Lois de probabilité issues de gaussiennes réitérées

Souad Elotmani, Armand Maul

Lois de probabilité issues de gaussiennes réitérées

Souad El Otmani & Armand Maul

Université Paul Verlaine-Metz, LMAM, CNRS UMR 7122, Ile du Saulcy, METZ, F-57045, France

Résumé

Probability distributions arising from nested Gaussians. We consider a random sample $X_1, \ldots, X_n$ of size $n \geq 1$ from an $\mathcal{N}(\mu, \sigma_1^2)$ Gaussian law. Then, conditionnaly on each $X_i,\ i = 1, \ldots, n$, we define a new random sample $X_{i,1}, \ldots, X_{i,n}$ from the $\mathcal{N}(X_i, \sigma_2^2)$ normal distribution ($\mathcal{N}(X_i, \sigma_2^2)$ is notation introduced for convenience). Assuming that the so obtained n new random samples are conditionnaly independent, we get a second step randomly generated set of points. The question is to investigate the properties of this set. We give a theorem precising the limiting density obtained when n approaches infinity, and we generalize this theorem by studying what occurs when repeating this process until, conditionnaly on each $X_{i_1,i_2,\ldots,i_{p-1}},\ i_1 = 1, \ldots, n_1, i_2 = 1, \ldots, n_2, \ldots, i_{p-1} = 1, \ldots, n_{p-1}$, we get new random samples $X_{i_1,i_2,\ldots,i_p},\quad i_p = 1, \ldots, n_p$, from the $\mathcal{N}(X_{i_1,i_2,\ldots,i_{p-1}}, \sigma_p^2)$ normal distribution.

On considère un échantillon aléatoire $X_1, \ldots, X_n$ suivant la loi normale $\mathcal{N}(\mu, \sigma_1^2)$, de taille $n \geq 1$. Conditionnellement à chaque $X_i,\ i = 1, \ldots, n$, on définit un nouvel échantillon aléatoire $X_{i,1}, \ldots, X_{i,n}$ suivant la loi normale $\mathcal{N}(X_i, \sigma_2^2)$ ($\mathcal{N}(X_i, \sigma_2^2)$ est une notation introduite par commodité). Sous l'hypothèse que les n nouveaux échantillons aléatoires ainsi obtenus sont conditionnellement indépendants, on obtient un ensemble de points aléatoires de seconde génération. La question est d'étudier les propriétés de cet ensemble. On donne un théorème précisant la densité limite obtenue lorsque n tend vers l'infini, et on généralise ce théorème en étudiant ce qui se produit lorsque que l'on répète cette procédure jusqu'à obtenir, conditionnellement à chaque $X_{i_1,i_2,\ldots,i_{p-1}},\ i_1 = 1, \ldots, n_1, i_2 = 1, \ldots, n_2, \ldots, i_{p-1} = 1, \ldots, n_{p-1}$, de nouveaux échantillons aléatoires $X_{i_1,i_2,\ldots,i_p},\quad i_p = 1, \ldots, n_p$ suivant la loi normale $\mathcal{N}(X_{i_1,i_2,\ldots,i_{p-1}}, \sigma_p^2)$.

Mots clés : Mélanges gaussiens, lois limites, densités limites, loi normale.

1 Un résultat préliminaire

Soient $\mu \in \mathbb{R}$ et $\sigma_1 > 0$, $\sigma_2 > 0$. On considère un échantillon aléatoire $X_1, \ldots, X_n$ de taille $n \geq 1$ issu d'une loi gaussienne $\mathcal{N}(\mu, \sigma_1^2)$. Conditionnellement à chaque $X_i,\ i = 1, \ldots, n$, définissons un nouvel échantillon aléatoire $X_{i,1}, \ldots, X_{i,n}$ issu d'une loi gaussienne $\mathcal{N}(X_i, \sigma_2^2)$.

Si nous supposons que les n nouveaux échantillons aléatoires ainsi obtenus sont conditionnellement indépendants, nous obtenons un ensemble de n^2 points de seconde génération. Nous prouvons le résultat suivant, relatif à ce nouvel ensemble de points.

Theorem 1 *Notons x_i une réalisation de X_i, $i = 1, \ldots, n$. L'ensemble généré à la seconde génération est, conditionnellement à l'échantillon initial, un mélange Gaussien (Arora et Kannan (2001), Dasgupta (1999), Everitt et Hand (1981), McLachlan et Peel (2000)) dont la densité conditionnelle, $X_1, \ldots, X_n$ étant donnés, est de la forme*

$$h_n(x) = \frac{1}{n} \frac{1}{\sqrt{2\pi}\sigma_2} \sum_{i=1}^{n} e^{-\frac{(x-x_i)^2}{2\sigma_2^2}},$$

et on a

$$\lim_{n\to\infty} h_n(x) = \lim_{n\to\infty} \left[\frac{1}{n} \frac{1}{\sqrt{2\pi}\sigma_2} \sum_{i=1}^{n} e^{-\frac{(x-x_i)^2}{2\sigma_2^2}} \right] = \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} e^{-\frac{(x-\mu)^2}{2(\sigma_1^2+\sigma_2^2)}},$$

qui est la densité au point x de la gaussienne de moyenne μ et de variance $\sigma_1^2 + \sigma_2^2$.

Preuve du Théorème 1 Par souci de simplicité, nous prouvons ce résultat pour $\mu = 0$ et $\sigma_1 = \sigma_2 = 1$. Pour des valeurs quelconques de μ, σ_1 et σ_2, la démonstration est identique avec des changements de variables évidents dans les intégrales ci-dessous.

Ainsi, considérons des réalisations $x_1, \ldots, x_n$ de $X_1, \ldots, X_n$ (X suit la gaussienne $\mathcal{N}(0,1)$) et, pour x fixé, définissons $y_i = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-x)^2}{2}}$. Alors, les y_i sont des réalisations de la variable aléatoire $\varphi(X)$, où $\varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(t-x)^2}{2}}$.

Posons $\overline{y_n} = (y_1 + \ldots + y_n)/n$. En vertu de la loi des grands nombres, on a $\lim_{n\to\infty} \overline{y_n} = E(\varphi(X))$.

Calculons $E(\varphi(X))$. on a $E(\varphi(X)) = \int_{\mathbb{R}} \varphi(t) f(t) dt = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-\frac{(t-x)^2}{2}} e^{-\frac{t^2}{2}} dt.$

Posant $s = t - \frac{x}{2}$, on obtient

$$\begin{aligned}
\lim_{n\to\infty} \overline{y_n} &= \frac{1}{2\pi} \int_{\mathbb{R}} e^{-\frac{(s-\frac{x}{2})^2}{2}} e^{-\frac{(s+\frac{x}{2})^2}{2}} ds = \frac{1}{2\pi} \int_{\mathbb{R}} e^{\frac{(-s^2+sx-x^2/4-s^2-sx-x^2/4)}{2}} ds, \\
&= \frac{1}{2\pi} \int_{\mathbb{R}} e^{-s^2} e^{-\frac{x^2}{4}} ds = \frac{1}{2\pi} e^{-\frac{x^2}{4}} \int_{\mathbb{R}} e^{-s^2} ds = \frac{1}{\sqrt{2\pi}} \frac{\sqrt{2}}{2} e^{-\frac{x^2}{4}}.
\end{aligned}$$

En conséquence, nous venons de prouver que lorsque n tend vers ∞, l'ensemble généré à la seconde génération peut être considéré comme constitué de réalisations d'une variable aléatoire qui suit la distribution gaussienne $\mathcal{N}(0,2)$.

■

2 Un théorème général

Soient $\mu \in \mathbb{R}$ et $\sigma_1 > 0, \ldots, \sigma_p > 0$. Considérons un échantillon aléatoire $X_1, \ldots, X_{n_1}$ de taille $n_1 \geq 1$, et issu d'une distribution gaussienne $\mathcal{N}(\mu, \sigma_1^2)$. Alors, conditionnellement à chaque $X_{i_1},\ i_1 = 1, \ldots, n_1$, définissons un nouvel échantillon aléatoire $X_{i_1,1}, \ldots, X_{i_1,n_2}$ issu de la gaussienne $\mathcal{N}(X_{i_1}, \sigma_2^2)$ (avec les mêmes notations que précédemment). Supposant que les n_1 nouveaux échantillons aléatoires ainsi obtenus sont conditionnellement indépendants, on obtient un ensemble de seconde génération, composé de $n_1 n_2$ points généré aléatoirement. Nous répétons cette opération jusqu'à ce que, conditionnellement à chaque $X_{i_1,i_2,\ldots,i_{p-1}},\ i_1 = 1, \ldots, n_1, i_2 = 1, \ldots, n_2, \ldots, i_{p-1} = 1, \ldots, n_{p-1}$, nous ayons défini un nouvel échantillon aléatoire $X_{i_1,i_2,\ldots,i_p},\quad i_p = 1, \ldots, n_p$, issu de la gaussienne $\mathcal{N}(X_{i_1,i_2,\ldots,i_{p-1}}, \sigma_p^2)$. Supposant que les $n_1 n_2 \ldots n_{p-1}$ nouveaux échantillons aléatoires sont conditionnellement indépendants, nous obtenons un ensemble de $p$ième génération, constitué de $n_1 n_2 \ldots n_p$ points générés aléatoirement. Nous prouvons le résultat suivant, relatif à cet ensemble de points.

Theorem 2 *Si x désigne une réalisation de X, l'ensemble généré aléatoirement après l'étape p est un mélange gaussien de densité conditionnelle*

$$h_{n_1,n_2,\ldots,n_{p-1}}(x) = \frac{1}{n_1 n_2 \ldots n_{p-1}} \frac{1}{\sqrt{2\pi}\sigma_{n_p}} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \cdots \sum_{i_{p-1}=1}^{n_{p-1}} e^{-\frac{(x - x_{i_1,i_2,\ldots,i_{p-1}})^2}{2\sigma_p^2}}.$$

Faisons tendre $n_1, n_2, \ldots, n_{p-1}$ vers ∞. On a

$$\lim_{n_1,n_2,\ldots,n_{p-1}\to\infty} h_{n_1,n_2,\ldots,n_{p-1}}(x) = \frac{1}{\sqrt{2\pi \sum_{i=1}^{p} \sigma_i^2}} e^{-\frac{(x-\mu)^2}{2\sum_{i=1}^{p} \sigma_i^2}}.$$

Preuve du théorème 2 L'expression de la densité conditionnelle est claire dans la mesure où l'ensemble de nombres de $p$ième génération est un mélange gaussien. Aussi nous employons-nous à prouver la seconde équation du théorème 2.

Si $p = 3$, la densité conditionnelle du mélange gaussien est

$$h_{n_1,n_2}(x) = \frac{1}{n_1 n_2} \frac{1}{\sqrt{2\pi}\sigma_3} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} e^{-\frac{(x - x_{i,j})^2}{2\sigma_3^2}}.$$

Lorsque n_1 et n_2 tendent vers ∞, on a

$$\lim_{n_1,n_2\to\infty} h_{n_1,n_2}(x) = \lim_{n_1\to\infty} \frac{1}{n_1} \sum_{i=1}^{n_1} \Big[\lim_{n_2\to\infty} \Big(\frac{1}{n_2} \frac{1}{\sqrt{2\pi}\sigma_3} \sum_{j=1}^{n_2} e^{-\frac{(x-x_{i,j})^2}{2\sigma_3^2}} \Big) \Big].$$

D'après le théorème 1, cette équation s'écrit

$$\lim_{n_1,n_2\to\infty} h_{n_1,n_2}(x) = \lim_{n_1\to\infty} \frac{1}{n_1} \sum_{i=1}^{n_1} \Big[\frac{1}{\sqrt{2\pi(\sigma_2^2+\sigma_3^2)}} e^{-\frac{(x-x_i)^2}{2(\sigma_2^2+\sigma_3^2)}} \Big] = \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2+\sigma_3^2)}} e^{-\frac{(x-\mu)^2}{2(\sigma_1^2+\sigma_2^2+\sigma_3^2)}}$$

En conséquence, pour tout $p \in \mathbb{N}$, on a

$$\begin{aligned}
&\lim_{n_1,n_2,\ldots,n_{p-1}\to\infty} h_{n_1,n_2,\ldots,n_{p-1}}(x) = \\
&= \lim_{n_1\to\infty} \frac{1}{n_1} \sum_{i_1=1}^{n_1} \Big[\lim_{n_2\to\infty} \frac{1}{n_2} \sum_{i_2=1}^{n_2} \Big[\ldots \frac{1}{n_{p-2}} \sum_{i_{p-2}=1}^{n_{p-2}} \Big[\lim_{n_{p-1}\to\infty} \frac{1}{n_{p-1}} \frac{1}{\sqrt{2\pi}\sigma_p} \sum_{i_{p-1}=1}^{n_{p-1}} e^{-\frac{(x-x_{i_1,i_2,\ldots,i_{p-1}})^2}{2\sigma_p^2}} \Big] \\
&= \lim_{n_1\to\infty} \frac{1}{n_1} \sum_{i_1=1}^{n_1} \Big[\lim_{n_2\to\infty} \frac{1}{n_2} \sum_{i_2=1}^{n_2} \Big[\ldots \frac{1}{n_{p-2}} \sum_{i_{p-2}=1}^{n_{p-2}} \Big[\frac{1}{\sqrt{2\pi(\sigma_{p-1}^2+\sigma_p^2)}} e^{-\frac{(x-x_{i_1,i_2,\ldots,i_{p-2}})^2}{2(\sigma_{p-1}^2+\sigma_p^2)}} \Big] \\
&= \lim_{n_1\to\infty} \frac{1}{n_1} \sum_{i_1=1}^{n_1} \Big[\lim_{n_2\to\infty} \frac{1}{n_2} \sum_{i_2=1}^{n_2} \Big[\ldots \frac{1}{n_{p-3}} \sum_{i_{p-3}=1}^{n_{p-3}} \Big[\frac{1}{\sqrt{2\pi(\sigma_{p-2}^2+\sigma_{p-1}^2+\sigma_p^2)}} e^{-\frac{(x-x_{i_1,i_2,\ldots,i_{p-3}})^2}{2(\sigma_{p-2}^2+\sigma_{p-1}^2+\sigma_p^2)}} \Big] \\
&= \ldots \\
&= \frac{1}{\sqrt{2\pi\sum_{i=1}^{p}\sigma_i^2}} e^{-\frac{(x-\mu)^2}{2\sum_{i=1}^{p}\sigma_i^2}}.
\end{aligned}$$

■

Remarque *Faisons tendre p vers ∞.*

- *Si les σ_i sont tels que $\sum_{i=1}^{\infty}\sigma_i^2 = \beta < \infty$, alors la densité est $h(x) = \frac{1}{\sqrt{2\pi\beta}} e^{-\frac{(x-\mu)^2}{2\beta}}$, c'est-à-dire la densité au point x d'une gaussienne de moyenne μ et de variance β.*
- *Si les $\sigma_i$sont tels que $\sum_{i=1}^{\infty}\sigma_i^2 = \infty$, alors l'expression limite de $h(x)$ est nulle.*

3 Conclusion

Cette étude a été fortement motivée par des observations pratiques. Dans le domaine de la botanique, nous nous sommes intéressés au mode de reproduction du chlorophytum (ou plante araignée), une plante qui se développe dans les forêts naturelles indiennes, de l'est de l'Assam au Gujarat. Elle produit de nouvelles petites plantes situées au bout de longues tiges arcées, dont la localisation peut être vue comme obéissant à une loi gaussienne (la plupart des plantelettes se disposent au voisinage immédiat de la plante d'origine). En parvenant à maturité, chaque plantelette produit à son tour de nombreuses nouvelles plantelettes, dont la distribution est conforme à notre modèle. Dans le domaine militaire, nous avons étudié les impacts des bombes à fragmentations. En explosant, ces bombes éjectent de multiples petites bombes, dites bombelettes. La distribution des impacts des bombelettes autour du point d'explosion de la bombe originale peut être vue comme une gaussienne. En frappant le sol, chaque bombelette explose à son point d'impact, éjectants des éclats qui se répandent à leur tour sur le sol environnant suivant une distribution gaussienne. Comme on le voit sur ces exemples, notre étude peut donc être utilisée pour modéliser certains phénomènes naturels ou artificiels.

Bibliographie

[1] S. Arora, R. Kannan (2001) Learning Mixtures of Arbitrary Gaussians, Symposium on Theory of Computing (STOC).
[2] S. Dasgupta (1999) Learning Mixtures of Gaussians, Proc. of Symposium on Foundations of Computer Science (FOCS).
[3] S. El Otmani, A. Maul (2009) Probability distributions arising from nested Gaussians, C. R. Acad. Sci. Paris, Ser. I 347.
[4] B. S. Everitt, D. J. Hand (1981) Finite Mixture Distributions, Chapman and Hall, New York.
[5] R. J. Herrnstein, C. Murray (1994) The Bell Curve : Intelligence and Class Structure in American Life, Free Press. ISBN 0-02-914673-9.
[6] G. J. McLachlan, D. Peel (2000) Finite Mixture Models, Wiley, New York.
[7] J. K. Patel, C. B. Read (1982) Handbook of the Normal Distribution, New York : Dekker.
[8] M. R. Sheldon (2003) Introduction to Probability Models, 8th ed., Academic Press, San Diego.