Online k-MLE for mixture modelling with exponential families

Christophe Saint-Jean Frank Nielsen

MATHEMATIQUES IMAGE APPLICATIONS
UNIVERSITÉ DE LA ROCHELLE

Geometry Science Information 2015

Oct 28-30, 2015 - Ecole Polytechnique, Paris-Saclay

We are interested in building a system (a model) which evolves when new data is available:

$$x_1, x_2, \ldots, x_N, \ldots$$

- The time needed for processing a new observation must be constant w.r.t the number of observations.
- The memory required by the system is bounded.
- Denote π the unknown distribution of X

Outline of this talk

1. Online learning exponential families
2. Online learning of mixture of exponential families
   - Introduction, EM, k-MLE
   - Recursive EM, Online EM
   - Stochastic approximations of k-MLE
   - Experiments
3. Conclusions

Reminder : (Regular) Exponential Family

Firstly, π will be approximated by a member of a (regular) exponential family (EF):

$$E_F = \{f(x;\theta) = exp\{\langle s(x), \theta\rangle + k(x) - F(\theta)|\theta \in \Theta\}$$

Terminology:

- λ source parameters.
- θ natural parameters.
- η expectation parameters.
- $s(x)$ sufficient statistic.
- $k(x)$ auxiliary carrier measure.

- $F(\theta)$ the log-normalizer: differentiable, strictly convex
- Θ $= \{\theta \in \mathbb{R}^D | F(\theta) < \infty\}$ is an open convex set

Almost all common distributions are EF members but uniform, Cauchy distributions.

Reminder : Maximum Likehood Estimate (MLE)

- Maximum Likehood Estimate for general p.d.f:

$$\hat{\theta}^{(N)} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{N} f(x_i;\theta) = \underset{\theta}{\operatorname{argmin}} -\frac{1}{N}\sum_{i=1}^{N} \log f(x_i;\theta)$$

assuming a sample $\chi = \{x_1, x_2, ..., x_N\}$ of i.i.d observations.

- Maximum Likehood Estimate for an EF:

$$\hat{\theta}^{(N)} = \underset{\theta}{\operatorname{argmin}} \left(- \left\langle \frac{1}{N}\sum_{i} s(x_i), \theta \right\rangle - cst(\chi) + F(\theta) \right)$$

which is exactly solved in H, the space of expectation parameters:

$$\hat{\eta}^{(N)} = \nabla F(\hat{\theta}^{(N)}) = \frac{1}{N}\sum_{i} s(x_i) \quad \equiv \quad \hat{\theta}^{(N)} = (\nabla F)^{-1}\left(\frac{1}{N}\sum_{i} s(x_i)\right)$$

Exact Online MLE for exponential family

- A recursive formulation is easily obtained

Algorithm 1: Exact Online MLE for EF

Input: a sequence $\mathcal{S}$ of observations

Input: Functions s and $(\nabla F)^{-1}$ for some EF

Output: a sequence of MLE for all observations seen before

$\hat{\eta}^{(0)} = 0; \quad N = 1;$

for $x_N \in \mathcal{S}$ **do**

  $\hat{\eta}^{(N)} = \hat{\eta}^{(N-1)} + N^{-1}(s(x_N) - \hat{\eta}^{(N-1)});$

  yield $\hat{\eta}^{(N)}$ or **yield** $(\nabla F)^{-1}(\hat{\eta}^{(N)});$

  $N = N + 1;$

Analytical expressions of $(\nabla F)^{-1}$ exist for most EF (but not all)

Case of Multivariate normal distribution (MVN)

- Probability density function of MVN:

$$\mathcal{N}(x;\mu,\Sigma) = (2\pi)^{-\frac{d}{2}}|\Sigma|^{-\frac{1}{2}}\exp^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

- One possible decomposition:

$$\begin{aligned}\mathcal{N}(x;\theta_1,\theta_2) = \exp\{&\langle\theta_1,x\rangle + \langle\theta_2,-xx^T\rangle_F \\ &-\frac{1}{4}{}^t\theta_1\theta_2^{-1}\theta_1 - \frac{d}{2}\log(\pi) + \frac{1}{2}\log|\theta_2|\}\end{aligned}$$

$$\Longrightarrow \begin{cases} s(x) = (x,-xx^T) \\ (\nabla F)^{-1}(\eta_1,\eta_2) = \left((-\eta_1\eta_1^T-\eta_2)^{-1}\eta_1, \frac{1}{2}(-\eta_1\eta_1^T-\eta_2)^{-1}\right) \end{cases}$$

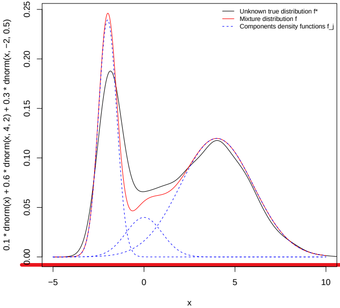
Case of the Wishart distribution

See details in the paper.

- Now, π will be approximated by a finite (parametric) mixture $f(\cdot;\theta)$ indexed by θ:

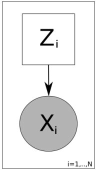
$$\pi(x) \approx f(x;\theta) = \sum_{j=1}^{K} w_j\ f_j(x;\theta_j), \quad 0 \leq w_j \leq 1, \sum_{j=1}^{K} w_j = 1$$

  where w_j are the mixing proportions, f_j are the component distributions.
- When all f_j's are EFs, it is called a Mixture of EFs (MEF).



incomplete		complete
observable	$\overset{deterministic}{\longleftarrow}$	unobservable
$\chi = \{x_1, \ldots, x_N\}$		$\chi_c = \{y_1 = (x_1, z_1), \ldots, y_N\}$

$$Z_i \sim cat_K(w)$$
$$X_i | Z_i = j \sim f_j(\cdot; \theta_j)$$



For a MEF, the joint density $p(x, z; \theta)$ is an EF:

$$\log p(x, z; \theta) = \sum_{j=1}^{K} [z = j]\{\log(w_j) + \langle \theta_j, s_j(x) \rangle + k_j(x) - F_j(\theta_j)\}$$

$$= \sum_{j=1}^{K} \left\langle \begin{pmatrix} [z = j] \\ [z = j]\ s_j(x) \end{pmatrix}, \begin{pmatrix} \log w_j - F_j(\theta_j) \\ \theta_j \end{pmatrix} \right\rangle + k(x, z)$$

The EM algorithm maximizes iteratively $\mathcal{Q}(\theta; \hat{\theta}^{(t)}, \chi)$.

Algorithm 2: EM algorithm

Input: $\hat{\theta}^{(0)}$ initial parameters of the model

Input: $\chi^{(N)} = \{x_1, \ldots, x_N\}$

Output: A (local) maximizer $\hat{\theta}^{(t^*)}$ of $\log f(\chi; \theta)$

t $\leftarrow$ 0;

repeat

  Compute $\mathcal{Q}(\theta; \hat{\theta}^{(t)}, \chi) := \mathbb{E}_{\hat{\theta}^{(t)}}[\log p(\chi_c; \theta)|\chi]$; `// E-Step`

  Choose $\hat{\theta}^{(t+1)} = \text{argmax}_\theta \, \mathcal{Q}(\theta; \hat{\theta}^{(t)}, \chi)$; `// M-Step`

  t $\leftarrow$ t +1;

until *Convergence of the complete log-likehood*;

- For a mixture, the E-Step is always explicit:

$$\hat{z}_{i,j}^{(t)} = \hat{w}_j^{(t)} f(x_i; \hat{\theta}_j^{(t)}) / \sum_{j'} \hat{w}_{j'}^{(t)} f(x_i; \hat{\theta}_{j'}^{(t)})$$

- For a MEF, the M-Step then reduces to:

$$\hat{\theta}^{(t+1)} = \operatorname*{argmax}_{\{w_j,\theta_j\}} \sum_{j=1}^{K} \left\langle \begin{pmatrix} \sum_i \hat{z}_{i,j}^{(t)} \\ \sum_i \hat{z}_{i,j}^{(t)} \; s_j(x_i) \end{pmatrix}, \begin{pmatrix} \log w_j - F_j(\theta_j) \\ \theta_j \end{pmatrix} \right\rangle$$

$$\hat{w}_j^{(t+1)} = \sum_{i=1}^{N} \hat{z}_{i,j}^{(t)} / N$$

$$\hat{\eta}_j^{(t+1)} = \nabla F(\hat{\theta}_j^{(t+1)}) = \frac{\sum_i \hat{z}_{i,j}^{(t)} s_j(x_i)}{\sum_i \hat{z}_{i,j}^{(t)}} \quad (\textit{weighted average of SS})$$

k-Maximum Likelihood Estimator (k-MLE) [2]

- The k-MLE introduces a geometric split $\chi = \bigsqcup_{j=1}^{K} \hat{\chi}_j^{(t)}$ to accelerate EM :

$$\tilde{z}_{i,j}^{(t)} = [\operatorname*{argmax}_{j'} w_{j'} f(x_i; \hat{\theta}_{j'}^{(t)}) = j]$$

- Equivalently, it amounts to maximize $\mathcal{Q}$ over partition Z [3]
- For a MEF, the M-Step of the k-MLE then reduces to:

$$\hat{\theta}^{(t+1)} = \operatorname*{argmax}_{\{w_j, \theta_j\}} \sum_{j=1}^{K} \left\langle \begin{pmatrix} |\hat{\chi}_j^{(t)}| \\ \sum_{x_i \in \hat{\chi}_j^{(t)}} s_j(x_i) \end{pmatrix}, \begin{pmatrix} \log w_j - F_j(\theta_j) \\ \theta_j \end{pmatrix} \right\rangle$$

$$\hat{w}_j^{(t+1)} = |\hat{\chi}_j^{(t)}|/N \qquad \hat{\eta}_j^{(t+1)} = \nabla F(\hat{\theta}_j^{(t+1)}) = \frac{\sum_{x_i \in \hat{\chi}_j^{(t)}} s_j(x_i)}{|\hat{\chi}_j^{(t)}|}$$

(*cluster-wise unweighted average of SS*)

- Consider now the online setting

$$x_1, x_2, \ldots, x_N, \ldots$$

- Denote $\hat{\theta}^{(N)}$ or $\hat{\eta}^{(N)}$ the parameter estimate after dealing N observations
- Denote $\hat{\theta}^{(0)}$ or $\hat{\eta}^{(0)}$ their initial values
- Remark: For a fixed-size dataset χ, one may apply multiple passes (with shuffle) on χ.
- The increase in the likelihood function is no more guaranteed after an iteration.

Two main approaches to online EM-like estimation:

- Stochastic M-Step : Recursive EM (1984) [5]

$$\hat{\theta}^{(N)} = \hat{\theta}^{(N-1)} + \{NI_c(\hat{\theta}^{(N-1)}\}^{-1}\nabla_\theta \log f(x_N; \hat{\theta}^{(N-1)})$$

where I_c is the Fisher Information matrix for the complete data:

$$I_c(\hat{\theta}^{(N-1)}) = -\mathbb{E}_{\hat{\theta}_j^{(N-1)}}\left[\frac{\log p(x, z; \theta)}{\partial\theta\partial\theta^T}\right]$$

A justification for this formula comes from the Fisher's Identity:

$$\nabla \log f(x; \theta) = \mathbb{E}_\theta[\log p(x, z; \theta)|x]$$

One can recognize a second order Stochastic Gradient Ascent which requires to update and invert I_c after each iteration.

- Stochastic E-Step : Online EM (2009) [7]

$$\hat{\mathcal{Q}}^{(N)}(\theta) = \hat{\mathcal{Q}}^{(N-1)}(\theta) + \alpha^{(N)} \left(\mathbb{E}_{\hat{\theta}^{(N-1)}}[\log p(x_N, z_N; \theta) | x_N] - \hat{\mathcal{Q}}^{(N-1)}(\theta) \right)$$

In case of a MEF, the algorithm works only with the cond. expectation of the sufficient statistics for complete data.

$$\hat{z}_{N,j} = \mathbb{E}_{\theta^{(N-1)}}[z_{N,j} | x_N]$$

$$\begin{pmatrix} \hat{S}_{w_j}^{(N)} \\ \hat{S}_{\theta_j}^{(N)} \end{pmatrix} = \begin{pmatrix} \hat{S}_{w_j}^{(N-1)} \\ \hat{S}_{\theta_j}^{(N-1)} \end{pmatrix} + \alpha^{(N)} \left(\begin{pmatrix} \hat{z}_{N,j} \\ \hat{z}_{N,j} \ s_j(x_N) \end{pmatrix} - \begin{pmatrix} \hat{S}_{w_j}^{(N-1)} \\ \hat{S}_{\theta_j}^{(N-1)} \end{pmatrix} \right)$$

The M-Step is unchanged:

$$\begin{aligned} \hat{w}_j^{(N)} &= \hat{\eta}_{w_j}^{(N)} = \hat{S}_{w_j}^{(N)} \\ \hat{\theta}_j^{(N)} &= (\nabla F_j)^{-1} (\hat{\eta}_{\theta_j}^{(N)} = \hat{S}_{\theta_j}^{(N)} / \hat{S}_{w_j}^{(N)}) \end{aligned}$$

Some properties:

- Initial values $\hat{S}^{(0)}$ may be used for introducing a "prior":

$$\hat{S}^{(0)}_{w_j} = w_j, \hat{S}^{(0)}_{\theta_j} = w_j \eta_j^{(0)}$$

- Parameters constraints are automatically respected
- No matrix to invert !
- Policy for $\alpha^{(N)}$ has to be chosen (see [7])
- Consistent, asymptotically equivalent to the recursive EM !!

Stochastic approximations of k-MLE(1)

In order to keep previous advantages of online EM for an online k-MLE, our only choice concerns the way to affect x_N to a cluster.

Strategy 1 Maximize the likelihood of the complete data (x_N, z_N)

$$\tilde{z}_{N,j} = [\operatorname*{argmax}_{j'} \hat{w}_{j'}^{(N-1)} f(x_N; \hat{\theta}_{j'}^{(N-1)}) = j]$$

Equivalent to Online CEM and similar to Mac-Queen iterative k-Means.

Stochastic approximations of k-MLE(2)

Strategy 2 Maximize the likelihood of the complete data (x_N, z_N) **after** the M-Step:

$$\tilde{z}_{N,j} = [\operatorname*{argmax}_{j'} \hat{w}_{j'}^{(N)} f(x_N; \hat{\theta}_{j'}^{(N)}) = j]$$

- Similar to Hartigan's method for k-means.
- Additional cost: pre-compute all possible M-Steps for the Stochastic E-Step.

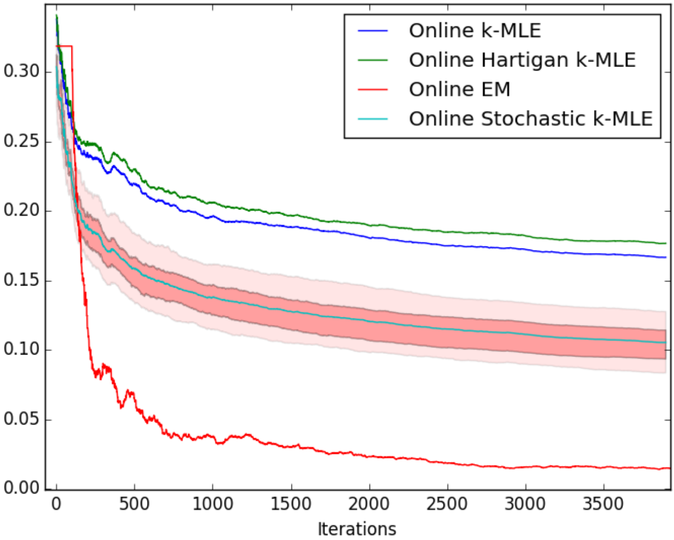
Strategy 3 Draw $\tilde{z}_{N,j}$ from the categorical distribution

$$\tilde{z}_N \text{ sampled from } Cat_K(\{p_j = \log(\hat{w}_j^{(N-1)} f_j(x_N; \hat{\theta}_j^{(N-1)}))\}_j)$$

- Similar to sampling in Stochastic EM [3]
- The motivation is to try to break the inconsistency of k-MLE.

For strategies 1 and 3, the M-Step reduces the update of the parameters for a single component.

- True distribution $\pi = 0.5\mathcal{N}(0,1) + 0.5\mathcal{N}(\mu_2, \sigma_2^2)$
- Different values for μ_2, σ_2 for more or less overlap between components.
- A small subset of observations has be taken for initialization (k-MLE++ / k-MLE).
- Video illustrating the inconsistency of online k-MLE.


Online k-MLE
Online Hartigan k-MLE
Online EM
Online Stochastic k-MLE
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
500
1000
1500
2000
2500
3000
3500
Iterations

- On consistency:
  - EM, Online EM are consistent
  - k-MLE, online k-MLE (Strategies 1,2) are inconsistent (due to the Bayes error in maximizing the classification likelihood)
  - Online stochastic k-MLE (Strategy 3) : consistency ?
- So, when components overlap, online EM $>$ k-MLE $>$ online k-MLE for parameter learning.
- Need to study how the dimension influences the inconstancy/convergence rate for online k-MLE.
- Convergence rate is lower for online methods (sub-linear convergence of the SGD)
- Time for an update vs sample size:

online k-MLE (1,3) $<$ online EM $<$ online k-MLE (2) $<<$ k-MLE *online EM appears to be the best compromise !!*

Dempster, A.P., Laird, N.M. and Rubin, D.B.:
Maximum likelihood from incomplete data via the EM algorithm.
Journal of the Royal Statistical Society. Series B (Methodological), pp. 1–38, 1977.

Nielsen, F.:
On learning statistical mixtures maximizing the complete likelihood
Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt 2014), AIP Conference Proceedings Publishing, 1641, pp. 238-245, 214.

Celeux, G. and Govaert, G.:
A classification EM algorithm for clustering and two stochastic versions.
Computational Statistics and Data Analysis, 14(3), pp. 315-332, 1992.

Samé, A., Ambroise, C., Govaert, G.:
An online classification EM algorithm based on the mixture model
Statistics and Computing, 17(3), pp. 209–218, 2007.

Titterington, D. M. :
Recursive Parameter Estimation Using Incomplete Data.
Journal of the Royal Statistical Society. Series B (Methodological), Volume 46, Number 2, pp. 257–267, 1984.

Amari, S. I. :
Natural gradient works efficiently in learning.
Neural Computation, Volume 10, Number 2, pp. 251?276, 1998.

Cappé, O., Moulines, E.:
On-line expectation-maximization algorithm for latent data models.
Journal of the Royal Statistical Society. Series B (Methodological), 71(3):593-613, 2009.

Neal, R. M., Hinton, G. E.:
A view of the EM algorithm that justifies incremental, sparse, and other variants.
In Jordan, M. I., editor, Learning in graphical models, pages 355-368. MIT Press, Cambridge, 1999.

Bottou, Léon :
Online Algorithms and Stochastic Approximations.
Online Learning and Neural Networks, Saad, David Eds.,Cambridge University Press, 1998.