

Parametric Manifold Learning of Gaussian Mixture Models - Supplementary

Ziquan Liu¹, Lei Yu¹, Janet H. Hsiao², and Antoni B. Chan¹

¹Department of Computer Science, City Univerisity of Hong Kong

²Department of Psychology, The University of Hong Kong

1 Learning objective and variational approximation

The reconstruction parameters $\Theta = \{\mathbf{a}_r, \{\mathbf{m}_{rl}\}_{l=1}^{d_z}, \{\mathbf{C}_{rl}\}_{l=1}^{d_y}, \mathbf{b}_r, \beta_r\}_{r=1}^{K_m}$ and latent variables for the training data $\Omega = \{\mathbf{w}_i, \mathbf{z}_i, \mathbf{y}_i\}_{i=1}^N$ are obtained by minimizing the reconstruction loss between $p_i(\mathbf{x})$ and $\hat{p}_i(\mathbf{x})$, measured with the Kullback-Leibler divergence (KLD; Kullback (1997)),

$$\{\Theta^*, \Omega^*\} = \underset{\Theta, \Omega}{\operatorname{argmin}} \sum_{i=1}^N D_{KL}(p_i \| \hat{p}_i), \quad (1)$$

where $D_{KL}(p \| q) = \int p(x) \log \frac{p(x)}{q(x)} dx$ is the KLD between p and q . The KL objective function $J_{KL}(\Theta, \Omega)$ can be decomposed as

$$\begin{aligned} J_{KL}(\Theta, \Omega) &= \sum_{i=1}^N D_{KL}(p_i \| \hat{p}_i) \\ &= \sum_{i=1}^N \int p_i(\mathbf{x}) \log p_i(\mathbf{x}) d\mathbf{x} - \int p_i(\mathbf{x}) \log \hat{p}_i(\mathbf{x} | \Theta, \Omega_i) d\mathbf{x}, \end{aligned} \quad (2)$$

where the first term is a constant w.r.t. $\{\Theta, \Omega\}$. Hence, minimizing J_{KL} is equivalent to minimizing the second term (i.e., cross-entropy),

$$\begin{aligned} J_{CE}(\Theta, \Omega) &= - \sum_{i=1}^N \int p_i(\mathbf{x}) \log \hat{p}_i(\mathbf{x} | \Theta, \Omega_i) d\mathbf{x} \\ &= - \sum_{i=1}^N \int \sum_{k=1}^{K_b} \pi_{ik} \mathcal{N}_{ik}(\mathbf{x}) \log \sum_{r=1}^{K_m} \hat{\pi}_{ir} \hat{\mathcal{N}}_{ir}(\mathbf{x}) d\mathbf{x}, \end{aligned} \quad (3)$$

where $\mathcal{N}_{ik}(\mathbf{x})$ and $\hat{\mathcal{N}}_{ir}(\mathbf{x})$ are the original and reconstructed Gaussian components, i.e. $\mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik})$ and $\mathcal{N}(\mathbf{x} | \hat{\boldsymbol{\mu}}_{ir}, \hat{\boldsymbol{\Sigma}}_{ir})$.

Since the cross-entropy between two GMMs is intractable, we derive a variational upper-bound as an approximation inspired by Yu et al. (2018). Introducing a set of variational parameters $q_{kr}^{(i)}$, the cross entropy J_{CE} is

rewritten as

$$\begin{aligned}
J_{CE}(\Theta, \Omega) &= \\
&= - \sum_{i=1}^N \int \sum_{k=1}^{K_b} \pi_{ik} \mathcal{N}_{ik}(\mathbf{x}) \log \sum_{r=1}^{K_m} \frac{\hat{\pi}_{ir} \hat{\mathcal{N}}_{ir}(\mathbf{x})}{q_{kr}^{(i)}} q_{kr}^{(i)} d\mathbf{x} \\
&\leq - \sum_{i=1}^N \int \sum_{k=1}^{K_b} \pi_{ik} \mathcal{N}_{ik}(\mathbf{x}) \sum_{r=1}^{K_m} q_{kr}^{(i)} \log \frac{\hat{\pi}_{ir} \hat{\mathcal{N}}_{ir}(\mathbf{x})}{q_{kr}^{(i)}} d\mathbf{x} \\
&= - \sum_{i=1}^N \sum_{k=1}^{K_b} \pi_{ik} \sum_{r=1}^{K_m} q_{kr}^{(i)} \left[\log \frac{\hat{\pi}_{ir}}{q_{kr}^{(i)}} + \mathbb{E}_{\mathbf{x} \sim \mathcal{N}_{ik}} [\log \hat{\mathcal{N}}_{ir}(\mathbf{x})] \right] \\
&\triangleq \tilde{J}_{CE}(\Theta, \Omega, \mathbf{q}),
\end{aligned}$$

2 Formulation using Virtual Samples

Suppose that we have N_v i.i.d virtual samples from the ground truth distribution $p_i(\mathbf{x})$, $\mathbf{X}_i = \{\mathbf{x}_1^{(i)}, \dots, \mathbf{x}_{N_v}^{(i)}\}$. The reconstruction parameters are $\Theta = \{\mathbf{a}_r, \{\mathbf{m}_{rl}\}_{l=1}^{d_z}, \{\mathbf{C}_{rl}\}_{l=1}^{d_y}, \mathbf{b}_r, \beta_r\}_{r=1}^{K_m}$ and latent variables for the training data are $\Omega = \{\mathbf{w}_i, \mathbf{z}_i, \mathbf{y}_i\}_{i=1}^N$. We can obtain the approximation distribution $\hat{p}_i(\mathbf{x})$ by minimizing the cross entropy between $p_i(\mathbf{X}_i)$ and $\hat{p}_i(\mathbf{X}_i)$,

$$J_{CE}(\Theta, \Omega) = - \sum_{i=1}^N \mathbb{E}_{\mathbf{X}_i \sim p_i} [\log \hat{p}_i(\mathbf{X}_i | \Theta, \Omega)] \quad (4)$$

$$= - \sum_{i=1}^N \sum_k \pi_{ik} \mathbb{E}_{\mathbf{X}_i \sim p_{ik}} [\log \hat{p}_i(\mathbf{X}_i | \Theta, \Omega)] \quad (5)$$

where p_{ik} is the k th component in p_i . The expectation is not tractable so we induce variational parameters $q_{kr}^{(i)}$

$$- \mathbb{E}_{\mathbf{X}_i \sim p_{ik}} [\log \hat{p}_i(\mathbf{X}_i | \Theta, \Omega)] \quad (6)$$

$$\leq - \sum_r q_{kr}^{(i)} \left\{ \log \frac{\hat{\pi}_{ir}}{q_{kr}^{(i)}} + \mathbb{E}_{\mathbf{X}_i \sim p_{ik}} [\log \hat{p}_{ir}(\mathbf{X}_i | \Theta, \Omega)] \right\} \quad (7)$$

The virtual samples are i.i.d so the expectation term in 7 can be written as

$$\mathbb{E}_{\mathbf{X}_i \sim p_{ik}} [\log \hat{p}_{ir}(\mathbf{X}_i | \Theta, \Omega)] \quad (8)$$

$$= \mathbb{E}_{\mathbf{X}_i \sim p_{ik}} \left[\sum_{v=1}^{N_v} \log \hat{p}_{ir}(\mathbf{x}_v^{(i)} | \Theta, \Omega) \right] \quad (9)$$

$$= \sum_{v=1}^{N_v} \mathbb{E}_{\mathbf{X}_i \sim p_{ik}} [\log \hat{p}_{ir}(\mathbf{x}_v^{(i)} | \Theta, \Omega)] \quad (10)$$

$$= N_v \mathbb{E}_{\mathbf{x} \sim p_{ik}} [\log \hat{p}_{ir}(\mathbf{x} | \Theta, \Omega)] \quad (11)$$

Substituting 11 and 7 into 5, we obtain

$$J_{CE}(\Theta, \Omega) \quad (12)$$

$$\leq - \sum_i \sum_k \sum_r \pi_{ik} q_{kr}^{(i)} \left\{ \log \frac{\hat{\pi}_{ir}}{q_{kr}^{(i)}} + N_v \mathbb{E}_{\mathbf{x} \sim p_{ik}} [\log \hat{p}_{ir}(\mathbf{x} | \Theta, \Omega)] \right\} \quad (13)$$

$$= \tilde{J}_{CE}(\Theta, \Omega, \mathbf{q}) \quad (14)$$

The upper bound has the same form as the upper bound in the paper, except it now includes a multiplication by N_v . The optimal values for $q_{kr}^{(i)}$ can also be obtained via the Lagrangian multiplier method (see derivation in

the next section),

$$\hat{q}_{kr}^{(i)} = \frac{\hat{\pi}_{ir} \{\mathcal{N}(\boldsymbol{\mu}_{ik} | \hat{\boldsymbol{\mu}}_{ir}, \hat{\boldsymbol{\Sigma}}_{ir})\}^{N_v} \exp\{-\frac{1}{2} N_v \text{tr}(\hat{\boldsymbol{\Sigma}}_{ir}^{-1} \boldsymbol{\Sigma}_{ik})\}}{\sum_{n=1}^{K_m} \hat{\pi}_{in} \{\mathcal{N}(\boldsymbol{\mu}_{ik} | \hat{\boldsymbol{\mu}}_{in}, \hat{\boldsymbol{\Sigma}}_{in})\}^{N_v} \exp\{-\frac{1}{2} N_v \text{tr}(\hat{\boldsymbol{\Sigma}}_{in}^{-1} \boldsymbol{\Sigma}_{ik})\}}. \quad (15)$$

From the point of view of deterministic annealing, we can obtain a similar form for $\hat{q}_{kr}^{(i)}$. Assume we do not sample those virtual points, the optimization objective is $\tilde{J}_{CE}(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \mathbf{q})$,

$$\min_{\boldsymbol{\Theta}, \boldsymbol{\Omega}, \mathbf{q}} \tilde{J}_{CE}(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \mathbf{q}) = - \sum_{i=1}^N \sum_{k=1}^{K_b} \pi_{ik} \sum_{r=1}^{K_m} q_{kr}^{(i)} \left[\log \frac{\hat{\pi}_{ir}}{q_{kr}^{(i)}} + \mathbb{E}_{\mathbf{x} \sim \mathcal{N}_{ik}} [\log \hat{\mathcal{N}}_{ir}(\mathbf{x})] \right].$$

The $q_{kr}^{(i)}$ here serve as the same effect as the association probability in deterministic annealing. Follow the idea of deterministic annealing, we add randomness into the objective function by maximizing the entropy of variational parameters. So the minimization objective changes to

$$\min_{\boldsymbol{\Theta}, \boldsymbol{\Omega}, \mathbf{q}} \tilde{J}_{CE}(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \mathbf{q}) + T_{DA} \sum_{i=1}^N \sum_{k=1}^{K_b} \pi_{ik} \sum_{r=1}^{K_m} q_{kr}^{(i)} \log q_{kr}^{(i)}, \quad (16)$$

where T_{DA} is the Lagrangian multiplier. The problem can be further written as

$$- \sum_{i=1}^N \sum_{k=1}^{K_b} \pi_{ik} \sum_{r=1}^{K_m} q_{kr}^{(i)} \left[\log \hat{\pi}_{ir} + \mathbb{E}_{\mathbf{x} \sim \mathcal{N}_{ik}} [\log \hat{\mathcal{N}}_{ir}(\mathbf{x})] \right] + (T_{DA} + 1) \sum_{i=1}^N \sum_{k=1}^{K_b} \pi_{ik} \sum_{r=1}^{K_m} q_{kr}^{(i)} \log q_{kr}^{(i)} \quad (17)$$

Similarly, the variational parameters can be solved as

$$\hat{q}_{kr}^{(i)} = \frac{\hat{\pi}_{ir}^{\frac{1}{T_{DA}+1}} \{\mathcal{N}(\boldsymbol{\mu}_{ik} | \hat{\boldsymbol{\mu}}_{ir}, \hat{\boldsymbol{\Sigma}}_{ir})\}^{\frac{1}{T_{DA}+1}} \exp\{-\frac{1}{2} \frac{1}{T_{DA}+1} \text{tr}(\hat{\boldsymbol{\Sigma}}_{ir}^{-1} \boldsymbol{\Sigma}_{ik})\}}{\sum_{n=1}^{K_m} \hat{\pi}_{in}^{\frac{1}{T_{DA}+1}} \{\mathcal{N}(\boldsymbol{\mu}_{ik} | \hat{\boldsymbol{\mu}}_{in}, \hat{\boldsymbol{\Sigma}}_{in})\}^{\frac{1}{T_{DA}+1}} \exp\{-\frac{1}{2} \frac{1}{T_{DA}+1} \text{tr}(\hat{\boldsymbol{\Sigma}}_{in}^{-1} \boldsymbol{\Sigma}_{ik})\}}. \quad (18)$$

Let $\frac{1}{T_{DA}+1} = N_v$, we observe that the only difference is the exponential term of reconstruction weights. So the number of virtual sample serves as the temperature in deterministic annealing, which controls the degree of randomness of the variational parameters. If N_v gets bigger, T_{DA} gets smaller, the variational parameters will become more deterministic and the assignment will become harder. If N_v gets smaller, T_{DA} gets bigger, the variational parameters will become more random and the assignment will get softer.

3 Variational EM Derivation

3.1 Variational E-step

In the E-step, we optimize $\tilde{J}_{CE}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Omega}}, \mathbf{q})$ with respect to $\mathbf{q}$. The problem is formulated as

$$\min_{\mathbf{q}} \tilde{J}_{CE}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Omega}}, \mathbf{q}) \quad (19)$$

$$\text{s.t. } \sum_r q_{kr}^{(i)} - 1 = 0, \forall i, k, \quad (20)$$

where the constraint enforces that the assignment variables sum to 1 (over r). Using Lagrangian multipliers, the Lagrangian function is

$$- \sum_i \sum_k \sum_r \pi_{ik} q_{kr}^{(i)} \left\{ \log \frac{\hat{\pi}_{ir}}{q_{kr}^{(i)}} + N_v \mathbb{E}_{\mathbf{x} \sim p_{ik}} [\log \hat{p}_{ir}(\mathbf{x} | \hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Omega}})] \right\} + \sum_i \sum_k \lambda_k^{(i)} \left(\sum_r q_{kr}^{(i)} - 1 \right) \quad (21)$$

Taking derivative with respect to $q_{kr}^{(i)}$ and setting to zero yields

$$\hat{q}_{kr}^{(i)} = \frac{\hat{\pi}_{ir} \exp(N_v \mathbb{E}_{\mathbf{x} \sim p_{ik}} [\log \hat{p}_{ir}(\mathbf{x} | \hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Omega}})])}{\sum_n \hat{\pi}_{in} \exp(N_v \mathbb{E}_{\mathbf{x} \sim p_{ik}} [\log \hat{p}_{in}(\mathbf{x} | \hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Omega}})])}. \quad (22)$$

Substituting for the expectation of a log-Gaussian, we obtain

$$\hat{q}_{kr}^{(i)} = \frac{\hat{\pi}_{ir} \{\mathcal{N}(\boldsymbol{\mu}_{ik} | \hat{\boldsymbol{\mu}}_{ir}, \hat{\boldsymbol{\Sigma}}_{ir})\}^{N_v} \exp\{-\frac{1}{2} N_v \text{tr}(\hat{\boldsymbol{\Sigma}}_{ir}^{-1} \boldsymbol{\Sigma}_{ik})\}}{\sum_{n=1}^{K_m} \hat{\pi}_{in} \{\mathcal{N}(\boldsymbol{\mu}_{ik} | \hat{\boldsymbol{\mu}}_{in}, \hat{\boldsymbol{\Sigma}}_{in})\}^{N_v} \exp\{-\frac{1}{2} N_v \text{tr}(\hat{\boldsymbol{\Sigma}}_{in}^{-1} \boldsymbol{\Sigma}_{ik})\}} \quad (23)$$

3.2 Variational M-step

In the M-step, we optimize $\tilde{J}_{CE}(\Theta, \Omega, \hat{q})$ with respect to Θ and Ω . The problem is

$$\min_{\Theta, \Omega} \tilde{J}_{CE}(\Theta, \Omega, \hat{q}) \quad (24)$$

We use L-BFGS method to optimize the parameters so we need to take derivative with respect to $\Theta = \{\mathbf{a}_r, \{\mathbf{m}_{rl}\}_{l=1}^{d_z}, \{\mathbf{C}_{rl}\}_{l=1}^{d_y}, \mathbf{b}_r, \beta_r\}_{r=1}^{K_m}$ and $\Omega = \{\mathbf{w}_i, \mathbf{z}_i, \mathbf{y}_i\}_{i=1}^N$.

(a) **Derivative with respect to Θ**

$$\frac{\partial \tilde{J}_{CE}(\Theta, \Omega, \hat{q})}{\partial \mathbf{a}_{rl}} \quad (25)$$

$$= - \sum_i \sum_k \pi_{ik} \frac{\partial}{\partial \mathbf{a}_{rl}} (\hat{q}_{kr}^{(i)} \log \frac{g(\mathbf{w}_i^T \mathbf{a}_r)}{\sum_m g(\mathbf{w}_i^T \mathbf{a}_r)} + \sum_{t \neq r} \hat{q}_{kt}^{(i)} \log \frac{g(\mathbf{w}_i^T \mathbf{a}_t)}{\sum_m g(\mathbf{w}_i^T \mathbf{a}_t)}) \quad (26)$$

$$= \sum_i \sum_k \pi_{ik} (\hat{q}_{kr}^{(i)} g(\mathbf{w}_i^T \mathbf{a}_r) w_{il} \exp(\mathbf{w}_i^T \mathbf{a}_r) - \frac{g(\mathbf{w}_i^T \mathbf{a}_r)^2 w_{il} \exp(\mathbf{w}_i^T \mathbf{a}_r)}{\sum_m g(\mathbf{w}_i^T \mathbf{a}_m)}) \quad (27)$$

$$= \sum_i \sum_k \pi_{ik} (\hat{q}_{kr}^{(i)} - \hat{\pi}_{ir}) g(-\mathbf{w}_i^T \mathbf{a}_r) w_{il} \quad (28)$$

$$\frac{\partial \tilde{J}_{CE}(\Theta, \Omega, \hat{q})}{\partial \mathbf{m}_{rl}} = \frac{1}{2} \frac{\partial}{\partial \mathbf{m}_{rl}} \sum_i \sum_k \sum_r \pi_{ik} \hat{q}_{kr}^{(i)} (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})^T \hat{\boldsymbol{\Sigma}}_{ir}^{-1} (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir}) \quad (29)$$

$$= - \sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} z_{il} \hat{\boldsymbol{\Sigma}}_{ir}^{-1} (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir}) \quad (30)$$

$$\frac{\partial \tilde{J}_{CE}(\Theta, \Omega, \hat{q})}{\partial \mathbf{b}_r} = - \sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \hat{\boldsymbol{\Sigma}}_{ir}^{-1} (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir}) \quad (31)$$

Denote $\mathbf{Y}_i^{\log}$ and $\mathbf{C}_r$ as follows

$$\mathbf{Y}_i^{\log} = \begin{bmatrix} \log(1 + \exp(y_{i1}))\mathbf{I} & 0 & \dots & 0 \\ 0 & \ddots & 0 & \vdots \\ \vdots & \ddots & \log(1 + \exp(y_{i(d_y-1)}))\mathbf{I} & \vdots \\ 0 & \dots & \dots & \log(1 + \exp(y_{id_y}))\mathbf{I} \end{bmatrix} \in R^{D_{d_y} \times D_{d_y}} \quad (32)$$

$$\mathbf{C}_r = [\mathbf{C}_{r1}, \dots, \mathbf{C}_{rd_y}] \quad (33)$$

So the derivative of $\mathbf{C}_r$ can be written as

$$\frac{\partial \tilde{J}_{CE}(\Theta, \Omega, \hat{q})}{\partial \mathbf{C}_r} \quad (34)$$

$$= \frac{1}{2} \frac{\partial}{\partial \mathbf{C}_r} \sum_i \sum_k \sum_r \pi_{ik} \hat{q}_{kr}^{(i)} \{(\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})^T \hat{\boldsymbol{\Sigma}}_{ir}^{(-1)} (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir}) + \text{tr}(\hat{\boldsymbol{\Sigma}}_{ir}^{(-1)} \boldsymbol{\Sigma}_{ik}) - \log |\hat{\boldsymbol{\Sigma}}_{ir}^{(-1)}| \} \quad (35)$$

$$= \sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \{ \boldsymbol{\Sigma}_{ik} \mathbf{C}_r \mathbf{Y}_i^{\log} + (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})(\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})^T \mathbf{C}_r \mathbf{Y}_i^{\log} - (\mathbf{C}_r \mathbf{Y}_i^{\log} \mathbf{C}_r^T + \beta_r^2 \mathbf{I})^{-1} \mathbf{C}_r \mathbf{Y}_i^{\log} \} \quad (36)$$

$$\frac{\partial \tilde{J}_{CE}(\Theta, \Omega, \hat{q})}{\partial \beta_r} = \beta_r \sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \{(\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})^T (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir}) - \text{tr}(\hat{\boldsymbol{\Sigma}}_{ir} - \boldsymbol{\Sigma}_{ik}) \} \quad (37)$$

(b) **Derivative with respect to Ω**

$$\frac{\partial \tilde{J}_{CE}(\Theta, \Omega, \hat{q})}{\partial w_{il}} \quad (38)$$

$$= -\frac{\partial}{\partial w_{il}} \sum_k \pi_{ik} \sum_r \hat{q}_{kr}^{(i)} \log \frac{g(\mathbf{w}_i^T \mathbf{a}_r)}{\sum_m g(\mathbf{w}_i^T \mathbf{a}_m)} \quad (39)$$

$$= \sum_k \pi_{ik} \sum_r \hat{q}_{kr}^{(i)} \left\{ g(\mathbf{w}_i^T \mathbf{a}_r) \exp(\mathbf{w}_i^T \mathbf{a}_r) a_{rl} - \frac{\sum_m a_{ml} \exp(\mathbf{w}_i^T \mathbf{a}_m) g(\mathbf{w}_i^T \mathbf{a}_m)^2}{\sum_m g(\mathbf{w}_i^T \mathbf{a}_m)} \right\} \quad (40)$$

$$= \sum_k \pi_{ik} \sum_r \hat{q}_{kr}^{(i)} \left\{ g(\mathbf{w}_i^T \mathbf{a}_r) \exp(\mathbf{w}_i^T \mathbf{a}_r) a_{rl} - \sum_m \hat{\pi}_{im} a_{ml} \exp(\mathbf{w}_i^T \mathbf{a}_m) g(\mathbf{w}_i^T \mathbf{a}_m) \right\} \quad (41)$$

$$= \sum_k \pi_{ik} \sum_r \hat{q}_{kr}^{(i)} \left\{ g(-\mathbf{w}_i^T \mathbf{a}_r) a_{rl} - \sum_m \hat{\pi}_{im} a_{ml} g(-\mathbf{w}_i^T \mathbf{a}_m) \right\} \quad (42)$$

$$\frac{\partial \tilde{J}_{CE}(\Theta, \Omega, \hat{q})}{\partial z_{il}} = - \sum_k \sum_r \pi_{ik} \hat{q}_{kr}^{(i)} \boldsymbol{\mu}_{rl}^T \hat{\Sigma}_{ir}^{-1} (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir}) \quad (43)$$

$$\frac{\partial \tilde{J}_{CE}(\Theta, \Omega, \hat{q})}{\partial y_{il}} \quad (44)$$

$$= -\frac{1}{2} \sum_k \sum_r \pi_{ik} \hat{q}_{kr}^{(i)} \frac{\exp(y_{il})}{1 + \exp(y_{il})} \{ \text{tr}(\mathbf{C}_{rl} \mathbf{C}_{rl}^T (\boldsymbol{\Sigma}_{ik} - \hat{\boldsymbol{\Sigma}}_{ir})) + (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})^T \mathbf{C}_{rl} \mathbf{C}_{rl}^T (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir}) \} \quad (45)$$

After we obtaining the derivative with respect to Θ and Ω , we use L-BFGS to optimize those parameters and latent variables at the same time by taking them as a long vector.

4 Optimization for Hierarchical Latent Space

We incorporate PCA into the optimization procedure to obtain the hierarchical latent vector $\mathbf{v}$ directly. Denote the original vector as $\mathbf{L}_i = [\mathbf{w}_i; \mathbf{y}_i; \mathbf{z}_i] \in \mathbb{R}^{d_w + d_y + d_z}$. PCA aims to find a linear combination of a set of orthonormal basis vectors to represent high-dimensional vectors so we assume a matrix $\mathbf{H} = [\mathbf{H}_1, \mathbf{H}_2, \dots, \mathbf{H}_{d_v}] \in \mathbb{R}^{(d_w + d_y + d_z) \times d_v}$ consisting of orthonormal basis vectors and linear combinations $\mathbf{v}_i$. We have the following relationship

$$\mathbf{L}_i = \mathbf{H} \mathbf{v}_i \quad (46)$$

We multiply $\mathbf{L}_i$ by $\mathbf{T}_w$ and $\mathbf{T}_z$ to obtain $\mathbf{w}_i$ and $\mathbf{z}_i$ and multiply $\mathbf{L}_i$ by $\mathbf{t}_{y_l}$ to obtain y_{il} . The reason why we want y_{il} is that we need elements of $\mathbf{y}_i$ to calculate the reconstruction precision matrix. So we can represent $\mathbf{w}_i$, y_{il} and $\mathbf{z}_i$ in the following way

$$\mathbf{w}_i = \mathbf{T}_w \mathbf{H} \mathbf{v}_i, \quad y_{il} = \mathbf{t}_{y_l} \mathbf{H} \mathbf{v}_i, \quad \mathbf{z}_i = \mathbf{T}_z \mathbf{H} \mathbf{v}_i \quad (47)$$

Further, $\hat{\pi}_{ir}$, $\hat{\boldsymbol{\mu}}_{ir}$, and $\hat{\Sigma}_{ir}^{-1}$ can be expressed as

$$\hat{\pi}_{ir} = \frac{g(\mathbf{v}_i^T \mathbf{H}^T \mathbf{T}_w^T \mathbf{a}_r)}{\sum_{n=1}^{K_m} g(\mathbf{v}_i^T \mathbf{H}^T \mathbf{T}_w^T \mathbf{a}_n)}, \quad \hat{\boldsymbol{\mu}}_{ir} = \mathbf{M}_r \mathbf{T}_z \mathbf{H} \mathbf{v}_i + \mathbf{b}_r, \quad \hat{\Sigma}_{ir}^{-1} = \sum_{l=1}^{d_y} \log(1 + \exp(\mathbf{t}_{y_l} \mathbf{H} \mathbf{v}_i)) \mathbf{C}_{rl} \mathbf{C}_{rl}^T + \beta_r^2 \mathbf{I}. \quad (48)$$

where $\mathbf{M}_r = [\mathbf{m}_{r1}, \dots, \mathbf{m}_{rl}]$. The objective function is,

$$\tilde{J}_{CE}(\Theta, \mathbf{H}, \mathbf{v}_i|_{i=1}^N, \mathbf{q}) + c_v \sum_{i=1}^N \|\mathbf{v}_i\|^2, \quad (49)$$

Now we need to take the derivative with respect to $\mathbf{v}_i$ and $\mathbf{H}$. Like the previous part, we tackle different parts of the objective function which are functions of $\mathbf{w}$, $\mathbf{y}$ and $\mathbf{z}$.

$$\frac{\partial \tilde{J}_{CE}(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \hat{\mathbf{q}})|_{\mathbf{w}_i}}{\partial \mathbf{v}_i} = \frac{\partial}{\partial \mathbf{v}_i} - \sum_k \sum_r \pi_{ik} \hat{q}_{kr}^{(i)} \log \hat{\pi}_{ir} \quad (50)$$

$$= \sum_k \sum_r \pi_{ik} \hat{q}_{kr}^{(i)} \{ \exp(\mathbf{w}_i^T \mathbf{a}_r) g(\mathbf{w}_i^T \mathbf{a}_r) \mathbf{H}^T \mathbf{T}_w^T \mathbf{a}_r - \sum_m \hat{\pi}_{im} g(\mathbf{w}_i^T \mathbf{a}_m) \exp(\mathbf{w}_i^T \mathbf{a}_m) \mathbf{H}^T \mathbf{T}_w^T \mathbf{a}_m \} \quad (51)$$

$$\frac{\partial \tilde{J}_{CE}(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \hat{\mathbf{q}})|_{\mathbf{w}}}{\partial \mathbf{H}} = \frac{\partial}{\partial \mathbf{v}_i} - \sum_i \sum_k \sum_r \pi_{ik} \hat{q}_{kr}^{(i)} \log \hat{\pi}_{ir} \quad (52)$$

$$= \sum_i \sum_k \sum_r \pi_{ik} \hat{q}_{kr}^{(i)} \{ \exp(\mathbf{w}_i^T \mathbf{a}_r) g(\mathbf{w}_i^T \mathbf{a}_r) \mathbf{T}_w^T \mathbf{a}_r \mathbf{v}_i^T - \sum_m \hat{\pi}_{im} g(\mathbf{w}_i^T \mathbf{a}_m) \exp(\mathbf{w}_i^T \mathbf{a}_m) \mathbf{T}_w^T \mathbf{a}_m \mathbf{v}_i^T \} \quad (53)$$

$$\frac{\partial \tilde{J}_{CE}(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \hat{\mathbf{q}})|_{\mathbf{z}_i}}{\partial \mathbf{v}_i} = \frac{\partial}{\partial \mathbf{v}_i} \frac{1}{2} \sum_k \sum_r \pi_{ik} \hat{q}_{kr}^{(i)} (\boldsymbol{\mu}_{ik} - \mathbf{M}_r \mathbf{T}_z \mathbf{H} \mathbf{v}_i - \mathbf{b}_r)^T \hat{\boldsymbol{\Sigma}}_{ir}^{-1} (\boldsymbol{\mu}_{ik} - \mathbf{M}_r \mathbf{T}_z \mathbf{H} \mathbf{v}_i - \mathbf{b}_r) \quad (54)$$

$$= \sum_k \sum_r \pi_{ik} \hat{q}_{kr}^{(i)} \{ \mathbf{H}^T \mathbf{T}_z^T \mathbf{M}_r^T \hat{\boldsymbol{\Sigma}}_{ir}^{-1} (\hat{\boldsymbol{\mu}}_{ir} - \boldsymbol{\mu}_{ik}) \} \quad (55)$$

$$\frac{\partial \tilde{J}_{CE}(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \hat{\mathbf{q}})|_{\mathbf{z}}}{\partial \mathbf{H}} = \sum_i \sum_k \sum_r \pi_{ik} \hat{q}_{kr}^{(i)} \{ \mathbf{T}_z^T \mathbf{M}_r^T \hat{\boldsymbol{\Sigma}}_{ir}^{-1} (\hat{\boldsymbol{\mu}}_{ir} - \boldsymbol{\mu}_{ik}) \mathbf{v}_i^T \} \quad (56)$$

$$\frac{\partial \tilde{J}_{CE}(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \hat{\mathbf{q}})|_{\mathbf{y}_i}}{\partial \mathbf{v}_i} = \frac{1}{2} \frac{\partial}{\partial \mathbf{v}_i} \sum_k \sum_r \pi_{ik} \hat{q}_{kr}^{(i)} \{ \sum_{l=1}^{d_y} (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})^T \mathbf{C}_{rl} \mathbf{C}_{rl}^T (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir}) \log(1 + \exp(\mathbf{t}_{y_l} \mathbf{H} \mathbf{v}_i)) + \quad (57)$$

$$+ \text{tr}(\sum_{l=1}^{d_y} \log(1 + \exp(\mathbf{t}_{y_l} \mathbf{H} \mathbf{v}_i)) \mathbf{C}_{rl} \mathbf{C}_{rl}^T \boldsymbol{\Sigma}_{ik}) \quad (58)$$

$$- \log | \sum_{l=1}^{d_y} \log(1 + \exp(\mathbf{t}_{y_l} \mathbf{H} \mathbf{v}_i)) \mathbf{C}_{rl} \mathbf{C}_{rl}^T + \beta^2 \mathbf{I} | \} \quad (59)$$

$$= \frac{1}{2} \sum_k \sum_r \pi_{ik} \hat{q}_{kr}^{(i)} \{ \sum_{l=1}^{d_y} (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})^T \mathbf{C}_{rl} \mathbf{C}_{rl}^T (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir}) \mathbf{H}^T \mathbf{t}_{y_l}^T \frac{\exp(y_{il})}{1 + \exp(y_{il})} \quad (60)$$

$$+ \sum_{l=1}^{d_y} \mathbf{H}^T \mathbf{t}_{y_l}^T \frac{\exp(y_{il})}{1 + \exp(y_{il})} \text{tr}(\mathbf{C}_{rl} \mathbf{C}_{rl}^T \boldsymbol{\Sigma}_{ik}) \quad (61)$$

$$- \sum_{l=1}^{d_y} \mathbf{H}^T \mathbf{t}_{y_l}^T \frac{\exp(y_{il})}{1 + \exp(y_{il})} \text{tr}(\hat{\boldsymbol{\Sigma}}_{ir} \mathbf{C}_{rl} \mathbf{C}_{rl}^T) \} \quad (62)$$

$$\frac{\partial \tilde{J}_{CE}(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \hat{\mathbf{q}})|_{\mathbf{y}}}{\partial \mathbf{H}} = \sum_i \sum_k \sum_r \pi_{ik} \hat{q}_{kr}^{(i)} \{ \sum_{l=1}^{d_y} (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})^T \mathbf{C}_{rl} \mathbf{C}_{rl}^T (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir}) \mathbf{t}_{y_l}^T \frac{\exp(y_{il})}{1 + \exp(y_{il})} \mathbf{v}_i^T \quad (63)$$

$$+ \sum_{l=1}^{d_y} \mathbf{t}_{y_l}^T \frac{\exp(y_{il})}{1 + \exp(y_{il})} \mathbf{v}_i^T \text{tr}(\mathbf{C}_{rl} \mathbf{C}_{rl}^T \boldsymbol{\Sigma}_{ik}) \quad (64)$$

$$- \sum_{l=1}^{d_y} \mathbf{t}_{y_l}^T \frac{\exp(y_{il})}{1 + \exp(y_{il})} \mathbf{v}_i^T \text{tr}(\hat{\boldsymbol{\Sigma}}_{ir} \mathbf{C}_{rl} \mathbf{C}_{rl}^T) \} \quad (65)$$

Algorithm 1 Optimization Algorithm for PML-GMM

Input: $\{\pi_{ik}, \mu_{ik}, \Sigma_{ik}\}_{k=1}^{K_b}, i = 1, \dots, N$ **Parameter:** N_v, c_w, c_z, c_y, T **Output:** $\{\hat{q}, \hat{\Theta}, \hat{\Omega}\}$

```
1: Initialize  $\{q, \Theta, \Omega\}$ 
2: while not converge do
3:   M-step: Fix  $\hat{q}$  and optimize  $\{\Theta, \Omega\}$ 
4:   E-step: Fix  $\{\hat{\Theta}, \hat{\Omega}\}$  and optimize  $\hat{q}$ , save  $\tilde{J}_{CE}$ 
5:   Randomly sample a GMM and its component, denoted as  $i_s, k_s$ 
6:   Randomize variational parameters of  $k_s$ th component of  $i_s$ th GMM as  $\hat{q}_{k_s}^{(i_s)*}$  and calculate  $\tilde{J}_{CE}^*$  using  $\hat{q}_{k_s}^{(i_s)*}$ 
7:   if  $p_{swap} = \frac{\exp(-\tilde{J}_{CE}^*/T)}{\exp(-\tilde{J}_{CE}/T)} > 1$  then
8:     Replace  $\hat{q}_{k_s}^{(i_s)}$  with  $\hat{q}_{k_s}^{(i_s)*}$ 
9:   else
10:    Do the above replacement with probability  $p_{swap}$ 
11:   end if
12: end while
13: return  $\{\hat{q}, \hat{\Theta}, \hat{\Omega}\}$ 
```

4.1 Stochastic Optimization

Our optimization starts from M-step, i.e., given an initialization of $\hat{q}$, we first optimize $\{\hat{\Theta}, \hat{\Omega}\}$. We do not start from E-step in that a random initialization of $\{\hat{\Theta}, \hat{\Omega}\}$ will lead to very poor $\hat{q}$ while a randomly or uniformly initialized $\hat{q}$ will benefit the exploration of the manifold parameters. However, if $\hat{q}$ is stuck to a local minimum in the optimization process, a deterministic optimization cannot let $\hat{q}$ as well as $\{\hat{\Theta}, \hat{\Omega}\}$ escape from that. To avoid local minima, we introduce randomness into our optimization by Metropolis-Hastings sampling for $\hat{q}$. See Algorithm 1 for the detail.

5 Ablation Study

PML-GMM can also be used to learn a GMM manifold directly from the data (denoted as “Direct-PML-GMM”). Let $\mathbf{x}_{ij}$ correspond to the j th sample for the i th subject (data subset), where the i th subject has N_i samples. We can represent the data from the i th subject as a GMM with N_i components, where the component means are the data points $\mathbf{x}_{ij}$, the covariance matrices are $\mathbf{0}$, and the component priors are uniform. Using these data GMMs, the optimization objective of PML-GMM is equivalent to maximizing the log-likelihood of all the data when fitting the GMM manifold. Conceptually, Direct-PML uses real data samples to estimate the manifold, whereas (indirect) PML-GMM uses virtual samples. Because it uses all samples, Direct-PML is too slow to be used on large datasets. PML-GMM addresses this problem by first learning an unconstrained GMM for each subject, and then projecting them to the GMM manifold, thus avoiding calculations using all data samples.

The input to the model is now data points $\mathbf{X}_i = \{\mathbf{x}_{i1}, \dots, \mathbf{x}_{iN_i}\}$ for each entity, where N_i is the number of data samples in data entity i . We put one delta function to each data sample and see one data entity as one GMM with N_i components where the means are data samples, weights are uniform and the covariance matrices are $\mathbf{0}$. The objective function is to minimize the negative cross entropy between input GMMs and approximated GMMs, which is equivalent to maximize cross entropy

$$\min - \sum_i^N E_{p_i}[\log \hat{p}_i] \simeq \max \sum_i^N E_{p_i}[\log \hat{p}_i] \quad (66)$$

where the p_i is a sum of delta functions and $\hat{p}_i$ is a GMM with K_m components. Since the expectation is integral over a sum of delta functions, it changes into a summation of log-likelihood

$$\sum_i^N E_{p_i}[\log \sum_r^{K_m} \hat{\pi}_{ir} \mathcal{N}(\mathbf{x} | \hat{\mu}_{ir}, \hat{\Sigma}_{ir})] = \sum_i^N \sum_j^{N_i} \log \sum_r^{K_m} \hat{\pi}_{ir} \mathcal{N}(\mathbf{x}_{ij} | \hat{\mu}_{ir}, \hat{\Sigma}_{ir}) \quad (67)$$

so maximizing cross entropy is equivalent to maximizing log-likelihood for the data samples. Again, we use Jensen's inequality to obtain a low bound for the log-likelihood

$$\sum_i^N \sum_j^{N_i} \log \sum_r^{K_m} \hat{\pi}_{ir} \mathcal{N}(\mathbf{x}_{ij} | \hat{\boldsymbol{\mu}}_{ir}, \hat{\boldsymbol{\Sigma}}_{ir}) \geq \sum_i^N \sum_j^{N_i} \sum_r^{K_m} q_{jr}^{(i)} \log \frac{\hat{\pi}_{ir} \mathcal{N}(\mathbf{x}_{ij} | \hat{\boldsymbol{\mu}}_{ir}, \hat{\boldsymbol{\Sigma}}_{ir})}{q_{jr}^{(i)}} \triangleq \tilde{J}_{LL}(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \mathbf{q}) \quad (68)$$

where $q_{jr}^{(i)}$ is the variational parameters, indicating the assignment value of j th sample to r th GMM component. The idea here is similar to HEM. We maximize the log-likelihood w.r.t. manifold parameters and latent variables to obtain the GMM manifold. Now we extend our framework so that it can be applied to raw data besides GMMs.

Variational E-step. We first optimize w.r.t. variational parameters. Like in Section 3.1, we add Lagrangian multipliers to the optimization problem and we can obtain a closed-form solution for $q_{jr}^{(i)}$

$$\hat{q}_{jr}^{(i)} = \frac{\hat{\pi}_{ir} \mathcal{N}(\mathbf{x}_{ij} | \hat{\boldsymbol{\mu}}_{ir}, \hat{\boldsymbol{\Sigma}}_{ir})}{\sum_n^{K_m} \hat{\pi}_{in} \mathcal{N}(\mathbf{x}_{ij} | \hat{\boldsymbol{\mu}}_{in}, \hat{\boldsymbol{\Sigma}}_{in})} \quad (69)$$

This result is the same as Equation 23 when covariance matrices are $\mathbf{0}$ and the number of virtual sample is 1. Since here we use the real samples as the training data, we set the number of virtual sample to be 1 so that this result is consistent with previous framework.

Variational M-step. After we obtain $\hat{q}_{jr}^{(i)}$, we can optimize w.r.t. $\boldsymbol{\Theta}$ and $\boldsymbol{\Omega}$. Bring the approximated variational parameters and expand the terms in $\tilde{J}_{LL}(\boldsymbol{\Theta}, \boldsymbol{\Omega})$

$$\tilde{J}_{LL}(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \hat{\mathbf{q}}) = \sum_i^N \sum_j^{N_i} \sum_r^{K_m} \hat{q}_{jr}^{(i)} \left\{ \log \hat{\pi}_{ir} - \frac{d}{2} \log 2\pi - \frac{1}{2} \log |\hat{\boldsymbol{\Sigma}}_{ir}| - \frac{1}{2} (\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_{ir})^T \hat{\boldsymbol{\Sigma}}_{ir}^{-1} (\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_{ir}) - \log \hat{q}_{jr}^{(i)} \right\} \quad (70)$$

$$\propto \sum_i^N \sum_j^{N_i} \sum_r^{K_m} \hat{q}_{jr}^{(i)} \left\{ \log \hat{\pi}_{ir} - \frac{1}{2} \log |\hat{\boldsymbol{\Sigma}}_{ir}| - \frac{1}{2} (\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_{ir})^T \hat{\boldsymbol{\Sigma}}_{ir}^{-1} (\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_{ir}) \right\} \quad (71)$$

a) Derivative w.r.t. $\boldsymbol{\Theta}$

$$\frac{\partial}{\partial a_{rl}} \tilde{J}_{LL}(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \hat{\mathbf{q}}) \quad (72)$$

$$= \frac{\partial}{\partial a_{rl}} \sum_i^N \sum_r^{K_m} \sum_j^{N_i} \hat{q}_{jr}^{(i)} \log \hat{\pi}_{ir} \quad (73)$$

$$= \frac{\partial}{\partial a_{rl}} \sum_i^N \sum_r^{K_m} \hat{q}_r^{(i)} \log \hat{\pi}_{ir} \quad (74)$$

$$= \sum_i^N \left\{ \frac{\partial}{\partial a_{rl}} \hat{q}_r^{(i)} \log \frac{g(\mathbf{w}_i^T \mathbf{a}_r)}{\sum_m g(\mathbf{w}_i^T \mathbf{a}_m)} + \frac{\partial}{\partial a_{rl}} \sum_{t \neq r} \hat{q}_t^{(i)} \log \frac{g(\mathbf{w}_i^T \mathbf{a}_t)}{\sum_m g(\mathbf{w}_i^T \mathbf{a}_m)} \right\} \quad (75)$$

$$= \sum_i^N \left\{ -\frac{\partial}{\partial a_{rl}} \hat{q}_r^{(i)} \log(1 + \exp(\mathbf{w}_i^T \mathbf{a}_r)) - \frac{\partial}{\partial a_{rl}} \sum_r \hat{q}_r^{(i)} \log \sum_m g(\mathbf{w}_i^T \mathbf{a}_m) \right\} \quad (76)$$

$$= \sum_i^N \left\{ -\frac{\partial}{\partial a_{rl}} \hat{q}_r^{(i)} \log(1 + \exp(\mathbf{w}_i^T \mathbf{a}_r)) - N_i \frac{\partial}{\partial a_{rl}} \log \sum_m g(\mathbf{w}_i^T \mathbf{a}_m) \right\} \quad (77)$$

$$= \sum_i^N \left\{ -\hat{q}_r^{(i)} \frac{\exp(\mathbf{w}_i^T \mathbf{a}_r)}{1 + \exp(\mathbf{w}_i^T \mathbf{a}_r)} w_{il} + N_i \frac{g(\mathbf{w}_i^T \mathbf{a}_r)^2 \exp(\mathbf{w}_i^T \mathbf{a}_r) w_{il}}{\sum_m g(\mathbf{w}_i^T \mathbf{a}_m)} \right\} \quad (78)$$

$$= \sum_i^N \left\{ N_i \frac{g(\mathbf{w}_i^T \mathbf{a}_r)^2 \exp(\mathbf{w}_i^T \mathbf{a}_r) w_{il}}{\sum_m g(\mathbf{w}_i^T \mathbf{a}_m)} - \hat{q}_r^{(i)} g(\mathbf{w}_i^T \mathbf{a}_r) \exp(\mathbf{w}_i^T \mathbf{a}_r) w_{il} \right\} \quad (79)$$

$$= \sum_i^N \left\{ N_i \hat{\pi}_{ir} g(\mathbf{w}_i^T \mathbf{a}_r) \exp(\mathbf{w}_i^T \mathbf{a}_r) w_{il} - \hat{q}_r^{(i)} g(\mathbf{w}_i^T \mathbf{a}_r) \exp(\mathbf{w}_i^T \mathbf{a}_r) w_{il} \right\} \quad (80)$$

$$\frac{\partial}{\partial \mathbf{m}_{rl}} \tilde{J}_{LL}(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \hat{\mathbf{q}}) \quad (81)$$

$$= \frac{\partial}{\partial \mathbf{m}_{rl}} - \frac{1}{2} \sum_i^N \sum_j^{N_i} \sum_r^{K_m} \hat{q}_{jr}^{(i)} (\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_{ir})^T \hat{\boldsymbol{\Sigma}}_{ir}^{-1} (\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_{ir}) \quad (82)$$

$$= \sum_i^N \sum_j^{N_i} \hat{q}_{jr}^{(i)} z_{il} \hat{\boldsymbol{\Sigma}}_{ir}^{-1} (\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_{ir}) \quad (83)$$

$$\frac{\partial}{\partial \mathbf{b}_r} \tilde{J}_{LL}(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \hat{\mathbf{q}}) \quad (84)$$

$$= \frac{\partial}{\partial \mathbf{b}_r} - \frac{1}{2} \sum_i^N \sum_j^{N_i} \sum_r^{K_m} \hat{q}_{jr}^{(i)} (\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_{ir})^T \hat{\boldsymbol{\Sigma}}_{ir}^{-1} (\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_{ir}) \quad (85)$$

$$= \sum_i^N \sum_j^{N_i} \hat{q}_{jr}^{(i)} \hat{\boldsymbol{\Sigma}}_{ir}^{-1} (\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_{ir}) \quad (86)$$

Use the same notation $\mathbf{Y}_i^{\log}$ and $\mathbf{C}_r$ in Section 3.2, the derivative with respect to $\mathbf{C}_r$ is

$$\frac{\partial}{\partial \mathbf{C}_r} \tilde{J}_{LL}(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \hat{\mathbf{q}}) \quad (87)$$

$$= \frac{\partial}{\partial \mathbf{C}_r} \frac{1}{2} \sum_i^N \sum_j^{N_i} \sum_r^{K_m} \hat{q}_{jr}^{(i)} \{ \log |\hat{\boldsymbol{\Sigma}}_{ir}^{-1}| - (\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_{ir})^T \hat{\boldsymbol{\Sigma}}_{ir}^{-1} (\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_{ir}) \} \quad (88)$$

$$= \sum_i^N \sum_j^{N_i} \hat{q}_{jr}^{(i)} \{ (\mathbf{C}_r \mathbf{Y}_i^{\log} \mathbf{C}_r^T + \beta_r^2 \mathbf{I})^{-1} \mathbf{C}_r \mathbf{Y}_i^{\log} - (\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_{ir})(\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_{ir})^T \mathbf{C}_r \mathbf{Y}_i^{\log} \} \quad (89)$$

$$\frac{\partial}{\partial \beta_r} \tilde{J}_{LL}(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \hat{\mathbf{q}}) \quad (90)$$

$$= \beta_r \sum_i^N \sum_j^{N_i} \hat{q}_{jr}^{(i)} \{ \text{tr}(\hat{\boldsymbol{\Sigma}}_{ir}) - (\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_{ir})^T (\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_{ir}) \} \quad (91)$$

b) Derivative w.r.t. $\boldsymbol{\Omega}$

$$\frac{\partial}{\partial w_{il}} \tilde{J}_{LL}(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \hat{\mathbf{q}}) \quad (92)$$

$$= \frac{\partial}{\partial w_{il}} \sum_r^{K_m} \hat{q}_r^{(i)} \log \hat{\pi}_{ir} \quad (93)$$

$$= \frac{\partial}{\partial w_{il}} \sum_r^{K_m} \hat{q}_r^{(i)} \log \frac{g(\mathbf{w}_i^T \mathbf{a}_r)}{\sum_m g(\mathbf{w}_i^T \mathbf{a}_m)} \quad (94)$$

$$= \frac{\partial}{\partial w_{il}} \sum_r^{K_m} \hat{q}_r^{(i)} \log g(\mathbf{w}_i^T \mathbf{a}_r) - \frac{\partial}{\partial w_{il}} \sum_r^{K_m} \hat{q}_r^{(i)} \log \sum_m g(\mathbf{w}_i^T \mathbf{a}_m) \quad (95)$$

$$= \frac{\partial}{\partial w_{il}} \sum_r^{K_m} \hat{q}_r^{(i)} \log g(\mathbf{w}_i^T \mathbf{a}_r) - \frac{\partial}{\partial w_{il}} N_i \hat{q}_r^{(i)} \log \sum_m g(\mathbf{w}_i^T \mathbf{a}_m) \quad (96)$$

$$= N_i \frac{\sum_m a_{ml} \exp(\mathbf{w}_i^T \mathbf{a}_m) g(\mathbf{w}_i^T \mathbf{a}_m)^2}{\sum_m g(\mathbf{w}_i^T \mathbf{a}_m)} - \sum_r^{K_m} \hat{q}_r^{(i)} g(\mathbf{w}_i^T \mathbf{a}_r) \exp(\mathbf{w}_i^T \mathbf{a}_r) a_{rl} \quad (97)$$

$$= N_i \sum_m \hat{\pi}_{im} a_{ml} \exp(\mathbf{w}_i^T \mathbf{a}_m) g(\mathbf{w}_i^T \mathbf{a}_m) - \sum_r^{K_m} \hat{q}_r^{(i)} g(\mathbf{w}_i^T \mathbf{a}_r) \exp(\mathbf{w}_i^T \mathbf{a}_r) a_{rl} \quad (98)$$

$$\frac{\partial}{\partial z_{il}} \tilde{J}_{LL}(\Theta, \Omega, \hat{q}) \quad (99)$$

$$= \frac{\partial}{\partial z_{il}} - \frac{1}{2} \sum_j^{N_i} \sum_r^{K_m} \hat{q}_{jr}^{(i)} (\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_{ir})^T \hat{\boldsymbol{\Sigma}}_{ir}^{-1} (\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_{ir}) \quad (100)$$

$$= \sum_j^{N_i} \sum_r^{K_m} \hat{q}_{jr}^{(i)} \boldsymbol{\mu}_{rl}^T \hat{\boldsymbol{\Sigma}}_{ir}^{-1} (\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_{ir}) \quad (101)$$

$$\frac{\partial}{\partial y_{il}} \tilde{J}_{LL}(\Theta, \Omega, \hat{q}) \quad (102)$$

$$= \frac{\partial}{\partial y_{il}} \frac{1}{2} \sum_j^{N_i} \sum_r^{K_m} \hat{q}_{jr}^{(i)} \{ \log |\hat{\boldsymbol{\Sigma}}_{ir}^{-1}| - (\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_{ir})^T \hat{\boldsymbol{\Sigma}}_{ir}^{-1} (\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_{ir}) \} \quad (103)$$

$$= \frac{1}{2} \sum_j^{N_i} \sum_r^{K_m} \hat{q}_{jr}^{(i)} \frac{\exp(y_{il})}{1 + \exp(y_{il})} \{ \text{tr}(\mathbf{C}_{rl} \mathbf{C}_{rl}^T \hat{\boldsymbol{\Sigma}}_{ir}) - (\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_{ir})^T \mathbf{C}_{rl} \mathbf{C}_{rl}^T (\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_{ir}) \} \quad (104)$$

6 PML-GMM Degenerates to PCA

In our parameterization, when GMM degenerates to Gaussian, $\mathbf{y}_i \ll \mathbf{0}$ and β_r is a constant, the KL divergence minimization is equivalent to minimizing $\|\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_i\|^2$. Here we prove the result. The objective function is

$$\tilde{J}_{CE} = \sum_i \{ (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})^T \hat{\boldsymbol{\Sigma}}_{ir}^{(-1)} (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir}) + \text{tr}(\hat{\boldsymbol{\Sigma}}_{ir}^{(-1)} \boldsymbol{\Sigma}_{ik}) - \log |\hat{\boldsymbol{\Sigma}}_{ir}^{(-1)}| \} \quad (105)$$

When $\mathbf{y} = \mathbf{0}$ and β_r is a constant, $\hat{\boldsymbol{\Sigma}}_{ir}^{(-1)} = \beta_r^2 \mathbf{I}$. The objective function becomes

$$\tilde{J}_{CE} \propto \sum_i \beta_r^2 (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})^T (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir}) \quad (106)$$

which is equivalent to minimize $\|\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_i\|^2$. If we give a constraint that $\{\mathbf{m}_{rl}\}_{l=1}^{d_z}$ is orthonormal, this optimization problem is the same as PCA.

7 Sub-Optimization Decomposition

7.1 Optimize Manifold Parameters

a) First look at $\mathbf{m}_{rl}$. We can write the term related with $\mathbf{m}_{rl}$ as

$$\sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})^T \hat{\boldsymbol{\Sigma}}_{ir}^{-1} (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir}) \quad (107)$$

Denote $\bar{\mathbf{M}}_r = [\mathbf{m}_{r1}; \dots; \mathbf{m}_{rd_z}] \in \mathbb{R}^{(D*d_z) \times 1}$ and $\mathbf{z}_i = [z_{i1} \mathbf{I}, \dots, z_{id_z} \mathbf{I}] \in \mathbb{R}^{D \times (D*d_z)}$. $\hat{\boldsymbol{\mu}}_{ir}$ can be written as

$$\hat{\boldsymbol{\mu}}_{ir} = \mathbf{z}_i \bar{\mathbf{M}}_r + \mathbf{b}_r \quad (108)$$

Take gradient w.r.t. $\bar{\mathbf{M}}_r$,

$$\frac{\partial}{\partial \bar{\mathbf{M}}_r} \sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})^T \hat{\boldsymbol{\Sigma}}_{ir}^{-1} (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir}) \quad (109)$$

$$= -2 \sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \mathbf{z}_i^T \hat{\boldsymbol{\Sigma}}_{ir}^{-1} (\boldsymbol{\mu}_{ik} - \mathbf{b}_r) + 2 \sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \mathbf{z}_i^T \hat{\boldsymbol{\Sigma}}_{ir}^{-1} \mathbf{z}_i \bar{\mathbf{M}}_r \quad (110)$$

So the optimal $\bar{\mathbf{M}}_r$ is

$$\bar{\mathbf{M}}_r = \left(\sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \mathbf{z}_i^T \hat{\boldsymbol{\Sigma}}_{ir}^{-1} \mathbf{z}_i \right)^{-1} \left\{ \sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \mathbf{z}_i^T \hat{\boldsymbol{\Sigma}}_{ir}^{-1} (\boldsymbol{\mu}_{ik} - \mathbf{b}_r) \right\} \quad (111)$$

b) For $\mathbf{b}_r$, the equations are about the same.

$$\frac{\partial}{\partial \mathbf{b}_r} \sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} (\boldsymbol{\mu}_{ik} - \mathbf{z}_i \bar{\mathbf{M}}_r - \mathbf{b}_r)^T \hat{\boldsymbol{\Sigma}}_{ir}^{-1} (\boldsymbol{\mu}_{ik} - \mathbf{z}_i \bar{\mathbf{M}}_r - \mathbf{b}_r) \quad (112)$$

$$= -2 \sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \hat{\boldsymbol{\Sigma}}_{ir}^{-1} (\boldsymbol{\mu}_{ik} - \mathbf{z}_i \bar{\mathbf{M}}_r) + 2 \sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \hat{\boldsymbol{\Sigma}}_{ir}^{-1} \mathbf{b}_r \quad (113)$$

So the optimal $\mathbf{b}_r$ is

$$\mathbf{b}_r = \left(\sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \hat{\boldsymbol{\Sigma}}_{ir}^{-1} \right)^{-1} \sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \hat{\boldsymbol{\Sigma}}_{ir}^{-1} (\boldsymbol{\mu}_{ik} - \mathbf{z}_i \bar{\mathbf{M}}_r) \quad (114)$$

$$= \left(\sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \hat{\boldsymbol{\Sigma}}_{ir}^{-1} \right)^{-1} \sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \hat{\boldsymbol{\Sigma}}_{ir}^{-1} (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir} + \mathbf{b}_r) \quad (115)$$

c) The derivative of $\mathbf{C}_r$ can be written as

$$\frac{\partial \tilde{J}_{CE}(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \hat{\mathbf{q}})}{\partial \mathbf{C}_r} \quad (116)$$

$$= \frac{\partial}{\partial \mathbf{C}_r} \frac{1}{2} \sum_i \sum_k \sum_r \pi_{ik} \hat{q}_{kr}^{(i)} \{ (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})^T \hat{\boldsymbol{\Sigma}}_{ir}^{(-1)} (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir}) + \text{tr}(\hat{\boldsymbol{\Sigma}}_{ir}^{(-1)} \boldsymbol{\Sigma}_{ik}) - \log |\hat{\boldsymbol{\Sigma}}_{ir}^{(-1)}| \} \quad (117)$$

$$= \sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \{ \boldsymbol{\Sigma}_{ik} \mathbf{C}_r \mathbf{Y}_i^{\log} + (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})(\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})^T \mathbf{C}_r \mathbf{Y}_i^{\log} - (\mathbf{C}_r \mathbf{Y}_i^{\log} \mathbf{C}_r^T + \beta_r^2 \mathbf{I})^{-1} \mathbf{C}_r \mathbf{Y}_i^{\log} \} \quad (118)$$

$$= \sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \{ \boldsymbol{\Sigma}_{ik} + (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})(\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})^T - (\mathbf{C}_r \mathbf{Y}_i^{\log} \mathbf{C}_r^T + \beta_r^2 \mathbf{I})^{-1} \} \mathbf{C}_r \mathbf{Y}_i^{\log} \quad (119)$$

$$(120)$$

Let the gradient be zero

$$\sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \{ \boldsymbol{\Sigma}_{ik} + (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})(\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})^T - (\mathbf{C}_r \mathbf{Y}_i^{\log} \mathbf{C}_r^T + \beta_r^2 \mathbf{I})^{-1} \} \mathbf{C}_r \mathbf{Y}_i^{\log} = 0 \quad (121)$$

$$\sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \{ \boldsymbol{\Sigma}_{ik} + (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})(\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})^T - (\mathbf{C}_r \mathbf{Y}_i^{\log} \mathbf{C}_r^T + \beta_r^2 \mathbf{I})^{-1} \} \mathbf{C}_r \mathbf{Y}_i^{\log} \mathbf{C}_r = 0 \quad (122)$$

$$\sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \{ \boldsymbol{\Sigma}_{ik} + (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})(\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})^T - (\mathbf{C}_r \mathbf{Y}_i^{\log} \mathbf{C}_r^T + \beta_r^2 \mathbf{I})^{-1} \} (\mathbf{C}_r \mathbf{Y}_i^{\log} \mathbf{C}_r + \beta_r^2 \mathbf{I} - \beta_r^2 \mathbf{I}) = 0 \quad (123)$$

$$\sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \{ \boldsymbol{\Sigma}_{ik} + (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})(\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})^T - (\mathbf{C}_r \mathbf{Y}_i^{\log} \mathbf{C}_r^T + \beta_r^2 \mathbf{I})^{-1} \} (\mathbf{C}_r \mathbf{Y}_i^{\log} \mathbf{C}_r + \beta_r^2 \mathbf{I}) = \quad (124)$$

$$\sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \{ \boldsymbol{\Sigma}_{ik} + (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})(\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})^T - (\mathbf{C}_r \mathbf{Y}_i^{\log} \mathbf{C}_r^T + \beta_r^2 \mathbf{I})^{-1} \} \beta_r^2 \mathbf{I} \quad (125)$$

$$\sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \{ \boldsymbol{\Sigma}_{ik} + (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})(\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})^T \} (\mathbf{C}_r \mathbf{Y}_i^{\log} \mathbf{C}_r^T + \beta_r^2 \mathbf{I}) - \sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \mathbf{I} = \quad (126)$$

$$\sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \{ \boldsymbol{\Sigma}_{ik} + (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})(\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})^T - (\mathbf{C}_r \mathbf{Y}_i^{\log} \mathbf{C}_r^T + \beta_r^2 \mathbf{I})^{-1} \} \beta_r^2 \mathbf{I} \quad (127)$$

The term $\{ \boldsymbol{\Sigma}_{ik} + (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})(\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})^T \} \beta_r^2 \mathbf{I}$ can be canceled.

$$\sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \{ \boldsymbol{\Sigma}_{ik} + (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})(\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})^T \} (\mathbf{C}_r \mathbf{Y}_i^{\log} \mathbf{C}_r^T) = \sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \mathbf{I} - \beta_r^2 \sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \hat{\boldsymbol{\Sigma}}_{ir} \quad (128)$$

We write the term $\mathbf{C}_r \mathbf{Y}_i^{\log}$ as

$$\mathbf{C}_r \mathbf{Y}_i^{\log} = \bar{\mathbf{Y}}_i^{\log} \bar{\mathbf{C}}_r \quad (129)$$

where $\bar{\mathbf{Y}}_i^{\log} = [\gamma(y_{i1}) \mathbf{I}, \dots, \gamma(y_{idy}) \mathbf{I}]$ and

$$\bar{\mathbf{C}}_r = \begin{bmatrix} \mathbf{C}_{r1} & 0 & \dots & 0 \\ 0 & \ddots & 0 & \vdots \\ 0 & \dots & \dots & \mathbf{C}_{r d_y} \end{bmatrix} \quad (130)$$

Bring this back to the equation, we have

$$\sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \{ \Sigma_{ik} + (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})(\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})^T \} (\bar{\mathbf{Y}}_i^{\log} \bar{\mathbf{C}}_r \mathbf{C}_r^T) = \sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \mathbf{I} - \beta_r^2 \sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \hat{\Sigma}_{ir} \quad (131)$$

Denote the term $\Sigma_{ik} + (\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})(\boldsymbol{\mu}_{ik} - \hat{\boldsymbol{\mu}}_{ir})^T = \boldsymbol{\Psi}_{kr}^{(i)}$

$$\sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \boldsymbol{\Psi}_{kr}^{(i)} \bar{\mathbf{Y}}_i^{\log} \bar{\mathbf{C}}_r \mathbf{C}_r^T = \sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \mathbf{I} - \beta_r^2 \sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \hat{\Sigma}_{ir} \quad (132)$$

It can be written by

$$\sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \boldsymbol{\Psi}_{kr}^{(i)} \bar{\mathbf{Y}}_i^{\log} \begin{bmatrix} \mathbf{C}_{r1} \mathbf{C}_{r1}^T \\ \vdots \\ \mathbf{C}_{rd_y} \mathbf{C}_{rd_y}^T \end{bmatrix} = \sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \mathbf{I} - \beta_r^2 \sum_i \sum_k \pi_{ik} \hat{q}_{kr}^{(i)} \hat{\Sigma}_{ir} \quad (133)$$

This problem can be solved by standard SPD solver.

d) $\mathbf{a}_r$ part. The gradient with respect to a_{rl} is

$$\frac{\partial \tilde{J}_{CE}(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \hat{\mathbf{q}})}{\partial a_{rl}} = \sum_i \sum_k \pi_{ik} (\hat{q}_{kr}^{(i)} - \hat{\pi}_{ir}) g(-\mathbf{w}_i^T \mathbf{a}_r) w_{il} \quad (134)$$

Use Newton-Raphson method, we need to calculate the second-order derivative, i.e., Hessian matrix of objective function w.r.t. $\mathbf{a}_r$.

$$\frac{\partial^2 \tilde{J}_{CE}(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \hat{\mathbf{q}})}{\partial a_{rl} \partial a_{rn}} = \frac{\partial}{\partial a_{rl}} \sum_i \sum_k \pi_{ik} (\hat{q}_{kr}^{(i)} - \hat{\pi}_{ir}) g(-\mathbf{w}_i^T \mathbf{a}_r) w_{in} \quad (135)$$

$$= \sum_i \sum_k \pi_{ik} \left[\hat{q}_{kr}^{(i)} g(-\mathbf{w}_i^T \mathbf{a}_r) g(\mathbf{w}_i^T \mathbf{a}_r) w_{il} - \hat{\pi}_{ir} g(-\mathbf{w}_i^T \mathbf{a}_r) g(\mathbf{w}_i^T \mathbf{a}_r) - g(-\mathbf{w}_i^T \mathbf{a}_r) \frac{\partial}{\partial a_{rl}} \hat{\pi}_{ir} \right] w_{in} \quad (136)$$

$$= \sum_i \sum_k \pi_{ik} \{ (\hat{q}_{kr}^{(i)} - \hat{\pi}_{ir}) g(-\mathbf{w}_i^T \mathbf{a}_r) g(\mathbf{w}_i^T \mathbf{a}_r) w_{il} + \hat{\pi}_{ir} g^2(-\mathbf{w}_i^T \mathbf{a}_r) w_{il} - \hat{\pi}_{ir}^2 g^2(-\mathbf{w}_i^T \mathbf{a}_r) w_{il} \} w_{in} \quad (137)$$

$$= \sum_i \sum_k \pi_{ik} \left[(\hat{q}_{kr}^{(i)} - \hat{\pi}_{ir}) g(\mathbf{w}_i^T \mathbf{a}_r) + \hat{\pi}_{ir} (1 - \hat{\pi}_{ir}) \right] g(-\mathbf{w}_i^T \mathbf{a}_r) w_{il} w_{in} \quad (138)$$

8 Implementation

8.1 Synthetic Data

When training PML-GMM, we set $N_v = 1, c_v = 0.2$. We initialize $\{\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Omega}}\}$ randomly and set initial values of $\hat{\mathbf{q}}$ as uniform, i.e. $\hat{q}_{kr}^{(i)} = 1/K_m$. Then alternate the optimization with respect to $\hat{\mathbf{q}}$ and $\{\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Omega}}\}$. The ground truth latent function is as follows

$$\Sigma_{ik} = y_{i1} \mathbf{C}_{k1}^{(gt)} + y_{i2} \mathbf{C}_{k2}^{(gt)}, \quad \boldsymbol{\mu}_{ik} = z_{i1} \boldsymbol{\mu}_{k1}^{(gt)} + z_{i2} \boldsymbol{\mu}_{k2}^{(gt)}, k = 1, \dots, 10, i = 1, \dots, 39, \quad (139)$$

where all $\mathbf{C}_{kl}^{(gt)}$ are positive definite matrix and y_{il} are positive numbers. This generation makes sure that the generated GMMs are valid. We run 5 random initializations and choose the one result with best objective function. In testing stage, we set the same hyperparameters as training stage and the same initialization method. We also run several random experiments and choose the best one.

8.2 Eye-Fixation Data

In eye-fixation data, we set $N_v = 6, c_v = 0.2$. Here we need to optimize w.r.t. hidden variables $\mathbf{v}$ so we need to regularize $\mathbf{v}$. The initialization method is the same as before. We run 5 random initializations and choose the best result with the best objective function.

8.3 Flow Cytometry Data

We model flow cytometry data as Gaussians and 2-component GMMs respectively. So for this dataset, we run experiments on Gaussians and GMMs. For learning on GMMs, We set $N_v = 4$. Other hyperparameters are the same as Section 8.2. We initialize parameters as before and run 5 experiments and choose the best one. In test stage, we use the same hyperparameters and initialization, and run several experiments to choose the best result. For learning on Gaussians, N_v is set to be 1 and other settings are the same as that of GMMs.

8.4 Social Check-in Data

Social checkin data are often modeled by kernel density estimation (KDE) and the number of components scales with checkin data (about 10^2 - 10^3). We apply PML-GMM on this data to test the scalability of PML w.r.t. the number of components. Using the Gowalla dataset Lichman and Smyth (2014), we sample subjects whose checkin numbers are larger than 100 and label their living cities by their frequent checkin locations. We obtain 56 subjects living in Los Angeles, Riverside, Irvine and San Diego and split them into 40 training and 16 testing GMMs. We reduce the number of GMM components to 100 by DPHEM Yu et al. (2018), which both simplifies the representation while preserving the density.

Figure 5a-d show latent spaces of the 4 methods. PML-GMM learns a latent space that is consistent with the fact that LA, Riverside and Irvine are close together and far from San Diego. While FINE can also give an organized latent space w.r.t. cities, it has worse classification accuracy than PML-GMM (see Table 1 in the paper). Figure 8 shows the reconstruction of GMMs by our method. The means are well-preserved in reconstructed GMMs, and the reconstructed covariances are at a reasonable scale.

8.5 Ablation Study

We run PML-GMM on data samples from 20 synthetic GMMs. We sample 300 points from these synthetic GMMs and run the experiment. N_v is set to be 1 and other hyperparameters are the same as Section 8.1. It turns out that the Direct-PML-GMM cannot learn the GMMs well.

References

- Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.
- Lei Yu, Tianyu Yang, and Antoni B Chan. Density-preserving hierarchical EM algorithm: Simplifying Gaussian mixture models for approximate inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (1):1–1, 2018.
- Moshe Lichman and Padhraic Smyth. Modeling human location data with mixtures of kernel densities. In *Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, pages 35–44, 2014.

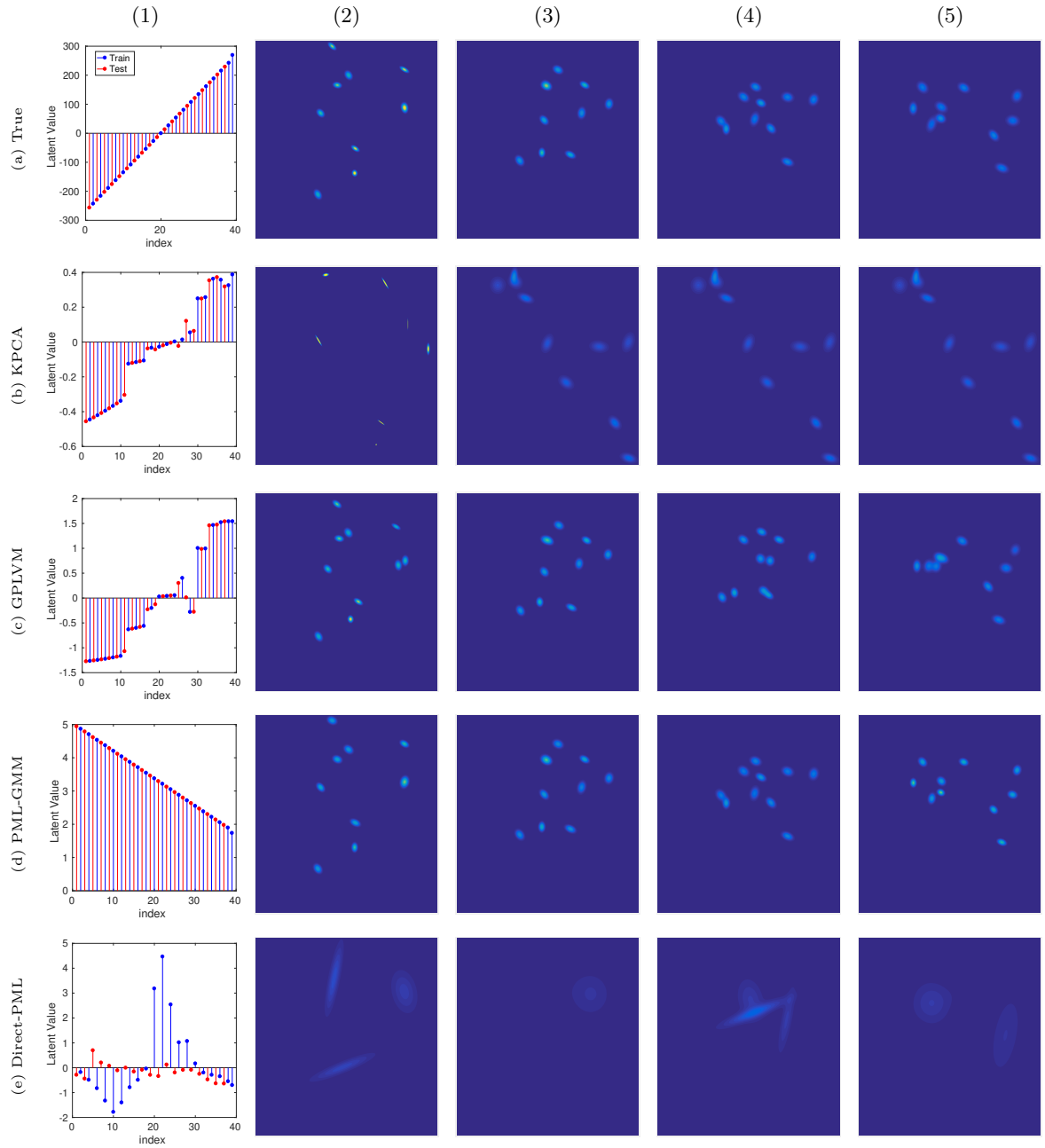


Figure 2: Experiment on Synthetic Data. The positions of latent variables are kept the same as that of ground truth. PML-GMM learns a smooth latent structure and achieve the best reconstruction.

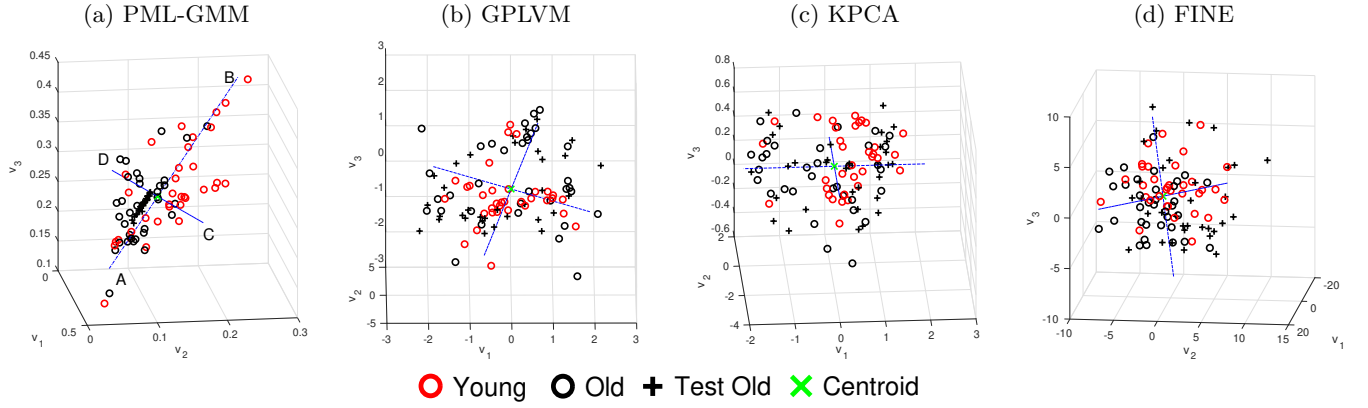


Figure 3: Latent spaces of Eye-Fixation Data. PML-GMM can learn meaningful regions corresponding to older (AD quadrant) and young (BC quadrant) adults, and embeds test data to the correct region.

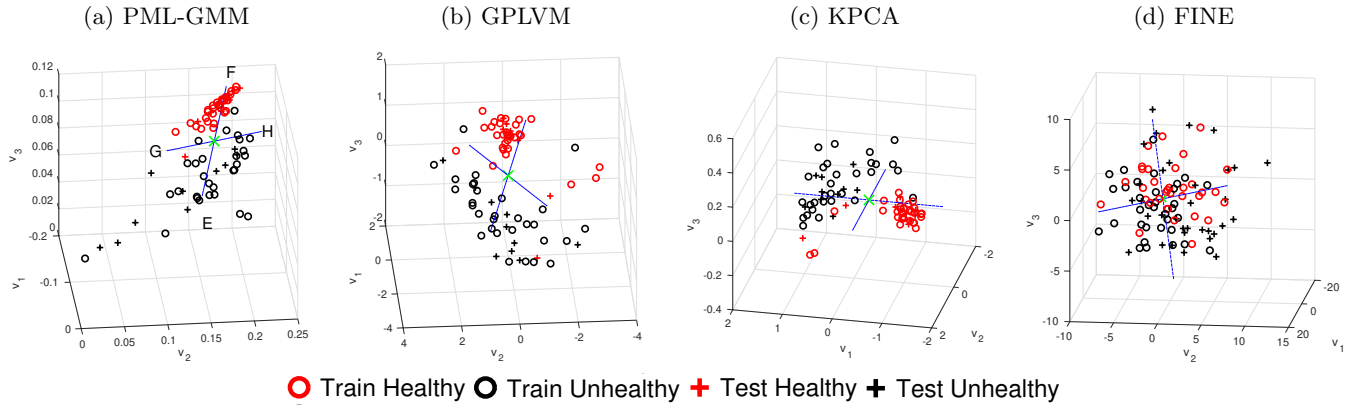


Figure 4: Latent spaces of flow cytometry data.

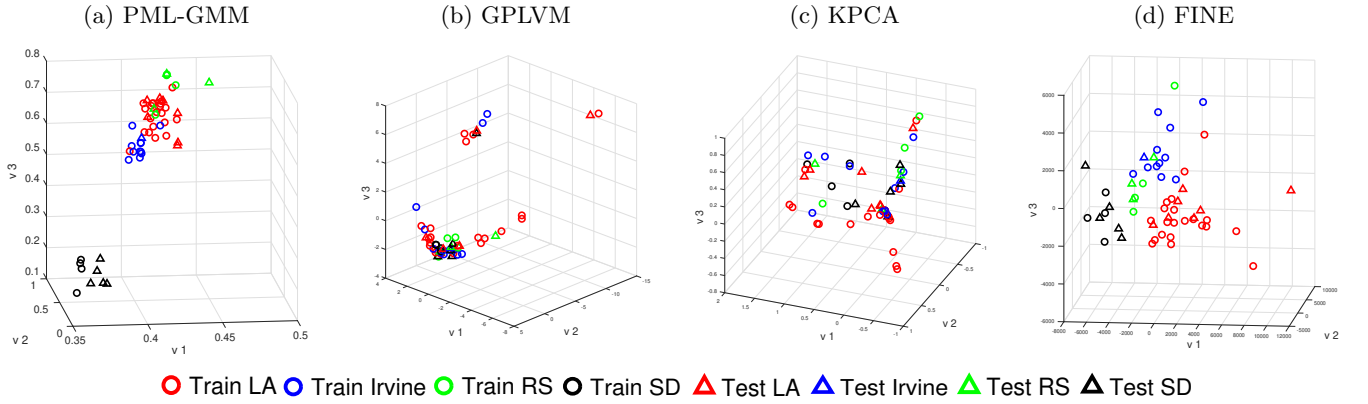


Figure 5: Latent spaces of Social Check-in Data.

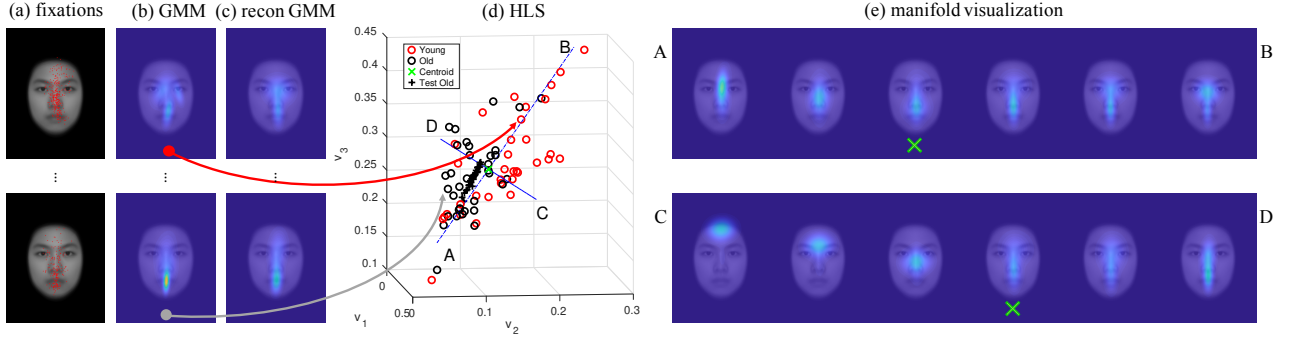


Figure 6: Manifold Visualization of Eye-Fixation Data. (a-c) show the reconstruction ability of PML-GMM. (e) shows the changes in eye-fixation regions of older adults (AD quadrant) to young ones (CB quadrant). Green crosses show GMMs nearest to the centroid.

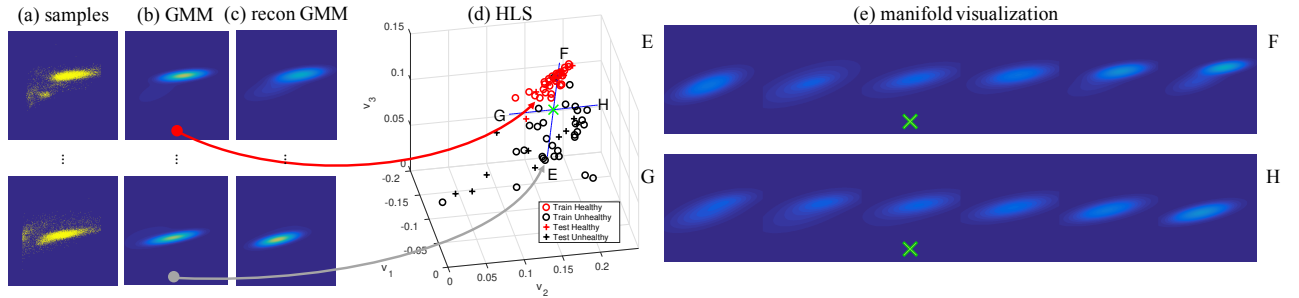


Figure 7: Manifold Visualization of Flow Cytometry Data. (a-c) show reconstruction ability of PML-GMM. (e) $\overline{EF}$ visualizes the change from unhealthy cells to healthy cells and $\overline{GH}$ visualizes the middle region between healthy and unhealthy.

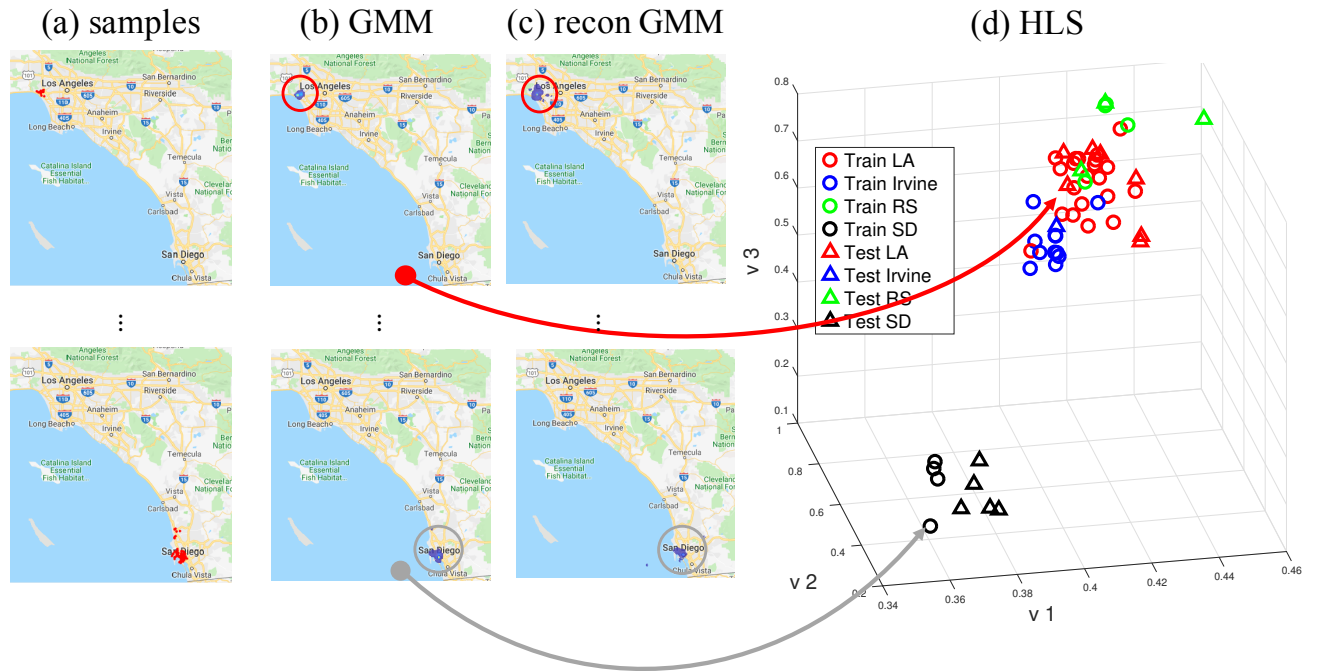


Figure 8: Reconstruction of Social Checkin Data.