Distributed Estimation of Mixture Models

Kamil Dedecius and Jan Reichl

Abstract The contribution deals with sequential distributed estimation of global parameters of normal mixture models, namely mixing probabilities and component means and covariances. The network of cooperating agents is represented by a directed or undirected graph, consisting of vertices taking observations, incorporating them into own statistical knowledge about the inferred parameters and sharing the observations and the posterior knowledge with other vertices. The aim to propose a computationally cheap online estimation algorithm naturally disqualifies the popular (sequential) Monte Carlo methods for the associated high computational burden, as well as the expectation-maximization (EM) algorithms for their difficulties with online settings requiring data batching or stochastic approximations. Instead, we proceed with the quasi-Bayesian approach, allowing sequential analytical incorporation of the (shared) observations into the normal inverse-Wishart conjugate priors. The posterior distributions are subsequently merged using the Kullback-Leibler optimal procedure.

Key words: Mixture estimation, distributed estimation, quasi-Bayesian estimation.

1 Introduction

The rapid development of ad-hoc networks and the emergence of the so-called big data phenomenon have brought new challenges for distributed statistical data processing. For instance, the processing often needs to be decentralized, i.e. without

Kamil Dedecius
Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 4, 182 08 Prague 8, Czech Republic, e-mail: dedecius@utia.cas.cz

Jan Reichl
Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 4, 182 08 Prague 8, Czech Republic, e-mail: reichja3@fjfi.cvut.cz

any dedicated unit in the network. Instead, *all* agents are responsible for (i) taking measurements, (ii) processing them, and (iii) sharing the statistical knowledge about the (usually global) inferred parameters. In addition, the estimation should run online in many cases. This means to take observations of a dynamic process and incorporate them sequentially into the shared knowledge. This often disqualifies the popular sequential Monte Carlo (MC) approaches for the associated high computational burden. Very rich surveys on distributed estimation are the recent papers by Sayed [10] (non-MC) and Hlinka et al. [5] (MC-based).

Despite the great potential of the Bayesian paradigm in this field, its adoption is still rather an exception than a rule. From the probabilistic viewpoint, the resulting "classical" (that is, non-Bayesian) algorithms often suffer statistical inconsistencies. For instance point estimators are often combined without reflecting the associated uncertainty, which may lead to statistically absurd situations. The first author's work [1] aims to partially fill this gap. It proposes a fully Bayesian approach to decentralized distributed estimation with a fusion based on minimization of the Kullback-Leibler divergence. The present contribution extends the results to the case of mixture models, otherwise covered for the static cases, e.g. in [3, 13, 8].

The novelty of the proposed framework lies in a fully analytical Bayesian processing of observations and shared knowledge about the estimated parameters. To this end, the underlying theory relies on the quasi-Bayesian approach, proposed by Smith, Makov and Titterington [11, 12] and followed by Kárný et al. [6], whose approach is adopted here. It provides analytical tractability of mixture inference by relying on point estimators where necessary. Though we focus on normal mixtures, the results are applicable to homogeneous mixtures of exponential family distributions.

2 Quasi-Bayesian Estimation of Mixture Models

Consider an observable time series $\{Y_t,\ t \in \mathbb{N}\}$ with $Y_t \in \mathbb{R}^n$ following a normal mixture distribution

$$\begin{aligned} Y_t|\phi,\theta &\sim \phi_1 \mathrm{N}(\mu_1,\Sigma_1) + \ldots + \phi_K \mathrm{N}(\mu_K,\Sigma_K) \\ &\sim \phi_1 \mathrm{N}(\theta_1) + \ldots + \phi_K \mathrm{N}(\theta_K), \end{aligned} \tag{1}$$

where $\mathrm{N}(\mu_k,\Sigma_k)$ denotes the kth component – a normal distribution with a mean vector $\mu_k \in \mathbb{R}^n$ and a covariance matrix $\Sigma_k \in \mathbb{R}^{n\times n}$, in the latter notation summarized by $\theta_k = \{\mu_k,\Sigma_k\}$. The nonnegative variables ϕ_k taking values in the unit K-simplex are the component probabilities. The number of components K is assumed known a priori. Furthermore, the notation $\theta = \{\theta_1,\ldots,\theta_K\}$, $\phi = \{\phi_1,\ldots,\phi_K\}$ is used.

Let $p_k(y_k|\theta_k)$ be the probability density function of the kth component, yielding the mixture density of the form

$$p(y_t|\theta,\phi) = \sum_{k=1}^{K} \phi_k p_k(y_t|\theta_k). \tag{2}$$

At each time instant t the observation y_t is generated by the k_tth component $p_k(y_t|\theta_k)$, selected with probability ϕ_k,

$$p(y_t|\theta,\phi,k_t) = \prod_{k=1}^{K} [\phi_k p_k(y_t|\mu_k,\Sigma_k)]^{S_{k,t}}, \tag{3}$$

where $S_{k,t}$ is the indicator function of the active component

$$S_{k,t} = \begin{cases} 1 & \text{if } S_{k,t} = k_t, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

In other words, $S_t = (S_{1,t},\dots,S_{K,t})$ can be viewed as a vector with 1 on the k_t'th position and zeros elsewhere, and hence follow the multinomial distribution $\text{Multi}(1,\phi)$.

From the Bayesian viewpoint the topological property of ϕ is crucial, as it allows its modelling with the Dirichlet distribution with parameters $\kappa_1,\dots,\kappa_K$,

$$\phi = (\phi_1,\dots,\phi_K) \sim \text{Dir}(\kappa_1,\dots,\kappa_K), \qquad \kappa_k > 0 \quad \text{for all} \quad k = 1,\dots,K,$$

conjugate to the multinomial distribution of S_t. The sequential estimation of each single component mean and covariance can then proceed with the conjugate normal inverse-Wishart distribution (or normal inverse-gamma in the univariate case),

$$\theta_k = \{\mu_k,\Sigma_k\} \sim \text{NiW}(m,s,a,b), \qquad m \in \mathbb{R}^n, \;\; s \in \mathbb{R}^{n\times n}, \;\; a,b > 0.$$

Exact knowledge of S_t would make the Bayesian inference of both the component parameters μ_k,Σ_k and mixing probabilities ϕ easily tractable, since the product (3) simplifies to a single density and a single component probability. Likewise, the Bayesian inference of mixing probabilities ϕ is easy under known components, as the detection of the active one is a relatively simple hypotheses testing problem, see e.g. [4]. However, our attention is shifted towards estimating both component parameters μ,Σ and mixing probabilities ϕ. For this sake, we need to derive the Bayesian update

$$\pi_{\phi,\theta}(\phi,\theta|y_{1:t},k_{1:t}) \propto \pi_{\phi,\theta}(\phi,\theta|y_{1:t-1},k_{1:t-1}) \prod_{k=1}^{K} [\phi_k p_k(y_t|\theta_k)]^{S_{k,t}}$$

where the joint prior distribution is assumed to be

$$\pi_{\phi,\theta}(\phi,\theta|y_{1:t-1},k_{1:t-1}) = \pi_\phi(\phi|y_{1:t-1},k_{1:t-1})\pi_\theta(\theta|y_{1:t-1},k_{1:t-1}).$$

The independence of ϕ and θ allows tractable computation of the posterior distribution. Indeed, this assumption is not quasi-Bayes specific.

In this case, Kárný et al. [6] propose to rely on the approach of Smith, Makov and Titterington [11, 12] and replace the latent indicators $S_{k,t}$ – Equation (4) – by their respective point estimates with respect to ϕ_k and θ_k of the form

$$\begin{aligned}\widehat{S}_{k,t} &= \mathbb{E}\left[S_{k,t}|y_{1:t},k_{1:t-1}\right] \\ &\propto \mathbb{E}\left[\phi_k|y_{1:t-1},k_{1:t-1}\right] p_k(y_t|y_{1:t-1},k_{1:t-1}),\end{aligned} \tag{5}$$

where

$$p_k(y_t|y_{1:t-1},k_{1:t-1}) = \int p_k(y_t|\theta_k)\pi_{\theta_k}(\theta_k|y_{1:t-1},k_{1:t-1})\mathrm{d}\theta_k \tag{6}$$

is the predictive distribution (under normal inverse-Wishart prior it is Student's t distribution). To summarize, the estimation of the indicator $S_{k,t}$ of the active component k is based on (i) testing the component membership based on the predictive likelihood (6), and (ii) the estimated probability of the particular component $\mathbb{E}[\phi_k|\cdot]$ in (5).

The quasi-Bayesian update then takes the weighted form of the regular update under known S_t,

$$\pi_\phi(\phi|y_{1:t},k_{1:t}) \propto \mathbb{E}\left[\widehat{S}_t|y_{1:t},k_{1:t-1}\right]\pi_\phi(\phi|y_{1:t-1},k_{1:t-1}), \tag{7}$$

$$\pi_{\theta_k}(\theta_k|y_{1:t},k_{1:t}) \propto \left[p_k(y_t|\theta_k)\right]^{\widehat{S}_{k,t}}\pi_{\theta_k}(\theta_k|y_{1:t-1},k_{1:t-1}). \tag{8}$$

If the component density is rewritten to the exponential family form and the prior density to its conjugate form as shown in Appendix, the update of the relevant hyperparameters is particularly easy.

3 Distributed Estimation

Assume that the distributed estimation runs in a network represented by a directed or undirected connected graph $\mathrm{G}(V,E)$ consisting of a set of vertices $V = \{1,\ldots,N\}$ (also called nodes or agents) and a set E of edges, defining the graph topology. The vertices $n \in V$ are allowed to communicate with adjacent vertices: For a fixed vertex n, these neighbors form a complete bipartite subgraph (every neighboring vertex is connected with n) with radius 1, diameter at most 2 and of type star (unless the vertex n is of degree 1), where n is then the central vertex and all other vertices peripheral. We denote the set of vertices of this subgraph by V_n.

The vertices independently observe the process $\{Y_t, t \in \mathbb{N}\}$, taking observations $y_t^{(n)}, n \in V$. These are shared within V_n in the sense that each vertex n has access to $y_t^{(j)}$ of vertices $j \in V_n$ and incorporates them according to the quasi-Bayesian estimation theory portrayed in the previous section. That is, each node n ends with the joint posterior density

$$\pi^{(n)}_{\phi,\theta}(\phi,\theta|\widetilde{y}_{1:t},\widetilde{k}_{1:t}), \tag{9}$$

resulting from the number of cardV_n updates of the form (7) and (8). Here tilde denotes the statistical knowledge comprising the V_n's information relevant to the particular variable. This step is called *adaptation*, e.g. [10].[1]

3.1 Combination of estimates

In the *combination* step [10], the vertices $n \in V$ access V_n's posterior distributions (9) resulting from the adaptation,

$$\pi^{(j)}_{\phi,\theta}(\phi,\theta|\widetilde{y}_{1:t},\widetilde{k}_{1:t}), \qquad j \in V_n.$$

Now the goal is to represent (i.e. approximate) them by a *single* joint posterior $\widetilde{\pi}^{(n)}_{\phi,\theta}$ parameterizing the mixture (1) in consideration. To this end, we adopt the Kullback-Leibler divergence [7] defined in the Appendix, and seek for $\widetilde{\pi}^{(n)}_{\phi,\theta}$ satisfying

$$\sum_{j\in V_n} \alpha_{nj} \mathrm{D}(\widetilde{\pi}^{(n)}_{\phi,\theta}||\pi^{(j)}_{\phi,\theta}) \to \min, \tag{10}$$

where $\alpha_{nj} = 1/(\mathrm{card}\, V_n)$ are nonnegative uniform weights assigned to nodes $j \in V_n$ summing to unity. Other weight choices, e.g. reflecting properties of the neighboring vertices are possible as well.

Let us impose an additional assumption simplifying the theory: We assume identical order of component parameters and significantly overlapping densities $\pi^{(j)}_{\phi,\theta}$ of all $j \in V_n$. This means that the order of components and their parameterization agrees at all vertices in V_n (and hence V). This assumption can be easily removed by incorporating detection of similar posterior distributions or enforced by starting from identical initial priors.

We exploit the following general proposition proved, e.g., in [1]. Thought we consider exponential family distributions (where it provides analytically tractable results), the proposition is not limited to them.

Proposition 1. *Let $\pi^{(j)}_{\phi,\theta}$ be the posterior probability density functions of vertices $j \in V_n$ and α_{nj} their weights from the unit* cardV_n*-simplex. Their approximation by a single density $\widetilde{\pi}^{(n)}_{\phi,\theta}$ optimal in the Kullback-Leibler sense* (10) *has the form*

$$\widetilde{\pi}_{\phi,\theta} \propto \prod_{j\in V_n} \left[\pi^{(j)}_{\phi,\theta}\right]^{\alpha_{nj}}. \tag{11}$$

The resulting approximate posterior density hence virtually parameterizes a much richer mixture, however, the individual densities overlap by the given assumption. Then Proposition (11) gives a method for reduction to the parametrization of

[1] The terms "adaptation" and "combination" were introduced by Sayed. We adopt them for our Bayesian counterparts.

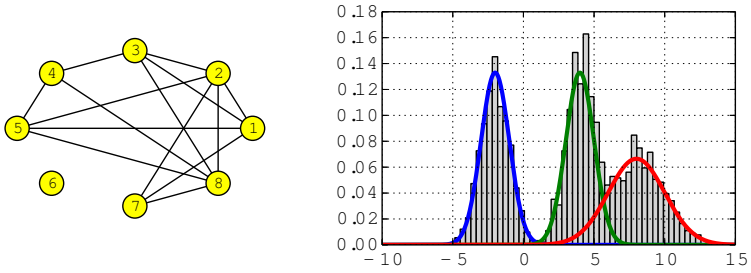

Fig. 1 Left: Layout of the graph with isolated node 6 for comparison. Right: Normalized histogram and true components of the mixture.

K components,

$$\widetilde{\pi}_{\phi}^{(n)} \propto \prod_{j \in V_n} \left[\pi_{\phi}^{(j)}\right]^{\alpha_{nj}} \qquad \text{and} \qquad \widetilde{\theta}_{\phi}^{(n)} \propto \prod_{j \in V_n} \left[\theta_{\phi}^{(j)}\right]^{\alpha_{nj}},$$

which, due to the structure of conjugate priors (see Appendix) and component ordering yields

$$\widetilde{\xi}_{k,t}^{(n)} = \sum_{j \in V_n} \alpha_{nj} \xi_{k,t}^{(j)}, \quad \widetilde{\nu}_{k,t}^{(n)} = \sum_{j \in V_n} \alpha_{nj} \nu_{k,t}^{(j)} \qquad \text{and} \qquad \widetilde{\kappa}_{k,t}^{(n)} = \sum_{j \in V_n} \alpha_{nj} \kappa_{k,t}^{(j)}.$$

for the hyperparameters ξ, ν and κ of the prior distributions for θ and ϕ, respectively. The resulting KL-optimal posterior is then again conjugate to the model and can be used for the subsequent adaptation step.

4 Simulation Example

The simulation example deals with estimating a three component normal model, for simplicity univariate of the form

$$Y \sim \frac{1}{3}\mathrm{N}(-2,1) + \frac{1}{3}\mathrm{N}(4,1) + \frac{1}{3}\mathrm{N}(8,2)$$

with unknown means and variances. The graph $\mathrm{G}(V,E)$, whose scheme is depicted together with the components and samples in Fig. 1, consists of a set of vertices $V = \{1,\ldots,8\} \setminus \{6\}$. The 6th vertex is disconnected and serves for comparison. The vertices $n \in V \cup \{6\}$ take observations $y_t^{(n)}$ with $t = 1,\ldots,300$. Clearly, one would expect relatively easy identification of the leftmost component, while the other two may be problematic due to their closeness. The quasi-Bayesian estimation of components $k \in \{1,2,3\}$ exploits the conjugate normal inverse-gamma prior

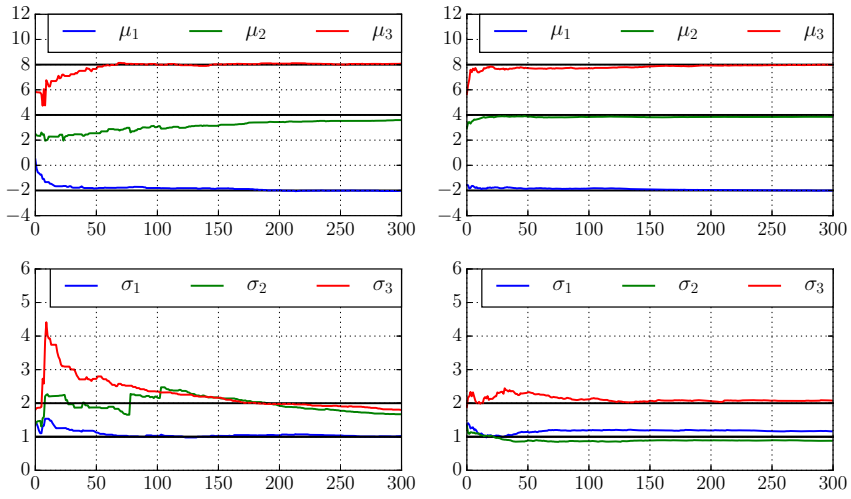

Fig. 2 Evolution of estimates of component means and standard deviations. Left: isolated vertex 6. Right: situation at a chosen cooperating vertex 4. Solid black lines depict true values.

$\mathrm{NiG}(\mu_k,\sigma_k;m_k,s_k,a_k,b_k) = \mathrm{N}(\mu_k|\sigma_k^2;m_k,\sigma^2 s_k) \times \mathrm{iG}(\sigma_k^2;a_k,b_k)$ with initial hyperparameters m_k set to 0, 3, and 6, respectively; the other hyperparameters $s_k = 1, a_k = 2, b_k = 2$ for all k. The prior for component probabilities $\phi \sim \mathrm{Dir}(\frac{1}{3},\frac{1}{3},\frac{1}{3})$. This initialization is identical across the graph.

The progress of the point estimates of μ_k and σ_k is depicted in Fig. 2 — isolated vertex 6 (left) and randomly chosen vertex 4 (right). The point estimates of μ_k converge relatively well in both cases, however, the variance estimates converge well only in the case of the distributed estimation (with the exception of σ_1^2). This is due to the much richer data available to the interconnected vertices. The mean squared errors (MSE) of final estimates are given in Table 1.

Table 1 Statistics of mean square errors (MSEs) of resulting estimates: distributed estimation and isolated vertex 6.

MSE	min (distr.)	max (distr.)	mean (distr.)	Vertex 6
Means μ_k	0.007	0.007	0.007	0.057
Variances σ_k^2	0.092	0.481	0.222	1.26
Comp. probabilities ϕ	0	0	0	0.001

5 Conclusion

The quasi-Bayesian method for analytically tractable sequential inference of parameters of probabilistic mixtures has been extended to the case of distributed estimation of normal mixture model with unknown mixing probabilities and component parameters. Here, *distributed* means that there is a graph (network) of cooperating vertices (nodes, agents) sharing their statistical knowledge (observations and estimates) with a limited subset of other vertices. This knowledge is combined at each vertex: the observations are incorporated by means of the Bayes' theorem, the estimates are combined via the Kullback-Leibler optimal rule.

The main advantage of the method is its simplicity and scalability. Unlike Monte Carlo approaches, it is computationally very cheap. The authors have recently shown [2] (*to appear*) that this method is suitable for the whole class of mixture models consisting of exponential family distributions and prior distributions conjugate to them.

The difficulty associated with the method is common for most mixture estimation methods: the initialization. Also, merging and splitting of components after the combination of estimates would significantly enhance the suitability of the approach for real dynamic cases. These topics remain for further research.

Appendix

Below we give several useful definitions and lemmas regarding the Bayesian estimation of exponential family distributions with conjugate priors [9]. The proofs are trivial. Their application to the normal model and normal inverse-gamma prior used in Section 4 follows.

Definition 1 (Exponential family distributions and conjugate priors). Any distribution of a random variable y parameterized by θ with the probability density function of the form

$$p(y|\theta) = f(y)g(\theta)\exp\left\{\eta(\theta)^\intercal T(y)\right\},$$

where f, g, η and T are known functions, is called an exponential family distribution. $\eta \equiv \eta(\theta)$ is its natural parameter, $T(y)$ is the (dimension preserving) sufficient statistic. The form is not unique.

Any prior distribution for θ is said to be conjugate to $p(y|\theta)$, if it can be written in the form

$$\pi(\theta|\xi,\nu) = q(\xi,\nu)g(\theta)^\nu \exp\left\{\eta(\theta)^\intercal \xi\right\}.$$

where q is a known function and the hyperparameters $\nu \in \mathbb{R}^+$ and ξ is of the same shape as $T(y)$.

Lemma 1 (Bayesian update with conjugate priors). *Bayes' theorem*

$$\pi(\theta|\xi_t,\nu_t) \propto p(y_t|\theta)\pi(\theta|\xi_{t-1},\nu_{t-1})$$

yields the posterior hyperparameters as follows:

$$\xi_t = \xi_{t-1} + T(y_t) \qquad \textit{and} \qquad \nu_t = \nu_{t-1} + 1.$$

Lemma 2. *The normal model*

$$p(y_t|\mu,\sigma^2) = \frac{(\sigma^2)^{-\frac{1}{2}}}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(y_t-\mu)^2\right\}$$

where μ,σ^2 are unknown can be written in the exponential family form with

$$\eta = \left(\frac{\mu}{\sigma^2}, \frac{-1}{2\sigma^2}, \frac{-\mu^2}{2\sigma^2}\right)^\intercal, \qquad T(y_t) = \left(y, y^2, 1\right)^\intercal, \qquad g(\eta) = \left(\sigma^2\right)^{-\frac{1}{2}}.$$

Lemma 3. *The normal inverse-gamma prior distribution for μ,σ^2 with the (nonnatural) real scalar hyperparameters m, and positive s,a,b and having the density*

$$p(\mu,\sigma^2|m,s,a,b) = \frac{b^a(\sigma^2)^{a+1+\frac{1}{2}}}{\sqrt{2\pi}s\Gamma(a)} \exp\left\{-\frac{1}{\sigma^2}\left[b + \frac{1}{2s}(m-\mu)^2\right]\right\}$$

can be written in the prior-conjugate form with

$$\xi_t = \left(\frac{m}{s}, \frac{m^2}{s} + 2b, \frac{1}{s}\right)^\intercal.$$

Lemma 4. *The Bayesian update of the normal inverse-gamma prior following the previous lemma indeed coincides with the 'ordinary' well-known update of the original hyperparameters,*

$$s_t^{-1} = s_{t-1}^{-1} + 1, \qquad a_t = a_{t-1} + \frac{1}{2},$$
$$m_t = s_t\left(\frac{m_{t-1}}{s_{t-1}} + y_t\right), \qquad b_t = b_{t-1} + \frac{1}{2}\left(\frac{m_{t-1}^2}{s_{t-1}} - \frac{m_t^2}{s_t} + y_t^2\right).$$

Definition 2 (Kullback-Leibler divergence). Let $f(x), g(x)$ be two probability density functions of a random variable x, f absolutely continuous with respect to g. The Kullback-Leibler divergence is the nonnegative functional

$$\mathrm{D}(f||g) = \mathbb{E}_f\left[\log\frac{f(x)}{g(x)}\right] = \int f(x)\log\frac{f(x)}{g(x)}dx, \tag{12}$$

where the integration domain is the support of f. The Kullback-Leibler divergence is a premetric; it is zero if $f = g$ almost everywhere, it does not satisfy the triangle inequality nor is it symmetric.

Acknowledgements This work was supported by the Czech Science Foundation, postdoctoral grant no. 14–06678P. The authors thank the referees for their valuable comments.

References

1. Dedecius, K., Sečkárová, V.: Dynamic Diffusion Estimation in Exponential Family Models. IEEE Signal Processing Letters **20**(11), 1114–1117 (2013).
2. Dedecius, K., Reichl, J., Djurić, P.M.: Sequential Estimation of Mixtures in Diffusion Networks. IEEE Signal Processing Letters **22**(2), 197–201 (2015).
3. Dongbing Gu: Distributed EM Algorithm for Gaussian Mixtures in Sensor Networks. IEEE Transactions on Neural Networks **19**(7), 1154–1166 (2008).
4. Frühwirth-Schnatter, S.: Finite Mixture and Markov Switching Models. Springer, London (2006)
5. Hlinka, O., Hlawatsch, F., Djurić, P.M.: Distributed Particle Filtering in Agent Networks: A Survey, Classification, and Comparison. IEEE Signal Processing Magazine **30**(1), 61–81 (2013).
6. Kárný, M., Böhm, J., Guy, T.V., Jirsa, L., Nagy, I., Nedoma, P., Tesař, L.: Optimized Bayesian Dynamic Advising: Theory and Algorithms. Springer, London (2006).
7. Kullback, S., Leibler, R.A.: On Information and Sufficiency. The Annals of Mathematical Statistics **22**(1), 79–86 (1951).
8. Pereira, S.S., Lopez-Valcarce, R., Pages-Zamora, A.: A Diffusion-Based EM Algorithm for Distributed Estimation in Unreliable Sensor Networks. IEEE Signal Processing Letters **20**(6), 595–598 (2013).
9. Raiffa, H., Schlaifer, R.: Applied Statistical Decision Theory (Harvard Business School Publications). Harvard University Press (1961).
10. Sayed, A.H.: Adaptive Networks. Proceedings of the IEEE **102**(4), 460–497 (2014).
11. Smith, A.F.M., Makov, U.E.: A Quasi-Bayes Sequential Procedure for Mixtures. Journal of the Royal Statistical Society. Series B (Methodological) **40**(1), 106–112 (1978).
12. Titterington, D.M., Smith, A.F.M., Makov, U.E.: Statistical Analysis of Finite Mixture Distributions. John Wiley (1985)
13. Weng, Y., Xiao, W., Xie, L.: Diffusion-Based EM Algorithm for Distributed Estimation of Gaussian Mixtures in Wireless Sensor Networks. Sensors **11**(6), 6297–316 (2011).