REVUE DE STATISTIQUE APPLIQUÉE

C. SOUBIRAN
G. CELEUX
J. DIEBOLT
C. ROBERT

Analyse de mélanges gaussiens pour de petits échantillons : application à la cinématique stellaire

Revue de statistique appliquée, tome 39, n° 3 (1991), p. 17-35
<http://www.numdam.org/item?id=RSA_1991__39_3_17_0>

Rev. Statistique Appliquée, 1991, **XXXIX** (3), 17-36

ANALYSE DE MÉLANGES GAUSSIENS POUR DE PETITS ÉCHANTILLONS : APPLICATION À LA CINÉMATIQUE STELLAIRE

C. SOUBIRAN[1], G. CELEUX[2], J. DIEBOLT[3], C. ROBERT[3]

(1) Observatoire de Paris
(2) INRIA-Rocquencourt
(3) L.S.T.A., Université Pierre et Marie Curie Paris 6

RÉSUMÉ

L'estimation des paramètres d'un mélange de distributions est un problème difficile à résoudre, particulièrement pour de petits échantillons. Nous présentons deux méthodes bien adaptées au traitement d'un petit nombre de données : l'algorithme SAEM et la méthode d'échantillonnage bayésien. SAEM est une version de type recuit simulé de l'algorithme EM, l'échantillonnage bayésien approche de manière itérative l'estimateur de Bayes, autrement incalculable, et les deux techniques sont utilisées conjointement : SAEM fournit un choix de conditions a priori pour la méthode bayésienne, qui, à son tour, valide les résultats de SAEM. L'article est centré sur une application astrophysique en cinématique stellaire. Les 2 algorithmes sont mis en œuvre sur 3 échantillons de vitesses spatiales d'étoiles pour lesquels on cherche à exhiber l'existence de 2 sous-populations gaussiennes.

Mots-clés : *mélange, petits échantillons, algorithme stochastique, algorithme EM, échantillonnage bayésien, estimateur de Bayes, cinématique stellaire.*

ABSTRACT

The estimation of the parameters of a mixture distribution is particularly delicate, especially when the sample size is small. We present in this paper two competing methods to deal with this problem, SAEM and Bayesian sampling. SAEM appears as a simulated annealing version of EM, while Bayesian sampling proposes a Monte-Carlo approximation of the Bayes estimator. We consider a joint use of the two approaches : SAEM provides some initial values for Bayesian sampling and Bayesian sampling somehow selects the best solution among these values. The model we consider in the paper is suggested by an astrophysical problem. We are comparing the two algorithms on three samples of bidimensional star velocities, in order to exhibit a mixture of two normal distributions.

Key-words : *mixture, small samples, stochastic algorithm, EM algorithm, Bayesian sampling, Bayes estimator, stellar kinematics.*

1. Motivations de l'étude

La distribution des étoiles dans la Galaxie est fonction de plusieurs paramètres : position par rapport au centre de la Galaxie, vitesse (résiduelle et systématique), composition chimique, âge, masse. Dans les modèles de représentation de notre galaxie, les étoiles sont séparées en 2 populations. La Population I, ou *population du disque*, comprend des étoiles de même métallicité que le Soleil, concentrées dans le plan galactique, avec des orbites peu excentriques autour du centre de la Galaxie. On trouve dans la Population II des étoiles vieilles, de faible métallicité, ayant une distribution spatiale pratiquement sphérique et des orbites très excentriques. Certains modèles prennent en compte l'existence d'une population ayant des propriétés intermédiaires, et que l'on nomme *disque épais*.

La cinématique stellaire joue un rôle très important dans l'étude des populations de notre Galaxie et permet de développer des théories sur sa formation et son évolution.

On considère le mouvement des étoiles dans un système de référence centré sur le Soleil. Les directions des 3 axes sont notés :

- U dans la direction du centre galactique,
- V dans la direction de la rotation galactique,
- W dans la direction perpendiculaire au plan galactique.

L'étude de la vitesse spatiale des étoiles du disque a montré que chacune de ces 3 vitesses a une distribution très proche d'une loi gaussienne, dont la variance diffère selon le type d'étoiles.

La cinématique des étoiles naines de type spectral A normal a été étudiée récemment par Gómez *et al.* (1990). Les étoiles A sont des étoiles jeunes, d'âge inférieur à 10^9 ans. Du fait de leurs caractéristiques physiques (composition chimique, température effective,...), de leur distribution spatiale et de leurs propriétés cinématiques, elles appartiennent à la Population I. L'histogramme de la vitesse U de ces étoiles suggérait la possibilité d'un mélange, et a incité les auteurs à utiliser l'algorithme SEM (voir Celeux et Diebolt (1986)), qui est une version stochastique de la méthode itérative EM de résolution des équations du maximum de vraisemblance, introduite par Dempster, Laird et Rubin (1977). Effectivement, SEM leur a permis de mettre en évidence l'existence de 2 ou 3 sous-populations, leur proportion étant fonction de l'âge moyen des étoiles considérées. La conclusion de cette étude était que chacune de ces 2 sous-populations est représentative d'une même génération d'étoiles qui seraient nées en même temps, dans une même "bouffée" de formation.

Il existe, parmi les étoiles A, des étoiles qui présentent des anomalies spectrales. Ces étoiles, bien que de composition chimique de surface différente, admettent une distribution spatiale identique à celle des étoiles de type spectral A normal, et devraient présenter le même comportement cinématique car elles ont le même âge.

Nous disposons de 2 échantillons de taille N de ces étoiles particulières :

- AM1 (N=51) défini par Gómez *et al.* (1981),

– Ap4-Ap5 (N=38) défini par Grenier *et al.* (1981).

A cause du petit nombre de données, les algorithmes EM et SEM n'ont pas permis de séparer de manière évidente d'éventuelles sous-populations gaussiennes dans ces 2 échantillons (voir Soubiran (1988)). Par contre, les 2 algorithmes (EB et SAEM) que nous présentons ici sont particulièrement adaptés à l'étude des petits échantillons.

Dans un premier temps, nous avons testé la fiabilité de ces 2 algorithmes sur un échantillon d'étoiles A normales A2V (N=97). Cet échantillon a déjà été analysé par des techniques graphiques, ainsi que par les algorithmes EM et SEM (voir Bougeard *et al.* (1989), Soubiran *et al.* (1989), Gómez *et al.* (1990)), et a toujours conduit à des résultats très stables car il a 2 composantes bien séparées.

Nous avons ensuite analysé la cinématique de ces étoiles A particulières à l'aide de l'algorithme SAEM. La méthode bayésienne nous a alors permis, en soumettant les données à un point de vue totalement différent, de confirmer ou d'infirmer les solutions proposées par SAEM.

Dans cette étude, nous avons travaillé sur les vitesses U et V, qui sont les plus discriminantes (voir Gómez *et al.* (1990)). Mais il est possible d'adapter ces méthodes au cas tri-dimensionnel et à des distributions autres que les lois gaussiennes.

2. L'algorithme SAEM

L'estimation par maximum de vraisemblance constitue l'approche la plus répandue pour l'identification de paramètres d'un mélange. Cette approche nécessite l'emploi d'algorithmes itératifs dont le plus utilisé est *l'algorithme EM* (voir Dempster *et al.* (1977) et Redner et Walker (1984)).

Cet algorithme conduit dans de nombreux cas à des estimations satisfaisantes des paramètres. Néanmoins, il peut s'avérer décevant (lenteur excessive, dépendance très forte de la position initiale, convergence vers un col de la vraisemblance, etc.), particulièrement dans les situations délicates (composantes du mélange imbriquées, proportions des composantes déséquilibrées, etc.).

Une modification stochastique de l'algorithme EM, *l'algorithme SEM* (voir, par exemple, Celeux et Diebolt (1986)), permet en général de dépasser les limitations précédentes. Cet algorithme comporte une étape stochastique (*étape 'S'*) qui vient s'intercaler entre *l'étape 'E'* de calcul des probabilités a posteriori d'appartenance des points aux composantes du mélange et *l'étape 'M'* de maximisation de la vraisemblance conditionnelle. L'étape 'S' est fondée sur un principe d'affectation aléatoire qui consiste à remplacer les provenances (inconnues) des points de l'échantillon de l'une des composantes du mélange par des affectations aléatoires des observations à ces composantes, suivant les probabilités a posteriori calculées à l'étape 'E'. L'étape 'M' se construit alors sur la base de ces affectations aléatoires. Les perturbations ainsi introduites par SEM lui permettent, en général, d'éviter les écueils de EM.

Les défauts de l'algorithme EM sont particulièrement gênants pour les petits échantillons. En effet, pour N petit, la fonction de vraisemblance est souvent entachée de nombreux maxima locaux, instables et sans intérêt statistique, mais qui accentuent la dépendance de EM vis-à-vis de sa position initiale et rendent difficile l'utilisation de cet algorithme. De plus, l'algorithme SEM ne pallie pas les limitations de EM dans ces situations car les perturbations aléatoires de l'étape 'S' prennent alors trop d'importance, et l'algorithme SEM risque alors de faire disparaître à tort une ou plusieurs des composantes du mélange analysé.

L'algorithme SAEM (Celeux et Diebolt, 1990) a été conçu précisément pour conserver les qualités de SEM tout en évitant les défauts de EM, même pour de petits échantillons. Cet algorithme occupe en fait une place intermédiaire entre les deux méthodes précédentes. Comme SEM, il utilise un principe d'affectation aléatoire pour produire des perturbations dans la suite des estimations des paramètres, mais, grâce à une suite de nombres réels positifs (γ_n), qui décroît vers 0 quand n tend vers l'infini, il réduit l'intensité des perturbations aléatoires au fil des itérations et est ainsi plus fiable pour traiter de petits ensembles de données. De plus, comme EM, il propose à la convergence un estimateur ponctuel des paramètres. Il est par conséquent d'une plus grande simplicité d'utilisation que SEM dont l'estimateur converge seulement en distribution. La suite (γ_n) joue en fait le même rôle que la *température* dans les algorithmes de type *recuit simulé* (voir, par exemple, van Laarhoven (1988)), d'où le nom SAEM : *Simulated Annealing EM*.

Nous décrivons maintenant l'algorithme SAEM en nous limitant au cadre qui nous intéresse ici, à savoir l'identification d'un mélange gaussien bidimensionnel. On dispose d'un échantillon $x_1, ..., x_N$ d'une variable aléatoire à valeurs dans R^2, de densité

$$h(x) = p_1\varphi(x|\theta_1, \Sigma_1) + p_2\varphi(x|\theta_2, \Sigma_2),$$

avec $0 < p_1 < 1, p_2 = 1 - p_1$, et $\varphi(.|\theta, \Sigma)$ désignant la densité d'une loi normale de moyenne θ et de matrice de variance Σ. Il s'agit d'estimer le paramètre $\phi = (p_1, \theta_1, \Sigma_1, \theta_2, \Sigma_2)$. Une itération de l'algorithme SAEM, qui à $\phi^{(n)}$ associe $\phi^{(n+1)}$, se décompose ainsi :

Etape E : pour $j = 1, 2$, calcul des probabilités a posteriori

$$t_j^n(x_i) = \frac{p_j^n \varphi(x_i|\theta_j^n, \Sigma_j^n)}{\sum_{\ell=1}^{2} p_\ell^n \varphi(x_i|\theta_\ell^n, \Sigma_\ell^n)},$$

d'appartenance des points x_i à chaque composante du mélange $(i = 1, ..., N)$.

Etape S : tirage pour chaque x_i de la variable aléatoire $z_1^n(x_i)$ suivant une loi de Bernoulli de paramètre $t_1^n(x_i)$. On pose $z_2^n(x_i) = 1 - z_1^n(x_i)$. Si $z_1^n(x_i) = 1$, x_i est affecté à la première composante, sinon il est affecté à la seconde composante. Si l'une des deux classes ainsi obtenues a un cardinal plus petit que 2, l'algorithme s'arrête. Sinon,

Etape A : calcul des quantités

$$r_j^n(x_i) = t_j^n(x_i) + \gamma_n \left(z_j^n(x_i) - t_j^n(x_i)\right)$$

pour $i = 1, ..., N$ et $j = 1, 2$.

Etape M : calcul de $\phi^{(n+1)}$, estimateur du maximum de vraisemblance de ϕ, les $r_j^n(x_i)$ étant considérés artificiellement comme les probabilités a posteriori d'appartenance des points x_i aux composantes du mélange. Pour $j = 1, 2$, on en déduit les actualisations

$$p_j^{n+1} = \frac{1}{N} \sum_{i=1}^{N} r_j^n(x_i),$$

$$\theta_j^{n+1} = \frac{\sum_{i=1}^{N} r_j^n(x_i) x_i}{\sum_{i=1}^{N} r_j^n(x_i)},$$

$$\Sigma_j^{n+1} = \frac{\sum_{i=1}^{N} r_j^n(x_i)(x_i - \theta_j^{n+1})(x_i - \theta_j^{n+1})^t}{\sum_{i=1}^{N} r_j^n(x_i)},$$

où a^t désigne le transposé de a.

La position intermédiaire de SAEM vis-à-vis de EM et de SEM se lit bien sur la formule

$$r_j^n(x_i) = (1 - \gamma_n) t_j^n(x_i) + \gamma_n z_j^n(x_i).$$

En effet, formellement, l'étape 'M' pour l'algorithme EM conduit exactement aux mêmes formules que pour SAEM, les $t_j^n(x_i)$ prenant la place des $r_j^n(x_i)$, et il en est de même pour SEM, où cette fois, les $z_j^n(x_i)$ remplacent les $r_j^n(x_i)$. Ainsi, lorsque $\gamma_n = 1$, une itération de SAEM est exactement une itération de SEM et, lorsque $\gamma_n = 0$, SAEM et EM coïncident.

Sous des hypothèses très générales et si la suite (γ_n) converge suffisamment lentement vers 0, la suite $\phi^{(n)}$ engendrée par l'algorithme SAEM converge presque sûrement vers un maximum local de la vraisemblance (voir Celeux et Diebolt (1990)). Notons qu'en toute généralité, on ne peut assurer un tel résultat pour une suite générée par l'algorithme EM car elle peut converger, si N n'est pas suffisamment grand, vers un col de la vraisemblance (voir Redner et Walker (1984)). Par ailleurs, les théorèmes relatifs à l'algorithme SEM donnent des résultats de convergence en distribution vers une loi normale centrée sur l'estimateur convergent du maximum de vraisemblance, mais ils nécessitent également une taille N suffisamment grande pour être valides.

D'un point de vue pratique, le choix du mode de décroissance de la suite (γ_n) vers 0 est très important et peut s'avérer délicat. Comme pour les algorithmes de type recuit simulé, les meilleurs résultats sont obtenus avec un mode de convergence assez lent vers 0. Des simulations ont montré qu'alors SAEM remplissait bien son rôle et pouvait être recommandé de préférence à EM et à SEM pour analyser des mélanges à partir de petits échantillons (Celeux et Diebolt, 1990) : il évite mieux que EM la convergence vers une solution erronée et risque moins que SEM de faire disparaître à tort une des composantes du mélange.

3. Estimation par échantillonnage bayésien (EB)

Lorsque, en pratique, on considère un problème d'estimation, le contexte peut apporter des renseignements supplémentaires sur les paramètres contrôlant le modèle probabiliste retenu. Quand ces informations *a priori* (i.e. antérieures à l'observation) peuvent être traduites en une distribution de probabilité, π, dite *distribution a priori*, *l'analyse statistique bayésienne* permet d'en faire usage. Si les informations sont plus 'vagues', on se ramène généralement à des distributions π classiques dites *conjuguées*, dont les paramètres sont déterminés à partir de l'information *a priori* disponible. Nous allons exposer ci-dessous les principes de la modélisation bayésienne dans le cas des mélanges, le lecteur pouvant se reporter à Berger (1985) ou à Robert (1991) pour un exposé général.

L'algorithme EM et ses généralisations fournissent une approximation de l'estimateur du maximum de vraisemblance, qui est fondé sur des considérations asymptotiques (cf Redner et Walker (1984)). Au contraire, l'approche bayésienne a l'avantage théorique de fournir un estimateur exact pour les paramètres du mélange et de disposer de justifications à taille d'échantillon finie. De plus, cette utilisation de l'information *a priori*, même imprécise, a un effet stabilisateur sur les estimateurs, qui manque parfois aux méthodes précédentes pour de petits échantillons. (L'avantage de l'approche bayésienne est en fait surtout apparent pour les petits échantillons, puisqu'elle est asymptotiquement équivalente au maximum de vraisemblance.) Enfin, dans les situations où on se refuse à prendre en compte les informations *a priori*, de peur de 'fausser' l'inférence, l'approche bayésienne fournit une méthode de vérification des résultats des algorithmes précédents.

Tout d'abord, signalons que les mélanges de distributions nécessitent un traitement spécial lorqu'on les aborde sous l'angle bayésien. En effet, bien qu'il soit possible de proposer formellement un estimateur des paramètres du modèle, cet estimateur ne peut être calculé en un temps raisonnable, même pour des tailles d'échantillon modestes (40 étant déjà trop grand...) Ce problème, exposé dans Diebolt et Robert (1990a), nécessite l'appel à des procédures itératives qui rapprochent, en pratique, la modélisation bayésienne des méthodes précédentes, pour lesquelles un calcul explicite de l'estimateur du maximum de vraisemblance est également impossible.

Dans le cas particulier d'un mélange de deux lois gaussiennes bidimensionnelles, nous supposerons que l'information *a priori* disponible se modélise au travers d'une loi *a priori* dite *normale-inverse Wishart*. Plus précisément, si

$$x \sim p_1 \mathcal{N}_2(\theta_1, \Sigma_1) + p_2 \mathcal{N}_2(\theta_2, \Sigma_2),$$

les lois *a priori* retenues sur les paramètres p_1, θ_j et Σ_j seront

$$\begin{aligned} \theta_j | \Sigma_j &\sim \mathcal{N}_2(\mu_j, \frac{1}{t_j}\Sigma_j), \\ \Sigma_j^{-1} &\sim \mathcal{W}_2(k_j, W_j), \\ p_1 &\sim \mathcal{B}e(\alpha, \beta), \end{aligned} \tag{3.1}$$

où $\mathcal{W}_2(k_j, W_j)$ est la *loi de Wishart*, généralisation multidimensionnelle de la loi du chi-deux (voir Eaton, 1983 ou Anderson, 1984) et $\mathcal{B}e(\alpha, \beta)$ la loi bêta. Les hyperparamètres (μ_j, t_j, k_j, W_j, α et β) de ces lois doivent bien sûr être déterminés pour pouvoir effectuer l'analyse bayésienne. Le choix de ces hyperparamètres se fait au moyen des moments *a priori* :

$$E^{\pi}[p_1] = \frac{\alpha}{\alpha+\beta} \quad , \qquad \mathrm{var}^{\pi}(p_1) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \quad ,$$

$$E^{\pi}[\theta_j] = \mu_j \quad , \qquad \mathrm{var}^{\pi}(\theta_j) = \frac{1}{t_j(k_j-3)}\, W_j^{-1} \quad ,$$

$$E^{\pi}[\Sigma_j] = \frac{1}{k_j-3}\, W_j^{-1} \quad , \qquad \mathrm{var}^{\pi}(\sigma_{tt}^j) = \frac{2(w_{tt}^j)^{-4}}{(k_j-2)^2(k_j-4)} \quad ,$$

sauf pour k_j qui est généralement choisi égal à 6 pour des raisons 'non informatives' (comme maximisant la variance *a priori* sur Σ_j). Nous reviendrons au § IV-2 sur l'influence d'une modification de ces hyperparamètres sur l'estimation de ϕ. On choisit ces familles particulières de lois car elles conduisent à des estimateurs analytiquement 'explicites' dans le cas de très petits échantillons (au sens où les lois *a posteriori* sont de la même famille, voir Diebolt et Robert (1990a)), et surtout permettent des approximations directes, grâce à des simulations, pour les plus grands échantillons (voir ci-dessous), les deux aspects étant liés.

De même que l'algorithme SAEM, la *méthode d'échantillonnage bayésien* (EB) propose une suite $\phi^{(n)}$ d'estimations des paramètres. On montre (voir Diebolt et Robert, 1990b) qu'elle converge géométriquement vers l'estimateur de Bayes du paramètre ϕ,

$$\begin{aligned}\phi^{\pi}(x_1, \ldots, x_N) &= E^{\pi}[\phi | x_1, \ldots, x_N] \\ &= \int \phi \pi(\phi | x_1, \ldots, x_N) d\phi,\end{aligned}$$

avec

$$\phi = (p_1, \theta_1, \Sigma_1, \theta_2, \Sigma_2).$$

L'actualisation de $\phi^{(n)}$ repose sur une séparation aléatoire de l'échantillon $x_1, \ldots, x_N$ en deux sous-échantillons, suivant leur vraisemblance d'appartenance à la première ou à la seconde composante. Chaque observation x_i $(1 \leq i \leq N)$ est donc affectée à la première composante avec la probabilité

$$t_1^n(x_i) = \frac{p_1^n \varphi(x_i | \theta_1^n, \Sigma_1^n)}{p_1^n \varphi(x_i | \theta_1^n, \Sigma_1^n) + p_2^n \varphi(x_i | \theta_2^n, \Sigma_2^n)}.$$

Les deux sous-échantillons ainsi formés, $y_1, \ldots, y_m$ et $z_1, \ldots, z_{N-m}$, servent alors à générer les nouvelles valeurs des paramètres $\phi^{(n+1)}$, suivant les lois :

$$\left(\Sigma_1^{n+1}\right)^{-1} \sim \mathcal{W}_2(k_1^*, W_1^*),$$
$$\theta_1^{n+1}|\Sigma_1^{n+1} \sim \mathcal{N}_2\left(\mu_1^*, \frac{1}{t_1^*}\Sigma_1^{n+1}\right),$$
$$p_1^{n+1} \sim \mathcal{B}e(\alpha + m, \beta + N - m),$$

avec

$$\mu_1^* = \frac{t_1\mu_1 + m\bar{y}}{t_1 + m},$$
$$k_1^* = m + k_1, \qquad t_1^* = m + t_1,$$
$$W_1^* = \left(W_1^{-1} + S_1^* + \frac{t_1 m}{t_1 + m}(\bar{y} - \mu_1)(\bar{y} - \mu_1)^t\right)^{-1},$$
$$\bar{y} = \frac{1}{m}\sum_{\ell=1}^{m} y_\ell, \quad S_1^* = \sum_{\ell=1}^{m}(y_\ell - \bar{y})(y_\ell - \bar{y})^t,$$

et de même pour θ_2 et Σ_2. Cette loi sur (θ_1, Σ_1) est en fait la loi *a posteriori* pour le sous-échantillon $y_1, \ldots, y_m$. L'affectation aléatoire des observations à chaque itération de l'algorithme permet donc, en décomposant ainsi l'échantillon en deux sous-échantillons gaussiens, d'éviter le calcul de la loi *a posteriori* pour le mélange, l'inconvénient étant évidemment que les valeurs actualisées $\phi^{(n)}$ ne sont pas exactement tirées de cette loi *a posteriori*, $\pi(\phi|x)$. On montre cependant que la suite $\phi^{(n)}$ peut être utilisée pour approcher l'estimateur de Bayes au moyen de

$$\tilde{\phi}(x) = \frac{1}{K}\sum_{n=n_0+1}^{n_0+K} \phi^n \quad \text{ou} \quad \hat{\phi}(x) = \frac{1}{K}\sum_{n=n_0+1}^{n_0+K} E^\pi[\phi|y_1^n, \ldots, y_{m^n}^n, z_1^n, \ldots, z_{N-m^n}^n],$$

m^n étant la taille du premier sous-échantillon à l'étape n, n_0 représentant la durée de la phase d'initialisation de l'algorithme, et K étant choisi assez grand. En raison du théorème de Rao-Blackwell, la seconde moyenne est préférable, au moins lorsque les $\phi^{(n)}$ sont effectivement indépendants.

On trouvera dans Diebolt et Robert (1990b, c) des résultats de convergence plus précis pour l'algorithme, ainsi que des commentaires sur la simulation des lois ci-dessus. Notons que la génération de lois de Wishart s'opère très aisément à partir d'un générateur normal.

4. Les résultats

4.1. Les résultats de SAEM

Dans les essais qui sont présentés, SAEM a été utilisé avec un mode de décroissance lente : $\gamma_n = \cos(n\pi/2I)$, I étant le nombre d'itérations de l'algorithme (ici I = 200). Pour chaque échantillon, nous avons fait les 2 essais suivants :

– (a) SAEM est lancé 100 fois en tirant au hasard les conditions initiales,

– (b) SAEM est lancé 100 fois en partant d'une position initiale "lue" sur le nuage (U,V).

La méthode (a) n'a donné des résultats satisfaisants que pour A2V. En effet, pour AM1 et Ap4-Ap5, la tendance générale de l'algorithme a été de faire disparaître une composante du mélange, les quelques réponses à 2 composantes oscillant entre plusieurs solutions. Par contre, (b) a donné des résultats très stables pour A2V et AM1. Pour l'échantillon Ap4-Ap5, sur 100 résultats on obtient 65 fois la même solution et 33 fois un autre maximum local de la vraisemblance. Ces différents résultats sont présentés sur les figures 1 à 3, où les ellipses représentent les régions de confiance à 61% (un écart-type) pour les valeurs estimées des paramètres.

D'après ces résultats, on peut conclure que SAEM permet de séparer 2 populations gaussiennes, même pour un petit échantillon. Plus le nombre de données est petit, plus l'algorithme est sensible aux conditions initiales. L'existence de 2 composantes est bien confirmée pour les trois échantillons, mais le plus petit des trois (Ap4-Ap5) pose un problème car SAEM peut converger vers 2 solutions, ceci à cause de quelques points de l'échantillon dont l'appartenance à l'une ou l'autre des composantes n'est pas clairement définie. Du point de vue astrophysique, l'existence d'une composante d'écart-type 2.4 km/s (première solution) n'est pas significative car cette valeur est du même ordre que les erreurs de mesures.

Pour cet échantillon, nous avons effectué la vérification suivante : en prenant comme conditions initiales chacune des solutions fournies par SAEM, nous avons appliqué l'algorithme SEM. Seule la deuxème solution a résisté aux perturbations aléatoires et a permis d'enregistrer 100 réponses identiques sur 100 essais. Cependant, la tendance à faire disparaître la plus petite des composantes (16 fois sur 100 essais) dans le cas de la première solution est un peu artificielle dans le mesure où cette composante, contenant un petit nombre de points, est peu susceptible de résister à des perturbations aléatoires.

4.2. Les résultats de EB

Nous allons détailler ici la démarche que nous avons choisi d'utiliser pour déterminer les lois a priori concernant l'échantillon A2V. Résumons l'information dont nous disposons : SAEM a proposé pour $\phi = (p_1, \theta_1, \Sigma_1, \theta_2, \Sigma_2)$ la solution

$$\widehat{\phi} = \left(0.65, \begin{pmatrix} -20 \\ -14 \end{pmatrix}, \begin{pmatrix} 182 & 15 \\ 15 & 105 \end{pmatrix}, \begin{pmatrix} 12 \\ 2 \end{pmatrix}, \begin{pmatrix} 45 & 2 \\ 2 & 33 \end{pmatrix}\right).$$

Il semble donc que l'échantillon est formé de 2 composantes dont les moyennes sont approximativement $U_1 = -20\ km/s$, $V_1 = -14\ km/s$, $U_2 = 12\ km/s$ et $V_2 = 2\ km/s$. On se libère de l'influence de SAEM en considérant une information beaucoup plus vague sur la proportion p_1 et sur les matrices de variance Σ_j. Ceci conduit à les choisir plus grandes que les matrices de variance de $\widehat{\phi}$, car on augmente ainsi la variabilité a priori des paramètres autour de leur moyenne. On

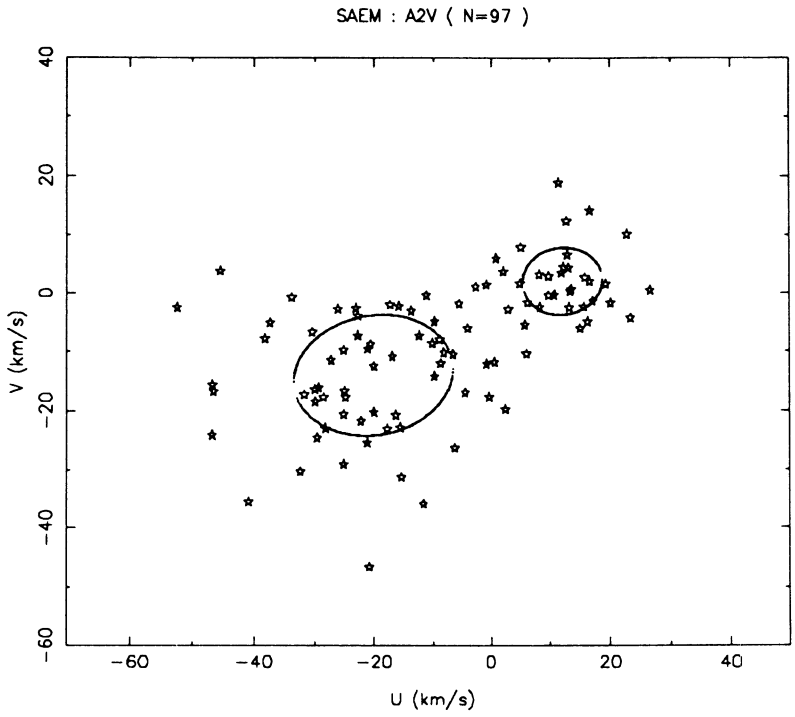

FIGURE 1. – Distribution de (U,V) pour l'échantillon A2V (N = 97)
Régions de confiance à 61% des paramètres calculés par SAEM.
La solution représentée correspond à 98 réponses sur 100

choisit de prendre comme conditions initiales de EB :

$$\phi^{(0)} = \left(0.50, \begin{pmatrix} -20 \\ -14 \end{pmatrix}, \begin{pmatrix} 400 & 50 \\ 50 & 400 \end{pmatrix}, \begin{pmatrix} 12 \\ 2 \end{pmatrix}, \begin{pmatrix} 150 & 20 \\ 20 & 150 \end{pmatrix}\right)$$

et comme lois a priori :

$$\theta_1|\Sigma_1 \sim \mathcal{N}_2\left(\begin{pmatrix} -20 \\ -14 \end{pmatrix}, \begin{pmatrix} 4 & 0.5 \\ 0.5 & 4 \end{pmatrix}\right) \simeq \mathcal{N}_2(\theta_1^{(0)}, \frac{1}{t_1}\Sigma_1^{(0)}) \quad \text{avec } t_1 = 10,$$

$$\theta_2|\Sigma_2 \sim \mathcal{N}_2\left(\begin{pmatrix} 12 \\ 2 \end{pmatrix}, \begin{pmatrix} 4 & 0.5 \\ 0.5 & 4 \end{pmatrix}\right) \simeq \mathcal{N}_2(\theta_2^{(0)}, \frac{1}{t_2}\Sigma_2^{(0)}) \quad \text{avec } t_2 = 37.5,$$

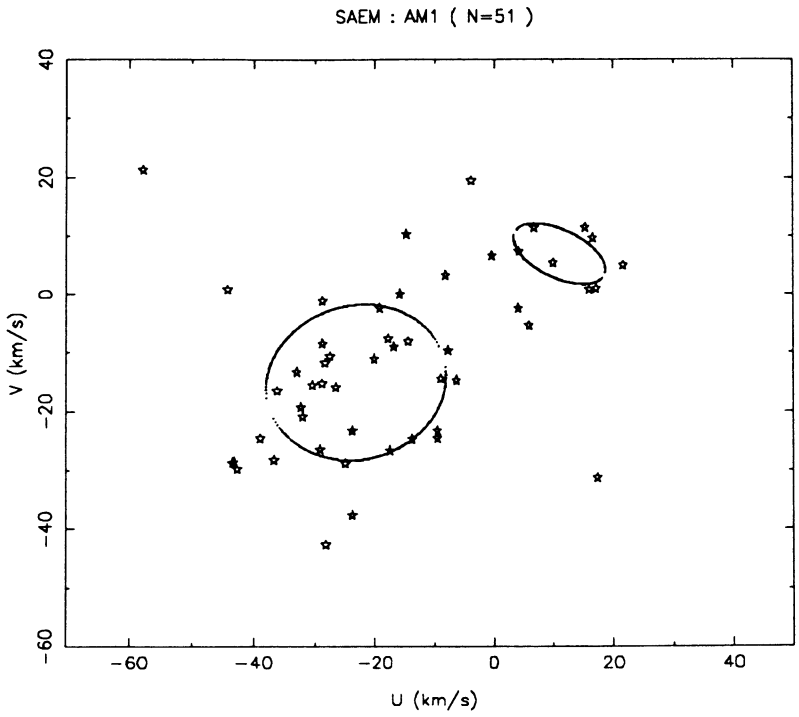

FIGURE 2. – Distribution de (U,V) pour l'échantillon AM1 (N = 51)
Régions de confiance à 61% des paramètres calculés par SAEM.
(91 réponses sur 100)

c'est-à-dire que l'on a choisi t_j en estimant connaître les moyennes des 2 composantes à 4 km/s près, et

$$\Sigma_1^{-1} \sim \mathcal{W}_2\left(6, \frac{1}{3}\begin{pmatrix} 400 & 50 \\ 50 & 400 \end{pmatrix}^{-1}\right),$$
$$\Sigma_2^{-1} \sim \mathcal{W}_2\left(6, \frac{1}{3}\begin{pmatrix} 150 & 20 \\ 20 & 150 \end{pmatrix}^{-1}\right),$$

Le paramètre $k_j = 6$ correspond au cas non-informatif car on préfère ne pas accorder trop de confiance à l'information concernant les variances, et induit le facteur $\frac{1}{3}$ puisque si $\Sigma^{-1} \sim \mathcal{W}_p(k, W)$, $E[\Sigma] = \frac{1}{k-3}W^{-1}$. On choisit $p_1 \sim \mathcal{B}e(0.5, 0.5)$, en considérant que nous n'avons aucune information a priori sur la proportion du mélange.

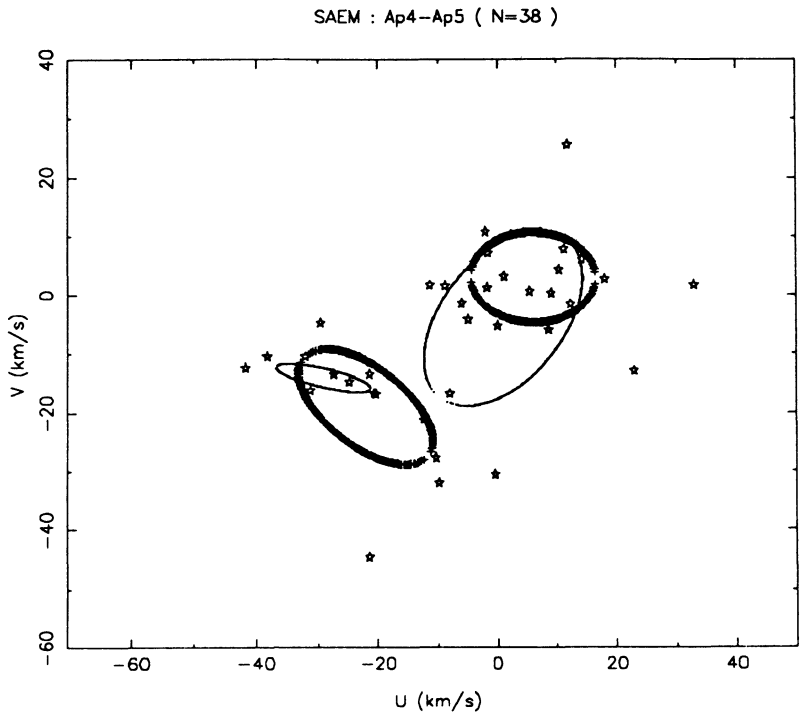

FIGURE 3. – Distribution de (U,V) pour l'échantillon Ap4-Ap5 (N = 38)
Régions de confiance à 61% des paramètres calculés par SAEM.
– en trait fin, 1ère solution (65 réponses sur 100)
– en trait gras, 2ème solution (33 réponses sur 100)

En calculant la moyenne des $\phi^{(n)}$ pour $n = 100, \ldots, 500$, on obtient

$$\widetilde{\phi} = \left(0.64, \begin{pmatrix} -20 \\ -14 \end{pmatrix}, \begin{pmatrix} 195 & 22 \\ 22 & 122 \end{pmatrix}, \begin{pmatrix} 12 \\ 2 \end{pmatrix}, \begin{pmatrix} 57 & 7 \\ 7 & 46 \end{pmatrix}\right)$$

On voit que $\widetilde{\phi}$ s'est nettement rapproché de $\widehat{\phi}$ par rapport aux conditions initiales. Les moyennes des 2 composantes sont très stables. Nous pouvons alors renforcer l'a priori et relancer EB avec :

$$\phi^{(0)} = \left(0.64, \begin{pmatrix} -20 \\ -14 \end{pmatrix}, \begin{pmatrix} 195 & 22 \\ 22 & 122 \end{pmatrix}, \begin{pmatrix} 12 \\ 2 \end{pmatrix}, \begin{pmatrix} 57 & 7 \\ 7 & 46 \end{pmatrix}\right)$$

et les lois a priori :

$$\theta_1|\Sigma_1 \sim \mathcal{N}_2\left(\begin{pmatrix}-20\\-14\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right) \ (t_1 = 150),$$
$$\theta_2|\Sigma_2 \sim \mathcal{N}_2\left(\begin{pmatrix}12\\2\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right) \ (t_2 = 50).$$

Ici, l'information sur les moyennes θ_1 et θ_2 est plus précise, on estime les connaître à $\pm 1 km/s$. On prend de plus :

$$\Sigma_1^{-1} \sim \mathcal{W}_2\left(10, \frac{1}{7}\begin{pmatrix}195 & 22\\22 & 122\end{pmatrix}^{-1}\right),$$
$$\Sigma_2^{-1} \sim \mathcal{W}_2\left(10, \frac{1}{7}\begin{pmatrix}57 & 7\\7 & 46\end{pmatrix}^{-1}\right),$$

$k_j = 10$ renforce l'a priori sur les variances, et $p_1 \sim \mathcal{B}e(17.78, 10)$, avec $\beta = 10$, stabilise l'algorithme sur la proportion $p_1 = 0.64$.

La solution est alors :

$$\widetilde{\phi} = \left(0.65, \begin{pmatrix}-20\\-14\end{pmatrix}, \begin{pmatrix}186 & 20\\20 & 108\end{pmatrix}, \begin{pmatrix}12\\2\end{pmatrix}, \begin{pmatrix}45 & 3\\3 & 36\end{pmatrix}\right)$$

On constate que la variance de la première composante a très peu bougé. On réinjecte ces nouvelles valeurs de p_1, Σ_1, Σ_2 dans l'algorithme. En agissant de cette manière itérative sur l'input (conditions initiales et lois a priori) de EB, on obtient une stabilisation complète sur p_1 et Σ_1 en 2 nouvelles exécutions. Il faut une exécution supplémentaire pour arriver à la convergence de Σ_2. La solution finale est la suivante :

$$\widetilde{\phi} = \left(0.66, \begin{pmatrix}-20\\-14\end{pmatrix}, \begin{pmatrix}190 & 23\\23 & 109\end{pmatrix}, \begin{pmatrix}12\\2\end{pmatrix}, \begin{pmatrix}41 & 1\\1 & 33\end{pmatrix}\right)$$

On peut constater qu'à la convergence, on obtient un estimateur de Bayes très proche du maximum de vraisemblance.

La solution de SAEM pour l'échantillon AM1 est :

$$\widehat{\phi} = \left(0.82, \begin{pmatrix}-23\\-15\end{pmatrix}, \begin{pmatrix}225 & 16\\16 & 178\end{pmatrix}, \begin{pmatrix}11\\7\end{pmatrix}, \begin{pmatrix}59 & -21\\-21 & 27\end{pmatrix}\right)$$

Nous avons utilisé ce résultat exactement de la même manière que pour l'échantillon A2V, en négligeant l'information sur p_1 et Σ_j, et en partant de matrices

de variance très grandes par rapport à $\widehat{\Sigma}_j$. En réinjectant de manière itérative le résultat de EB en input, on arrive en 6 exécutions à la solution :

$$\widetilde{\phi} = \left(0.81, \begin{pmatrix} -22 \\ -15 \end{pmatrix}, \begin{pmatrix} 222 & 20 \\ 20 & 183 \end{pmatrix}, \begin{pmatrix} 11 \\ 7 \end{pmatrix}, \begin{pmatrix} 56 & -15 \\ -15 & 26 \end{pmatrix}\right).$$

Le cas de l'échantillon Ap4-Ap5 est plus complexe dans la mesure où SAEM propose les 2 solutions :

$$\widehat{\phi}_1 = \left(0.23, \begin{pmatrix} -29 \\ -14 \end{pmatrix}, \begin{pmatrix} 60 & -14 \\ -14 & 6 \end{pmatrix}, \begin{pmatrix} 1 \\ -4 \end{pmatrix}, \begin{pmatrix} 177 & 92 \\ 92 & 220 \end{pmatrix}\right)$$
$$\widehat{\phi}_2 = \left(0.42, \begin{pmatrix} -22 \\ -19 \end{pmatrix}, \begin{pmatrix} 124 & -65 \\ -65 & 97 \end{pmatrix}, \begin{pmatrix} 6 \\ 3 \end{pmatrix}, \begin{pmatrix} 110 & -1 \\ -1 & 58 \end{pmatrix}\right)$$

On doit chercher un compromis entre ces 2 solutions pour en tirer une information a priori. Il paraît logique de démarrer l'algorithme sur une valeur de (θ_1, θ_2) à mi-chemin entre ces 2 solutions, mais il faut alors adapter le paramètre t_j, qui permet de régler l'incertitude sur les moyennes des composantes du mélange, de manière à ce que la loi de μ_j permette d'englober les 2 valeurs $\widehat{\theta}_j$ obtenues avec SAEM. Comme pour les 2 échantillons précédents, on choisit la loi non-informative sur p_1 et Σ_j, en partant de variances majorant très nettement les $\widehat{\Sigma}_j$ trouvées par SAEM.

Les conditions de départ de EB seront :

$$\phi^{(0)} = \left(0.50, \begin{pmatrix} -25.5 \\ -16.5 \end{pmatrix}, \begin{pmatrix} 200 & -60 \\ -60 & 200 \end{pmatrix}, \begin{pmatrix} 3.5 \\ -0.5 \end{pmatrix}, \begin{pmatrix} 400 & 60 \\ 60 & 400 \end{pmatrix}\right)$$
$$\theta_1|\Sigma_1 \sim \mathcal{N}_2\left(\begin{pmatrix} -25.5 \\ -16.5 \end{pmatrix}, \begin{pmatrix} 9 & -3 \\ -3 & 9 \end{pmatrix}\right) \simeq \mathcal{N}_2(\theta_1^{(0)}, \frac{1}{t_1}\Sigma_1^{(0)}) \quad \text{avec } t_1 = 22,$$
$$\theta_2|\Sigma_2 \sim \mathcal{N}_2\left(\begin{pmatrix} 3.5 \\ -0.5 \end{pmatrix}, \begin{pmatrix} 9 & 1 \\ 1 & 9 \end{pmatrix}\right) \simeq \mathcal{N}_2(\theta_2^{(0)}, \frac{1}{t_2}\Sigma_2^{(0)}) \quad \text{avec } t_2 = 44,$$
$$\Sigma_1^{-1} \sim \mathcal{W}_2\left(6, \frac{1}{3}\begin{pmatrix} 200 & -60 \\ -60 & 200 \end{pmatrix}^{-1}\right),$$
$$\Sigma_2^{-1} \sim \mathcal{W}_2\left(6, \frac{1}{3}\begin{pmatrix} 400 & 60 \\ 60 & 400 \end{pmatrix}^{-1}\right),$$
$$p_1 \sim \mathcal{B}e(0.5, 0.5).$$

Notons ici que la matrice de départ choisie pour Σ_1 appuie l'information donnée par la deuxième solution de SAEM, alors que la matrice choisie pour Σ_2 favorise la première solution.

Le premier résultat de EB montre une certaine hésitation de l'algorithme à se rapprocher d'une des 2 solutions de SAEM. $\widetilde{\theta}_1$, $\widetilde{\Sigma}_1$, $\widetilde{\theta}_2$ manifestent une tendance à aller vers la solution 2 alors que $\widetilde{p}_1$ et $\widetilde{\Sigma}_2$ sont très proches de la solution 1 :

$$\widetilde{\phi} = \left(0.35, \begin{pmatrix} -25 \\ -17 \end{pmatrix}, \begin{pmatrix} 132 & -63 \\ -63 & 113 \end{pmatrix}, \begin{pmatrix} 4 \\ 0 \end{pmatrix}, \begin{pmatrix} 171 & 45 \\ 45 & 155 \end{pmatrix}\right)$$

On réinjecte ces valeurs dans EB, mais sans renforcer l'a priori, car l'information que l'on vient d'obtenir est encore vague. Cette nouvelle exécution montre un très net avantage en faveur de la deuxième solution de SAEM. On choisit alors de renforcer l'a priori en prenant t_j tel que $\frac{\Sigma_j}{t_j}$ soit a peu près égale à $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, et en prenant $k_j = 10$, $\beta = 10$, et de manière itérative, le résultat de EB comme input. En relançant 6 nouvelles fois EB, l'algorithme se stabilise sur :

$$\widetilde{\phi} = \left(0.41, \begin{pmatrix} -23 \\ -18 \end{pmatrix}, \begin{pmatrix} 122 & -65 \\ -65 & 98 \end{pmatrix}, \begin{pmatrix} 5 \\ 2 \end{pmatrix}, \begin{pmatrix} 114 & 2 \\ 2 & 62 \end{pmatrix}\right)$$

La solution du maximum de vraisemblance la plus fréquente (65% d'occurences) a été rejetée par EB itératif. On n'a pourtant considéré qu'une information très vague sur les moyennes des 2 composantes.

Les résultats obtenus avec l'algorithme EB sur les 3 échantillons sont présentés sur les figures 4 à 6, dans les mêmes conditions que les solutions de SAEM (régions de confiance à 61 %).

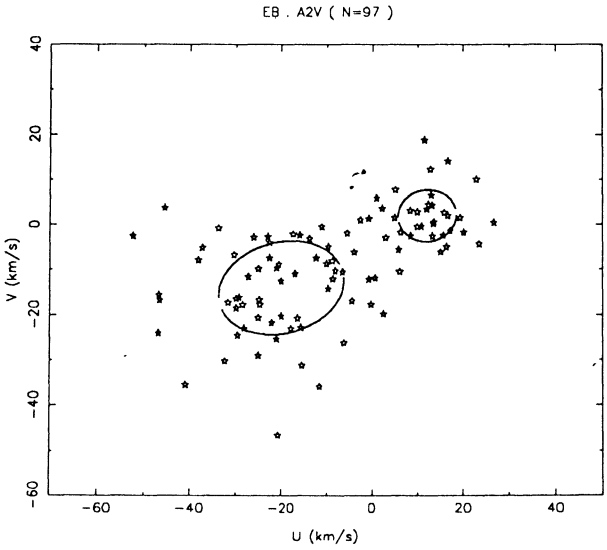

FIGURE 4. – Distribution de (U,V) pour l'échantillon A2V (N = 97)
Régions de confiance à 61% des paramètres calculés par EB itératif.

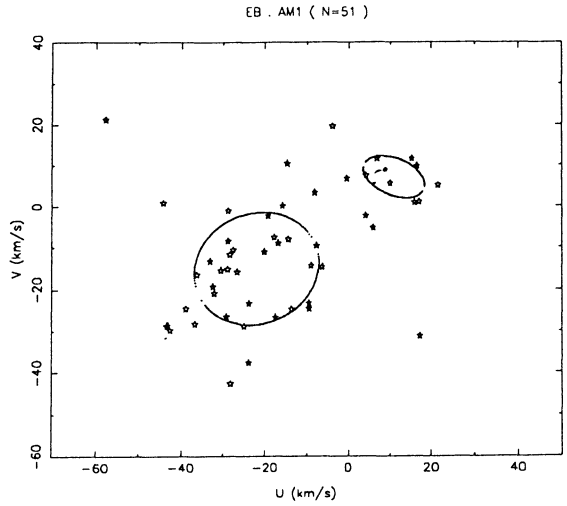

FIGURE 5. – Distribution de (U,V) pour l'échantillon AM1 (N=51) Régions de confiance à 61% des paramètres calculés par EB itératif.

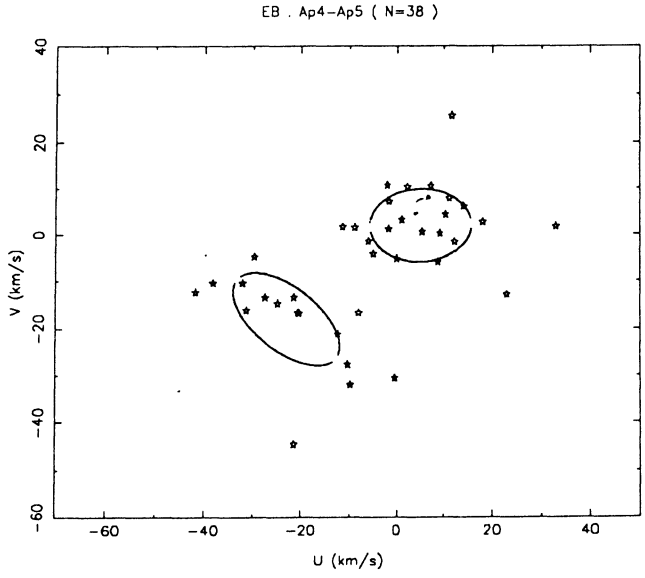

FIGURE 6. – Distribution de (U,V) pour l'échantillon Ap4-Ap5 (N=38) Régions de confiance à 61% des paramètres calculés par EB itératif.

5. Discussion

Nous avons récapitulé dans le tableau 1 les différents résultats obtenus par SAEM et EB sur les 3 échantillons.

L'échantillon A2V a servi à montrer la fiabilité de SAEM et EB. Les solutions des 2 algorithmes sont cohérentes et très proches de la solution trouvée précédemment par Gómez *et al.* (1990). L'échantillon AM1 a donné des résultats très stables et pratiquement identiques par les deux méthodes. L'échantillon Ap4-Ap5 montre l'interêt d'une approche mixte dans les cas les plus délicats. EB itératif a permis de choisir parmi 2 solutions proposées par SAEM. On aurait pu choisir la solution donnée la plus fréquemment par SAEM, mais nous n'y voyons aucune justification statistique. Par contre, EB itératif confronte les deux solutions selon un schéma clairement argumenté, et permet sans doute de choisir la solution la plus satisfaisante.

Nous pouvons tirer 2 conclusions de cette étude.

Du point de vue astrophysique, nous avons retrouvé dans la cinématique des étoiles A particulières les mêmes caractéristiques qui avaient été trouvées dans les étoiles A normales. L'existence de 2 sous-populations est confirmée.

Du point de vue statistique, nous avons montré qu'il est possible de traiter un mélange de distributions gaussiennes bi-dimensionnelles à partir d'un échantillon de moins de 40 points. Dans le cadre de l'analyse des petits échantillons, nous proposons la stratégie suivante :

– faire tourner SAEM afin de trouver les zones de stabilité de l'échantillon,

– définir les conditions initiales et lois a priori de EB à partir d'une moyenne des paramètres de position des solutions de SAEM et de matrices de variances notablement plus grandes que celles obtenues par SAEM,

– relancer EB jusqu'à stabilisation, en y réinjectant les solutions précédentes comme input, tout en renforçant progressivement la confiance dans les lois a priori.

	A2V (N=97)			AM1 (N=51)		Ap4-Ap5 (N=38)		
	SEM(*)	SAEM	EB	SAEM	EB	SAEM 1	SAEM 2	EB
p_1	0.64	0.65	0.66	0.82	0.81	0.23	0.42	0.41
U_1	-21	-20	-20	-23	-22	-29	-22	-23
V_1	-14	-14	-14	-15	-15	-14	-19	-18
σ_{U_1}	13	13	14	15	15	8	11	11
σ_{V_1}	10	10	10	13	14	2	10	10
U_2	12	12	12	11	11	1	6	5
V_2	2	2	2	7	7	-4	3	2
σ_{U_2}	6	7	6	8	7	13	10	11
σ_{V_2}	5	6	6	5	5	15	8	8

Tableau 1 : Récapitulation des résultats
(*) : voir Gómez *et al.* (1990)

Références

Anderson T.W. (1984) *An Introduction to Multivariate Statistical Analysis* (2nd edition). J. Wiley, NY.

Berger J.O. (1985) *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, NY.

Bougeard M.L., Arenou F., Soubiran C. et Grenier S. (1989) in *Errors, Bias and Uncertainties in Astronomy*, eds. F. Murtagh et C. Jaschek, Strasbourg, 11-14 September, 281-284.

Celeux G. et Diebolt J. (1986) L'algorithme SEM : un algorithme d'apprentissage probabiliste pour la reconnaissance des mélanges de densité. *Revue de Statistique Appliquée*, **34**(2), 35-52.

Celeux G. et Diebolt J. (1990) Une version de type recuit simulé de l'algorithme EM. *Notes Comptes Rendus Acad. Sciences Paris*, **310**, I, 119-124.

Dempster A., Laird N. et Rubin D. (1977) Maximum likelihood from incomplete data via the EM algorithm. *JRSS* (Ser. B), **39**, 1-38.

Diebolt J. et Robert C. (1990a) Bayesian estimation of finite mixture distributions. Part I : Theoretical aspects. Rapport tech. 110, LSTA, Université Paris VI.

Diebolt J. et Robert C. (1990b) Bayesian estimation of finite mixture distributions. Part II : Sampling implementation. Rapport tech. 111, LSTA, Université Paris VI.

Diebolt J. et Robert C. (1990c) Estimation des paramètres d'un mélange par échantillonnage bayésien. *Notes Comptes Rendus Acad. Sciences Paris*, **311**, I, 653-658.

Eaton M. (1983) *Multivariate Statistics*. J. Wiley, NY.

Gómez A.E., Grenier S., Jaschek M. et Heck A. (1981) The absolute magnitude of the Am stars. *Astron. Astrophys.*, **93**, 155-159.

Gómez A.E., Delhaye J., Grenier S., Jaschek C., Arenou F. et Jaschek M. (1990) Local kinematic properties of Population I (B5-F5)-type stars and galactic disk evolution. *Astron. Astrophys.*, **236**, 95-98.

Grenier S., Jaschek M., Gómez A.E., Jaschek C. et Heck A. (1981) The absolute magnitude of the Ap stars. *Astron. Astrophys.*, **100**, 24-27.

van Laarhoven P. (1988) *Theoretical and Computational Aspects of Simulated Annealing*. CWI Tract 51, Amsterdam.

Redner R. et Walker H. (1984) Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, **26**(2), 195-239.

Robert C. (1991) L'Analyse statistique bayésienne. Notes de D.E.A., ISUP, Université Paris VI. (à paraître chez Economica)

Soubiran C. (1988) Stage de D.E.A., Observatoire de Paris.

Soubiran C., Gómez A.E., Arenou F. et Bougeard M.L. (1989) in *Errors, Bias and Uncertainties in Astronomy*, eds. F. Murtagh et C. Jaschek, Strasbourg, 11-14 September, 407-410.