# 2 Apprentissage non-supervisé

## 2.1 Introduction à l'apprentissage non-supervisé

❒ **Motivation** – Le but de l'apprentissage non-supervisé est de trouver des formes cachées dans un jeu de données non-labelées $\{x^{(1)},...,x^{(m)}\}$.

❒ **Inégalité de Jensen** – Soit $f$ une fonction convexe et $X$ une variable aléatoire. On a l'inégalité suivante :

$$\boxed{E[f(X)] \geqslant f(E[X])}$$

## 2.2 Partitionnement

### 2.2.1 Espérance-Maximisation

❒ **Variables latentes** – Les variables latentes sont des variables cachées/non-observées qui posent des difficultés aux problèmes d'estimation, et sont souvent notées $z$. Voici les cadres dans lesquelles les variables latentes sont le plus fréquemment utilisées :

| Cadre | Variance latente $z$ | $x\|z$ | Commentaires |
|---|---|---|---|
| Mixture de k gaussiennes | Multinomial($\phi$) | $\mathcal{N}(\mu_j,\Sigma_j)$ | $\mu_j \in \mathbb{R}^n, \phi \in \mathbb{R}^k$ |
| Analyse factorielle | $\mathcal{N}(0,I)$ | $\mathcal{N}(\mu+\Lambda z,\psi)$ | $\mu_j \in \mathbb{R}^n$ |

❒ **Algorithme** – L'algorithme d'espérance-maximisation (EM) est une méthode efficace pour estimer le paramètre $\theta$. Elle passe par le maximum de vraisemblance en construisant un borne inférieure sur la vraisemblance (E-step) et optimisant cette borne inférieure (M-step) de manière successive :

— E-step : Évaluer la probabilité postérieure $Q_i(z^{(i)})$ que chaque point $x^{(i)}$) provienne d'une partition particulière $z^{(i)}$ de la manière suivante :

$$\boxed{Q_i(z^{(i)}) = P(z^{(i)}|x^{(i)};\theta)}$$

— M-step : Utiliser les probabilités postérieures $Q_i(z^{(i)})$ en tant que coefficients propres aux partitions sur les points $x^{(i)}$ pour ré-estimer séparemment chaque modèle de partition de la manière suivante :

$$\boxed{\theta_i = \underset{\theta}{\text{argmax}} \sum_i \int_{z^{(i)}} Q_i(z^{(i)}) \log\left(\frac{P(x^{(i)},z^{(i)};\theta)}{Q_i(z^{(i)})}\right) dz^{(i)}}$$