_1 = 0, \mu_2 = 50$ are mean values and $r_1 = 1, r_2 = 400$ are respective variances; the weight $\alpha = 0.9$. The alternative pdf $f_0 \equiv \beta_\alpha(10^{-6}, 10^{-6})$. The estimation was run in two scenarios:

Scenario 1: The estimation was initialized with wrong prior pdf $\hat{f}_0 \equiv \beta_\alpha(10, 90)$. It was expected that forgetting will gradually suppress this invalid information by stressing the available data.

Scenario 2: The estimation was initialized with credible prior pdf $\hat{f}_0 \equiv \beta_\alpha(90, 10)$. Forgetting was expected to *slightly* decrease the estimation quality due to the incorporation of flat alternative pdf.

Three approaches were tested: (i) estimation without forgetting, (ii) exponential forgetting with constant factor $\lambda_t = 0.985$ and (iii) adaptive exponential forgetting.

Table 2 depicts the estimation results in terms of a root mean square error (RMSE) for both scenarios. Figures 1 and 2 show the evolution of the absolute estimation error for estimation without forgetting and with adaptive forgetting. Fig. 3 depicts the evolution of forgetting factor of adaptive forgetting.

To conclude, the results were consistent with our expectation. If the decision making (estimation) is run under the proper knowledge, any forgetting leads to slightly worse results. However, the proper knowledge is rather rare in practice and the user faces uncertainty. Then, forgetting provides a way to reflect it and to improve estimation quality.

	No forgetting	EF (0.985)	AF
Scenario 1	0.456	0.366	0.320
Scenario 2	0.004	0.011	0.033

Table 2: Example: RMSE for two presented scenarios with no forgetting, exponential forgetting (EF) with factor 0.985 and adaptive forgetting (AF).

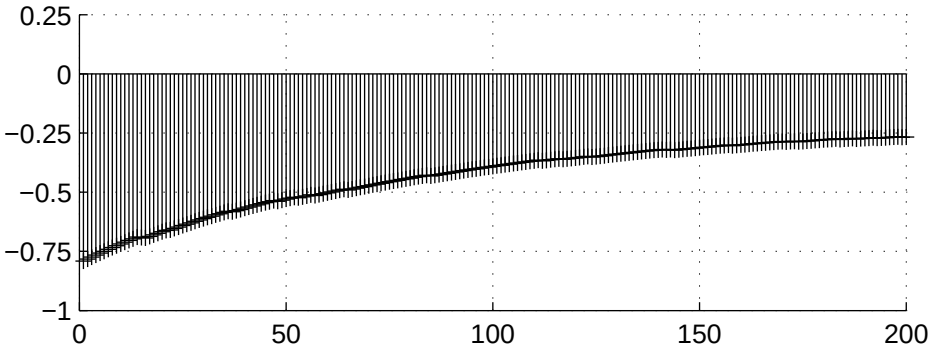

Figure 1: Estimation without forgetting – evolution of absolute estimation error.

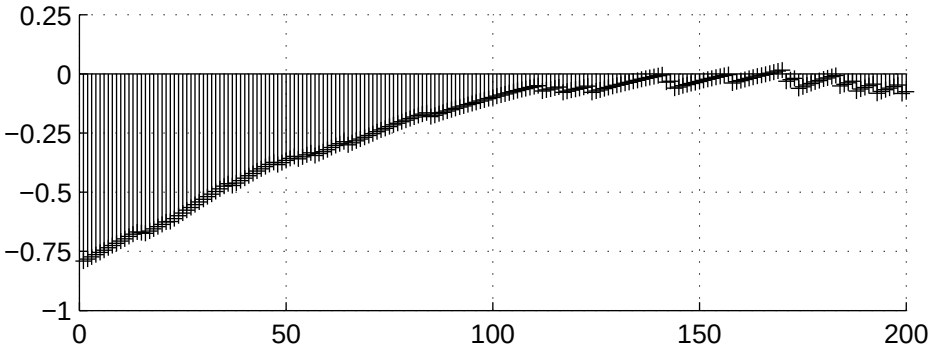

Figure 2: Estimation with adaptive forgetting – evolution of absolute estimation error.

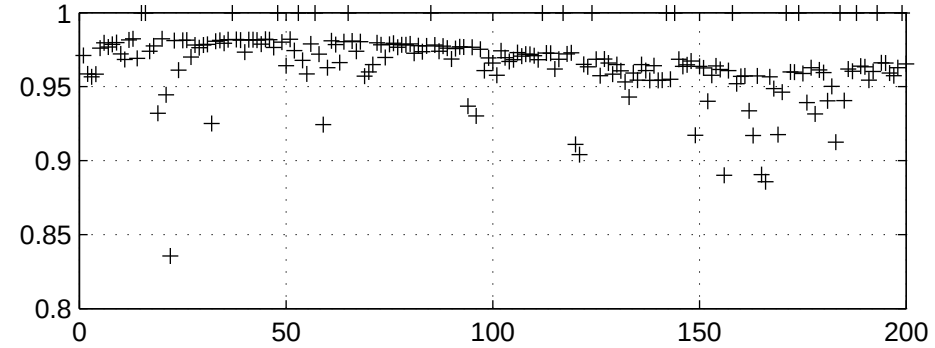

Figure 3: Estimation with adaptive forgetting – evolution of forgetting factor.

References

[1] S. Kullback and R. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[2] A. Lazo and P. Rathie, "On the entropy of continuous probability distributions (corresp.)," *Information Theory, IEEE Transactions on*, vol. 24, no. 1, pp. 120–122, 1978.

[3] J. Jensen, "Sur les fonctions convexes et les inégalités entre les valeurs moyennes," *Acta Mathematica*, vol. 30, no. 1, pp. 175–193, 1906.