REVUE DE STATISTIQUE APPLIQUÉE

E. BRUNEL

Application d'estimateurs de la densité à la simulation d'épisodes pluvieux extrêmes en Languedoc-Roussillon

Revue de statistique appliquée, tome 46, n° 4 (1998), p. 45-58

<http://www.numdam.org/item?id=RSA_1998__46_4_45_0>

Rev. Statistique Appliquée, 1998, **XLVI** (4), 45-58

APPLICATION D'ESTIMATEURS DE LA DENSITÉ À LA SIMULATION D'ÉPISODES PLUVIEUX EXTRÊMES EN LANGUEDOC-ROUSSILLON

E. Brunel

Laboratoire de Probabilités et Statistique, Université Montpellier II,
Place Eugène Bataillon, 34095 Montpellier cedex 5

RÉSUMÉ

Nous nous proposons d'établir un modèle de simulation d'épisodes pluvieux extrêmes. Leur modélisation spatiale s'appuie sur l'estimation de la densité de probabilité pour laquelle deux approches sont envisagées. Une méthode non-paramétrique qui ne nécessite aucune information *a priori* sur la loi est tout d'abord mise en œuvre. Nous ajustons ensuite aux données un modèle de mélange dont les paramètres sont estimés à l'aide de l'algorithme SEM.

Mots-clés : *Estimateur à noyau, densité bivariée, mélange de lois, algorithmes EM et SEM, épisodes pluvieux extrêmes.*

ABSTRACT

Our aim is to provide a random generator of the location of an extreme rainfall process from a probability density estimator. A nonparametric method, which does not require any distributional assumption is first proposed. An alternative parametric fitting method is also investigated via the SEM algorithm.

Keywords : *Kernel estimation, bivariate density, mixture density, EM and SEM algorithms, extreme rainfall data.*

1. Introduction

Les récentes catastrophes de Nîmes (1988), de Vaison-la-Romaine (1992), de la région biterroise (1995) et de Montpellier (1997) ont sensibilisé l'opinion publique aux risques encourus par les agglomérations urbaines face à des phénomènes pluvieux extrêmes dus aux cumuls de précipitations. L'une des préoccupations de l'hydrologue est d'évaluer le risque pluvial. Différents paramètres interviennent dans la détermination des zones à risque comme le relief, la nature des sols ou encore l'aléa climatique. C'est ce dernier point qui suscite notre intérêt. Le traitement statistique d'observations pluviométriques (hauteurs d'eau en mm relevées généralement toutes les 24 heures à un poste pluviométrique donné) a été envisagé de multiples façons

dans la littérature. Il est possible de considérer ces observations comme une série chronologique indexée par le temps dans une optique prévisionnelle par exemple. Une autre façon d'aborder le problème consiste à estimer la densité de probabilité des hauteurs d'eau en accordant un intérêt particulier aux queues de distributions (valeurs extrêmes). Ce point de vue a été abondamment étudié dans la littérature d'hydrologie de ces dix dernières années avec des techniques d'ajustement à des lois de type Gumbel, exponentielle, etc. Adamowski (1989) et plus récemment Thao, Bois & Villasenor (1993) ont comparé ces méthodes paramétriques à l'estimateur à noyau dans un cadre univarié.

Notre étude s'appuie sur une approche différente de celles que nous venons d'évoquer. En effet, les hydrologues étudient d'une part la structure des épisodes pluvieux (superficie des surfaces pluvieuses, évolution de ces surfaces en fonction d'un seuil de pluie donné, hauteur d'eau maximale observée), mais aussi leur position sur la région (Neppel, 1997). Au lieu de considérer les hauteurs d'eau précipitées, on va donc plutôt s'intéresser à la localisation géographique des épisodes pluvieux. Cette idée est basée sur le modèle de Huff (1958) dont nous donnons brièvement le principe. On suppose qu'un épisode pluvieux peut être représenté par un épicentre, point théorique autour duquel s'organisent les isohyètes (courbes de niveau des hauteurs d'eau). Ainsi, la position d'un épisode est entièrement déterminée par la localisation de son épicentre. Le jeu de points fourni par le Laboratoire d'Hydrologie de Montpellier est constitué de la façon suivante : 93 épisodes pluvieux de la région qui ont dépassé un seuil donné (190 mm par jour) ont été sélectionnés sur une période de 36 années de 1958 à 1993 (*cf.* carte des relevés, figure 1). La vraie valeur du maximum de l'épisode pluvieux n'est jamais mesurée, mais approchée par la quantité maximale enregistrée à une station proche du vrai maximum. Comme l'ont souligné Richard & al. (1988), cette approximation introduit une erreur dans la modélisation.

L'évaluation analytique de cette erreur s'avère délicate car elle est liée d'une part à la structure de l'épisode pluvieux (notamment à la forme géométrique des isohyètes), mais aussi au réseau des stations d'enregistrement dont la répartition est très hétérogène suivant les régions. La validité de cette approximation repose en partie sur des résultats de simulations ou des comparaisons expérimentales (*cf.* Foufoula-Georgiou, 1989). Nous ferons ici l'hypothèse que la position de l'épicentre peut être identifiée avec celle du poste pluviométrique touché par l'épisode et qui a enregistré la hauteur d'eau maximale. Le problème de la modélisation spatiale de ces épisodes pluvieux extrêmes se ramène donc à l'estimation de la densité de probabilité du couple de variables (X, Y), coordonnées dans le plan des épicentres. C'est l'objet de la partie 2, où nous mettons en œuvre différents estimateurs à noyau. Dans la partie 3, nous envisageons un ajustement paramétrique des données à un modèle de mélange motivé par l'objectif poursuivi : établir un modèle de simulation.

2. Approche non-paramétrique

2.1. Estimateur à noyau de la densité

Rares sont les cas où l'on connaît avec exactitude la forme de la loi inconnue. L'intérêt de l'approche non-paramétrique tient essentiellement à sa souplesse d'application : aucune hypothèse n'est à imposer concernant l'appartenance de la densité

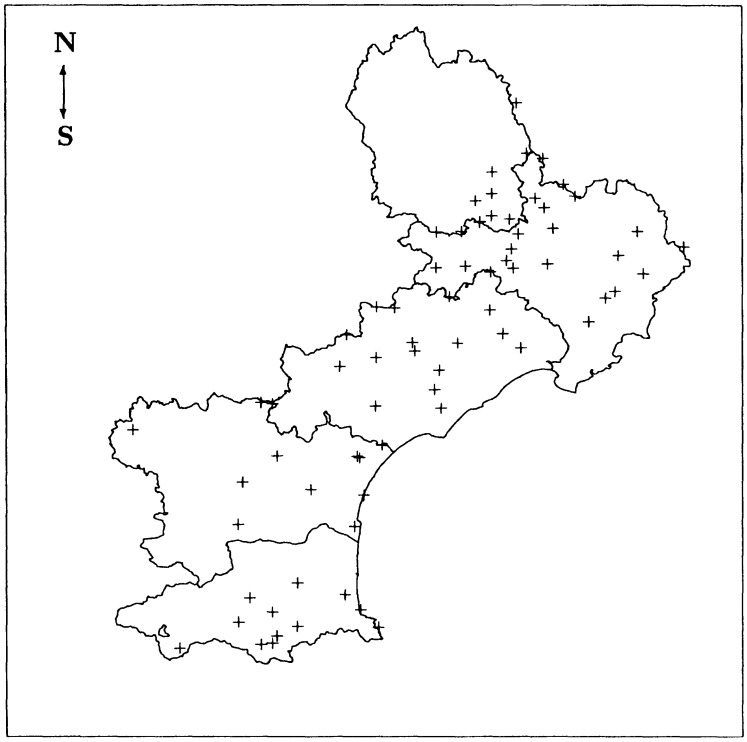

FIGURE 1
Episodes pluvieux extrêmes en Languedoc-Roussillon (1958 à 1993)

f inconnue à une classe paramétrée de densités de probabilité. A partir de l'échantillon $(X_1, Y_1), ..., (X_n, Y_n)$ du couple de variables aléatoires (X, Y) coordonnées dans le plan des épicentres, la densité bivariée f est estimée au point (x, y) de $\mathbb{R}^2$ par l'estimateur à noyau :

$$\hat{f}_n(x,y) = \frac{1}{nh_1h_2} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h_1}\Big) K\Big(\frac{y - Y_i}{h_2}\Big)$$

où le noyau K est une fonction mesurable dont l'intégrale de Lebesgue vaut 1 et où le paramètre de lissage $h = (h_1, h_2)$ est une suite de $\mathbb{R}_+^2$ qui tend vers zéro.

Si le noyau K est une densité de probabilité, $\hat{f}_n$ en est une également et des résultats théoriques assurent la convergence dans L^2 (Tiago de Oliveira, 1963), dans L^1 (Devroye, 1983) ou la convergence uniforme presque complète (Deheuvels, 1974) de l'estimateur à noyau, sous les hypothèses habituelles $h \to 0$, $nh \to +\infty$ quand $n \to +\infty$ et pourvu que la densité f et le noyau K satisfassent certaines conditions de régularité. Le chapitre 4 du livre de Bosq et Lecoutre (1987) est consacré à ces théorèmes de convergence de l'estimateur à noyau multivarié; on s'y référera pour de plus amples développements.

Nous devons alors choisir le noyau K et le paramètre de lissage h de façon optimale au sens d'un critère qui mesure la qualité de l'estimation. L'erreur quadratique moyenne intégrée (MISE) se décompose en termes de biais et de variance de l'estimateur. Ainsi, choisir le MISE comme critère d'erreur permet une interprétabilité statistique.

2.2. Fenêtre optimale pour un noyau positif

Dans ce paragraphe, nous considérons un noyau K positif, symétrique, de carré intégrable et de moment d'ordre 2 fini. Posons alors $R(K) = ||K||_{L^2}^2 = \int_{-\infty}^{+\infty} K^2(x)\,dx$ et $\sigma^2(K) = \int_{-\infty}^{+\infty} x^2\,K(x)\,dx$. Le développement asymptotique du MISE en dimension 2 lorsque $\max(h_1, h_2)$ tend vers 0 et $n\min(h_1, h_2)$ tend vers $+\infty$ s'écrit :

$$\text{MISE}\,(\hat{f}_n, f) = E\int_{\mathbb{R}^2}\left(\hat{f}_n(x) - f(x)\right)^2 dx = \text{AMISE}\,(\hat{f}_n, f) + o\left(h_1^2 h_2^2 + \frac{1}{nh_1h_2}\right)$$

où le terme d'AMISE, partie principale du MISE dans le développement asymptotique, est la somme du biais au carré et de la variance asymptotiques et est égal à :

$$\frac{\sigma^4(K)}{4}\sum_{i,j=1}^{2} h_i^2 h_j^2 \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{\partial^2 f}{\partial x_i^2}(x_1, x_2)\,\frac{\partial^2 f}{\partial x_j^2}(x_1, x_2)\,dx_1\,dx_2 + \frac{R^2(K)}{nh_1h_2}$$

Le biais asymptotique dont le carré est le premier terme de l'AMISE sera d'autant plus petit que h sera petit. Par contre, la variance asymptotique est d'autant plus petite que l'on «moyenne» sur beaucoup de points, c'est-à-dire que la largeur de la fenêtre est grande. Cette interprétation montre à quel point le choix de la fenêtre h est important puisqu'il faut trouver un équilibre entre biais et variance. La valeur asymptotiquement optimale h^* du paramètre h qui rend l'AMISE minimale est :

$$h_i^* = C_i(f)\,n^{-\frac{1}{6}}, \text{ pour } i = 1, 2$$

et la valeur correspondante de l'AMISE est d'ordre $O(n^{-\frac{2}{3}})$. Le calcul de la fenêtre optimale $h^* = (h_1^*, h_2^*)$ se ramène alors à l'évaluation des constantes réelles $C_1(f)$ et $C_2(f)$. Leur expression est donnée en fonction des quantités inconnues $Rf_{ij} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{\partial^2 f}{\partial x_i^2}(x_1, x_2)\,\frac{\partial^2 f}{\partial x_j^2}(x_1, x_2)\,dx_1\,dx_2$, pour i, j variant de 1 à 2, par :

$$\begin{cases} C_1(f) = \left(\dfrac{Rf_{22}}{Rf_{11}}\right)^{\frac{1}{4}} C_2(f) \\ C_2(f) = \left(\dfrac{R^2(K)/\sigma^4(K)}{(Rf_{22})^{5/4}\,(Rf_{11})^{-1/4} + Rf_{12}\,(Rf_{22})^{3/4}\,(Rf_{11})^{-3/4}}\right)^{\frac{1}{6}} \end{cases}$$

Une méthode de référence consiste à remplacer $C_1(f)$ et $C_2(f)$ par $C_1(f_0)$ et $C_2(f_0)$ où f_0 est une densité appartenant à une classe paramétrée connue (*cf.* Deheuvels, 1977 et Scott, 1992).

Densité de référence gaussienne

On choisit comme référence la densité gaussienne unimodale :

$$f_0^{unim}(x,y) = \frac{1}{2\pi\sigma_1\sigma_2}\exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_1^2}+\frac{y^2}{\sigma_2^2}\right)\right]$$

Pour ce choix et si le noyau K est la densité gaussienne, un rapide calcul donne :

$$h_i^* = \sigma_i\, n^{-\frac{1}{6}}, \text{ pour } i = 1,2$$

et la valeur correspondante de l'AMISE est $\text{AMISE}^* = \frac{3}{8\pi}(\sigma_1\sigma_2)^{-1}\, n^{-\frac{2}{3}}$. Il suffit alors d'estimer σ_1^2 et σ_2^2 par les variances empiriques marginales $\hat{\sigma}_1^2$ et $\hat{\sigma}_1^2$. Ce choix qui privilégie la loi normale se justifie puisque pour d'autres densités de référence l'influence sur la valeur de l'AMISE est peu sensible. Remarquons que pour un autre noyau positif, centré, il suffit de diviser h_1^* et h_2^* par l'écart-type de ce noyau pour obtenir les valeurs de la fenêtre optimale qui lui sont associées (*cf.* Scott, 1992, page 142). La figure 2 montre l'estimateur à noyau construit à partir du noyau gaussien pour la densité de référence gaussienne f_0^{unim}.

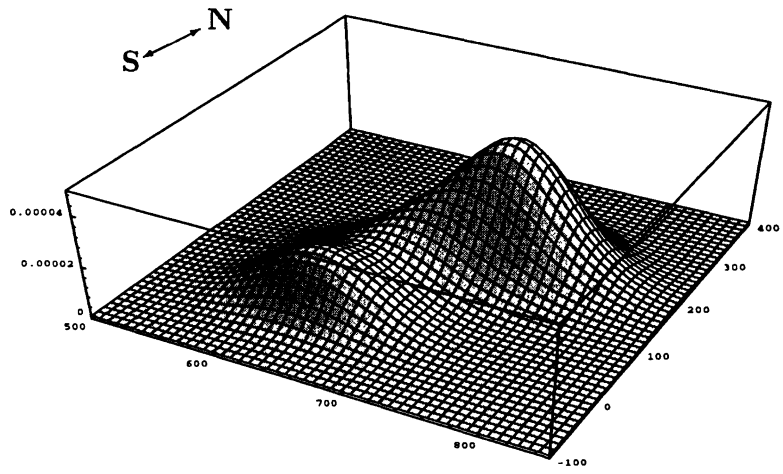

FIGURE 2
Estimateur à noyau, $\hat{h}_1^* = 35$, $\hat{h}_2^* = 34$ *(noyau gaussien)*

Densité de référence bimodale

Une méthode de référence consiste à utiliser une densité f_0 connue à la place de de la densité f pour calculer le paramètre de lissage. Par conséquent, l'estimation doit être d'autant meilleure que f_0 est «proche» de f. Or, les estimateurs obtenus avec la densité gaussienne laissent clairement apparaître deux modes qui correspondent à deux zones de relief important de la région, d'une part le massif des Cévennes au Nord-Est (qui correspond au mode le plus important) et d'autre part les Pyrénées Orientales. Guidés par cette information, il nous semble donc plus adapté d'utiliser une densité de référence bimodale dont les deux modes seront localisés sur ces deux zones géographiques plutôt que la densité gaussienne f_0^{unim} qui conduit à un surlissage. Notre choix se porte alors sur la densité de référence bimodale, mélange de deux gaussiennes :

$$f_0^{bim}(x,y) = \alpha\, g(x,y,\mu_1,\Sigma_1) + (1-\alpha)\, g(x,y,\mu_2,\Sigma_2)$$

où les paramètres $\alpha, \mu_1, \Sigma_1, \mu_2$ et Σ_2 doivent être estimés (*cf.* algorithme EM, paragraphe 3.2). Finalement, nous pourrions résumer la méthode de la façon suivante :

1. On estime paramétriquement la densité f par f_0^{bim} mélange de deux gaussiennes.
2. On injecte l'estimateur paramétrique f_0^{bim} dans l'expression du paramètre de lissage optimal pour construire l'estimateur à noyau.

En supposant les matrices de variance-covariances de chaque composante du mélange diagonales, les paramètres estimés par l'algorithme EM sont :

$$\hat{\alpha} = 0.419$$

$$\hat{\mu}_{11} = 638.812 \quad \hat{\mu}_{12} = 73.155 \quad \hat{\Sigma}_1 = \begin{pmatrix} 29.3 & 0 \\ 0 & 51.7 \end{pmatrix}$$

$$\hat{\mu}_{21} = 721.115 \quad \hat{\mu}_{22} = 200.407 \quad \hat{\Sigma}_2 = \begin{pmatrix} 26.3 & 0 \\ 0 & 28.9 \end{pmatrix}$$

Cette densité f_0^{bim}, qui est un estimateur paramétrique de f, a la même allure que l'estimateur à noyau obtenu précédemment avec la densité de référence unimodale f_0^{unim}. D'ailleurs, nous pourrions choisir d'utiliser l'estimateur à noyau de la figure 2 comme densité de référence à la place de l'estimateur paramétrique f_0^{bim}. La figure 3 représente l'estimateur à noyau construit avec un noyau gaussien et le paramètre de lissage $(\hat{h}_1^*, \hat{h}_2^*) = (19, 17)$ basé sur f_0^{bim}.

Nous détectons alors la présence de plusieurs autres modes (4 ou 5) : les deux modes déjà identifiés précédemment et deux modes au Nord de la région.

La cinquième «bosse» située au sud-est du mode principal des Cévennes ne correspond pas à une zone de relief mais peut être expliquée par la présence de valeurs extrêmes comme celle de l'inondation de Nîmes.

La comparaison des estimateurs basés sur deux densités de référence différentes illustre le rôle très important du paramètre de lissage. La densité de référence entraîne

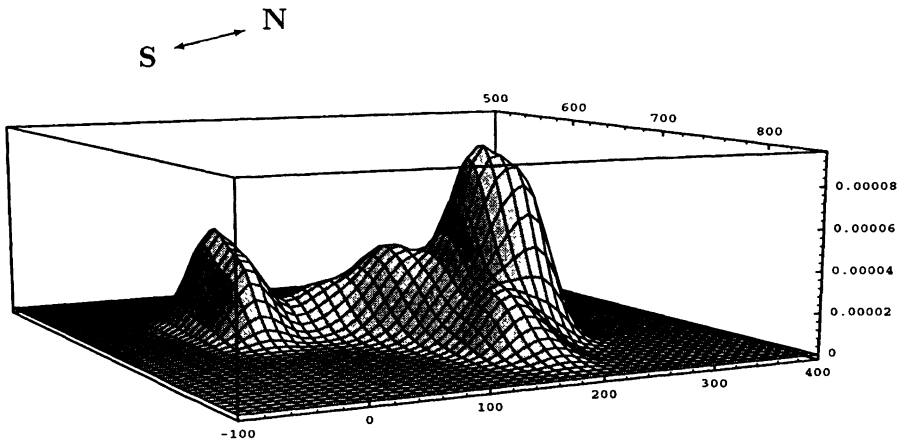

FIGURE 3
Estimateur à noyau : $\hat{h}_1^* = 19$, $\hat{h}_2^* = 17$ *(noyau gaussien)*

des sous ou sur-lissages si elle est mal adaptée aux données. Un noyau gaussien unimodal, bien que souvent employé, n'est donc pas un choix judicieux pour une densité multimodale.

2.3 Noyaux d'ordre supérieur

Un noyau K est dit d'ordre r $(r \geqslant 2)$, si :

$$\int_{-\infty}^{+\infty} x^i K(x)\,dx = 0, \text{ pour } i = 1, ..., r-1 \quad \text{et} \quad \int_{-\infty}^{+\infty} x^r K(x)\,dx \neq 0.$$

Dès que r est strictement supérieur à 2, ces noyaux ne sont plus des densités de probabilité. Par conséquent, nous n'obtiendrons pas nécessairement des estimateurs stricts de la densité. Cependant, l'annulation de leurs moments jusqu'à l'ordre $r-1$ conduit à une réduction du terme de biais asymptotique. Le principe d'une méthode multi-noyaux (Berlinet, 1993) est de parcourir une famille de noyaux associée à une densité K_0, appelée noyau de base, et de calculer pour chaque noyau d'ordre r le paramètre $\hat{h}_r^*$ optimal au sens d'un critère donné. Si l'on choisit de minimiser l'AMISE, on retiendra le couple $(K_r, \hat{h}_r^*)$ pour lequel l'AMISE est minimale.

L'analyse révèle que les noyaux d'ordre 4 ou 6 de la hiérarchie gaussienne ne conduisent pas à un gain significatif de l'AMISE. La valeur calculée de l'AMISE pour le paramètre de lissage optimal $(\hat{h}_1^*, \hat{h}_2^*)$ vaut $3.593.10^{-6}$ pour le noyau gaussien $K_0(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$ tandis qu'elle est égale à $3.395.10^{-6}$ pour $K_2(x) = \left(\frac{3}{2} - \frac{x^2}{2}\right) K_0(x)$, noyau d'ordre 2 de la hiérarchie gaussienne. Ce gain non significatif s'explique par le fait qu'une légère diminution du biais asymptotique

(on passe de $1.278.10^{-6}$ pour le noyau gaussien K_0 à $8.410.10^{-7}$ pour le noyau K_2 d'ordre 2) est compensée par une augmentation de la variance asymptotique ($2.315.10^{-6}$ pour K_0 à $2.554.10^{-6}$ pour K_2). De même, l'AMISE calculée pour $K_6(x) = \left(\frac{15}{8} - \frac{5}{4}x^2 - \frac{1}{8}x^4\right) K_0(x)$, noyau d'ordre 6 ne donne pas un gain satisfaisant ($3.274.10^{-6}$ dont $3.607.10^{-7}$ pour le biais et $2.913.10^{-6}$ pour la variance). Pour notre taille d'échantillon les noyaux d'ordre 2 ou 4 suffisent pour construire l'estimateur non-paramétrique.

3. Ajustement à un modèle de mélange

Nous allons maintenant ajuster un modèle paramétrique aux données. Notre objectif n'est pas de présenter ce modèle comme un compétiteur de l'estimateur à noyau. Il s'agit plutôt d'utiliser les informations de l'approche non-paramétrique pour le choix d'un modèle de mélange plus adapté à l'élaboration d'un algorithme rapide de simulation.

3.1. Motivation pour un modèle de mélange

Les modèles de mélange ont connu un intérêt croissant depuis les années 50 car ils interviennent dans de nombreux domaines d'application. Leur rôle apparaît principalement lorsque les distributions conditionnelles ne sont pas directement accessibles à partir des observations et que seule la distribution «mélangée» de tout l'échantillon peut être étudiée. De telles situations sont courantes quand la variable sous-jacente qui partitionne les observations en groupes est inobservable. Des exemples issus des sciences expérimentales sont donnés dans Everitt & Hand (1981). Les modèles les plus étudiés sont sans doute les mélanges de gaussiennes que nous utiliserons ici.

Considérons le modèle de mélange fini de densités de probabilité dans $\mathbb{R}^2$:

$$f(x,y) = \sum_{k=1}^{\kappa} p_k \, g(x,y,\mu_k,\Sigma_k)$$

avec les notations suivantes :

- $0 < p_k < 1$, $k = 1, ..., \kappa$ et $\sum_{k=1}^{\kappa} p_k = 1$ (proportions du mélange).
- La $k^{\text{ième}}$ composante du mélange $g(x,y,\mu_k,\Sigma_k)$ est la densité gaussienne bivariée de moyenne $\mu_k = (\mu_{k1}, \mu_{k2})^t$ et de matrice de covariance Σ_k.
- L'entier κ désigne le nombre de composantes du mélange.

L'examen de l'estimateur non-paramétrique donne une indication sur le nombre de modes. En effet, l'estimateur à noyau étudié dans le paragraphe 2 laisse apparaître 4 modes ce qui nous amène à conjecturer un mélange de 4 gaussiennes. Cependant, un mélange de deux ou plusieurs densités unimodales ne conduit pas nécessairement à une densité multimodale (*cf.* Behboodian 1970).

3.2. Estimateurs du Maximum de Vraisemblance

Il s'agit d'estimer le paramètre d'intérêt $\theta = (p_k, \mu_k, \Sigma_k)_{k=1}^{\kappa}$ lorsque κ est supposé connu. En supposant les matrices de variance-covariances diagonales, le nombre de paramètres réels à estimer est égal à $5\kappa - 1$. Dès que l'on a plus de 3 composantes ou dans le cas multivarié, quelle que soit la méthode utilisée pour le calcul des estimateurs, le coût numérique est important. Le choix de la méthode du maximum de vraisemblance se justifie plutôt par le fait que les estimateurs obtenus possèdent de bonnes propriétés statistiques. Ils convergent en probabilité vers les estimés et sont asymptotiquement normalement distribués.

Soit $(X_1, Y_1), ..., (X_n, Y_n)$ un n-échantillon de densité f. La fonction de vraisemblance du modèle s'ecrit :

$$L(\theta) = \prod_{j=1}^{n} \sum_{k=1}^{\kappa} p_k \, g(X_j, Y_j, \mu_k, \Sigma_k)$$

La maximisation du logarithme de la vraisemblance sous la contrainte que $\sum_{k=1}^{\kappa} p_k = 1$ revient en introduisant le multiplicateur de Lagrange λ à maximiser :

$$\mathcal{L}(\theta) = \ln L(\theta) - \lambda \left(\sum_{k=1}^{\kappa} p_k - 1 \right)$$

Ecrivant que les dérivées partielles de $\mathcal{L}(\theta)$ par rapport aux différentes composantes de θ sont nulles, on obtient les équations suivantes pour k variant de 1 à κ et $\ell = 1, 2$:

$$\begin{cases} \dfrac{\partial \ln \mathcal{L}}{\partial p_k} = \displaystyle\sum_{j=1}^{n} \frac{g(X_j, Y_j, \mu_k, \Sigma_k)}{f(X_j, Y_j)} - \lambda = 0 \\ \dfrac{\partial \ln \mathcal{L}}{\partial \mu_{k\ell}} = \displaystyle\sum_{j=1}^{n} p_k \frac{\partial g(X_j, Y_j, \mu_k, \Sigma_k) \,/\, \partial \mu_{k\ell}}{f(X_j, Y_j)} = 0 \\ \dfrac{\partial \ln \mathcal{L}}{\partial \sigma_{k\ell}} = \displaystyle\sum_{j=1}^{n} p_k \frac{\partial g(X_j, Y_j, \mu_k, \Sigma_k) \,/\, \partial \sigma_{k\ell}}{f(X_j, Y_j)} = 0 \end{cases}$$

avec $\mu_k = (\mu_{k1}, \mu_{k2})^t$ et $\Sigma_k = \begin{pmatrix} \sigma_{k1}^2 & 0 \\ 0 & \sigma_{k2}^2 \end{pmatrix}$, pour $k = 1, ..., \kappa$.

On obtient alors facilement l'expression des estimateurs de p_k, μ_k, Σ_k, pour k variant de 1 à κ :

$$\hat{p}_k = \frac{1}{n} \sum_{j=1}^{n} p_{kj} \tag{1}$$

$$\hat{\mu}_k = (\hat{\mu}_{k1}, \hat{\mu}_{k2})^t = \Big(\frac{1}{n\,\hat{p}_k}\sum_{j=1}^{n} p_{kj}\,X_j,\ \frac{1}{n\,\hat{p}_k}\sum_{j=1}^{n} p_{kj}\,Y_j\Big)^t \tag{2}$$

$$\hat{\Sigma}_k = \begin{pmatrix} \dfrac{1}{n\,\hat{p}_k}\displaystyle\sum_{j=1}^{n} p_{kj}\,(X_j - \hat{\mu}_{k1})^2 & 0 \\ 0 & \dfrac{1}{n\,\hat{p}_k}\displaystyle\sum_{j=1}^{n} p_{kj}\,(Y_j - \hat{\mu}_{k2})^2 \end{pmatrix} \tag{3}$$

où la quantité p_{kj} désigne la probabilité *a posteriori* pour que l'observation (X_j, Y_j) soit issue de la k$^{\text{ième}}$ sous-population et est définie par :

$$p_{kj} = p_k\, g(X_j, Y_j, \mu_k, \Sigma_k)/f(X_j, Y_j). \tag{4}$$

La forme de ces estimateurs n'est pas sans nous rappeler celle des estimateurs classiques du maximum de vraisemblance d'un modèle gaussien à savoir les moyennes et variances empiriques excepté que chaque observation (X_j, Y_j) est pondérée par la probabilité a posteriori p_{kj}. D'autre part, écrits sous cette forme, les estimateurs ne sont pas connus explicitement puisque les probabilités p_{kj} sont inconnues. Une procédure itérative est donc nécessaire pour pouvoir les calculer à partir des équations (1), (2) et (3). Dempster, Laird et Rubin (1977) proposent d'appliquer leur algorithme EM au cas de l'estimation des paramètres d'un mélange de densités. A partir d'une valeur initiale $\theta^0 = (p_k^0, \mu_k^0, \Sigma_k^0)_{k=1}^{\kappa}$, on calcule les premiers estimateurs $\hat{p}_{kj}$ des probabilités *a posteriori* p_{kj} grâce à l'équation (4) : c'est l'étape d'Estimation. Puis, les estimateurs $\hat{p}_{kj}$ sont injectés dans les équations (2) et (3) pour donner des estimateurs «revisités» des moyennes et variances : c'est l'étape de Maximisation. Ces deux étapes sont répétées jusqu'à ce qu'un critère de convergence soit satisfait.

3.3. Application aux données hydrologiques

Dans le paragraphe 3.1, nous avons déjà évoqué le problème du choix du nombre κ de composantes du mélange. C'est là le point clé de la démarche du praticien. Nous avons vu également que l'algorithme EM ne permet pas de résoudre ce problème puisqu'il nécessite que κ soit connu. Nous allons mettre en œuvre l'algorithme SEM qui est une version améliorée de EM due à Celeux & Diebolt (1985). La modification consiste en une procédure d'apprentissage probabiliste (Etape Stochastique), incorporée avant les procédures d'Estimation et de Maximisation. On définit une borne supérieure $\kappa_{\max}$ du nombre κ inconnu et à l'issue de la phase d'«apprentissage», l'algorithme propose un nombre $\kappa \leqslant \kappa_{\max}$ de composantes. D'autre part, la suite $\theta^n = (p_k^n, \mu_k^n, \Sigma_k^n)_{k=1}^{\kappa}$ des paramètres obtenus à l'itération n, est une chaîne de Markov ergodique qui converge en probabilité vers l'unique point stationnaire Θ (*cf.* Redner & Walker, 1984). L'étape Stochastique évite que la suite θ^n ne reste dans un état stationnaire instable de la chaîne.

Initialisation de l'algorithme

Les résultats de l'algorithme SEM sont fortement liés aux valeurs initiales. En général, les équations de vraisemblance n'admettent pas une unique solution. Un algorithme itératif de calcul ne nous garantit pas que les estimateurs associés au maximum global seront atteints. Dans le cas multivarié, nous possédons peu de techniques pour trouver des valeurs initiales. Nous proposons d'utiliser les modes de l'estimateur à noyau comme valeurs initiales des moyennes de chacune des composantes. Grâce à un algorithme d'optimisation (algorithme de descente la plus rapide), nous avons calculé les 5 maxima de l'estimateur à noyau.

TABLEAU 1
Maxima de l'estimateur à noyau

k	1	2	3	4	5
μ_{k1}^0	706.055	671.610	617.424	658.147	766.023
μ_{k2}^0	203.571	152.381	22.315	101.172	177.064

Critère d'arrêt

Nous ne disposons d'aucun critère théorique capable de tester si la suite θ^n a atteint ou non la stationnarité. En pratique, ceci revient à s'interroger sur le nombre d'itérations nécessaires pour que l'algorithme converge. Celeux & Diebolt préconisent de procéder de la façon suivante :

1. L'algorithme est itéré jusqu'à ce qu'une valeur de κ soit acceptée (étape d'apprentissage).
2. Puis, après chaque itération n, on stocke les valeurs $p_k^n, \mu_{k1}^n, \mu_{k2}^n, \sigma_{k1}^n$, et σ_{k2}^n afin de calculer les moyennes et écart-types de chacun de ces paramètres.

Les résultats consignés dans le tableau 2 ont été obtenus après 50 itérations. Chaque ligne correspond à l'un des paramètres du mélange. La première colonne donne la valeur moyenne $\overline{x}$, la seconde l'écart-type s qui ont été calculés sur 40 itérations après l'étape d'apprentissage. La troisième et la quatrième indiquent les valeurs de $\overline{x} - s$ et $\overline{x} + s$. Enfin, dans la dernière colonne nous avons reporté les valeurs finales des estimateurs.

Toutes ces valeurs tombent dans l'intervalle $(\overline{x} - s, \overline{x} + s)$; ceci est un argument pour dire que la stabilité a été atteinte. De plus, il s'avère que si l'on poursuit l'algorithme jusqu'à la 100-ième itération, les valeurs finales sont identiques à celles obtenues après 50 itérations à 10^{-4} près. Or, si après ces 50 itérations l'algorithme n'avait atteint qu'un maximum local (cette éventualité restant toujours possible), l'étape stochastique aurait probablement permis à l'algorithme d'explorer une autre région. Nous remarquons également que la convergence des variances est beaucoup moins rapide; mais l'initialisation était beaucoup plus approximative que pour les moyennes puisque nous nous sommes basés sur les variances marginales empiriques

TABLEAU 2
Résultats de SEM après 50 itérations

	$\overline{x}$	s	$\overline{x}-s$	$\overline{x}+s$	valeur après 50 itérations
p_1	0.532	0.006	0.527	0.538	0.529
μ_{11}	724.877	0.447	724.430	725.324	725.059
μ_{12}	204.790	0.591	204.198	205.382	205.077
σ^2_{11}	23.875	0.290	23.585	24.165	23.776
σ^2_{12}	25.857	0.454	25.403	26.311	25.633
p_2	0.195	0.003	0.192	0.198	0.196
μ_{21}	630.448	1.509	628.938	631.958	630.946
μ_{22}	26.205	6.955	19.249	33.161	23.491
σ^2_{21}	25.191	0.833	24.358	26.025	24.958
σ^2_{22}	12.263	9.028	3.235	21.291	8.526
p_3	0.155	0.007	0.148	0.162	0.152
μ_{31}	669.175	2.180	666.995	671.356	670.240
μ_{32}	142.439	7.042	135.397	149.481	145.788
σ^2_{31}	16.794	0.992	15.801	17.787	16.934
σ^2_{32}	19.093	7.466	11.627	26.559	15.567
p_4	0.117	0.008	0.109	0.125	0.122
μ_{41}	629.493	3.466	626.028	632.959	629.964
μ_{42}	92.130	7.444	84.685	99.575	95.976
σ^2_{41}	31.350	0.630	30.720	31.980	31.701
σ^2_{42}	25.470	9.598	15.872	35.066	20.648

de l'échantillon. La densité de mélange dont les paramètres sont donnés dans le tableau précédent est représentée figure 4. Des algorithmes classiques de simulation d'un mélange de lois connues sont disponibles dans la littérature (cf. Devroye, 1986) et peuvent être aisément appliqués à partir de notre estimateur.

4. Conclusion

Les méthodes paramétriques et non-paramétriques, bien qu'elles identifient les mêmes modes ne donnent pas le même «poids» à chacun d'entre eux. En effet, l'estimateur à noyau met en évidence un mode beaucoup plus important sur la région des Cévennes alors que les deux modes principaux du modèle paramétrique sont comparables. L'estimateur à noyau est affecté par les effets de bords qui sont inhérents à la méthode non-paramétrique. Le deuxième mode sur les Pyrénées Orientales correspond à une région de frontière : les observations ne sont pas disponibles au delà. Cependant, l'estimateur à noyau que l'on peut interpréter comme un mélange de n

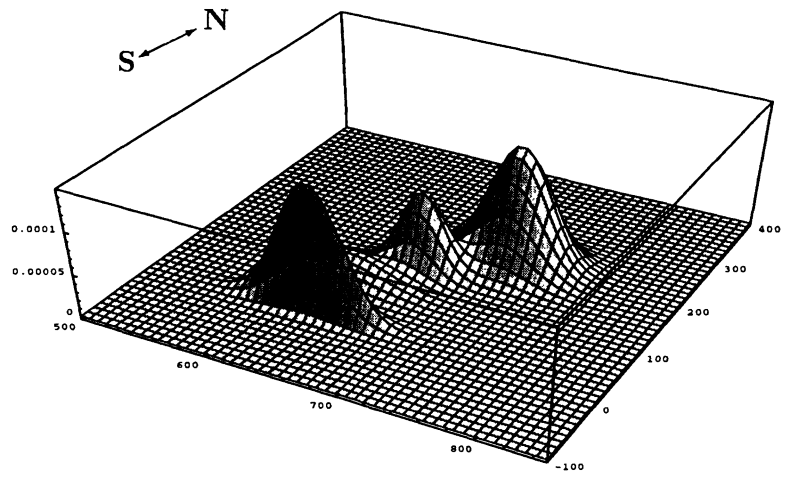

FIGURE 4
Mélange de 4 gaussiennes

gaussiennes (centrées en chaque point de l'échantillon) opère une première réduction de dimension en identifiant 4 ou 5 modes principaux. Les réponses apportées par cette étude ne sont pas les seules et nous pourrions envisager d'autres techniques de réduction de dimension (projection-pursuit par exemple). Cependant, les estimateurs présentés ici sont encore méconnus des hydrologues. Leur application à des données réelles n'est pas sans poser des problèmes pratiques. En effet, quelle que soit la méthode, la mise en œuvre n'est pas automatique et l'on est confronté à des choix directement liés au jeu de données en présence.

Remerciements

Je tiens à remercier le Professeur Alain Berlinet dont les conseils m'ont été précieux tout au long de ce travail. Les discussions que j'ai eues avec Luc Neppel du Laboratoire d'Hydrologie de Montpellier m'ont beaucoup aidée dans la compréhension des phénomènes physiques sous-jacents à cette étude et je l'en remercie également.

Références

[1] ADAMOWSKI K., (1989), A Monte Carlo Comparison of Parametric and Nonparametric Estimation of Flood Frequencies, Journal of Hydrology, 108, 295-308.

[2] BEHBOODIAN J., (1970), On the Modes of a Mixture of two Normal Distributions, Technometrics, 12, 131-39.

[3] BERLINET A., (1993), Hierarchies of Higher Order Kernels, Probab. Theory Relat. Fields , 94, 489-504.

[4] BHATTACHARAYA C.J., (1967), A Simple Method of Resolution of a Distribution into Gaussian Components, Biometrics, 23, 115-35.

[5] BOSQ D., LECOUTRE J.P., (1987), Théorie de l'Estimation Fonctionnelle, Economica.

[6] CELEUX G., DIEBOLT J., (1985), The SEM Algorithm : A Probabilistic Teacher Algorithm Derived from the EM Algorithm for the Mixture Problem, Computational Statistics Quarterly, vol. 2, Issue 1, 73-82.

[7] DEHEUVELS P., (1974), Conditions Nécessaires et Suffisantes de Convergence Presque Sûre et Uniforme Presque Sûre des Estimateurs de la Densité, C.R. Acad. Sci. Paris A, 278, 1217-1220.

[8] DEHEUVELS P., (1977), Estimation Non Paramétrique de la Densité par Histogrammes Généralisés, Rev. Statist. Appl. , n°3, 35, 5-42.

[9] DEMPSTER A.P., LAIRD N.M., RUBIN D.B., (1977), Maximum Likelihood from Incomplete Data via the EM Algorithm, J. Royal Statist. Soc., Serie B, 39, 1-38.

[10] DEVROYE L., (1983), The Equivalence of Weak, Strong and Complete Convergence in L^1 for Kernel Density Estimates, Ann. Statist. 11, 896-904.

[11] DEVROYE L., (1986), Non-Uniform Random Variate Generation, Springer-Verlag.

[12] EVERITT B.S., HAND D.J., (1981), Finite Mixture Distributions in Monographs on Applied Probability and Statistics, Chapman & Hall, London.

[13] FOUFOULA-GEORGIOU E., (1989), On the Accuracy of the Maximum Recorded Depth in Extreme Rainstorms, in Proceedings of the IAHS Third Scientific Assembly, Baltimore, 1989.

[14] HUFF F.A., (1958), Hydrometeorological Analysis of Severe Rainstorms in Illinois, Rep. Invest. 35, State Water Survey, Urbana.

[15] NEPPEL L., (1997), Caractérisation de l'aléa climatique en Languedoc-Roussillon, Thèse de Docteur-Ingénieur de l'Université de Montpellier II.

[16] REDNER R.A., WALKER H.F., (1984), Mixture Densities, Maximum Likelihood and the EM Algorithm, SIAM Review, vol. 26, 2, 195-239.

[17] RICHARD F., HANSEN E.M., WOODWARD D., (1988), Estimation of Precipitation Distribution and Amount in Extreme Events, Eos Trans., AGU, 69, 351.

[18] SCOTT D.W., (1992), Multivariate Density Estimation, Wiley Series in Probability and Mathematical Statistics.

[19] THAO N.T.P., BOIS P., VILLASENOR J.A., (1993), Simulation in Order to Choose a Fitting Method for Extreme Rainfall Data, Atmospheric Research, 30, 13-36.

[20] TIAGO DE OLIVEIRA J., (1963), Estatistica de Densidades. Resultados Assintoticos. Rev. Fac. Ci. Univ. Lisboa, ser. A, 9, 111-206.