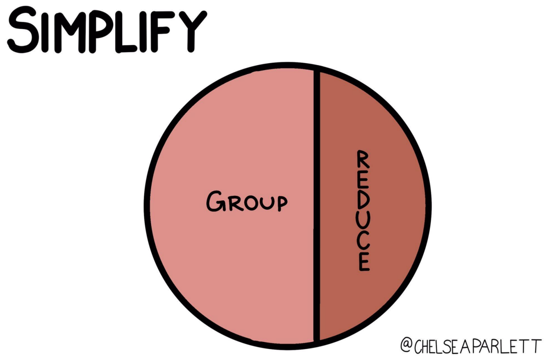
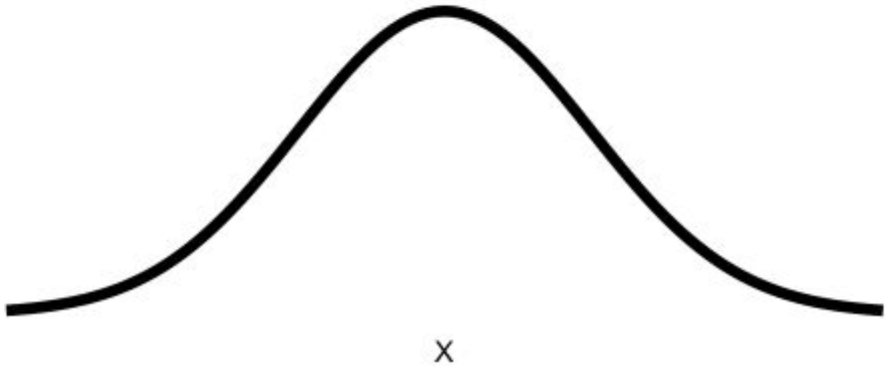


Gaussian Mixtures

Dr. Chelsea Parlett-Pelleriti

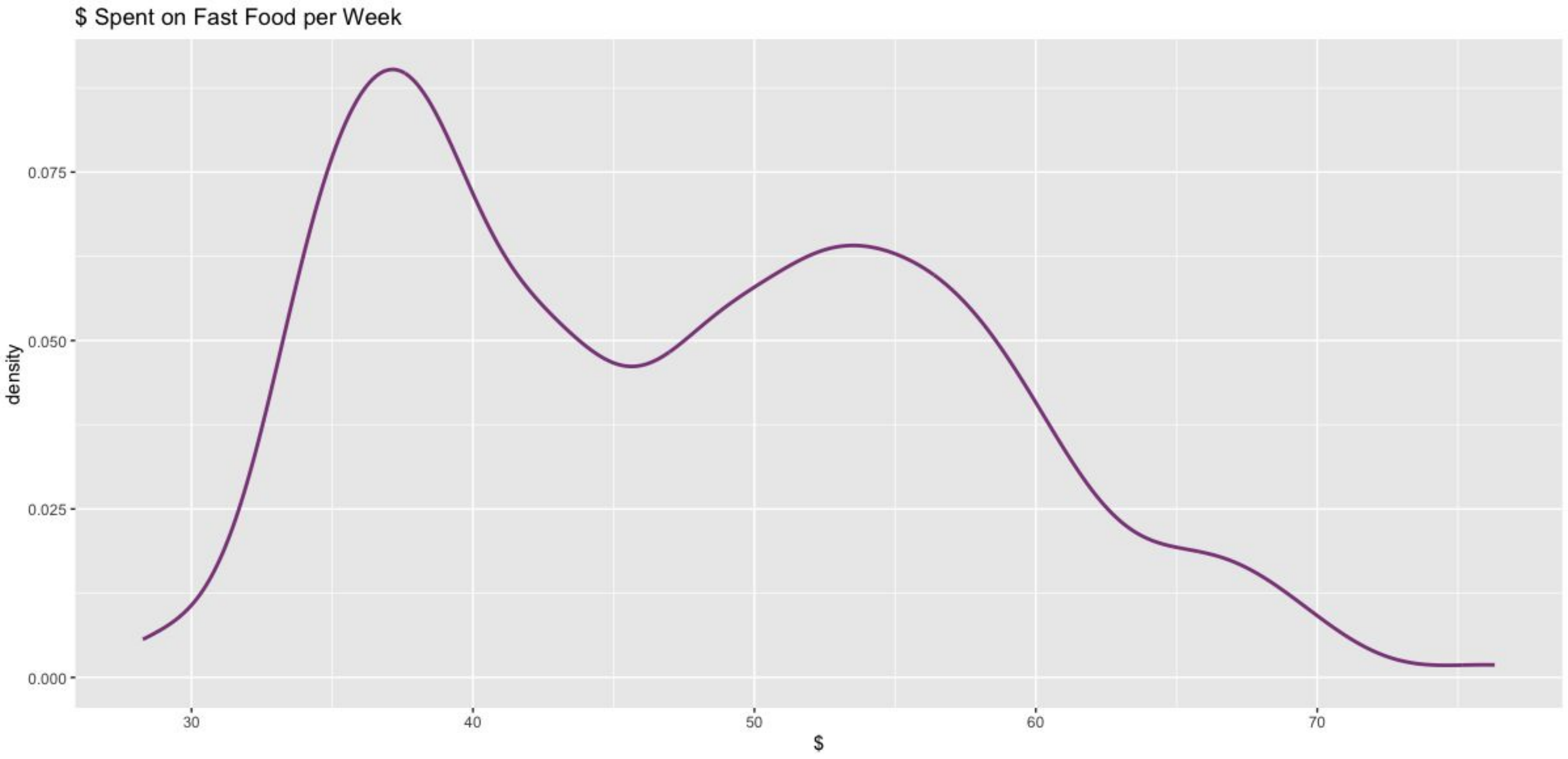
Normal (Gaussian) Distribution



$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

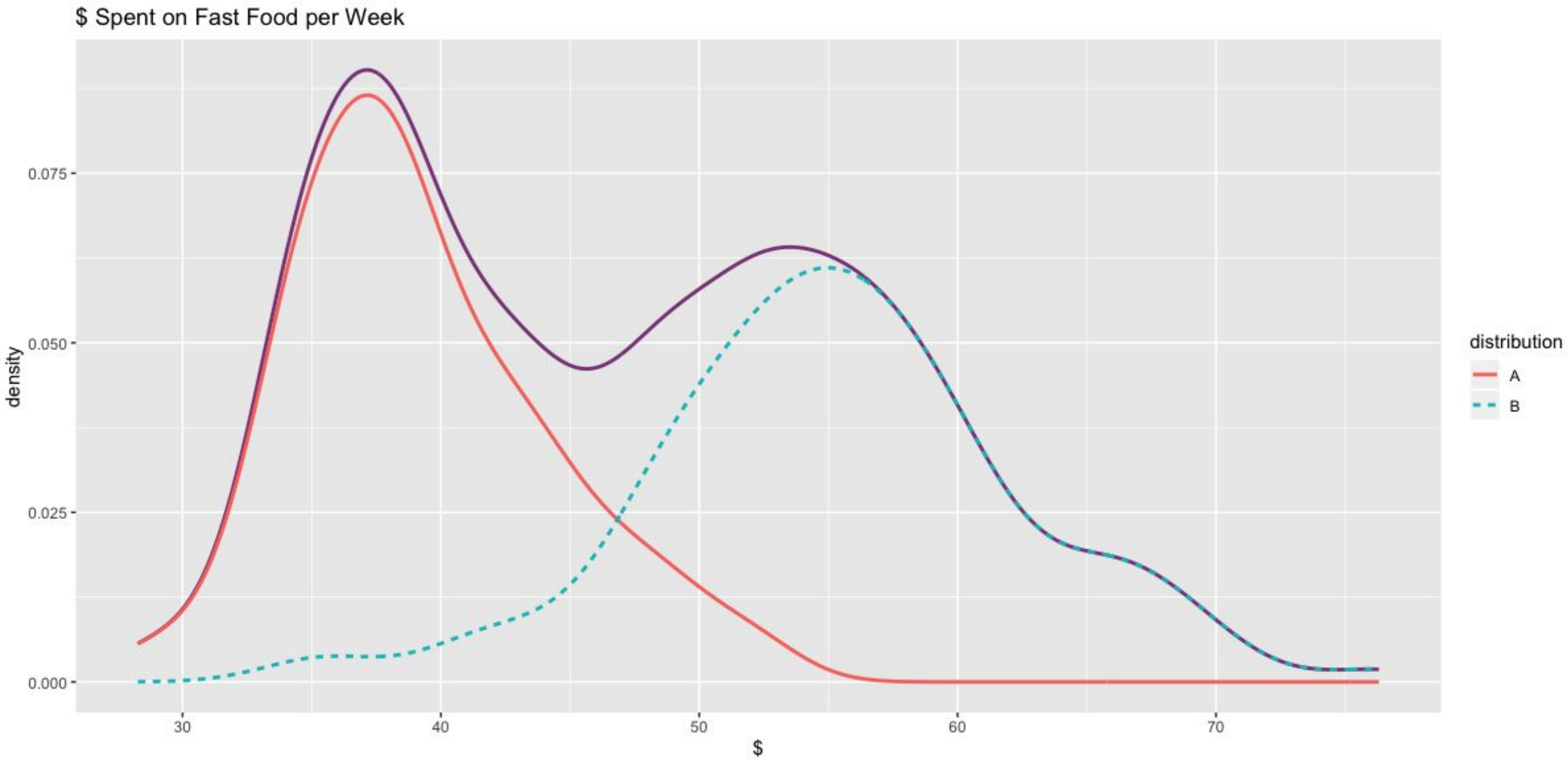
$\mu =$ Mean
$\sigma =$ Standard Deviation
$\pi \approx 3.14159\cdots$
$e \approx 2.71828\cdots$

GMM



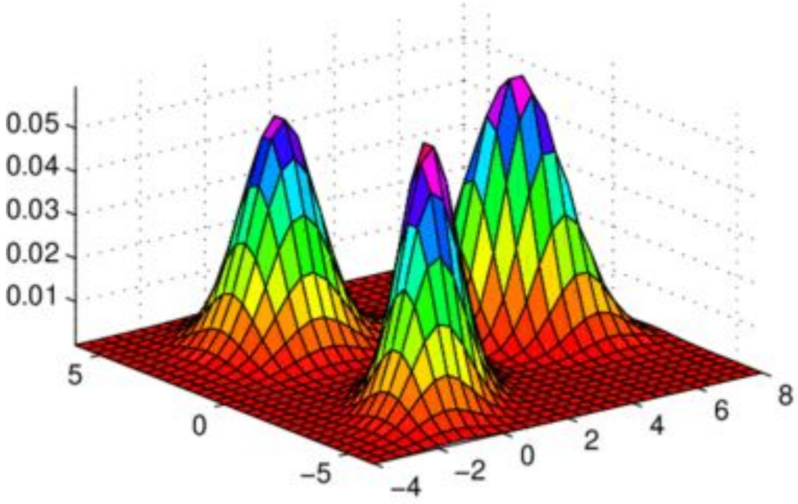
$ Spent on Fast Food per Week
0.075
0.050
0.025
0.000
density
30
40
50
60
70
$

GMM

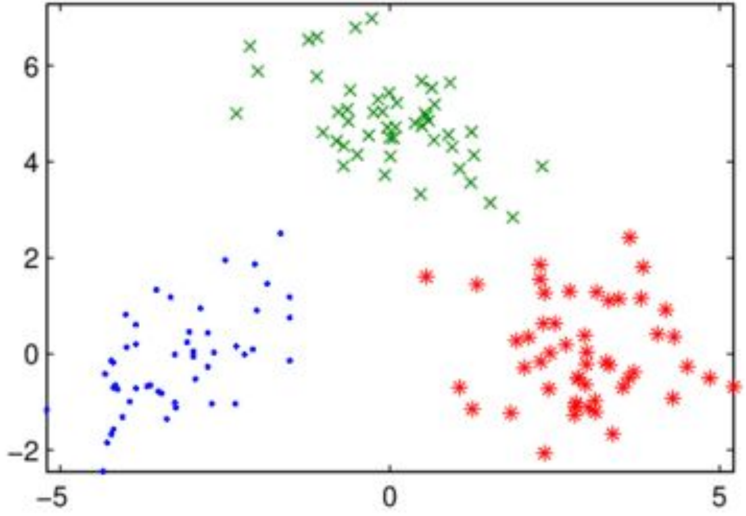

$ Spent on Fast Food per Week
density
0.000
0.025
0.050
0.075
30
40
50
60
70
$
distribution
A
B

Multivariate Normal Distributions

(a) 3 components Gaussian mixture density



(b) Data from 3 components Gaussian mixture density



Assume:

- Latent Groups
- Groups are Gaussian

GMM

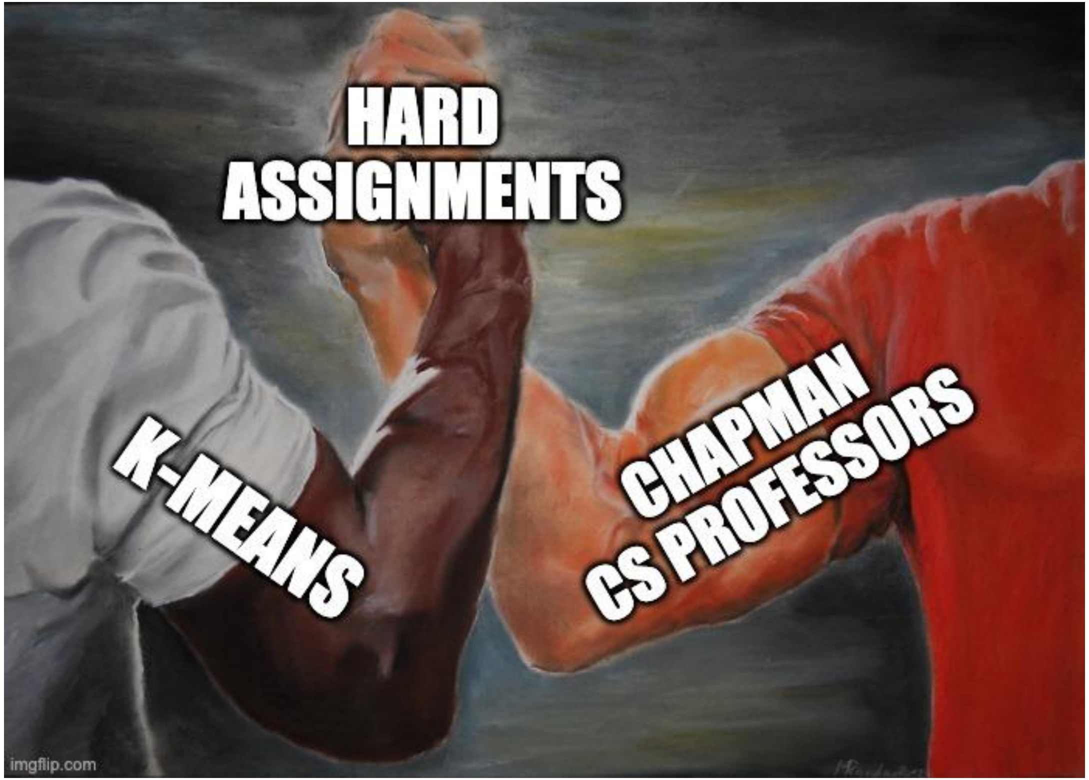
K means

- Hard Assignment
- All Variances the Same
- Roughly same # of datapoints

GMM

- Soft (probabilistic) Assignment
- Variances can be different
- Explicitly model # of data points

GMM


HARD
ASSIGNMENTS
K-MEANS
CHAPMAN
CS PROFESSORS
imgflip.com

GMM

$$\underbrace{J}_{\text{distortion}} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2$$

Goal: choose r_{nk} and \mu_k that minimize J

GMM

$$\underbrace{J}_{\text{distortion}} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \, \|x_n - \mu_k\|^2$$

$$r_{nk} = \begin{cases} 1 & \text{if } \; k = \arg\min_j \|x_n - \mu_j\|^2 \\ 0 & \text{otherwise} \end{cases}$$

GMM

$$\underbrace{J}_{\text{distortion}} = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk} \left\|x_n - \mu_k\right\|^2$$

$$\underbrace{2\sum_{n=1}^{N} r_{nk}(x_n - \mu_k) = 0}_{\text{derivative of} J}$$

GMM

$$\underbrace{J}_{\text{distortion}} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2$$

$$\underbrace{2 \sum_{n=1}^{N} r_{nk}(x_n - \mu_k) = 0}_{\text{derivative of} J}$$

$$\mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}}$$

GMM

$$J = \sum^{N} \sum^{K} r_{nk} \|x_n - \mu_k\|^2$$

Surprise, it's K-Means

derivative of J

$$\mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}}$$

GMM

$$\underbrace{J}_{\text{distortion}} = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\,\|x_n - \mu_k\|^2$$

when r_{nk} can only be 0 or 1 **(hard assignment)**

$$\underbrace{2\sum_{n=1}^{N} r_{nk}(x_n - \mu_k) = 0}_{\text{derivative of} J}$$

$$\mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}}$$

GMM

$$\underbrace{J}_{\text{distortion}} = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|x_n - \mu_k\|^2$$

when r_{nk} can only be 0 or 1 **(hard assignment)**

$$\underbrace{2\sum_{n=1}^{N} r_{nk}(x_n - \mu_k) = 0}_{\text{derivative of} J}$$

The value of \mu_{k} that minimizes our loss is the mean of all data points belonging to a cluster

$$\mu_k = \frac{\sum_n r_{nk}x_n}{\sum_n r_{nk}}$$

GMM

$$\underbrace{J}_{\text{distortion}} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2$$

when r_{nk} can only be 0 or 1 (**hard assignment**)

$$\underbrace{2\sum_{n=1}^{N} r_{nk}(x_n - \mu_k) = 0}_{\text{derivative of} J}$$

The value of \mu_{k} that minimizes our loss is the mean of all data points belonging to a cluster

$$\mu_k = \frac{\sum_n r_{nk} x_n}{N_k} = \frac{1}{N_k} \sum_n r_{nk} x_n$$

K-Means Algorithm

1. Choose **k** random points to be cluster centers
2. For each data point, assign it to the cluster whose center is closest
3. Using these assignments, recalculate the **centers**
4. Repeat 2 and 3 until either:
   a. Cluster membership does not change
   b. Centers change only a tiny amount

Gaussian Mixture Model (EM Algorithm)

1. Choose **k** random points to be cluster centers (**or estimate using k-means...etc)**
2. For each data point, calculate the **probability** of belonging to each cluster
3. Using these probability weights, recalculate the **means + variances** (and weights)
4. Repeat 2 and 3 until **distributions converge.**

Mixtures

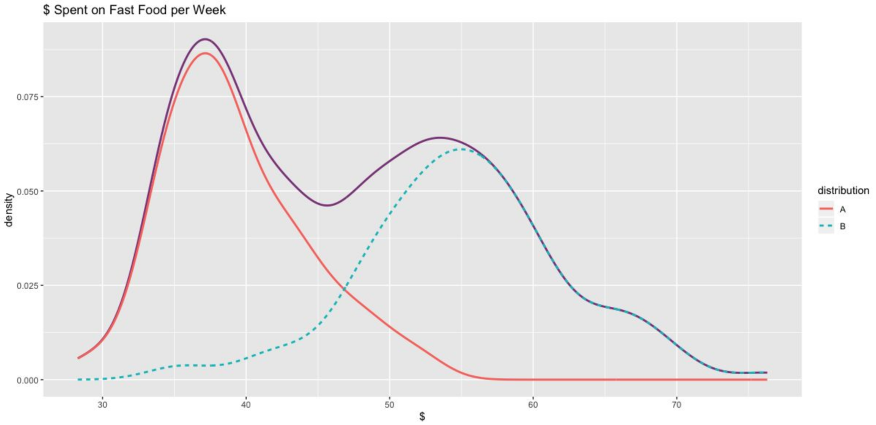
$$p(x) = \sum_{k=1}^{K} w_k \, p_k(x)$$

Mixtures

$$p(x) = \sum_{k=1}^{K} w_k \boxed{p_k(x)}$$

$$p(x) = \sum_{k=1}^{K} w_k \boxed{\mathcal{N}(x|\mu_k, \Sigma_k)}$$

Mixtures



$$p(x) = \sum_{k=1}^{K} w_k \, \mathcal{N}(x|\mu_k, \Sigma_k)$$

Mixtures

Probability of being in group k

Likelihood of seeing x in group k

$$p(x) = \sum_{k=1}^{K} \boxed{w_k} \boxed{\mathcal{N}(x|\mu_k, \Sigma_k)}$$

Mixtures

Probability of being in group k

Likelihood of seeing x in group k

$$w_k \, \mathcal{N}(x|\mu_k, \Sigma_k)$$

Posterior Probabilities

Prior probability of being in cluster k

Likelihood of seeing x in cluster k

$$p(\text{cluster } k|x) = \frac{w_k \, \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} w_j \, \mathcal{N}(x|\mu_j, \Sigma_j)}$$

Posterior probability of being in cluster k

Maximum Likelihood Estimation

Probability of being in group k

Likelihood of seeing x in group k

$$p(x) = \sum_{k=1}^{K} \boxed{w_k} \boxed{\mathcal{N}(x|\mu_k, \Sigma_k)}$$

$$p(\mathbf{X}|\mathbf{w}, \mu, \mathbf{\Sigma}) = p(x_1, x_2, ..., x_n|\mathbf{w}, \mu, \mathbf{\Sigma}) =$$

$$\prod_{n=1}^{N} \sum_{k=1}^{K} \boxed{w_k} \boxed{\mathcal{N}(x_n|\mu_k, \Sigma_k)}$$

Maximum Likelihood Estimation

$$p(x) = \sum_{k=1}^{K} \underbrace{w_k}_{\text{Probability of being in group k}} \underbrace{\mathcal{N}(x|\mu_k, \Sigma_k)}_{\text{Likelihood of seeing x in group k}}$$

$$p(\mathbf{X}|\mathbf{w}, \mu, \mathbf{\Sigma}) = p(x_1, x_2, ..., x_n|\mathbf{w}, \mu, \mathbf{\Sigma}) =$$

$$\prod_{n=1}^{N} \sum_{k=1}^{K} w_k \mathcal{N}(x_n|\mu_k, \Sigma_k)$$

$$log(p(\mathbf{X}|\mathbf{w}, \mu, \mathbf{\Sigma})) = \sum_{n=1}^{N} log \left\{ \sum_{k=1}^{K} w_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \right\}$$

Maximum Likelihood Estimation

Probability of being in group k

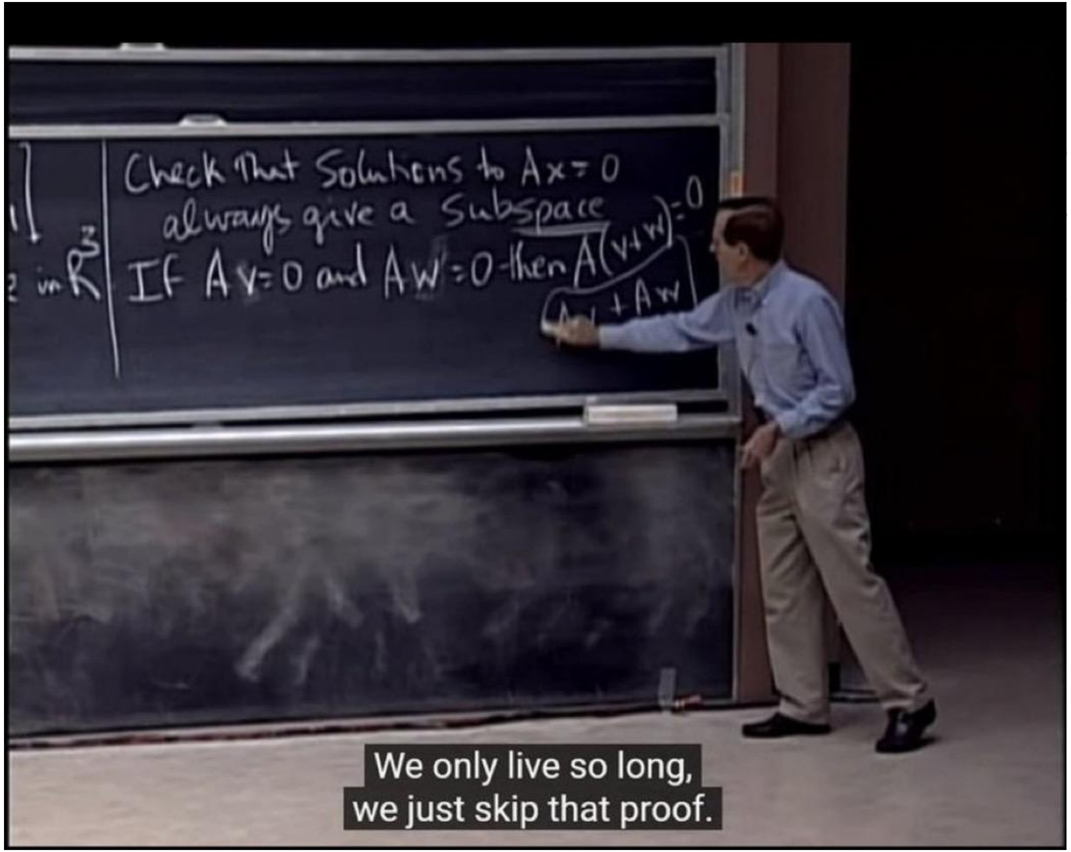
Likelihood of seeing x in group k

$$p(x) = \sum_{k=1}^{K} \boxed{w_k} \boxed{\mathcal{N}(x|\mu_k, \Sigma_k)}$$

$$log(p(\mathbf{X}|\mathbf{w}, \mu, \mathbf{\Sigma})) = \sum_{n=1}^{N} log\left\{\sum_{k=1}^{K} w_k \, \mathcal{N}(x_n|\mu_k, \Sigma_k)\right\}$$

Goal: choose w, μ, Σ that maximize the log likelihood

Maximum Likelihood Estimation



Probability of being in group k

Likelihood of seeing x in group k

$$p(x) = \sum_{k=1}^{K} \boxed{w_k} \boxed{\mathcal{N}(x|\mu_k, \Sigma_k)}$$

$$w_k \, \mathcal{N}(x_n|\mu_k, \Sigma_k)$$

w_k $N(x|\mu_k, \Sigma_k)$

Formulas (E-Step)

$$r_{nk} = \frac{w_k N(x_n|\mu_k, \Sigma_k)}{\sum_j w_j N(x_n|\mu_j, \Sigma_j)}$$

Formulas (E-Step)

$$w_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

$$r_{nk} = \frac{w_k N(x_n|\mu_k, \Sigma_k)}{\sum_j w_j N(x_n|\mu_j, \Sigma_j)}$$

Responsibilities are the posterior probability of a data point being in cluster k

Prior probability of being in cluster k

Likelihood of seeing x in cluster k

$$p(\text{cluster } k|x) = \frac{w_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} w_j \mathcal{N}(x|\mu_j, \Sigma_j)}$$

Posterior probability of being in cluster k

Formulas (M-Step)

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} x_n$$

$$N_k = \sum_{n=1}^{N} r_{nk}$$

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (x_n - \mu_k)(x_n - \mu_k)^T$$

Formulas (M-Step)

$$\mu_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}x_n \qquad N_k = \sum_{n=1}^{N} r_{nk}$$

$$\Sigma_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}(x_n - \mu_k)(x_n - \mu_k)^T$$

Formulas (M-Step)

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \boxed{r_{nk}} \boxed{x_n}$$

$$N_k = \sum_{n=1}^{N} r_{nk}$$

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \boxed{r_{nk}} \boxed{(x_n - \mu_k)(x_n - \mu_k)^T}$$

Formulas (M-Step)

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} x_n$$

$$N_k = \sum_{n=1}^{N} r_{nk}$$

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (x_n - \mu_k)(x_n - \mu_k)^T$$

Formulas (M-Step)

$$\mu_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk} x_n$$

$$N_k = \sum_{n=1}^{N} r_{nk}$$

$$\Sigma_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}(x_n - \mu_k)(x_n - \mu_k)^T$$

$$\pi_k = \frac{N_k}{N}$$

Comparison to K-Means

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} x_n$$

$$N_k = \sum_{n=1}^{N} r_{nk}$$

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (x_n - \mu_k)(x_n - \mu_k)^T$$

$$\pi_k = \frac{N_k}{N}$$

Math Summary

- GMM does **soft assignment**, every data point belongs to every cluster with some probability
- Data points that are more likely to be in a cluster have **more influence** over its parameters
- GMM uses the EM algorithm to iteratively update the cluster distributions. It first assigning a responsibility to each data point (**E-step**), and then using them to calculate weighted means and variances for each cluster (**M-step**)
- Responsibilities measure the **probability of a data point being in each cluster** (technically the **posterior probability**).
- Responsibilities contain information about **how common a cluster is** as well as the l**ikelihood of a data point belonging to that cluster**

Math Summary

- GMM does **soft assignment**, every data point belongs to every cluster with some probability
- Data points that are more likely to be in a cluster have **more influence** over its parameters
- GMM uses the EM algorithm to iteratively update the cluster distributions. It first assigning a responsibility to each data point (**E-step**), and then using them to calculate weighted means and variances for each cluster (**M-step**)
- Responsibilities measure the **probability of a data point being in each cluster** (technically the **posterior probability**).
- Responsibilities contain information about **how common a cluster is** as well as the l**ikelihood of a data point belonging to that cluster**

Math Summary

- GMM does **soft assignment**, every data point belongs to every cluster with some probability
- Data points that are more likely to be in a cluster have **more influence** over its parameters
- GMM uses the EM algorithm to iteratively update the cluster distributions. It first assigning a responsibility to each data point (**E-step**), and then using them to calculate weighted means and variances for each cluster (**M-step**)
- Responsibilities measure the **probability of a data point being in each cluster** (technically the **posterior probability**).
- Responsibilities contain information about **how common a cluster is** as well as the l**ikelihood of a data point belonging to that cluster**

Math Summary

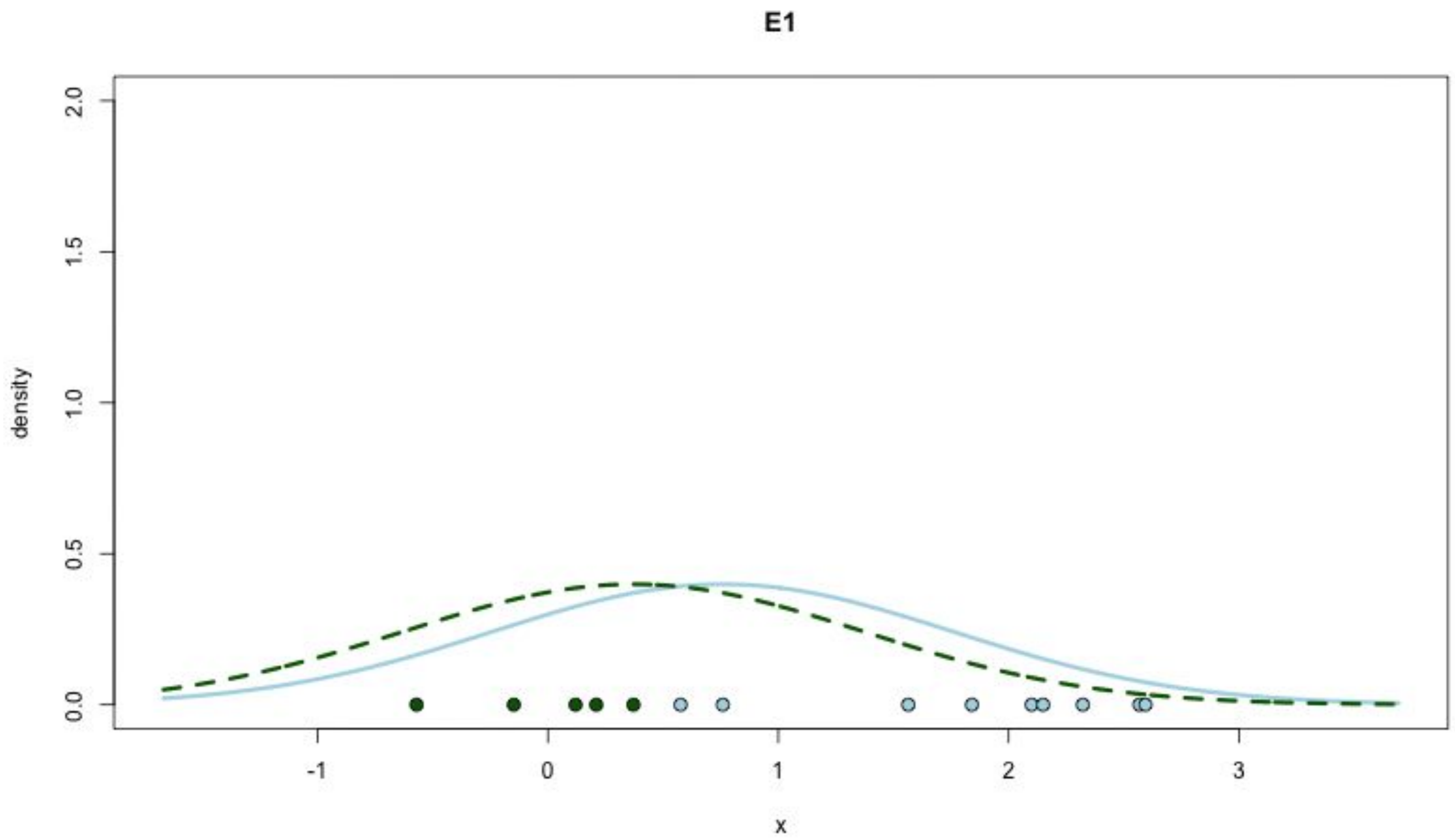
- GMM does **soft assignment**, every data point belongs to every cluster with some probability
- Data points that are more likely to be in a cluster have **more influence** over its parameters
- GMM uses the EM algorithm to iteratively update the cluster distributions. It first assigning a responsibility to each data point (**E-step**), and then using them to calculate weighted means and variances for each cluster (**M-step**)
- Responsibilities measure the **probability of a data point being in each cluster** (technically the **posterior probability**).
- Responsibilities contain information about **how common a cluster is** as well as the l**ikelihood of a data point belonging to that cluster**

Math Summary

- GMM does **soft assignment**, every data point belongs to every cluster with some probability
- Data points that are more likely to be in a cluster have **more influence** over its parameters
- GMM uses the EM algorithm to iteratively update the cluster distributions. It first assigning a responsibility to each data point (**E-step**), and then using them to calculate weighted means and variances for each cluster (**M-step**)
- Responsibilities measure the **probability of a data point being in each cluster** (technically the **posterior probability**).
- Responsibilities contain information about **how common a cluster is** as well as the l**ikelihood of a data point belonging to that cluster**

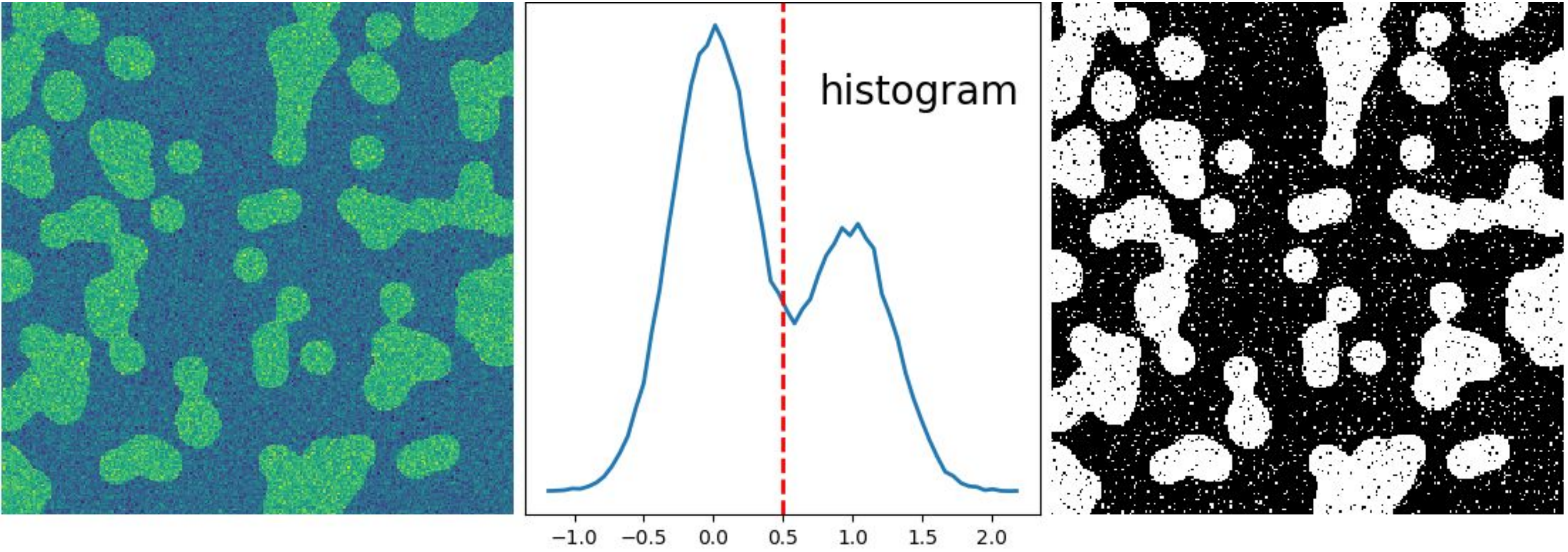
GMM

K means	**GMM**
Hard Assignment	Soft (probabilistic) Assignment
All Variances the Same	Variances can be different
Roughly same # of datapoints	Explicitly model # of data points



E1
density
x
0.0
0.5
1.0
1.5
2.0
-1
0
1
2
3

Applications



https://scipy-lectures.org/advanced/image_processing/auto_examples/plot_GMM.html

Applications

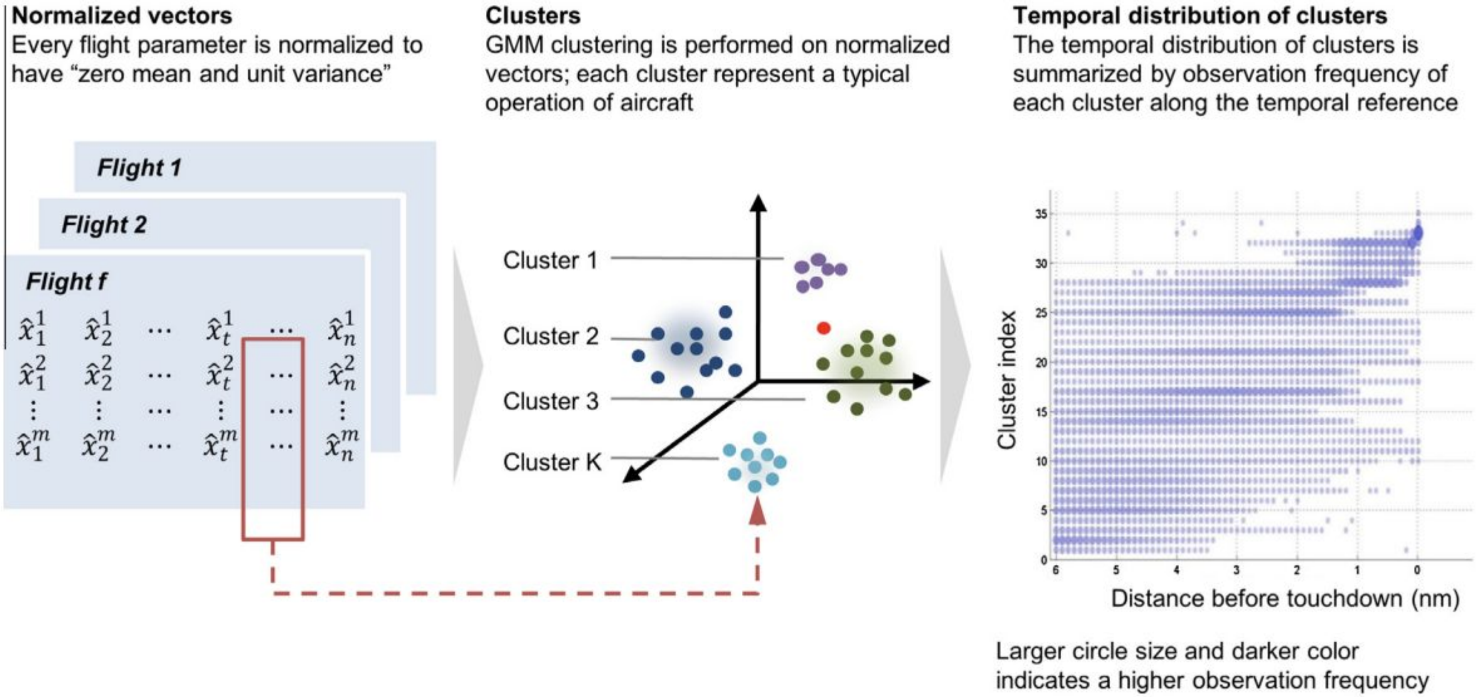


Fig. 3. Cluster analysis: identify typical operations and temporal distribution.

https://www.sciencedirect.com/science/article/abs/pii/S0968090X16000188

Applications

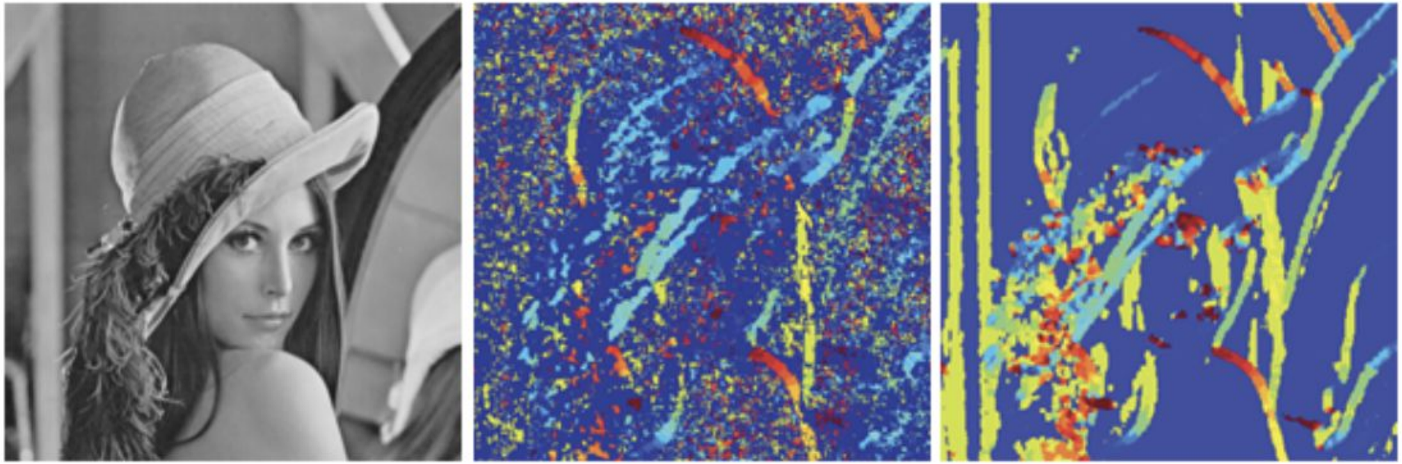


Fig. 1. Illustration of clustering of patches in the PLE method for the Lena image. LEFT: Original image; RIGHT: Clustered image; The pixels in the same color indicate that 8×8 patches around them are in the same cluster. It can be seen that patches from different parts of image are grouped into one cluster [17].

https://ieeexplore.ieee.org/document/7128671

Applications

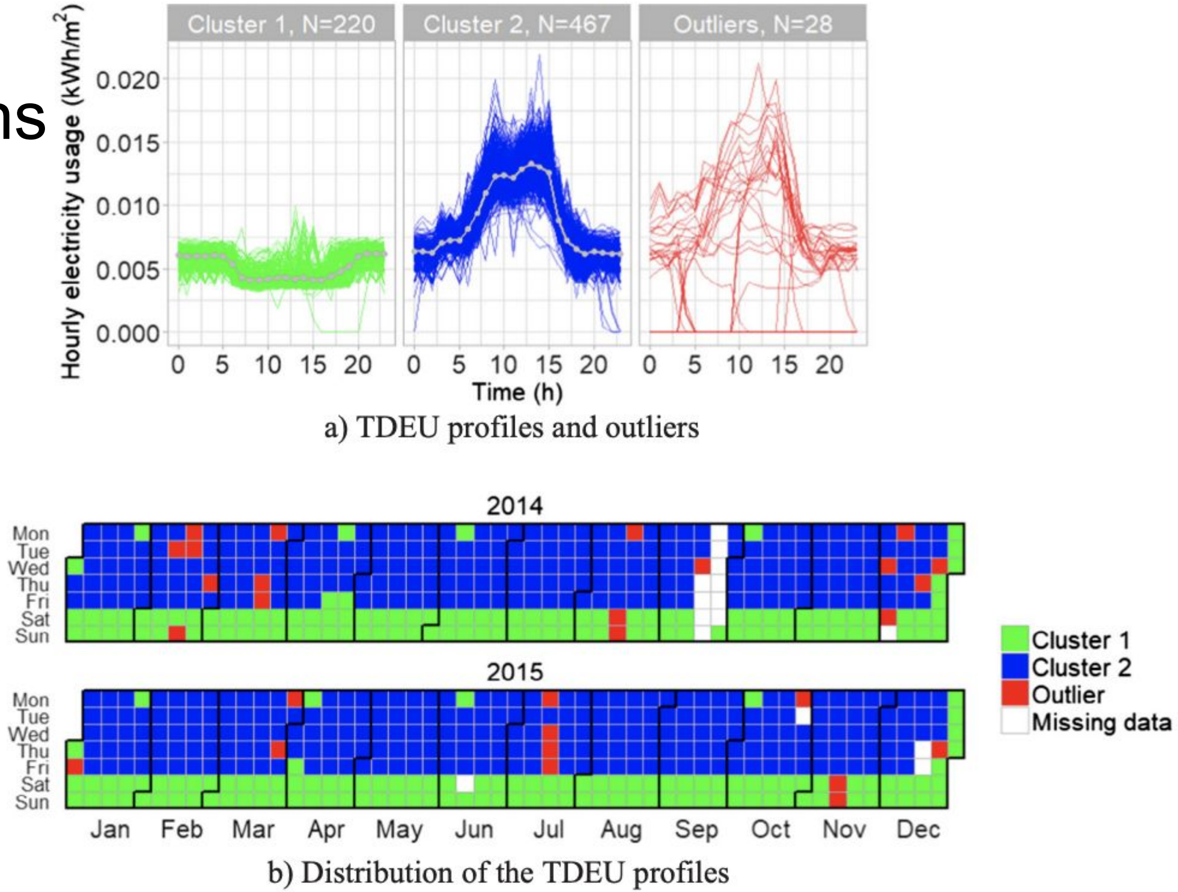


Fig. 7. Visualisation of the intra-building clustering result of Building #16.

https://www.sciencedirect.com/science/article/pii/S0306261918313606?casa_token=ISbTRYcOzcYAAAAA:p2OBUhbFBenV K6aXGaVbJ4p2WRsmpa0y6bB8Uw3LSmjx8kPkhLyKVCx8gtOJ0cXDUFf07fNdDdHr