ロバストなemアルゴリズム

総合研究大学院大学　川島　孝行
統計数理研究所　藤澤　洋徳

EM アルゴリズムは，正規混合分布などを対象とした場合に，隠れ変数を考えることで最尤推定を行う手法である．em アルゴリズム (Amari, 1995) はこれを幾何学的に解釈したものである．具体的には，我々があてはめたいパラメトリックな確率モデルからなるモデル多様体 $p(x,z;\theta)$，観測データからなるデータ多様体 $q(x,z;\eta)$ 間の KL ダイバージェンスを交互に最小化を行うことで実現される．

$$\text{e-step}: \qquad \min_{\eta} D_{KL}(q(x,z;\eta), p(x,z;\theta)).$$

$$\text{m-step}: \qquad \min_{\theta} D_{KL}(q(x,z;\eta), p(x,z;\theta)).$$

ここで，$q(x,z;\eta)=\bar{q}(x)q(z|x;\eta)$ という分解を仮定した場合には，EM アルゴリズムと一致することが知られている．

我々は，KL ダイバージェンス最小化の代わりに，ロバストにパラメータ推定が可能な γ ダイバージェンス (Fujisawa and Eguchi, 2008) 最小化に基づく em アルゴリズムを提案する．これを em アルゴリズムに導入するには，具体的に以下のように交互に最小化を行う．

$$\text{Robust} \quad \text{e-step}: \qquad \min_{\eta} D_{\gamma}(\bar{q}(x)q(z;\eta), p(x,z;\theta)).$$

$$\text{Robust} \quad \text{m-step}: \qquad \min_{\theta} D_{\gamma}(\bar{q}(x)q(z;\eta), p(x,z;\theta)).$$

このアルゴリズムには次のような単調性が成り立つ．

$$D_{\gamma}(\bar{q}(x)q(z;\eta^{(0)}), p(x,z;\theta^{(0)}) \geq D_{\gamma}(\bar{q}(x)q(z;\eta^{(1)}), p(x,z;\theta^{(0)}) \geq D_{\gamma}(\bar{q}(x)q(z;\eta^{(1)}), p(x,z;\theta^{(1)}) \geq \ldots.$$

正規混合分布を対象とした場合のアルゴリズムは以下で表される．

$$\begin{aligned}
\mu_k^{(t+1)} &= \frac{\sum_{i=1}^{N} x_i p(x_i|k;\theta^{(t)})^{\gamma}}{\sum_{i=1}^{N} p(x_i|k;\theta^{(t)})^{\gamma}}. \\
\sigma_k^{2(t+1)} &= (1+\gamma)\left\{\frac{\sum_{i=1}^{N} {x_i}^2 p(x_i|k;\theta^{(t)})^{\gamma}}{\sum_{i=1}^{N} p(x_i|k;\theta^{(t)})^{\gamma}} - (\mu_k^{(t+1)})^2\right\}. \\
\pi_k^{(t+1)} &= \frac{\left(\sigma_k^{2(t+1)}\right)^{\frac{\gamma}{2(1+\gamma)}} \left(\sum_{i=1}^{N} p(x_i,k;\theta^{(t)})^{\gamma}\right)^{\frac{1}{\gamma}}}{\sum_{j=0}^{K} \left(\sigma_j^{2(t+1)}\right)^{\frac{\gamma}{2(1+\gamma)}} \left(\sum_{i=1}^{N} p(x_i,j;\theta^{(t)})^{\gamma}\right)^{\frac{1}{\gamma}}}.
\end{aligned}$$

いま，ここで各コンポーネントの平均，分散パラメータの更新式に注目すると各 k については $D_{\gamma}(\bar{q}(x), p(x|k;\theta))$ の更新式と同じであることがわかる．そのため各コンポーネントが離れていた場合，γ ダイバージェンスの性質から，適切な初期値 $\theta^{(0)}$ を選べば，平均および分散パラメータは真の値に収束することがわかる．データ多様体のほうに $q(x,z;\eta)=\bar{q}(x)q(z;\eta)$ という分解の仮定をおいているため，この独立性の仮定が推定に与える影響，正規混合分布を対象とした場合の数値実験の結果については発表当日に報告を行う．