Clustering of Non-Parametric hidden Markov models observations

Ibrahim KADDOURI

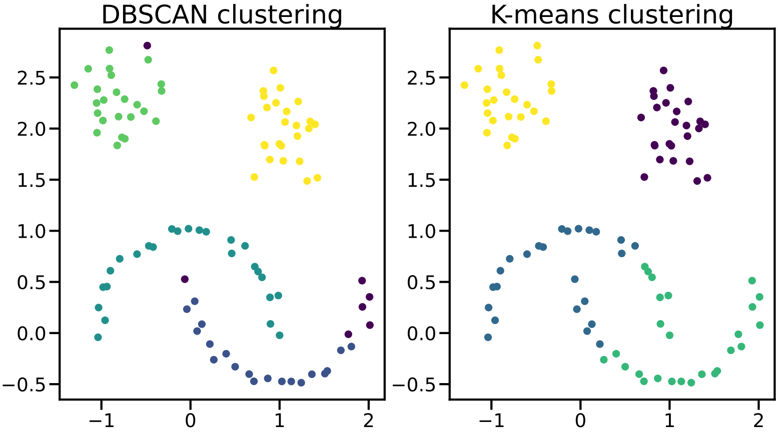
Joint work with Elisabeth Gassiat and Zacharie Naulet

Université Paris-Saclay, Laboratoire de mathématiques d'Orsay

May 27, 2024

Clustering

Clustering is an ill-posed problem which aims to find out interesting structures in the data or to derive a useful grouping of the observations.



Applications of clustering

- Recommender system in social network
- Statistical data analysis
- Anomaly detection
- Image segmentation and object detection
- ...

Model-based clustering: Mixture models

Observations $Y = (Y_k)_{1 \leq k \leq n}$ coming from J populations.
Define latent variables $X = (X_k)_{1 \leq k \leq n}$ such that: for each k,

$$Y_k \mid X_k = j \sim f_j$$

Model-based clustering: Mixture models

Observations $Y = (Y_k)_{1 \leq k \leq n}$ coming from J populations.
Define latent variables $X = (X_k)_{1 \leq k \leq n}$ such that: for each k,

$$Y_k \mid X_k = j \sim f_j$$

Then Y_k has distribution

$$\sum_{j=1}^{J} \pi_j f_j$$

π_j: Probability to come from population j

Useful to model data coming from heterogeneous populations.

Mixture models: Identifiability

Mixture models are not identifiable :

$$\sum_{j=1}^{J} \pi_j f_j = \frac{\pi_1}{2} f_1 + \left(\frac{\pi_1}{2} + \pi_2\right) \left(\frac{\frac{\pi_1}{2} f_1 + \pi_2 f_2}{\frac{\pi_1}{2} + \pi_2}\right) + \sum_{j=3}^{J} \pi_j f_j$$

Mixture models: Identifiability

Mixture models are not identifiable :

$$\sum_{j=1}^{J} \pi_j f_j = \frac{\pi_1}{2} f_1 + \left(\frac{\pi_1}{2} + \pi_2\right) \left(\frac{\frac{\pi_1}{2} f_1 + \pi_2 f_2}{\frac{\pi_1}{2} + \pi_2}\right) + \sum_{j=3}^{J} \pi_j f_j$$

Learning of population components possible only under additional structural assumptions such as:

- Parametric mixtures
- Shape restrictions (gaussian, multinomial, ...)

$\rightarrow$ Might lead to poor results in practice

Hidden Markov Models and why they are useful

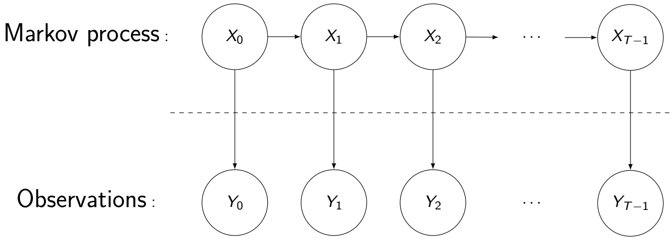


Figure: A Hidden Markov Model.

Latent (unobserved) variables $(X_k)_k$ form a Markov chain.
Observations $(Y_k)_k$ are independent conditionnally to $(X_k)_k$.

Hidden Markov Models and why they are useful

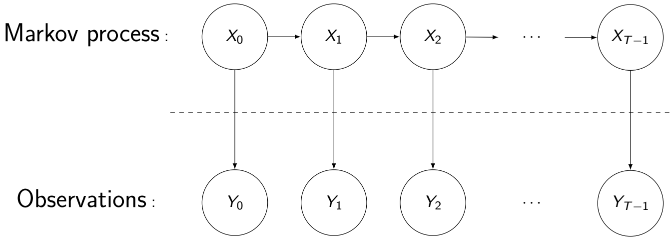


Figure: A Hidden Markov Model.

Latent (unobserved) variables $(X_k)_k$ form a Markov chain.
Observations $(Y_k)_k$ are independent conditionnally to $(X_k)_k$.

HMMs are identifiable without any shape restriction!

Outline

Inference in Hidden Markov Models

The HMM parameters are:

- The initial distribution ν.
- The transition matrix Q.
- The emission distributions $F = (f_i)_{1 \leq i \leq J}$

Purpose: Estimate the model parameters and the hidden states associated to the observations.

Inference in Hidden Markov Models

Many estimators have been studied in the HMM framework:

- **Kernel estimators**
- **Wavelet estimators**
- **Projection estimators**

The associated optimal rates of convergence were derived.
Fundamental limits for learning these models were also identified.

Outline

Online vs offline clustering

We study the risk of clustering observations in two frameworks:

- **Offline**: All observations are used in the clustering procedures. Clustering rules are of the form: $h(Y_{1:n}) = (h_i(Y_{1:n}))_{1 \leq i \leq n}$

Online vs offline clustering

We study the risk of clustering observations in two frameworks:

- **Offline**: All observations are used in the clustering procedures. Clustering rules are of the form: $h(Y_{1:n}) = (h_i(Y_{1:n}))_{1 \leq i \leq n}$
- **Online**: Clustering can use only past (and current) observations. Clustering rules are of the form: $h(Y_{1:n}) = (h_i(Y_{1:i}))_{1 \leq i \leq n}$

Online vs offline clustering

We study the risk of clustering observations in two frameworks:

- **Offline**: All observations are used in the clustering procedures. Clustering rules are of the form: $h(Y_{1:n}) = (h_i(Y_{1:n}))_{1 \leq i \leq n}$
- **Online**: Clustering can use only past (and current) observations. Clustering rules are of the form: $h(Y_{1:n}) = (h_i(Y_{1:i}))_{1 \leq i \leq n}$

For the moment, we focus on the offline case.

Risk of clustering

Consider the loss function:

$$L_1(x'_{1:n}, x_{1:n}) = \inf_{\tau \in \mathcal{S}} \frac{1}{n} \sum_{k=1}^{n} 1_{x'_k \neq \tau(x_k)}$$

Risk of clustering

Consider the loss function:

$$L_1(x'_{1:n}, x_{1:n}) = \inf_{\tau \in \mathcal{S}} \frac{1}{n} \sum_{k=1}^{n} 1_{x'_k \neq \tau(x_k)}$$

The risk associated to a classifier h is:

$$\mathcal{R}_{n,HMM}^{cluster}(h) = \mathbb{E}_\theta[L_1(h(Y_{1:n}), X_{1:n})] = \mathbb{E}_\theta \left[\inf_{\tau \in \mathcal{S}} \frac{1}{n} \sum_{i=1}^{n} 1_{[h(Y_{1:n})]_i \neq \tau(X_i)} \right]$$

Risk of clustering

Consider the loss function:

$$L_1(x'_{1:n}, x_{1:n}) = \inf_{\tau \in \mathcal{S}} \frac{1}{n} \sum_{k=1}^{n} 1_{x'_k \neq \tau(x_k)}$$

The risk associated to a classifier h is:

$$\mathcal{R}_{n,HMM}^{cluster}(h) = \mathbb{E}_\theta[L_1(h(Y_{1:n}), X_{1:n})] = \mathbb{E}_\theta \left[\inf_{\tau \in \mathcal{S}} \frac{1}{n} \sum_{i=1}^{n} 1_{[h(Y_{1:n})]_i \neq \tau(X_i)} \right]$$

The purpose is to exhibit bounds on the quantity:

$$\mathcal{R}_{n,HMM}^{\star,cluster} = \inf_h \mathcal{R}_{n,HMM}^{cluster}(h)$$

Upper bound

A straightforward upper-bound on the risk of clustering is:

$$\mathcal{R}^{\star,cluster}_{n,HMM} \leq \inf_h \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} 1_{[h(Y_{1:n})]_i \neq X_i}\right] = \mathcal{R}^{\star,classif}_{n,HMM}$$

where $\mathcal{R}^{\star,classif}_{n,HMM} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\theta^\star}\left[\min_{x\in\mathbb{X}} \mathbb{P}\left(X_i \neq x \mid Y_{1:n}\right)\right]$ corresponds to the Bayes risk of classification of HMM observations.

Lower bound

Theorem

Assume $\delta = \min_{i,j} Q_{i,j} > 0$. *Then, the risk of clustering and classification ensure the following inequalities:*

- *For iid observations:*

$$\mathcal{R}^{\star,\mathit{Offline}}_{\theta^\star,n}(L_1) - \sqrt{\frac{\log(J!)}{2n}} \leq \mathcal{R}^{\star,\mathit{Offline}}_{\theta^\star,n}(L_2) \leq \mathcal{R}^{\star,\mathit{Offline}}_{\theta^\star,n}(L_1)$$

- *For HMM observations:*

$$\mathcal{R}^{\star,\mathit{Offline}}_{\theta^\star,n}(L_1) - \frac{1}{1-\rho_0}\sqrt{\frac{\log(J!)}{2n}} \leq \mathcal{R}^{\star,\mathit{Offline}}_{\theta^\star,n}(L_2) \leq \mathcal{R}^{\star,\mathit{Offline}}_{\theta^\star,n}(L_1)$$

where J *is the number of classes,* $\rho_0 = \frac{1-J\delta}{1-(J-1)\delta}$.
Exactly similar inequalities hold for the risk of online clustering.

Outline

Stationarized Bayes risk

Consider the following quantities:

$$\mathcal{R}^{\star,\mathit{Offline}}_{\theta^\star,n}(L_1) = \inf_h \mathcal{R}^{\mathit{Offline}}_{\theta^\star,n}(L_1,h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\theta^\star}\left[\min_{x\in\mathbb{X}} \mathbb{P}\left(X_i \neq x \mid Y_{1:n}\right)\right]$$

$$\mathcal{R}^{\star,\mathit{Offline}}_{\theta^\star,\mathit{stat}}(L_1) = \mathbb{E}_{\theta^\star}\left[\min_{x\in\mathbb{X}} \mathbb{P}_{\theta^\star}\left(X_0 \neq x \mid Y_{-\infty:+\infty}\right)\right]$$

$$\mathcal{R}^{\star,\mathit{Online}}_{\theta^\star,n}(L_1) = \inf_h \mathcal{R}^{\mathit{Online}}_{\theta^\star,n}(L_1,h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\theta^\star}\left[\min_{x\in\mathbb{X}} \mathbb{P}_{\theta^\star}\left(X_i \neq x \mid Y_{1:i}\right)\right]$$

$$\mathcal{R}^{\star,\mathit{Online}}_{\theta^\star,\mathit{stat}}(L_1) = \mathbb{E}_{\theta^\star}\left[\min_{x\in\mathbb{X}} \mathbb{P}_{\theta^\star}\left(X_0 \neq x \mid Y_{-\infty:0}\right)\right]$$

Exponential forgetting

Proposition

Assume:

- *The initial distribution is the stationary distribution*
- *The HMM model is comprised of two hidden states*
- $\delta = \min_{i,j} Q_{i,j} > 0$
- $\rho_0 = \frac{1-2\delta}{1-\delta}$ *and* $\rho_1 = 1 - 2\delta$

Exponential forgetting

Proposition

Assume:

- *The initial distribution is the stationary distribution*
- *The HMM model is comprised of two hidden states*
- $\delta = \min_{i,j} Q_{i,j} > 0$
- $\rho_0 = \frac{1-2\delta}{1-\delta}$ *and* $\rho_1 = 1 - 2\delta$

Then, for $0 \leq j' \leq j$*,* $k \geq 0$ *and* $n \geq 0$*:*

$$\|\mathbb{P}_\theta(X_k \in . \mid Y_{-j:n}) - \mathbb{P}_\theta(X_k \in . \mid Y_{-j':n})\|_{TV} \leq 2\rho_0^{k \wedge n + j'} \rho_1^{k - k \wedge n}$$

Exponential forgetting

Proposition

Assume:

- *The initial distribution is the stationary distribution*
- *The HMM model is comprised of two hidden states*
- $\delta = \min_{i,j} Q_{i,j} > 0$
- $\rho_0 = \frac{1-2\delta}{1-\delta}$ *and* $\rho_1 = 1 - 2\delta$

Then, for $0 \leq j' \leq j$, $k \geq 0$ *and* $n \geq 0$:

$$\|\mathbb{P}_\theta(X_k \in . \mid Y_{-j:n}) - \mathbb{P}_\theta(X_k \in . \mid Y_{-j':n})\|_{TV} \leq 2\rho_0^{k\wedge n+j'}\rho_1^{k-k\wedge n}$$

Similarly, for $0 \leq k \leq j' \leq j$ *and* $n \geq 0$ *one has:*

$$\|\mathbb{P}_\theta(X_k \in . \mid Y_{-n:j}) - \mathbb{P}_\theta(X_k \in . \mid Y_{-n:j'})\|_{TV} \leq 2\rho_0^{-k+j'}$$

Stationarized Bayes risk

Theorem

Under the same assumptions:

$$\left|\mathcal{R}^{\star,Offline}_{n,HMM} - \mathcal{R}^{\star,Offline}_{stat,HMM}\right| \leq \frac{2}{n(1-\rho_0)}$$

$$\left|\mathcal{R}^{\star,Online}_{n,HMM} - \mathcal{R}^{\star,Online}_{stat,HMM}\right| \leq \frac{\rho_1}{2n}\frac{1}{1-\rho_0}$$

where $\rho_0 = \frac{1-2\delta}{1-\delta}$, $\rho_1 = 1-2\delta$ *and* $\delta = \min_{i,j} Q_{i,j} > 0$.

Bounds on asymptotic Bayes risk

Theorem

Assume the initial distribution is the stationary distribution. where $\delta = \min_{i,j} Q_{i,j} > 0$. *One has:*

$$\frac{\delta}{1-\delta}\mathcal{R}^{\star,Online}_{\theta^\star,\infty} \leq \mathcal{R}^{\star,Offline}_{\theta^\star,\infty} \leq \mathcal{R}^{\star,Online}_{\theta^\star,\infty}$$

Bounds on asymptotic Bayes risk

Theorem

Assume the initial distribution is the stationary distribution. where $\delta = \min_{i,j} Q_{i,j} > 0$. *One has:*

$$\frac{\delta}{1-\delta}\mathcal{R}^{\star,Online}_{\theta^\star,\infty} \leq \mathcal{R}^{\star,Offline}_{\theta^\star,\infty} \leq \mathcal{R}^{\star,Online}_{\theta^\star,\infty}$$

$$\delta \int_{\mathbb{R}} [f_0 \wedge f_1](z)\mu(dz) \leq \mathcal{R}^{\star,Online}_{\theta^\star,\infty} \leq (1-\delta)\int_{\mathbb{R}} [f_0 \wedge f_1](z)\mu(dz)$$

Appropriate Signal-to-Noise ratio

Corollary

Assume the initial distribution of the hidden states is the stationary distribution of Q and in the case of multidimensional gaussian emission distributions having the same covariance matrix Σ and means μ_1 and μ_2. Let $SNR = (\mu_0 - \mu_1)^\intercal \Sigma^{-1} (\mu_0 - \mu_1)$:

$$\frac{\delta}{2} \exp\left(-\frac{SNR}{2}\right) \leq \mathcal{R}_{\theta^\star,\infty}^{\star,Online} \leq (1-\delta) \exp\left(-\frac{SNR}{8}\right)$$

Outline

Notations

$$(f_0, f_1) = \text{Emission densities}$$

$$\theta = (\nu, Q, (f_x)_{x=0,1}) \text{ true parameters}$$

$$\hat{\theta} = (\hat{\nu}, \hat{Q}, (\hat{f}_x)_{x=0,1}) \text{ estimators of the true parameters}$$

$$\mathbb{P}_\theta(X_i \in . \mid Y_{1:n}) \quad \text{smoothing distribution under true parameters } \theta$$

$$\mathbb{P}_{\hat{\theta}}(X_i \in . \mid Y_{1:n}) \quad \text{smoothing distribution under estimated parameters } \hat{\theta}$$

$$h_\theta^{Offline}(Y_{1:n}) = (1_{\mathbb{P}_\theta(X_i=1|Y_{1:n})>1/2})_{1\le i\le n} \quad \text{Bayes classifier}$$

$$h_{\hat{\theta}}^{Offline}(Y_{1:n}) = (1_{\mathbb{P}_{\hat{\theta}}(X_i=1|Y_{1:n})>1/2})_{1\le i\le n} \quad \text{plug-in Bayes classifier}$$

$$h_\theta^{Online}(Y_{1:n}) = (1_{\mathbb{P}_\theta(X_i=1|Y_{1:i})>1/2})_{1\le i\le n} \quad \text{Bayes classifier}$$

$$h_{\hat{\theta}}^{Online}(Y_{1:n}) = (1_{\mathbb{P}_{\hat{\theta}}(X_i=1|Y_{1:i})>1/2})_{1\le i\le n} \quad \text{plug-in Bayes classifier}$$

Reconstruction algorithm

In practice θ is unknown. One rather uses an estimator $\hat{\theta}$ and the algorithm yields:

$$\hat{h}(Y_{1:n}) = \left(\arg\max_{x_k \in \mathbb{X}} \mathbb{P}_{\hat{\theta}}(X_k = x_k \mid Y_{1:n}) \right)_{1 \leq k \leq n}$$

Reconstruction algorithm

In practice θ is unknown. One rather uses an estimator $\hat{\theta}$ and the algorithm yields:

$$\hat{h}(Y_{1:n}) = \left(\arg\max_{x_k \in \mathbb{X}} \mathbb{P}_{\hat{\theta}}(X_k = x_k \mid Y_{1:n}) \right)_{1 \leq k \leq n}$$

Algorithm 2: MAP classifier algorithm

Assume $\mathbb{X} = \{0, ..., r-1\}$, $\theta = (\nu, Q, F)$ is given.;
Using the **Forward-Backward** algorithm, compute
$\mathbb{P}_\theta(X_1 = . \mid Y_{1:n}), .., \mathbb{P}_\theta(X_n = . \mid Y_{1:n})$.;
for $k \in \{1, .., n\}$ **do**
$\quad x_k = \arg\max_{0 \leq x \leq r-1} \mathbb{P}_\theta(X_k = x \mid Y_{1:n})$

Efficiency of plug-in empirical Bayes classifier

Theorem

Assume in addition that the emission densities f_0 *and* f_1 *are lower-bounded by* $c^\star > 0$*. Let* $\delta = \min_{i,j} Q_{i,j} > 0$ *and* $\rho = \frac{1-2\delta}{1-\delta}$*. Then:*

$$\mathcal{R}_{\theta^\star,n}^{Online}(h_{\hat{\theta}}^{Online}) - \mathcal{R}_{\theta^\star,n}^{\star,Online} \leq \frac{4(1-\delta)^2}{\delta^3} \inf_{\tau \in \mathcal{S}} \mathbb{E}_{\theta^\star}\left[\frac{1}{n}\|\nu^\tau - \hat{\nu}\|_2 + \|Q^\tau - \hat{Q}\|_F + \frac{1}{c^\star}\max_{x=0,1}||f_{\tau(x)} - \hat{f}_x||_\infty\right]$$

$$\mathcal{R}_{\theta^\star,n}^{Offline}(h_{\hat{\theta}}^{Offline}) - \mathcal{R}_{\theta^\star,n}^{\star,Offline} \leq \frac{4(1-\delta)}{\delta^2} \inf_{\tau \in \mathcal{S}} \mathbb{E}_{\theta^\star}\left[\frac{1}{n(1-\rho)}\|\nu^\tau - \hat{\nu}\|_2 + \Big(1/(1-\rho) + 1/(1-\hat{\rho})\Big)\Big(\|Q^\tau - \hat{Q}\|_F + \frac{2}{c^\star}\max_{x=0,1}||f_{\tau(x)} - \hat{f}_x||_\infty\Big)\right]$$

Rate of convergence

Corollary

Assume $f_0 \neq f_1$ and that they belong to $C^s(\mathbb{R})$, the usual space of s-Hölder-continuous functions.
Assume Q is full-rank, irreducible and aperiodic.
Let $M_n \to +\infty$ arbitrarily slowly and let $k_n = \left(\frac{\log(n)}{n}\right)^{\frac{s}{2s+1}}$.
There exists an estimator $\hat{\theta} = \left(\hat{\pi}, \hat{Q}, (\hat{f}_i)_{i=0,1}\right)$ of θ and a sequence of random permutations $(\tau_n)_n$ of $\{0,1\}$ and $c, c' \geq 0$ such that:

$$\mathcal{R}^{Online}_{\theta^\star,n}(h^{Online}_{\hat{\theta}^{\tau_n}}) - \mathcal{R}^{\star,Online}_{\theta^\star,n} \leq c M_n^3 k_n$$
$$\mathcal{R}^{Offline}_{\theta^\star,n}(h^{Offline}_{\hat{\theta}^{\tau_n}}) - \mathcal{R}^{\star,Offline}_{\theta^\star,n} \leq c' M_n^3 k_n$$

Two examples

Data are generated through the same transition matrix $Q = \begin{pmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{pmatrix}$.

- **First example:** A sample of size $n = 5.10^4$ is generated from two gaussian mixtures :$\frac{1}{2}\left(\mathcal{N}(1.7, 0.2) + \mathcal{N}(7, 0.15)\right)$ and $\frac{1}{2}\left(\mathcal{N}(3.5, 0.2) + \mathcal{N}(5, 0.4)\right)$.
- **Second example:** A sample of size $n = 10^5$ is generated from two gaussian mixtures :$\frac{1}{2}\left(\mathcal{N}(3, 0.6) + \mathcal{N}(7, 0.4)\right)$ and $\frac{1}{2}\left(\mathcal{N}(5, 0.3) + \mathcal{N}(9, 0.4)\right)$.

Purpose: Retrieve the sequence of hidden states using only the observations.

Example 1

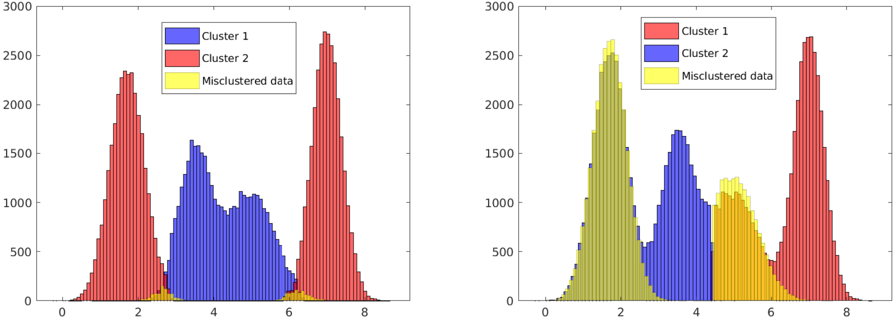


Figure: Histograms of the clusters. Left: clustering using plug-in classifier. Right: K-means clustering

Example 2

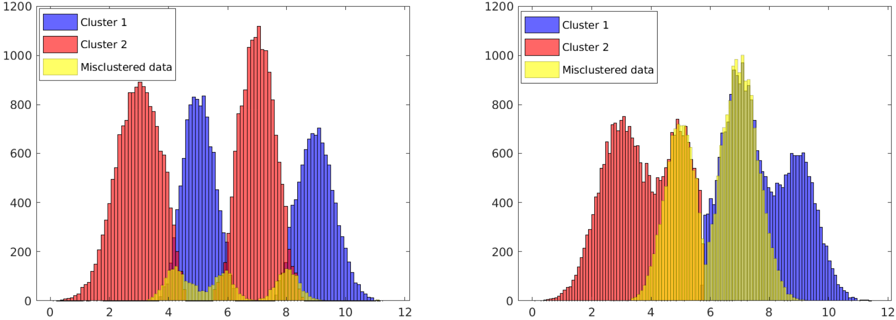


Figure: Histograms of the clusters. Left: clustering using plug-in classifier. Right: K-means clustering

Clustering errors

	Bayes classifier	Plug-in classifier	K-means algorithm
Example 1	1.56%	1.61%	46.7%
Example 2	6.42%	6.51%	47.3%

Table: Errors of clustering using 3 algorithms: the Bayes classifier (using the true model parameters), the plug-in classifier (using the estimated parameters) and the K-means algorithm.