Introduction to Gaussian Mixture Models

March 3, 2021

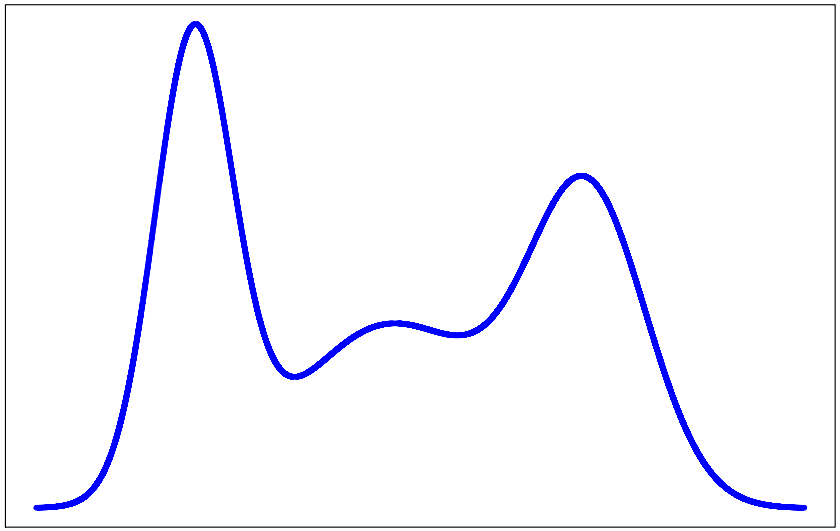
Joshua Tobin

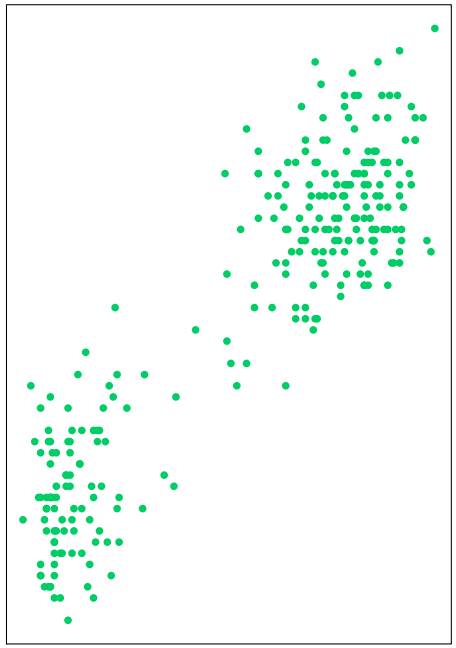
Outline

1. What is a Gaussian Mixture Model (GMM)?
2. How can we use GMMs to cluster data?
3. What are the prominent methods for clustering data with GMMs?
4. Do these methods have drawbacks?
5. Can we improve them?

1 What is a Gaussian Mixture Model?

Linear superposition of Gaussian components, aimed to provide richer class of density models.

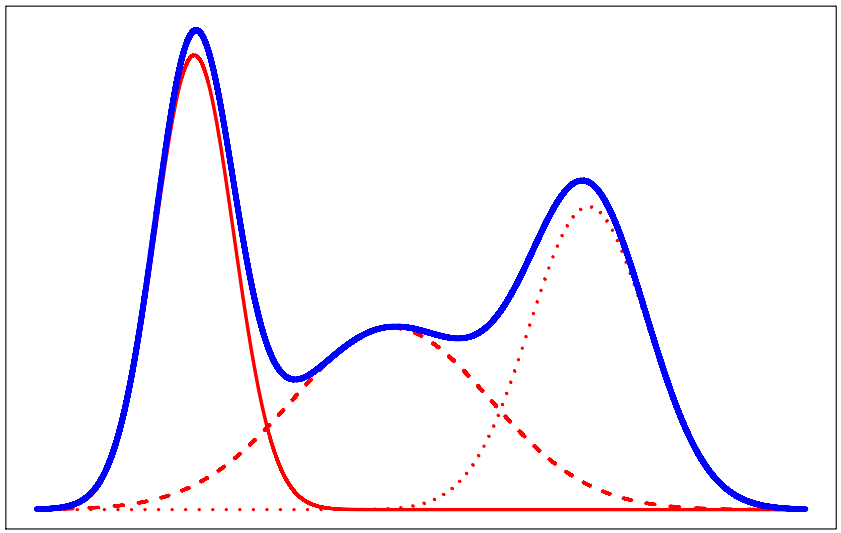


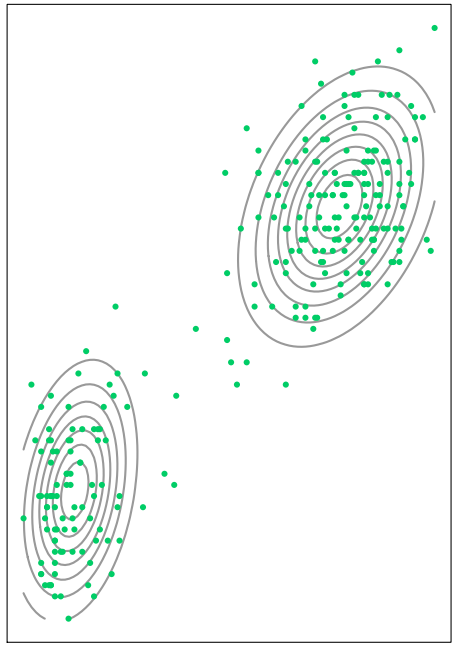


Aim is approximation of complex densities by adjusting means $\boldsymbol{\mu}_k$ and covariances $\boldsymbol{\Sigma}_k$ of K *component* Gaussians.

1 What is a Gaussian Mixture Model?

Linear superposition of Gaussian components, aimed to provide richer class of density models.

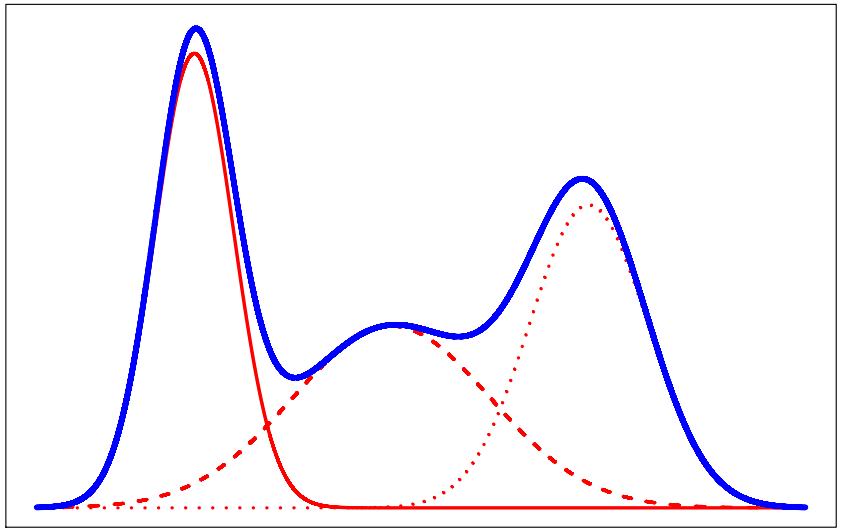


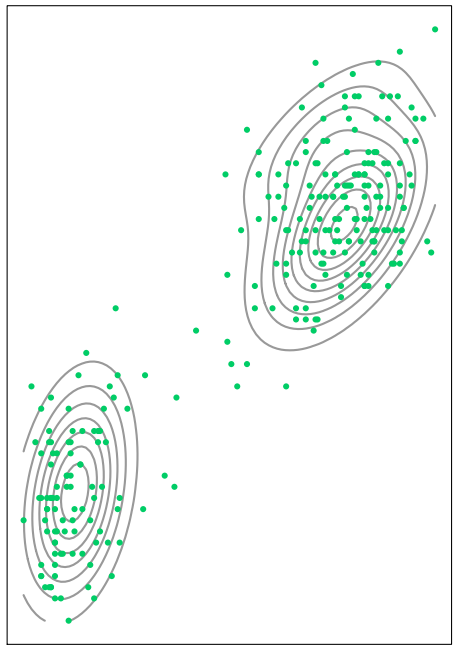


Aim is approximation of complex densities by adjusting means $\boldsymbol{\mu}_k$ and covariances $\boldsymbol{\Sigma}_k$ of K *component* Gaussians.

1 What is a Gaussian Mixture Model?

Linear superposition of Gaussian components, aimed to provide richer class of density models.





Aim is approximation of complex densities by adjusting means $\boldsymbol{\mu}_k$ and covariances $\boldsymbol{\Sigma}_k$ of K *component* Gaussians.

1 What is a Gaussian Mixture Model?

So we consider a superposition of K Gaussian densities of the form:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

which is equivalent to:

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(k)p(\mathbf{x}|k)$$

- $\pi_k = p(k)$ - prior probability of picking the kth component.
- $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = p(\mathbf{x}|k)$ - probability of $\mathbf{x}$ conditioned on k.

We seek $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ and $p(k|\mathbf{x})$

2 How can we use GMMs to cluster data?

If we have data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ which we wish to model using a mixture of Gaussians for group assignment:

- Introduce K-dimensional 1-of-K variable $\mathbf{z}$ with marginal $p(z_k = 1) = \pi_k$.
- Now the conditional distribution of $\mathbf{x}$ given $\mathbf{z}$ is

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}.$$

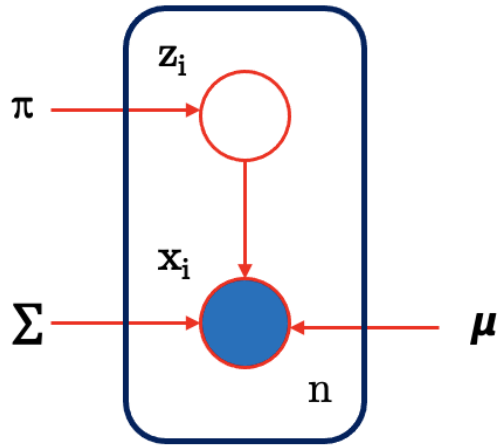
- So taking $p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$ and summing over $\mathbf{z}$ yields

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

2 How can we use GMMs to cluster data?

We have

- $\mathbf{X}$ - $n \times d$ data matrix
- $\mathbf{Z}$ - $n \times K$ latent variable matrix
- π - prior component probabilities
- $\boldsymbol{\mu}$ - d-dimensional component mean vectors
- $\boldsymbol{\Sigma}$ - $d \times d$ component covariance matrices



Popular approach formulates the log-likelihood function:

$$\ln p(\mathbf{X}|\pi, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{n} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}.$$

Problem: This likelihood is very difficult to maximise.

3. Some prominent methods for clustering data with GMMs?

Expectation Maximisation (EM) - a powerful and popular approach.

1. Initialise $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ and π_k and evaluate the log likelihood.
2. **E step** - Compute the responsibilities:

$$p(z_{ik} = 1|\boldsymbol{x}_i) = \gamma(z_{ik}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

3. Some prominent methods for clustering data with GMMs?

Expectation Maximisation (EM) - a powerful and popular approach.

3. **M step** - Update the parameters:

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{n_k} \sum_{i=1}^{n} \gamma(z_{ik}) \mathbf{x}_i$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{n_k} \sum_{i=1}^{n} \gamma(z_{ik})(\mathbf{x}_i - \boldsymbol{\mu}_k^{\text{new}})(\mathbf{x}_i - \boldsymbol{\mu}_k^{\text{new}})^T$$

$$\pi_k^{\text{new}} = \frac{n_k}{n}$$

where $n_k = \sum_{i=1}^{n} \gamma(z_{ik})$.

4. Evaluate the log likelihood and check for convergence of either the parameters or the log likelihood.

3. Some prominent methods for clustering data with GMMs?

A vast literature exists describing ways to find GMMs including

- **Variational Inference**
  - Deterministic approximation scheme which assumes latent variable and parameter distributions can be factorised.
  - Similar computational burden to EM, but various improvements in approach.

- Agglomerative approaches, based on HAC and HDBSCAN.
- Spectral methods involving decomposition.
- Methods which aim to maximise log-likelihood numerically.

3. Some prominent methods for clustering data with GMMs?

A vast literature exists describing ways to find GMMs including

- **Variational Inference**
  - Deterministic approximation scheme which assumes latent variable and parameter distributions can be factorised.
  - Similar computational burden to EM, but various improvements in approach.

- Agglomerative approaches, based on HAC and HDBSCAN.
- Spectral methods involving decomposition.
- Methods which aim to maximise log-likelihood numerically.

4. Do these methods have drawbacks?

EM struggles if initialisation is poor.

- Non-convexity of likelihood comes from the parametrisation of the model components.
- Set of all mixture models is not convex when the distribution has free parameters.

Escape Route:

Assuming the data is dense enough that there is always a data point close to the real centre ...

... we can restrict possible centres to the set of data points ...

... leading to a convex cost function which unconditionally converges to global minimum.

4. Do these methods have drawbacks?

Lashkari & Golland (2007) formulate a mixture model

$$Q(\mathbf{x}) = \sum_{j=1}^{n} q_j \mathcal{N}(\mathbf{x}|\mathbf{x}_j)$$

where

- q_j - prior probability of the jth component.
- $\mathcal{N}(\mathbf{x}|\mathbf{x}_j)$ - Normal distribution with expectation parameter equal to the jth data point.

which yields the normalised log likelihood over q_j:

$$\begin{aligned} L(\{q_j\}; \mathbf{X}) &= \frac{1}{n} \sum_{i=1}^{n} \ln \left\{ \sum_{j=1}^{n} q_j \mathcal{N}(\mathbf{x}|\mathbf{x}_j) \right\} \\ &= \frac{1}{n} \sum_{i=1}^{n} \ln \left\{ \sum_{j=1}^{n} q_j e^{-\beta \|\mathbf{x}_i - \mathbf{x}_j\|_2^2} \right\} \end{aligned}$$

4. Do these methods have drawbacks?

We can represent this likelihood in terms of KL-Divergence:

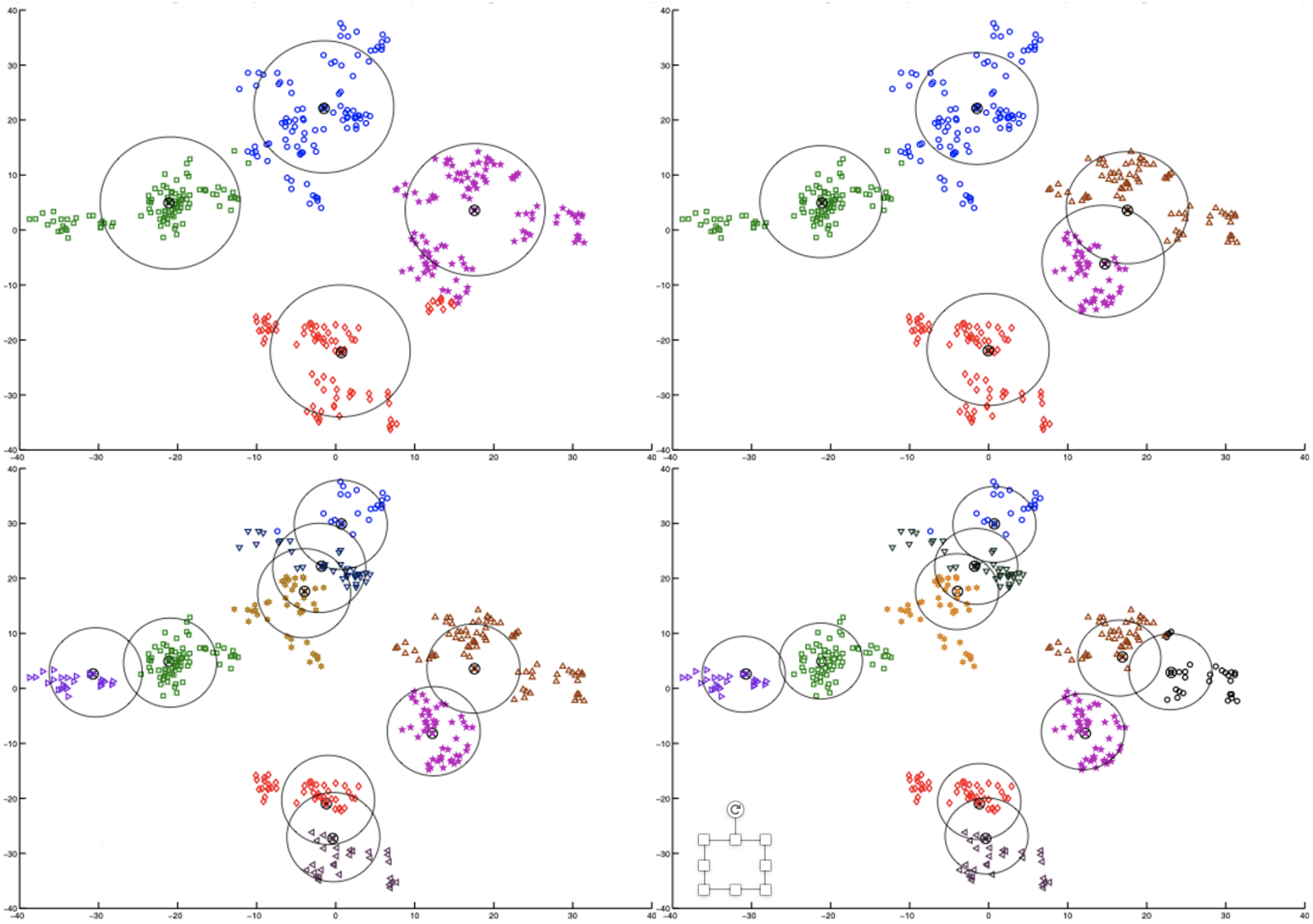
$$D(\hat{P}\|Q) = -\sum_{x\in\mathbf{X}} \hat{P}(\mathbf{x}) \ln Q(\mathbf{x}) - \mathbb{H}(\hat{P}) = -L(\{q_j\}; \mathbf{X}) + c,$$

where $\hat{P}(\mathbf{x}) = 1/n$, the empirical distribution.

We update the component prior probabilities with

$$q_j^{(t+1)} = q_j^{(t)} \sum_{x\in\mathbf{X}} \frac{\hat{P}(\mathbf{x})\mathcal{N}(\mathbf{x}|\mathbf{x}_j)}{\sum_{j'=1}^{n} q_{j'}^{(t)} \mathcal{N}(\mathbf{x}|\mathbf{x}_{j'})}.$$

This is guaranteed to converge to a global optimum.



4. Do these methods have drawbacks?

Pilanci et. al. improve on this formulation with a cardinality penalty on $\{q_j\}$:

$$\max_{1^T\mathbf{q}=1,\mathbf{q}\geq 0} \sum_{i=1}^{n} \ln \left\{ \sum_{j=1}^{n} q_j e^{-\beta\|\mathbf{x}_i-\mathbf{x}_j\|_2^2} \right\} - \lambda \text{card}(\mathbf{q}),$$

where the parameter λ can easily return a specified number of components.

So by using convex mixture models:

- We have gained the ability to locate global optimum.
- We have surrendered varying Σ.
- Still have a problem with slow convergence (Takahashi, 2011).

4. Do these methods have drawbacks?

Pilanci et. al. improve on this formulation with a cardinality penalty on $\{q_j\}$:

$$\max_{1^T\mathbf{q}=1,\mathbf{q}\geq 0} \sum_{i=1}^{n} \ln\left\{\sum_{j=1}^{n} q_j e^{-\beta\|\mathbf{x}_i-\mathbf{x}_j\|_2^2}\right\} - \frac{\lambda}{\max_i q_i},$$

where the parameter λ can easily return a specified number of components.

So by using convex mixture models:

- We have gained the ability to locate global optimum.
- We have surrendered varying Σ.
- Still have a problem with slow convergence (Takahashi, 2011).

5. Can we improve them?

Aim is to build a fast exemplar-based GMM in which the covariance matrices are free from constraints

Propose selecting initial K exemplars using peak-finding

- Set local density $\rho_i = \sum_{l=1}^{K} \exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_i^{(l)}\|_2)$
- Compute minimum distances to points that have higher local density values

$$\delta_i = \begin{cases} \max\{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 : 1 \leq j \leq n\}, & \text{if } \rho_i \text{ is the largest;} \\ \min\{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 : 1 \leq j \leq n,\ \rho_j > \rho_i\}, & \text{otherwise.} \end{cases}$$

For each exemplar, we calculate a rough estimate of the covariance matrix, Σ_k using a set of nearest neighbours.

5. Can we improve them?

Given the exemplar set and covariance estimates:

- **E** - $K \times p$ exemplar matrix
- **D** - distance matrix, $d_{ij} = \sqrt{(\mathbf{x}_i - \mathbf{e}_j)^T \Sigma_j^{-1} (\mathbf{x}_i - \mathbf{e}_j)}$
- **Q** - responsibility matrix.

Specifying the following optimisation problem:

$$\min_{\{\boldsymbol{Q}_{i\cdot} \in \Delta\}_{i=1}^n} \sum_{j=1}^K \boldsymbol{D}_{\cdot j}^T \boldsymbol{Q}_{\cdot j} + \frac{\rho}{2} \|\boldsymbol{E}^T \boldsymbol{Q}^T \mathbf{1}_n - \boldsymbol{X}^T \mathbf{1}_n\|_2^2 + \lambda \text{card}(\boldsymbol{Q}^T \mathbf{1}_n),$$

This objective is three-fold.

5. Can we improve them?

Given the exemplar set and covariance estimates:

- **E** - $K \times p$ exemplar matrix
- **D** - distance matrix, $d_{ij} = \sqrt{(\mathbf{x}_i - \mathbf{e}_j)^T \Sigma_j^{-1} (\mathbf{x}_i - \mathbf{e}_j)}$
- **Q** - responsibility matrix.

Specifying the following optimisation problem:

$$\min_{\{\boldsymbol{Q}_{i\cdot} \in \Delta\}_{i=1}^n} \sum_{j=1}^K \boldsymbol{D}_{\cdot j}^T \boldsymbol{Q}_{\cdot j} + \frac{\rho}{2} \|\boldsymbol{E}^T \boldsymbol{Q}^T \mathbf{1}_n - \boldsymbol{X}^T \mathbf{1}_n\|_2^2 + \lambda \text{card}(\boldsymbol{Q}^T \mathbf{1}_n),$$

1. Minimising total within-cluster variance.

5. Can we improve them?

Given the exemplar set and covariance estimates:

- **E** - $K \times p$ exemplar matrix
- **D** - distance matrix, $d_{ij} = \sqrt{(\mathbf{x}_i - \mathbf{e}_j)^T \Sigma_j^{-1} (\mathbf{x}_i - \mathbf{e}_j)}$
- **Q** - responsibility matrix.

Specifying the following optimisation problem:

$$\min_{\{\boldsymbol{Q}_{i\cdot} \in \Delta\}_{i=1}^n} \sum_{j=1}^K \boldsymbol{D}_{\cdot j}^T \boldsymbol{Q}_{\cdot j} + \frac{\rho}{2} \|\boldsymbol{E}^T \boldsymbol{Q}^T \mathbf{1}_n - \boldsymbol{X}^T \mathbf{1}_n\|_2^2 + \lambda \text{card}(\boldsymbol{Q}^T \mathbf{1}_n),$$

3. Matching empirical moments to population moments.

5. Can we improve them?

Given the exemplar set and covariance estimates:

- **E** - $K \times p$ exemplar matrix
- **D** - distance matrix, $d_{ij} = \sqrt{(\mathbf{x}_i - \mathbf{e}_j)^T \Sigma_j^{-1} (\mathbf{x}_i - \mathbf{e}_j)}$
- **Q** - responsibility matrix.

Specifying the following optimisation problem:

$$\min_{\{\boldsymbol{Q}_{i\cdot} \in \Delta\}_{i=1}^n} \sum_{j=1}^K \boldsymbol{D}_{\cdot j}^T \boldsymbol{Q}_{\cdot j} + \frac{\rho}{2} \|\boldsymbol{E}^T \boldsymbol{Q}^T \mathbf{1}_n - \boldsymbol{X}^T \mathbf{1}_n\|_2^2 + \lambda \text{card}(\boldsymbol{Q}^T \mathbf{1}_n),$$

3. Purifying the exemplar set.

5. Can we improve them?

The objective can be split into K convex programs, each solved in parallel.

$$\min_{\nu=1,\dots,K} \left\{ \min_{\{\boldsymbol{Q}_{i\cdot}\in\Delta\}_{i=1}^{n}} \sum_{j=1}^{K} \boldsymbol{D}_{\cdot j}^{T}\boldsymbol{Q}_{\cdot j} + \frac{\rho}{2}\|\boldsymbol{E}^{T}\boldsymbol{Q}^{T}\mathbf{1}_n - \boldsymbol{X}^{T}\mathbf{1}_n\|_2^2 + \frac{\lambda}{\mathbf{1}_n^{T}\boldsymbol{Q}_{\cdot\nu}} \right\},$$

Once the responsibility matrix $\mathbf{Q}$ is returned:

- Still need to obtain clustering results
- Propose using DA-EM to compute updated component priors and update covariance matrices
- Re-running for different values of λ and use criteria to select best model.

5. Can we improve them?

Questions still to be answered:

1. Trade-off between limiting number of centres with freer covariance matrices?
2. Better approach for updating component priors and covariance matrices?
3. Can we incorporate the different covariance structures of Celeux & Govaert?
4. What is overall complexity?

Thanks for listening, any advice or recommended reading would be greatly appreciated!