Gaussian Mixture Models

jhsh
jhshi@vt.edu

April 8, 2021

Introduction: Gaussian as Mixtures

Model: Data $x \in \mathbf{R}^M$ disbursed *normally* with K *centroids*

$$p(x_i \,|\, (\mu,\sigma)_k) = \mathcal{N}(x_i, (\mu,\sigma)_k) \equiv g_{ik} \quad x_i = \text{Observable Data Point i}$$
$$(\mu,\sigma)_k = \text{mean, covariance k}$$

Hidden DOF $(z_j) \equiv$ "one hot" at some $k := \mathbf{1}(k)$ **k-th centroid**

$$p(z_j \,|\, \pi) = \prod_{s \in [1,\mathscr{K}]} \pi_s^{(z_j)^s} \quad \text{"simplex:"} \sum_{s \in [1,\mathscr{K}]} \pi_s = 1$$
$$p(x_i \,|\, z_i(\mu,\sigma)) = \prod_{t \in [1,\mathscr{K}]} g_{it}^{(z_i)^t} \quad p(x \,|\, z(\mu,\sigma)) = \prod_{i \in [1,N]} p(x_i \,|\, z_i(\mu,\sigma))$$

Introduction: z_j unobserved *vs.* observed

$\mathscr{L}_N(\theta \,|\, x, z) \equiv$ "log likelihood" : $\arg\max_\theta p(x, z \,|\, \theta) \simeq \arg\max_\theta \mathscr{L}_N(\theta \,|\, x, z)$

- if z_j remained *hidden*:

$$\mathscr{L}(\theta, x, z) = \sum_{i \in [1,N]} \log \sum_{z_i} p(x_i \,|\, z_i \theta) p(z_i \,|\, \theta) = \sum_{i \in [1,N]} \log \sum_{z_i} \prod_{s \in [1,\mathscr{K}]} (\pi_s g_{is})^{(z_i)^s}$$

  $\sum_{z_i}$ **Hard to Compute...**

- assume *complete* observability $D = (x, z)$: $z_j^k := \mathbf{1}(k)$ k-th cluster

$$\mathscr{L}_N(\theta \,|\, x, z) = \sum_{i \in [1,N]} \sum_{j \in [1,\mathscr{K}]} (z_i)^j \cdot \log(\pi_j g_{ij})$$

Q-Learning(?)

convenient to define $Q(\theta^{(t)}) \equiv \langle \mathscr{L}(\theta, x, z) \rangle_{p\sim z|x\theta(t)}$ **iteration** t

$$Q(\theta^{(t+1)}) \leftarrow \left\langle \sum_{i\in[1,N]} \sum_{j\in[1,\mathscr{K}]} (z_i)^j \log(\pi_j g_{ij})^{(t)} \right\rangle_{p\sim z|x\theta(t)}$$

$$= \sum_{i\in[1,N]} \sum_{j\in[1,\mathscr{K}]} r_{ij}^{(t)} \log(\pi_j g_{ij})^{(t)}$$

$r_{ij} = \left\langle (z_i)^j \right\rangle_{p\sim z|x\theta(t)} = p((z_i)^j \mid \theta, x_i) = (\pi_k g_{ik})(\sum_{s\in[1,\mathscr{K}]}(\pi_s g_{is}))^{(-)}$
"responsibilities"

Optimization:

- update (r, Q):

$$r_{ik}^{(t+1)} \leftarrow (\pi_k g_{ik})(\sum_{s\in[1,\mathscr{K}]} (\pi_s g_{is})^{(t)})^{(-)}$$

- update parameters θ:

$$\pi_k^{t+1} \leftarrow N^{(-)}(\sum_{i\in[1,N]} r_{ik}^t)$$

$$\mu_k^{t+1} \leftarrow (\sum_{i\in[1,N]} r_{ik}^t)^{(-)}(\sum_{i\in[1,N]} r_{ik}^t x_i)$$

$$\sigma_k^{t+1} \leftarrow \frac{\sum_i r_{ik} \|x_i - \mu_k\|^2}{2\sum_i r_{ik}}$$

eg: $0 \equiv \frac{\partial}{\partial \pi_k}\Big(Q(\theta^{(t)}) + \lambda(\sum_{s\in[1,\mathscr{K}]} \pi_s - 1)\Big) \rightarrow \lambda = -N$
*** derivations ***

Blank Page

review: EM

- Let x be observed data and θ be the underlying parameters, and z be the latent variables.
- "incomplete" log-likelihood: $\mathscr{L}_N(x, \theta) = \log p(x\theta)$
- "complete" log-likelihood: $\mathscr{L}_N(x, \theta)^{\textsf{comp}} = \log p(zx\theta)$
- But $\mathscr{L}_N(x, \theta)^{\textsf{comp}}$ is not physically observable, since z is still hidden from view. One proposal, i.e EM, is to trace off z along with its posterior.
- EM algorithm:
  - At present t, calculate the Q-function:

$$Q(\theta \mid \theta') \equiv \left\langle \log p(x, z, \theta) \middle| x\theta' \right\rangle \quad \text{where} \quad \left\langle \bullet \middle| x\theta' \right\rangle \sim \sum_z p(z \mid x\theta')(\bullet) \quad \text{"E Step"}$$

  - maximize Q with respect to θ:

$$\theta_t \leftarrow \arg\max_{\theta_t} Q(\theta_{t+1} \mid \theta_t) \quad \text{"M Step"}$$

- "Likelihood never decreases in EM." $\mathscr{L}_N(x, \theta_t) \leq \mathscr{L}_N(x, \theta_{t+1})$
  i.e non-negative rest term: $\Delta(x, \theta, q) \equiv \mathscr{L}_N(x, \theta) - F(\theta, x, q)$
  - $\Delta(x, \theta, q) = D_{\textsf{KL}}(q \parallel w)$

continued

- Rate of Convergence: assume $\lim_{t\gg 0}(\theta_t, \theta_{t+1}) \equiv \theta_{(*)}$ and vanishing $\left(\partial_\theta Q(\theta, \theta')\right)_{\theta_{(*)}} \sim 0 \sim \left(\partial_\theta Q(\theta_t, \theta_{t+1})\right)$

$$R \equiv \lim_{t\gg 0}(\theta_{t+1} - \theta_{(*)})(\theta_t - \theta_{(*)})^{(-)} = \mathscr{I}_{\textsf{com}}(\theta_{(*)} \,|\, x)\mathscr{I}^{(-)}_{\textsf{cond}}(\theta_{(*)} \,|\, x)$$

- Note: $\mathscr{I}_{\textsf{com}} \equiv (-)\left\langle \partial^2_{\theta_*} \log p(z, x, \theta_*) \middle| x, \theta_* \right\rangle$ complete Fisher metric
$\mathscr{I}_{\textsf{cond}} \equiv (-)\left\langle \partial^2_{\theta_*} \log p(z \,|\, x, \theta_*) \middle| x, \theta_* \right\rangle$ observed Fisher metric

Derivation

"$R \equiv \lim_{t \gg 0} (\theta_{t+1} - \theta_{(*)})(\theta_t - \theta_{(*)})^{(-)} = \mathscr{I}_{\textsf{com}}(\theta_{(*)} \,|\, x) \mathscr{I}^{(-)}_{\textsf{cond}}(\theta_{(*)} \,|\, x)$"?

- Taylor expand $\lim_{(\theta_1 \theta_2) = (\theta_t \theta_{t+1})} \partial_{\theta_2} Q(\theta_2 \,|\, \theta_1)$ around convergent $\theta_{(*)}$:

$$\begin{aligned}\lim_{(\theta_1 \theta_2) = (\theta_t \theta_{t+1})} \partial_{\theta_2} Q(\theta_2 \,|\, \theta_1) \approx \lim_{(\theta_1 \theta_2) = (\theta_t \theta_{t+1})} \Big[&\lim_{(\theta_1 \theta_2) = (\theta_*)} \partial_{\theta_2} Q(\theta_2 \,|\, \theta_1) \\ &+ (\theta_1 - \theta_*) \partial_{\theta_1} \partial_{\theta_2} Q(\theta_2 \,|\, \theta_1) + (\theta_2 - \theta_*) \partial^2_{\theta_2} Q(\theta_2 \,|\, \theta_1) \\ &+ \mathcal{O}(\sim \theta^2) \Big]\end{aligned}$$

- Employ "$\partial_{\theta_2} Q(\theta_2 \,|\, \theta_1) \Big|_{(\theta_t \theta_{t+1})} = 0 = \partial_{\theta_2} Q(\theta_2 \,|\, \theta_1) \Big|_{(\theta_* \theta_*)}$"?

$$\lim_{(\theta_1 \theta_2) = (\theta_t \theta_{t+1})} \left[(\theta_1 - \theta_*) \partial_{\theta_1} \partial_{\theta_2} Q(\theta_2 \,|\, \theta_1) + (\theta_2 - \theta_*) \partial^2_{\theta_2} Q(\theta_2 \,|\, \theta_1) \right] = 0$$

$$R = \lim_{t \gg 0} \frac{\theta_{t+1} - \theta_*}{\theta_t - \theta_*} = (-) \lim_{(\theta_1 \theta_2) \to (\theta_* \theta_*)} \frac{\partial_{\theta_1} \partial_{\theta_2} Q(\theta_2 \,|\, \theta_1)}{\partial^2_{\theta_2} Q(\theta_2 \,|\, \theta_1)}$$

- $\lim_{(\theta_1 \theta_2) \to (\theta_* \theta_*)} \partial_{\theta_1} \partial_{\theta_2} Q(\theta_2 \,|\, \theta_1) \underset{?}{=} (-) \cdot \lim_{(\theta_1 \theta_2) \to (\theta_* \theta_*)} \partial^2_{\theta_1} D(\theta_2 \,|\, \theta_1)$

$$\begin{aligned}\lim_{(\theta_1 \theta_2) \to (\theta_* \theta_*)} \partial_{\theta_1} \partial_{\theta_2} Q(\theta_2 \,|\, \theta_1) &= \lim_{(\theta_1 \theta_2) \to (\theta_* \theta_*)} \partial_{\theta_1} \partial_{\theta_2} \Big(\sum_z p(z \,|\, x, \theta_1) \textsf{log}\, p(x, z, \theta_2) \Big) \\ &= \lim_{(\theta_1 \theta_2) \to (\theta_* \theta_*)} \partial_{\theta_1} \partial_{\theta_2} \Big(\sum_z p(z \,|\, x, \theta_1) \textsf{log}\, p(z \,|\, x, \theta_2) \Big)\end{aligned}$$

- continued:

$$\lim_{(\theta_1\theta_2)\to(\theta_*\theta_*)} \partial_{\theta_1}\partial_{\theta_2} Q(\theta_2\,|\,\theta_1) = \lim_{(\theta_1\theta_2)\to(\theta_*\theta_*)} \sum_z \big(\partial_{\theta_1} p(z\,|\,x,\theta_1)\big)\cdot \partial_{\theta_2}\log p(z\,|\,x,\theta_2)$$

$$= \lim_{(\theta_1\theta_2)\to(\theta_*\theta_*)} \sum_z \big(\partial_{\theta_*} p(z\,|\,x,\theta_*)\big)\cdot \partial_{\theta_*}\log p(z\,|\,x,\theta_*)$$

$$= \lim_{(\theta_1\theta_2)\to(\theta_*\theta_*)} \sum_z \partial_{\theta_*}\Big[\big(p(z\,|\,x,\theta_*)\big)\cdot \partial_{\theta_*}\log p(z\,|\,x,\theta_*)\Big]$$

$$+ (-) \lim_{(\theta_1\theta_2)\to(\theta_*\theta_*)} \sum_z \big(p(z\,|\,x,\theta_*)\big)\cdot \partial^2_{\theta_*}\log p(z\,|\,x,\theta_*)$$

- The first term vanish under the rule of total derivative.

$$\sum_z \partial_{\theta_*}\Big[\big(p(z\,|\,x,\theta_*)\big)\cdot \partial_{\theta_*}\log p(z\,|\,x,\theta_*)\Big] = \sum_z \partial_{\theta_*}\Big[p(z\,|\,x,\theta_*)\cdot \frac{\partial_{\theta_*} p(z\,|\,x,\theta_*)}{p(z\,|\,x,\theta_*)}\Big]$$

$$= \sum_z \partial_{\theta_*}\Big[\partial_{\theta_*} p(z\,|\,x,\theta_*)\Big] = \partial^2_{\theta_*} \sum_z \Big[p(z\,|\,x,\theta_*)\Big] = \partial^2_{\theta_*}[1] = 0$$

- WLOG: $\lim_{\theta_1\theta_2\to\theta_*}(\bullet) \sim \lim_{t\gg 0}\lim_{\theta_1\theta_2\to\theta_t\theta_{t+1}}(\bullet)$

$$\lim_{(\theta_1\theta_2)\to(\theta_*\theta_*)} \partial_{\theta_1}\partial_{\theta_2} Q(\theta_2\,|\,\theta_1) = (-) \lim_{(\theta_1\theta_2)\to(\theta_*\theta_*)} \sum_z p(z\,|\,x,\theta_*)\cdot \partial^2_{\theta_*}\log p(z\,|\,x,\theta_*)$$

$$= (-) \lim_{t\gg 0} \sum_z p(z\,|\,x,\theta_t)\cdot \partial^2_{\theta_{t+1}}\log p(z\,|\,x,\theta_{t+1})$$

$$= (-) \lim_{t\gg 0} \partial^2_{\theta_{t+1}} \langle \log p(z\,|\,x,\theta_{t+1}) | x,\theta_t\rangle$$

- Substitute definitions for various Fisher metrics:

$$\lim_{(\theta_1 \theta_2)\to(\theta_* \theta_*)} \partial_{\theta_1} \partial_{\theta_2} Q(\theta_2 \mid \theta_1) = (-) \lim_{t \gg 0} \partial^2_{\theta_{t+1}} \langle \log p(z \mid x, \theta_{t+1}) | x, \theta_t \rangle$$

$$= (-) \left\langle \partial^2_{\theta_*} \log p(z \mid x, \theta_*) \middle| x, \theta_* \right\rangle = \mathscr{I}_{\mathsf{cond}}(\theta_* \mid x)$$

- Therefore, Rate of Convergence R of the parameter θ is proportional to ratios of Fisher matrices:

$$R = \lim_{t \gg 0} \frac{\theta_{t+1} - \theta_*}{\theta_t - \theta_*} = \frac{\mathscr{I}_{\mathsf{com}}(\theta_* \mid x)}{\mathscr{I}_{\mathsf{cond}}(\theta_* \mid x)}$$

Derivation

"$(\partial_\theta Q(\theta \,|\, \theta'))_{\theta_{(*)}} = 0 = \big(\partial_{\theta_{t+1}} Q(\theta_{t+1} \,|\, \theta_t)\big)$"?

- $\partial_{\theta_{t+1}} Q(\theta_t, \theta_{t+1}) = 0$ due to the definition of M Step:

$$\partial_{\theta_{t+1}} Q(\theta_t, \theta_{t+1}) = \lim_{\theta \to \theta_{t+1}} \partial_\theta \max_\theta Q(\theta \,|\, \theta_t) = 0$$

- θ_* maximizes $Q(\theta \,|\, \theta_*)$ for all θ:

$$(\partial_\theta Q(\theta \,|\, \theta'))_{\theta_{(*)}} = \lim_{\theta \to \theta_*} \partial_\theta \max_\theta Q(\theta \,|\, \theta_*) = 0$$

Derivation

"$\mathscr{L}_N(\theta_t \,|\, x) \leq \mathscr{L}_N(\theta_{t+1} \,|\, x)$"?

- Suppose decomposition: $\mathscr{L}_N(\theta_t \,|\, x) \overset{!}{=} F(x, \theta_t, q) + \Delta(x, \theta_t, q)$ where Δ is a non-negative term, with at least one zero $\Delta(q_*) = 0$.
- Suppose q_* sets $\Delta(x, \theta_t, q_*) = 0$:

$$\mathscr{L}_N(\theta_t \,|\, x) = F(x, \theta_t, q_*) \quad \text{"E Step"}$$

- Update $\theta_{t+1} \leftarrow \arg\max_{\theta_t} F(x, \theta_t, q_*)$ "M Step":

$$\mathscr{L}_N(\theta_t \,|\, x) \leq F(x, \theta_{t+1}, q_*) \leq \mathscr{L}_N(\theta_{t+1} \,|\, x) + (-)\Delta(x, \theta_{t+1}, q_*)$$

- Therefore, since Δ is non-negative:

$$\mathscr{L}_N(\theta_t \,|\, x) \leq \mathscr{L}_N(\theta_{t+1} \,|\, x)$$

Derivation

"$\Delta(x, \theta_t, q) = D_{\mathsf{KL}}(q(z) \parallel p(z \mid x, \theta))$"?

- Log-Likelihood fractures into various "entropies":

$$\begin{aligned}
\mathscr{L}_N(\theta \mid x) \equiv \sum_z q(z) \log p(x, \theta) &= \sum_z q(z) \log \left(p(z, x, \theta) p(z \mid x, \theta)\right) \\
&= \sum_z q(z) \log \left(p(z, x, \theta) p(z \mid x, \theta) q(z) q^{(-)}(z)\right) \\
&= \sum_z q(z) \log q(z) + \sum_z q(z) \log \left(p(z, x, \theta)\right) \\
&\qquad + \sum_z q(z) \log \left(p(z \mid x, \theta) q^{(-)}(z)\right) \\
&= (-)H(q) + Q(q(z) \parallel p(z, x, \theta)) + (-)D_{\mathsf{KL}}(q(z) \parallel p(z \mid x, \theta))
\end{aligned}$$

- Where $H(q) = (-)\sum_z q(z) \log q(z)$ is Shannon entropy. Assign:

$$\begin{aligned}
F(x, \theta, q) &:= (-)H(q) + Q(q(z) \parallel p(z, x, \theta)) \\
\Delta(x, z, q) &:= D_{\mathsf{KL}}(q(z) \parallel p(z \mid x, \theta))
\end{aligned}$$

- Δ is non-negative since it's equivalent to Kullback–Leibler divergence:

$$\begin{aligned}
\Delta(x, z, q) &= D_{\mathsf{KL}}(q(z) \parallel p(z \mid x, \theta)) \\
&= \sum_z q(z) \log q(z) p^{(-)}(z \mid x, \theta)
\end{aligned}$$

- continued:

$$\sum_z q(z)\text{log}\, q(z)p^{(-)}(z\,|\,x,\theta) = \Big\langle \text{log}\, q(z)p^{(-)}(z\,|\,x,\theta) \Big\rangle_{z\sim q}$$

- Jensens' inequality for $\phi(\bullet) = (-)\text{log}\,(\bullet)$ which is convex:

$$\Big\langle \text{log}\, q(z)p^{(-)}(z\,|\,x,\theta) \Big\rangle_{z\sim q} = (-)\Big\langle (-)\text{log}\, q(z)p^{(-)}(z\,|\,x,\theta) \Big\rangle = \Big\langle \phi\big(q^{(-)}(z)p(z\,|\,x,\theta)\big) \Big\rangle$$

- By that identity, KL divergences are non-negative:

$$\begin{aligned}
\Big\langle \phi\big(q^{(-)}(z)p(z\,|\,x,\theta)\big) \Big\rangle &\geq \phi\big(\Big\langle q^{(-)}(z)p(z\,|\,x,\theta) \Big\rangle \big) \\
&\geq \text{log}\,\big(\sum_z q(z)q^{(-)}(z)p(z\,|\,x,\theta)\big) \\
&\geq \text{log}\,\big(\sum_z p(z\,|\,x,\theta)\big) \\
&\geq \text{log}\,(1) \\
&\geq (0)
\end{aligned}$$

- Therefore $\Delta(x, z, q) \geq (0)$
- To achieve the minimum: $q(z) \leftarrow p(z\,|\,x,\theta)$ "E Step":

$$\begin{aligned}
\Delta(x, z, q) &= \sum_z q(z)\text{log}\, q(z)p^{(-)}(z\,|\,x,\theta) \overset{!}{=} 0 \quad \forall q(z) \neq 0 \\
0 &= \text{log}\,\big(\sum_z q(z)p^{(-)}(z\,|\,x,\theta) \\
q(z) &= p(z\,|\,x,\theta)
\end{aligned}$$

summarize GMM:

- Log-likelihood of "hard clustering": $\mathscr{L}_N^{\mathsf{GMM}} \equiv \sum_{\alpha\in[1,N]} \log p(x_\alpha \,|\, \theta)$
- Assume "latent" variables in "one-hot" representations: $z_\alpha \sim \mathbf{1}_\alpha(J_*)$ with $J_* \in [1, \mathscr{K}]$.
- z_α encodes cluster indices for data $x_\alpha, \forall \alpha \in [1, N]$. This means z_α distributes in accord with a multinomial distribution:

$$p(z \,|\, \pi) = \prod\nolimits_{\alpha\in[1,N]} p(z_\alpha \,|\, \pi_I) \; = \prod\nolimits_{\alpha\in[1,N]} \prod\nolimits_{I\in[1,\mathscr{K}]} (\pi_I)^{z_\alpha^I}$$

  - This can be derived by substituting "one hot" $z_{\alpha I}$ into a general multinomial PD:

$$p(z_\alpha \,|\, \pi) = \frac{\mathscr{N}!}{\prod_{\alpha\in[1,N]} z_{\alpha I}} \prod\nolimits_{\alpha\in[1,N]} (\pi_I)^{z_{\alpha I}} \quad \text{where} \quad \mathscr{N} := \sum_{\alpha\in[1,N]} z_{\alpha I}$$

- Given a label z_α, assume x_α follows a Gaussian distribution, and the joint PD:

$$p(x \,|\, z\theta) \equiv \prod\nolimits_\alpha \prod\nolimits_I \mathcal{N}(x_\alpha, (\mu\sigma)_{\alpha I})^{(z_\alpha)^I} := \prod\nolimits_\alpha \prod\nolimits_I g_{\alpha I}^{(z_\alpha)^I}$$

- Primitive attempt leads to "incomplete" likelihood:

$$\mathscr{L}_N(\theta \,|\, x) \equiv \log \prod\nolimits_{\alpha\in[1,N]} p(x_\alpha, \theta) = \sum_{\alpha\in[1,N]} \log \sum_{z_\alpha} p(x_\alpha, z_\alpha, \theta)$$

GMM, continued:

- However, the task now involves "summing over all categories of z," which is hard (Why?). Suppose z be chosen, $\sum_z \sim 1$ and it leads to "complete" likelihood:

$$\mathscr{L}_N^{\text{comp}}(\theta \,|\, x) = \sum_{\alpha\in[1,N]} \log p(x_\alpha, z_\alpha, \theta) = \sum_{\alpha\in[1,N]} \sum_{I\in[1,\mathscr{K}]} (z_\alpha)^I \log \pi_I \cdot g_{\alpha I}$$

- Define Q-function:

$$Q(\theta_{t+1} \,|\, \theta_t) \equiv \left\langle \mathscr{L}_N^{\text{comp}} \middle| x\theta_t \right\rangle = \sum_{I\in[1,\mathscr{K}]} \sum_{\alpha\in[1,N]} \Big[\sum_{z_\alpha} p(z_\alpha \,|\, x\theta_t)(z_\alpha)^I\Big] \log\left(\pi_I \cdot g_{\alpha I}\right)$$

- The average over $(z_\alpha)^I$ is called Responsibility, and amounts to a probability weight for Q:

$$r_{\alpha I} \equiv \sum_{I\in[1,\mathscr{K}]} p(z_\alpha \,|\, x\theta^t)(z_\alpha)^I = p(z_\alpha^I \,|\, x\theta^t) = \frac{\pi_I g_{I\alpha}}{\sum_{J\in[1,\mathscr{K}]} \pi_J g_{J\alpha}}$$

- Thus, Q-function works out to be:

$$Q(\theta_{t+1} \,|\, \theta_t) \equiv \left\langle \mathscr{L}_N^{\text{comp}} \middle| x\theta_t \right\rangle = \sum_{I\in[1,\mathscr{K}]} \sum_{\alpha\in[1,N]} r_{\alpha I} \cdot \log\left(\pi_I \cdot g_{\alpha I}\right)$$

- EM algorithm optimizes Q-function on θ:
  - E Step consists of evaluating $r_{I\alpha}$ and then $Q(\theta^{t+1} \,|\, \theta^t)$.
  - M Step consists of updating the parameters θ by max'ing out Q:

$$\theta_I^{t+1} \leftarrow \arg\max_\theta Q(\theta \,|\, \theta^t)$$

GMM, continued:

- The update rules work out to be:

$$\pi_I^{t+1} \leftarrow \frac{N_I}{N} \quad N_I \equiv \sum_{i \in [1,N]} r_{\alpha I}^t$$

$$\mu_I^{t+1} \leftarrow \frac{1}{N_I}\left(\sum_{\alpha \in [1,N]} r_{I\alpha}^t x_\alpha\right) \quad \sigma_I^{t+1} \leftarrow \frac{\sum_\alpha r_{\alpha I} \|x_\alpha - \mu_I\|^2}{2\sum_\alpha r_{\alpha I}}$$

Derivation

"$\pi_I^{t+1} \leftarrow \frac{N_I}{N} \quad N_I \equiv \sum_{\alpha\in[1,N]} r^t_{\alpha I}$"?

- Remind that the task is to update π as the maximal solution to Q:

$$\pi_I^{t+1} \leftarrow \underset{\pi_I}{\arg\max}\Big[Q(\theta\,|\,\theta^t) + \lambda\cdot\big(\sum_{I\in[1,\mathscr{K}]}\pi_I - 1\big)\Big]$$

- In practice, differentiate Q with the Lagrange multiplier λ, and set to 0:

$$0 \stackrel{!}{=} \partial_{\pi_I}\Big(Q(\theta\,|\,\theta^t) + \lambda\cdot\big(\sum_{J\in[1,\mathscr{K}]}\pi_J - 1\big)\Big) = \partial_{\pi_J}Q(\theta\,|\,\theta^t) + \lambda\sum_{J\in[1,\mathscr{K}]}\partial_{\pi_I}\pi_J$$
$$= \partial_{\pi_J}Q(\theta\,|\,\theta^t) + \lambda\sum_{J\in[1,\mathscr{K}]}\delta_{IJ} = \partial_{\pi_J}Q(\theta\,|\,\theta^t) + \lambda$$

- The derivative of Q:

$$\partial_{\pi_J}Q(\theta\,|\,\theta^t) = \partial_{\pi_J}\sum_{I\in[1,\mathscr{K}]}\sum_{\alpha\in[1,N]} r_{\alpha I}\log\left(\pi_I\cdot g_{\alpha I}\right) = \sum_{I\alpha} r_{\alpha I}\partial_{\pi_J}\log\left(\pi_I\cdot g_{\alpha I}\right)\sum_{I\alpha} r_{\alpha I}(\pi_I$$
$$= \sum_{I\alpha}\pi_I^{(-)} r_{\alpha I}\delta_{IJ} = \sum_{\alpha}\pi_J^{(-)} r_{\alpha J} = \pi_J^{(-)} N_J$$

- Solving the differential equation:

$$0 \stackrel{!}{=} \partial_{\pi_J}Q(\theta\,|\,\theta^t) + \lambda = \pi_J^{(-)} N_J + \lambda \implies \pi_J = (-)\frac{N_J}{\lambda}$$

- To obtain λ, trace over that idenity in the last item $\sum_J$:

$$
\begin{aligned}
(1) &= \sum_{J\in[1,\mathscr{K}]} \pi_J = \sum_{J\in[1,\mathscr{K}]} (-)\frac{N_J}{\lambda} = (-)\frac{1}{\lambda} \sum_{J\in[1,\mathscr{K}]} \Big(\sum_{\alpha\in[1,N]} r_{\alpha J}\Big) \\
&= (-)\frac{1}{\lambda} \sum_{\alpha\in[1,N]} \Big(\sum_{J\in[1,\mathscr{K}]} r_{\alpha J}\Big) = (-)\frac{1}{\lambda} \sum_{\alpha\in[1,N]} \big(1\big) = (-)\frac{1}{\lambda}(N)
\end{aligned}
$$

- WLOG, notice $r_{\alpha J}$ is the poterior PD, which is normalized to 1:

$$
\sum_{J\in[1,\mathscr{K}]} r_{\alpha J} = \sum_{J\in[1,\mathscr{K}]} p(z_\alpha^J \,|\, x, \theta) \stackrel{!}{=} 1
$$

- Therefore, the update rule for π_J:

$$
\lambda = (-)N \implies \pi_J = (-)\frac{N_J}{\lambda} = \frac{N_J}{N}
$$

Derivation

"$\mu_I^{t+1} \leftarrow \frac{1}{N_I}(\sum_{\alpha\in[1,N]} r_{I\alpha}^t x_\alpha)$"?

- Similarly, the update rule for μ_I can be obtained taking a derivative of Q over it:

$$\begin{aligned}\partial_{\mu_I} Q(\theta \,|\, \theta^t) &= \partial_{\mu_I} \sum_{J\in[1,\mathscr{K}]} \sum_{\alpha\in[1,N]} r_{\alpha J} \log(\pi_J \cdot g_{\alpha J}) \\ &= \sum_{J\in[1,\mathscr{K}]} \sum_{\alpha\in[1,N]} r_{\alpha J} \partial_{\mu_I} \log(\pi_J \cdot g_{\alpha J})\end{aligned}$$

- evaluate the log content first, and then evaluate the derivative:

$$\begin{aligned}\partial_{\mu_I} \log(\pi_J \cdot g_{\alpha J}) &= \partial_{\mu_I} \log\left[\mathscr{Z}_J e^{(-)2^{(-)}\mathsf{Tr}\left((x_\alpha-\mu_J)^\intercal \sigma_J^{(-)}(x_\alpha-\mu_J)\right)}\right] \\ &= \partial_{\mu_I}\left[\log \mathscr{Z}_J + (-)2^{(-)}\mathsf{Tr}\left((x_\alpha-\mu_J)^\intercal \sigma_J^{(-)}(x_\alpha-\mu_J)\right)\right] \\ &= \partial_{\mu_I}\Big[(2^{(-)}\dim x_\alpha)\log(2\pi\|\sigma_J\|)^{(-)} \\ &\qquad\qquad + (-)2^{(-)}\mathsf{Tr}\left((x_\alpha-\mu_J)^\intercal \sigma_J^{(-)}(x_\alpha-\mu_J)\right)\Big] \\ &= \partial_{\mu_I}\left[(0) + (-)2^{(-)}\mathsf{Tr}\left((x_\alpha-\mu_J)^\intercal \sigma_J^{(-)}(x_\alpha-\mu_J)\right)\right] \\ &= (-)2^{(-)}\Big((x_{\alpha m}-\mu_{Jm})[\sigma_J^{(-)}]^{mn}(x_{\alpha n}-\mu_{Jn})\Big) \\ &= (x_{\alpha m}-\mu_{Jm})[\sigma_J^{(-)}]^{mn}(\delta_{IJ}\mathbf{1}_n) \\ &= \delta_{IJ}\cdot(x_\alpha-\mu_J)^\intercal \sigma_J^{(-)}\mathbf{1}\end{aligned}$$

- Thus, substitute and evaluate the derivative of Q:

$$\partial_{\mu_I} Q(\theta \mid \theta^t) = \sum_{J\in[1,\mathscr{K}]} \sum_{\alpha\in[1,N]} r_{\alpha J} \partial_{\mu_I} \log(\pi_J \cdot g_{\alpha J})$$
$$= \sum_{J\in[1,\mathscr{K}]} \sum_{\alpha\in[1,N]} r_{\alpha J} \delta_{IJ} \cdot (x_\alpha - \mu_J)^\intercal \sigma_J^{(-)} \mathbf{1}$$
$$= \sum_{\alpha\in[1,N]} r_{\alpha I} (x_\alpha - \mu_I)^\intercal \sigma_I^{(-)} \mathbf{1}$$

- setting to 0 and invert the differential equation:

$$0 \stackrel{!}{=} \partial_{\mu_I} Q(\theta \mid \theta^t) = \sum_{\alpha\in[1,N]} r_{\alpha J} (x_\alpha - \mu_J)^\intercal \sigma_J^{(-)} \mathbf{1}$$
$$= \sum_{\alpha\in[1,N]} r_{\alpha J} x_\alpha^\intercal \sigma_J^{(-)} \mathbf{1} + (-) \sum_{\alpha\in[1,N]} r_{\alpha J} \mu_J^\intercal \sigma_J^{(-)} \mathbf{1}$$

- Therefore, the update rule for μ_J:

$$\sum_{\alpha\in[1,N]} r_{\alpha J} x_\alpha^\intercal \sigma_J^{(-)} \mathbf{1} = \sum_{\alpha\in[1,N]} r_{\alpha J} \mu_J^\intercal \sigma_J^{(-)} \mathbf{1}$$
$$\mu_J^{t+1} \leftarrow \frac{\sum_{\alpha\in[1,N]} r_{\alpha J} x_\alpha}{\sum_{\alpha\in[1,N]} r_{\alpha J}}$$

Derivation

"$\sigma_I^{t+1} \leftarrow \frac{\sum_\alpha r_{\alpha I}\|x_\alpha - \mu_I\|^2}{2\sum_\alpha r_{\alpha I}}$"?

- For ease of notation, denote $\partial_{\sigma_I^{(-)}} \to \partial_I$
- As a rhyme to the two derivations above, differentiate Q over $\sigma_I^{(-)}$:

$$\partial_I Q(\theta \,|\, \theta^t) = \sum_{\alpha,I} r_{\alpha I} \partial_I \left[\log \mathscr{Z}_J^{(-)} + (-)\frac{1}{2} \operatorname{Tr}\left((x_\alpha - \mu_J)^\intercal \sigma_J^{(-)} (x_\alpha - \mu_J)\right)\right]$$
$$= \sum_{\alpha,I} r_{\alpha I} \left[\partial_I \log \mathscr{Z}_J^{(-)} + (-)\frac{1}{2} \partial_I \operatorname{Tr}\left((x_\alpha - \mu_J)^\intercal \sigma_J^{(-)} (x_\alpha - \mu_J)\right)\right]$$

- Applying a well known matrix identity, i.e $\frac{\delta \det g}{\det g} = \delta g g^{(-)}$

$$\partial_I \log \mathscr{Z}_J^{(-)} = \partial_I \log \left(2\pi \|\sigma_J\|\right)^{(-)} = \partial_I \log \left((2\pi)^{(-)} \det \sigma_J^{(-)}\right) = \partial_I \log \left((1) \det \sigma_J^{(-)}\right)$$
$$= \partial_I \log \left(\det \sigma_J^{(-)}\right)^{(-)} = \sigma_J \delta_{IJ} \mathbf{1}_{\zeta\times\zeta} \quad \text{where} \quad \zeta := \dim x_\alpha$$

- Cyclic permute the trace content, and apply $\partial_A \operatorname{Tr}(AB) = B$:

$$(-)\frac{1}{2}\partial_I \operatorname{Tr}\left((x_\alpha - \mu_J)^\intercal \sigma_J^{(-)} (x_\alpha - \mu_J)\right) = (-)\frac{1}{2}\partial_I \operatorname{Tr}\left(\sigma_J^{(-)} (x_\alpha - \mu_J)(x_\alpha - \mu_J)^\intercal\right)$$
$$= (-)\frac{1}{2} \operatorname{Tr}\left(\partial_I \sigma_J^{(-)} (x_\alpha - \mu_J)(x_\alpha - \mu_J)^\intercal\right)$$
$$= (-)\frac{1}{2} \operatorname{Tr}\left(\delta_{IJ} \mathbf{1}_{\zeta\times\zeta} (x_\alpha - \mu_J)(x_\alpha - \mu_J)^\intercal\right)$$
$$= (-)\frac{1}{2} \delta_{IJ} \|x_\alpha - \mu_J\|^2$$

- Finish evaluating the differentiation on Q, set to 0:

$$\begin{aligned}\partial_I Q(\theta \,|\, \theta^t) &= \sum_{\alpha,I} r_{\alpha I} \partial_I \log \mathscr{Z}_J + \sum_{\alpha,I} r_{\alpha I} (-) \frac{1}{2} \partial_I \operatorname{Tr}\left((x_\alpha - \mu_J)^\mathsf{T} \sigma_J^{(-)} (x_\alpha - \mu_J) \right) \\ &= \sum_{\alpha} r_{\alpha I} \sigma_I \mathbf{1}_{\zeta \times \zeta} + \sum_{\alpha} r_{\alpha I} (-) \frac{1}{2} \|x_\alpha - \mu_I\|^2 \overset{!}{=} 0\end{aligned}$$

- Invert the differential equation results in:

$$\frac{1}{2} \sum_{\alpha} r_{\alpha I} \|x_\alpha - \mu_I\|^2 = \sum_{\alpha} r_{\alpha I} \sigma_I$$

- Thus, the update rule on σ_I:

$$\sigma_I^{t+1} \leftarrow \frac{\sum_\alpha r_{\alpha I} \|x_\alpha - \mu_I\|^2}{2 \sum_\alpha r_{\alpha I}}$$

K-means

- Duality K-means $\sim$ GMM: Introduce generalized log-likelihood $\mathscr{L}_\epsilon(x^{\otimes N}, I) \equiv \epsilon^{(-)} \cdot \sum_{\alpha \in (1,\ldots,N)} \log \sum_{I \in (1,\ldots,\mathscr{K})} p(x_\alpha, I)^\epsilon$
  - $\lim_{\epsilon \gg 0} \mathscr{L}_\epsilon(x^{\otimes N}, I) \approx (-)2^{(-)} \sum_{J \in [1,\mathscr{K}]} \sum_{\alpha_J \in [1,N_J]} \|x_\alpha - \mu_J\|^2$
  - Both GMM and K-Means use EM algorithms. (i.e reducing to K-Means by substituting $\sigma_J \sim 1$).
- For *soft clustering (def.?)*, introduce tight breaker r as an indicator function:

$$r_{\alpha I} = \mathbf{1}(\|x_\alpha - \mu_I\|^2 \leq \|x_\alpha - \mu_J\|^2) \quad \forall J \in [1, \mathscr{K}]$$

$$\mathscr{L}_N(x, I) := \sum_{\alpha=[1,N]} \sum_{J=[1,\mathscr{K}]} r_{\alpha I} \|x_\alpha - \mu_I\|^2$$

- algorithm:
  - assign clusters $\{c_I\} \leftarrow \arg\min \|x_\alpha - \mu_I\|^2 \quad \forall \alpha \in [1, N]$
  - $\mu_I \leftarrow \arg\max \mathscr{L}_N(x^{\otimes N}, I) \leftarrow (\sum_{\beta \in [1,N]} r_{\beta I})^{(-)} \cdot \sum_{\alpha \in [1,N]} r_{\alpha I} x_I$
- Aside: Suppose sample average $\bar{x}$: $\arg\max_{\mu_I} \mathscr{L}_N(x, I) \approx \arg\min_{\mu_I} B(x)$ i.e $\sum_{\alpha, I} r_{\alpha I} W_I(x) + N \cdot B(x) \propto N^2 T(x) + \mathcal{O}(x, \mu, \bar{x}) \quad T \perp \{\mu_I\}$

$$T = N^{(-)} \sum_\alpha \|x_\alpha - \bar{x}\| \quad \text{total deviation}$$

$$W_I = (\sum_\alpha r_{\alpha I} \|x_\alpha - \mu_I\|^2)(\sum_\beta r_{\beta I})^{(-)} \quad \text{intra-cluster deviation}$$

$$B = \sum_I N^{(-)} \sum_\alpha r_{\alpha I} \|\mu_I - \bar{x}\|^2 \quad \text{inter-cluster deviation}$$

Derivation

" $\lim_{\epsilon\gg 0}\mathscr{L}_\epsilon(x^{\otimes N}, I) = \sum_{J,\alpha_J}\log p(J, x_{\alpha_J})$

- Note: $\mathscr{L}_\epsilon(x^{\otimes N}, I)$ at $\epsilon\to 1$ corresponds to GMM:

$$\lim_{\epsilon\to 1}\mathscr{L}_\epsilon(x^{\otimes N}, I) \equiv \lim_{\epsilon\to 1}\epsilon^{(-)}\cdot\sum_{\alpha\in[1,N]}\log\sum_{I\in[1,\mathscr{K}]}p(x_\alpha, I)^\epsilon \approx \sum_{\alpha\in[1,N]}\log\sum_{I\in[1,\mathscr{K}]}p(x_\alpha, I)$$

- The promotion $\mathscr{L}(I, x_\alpha)\mapsto\mathscr{L}_\epsilon(I, x_\alpha)$ requires $p(I\,|\,x_\alpha)\mapsto p_\epsilon(I\,|\,x_\alpha)$:

$$p(I\,|\,x_\alpha) = (p(x_\alpha, I))(p(x_\alpha))^{(-)} = (p(x_\alpha, I))(\sum_{J\in[1,\mathscr{K}]}p(x_\alpha, J))^{(-)}$$

$$p_\epsilon(I\,|\,x_\alpha) \equiv (p^\epsilon(x_\alpha, I))(\sum_{J\in[1,\mathscr{K}]}p^\epsilon(x_\alpha, J))^{(-)}$$

- Assume there is a cluster I so that $p(x_\alpha, I) > p(x_\alpha, J)$ for all $J\neq I, J\in[1,\mathscr{K}]$ "E Step" (Why?)

$$\lim_{\epsilon\gg 0}p_\epsilon^{(-)}(I\,|\,x_\alpha) = \lim_{\epsilon\gg 0}(p^\epsilon(x_\alpha, I))^{(-)}(\sum_{J\in[1,\mathscr{K}]}p^\epsilon(x_\alpha, J)) = \lim_{\epsilon\gg 0}\sum_{J\in[1,\mathscr{K}]}\left(\frac{p(x_\alpha, J)}{p(x_\alpha, I)}\right)^\epsilon$$

$$= \lim_{\epsilon\gg 0}\left(1 + \sum_{J\neq I}\left(\frac{p(x_\alpha, J)}{p(x_\alpha, I)}\right)^\epsilon\right) = 1 + \sum_{J\neq I}\lim_{\epsilon\gg 0}\left(\frac{p(x_\alpha, J)}{p(x_\alpha, I)}\right)^\epsilon$$

$$= 1 + \lim_{\epsilon\gg 0}\sum_{J\neq I}(0) = 1$$

- The following items assume some simple limits:

$$\lim_{\epsilon \gg 0} p_\epsilon(I \mid x_\alpha) = 1 \quad \lim_{\epsilon \gg 0} p_\epsilon(I \mid x_\alpha) \log p_\epsilon(I \mid x_\alpha) = 1 \log 1 = 0$$

- In the large ϵ limit, introducing a (1) in the generalized likelihood:

$$\begin{aligned}
\lim_{\epsilon \gg 0} \mathscr{L}_\epsilon(x^{\otimes N}, I) &= \lim_{\epsilon \gg 0} \epsilon^{(-)} \sum_{\alpha \in [1,N]} \log \sum_{I \in [1,\mathscr{K}]} p(x_\alpha, I)^\epsilon \\
&= \lim_{\epsilon \gg 0} \epsilon^{(-)} \sum_{\alpha \in [1,N]} \Big(\sum_{J \in [1,\mathscr{K}]} p_\epsilon(J \mid x_\alpha) \Big) \log \sum_{I \in [1,\mathscr{K}]} p^\epsilon(x_\alpha, I) \\
&= \lim_{\epsilon \gg 0} \epsilon^{(-)} \sum_{\alpha \in [1,N]} \Big(\sum_{J \in [1,\mathscr{K}]} p_\epsilon(J \mid x_\alpha) \Big) \log \frac{p_\epsilon(J, x_\alpha)}{p^\epsilon(J \mid x_\alpha)} \\
&= \lim_{\epsilon \gg 0} \epsilon^{(-)} \sum_{\alpha \in [1,N]} \sum_{J \in [1,\mathscr{K}]} p_\epsilon(J \mid x_\alpha) \log \frac{p_\epsilon(J, x_\alpha)}{p^\epsilon(J \mid x_\alpha)}
\end{aligned}$$

- If for each $J \in \mathscr{K}$, $N(J)$ is not uniform, i.e $N(J) = N_J$, the double sum is replaced by:

$$\sum_{\alpha \in [1,N]} \sum_{J \in [1,\mathscr{K}]} \to \sum_{J \in [1,\mathscr{K}]} \sum_{\alpha_J \in [1,N_J]} = \sum_{J, \alpha_J}$$

- Therefore, with ϵ being very large, the likelihood limits to:

$$\lim_{\epsilon \gg 0} \mathscr{L}_\epsilon(x^{\otimes N}, I) = \lim_{\epsilon \gg 0} \epsilon^{(-)} \sum_{J, \alpha_J} p_\epsilon(J \mid x_{\alpha_J}) \log \frac{p_\epsilon(J, x_{\alpha_J})}{p^\epsilon(J \mid x_{\alpha_J})}$$

- continued:

$$
\begin{aligned}
\lim_{\epsilon \gg 0} \mathscr{L}_\epsilon(x^{\otimes N}, I) &= \lim_{\epsilon \gg 0} \epsilon^{(-)} \sum_{J,\alpha_J} p_\epsilon(J \mid x_{\alpha_J}) \log p_\epsilon(J, x_{\alpha_J}) \\
&\qquad + (-) \lim_{\epsilon \gg 0} \epsilon^{(-)} \sum_{J,\alpha_J} p_\epsilon(J \mid x_{\alpha_J}) \log p^\epsilon(J \mid x_{\alpha_J}) \\
&= \lim_{\epsilon \gg 0} \epsilon^{(-)} \sum_{J,\alpha_J} (1) \log p^\epsilon(J, x_{\alpha_J}) + (-) \lim_{\epsilon \gg 0} \epsilon^{(-)} \sum_{J,\alpha_J} (0) \\
&= \lim_{\epsilon \gg 0} \epsilon^{(-)} \sum_{J,\alpha_J} \log p^\epsilon(J, x_{\alpha_J}) \\
&= \lim_{\epsilon \gg 0} \sum_{J,\alpha_J} \log p^{\epsilon \cdot \epsilon^{(-)}}(J, x_{\alpha_J}) \\
&= \sum_{J,\alpha_J} \log p(J, x_{\alpha_J})
\end{aligned}
$$

Derivation

"$\mathscr{L}^{\mathsf{GMM}}_{\epsilon\gg 0}(I,\theta\,|\,x^{\otimes N}) = (-)2^{(-)}\sum_{I\in[1,\mathscr{K}]}\sum_{\alpha_I\in[1,N_I]}\|x_{\alpha_I}+(-)\mu_I\|^2$"?

- Factorize the joint PD by Bayes' rule, and substitute in Normal Mixtures (GMM) WLOG:

$$p(J,x_{\alpha_J}) = p(J\,|\,x_{\alpha_J},\theta)p(x_{\alpha_J}\,|\,\theta) = \mathscr{Z}_J^{(-)}e^{(-)\frac{1}{2}\mathsf{Tr}\,(x_{\alpha_J}-\mu_J)^{\mathsf{T}}\sigma_J^{(-)}(x_{\alpha_J}-\mu_J)}\pi_J$$

- Substitute the aforementioned item into the likelihood:

$$\begin{aligned}\lim_{\epsilon\gg 0}\mathscr{L}_\epsilon(x^{\otimes N},I) &= \sum_{J,\alpha_J}\mathsf{log}\,\mathscr{Z}_J^{(-)}e^{(-)\frac{1}{2}\mathsf{Tr}\,(x_{\alpha_J}-\mu_J)^{\mathsf{T}}\sigma_J^{(-)}(x_{\alpha_J}-\mu_J)}\pi_J \\ &= \sum_{J,\alpha_J}\mathsf{log}\,\mathscr{Z}_J^{(-)} + \sum_{J,\alpha_J}(-)\frac{1}{2}\mathsf{Tr}\,(x_{\alpha_J}-\mu_J)^{\mathsf{T}}\sigma_J^{(-)}(x_{\alpha_J}-\mu_J)\end{aligned}$$

- The first term above:

$$\mathscr{Z}_J(x_{\alpha_J})^{(-)} = \left((2\pi)^\zeta\det\{\sigma_J\}\right)^{(-)2^{(-)}} \quad \zeta := \dim x_{\alpha_J}$$

$$\sum_{J,\alpha_J}\mathsf{log}\,\mathscr{Z}_J^{(-)} = \sum_{J,\alpha_J}\mathsf{log}\left((2\pi)^\zeta\det\{\sigma_J\}\right)^{(-)2^{(-)}} = (-)\frac{1}{2}\sum_{J,\alpha_J}\mathsf{log}\left((2\pi)^\zeta\det\{\sigma_J\}\right)$$

- Assume that $\sigma_J \stackrel{!}{=} \mathbf{1}_{\zeta\times\zeta}$
- Assume also that $\mathscr{Z}_J^{(-)}(x_{\alpha_J}) \sim (0)$ (WLOG?)

- continued:

$$\lim_{\epsilon \gg 0} \mathscr{L}_\epsilon(x^{\otimes N}, I) \propto (0) + \sum_{J,\alpha_J} (-)\frac{1}{2} \mathsf{Tr}\,(x_{\alpha_J} - \mu_J)^\mathsf{T}(1)(x_{\alpha_J} - \mu_J)$$

$$\propto (-)\frac{1}{2} \sum_{J,\alpha_J} \|x_{\alpha_J} - \mu_J\|^2$$

Derivation

"$\mu_I^{t+1} \leftarrow \sum_{\alpha_I} x_{\alpha_I}$"?

- Only M Step requires derivation.
- For *hard clustering*, $r_{\alpha_I I} \equiv 1$. π_I are stationary:

$$\sum_{\alpha \in [1, N_I]} r_{\alpha_I I} = \sum_{\alpha \in [1, N_I]} (1) = N_I$$

- WLOG: $\sigma_I(x_{\alpha_I}) \sim 1$. The remaining update:

$$\mu_I^{t+1} \leftarrow \underset{\mu_I}{\arg\max} \mathscr{L}_{\epsilon \sim 1}(x^{\otimes N}, I) = N_I^{(-)} \cdot \sum_{i \in [1, N_I]} x_{\alpha_I}$$

Derivation

"$\mu_I \leftarrow N_I^{(-)} \cdot \sum_{\alpha\in[1,\mathscr{K}]} r_{\alpha I} x_I$"? where $N_I \equiv \sum_{\beta\in[1,\mathscr{K}]} r_{\beta I}$"?

- Assume the task: $\mu_I \leftarrow \arg\max_{\mu_I} \mathscr{L}_N(x^{\otimes N}, I)$:

$$0 \stackrel{!}{=} \partial_{\mu_I} \mathscr{L}_N(x^{\otimes N}, I) = \partial_{\mu_I} \sum_{\alpha\in[1,N]} r_{\alpha I} \|x_\alpha - \mu_I\|^2$$
$$= \sum_{\alpha\in[1,N]} r_{\alpha I}(x_\alpha - \mu_I)(-2)$$

- Therefore:

$$\mu_I := \frac{\sum_{\alpha\in[1,N]} r_{\alpha I} x_\alpha}{\sum_{\beta\in[1,N]} r_{\beta I}}$$

Derivation

"$\sum_{\alpha,I} r_{\alpha I} W_I(x) + N \cdot B(x) \propto N^2 T(x) + \mathcal{O}(x,\mu,\bar{x}) \quad T \perp \{\mu_I\}$"?"

- The first piece on the left is equivalent to the Score function.

$$\begin{aligned}
\sum_{\alpha,I} r_{\alpha I} W_I(x) &= \sum_{\alpha\in[1,N]} \sum_{I\in[1,\mathscr{K}]} r_{\alpha I} (\sum_{\gamma\in[1,N]} r_{\gamma I} \|x_\gamma - \mu_I\|^2)(\sum_\beta r_{\beta I})^{(-)} \\
&= \sum_{I\in[1,\mathscr{K}]} \sum_{\gamma\in[1,N]} r_{\gamma I} (\sum_{\alpha\in[1,N]} r_{\alpha I} \|x_\alpha - \mu_I\|^2)(\sum_{\beta\in[1,N]} r_{\beta I})^{(-)} \\
&= \sum_{I\in[1,\mathscr{K}]} \sum_{\gamma\in[1,N]} r_{\gamma I} \|x_\gamma - \mu_I\|^2 = \mathscr{L}_N(x,\mu)
\end{aligned}$$

- A term proportional to T can be factored out from the Left Hand Side:

$$\begin{aligned}
\sum_{\alpha,I} r_{\alpha I} W_I(x) + N \cdot B(x) &= \sum_{I\in[1,\mathscr{K}]} \sum_{\gamma\in[1,N]} r_{\gamma I} \|x_\gamma - \mu_I\|^2 + \frac{N}{N} \cdot \sum_{I,\alpha} r_{\alpha I} \|\mu_I - \bar{x}\|^2 \\
&= \sum_{I\in[1,\mathscr{K}]} \sum_{\alpha\in[1,N]} r_{\alpha I} \left(\|x_\alpha - \mu_I\|^2 + \|\mu_I - \bar{x}\|^2\right) \\
&= \sum_{I\in[1,\mathscr{K}]} \sum_{\alpha\in[1,N]} r_{\alpha I} \left(x_\alpha^2 + (2)\mu_I^2 + \bar{x}^2 + (-2)x_\alpha \mu_I + (-2)\mu_I \bar{x}\right) \\
&= \sum_{I\in[1,\mathscr{K}]} \sum_{\alpha\in[1,N]} r_{\alpha I} \left(\|x_\alpha + (-)\bar{x}\|^2 + (-2)x_\alpha \mu_I + (2)\mu_I^2 + x_\alpha\right) \\
&= \sum_{I\in[1,\mathscr{K}]} \sum_{\alpha\in[1,N]} r_{\alpha I} \|x_\alpha + (-)\bar{x}\|^2 + \mathcal{O}(x_\alpha, \bar{x}, \mu_I)
\end{aligned}$$

- WLOG: "$\sum_{I\in[1,\mathscr{K}]} r_{I\alpha} \overset{?}{=} N \quad \forall r \propto N \times N$"

$$\sum_{\alpha,I} r_{\alpha I} W_I(x) + N \cdot B(x) = \sum_{\alpha\in[1,N]} (N)\|x_\alpha + (-)\bar{x}\|^2 + \mathcal{O}(x_\alpha, \bar{x}, \mu_I)$$
$$= (N)^2 \cdot T(x) + \mathcal{O}(x_\alpha, \bar{x}, \mu_I)$$

- Where $\mathcal{O}(x_\alpha, \bar{x}, \mu_I) \equiv \sum_{I\in[1,\mathscr{K}]} \sum_{\alpha\in[1,N]} r_{I,\alpha}\big(x_\alpha + (-2)x_\alpha\mu_I + (2)\mu_I^2\big)$
- "Bias-Variance Decomp?" in disguise?

Derivation

"$p(z_\alpha^I \,|\, x\theta^t) = \frac{\pi_I g_{I\alpha}}{\sum_{J\in[1,\mathscr{K}]} \pi_J g_{J\alpha}}$"?

- Factorize x into x_β, and note $p(z_\alpha^I \,|\, x_\beta\theta^t) \stackrel{!}{=} 0 \quad \text{if} \quad \alpha \neq \beta$:

$$p(z_\alpha^I \,|\, x\theta^t) = \prod\nolimits_{\beta\in[1,N]} p(z_\alpha^I \,|\, x_\beta\theta^t) = \prod\nolimits_{\beta\in[1,N]} p(z_\alpha^I \,|\, x_\alpha\theta^t)^{\delta_{\alpha\beta}} = p(z_\alpha^I \,|\, x_\alpha\theta^t)$$

- Apply Bayes rule:

$$p(z_\alpha^I \,|\, x_\alpha\theta^t)\, p(x_\alpha \,|\, \theta^t) = p(z_\alpha^I, x_\alpha \,|\, \theta^t) = p(x_\alpha \,|\, z_\alpha^I, \theta^t) p(z_\alpha^I \,|\, \theta^t) = \pi_I g_{\alpha I}$$

- solving for the posterior:

$$p(z_\alpha^I \,|\, x_\alpha\theta^t) = \frac{\pi_I g_{\alpha I}}{p(x_\alpha \,|\, \theta^t)}$$

- For $p(x_\alpha \,|\, \theta^t)$, introduce $\{z_\alpha^J\}$:

$$p(x_\alpha \,|\, \theta^t) = \sum_{J\in[1,\mathscr{K}]} p(x_\alpha z_\alpha^J \,|\, \theta^t) = \sum_{J\in[1,\mathscr{K}]} \pi_J g_{\alpha J}$$

- Therefore:

$$\therefore p(z_\alpha^I \,|\, x\theta^t) = \frac{\pi_I g_{I\alpha}}{\sum_{J\in[1,\mathscr{K}]} \pi_J g_{J\alpha}}$$

Silhouette, AIC, & BIC

- Denote *subcluster* to be $\mathcal{C}_I$. For all points $(x_{\alpha_I}, y_{\alpha_I}) \in \cup_{I\in[1,\mathscr{K}]}\mathcal{C}_I$, **Silhouette Coeffcient** S is the **net discrapancy** (mod normalization) between the intra-cluster mean distances and the *minimal* inter-cluster mean distances:

$$S \equiv \sum_{I=1}^{\mathscr{K}} \sum_{\alpha_I=1}^{\mathscr{N}_I} \hat{s}(y_{\alpha_I}) \quad \text{where} \quad \hat{s}(y_{\alpha_I}) \equiv \frac{\hat{b}_{\alpha_I} - \hat{a}_{\alpha_I}}{\max(\hat{a}_{\alpha_I}, \hat{b}_{\alpha_I})}$$

$$\hat{b}_{\alpha_I} \equiv \min_{J\in\mathcal{C}, J\neq I} \sum_{\beta_J=1}^{\mathscr{N}_J} \frac{d(\alpha_I, \beta_J)}{\mathscr{N}_I - 1} \quad \hat{a}_{\alpha_I} \equiv \sum_{\substack{\beta_I=1 \\ \beta_I\neq\alpha_I}}^{\mathscr{N}_I} \frac{d(\alpha_I\beta_I)}{\mathscr{N}_I}$$

- **AIC**: For $X \equiv \{X_\alpha\}_{\alpha\in[1,\mathscr{N}]}$ random IID, AIC is the **unbiased** estimator of the **true risk** of the log-likelihood $\mathscr{L}_{\mathscr{N}}$ (mod a factor $\times 2$):

$$\text{AIC} = 2\mathscr{K} - 2\max_\theta \mathscr{L}_{\mathscr{N}}$$

- **BIC**: For $X \equiv \{X_\alpha\}_{\alpha\in[1,\mathscr{N}]}$ random IID, BIC is the first order approximation ($\sim \mathcal{O}(\mathscr{N})$) of the **log evidence** $\log(X \,|\, m)$ for all models m_I, $\forall I \in [1, \mathscr{K}]$:

$$\text{BIC} = \mathscr{K} \log \mathscr{N} - \max_\theta \mathscr{L}_{\mathscr{N}}$$

Derivation

"AIC $\stackrel{?}{=} 2\mathscr{K} - 2\mathscr{L}_{\mathscr{N}}$"

- Review MLE
- Consider log-likelihood $\mathscr{L}$ and mean log-likelihood $\overline{\mathscr{L}}$

$$\mathscr{L}(\theta\,|\,x) \equiv \log p(x\,|\,\theta) = \log \prod_{\alpha=1}^{\mathscr{N}} p(x_\alpha\,|\,\theta) = \sum_{\alpha=1}^{\mathscr{N}} \log p(x_\alpha\,|\,\theta)$$

$$\overline{\mathscr{L}}(\theta\,|\,x) \equiv \langle \mathscr{L}(\theta\,|\,x)\rangle_{x\sim p} = \int \mathrm{d}p(x)\,\mathscr{L}(\theta\,|\,x)$$

- goals of MLE are to optimize $\mathscr{L}$ and $\overline{\mathscr{L}}$

$$\hat{\theta} \equiv \operatorname*{arg\,max}_\theta \mathscr{L} \quad \theta^* \equiv \operatorname*{arg\,max}_\theta \overline{\mathscr{L}} \quad \text{i.e} \quad \max_\theta \mathscr{L}(\theta\,|\,x) = \mathscr{L}(\hat{\theta})$$

- Introduce True Risk $\mathfrak{R}(\mathscr{L})$ and Empirical Risk $\hat{\mathfrak{R}}(\mathscr{L})$

$$\mathfrak{R}(\mathscr{L}(\hat{\theta}\,|\,x)) \equiv [-]\mathscr{N}\,\overline{\mathscr{L}}(\hat{\theta}\,|\,x) \quad \hat{\mathfrak{R}}(\mathscr{L}(\hat{\theta}\,|\,x)) \equiv \sum_{\alpha=1}^{\mathscr{N}} [-]\mathscr{L}(\hat{\theta}\,|\,x_\alpha)$$

- $\hat{\mathfrak{R}}$ serves as the estimator for $\mathfrak{R}$
- AIC is $\hat{\mathfrak{R}}$ mod bias. To find the bias of $\mathfrak{R}$

$$\text{bias}(\mathfrak{R}) \equiv \mathbb{E}(\hat{\mathfrak{R}}) - \mathfrak{R}$$

- Approximate $\mathscr{L}(\hat{\theta})$ near the fixed point θ^* via Taylor Expansion

$$\mathscr{L}(\hat{\theta}\,|\,x) \equiv \sum_{\alpha=1}^{\mathscr{N}} \mathscr{L}(\hat{\theta}\,|\,x_\alpha) \quad \mathscr{L}(\hat{\theta}\,|\,x_\alpha) = \log p(\hat{\theta}\,|\,x_\alpha)$$

Derivation

- continued. Taylor expand around $\hat{\theta} \sim \theta^*$

$$\mathscr{L}(\hat{\theta}\,|\,x) \approx \sum_{\alpha=1}^{\mathscr{N}} \mathscr{L}(\theta^*\,|\,x_\alpha) + (\hat{\theta}-\theta^*) \sum_{\alpha=1}^{\mathscr{N}} \Big[\boldsymbol{\nabla}_{\hat{\theta}} \mathscr{L}(\hat{\theta}\,|\,x_\alpha)\Big]^{\mathsf{T}}_{\theta^*=\hat{\theta}}$$
$$+ \frac{1}{2}(\hat{\theta}-\theta^*)^{\mathsf{T}} \left[\sum_{\alpha=1}^{\mathscr{N}} \boldsymbol{\nabla}_{\hat{\theta}} \boldsymbol{\nabla}_{\hat{\theta}} \mathscr{L}(\theta^*\,|\,x)\right]_{\theta^*=\hat{\theta}} (\hat{\theta}-\theta^*)$$

- Used below (since $\hat{\theta} \equiv \arg\max_\theta \mathscr{L}(\theta\,|\,x)$)

$$0 = \boldsymbol{\nabla}_{\hat{\theta}} \mathscr{L}(\theta\,|\,x) = \boldsymbol{\nabla}_{\hat{\theta}} \log \prod\nolimits_{\alpha=1}^{\mathscr{N}} p(\theta\,|\,x_\alpha) = \sum_{\alpha=1}^{\mathscr{N}} \boldsymbol{\nabla}_{\hat{\theta}} \mathscr{L}(\theta\,|\,x_\alpha)$$

- Consider $\mathcal{O}(\theta)$ term

$$\begin{aligned}
&\sum_{\alpha=1}^{\mathscr{N}} (\hat{\theta}-\theta^*) \Big[\boldsymbol{\nabla}_{\hat{\theta}} \mathscr{L}(\hat{\theta}\,|\,x_\alpha)\Big]^{\mathsf{T}}_{\theta^*=\hat{\theta}} \\
&= \sum_{\alpha=1}^{\mathscr{N}} (\hat{\theta}-\theta^*) \Big[\boldsymbol{\nabla}_{\hat{\theta}} \mathscr{L}(\hat{\theta}\,|\,x_\alpha)\Big]^{\mathsf{T}}_{\theta^*=\hat{\theta}} - 0 \\
&= \sum_{\alpha=1}^{\mathscr{N}} (\hat{\theta}-\theta^*) \Big[\boldsymbol{\nabla}_{\hat{\theta}} \mathscr{L}(\hat{\theta}\,|\,x_\alpha)\Big]^{\mathsf{T}}_{\theta^*=\hat{\theta}} - \sum_{\alpha=1}^{\mathscr{N}} (\hat{\theta}-\theta^*) [\boldsymbol{\nabla}_{\hat{\theta}} \mathscr{L}(\theta\,|\,x_\alpha)]^{\mathsf{T}}
\end{aligned}$$

Derivation

- conitnued.

$$
\begin{aligned}
&\sum_{\alpha=1}^{\mathscr{N}}(\hat{\theta}-\theta^*)\left[\boldsymbol{\nabla}_{\hat{\theta}}\mathscr{L}(\hat{\theta}\,|\,x_\alpha)\right]^{\mathsf{T}}_{\theta^*=\hat{\theta}} \\
&=\sum_{\alpha=1}^{\mathscr{N}}(\hat{\theta}-\theta^*)\Big\{\left[\boldsymbol{\nabla}_{\hat{\theta}}\mathscr{L}(\hat{\theta}\,|\,x_\alpha)\right]_{\theta^*=\hat{\theta}}-\boldsymbol{\nabla}_{\hat{\theta}}\mathscr{L}(\theta\,|\,x_\alpha)\Big\}^{\mathsf{T}} \\
&\approx\sum_{\alpha=1}^{\mathscr{N}}(\hat{\theta}-\theta^*)\Big\{\left[\boldsymbol{\nabla}_{\hat{\theta}}\mathscr{L}(\hat{\theta}\,|\,x_\alpha)\right]_{\theta^*=\hat{\theta}}-(\hat{\theta}-\theta^*)^{\mathsf{T}}\left[\boldsymbol{\nabla}_{\hat{\theta}}\mathscr{L}(\hat{\theta}\,|\,x_\alpha)\right]_{\theta^*=\hat{\theta}} \\
&\qquad\qquad-\left[\boldsymbol{\nabla}_{\hat{\theta}}\boldsymbol{\nabla}_{\hat{\theta}}\mathscr{L}(\hat{\theta}\,|\,x_\alpha)\right]_{\theta^*=\hat{\theta}}\Big\}^{\mathsf{T}} \\
&\approx[-](\hat{\theta}-\theta^*)^{\mathsf{T}}\left[\sum_{\alpha=1}^{\mathscr{N}}\boldsymbol{\nabla}_{\hat{\theta}}\boldsymbol{\nabla}_{\hat{\theta}}\mathscr{L}(\hat{\theta}\,|\,x_\alpha)\right]_{\theta^*=\hat{\theta}}(\hat{\theta}-\theta^*)
\end{aligned}
$$

- Take the large $\mathscr{N}$ limit

$$
\lim_{\mathscr{N}\gg 1}\left[\sum_{\alpha=1}^{\mathscr{N}}\boldsymbol{\nabla}_{\hat{\theta}}\boldsymbol{\nabla}_{\hat{\theta}}\mathscr{L}(\hat{\theta}\,|\,x_\alpha)\right]_{\theta^*=\hat{\theta}}\approx\mathscr{N}\left\langle\boldsymbol{\nabla}_{\hat{\theta}}\boldsymbol{\nabla}_{\hat{\theta}}\mathscr{L}(\hat{\theta}\,|\,x_\alpha)\right\rangle_{\theta^*=\hat{\theta}}\approx[-]\mathscr{N}g(\theta^*)
$$

- Thus

$$
\begin{aligned}
\lim_{\mathscr{N}\gg 1}\sum_{\alpha=1}^{\mathscr{N}}(\hat{\theta}-\theta^*)^{\mathsf{T}}\left[\boldsymbol{\nabla}_{\hat{\theta}}\mathscr{L}(\hat{\theta}\,|\,x_\alpha)\right]_{\theta^*=\hat{\theta}}&\approx(\hat{\theta}-\theta^*)^{\mathsf{T}}\mathscr{N}g(\theta^*)(\hat{\theta}-\theta^*) \\
&\approx\mathscr{N}\left\|\hat{\theta}-\theta^*\right\|^2_{g(\theta^*)}
\end{aligned}
$$

Derivation

- Therefore

$$\mathscr{L}(\hat{\theta} \,|\, x) \approx \sum_{\alpha=1}^{\mathscr{N}} \mathscr{L}(\theta^* \,|\, x_\alpha) + \mathscr{N} \left\| \hat{\theta} - \theta^* \right\|_{g(\theta^*)}^2 + [-] \frac{\mathscr{N}}{2} \left\| \hat{\theta} - \theta^* \right\|_{g(\theta^*)}^2$$
$$\approx \mathscr{L}(\theta^* \,|\, x) + \frac{\mathscr{N}}{2} \left\| \hat{\theta} - \theta^* \right\|_{g(\theta^*)}^2$$

- Next, approximate $\overline{\mathscr{L}}(\hat{\theta} \,|\, x)$

$$\overline{\mathscr{L}}(\hat{\theta} \,|\, x) \approx \overline{\mathscr{L}}(\theta^* \,|\, x) + (\hat{\theta} - \theta^*) \left[\boldsymbol{\nabla}_{\hat{\theta}} \overline{\mathscr{L}}(\hat{\theta} \,|\, x) \right]_{\hat{\theta}=\theta^*}^{\mathsf{T}}$$
$$+ \frac{1}{2} (\hat{\theta} - \theta^*)^{\mathsf{T}} \left[\boldsymbol{\nabla}_{\hat{\theta}} \boldsymbol{\nabla}_{\hat{\theta}} \overline{\mathscr{L}}(\hat{\theta} \,|\, x) \right]_{\hat{\theta}=\theta^*} (\hat{\theta} - \theta^*)$$

- The term $\sim \mathcal{O}(\theta)$ correspond with MLE goals

$$(\hat{\theta} - \theta^*) \left[\boldsymbol{\nabla}_{\hat{\theta}} \overline{\mathscr{L}}(\hat{\theta} \,|\, x) \right]_{\hat{\theta}=\theta^*}^{\mathsf{T}} \approx (\hat{\theta} - \theta^*)(0) \approx (0) \quad \because \theta^* = \arg\max \overline{\mathscr{L}}$$

- The term $\sim \mathcal{O}(\theta^2)$ at $\mathscr{N} \gg 1$

$$\lim_{\mathscr{N} \gg 1} \frac{1}{2} (\hat{\theta} - \theta^*)^{\mathsf{T}} \left[\boldsymbol{\nabla}_{\hat{\theta}} \boldsymbol{\nabla}_{\hat{\theta}} \overline{\mathscr{L}}(\hat{\theta} \,|\, x) \right]_{\hat{\theta}=\theta^*} (\hat{\theta} - \theta^*)$$
$$\approx \frac{1}{2} (\hat{\theta} - \theta^*)^{\mathsf{T}} \left[\overline{\boldsymbol{\nabla}_{\hat{\theta}} \boldsymbol{\nabla}_{\hat{\theta}} \mathscr{L}(\hat{\theta} \,|\, x)} \right]_{\hat{\theta}=\theta^*} (\hat{\theta} - \theta^*)$$

Derivation

- continued

$$\lim_{\mathscr{N}\gg 1}\frac{1}{2}(\hat\theta-\theta^*)^\intercal\left[\boldsymbol{\nabla}_{\hat\theta}\boldsymbol{\nabla}_{\hat\theta}\overline{\mathscr{L}}(\hat\theta\,|\,x)\right]_{\hat\theta=\theta^*}(\hat\theta-\theta^*)$$
$$\approx\frac{1}{2}(\hat\theta-\theta^*)^\intercal([-]g(\theta^*))(\hat\theta-\theta^*)\approx[-]\frac{1}{2}\left\|\hat\theta-\theta^*\right\|^2_{g(\theta^*)}$$

- Collecting results for $\mathscr{L}$ and $\overline{\mathscr{L}}$

$$\overline{\mathscr{L}}(\hat\theta\,|\,x)\approx\overline{\mathscr{L}}(\theta^*\,|\,x)+[-]\frac{1}{2}\left\|\hat\theta-\theta^*\right\|^2_{g(\theta^*)}$$
$$\mathscr{L}(\hat\theta\,|\,x)\approx\mathscr{L}(\theta^*\,|\,x)+\frac{\mathscr{N}}{2}\left\|\hat\theta-\theta^*\right\|^2_{g(\theta^*)}$$

- Next, compute risks $\hat{\mathfrak{R}}$ and $\mathfrak{R}$

$$\mathfrak{R}(\mathscr{L}(\hat\theta\,|\,x))=[-]\mathscr{N}\,\mathbb{E}_{\hat\theta}[\overline{\mathscr{L}}]=[-]\mathscr{N}\left(\mathbb{E}_{\hat\theta}\left[\overline{\mathscr{L}}(\theta^*\,|\,x)\right]+[-]\frac{1}{2}\mathbb{E}_{\hat\theta}\left[\left\|\hat\theta-\theta^*\right\|^2_{g(\theta^*)}\right]\right)$$
$$\hat{\mathfrak{R}}(\mathscr{L}(\hat\theta\,|\,x))=\sum_{\alpha=1}^{\mathscr{N}}[-]\mathscr{L}(\hat\theta\,|\,x_\alpha)\approx[-]\left[\mathscr{L}(\theta^*\,|\,x)+\frac{\mathscr{N}}{2}\left\|\hat\theta-\theta^*\right\|^2_{g(\theta^*)}\right]$$

- Evaluate bias

$$\text{bias}\approx\left\langle\hat{\mathfrak{R}}(\mathscr{L}(\hat\theta\,|\,x_\alpha))\right\rangle_{\hat\theta}-\mathfrak{R}(\mathscr{L}(\hat\theta\,|\,x_\alpha))$$
$$\approx\left\langle[-]\left[\mathscr{L}(\theta^*\,|\,x)+\frac{\mathscr{N}}{2}\left\|\hat\theta-\theta^*\right\|^2_{g(\theta^*)}\right]\right\rangle_{\hat\theta}-[-]\mathscr{N}\left(\overline{\mathscr{L}}(\theta^*\,|\,x)+[-]\frac{1}{2}\|\hat\theta-\theta^*\|^2_{g(\theta^*)}\right)$$

Derivation

- continued

$$\text{bias} \approx \langle [-]\mathscr{L}(\theta^* \mid x)\rangle_{\hat\theta} + \left\langle [-]\frac{\mathscr{N}}{2}\left\|\hat\theta - \theta^*\right\|^2_{g(\theta^*)}\right\rangle_{\hat\theta}$$
$$+ \mathscr{N}\mathbb{E}_{\hat\theta}\left[\overline{\mathscr{L}}(\theta^* \mid x)\right] + [-]\mathscr{N}\frac{1}{2}\mathbb{E}_{\hat\theta}\left[\left\|\hat\theta - \theta^*\right\|^2_{g(\theta^*)}\right]$$

- Note: 2nd term=4th term.
- 1st term is similar to 3th term

$$\langle [-]\mathscr{L}(\theta^* \mid x)\rangle_{\hat\theta} \approx [-]\mathbb{E}_{\hat\theta}[\mathscr{L}(\theta^* \mid x)] \approx [-]\mathscr{N}\mathscr{N}^{[-]}\mathbb{E}_{\hat\theta}\left[\sum_{\alpha=1}^{\mathscr{N}}\mathscr{L}(\theta^* \mid x_\alpha)\right]$$
$$\approx [-]\mathscr{N}\mathbb{E}_{\hat\theta}\left[\mathscr{N}^{[-]}\sum_{\alpha=1}^{\mathscr{N}}\mathscr{L}(\theta^* \mid x_\alpha)\right] \approx [-]\mathscr{N}\mathbb{E}_{\hat\theta}\left[\overline{\mathscr{L}}(\theta^* \mid x_\alpha)\right]$$

- Hence

$$\text{bias} \approx [-]\mathscr{N}\mathbb{E}_{\hat\theta}\left[\overline{\mathscr{L}}(\theta^* \mid x_\alpha)\right] + \left\langle [-]\frac{\mathscr{N}}{2}\left\|\hat\theta - \theta^*\right\|^2_{g(\theta^*)}\right\rangle_{\hat\theta}$$
$$\mathscr{N}\mathbb{E}_{\hat\theta}\left[\overline{\mathscr{L}}(\theta^* \mid x_\alpha)\right] + [-]\mathscr{N}\frac{1}{2}\mathbb{E}_{\hat\theta}\left[\left\|\hat\theta - \theta^*\right\|^2_{g(\theta^*)}\right]$$
$$\approx (0) + [-]\mathscr{N}\mathbb{E}_{\hat\theta}\left[\left\|\hat\theta - \theta^*\right\|^2_{g(\theta^*)}\right]$$

Derivation

- By definition of Central Limit Theorem (CLT) and Law of Large Numbers (LLN) IID $\hat{\theta}$ distribute as Gaussians, and $(\hat{\theta} - \theta^*)^\mathsf{T} \mathscr{N} g(\theta^*)(\hat{\theta} - \theta^*)$ as χ^2 (prove it elsewhere?):

$$\sqrt{\mathscr{N}}(\hat{\theta} - \theta^*) \sim \mathcal{N}(0, g^{[-]}(\theta^*)) \quad \mathscr{N}(\hat{\theta} - \theta^*)^\mathsf{T} g(\theta^*)(\hat{\theta} - \theta^*) \sim \chi^2_{\mathscr{K}}$$
$$\mathbb{E}_{\hat{\theta}}\Big[\mathscr{N}(\hat{\theta} - \theta^*)^\mathsf{T} g(\theta^*)(\hat{\theta} - \theta^*)\Big] \sim \mathscr{K}$$

- Thus, bias of $\hat{\mathfrak{R}}$ is

$$\text{bias} \approx [-]\mathscr{K}$$

- AIC is $\times 2$ the empirical risk (mod the bias)

$$\text{AIC} = 2(\hat{\mathfrak{R}}(\mathscr{L}(\hat{\theta} \mid x)) - \text{bias}) = 2\mathscr{K} - 2\mathscr{L}(\hat{\theta} \mid x)$$

Derivation

"$(\hat{\theta}-\theta^*)^{\mathsf{T}}\mathscr{N} g(\theta^*)(\hat{\theta}-\theta^*) \sim \chi^2_{\mathscr{K}}$"

- Useful fact: Since Fisher g is real, symmetric (and semi-definite), a rotation S can diagonalize it.
- For IID $\hat{\theta}$, Central Limit Theorem implies: $\sqrt{\mathscr{N}}(\hat{\theta}-\theta^*) \sim \mathcal{N}(0, g(\theta^*))$ (proof?)

$$\begin{aligned}
G &\equiv \int \mathrm{d}\hat{\theta}\, \mathscr{Z}^{[-]} e^{[-]\frac{1}{2}\|\hat{\theta}-\theta^*\|^2_{g(\theta^*)}} = \mathscr{Z}^{[-]} \int \mathrm{d}(\hat{\theta}-\theta^*)\, e^{[-]\frac{1}{2}\|\hat{\theta}-\theta^*\|^2_{g(\theta^*)}} \\
&= \mathscr{Z}^{[-]} \prod_{M=1}^{\mathscr{K}} \int \mathrm{d}(\hat{\theta}_M-\theta^*_M)\, e^{[-]\frac{1}{2}(\hat{\theta}_I-\theta^*_I)(\hat{\theta}_J-\theta^*_J) g(\theta^*)_{IJ}} \\
&= \mathscr{Z}^{[-]} \prod_{M=1}^{\mathscr{K}} \int \mathrm{d}(\hat{\theta}_M-\theta^*_M)\, e^{[-]\frac{1}{2}(\hat{\theta}_I-\theta^*_I)(\hat{\theta}_J-\theta^*_J)(S^{\mathsf{T}}\mathbf{diag}(g(\theta^*))S)_{IJ}} \\
&= \mathscr{Z}^{[-]} \prod_{M=1}^{\mathscr{K}} \int \mathrm{d}(\hat{\theta}_M-\theta^*_M)\, e^{[-]\frac{1}{2}(\hat{\theta}_I-\theta^*_I)(\hat{\theta}_J-\theta^*_J)(S^{\mathsf{T}}_{IL}\mathbf{diag}(g(\theta^*))_{LN} S_{NJ})} \\
&= \mathscr{Z}^{[-]} \prod_{M=1}^{\mathscr{K}} \int \mathrm{d}(\hat{\theta}_M-\theta^*_M)\, e^{[-]\frac{1}{2}\sum_{I=1}^{\mathscr{K}}[S_{IJ}(\hat{\theta}_J-\theta^*_J)]^2} \\
&= \mathscr{Z}^{[-]} \prod_{M=1}^{\mathscr{K}} \left(\int \mathrm{d}(\hat{\theta}_M-\theta^*_M)\, e^{[-]\frac{1}{2}\mathbf{diag}(g(\theta^*))_{MM}[S_{MN}(\hat{\theta}_N-\theta^*_N)]^2} \right)
\end{aligned}$$

- Change of variable $\sqrt{\mathrm{diag}(g(\theta^*))_{MM}}S_{MN}(\hat{\theta}_N-\theta^*_N) \to \phi_M$

$$G = \prod_{M=1}^{\mathscr{K}} \left(\int \mathrm{d}(\hat{\theta}_M-\theta^*_M)\, e^{[-]\frac{1}{2}\sum_{M=1}^{\mathscr{K}}\phi_M^2} \right)$$

Derivation

- Transform integration measures $(\hat{\theta}_M - \theta^*_M), \forall M \in [1, \mathscr{K}]$
$$\int \mathrm{d}(\hat{\theta}_M - \theta^*_M) = \frac{1}{\sqrt{\mathsf{diag}(g(\theta^*))_{MM}}} \int \mathrm{d}(S^{\mathsf{T}}_{NM}\phi_N) = \int \mathsf{det}(S^{\mathsf{T}})\, \mathrm{d}\phi_N = \int \mathrm{d}\phi_N$$
- Whence S is orthonormal (i.e.$\|S\|^2 = S^{\mathsf{T}}S = 1 = SS^{\mathsf{T}}$), and as such $\mathsf{det}S = 1 = \frac{1}{\mathsf{det}S^{\mathsf{T}}} = \mathsf{det}S^{\mathsf{T}}$
- Recollect $\mathscr{Z}^{[-]} = \frac{\sqrt{\mathsf{det}(\mathscr{N}\, g(\theta^*))}}{(2\pi)^{\frac{\mathscr{K}}{2}}}$ (Jaynes [1] App E)
- $$\begin{aligned} G &= \prod_{M=1}^{\mathscr{K}} \left(\int \mathrm{d}(\hat{\theta}_M - \theta^*_M)\, e^{[-]\frac{1}{2}\sum_{M=1}^{\mathscr{K}} \phi^2_M} \right) \\ &= \frac{1}{\sqrt{\mathsf{diag}(\mathscr{N}\, g(\theta^*))_{MM}}\,\mathscr{Z}} \prod_{M=1}^{\mathscr{K}} \left(\int \mathrm{d}\phi_M\, e^{[-]\frac{1}{2}\sum_{M=1}^{\mathscr{K}} \phi^2_M} \right) \\ &= \frac{\sqrt{\mathsf{diag}(\mathscr{N}\, g(\theta^*))_{MM}}}{\sqrt{\mathsf{diag}(\mathscr{N}\, g(\theta^*))_{MM}}(2\pi)^{\frac{\mathscr{K}}{2}}} \prod_{M=1}^{\mathscr{K}} \left(\int \mathrm{d}\phi_M\, e^{[-]\frac{1}{2}\sum_{M=1}^{\mathscr{K}} \phi^2_M} \right) \\ &= \frac{1}{(2\pi)^{\frac{\mathscr{K}}{2}}} \prod_{M=1}^{\mathscr{K}} \left(\int \mathrm{d}\phi_M\, e^{[-]\frac{1}{2}\sum_{M=1}^{\mathscr{K}} \phi^2_M} \right) \end{aligned}$$
- Change of variable $\omega^2 := \phi_M \phi^M = \sum_{M=1}^{\mathscr{K}} \phi^2_M$ (i.e Path to polar coordinate)
$$\prod_{M=1}^{\mathscr{K}} \int \mathrm{d}\phi_M = \prod_{M=1}^{\mathscr{K}} \int_{\mathbb{R}} \mathrm{d}\big[\sqrt{\mathscr{N}} S(\hat{\theta} - \theta^*)\big]_M = \int_{\mathbb{R}^+} \mathrm{d}\omega\, A_{\mathscr{K}-1}$$

[1] "Probability Theory: The Logic of Science" by E. Jaynes

Derivation

- Where $A_{\mathscr{K}-1} \equiv$ Surface Area of $(\mathscr{K}-1)$-Shell $= \frac{2\pi^{\mathscr{K}/2}}{\Gamma(\mathscr{K}/2)}\omega^{\mathscr{K}-1}$

$$G = \frac{1}{(2\pi)^{\frac{\mathscr{K}}{2}}} \int_{\mathbb{R}+} \mathrm{d}\omega\, A_{\mathscr{K}-1} e^{[-]\frac{1}{2}\omega^2} = \frac{1}{(2\pi)^{\frac{\mathscr{K}}{2}}} \int_{\mathbb{R}+} \mathrm{d}\omega\, \frac{2\pi^{\mathscr{K}/2}}{\Gamma(\mathscr{K}/2)} \omega^{\mathscr{K}-1} e^{[-]\frac{1}{2}\omega^2}$$

- Change variable $\omega^2 := \zeta$

$$\begin{aligned}
G &:= \frac{1}{(2\pi)^{\frac{\mathscr{K}}{2}}} \int_{\mathbb{R}+} \mathrm{d}\sqrt{\zeta}\, \frac{2\pi^{\mathscr{K}/2}}{\Gamma(\mathscr{K}/2)} \zeta^{\frac{1}{2}(\mathscr{K}-1)} e^{[-]\frac{1}{2}\zeta} \\
&= \frac{2\pi^{\mathscr{K}/2}}{\Gamma(\mathscr{K}/2)(2\pi)^{\frac{\mathscr{K}}{2}}} \int_{\mathbb{R}+} \frac{\mathrm{d}\zeta}{2\sqrt{\zeta}} \zeta^{\frac{1}{2}(\mathscr{K}-1)} e^{[-]\frac{1}{2}\zeta} \\
&= \frac{1}{\Gamma(\mathscr{K}/2)(2)^{\frac{\mathscr{K}}{2}}} \int_{\mathbb{R}+} \mathrm{d}\zeta\, \zeta^{\frac{\mathscr{K}}{2}-1} e^{[-]\frac{1}{2}\zeta} \\
&= \int_{\mathbb{R}+} \mathrm{d}\zeta\, \frac{\zeta^{\frac{\mathscr{K}}{2}-1} e^{[-]\frac{1}{2}\zeta}}{\Gamma(\mathscr{K}/2)(2)^{\frac{\mathscr{K}}{2}}} \\
&\equiv \int \mathrm{d}\zeta\, \chi^2[\mathscr{K}]
\end{aligned}$$

- Thus, chi-squared of degree $\mathscr{K}$ is

$$\therefore \chi^2[\mathscr{K}] \equiv \frac{\zeta^{\frac{\mathscr{K}}{2}-1} e^{[-]\frac{1}{2}\zeta}}{\Gamma(\mathscr{K}/2)(2)^{\frac{\mathscr{K}}{2}}}$$

Derivation

"$\mathbb{E}_{\hat{\theta}}\Big[(\hat{\theta}-\theta^*)^\intercal \mathscr{N} g(\theta^*)(\hat{\theta}-\theta^*)\Big] \sim \mathscr{K}$"

- Introduce χ^2 of degree 1

$$\chi^2[\mathscr{K}=1] \equiv \frac{1}{\sqrt{2\pi\omega e^{\omega}}}$$

- Compute the moment generator $\mathfrak{m}$ for a Γ distribution, on $\omega \in [-\infty, \infty]$ (not dom$(\chi^2) = (0, \infty)$)

$$\mathfrak{m} \equiv \int_{\mathbb{R}} \mathrm{d}\omega \, e^{t\omega} \chi^2(\omega) = \int_{\mathbb{R}} \mathrm{d}\omega \, e^{t\omega} \frac{1}{\sqrt{2\pi\omega e^{\omega}}}$$

- Change of variable $\omega \mapsto \tilde{\omega}^2$

$$\begin{aligned}\mathfrak{m} &= 2 \cdot \frac{1}{\sqrt{2\pi}} \int_0^\infty \mathrm{d}\tilde{\omega}^2 \, e^{t\tilde{\omega}^2} \frac{1}{\sqrt{\tilde{\omega}^2 e^{\tilde{\omega}^2}}} = 2 \cdot \frac{1}{\sqrt{2\pi}} \int_0^\infty \mathrm{d}\tilde{\omega} \, e^{(-)\frac{1}{2}(1-2t)\tilde{\omega}^2} \\ &= 2 \cdot \frac{1}{\sqrt{2\pi}} \frac{1}{2} \sqrt{\frac{\pi}{\frac{1}{2}(1-2t)}} = \sqrt{\frac{\pi}{(1-2t)}}\end{aligned}$$

Derivation cont

- For a random IID $\rho \equiv \sum_{I=1}^{\mathscr{K}} \omega_I$ such that $\rho \sim p(\rho) \equiv \prod_{I=1}^{\mathscr{K}} p(\omega_I)$

$$\mathfrak{m}_{\mathscr{K}} = \int \mathrm{d}\rho_{\mathbb{R}+}\, e^{t\rho} p(\rho) = \int \mathrm{d}\omega_1 \ldots \int \mathrm{d}\omega_{\mathscr{K}}\, e^{t\sum_{J=1}^{\mathscr{K}} \omega_J} \prod_{I=1}^{\mathscr{K}} p(\omega_I)$$
$$= \prod_{J=1}^{\mathscr{K}} \left[\int \mathrm{d}\omega_J\, e^{t\omega_J} p(\omega_J) \right] = \left[\int \mathrm{d}\omega_1\, e^{t\omega_1} p(\omega_{J=1}) \right]^{\mathscr{K}} = \mathfrak{m}^{\mathscr{K}}$$

- Apply on χ^2

$$\mathfrak{m}_{\mathscr{K}} = \left[\int \mathrm{d}\omega_1\, e^{t\omega_1} \chi^2 \right]^{\mathscr{K}} = \left[(1-2t) \right]^{(-)\frac{\mathscr{K}}{2}}$$

- Useful fact

$$\left[\frac{\partial}{\partial \xi} \mathfrak{m}_\xi [\chi^2(X)] \right]_{\xi=0} = \left[\frac{\partial}{\partial \xi} \int \mathrm{d}x\, e^{\xi x} \chi^2 \right]_{\xi=0} = \left[\int \mathrm{d}x\, \frac{\partial}{\partial \xi} e^{\xi x} \chi^2 \right]_{\xi=0} = \left[\int \mathrm{d}x\, x e^{\xi x} \chi^2 \right]_{\xi=0}$$
$$= \mathbb{E}_{x \sim \chi^2}(X)$$

- Let $x \equiv \omega = \|\hat{\theta} - \theta^*\|^2_{\mathscr{N} g(\theta^*)}$

$$\mathbb{E}_{x\sim\chi^2}(X) = \left\langle \|\hat{\theta} - \theta^*\|^2_{\mathscr{N} g(\theta^*)} \right\rangle = \left[\frac{\partial}{\partial \xi} \mathfrak{m}_\xi \left[\chi^2 (x = \|\hat{\theta} - \theta^*\|^2_{\mathscr{N} g(\theta^*)}) \right] \right]_{\xi=0}$$
$$= \left[\frac{\partial}{\partial \xi} \mathfrak{m}_\xi \left[\chi^2 (x = \|\hat{\theta} - \theta^*\|^2_{\mathscr{N} g(\theta^*)}) \right] \right]_{\xi=0} = \left[\frac{\partial}{\partial \xi} (1-2t)^{(-)\frac{1}{2}\mathscr{K}} \right]_{\xi=0}$$
$$= \left[\mathscr{K} (1-2\xi)^{(-)\frac{3}{2}\mathscr{K}} \right]_{\xi=0} = \mathscr{K}$$

Derivation

"BIC $= \mathscr{K}\log\mathscr{N} - \mathscr{L}(\hat\theta\,|\,x)$"

- Assume models $m_J, \forall J \in [1, \mathscr{K}]$ *Model Probability* defines to be

$$p(m_J\,|\,x) = \frac{p(m_J, x)}{p(x)} = \int \mathrm{d}\theta_J\, \frac{p(m_J, \theta_J, x)}{p(x)} = \int \mathrm{d}\theta_J\, \frac{p(x\,|\,m_J, \theta_J)p(\theta_J\,|\,m_J)p(m_J)}{p(x)}$$
$$= \frac{p(m_J)}{p(x)} \int \mathrm{d}\theta_J\, p(x\,|\,m_J, \theta_J)p(\theta_J\,|\,m_J) = \frac{p(m_J)}{p(x)} \int \mathrm{d}\theta_J\, \mathscr{L}(\hat\theta)$$

- Let $p(m_J\,|\,x)$ assumes the role of likelihood. Energy E defines to be

$$E = [-]\log p(m_J\,|\,x) = [-]\log\frac{p(m_J)}{p(x)}\int \mathrm{d}\theta_J\, \mathscr{L}(\hat\theta) = [-]\log\frac{p(m_J)}{p(x)} + [-]\log\int \mathrm{d}\theta_J\, \mathscr{L}(\hat\theta)$$

- Log-Likelihood *completed* with $\{\theta_J\}$ defines to be

$$\mathscr{L}(\theta_J) \equiv \log p(x\,|\,m_J\theta_J)$$

- Taylor expand around $\sim \hat\theta_J$

$$\mathscr{L}(\theta_J) \approx \mathscr{L}(\hat\theta_J) + (\theta_J - \hat\theta_J)^\intercal\Big(\boldsymbol{\nabla}_{\theta_J}\mathscr{L}(\theta_J)\Big)_{\theta=\hat\theta} + \frac{1}{2}(\theta_J - \hat\theta_J)^\intercal\Big(\boldsymbol{\nabla}_{\theta_J}\boldsymbol{\nabla}_{\theta_J}\mathscr{L}(\theta_J)\Big)_{\theta=\hat\theta}(\theta_J - \hat\theta_J)$$
$$\approx \mathscr{L}(\hat\theta_J) + (\theta_J - \hat\theta_J)^\intercal(0) + \frac{1}{2}(\theta_J - \hat\theta_J)^\intercal\Big(\boldsymbol{\nabla}_{\theta_J}\boldsymbol{\nabla}_{\theta_J}\mathscr{L}(\theta_J)\Big)_{\theta=\hat\theta}(\theta_J - \hat\theta_J)$$

- Where $\boldsymbol{\nabla}_{\theta_J}\mathscr{L}(\theta_J) = 0$ for $\hat\theta_J \equiv \arg\max_\theta \mathscr{L}$

Derivation cont

- Term $\sim \mathcal{O}(\theta^2)$

$$\begin{aligned}\Big(\boldsymbol{\nabla}_{\theta_J}\boldsymbol{\nabla}_{\theta_J}\mathscr{L}(\theta_J)\Big)_{\theta=\hat{\theta}} &= \Big[\frac{\mathscr{N}}{\mathscr{N}}\boldsymbol{\nabla}_{\theta_J}\boldsymbol{\nabla}_{\theta_J}\mathscr{L}(\theta_J)\Big]_{\theta=\hat{\theta}} = \Big[\frac{\mathscr{N}}{\mathscr{N}}\sum_{\alpha=1}^{\mathscr{N}}\boldsymbol{\nabla}_{\theta_J}\boldsymbol{\nabla}_{\theta_J}\mathscr{L}(\theta_J)\Big]_{\theta=\hat{\theta}} \\ &= \mathscr{N}\Big[\frac{1}{\mathscr{N}}\sum_{\alpha=1}^{\mathscr{N}}\boldsymbol{\nabla}_{\theta_J}\boldsymbol{\nabla}_{\theta_J}\mathscr{L}(\theta_J)\Big] \approx \mathscr{N}\Big[\big\langle\boldsymbol{\nabla}_{\theta_J}\boldsymbol{\nabla}_{\theta_J}\mathscr{L}(\theta_J)\big\rangle_{x\sim p(x\theta)}\Big]_{\theta=\hat{\theta}} \\ &\approx [-]\mathscr{N}\Big[g(\theta)\Big]_{\theta=\hat{\theta}} \approx [-]\mathscr{N}g(\hat{\theta})\end{aligned}$$

- Thus, log-likelihood around $\hat{\theta}_J$ approximates to be

$$\mathscr{L}(\theta_J) \approx \mathscr{L}(\hat{\theta}_J) + \frac{\mathscr{N}}{2}(\theta-\hat{\theta})^{\intercal}g(\hat{\theta})(\theta-\hat{\theta}) \approx \mathscr{L}(\hat{\theta}_J) + [-]\frac{1}{2}\|\theta-\hat{\theta}\|^2_{g(\hat{\theta})}$$

- Exponentiate that approximation $\mathscr{L}$ and put it into E

$$\begin{aligned}[-]\log p(m_J\,|\,x) &\approx [-]\log\frac{p(m_J)}{p(x)} + [-]\log\int \mathrm{d}\theta_J\, e^{\mathscr{L}(\hat{\theta}_J)+[-]\frac{1}{2}\|\theta-\hat{\theta}\|^2_{g(\hat{\theta})}} \\ &\approx [-]\log\frac{p(m_J)}{p(x)} + [-]\log\Big[e^{\mathscr{L}(\hat{\theta}_J)}\int \mathrm{d}\theta_J\, e^{[-]\frac{1}{2}\|\theta-\hat{\theta}\|^2_{g(\hat{\theta})}}\Big] \\ &\approx [-]\log\frac{p(m_J)}{p(x)} + [-]\mathscr{L}(\hat{\theta}_J) + [-]\log\int \mathrm{d}\theta_J\, e^{[-]\frac{1}{2}\|\theta-\hat{\theta}\|^2_{g(\hat{\theta})}}\end{aligned}$$

Derivation cont

- For Multivariate Gaussian

$$\int \mathrm{d}\theta_J \, e^{[-]\frac{1}{2}\|\theta-\hat{\theta}\|^2_{g(\hat{\theta})}} = \frac{1}{\sqrt{2\pi \mathsf{det}(\mathscr{N} g(\hat{\theta}))}^{\mathscr{K}}} = \frac{1}{\sqrt{2\pi \mathscr{N} \mathsf{det}(g(\hat{\theta}))}^{\mathscr{K}}}$$

- Energy E approximates to be

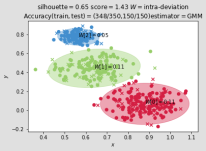
$$\begin{aligned} E &\approx [-]\log\frac{p(m_J)}{p(x)} + [-]\mathscr{L}(\hat{\theta}_J) + [-]\log\frac{1}{\sqrt{2\pi\mathscr{N}\mathsf{det}(g(\hat{\theta}))}^{\mathscr{K}}} \\ &\approx [-]\log\frac{p(m_J)}{p(x)} + [-]\mathscr{L}(\hat{\theta}_J) + [-]\log\frac{1}{\sqrt{2\pi\mathsf{det}(g(\hat{\theta}))}^{\mathscr{K}}} + [-]\log\frac{1}{\sqrt{\mathscr{N}}^{\mathscr{K}}} \end{aligned}$$

- Keep terms $\sim \mathcal{O}(\mathscr{N})$. Disregard the rest insensitive, or subleading, to batch size $\mathscr{N}$

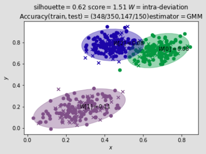
$$\mathsf{BIC} = [-]\mathscr{L}(\hat{\theta}_J) + [-]\log\frac{1}{\sqrt{\mathscr{N}}^{\mathscr{K}}} = \mathscr{K}\log\mathscr{N} + [-]\mathscr{L}(\hat{\theta}_J)$$

Case Study: GMM

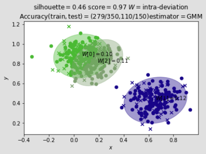
- 4 types of Gaussian clusters estimated by GMM [2] (with #labels = fixed.)



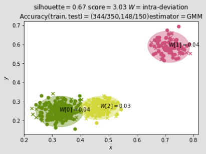
(a) Sizes



(b) Slender

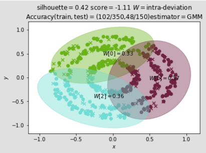


(c) Spherical

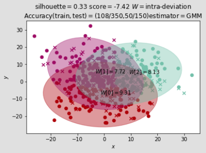


(d) Populations [3]

- 2 cases where GMM fails to cluster correctly.



(e) Rings

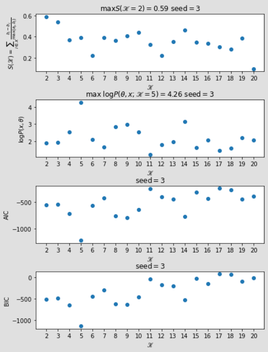


(f) Blur

[2] https://github.com/scikit-learn/ scikit-learn/tree/main/sklearn
[3] #points for each sub-cluster $\mathcal{C}_I$ are distributed by random weights.

Case Study: GMM continued

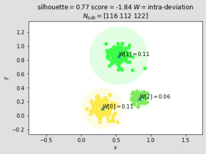
- Evaluating different scores on a parameter set of models $\mathscr{K} \equiv$ #initial clusters.



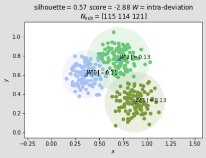
- seed $\equiv$ #clusters used for data generation

Case Study: K-Means

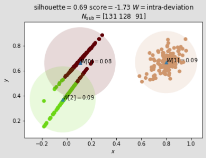
- K-means cluster different geometries of Gaussian blobs.



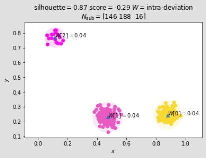
(g) Sizes



(h) Identical

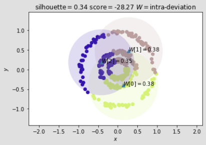


(i) Slender

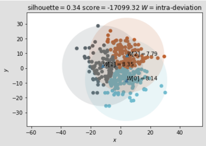


(j) Populations

- K-means failed to cluster rings (Gaussian error bars) and blurry blobs.



(k) Rings



(l) Blur