Generalized EM Algorithms for Minimum Divergence Estimation

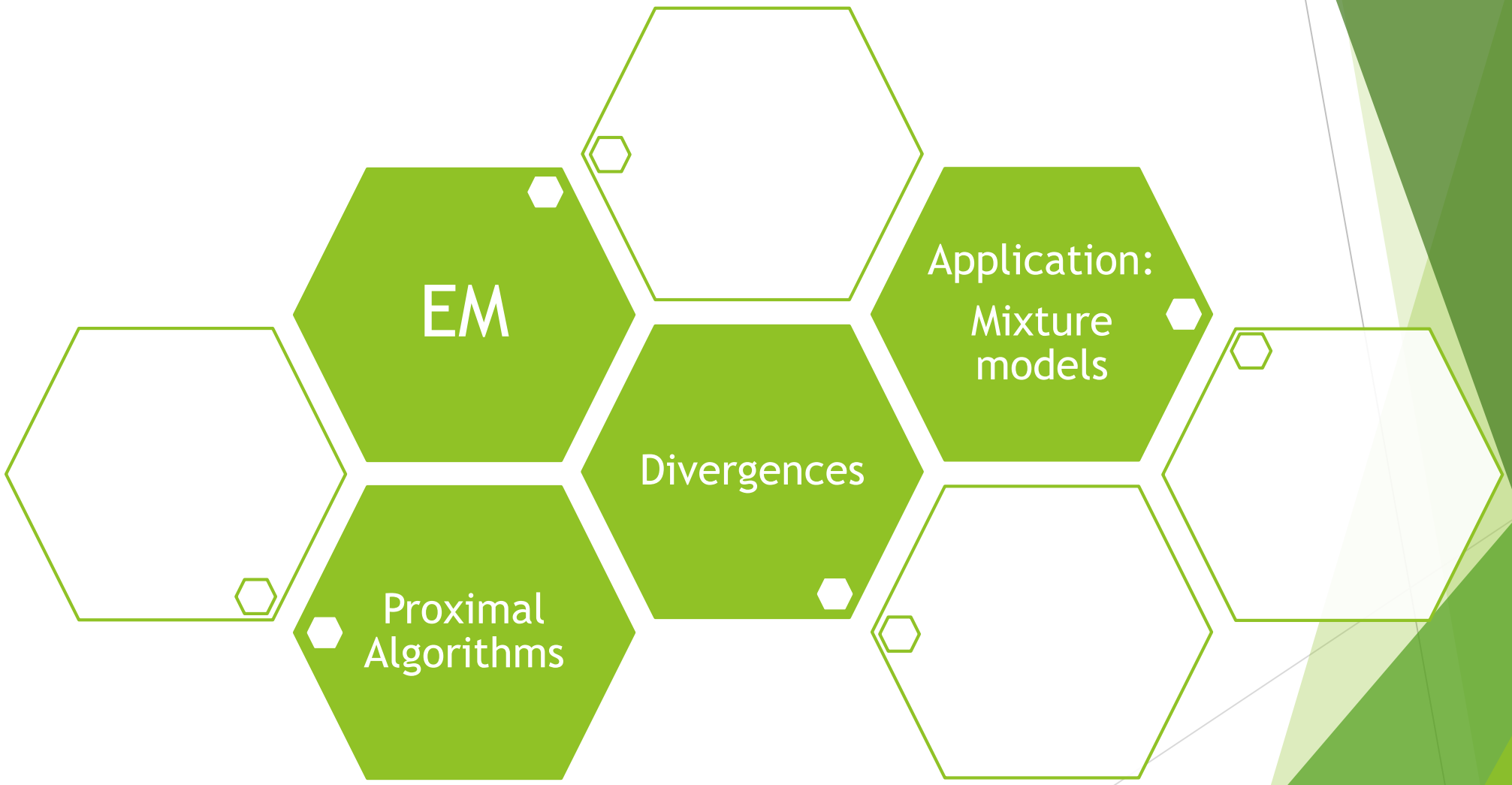
Diaa AL MOHAMAD and Michel BRONIATOWSKI

LSTA, Université de Pierre et Marie Curie, Paris 6

2nd Conference on Geometry Science of Information (GSI)

30 Oct 2015

The Title


EM
Application:
Mixture
models
Divergences
Proximal
Algorithms

φ –Divergences

- A φ –divergence is a measure of discrepancy. For P and Q two probability measures s.t. $Q \ll P$:

$$D_\varphi(Q,P) = \int \varphi\left(\frac{dQ}{dP}(x)\right) dP(x)$$

- Basic property: Identification, i.e.

$$D_\varphi(Q,P) = 0 \Leftrightarrow P = Q$$

  φ needs to be strictly convex.

- Standard Divergences: Kullback-Leibler $\varphi(t) = t \log t - t + 1$, Pearson's χ^2, $\varphi(t) = \frac{1}{2}(t-1)^2$ and Hellinger distance $\varphi(t) = \frac{1}{2}\left(\sqrt{t}-1\right)^2$.
- Robustness of obtained estimators.

Proximal algorithms and WHY ?!

- Idea of perturbing the objective function little in order to avoid some of the local optima

$$x^{k+1} = \underset{x}{\operatorname{arginf}}\, D(x) + \beta_k\, P(x, x^k)$$

  with $P(x, x) = 0$ to ensure convergence towards the optimum of D.

- Beginning with Martinet 1963 with the choice $P\left(x, x^k\right) = \left\|x - x^k\right\|$.
- Gain:
  - a better convergence speed (Goldstein and Russak 1987, Chrétien and Hero 1998).
  - avoiding saddle points (Chrétien and Hero 2008).

EM algorithm

- Dempster 1977: Complete data (X, Y) with joint density f

$$\phi^{k+1} = \underset{\phi}{\operatorname{argmax}}\, \mathbb{E}[\log f(X, Y)\,|Y = y, \phi^k]$$

X: Labels
Y: Observations
p_ϕ: model over Y
ϕ: parameters

- Proximal formulation $y = (y_1, \dots, y_n)$:

$$\phi^{k+1} = \underset{\phi}{\operatorname{argmax}} \sum_{i=1}^{n} \log(p_\phi)(y_i) + \sum_{i=1}^{n} \int \log h_i(x|\phi)\, h_i(x|\phi^k)\, dx$$

with $h_i(x|\phi) = \frac{f(x, y_i)}{p_\phi(y_i)}$ is the conditional density of class x provided observation y_i.

- First generalization (Tseng 2004): $-\log(t) = \psi(t)$

$$\phi^{k+1} = \underset{\phi}{\operatorname{argmax}} \sum_{i=1}^{n} \log(p_\phi)(y_i) - \sum_{i=1}^{n} \int \psi\left(\frac{h_i(x|\phi)}{h_i(x|\phi^k)}\right) h_i(x|\phi^k)\, dx$$

From Log-Likelihood to Divergences

- φ −divergence can be estimated for example using Fenchel-duality.
  - Dual formula (Broniatowski and Keziou 2006, Liese and Vajda 2006):

$$\widehat{D}(P_\phi, P_T) = \sup_\alpha \int \varphi\left(\frac{p_\phi}{p_\alpha}\right)(x)p_\phi(x)dx - \frac{1}{n}\sum_{i=1}^{n}\left[\frac{p_\phi}{p_\alpha}\varphi'\left(\frac{p_\phi}{p_\alpha}\right) - \varphi\left(\frac{p_\phi}{p_\alpha}\right)\right](y_i)$$

Can be Efficient but NOT robust

  - Kernel-based dual formula (AL MOHAMAD 2015):

$$\widehat{D}(P_\phi, P_T) = \int \varphi\left(\frac{p_\phi}{K_{n,w}}\right)(x)p_\phi(x)dx - \frac{1}{n}\sum_{i=1}^{n}\left[\frac{p_\phi}{K_{n,w}}\varphi'\left(\frac{p_\phi}{K_{n,w}}\right) - \varphi\left(\frac{p_\phi}{K_{n,w}}\right)\right](y_i)$$

Robust and Can be Efficient

- For $\varphi(t) = -\log(t) + t - 1$:

$$\operatorname*{arginf}_\phi \widehat{D}(P_\phi, P_n) = MLE$$

Our proposed Algorithm

- For $\varphi_{KLm}(t) = -\log(t) + t - 1$, Tseng's generalization can be rewritten:

$$\begin{aligned}\phi^{k+1} &= \underset{\phi}{\operatorname{argmax}} \sum_{i=1}^{n} \log(p_\phi)(y_i) - \sum_{i=1}^{n} \int \psi\left(\frac{h_i(x|\phi)}{h_i(x|\phi^k)}\right) h_i(x|\phi^k)\, dx \\ &= \underset{\phi}{\operatorname{argmax}} -n\widehat{D}_{\varphi_{KLm}}(P_\phi, P_T) - nD_\psi(\phi|\phi^k) \\ &= \underset{\phi}{\operatorname{arginf}}\, \widehat{D}_{\varphi_{KLm}}(P_\phi, P_T) + D_\psi(\phi|\phi^k)\end{aligned}$$

- For any φ (strictly convex) which defines a φ −divergence:

$$\phi^{k+1} = \underset{\phi}{\operatorname{arginf}}\, \widehat{D}_\varphi(P_\phi, P_T) + D_\psi(\phi, \phi^k)$$

A brief summary of what we have done !

EM

- $\phi^{k+1} = \operatorname{argsup}_{\phi} LogLikelihood(y|\phi) - nD_{-\log t+t-1}(\phi, \phi^k)$

Tseng

- $\phi^{k+1} = \operatorname{argsup}_{\phi} LogLikelihood(y|\phi) - nD_{\psi}(\phi, \phi^k)$

Ours

- $\phi^{k+1} = \operatorname{arginf}_{\phi} \widehat{D}_{\varphi}(P_{\phi}, P_T) + D_{\psi}(\phi, \phi^k)$

Wish for

- Separation into two optimizations for mixture models as in EM
  - Proportion
  - Mixture Parameters

What does the algorithm ensure ?

- **Decrease and convergence** of the objective function at each iteration, i.e.

$$\widehat{D}_{\varphi}\left(P_{\phi^{k+1}}, P_T\right) \leq \widehat{D}_{\varphi}\left(P_{\phi^{k}}, P_T\right)$$

- **Proposition** :

  I. Assume that both $\widehat{D}_{\varphi}$ and D_{ψ} are lower semicontinuous w.r.t ϕ;

  II. Assume that the set $\Phi^0 = \left\{\phi \in \Phi : \widehat{D}_{\varphi}\left(P_{\phi}, P_T\right) \leq \widehat{D}_{\varphi}\left(P_{\phi^0}, P_T\right)\right\}$ is a compact subset of $int(\Phi)$ for some initial point ϕ^0,

  then, the sequence ϕ^k is defined and bounded. Moreover, the sequence $\widehat{D}_{\varphi}\left(P_{\phi^{k}}, P_T\right)$ is decreasing and convergent.

What does the algorithm ensure ? (continued)

- **Convergence towards stationary points** and with more assumptions towards local minima.

- **Proposition** :

  I. Suppose that both $\phi \mapsto \widehat{D}_{\varphi}(P_{\phi}, P_T)$ and D_{ψ} are of class $\mathcal{C}^1$ w.r.t. its first argument;

  II. Suppose also that the set $\Phi^0 = \{\phi \in \Phi : \widehat{D}_{\varphi}(P_{\phi}, P_T) \leq \widehat{D}_{\varphi}(P_{\phi^0}, P_T)\}$ is a compact subset of $int(\Phi)$ for some initial point ϕ^0,

  then any limit point of the sequence ϕ^k is a stationary point of the objective function $\phi \mapsto \widehat{D}_{\varphi}(P_{\phi}, P_T)$. If $\phi \mapsto \widehat{D}_{\varphi}(P_{\phi}, P_T)$ is not differentiable, then 0 belongs to its subgradient calculated at the limit point.

Convergence of the sequence ϕ^k

- **Proposition :**

  I. Suppose that both $\phi \mapsto \widehat{D}_\varphi(P_\phi, P_T)$ and D_ψ are continuous;

  II. Suppose that the set $\Phi^0 = \{\phi \in \Phi: \widehat{D}_\varphi(P_\phi, P_T) \leq \widehat{D}_\varphi(P_{\phi^0}, P_T)\}$ is a compact subset of $int(\Phi)$ for some initial point ϕ^0;

  III. Suppose $D_\psi(\phi|\phi^k) = 0$ iff $\phi = \phi^k$,

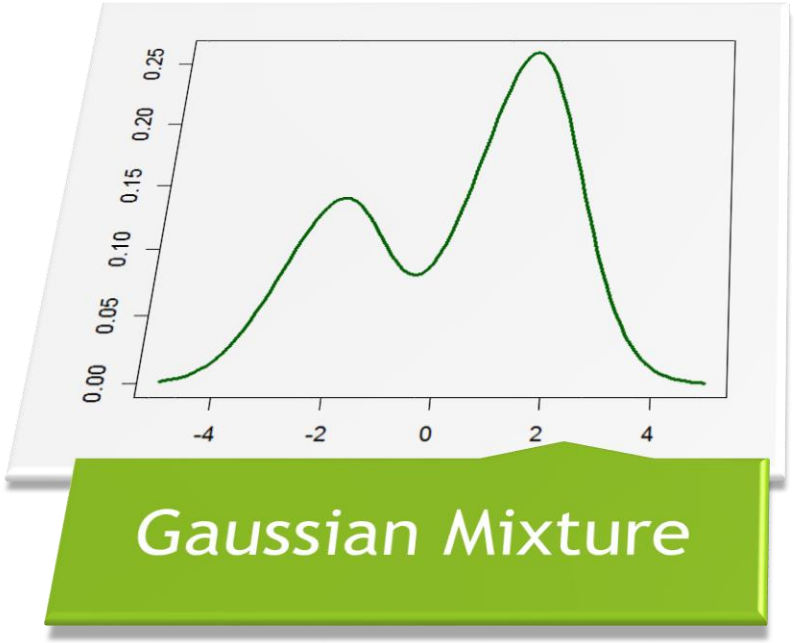
  then $\{\phi^{k+1} - \phi^k\} \to 0$.

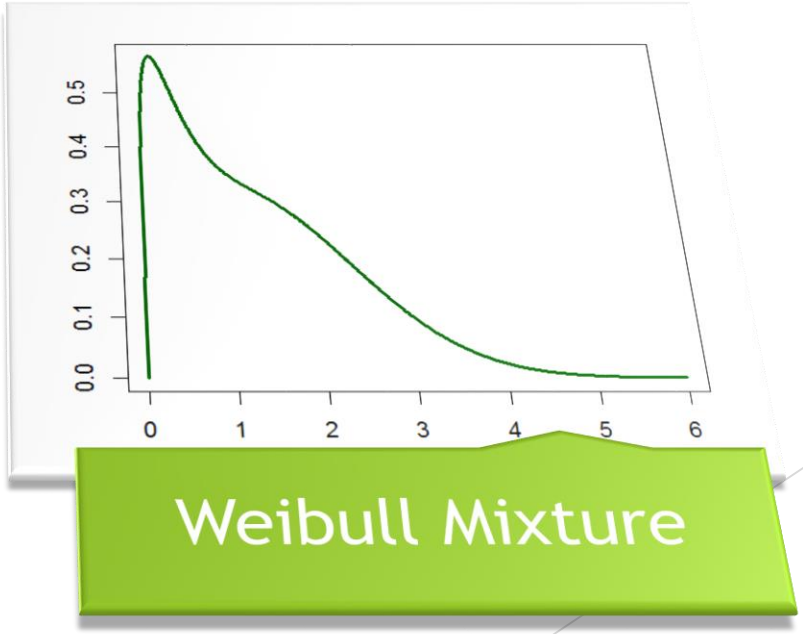
- **Corollary :** Uner the same assumptions, the set of accumulation points of the sequence ϕ^k is a connected set. Moreover, if $\phi \mapsto \widehat{D}_\varphi(P_\phi, P_T)$ is strictly convex in a neighborhood of a limit point of ϕ^k, then the **whole sequence** ϕ^k **converges** to a local minimum of the objective function $\phi \mapsto \widehat{D}_\varphi(P_\phi, P_T)$.

Simulations

- Two component mixture model : $p_\phi(x) = \lambda f(x;\theta_1) + (1-\lambda) f(x;\theta_2)$.
- 100 observations.
- Hellinger divergence $\varphi(t) = \frac{1}{2}\left(\sqrt{t}-1\right)^2$. $\psi(t) = \frac{1}{2}(t-1)^2$.
- Estimation error calculated using the total variation distance defined as:

$$TVD = \sup_{A \in \mathfrak{B}(\mathbb{R})} \left| P_\phi(A) - P_T(A) \right|$$





The two gaussian mixture

True set of parameters : $\lambda = 0.35, \mu_1 = -2, \mu_2 = 1.5, \sigma_1 = \sigma_2 = 1$

Estimation method	λ	sd(λ)	μ_1	sd(μ_1)	μ_2	sd(μ_2)	TVD	sd(TVD)
Without outliers								
Classical MDϕDE	0.349	0.049	−1.989	0.207	1.511	0.151	**0.061**	0.029
Kernel-Based MDϕDE	0.349	0.049	-1.987	0.208	1.520	0.155	0.062	0.029
EM (MLE)	0.360	0.054	−1.989	0.204	1.493	0.136	0.064	0.025
With 10% outliers								
Classical MDϕDE	0.357	0.022	−2.629	0.094	1.734	0.111	0.146	0.034
Kernel-Based MDϕDE	0.352	0.057	−1.756	0.224	1.358	0.132	**0.087**	0.033
EM (MLE)	0.342	0.064	-2.617	0.288	1.713	0.172	0.150	0.034

The two-component Weibull mixture

True set of parameters: $\lambda = 0.35, \nu_1 = 1.2, \nu_2 = 2, \sigma_1 = 0.5, \sigma_2 = 2$

Estimation method	λ	sd(λ)	ν_1	sd(ν_1)	ν_2	sd(ν_2)	TVD	sd(TVD)
Without outliers								
Classical MDϕDE	0.356	0.066	1.245	0.228	2.055	0.237	**0.052**	0.025
Kernel-Based MDϕDE	0.387	0.067	1.229	0.241	2.145	0.289	0.058	0.029
EM (MLE)	0.355	0.066	1.245	0.228	2.054	0.237	**0.052**	0.025
With 10% outliers								
Classical MDϕDE	0.250	0.085	1.089	0.300	1.470	0.335	0.092	0.037
Kernel-Based MDϕDE	0.349	0.076	1.122	0.252	1.824	0.324	**0.067**	0.034
EM (MLE)	0.259	0.095	0.941	0.368	1.565	0.325	0.095	0.035

Simple conclusions

1. It works !
2. Kernel-based MDφDE is clearly robust
3. Classical (supremal) MDφDE has best results under the model even compared to EM.
4. The proximal algorithm has a better stability with respect to initialization than EM

Thanks for your attention