Modèles hybrides de sources pour la quantification de l'aléa sismique

Bayesian modeling of earthquakes

Ibrahim SEYDI

14 septembre 2025

Restricted case

Spatial data without aftershocks or foreshocks

Let observed seismic events be :

$$x_1, \ldots, x_n \in \mathbb{R}^2.$$

We assume that each observation $x_i = (x_i^{(1)}, x_i^{(2)})$ is a spatial coordinate drawn independently from a density f which is unknown.

We set that :

$$\begin{aligned} x_i \mid \theta_i &\sim \mathcal{N}(x_i \mid \mu_i, \Sigma_i), \quad i = 1, \ldots, n \\ \theta_i = (\mu_i, \Sigma_i) \mid G &\sim G \\ G \mid \alpha, G_0 &\sim \mathrm{DP}(\alpha, G_0) \\ G_0 \mid m_0, \Lambda_0, \psi_0, \nu_0 &= \mathcal{NIW}(m_0, \Lambda_0, \psi_0, \nu_0) \end{aligned}$$

where :

- $\mathcal{N}(\cdot \mid \mu_i, \Sigma_i)$ is a bivariate normal distribution
- $\theta_i = (\mu_i, \Sigma_i) \in \mathbb{R}^2 \times \mathcal{S}_+^2$, with $\mathcal{S}_+^2$ the set of symmetric positive definite 2×2 covariance matrices
- G is a random probability measure over the parameter space θ, drawn from a **Dirichlet process**
- G_0 is the **base measure** following a normal-inverse-Wishart on θ_i.

Liste de papiers pour méthodes zoneless (à filtrer) :

- Woessner et al (2015) The 2013 European Seismic hazard model : key components and results.
- Petersen MD, Harmsen SC, Jaiswal KS, Rukstales KS, Luco N, Haller KM, Mueller CS, Shumway AM (2018) Seismic hazard, risk, and design for south America.
- Helmstetter A, Werner MJ (2012) Adaptive spatiotemporal smoothing of seismicity for long-term earthquake forecasts in California.
- Woo G (1996) Kernel estimation methods for seismic hazard area source modeling.
- S. Molina, C. Lindholm, H. Bungum (2001) Probabilistic seismic hazard analysis : zoning free versus zoning methodology.
- Chethanamba Kempanna Ramanna, G. Dodagoudar (2012) Probabilistic seismic hazard analysis using kernel density estimation technique for Chennai, India.
- S. Lasocki (2021) Kernel Density Estimation in Seismology
- M. Danese, M. Lazzari, B. Murgante (2008) Kernel Density Estimation Methods for a Geostatistical Approach in Seismic Risk Analysis : The Case Study of Potenza Hilltop Town (Southern Italy)
- C. Stock, Euan Smith (2002) Adaptive Kernel Estimation and Continuous Probability Representation of Historical Earthquake Catalogs
- C. Stock, Euan Smith (2002) Comparison of Seismicity Models Generated by Different Kernel Estimations
- G. Estévez-Pérez, H. L. Cimadevila, A. Quintela-del-Río (2002) Nonparametric analysis of the time structure of seismicity in a geographic region
- M. Crespo, F. Martínez, J. Martí (2014) Seismic hazard of the Iberian Peninsula : evaluation with kernel functions
- Francis Tong, Stanisław Lasocki, Beata Orlecka-Sikora (2025) Non-parametric kernel density estimation of magnitude distribution for the analysis of seismic hazard posed by anthropogenic seismicity
- Karaburun, A. ; Demirci, A. (2016) Spatio-temporal cluster analysis of the earthquake epicenters in Turkey and its surrounding area between 1900 and 2014
- Kernel Density Estimation for the Interpretation of Seismic Big Data in Tectonics Using QGIS : The Türkiye–Syria Earthquakes (2023)
-

Simulation de processus de Dirichlet

Simulation par Stick-Breacking

Générer une (approximation) de la densité sous la forme :

$$f(x) = \sum_{k=1}^{K} w_k \, \delta_{\theta_k}(x)$$

où : $\theta_k \sim G_0$.

Tronquer le modèle à un nb K fixé de composantes du mélange.

Input

- Nombre de composantes K
- Param de concentration $\alpha > 0$

Étapes :

1. Initialisation : Créer liste vide **poids** = **[]** et le reste du bâton : $r = 1.0$; Créer liste vide $\boldsymbol{\theta}$
2. Pour $k = 1$ à $K - 1$:
   - Tirer $v_k \sim \text{Beta}(1, \alpha)$
   - Calculer $w_k = v_k \cdot r$
   - Ajouter w_k à **poids**
   - Mise à jour du baton $r = r \cdot (1 - v_k)$
   - Simuler $\theta_k \sim G_0$
   - Ajouter θ_k à $\boldsymbol{\theta}$
3. Ajouter $w_K = r$ à **poids**
4. Simuler $\theta_K \sim G_0$ et l'ajouter à $\boldsymbol{\theta}$

Output : Approximation d'un DP avec : $\mathcal{P} = \sum_{k=1}^{K} w_k \delta_{\theta_k}$

Simulation par Stick-Breaking (avec seuil τ)

Générer une approximation de la densité sous la forme :

$$f(x) = \sum_{k=1}^{\infty} w_k \, \delta_{\theta_k}(x)$$

où $\theta_k \sim G_0$, et les poids sont générés par le procédé de Stick-Breaking.

Input :

- Param de concentration $\alpha > 0$
- Seuil $\tau > 0$

Étapes :

1. Initialisation :
   - Liste vide des poids : `poids = [ ]`
   - Liste vide des paramètres : $\boldsymbol{\theta} = []$
   - Reste du bâton : $r \leftarrow 1.0$
   - Indice : $k \leftarrow 1$
2. Tant que $r > \tau$, faire :
   - Tirer $v_k \sim \text{Beta}(1, \alpha)$
   - Calculer $w_k = v_k \cdot r$
   - Ajouter w_k à `poids`
   - Mettre à jour : $r \leftarrow r \cdot (1 - v_k)$
   - Simuler $\theta_k \sim G_0$
   - Ajouter θ_k à $\boldsymbol{\theta}$
   - Incrément $k \leftarrow k + 1$

Output : Approximation d'un DP avec :

$$\mathcal{P} = \sum_{k=1}^{K} w_k \, \delta_{\theta_k}(x), \quad \text{où } K \text{ est déterminé en fonction de } \tau$$

- G_0 encode les connaissances a priori globales. On pourrait utiliser un mélange sur G_0 :

$$G_0(\cdot) \;=\; \sum_{j=1}^{J} \omega_j \cdot G_{0,j}(\cdot) \;, \quad \text{avec } \omega_j \propto \text{connaissance sur zone } j$$

- Rendre G ou G_0 dépendant de la zone géographique :

$$G \;\sim\; DP(\alpha(s), G_0(s))$$

Intégrer l'information des zonages sismotectoniques sous forme de prior informatif

Nous avons accès à un nombre n de positions de séismes sur un lieu Ω :

$$x_1, \dots, x_n \quad \sim \quad f \quad (f \text{ densité})$$

où $x_i = (x_i^{(1)}, x_i^{(2)})$ pour tout $i \in [1, n]$.
Notre approche : Estimation bayésienne non paramétrique de f

$$\begin{cases} f(x) \;=\; \int \mathcal{N}(\mu, \Sigma) \;\; dG(\mu, \Sigma) \\ \qquad G \;\sim\; \mathrm{DP}(\alpha, G_0) \end{cases}$$

Autre formulation :

$$\begin{aligned} f(x) \;&=\; \sum_{k=1}^{\infty} w_k \, \mathcal{N}(\mu_k, \Sigma_k) \\ (w_k)_k \;&\sim\; \mathrm{SB}(\alpha) \\ (\mu_k, \Sigma_k)_k \;&\sim\; G_0 = \mathcal{N}(\mu_k \mid \mu_0, \frac{\Sigma_k}{\lambda_0}) \, \mathcal{IW}(\Sigma_k \mid \psi_0, \nu_0) \end{aligned}$$

But : Intégrer prior informatif du zonage sismotectonique

On note f_0 la densité de la distribution du zonage. on a :

$$f_0(x) \;=\; \frac{\sum_{j=1}^{J} w_{0,j} \; \mathbb{1}_{S_{0,j}}(x)}{\sum_{j=1}^{J} w_{0,j} \; A_{0,j}}$$

où $S_{0,1}, \dots, S_{0,J}$ est une partition de Ω et représente les zones d'un zonage sismotectonique et chaque $A_{0,j}$ est la surface de $S_{0,j}$.

Une idée serait d'utiliser des gaussiennes pour approcher les découpes du zonage avec $\mu_{0,j}$ des centroides des zones $S_{0.j}$ et $\Sigma_{0,j}$ des diamètres d'ellipses (?). On aurait :

$$\tilde{f}_0(x) \;=\; \frac{\sum_{j=1}^{J} w_{0,j} \; \mathcal{N}(\mu_{0,j}, \Sigma_{0,j})}{\sum_{j=1}^{J} w_{0,j} \; A_{0,j}}$$

Ainsi, on aurait la mesure de base a priori informative suivante :

$$G_0^{\inf}(\cdot) \;=\; \sum_{j=1}^{J} w_{0,j} \, \mathcal{N}(\cdot \mid \mu_{0,j}, \frac{\Sigma_j}{\lambda_0}) \, \mathcal{IW}(\cdot \mid \Sigma_{0,j}, \nu_0)$$

Première étape : Évaluation de la qualité de la version informative

---

On cherche à construire une densité spatiale sur la carte de France Ω à partir d'un zonage sismo, c'est-à-dire :

$$\int_{\text{France}} f_0(x,y)\,dx\,dy \quad = \quad 1$$

où $f_0(x,y)$ est constante sur chaque zone $S_{0,j}$.

Soit $\{S_{0,1},\ldots,S_{0,J}\}$ un zonage sismotectonique de la France Ω, tel que $\Omega = \bigcup_{j=1}^{J} S_{0,j}$, avec $S_{0,j} \cap S_{0,i} = \emptyset$ si $i \neq j$. Chaque $S_{0,j}$ a :

- une surface $A_{0,j} = \text{Surf}(S_{0,j})$
- une poids associé $w_{0,j} \geq 0$, avec : $\sum_{j=1}^{J} w_{0,j} \;=\; 1$

On peut définir :

$$f_0(x,y) = \sum_{j=1}^{J} \frac{w_{0,j}}{A_{0,j}} \cdot \mathbb{1}_{S_{0,j}}(x,y)$$

On a bien une densité spatiale sur la France.
Si on a une catégorisation de chaque zone selon un niveau de sismicité, on peut attribuer un facteur $\lambda_{0,j}$ proportionnel à chaque caté pour obtenir un poids normalisé comme suit :

$$w_{0,j} = \frac{\lambda_{0,j}}{\sum_j \lambda_{0,j}}$$

On a :

$$\begin{aligned}\int_{\Omega} f_0(x,y)\;dx\,dy \;&=\; \int_{\Omega} \sum_{j=1}^{J} \frac{w_{0,j}}{A_{0,j}} \cdot \mathbb{1}_{S_{0,j}}(x,y)\;dx\,dy \;=\; \sum_{j=1}^{J} \frac{w_{0,j}}{A_{0,j}} \cdot \int_{\Omega} \mathbb{1}_{S_{0,j}}(x,y)\;dx\,dy \\ &=\; \sum_{j=1}^{J} \frac{w_{0,j}}{A_{0,j}} \cdot A_{0,j} \;=\; \sum_{j=1}^{J} w_{0,j} \;=\; 1\end{aligned}$$

Donc $f_0(x,y)$ est bien une densité sur Ω.

La loi associée $\mathbb{P}_{X,Y} : \mathcal{B}(\mathbb{R}^2) \to [0,1]$ associée à cette densité serait donnée par :

$$\begin{aligned}\mathbb{P}_{X,Y}(B) \;&=\; \mathbb{P}((X,Y) \in B) \;=\; \int_B f_0(x,y)\;dx\,dy \\ &=\; \sum_{j=1}^{J} \frac{w_{0,j}}{A_{0,j}} \cdot \int_{B \cap S_{0,j}} dx\,dy \qquad \text{pour tout borélien } B\end{aligned}$$

La loi Normale-Inverse Wishart est une loi jointe sur : la moyenne d'une loi normale multivariée $\boldsymbol{\mu}$ et la matrice de covariance à cette loi normale multivariée $\boldsymbol{\Sigma}$. Cette loi est caractérisée par quatre hyperparamètres :

- $\boldsymbol{\mu_0}$ (vect de dim d) : la moyenne prior sur $\boldsymbol{\mu}$
- λ_0 (scalaire positif) : un facteur d'échelle sur la précision de la moyenne
- $\boldsymbol{\Psi_0}$ (matrice d x d, sym et def pos) : un paramètre d'échelle pour la matrice de covariance
- ν_0 (degré de liberté $>$ d - 1) : un paramètre qui contrôle la concentration de la loi Inverse Wishart sur $\boldsymbol{\Sigma}$

La NIW est donnée ainsi :

$$\boldsymbol{\Sigma} \sim \mathcal{IW}(\Psi_0, \nu_0), \quad \boldsymbol{\mu} \mid \boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{\mu_0}, \frac{\boldsymbol{\Sigma}}{\lambda_0})$$

où $\mathcal{IW}$ est la loi inverse Wishart et $\boldsymbol{\mu}$ suit une normale multivariée avec covariance $\boldsymbol{\Sigma}/\lambda_0$.

La deensité de l'Inverse Wishart est :

$$f(\boldsymbol{\Sigma} \mid \boldsymbol{\Psi_0}, \nu_0) = \frac{|\boldsymbol{\Psi_0}|^{\frac{\nu_0}{2}}}{2^{\frac{\nu_0 d}{2}} \Gamma_d\left(\frac{\nu_0}{2}\right)} |\boldsymbol{\Sigma}|^{-\frac{\nu_0+d+1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi_0}\boldsymbol{\Sigma}^{-1})\right)$$

avec $\Gamma_d(\cdot)$ la fonction gamma multivariée dim d, $|\cdot|$ le déterminant et tr la trace.

La densité de la loi normale conditionnelleest :

$$f\left(\boldsymbol{\mu} \mid \boldsymbol{\mu_0}, \frac{\boldsymbol{\Sigma}}{\lambda_0}\right) = \frac{1}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}/\lambda_0|^{\frac{1}{2}}} \exp\left(-\frac{\lambda_0}{2}(\boldsymbol{\mu} - \boldsymbol{\mu_0})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu_0})\right)$$

Donc la densité NIW est :

$$\begin{aligned} f(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= f(\boldsymbol{\Sigma} \mid \boldsymbol{\Psi_0}, \nu_0) \cdot f\left(\boldsymbol{\mu} \mid \boldsymbol{\mu_0}, \frac{\boldsymbol{\Sigma}}{\lambda_0}\right) \\ &= \frac{|\boldsymbol{\Psi_0}|^{\frac{\nu_0}{2}} \lambda_0^{\frac{d}{2}}}{(2\pi)^{\frac{d}{2}} 2^{\frac{\nu_0 d}{2}} \Gamma_d\left(\frac{\nu_0}{2}\right)} |\boldsymbol{\Sigma}|^{-\frac{\nu_0+d+2}{2}} \\ &\quad \times \exp\left(-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi_0}\boldsymbol{\Sigma}^{-1}) - \frac{\lambda_0}{2}(\boldsymbol{\mu} - \boldsymbol{\mu_0})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu_0})\right) \end{aligned}$$

$\Gamma_d(\cdot)$ est la fonction gamma multivariée dim d, définie par :

$$\Gamma_d(a) = \pi^{\frac{d(d-1)}{4}} \prod_{i=1}^{d} \Gamma\left(a + \frac{1-i}{2}\right)$$

Paramètre	**Effet/Interprétation**
$\boldsymbol{\mu_0}$	• Moyenne de la loi de $\boldsymbol{\mu}$ • Plus λ_0 est grand, plus $\boldsymbol{\mu}$ est concentré autour de $\boldsymbol{\mu_0}$ • Plus λ_0 est faible, $\boldsymbol{\mu}$ s'écarte de $\boldsymbol{\mu_0}$
λ_0	• C'est un facteur d'échelle sur la variance de $\boldsymbol{\mu}$ • Contrôle l'incertitude a priori sur la moyenne $\boldsymbol{\mu}$ • $\lambda_0 \to 0$ signifie une incertitude infinie sur $\boldsymbol{\mu}$
$\boldsymbol{\Psi_0}$	• Contrôle la taille et l'orientation moyennes des matrices de covariance $\boldsymbol{\Sigma}$ • Plus les valeurs propres de $\boldsymbol{\Psi_0}$ sont grandes, plus les réalisations de $\boldsymbol{\Sigma}$ sont grandes (variances plus larges) et/ou des corrélations plus marquées
ν_0	• Contrôle la concentration de la loi de $\boldsymbol{\Sigma}$ • Doit être strictement supérieur à $d-1$ pour que la moyenne existe • Plus ν_0 est grand, plus $\boldsymbol{\Sigma}$ est concentré autour de sa moyenne $\boldsymbol{\Psi_0}/(\nu_0 - d - 1)$ • Si ν_0 est faible (proche de d), la dispersion des matrices $\boldsymbol{\Sigma}$ est grande (forte incertitude)

On cherche à approximer la distance L^2 entre deux densités f et g de $\mathbb{R}^2$, définie par :

$$\|f-g\|_{L^2} = \left(\int_\Omega \left(f(x,y) - g(x,y)\right)^2 \, dx\, dy \right)^{1/2}$$

où Ω est domaine d'étude.

On peut utiliser l'approche des somme discrète pour approximer l'intégrale. On utilise la formule d'approximation :

$$\int_\Omega \left(f(x,y) - g(x,y)\right)^2 \, dx\, dy \quad \approx \quad \sum_{i,j} \left(f(x_i,y_j) - g(x_i,y_j)\right)^2 \cdot \Delta x \cdot \Delta y$$

où Δx, Δy sont les pas de grille.

- Deux tâches pour l'inférence :
  1. Inférer G (combien de clusters et caractéristiques des clusters)
  2. Inférer les observations x_i (dans quel cluster)
- Modèle conjugué $\longrightarrow$ Algo 1, 2 et 3 de Neal (2000) [Du Gibbs]

On cherche à approximer la densité moyenne d'un DPMM avec un prior informatif. L'approximation se fait par moyenne de Monte Carlo sur N densités générées. Chaque composante k du mélange est tirée sous une NIW :

$$(\mu_k, \Sigma_k) \sim \text{NIW}(\mu_0^{(j)}, \lambda_0, \Psi_0, \nu_0) \quad \text{où} \quad \mu_0^{(j)} \in \{[0.5, 0.5], [1.5, 0.5], [0.5, 1.5], [1.5, 1.5]\},$$

$$\Psi_0 = \begin{pmatrix} 0.26 & 0 \\ 0 & 0.26 \end{pmatrix}, \quad \lambda_0 = 50.0, \quad \text{et} \quad \nu_0 = 4.$$

Les poids du mélange sont générés via stick-breaking avec paramètre de concentration α et seuil de troncature τ :

$$v_k \sim \text{Beta}(1, \alpha), \qquad w_k = v_k \prod_{i=1}^{k-1}(1 - v_i)$$

$$\text{On arrête quand } \prod_{i=1}^{k}(1 - v_i) < \tau$$

Les poids sont normalisés ensuite : $\sum_{k=1}^{K} w_k = 1$

Chaque densité générée s'écrit comme un mélange de normales :

$$f^{(i)}(\cdot) = \sum_{k=1}^{K^{(i)}} w_k^{(i)} \cdot \mathcal{N}(\cdot \mid \mu_k^{(i)}, \Sigma_k^{(i)})$$

avec les composantes $(\mu_k^{(i)}, \Sigma_k^{(i)}) \sim \text{NIW}(\mu_0^{(j)}, \lambda_0, \Psi_0, \nu_0)$

Enfin, on calcule la moyenne empirique des densités :

$$\bar{f}_N(\cdot) = \frac{1}{N} \sum_{i=1}^{N} f^{(i)}(\cdot) = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K^{(i)}} w_k^{(i)} \cdot \mathcal{N}(\cdot \mid \mu_k^{(i)}, \Sigma_k^{(i)})$$

Formule complète de la densité d'un DPMM tronqué sur le carré $[0,2]^2$:

$$f(\cdot) = \sum_{k=1}^{K} w_k \cdot \mathcal{N}(\cdot \mid \mu_k, \Sigma_k) \cdot \frac{\mathbb{1}_{[0,2]^2}(\cdot)}{Z_k}$$

où : $Z_k = \int_{[0,2]^2} \mathcal{N}(u \mid \mu_k, \Sigma_k)\, du$

- *'define_zonage_grid(n_rows, n_cols, x_range, y_range)'* : Crée une partition régulière de l'espace $\Omega = [x_{\min}, x_{\max}] \times [y_{\min}, y_{\max}]$ en $J = n_{\text{rows}} \times n_{\text{cols}}$ sous-zones. On définit des rectangles $S_j = [x_j^{(0)}, x_j^{(1)}] \times [y_j^{(0)}, y_j^{(1)}] \subset \Omega$, pour $j = 1, \ldots, J$, tels que :

$$\bigcup_{j=1}^{J} S_j = \Omega \quad \text{et} \quad S_j \cap S_k = \emptyset \text{ pour } j \neq k$$

- *'compute_f0_density(x, y, zones, weights, areas)'* : Définit une densité par morceaux $f_0(x, y)$, constante sur chaque zone. Soient :
  - w_j un poids associé à la zone S_j
  - $A_j = \text{aire}(S_j)$
  - La densité est définie par :

$$f_0(x) \;=\; \frac{\sum_{j=1}^{J} w_j \; \mathbb{1}_{S_j}(x)}{\sum_{j=1}^{J} w_j \; A_j}$$

    et $f_0(x, y) = 0$ sinon.

- *'sample_from_f0(n, zones, weights, areas)'* : Génère des échantillons selon la loi f_0 définie plus haut.

  1. Tirer une zone S_j avec probabilité :

$$\mathbb{P}(S_j) = \frac{w_j A_j}{\sum_{k=1}^{J} w_k A_k}$$

  2. Tirer $(X, Y) \sim \mathcal{U}(S_j)$.

- *'compute_zone_gaussian_parameters(zones)'* : Approxime chaque zone S_j par une gaussienne centrée sur son centroïde avec une covariance isotrope. On a :
  - $\mu_j = \text{centre de } S_j = \left(\frac{x_j^{(0)} + x_j^{(1)}}{2}, \frac{y_j^{(0)} + y_j^{(1)}}{2} \right)$

– $\Sigma_j = \sigma_j^2 I_2$, où σ_j est proportionnel au rayon de la zone :

$$\sigma_j = \frac{\text{diam}(S_j)}{2 \cdot 1.96} \quad \Rightarrow \quad \Sigma_j = \sigma_j^2 \cdot I$$

- *'compute_f0tilde_density(x, y, mus, covariances, weights)'* : Calcul la densité d'un mélange de lois normales pondérées. On a :

$$\tilde{f_0}(x, y) = \sum_{j=1}^{J} w_j \cdot \mathcal{N}\left((x, y) \mid \mu_j, \Sigma_j\right)$$

où $\mathcal{N}(\cdot \mid \mu_j, \Sigma_j)$ est la densité de la gaussienne centrée en μ_j avec covariance Σ_j.

- *'sample_from_f0tilde(n, mus, covariances, weights, areas)'* : Génère des échantillons suivant la loi $\tilde{f_0}$, selon le processus suivant :

1. Tirer un indice $j \in \{1, \ldots, J\}$ avec probabilité :

$$\mathbb{P}(j) = \frac{w_j A_j}{\sum_{k=1}^{J} w_k A_k}$$

2. Tirer un échantillon $(X, Y) \sim \mathcal{N}(\mu_j, \Sigma_j)$

Développement calcul de la matrice de covariance Σ_j

Soit, chaque zone S_j est un rectangle de forme :

$$S_j = [x_j^{(0)}, x_j^{(1)}] \times [y_j^{(0)}, y_j^{(1)}]$$

On approxime la densité sur cette zone par une loi normale bidimensionnelle centrée au centroïde de la zone, et avec une matrice de covariance diagonale $\Sigma_j = \sigma_j^2 I$, où I est la matrice identité 2×2.

Le diamètre est la distance maximale entre deux points dans la zone rectangulaire S_j. Ici, comme la zone est rectangulaire, ce diamètre correspond à la diagonale :

$$\text{diam}(S_j) = \sqrt{(x_j^{(1)} - x_j^{(0)})^2 + (y_j^{(1)} - y_j^{(0)})^2}$$

L'enjeu ici est d'approximer la densité uniforme sur S_j par une densité normale $\mathcal{N}(\mu_j, \Sigma_j)$, et de calibrer Σ_j pour que cette gaussienne couvre 95% de la masse dans la zone. À cet objectif, on choisit un écart-type σ_j tel que le disque de rayon $\text{diam}(S_j)/2$ corresponde à environ 1.96 écarts-types (quantile pour avoir une confiance à 95% : $\mathbb{P}(1.96 < Z < 1.96) \approx 0.95$ pour $Z \sim \mathcal{N}(0, 1)$; quantile pour confiance à 99% 2.576).

Ainsi, on a :

$$\sigma_j = \frac{\text{diam}(S_j)}{2 \cdot 1.96}$$

On suppose que la covariance est isotrope, donc diagonale proportionnelle à l'identité : $\Sigma_j = \sigma_j^2 \cdot I$.

Autrement dit :

$$\Sigma_j = \left(\frac{\text{diam}(S_j)}{2 \cdot 1.96} \right)^2 \cdot I$$

Application dans cas jouet : Soit zone $S_j = [0,1] \times [0,1]$. On a :

- Le diamètre : $\text{diam}(S_j) = \sqrt{1^2 + 1^2} = \sqrt{2}$
- L'écart-type : $\sigma_j = \frac{\sqrt{2}}{2 \cdot 1.96}$
- La matrice de covariance est donc :

$$\Sigma_j = \left(\frac{\sqrt{2}}{2 \cdot 1.96} \right)^2 \cdot I \approx (0.255)^2 \cdot I \approx 0.065 \cdot I$$

Z loi normale standard : $\mathbb{P}(-1.96 \leq Z \leq 1.96) \approx 0.95$
95% de la masse d'une loi normale univariée est contenue dans l'intervalle $[-1.96, +1.96]$.

On veut approximer une densité uniforme sur une zone rectangulaire S_j par une densité normale centrée au centroïde de la zone. On veut que cette normale couvre environ 95%. On choisit alors un écart-type σ_j tel que :

$$\text{Rayon} = 1.96 \cdot \sigma_j \quad \Rightarrow \quad \sigma_j = \frac{\text{diam}(S_j)}{2 \cdot 1.96}$$

Code Julien - Étude de la capacité à bien approcher un pavage uniforme par des gaussiennes :

- *gmm_em(X, G, mu_init, sigma_init = NULL, max_iter = 100, tol = 1e-6) :* Estimation des paramètres d'un mélange de G lois normales multivariées à partir d'un jeu de données X, à l'aide de l'algo EM.

  Paramètres d'entrée :

  - X : matrice $n \times d$ de données (n observations, d variables)
  - G : nombre de composantes du mélange
  - mu_init : matrice initiale $G \times d$ des moyennes

- sigma_init : tableau $d \times d \times G$ des matrices de covariance initiales
- max_iter : nombre maximal d'itérations
- tol : tolérance pour l'arrêt basé sur la variation du log-vraisemblance

Grandes lignes :

1. Initialisation (n : nombre d'observations ; d : dim des données ; mu : initialisation des moyennes ; pi_k : proportions initiales (toutes égales)
2. Initialisation des matrices de covariance (Si sigma_init est NULL → utiliser la matrice de covariance globale de X pour toutes les composantes)
3. **Algo EM :** *Pour chaque itération jusqu'à max_iter :*
   - **E-Step :** Calcule γ_{ik}, i.e. la proba que l'observation x_i provienne de la composante k :
   $$\gamma_{ik} = \frac{\pi_k \cdot \mathcal{N}(x_i \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{G} \pi_j \cdot \mathcal{N}(x_i \mid \mu_j, \Sigma_j)}$$
   - **M-Step :** Met à jour les paramètres π_k, μ_k, Σ_k à partir de γ :
     * $N_k = \sum_{i=1}^{n} \gamma_{ik} \quad \longrightarrow \quad \pi_k = N_k/n$
     * $\mu_k = \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{ik} x_i$
     * $\Sigma_k = \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{ik}(x_i - \mu_k)(x_i - \mu_k)^T$
   - **Calcul log-vraisemblance :**
   $$\log L = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{G} \pi_k \cdot \mathcal{N}(x_i \mid \mu_k, \Sigma_k) \right)$$
   Si la variation de la log-vraisemblance est inférieure à *tol*, on stop.
4. La fonction renvoie une liste avec les paramètres estimés : $G, \pi_k, \mu, \sigma, \log L$

- *rgmm(n, prob, mean, sigma) :* Génère n points aléatoires selon un mélange normales multivariées
- *dgmm(x, prob, mean, sigma) :* Calcule la densité totale d'un mélange gaussien multivarié
- *d_mar_gmm(x, mar, prob, mean, sigma) :* Calcul la densité marginale (1d)
- *stick_breaking(alpha, tau = 1e-3) :* SB
- *g_0(prob, mean, sigma, nu_0, lambda_0) :* Tire un param de composante (μ, Σ) à partir d'un DP
- *sample_f(alpha, prob, mean, sigma, nu_0, lambda_0, tau = 1e-3) :* Simule DP

Il est important d'adopter une vision plus large et globale lorsque je code. Mon code doit être à la fois adaptable et facilement débogable. Il paraît pertinent de viser une uniformisation maximale dans la structure des fonctions. Par ailleurs, écrire mathématiquement ce que je veux implémenter, semble être une approche judicieuse (surtout d'un point de vue longtermiste).

We are pleased to present the Epos-France database, which includes more than 16,000 high-quality seismic records from seismological stations in mainland France. They correspond to earthquakes with local magnitude between 2 and 5.6 that occurred between 1996 and the end of 2021. (Sur site SIGMA https://sigma-programs.com/newsletter-1-february-2025-copy-2/)

Un processus de Dirichlet dépendant est une extension du DP qui permet à la distribution aléatoire de varier en fonction d'un ensemble de covariables ou d'indices (par exemple, l'espace, le temps, ou d'autres facteurs). Autrement dit, au lieu d'avoir une unique mesure aléatoire $G \sim DP(\alpha, G_0)$, on considère une famille de mesures $\{G_x : x \in \mathcal{X}\}$ indexées par x, où chaque G_x suit une distribution similaire à un DP, mais dépendant de x. Cela permet de modéliser la dynamique ou la variation locale des distributions. Par exemple, dans un contexte spatio-temporel, on pourrait avoir une distribution de points (ou d'événements) qui varie selon la localisation géographique ou le moment.

Modélisation adaptative de la densité spatiale : un DDP pourrait modéliser la distribution spatiale des événements de manière flexible, permettant à la forme de la densité de changer localement selon la position géographique ou d'autres covariables (ex. caractéristiques géologiques).

'compute_zone_gaussian_parameters' : Approxime mélange d'uniforme X par un mélange de gaussiennes initialisé avec k-means, afin d'estimer : les centroïdes et les covariances (et éventuellement les poids) de chaque composante gaussienne.

1. Initialisation centroïdes avec KMeans
2. Initialisation covariances : *GaussianMixture* accepte l'initialisation via précisions (inverses des covariances), d'où utilisation de *pinv* donne version robuste de l'inverse (pseudo-inverse de Moore-Penrose)

3. Entrainement GMM (par algo EM)
4. Renvoie des centroïdes et covariances estimés

Objectif : Construire une loi NIW$(\mu_0, \lambda_0, \Psi_0, \nu_0)$ telle que :

- la moyenne a priori de la distribution $\mu \sim \mathcal{N}(\mu_0, \Sigma/\lambda_0)$ soit proche d'un centroïde estimé μ^* (supposé connu à ce stade),
- la valeur centrale de $\Sigma \sim$ Inv-Wishart(Ψ_0, ν_0) soit proche d'une matrice covariance estimée Σ^* (supposée connue à ce stade).

Caractéristiques de la NIW : Soit $(\mu, \Sigma) \sim \text{NIW}(\mu_0, \lambda_0, \Psi_0, \nu_0)$, alors :

$$\Sigma \sim \mathcal{IW}(\Psi_0, \nu_0)$$
$$\mu \mid \Sigma \sim \mathcal{N}(\mu_0, \Sigma/\lambda_0)$$

Moyennes a priori :

$$\mathbb{E}[\mu] = \mu_0 \qquad ; \qquad \mathbb{E}[\Sigma] = \frac{\Psi_0}{\nu_0 - d - 1} \qquad \text{si } \nu_0 > d + 1 \text{ (en dim d = 2 ici)}$$

Construction à partir de μ^* et Σ^* - Étapes :

- Choisir $\mu_0 = \mu^*$, on place la moyenne de la NIW exactement sur l'estimation empirique du centroïde
- Fixer $\nu_0 > d + 1$ (e.g. $\nu_0 = d + 3 = 5$ pour rester modéré)
- Fixer $\Psi_0 = \Sigma^* \cdot (\nu_0 - d - 1)$. De sorte que :

$$\mathbb{E}[\Sigma] = \frac{\Psi_0}{\nu_0 - d - 1} = \Sigma^*$$

- Fixer λ_0 selon la confiance en la moyenne μ_0. Si on veut un prior fortement concentré autour de μ_0, on prend un λ_0 élevé. Si on veut queque chose de plus vague : $\lambda_0 \approx 0.01$.

Ainsi, on construit un prior NIW centré sur une gaussienne empirique, avec une force de concentration ajustable via α, λ_0 et ν_0.

Objective

Let $\mathcal{X} \subset \mathbb{R}^2$. We aim to estimate an unknown density $f : \mathcal{X} \to \mathbb{R}_+$ using a nonparametric Bayesian model based on a Dirichlet Process Mixture Model (DPMM), defined as :

$$f(x) = \int \mathcal{N}(x \mid \mu, \Sigma)\, dG(\mu, \Sigma)$$

where $G \sim \mathrm{DP}(\alpha, G_0)$ is a Dirichlet process with concentration parameter $\alpha > 0$ and base distribution G_0.

Stick-Breaking Approximation

The Dirichlet process is approximated using stick-breaking construction :

$$\begin{aligned}
\beta_k &\sim \mathrm{Beta}(1, \alpha) \\
\pi_k &= \beta_k \prod_{j=1}^{k-1} (1 - \beta_j), \quad k = 1, 2, \dots \\
\theta_k &\sim G_0, \quad \text{with } \theta_k = (\mu_k, \Sigma_k)
\end{aligned}$$

Truncating the process at a small threshold $\tau \ll 1$, we obtain :

$$G \approx \sum_{k=1}^{K} \pi_k \delta_{\theta_k}, \quad \text{with } \sum_{k=1}^{K} \pi_k \approx 1$$

Choice of the concentration parameter α

Work in progress ...

Choice of the Prior G_0

Non-informative Prior

A standard Normal-Inverse-Wishart prior can be used :

$$G_0 = \mathcal{NIW}(\mu_0, \lambda_0, \Psi_0, \nu_0)$$

which implies :

$$\begin{aligned}
\Sigma &\sim \mathcal{W}^{-1}(\Psi_0, \nu_0) \\
\mu \mid \Sigma &\sim \mathcal{N}(\mu_0, \Sigma/\lambda_0)
\end{aligned}$$

Informative Prior (Zonage)

In an informative prior setting, we exploit knowledge about the spatial structure of the domain. We divide the input space (in our simplified case we have taken $[0,2] \times [0,2]$) into a regular grid of $n_{\text{rows}} \times n_{\text{cols}}$ rectangular zones. Let the zones be denoted as $\{Z_j\}_{j=1}^{J}$ with corresponding areas A_j and weights w_j. We can define a zonal uniform density f_0 as :

$$f_0(x) = \sum_{j=1}^{J} \frac{w_j}{A_j} \mathbb{1}_{x \in Z_j}$$

From this piecewise-constant density f_0, we generate synthetic data and fit a Gaussian Mixture Model using an EM algo. Initialization is performed via KMeans to provide stable estimates of the component means. The fitted GMM yields a smoothed approximation $\tilde{f}_0$ of the zonal density :

$$\tilde{f}_0(x) = \sum_{j=1}^{J} \tilde{w}_j \cdot \mathcal{N}(x \mid \tilde{\mu}_j, \tilde{\Sigma}_j)$$

We then define an informative prior G_0^{inf} using the estimated GMM parameters :

- The component means $\tilde{\mu}_j$ are used as prior centers μ_j
- The covariance matrices $\tilde{\Sigma}_j$ are multiplied by $\nu_0 - d - 1$ to define the NIW scale matrices Ψ_j
- The estimated weights $\tilde{w}_j$ are used as discrete probas over the base

Thus, G_0^{inf} is a mixture of NIW components written as :

$$G_0^{\text{inf}} = \sum_{j=1}^{J} \tilde{w}_j \ \cdot \ \mathcal{NIW}(\mu_j, \lambda_0, \Psi_j, \nu_0)$$

This informative base measure allows the Dirichlet process to generate more plausible components in regions of high prior mass, effectively incorporating structured domain knowledge into the prior.

Posterior predictive density

The resulting posterior predictive density becomes :

$$f(x) \approx \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(x \mid \mu_k, \Sigma_k)$$

Cadre

Soit $(X_1, X_2, \ldots, X_d)$ un vecteur aléatoire de dimension d défini sur un espace $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ avec loi jointe $\pi(x_1, \ldots, x_d)$ connue à une constante multiplicative près :

$$\pi(x_1, \ldots, x_d) = \frac{\tilde{\pi}(x_1, \ldots, x_d)}{Z}$$

où $\tilde{\pi}$ est connue et $Z = \int_{\mathcal{X}} \tilde{\pi}(x)\, dx$ est inconnu.

Objectif : Générer une suite $\{X^{(t)}\}_{t \geq 0}$ telle que la loi stationnaire soit π, afin d'estimer des espérances

$$\mathbb{E}_\pi[h(X)] \;\approx\; \frac{1}{T} \sum_{t=1}^{T} h(X^{(t)}).$$

Idée clé du Gibbs sampling : Plutôt que d'échantillonner directement selon π, on utilise les lois conditionnelles complètes :

$$\pi(x_i \mid x_{-i}) \quad \text{où} \quad x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d).$$

Hypothèse essentielle :

$$\pi(x_i \mid x_{-i}) \text{ est simulable explicitement pour tout } i.$$

Algo de Gibbs

Soit $X^{(t)} = (X_1^{(t)}, \ldots, X_d^{(t)})$ l'état à l'itération t.

1. Initialisation : choisir $X^{(0)} \in \mathcal{X}$ arbitrairement (ou selon une loi de départ)
2. Itération : pour $t = 0, 1, \ldots$
   Pour $i = 1, \ldots, d$:
   $$X_i^{(t+1)} \sim \pi\big(x_i \,\big|\, X_1^{(t+1)}, \ldots, X_{i-1}^{(t+1)}, X_{i+1}^{(t)}, \ldots, X_d^{(t)}\big)$$

Autrement dit :

- On balaye les coordonnées une par une
- Pour la coordonnée i, on utilise les valeurs déjà mises à jour pour $1, \ldots, i-1$ et les valeurs de l'itération précédente pour $i+1, \ldots, d$

Propriétés théoriques

Chaîne de Markov

La suite $\{X^{(t)}\}$ est une chaîne de Markov sur $\mathcal{X}$ avec noyau de transition

$$K(x, dy) = \prod_{i=1}^{d} \pi\big(y_i \mid y_{1:i-1}, x_{i+1:d}\big)\, dy_i$$

où $y_{1:i-1} = (y_1, \ldots, y_{i-1})$ et $x_{i+1:d} = (x_{i+1}, \ldots, x_d)$.

Invariance

π est loi invariante pour K :

$$\int_{\mathcal{X}} \pi(x)\, K(x, dy)\, dx = \pi(dy).$$

Cela vient du fait que l'échantillonnage séquentiel selon les conditionnelles reproduit la loi jointe.

Réversibilité

Le noyau K est π-réversible :

$$\pi(dx)\, K(x, dy) = \pi(dy)\, K(y, dx).$$

Donc π est bien stationnaire.

Convergence

Si la chaîne est :

- irréductible (toute région de $\mathcal{X}$ atteignable)
- apériodique (si la chaîne n'évolue pas en cycles fixes)
- et positive récurrente (récurrente : la proba de revisiter chaque état est 1 ; positive récurrente : le temps moyen de retour de chaque état est fini)

alors :

$$\lim_{t\to\infty} \mathcal{L}(X^{(t)}) = \pi.$$

On a convergence en loi.

Version bloc (blocked Gibbs)

On peut regrouper certaines coordonnées en blocs $X = (Y_1, \ldots, Y_m)$ et mettre à jour chaque bloc selon sa loi conditionnelle jointe :

$$Y_j^{(t+1)} \sim \pi\big(y_j \,\big|\, Y_{-j}\big).$$

Avantage : réduit la corrélation entre itérations quand les variables sont fortement dépendantes.

Conjugaison de la NIW à gaussienne multivariée

On considère un vecteur aléatoire $\mathbf{x} \in \mathbb{R}^d$ tel que :

$$\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

On suppose un échantillon i.i.d. $\mathbf{x}_1, \ldots, \mathbf{x}_n$, et on cherche $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. On place une loi normale-inverse-Wishart a priori :

$$\begin{aligned} \boldsymbol{\Sigma} &\sim \mathcal{IW}(\nu_0, \boldsymbol{\Psi}_0) \\ \boldsymbol{\mu} \mid \boldsymbol{\Sigma} &\sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}/\lambda_0) \end{aligned}$$

Les densités sont :

$$\begin{aligned} p(\boldsymbol{\Sigma}) &\propto |\boldsymbol{\Sigma}|^{-(\nu_0+d+1)/2} \exp\left(-\frac{1}{2}\operatorname{Tr}(\boldsymbol{\Psi}_0\boldsymbol{\Sigma}^{-1})\right) \\ p(\boldsymbol{\mu} \mid \boldsymbol{\Sigma}) &\propto |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{\lambda_0}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\right) \end{aligned}$$

Donc :

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(\nu_0+d+2)/2} \exp\left(-\frac{1}{2}\left[\operatorname{Tr}(\boldsymbol{\Psi}_0\boldsymbol{\Sigma}^{-1}) + \lambda_0(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\right]\right)$$

La vraisemblance des données est :

$$p(\mathbf{x}_{1:n} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n} \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\right)$$

En posant $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ et $\mathbf{S} = \sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^\top$, on réécrit :

$$\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}) = \operatorname{Tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1}) + n(\bar{\mathbf{x}}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})$$

Ainsi, on a :

$$p(\mathbf{x}_{1:n} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-n/2} \exp\left(-\frac{1}{2}\left[\operatorname{Tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1}) + n(\bar{\mathbf{x}}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})\right]\right)$$

Calcul postérieur

Le calcul du posterior est :

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \mathbf{x}_{1:n}) \propto p(\mathbf{x}_{1:n} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \cdot p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(\nu_0+n+d+2)/2} \exp\left(-\frac{1}{2}Q(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1})\right)$$

avec :

$$Q(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}) = \operatorname{Tr}((\boldsymbol{\Psi}_0 + \mathbf{S})\boldsymbol{\Sigma}^{-1}) + \lambda_0(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0) + n(\boldsymbol{\mu}-\bar{\mathbf{x}})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\bar{\mathbf{x}})$$

En regroupan les termes quadratiques en $\boldsymbol{\mu}$, on a :

$$\lambda_n = \lambda_0 + n, \quad \boldsymbol{\mu}_n = \frac{\lambda_0\boldsymbol{\mu}_0 + n\bar{\mathbf{x}}}{\lambda_0 + n}$$

$$\lambda_0(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^2 + n(\boldsymbol{\mu}-\bar{\mathbf{x}})^2 = \lambda_n(\boldsymbol{\mu}-\boldsymbol{\mu}_n)^2 + \frac{\lambda_0 n}{\lambda_n}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)^2$$

En écrivant :

$$\begin{aligned}\boldsymbol{\Psi}_n &= \boldsymbol{\Psi}_0 + \mathbf{S} + \frac{\lambda_0 n}{\lambda_0+n}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)^\top \\ \nu_n &= \nu_0 + n\end{aligned}$$

En conclusion, on a :

$$\begin{aligned}p(\boldsymbol{\mu},\boldsymbol{\Sigma} \mid \mathbf{x}_{1:n}) &\propto |\boldsymbol{\Sigma}|^{-(\nu_n+d+2)/2}\exp\left(-\frac{1}{2}\left[\,\mathrm{Tr}(\boldsymbol{\Psi}_n\boldsymbol{\Sigma}^{-1}) + \lambda_n(\boldsymbol{\mu}-\boldsymbol{\mu}_n)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_n)\,\right]\right) \\ &\sim \mathcal{NIW}\left(\boldsymbol{\mu}_n,\lambda_n,\nu_n,\boldsymbol{\Psi}_n\right)\end{aligned}$$

Ainsi, laa loi NIW est conjuguée à la loi normale multivariée.

Loi Student

La loi de Student t est une loi de proba continue, symétrique autour de 0, similaire à une gaussienne, mais avec des queues plus épaisses. Elle apparaît naturellement dans le contexte suivant :

- On a un échantillon $X_1,\ldots,X_n$ tiré d'une population normale avec moyenne μ et variance σ^2.
- La variance σ^2 est inconnue.
- On veut construire une statistique centrée et normalisée :

$$t = \frac{\bar{X}-\mu}{S/\sqrt{n}}$$

où :

- $\bar{X}$: moyenne
- $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i-\bar{X})^2$: variance

Alors, cette statistique t suit exactement une loi de Student à $n-1$ degrés de liberté. La loi de Student t est caractérisée par :

1. ν : degrés de liberté ; Typiquement $\nu = n-1$ pour un échantillon de taille n. Plus ν est grand, plus la loi t ressemble à une normale standard.

2. Moyenne et variance : μ et σ^2 ; Moyenne $= 0$ si $\nu > 1$ et Variance $= \nu/(\nu-2)$ si $\nu > 2$

Les queues plus épaisses reflètent l'incertitude supplémentaire due à l'estimation de la variance à partir de l'échantillon. La loi t s'élargit pour refléter l'incertitude sur la variance. Si on avait la variance exacte, on aurait juste une normale. À partir d'une certaine valeur de degré de liberté, la student est assimilable à une gaussienne.

La densité de la loi t pour $x \in \mathbb{R}$ et ν degrés de liberté :

$$f(x|\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

où Γ est la fonction gamma.

Quand $\nu \to \infty$, t converge vers la normale standard $N(0,1)$.

Apparition de la Student

On considère :

$$\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad (\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \mathcal{NIW}(\boldsymbol{\mu}_0, \lambda_0, \boldsymbol{\Psi}_0, \nu_0)$$

On veut la distribution prédictive marginale $p(\mathbf{x})$, c'est-à-dire la distribution de $\mathbf{x}$ après avoir intégré $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$$p(\mathbf{x}) = \int \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\, \mathcal{NIW}(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{\mu}_0, \lambda_0, \boldsymbol{\Psi}_0, \nu_0)\, d\boldsymbol{\mu}\, d\boldsymbol{\Sigma}$$

Intégration sur $\boldsymbol{\mu}$ conditionnellement à $\boldsymbol{\Sigma}$

On a :

$$\boldsymbol{\mu} \mid \boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}/\lambda_0)$$

Du couo, la convolution avec la normale de $\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}$ donne :

$$\mathbf{x} \mid \boldsymbol{\Sigma} \;\sim\; \mathcal{N}(\boldsymbol{\mu}_0, (1+1/\lambda_0)\boldsymbol{\Sigma}) \;=\; \mathcal{N}\left(\boldsymbol{\mu}_0, \frac{\lambda_0+1}{\lambda_0}\boldsymbol{\Sigma}\right)$$

Intégration sur $\boldsymbol{\Sigma}$

On a :

$$\boldsymbol{\Sigma} \sim \mathcal{IW}(\nu_0, \boldsymbol{\Psi}_0)$$

La distribution de $\mathbf{x}$ après avoir intégré $\boldsymbol{\Sigma}$ devient une Student-t multivariée :

$$\mathbf{x} \sim t_{\nu_0-d+1}(\boldsymbol{\mu}_0, \frac{\lambda_0+1}{\lambda_0(\nu_0-d+1)}\boldsymbol{\Psi}_0)$$

avec :

- $\nu = \nu_0 - d + 1$: degrés de liberté
- $\boldsymbol{\mu}_0$: centre de la Student-t
- $\boldsymbol{\Sigma}_t = \frac{\lambda_0+1}{\lambda_0(\nu_0-d+1)}\boldsymbol{\Psi}_0$: matrice de dispersion

Dans l'algorithme Gibbs pour l'inférence du DPMM :

- Pour un nouveau cluster, la predictive est donnée par :

$$\mathbf{x}_i \mid \boldsymbol{\mu_0}, \lambda_0, \boldsymbol{\Phi_0}, \nu_0 \quad \sim \quad t_{\nu_0-d+1}(\boldsymbol{\mu}_0, \frac{\lambda_0+1}{\lambda_0(\nu_0-d+1)}\boldsymbol{\Psi}_0)$$

C'est ce qu'on utilise pour calculer les poids prédictifs t_{new}.

On veut la distribution prédictive $p(\mathbf{x} \mid G_0^{inf})$, c'est-à-dire la distribution de $\mathbf{x}$ après avoir intégré $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$$p(\mathbf{x}) = \int \mathcal{N}\Big(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\Big) \cdot \sum_{j=1}^{J} \tilde{w}_j \, \mathcal{NIW}\Big(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{\mu}_j, \lambda_j, \boldsymbol{\Psi}_j, \nu_j\Big) \, d\boldsymbol{\mu} \, d\boldsymbol{\Sigma}$$

On décompose :

$$\mathcal{NIW}(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{\mu}_j, \lambda_j, \boldsymbol{\Psi}_j, \nu_j) = \mathcal{N}\Big(\boldsymbol{\mu} \mid \boldsymbol{\mu}_j, \tfrac{1}{\lambda_j}\boldsymbol{\Sigma}\Big) \cdot \mathcal{IW}(\boldsymbol{\Sigma} \mid \boldsymbol{\Psi}_j, \nu_j).$$

Les densités étants positives par Fubini-Tonelli, on échange somme et intégrale :

$$p(\mathbf{x}) = \sum_{j=1}^{J} \tilde{w}_j \int\int \mathcal{N}\Big(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\Big) \, \mathcal{N}\Big(\boldsymbol{\mu} \mid \boldsymbol{\mu}_j, \tfrac{1}{\lambda_j}\boldsymbol{\Sigma}\Big) \, \mathcal{IW}(\boldsymbol{\Sigma} \mid \boldsymbol{\Psi}_j, \nu_j) \;\; d\boldsymbol{\mu} \, d\boldsymbol{\Sigma}.$$

Intégration sur $\boldsymbol{\mu}$

Conditionnellement à $\boldsymbol{\Sigma}$, on a (résultat à étayer) :

$$\int \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \, \mathcal{N}\Big(\boldsymbol{\mu} \mid \boldsymbol{\mu}_j, \tfrac{1}{\lambda_j}\boldsymbol{\Sigma}\Big) \, d\boldsymbol{\mu} \;=\; \mathcal{N}\Big(\mathbf{x} \mid \boldsymbol{\mu}_j, \; \tfrac{\lambda_j+1}{\lambda_j}\,\boldsymbol{\Sigma}\Big).$$

Donc :

$$p(\mathbf{x}) \;=\; \sum_{j=1}^{J} \tilde{w}_j \int \mathcal{N}\Big(\mathbf{x} \mid \boldsymbol{\mu}_j, \; \tfrac{\lambda_j+1}{\lambda_j}\,\boldsymbol{\Sigma}\Big) \, \mathcal{IW}(\boldsymbol{\Sigma} \mid \boldsymbol{\Psi}_j, \nu_j) \; d\boldsymbol{\Sigma}$$

Intégration sur $\boldsymbol{\Sigma}$

L'intégrale d'une normale conditionnelle en $\boldsymbol{\Sigma}$ contre une inverse-Wishart donne une Student multivariée (résultat à étayer) :

$$\int \mathcal{N}\Big(\mathbf{x} \mid \boldsymbol{\mu}_j,\ \tfrac{\lambda_j+1}{\lambda_j}\,\boldsymbol{\Sigma}\Big)\,\mathcal{IW}(\boldsymbol{\Sigma} \mid \boldsymbol{\Psi}_j, \nu_j)\,d\boldsymbol{\Sigma} \;=\; t_{\nu_j-d+1}\Big(\mathbf{x} \,\Big|\, \boldsymbol{\mu}_j,\ \frac{\lambda_j+1}{\lambda_j\,(\nu_j-d+1)}\,\boldsymbol{\Psi}_j\Big), \qquad \nu_j > d-1$$

Résultat

Au final, on a :

$$\boxed{p(\mathbf{x}) \;=\; \sum_{j=1}^{J} \tilde{w}_j\; t_{\nu_j-d+1}\Big(\mathbf{x} \,\Big|\, \boldsymbol{\mu}_j,\ \tfrac{\lambda_j+1}{\lambda_j(\nu_j-d+1)}\,\boldsymbol{\Psi}_j\Big).}$$

Dans l'algorithme Gibbs pour l'inférence du DPMM :

- Pour un nouveau cluster, la predictive est donnée par :

$$\mathbf{x}_i \mid \boldsymbol{\mu_0}, \lambda_0, \boldsymbol{\Psi_0}, \nu_0 \;\sim\; \sum_{j=1}^{J} \tilde{w}_j\; t_{\nu_j-d+1}\Big(\mathbf{x} \,\Big|\, \boldsymbol{\mu}_j,\ \tfrac{\lambda_j+1}{\lambda_j(\nu_j-d+1)}\,\boldsymbol{\Psi}_j\Big)$$

C'est ce qu'on utilise pour calculer les poids prédictifs t_{new}.

Modèle d'étude

On observe $x_1, \ldots, x_n \in \mathbb{R}^2$.

$$\begin{aligned}
x_i \mid \theta_i &\sim F(x_i \mid \theta_i)\ , \qquad F(\cdot \mid \theta_i) \;=\; \sum_{k=1}^{\infty} \pi_k\ \mathcal{N}(\cdot \mid \mu_k, \Sigma_k) \\
\theta_i \mid G &\overset{i.i.d.}{\sim} G \\
G &\sim \mathrm{DP}(\alpha, G_0) \\
G_0 &= \mathcal{NIW}(m_0, \kappa_0, \Psi_0, \nu_0)
\end{aligned}$$

Dans le cas prior informatif :

$$G_0^{inf} \;=\; \sum_{j=1}^{J} \omega_j\ \mathcal{NIW}(m_{0j}, \kappa_{0j}, \Psi_{0j}, \nu_{0j})$$

Reformulation avec variables latentes :

$$
\begin{aligned}
x_i \mid \phi, c_i &\sim F(x_i \mid \phi_{c_i}) \,, \qquad F(\cdot \mid \phi_{c_i}) = \sum_{k=1}^{\infty} \pi_k \, f(\cdot \mid \phi_{c_k}) \\
c_i \mid \boldsymbol{\pi} &\sim \text{Discrete}(\pi_1, \ldots, \pi_K) \\
\phi_c &\sim G_0 \\
\pi_k &= \beta_k \prod_{l=1}^{k-1} (1 - \beta_l) = \beta_k \left(1 - \sum_{l=1}^{k-1} \pi_l\right) \\
\beta_k &\sim \text{Beta}(1, \alpha)
\end{aligned}
$$

On a :

- $c_i \in \{1, 2, \dots\}$: indice du cluster auquel appartient x_i.
- (μ_k, Σ_k) : paramètres du cluster k

Éléments de l'algorithme de Gibbs

(I) Échantillonnage du cluster c_i

- **Pour chaque cluster existant k avec $i-1$ observations.** On pose les stats suffisantes suivantes : $\bar{x}$ et $S = \sum_{j=1}^{n} (x_j - \bar{x})(x_j - \bar{x})^\top$. On a

$$
p\Big(c_i = k \mid c_{-i}, \mu_k, \Sigma_k, \alpha, x_i\Big) \quad \propto \quad n_{k,-i} \;\cdot\; p(x_i \mid \mu_k, \Sigma_k) \tag{1}
$$

  où $p(x_i \mid \mu_k, \Sigma_k)$ est la vraisemblance :

$$
p(x_i \mid \mu_k, \Sigma_k) \quad = \quad \mathcal{N}\Big(x_i \mid \mu_k,\; \Sigma_k\Big)
$$

- **(Non informatif) Pour un nouveau cluster**, on a :

$$
p\Big(c_i \neq c_j \,,\; \forall\, j \neq i \;\mid\; c_{-i}, \alpha, x_i, \boldsymbol{\mu_0}, \lambda_0, \boldsymbol{\Phi_0}, \nu_0\Big) \quad \propto \quad \alpha \;\cdot\; p(x_i \mid G_0) \tag{2}
$$

  $p(x_i \mid G_0)$ est la prior predictive où :

$$
p(x_i \mid G_0) \quad = \quad t_{\nu_0 - d + 1}(\boldsymbol{\mu}_0, \frac{\lambda_0 + 1}{\lambda_0(\nu_0 - d + 1)} \boldsymbol{\Psi}_0)
$$

- **(Informatif) Pour un nouveau cluster**, on a :

$$
p\Big(c_i \neq c_j \,,\; \forall\, j \neq i \;\mid\; c_{-i}, \alpha, x_i, \boldsymbol{\mu_0}, \lambda_0, \boldsymbol{\Phi_0}, \nu_0\Big) \quad \propto \quad \alpha \;\cdot\; p(x_i \mid G_0) \tag{3}
$$

  $p(x_i \mid G_0)$ est la prior predictive où :

$$
p(x_i \mid G_0) \quad = \quad \sum_{j=1}^{J} \tilde{w_j} \;\cdot\; t_{\nu_j - d + 1}\Big(\mathbf{x} \;\mid\; \boldsymbol{\mu}_j,\; \tfrac{\lambda_j + 1}{\lambda_j(\nu_j - d + 1)} \boldsymbol{\Psi}_j\Big)
$$

Par ailleurs, pour $x \in \mathbb{R}^d$, $\nu > 0$ degrés de liberté, un centre $\mu \in \mathbb{R}^d$ et une matrice d'échelle $\Sigma \in \mathbb{R}^{d \times d}$, la densité de la Student multivariée $t_\nu(\mu, \Sigma)$ est :

$$t_\nu(x \mid \mu, \Sigma) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{d/2}\,|\Sigma|^{1/2}} \left[1 + \frac{1}{\nu}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right]^{-\frac{\nu+d}{2}}.$$

où $\Gamma(\cdot)$ est la fonction gamma, μ est le vecteur moyenne et la matrice de variance-covariance est $\mathrm{Cov}[X] = \frac{\nu}{\nu-2}\,\Sigma$.

(II) Échantillonnage des paramètres de cluster (μ_k, Σ_k)

Supposons un cluster $c_i = k$. On pose les stats suffisantes suivantes :

- $n_k = |c_i|$
- $\bar{x}_k = \frac{1}{n_k} \sum_{x \in c_i} x$
- $S_k = \sum_{x \in c_i} (x - \bar{x}_i)(x - \bar{x}_i)^\top$.

En écrivant :

$$\begin{aligned}
\lambda_k &= \lambda_0 + n_k \\
\boldsymbol{\mu}_k &= \frac{\lambda_0 \boldsymbol{\mu}_0 + n_k \bar{x}_k}{\lambda_0 + n_k} \\
\nu_k &= \nu_0 + n_k \\
\Psi_k &= \Psi_0 + S_k + \frac{\lambda_0 n_k}{\lambda_0 + n_k}(\bar{x}_k - \boldsymbol{\mu}_0)(\bar{x}_k - \boldsymbol{\mu}_0)^\top
\end{aligned}$$

Le posterior est donc donné par :

$$\boldsymbol{\Sigma_k} \sim \mathcal{IW}(\nu_k, \boldsymbol{\Psi_k}), \qquad \boldsymbol{\mu}_k \mid \boldsymbol{\Sigma_k} \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma_k}/\lambda_k) \tag{4}$$

(III) Échantillonnage du paramètre de concentration α

Work in progress ...

Algorithm 1 Gibbs pour DPMM

Input :
  Observations $X = \{x_i\}_{i=1\dots n} \subset \mathbb{R}^2$
  Hyperparamètres : $\alpha; (\boldsymbol{\mu_0}, \lambda_0, \boldsymbol{\Psi_0}, \nu_0)$
  Iterations T, burn-in B ?, thinning L ?

Initialisation :
  Assigner aléatoirement chaque x_i à un cluster
  Calculer stats suffisantes $\{n_k, \bar{x}_k, S_k\}$ pour chaque cluster non vide
  $K \leftarrow$ nombre de clusters non vides

for $t = 1$ **to** T **do**

  (I) Mise à jour des clusters $c_i = k$:

  for $i = 1$ **to** n **do**
    Retirer x_i de son cluster k_{old} : $n_{k_{\text{old}}} \leftarrow n_{k_{\text{old}}} - 1$
    if $n_{k_{\text{old}}} = 0$ **then**
      Supprimer le cluster k_{old} : $K \leftarrow K - 1$
    end if
    Pour chaque cluster existant $k = 1, \dots, K$ **:**
      Calculer
    Pour un nouveau cluster :
      $t_{\text{new}} \leftarrow$
  end for
end for