

# Probabilistic Graphical Models and Deep Generative Models

D'après Pierre Latouche et Pierre-Alexandre Mattei

9 octobre 2025



## Table des matières

<b>1</b>	<b>Notion de Modèle Graphique Orienté</b>	<b>1</b>
1.1	Modèle Statistique . . . . .	1
1.2	Statistiques et Information . . . . .	2
1.3	Estimation Fréquentistes . . . . .	3
1.4	Approche Bayésienne . . . . .	4
1.5	Modèles Graphiques . . . . .	5
<b>2</b>	<b>Modèles de Mixtures Gaussiennes</b>	<b>6</b>
2.1	Mixtures Gaussiennes . . . . .	6
2.2	Algorithme EM . . . . .	6

## Résumé

<mailto:pierre.latouche@uca.fr> Page du cours sur <https://lmbp.uca.fr/~latouche/mva/IntroductiontoProbabilisticGraphicalModels.html>

## 1 Notion de Modèle Graphique Orienté

### 1.1 Modèle Statistique

**Définition 1.1** Un *échantillon* aléatoire  $(X_1, \dots, X_n)$  de taille  $n$  est un vecteur de  $n$  variables aléatoires indépendantes et identiquement distribuées suivant la même loi qu'une variable aléatoire  $X_0$ .

**Définition 1.2** Un *modèle statistique* est une famille de lois de probabilités sur un espace  $\chi$ . Le modèle est dénoté par :

$$\mathcal{M} = \{p(\cdot | \theta), \theta \in \Theta\}$$

- Si  $\Theta \subseteq \mathbb{R}^d$ , le modèle est dit *paramétrique*.
- Sinon, on dira que le modèle est *non paramétrique*.

Pour une variable aléatoire discrète  $X$ , on notera  $p(x | \theta) = \mathbb{P}_\theta(X = x)$ . Pour une variable aléatoire continue  $X$ , on notera  $p(x | \theta) = f_\theta(x)$  où  $f_\theta$  est une densité de probabilité. Dans la suite, on dira que  $X$  est aléatoire et que  $x$  est une réalisation ou observation de  $X$  (même chose pour  $X_i$  et  $x_i$ ). Le modèle  $\{p(x | \theta)\}$  est l'unique lien entre  $\theta$  et l'observation  $x$ .

**Définition 1.3** La *vraisemblance* (likelihood) de  $\theta$  de l'observation  $x$  est la fonction :

$$l_x(\theta) = p(x | \theta)$$

La *log-vraisemblance* est la fonction :

$$L_x(\theta) = \log l_x(\theta)$$

Dans la suite, on va chercher à maximiser la vraisemblance par rapport au paramètre  $\theta$  étant donné un échantillon et un modèle statistique. Puisqu'on a une hypothèse d'indépendance :

$$l_{(x_1, \dots, x_n)}(\theta) = \prod_{i=1}^n p(x_i | \theta)$$

et on a :

$$L_{(x_1, \dots, x_n)}(\theta) = \sum_{i=1}^n \log p(x_i | \theta)$$

La log-vraisemblance d'un échantillon est la somme des log-vraisemblances des variables partielles d'observation. C'est une fonction aléatoire (définissant une variable aléatoire) en  $\theta$ .

Il y a trois versions principales du paramètre  $\theta$  qui vont être utilisées :

1.  $\theta^*$ , une constante inconnue qui est le véritable paramètre du modèle ;
2.  $\theta_0$ , une constante connue qui est un candidat pour le paramètre  $\theta$  et sert à tester une hypothèse sur la valeur de  $\theta$  ;
3.  $\hat{\theta}$ , une variable aléatoire fonction de l'échantillon aléatoire, et qui sert à estimer  $\theta$ .

## 1.2 Statistiques et Information

**Définition 1.4** Une *statistique*  $T$  est une fonction de l'échantillon aléatoire  $(X_1, \dots, X_n)$ .

Par la suite, on supposera que les lois  $p(x | \theta)$  dans le modèle ont même support et sont deux fois dérivables en  $\theta$ , de sorte que les opérateurs de dérivation par rapport à  $\theta$  et d'intégration par rapport à  $x$  commutent.

**Définition 1.5** Le *score* d'un échantillon est défini par :

$$L'_{(X_1, \dots, X_n)}(\theta) = \sum_{i=1}^n L'_{X_i}(\theta)$$

C'est une fonction aléatoire de  $\theta$ .

**Théorème 1.1** En  $\theta^*$ , l'espérance du score est nulle :

$$\mathbb{E}^*[L'_{(X_1, \dots, X_n)}(\theta^*)] = 0$$

Ici,  $\mathbb{E}^*$  désigne l'espérance par rapport à la loi  $p(\cdot | \theta^*)$ , l'espérance par rapport à la vraie distribution. Dans la suite, sauf précisé autrement, c'est toujours par rapport à cette distribution qu'on calculera l'espérance.

*Démonstration.* Par définition d'une mesure de probabilité, on a :

$$\int l_x(\theta) dx = \int p(x | \theta) dx = 1$$

En dérivant sous le signe intégral :

$$\int l'_x(\theta) dx = \int l_x(\theta) L'_x(\theta) dx = 0$$

Par définition de  $l_x(\theta)$  et par définition de l'espérance :

$$\mathbb{E}^*[L'_x(\theta^*)] = \int L'_x(\theta^*) p(x | \theta^*) dx = 0$$

■

**Théorème 1.2** L'information de Fisher de  $X$  en  $\theta^*$  est :

$$I_X(\theta^*) = \mathbb{E}^*[-L''_X(\theta^*)] = \mathbb{V}^*(L'_X(\theta^*))$$

De même que précédemment,  $\mathbb{V}^*$  désigne la variance calculée par rapport à la distribution  $\theta^*$ .

*Démonstration.* Par définition :

$$L''_X(\theta) = \frac{l''_X(\theta)l_X(\theta) - l'_X(\theta)^2}{l_X(\theta)^2} = \frac{l''_X(\theta)}{l_X(\theta)} - \left(\frac{l'_X(\theta)}{l_X(\theta)}\right)^2$$

Donc, en  $\theta = \theta^*$ , et en passant à l'espérance :

$$I_X(\theta^*) = -0 + \mathbb{E}\left[\left(\frac{l'_X(\theta^*)}{l_X(\theta^*)}\right)^2\right] = \mathbb{E}[L'_X(\theta^*)^2] = \mathbb{V}(L'_X(\theta^*))$$

■

**Proposition 1.1** L'information de Fisher en  $\theta^*$  pour un échantillon  $(X_1, \dots, X_n)$  est  $I_n(\theta)n \times I_{X_i}(\theta), \forall i$ .

On notera  $I_{X_i}(\theta) = I(\theta)$ .

**R** On a  $\mathbb{E}^*[L''_X(\theta^*)] < 0$ , en supposant la fonction deux fois continuellement dérivable sur un voisinage de  $\theta^*$ , elle y est concave et donc  $\theta^*$  est le lieu d'un maximum.

### 1.3 Estimation Fréquentistes

**Définition 1.6** Un *estimateur*  $\hat{\theta}(X_1, \dots, X_n)$  est une statistique. On définit le *biais* d'un estimateur par :

$$b_{\theta^*}(\hat{\theta}) = \mathbb{E}[\hat{\theta}(X_1, \dots, X_n)] - \theta^*$$

Un estimateur est *non biaisé* si son biais est nul.

**Définition 1.7** Un estimateur  $\hat{\theta}$  est dit *asymptotiquement non-biaisé* si pour tout  $\theta^*$  :

$$\mathbb{E}^*[\hat{\theta}(X_1, \dots, X_n)] \xrightarrow{n \rightarrow \infty} \theta^*$$

**Définition 1.8** La *déviat*ion carrée moyenne d'un estimateur  $\hat{\theta}$  est :

$$Q = \mathbb{E}^*[\left(\hat{\theta}(X_1, \dots, X_n) - \theta^*\right)^2]$$

**Proposition 1.2** On a :

$$Q = b_{\theta^*}(\hat{\theta})^2 + \mathbb{V}^*(\hat{\theta})$$

*Démonstration.* C'est la formule de König-Huygens pour des variables i.i.d. ■

Dans la suite, on va s'intéresser à des estimateurs dont le comportement est asymptotiquement bon quand  $n \rightarrow \infty$ . On notera  $\hat{\theta}(X_1, \dots, X_n) = \hat{\theta}_n$

**Définition 1.9** Une suite d'estimateurs  $\hat{\theta}_n$  est dite *consistente* quand  $\hat{\theta}_n \xrightarrow[n \rightarrow \infty]{C} \theta^*$ , pour une propriété de convergence  $C$  à préciser.

- On parle de *consistence forte* quand  $C$  est la convergence presque sûre ;
- On parle de *consistence en moyenne carrée* quand  $Q(\hat{\theta}_n) \xrightarrow[n \rightarrow \infty]{} 0$  ;
- On parle de *consistence en probabilité* quand :

$$\forall \varepsilon > 0, \mathbb{P}\left(\theta^* - \varepsilon < \hat{\theta}_n < \theta^* + \varepsilon\right) \xrightarrow[n \rightarrow \infty]{} 1$$

En pratique, on utilise majoritairement la consistance en moyenne carrée, puisque faire tendre  $Q$  vers 0 revient à trouver un estimateur non-biaisé dont la variance tend vers 0, asymptotiquement.

**Définition 1.10** L'*estimateur de vraisemblance maximale* (EVM, ou MLE (*maximum likelihood estimator*)) pour le modèle statistique  $\mathcal{M} = \{p(\cdot | \theta) | \theta \in \Theta\}$  est :

$$\hat{\theta} = \arg \max_{\theta \in \Theta} L_{(X_1, \dots, X_n)}(\theta) \quad (\text{EVM})$$

**R** Si  $\hat{\theta}$  est un EVM de  $\theta^*$  pour le modèle statistique  $\mathcal{M}$ , et  $\psi$  est une bijection, alors  $\psi(\hat{\theta})$  est un EVM de  $\psi(\theta^*)$ .

Rien ne présuppose en général qu'un EVM existe ou est unique. On va donc souvent avoir besoin d'optimiser numériquement, sans garantie (en général) de trouver un maximum global. Par concavité de la vraisemblance dans un petit voisinage de  $\theta^*$ , on sait que l'EVM y est unique.

## 1.4 Approche Bayésienne

Dans l'approche bayésienne, au lieu d'introduire des estimateurs, on introduit une loi  $p(\theta)$  avant d'observer les données, puis on calcule la loi postérieure  $p(\theta | x_1, \dots, x_n)$ . Par la formule de Bayes :

$$\underbrace{p(\theta | x_1, \dots, x_n)}_{\text{Postérieur}} = \frac{\underbrace{p(x_1, \dots, x_n | \theta)}_{\text{Vraisemblance prior}} \underbrace{p(\theta)}_{\text{marginalisation}}}{\underbrace{p(x_1, \dots, x_n)}_{\text{marginalisation}}} \quad (1)$$

Cette formule s'applique notamment dans le cas de l'apprentissage supervisé.

### 1.4.0.1 Bernoulli et Beta

Faisons un petit exemple. On prend  $\text{support}(X_i) = \{0, 1\}$ ,  $p(x|\mu) = \mu^x(1-\mu)^{1-x}$ . On considère *a priori*  $p(\mu) = \text{Beta}(\mu, a_0, b_0) = \frac{1}{B(a,b)}\mu^{a-1}(1-\mu)^{b-1}\forall\mu \in [0, 1]$ . On a  $\mathbb{E}[\mu] = \frac{a}{a+b}$ ,  $\mathbb{V}(\mu) = \frac{ab}{(a+b)^2(a+b+1)}$  et  $\text{mode}(\mu) = \frac{a-1}{a+b-2}$ .

**Proposition 1.3** La distribution posérieure de  $\mu$  dans le modèle de Bernoulli est donnée par :

$$p(\mu | x_1, \dots, x_n) = \text{Beta}(\mu, a_n, b_n)$$

avec  $a_n = a_0 + \sum_{i=1}^n x_i$  et  $b_n = b_0 + n - \sum_{i=1}^n x_i$ .

**Définition 1.11** La distribution Beta est appelée la *distribution conjuguée a priori* de la loi de Bernoulli puisque la distribution postérieure est aussi une distribution Beta.

Dans le modèle Bayésien, l'estimateur a posteriori maximum est défini par :

$$\hat{\mu}_{\text{MAP}} = \frac{a_0 + \sum_{i=1}^n x_i - 1}{a_0 + b_0 + n - 2}$$

L'estimateur basé sur la distribution prédictive est :

$$\hat{\mu} = \frac{a_0 + \sum_{i=1}^n x_i}{a_0 + b_0 + n}$$

**Proposition 1.4** On a :  $\hat{\mu}_{\text{MAP}} \rightarrow \hat{\mu}_{\text{ML}}$  et  $\hat{\mu} \rightarrow \hat{\mu}_{\text{ML}}$  quand  $n \rightarrow \infty$ .

*Démonstration.* ■

## 1.5 Modèles Graphiques

On cherche à donner des propriétés probabilistes sur les données, et particulièrement les modéliser par des distributions conditionnelles. Les modèles graphiques sont un formalisme permettant l'analyse de distributions conditionnelles grâce à de la théorie des graphes. On ne fera pas de rappels des notions de théorie des graphes dans ce polycopié, sauf pour introduire des notations. En général on considèrera des graphes orientés acycliques (GAO, ou DAG en anglais).

**Définition 1.12** Un *modèle graphique* est une famille de distribution de probabilités factorisables d'une manière définie par un GAO donné. Chaque sommet du graphe va correspondre à une variable aléatoire, les arêtes correspondent à des dépendances entre variables.

Les modèles graphiques orientés correspondent aux modèles statistiques les plus communément utilisés.

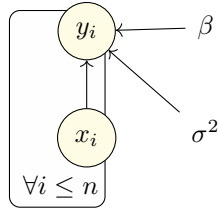
On va s'intéresser à la notion de factorisation dans un DAO :

**Définition 1.13** Une distribution de probabilité  $p$  se *factorise dans*  $G$  lorsque pour tout  $x$  :

$$p(x) \prod_{i=1}^d p(x_i | x_{pa_i})$$

où les  $pa_i$  parcourent l'ensemble des parents du sommet  $i$ .

Par exemple, le graphe suivant exprime la régression linéaire fréquentiste :



avec :

$$p(y_i | x_i, \beta, \sigma^2) = \mathcal{N}(y_i; x_i \beta, \sigma^2)$$

## 2 Modèles de Mixtures Gaussiennes

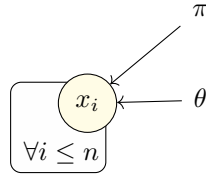
### 2.1 Mixtures Gaussiennes

**Définition 2.1** Une *mixture gaussienne* à  $K$  composante est définie par la densité :

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

où  $\theta = (\mu_k, \Sigma_k)_k$  et  $\pi = {}^t(\pi_1, \dots, \pi_K) \in \Delta_K$  est le vecteur des poids.

En modèle graphique :

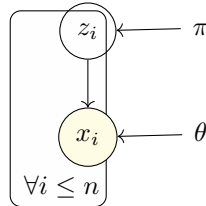


**Proposition 2.1** La log-vraisemblance d'une mixture Gaussienne est donnée par :

$$\begin{aligned} L_{x_1, \dots, x_n}(\pi, \theta) &= \sum_{i=1}^n \log p(x_i | \pi, \theta) \\ &= \sum_{i=1}^n \log \left( \sum_{k=1}^K \pi_k \mathcal{N}(x_i; \mu_k, \Sigma_k) \right) \end{aligned}$$

Optimiser ceci n'est pas trivial, puisqu'on n'a pas d'expression analytique pour les estimateurs de  $\pi$  et  $\theta$ .

En écrivant  $Z_i \sim \mathcal{M}(1, \pi)$  et  $X_i | Z_{ik} = 1 \sim \mathcal{N}(\mu_k, \Sigma_k)$ , on a le modèle suivant :



### 2.2 Algorithme EM

**Définition 2.2** En considérant l'échantillon dit complet des couples  $(X_i, Z_i)$ , la log-vraisemblance *à données complètes*

est donnée par :

$$\begin{aligned} L_{(x_i, z_i)_i}(\pi, \theta) &= \log p((x_i, z_i)_i | \pi, \theta) \\ &= \sum_{i=1}^n \sum_{k=1}^K z_{ik} \log(\pi_k \mathcal{N}(x_i; \mu_k, \Sigma_k)). \end{aligned}$$

**Proposition 2.2** Les estimateurs de  $\pi$  et  $\theta$  maximisant la log-vraisemblance à données complètes sont :

- $\hat{\pi}_k = (1/n) \sum z_{ik}$
- $\hat{\mu}_k = (1/n_k) \sum z_{ik} x_i$
- $\hat{\Sigma}_k = (1/n_k) \sum z_{ik} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$

Puisqu'on ne connaît pas les  $z_i$  et donc l'information de groupement qu'on cherche, on utilise l'algorithme EM (expectation maximisation), qui est une variante pondérée de l'algorithme des  $k$ -moyennes.

L'algorithme se fonde sur deux propriétés :

**Lemme 2.1** Étant données les observations et les paramètres, les  $Z_i$  sont indépendants.

**Lemme 2.2** Les probabilités  $p(z_i | x_i, \pi, \theta) = \mathcal{M}(z_i, 1, \tau_i)$  où :

$$\tau_{ik} = \frac{\pi_k \mathcal{N}(x_i, \mu_k, \Sigma_k)}{\sum_{l=1}^K \pi_l \mathcal{N}(x_i, \mu_l, \Sigma_l)}$$

$\tau_{ik}$  est la probabilité que l'observation  $i$  soit dans le groupe  $k$ .

On peut donc calculer la log-vraisemblance par rapport aux  $Z_i$  considérés comme variables aléatoires :

**Proposition 2.3**

$$\mathbb{E}_{(Z_i)_i} [L_{(x_i, Z_i)_i}(\pi, \theta)] = \sum_{i=1}^n \sum_{k=1}^K \tau_{ik} \log(\pi_k \mathcal{N}(x_i, \mu_k, \Sigma_k))$$

**Proposition 2.4** Les estimateurs de  $\pi$  et  $\theta$  maximisant la log-vraisemblance à données complètes sont :

- $\hat{\pi}_k = (1/n) \sum \tau_{ik}$
- $\hat{\mu}_k = (1/n_k) \sum \tau_{ik} x_i$
- $\hat{\Sigma}_k = (1/n_k) \sum \tau_{ik} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$

On a alors l'algorithme suivant :

---

#### Algorithme 1 EM

---

On initialise les groupes, par exemple avec  $k$ -moyennes. Ensuite, tant que la log-vraisemblance change :

M Calculer  $\hat{\pi}$  et  $\hat{\theta}$  par rapport aux  $\tau_i$

E Calculer les  $\tau_i$  par rapport à  $\pi$  et  $\theta$ .

---

On vérifie aisément qu'une itération d'EM augmente la log-vraisemblance, mais on n'a pas de garantie, en général, de converger vers un maximum global.

Pour estimer le nombre  $K$  de groupes à partir des données, on applique l'algorithme pour différentes valeurs de  $K$  et on choisit un critère, basé sur  $M_K$  le nombre de paramètre libres :

- Le critère d'information bayésienne :  $L_{x_i}(\hat{\pi}, \hat{\theta}) - (M_K/2) \log n$
- Le critère d'information d'Akaike :  $L_{x_i}(\hat{\pi}, \hat{\theta}) - M_K$ .