# Discussion: A Generalized Bayes Framework for Probabilistic Clustering

Jizhou Kang

University of California, Santa Cruz
Department of Statistics

May 20th, 2024

# Motivation

- Focus on clustering, finding a partition of the observed data.
- Algorithm-based approaches: K-means and its generalizations; Key disadvantage: lack of uncertainty quantification.
- Model-based approaches: mixture models; Key disadvantage: sensitive to misspecification of the kernel, and also computationally expensive.
- Bridge these two approaches, providing a probability distribution that characterizes uncertainty in clustering without requiring specification of a likelihood.

# Generalized Bayes Product Partition Models (GB-PPM)

- Assume factorized loss

$$\ell(c; X) = \sum_{k=1}^{K} \sum_{i \in C_k} \mathcal{D}(x_i; X_k), \quad c : |C| = K.$$

  - $\mathcal{D}(x_i; X_k)$ quantifies the discrepancy between the $i$th unit and the $k$th cluster.
  - $K$ is predetermined, and we only consider partitions of $n$ data into $K$ groups.

- Focus on uniform clustering prior: $\pi(c) = 1/\mathcal{S}(n, K)$.
- GB-PPM is defined as

$$\pi(c \mid \lambda, X) \propto \pi(c) \prod_{k=1}^{K} \rho(C_k; \lambda, X_k) \propto \prod_{k=1}^{K} \exp\{-\lambda \sum_{i \in C_k} \mathcal{D}(x_i; X_k)\}$$

# Decision-theoretic Justification

- Target of inference: $c_{opt}$, a partition of $n$ data into $K$ groups, such that the integrated loss $\sum_{k=1}^{K}\sum_{i\in C_k}\mathrm{E}_{\pi_0}\{\mathcal{D}(x_i;X_k)\}$ is minimized. Expectation taken w.r.t. DGP $\pi_0(X)$.
- Goal: find an optimal way to quantify our subjective beliefs about $c_{opt}$, conditioning on data.
- Define a loss function regarding one's posterior belief about $c_{opt}$,

$$\mathcal{L}\{\nu(c)\}=\lambda E_{\nu}\{\ell(c;X)\}+KL\{\nu(c)\|\pi(c)\}$$

  Then GB-PPM is the unique minimizer of $\mathcal{L}\{\nu(c)\}$.
- Why define $\mathcal{L}$ like this? It balances two components: discrepancy w.r.t. the observed data, and closeness to the prior.

# Choice of loss

- A given loss favors the formation of specific cluster shapes, leading to a specific optimal partition $c_{opt}$.
- Choosing the loss function align with the goal, e.g., if we want to identify spherical clusters, then may consider using the k-means loss.
- What if the loss is misspecified? From GB-PPM we still get the optimal posterior belief update about $c_{\text{opt}}$, but that $c_{\text{opt}}$ may be a bad summary of important groups in the data.
- Choose the number of clusters $K$ is part of specifying the loss. At which resolution do we want to partition the observations?

# Posterior Point Estimate

- Let $\hat{c}_{MAP} = \arg\max_c \pi(c \mid \lambda, X)$, then $\hat{c}_{MAP} = \arg\min_c \sum_{k=1}^{K} \sum_{i \in C_k} \mathcal{D}(x_i; X_k)$. That is, $\hat{c}_{MAP}$ is the empirical risk minimizer.

- A trivial result relies on the uniform prior.

# Gibbs sampler

- Let $c_{-i}$ denote the cluster indicators without the $i$th unit, $C_{k,-i}$ be the associated partition, $X_{k,-i}$ denote the data, and $\rho(C_{k,-i}; \lambda, X_{k,-i}) = \exp\{-\lambda \sum_{j \in C_{k,-i}} D(x_j; X_{k,-i})\}$ the cohesion functions.

- The conditional distribution of $c_i$ given $c_{-i}$ is

$$\Pr(c_i = k \mid c_{-i}, \lambda, X) \propto \frac{\rho(C_k; \lambda, X_k)}{\rho(C_{k,-i}; \lambda, X_{k,-i})}.$$

- Intuitively, the $i$th unit is likely to be allocated to the $k$th cluster if the cohesion of the newly created cluster is higher.

- If a new data point $x_{n+1}$ becomes available, the allocation is done in a similar manner.

# Calibration of $\lambda$

- Consider a prior $\pi(\lambda)$ and a joint loss $\ell(c, \lambda; X) = \lambda \ell(c; X) - \xi \log(\lambda)$, the joint Gibbs posterior is
$$\pi(c, \lambda \mid \tilde{\lambda}, X) \propto \pi(c)\pi(\lambda) \exp\{-\tilde{\lambda}\ell(c, \lambda; X)\}$$

- Transforms the problem from choosing $\lambda$ to calibrating $\tilde{\lambda}$ and $\xi$.

- For a broad class of loss functions, there are default choices of $\tilde{\lambda}$ and $\xi$, obtained by the connection of losses and modified loglikelihoods.

# Bergman Distance and Mixture of Exponential Dispersion

- Write the likelihood of data as mixture of exponential dispersion distributions, with location parameter $\theta_k$ and scale parameter $\lambda$;

- Plug in the MLE of $\theta_k$ (or an adjusted version) to obtain the profile likelihood.

- There is a connection between that profile likelihood and GB-PPM with Bregman divergence, and we use that connection to specify $\pi(\lambda)$, $\tilde{\lambda}$ and $\xi$.

# Pairwise dissimilarities and Spherical Distributions

- Write the composite likelihood of data using pairwise dissimilarities and spherical distribution;
- There is a connection between that composite likelihood and GB-PPM with pairwise dissimilarities.
- We can use that connection to specify $\pi(\lambda)$, $\tilde{\lambda}$ and $\xi$.

# Clustering with Minkowski distance

- Consider model data through pairwise difference likelihood (a type of composite likelihood), and the $L^p$ spherical distribution,

$$p(X \mid \lambda) = \prod_{k=1}^{K} \prod_{i \in C_k} \{ \prod_{i' \in C_k} \pi_{sp}(x_i - x_i \mid \lambda) \}^{1/n_k}, \; \pi_{sp}(x_i - x_i \mid \lambda) \propto \lambda^d \exp\{-\frac{\lambda}{2} \|x_i - x_{i'}\|_p\}$$

- Related to a GB-PPM with pairwise dissimilarity cohesion:

$$\pi(c \mid \lambda, X) \propto \prod_{k=1}^{K} \exp\{-\frac{\lambda}{2} \sum_{i \in C_k} \frac{1}{n_k} \sum_{i' \in C_k} \|x_i - x_{i'}\|_p\}$$

- Therefore, we can set $\tilde{\lambda} = 1$, $\xi = nd$, and $\pi(\lambda) \sim Gamma(a_\lambda, b_\lambda)$.

# Clustering Collection of Binary Indicators

- Consider data $\mathbf{x}_i = (x_{i1}, \cdots, x_{id})^\top$, where $x_{ij} \in \{0, 1\}$.
- Model with a $K$ component mixture of multivariate Binomial (Bernoulli) distribution,

$$p(X \mid \boldsymbol{\theta}) = \prod_{k=1}^{K} \prod_{i \in C_k} \exp[\sum_{j=1}^{d} \{x_{ij} \log(\frac{\theta_{kj}}{1 - \theta_{kj}}) + \log(1 - \theta_{kj})\}]$$

- Consider the MLE of $\theta_{kj}$, given by $\bar{x}_{kj}$, the within-cluster proportion.
- We have a well-defined GB-PPM, correspond to the profile likelihood:

$$\pi(c \mid \lambda, X) \propto \prod_{k=1}^{K} \prod_{i \in C_k} \exp[\sum_{j=1}^{d} \{x_{ij} \log(\frac{\bar{x}_{kj}}{1 - \bar{x}_{kj}}) + \log(1 - \bar{x}_{kj})\}].$$

- Use adjusted centroids $\tilde{x}_{kj}$ to avoid boundary issues.
- Fix $\lambda = 1$ because there is no scale parameter.

# Clustering with Squared Euclidean distance (Bergman divergence perspective)

- Consider model $X$ with mixture of Gaussians:

$$p(X \mid \theta, \lambda) \propto \prod_{k=1}^{K} \prod_{i \in C_k} \exp[-\lambda \|x_i - \theta_k\|^2].$$

- Plug in MLE of $\theta_k$, given by $\bar{x}_k$, we obtain

$$p(c \mid \lambda, X) \propto \prod_{k=1}^{K} \prod_{i \in C_k} \exp[-\lambda \|x_i - \bar{x}_k\|^2].$$

- Therefore, we can set $\tilde{\lambda} = 1$, $\xi = nd/2$ and $\lambda \sim Gamma(a_\lambda, b_\lambda)$.

# Clustering with Squared Euclidean distance (pairwise dissimilarity perspective)

- Consider model $X$ with pairwise difference likelihood, and a Gaussian density:

$$p(X \mid \theta, \lambda) \propto \prod_{k=1}^{K} \prod_{i \in C_k} \{ \prod_{i' \in C_k} \sqrt{\lambda} \exp(-\frac{\lambda}{2} \|x_i - x_{i'}\|^2) \}^{1/n_k}.$$

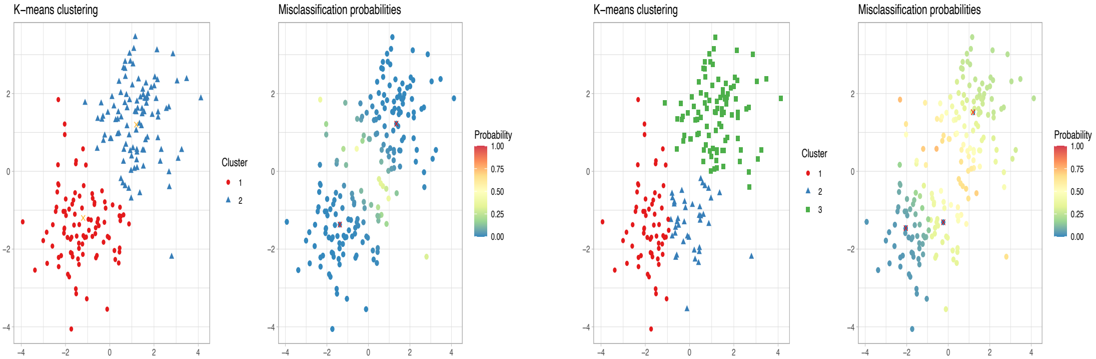
- Related to a GB-PPM with pairwise dissimilarity cohesion:

$$p(c \mid \lambda, X) \propto \prod_{k=1}^{K} \exp\{-\frac{\lambda}{2} \sum_{i \in C_k} \frac{1}{n_k} \sum_{i' \in C_k} \|x_i - \bar{x}_k\|^2\}.$$

- Therefore, we can set $\tilde{\lambda} = 1$, $\xi = nd/2$ and $\lambda \sim Gamma(a_\lambda, b_\lambda)$.

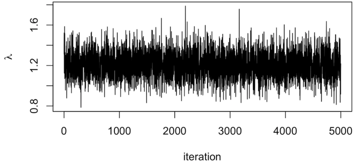
# A toy example

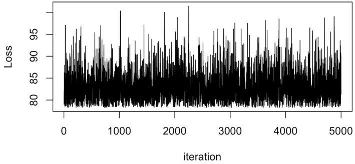
- We generate $n = 200$ data, each data is of the form $\mathbf{x}_i = (x_{i1}, x_{i2})$.
- For $i = 1, \ldots, 100$, $\mathbf{x}_i \sim N(\boldsymbol{\mu}_1, \mathbf{I}_2)$, and for $i = 101, \ldots, 200$, $\mathbf{x}_i \sim N(\boldsymbol{\mu}_2, \mathbf{I}_2)$.
- Try $K = 2$ and $K = 3$.



# *USArrests* data

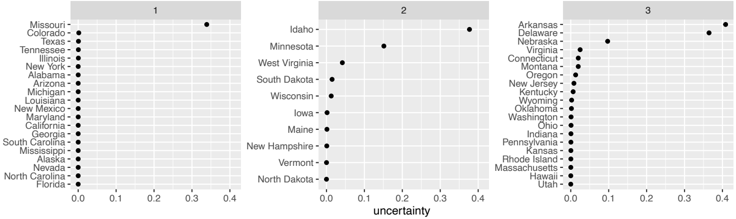
- This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape, in each of the 50 US states in 1973, together with the population living in urban areas. $X \in \mathbb{R}^{50\times 4}$
- Model-based approach based on finite mixture of multivariate Gaussian distributions. Using "Mclust" function in R, the optimal model selected has $K = 3$ clusters, and mixing kernel $N(\boldsymbol{\mu}_k, \lambda_k \mathbf{A})$, where $\mathbf{A}$ is diagonal and $\det(\mathbf{A}) = 1$.
- Loss-based generalized Bayesian clustering based on squared Euclidean distance. Prior on $\lambda$ is $\pi(\lambda) \propto \frac{1}{\lambda}$.
- Posterior trace plot of $\lambda$ and $\ell(c; X)$:



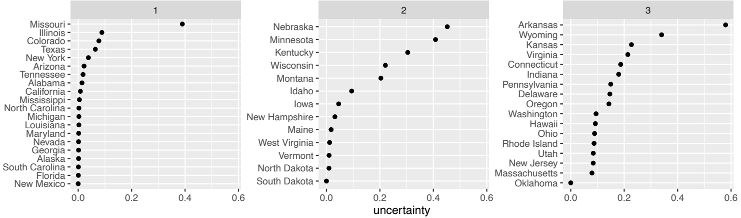


# Results (misclassification probabilities)

- Model-based clustering:

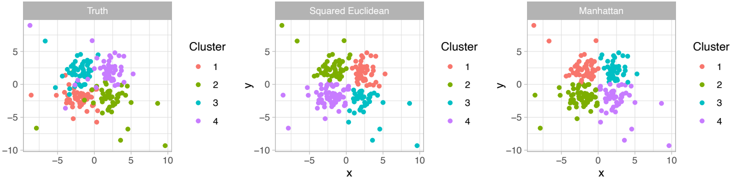


- Generalized Bayesian clustering:

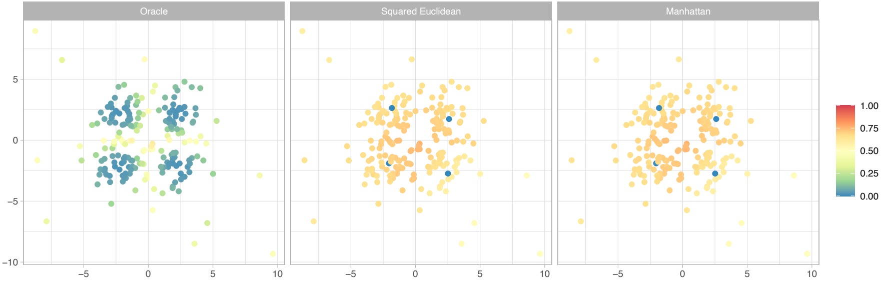


# Robust clustering (Section 7.2)

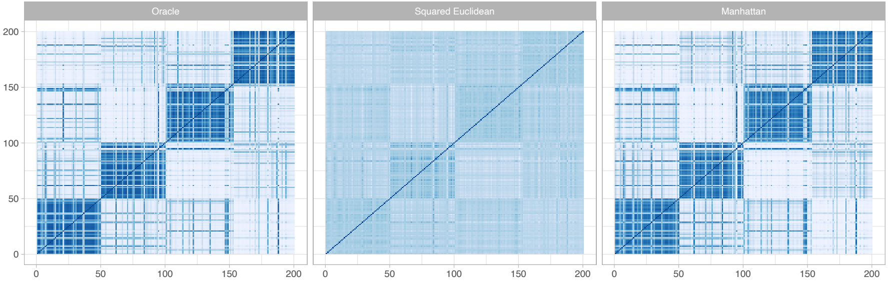
- Generate $n = 200$ data, each data is of the form $\mathbf{x}_i = (x_{i1}, x_{i2})$.
- Evenly divided into $K = 4$ clusters, each with 50 data points.
- Within each partition we assume $(x_i \mid \boldsymbol{\mu}_k, \sigma^2, c_i = k) \overset{i.i.d.}{\sim} t_2(\boldsymbol{\mu}_k, \sigma^2 \mathbf{I}_2)$.
- Two GB methods: Squared Euclidean distance or Manhattan distance.
- Cluster membership:



# Results (misclassification probabilities)



# Results (coclustering probabilities)



# Binary clustering

- Mimic the real data example presented in the paper. We consider clustering vectors with binary entries.

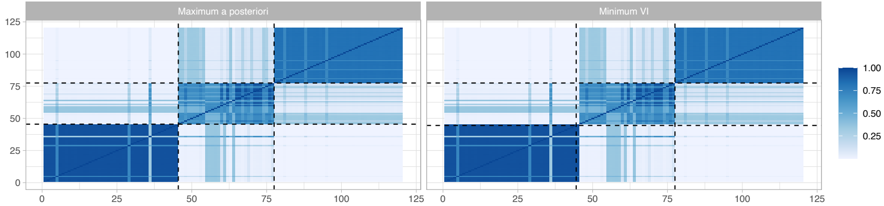
- Suppose $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})$, each entry is either 0 or 1.

- Consider $K = 3$ clusters:
    - For cluster 1, generate $N_1 = 50$ data from $\mathbf{p}_1 = (0.99, 0.98, 0.97, 0.86)$;
    - For cluster 2, generate $N_2 = 20$ data from $\mathbf{p}_2 = (0.56, 0.65, 0.72, 0.48)$;
    - For cluster 3, generate $N_3 = 50$ data from $\mathbf{p}_3 = (0.01, 0.02, 0.05, 0.06)$.

- Case 1 indicates agreement on presence; Case 2 indicates unclear diagnoses; Case 3 indicates agreement on absence.

# Results

- Coclustering probability:



- Adjusted centroids associated with each cluster:

Table: Maximum a posterior

| | 1 | 2 | 3 | 4 | $N_k$ |
|---|---|---|---|---|---|
| 1 | 0.99 | 0.97 | 0.99 | 0.95 | 45 |
| 2 | 0.59 | 0.41 | 0.08 | 0.77 | 32 |
| 3 | 0.01 | 0.01 | 0.08 | 0.01 | 43 |

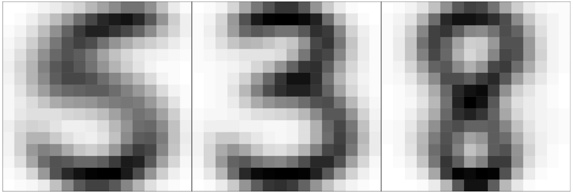
Table: Minimum VI

| | 1 | 2 | 3 | 4 | $N_k$ |
|---|---|---|---|---|---|
| 1 | 0.99 | 0.99 | 0.99 | 0.94 | 44 |
| 2 | 0.60 | 0.40 | 0.10 | 0.78 | 33 |
| 3 | 0.01 | 0.01 | 0.08 | 0.01 | 43 |

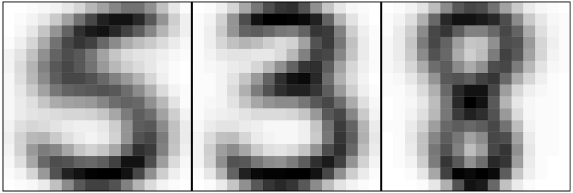
# High dimensional clustering

- Clustering on the USPS handwritten digit dataset. We will use a sample of the handwritten digits consisting of only the number 3, 5, and 8.
- The data set contains 1756 samples. Each of them is a vector with 256 entries, corresponding to the pixels in the $8 \times 8$ image.
- Model based clustering using GMM, with covariance matrix of the form $\lambda_k \mathbf{I}$.

**True centroid**



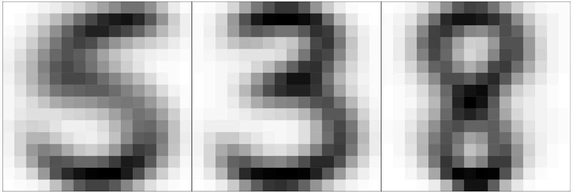
**Estimated centroid**



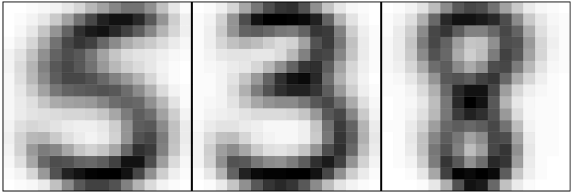
# Results (Manhattan distance)

- Focus on the posterior estimate of cluster centroids (arithmetic mean).

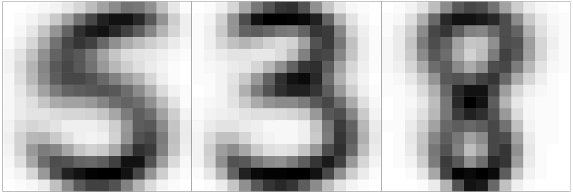
**Posterior mean**



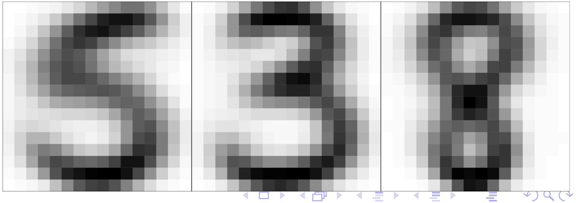
**posterior median**



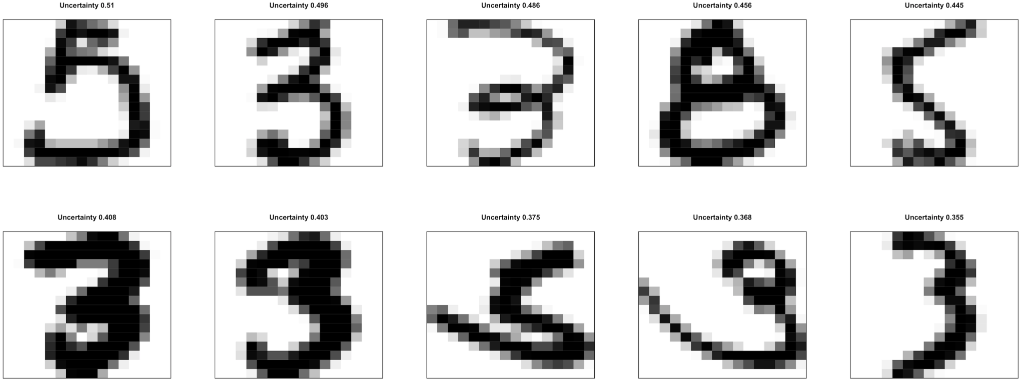
**Posterior** 5% **percentile**



**Posterior** 95% **percentile**



# Hard cases (top 10 largest uncertainty)



Uncertainty 0.51

Uncertainty 0.496

Uncertainty 0.486

Uncertainty 0.456

Uncertainty 0.445

Uncertainty 0.408

Uncertainty 0.403

Uncertainty 0.375

Uncertainty 0.368

Uncertainty 0.355

# Thanks!