Choice of the number of component clusters in mixture models by information criteria

Christian Olivier
SIC-IRCOM, UMR CNRS 6615
Université de Poitiers, France
Olivier@sic.univ-poitiers.fr

Frédéric Jouzel
laboratoire PSI
Université de Rouen, France
Frederic.Jouzel@univ-rouen.fr

Abdelaziz El Matouat
ENS de Fès, Maroc
Abdelaziz.El-Matouat@
univ-rouen.fr

Abstract

This paper considers the problem of choosing the number of component clusters within the context of the standard mixture of multivariate normal distributions. The problem to choose the number of clusters in a clustering procedure has already been dealt with, but still remains opened. We propose to use information criteria to solve this problem in the Gaussian mixture-model approach, which is nowadays a standard approach in clustering. The different criteria are presented and then compared with other well-known criteria on synthetic data sets. Often, the number of clusters k is unknown and needs to be estimated. A two-stage iterative maximum-likelihood procedure is used as a clustering technique to estimate the parameters of the mixture-model. A new criterion φ_β is derived and proposed as a criterion for choosing the number of clusters in the mixture-model context. For comparative purposes, Akaike's information criterion AIC *(1973) and Rissanen's 1978* MDL *criterion are also introduced in the mixture-model context. Numerical examples are shown on simulated normal data sets with a known number of mixture clusters to illustrate the significance of φ_β in choosing the number of clusters and the best fitting model. We demonstrate its efficiency and robustness through experimental results for synthetic mixture data sets.*

1 Introduction and statement of the problem

A general common problem in all clustering techniques is the difficulty of deciding of the number of clusters present in a given data set, cluster validity, and the identification of the appropriate number of clusters. This paper considers the problem of choosing the number of clusters within the context of unsupervised clustering in the framework of pattern recognition.

Most of the existing clustering procedures require prior knowledge of the number of clusters which is often unavailable and has to be estimated [17]. In the mixture-model cluster analysis, often the number of component clusters k is not known and needs to be estimated from the available observed data. This problem is known as the *cluster validation* problem [8]. Nevertheless, this aspect is often omitted in mixture type approaches [5, 16]. Despite the increased number of books appearing on finite mixture distributions such as [5], a relatively little work has been done concerning the choice of the number of component mixture clusters.

In this paper, the clustering viewpoint consists in identifying and describing the class distribution using a sample drawn from the mixture-model, and estimating the number of mixture clusters k in a non-subjective manner. To achieve this, we will use a two stage iterative maximum likelihhod (ML) procedure (EM algorithm) to estimate the parameters in the mixture-model, and develop a new information criterion φ_β as a new index for cluster validation.

The paper is organized as follows. In the next section, we introduce the ML procedure to estimate the parameters of the model. In section 3 we introduce the φ_β criterion to perform order estimate (choice of the number of clusters). We explain in part 3.2 the analytical expression of the criteria in the mixture-model case. Finally, we give in section 4 some results on synthetic datasets.

2 The standard mixture-model cluster analysis

There are a lot of clustering algorithms to perform unsupervised classification, for instance the k-means algorithm or hierarchical classification. A standard and powerful approach is available through a multivariate Gaussian mixture hypothesis. Data are assumed to

be independent realizations $x_1, x_2, ..., x_N$ of $\mathbb{R}^d$ coming from a mixture of k sub-populations $P_1, P_2, ..., P_k$, each cluster $P_j, 1 \leq j \leq k$, being characterized by a multivariate Gaussian probability density function (PDF):

$$f_j(x_i) = \frac{1}{(2\pi)^{\frac{d}{2}}} \times \frac{1}{\sqrt{|\Sigma_j|}} \exp -\frac{1}{2}(x_i - \mu_j)^T \Sigma_j^{-1}(x_i - \mu_j) \quad (1)$$

The finite mixture-model is the density composed of a sum of component densities. Throughout this work we will assume that the mixture densities are from the same family (ie Gaussian). Thus a typical mixture model is of the form:

$$g(x) = \sum_{j=1}^{k} p_j f_j(x) \quad (2)$$

where the $f_j, 1 \leq j \leq K$ are the mixture components, for example normal densities, and the $p_j, 1 \leq j \leq k$, are the mixture proportions such that $p_j > 0$ and $\sum_{j=1}^{k} p_j = 1$. Given a model of the form (2) and data $x_1, ..., x_N$, we must fit the parameters of the model to the data. This is done (most often) using the Expectation - Maximisation (EM) algorithm [3, 16]. This is an iterative method, which assumes an adequate initialization of the parameters values. First the expectation of component membership for each data and each component is computed (w_{ij}), given the current estimated value of the parameters (the E step), then the parameters are updated as the ML solutions conditioned on the w_{ij} (the M step):

- E (expectation) step: estimate w_{ij} by $\widehat{w}_{ij} = \frac{\widehat{p}_j f_j(x_i)}{\sum_{l=1}^{k} \widehat{p}_l f_l(x_i)}$
- M (maximisation) step: update the estimated parameters $\widehat{p}_j$, $\widehat{\mu}_j$, $\widehat{\Sigma}_j$:

$$\begin{aligned} \widehat{p}_j &= \frac{1}{N} \sum_{i=l}^{N} \widehat{w}_{ij} \\ \widehat{\mu}_j &= \frac{1}{N\widehat{p}_j} \sum_{i=l}^{N} \widehat{w}_{ij} x_i \\ \widehat{\Sigma}_j &= \frac{1}{N\widehat{p}_j} \sum_{i=l}^{N} \widehat{w}_{ij}(x_i - \widehat{\mu}_j)(x_i - \widehat{\mu}_j)^T \end{aligned}$$

This procedure is repeated until the parameters (or likelihood) stop changing by more than a given fixed amount.

The objective of the EM algorithm is to iteratively maximizing the log-likelihood function defined by:

$$L(k) = \sum_{i=1}^{N} \log g(x_i) = \sum_{i=1}^{N} \log \sum_{j=1}^{k} p_j f_j(x_i) \quad (3)$$

Here we assume that the number of components k is known. The estimation of the number of components k will be discussed in the next section. It should be noted that the mixtures we are considering, normal mixtures, are identifiable. This means that for finite k, there is only one way (up to reordering the terms) of writing the density as a mixture. This is not true for all component densities, for example uniform densities.

Unfortunately, the ML solution is not guaranteed to be unique for a given data set, nor is there necessarily a single local maximum.

We can notice that the EM algorithm performs a kind of *fuzzy partition* of the data. A *hard partition* can be obtained by assigning each observation vector to its component membership in the model in such a way:

$$\forall j, j \neq i, p_i f_i(x) > p_j f_j(x) \Rightarrow x \in P_i \quad (4)$$

so that each observation is assigned to the nearest component in the sense of the Mahalanobis distance.

3 Model selection criteria

The philosophy behind the choice of a model is to strike a balance between a good modeling and a reasonable number of parameters, which observes the principle of *parsimony*. Indeed, as the number of component clusters grows, the PDF approximation fits the empirical density increasingly tightly at the expense of its capacity to generalize. Moreover, a large number of clusters requires a great deal of computational ressources and time [12]. The idea is to use one objective decision which penalizes models with too many parameters. One way of doing so is to use a type of log-likelihood criterion usually called information criterion (IC), in reference to the information theory. These criteria take the form of a penalized likelihood function, that is the negative log-likelihood plus a penalty term which increases with the number of parameters. The most current of these criteria is Akaike´s information criterion (AIC) [1]:

$$AIC(k) = -2 \sum_{i=1}^{N} \log f\left(x_i \mid \hat{\theta}_k\right) + 2k \quad (5)$$

where $\hat{\theta}_k$ is the ML estimator for the unknown parameter θ_k based on the sample $x_1, ..., x_N$, and the order choice is such that $\hat{k} = \arg\min_k AIC(k)$. Theoretical researches in the Akaike's criterion has helped to specify the asymptotic behaviour of AIC. Akaike's criterion is then unsatisfactory since it asymptotically leads to a strict overparametrization of the model order [15]. In order to palliate the inconsistency of Akaike's criterion, G. Schwarz (1978) proposed a new criterion for

an exponential family founded on a bayesian justification. He suggested the Bayesian information criterion (BIC) [13]. In a different way, J. Rissanen (1978) came up with an equivalent criterion using a coding technique (minimizing the codelength in relation to the observations) for a parametrized density, which is referred as minimum description length (MDL) principle:

$$MDL(k) = -2\sum_{i=1}^{N} \log f\left(x_i \mid \hat{\theta}_k\right) + k \log N \qquad (6)$$

This criterion is asymptotically convergent in that it helps in finding the appropriate model when $N \to \infty$ (strong consistency). Note that the latter criterion penalizes more stringently the log-likelihood as the number of observations increases in comparison with AIC.

A third criterion was introduced by E. J. Hannan and B. G. Quinn [6] in the case of an autoregressive process. It substitutes $k \log \log N$ for the preceding penalty and leads to convergence in probability of the order estimator (weak consistency); This criterion is written as φ and stands as a compromise between AIC and MDL.

Finally, let us give a more recent criterion [4, 9] drawn on Rissanen's [11] works on *stochastic complexity* ending up in a criterion written as φ_β in the general case of parametrized PDF (see [4] for the detailed derivation of the criterion):

$$\varphi_\beta(k) = -2\sum_{i=1}^{N} \log f\left(x_i \mid \hat{\theta}_k\right) + kN^\beta \log \log N, 0 < \beta < 1 \qquad (7)$$

The selection is obtained by minimizing φ_β. This selection is strongly consistent [10].

The criteria presented here are of the same type: they all have the log-likelihood part depending on data and penalization or compensation depending on the number of free parameters and sample size. The general form of these criteria is then (IC stands for Information Criterion):

$$IC(k) = -2L\left(\hat{\theta}_k\right) + c_N k \qquad (8)$$

where $\hat{\theta}_k$ is the ML estimator of the parameter vector relating to the model with order k, N the number of observations, c_N is an increasing function of N and L is the log-likelihood of the model of order k: $L\left(\hat{\theta}_k\right) = \log f\left(x^N \mid \hat{\theta}_k\right) = \sum_{i=1}^{N} \log f\left(x_i \mid \hat{\theta}_k\right)$. The optimal model order is the one which minimises IC, viz:

$$\hat{k} = \arg\min_k IC(k)$$

For competitive models with the same number of freedom degrees, IC is simply the ML estimator. We are making reference to [14] or [7, chap. 7] for a detailed review of the different criteria developed in literature.

In what follows, we will discusses the choice of the β value. In section 3.2, we will give the expression of the criteria mentioned above for the estimation of the number of components of a multivariate Gaussian mixture-model.

3.1 Choice of β

For $0 < \beta < 1$, the φ_β criterion given in Eq. (7) is strongly consistent. Unfortunately, we observe in practice an underparametrization of the model for large values of β ($\beta > 0.5$). To palliate this drawback, we consider the choice of β in a prescribed range which depends on the sample size N. This choice, for N sufficiently large, is defined using the following inequalities: $N^\beta \geq \log N$ and $N^{1-\beta} \geq \log N$. Thus, one can obtain the following range for the value of β:

$$\frac{\log \log N}{\log N} \leq \beta \leq 1 - \frac{\log \log N}{\log N} \qquad (9)$$

In practice, we choose β near $\frac{\log \log N}{\log N}$ to obtain an adequate penalization form [1].

3.2 Expression of the criteria in the mixture-model case

The likelihood function is optimized via the EM algorithm given in section 2. To choose the number of components, we use the penalized log-likelihood criterion:

$$IC(k) = -2L(k) + c_N \alpha(k) \qquad (10)$$

where L is the likelihood function given in (3) and $\alpha(k)$ is the number of free parameters of the mixture-model with k components. The number of free parameters for each covariance matrix equals $\frac{d(d+1)}{2}$ since each matrix is symmetric. The remaining parameters are the d free parameters per mean vector and the $k-1$ mixture proportions. Thus, the number of degrees of freedom is:

$$\alpha(k) = kd + k - 1 + k\frac{d(d+1)}{2} = \frac{k(d+1)(d+2)}{2} - 1 \qquad (11)$$

To choose the maximal number of component K, several heuristics have been proposed:

- the number of observations N must be greater than the total number of parameters, which leads to $K \leq \frac{2N}{(d+1)(d+2)}$ [12];

[1] for instance, $\frac{\log \log N}{\log N} \simeq 0,28$ for $N = 1000$ and $\frac{\log \log N}{\log N} \simeq 0,24$ for $N = 10000$, thus choosing β such as $0.2 \leq \beta \leq 0.3$ seems to be adequate for a moderate or even small sample size.

- another rule consists in choosing K such as $K \simeq \sqrt{\frac{N}{2}}$ or $(\frac{N}{\log N})^{\frac{1}{3}}$ [2].

4 Numerical simulations

In this section, we propose to test the efficiency of the φ_β criterion for the choice of the number of component clusters on simulated datasets. The different examples given below consist in samples drawn from mixture-models with a known number of component clusters.

The EM algorithm is initialized randomly. Thus, we iterate the algorithm until convergence. We repeat the experiment for each value of $k, 1 \leq k \leq K$, where K is fixed. The datasets are 1 or 2 dimensional. For comparison purpose, we also plotted the value of a variant of AIC introduced in [2] in the mixture-model context:

$$AIC_3(k) = -2L(k) + 3k \tag{12}$$

EXAMPLE 4.1 *This example concerns a simulated mixture-model with* 4 *components. The parameters are given in table 1. Figure 1 shows the* PDF *estimated by the standard* EM *algorithm for an increasing number of components.*

Table 1: parameters of the mixture-model (example 4.1)

$n_1 = 200$	$\mu_1 = -10$	$\sigma_1 = 2$
$n_2 = 120$	$\mu_2 = -4$	$\sigma_2 = 1$
$n_3 = 250$	$\mu_3 = 5$	$\sigma_3 = 7$
$n_4 = 100$	$\mu_4 = 10$	$\sigma_4 = 1$

EXAMPLE 4.2 *This numeric simulation concerns the* 3 *components mixture given in table 2. Components* 1 *et* 3 *have the same covariance matrix (identity matrix) and different means, while the covariance matrix of the second component is non-diagonal. Figure 2(b) shows the number of clusters selected by the minimum* IC *estimate (*MICE*), and figures 2(c) to 2(f) give the clusters shape*[2] *obtained by the* EM *algorithm.*

EXAMPLE 4.3 *This numeric simulation concerns data where the "true"* PDF *is not Gaussian. The results are given in figure 3. Heuristically, the models with more than* 5 *components seem more adequate for this dataset. Figure 3(f) shows the estimated number of components.*

[2]Each ellipse represent the contour line with same density for each Gaussian component.

Table 2: mixture parameters used for simulation (example 4.2)

$n_1 = 300$	$\mu_1 = \begin{pmatrix} 5 \\ 0 \end{pmatrix}$	$\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
$n_2 = 300$	$\mu_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$	$\Sigma_2 = \begin{pmatrix} 5 & 2 \\ 2 & 3 \end{pmatrix}$
$n_3 = 300$	$\mu_3 = \begin{pmatrix} -5 \\ 0 \end{pmatrix}$	$\Sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

The different examples given above show the better cluster validation obtained with MDL and φ_β criteria (see figures 1, 2 and 3), avoiding the overparametrization of the model, whereas AIC (and AIC_3) generally tends to overestimate the number of clusters. Moreover, the better minimum visualization (see figure 2(b) for instance) of the criteria $\varphi_\beta, \beta = 0.1, 0.2, 0.3$, justifies their benefit, even if the value $\beta = 0.3$ seems too large in some cases (see figure 1(f)).

5 Conclusion

Clustering is helpful to improve the synthetic comprehension of a large data set, particularly in high dimension. To choose the number of clusters within a partition of these data is an old and difficult problem still opened. In this paper, we solved this problem by considering at first a strict hypothesis about the clustering procedure (Gaussian mixture-model). We used the EM algorithm to obtain the ML estimates of the parameters of the mixture-model, for each number of components k. Finally, we studied the behaviour of the most representative criteria AIC and MDL for the choice of the number of components on simulated data sets. We also introduce a new information criterion φ_β. We compared and discuss about the behaviour of the three criteria to choose the number of clusters and showed the good behavior of $\varphi_\beta, \beta = 0.1, 0.2$ on simulated $1D$ and $2D$ datasets. Finally, we can notice that MICE can be applied to other well-known clustering procedures such as the k-means algorithm or hierarchical classifications.

References

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Cáski, editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest, 1973. Akademiai Kaidó.

[2] H. Bozdogan. Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity. In H. Bozdogan et al., editors, *Proc. of the First US/Japan Conf. on the Frontiers of Statistical Modeling: An Informational Approach*, pages 69–113. Kluwer Academic Publishers, 1994.

[3] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Stat. Soc. B*, 39:185–197, 1977.

[4] A. El Matouat and M. Hallin. Order selection, stochastic complexity and Kullback-Leibler information. In P.M. Robinson and M. Rosenblatt, editors, *Time Series Analysis*, volume 2, in memory of E.J. Hannan, pages 291–299. Springer Verlag, New York, 1996.

[5] B.S. Everitt and D.J. Hand. *Finite Mixture Distributions*. Chapman and Hall, London, UK, London, 1981.

[6] E .J. Hannan and B. G. Quinn. The determination of the order of an autoregression. *Journal of the Roy. Stat. Soc. B*, 41(2):190–195, 1979.

[7] E.J. Hannan and M. Deistler. *The Statistical Theory of Linear Systems*. John Wiley and Sons, New York, 1988.

[8] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, Englewood Cliffs, NJ, 1988.

[9] F. Jouzel, C. Olivier, and A. El Matouat. Information criteria based edge detection. In *Proc. of the 9 th Int. European Signal Processing Conference EUSIPCO'98, Rhodes, Greece*. EURASIP, September 1998. A paraître.

[10] R. Nishii. Maximum likelihood principle and model selection when the true model is unspecified. *Journal of Multivariate analysis*, 27:392–403, 1988.

[11] J. Rissanen. Stochastic complexity and modeling. *The Annals of Statistics*, 14(3):1080–1100, 1986.

[12] L. Sardo and J. Kittler. Minimum complexity PDF estimation for correlated data. In *Proc. Of ICPR96*, pages 750–754. IEEE, 1996.

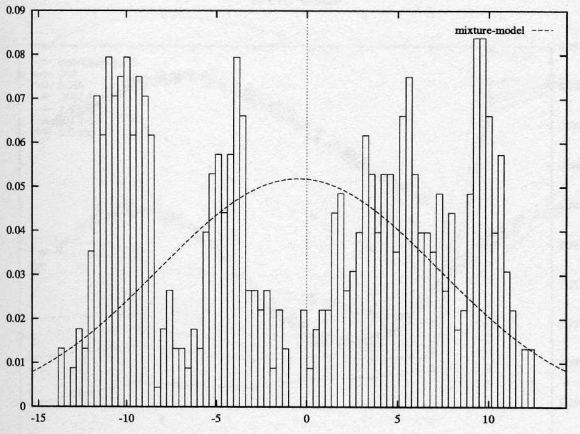
[13] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[14] S. Sclove. Applications of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52(3):333–343, 1987.

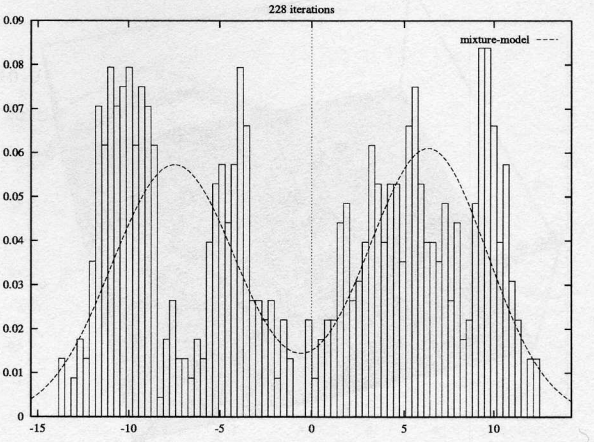
[15] R. Shibata. Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, 63:117–126, 1976.

[16] D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical Analysis of Finite Mixture Distributions*. John Wiley and Sons, 1985.

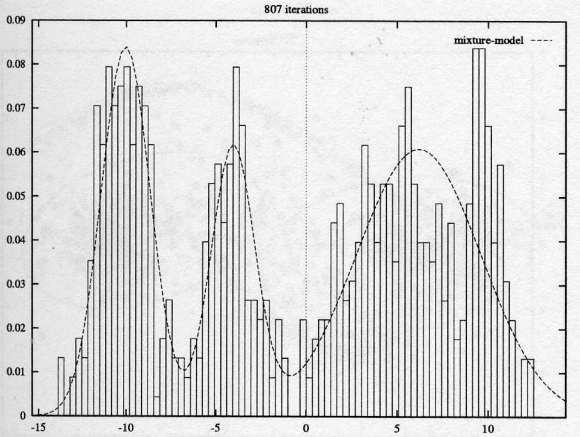
[17] J. Zhang and J.-M. Modestino. A model-fitting approach to cluster validation with application to stochastic model-based image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 12(10), 1990.



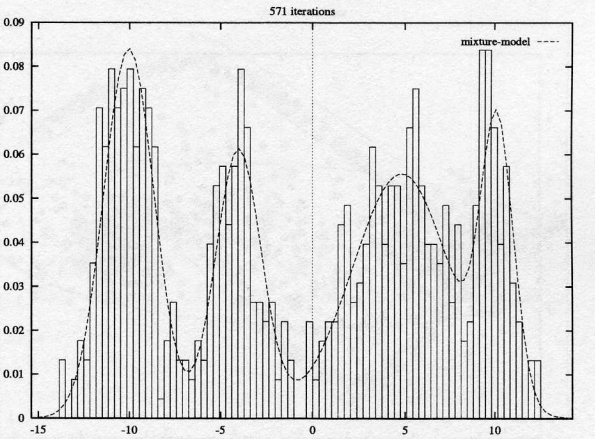
(a) $k = 1$



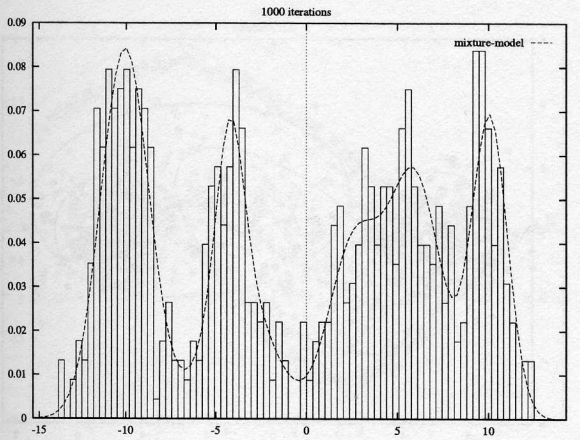
(b) $k = 2$



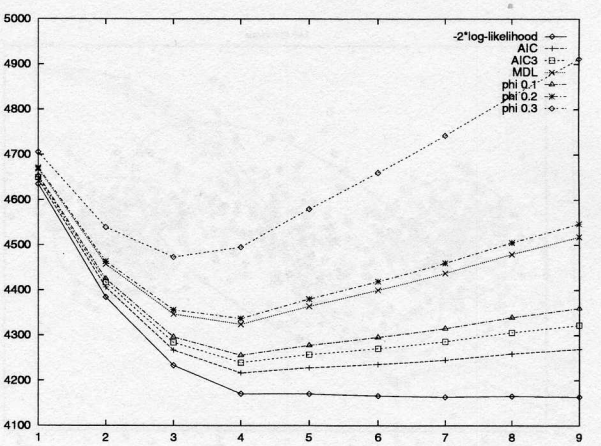
(c) $k = 3$



(d) $k = 4$

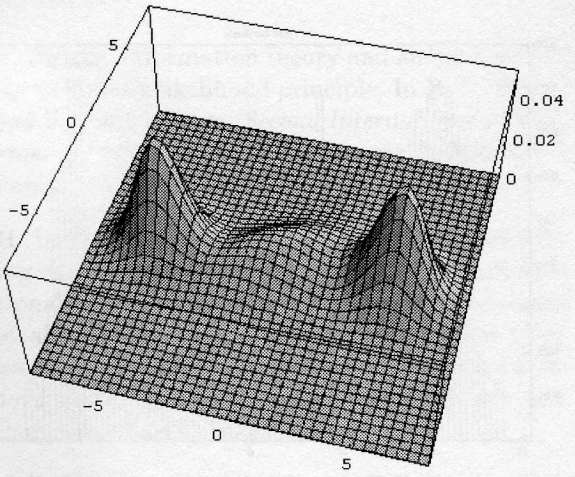


(e) $k = 6$

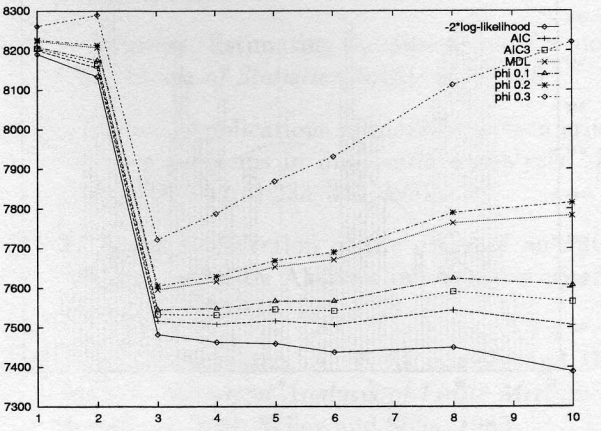


(f) estimated number of components $\hat{k}$

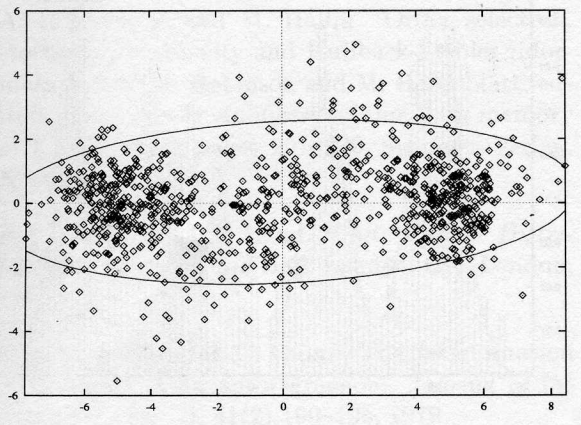
Figure 1: identified mixture-model for different values of k



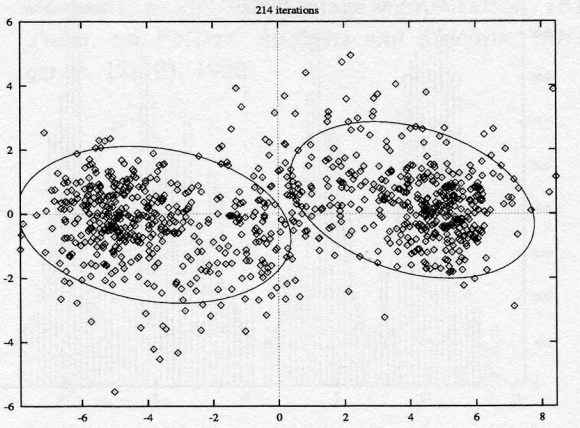
(a) mixture-model PDF



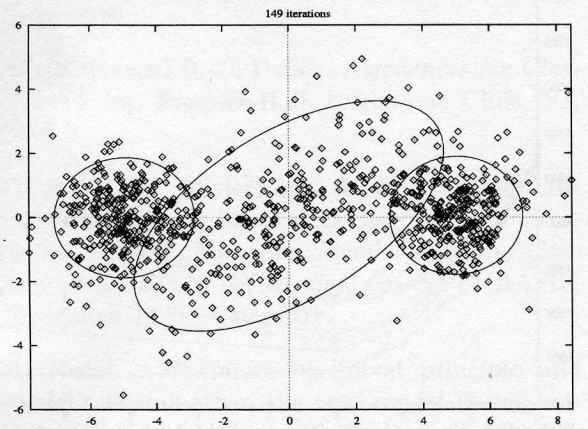
(b) estimated number of clusters



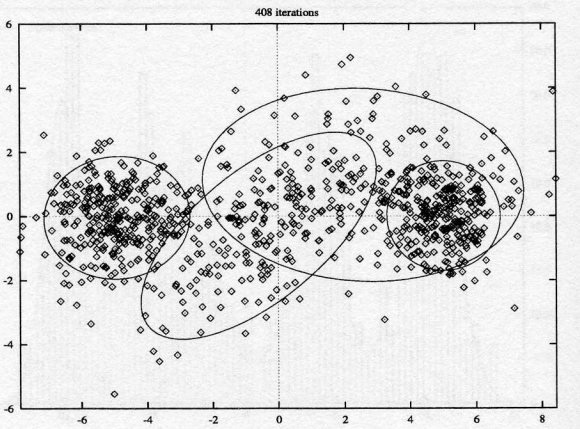
(c) estimated cluster ($k = 1$)



(d) estimated clusters for $k = 2$

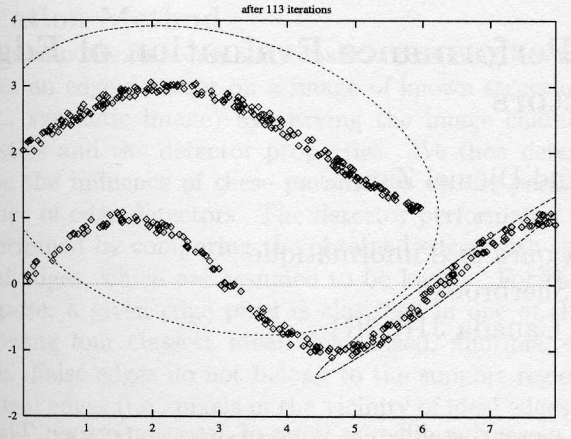


(e) estimated clusters for $k = 3$

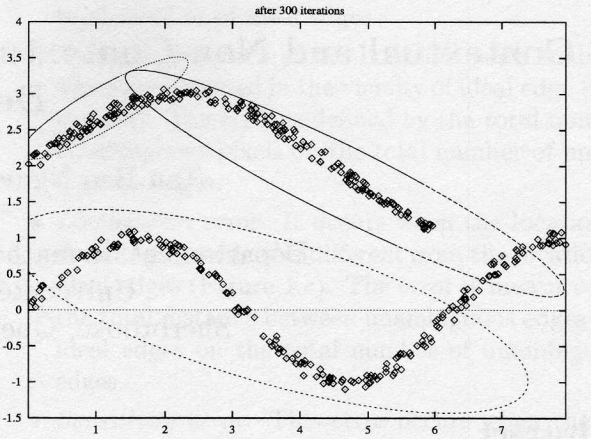


(f) estimated clusters for $k = 4$

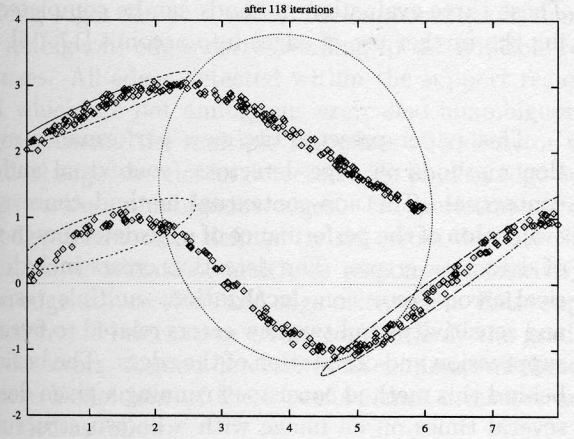
Figure 2: simulated data of example 4.2



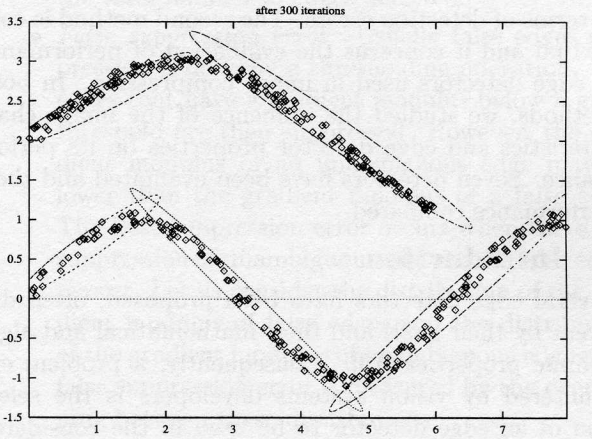
(a) $k = 2$



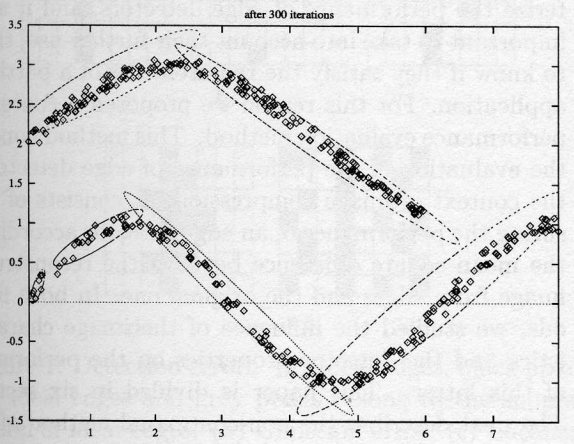
(b) $k = 3$



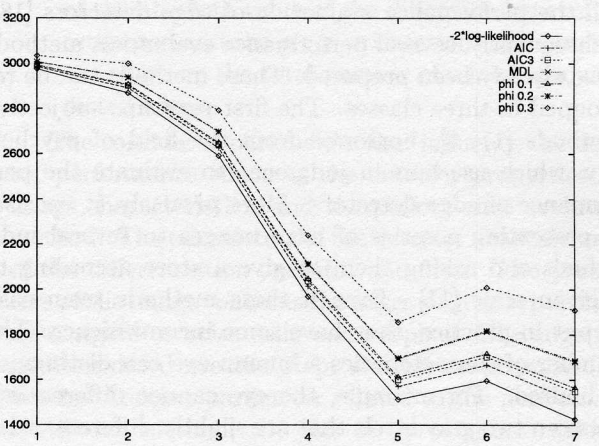
(c) $k = 4$



(d) $k = 5$



(e) $k = 6$



(f) estimated number of components

Figure 3: estimated clusters for each value of k and estimated number of components