On learning statistical mixtures maximizing the complete likelihood

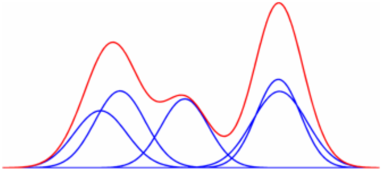
The k-MLE methodology using geometric hard clustering

Frank NIELSEN

École Polytechnique
Sony Computer Science Laboratories

MaxEnt 2014
September 21-26 2014
Amboise, France

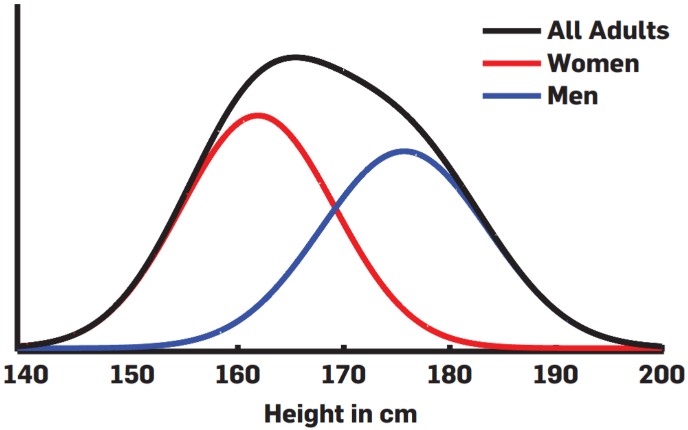
Finite mixtures: Semi-parametric statistical models



- Mixture $M \sim \mathrm{MM}(W, \Lambda)$ with density $\boxed{m(x) = \sum_{i=1}^{k} w_i p(x|\lambda_i)}$
  not sum of RVs!. $\Lambda = \{\lambda_i\}_i$, $W = \{w_i\}_i$
- Multimodal, universally modeling smooth densities
- Gaussian MMs with support $\mathcal{X} = \mathbb{R}$, Gamma MMs with support $\mathcal{X} = \mathbb{R}^+$ (modeling distances [34])
- Pioneered by Karl Pearson [29] (1894). precursors: Francis Galton [13] (1869), Adolphe Quetelet [31] (1846), etc.
- Capture **sub-populations** within an overall population ($k = 2$, crab data [29] in Pearson)

Example of $k = 2$-component mixture [17]

Sub-populations ($k = 2$) within an overall population...



Sub-species in species, etc.

Truncated distributions (what is the support! black swans ?!)

Sampling from mixtures: Doubly stochastic process

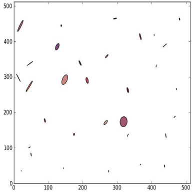
To sample a variate x from a MM:

- Choose a component l according to the weight distribution $w_1, ..., w_k$ (multinomial),
- Draw a variate x according to $p(x|\lambda_l)$.

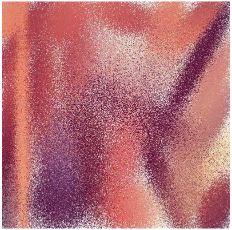
Statistical mixtures: Generative data models

Image = 5D xyRGB point set
GMM = *feature descriptor* for information retrieval (IR)
Increase dimension d using color image $s \times s$ *patches*: $d = 2 + 3s^2$


Source



GMM


Sample (stat img)

Low-frequency information encoded into compact statistical model.

Mixtures: ϵ-statistically learnable and ϵ-estimates

Problem statement: Given n IID d-dimensional observations $x_1, ..., x_n \sim \mathrm{MM}(\Lambda, W)$, estimate $\mathrm{MM}(\hat{\Lambda}, \hat{W})$:

- **Theoretical Computer Science** (TCS) approach: ϵ-*closely parameter recovery* (π: permutation)
  - $|w_i - \hat{w}_{\pi(i)}| \leq \epsilon$
  - $\mathrm{KL}(p(x|\lambda_i) : p(x|\hat{\lambda}_{\pi(i)})) \leq \epsilon$ (or other divergences like TV, etc.)

  Consider ϵ-*learnable* MMs:

  - $\min_i w_i \geq \epsilon$
  - $\mathrm{KL}(p(x|\lambda_i) : p(x|\lambda_i)) \geq \epsilon, \ \forall i \neq j$ (or other divergence)
- **Statistical approach**:
  Define the **best** model/MM as the one maximizing the *likelihood function* $l(\Lambda, W) = \prod_i m(x_i|\Lambda, W)$.

Mixture inference: Incomplete versus complete likelihood

- Sub-populations within an overall population: observed data x_i *does not* include the subpopulation label l_i
- $k = 2$: Classification and Bayes error (upper bounded by Chernoff information [24])
- Inference: Assume IID, maximize (log)-likelihood:
  - **Complete** using **indicator variables** $z_{i,j}$ (for l_i: $z_{i,l_i} = 1$):

$$l_c = \log \prod_{i=1}^{n} \prod_{j=1}^{k} (w_j p(x_i|\theta_j))^{z_{i,j}} = \sum_i \sum_j z_{i,j} \log(w_j p(x_i|\theta_j))$$

  - **Incomplete** (hidden/latent variables) and *log-sum intractability*:

$$l_i = \log \prod_i m(x|W,\Lambda) = \sum_i \log \left(\sum_j w_j p(x_i|\theta_j) \right)$$

Mixture learnability and inference algorithms

- **Which criterion to maximize? incomplete or complete likelihood? What kind of evaluation criteria?**
- From Expectation-Maximization [8] (1977) to *TCS methods*: Polynomial learnability of mixtures [22, 15] (2014), mixtures and core-sets [10] for massive data sets, etc.

Some technicalities:

- Many local maxima of *likelihood functions* l_i and l_c (EM converges locally and *needs* a stopping criterion)
- Multimodal density ($\#modes > k$ [9], *ghost modes* even for isotropic GMMs)
- Identifiability (permutation of labels, parameter distinctness)
- Irregularity: Fisher information may be zero [6], convergence speed of EM
- etc.

Learning MMs: A geometric hard clustering viewpoint

$$
\begin{aligned}
\max_{W,\Lambda} l_c(W,\Lambda) &= \max_{\Lambda} \sum_{i=1}^{n} \max_{j=1}^{k} \log(w_j p(x_i|\theta_j)) \\
&\equiv \min_{W,\Lambda} \sum_{i} \min_{j} (-\log p(x_i|\theta_j) - \log w_j) \\
&= \boxed{\min_{W,\Lambda} \sum_{i=1}^{n} \min_{j=1}^{k} D_j(x_i)},
\end{aligned}
$$

where $c_j = (w_j, \theta_j)$ (**cluster prototype**) and $D_j(x_i) = -\log p(x_i|\theta_j) - \log w_j$ are **potential distance-like functions**.

- Maximizing the complete likelihood amounts to a *geometric hard clustering* [37, 11] for fixed w_j's (distance $D_j(\cdot)$ depends on cluster prototypes c_j): $\min_\Lambda \sum_i \min_j D_j(x_i)$.
- Related to classification EM [5] (CEM), hard/truncated EM
- Solution of $\arg\max l_c$ to initialize l_i (optimized by EM)

The k-MLE method: k-means type clustering algorithms

k-MLE:

1. Initialize weight W (in open probability simplex Δ_k)
2. Solve $\min_\Lambda \sum_i \min_j D_j(x_i)$ (**center-based clustering**, W fixed)
3. Solve $\min_W \sum_i \min_j D_j(x_i)$ (Λ fixed)
4. Test for convergence and go to step 2) otherwise.

$\Rightarrow$ group coordinate ascent (ML)/descent (distance) optimization.

k-MLE: Center-based clustering, W fixed

Solve $\boxed{\min_{\Lambda} \sum_i \min_j D_j(x_i)}$

k-means type convergence proof for assignment/relocation:

- **Data assignment**:
  $\forall i, l_i = \arg\max_j w_j p(x|\lambda_j) = \arg\min_j D_j(x_i)$, $\mathcal{C}_j = \{x_i | l_i = j\}$
- **Center relocation**: $\forall j, \lambda_j = \mathrm{MLE}(\mathcal{C}_j)$

Farthest Maximum Likelihood (FML) Voronoi diagram:

$$\begin{aligned}
\mathrm{Vor}_{\mathrm{FML}}(c_i) &= \{x \in \mathcal{X} : w_i p(x|\lambda_i) \geq w_j p(x|\lambda_j),\ \forall i \neq j\} \\
\mathrm{Vor}(c_i) &= \{x \in \mathcal{X} : D_i(x) \leq D_j(x),\ \forall i \neq j\}
\end{aligned}$$

FML Voronoi $\equiv$ **additively weighted Voronoi** with:

$$D_l(x) = -\log p(x|\lambda_l) - \log w_l$$

k-MLE: Example for mixtures of exponential families

Exponential family:

Component density $\boxed{p(x|\theta) = \exp(t(x)^\top \theta - F(\theta) + k(x))}$ is *log-concave* with:

- $t(x)$: sufficient statistic in $\mathbb{R}^D$, D: family order.
- $k(x)$: auxiliary carrier term (wrt Lebesgue/counting measure)
- $F(\theta)$: log-normalized, cumulant function, log-partition.

$D_j(x)$ **is convex: Clustering** k**-means wrt convex "distances".**

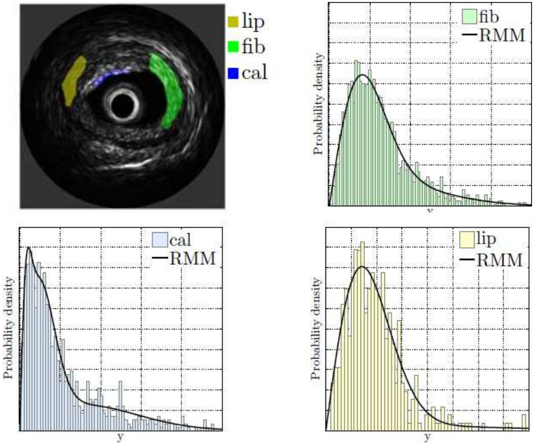
Farthest ML Voronoi $\equiv$ *additively-weighted Bregman Voronoi* [4]:

$$
\begin{aligned}
-\log p(x;\theta) - \log w &= F(\theta) - t(x)^\top \theta - k(x) - \log w \\
&= B_{F^*}(t(x) : \eta) + F^*(t(x)) + k(x) - \log w
\end{aligned}
$$

$F^*(\eta) = \max_\theta(\theta^\top \eta - F(\theta))$: Legendre-Fenchel convex conjugate

Exponential families: Rayleigh distributions [36, 25]

Application: IntraVascular UltraSound (IVUS) imaging:



Rayleigh distribution:

$p(x; \lambda) = \frac{x}{\lambda^2} e^{-\frac{x^2}{2\lambda^2}}$

$x \in \mathbb{R}^+ = \mathbb{X}$

$d = 1$ (univariate)

$D = 1$ (order 1)

$\theta = -\frac{1}{2\lambda^2}$

$\Theta = (-\infty, 0)$

$F(\theta) = -\log(-2\theta)$

$t(x) = x^2$

$k(x) = \log x$

(Weibull for $k = 2$)

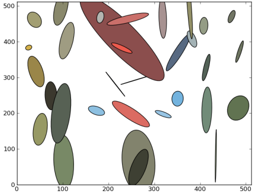
Coronary plaques: fibrotic tissues, calcified tissues, lipidic tissues

Rayleigh Mixture Models (**RMMs**):

for *segmentation* and *classification* tasks

Exponential families: Multivariate Gaussians [14, 25]

Gaussian Mixture Models (GMMs).
(Color image interpreted as a *5D xyRGB* point set)



Gaussian distribution $p(x;\mu,\Sigma)$:
$$\frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|\Sigma|}} e^{-\frac{1}{2} D_{\Sigma^{-1}}(x-\mu,x-\mu)}$$
Squared Mahalanobis distance:
$D_Q(x,y) = (x-y)^T Q(x-y)$
$x \in \mathbb{R}^d = \mathbb{X}$
d (multivariate)
$D = \frac{d(d+3)}{2}$ (order)
$\theta = (\Sigma^{-1}\mu, \frac{1}{2}\Sigma^{-1}) = (\theta_v, \theta_M)$
$\Theta = \mathbb{R} \times S^d_{++}$
$F(\theta) = \frac{1}{4}\theta_v^T \theta_M^{-1}\theta_v - \frac{1}{2}\log|\theta_M| + \frac{d}{2}\log\pi$
$t(x) = (x, -xx^T)$
$k(x) = 0$

The k-MLE method for exponential families

k-MLEEF:

1. Initialize weight W (in open probability simplex Δ_k)
2. Solve $\min_\Lambda \sum_i \min_j (B_{F^*}(t(x) : \eta_j) - \log w_j)$
3. Solve $\min_W \sum_i \min_j D_j(x_i)$
4. Test for convergence and go to step 2) otherwise.

Assignment condition in Step 2: additively-weighted Bregman Voronoi diagram.

k-MLE: Solving for weights given component parameters

Solve $\boxed{\min_W \sum_i \min_j D_j(x_i)}$

Amounts to $\arg\min_W -n_j \log w_j = \arg\min_W -\frac{n_j}{n} \log w_j$ where $n_j = \#\{x_i \in \mathrm{Vor}(c_j)\} = |\mathcal{C}_j|$.

$$\boxed{\min_{W \in \Delta_k} H^\times(N : W)}$$

where $N = (\frac{n_1}{n}, ..., \frac{n_k}{n})$ is *cluster point proportion vector* $\in \Delta_k$.
Cross-entropy $H^\times$ is minimized when $H^\times(N : W) = H(N)$ that is $W = N$.
Kullback-Leibler divergence:
$\mathrm{KL}(N : W) = H^\times(N : W) - H(N) = 0$ when $W = N$.

MLE for exponential families

Given a ML farthest Voronoi partition, computes MLEs θ_j's:

$$\hat{\theta}_j = \arg\max_{\theta\in\Theta} \prod_{x_i\in\mathrm{Vor}(c_j)} p_F(x_i;\theta)$$

is *unique* (***) maximum since $\nabla^2 F(\theta) \succ 0$:

$$\textbf{Moment equation} : \nabla F(\hat{\theta}_j) = \eta(\hat{\theta}_j) = \frac{1}{n_j} \sum_{x_i\in\mathrm{Vor}(c_j)} t(x_i) = \bar{t} = \hat{\eta}$$

MLE is *consistent*, *efficient* with *asymptotic normal distribution*:

$$\hat{\theta}_j \sim N\left(\theta_j, \frac{1}{n_j} I^{-1}(\theta_j)\right)$$

Fisher information matrix

$$I(\theta_j) = \mathrm{var}[t(X)] = \nabla^2 F(\theta_j) = (\nabla^2 F^*)^{-1}(\eta_j)$$

MLE may be biased (eg, normal distributions).

Existence of MLEs for exponential families (***)

For minimal and full EFs, MLE guaranteed to exist [3, 21] provided that matrix:

$$T = \begin{bmatrix} 1 & t_1(x_1) & \ldots & t_D(x_1) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_1(x_n) & \ldots & t_D(x_n) \end{bmatrix} \tag{1}$$

of dimension $n \times (D+1)$ has *rank* $D+1$ [3].
For example, problems for MLEs of MVNs with $n < d$ observations (undefined with likelihood ∞).

Condition: $\bar{t} = \frac{1}{n_j} \sum_{x_i \in \mathrm{Vor}(c_j)} t(x_i) \in \mathrm{int}(C)$, where C is *closed convex support*.

MLE of EFs: Observed point in IG/Bregman 1-mean

$$\hat{\theta} = \arg\max_\theta \prod_{i=1}^n p_F(x_i;\theta) = \arg\max_\theta \sum_{i=1}^n \log p_F(x_i;\theta)$$

$$\mathrm{argmax}_\theta \quad \sum_{i=1}^n -B_{F^*}(t(x_i):\eta) + \underbrace{F^*(t(x_i)) + k(x_i)}_{\text{constant}}$$

$$\equiv \mathrm{argmin}_\theta \quad \sum_{i=1}^n B_{F^*}(t(x_i):\boxed{\eta})$$

Right-sided *Bregman centroid* = *center of mass*: $\boxed{\hat{\eta} = \frac{1}{n}\sum_{i=1}^n t(x_i)}$.

$$\begin{aligned} \bar{l} &= \frac{1}{n}\sum_{i=1}^n(-B_{F^*}(t(x_i):\hat{\eta}) + F^*(t(x_i)) + k(x_i)) \\ &= \langle\hat{\eta},\hat{\theta}\rangle - F(\hat{\theta}) + \bar{k} = \boxed{F^*(\hat{\eta}) + \bar{k}} \end{aligned}$$

The k-MLE method: Heuristics based on k-means

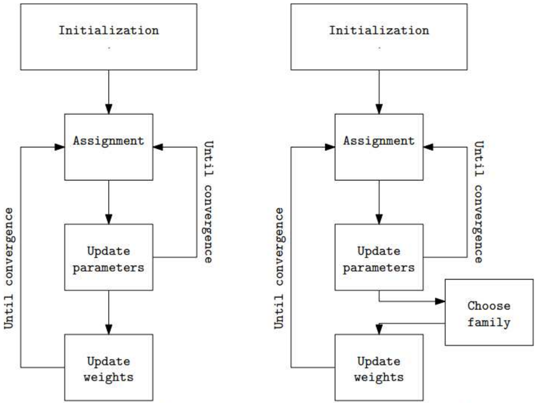
k-means is NP-hard (non-convex optimization) when $d > 1$ and $k > 1$ and solved exactly using dynamic programming [26] in $O(n^2 k)$ and $O(n)$ memory when $d = 1$.

Heuristics:

- Kanungo et al. [18] swap: yields a $(9 + \epsilon)$-approximation
- Global seeds: random seed (Forgy [12]), k-means++ [2], global k-means initialization [38],
- Local refinements: Lloyd batched update [19], MacQueen iterative update [20], Hartigan single-point swap [16], etc.
- etc.

Generalized k-MLE

Weibull or generalized Gaussians are *parametric families of exponential families* [35]: $F(\gamma)$.
Fixing some parameters yields *nested families* of (sub)-exponential families [34]: obtain one free parameter with convex conjugate F^* approximated by line search (Gamma distributions/generalized Gaussians).



(a) Original k-MLE (b) Extended k-MLE

Generalized k-MLE

k-GMLE:

1. Initialize weight $W \in \Delta_k$ and family type $(F_1, ..., F_k)$ for each cluster
2. Solve $\min_\Lambda \sum_i \min_j D_j(x_i)$ (center-based clustering for W fixed) with potential functions:
   $D_j(x_i) = -\log p_{F_j}(x_i|\theta_j) - \log w_j$
3. Solve family types maximizing the MLE in each cluster $\mathcal{C}_j$ by choosing the parametric family of distributions $F_j = F(\gamma_j)$ that yields the best likelihood:
   $\min_{F_1=F(\gamma_1),...,F_k=F(\gamma_k)\in F(\gamma)} \sum_i \min_j D_{w_j,\theta_j,F_j}(x_i)$.
4. Update W as the cluster point proportion
5. Test for convergence and go to step 2) otherwise.

$$\boxed{D_{w_j,\theta_j,F_j}(x) = -\log p_{F_j}(x;\theta_j) - \log w_j}$$

Generalized k-MLE: Convergence

- Lloyd's batched generalized k-MLE maximizes **monotonically** the complete likelihood
- Hartigan single-point relocation generalized k-MLE maximizes **monotonically** the complete likelihood [32], improves over Lloyd local maxima, and avoids the problem of the existence of MLE inside clusters by ensuring $n_j \geq D$ in general position (T rank $D+1$).
- **Model selection**: Learn k automatically using DP k-means [32] (Dirichlet Process)

k-MLE [23] versus EM for Exponential Families [1]

	k-MLE/Hard EM [23] (2012-) = Bregman hard clustering	Soft EM [1] (1977) = Bregman soft clustering
Memory	lighter $O(n)$	heavier $O(nk)$
Assignment	NNs with VP-trees [27], BB-trees [30]	all k-NNs
Conv.	always finitely	∞, need stopping criterion

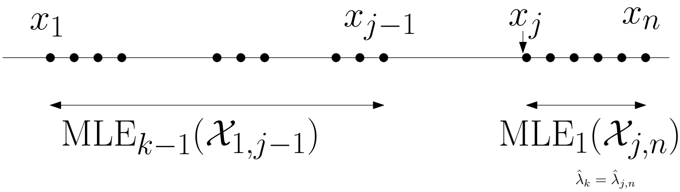
Many (probabilistically) guaranteed initialization for k-MLE [18, 2, 28]

k-MLE: Solving for $D = 1$ exponential families

- Rayleigh, Poisson or (nested) univariate normal with constant σ are order 1 EFs ($D = 1$).
- Clustering problem: Dual 1D Bregman clustering [1] on 1D scalars $y_i = t(x_i)$.
- FML Voronoi diagrams have **connected cells**: Optimal clustering yields *interval clustering*.
- 1D k-means (with additive weights) can be solved exactly using **dynamic programming** in $O(n^2 k)$ time [26]. Then update the weights W (cluster point proportion) and reiterate...

Dynamic programming for $D = 1$-order mixtures [26]

Consider W fixed. k-MLE cost: $\sum_{j=1}^{k} l(\mathcal{C}_j)$ where $\mathcal{C}_j$ are clusters.



Dynamic programming optimality equation:

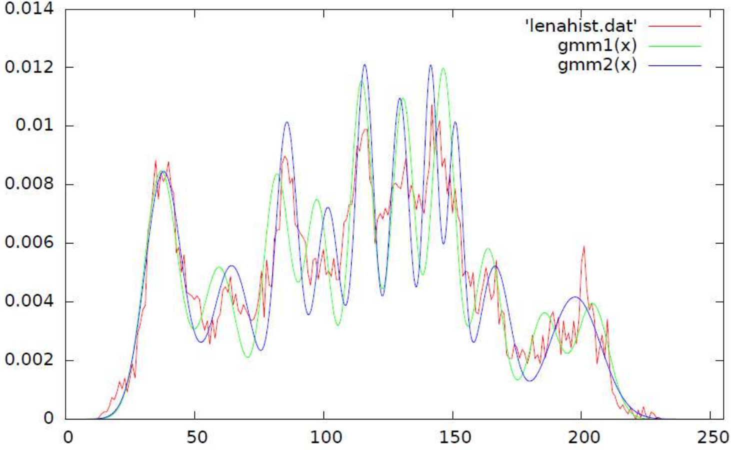
$$\mathrm{MLE}_k(x_1, ..., x_n) = \max_{j=2}^{n} (\mathrm{MLE}_{k-1}(\mathcal{X}_{1,j-1}) + \mathrm{MLE}_1(\mathcal{X}_{j,n}))$$

$\mathcal{X}_{l,r} : \{x_l, x_{l+1}, ..., x_{r-1}, x_r\}$.

- Build dynamic programming *table* from $l = 1$ to $l = k$ columns, $m = 1$ to $m = n$ rows.
- Retrieve $\mathcal{C}_j$ from DP table by *backtracking* on the $\arg\max_j$.
- For $D = 1$ EFs, $O(n^2 k)$ time [26].

Experiments with: 1D Gaussian Mixture Models (GMMs)

gmm_1 score $= -3.075$ (Euclidean k-means, σ fixed)
gmm_2 score $= -3.038$ (Bregman k-means, σ fitted, better)



Summary: k-MLE methodology for learning mixtures

Learn MMs from **sequences of geometric hard clustering** [11].

- Hard k-MLE ($\equiv$ dual Bregman hard clustering for EFs) versus soft EM ($\equiv$ soft Bregman clustering [1] for EFs):
  - k-MLE maximizes the **complete likelihood** l_c.
  - EM maximizes locally the **incomplete likelihood** l_i.
- The component parameters η geometric clustering (Step 2.) can be implemented using **any** Bregman k-means heuristic on conjugate F^*
- Consider *generalized k-MLE* when F^* not available in closed form: nested exponential families (eg., Gamma)
- Initialization can be performed using k-means initialization: k-MLE++, etc.
- Exact solution with dynamic programming for order 1 EFs (with prescribed weight proportion W).
- Avoid **unbounded likelihood** (eg., ∞ for location-scale member with $\sigma \rightarrow 0$: Dirac) using Hartigan's heuristic [32]

Discussion: Learning statistical models FAST!

- (EF) Mixture Models allow one to approximate universally smooth densities
- A single (multimodal) EF can approximate any smooth density too [7] but F not in closed-form
- Which criterion to maximize is best/realistic: incomplete or complete, or parameter distortions? **Leverage** many recent results on k-means clustering to learning mixture models.
- Alternative approach: Simplifying mixtures from kernel density estimators (KDEs) is one **fine-to-coarse** solution [33]
- *Open problem*: How to constrain the MMs to have a prescribed number of modes/antimodes?

Thank you.

Experiments and performance evaluations on generalized k-MLE:

- k-GMLE for generalized Gaussians [35]
- k-GMLE for Gamma distributions [34]
- k-GMLE for singly-parametric distributions [26]

(compared with Expectation-Maximization [8])

Frank Nielsen (5793b870).

Bibliography I

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh.
Clustering with Bregman divergences.
Journal of Machine Learning Research, 6:1705–1749, 2005.

Anup Bhattacharya, Ragesh Jaiswal, and Nir Ailon.
A tight lower bound instance for k-means++ in constant dimension.
In T.V. Gopal, Manindra Agrawal, Angsheng Li, and S.Barry Cooper, editors, *Theory and Applications of Models of Computation*, volume 8402 of *Lecture Notes in Computer Science*, pages 7–22. Springer International Publishing, 2014.

Krzysztof Bogdan and Małgorzata Bogdan.
On existence of maximum likelihood estimators in exponential families.
Statistics, 34(2):137–149, 2000.

Jean-Daniel Boissonnat, Frank Nielsen, and Richard Nock.
Bregman Voronoi diagrams.
Discrete Comput. Geom., 44(2):281–307, September 2010.

Gilles Celeux and Gérard Govaert.
A classification EM algorithm for clustering and two stochastic versions.
Comput. Stat. Data Anal., 14(3):315–332, October 1992.

Jiahua Chen.
Optimal rate of convergence for finite mixture models.
The Annals of Statistics, pages 221–233, 1995.

Loren Cobb, Peter Koppstein, and Neng Hsin Chen.
Estimation and moment recursion relations for multimodal distributions of the exponential family.
Journal of the American Statistical Association, 78(381):124–130, 1983.

Bibliography II

Arthur Pentland Dempster, Nan M. Laird, and Donald B. Rubin.
Maximum likelihood from incomplete data via the EM algorithm.
Journal of the Royal Statistical Society. Series B (Methodological), 39(1):1–38, 1977.

Herbert Edelsbrunner, Brittany Terese Fasy, and Günter Rote.
Add isotropic Gaussian kernels at own risk: more and more resilient modes in higher dimensions.
In *Proceedings of the 2012 symposuim on Computational Geometry*, SoCG '12, pages 91–100, New York, NY, USA, 2012. ACM.

Dan Feldman, Matthew Faulkner, and Andreas Krause.
Scalable training of mixture models via coresets.
In J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2142–2150. Curran Associates, Inc., 2011.

Dan Feldman, Morteza Monemizadeh, and Christian Sohler.
A PTAS for k-means clustering based on weak coresets.
In *Proceedings of the twenty-third annual symposium on Computational geometry*, pages 11–18. ACM, 2007.

Edward W. Forgy.
Cluster analysis of multivariate data: efficiency vs interpretability of classifications.
Biometrics, 1965.

Francis Galton.
Hereditary genius.
Macmillan and Company, 1869.

Vincent Garcia and Frank Nielsen.
Simplification and hierarchical representations of mixtures of exponential families.
Signal Processing (Elsevier), 90(12):3197–3212, 2010.

Bibliography III

Moritz Hardt and Eric Price.
Sharp bounds for learning a mixture of two gaussians.
CoRR, abs/1404.4997, 2014.

John A. Hartigan.
Clustering Algorithms.
John Wiley & Sons, Inc., New York, NY, USA, 99th edition, 1975.

Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant.
Disentangling gaussians.
Communications of the ACM, 55(2):113–120, 2012.

Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu.
A local search approximation algorithm for k-means clustering.
Computational Geometry: Theory & Applications, 28(2-3):89–112, 2004.

Stuart P. Lloyd.
Least squares quantization in PCM.
Technical report, Bell Laboratories, 1957.

James B. MacQueen.
Some methods of classification and analysis of multivariate observations.
In L. M. Le Cam and J. Neyman, editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, CA, USA, 1967.

Weiwen Miao and Marjorie Hahn.
Existence of maximum likelihood estimates for multi-dimensional exponential families.
Scandinavian Journal of Statistics, 24(3):371–386, 1997.

Bibliography IV

Ankur Moitra and Gregory Valiant.
Settling the polynomial learnability of mixtures of Gaussians.
In *51st IEEE Annual Symposium on Foundations of Computer Science*, pages 93–102. IEEE, 2010.

Frank Nielsen.
k-MLE: A fast algorithm for learning statistical mixture models.
CoRR, abs/1203.5181, 2012.

Frank Nielsen.
Generalized Bhattacharyya and Chernoff upper bounds on Bayes error using quasi-arithmetic means.
Pattern Recognition Letters, 42(0):25 – 34, 2014.

Frank Nielsen and Vincent Garcia.
Statistical exponential families: A digest with flash cards, 2009.
arXiv.org:0911.4863.

Frank Nielsen and Richard Nock.
Optimal interval clustering: Application to bregman clustering and statistical mixture learning.
Signal Processing Letters, IEEE, 21(10):1289–1292, Oct 2014.

Frank Nielsen, Paolo Piro, and Michel Barlaud.
Bregman vantage point trees for efficient nearest neighbor queries.
In *Proceedings of the 2009 IEEE International Conference on Multimedia and Expo (ICME)*, pages 878–881, 2009.

Rafail Ostrovsky, Yuval Rabani, Leonard J. Schulman, and Chaitanya Swamy.
The effectiveness of Lloyd-type methods for the k-means problem.
In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 165–176, Washington, DC, USA, 2006. IEEE Computer Society.

Bibliography V

Karl Pearson.
Contributions to the mathematical theory of evolution.
Philosophical Transactions of the Royal Society A, 185:71–110, 1894.

Paolo Piro, Frank Nielsen, and Michel Barlaud.
Tailored Bregman ball trees for effective nearest neighbors.
In *European Workshop on Computational Geometry (EuroCG)*, LORIA, Nancy, France, March 2009. IEEE.

Adolphe Quetelet.
Lettres sur la théorie des probabilités, appliquée aux sciences morales et politiques.
Hayez, 1846.

Christophe Saint-Jean and Frank Nielsen.
Hartigan's method for k-MLE: Mixture modeling with Wishart distributions and its application to motion retrieval.
In *Geometric Theory of Information*, pages 301–330. Springer International Publishing, 2014.

Olivier Schwander and Frank Nielsen.
Model centroids for the simplification of kernel density estimators.
In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 737–740, 2012.

Olivier Schwander and Frank Nielsen.
Fast learning of Gamma mixture models with k-mle.
In *Similarity-Based Pattern Recognition (SIMBAD)*, pages 235–249, 2013.

Olivier Schwander, Aurélien J. Schutz, Frank Nielsen, and Yannick Berthoumieu.
k-MLE for mixtures of generalized Gaussians.
In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR)*, pages 2825–2828, 2012.

Bibliography VI

Jose Seabra, Francesco Ciompi, Oriol Pujol, Josepa Mauri, Petia Radeva, and Joao Sanchez.
Rayleigh mixture model for plaque characterization in intravascular ultrasound.
IEEE Transaction on Biomedical Engineering, 58(5):1314–1324, 2011.

Marc Teboulle.
A unified continuous optimization framework for center-based clustering methods.
Journal of Machine Learning Research, 8:65–102, 2007.

Juanying Xie, Shuai Jiang, Weixin Xie, and Xinbo Gao.
An efficient global k-means clustering algorithm.
Journal of computers, 6(2), 2011.