$$\nabla\psi(\bar\theta(s)) + \nabla\mathcal{R}(\bar\theta(s)) - J_\phi^\theta(\bar\theta(s))s = 0.$$

How to apply the implicit function theorem.

$s \mapsto \bar\theta(s)$ is the unique root of the function:

$(s,\theta) \mapsto \nabla\psi(\theta) + \nabla\mathcal{R}(\theta) - \left[J_\phi^\theta(\theta)\right]^T s = 0 = F(s,\theta)$.

↳ more coherent with the definition of the Jacobian.

Implicit function theorem:

$$\odot \quad D_s F(s,\theta) = -\left\{J_\phi^\theta(\theta)\right\}^T$$

$$D_\theta F(s,\theta) = H_\psi^\theta(\theta) + H_{\mathcal{R}}^\theta(\theta) - \sum_{i=1}^{d} s_i H_{\phi_i}^\theta(\theta)$$

where $H_g^\theta(\theta)$ is the Hessian of the scalar function g at θ.

$D_\theta F(s,\theta)$ is also the Hessian of the function: $D_\theta F(s,\theta) = H_\ell(s,\theta)$

$$\theta \mapsto \psi(\theta) + \mathcal{R}(\theta) - \langle\phi(\theta), s\rangle = \ell(s,\theta).$$

Implicit function theorem: if $D_\theta F(s,\bar\theta(s))$ is invertible, then

$s \mapsto \bar\theta(s)$ is differentiable and

$$J_{\bar\theta}^\theta(s) = \left\{D_\theta F(s,\bar\theta(s))\right\}^{-1}\left\{J_\phi^\theta(\theta)\right\}^T.$$

If $\theta \in \mathbb{R}^d$ and $s \in \mathbb{R}^\ell$, then

$J_{\bar\theta}^\theta(s)$ is $d \times \ell$

$J_\phi^\theta(\theta)$ is $\ell \times d$

$D_\theta F(s,\theta)$ is $d \times d$.

Set $w(s) = KL(\pi \| g_{\bar\theta(s)}) + \mathcal{R}(\bar\theta(s))$.

$$\nabla_s w(s) = -J_\phi^\theta(\bar\theta(s))\left[H_\ell^\theta(s,\bar\theta(s))\right]\left\{J_\phi^\theta(\bar\theta(s))\right\}^{-1} h(s)$$

{ see Notes Fruc. pdf on the dropbo

where $h(s) = \mathbb{E}_\pi\left[\bar S(Y;\bar\theta(s))\right] - s$.

Conditions upon which $\nabla_s w(s)$ is Lipschitz:

2.1. Application to Online Expectation Maximization for a GMM

Consider a mixture of M Gaussian distributions with known variance $\sigma^2 > 0$:

$$g_{\boldsymbol{\theta}}(y) = g(y, \boldsymbol{\theta}) = \sum_{m=1}^{M} \frac{\omega_m}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu_m)^2\right) \tag{12}$$

We denote by $\boldsymbol{\theta} = (\omega_1, \ldots, \omega_M, \mu_1, \ldots, \mu_M)$. Note that for any $i \in \{1, \ldots, M\}$, $\omega_i \geq 0$ and $\sum_{m=1}^{M} \omega_m = 1$. To apply the online EM method, we can augment the incomplete data Y_k with a latent variable Z_k taking value in $\{1, \ldots, M\}$. We denote the complete data as $X_k = (Y_k, Z_k)$. The complete data log-likelihood is given by

$$\log f(y, z, \boldsymbol{\theta}) = \sum_{m=1}^{M} \mathbb{1}_{\{z=m\}} \left\{-\frac{1}{2\sigma^2}(y-\mu_m)^2 + \log(\omega_m)\right\} . \tag{13}$$

better to set $\theta = (\omega_1, \ldots, \omega_{M-1}, \mu_1, \ldots, \mu_M)$, $\Theta = \mathcal{S}_M \times \mathbb{R}^M$ where $\mathcal{S}_M$ is the probability simplex

This complete likelihood belongs to the curved exponential family with $S(y, z) = (S_1(y, z), \ldots, S_M(y, z))$ and $\phi(\boldsymbol{\theta}) = (\phi_1(\boldsymbol{\theta}), \ldots, \phi_M(\boldsymbol{\theta}))$ with:

$$S_m(y, z) = \begin{bmatrix} S_m^{(1)}(y, z) \\ S_m^{(2)}(y, z) \end{bmatrix} = \begin{bmatrix} \mathbb{1}_{\{z=m\}} \\ \mathbb{1}_{\{z=m\}} y \end{bmatrix}$$

$$\phi_m(\boldsymbol{\theta}) = \begin{bmatrix} \log(\omega_m) - \frac{\mu_m^2}{2\sigma^2} \\ \frac{\mu_m}{\sigma^2} \end{bmatrix}$$

and $\psi(\boldsymbol{\theta}) \equiv 0$. The conditional expectation of the complete data sufficient statistics is determined by the posterior probabilities of the mixture components given by:

$$w_m(y, \boldsymbol{\theta}) := \mathbb{P}_{\boldsymbol{\theta}}(Z = m \,|\, Y = y) = \frac{\omega_m \exp(-\frac{1}{2\sigma^2}(y-\mu_m)^2)}{\sum_{j=1}^{M} \omega_j \exp(-\frac{1}{2\sigma^2}(y-\mu_j)^2)} \tag{14}$$

The $(n+1)$th step of the online EM is given by:

$$\begin{aligned}
\hat{s}_{m,n+1}^{(1)} &= \hat{s}_{m,n}^{(1)} + \gamma_{n+1}\left(w_m(Y_{n+1}, \boldsymbol{\theta}_n) - \hat{s}_{m,n}^{(1)}\right) \\
\hat{s}_{m,n+1}^{(2)} &= \hat{s}_{m,n}^{(2)} + \gamma_{n+1}\left(Y_{n+1} w_m(Y_{n+1}, \boldsymbol{\theta}_n) - \hat{s}_{m,n}^{(2)}\right) \\
\hat{\omega}_{m,n+1} &= \bar{w}_m(\hat{s}_{n+1}) \\
\hat{\mu}_{m,n+1} &= \bar{\mu}_m(\hat{s}_{n+1})
\end{aligned}$$

$m \in [\![1, M-1]\!]$

We set: $\ell(s;\theta) = \sum_{m=1}^{M}\left\{\log(\omega_m) - \frac{\mu_m^2}{2\sigma^2}\right\} s_m^{(1)} + \frac{1}{\sigma^2}\sum_{m=1}^{M} s_m^{(2)}\mu_m - \underbrace{\frac{\varepsilon^{(2)}}{2\sigma^2}\sum_{m=1}^{M}\mu_m^2 + \varepsilon^{(1)}\sum_{m=1}^{n}\log(\omega_m)}_{-R(\theta)}$

$$= \sum_{m=1}^{M-1}\left\{\log(\omega_m) - \frac{\mu_m^2}{2\sigma^2}\right\} s_m^{(1)} + \left\{\log\left(1 - \sum_{j=1}^{M-1}\omega_j\right) - \frac{\mu_M^2}{2\sigma^2}\right\}\left(1 - \sum_{j=1}^{M-1} s_j^{(1)}\right) + \ldots$$

and thus $\phi_m(\theta) = \left\{\log(\omega_m) - \frac{\mu_m^2}{2\sigma^2}\right\} - \left\{\log\left(1 - \sum_{j=1}^{M-1}\omega_j\right) - \frac{\mu_M^2}{2\sigma^2}\right\}$

$$\bar{w}_m(\underline{s}) = \frac{s_m^{(1)} + \varepsilon^{(1)}}{\sum_{j=1}^{M}\{s_j^{(1)} + \varepsilon^{(1)}\}} = \frac{s_m^{(1)} + \varepsilon^{(1)}}{1 + M\varepsilon^{(1)}} \qquad \varepsilon^{(1)} > 0$$

$$\bar{\mu}_m(\underline{s}) = \frac{s_m^{(2)}}{s_m^{(1)} + \varepsilon^{(2)}} \qquad \varepsilon^{(2)} > 0$$

$(s_1^{(1)}, \ldots, s_{M-1}^{(1)}) \in \mathcal{S}_M$ (M-dimensional open simplex $= \{(a_1, \ldots, a_{M-1}) \in \mathbb{R}^{M-1},\ a_i > 0, \sum_{j=1}^{M-1} a_j < 1\}$.)

$(s_1^{(2)}, \ldots, s_M^{(2)}) \in \mathbb{R}^M$ (the means are unconstrained).

In this case, $\underline{s} \mapsto \bar{\theta}(s)$ is differentiable over $\mathcal{S}_M$. This is a case where the computations are explicit, but it can also be derived from our general result.

$M=2$

$$\left(\log(\omega) - \frac{\mu_1^2}{2\sigma^2}\right) s^{(1)} + \left(\log(1-\omega) - \frac{\mu_2^2}{2\sigma^2}\right)\left(1-s^{(1)}\right) + \frac{1}{\sigma^2}\mu_1 s_1^{(2)} + \mu_2 s_2^{(2)}$$

$\theta = (\omega, \mu_1, \mu_2)$.

$$\phi_1(\theta) = \left(\log(\omega) - \frac{\mu_1^2}{2\sigma^2}\right) - \left(\log(1-\omega) - \frac{\mu_2^2}{2\sigma^2}\right) \qquad \psi(\theta) = \log(1-\omega) - \frac{\mu_2^2}{2\sigma^2}$$

$$\phi_2(\theta) = \frac{1}{\sigma^2}\mu_1$$

$$\phi_3(\theta) = \frac{\mu_2}{\sigma^2}.$$

$$J_\phi^\theta(\theta) = \begin{bmatrix} \left\{\frac{1}{\omega} + \frac{1}{1-\omega}\right\} & -\frac{\mu_1}{\sigma^2} & \frac{\mu_2}{\sigma^2} \\ 0 & \frac{1}{\sigma^2} & 0 \\ 0 & 0 & \frac{1}{\sigma^2} \end{bmatrix}$$

$$\ell(s,\theta) = \left(\log(\omega) - \frac{\mu_1^2}{2\sigma^2}\right) s^{(1)} + \left(\log(1-\omega) - \frac{\mu_2^2}{2\sigma^2}\right)\left(1-s^{(1)}\right) + \frac{1}{\sigma^2}\mu_1 s_1^{(2)} + \mu_2 s_2^{(2)} - \frac{\varepsilon^{(1)}\mu_1^2}{2\sigma^2} - \frac{\varepsilon^{(2)}\mu_2^2}{2\sigma^2} + \varepsilon^{(1)}\left(\log(\omega) + \log(1-\omega)\right)$$

$$\frac{\partial \ell(s,\theta)}{\partial \omega} = \frac{s^{(1)}}{\omega} - \frac{(1-s^{(1)})}{1-\omega} + \frac{\varepsilon^{(1)}}{\omega} - \frac{\varepsilon^{(1)}}{1-\omega} = \frac{s^{(1)}+\varepsilon^{(1)}}{\omega} - \frac{(1-s^{(1)}+\varepsilon^{(1)})}{1-\omega}.$$

$$\frac{\partial \ell(s,\theta)}{\partial \mu_1} = -\frac{\mu_1 s^{(1)}}{\sigma^2} + \frac{s_1^{(2)}}{\sigma^2} - \frac{\varepsilon^{(2)}\mu_1}{\sigma^2}$$

$$\frac{\partial \ell(s,\theta)}{\partial \mu_2} = -\frac{\mu_2}{\sigma^2}\left(1-s^{(1)}\right) + \frac{s_2^{(2)}}{\sigma^2} - \frac{\varepsilon^{(2)}\mu_2}{\sigma^2}.$$

$$\frac{\partial^2 \ell(s,\theta)}{\partial \omega^2} = -\frac{s^{(1)}+\varepsilon^{(1)}}{\omega^2} - \frac{(1-s^{(1)}+\varepsilon^{(1)})}{(1-\omega)^2}. \qquad \frac{\partial^2 \ell(s,\theta)}{\partial s \partial \mu_1} = 0 \qquad \frac{\partial^2 \ell(s,\theta)}{\partial s \partial \mu_2} = 0$$

$$\frac{\partial^2 \ell(s,\theta)}{\partial \mu_1 \partial \omega} = 0 \qquad \frac{\partial^2 \ell(s,\theta)}{\partial \mu_1^2} = -\frac{s^{(1)}+\varepsilon^{(2)}}{\sigma^2} \qquad \frac{\partial^2 \ell(s,\theta)}{\partial \mu_1 \partial \mu_2} = 0$$

$$\frac{\partial^2 \ell(s,\theta)}{\partial \mu_2 \partial \omega} = 0 \qquad \frac{\partial^2 \ell(s,\theta)}{\partial \mu_2 \partial \mu_1} = 0 \qquad \frac{\partial^2 \ell(s,\theta)}{\partial \mu_2^2} = -\frac{s^{(1)}+\varepsilon^{(2)}}{\sigma^2}.$$

The Hessian matrix is diagonal and bounded.

Therefore ... (check what is written above and conclude the computations).

$\nabla_s w(s) =$ com