# Clustering Report

Zijie Pan

# 1 Clustering

## 1.1 Experimental setup

**Notations** See [1].

**Dataset and models** Label column: "dec". Feature column: "attr", "sinc", "intel", "fun", "amb", "shar". Models: k-means and GMM with 3 types of inference — expectation maximization (GMM-EM), variational inference (GMM-VI), and Gibbs sampling (GMM-GS).

**Visualization via PCA** Fig. 1 shows the distribution of the dataset, where pc1 denotes the first principal component and pc2 denotes the second.

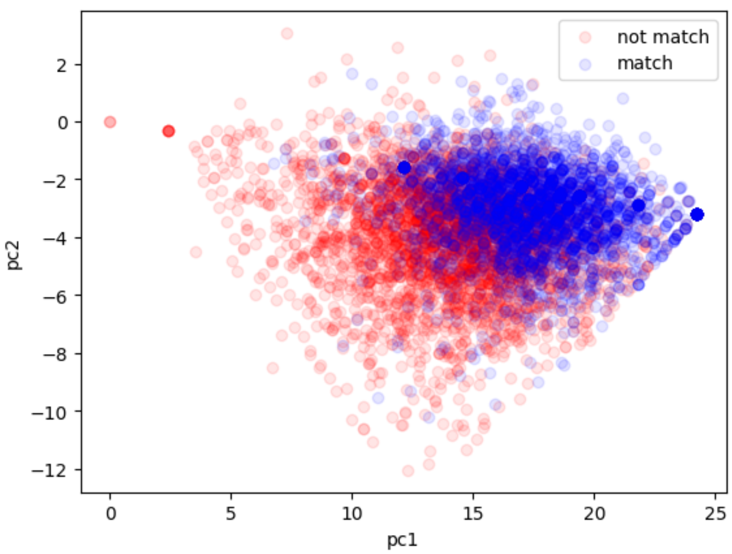



Figure 1: **Visualization via PCA.**

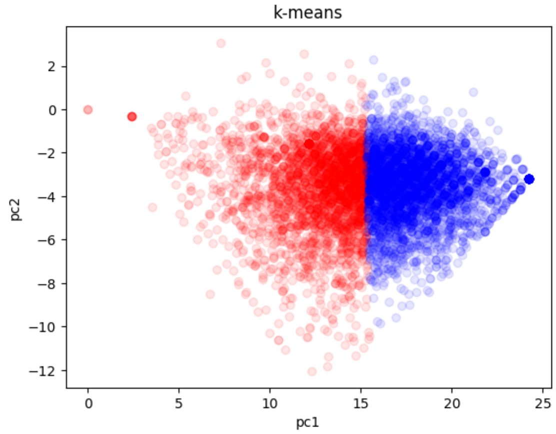
## 1.2 Results

**K-means**

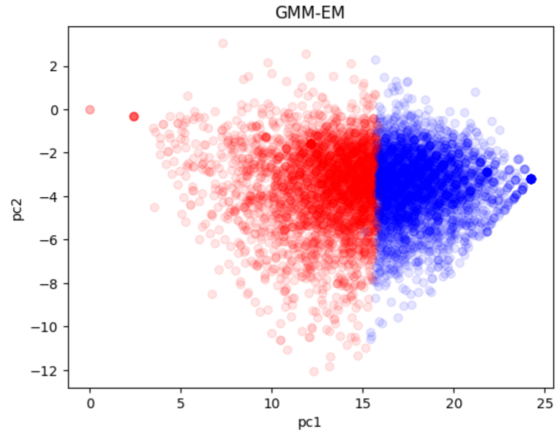
**GMM-EM**

**GMM-VI**

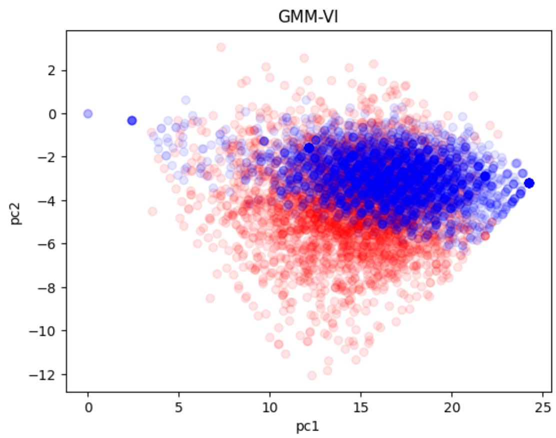
**GMM-GS**



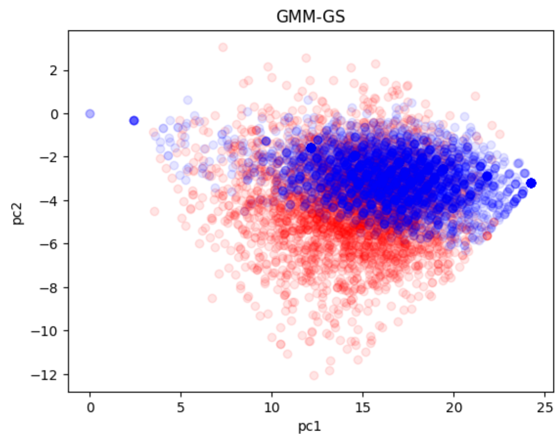

(a) K-means




(b) GMM-EM




(c) GMM-VI




(d) GMM-GS

Figure 2: **Clustering results.**

# References

[1] Kevin P Murphy. *Machine learning: a probabilistic perspective.* MIT press, 2012.

**Algorithm 1** GMM-EM

**input:** initial random labels $r_{ik}^0$
**for** $t = 0, 1, \cdots, T-1$
  M-step
  $\pi_l^{t+1} = \frac{\sum_i r_{il}^t}{n}$
  $\mu_l^{t+1} = \frac{\sum_i r_{il}^t x_i}{\sum_i r_{il}^t x_i}$
  $\Sigma_l^{t+1} = \frac{\sum_i r_{il}^t (x_i - \mu_l^{t+1})(x_i - \mu_l^{t+1})^\top}{\sum_i r_{il}^t}$
  E-step
  $r_{ik}^{t+1} = \frac{\pi_k^t N(x_i|\mu_k^t, \Sigma_k^t)}{\sum_l \pi_l^t N(x_i|\mu_l^t, \Sigma_l^t)}$
**end for**

**Algorithm 2** GMM-VI

**input:** initial random labels $r_{ik}$ and prior parameters: Dirichlet parameter $\alpha_0$, Gaussian parameter $m_0, \beta_0$, Wishart parameter $L_0, v_0$
**for** $t = 0, 1, \cdots, T-1$
  M-like step
  $N_k = \sum_i r_{ik}$
  $\bar{x}_k = \frac{1}{N_k} \sum_i r_{ik} x_i$
  $S_k = \frac{1}{N_k} \sum_i r_{ik}(x_i - \bar{x}_k)(x_i - \bar{x}_k)^\top$
  $\beta_k = \beta_0 + N_k$
  $m_k = \frac{1}{\beta_k}(\beta_0 m_0 + N_k \bar{x}_k)$
  $L_k^{-1} = L_0^{-1} + N_k S_k + \frac{\beta_0 N_k}{\beta_0 + N_k}(\bar{x}_k - m_0)(\bar{x}_k - m_0)^\top$
  $\nu_k = \nu_0 + N_k$
  $\alpha_k = \alpha_0 + N_k$
  E-like step
  $\tilde{\pi}_k = \exp\{\psi(\alpha_k) - \psi(\sum_l \alpha_l)\}$
  $\tilde{\Lambda}_k = \exp\left\{\sum_j \psi(\frac{\nu_k + 1 - j}{2}) + d \log 2 + \log\det L_k\right\}$
  $r_{ik} \propto \tilde{\pi}_k \tilde{\Lambda}_k^{\frac{1}{2}} \exp\left\{-\frac{1}{2} d \beta_k^{-1} - \frac{1}{2}\nu_k (x_i - m_k)^\top L_k (x_i - m_k)\right\}$
**end for**

**Algorithm 3** GMM-GS

**input:** initial $\alpha, m_0, V_0, S_0, \nu_0$ and $\pi, \mu_k, \Sigma_k$
**for** $t = 0, 1, \cdots, T-1$
  Sampling z
  $z_i \sim Multinomial(\pi_k N(x_i|\mu_k, \Sigma_k), k = 1, \cdots, K)$
  Sampling $\pi$
  $N_k = \sum_i 1(z_i = k)$
  $\pi \sim Dir(\alpha_1 + N_1, \cdots, \alpha_k + N_k)$
  Sampling $\mu_k$
  $V_k^{-1} = V_0^{-1} + N_k \Sigma_k$
  $m_k = V_k(\Sigma_k^{-1} \sum_{i:z_i=k} x_i + V_0^{-1} m_0)$
  $\mu_k \sim N(m_k, V_k)$
  Sampling $\Sigma_k$
  $\bar{x}_k = \frac{1}{N_k} \sum_{i:z_i=k} x_i$
  $S_k = \sum_{i:z_i=k}(x_i - \bar{x}_k)(x_i - \bar{x}_k)^\top$
  $\Sigma_k \sim IW(S_0 + S_k, \nu_0 + N_k)$
**end for**