

Apprentissage automatique

Mélange de gaussiennes - motivation

APPRENTISSAGE AUTOMATIQUE

Sujets: types d'apprentissage

RAPPEL

- Il existe différents types d'apprentissage
 - apprentissage supervisé : il y a une cible à prédire

$$\mathcal{D} = \{(\mathbf{x}_1, t_1), \dots, (\mathbf{x}_N, t_N)\}$$

- apprentissage non-supervisé : cible n'est pas fournie

$$\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$$

- apprentissage par renforcement (non couvert dans ce cours)

APPRENTISSAGE AUTOMATIQUE

Sujets: types d'apprentissage

RAPPEL

- Il existe différents types d'apprentissage
 - apprentissage supervisé : il y a une cible à prédire

$$\mathcal{D} = \{(\mathbf{x}_1, t_1), \dots, (\mathbf{x}_N, t_N)\}$$

- apprentissage non-supervisé : cible n'est pas fournie

$$\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$$

- apprentissage par renforcement (non couvert dans ce cours)

TYPES D'APPRENTISSAGE

Sujets: apprentissage non-supervisé, estimation de densité

RAPPEL

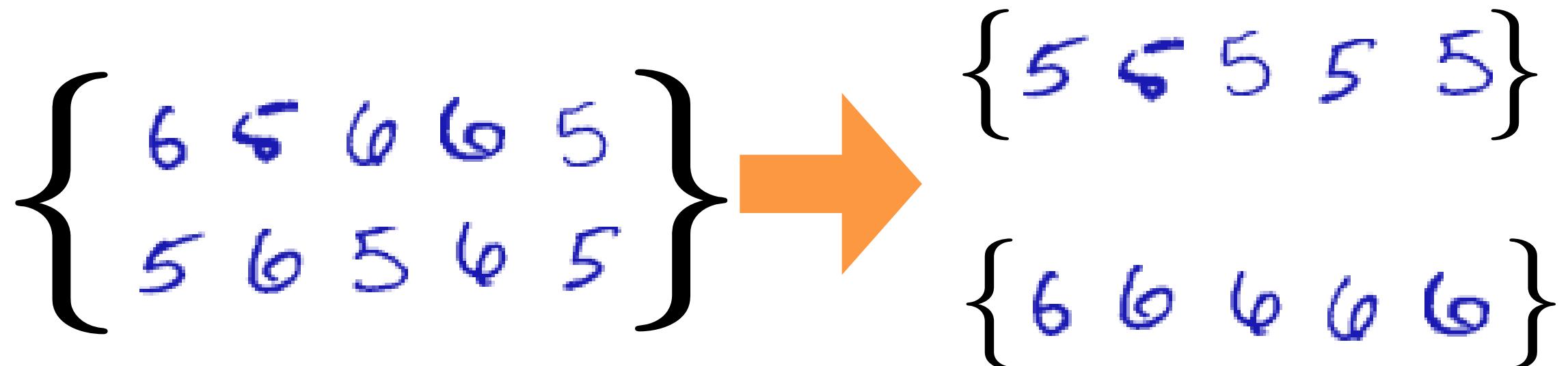
- L'apprentissage non-supervisé est lorsqu'une cible n'est pas explicitement donnée
 - estimation de densité : apprendre la loi de probabilité $p(\mathbf{x})$ ayant généré les données
 - pour générer de nouvelles données réalistes
 - pour distinguer les «vrais» données des «fausses» données (*spam filtering*)
 - compression de données

TYPES D'APPRENTISSAGE

Sujets: apprentissage non-supervisé, partitionnement

RAPPEL

- L'apprentissage non-supervisé est lorsqu'une cible n'est pas explicitement donnée
 - partitionnement de données / *clustering*



Apprentissage automatique

Mélange de gaussiennes - modèle

MÉLANGE DE GAUSSIENNES

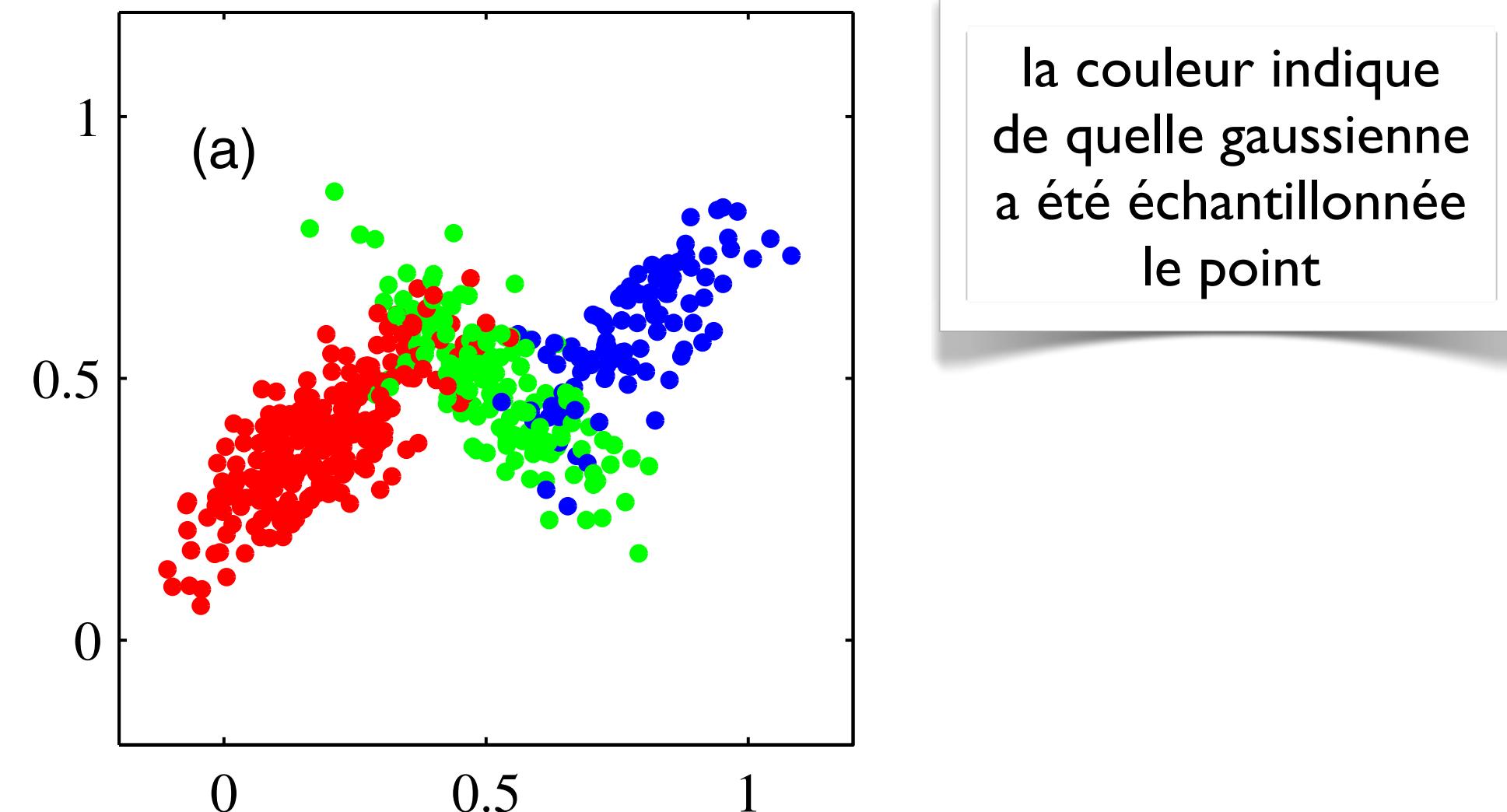
Sujets: mélange de gaussiennes

- Un mélange de gaussiennes suppose que les données ont été générées comme suit :
 - pour $n = 1 \dots N$
 - choisit un entier $k \in \{1, \dots, K\}$ selon probabilités π_1, \dots, π_K
 - génère \mathbf{x}_n d'une loi de la loi de probabilité $\mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k)$
 - En mots : les entrées sont des échantillons d'une de K différentes lois gaussiennes, ayant chacune des moyennes et covariances différentes

MÉLANGE DE GAUSSIENNES

Sujets: mélange de gaussiennes

- Exemple de données générées d'un mélange de gaussiennes ($K=3$)



MÉLANGE DE GAUSSIENNES

Sujets: probabilité a priori du choix de la gaussienne

- On va noter z la variable aléatoire correspondant à l'identité de la gaussienne qui a généré une entrée x
 - format *one-hot* : $z_k=1$ si x a été générée par la k^e gaußsienne
- La probabilité de choisir la k^e gaußsienne est donc

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

MÉLANGE DE GAUSSIENNES

Sujets: probabilité a priori du choix de la gaussienne

- On va noter z la variable aléatoire correspondant à l'identité de la gaussienne qui a généré une entrée x
 - format *one-hot* : $z_k=1$ si x a été générée par la k^e gaussienne
- Puisque z est *one-hot*, on peut aussi écrire

$$p(\mathbf{z}) = \prod_{k=1}^K \pi_k^{z_k}$$

MÉLANGE DE GAUSSIENNES

Sujets: fonction de densité conditionnelle de l'entrée

- Sachant z la probabilité (fonction de densité) de x est

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

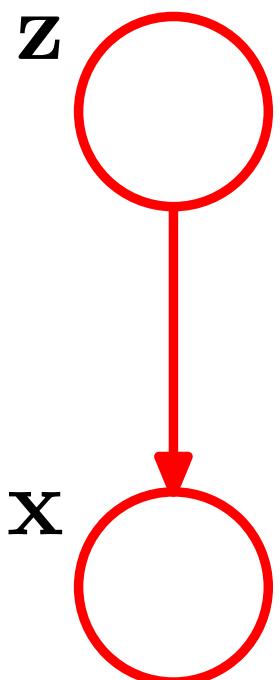
qu'on peut aussi écrire

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

MÉLANGE DE GAUSSIENNES

Sujets: réseau bayésien

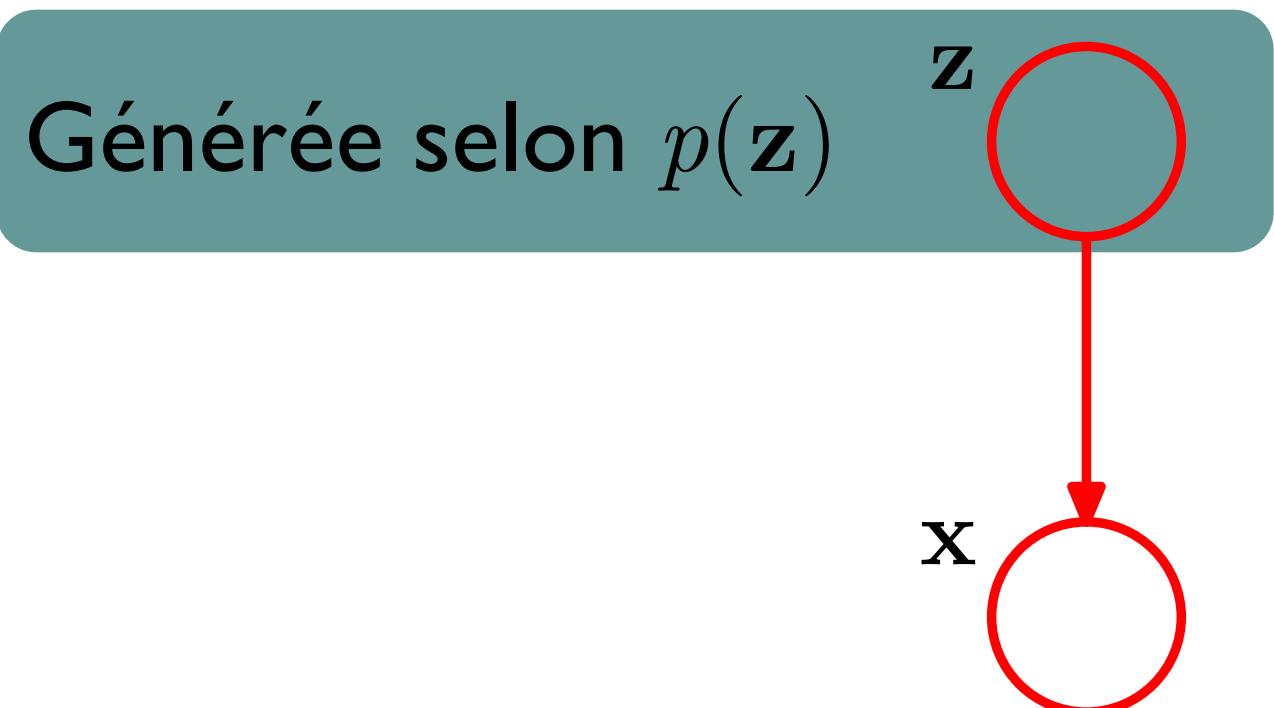
- On peut illustrer un mélange de gaussiennes sous la forme du réseau bayésien (modèle graphique) suivant :



MÉLANGE DE GAUSSIENNES

Sujets: réseau bayésien

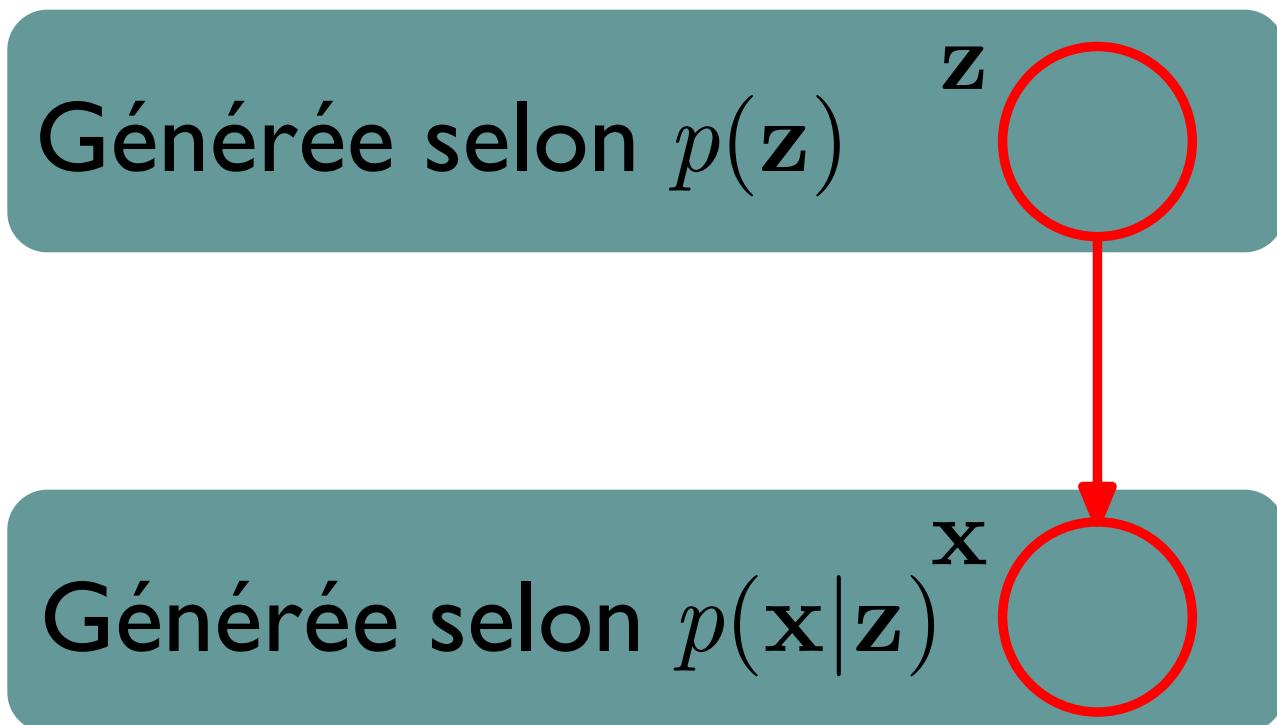
- On peut illustrer un mélange de gaussiennes sous la forme du réseau bayésien (modèle graphique) suivant :



MÉLANGE DE GAUSSIENNES

Sujets: réseau bayésien

- On peut illustrer un mélange de gaussiennes sous la forme du réseau bayésien (modèle graphique) suivant :



APPROCHE PROBABILISTE GÉNÉRATIVE

Sujets: approche probabiliste générative

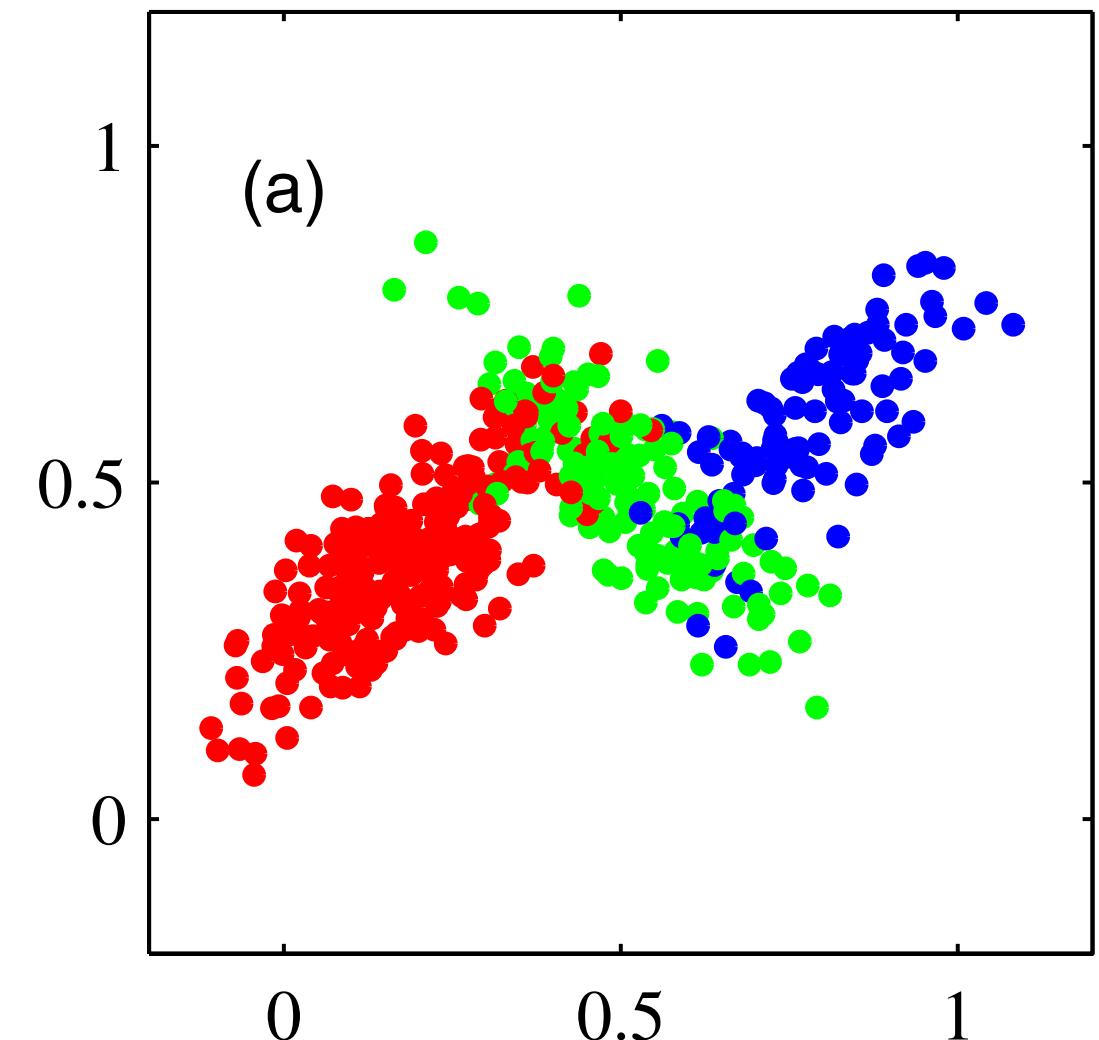
RAPPEL

- On va supposer que les données ont été générées selon le processus suivant (cas binaire) :
 - pour $n = 1 \dots N$
 - assigne $t_n=1$ avec probabilité $p(\mathcal{C}_1) = \pi$ et $t_n=0$ avec probabilité $p(\mathcal{C}_2) = 1 - \pi$
 - si $t_n=1$, génère \mathbf{x}_n de la loi de probabilité $p(\mathbf{x}_n|\mathcal{C}_1) = \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$
 - sinon ($t_n=0$), génère \mathbf{x}_n de la loi de probabilité $p(\mathbf{x}_n|\mathcal{C}_2) = \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$
 - En mots : les entrées sont des échantillons d'une loi gaussienne, mais de moyennes différentes pour les différentes classes

MÉLANGE DE GAUSSIENNES

Sujets: mélange de gaussiennes

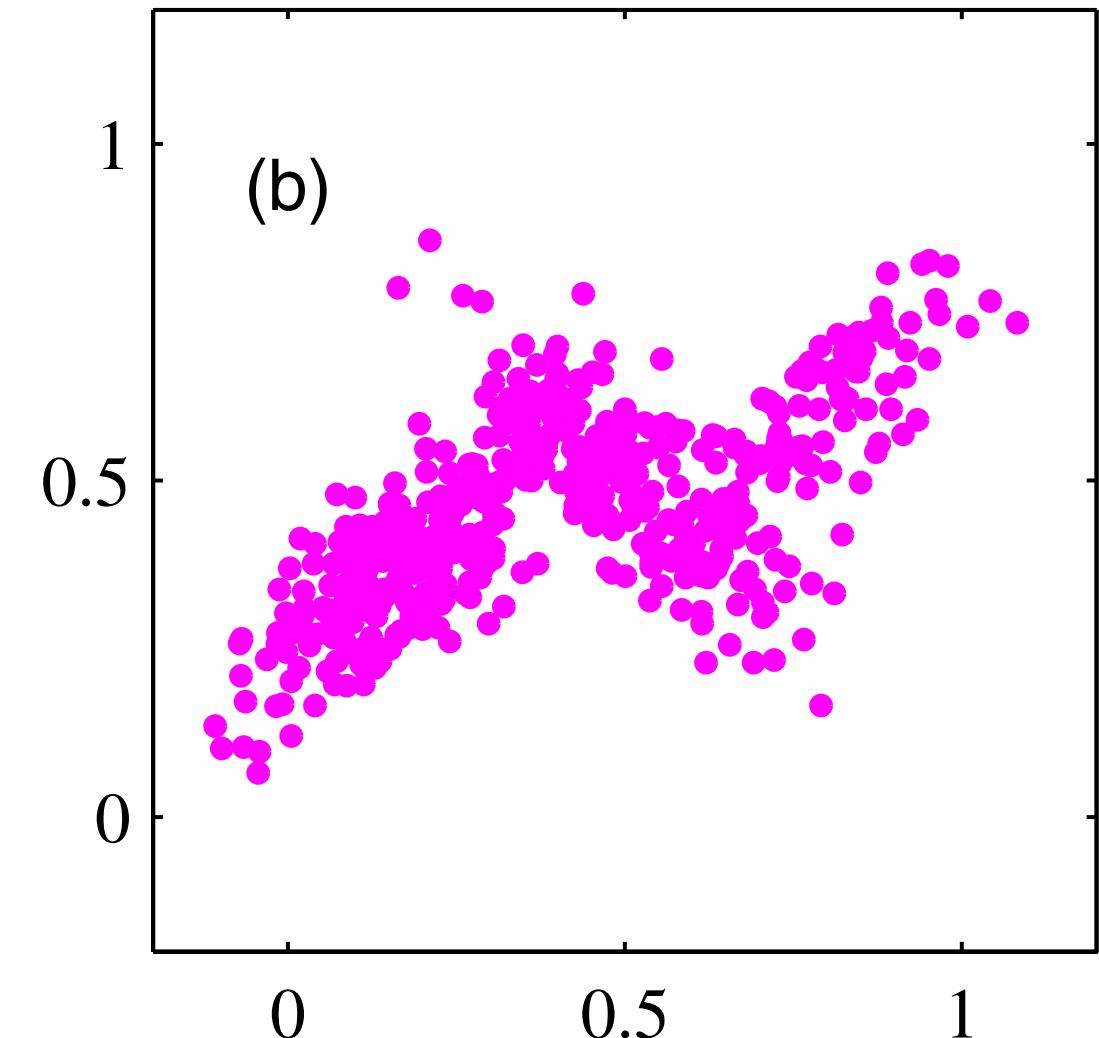
- Dans un mélange de gaussienne, l'appartenance aux K gaussiennes («classes») n'est pas connue



MÉLANGE DE GAUSSIENNES

Sujets: mélange de gaussiennes

- Dans un mélange de gaussienne, l'appartenance aux K gaussiennes («classes») n'est pas connue



MÉLANGE DE GAUSSIENNES

Sujets: fonction de densité marginale des entrées

- Puisqu'on ne connaît pas l'appartenance aux gaussiennes (\mathbf{z}), on va s'intéresser à la probabilité marginale :

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

- C'est de cette façon qu'on va mesurer la performance de notre modèle

Apprentissage automatique

Mélange de gaussiennes - partitionnement de données

MÉLANGE DE GAUSSIENNES

Sujets: mélange de gaussiennes

RAPPEL

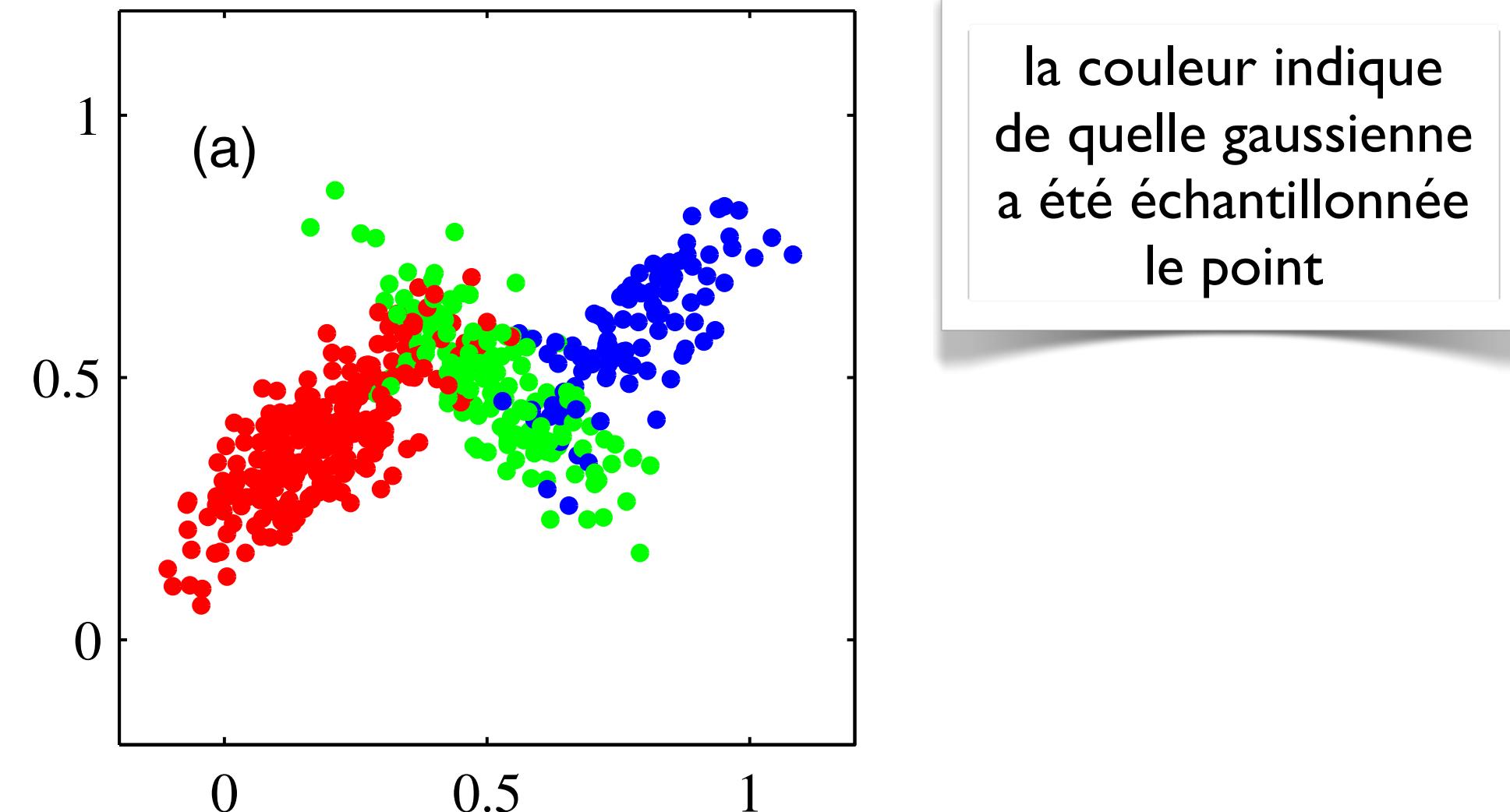
- Un mélange de gaussiennes suppose que les données ont été générées comme suit :
 - pour $n = 1 \dots N$
 - choisit un entier $k \in \{1, \dots, K\}$ selon probabilités π_1, \dots, π_K
 - génère \mathbf{x}_n d'une loi de la loi de probabilité $\mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k)$
 - En mots : les entrées sont des échantillons d'une de K différentes lois gaussiennes, ayant chacune des moyennes et covariances différentes

MÉLANGE DE GAUSSIENNES

Sujets: mélange de gaussiennes

RAPPEL

- Exemple de données générées d'un mélange de gaussiennes ($K=3$)

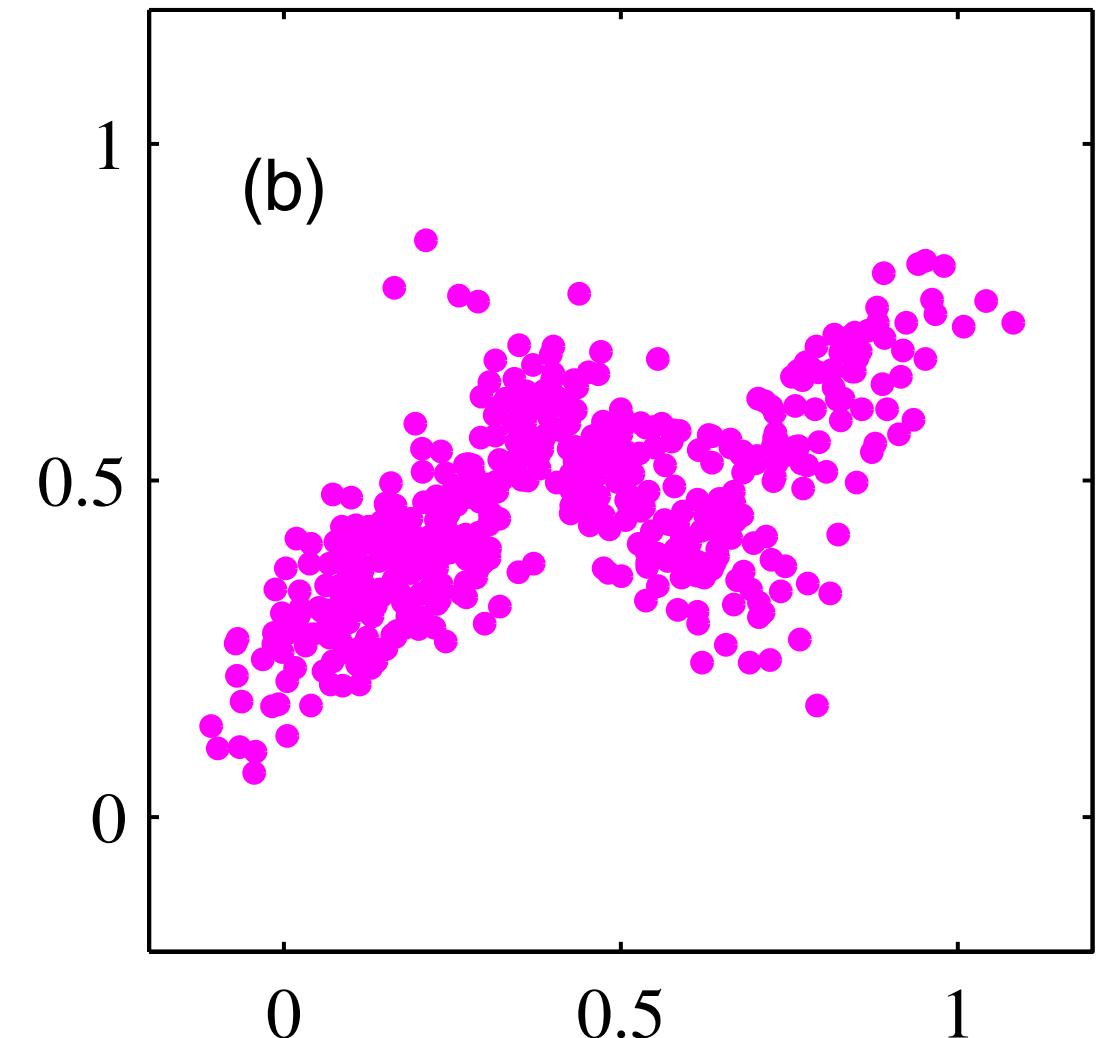


MÉLANGE DE GAUSSIENNES

Sujets: mélange de gaussiennes

RAPPEL

- Dans un mélange de gaussienne, l'appartenance aux K gaussiennes («classes») n'est pas connue



PARTITIONNEMENT DE DONNÉES

Sujets: partitionnement de données, *clustering*

- À partir d'un mélange de gaussiennes entraîné, on pourrait inférer à quelle gaussienne appartiennent les entrées
 - on pourrait alors automatiquement catégoriser nos données en fonction des probabilités d'appartenance à chacune des gaussiennes
- Cette application s'appelle le **partitionnement de données (*clustering*)**
 - permet de «mettre de l'ordre» dans les données
 - permet de visualiser les données une partition à la fois

PARTITIONNEMENT DE DONNÉES

Sujets: probabilité d'appartenance, *responsability*

- À l'aide de la règle de Bayes, on obtient la **probabilité d'appartenance** à la k^{e} gaussienne suivante :

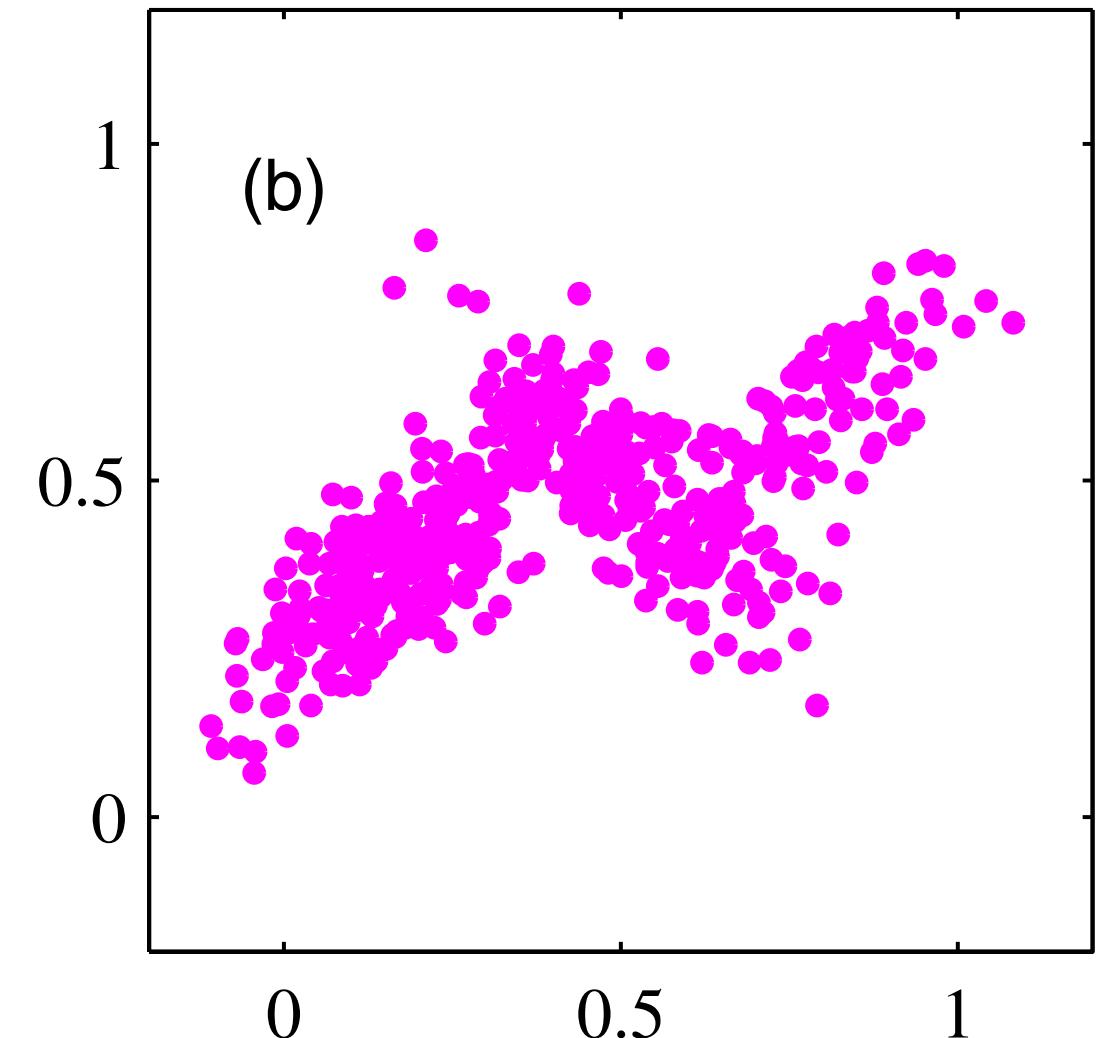
$$\begin{aligned}\gamma(z_k) \equiv p(z_k = 1 | \mathbf{x}) &= \frac{p(z_k = 1)p(\mathbf{x} | z_k = 1)}{\sum_{j=1}^K p(z_j = 1)p(\mathbf{x} | z_j = 1)} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}\end{aligned}$$

MÉLANGE DE GAUSSIENNES

Sujets: mélange de gaussiennes

RAPPEL

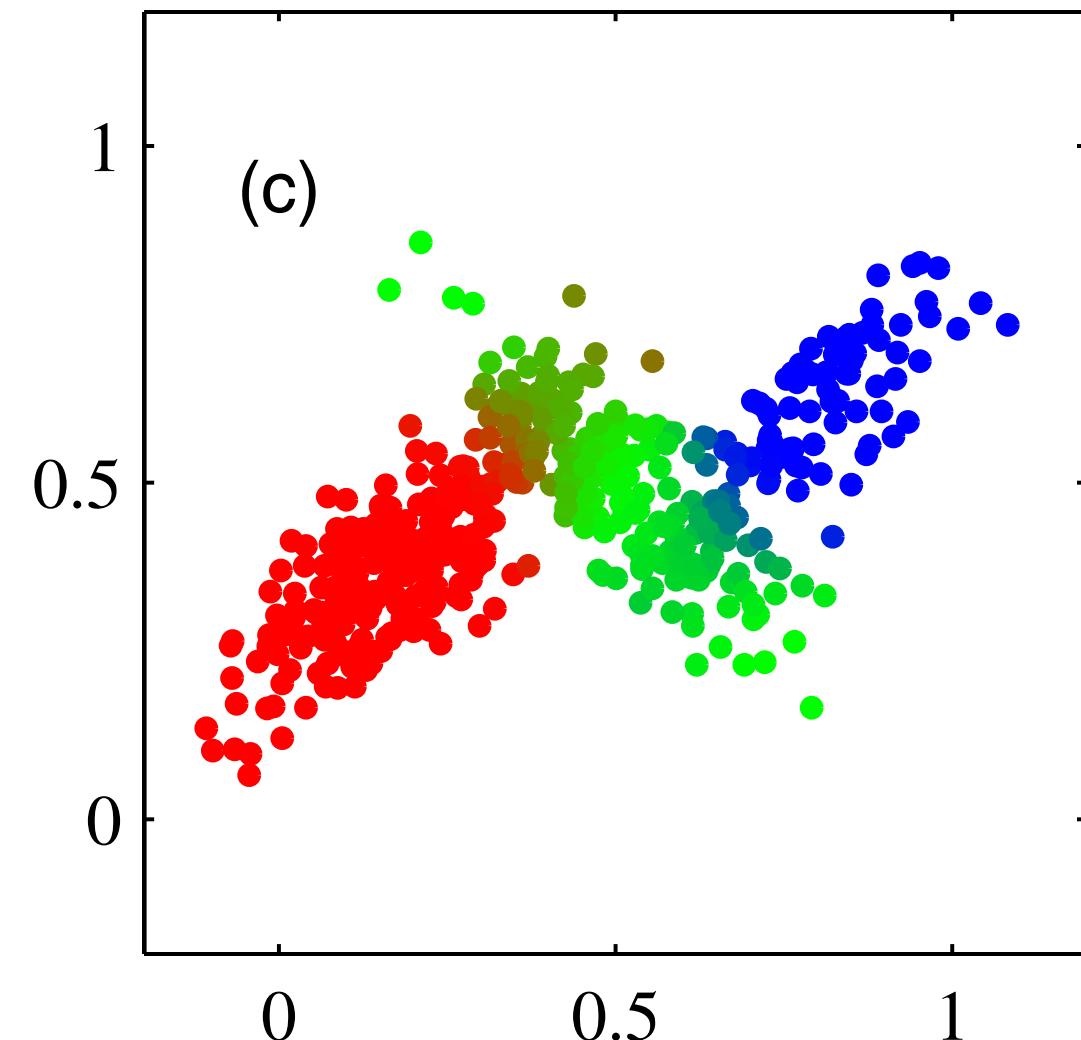
- Dans un mélange de gaussienne, l'appartenance aux K gaussiennes («classes») n'est pas connue



MÉLANGE DE GAUSSIENNES

Sujets: mélange de gaussiennes

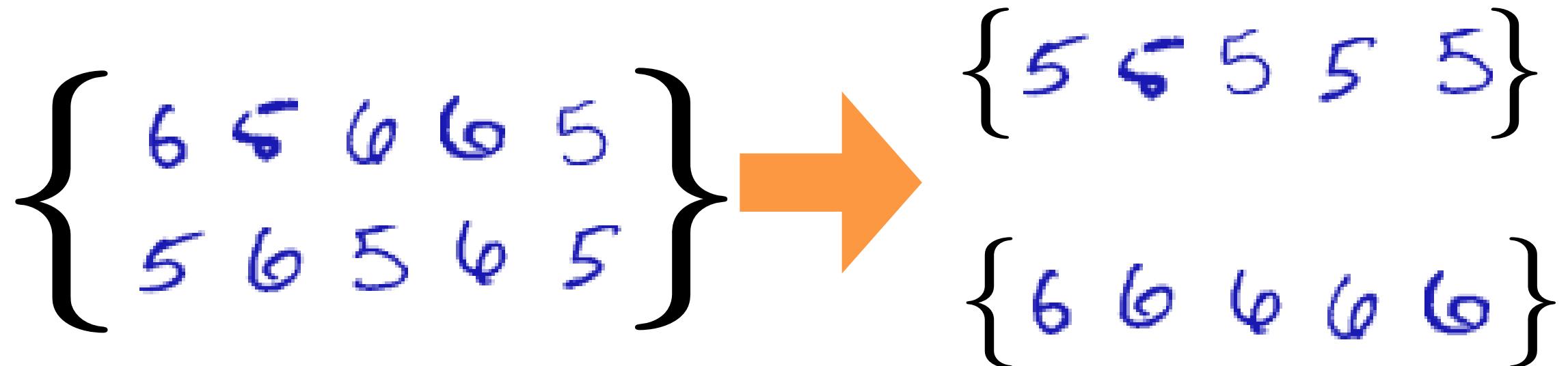
- Illustration des probabilités d'appartenance, pour chaque entrée



PARTITIONNEMENT DE DONNÉES

Sujets: partitionnement de données, *clustering*

- Lors du partitionnement, on assigne chaque entrée x à la gaussienne ayant associée à la plus grande probabilité d'appartenance

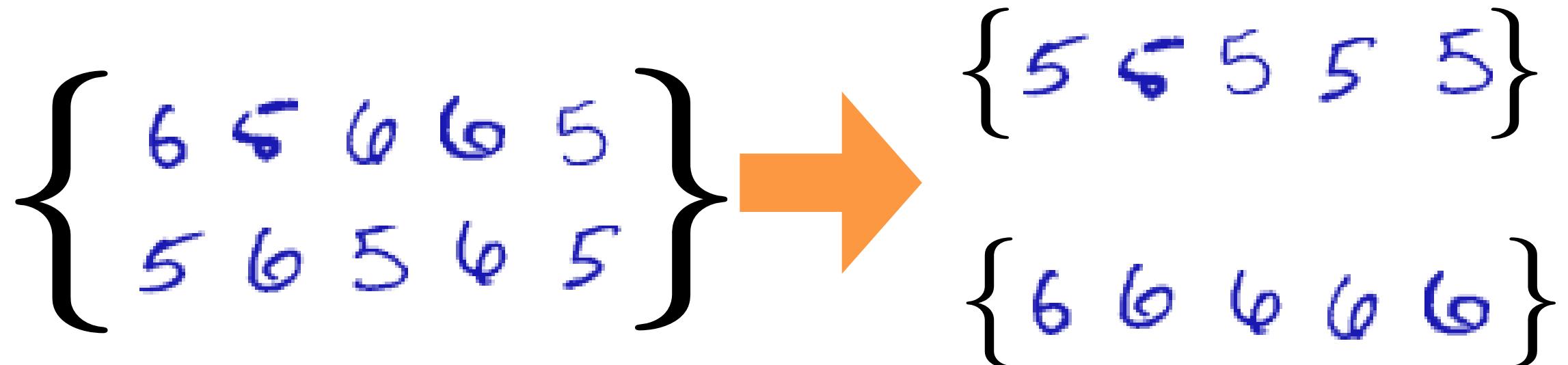


PARTITIONNEMENT DE DONNÉES

Sujets: partitionnement de données, *clustering*

- Lors du partitionnement, on assigne chaque entrée x à la gaussienne ayant associée à la plus grande probabilité d'appartenance

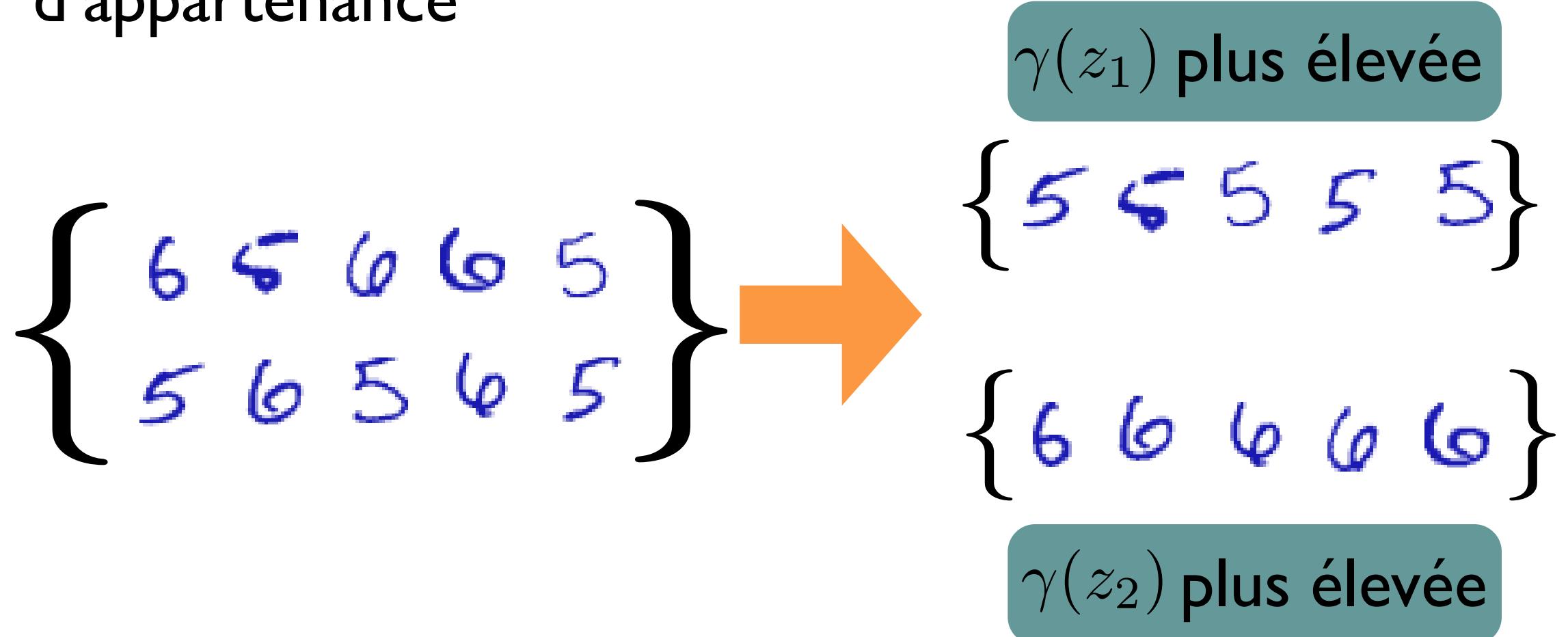
$\gamma(z_1)$ plus élevée



PARTITIONNEMENT DE DONNÉES

Sujets: partitionnement de données, *clustering*

- Lors du partitionnement, on assigne chaque entrée x à la gaussienne ayant associée à la plus grande probabilité d'appartenance



PARTITIONNEMENT DE DONNÉES

Sujets: partitionnement de données, *clustering*

- On n'a pas de garanties qu'on va retrouver les «vraies» catégories ?
 1. les données de chaque catégorie ne sont peut-être pas gaussiennes
 2. le modèle de mélange entraîné n'est peut-être pas bon
(plus là-dessus plus tard)
- Plus les données des différentes catégories seront bien séparées (pas entrelacées), meilleurs seront les résultats

Apprentissage automatique

Mélange de gaussiennes - maximum de vraisemblance (EM)

PARTITIONNEMENT DE DONNÉES

Sujets: partitionnement de données, *clustering*

RAPPEL

- À partir d'un mélange de gaussiennes entraîné, on pourrait inférer à quelle gaussienne appartiennent les entrées
 - on pourrait alors automatiquement catégoriser nos données en fonction des probabilités d'appartenance à chacune des gaussiennes
- Cette application s'appelle le **partitionnement de données (*clustering*)**
 - permet de «mettre de l'ordre» dans les données
 - permet de visualiser les données une partition à la fois

PARTITIONNEMENT DE DONNÉES

Sujets: partitionnement de données, *clustering*

RAPPEL

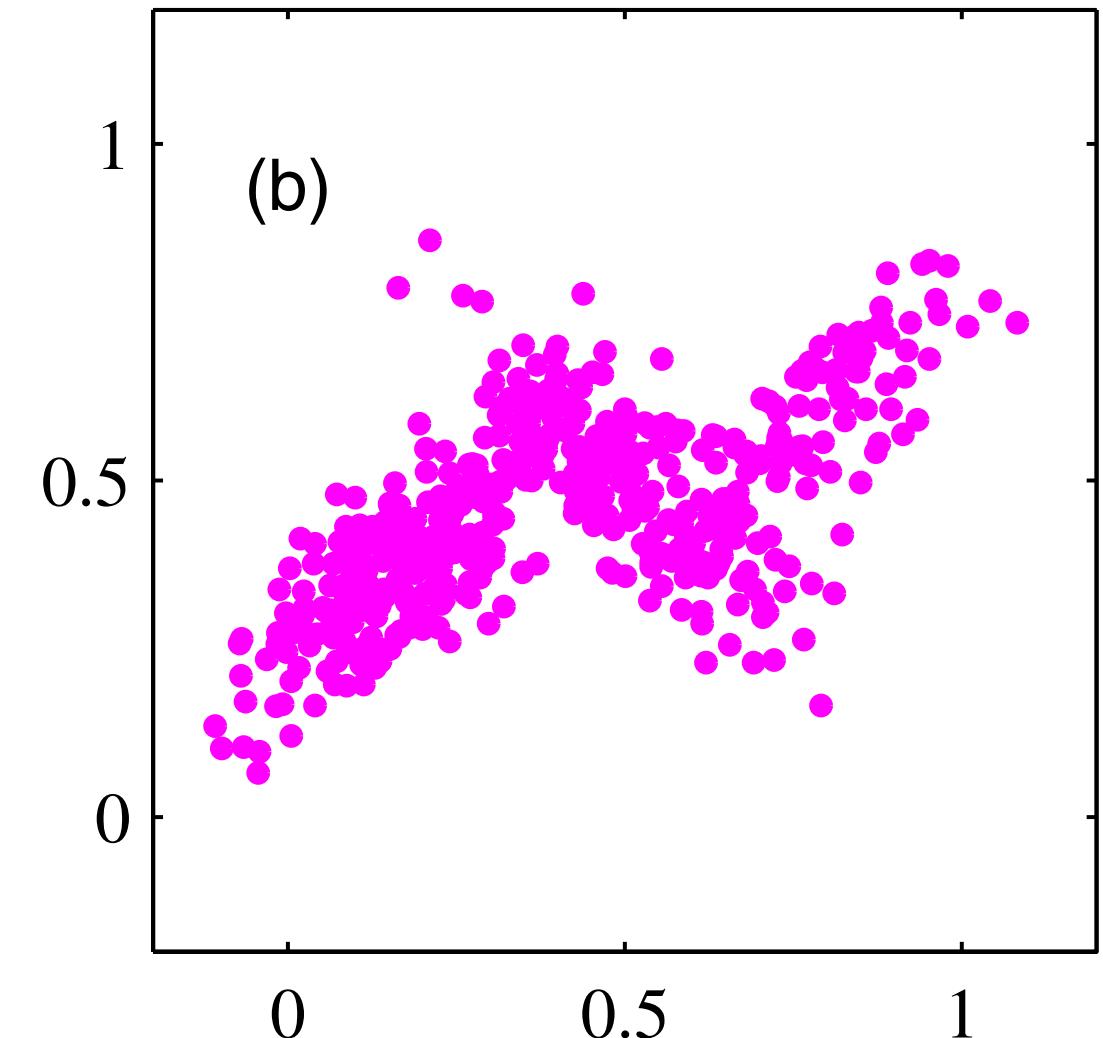
- À partir d'un mélange de gaussiennes entraîné, on pourrait inférer à quelle gaussienne appartiennent les entrées
 - on pourrait alors automatiquement catégoriser nos données en fonction des probabilités d'appartenance à chacune des gaussiennes
- Cette application s'appelle le **partitionnement de données (*clustering*)**
 - permet de «mettre de l'ordre» dans les données
 - permet de visualiser les données une partition à la fois

MÉLANGE DE GAUSSIENNES

Sujets: mélange de gaussiennes

RAPPEL

- Dans un mélange de gaussienne, l'appartenance aux K gaussiennes («classes») n'est pas connue



MÉLANGE DE GAUSSIENNES

Sujets: fonction de densité marginale des entrées

RAPPEL

- Puisqu'on ne connaît pas l'appartenance aux gaussiennes (\mathbf{z}), on va s'intéresser à la probabilité marginale :

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

- C'est de cette façon qu'on va mesurer la performance de notre modèle

MAXIMUM DE VRAISEMBLANCE

Sujets: maximum de vraisemblance

- On va entraîner un mélange de gaussiennes par maximum de vraisemblance

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

- on va maximiser la (log-)vraisemblance marginale des données d'entraînement

MAXIMUM DE VRAISEMBLANCE

Sujets: maximum de vraisemblance

- On va entraîner un mélange de gaussiennes par maximum de vraisemblance

$$\underbrace{\ln p(\mathbf{X}|\pi, \mu, \Sigma)}_{\text{probabilité de tous les } \mathbf{x}_n} = \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k) \right\}$$

- on va maximiser la (log-)vraisemblance marginale des données d'entraînement

MAXIMUM DE VRAISEMBLANCE

Sujets: maximum de vraisemblance

- On va entraîner un mélange de gaussiennes par maximum de vraisemblance

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

- Cas $\boldsymbol{\mu}_k$: la solution doit satisfaire

$$0 = - \sum_{n=1}^N \frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \boldsymbol{\Sigma}_k (\mathbf{x}_n - \boldsymbol{\mu}_k)$$

MAXIMUM DE VRAISEMBLANCE

Sujets: maximum de vraisemblance

- On va entraîner un mélange de gaussiennes par maximum de vraisemblance

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

- Cas $\boldsymbol{\mu}_k$: la solution doit satisfaire

$$0 = - \sum_{n=1}^N \gamma(z_{nk}) \boldsymbol{\Sigma}_k (\mathbf{x}_n - \boldsymbol{\mu}_k)$$

MAXIMUM DE VRAISEMBLANCE

Sujets: maximum de vraisemblance

- On va entraîner un mélange de gaussiennes par maximum de vraisemblance

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

- Cas $\boldsymbol{\mu}_k$: si on suppose que les $\gamma(z_{nk})$ sont fixes

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}_n , \quad \text{où } N_k = \sum_{n=1}^N \gamma(z_{nk})$$

MAXIMUM DE VRAISEMBLANCE

Sujets: maximum de vraisemblance

- On va entraîner un mélange de gaussiennes par maximum de vraisemblance

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

- Cas π_k : on utilise un multiplicateur de Lagrange

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \lambda \left(\sum_{k=1}^K \pi_k - 1 \right)$$

MAXIMUM DE VRAISEMBLANCE

Sujets: maximum de vraisemblance

- On va entraîner un mélange de gaussiennes par maximum de vraisemblance

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

- Cas π_k : la solution doit satisfaire

$$0 = \sum_{n=1}^N \frac{\mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} + \lambda , \forall k \quad \text{et} \quad \sum_{k=1}^K \pi_k = 1$$

MAXIMUM DE VRAISEMBLANCE

Sujets: maximum de vraisemblance

- On va entraîner un mélange de gaussiennes par maximum de vraisemblance

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

- Cas π_k : si on suppose que les $\gamma(z_{nk})$ sont fixes

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

MAXIMUM DE VRAISEMBLANCE

Sujets: maximum de vraisemblance

- On va entraîner un mélange de gaussiennes par maximum de vraisemblance

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

- Cas $\boldsymbol{\Sigma}_k$: si on suppose que les $\gamma(z_{nk})$ sont fixes

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T$$

MAXIMUM DE VRAISEMBLANCE

Sujets: Algorithme EM

- Les solutions pour μ_k, π_k, Σ_k supposent que les $\gamma(z_{nk})$ sont fixes
 - par contre, changer μ_k, π_k et Σ_k va changer $\gamma(z_{nk})$
 - la supposition que les $\gamma(z_{nk})$ ne changeront pas est donc fausse
- Solution : on alterne entre calculer $\gamma(z_{nk})$ et μ_k, π_k, Σ_k
 - I. Étape **Estimation** : calcul de tous les $\gamma(z_{nk})$
 2. Étape **Maximisation** : calcul des μ_k, π_k et Σ_k
- On appelle cela l'**algorithme EM**

ALGORITHME EM

Sujets: Algorithme EM

- Pseudocode
 1. Initialize the means μ_k , covariances Σ_k and mixing coefficients π_k , and evaluate the initial value of the log likelihood.

ALGORITHME EM

Sujets: Algorithme EM

- Pseudocode
2. E step. Evaluate the responsibilities using the current parameter values

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

ALGORITHME EM

Sujets: Algorithme EM

- Pseudocode

3. M step. Re-estimate the parameters using the current responsibilities

$$\mu_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}_n$$

$$\Sigma_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) (\mathbf{x}_n - \mu_k^{\text{new}}) (\mathbf{x}_n - \mu_k^{\text{new}})^T$$

$$\pi_k^{\text{new}} = \frac{N_k}{N}$$

ALGORITHME EM

Sujets: Algorithme EM

- Pseudocode

4. Evaluate the log likelihood

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

and check for convergence of either the parameters or the log likelihood. If the convergence criterion is not satisfied return to step 2.

ALGORITHME EM

Sujets: Algorithme EM

- Pseudocode
4. Evaluate the log likelihood on a validation set

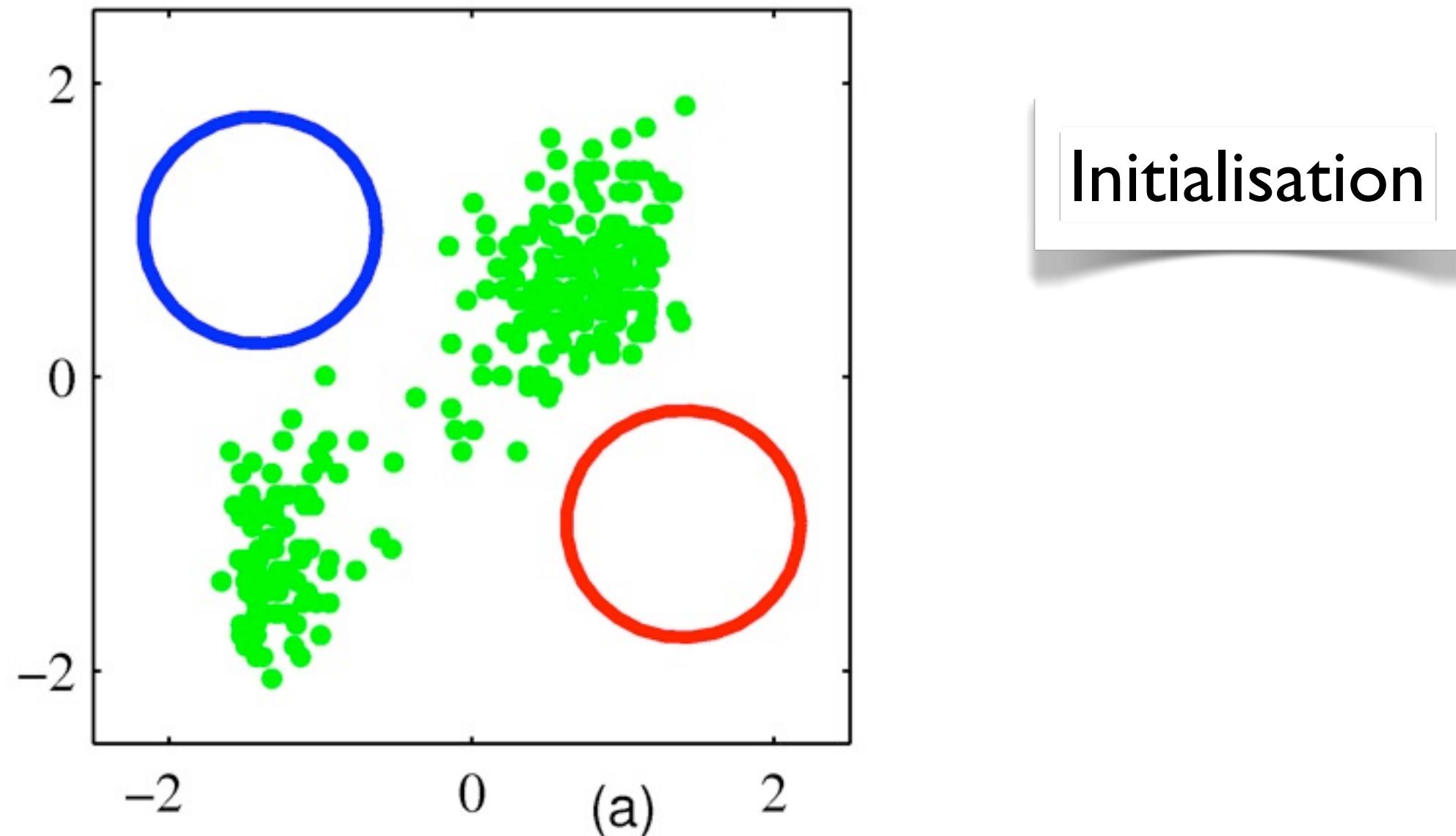
$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

and check for convergence of either the parameters or the log likelihood. If the convergence criterion is not satisfied return to step 2.

ALGORITHME EM

Sujets: Algorithme EM

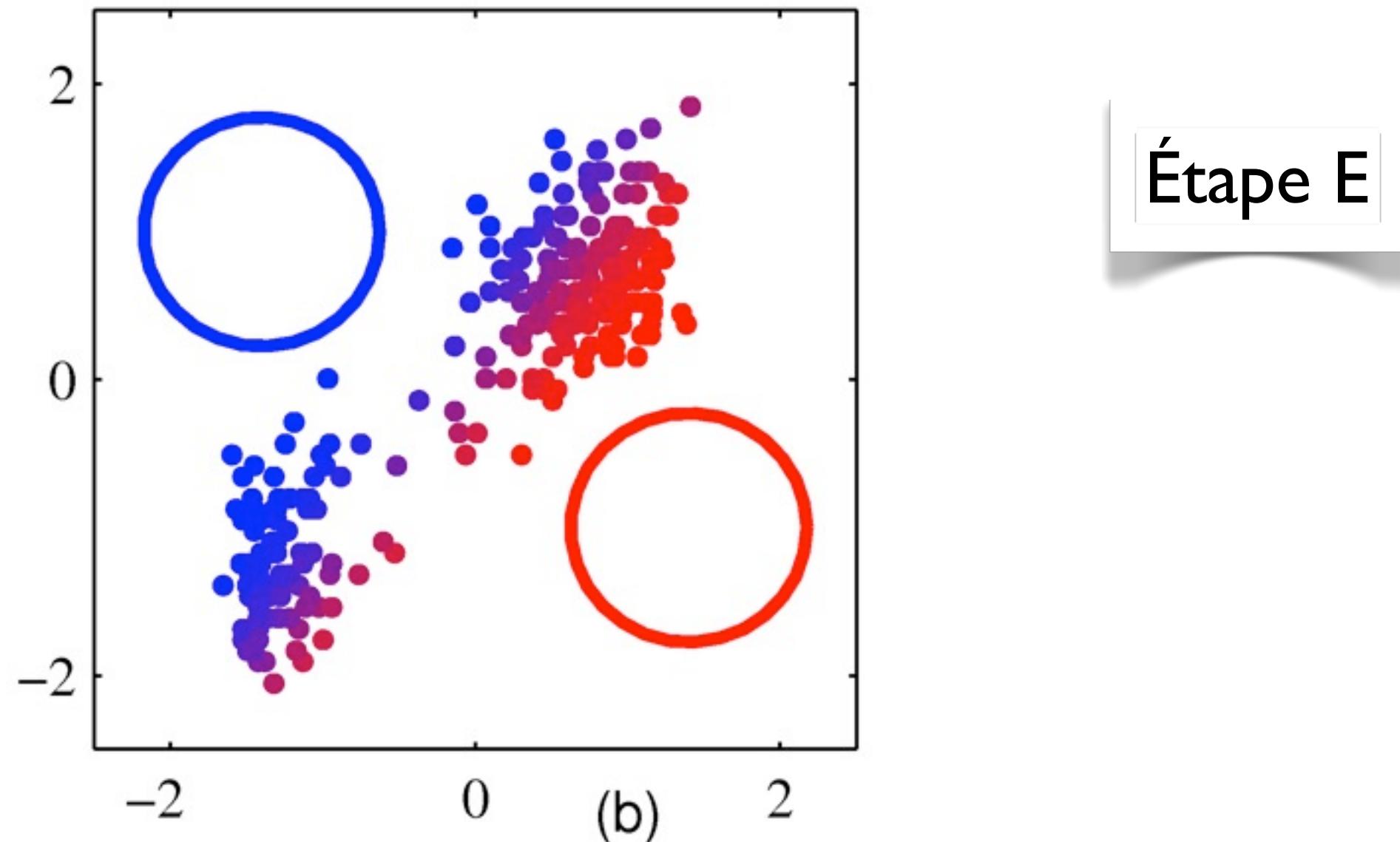
- Exemple d'exécution de l'algorithme EM



ALGORITHME EM

Sujets: Algorithme EM

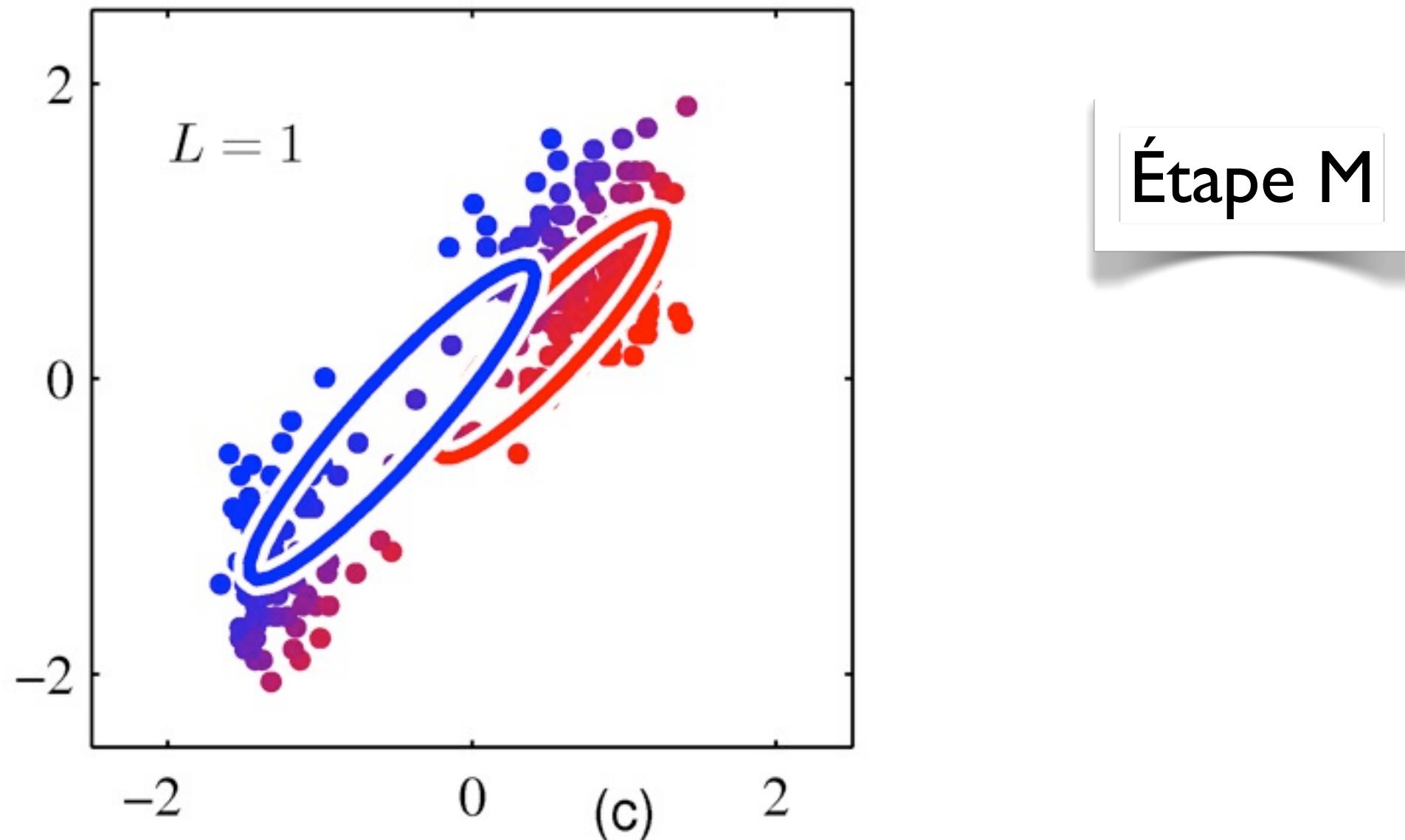
- Exemple d'exécution de l'algorithme EM



ALGORITHME EM

Sujets: Algorithme EM

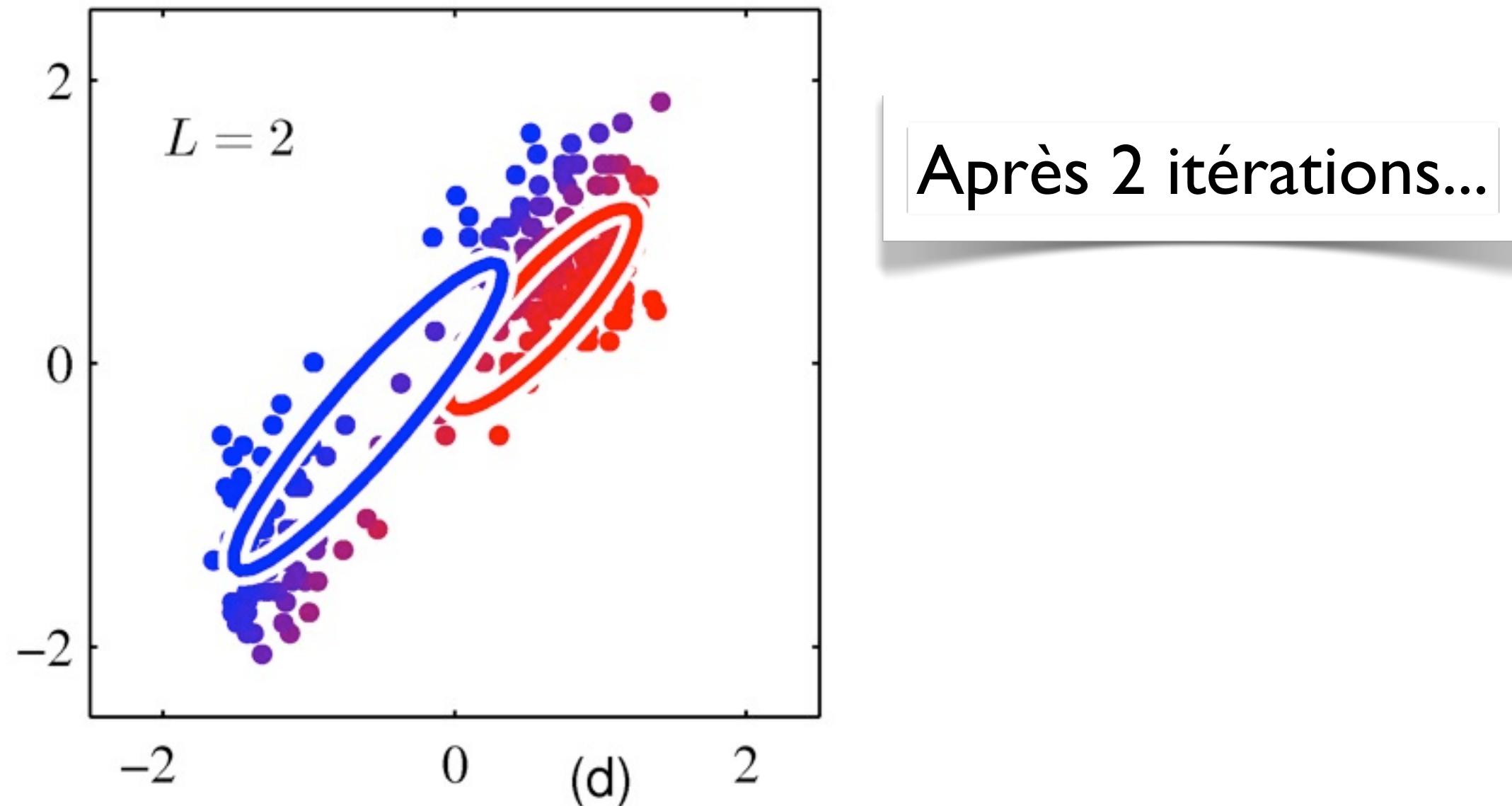
- Exemple d'exécution de l'algorithme EM



ALGORITHME EM

Sujets: Algorithme EM

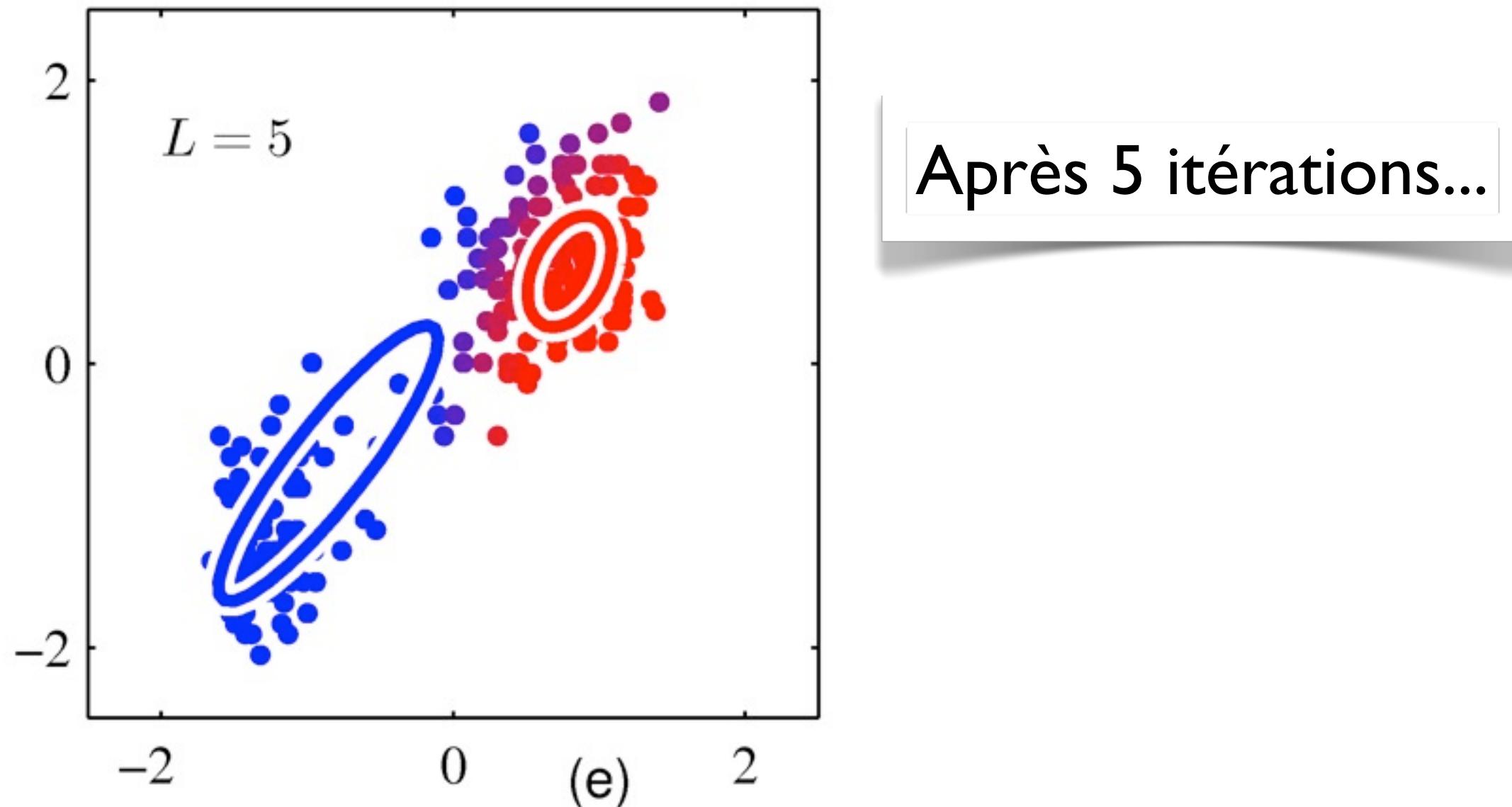
- Exemple d'exécution de l'algorithme EM



ALGORITHME EM

Sujets: Algorithme EM

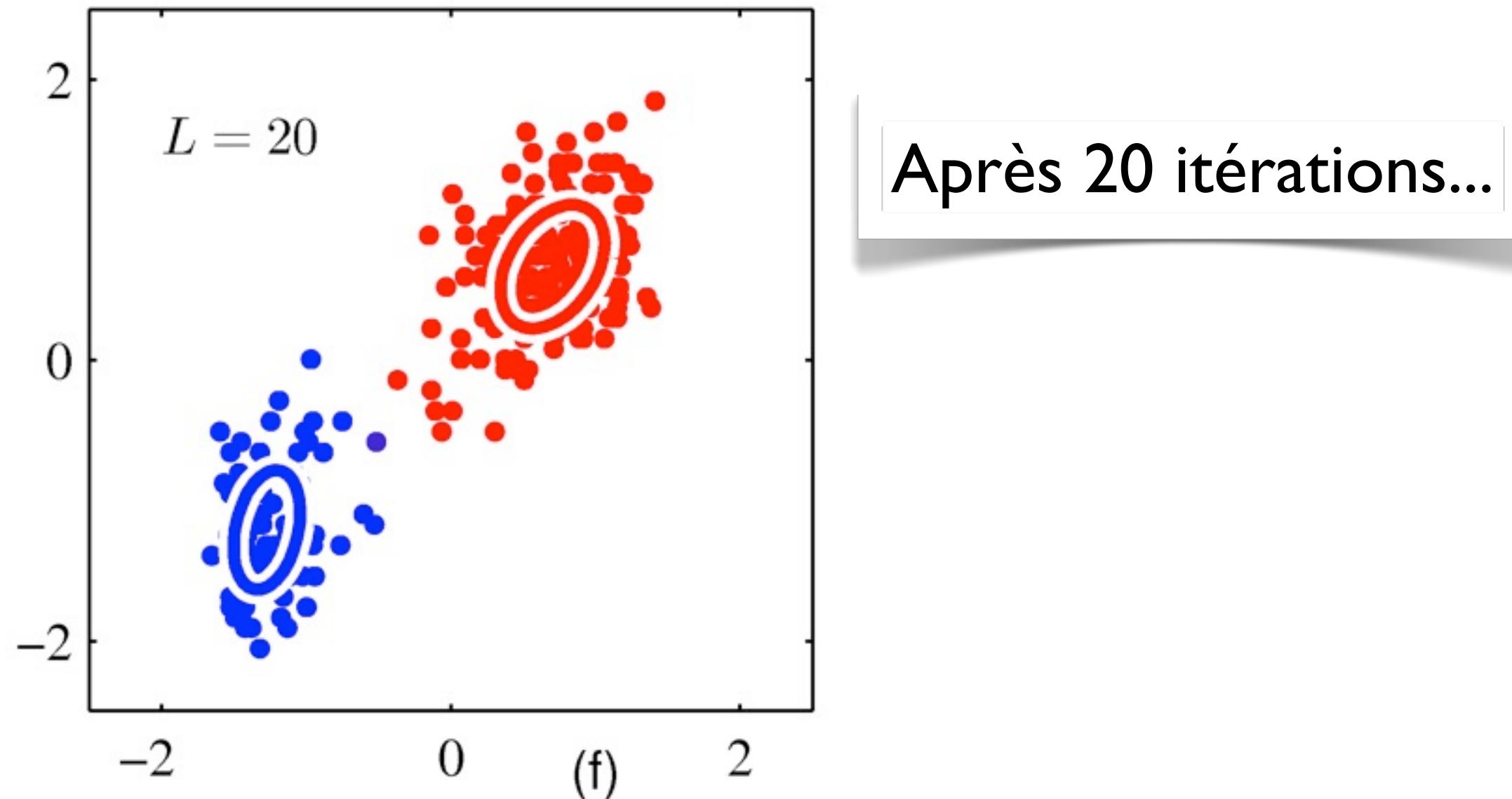
- Exemple d'exécution de l'algorithme EM



ALGORITHME EM

Sujets: Algorithme EM

- Exemple d'exécution de l'algorithme EM



Apprentissage automatique

Mélange de gaussiennes - dérivation générale de EM

MAXIMUM DE VRAISEMBLANCE

Sujets: maximum de vraisemblance

RAPPEL

- On va entraîner un mélange de gaussiennes par maximum de vraisemblance

$$\underbrace{\ln p(\mathbf{X}|\pi, \mu, \Sigma)}_{\text{probabilité de tous les } \mathbf{x}_n} = \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k) \right\}$$

- on va maximiser la (log-)vraisemblance marginale des données d'entraînement

MAXIMUM DE VRAISEMBLANCE

Sujets: Algorithme EM

RAPPEL

- Les solutions pour μ_k, π_k, Σ_k supposent que les $\gamma(z_{nk})$ sont fixes
 - par contre, changer μ_k, π_k et Σ_k va changer $\gamma(z_{nk})$
 - la supposition que les $\gamma(z_{nk})$ ne changeront pas est donc fausse
- Solution : on alterne entre calculer $\gamma(z_{nk})$ et μ_k, π_k, Σ_k
 - I. Étape **Estimation** : calcul de tous les $\gamma(z_{nk})$
 2. Étape **Maximisation** : calcul des μ_k, π_k et Σ_k
- On appelle cela l'**algorithme EM**

ALGORITHME EM

Sujets: Algorithme EM

- On va faire une dérivation alternative de EM qui va nous permettre de
 - comprendre pourquoi on peut s'attendre à ce que l'algorithme converge
 - voir comment on pourrait le généraliser à d'autres modèles de mélange

ALGORITHME EM

Sujets: Algorithme EM

- On commence par remarquer qu'on peut écrire

$$\ln p(\mathbf{X}|\theta) = \mathcal{L}(q, \theta) + \text{KL}(q\|p)$$

où

$$\mathcal{L}(q, \theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \right\}$$

$$\text{KL}(q\|p) = - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \theta)}{q(\mathbf{Z})} \right\}$$

ALGORITHME EM

Sujets: Algorithme EM

toutes les appartenances \mathbf{z}_n

- On commence par remarquer qu'on peut écrire

$$\ln p(\mathbf{X}|\theta) = \mathcal{L}(q, \theta) + \text{KL}(q\|p)$$

où

$$\mathcal{L}(q, \theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \right\}$$

$$\text{KL}(q\|p) = - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \theta)}{q(\mathbf{Z})} \right\}$$

ALGORITHME EM

Sujets: Algorithme EM

- On commence par remarquer qu'on peut écrire

$$\ln p(\mathbf{X}|\theta) = \mathcal{L}(q, \theta) + \text{KL}(q\|p)$$

où

$$\mathcal{L}(q, \theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \right\}$$

$$\text{KL}(q\|p) = - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \theta)}{q(\mathbf{Z})} \right\}$$

ALGORITHME EM

Sujets: Algorithme EM

- On commence par remarquer qu'on peut écrire

$$\ln p(\mathbf{X}|\theta) = \mathcal{L}(q, \theta) + \text{KL}(q\|p)$$

probabilité de toutes
les paires $\mathbf{x}_n, \mathbf{z}_n$

$$\prod_n p(\mathbf{x}_n|\mathbf{z}_n)p(\mathbf{z}_n)$$

où

$$\mathcal{L}(q, \theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \right\}$$

$$\text{KL}(q\|p) = - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \theta)}{q(\mathbf{Z})} \right\}$$

ALGORITHME EM

Sujets: Algorithme EM

- On commence par remarquer qu'on peut écrire

$$\ln p(\mathbf{X}|\theta) = \mathcal{L}(q, \theta) + \text{KL}(q\|p)$$

où

$$\mathcal{L}(q, \theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \right\}$$

$$\text{KL}(q\|p) = - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \theta)}{q(\mathbf{Z})} \right\}$$

ALGORITHME EM

Sujets: Algorithme EM

- On commence par remarquer qu'on peut écrire

$$\ln p(\mathbf{X}|\theta) = \mathcal{L}(q, \theta) + \text{KL}(q\|p)$$

où

$$\mathcal{L}(q, \theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \right\}$$

$$\text{KL}(q\|p) = - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \theta)}{q(\mathbf{Z})} \right\}$$

probabilité conditionnelle
d'appartenance de toutes
les paires $\mathbf{x}_n, \mathbf{z}_n$

$$\prod_n p(\mathbf{z}_n|\mathbf{x}_n) = \prod_n \prod_k \gamma(z_{nk})^{z_{nk}}$$

ALGORITHME EM

Sujets: Algorithme EM

- On commence par remarquer qu'on peut écrire

$$\ln p(\mathbf{X}|\theta) = \mathcal{L}(q, \theta) + \text{KL}(q\|p)$$

où

$$\mathcal{L}(q, \theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \right\}$$

$$\text{KL}(q\|p) = - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \theta)}{q(\mathbf{Z})} \right\}$$

ALGORITHME EM

Sujets: Algorithme EM

- On commence par remarquer qu'on peut écrire

$$\ln p(\mathbf{X}|\theta) = \mathcal{L}(q, \theta) + \text{KL}(q\|p)$$

où

$$\mathcal{L}(q, \theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \right\}$$

$$\text{KL}(q\|p) = - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \theta)}{q(\mathbf{Z})} \right\}$$

n'importe quelle distribution
sur toutes les appartenances \mathbf{z}_n

ALGORITHME EM

Sujets: Algorithme EM

- On commence par remarquer qu'on peut écrire

$$\ln p(\mathbf{X}|\theta) = \mathcal{L}(q, \theta) + \text{KL}(q\|p)$$

où

$$\mathcal{L}(q, \theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \right\}$$

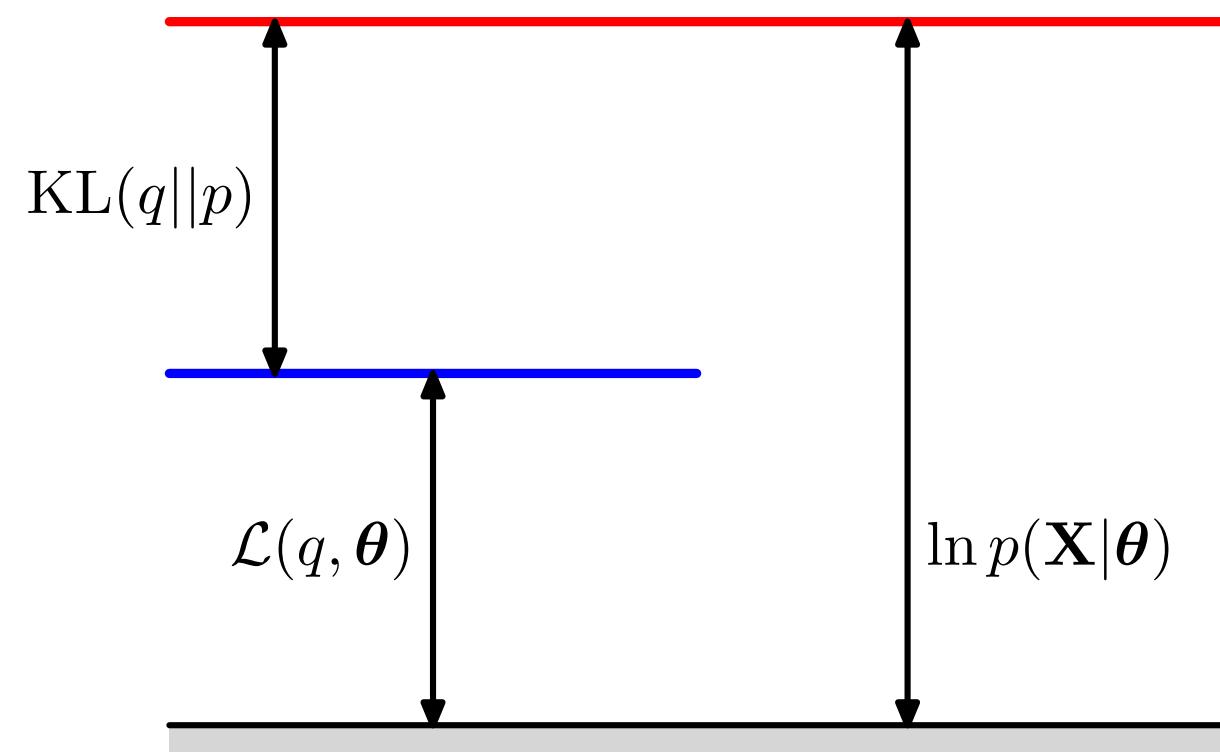
$$\text{KL}(q\|p) = - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \theta)}{q(\mathbf{Z})} \right\}$$

ALGORITHME EM

Sujets: Algorithme EM

- On commence par remarquer qu'on peut écrire

$$\ln p(\mathbf{X}|\theta) = \mathcal{L}(q, \theta) + \text{KL}(q\|p)$$



L'idée générale derrière EM est d'alterner entre changer avec $q(\mathbf{Z})$ (étape E) et θ (étape M)

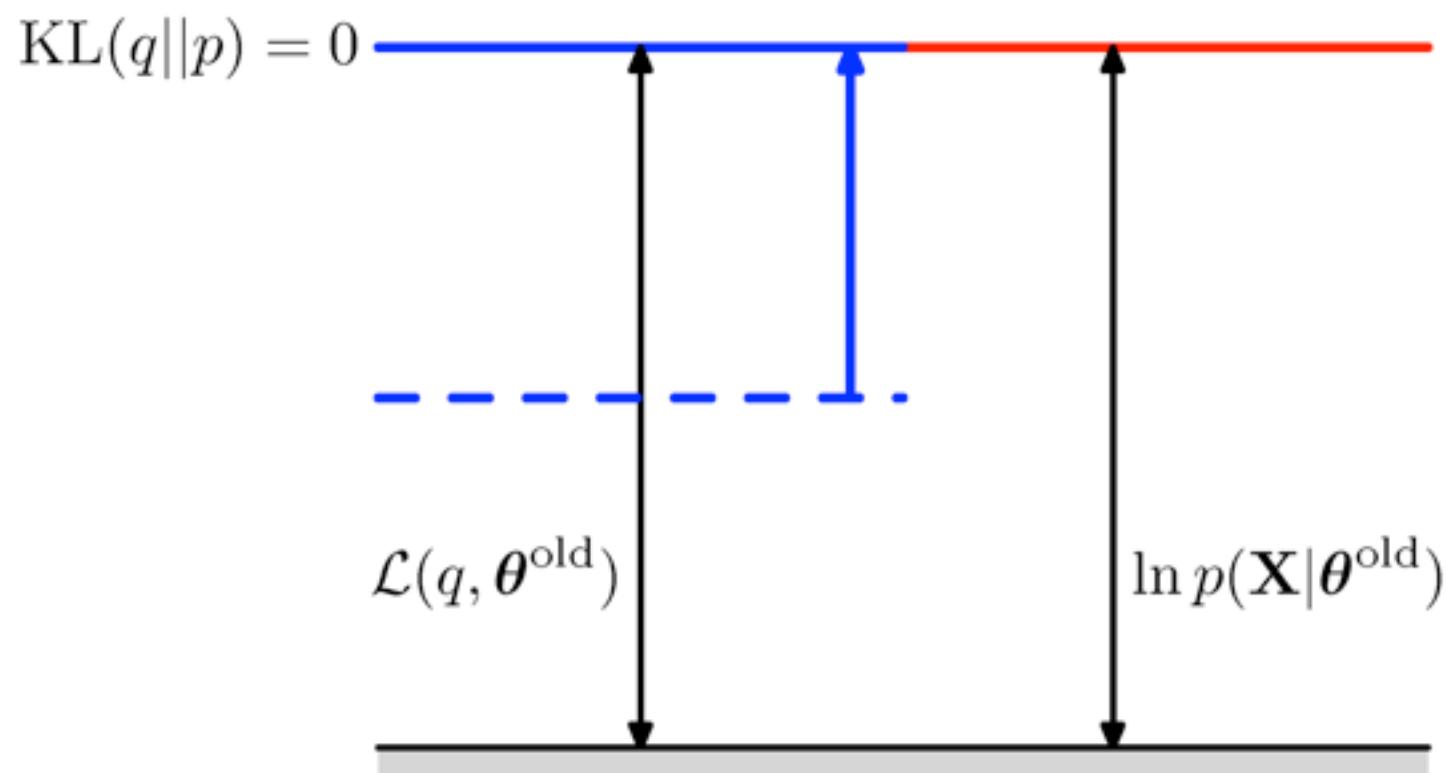
ALGORITHME EM

Sujets: Algorithme EM

- On commence par remarquer qu'on peut écrire

$$\ln p(\mathbf{X}|\theta) = \mathcal{L}(q, \theta) + \text{KL}(q\|p)$$

Étape E : changer $q(\mathbf{Z})$



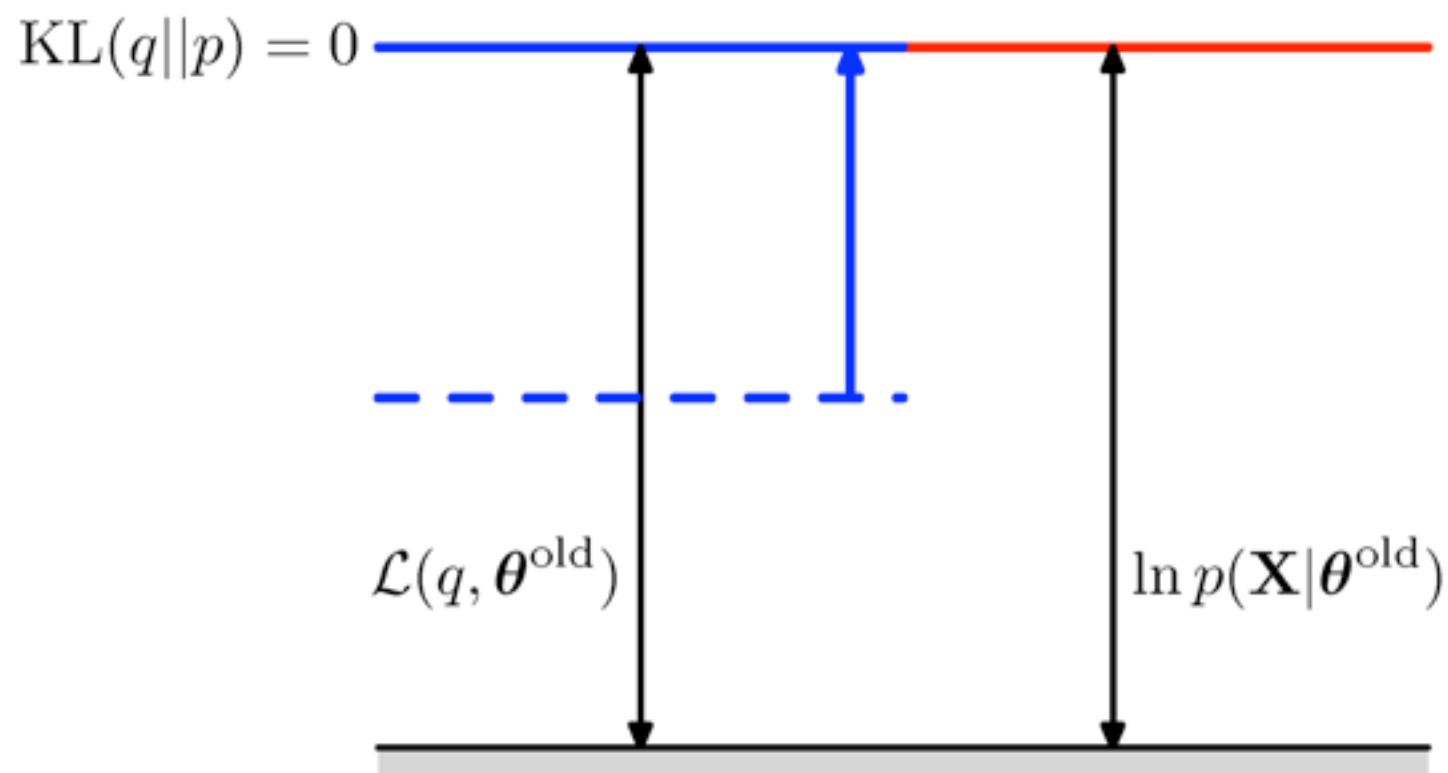
si $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \theta)$,
alors $\text{KL}(q\|p) = 0$

ALGORITHME EM

Sujets: Algorithme EM

- On commence par remarquer qu'on peut écrire

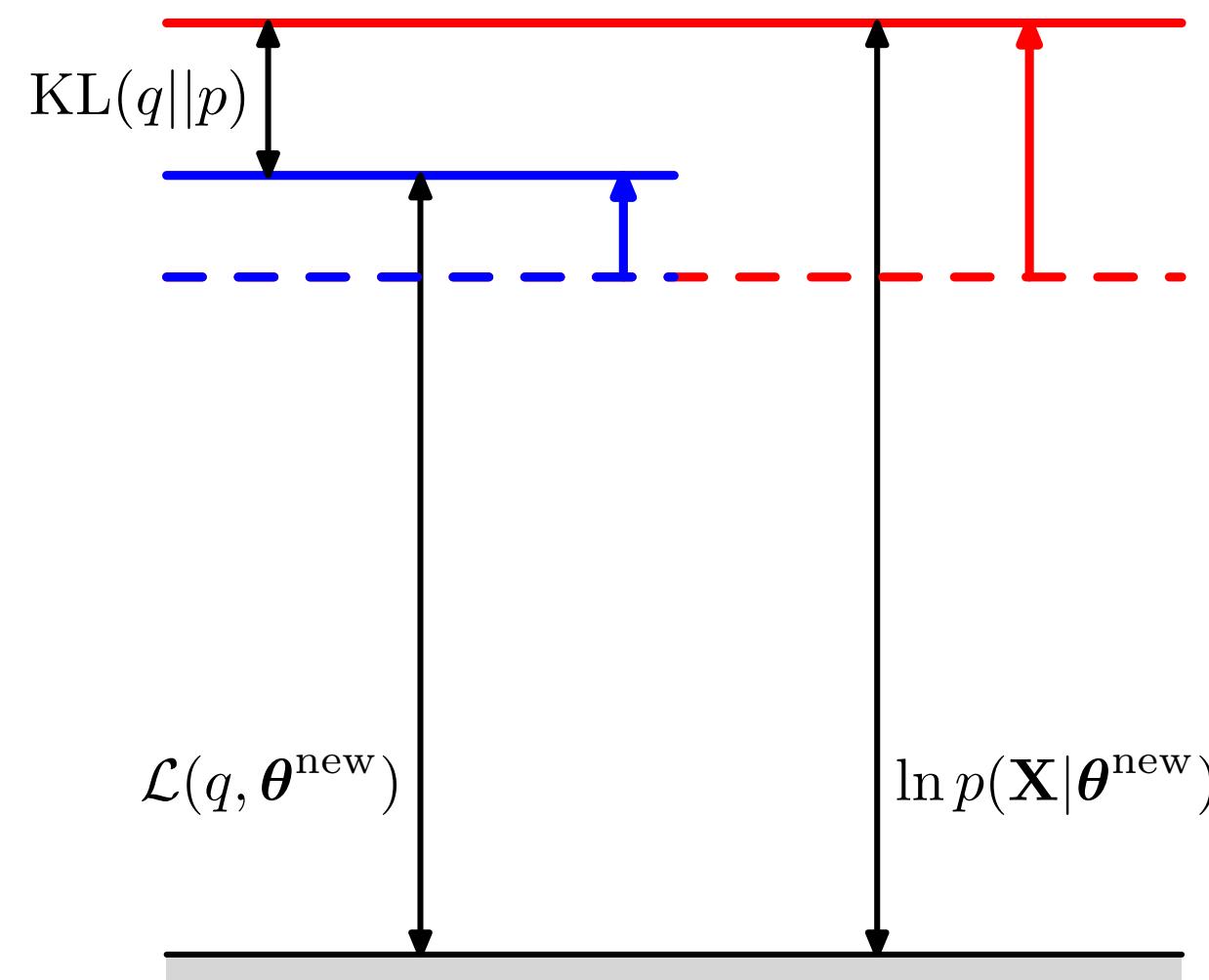
$$\ln p(\mathbf{X}|\theta) = \mathcal{L}(q, \theta) + \text{KL}(q\|p)$$



puisque $\text{KL}(q\|p) \geq 0$,
alors $\mathcal{L}(q, \theta) \leq \ln p(\mathbf{X}|\theta)$

ALGORITHME EM

Sujets: Algorithme EM



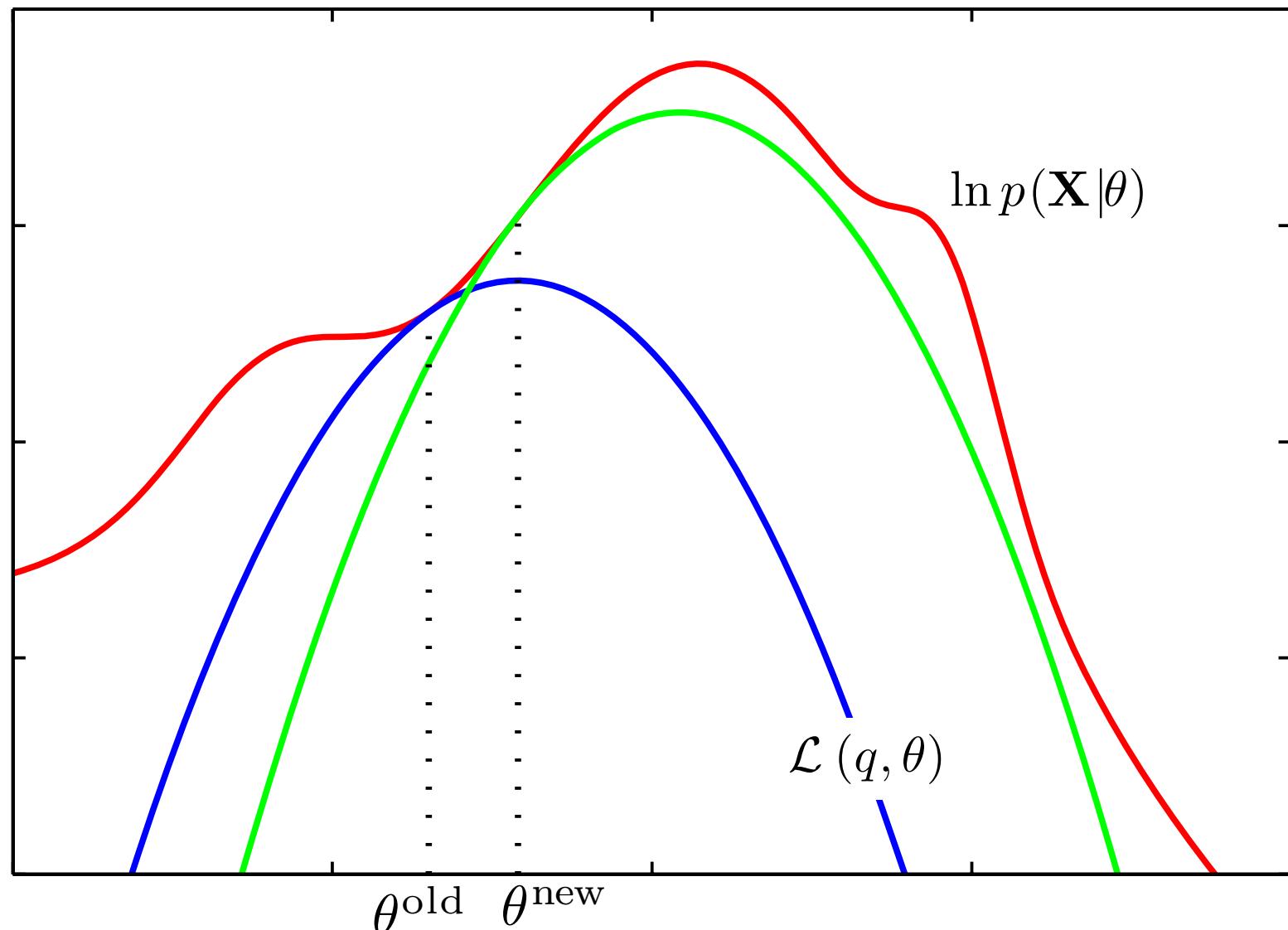
Étape M : maximiser $\mathcal{L}(q, \theta)p/r$ à θ .

$\ln p(\mathbf{X}|\theta)$ ne peut pas diminuer,
donc la performance doit
éventuellement converger

ALGORITHME EM

Sujets: Algorithme EM

- Visualisation alternative :



ALGORITHME EM

Sujets: Algorithme EM

- Lorsqu'on maximise, par rapport à θ

$$\mathcal{L}(q, \theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln p(\mathbf{X}, \mathbf{Z} | \theta) - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln q(\mathbf{Z})$$

on maximise une log-vraisemblance «complétée», où on suppose qu'une entrée \mathbf{x}_n appartient $\gamma(z_{nk})\%$ du temps à chaque gaussienne k

ALGORITHME EM

Sujets: Algorithme EM

- Lorsqu'on maximise, par rapport à θ

$$\begin{aligned}\mathcal{L}(q, \theta) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln p(\mathbf{X}, \mathbf{Z} | \theta) - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln q(\mathbf{Z}) \\ &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln p(\mathbf{X}, \mathbf{Z} | \theta) + \text{const}\end{aligned}$$

on maximise une log-vraisemblance «complétée», où on suppose qu'une entrée \mathbf{x}_n appartient $\gamma(z_{nk})\%$ du temps à chaque gaussienne k

ALGORITHME EM

Sujets: Algorithme EM

- Lorsqu'on maximise, par rapport à θ

$$\begin{aligned}\mathcal{L}(q, \theta) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln p(\mathbf{X}, \mathbf{Z} | \theta) - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln q(\mathbf{Z}) \\ &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln p(\mathbf{X}, \mathbf{Z} | \theta) + \text{const} \\ &= \sum_n \sum_k \gamma(z_{nk}) \ln(p(\mathbf{x}_n | z_{nk} = 1)p(z_{nk} = 1)) + \text{const}\end{aligned}$$

on maximise une log-vraisemblance «complétée», où on suppose qu'une entrée \mathbf{x}_n appartient $\gamma(z_{nk})\%$ du temps à chaque gaussienne k

ALGORITHME EM

Sujets: Algorithme EM

- Lorsqu'on maximise, par rapport à θ

$$\begin{aligned}\mathcal{L}(q, \theta) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln p(\mathbf{X}, \mathbf{Z} | \theta) - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln q(\mathbf{Z}) \\ &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln p(\mathbf{X}, \mathbf{Z} | \theta) + \text{const} \\ &= \sum_n \sum_k \gamma(z_{nk}) \ln(p(\mathbf{x}_n | z_{nk} = 1)p(z_{nk} = 1)) + \text{const}\end{aligned}$$

on obtient les mises à jours de μ_k , π_k et Σ_k vues précédemment

Apprentissage automatique

Mélange de gaussiennes - résumé

MÉLANGE DE GAUSSIENNES

Sujets: résumé du modèle de mélange de gaussiennes

- Modèle :

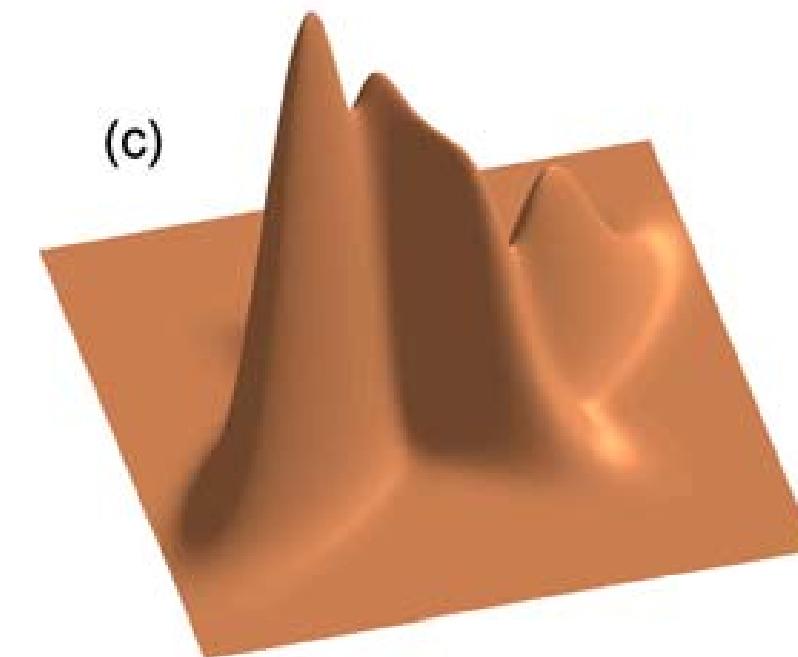
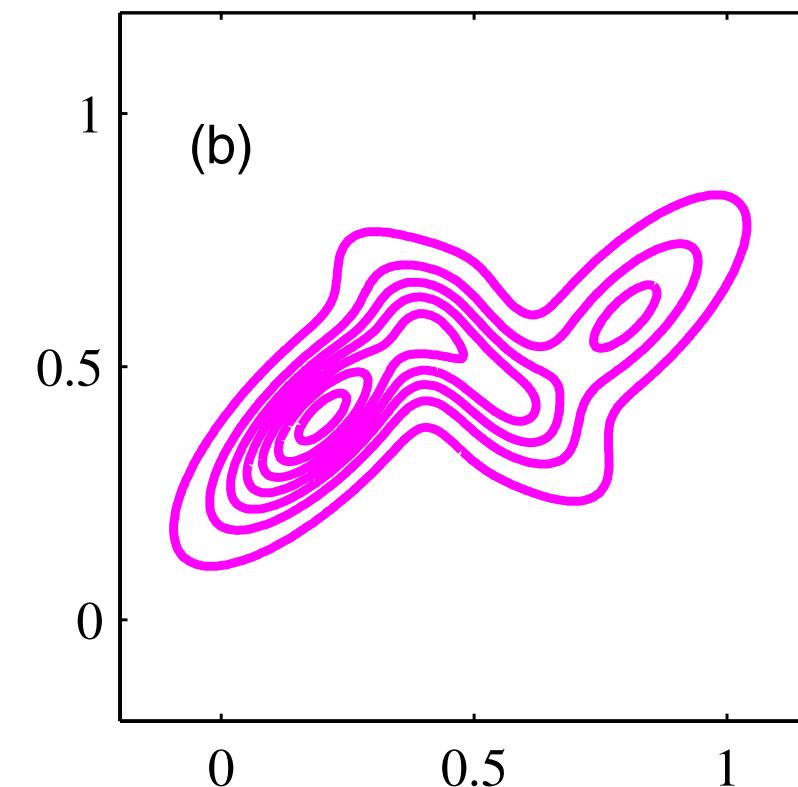
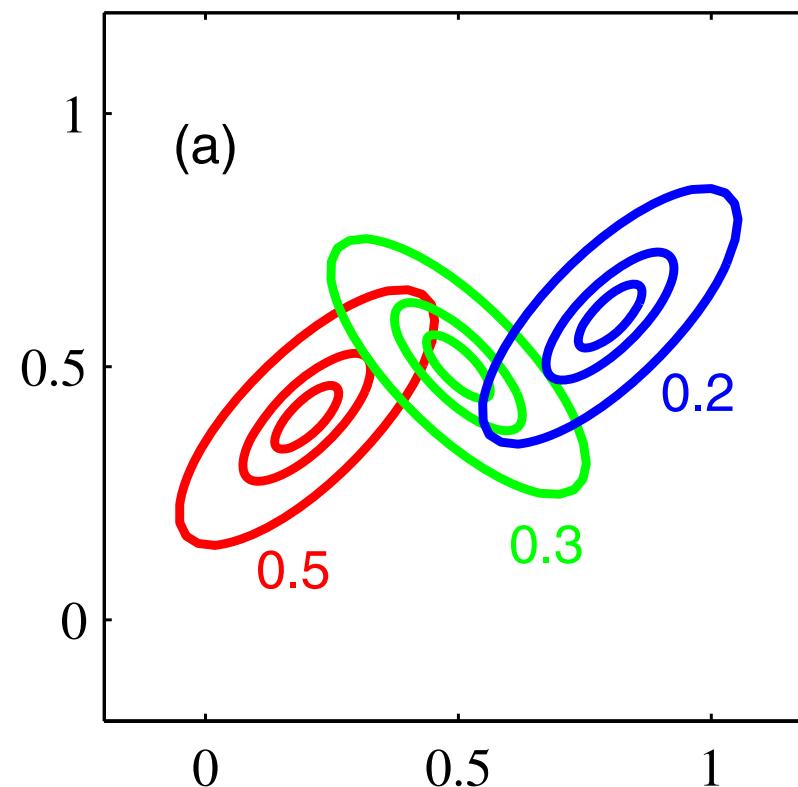
$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- Entraînement : algorithme EM (maximum vraisemblance)
 - Étape E : calcul des $\gamma(z_{nk})$ (colle la borne $\mathcal{L}(q, \theta)$ sur $\ln p(\mathbf{X}|\theta)$)
 - Étape M : maximise la borne par rapport à $\mathcal{L}(q, \theta)$
- Hyper-paramètre : K , nb. d'itérations d'EM
- Prédiction : $p(\mathbf{x})$ ou assignation à une des K gaussiennes

MÉLANGE DE GAUSSIENNES

Sujets: résumé du modèle de mélange de gaussiennes

- Visualisation



MÉLANGE DE GAUSSIENNES

Sujets: détails sur utilisation

- Hyper-paramètres vs capacité
 - plus le nombre de gaussiennes K est grand, plus la capacité augmente
 - plus le nombre d'itérations d'EM augmente, plus la capacité augmente
- Pour la sélection de modèle, on utilise $\ln p(\mathbf{X})$ comme mesure de performance
 - plus elle est élevée, meilleure la performance
 - de façon équivalente, $-\ln p(\mathbf{X})$ est l'erreur du modèle

MÉLANGE DE GAUSSIENNES

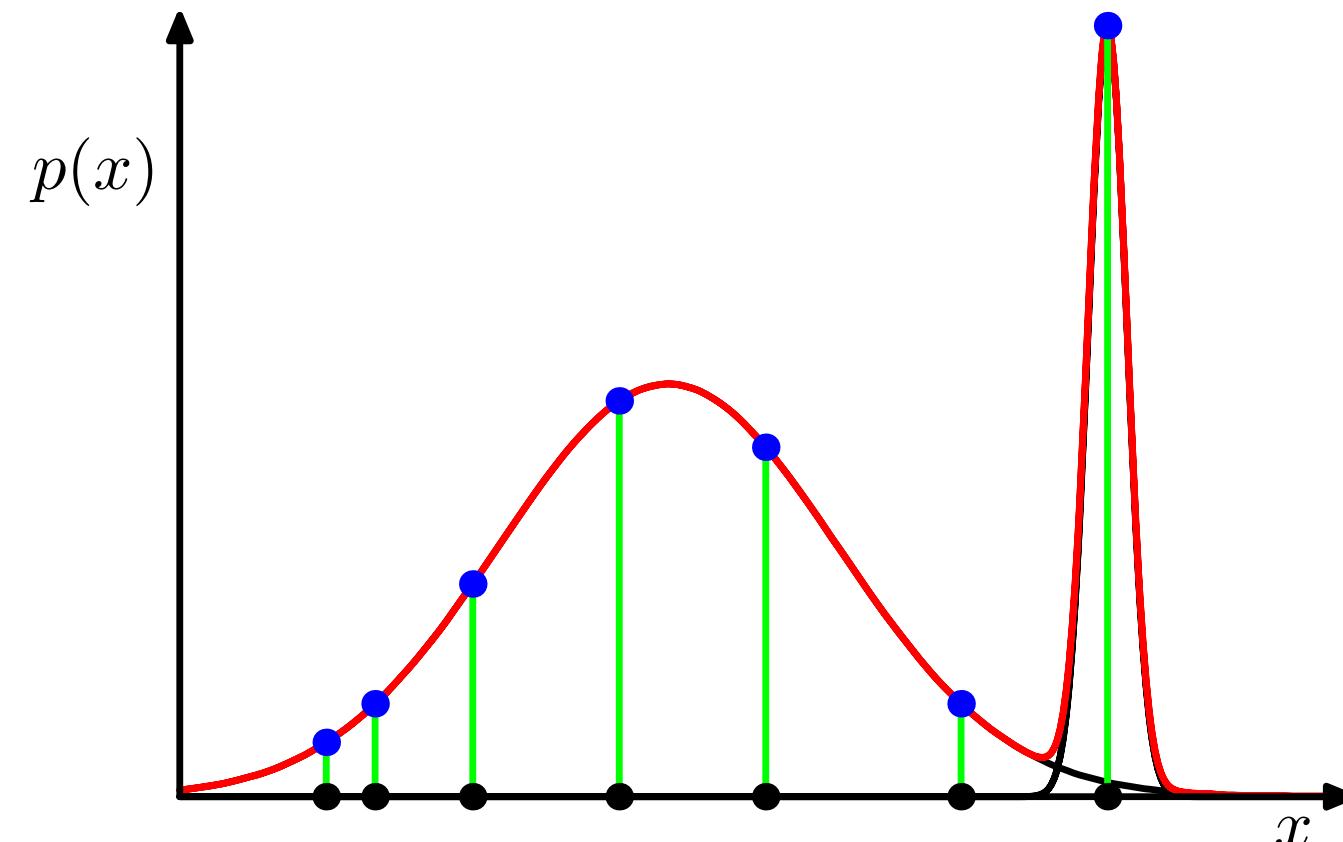
Sujets: détails sur utilisation

- Existe plusieurs optima locaux (problème non-convexe)
 - initialisation est importante
 - approche : initialiser les moyennes à K exemples \mathbf{x}_n aléatoires, auxquels on ajoute du bruit gaussien

MÉLANGE DE GAUSSIENNES

Sujets: détails sur utilisation

- Le problème d'optimisation n'est pas bien défini
 - si une gaussienne est centrée sur un exemple \mathbf{x}_n , la probabilité de \mathbf{x}_n peut devenir infinie en faisant tendre la covariance vers 0



FENÊTRE DE PARZEN

Sujets: fenêtre de Parzen

- La fenêtre de Parzen est un mélange de gaussienne où
 - $K = N$: il y a une gaussienne par exemple d'entraînement
 - $\pi_k = \frac{1}{N}$: chaque gaussienne a la même probabilité a priori
 - $\mu_n = \mathbf{x}_n$: chaque gaussienne est centrée autour d'un exemple
 - $\Sigma_n = \sigma^2 \mathbf{I}$: la covariance est sphérique (σ^2 est un hyper-param.)
- C'est un modèle très simple (voire simpliste), intéressant si on veut seulement estimer $p(\mathbf{x})$, sans partitionnement