



Algorithms for initialization of Gaussian Mixture Models

Una Radojčić

**UNIVERSITY J. J. STROSSMAYER OF OSIJEK
DEPARTMENT OF MATHEMATICS**

Trg Ljudevita Gaja 6

31000 Osijek, Croatia

<http://www.mathos.unios.hr>

uradojic@mathos.hr



HRZZ

Hrvatska zaklada
za znanost

[SUPPORTED BY CSF THROUGH RESEARCH GRANT IP-2016-06-6545]

30.7.2019.



What is a Gaussian Mixture model?

Definition

The Gaussian Mixture Model (GMM) with m components is a parametric probability density function represented as a convex combination of m Gaussian densities as given by the equation,

$$p(x|\lambda) = \sum_{i=1}^m w_i \phi_{\mu_i, \Sigma_i}(x),$$

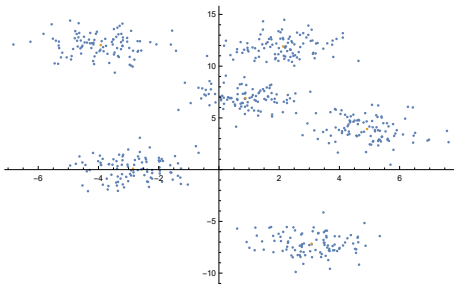
where $x \in \mathbb{R}^d$ is a data point, w_i are the mixture weights satisfying the constraint $\sum_{i=1}^m w_i = 1$, and ϕ_{μ_i, Σ_i} , $i = 1, \dots, m$, are the Gaussian densities of the form,

$$\phi_{\mu_i, \Sigma_i} = \frac{1}{(2\pi)^{d/2} |\Sigma_i|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)}.$$



Motivation

- GMM-s are used for modeling data in case of multiple underlying populations which each can be explained by a Gaussian distributions.
- Such data sets arise in various natural phenomena.





High dimensional problems

- Market research
- Document analysis
- Image processing
- Gene categorization
- Fraud detection



Model

The goal is to estimate location parameters $\underline{\Theta} = (\theta_1, \dots, \theta_k)$, $\theta_i \in \mathbb{R}^d$, in Radial Basis Gaussian Mixture Model (GRBMM),

$$\prod_{i=1}^n p_{\theta}(x_i) \approx \prod_{i=1}^n \sum_{j=1}^k e^{\frac{-1}{2\sigma^2} \|x_i - \theta_j\|^2}. \quad (1)$$

where $\underline{X} = \{x_1, \dots, x_n\}$, $x_i \in \mathbb{R}^d$ is data point set coming from a mixture distribution.

We assume that σ is known.



- Parameter estimation is done by maximization of the likelihood function
- In case of GMM the likelihood function of the model is not convex
- One of the most widely used algorithms for parameter estimation in GMM is Expectation-Maximization algorithm
- Problem is in convergence towards local instead of global maximum
- The key in solving the problem is finding a good initial approximation to such method



How to choose an initial approximation such that the MLE gives a reasonable estimation for $\underline{\Theta}$?

- Randomly choose location parameters from the set of data points
- Generate location parameters $\underline{\Theta}$ from, let's say, Gaussian distribution
- A sophisticated, way of initializing $\underline{\Theta}$ would be by generating it from a distribution of $\underline{\Theta}$ that has (normalized) likelihood as a density

Use Markov Chain Monte Carlo (MCMC) methods

Time until the steady state can be exponentially large



How to choose an initial approximation such that the MLE gives a reasonable estimation for $\underline{\Theta}$?

- Randomly choose location parameters from the set of data points
- Generate location parameters $\underline{\Theta}$ from, let's say, Gaussian distribution
- A sophisticated, way of initializing $\underline{\Theta}$ would be by generating it from a distribution of $\underline{\Theta}$ that has (normalized) likelihood as a density

Use Markov Chain Monte Carlo (MCMC) methods

Time until the steady state can be exponentially large



How to choose an initial approximation such that the MLE gives a reasonable estimation for $\underline{\Theta}$?

- Randomly choose location parameters from the set of data points
- Generate location parameters $\underline{\Theta}$ from, let's say, Gaussian distribution
- A sophisticated, way of initializing $\underline{\Theta}$ would be by generating it from a distribution of $\underline{\Theta}$ that has (normalized) likelihood as a density

Use Markov Chain Monte Carlo (MCMC) methods

Time until the steady state can be exponentially large



Swapping Algorithm

For given data points $\underline{X} = \{x_1, \dots, x_n\}$, $x_i \in \mathbb{R}^d$, we wish to generate location parameters $\underline{\Theta} = (\theta_1, \dots, \theta_k)$, $\theta_i \in \mathbb{R}^d$, from distribution with density equal to normalized likelihood

$$\prod_{i=1}^n p_{\underline{\Theta}}(x_i) \approx \prod_{i=1}^n \sum_{j=1}^k e^{\frac{-1}{2\sigma^2} \|x_i - \theta_j\|^2}. \quad (2)$$

Definition

Let $\mathcal{V} = \{1, \dots, k\}^n$ be the set of all k^n possible assignments of n data points to k clusters. Let $\underline{Z} = (Z_1, \dots, Z_n) \in \mathcal{V}$ be one labeling of data points where $Z_i = j \in \{1, \dots, k\}$ labels point x_i as a point from j -th cluster.



Using introduced labelings, we can rewrite likelihood of model as

$$\begin{aligned} \prod_{i=1}^n p_{\underline{\theta}}(x_i) &\approx \prod_{i=1}^n \sum_{j=1}^k e^{\frac{-1}{2\sigma^2} \|x_i - \theta_j\|^2} = \sum_{\underline{Z} \in \mathcal{V}} \prod_{i=1}^n e^{\frac{-1}{2\sigma^2} \|x_i - \theta_{Z_i}\|^2} \\ &= \sum_{\underline{Z} \in \mathcal{V}} \left(\prod_{j=1}^k n_{\underline{Z},j}^{-d/2} e^{\frac{-1}{2\sigma^2} \sum_{i:Z_i=j} \|x_i - \bar{X}_{\underline{Z},j}\|^2} \right) \\ &\quad \times \left(\prod_{j=1}^k n_{\underline{Z},j}^{d/2} e^{\frac{-1}{2\sigma^2} n_{\underline{Z},j} \|\bar{X}_{\underline{Z},j} - \theta_{Z_i}\|^2} \right), \end{aligned}$$

where $n_{\underline{Z},j}$ is number of data points assigned to j -th cluster according to labeling \underline{Z} , and $\bar{X}_{\underline{Z},j}$ is the corresponding mean.



- In many labelings, there are clusters with no observations
- In a Bayesian framework, we can rewrite the posterior as a proper Gaussian mixture by using independent Gaussian priors $\mathcal{N}(\alpha, \frac{\sigma^2}{\beta} I_d)$ for $\theta_1, \dots, \theta_k$, where α and β are fixed parameters

$$p(\underline{\Theta} | \underline{X}) \approx \sum_{\underline{Z} \in \mathcal{V}} \prod_{j=1}^k \frac{1}{(\beta + n_{\underline{Z},j})^{d/2}} e^{-\frac{1}{2\sigma^2} \left(\frac{n_{\underline{Z},j}\beta}{n_{\underline{Z},j} + \beta} \|\bar{X}_{\underline{Z},j} - \alpha\|^2 + \sum_{i: Z_i=j} \|x_i - \bar{X}_{\underline{Z},j}\|^2 \right)}$$

$$\times \prod_{j=1}^k (\beta + n_{\underline{Z},j})^{d/2} e^{-\frac{\beta + n_{\underline{Z},j}}{2\sigma^2} \|\theta_j - \tilde{\theta}_j\|^2},$$

where $\tilde{\theta}_j = \frac{n_{\underline{Z},j} \bar{X}_{\underline{Z},j} + \beta \alpha}{n_{\underline{Z},j} + \beta}$.



- In many labelings, there are clusters with no observations
- In a Bayesian framework, we can rewrite the posterior as a proper Gaussian mixture by using independent Gaussian priors $\mathcal{N}(\alpha, \frac{\sigma^2}{\beta} I_d)$ for $\theta_1, \dots, \theta_k$, where α and β are fixed parameters

$$p(\underline{\Theta} | \underline{X}) \approx \sum_{\underline{Z} \in \mathcal{V}} \prod_{j=1}^k \frac{1}{(\beta + n_{\underline{Z},j})^{d/2}} e^{-\frac{1}{2\sigma^2} \left(\frac{n_{\underline{Z},j}\beta}{n_{\underline{Z},j} + \beta} \|\bar{X}_{\underline{Z},j} - \alpha\|^2 + \sum_{i: Z_i=j} \|x_i - \bar{X}_{\underline{Z},j}\|^2 \right)}$$

$$\times \prod_{j=1}^k (\beta + n_{\underline{Z},j})^{d/2} e^{-\frac{\beta + n_{\underline{Z},j}}{2\sigma^2} \|\theta_j - \tilde{\theta}_j\|^2},$$

where $\tilde{\theta}_j = \frac{n_{\underline{Z},j} \bar{X}_{\underline{Z},j} + \beta \alpha}{n_{\underline{Z},j} + \beta}$.



- In many labelings, there are clusters with no observations
- In a Bayesian framework, we can rewrite the posterior as a proper Gaussian mixture by using independent Gaussian priors $\mathcal{N}(\alpha, \frac{\sigma^2}{\beta} I_d)$ for $\theta_1, \dots, \theta_k$, where α and β are fixed parameters

$$p(\underline{\Theta} | \underline{X}) \approx \sum_{\underline{Z} \in \mathcal{V}} \prod_{j=1}^k \frac{1}{(\beta + n_{\underline{Z},j})^{d/2}} e^{-\frac{1}{2\sigma^2} \left(\frac{n_{\underline{Z},j}\beta}{n_{\underline{Z},j} + \beta} \|\bar{X}_{\underline{Z},j} - \alpha\|^2 + \sum_{i: Z_i=j} \|x_i - \bar{X}_{\underline{Z},j}\|^2 \right)}$$

$$\times \prod_{j=1}^k (\beta + n_{\underline{Z},j})^{d/2} e^{-\frac{\beta + n_{\underline{Z},j}}{2\sigma^2} \|\theta_j - \tilde{\theta}_j\|^2},$$

where $\tilde{\theta}_j = \frac{n_{\underline{Z},j} \bar{X}_{\underline{Z},j} + \beta \alpha}{n_{\underline{Z},j} + \beta}$.



Let us denote

$$w(\underline{Z}) = \prod_{j=1}^k \frac{1}{(\beta + n_{\underline{Z},j})^{d/2}} e^{-\frac{1}{2\sigma^2} \left(\frac{n_{\underline{Z},j} \beta}{n_{\underline{Z},j} + \beta} \|\bar{X}_{\underline{Z},j} - \alpha\|^2 + \sum_{i: Z_i=j} \|x_i - \bar{X}_{\underline{Z},j}\|^2 \right)},$$

$$f(\underline{\Theta} | \underline{Z}) = \prod_{j=1}^k (\beta + n_{\underline{Z},j})^{d/2} e^{-\frac{\beta + n_{\underline{Z},j}}{2\sigma^2} \|\theta_j - \tilde{\theta}_j\|^2},$$

Sampling from $p(\underline{\Theta} | \underline{X})$ can be achieved by first sampling \underline{Z} from distribution proportional to $w(\underline{Z})$, and then, given the sampled \underline{Z} , generating $\underline{\Theta}$ from Gaussian distribution proportional to $f(\underline{\Theta} | \underline{Z}, \underline{X})$.



Lemma

Let R and Q be distributions on discrete set \mathcal{V} and let $V, V' \sim Q$ be independent random variables. Given r.v. V and V' and $a \in \mathbb{R}$, define a Bernoulli random variable $B \sim \text{Bern}(a + \frac{r(V) - q(V)}{q(V)})$ assuming that

$a + \frac{r(V) - q(V)}{q(V)} \in [0, 1]$. Then the random variable

$\tilde{V} := BV + (1 - B)V'$ has a distribution R .

- Use an internal annealing technique by introducing an annealing parameter $t \in [0, 1]$.
- The result is smoothly-time-parameterized family $p_t(\underline{\Theta} | \underline{X})$, such that it is easy (known) how to sample from p_0 , and transition from p_t to p_{t+h} , for $h > 0$ small enough, is done using Lemma 1.



Time-parameterized posterior as mixture of Gaussians

$$p_t(\underline{\Theta}|\underline{X}) = \prod_{j=1}^k e^{\frac{-\beta}{2\sigma^2} \|\theta_j - \alpha\|^2} \prod_{i=1}^n p_{\underline{\Theta}}^{(t)}(x_i) = \sum_{\underline{Z} \in \mathcal{V}} w_t(\underline{Z}) f_t(\underline{\Theta}|\underline{Z}),$$

where

$$w_t(\underline{Z}) = \prod_{j=1}^k \frac{1}{(\beta + tn_{\underline{Z},j})^{d/2}} e^{\frac{-1}{2\sigma^2} \left(\frac{tn_{\underline{Z},j}\beta}{tn_{\underline{Z},j} + \beta} \|\bar{X}_{\underline{Z},j} - \alpha\|^2 + t \sum_{i:Z_i=j} \|x_i - \bar{X}_{\underline{Z},j}\|^2 \right)},$$

$$f_t(\underline{\Theta}|\underline{Z}) = \prod_{j=1}^k (\beta + tn_{\underline{Z},j})^{d/2} e^{-\frac{\beta + tn_{\underline{Z},j}}{2\sigma^2} \|\theta_j - \tilde{\theta}_j\|^2},$$

and

$$\tilde{\theta}_j^{(t)} = \frac{tn_{\underline{Z},j}}{tn_{\underline{Z},j} + \beta} \bar{X}_{\underline{Z},j} + \frac{\beta}{tn_{\underline{Z},j} + \beta} \alpha = \frac{tn_{\underline{Z},j} \bar{X}_{\underline{Z},j} + \beta \alpha}{tn_{\underline{Z},j} + \beta}.$$



Coin-flip probability

- Let $\underline{Z}, \underline{Z}'$ be independent labelings from q_t .
- In order to generate labeling from q_{t+h} , we need to imitate a coin-flip with probability

$$a + \frac{q_{t+h}(\underline{Z}) - q_t(\underline{Z})}{q_t(\underline{Z})} \approx a + h\partial_t \log(q_t(\underline{Z})),$$

where $a \in \mathbb{R}$ is such that $a + h\partial_t \log(q_t(\underline{Z})) \in [0, 1]$.

Moreover,

$$\partial_t \log(q_t(\underline{Z})) = \partial_t \log(w_t(\underline{Z})) - \mathbb{E} [\partial_t \log(w_t(\underline{Z}'))] =: \delta_{\underline{Z}, t}.$$



Swapping Algorithm

- The proposed algorithm begins by drawing m independent, uniformly random labelings, which serve as an initialization for parallel MC
- For i -th chain (initialized by a labeling $\underline{Z}_{0,i}$), calculate at $t = 0$ $\partial_t \log(w_t(\underline{Z}_{0,i}))$
- Set $\delta_{\underline{Z}_{0,i}}$ to be calculated $\partial_t \log(w_t(\underline{Z}_{0,i}))$ at $t = 0$, minus the average of those across all of the m chains
- Search for an $a \in \mathbb{R}$ and $h > 0$ such that $p_i(t) = a + h\delta_{\underline{Z}_{0,i}} \in [0, 1]$, for every $i = 1, \dots, m$.
- With probability $p_i(t)$ leave the i -th chain alone, and with probability $1 - p_i(t)$ replace current labeling in i -th chain with one randomly selected from the $m - 1$ other chains.



Algorithm 1 SWAPPING Algorithm

Input: $\underline{X} = \{x_1, \dots, x_n\} \subset \mathbb{R}^d, \alpha \in \mathbb{R}^d, \beta > 0, \sigma > 0, h > 0, k \in \mathbb{N}, m \in \mathbb{N},$

$\underline{Z}_0 = (\underline{Z}_{0,1}, \dots, \underline{Z}_{0,m})$ from uniform distribution on $\mathcal{V} = \{1, \dots, k\}^n, t = 0;$

While $t < 1$ **do**

1: For each $\underline{Z}_{t,i}, i = 1, \dots, m$ calculate

$$\begin{aligned} \partial_t \log(w_t(\underline{Z}_{t,i})) = & \frac{-1}{2} \sum_{j=1}^k \left(\frac{dn_{\underline{Z}_{t,i},j}}{\beta + tn_{\underline{Z}_{t,i},j}} + \frac{\beta n_{\underline{Z}_{t,i},j}}{\sigma^2 (tn_{\underline{Z}_{t,i},j} + \beta)^2} \|\bar{X}_{\underline{Z}_{t,i},j} - \alpha\|^2 \right. \\ & \left. + \frac{1}{\sigma^2} \sum_{l: \underline{Z}_{t,i,l}=j} \|x_l - \bar{X}_{\underline{Z}_{t,i},j}\|^2 \right) \end{aligned} \quad (3)$$

2: Estimate the expectation of the quantity (3) by mean $\overline{\partial_t \log(w_t(\underline{Z}))}$ of (3) over all of the labelings, and the for each $\underline{Z}_{t,i}, i = 1, \dots, m$ calculate $\delta_{\underline{Z}_{t,i}}$ as

$$\delta_{\underline{Z}_{t,i}} = \partial_t \log(w_t(\underline{Z}_{t,i})) - \overline{\partial_t \log(w_t(\underline{Z}))} \quad (4)$$



3: Let $M(t) = \max\{|\delta_{\underline{Z}_{t,i}}|, i = 1 \dots, m\}$.

3.1 Set $h(t) = \frac{h}{M(t)}$, which ensures that $-h \geq h(t)\delta_{\underline{Z}_{t,i}} \leq h \forall i = 1, \dots, m$

3.2 Set $p_i(t) = 1 - h(1 - \frac{\delta_{\underline{Z}_{t,i}}}{M(t)}) \in [0, 1]$

4: Generate random number u from $\mathcal{U}[0, 1]$;

5: **For** $i = 1, \dots, m$

If $u \leq p_i(t)$ set $\underline{Z}_{t+h,i} = \underline{Z}_{t,i}$;

Else generate random number l from $\{1, \dots, m\} \setminus i$ and set $\underline{Z}_{t+h,i} = \underline{Z}_{t,l}$;

6: **Iterate**

1. For every $i = 1, \dots, m$ generate $\underline{\Theta}_{t+h,i}$ from multivariate normal distribution

proportional to f_{t+h}

2. For every $i = 1, \dots, m$. given $\underline{\Theta}_{t+h,i}$, set the probability of l -th component of $\underline{Z}_{t+h,i}$

being equal to j , i.e the probability of observation \underline{x}_l belonging to j -th cluster w.r.t. labeling $\underline{Z}_{t+h,i}$ to be

$$\mathbb{P}(Z_{t+h,i_l} = j | \underline{\Theta}_{t+h,i}) = \frac{e^{-\frac{1}{2\sigma^2} \|\underline{x}_l - \underline{\theta}_{t+h,i_j}\|^2}}{\sum_{l'=1}^n e^{-\frac{1}{2\sigma^2} \|\underline{x}_{l'} - \underline{\theta}_{t+h,i_j}\|^2}}. \quad (5)$$

7: Set $t = t + h(t)$;

End while

Output: $\{\underline{\Theta}_{1,1}, \dots, \underline{\Theta}_{1,m}\}$.



Algorithm Analysis

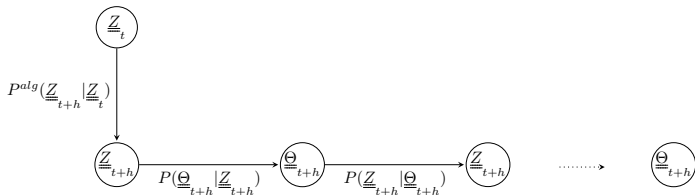


Figure: Transition from \underline{Z}_t to \underline{Z}_{t+h}

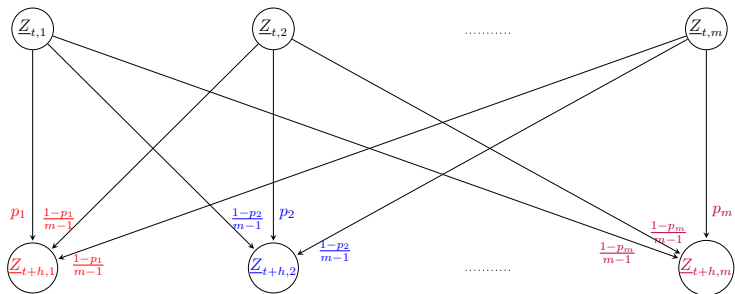


Figure: Swapping step



Transition probabilities

- Let $\underline{Z}_t = (\underline{Z}_{t,1}, \dots, \underline{Z}_{t,m})$ be an IID sample from distribution q_t of labelings at the time t
- $P_t(\underline{Z}_t) = \prod_{j=1}^m q_t(\underline{Z}_{t,j})$ is a joint distribution of \underline{Z}_t
- Transition probability from \underline{Z}_t to \underline{Z}_{t+h} used by an algorithm (1) is given by

$$P^{alg}(\underline{Z}_{t+h} | \underline{Z}_t) = \prod_{j=1}^m P^{alg}(\underline{Z}_{t+h,j} | \underline{Z}_t) =$$
$$\prod_{j=1}^m \left(p_j \delta_{\underline{Z}_{t,j}}(\underline{Z}_{t+h,j}) + (1 - p_j) \frac{1}{m-1} \sum_{i=1; i \neq j}^m \delta_{\underline{Z}_{t,i}}(\underline{Z}_{t+h,j}) \right)$$



Transition probabilities

- Conditional probability of $\underline{\Theta}_{t+h}$ given \underline{Z}_t obtained by the algorithm (1) is given by

$$P^{alg}(\underline{\Theta}_{t+h} | \underline{Z}_t) = \prod_{j=1}^m \left(p_j P(\Theta_{t+h,j} | \underline{Z}_{t,j}) + \frac{1-p_j}{m-1} \sum_{i=1, i \neq j}^m P(\Theta_{t+h,j} | \underline{Z}_{t,i}) \right)$$

- Sample $\underline{\Theta}_{t+h}$ obtained using upper conditional probability given \underline{Z}_t will be marginally from a target distribution at the time $t+h$, but will be no longer independent.



- We define another transitional probability

$$P^{mem}(\underline{Z}_{t+h} | \underline{Z}_t) = \prod_{j=1}^m (p_j \delta_{\underline{Z}_{t,j}}(\underline{Z}_{t+h,j}) + (1 - p_j) P_{t+h}(\underline{Z}_{t+h,j}))$$

- With respect to $P^{mem}(\underline{Z}_{t+h} | \underline{Z}_t)$, we define

$$\begin{aligned} P^{mem}(\underline{\Theta}_{t+h} | \underline{Z}_t) &= \prod_{j=1}^m \sum_{\underline{Z}_{t+h,j}} P(\underline{\Theta}_{t+h,j} | \underline{Z}_{t+h,j}) P^{mem}(\underline{Z}_{t+h,j} | \underline{Z}_t) \\ &= \prod_{j=1}^m \left(p_j P(\underline{\Theta}_{t+h,j} | \underline{Z}_{t,j}) + (1 - p_j) P_{t+h}(\underline{\Theta}_{t+h,j}) \right). \end{aligned}$$

If we could use $P^{mem}(\underline{Z}_{t+h} | \underline{Z}_t)$ as a transition probability from time t to $t + h$, and then sample $\underline{\Theta}_{t+h}$ from corresponding Gaussians, obtained sample would indeed be an IID sample from a target distribution at time $t + h$.



Theorem

$$\text{Let } P^{mem}(\underline{\Theta}_{t+h} | \underline{Z}_t) = \prod_{j=1}^m (p_j P(\Theta_{t+h,j} | Z_{t,j}) + (1 - p_j) P_{t+h}(\Theta_{t+h,j})),$$

$$P^{alg}(\underline{\Theta}_{t+h} | \underline{Z}_t) = \prod_{j=1}^m (p_j P(\Theta_{t+h,j} | Z_{t,j}) + \frac{1 - p_j}{m - 1} \sum_{i=1, i \neq j}^m P(\Theta_{t+h,j} | Z_{t,i}))$$

be two conditional PDF-s of $\underline{\Theta}_{t+h}$ given an IID r.v. \underline{Z}_t from q_t . Then,

$$D(P^{alg}(\underline{\Theta}_{t+h} | \underline{Z}_t) || P^{mem}(\underline{\Theta}_{t+h} | \underline{Z}_t)) \leq$$

$$\sum_{j=1}^m (1 - p_j) \frac{1}{m - 1} \sum_{i=1, i \neq j}^m D(P(\Theta_{t+h,j} | Z_{t,i}) || P_{t+h}(\Theta_{t+h,j})).$$



Theorem

Let $P^{alg}(\underline{\Theta}_{t+h})$ be a PDF of a r.v. $\underline{\Theta}_{t+h}$ obtained by applying one step of the swapping algorithm started at an IID sample \underline{Z}_t from q_t , and let $P_{t+h}(\underline{\Theta}_{t+h})$ be target distribution of $\underline{\Theta}_{t+h}$ at the time $t + h$. Then the following bound holds:

$$\begin{aligned}
 & D(P^{alg}(\underline{\Theta}_{t+h}) \parallel P_{t+h}(\underline{\Theta}_{t+h})) \\
 & \leq h C_t \mathbb{E}_{\underline{Z}_t, \tilde{\underline{Z}}_t \sim q_t} \sum_{j=1}^m \frac{1}{2} \left(d \left(\frac{\sqrt{tn_{\tilde{\underline{Z}}_t, j} + \beta}}{\sqrt{tn_{\underline{Z}_t, j} + \beta}} - \frac{1}{2} \log \frac{tn_{\tilde{\underline{Z}}_t, j} + \beta}{tn_{\underline{Z}_t, j} + \beta} - 1 \right) \right. \\
 & \quad \left. + \frac{\sigma}{\sqrt{tn_{\tilde{\underline{Z}}_t, j} + \beta}} \|\tilde{\Theta}_{j, \tilde{\underline{Z}}_t} - \Theta_{j, \underline{Z}_t}\|^2 \right),
 \end{aligned}$$

where C_t depends on the ratio of $\min_i |\delta_{\underline{Z}_t, i}|$ and $\max_i |\delta_{\underline{Z}_t, i}|$.





Further research

1. Adapt prior to imitate posterior in each step of the algorithm in order to decrease the variability among the labelings and increase the step size
2. In swapping step of the algorithm, instead of randomly sampling from the rest of the labelings, sample from some other distribution with support on subset of those remaining labelings. Do so in order to lower the dependence between MC-s.



Bibliography

-  Brinda, W. D. (2018). Adaptive Estimation with Gaussian Radial Basis Mixtures. PhD thesis, *Yale University*
-  Moon, T. K. (1996). The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6), 47-60.