Multivariate mixture distributions in stellar kinematics: *Statistical and numerical stability of the SEM algorithm*

Bougeard M.L.(1)(2), Arenou F.(3)

(1) Univ. Paris X, IUT, 1 chemin Desvallières, F-92410 Ville d'Avray
(2) URA 1125 CNRS, Observatoire de Paris, F-75014 Paris
(3) URA D0335 CNRS, Observatoire de Paris-Meudon, F-92195 Meudon

Summary :*In this paper, we are concerned with the problem of estimating the parameters of a gaussian mixture density. Here we tackle the problem of analysing the convergence stability of the SEM process by performing several independent runs. Then, the results of the most stable SEM solution are compared to classical clustering and classification techniques. The method is applied to samples of A type population I stars.*

Keywords: - *Gaussian mixture - Maximum likelihood - Stellar kinematics*

1. Introduction, notations and statistical background

Of interest in this paper is the parametric family of mixture of k normal multivariate densities, i.e the family of density functions of the form

$$f(x,\theta) = p_1 f_1(x|m_1,\Sigma_1) + \ldots\ldots + p_k f_k(x|m_k,\Sigma_k), \quad x \in \mathbb{R}^n$$

where the proportions p_i, i=1-k are constrained to be nonnegative and to amount to one. Each component f_i, i=1-k of the mixture is a n-multivariate gaussian density with n-vector mean $m_i \in \mathbb{R}^n$ and (n x n) covariance matrix Σ_i.. Much has been written on methodology for estimating the unknown parameters $\theta(k) = (p_i, m_i, \Sigma_i, i=1\text{-}k)$. For a review, we refer to (Titterington & al 1985; Redner &Walker 1984). A class of iterative procedures for numerically approximating maximum likelihood estimates is known as EM algorithm.which is an algorithm used for incomplete data problems (Dempster & al 1977).

It acts as follows for a N-sample of observations $(x_j, j=1\text{-}N)$, N>>k, $x_j \in \mathbb{R}^n$. Let $\theta^c = (p_i^c, m_i^c, \Sigma_i^c, i=1\text{-}k)$ be a current approximate maximizer of the log-likelihood function of the sample $L(\theta)$, and θ^{c+1} the next one.The **Expectation** step computes for i=1-k , j=1-N,

$$p^c(i,x_j) = p_i^c f_i(x_j|m_i^c, \Sigma_i^c) / f(x_j,\theta^c)$$

that is an estimate of the posterior probability that x_j belongs to the ith component given the approximate estimate θ^c. Then, the **Maximization** step of the EM algorithm yields θ^{c+1} maximizing L(.), given by

$$p_i^{c+1} = (1/N)\sum_{j=1}^{N} p^c(i,x_j) \qquad m_i^{c+1} = \left(\sum_{j=1}^{N} p^c(i,x_j).x_j \right) / \left(N p_i^{c+1} \right)$$

$$\Sigma_i^{c+1} = \left(\sum_{j=1}^{N} p^c(i,x_j).(x_j - m_i^{c+1})(x_j - m_i^{c+1})^t \right) / \left(N p_i^{c+1} \right)$$

The **EM** algorithm (Redner & Walker 1984) possesses several attractive properties (low computational cost, convergence, **constraints** on θ satisfied) compared to that of several alternative methods (Newton, scoring, quasi-Newton) for numerically approximating maximum likelihood estimates. The **SEM** algorithm, described in (Celeux &Diebolt, 1986), is a recent improvement of this algorithm : it incorporates a Stochastic step to accelerate the (a priori low) convergence. Nevertheless, even in the context of gaussian univariate mixture, the resulting likelihood surface is littered with **singularities** (Titterington & al 1985, ex. 4.3.2, p83). However, with a *good initialization* - obtained, for instance through graphical methods (Bougeard & al, 1989b)- and reasonable *sample size*, one can expect a (S.)E.M. iteration sequence to converge to a *local maximizer* of the log-likelihood function.

Of interest here is the convergence stability of the SEM algorithm in application to the study of the 3 dimensional x=(U,V,W) velocity distribution of 2 samples of A type population I stars, defined in Grenier & al (1985): an A2V sample (N=97) and an Ap sample (N=36).

2. Pertinence of a gaussian mixture model for the A2V sample

On a bidimensional graph UxV, one can foresee the presence of a potential mixture of two populations. The pertinence of the use of a parametric *gaussian* mixture model was shown in (Bougeard & al, 1989c) using the U component of the velocity. Nevertheless, it is known to be unsufficient to check univariate k' gaussian mixture for each variable U,V,W in order to be able to reject the possibility of a k mixture, k>k' (for example, see Titterington & al, 1985, fig 4.10, p68). So, a **multivariate** analysis has to be performed.

3. Numerical stability of the SEM algorithm

Assuming that the distribution of the (U,V,W) velocity sample is a mixture of multivariate gaussian components, we study the convergence stability of the SEM algorithm by performing several independent runs : each run represents a 200 iteration sequence.

3.1. Firstly, 31 runs of the SEM algorithm have been performed for the A2V sample using an initialization with **K=3** as upper bound of the number of components. Fig 1 shows the respective estimates in U found at each run. A 3 mixture solution appears as very unstable : 19 runs lead to a two mixture solution, 6 runs find no mixture.

3.2. At this stage, the same process has been performed by initializing SEM with **K=2.** The results are summarized on Fig 2 for the proportions (p_1>p_2) found at each run and on Fig 3 for the distributions in U,V,W.Table 1 gives the most stable solution (21 runs over 31) .Due to the fact that it is a multidimensional analysis, the result is slightly different in the U estimations from those obtained in an univariate context by Soubiran & al (1989), Bougeard & al (1989c). For interpretation in terms of star formation bursts, see Gómez & al (1989).

4. Statistical stability of the SEM estimates

The SEM algorithm provides also the probability for each star to belong to one of the estimated components (see Section 1). We compare the resulting classification with the results of classical multivariate data analysis methods.

Firstly, a Principal Component analysis (PCA) was performed on the correlation matrix (variables:U,V,W), by which it became apparent that the first axis (53% of the variance) was highly correlated with U,V lying in the galactic plane. Axis 2 (33.5% of the variance) is correlated with W perpendicular to this plane.The centers of the two gaussian components, projected as supplementary points, are highly correlated with the first axis and 7 stars are not in the same class if we perform a SEM univariate classification only on the U component.

A hierarchical classification was also performed with the reciprocal neighbour algorithm (Lebeaux, 1986; Lebart & al, 1984). The two clusters obtained by the top-level of the hierarchy are in good agreement with the SEM clusters (Bougeard & al, 1989a,b).

Finally, a linear discriminant analysis based on Fisher linear discriminant function and Mahalanobis distance was also performed to assess the discrimination between the two groups found by SEM . Only 15 stars which were on group 1 according to SEM are affected to group 2 ; this yields to an agreement of 84.5% of well classified stars.

5. Sensitivity of the SEM algorithm to the sample size

Finally, 31 SEM runs have been performed on the Ap sample (N=36), using an initialization with K=3 as upper bound of the number of the components. Two components were expected (Gómez & al,1989), but Fig 4 shows a high instability (no mixture is found in 20 runs over 31). The main reason is that the sample size is far too small and components are overlapping too much. We note, in the studied application, that a 3 (resp. 2) mixture model yields a SEM estimation of 2+3x3+3x6=29 (resp. 19) unconstrained parameters.

6. Conclusion

If the sample size is large enough and if the components are well separated, the SEM algorithm has been seen to provide a reasonable good convergence in the estimation of the parameters of gaussian mixtures in stellar kinematics. But it cannot be used rashly in other cases. In the particular case of the A2V sample studied here, SEM results have appeared as nearly stable and in good agreement with other clustering and classification techniques.

7. Acknowledgements

Principal component analysis, hierarchical classification and discriminant analysis were performed with SAS-ADDAD software on an IBM computer at CIRCE (F-Orsay). We thank Dr Celeux and Dr Diebolt (INRIA, F-Rocquencourt) for allowing the use of SEM software.

8. References

Bougeard M.L., Arenou F., Gómez A.; 1989a Bull.47th Int. Stat. Inst.: contr. papers, vol1,p161-162 , Paris Bougeard M.L., Arenou F., Gómez A.;1989b, "Mélanges gaussiens en cinématique stellaire. Une approche comparative de méthodes paramétriques et non paramétriques"(in preparation)
Bougeard M.L., Arenou F., Soubiran C., Gomez A., Grenier S.;1989c, (this issue)
Celeux G., Diebolt J.;1986, Revue Stat. Appl., 34, n°2,
Dempster A., Laird N., Rubin D.;1977,J. Royal Stat. Soc. B, 39, p1-38
Gómez, Delhaye, Grenier, Jaschek, Jaschek 1989: Astr. &Astroph. (soumis)
Grenier S. ,Gómez A., Jaschek C., Jaschek M., Heck A.: 1985 Astron. & Astrophys. 145, 331
Lebart L., Morineau A., Warwick K.;1984," Multivariate Descriptive Statistical Analysis", Wiley
Lebeaux M-O.;1986 Manuel de référence ADDADSAS, CIRCE, Orsay, france
Redner R., Walker H.;1984 SIAM,26,n°2, p195-239
Soubiran C., Bougeard M.L., Gomez A., Arenou F.; 1989 (this issue)
Titterington D., Smith A., Makov H.;1985, "Statistical Analysis of finite mixtures", Wiley

Table 1 : *Sample A2V (U,V,W) - SEM stable solution*

Component #1	**Component #2**
Proportion : 0.64	Proportion : 0.36
$m_1(u)$ $m_1(v)$ $m_1(w)$ -20.8 -14.3 -6.8	$m_2(u)$ $m_2(v)$ $m_2(w)$ 11.4 1.7 -7.4
variance-covariance matrix 176.9 10.8 17.5 10.8 103.4 -21.6 17.5 -21.6 75.3	variance-covariance matrix 50.2 3.8 -13.2 3.8 33.6 -3.8 - 13.2 -3.8 51.2

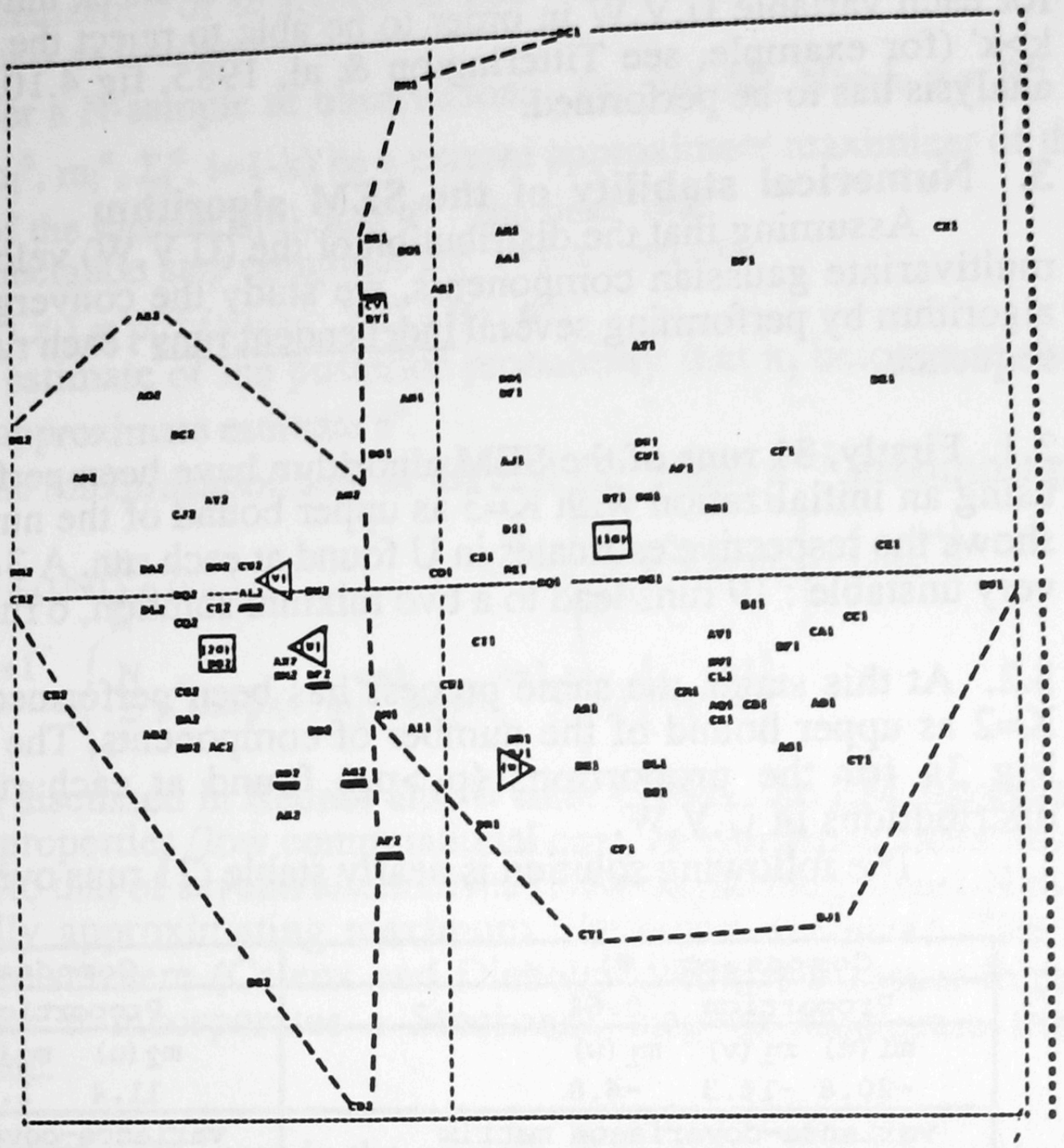

Fig 1 : Sample A2V (U,V,W). Factor pattern for Principal component analysis. Horizontal axis 1 - Vertical axis 2 (see text)

Fig 2 : Sample A2V (U,V,W) - SEM results per run
initialization with K=3 as upper bound of the number of components
Graph of mi(u) per run, the error bar is the square root of the respective variance vi(u) in the covariance matrix Σi

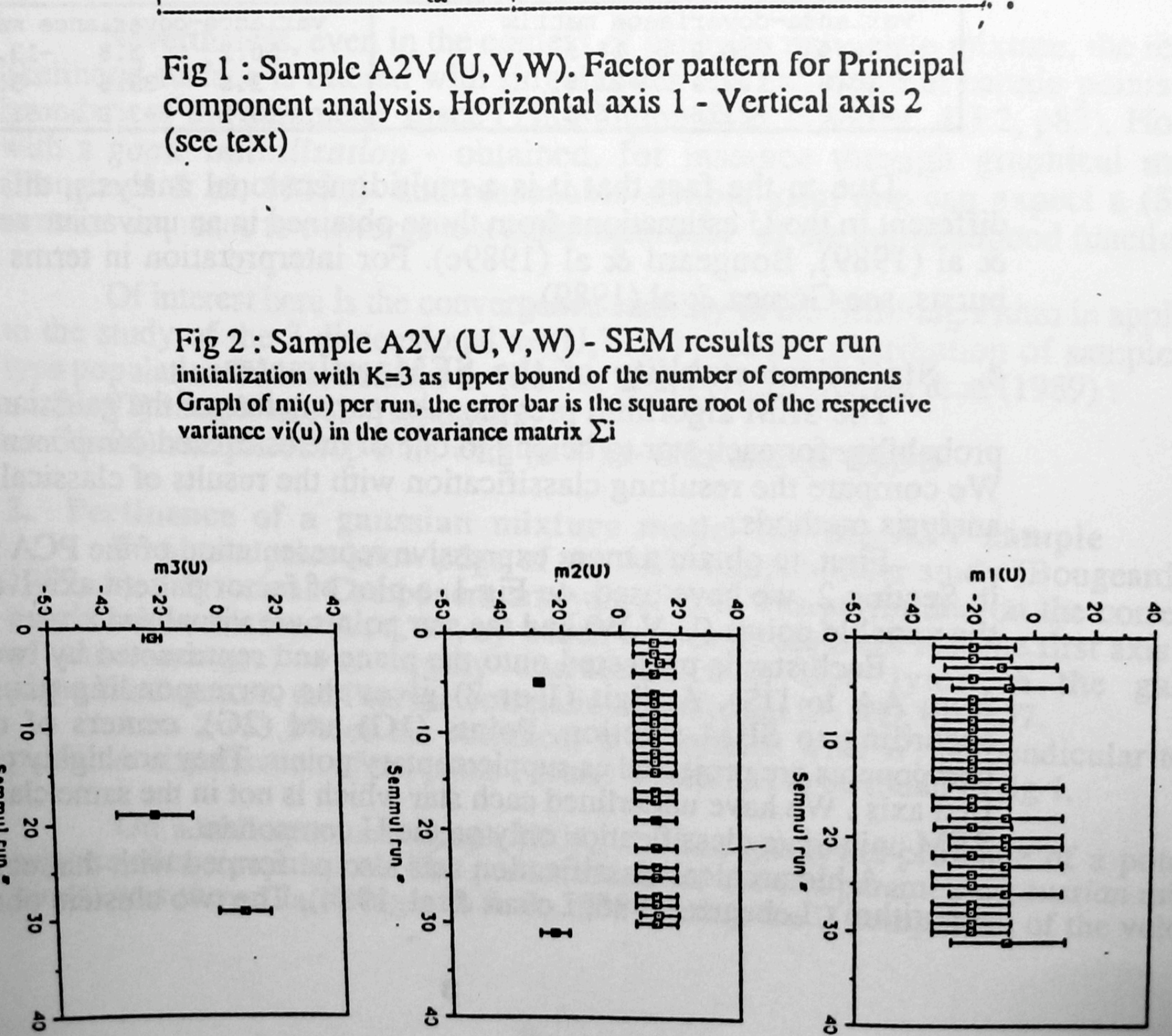

Fig 3b : same as Fig 2 ; initialization with K=2.

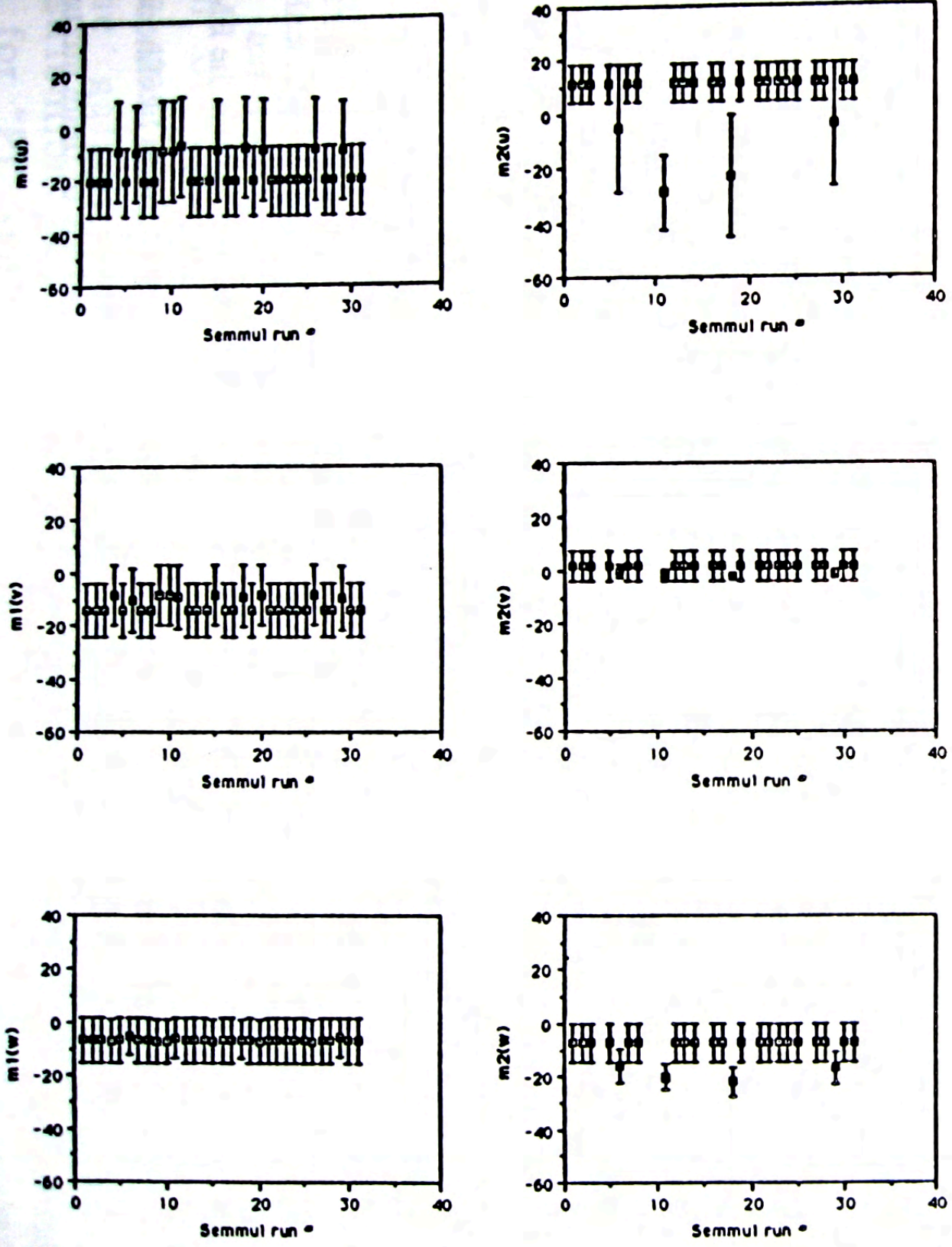

Fig 3a : Sample A2V (U,V,W) - SEM results per run
initialization with K=2 ; Graph of each proportion pi per run

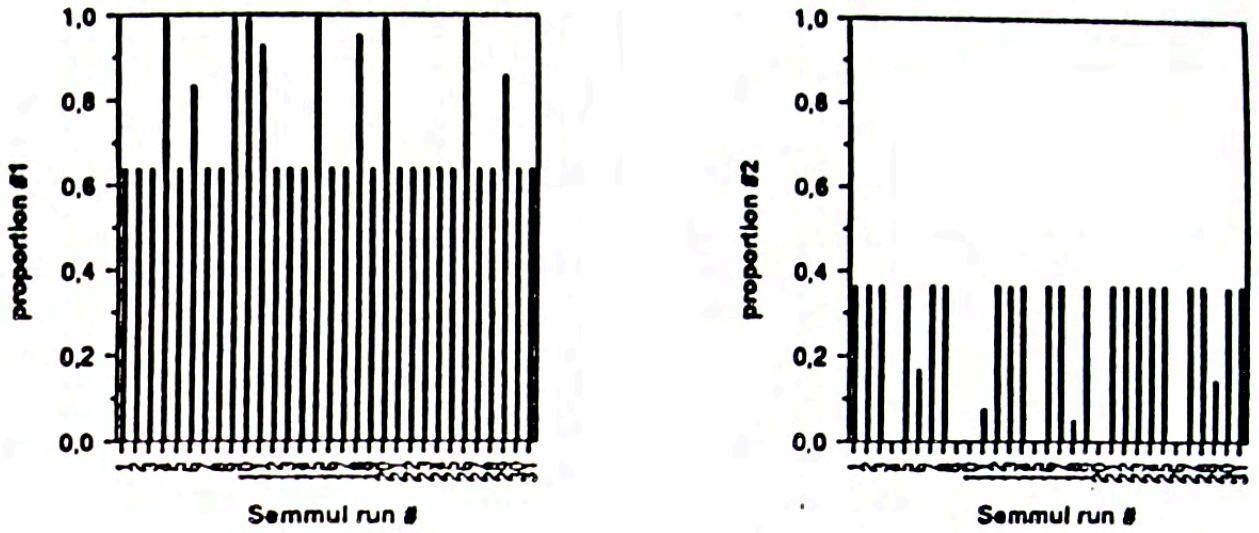

Fig 4 : Sample Ap (U,V,W) - SEM results per run
initialization with K=3 ; Graph of the proportions pi per run