Robustifying mixture models using maximum mean discrepancy

Adam Farooq[(1)], Yordan P. Raykov[(2)]
(1) Aston University, farooq.adam@outlook.com (2) University of Nottingham, yordan.raykov@nottingham.ac.uk

Introduction

Bayesian inference delivers principled rules for learning from data and integrating out uncertainty. However, standard Bayesian inference can be meaningfully interpreted only when the data generating mechanism is within the family of models defined by prior and likelihood. The growing complexity of data problems, has motivated intuitive generalisations of the Bayesian paradigm for dealing with situations when such an assumption does not necessarily hold. More notable strategies include:

(i) likelihood tempering and coarsened posteriors;
(ii) the Rule of Three;
(iii) Posterior Bootstrapping

Common conjectures in deriving such strategies are that:

- either the likelihood or the prior **only** are mis-specified and the full joint likelihood is mis-specified with a single parameter: $\mathrm{P}(\theta|y) \propto P(\theta)P(y|\theta)^{\epsilon}$
- the Kullback-Leibler divergence is an appropriate measure of discrepancy between the modelling and observed densities
- assuming correctly specified model facilitates optimal inference

In this work, we study a generic framework for doing inference in mixture models, under the assumption that the different random variables in the graphical model are mis-specified and only specified within a maximum mean discrepancy r-neighbourhood of the observed realizations. We demonstrate some examples where incorporating an assumption that component distributions are mis-specified leads to more efficient inference and better maximum-a-posteriori clustering. Finally, we propose a mixture modelling framework which uses 'pseudo-points' to define the density of each component, rather then exponential family parametric models.

Maximum mean discrepancy

The maximum mean discrepancy (MMD) can be used to find the statistical distance between two distributions by comparing the family of summary statistics between the two. Let P and Q be two Borel probability measures on some topological space $\mathcal{Y}$, then the MMD between the two is:

$$\mathrm{MMD}_{\mathcal{H}}[P,Q] = \left\|\mu_P - \mu_Q\right\|_{\mathcal{H}}$$

where μ_P and μ_Q are the 'mean embeddings' of the respective distributions in the reproducing kernel Hilbert space (RKHS) $\mathcal{H}$. One advantage of the MMD is that it can be approximated :

$$\mathrm{MMD}_{\mathcal{H}}[P,Q]^2 \approx \frac{1}{N_P(N_P-1)}\sum_{i\neq j} k\left(y_i^{(P)}, y_j^{(P)}\right) + \frac{1}{N_Q(N_Q-1)}\sum_{i\neq j} k\left(y_i^{(Q)}, y_j^{(Q)}\right) - \frac{2}{N_P N_Q}\sum_{i=1}^{N_P}\sum_{j=1}^{N_Q} k\left(y_i^{(P)}, y_j^{(Q)}\right)$$

Where $k(\cdot,\cdot)$ is a 'kernel function' in the RKHS, a common choice is the Gaussian kernel: $k(y_i, y_j) = \exp\left(-\frac{\|y_i - y_j\|_2^2}{\gamma}\right), \gamma > 0$

$$y_{1,..,N_P}^{(P)} \overset{\text{i.i.d.}}{\sim} P \qquad y_{1,..,N_Q}^{(Q)} \overset{\text{i.i.d.}}{\sim} Q$$

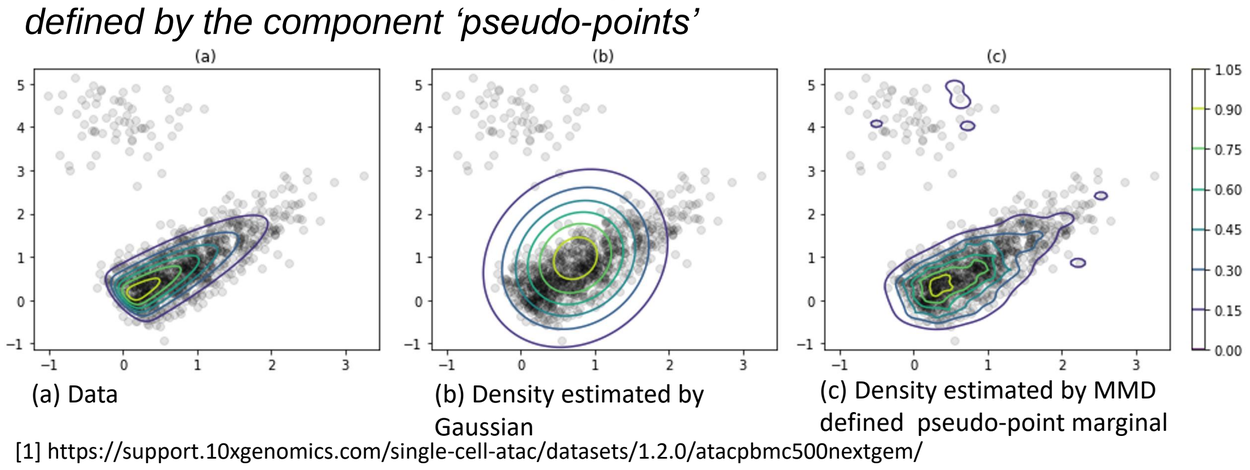
Robust MMD mixtures: Assume the augmented complete data likelihood:

$$\mathrm{P}(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{n=1}^{N}\sum_{k} \pi_k \mathrm{P}(\boldsymbol{y}_n | \mathrm{MMD}_{\mathcal{H}}[\boldsymbol{\theta}_k^*, \boldsymbol{\theta}_k]^2 < r)$$

Mixtures of MMD defined pseudo-point marginals

Assume each component k is parametrized by M pseudo-points $u_{1,\dots,M}^{(k)}$ and the assignment probabilities depends on the distributional distance: $\mathrm{MMD}_{\mathcal{H}}\left[P_{\boldsymbol{y}}, P_{\boldsymbol{u}}\right]^2$, where $P_{\boldsymbol{u}} \approx \frac{1}{M}\sum_{m=1}^{M}\delta_{\boldsymbol{u}_m}$ and $P_{\boldsymbol{y}} \approx \frac{1}{N}\sum_{n=1}^{N}\delta_{\boldsymbol{y}_n}$

Define component likelihood using MMD between the distribution defined by the component 'pseudo-points'



(a) Data (b) Density estimated by Gaussian (c) Density estimated by MMD defined pseudo-point marginal

[1] https://support.10xgenomics.com/single-cell-atac/datasets/1.2.0/atacpbmc500nextgem/

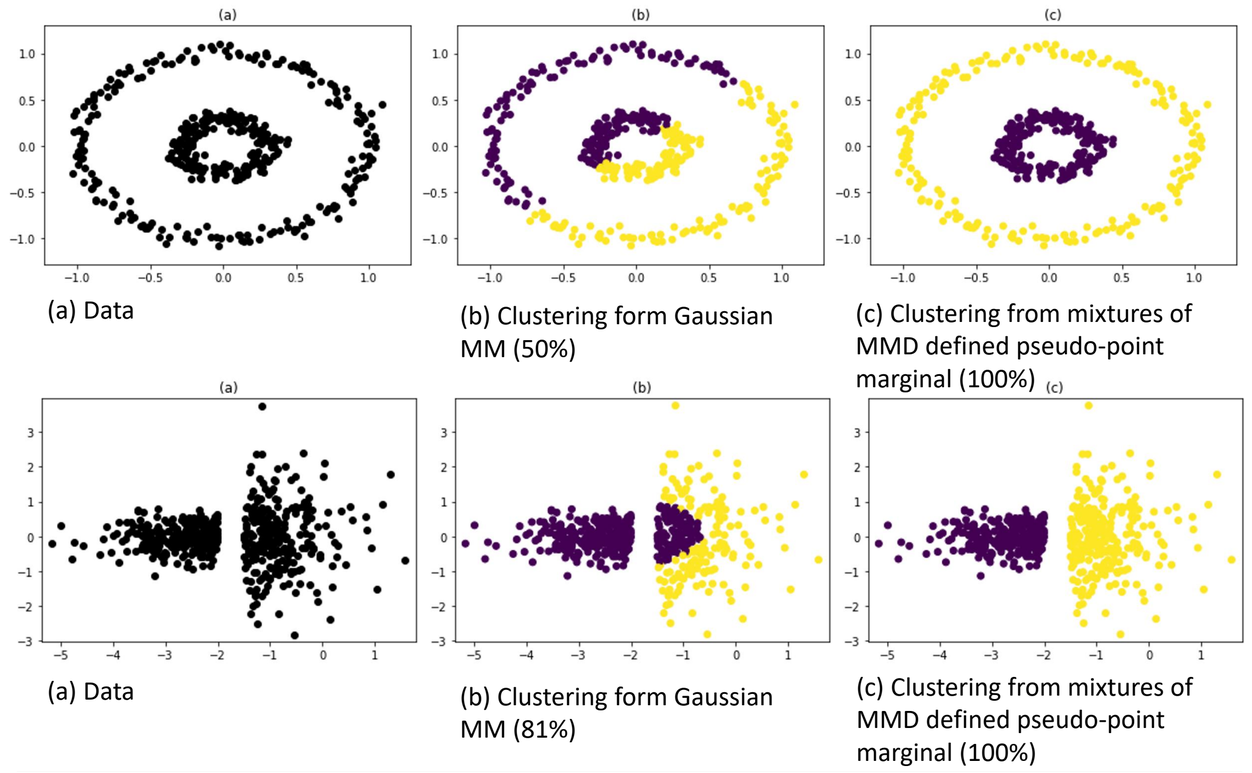
Our inference can iterate between updating the assignment probabilities given the robust component density terms and the mixing parameters and updating the component parameters or 'pseudo-points'.
If we denote the component assignments with $c_n \in \{1,\dots,K\}$, the assignment probabilities are updated using:

$$\mathrm{P}(c_n = k|\boldsymbol{\rho}, \boldsymbol{u}) = \frac{\mathrm{P}(c_n = k) \times \rho\left(\mathrm{MMD}_{\mathcal{H}}\left[\delta_{y_n}, P_{\boldsymbol{u}}^{(k)}\right]^2\right)}{\sum_{\boldsymbol{k}} \mathrm{P}(c_n = k) \times \rho\left(\mathrm{MMD}_{\mathcal{H}}\left[\delta_{y_n}, P_{\boldsymbol{u}}^{(k)}\right]^2\right)}$$

if we are not assuming a relaxation on the mixing probabilities $P(c_n = k)$.

In the proposed mixture density framework, the modeller has two make two distinctive assumptions: (i) one for the expected component density/basis $\rho(\cdot)$ and (ii) one for the most appropriate RKHS $\mathcal{H}$ to incorporate parameter invariance.

Defining flexible components using appropriate reproducing kernel Hilbert space



(a) Data (b) Clustering form Gaussian MM (50%) (c) Clustering from mixtures of MMD defined pseudo-point marginal (100%)

(a) Data (b) Clustering form Gaussian MM (81%) (c) Clustering from mixtures of MMD defined pseudo-point marginal (100%)

Results

We demonstrate a clustering application on assay for transposase-accessible chromatin using sequencing (ATAC-seq) data from 10X Genomes [1] internal single-cell demonstration data set of peripheral blood mononuclear cells (PBMCs) from a healthy donor with N = 482 cells and D = 26, 216 peaks. To avoid cluttering the sequencing depth we adopt a cosine kernel measures to define the mixtures of MMD defined pseudo-points marginal.

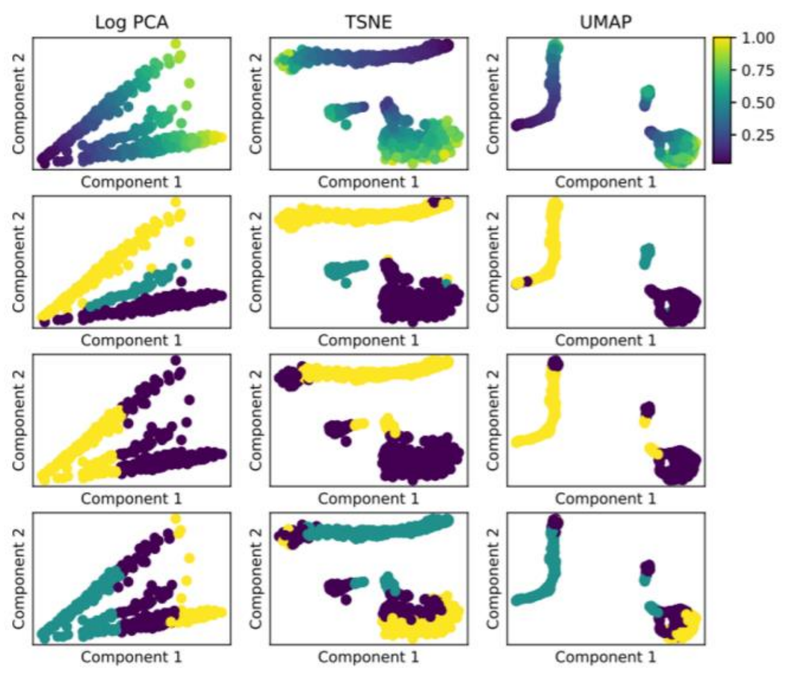


Figure 1: Plot of single-cell ATAC-seq data used visualised using PCA (left column), TSNE (middle column), and UMAP (right column). Data is coloured according to cell-specific sequencing (1st row). Data is coloured by clusters inferred from mixtures of MMD defined pseudo-points marginal (2nd row), negative-binomial mixture model (3rd row), and K-means (4th row).

Conclusion

- A simple framework for making explicit robustness assumptions on the different parameters of mixture models.
- Flexible and interpretable component densities can be captured using only component 'pseudo-points'
- With appropriate RKHS we can incorporate different invariance properties during inference.
- Main uses: (i) robust inference when our model is mis-specified; (ii) scalable inference even in when our model is well specified.

@_adamfarooq
@JordanRaykov