Clustering

Anita Faul

Laboratory for Scientific Computing, University of Cambridge

How can we sort objects?

- size,
- shape,
- colour,
- texture,
- ingredients.

Examples of Clustering

Examples of clustering problems:

- articles with similar content,
- search engines,
- suggestions from streaming sites,
- image segmentation,
- lossy data compression,
- bio informatics.
- Any suggestion?

K Means Clustering

K Means Clustering:

- Let $\mathbf{v}_1, \ldots, \mathbf{v}_N$ be the feature vectors of N data samples.
- Number of clusters fixed, K.
- *Hard clustering* assigns sample to the cluster with the nearest centre.
- Find cluster centres $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$ such that the sum of the squared distances of each data sample to its assigned cluster centre is minimal.
- Minimize

$$J = \sum_{n=1}^{N} \min_k \|\mathbf{v}_n - \boldsymbol{\mu}_k\|^2.$$

- *NP hard*: no known algorithm to solve this in polynomial time, since as cluster centres move around, for each sample its nearest cluster centre can change.

K Means Clustering

- Separate interdependency.
- *Hidden (latent)* variables $\mathbf{z}_n$, one for each data sample $\mathbf{v}_n$.
- *1-of-K representation*: $\mathbf{z}_n \in \{0,1\}^K$, one entry 1 and the others have to be 0.
- 1 in the k^{th} entry indicates that $\boldsymbol{\mu}_k$ is the nearest cluster centre to $\mathbf{v}_n$.
- Let z_{nk} be the k^{th} component of $\mathbf{z}_n$,

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \|\mathbf{v}_n - \boldsymbol{\mu}_k\|^2.$$

K Means Clustering

- Quadratic in $\boldsymbol{\mu}_k$.
- Find minimum by differentiating with respect to $\boldsymbol{\mu}_k$ and setting this to zero,

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} z_{nk}\mathbf{v}_n}{\sum_{n=1}^{N} z_{nk}}.$$

- $\sum_{n=1}^{N} z_{nk}$ number of data samples for which $\boldsymbol{\mu}_k$ is the closest centre.
- $\sum_{n=1}^{N} z_{nk}\mathbf{v}_n$ is the sum of those samples.
- $\boldsymbol{\mu}_k$ is the mean of the samples assigned to this particular cluster.
- Adjust indicator vectors $\mathbf{z_n}$.
- Alternate.
- Terminate, when after moving the centres, none of the indicator vectors changes.

K Means Clustering

- Local minimum.
- Highly dependent on the initialization of the cluster centres.
- At the start centres chosen randomly, or samples randomly assigned to clusters.
- Algorithm is run (possibly in parallel) with many different initializations. After convergence, the result with the lowest value of J is chosen.

K Medoids Algorithm

K Medoids Algorithm:

- Use any dissimilarity measure.
- Minimization depends on the differentiability of the dissimilarity measure, and whether it is possible to find where the derivative vanishes.
- If this is not possible, we require each cluster centre to be one of the data samples.
- The minimization with respect to $\boldsymbol{\mu}_k$ is then a search among the data samples assigned to the k^{th} cluster.

K Means Clustering

4 clusters

16 clusters

Mixture Models

Mixture Models:

- $\mathbf{z}_n$ indicates which process generated $\mathbf{v}_n$.
- Let $p_k(\mathbf{v})$ be the probability distribution of process k.
- Let π_k be the probability that process k generates a sample,

$$0 \leq \pi_k \leq 1 \text{ and } \sum_{k=1}^{K} \pi_k = 1.$$

- Latent variables $\mathbf{z}_1, \ldots, \mathbf{z}_N$ are drawn from from a probability distribution $p(\mathbf{z})$,

$$p(z_k = 1) = \pi_k.$$

Mixture Models

- Probability of $\mathbf{v}$ given $\mathbf{z}$ is the conditional probability distribution

$$p(\mathbf{v}|\mathbf{z}) = p(\mathbf{v}|z_k = 1) = p_k(\mathbf{v}).$$

- *Mixture distribution*:

$$\begin{aligned} p(\mathbf{v}) &= \sum_{\mathbf{z}} p(\mathbf{v}, \mathbf{z}) = \sum_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{v}|\mathbf{z}) \\ &= \sum_{k=1}^{K} p(z_k = 1) p(\mathbf{v}|z_k = 1) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{v}). \end{aligned}$$

- π_k are known as *mixing coefficients*.

Bayes Rule

Bayes Rule: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

- A and B are events.
- $P(A)$ and $P(B)$ are the probabilities of A and B without regard to each other.
- $P(A|B)$ is the *conditional probability* of A given that B is true. $P(B|A)$ is the conditional probability of B given that A is true.
- Often expressed as $P(A|B) \propto P(B|A)P(A)$ where $\propto$ means that the two sides are proportional to each other.

Mixture Models

- *Responsibility* process k takes for explaining the sample $\mathbf{v}$ is the probability that it was generated by process k:

$$p(z_k = 1|\mathbf{v}) = \frac{p(z_k = 1)p(\mathbf{v}|z_k = 1)}{p(\mathbf{v})} = \pi_k \frac{p_k(\mathbf{v})}{p(\mathbf{v})}$$

  by Bayes rule.
- Responsibilities sum to 1.
- *Soft clustering* makes cluster assignments according to the values $p(z_{nk} = 1|\mathbf{v}_n)$ for $k = 1, \ldots, K$.
- Possible that for a particular sample the probabilities are the same for two (or even more) values of k. These are samples which lie between clusters.

Mixture Models

- Maximize joint likelihood of the data samples $\mathbf{v}_1, \ldots, \mathbf{v}_n$:

$$\prod_{n=1}^{N} p(\mathbf{v}_n) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k p_k(\mathbf{v}_n).$$

- Or alternatively its logarithm

$$\mathcal{L} = \sum_{n=1}^{N} \log p(\mathbf{v}_n) = \sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k p_k(\mathbf{v}_n) \right).$$

- Subject to $\sum_{k=1}^{K} \pi_k = 1$.
- Using a Lagrange multiplier λ, we maximize

$$\sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k p_k(\mathbf{v}_n) \right) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right).$$

Mixture Models

- Differentiating with respect to π_k and setting to zero gives

$$\sum_{n=1}^{N} \frac{1}{\sum_{k=1}^{K} \pi_k p_k(\mathbf{v}_n)} p_k(\mathbf{v}_n) + \lambda = 0.$$

- Multiplying through by π_k and summing over all k, gives $\lambda = -N$.
- Inserting this and again multiplying by π_k, results in

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} p(z_{nk} = 1 | \mathbf{v}_n).$$

- The mixing coefficient π_k is the average responsibility that all data samples are generated by process k.
- Note: $p(z_{nk} = 1 | \mathbf{v}_n)$ depend on π_k itself $\Rightarrow$ iterative procedure.

Mixture Models

Caution:

- Assume that $K = 2$ and that all data samples are roughly grouped together apart from one outlier.
- π_1 tends to $(N-1)/N$ while π_2 tends to $1/N$.
- $p_1(\mathbf{v})$ roughly describes the distribution of $N-1$ samples.
- Likelihood can be increased again and again by concentrating the probability mass of $p_2(\mathbf{v})$ more and more tightly around the outlier.
- Evaluation of $p_2(\mathbf{v})$ at the outlier tends to infinity.
- Area where $p_2(\mathbf{v})$ is zero or close to zero tends to zero.
- $K!$ equivalent solutions.

Gaussian Mixture Models

- $p_k(\mathbf{v})$, $k = 1, \ldots, K$, are normal distributions, $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$,

$$p_k(\mathbf{v}) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_k|}} \exp\left(-\frac{1}{2}(\mathbf{v}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{v}-\boldsymbol{\mu}_k)\right).$$

- Derivative of $p_k(\mathbf{v})$ with respect to $\boldsymbol{\mu}_k$:

$$\frac{\partial}{\partial\boldsymbol{\mu}_k}p_k(\mathbf{v}) = p_k(\mathbf{v})\boldsymbol{\Sigma}_k^{-1}(\mathbf{v}-\boldsymbol{\mu}_k).$$

- Derivative of $p_k(\mathbf{v})$ with respect to $\boldsymbol{\Sigma}_k$:

$$\frac{\partial}{\partial\boldsymbol{\Sigma}_k}p_k(\mathbf{v}) = -\frac{1}{2}p_k(\mathbf{v})\left[\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{-1}(\mathbf{v}-\boldsymbol{\mu}_k)(\mathbf{v}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}\right].$$

$$\frac{\partial}{\partial \boldsymbol{\mu}_k}\mathcal{L} = \sum_{n=1}^{N} \frac{1}{p(\mathbf{v}_n)} \pi_k p_k(\mathbf{v}_n) \boldsymbol{\Sigma}_k^{-1}(\mathbf{v}_n - \boldsymbol{\mu}_k)$$
$$= \sum_{n=1}^{N} p(z_{nk} = 1|\mathbf{v}_n) \boldsymbol{\Sigma}_k^{-1}(\mathbf{v}_n - \boldsymbol{\mu}_k).$$

- Expected number of samples in cluster k: $N_k = \sum_{n=1}^{N} p(z_{nk} = 1|\mathbf{v}_n)$.
- $\boldsymbol{\mu}_k$ is a weighted average of all samples in the data set where the weights are the responsibilities that the sample was generated by process k,

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} p(z_{nk} = 1|\mathbf{v}_n)\mathbf{v}_n.$$

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_k}\mathcal{L} = \frac{1}{2}\sum_{n=1}^{N} p(z_{nk}=1|\mathbf{v}_n)\left[\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{-1}(\mathbf{v}_n - \boldsymbol{\mu}_k)(\mathbf{v}_n - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}\right].$$

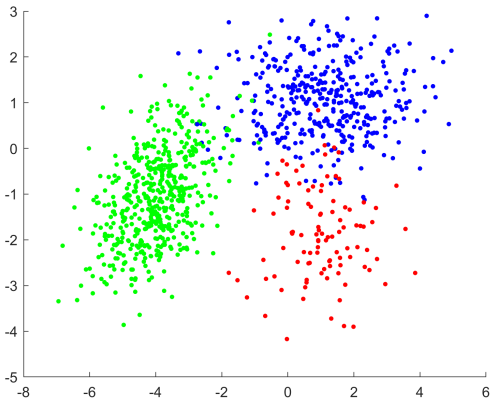
- Set to zero and multiply through with 2 and $\boldsymbol{\Sigma}_k$ from both sides.
- Similar to sample covariance,

$$\boldsymbol{\Sigma_k} = \frac{1}{N_k}\sum_{n=1}^{N} p(z_{nk}=1|\mathbf{v}_n)(\mathbf{v}_n - \boldsymbol{\mu}_k)(\mathbf{v}_n - \boldsymbol{\mu}_k)^T.$$

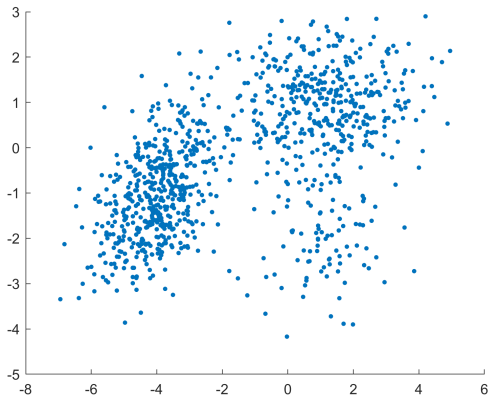
Gaussian Mixture Models

1. Choose K and convergence threshold.
2. Initialize means $\boldsymbol{\mu}_k$, covariances $\boldsymbol{\Sigma}_k$, and mixing coefficients π_k for $k = 1, \ldots, K$, (e.g from K means) and calculate the initial value of the logarithm of the likelihood.
3. For $n = 1, \ldots, N$ and $k = 1, \ldots, K$, calculate all the responsibilities $p(z_{nk} = 1|\mathbf{v}_n)$.
4. Use these responsibilities to update means $\boldsymbol{\mu}_k$, covariances $\boldsymbol{\Sigma}_k$, and mixing coefficients π_k for $k = 1, \ldots, K$.
5. Evaluate the change in the logarithm of the likelihood and terminate if this is below the convergence threshold (or if the change in parameters is below the convergence threshold). Otherwise return to step 3.

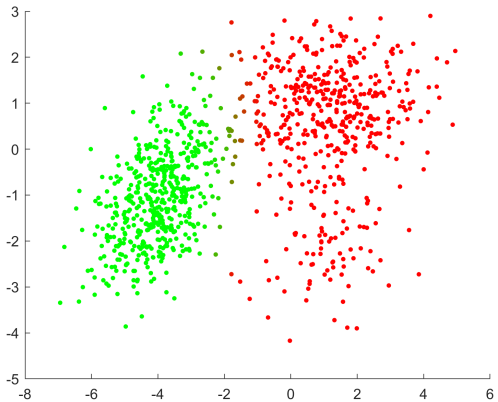
Gaussian Mixture Models



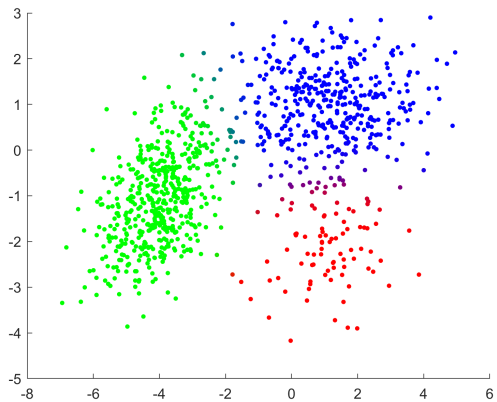
Data samples



from three processes.



Two clusters fitted.



Three clusters fitted.

Expectation-Maximization

- How to generally maximize

$$\mathcal{L} = \sum_{n=1}^{N} \log p(\mathbf{v}_n) = \sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k p_k(\mathbf{v}_n) \right) ?$$

- Note: When normal distributions are used, maximizing

$$\begin{aligned} \widehat{\mathcal{L}} &= \sum_{n=1}^{N} \sum_{k=1}^{K} p(z_{nk} = 1 | \mathbf{v}_n) \left[\log \pi_k + \log p_k(\mathbf{v}_n) \right] \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} p(z_{nk} = 1 | \mathbf{v}_n) \log \Big(p(z_{nk} = 1) p(\mathbf{v}_n | z_{nk} = 1) \Big) \end{aligned}$$

leads to the same update formulae.

Expectation-Maximization

- $\widehat{\mathcal{L}}$ is the expectation of the logarithm of the complete data likelihood

$$\sum_{n=1}^{N} \log p(\mathbf{v}_n, \mathbf{z}_n),$$

  where the expectation is taken with respect to the responsibilites, that is the posterior probabilities of the latent variables.
- How are $\mathcal{L}$ and $\widehat{\mathcal{L}}$ related and why do the parameters where their derivatives vanish coincide?

Expectation-Maximization

- Using the product rule for probabilities,

$$\sum_{n=1}^{N} \log p(\mathbf{v}_n, \mathbf{z}_n) = \sum_{n=1}^{N} \log p(\mathbf{v}_n) + \log p(\mathbf{z}_n|\mathbf{v}_n).$$

- Both sides are as functions of the random variables $\mathbf{z}_n$, and the expectation with respect to any distribution $q(\mathbf{z}_n)$ can be taken.
- Since $\mathbf{z}_n$ is a 1-of-K representation, the expectation is calculated by summing over all possible values for $\mathbf{z}_n$.

Expectation-Maximization

$$\sum_{k=1}^{K}\sum_{n=1}^{N} q(z_{nk}=1)\log p(\mathbf{v}_n, z_{nk}=1)$$

$$= \sum_{k=1}^{K}\sum_{n=1}^{N} q(z_{nk}=1)\log p(\mathbf{v}_n) + \sum_{k=1}^{K}\sum_{n=1}^{N} q(z_{nk}=1)\log p(z_{nk}=1|\mathbf{v}_n)$$

$$= \underbrace{\sum_{n=1}^{N} \log p(\mathbf{v}_n)}_{\mathcal{L}} + \sum_{k=1}^{K}\sum_{n=1}^{N} q(z_{nk}=1)\log p(z_{nk}=1|\mathbf{v}_n),$$

because of $\sum_{k=1}^{K} q(z_{nk}=1) = 1.$

Expectation-Maximization

- Subtracting and adding the term $\sum_{n=1}^{N}\sum_{k=1}^{K} q(z_{nk}=1)\log q(z_{nk}=1)$ gives

$$\begin{aligned}\mathcal{L} &= \sum_{k=1}^{K}\sum_{n=1}^{N} q(z_{nk}=1)\log\frac{p(\mathbf{v}_n, z_{nk}=1)}{q(z_{nk}=1)} \\ &\quad + \sum_{n=1}^{N}\sum_{k=1}^{K} q(z_{nk}=1)\log\frac{q(z_{nk}=1)}{p(z_{nk}=1|\mathbf{v}_n)},\end{aligned}$$

- In the last line, each sum over $k=1,\ldots,K$ is the *Kullback–Leibler divergence (KL divergence)* from the discrete distribution $p(\mathbf{z}_n|\mathbf{v}_n)$ to the discrete distribution $q(\mathbf{z}_n)$.

$$\mathcal{L} = \widetilde{\mathcal{L}} + D_{KL}(q(\mathbf{z}_n)\|p(\mathbf{z}_n|\mathbf{v}_n)).$$

Kullback–Leibler divergence

- Given two discrete probability distributions P and Q, the Kullback–Leibler divergence from Q to P is defined as

$$D_{KL}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}.$$

- Only defined if $Q(i) = 0$ implies $P(i) = 0$ to avoid a division by zero.
- If both distributions are the same, then the Kullback–Leibler divergence is zero.
- It is also non-negative, since $\log x \leq x - 1$ and therefore

$$\begin{aligned} D_{KL}(P\|Q) &= -\sum_i P(i) \log \frac{Q(i)}{P(i)} \geq -\sum_i P(i) \left(\frac{Q(i)}{P(i)} - 1 \right) \\ &= -\sum_i Q(i) + \sum_i P(i) = 0. \end{aligned}$$

Expectation-Maximization

- $\widetilde{\mathcal{L}}$ is a lower bound for $\mathcal{L}$, since $D_{KL}(q(\mathbf{z}_n)\|p(\mathbf{z}_n|\mathbf{v}_n))$ is non-negative.

$$\begin{aligned}\widetilde{\mathcal{L}} &= \sum_{k=1}^{K}\sum_{n=1}^{N} q(z_{nk}=1)\log\frac{p(\mathbf{v}_n, z_{nk}=1)}{q(z_{nk}=1)} \\ &= \overbrace{\sum_{k=1}^{K}\sum_{n=1}^{N} q(z_{nk}=1)\log\Big(p(z_{nk}=1)p(\mathbf{v}_n|z_{nk}=1)\Big)}^{\widehat{\mathcal{L}}} \\ &\quad -\sum_{k=1}^{K}\sum_{n=1}^{N} q(z_{nk}=1)\log q(z_{nk}=1).\end{aligned}$$

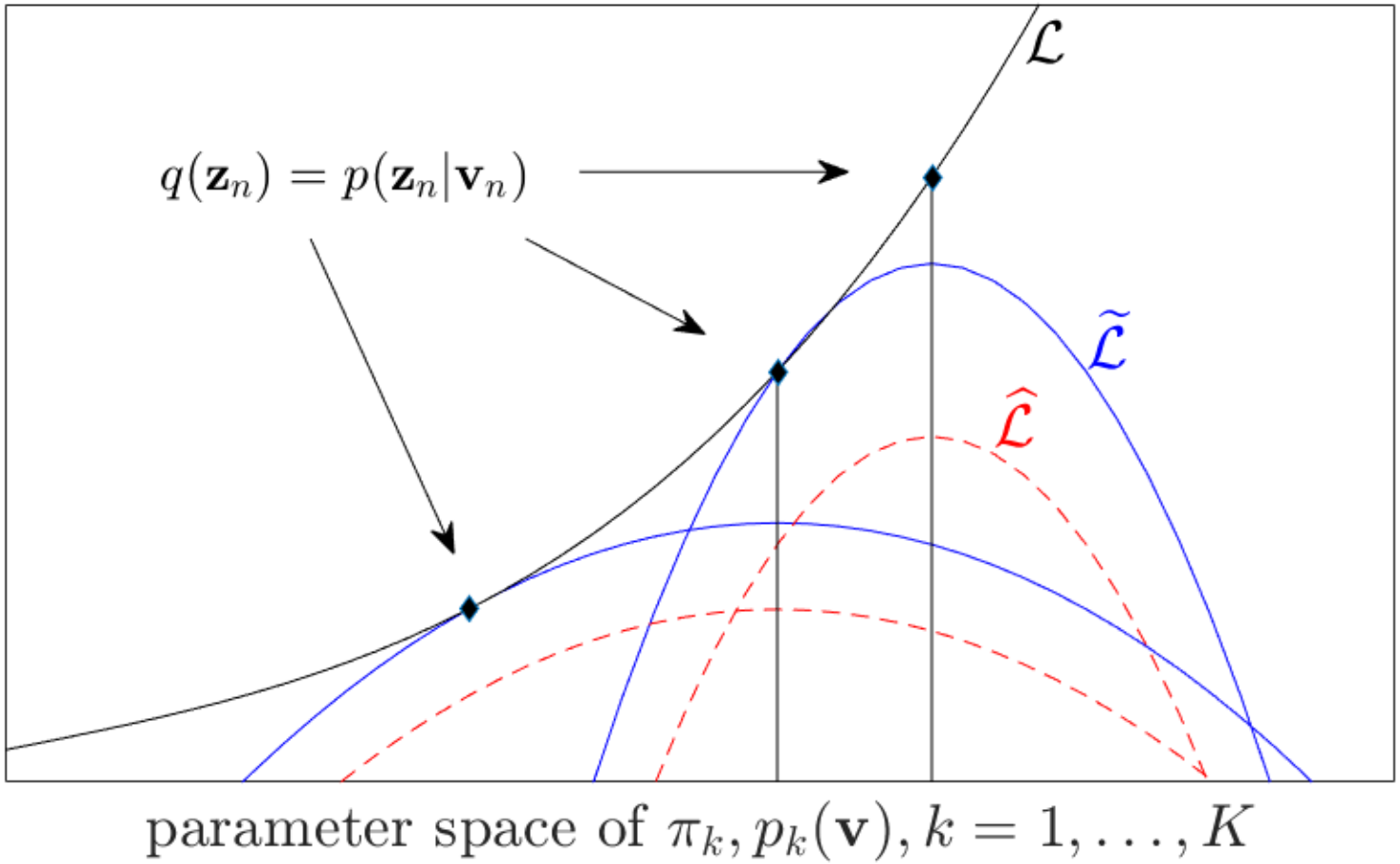
- It has the same value as $\mathcal{L}$, if $q(\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{v}_n)$, because then $D_{KL}(q(\mathbf{z}_n)\|p(\mathbf{z}_n|\mathbf{v}_n)) = 0$.

Expectation-Maximization

Maximize by alternating between

- maximizing with respect to $q(\mathbf{z}_n)$ which means setting $q(\mathbf{z}_n)$ to $p(\mathbf{z}_n|\mathbf{v}_n)$, where the responsibilities are evaluated using the current parameters of $p_k(\mathbf{v}) = p(\mathbf{v}|z_k = 1)$ and and mixing coefficients $\pi_k = p(z_{nk} = 1)$,
- maximizing with respect to the parameters of $p_k(\mathbf{v})$, $k = 1, \ldots, K$ and mixing coefficients π_k.

Expectation-Maximization



Expectation-Maximization

1. Choose K and convergence threshold.
2. Initialize all parameters of $p_k(\mathbf{v}) = p(\mathbf{v}|z_k = 1)$ and mixing coefficients $\pi_k = p(z_k = 1)$.
3. *E-step*: Evaluate the responsibilities $p(z_{nk} = 1|\mathbf{v}_n)$ for $n = 1, \ldots, N$ and $k = 1, \ldots, K$.
4. *M-step*: Maximize

$$\sum_{k=1}^{K}\sum_{n=1}^{N} p(z_{nk} = 1|\mathbf{v}_n) \log \Big(p(z_{nk} = 1) p(\mathbf{v}_n|z_{nk} = 1) \Big)$$

   with respect to the parameters of $p_k(\mathbf{v}) = p(\mathbf{v}|z_k = 1)$ and mixing coefficients $\pi_k = p(z_k = 1)$.
5. Terminate, if all changes are below the convergence threshold. Otherwise return to step 3

Distributions of parameters

Prior assumptions:

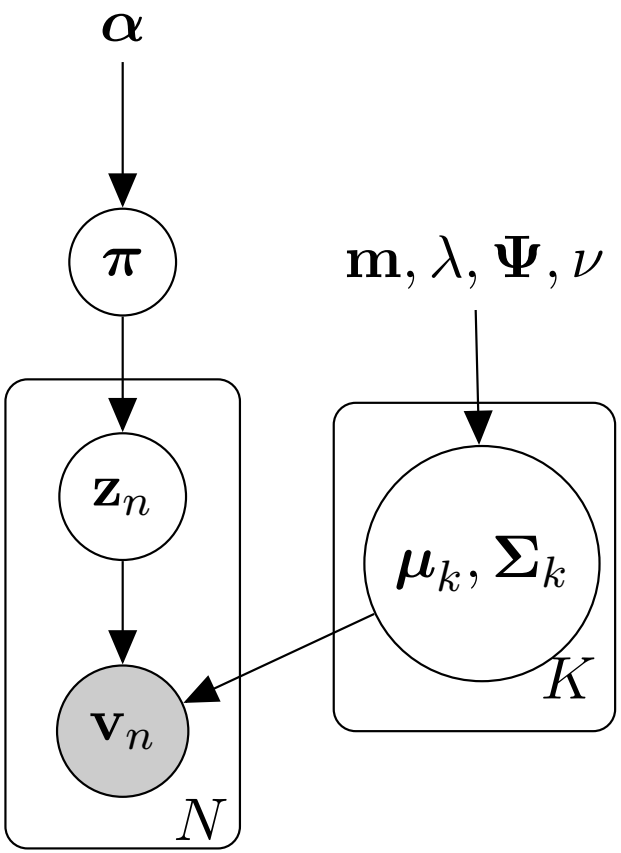
- K remains fixed.
- $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ follows a Dirichlet distribution with parameter $\boldsymbol{\alpha} = (\alpha/K, \ldots, \alpha/K)^T$.
- α is a hyperparameter.
- Probability distributions p_k are drawn themselves from a probability distribution over distributions, known as *base distribution* G_0.

Distributions of parameters

If each process is a normal distribution with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$, these can be drawn from the *normal inverse Wishart distribution*, $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \sim \mathrm{NIW}(\mathbf{m}, \lambda, \boldsymbol{\Psi}, \nu)$, with four hyperparameters:

- the *location vector* $\mathbf{m}$ lying in the feature space,
- the *mean fraction* λ,
- the *inverse scale matrix* $\boldsymbol{\Psi}$, which has to be symmetric and positive definite,
- and ν, which has to be at least the number of dimensions d of the feature space and regulates the degrees of freedom.

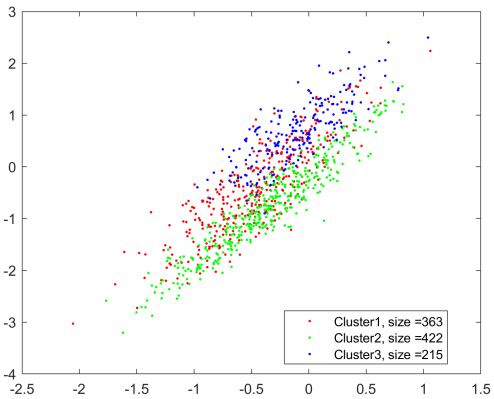
Distributions of parameters



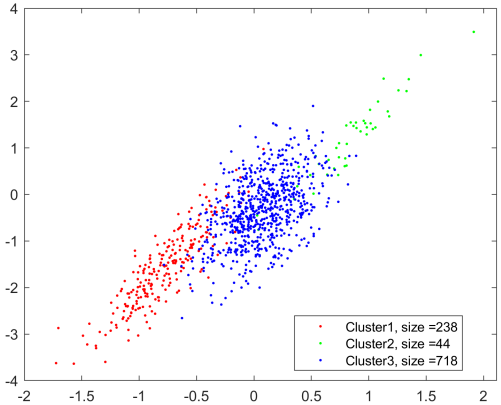
Effect of hyperparameters

- $\alpha = K$, then $\boldsymbol{\alpha} = (1, \ldots, 1)^T$ and the Dirichlet distribution is the uniform distribution over the simplex in which $\boldsymbol{\pi}$ lies. All possibilities for $\boldsymbol{\pi}$ are equally likely.
- As α increases, the Dirichlet probability density functions get more and more peaked at $\boldsymbol{\pi} = (1/K, \ldots, 1/K)^T$ with any other vectors becoming less likely.

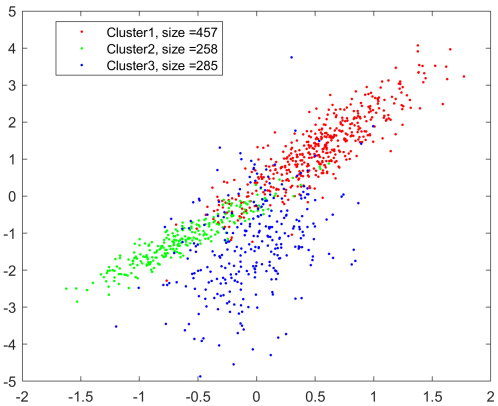
Effect of hyperparameters



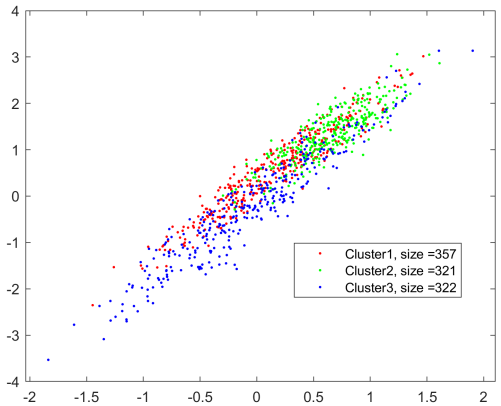
$\alpha = K = 3.$



$\alpha = K = 3.$



$\alpha = 10.$



$\alpha = 100.$

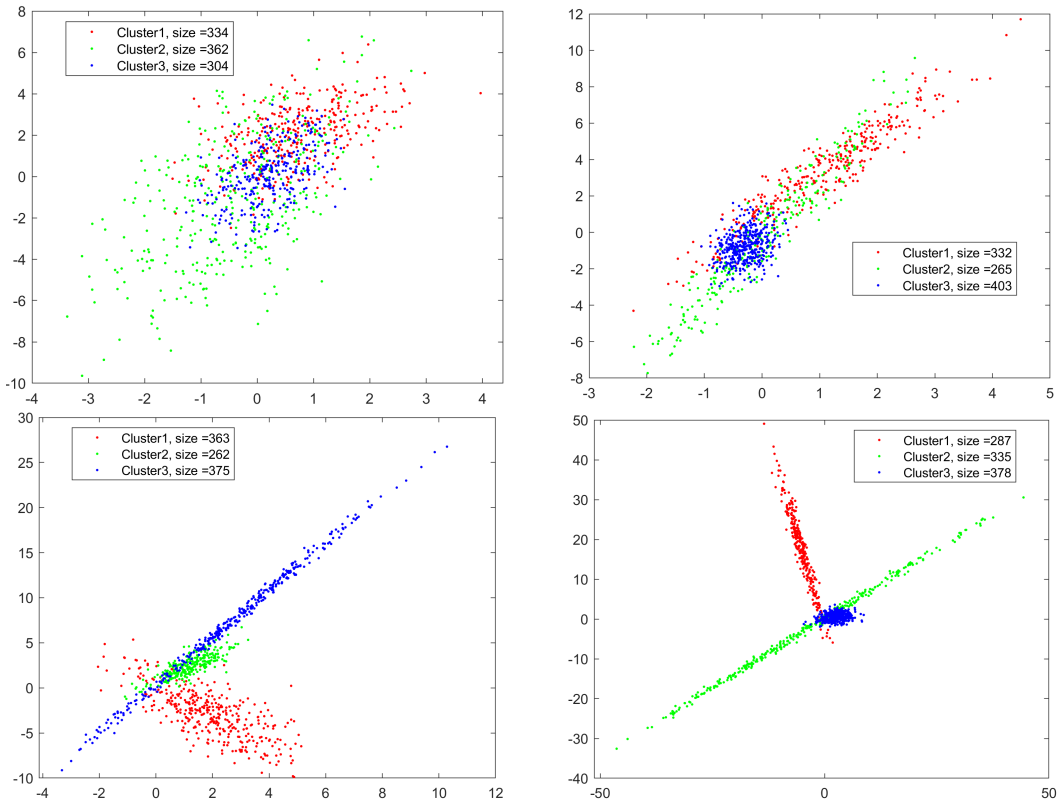
As α increases, cluster sizes approach N/K.

Effect of hyperparameters

The covariance matrix of each cluster follows an *Inverse Wishart distribution*, $\boldsymbol{\Sigma}_k \sim \mathcal{W}^{-1}(\boldsymbol{\Psi}, \nu)$.

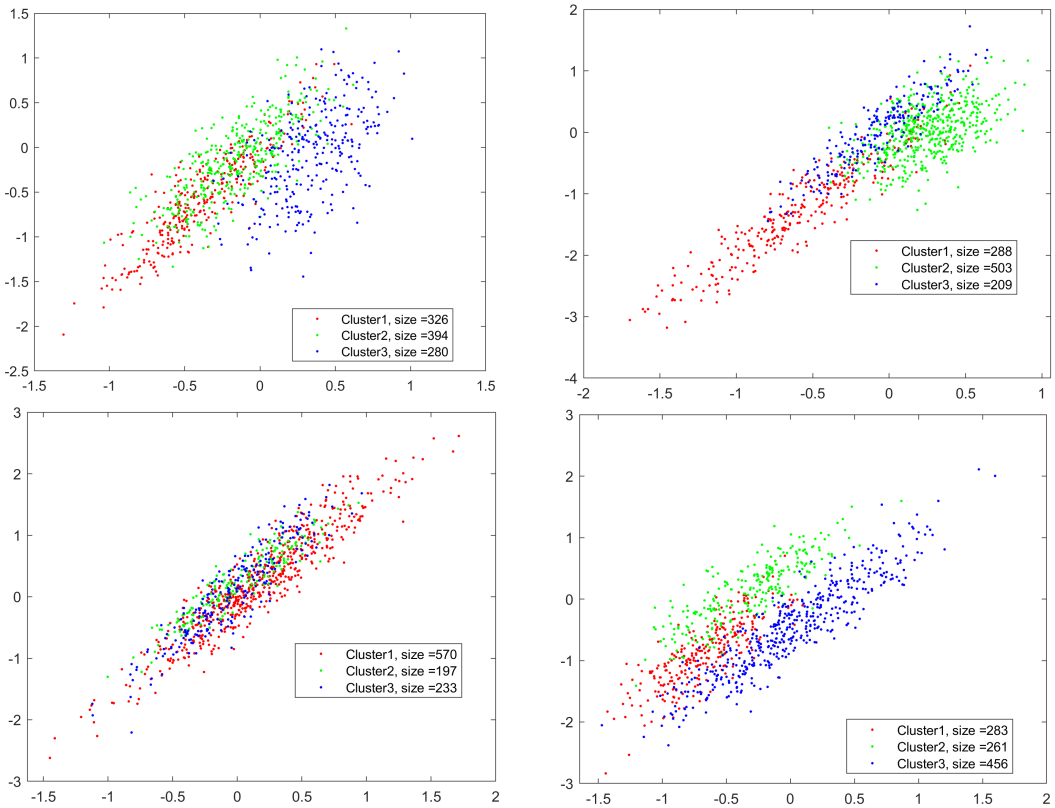
- The expectation is $\mathbb{E}[\boldsymbol{\Sigma}_k] = \dfrac{\boldsymbol{\Psi}}{\nu - d - 1}$.
- The variance of each element Σ_{ij} of $\boldsymbol{\Sigma}_k$ is
  $$\text{var}[\Sigma_{ij}] = \frac{(\nu - d + 1)\Psi_{ij}^2 + (\nu - d - 1)\Psi_{ii}\Psi_{jj}}{(\nu - d)(\nu - d - 1)^2(\nu - d - 3)} = \text{var}[\Sigma_{ji}],$$
  since $\boldsymbol{\Psi}$ is symmetric.
- On the diagonal, $\text{var}[\Sigma_{ii}] = \dfrac{2\Psi_{ii}^2}{(\nu - d - 1)^2(\nu - d - 3)}$.
- Since the power of ν is larger in the denominator than the numerator, it regulates the variability in $\boldsymbol{\Sigma}_k$, the larger ν, the more similar $\boldsymbol{\Sigma}_k$ to $\boldsymbol{\Psi}$.

Effect of hyperparameters



$\nu = 2$, $\alpha = 10K$, $\mathbf{m} = (0, 0)^T$, $\lambda = 1$ and $\mathbf{\Psi} = \begin{pmatrix} 1 & 3/2 \\ 3/2 & 1 \end{pmatrix}$.

Effect of hyperparameters



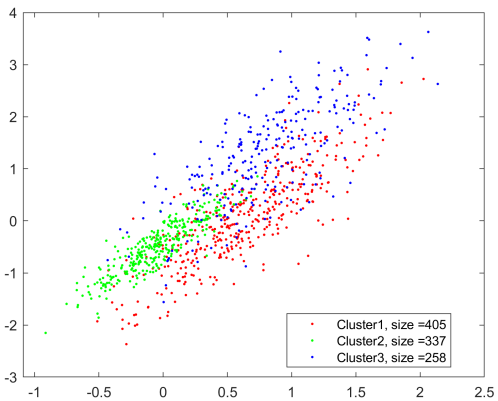
$\nu = 10$, $\alpha = 10K$, $\mathbf{m} = (0, 0)^T$, $\lambda = 1$ and $\mathbf{\Psi} = \begin{pmatrix} 1 & 3/2 \\ 3/2 & 1 \end{pmatrix}$.

Effect of hyperparameters

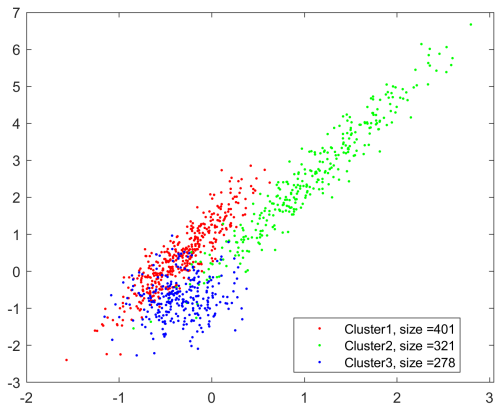
Having drawn $\mathbf{\Sigma}_k$, the cluster centre is drawn from a normal distribution, $\mathcal{N}(\mathbf{m}, \frac{1}{\lambda}\mathbf{\Sigma}_k)$.

- The parameter λ controls the spacing of the generated cluster centres.
- As λ decreases, the clusters separate.
- As λ increases, they overlap more and more.

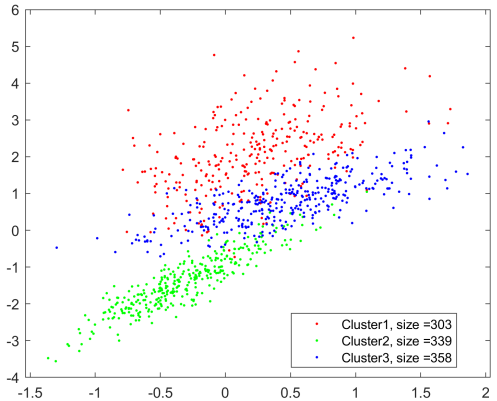
Effect of hyperparameters



$\lambda = 1,$



$\lambda = 1/3,$



$\lambda = 1/5.$

$$\nu = 6,\ \alpha = 10K,\ \mathbf{m} = (0,0)^T, \text{ and } \mathbf{\Psi} = \begin{pmatrix} 1 & 3/2 \\ 3/2 & 1 \end{pmatrix}.$$

Data generation

To summarize, the data is generated following the distributions:

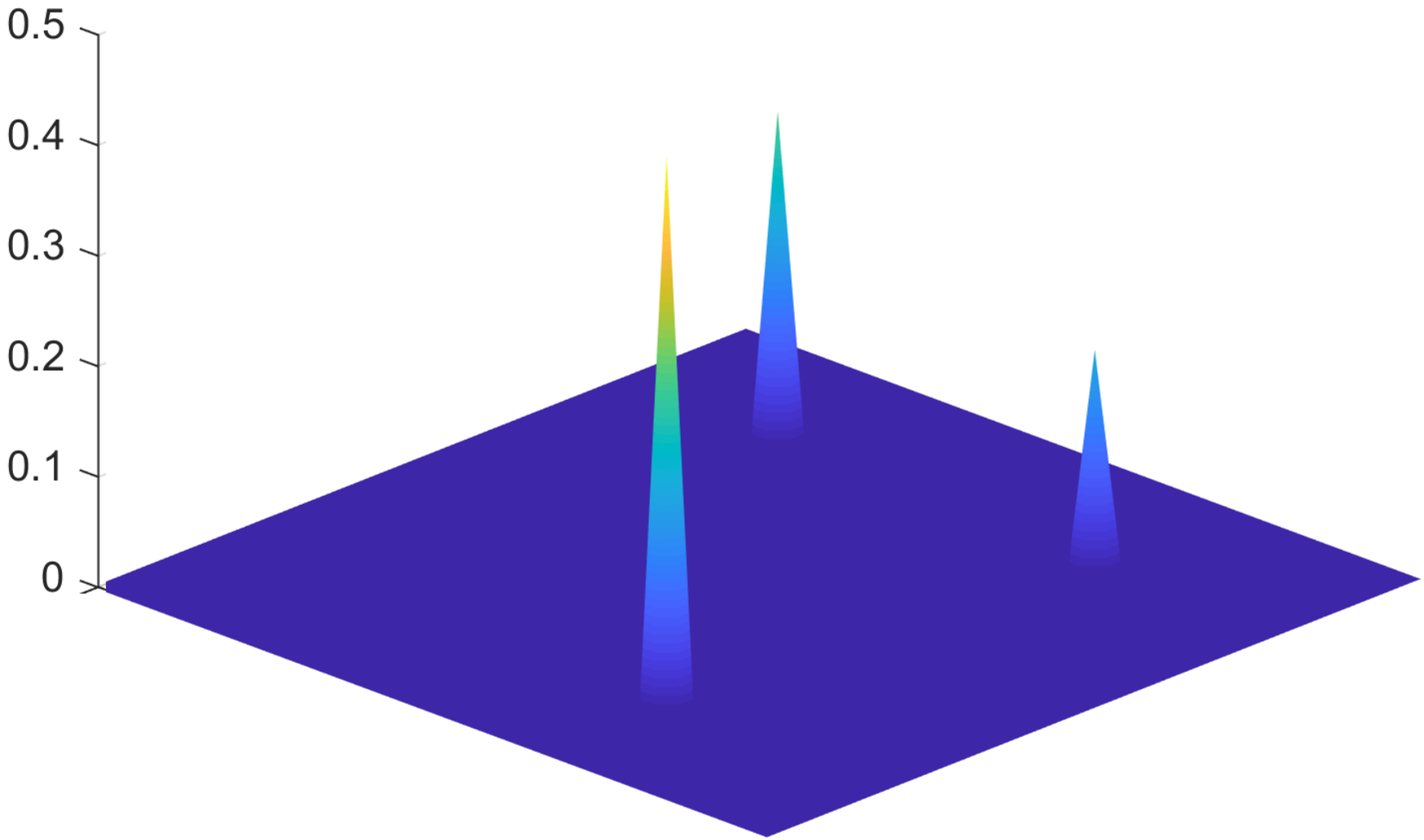
$$\begin{aligned} \mathbf{v}_n|\mathbf{z}_n &\sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ p(z_{n,k}=1) &= \pi_k, \\ \boldsymbol{\pi} &\sim \mathrm{Dir}(\boldsymbol{\alpha}), \\ \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k &\sim G_0 = \mathrm{NIW}(\boldsymbol{m}, \lambda, \boldsymbol{\Psi}, \nu). \end{aligned}$$

Remove latent variables $\mathbf{z}_n$ by imagining a distribution G over all possible pairs of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ which is zero everywhere apart from the specific pairs $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ where it has a point probability mass of π_k:

$$\begin{aligned} \mathbf{v}_n &\sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ \boldsymbol{\pi} &\sim \mathrm{Dir}(\boldsymbol{\alpha}), \\ \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k &\sim G. \end{aligned}$$

G_0 is indirectly part of G, since the points where G is nonzero are drawn from G_0, while the probability masses there are drawn from $\mathrm{Dir}(\boldsymbol{\alpha})$.

Illustration



Reverse

After playing with the model generating the data, we are now ready to infer the model parameters from the data.

- That is $\mathbf{m}_k, \lambda_k, \mathbf{\Psi}_k, \nu_k$ of K normal inverse Wishart distributions for each cluster.
- $\boldsymbol{\mu}_k, \mathbf{\Sigma}_k$ are drawn from these distributions.
- Only K and α are fixed user choices.

Initialization

Initialize the prior with the most uninformed parameters:

- $\mathbf{m}$ is the mean of *all* data.
- $s = \sum_{n=1}^{N} \|\mathbf{v}_n - \mathbf{m}\|^2/(d * N)$ is the average squared distance from the mean per dimension.
- $\mathbf{\Psi} = s\mathbf{I}$.
- $\lambda = 1$, clusters are overlapping.
- $\nu = d$, largest possible variability in the covariance matrices.

Initialization

- Assign clusters randomly.
- Calculate posterior $\mathbf{m}_k, \lambda_k, \mathbf{\Psi}_k, \nu_k$ for these assignments: For the k^{th} cluster, let N_k be the number of samples, $\bar{\mathbf{v}}_k$ its sample mean and $\mathbf{S}_k$ its sample covariance matrix.

$$\begin{aligned}
\mathbf{m}_k &= \frac{\lambda \mathbf{m} + N_k \bar{\mathbf{v}}_k}{\lambda + N_k}, \\
\lambda_k &= \lambda + N_k, \\
\nu_k &= \nu + N_k, \\
\mathbf{\Psi}_k &= \mathbf{\Psi} + N_k \mathbf{S}_k + \frac{\lambda N_k}{\lambda + N_k} (\bar{\mathbf{v}}_k - \mathbf{m})(\bar{\mathbf{v}}_k - \mathbf{m})^T.
\end{aligned}$$

- Draw $\boldsymbol{\mu}_k, \mathbf{\Sigma}_k$ from these posteriors.

Iterations

Adding a single data sample $\mathbf{v}_n$ to the k^{th} cluster. Then $\bar{\mathbf{v}}_k = \mathbf{v}_n$ and $\mathbf{S}_k = 0$.

$$
\begin{aligned}
\mathbf{m}_k^{\text{post}} &= \frac{\lambda_k^{\text{prior}} \mathbf{m}_k^{\text{prior}} + \mathbf{v}_n}{\lambda_k^{\text{prior}} + 1}, \\
\lambda_k^{\text{post}} &= \lambda_k^{\text{prior}} + 1, \\
\nu_k^{\text{post}} &= \nu_k^{\text{prior}} + 1, \\
\mathbf{\Psi}_k^{\text{post}} &= \mathbf{\Psi}_k^{\text{prior}} + \frac{\lambda_k^{\text{prior}}}{\lambda_k^{\text{prior}} + 1} (\mathbf{v}_n - \mathbf{m}_k^{\text{prior}})(\mathbf{v}_n - \mathbf{m}_k^{\text{prior}})^T.
\end{aligned}
$$

Iterations

Removing a single data sample $\mathbf{v}_n$ from the l^{th} cluster.

$$
\begin{aligned}
\mathbf{m}_l^{\text{prior}} &= \frac{\lambda_l^{\text{post}}\mathbf{m}_l^{\text{post}} - \mathbf{v}_n}{\lambda_l^{\text{post}} - 1}, \\
\lambda_l^{\text{prior}} &= \lambda_l^{\text{post}} - 1, \\
\nu_l^{\text{prior}} &= \nu_l^{\text{post}} - 1, \\
\mathbf{\Psi}_l^{\text{prior}} &= \mathbf{\Psi}_l^{\text{post}} - \frac{\lambda_l^{\text{post}}}{\lambda_l^{\text{post}} - 1}(\mathbf{v}_n - \mathbf{m}_l^{\text{post}})(\mathbf{v}_n - \mathbf{m}_l^{\text{post}})^T.
\end{aligned}
$$

Iterations

- Let $\mathcal{D} \setminus \{\mathbf{v}_n\}$ denote the set of data samples excluding $\mathbf{v}_n$ and their current cluster assignments.
- The current cluster assignments determine the current values of $N_k, \mathbf{m}_k, \lambda_k, \mathbf{\Psi}_k, \nu_k$.
- From these the current $\boldsymbol{\mu}_k, \mathbf{\Sigma}_k$ are drawn.
- Calculate the probability of $\mathbf{v}_n$ belonging to any of the K clusters, $p(z_{n,k} = 1 | \mathbf{v}_n, \mathcal{D} \setminus \{\mathbf{v}_n\}, \alpha)$.

Iterations

$$p(z_{n,k}=1|\mathbf{v}_n, \mathcal{D}\setminus\{\mathbf{v}_n\}, \alpha) = p(z_{n,k}=1|\mathcal{D}\setminus\{\mathbf{v}_n\}, \alpha) \times p(\mathbf{v}_n|\mathcal{D}\setminus\{\mathbf{v}_n\}, z_{n,k}=1)$$

- The first factor is the probability of $\mathbf{v}_n$ belonging to cluster k which is governed by the posterior Dirichlet distribution, given the prior Dirichlet distribution and all other cluster assignments with parameter $\boldsymbol{\alpha} = \left(\frac{N_1+\alpha/K}{N-1+\alpha}, \ldots, \frac{N_k+\alpha/K}{N-1+\alpha}\right)$.
- The second factor is the likelihood of seeing sample $\mathbf{v}_n$ given $\mathcal{D}\setminus\{\mathbf{v}_n\}, z_{n,k}=1$, and the parameters of the posterior, normal, inverse Wishart distribution for cluster k.

Iterations

$$p(z_{n,k} = 1 | \mathbf{v}_n, \mathcal{D} \setminus \{\mathbf{v}_n\}, \alpha) \approx$$

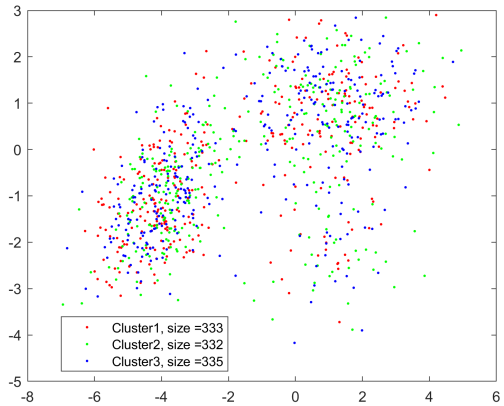
$$\frac{N_k + \alpha/K}{N - 1 + \alpha} \times \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_k|}} \exp\left(\frac{1}{2}(\mathbf{v}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{v}_n - \boldsymbol{\mu}_k)\right).$$

Since these are approximations we need to divide by the sum of these probabilities to ensure they add to one.

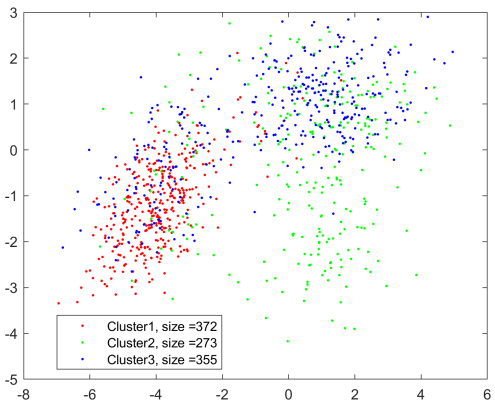
Iterations

- Sample cluster assignment for $\mathbf{v}_n$.
- Update $N_k, \mathbf{m}_k, \lambda_k, \mathbf{\Psi}_k, \nu_k$ of that cluster.
- Choose next random sample, remove it from its current cluster, sample new cluster assignment, update.
- One iteration is complete, if all samples have been considered in a random order.
- Repeat for a fixed number of iterations.

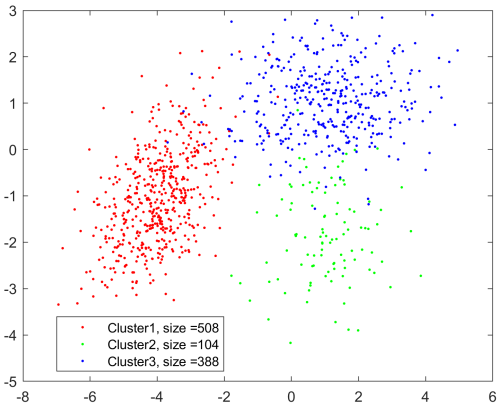
Illustration



Initial random cluster assignment.

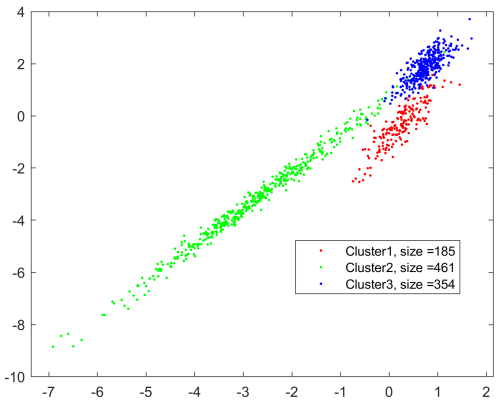


After 5 iterations.

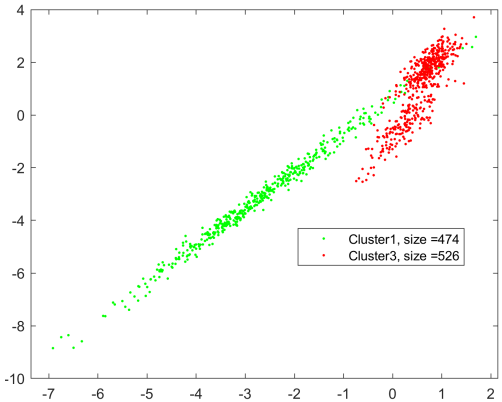


After 100 iterations.

Illustration



Original data.



Final assignments.

Indefinite Number of Clusters

- Fixing the number of clusters is undesirable.
- A new sample can be generated from existing processes or a completely new process.
- *Base distribution*: A distribution from which the distributions of the processes are generated.
- *Dispersion, concentration, scaling parameter or strength*: α.
- Probability that the second sample is generated by the first process: $1/(1+\alpha)$.
- Probability that it is generated by a new process: $\alpha/(1+\alpha)$.
- If $\alpha = 1$, both are one half.
- If $\alpha > 1$, then a new process is favoured.
- If $\alpha < 1$, the existing process is more likely to generate it.

Indefinite Number of Clusters

- The n^{th} sample is generated by

$$\begin{cases} \text{process } k \text{ with probability} & \dfrac{n_k}{n-1+\alpha} \\ \text{a new process with probability} & \dfrac{\alpha}{n-1+\alpha} \end{cases},$$

  where n_k is the number of samples generated by process k so far.
- When summing the samples generated by each process over all processes, the result is $n-1$. Therefore the probabilities sum to 1.
- Note that as more and more samples are generated by a particular process, it gets more likely that this process will generate further samples, since $n_k/(n-1+\alpha)$ increases. This is known as *rich-get-richer*.

Chinese Restaurant Process

Chinese Restaurant Process (CRP)

Reverse

Initialize the prior with the most uninformed parameters:

- $\mathbf{m}$ is the mean of *all* data.
- $s = \sum_{n=1}^{N} \|\mathbf{v}_n - \mathbf{m}\|^2/(d * N)$ is the average squared distance from the mean per dimension.
- $\mathbf{\Psi} = s\mathbf{I}$.
- $\lambda = 1$, clusters are overlapping.
- $\nu = d$, largest possible variability in the covariance matrices.
- $\alpha = 0.5$.

Initialization

- Consider data samples in a random order.
- Let $\mathbf{v}_1$ be first sample and thus the first cluster.

$$\begin{aligned}
\mathbf{m}_1 &= \frac{\lambda \mathbf{m} + \mathbf{v}_1}{\lambda + 1}, \\
\lambda_1 &= \lambda + 1, \\
\nu_1 &= \nu + 1, \\
\mathbf{\Psi}_1 &= \mathbf{\Psi} + \frac{\lambda}{\lambda + 1}(\mathbf{v}_1 - \mathbf{m})(\mathbf{v}_1 - \mathbf{m})^T.
\end{aligned}$$

- Draw $\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1$ from this posterior.

Initialization

- Let $\mathbf{v}_n$ be the n^{th} sample.
- Let $\mathcal{D}_n = \{\mathbf{v}_1, \ldots, \mathbf{v}_{n-1}\}$.
- Let K be the current number of clusters.
- For each cluster, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ have been drawn.
- Estimate the probability $\mathbf{v}_n$ belongs to existing cluster k as

$$p(z_{n,k} = 1 | \mathcal{D}_n, \alpha) \approx \frac{N_k}{n - 1 + \alpha} \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_k|}} \exp\left(\frac{1}{2}(\mathbf{v}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{v}_n - \boldsymbol{\mu}_k)\right).$$

Initialization

- Estimate the probability $\mathbf{v}_n$ belongs to a new cluster as

$$p(z_{n,K+1} = 1 | \mathcal{D}_n, \alpha) \approx \frac{\alpha}{n-1+\alpha} \frac{1}{\sqrt{|2\pi\mathbf{\Psi}|}} \exp\left(\frac{1}{2}(\mathbf{v}_n - \mathbf{m})^T \mathbf{\Psi}^{-1} (\mathbf{v}_n - \mathbf{m})\right).$$

- Divide by the sum of these probabilities to ensure they add to one.
- Draw a cluster assignment.

Initialization

- If assigned to existing cluster k, update $N_k, \mathbf{m}_k, \lambda_k, \mathbf{\Psi}_k, \nu_k$ and draw new $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$.
- If assigned to new cluster $K+1$, $N_{K+1} = 1$ and calculate

$$\begin{aligned}
\mathbf{m}_{K+1} &= \frac{\lambda \mathbf{m} + \mathbf{v}_n}{\lambda + 1}, \\
\lambda_{K+1} &= \lambda + 1, \\
\nu_{K+1} &= \nu + 1, \\
\mathbf{\Psi}_{K+1} &= \mathbf{\Psi} + \frac{\lambda}{\lambda + 1}(\mathbf{v}_n - \mathbf{m})(\mathbf{v}_n - \mathbf{m})^T.
\end{aligned}$$

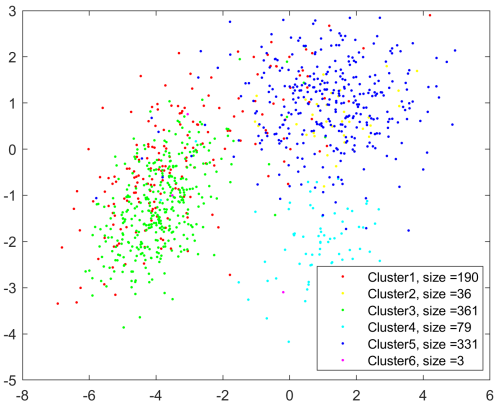
  and draw $\boldsymbol{\mu}_{K+1}$ and $\boldsymbol{\Sigma}_{K+1}$.
- Once all samples are assigned a cluster, the initialization is complete.

Initialization

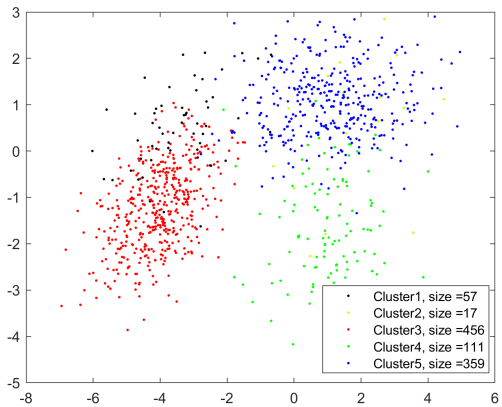
- Each iteration consider all samples in a random order.
- Remove sample $\mathbf{v}_n$ from its current cluster k.
- If the cluster is empty, remove it.
- Otherwise, update parameters $N_k, \mathbf{m}_k, \lambda_k, \mathbf{\Psi}_k, \nu_k$.
- Draw new $\boldsymbol{\mu}_k$ and $\mathbf{\Sigma}_k$.
- Approximate probabilities of $\mathbf{v}_n$ belonging to an existing cluster, or an unseen cluster.
- Divide by the sum of these probabilities to ensure they add to one.
- Draw a cluster assignment.
- Update.....

Dirichlet Process Method

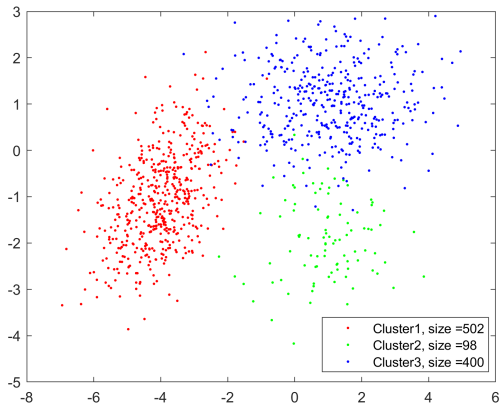
This algorithm is known as *Dirichlet Process Method*.



Initial cluster assignment.

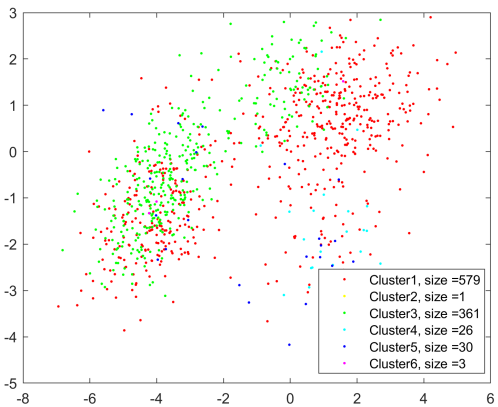


After 10 iterations

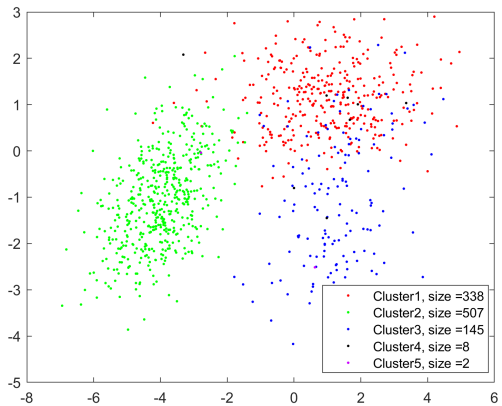


After 20 iterations.

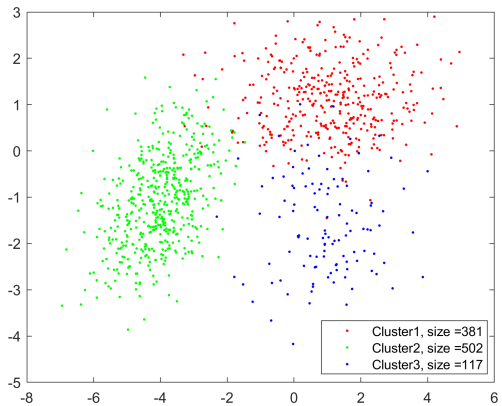
Dirichlet Process Method



Initial cluster assignment.



After 10 iterations



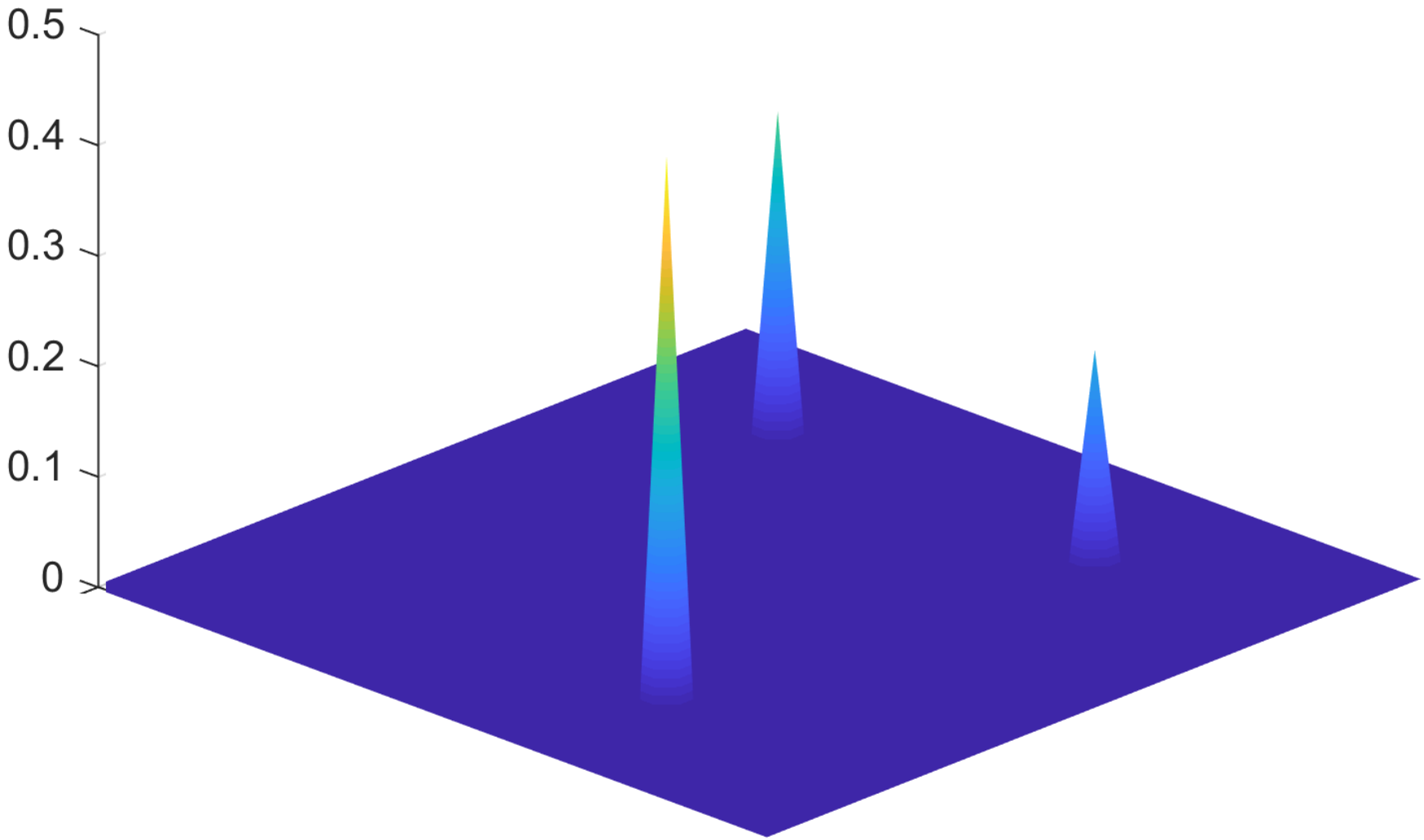
After 20 iterations.

Data generation

Recall for fixed K the distribution G over all possible pairs of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ which is zero everywhere apart from the specific pairs $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ where it has a point probability mass of π_k. The probability masses for these K pairs follow a Dirichlet distribution with parameter $\boldsymbol{\alpha}$

$$\begin{aligned} \mathbf{v}_n &\sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ \boldsymbol{\pi} &\sim \mathrm{Dir}(\boldsymbol{\alpha}), \\ \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k &\sim G. \end{aligned}$$

Illustration



Dirichlet Process

A *Dirichlet process* extends the concept to an unknown variable number of clusters K. The notation is

$$\begin{aligned} \mathbf{v}_n &\sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k &\sim G, \\ G &\sim \mathrm{DP}(\alpha, G_0). \end{aligned}$$

It is a distribution of distributions.

Dirichlet Process

- Infinite many possibilities for pairs $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- Length of $\boldsymbol{\pi}$ is often described as infinite.
- In practice, the number of clusters K is at most the number of samples N.
- Therefore G gives a probability mass of $N_k/(N-1+\alpha)$ to at most N pairs $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ drawn from the base distribution G_0, and assigns the probability of $\alpha/((N-1+\alpha)$ to the set of all other possible pairs.
- G is a discrete distribution defined on a finite partition of the space S of all pairs $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Dirichlet Process

- More formally, a distribution G is drawn from a *Dirichlet process* with parameters α and G_0, if for *any finite, disjoint partition* $S_1, \ldots, S_L$ of S, where L can be any finite number, the vector $(G(S_1), \ldots, G(S_L))^T$ follows a Dirichlet distribution with parameters $\alpha G_0(S_1), \ldots, \alpha G_0(S_L)$.
- *Bayesian nonparametrics* which are Bayesian models operating on an infinite-dimensional parameter space. In our case, this is the space of all possible pairs $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- See "Bayesian Nonparametrics" by J.K. Ghosh and R.V. Ramamoorthi.