MIXFIT: An Algorithm for the Automatic Fitting and Testing of Normal Mixture Models

G.J. McLachlan and D. Peel
Centre for Statistics, Department of Mathematics,
University of Queensland, St. Lucia, Queensland 4072, AUSTRALIA
FAX: +61 7 33651477
EMAIL: gjm@maths.uq.edu.au

Abstract

We consider the fitting of normal mixture models to multivariate data, using maximum likelihood via the EM algorithm. This approach requires the specification of an initial estimate of the vector of unknown parameters, or equivalently, of an initial classification of the data with respct to the components of the mixture model under fit. We describe an algorithm called MIXFIT that automatically undertakes this fitting, including the specification of suitable initial values if not supplied by the user. The MIXFIT algorithm has several options, including the provision to carry out a resampling-based test for the number of components in the mixture model.

EM algorithm; Cluster analysis; Likelihood ratio test

1. Introduction

Finite mixtures models are being increasingly used to model the distributions of a wide variety of random phenomena (McLachlan [1]). For multivariate data of a continuous nature, attention has focussed on the use of multivariate normal components because of their computational convenience. They can be easily fitted iteratively by maximum likelihood (ML) via the expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin [2], McLachlan and Krishnan [3]), as the iterates on the M-step are given in closed form. Also, in cluster analysis where a mixture model-based approach is widely adopted, the clusters in the data are often essentially elliptical in shape, so that it is reasonable to consider fitting mixtures of elliptically symmetric component densities. Within this class of component densities, the multivariate normal density is a convenient choice given its above-mentioned computational tractability.

We let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ denote an observed p-dimensional sample of size n. With a normal mixture model-based approach to drawing inferences from these data, each data point is assumed to be a realization of the random p-dimensional vector $\boldsymbol{Y}$ with the g-component normal mixture probability density function (p.d.f.),

$$f(\boldsymbol{y}; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i \phi(\boldsymbol{y}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{1}$$

where the mixing proportions π_i are nonnegative and sum to one and where $\phi(\boldsymbol{y}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ denotes the multivariate normal p.d.f. with mean (vector) $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Here $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{g-1}, \boldsymbol{\theta}^T)^T$, where $\boldsymbol{\theta}$ contains the elements of the $\boldsymbol{\mu}_i$ and the distinct elements of $\boldsymbol{\Sigma}_i$ $(i = 1, \ldots, g)$. Often, in order to reduce the number of unknown parameters, the component-covariance matrices are restricted to being equal, or even diagonal as in the AutoClass program of Cheeseman and Stutz [4]. Other packages for the fitting of finite mixtures are reviewed recently by Haughton [5]; see also Wallace and Dowe [6]. We shall see in the example to be presented that the restriction of equality of the component-covariance matrices can have a detrimental effect on the implied clustering, even when the true grouping is apparently homoscedastic.

Under the assumption that $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are independent realizations of the feature vector $\boldsymbol{Y}$, the log likelihood function for $\boldsymbol{\Psi}$ is given by

$$\log L(\boldsymbol{\Psi}) = \sum_{j=1}^{n} \log \sum_{i=1}^{g} \pi_i \phi(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i). \tag{2}$$

With the maximum likelihood approach to the estimation of $\boldsymbol{\Psi}$, an estimate is provided by an appropriate root of the likelihood equation,

$$\partial \log L(\boldsymbol{\Psi}) / \partial \boldsymbol{\Psi} = \boldsymbol{0}. \tag{3}$$

In this paper, we describe an algorithm called MIXFIT that has been developed using the EM algorithm to find solutions of (1) corresponding to local maxima. In the appendix of their monograph, McLachlan and Basford [7] gave the listing of FORTRAN programs that they had written for the maximum likelihood fitting of multivariate normal mixture models under a variety of experimental conditions. Over the years, these programs have undergone continued refinement and development, leading to an interim version known as the NMM algorithm (McLachlan and Peel [8]). Since then, there has been much further development, culminating in the present version of the algorithm known as MIXFIT.

For the mixture programs in [7], an initial specification had to be given by the user either for the parameter vector $\boldsymbol{\Psi}$ or for the classification of the data with respect to the components of the normal mixture model. With the MIXFIT algorithm, the user does not have to provide this specification. In the absence of a user-provided specification, the MIXFIT algorithm can be run for a specified number of random starts and/or for starts corresponding to classifications of the data by specified clustering procedures from a wide choice class that includes k-means and commonly used hierarchical methods. The MIXFIT algorithm also has several other options. One such option is the use of multivariate t-components instead of normal components. The t-distribution is a longer-tailed alternative to the normal. Hence the t-mixture model provides a more robust approach to the fitting of normal mixture models, as observations that are atypical of a component are given reduced weight in the calculation of its parameters.

Another major option of the MIXFIT algorithm allows the user to automatically carry out a test for the smallest number of components compatible with the data. This likelihood-based test uses the resampling approach of McLachlan [9] to assess the associated P-value. This option is based on the MMRESAMP subroutine of McLachlan et al. [10]. In addition to the field of cluster analysis where there is no available knowledge of the underlying group-structure, normal mixture models are also being increasingly used in discriminant analysis (McLachlan [11]), where the available data are classified with respect to externally existing groups; see, for example, Hastie and Tibshirani [12]. In this application, normal mixture distributions are used to model those group-conditional distributions that are unable to be modelled adequately by a single normal distribution.

2. Application of EM Algorithm

It is straightforward to find solutions of (3) using the EM algorithm [2]. For the purpose of the application of the EM algorithm, the observed-data vector $\boldsymbol{y}_{obs} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$ is regarded as being incomplete. The component-label variables z_{ij} are consequently introduced, where z_{ij} is defined to be one or zero according as $\boldsymbol{y}_j$ did or did not arise from the ith component of the mixture model $(i = 1, \ldots, g\,;\, j = 1, \ldots, n)$. This complete-data framework in which each observation is conceptualized as having arisen from one of the components of the mixture is directly applicable in those situations where $\boldsymbol{Y}$ can be physically identified as having come from a population which is a mixture of g groups. On putting $\boldsymbol{z}_j = (z_{1j}, \ldots, z_{gj})^T$, the complete-data vector $\boldsymbol{x}_{\rm c}$ is therefore given by

$$\boldsymbol{x}_{\rm c} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_n^T)^T,$$

where $\boldsymbol{x}_1 = (\boldsymbol{y}_1^T, \boldsymbol{z}_1^T)^T, \ldots, \boldsymbol{x}_n = (\boldsymbol{y}_n^T, \boldsymbol{z}_n^T)^T$ are taken to be independent and identically distributed with $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ being independent realizations from a multinomial distribution consisting of one draw on g categories with respective probabilities $\pi_1, \ldots, \pi_g$. That is,

$$\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n \overset{iid}{\sim} \text{Mult}_g(1, \boldsymbol{\pi}),$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)^T$.

For this specification, the complete-data log likelihood is

$$\log L_{\rm c}(\boldsymbol{\Psi}) = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij} \log\{\pi_i \phi(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\}. \qquad (4)$$

The EM algorithm is easy to program and proceeds iteratively in two steps, E (for expectation) and M (for maximization); see McLachlan and Krishnan [3] for a recent account of the EM algorithm in a general context. On the $(k+1)$th iteration, the E-step requires the calculation of

$$Q(\boldsymbol{\Psi}\,;\, \boldsymbol{\Psi}^{(k)}) = E_{\boldsymbol{\Psi}^{(k)}}\{\log L_{\rm c}(\boldsymbol{\Psi}) \mid \boldsymbol{y}_{obs}\},$$

the conditional expectation of the complete-data log likelihood $\log L_{\rm c}(\boldsymbol{\Psi})$, given the observed data $\boldsymbol{y}_{obs}$, using the current fit $\boldsymbol{\Psi}^{(k)}$ for $\boldsymbol{\Psi}$. Since $\log L_{\rm c}(\boldsymbol{\Psi})$ is a linear function of the unobservable component-label variables z_{ij}, the E-step is effected simply by replacing z_{ij} by its conditional expectation given $\boldsymbol{y}_j$, using $\boldsymbol{\Psi}^{(k)}$ for $\boldsymbol{\Psi}$. That is, z_{ij} is replaced by

$$\begin{aligned}
\tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k)}) &= E_{\boldsymbol{\Psi}^{(k)}}\{Z_{ij} \mid \boldsymbol{y}_j\} \\
&= \text{pr}_{\boldsymbol{\Psi}^{(k)}}\{Z_{ij} = 1 \mid \boldsymbol{y}_j\} \\
&= \frac{\pi_i^{(k)} \phi(\boldsymbol{y}_j;\, \boldsymbol{\mu}_i^{(k)}, \boldsymbol{\Sigma}_i^{(k)})}{\sum_{h=1}^{g} \pi_h^{(k)} \phi(\boldsymbol{y}_j;\, \boldsymbol{\mu}_h^{(k)}, \boldsymbol{\Sigma}_h^{(k)})}
\end{aligned}$$

for $i = 1, \ldots, g$, $j = 1, \ldots, n$), and where $\tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k)})$ is the estimate after the kth iteration of the posterior probability that the jth entity with feature vector $\boldsymbol{y}_j$ belongs to the ith component $(i = 1, \ldots, g\,;\, j = 1, \ldots, n)$.

On the M-step on the $(k+1)$th iteration, the intent is to choose the value of $\boldsymbol{\Psi}$, say $\boldsymbol{\Psi}^{(k+1)}$, that maximizes $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})$. It follows that on the M-step of the $(k+1)$th iteration, the current fit for the mixing proportions, the component means, and the covariance matrices is given explicitly by

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} \tau_i^{(k)}(\boldsymbol{y}_j)/n,$$

$$\boldsymbol{\mu}_i^{(k+1)} = \sum_{j=1}^{n} \tau_i^{(k)}(\boldsymbol{y}_j)\boldsymbol{y}_j / \sum_{i=1}^{n} \tau_i^{(k)}(\boldsymbol{y}_j),$$

and

$$\boldsymbol{\Sigma}_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_i^{(k)}(\boldsymbol{y}_j)(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})^T}{\sum_{i=1}^{n} \tau_i^{(k)}(\boldsymbol{y}_j)} \tag{5}$$

for $i = 1, \ldots, g$, where $\tau_i^{(k)}(\boldsymbol{y}_j) = \tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}^{(k)})$. In the homoscedastic case where the normal components are taken to have a common covariance matrix $\boldsymbol{\Sigma}$, the component-specific estimate (5) is replaced by

$$\boldsymbol{\Sigma}^{(k+1)} = \sum_{i=1}^{g} \pi_i^{(k+1)} \boldsymbol{\Sigma}_i^{(k)}.$$

A nice feature of the EM algorithm is that the mixture likelihood $L(\boldsymbol{\Psi})$ can never be decreased after the EM sequence. Hence

$$L(\boldsymbol{\Psi}^{(k+1)}) \geq L(\boldsymbol{\Psi}^{(k)}),$$

which implies that $L(\boldsymbol{\Psi}^{(k)})$ converges to some L^* for a sequence of likelihood values bounded above. The E- and M-steps are alternated repeatedly until the likelihood (or the parameter estimates) change by an arbitrarily small amount in the case of convergence.

3 Choice of Local Maximizer

Let $\hat{\boldsymbol{\Psi}}$ be the chosen solution of the likelihood equation. The likelihood function $L(\boldsymbol{\Psi})$ tends to have multiple local maxima for normal mixture models. In the homoscedastic case of normal components with a common covariance matrix, the likelihood function $L(\boldsymbol{\Psi})$ has a global maximum in the interior of the parameter space. Hence in this case in the absence of any information apart from the observed data, $\hat{\boldsymbol{\Psi}}$ is usually taken to be the root of (3) corresponding to the largest of the local maxima located.

The consistency of the global maximizer for finite mixture distributions has been established under the usual regularity conditions; see [7] for further details. In the heteroscedastic case of unrestricted component covariance matrices, $L(\boldsymbol{\Psi})$ is unbounded, as each data point gives rise to a singularity on the edge of the parameter space. But in the heteroscedastic case, attention can be directed to local maxima in the interior of the parameter space, since under essentially the usual regularity conditions there will exist a sequence of roots of the likelihood equation that is consistent and asymptotically efficient and normally distributed. With probability tending to one, these roots correspond to local maxima in the interior of the parameter space ([7]). In practice, however, consideration has to be given to the problem of relatively large local maxima that occur as a consequence of a fitted component having a very small (but nonzero) variance for univariate data or generalized variance (the determinant of the covariance matrix) for multivariate data. Such a component corresponds to a cluster containing a few data points either relatively close together or almost lying in a lower dimensional subspace in the case of multivariate data. There is thus a need to monitor the relative size of the fitted mixing proportions and of the component variances for univariate observations and of the generalized component variances for multivariate data in an attempt to identify these spurious local maximizers. There is also a need to monitor the Euclidean distances between the fitted component means.

4. Specification of Initial Values

It follows from above that care must be taken in the choice of the root of the likelihood equation, particularly in the case of unrestricted covariance matrices where $L(\boldsymbol{\Psi})$ is unbounded. In order to fit a mixture model using the EM algorithm, an initial value has to be specified for the vector $\boldsymbol{\Psi}$ of unknown parameters for use on the E-step on the first iteration of the EM algorithm. Equivalently, initial values must be specified for the posterior probabilities of component membership of the mixture, $\tau_1(\boldsymbol{y}_j; \boldsymbol{\Psi}^{(0)}), \ldots, \tau_g(\boldsymbol{y}_j; \boldsymbol{\Psi}^{(0)})$, for each $\boldsymbol{y}_j$ $(j = 1, \ldots, n)$ for use on commencing the EM algorithm on the M-step the first time through. The latter posterior probabilities can be specified as zero-one values, corresponding to an outright classification of the data with respect to the g components of the mixture. In this case, it suffices to specify the initial partition of the data. In a cluster analysis context it is usually more appropriate to do this rather than to specify an initial value for $\boldsymbol{\Psi}$.

5. MIXFIT Algorithm

We now describe an algorithm called MIXFIT, which automatically provides a selection of starting values for this purpose if the user does not provide any. The MIXFIT algorithm automatically provides starting values for the application of the EM algorithm by considering a selection

obtained from three sources: (a) random starts, (b) hierarchical clustering-based starts, and (c) k-means clustering-based starts. Concerning (b), the user has the option of using in either standardized or unstandardized form, the results from seven hierarchical methods (nearest neighbour, farthest neighbour, group average, median, centroid, flexible sorting, and Ward's method). There are several algorithm parameters that the user can optionally specify; alternatively, default values are used. The program fits the normal mixture model for each of the initial grouping specified from the three sources (a) to (c). All these computations are automatically carried out by the program. The user only has to provide the data set the restrictions on the component-covariance matrices (equal, unequal, or diagonal), the extent of the selection of the initial groupings to be used to determine starting values, and the number of components that are to be fitted. Summary information is automatically given as output for the final fit. However, it is not suggested that the clustering of a data set should be based solely on a single solution of the likelihood equation, but rather on the various solutions considered collectively. The default final fit is taken to be the one corresponding to the largest of the local maxima located. However, the summary information can be recovered for any distinct fit.

As well as the options pertaining to the automatic provision of starting values covered above, several other options are available, including the provision of standard errors for the fitted parameters in the mixture model, and the bootstrapping of the likelihood ratio statistic λ for testing $g = g_0$ versus $g = g_0 + 1$ components in the mixture model, where the value g_0 is specified by the user. With the latter option, the bootstrap samples are generated parametrically from the g_0-component normal mixture model with $\mathbf{\Psi}$ set equal to the fit $\hat{\mathbf{\Psi}}_{g_0}$ for $\mathbf{\Psi}$ under the null hypothesis of g_0 components.

6. Example

To illustrate the use of the MIXFIT algorithm, we consider the clustering of the crab data set of Campbell and Mahon [13] on the genus *Leptograpsus*, which has been analysed further in Ripley [14]. Attention is focussed on the sample of $n = 100$ blue crabs, there being $n_1 = 50$ males and $n_2 = 50$ females corresponding to groups G_1 and G_2 respectively. Each specimen has measurements on the width of the frontal lip *FL*, the rear width *RW*, and length along the midline *CL* and the maximum width *CW* of the carapace, and the body depth *BD* in mm. In Figure 1, we give the scatter plot of the second and third variates with their group of origin noted. Hawkins' [15 simultaneous test for multivariate normality and equal covariance matrices (homoscedasticity) suggests it is reasonable to assume that the group-conditional distributions are normal with a common covariance matrix. Consistent with this, it was found that the sample linear discriminant function (formed using the known classification of the data) misallocates only two more observations than the quadratic version formed without the restriction of equal group-covariance matrices. There are 6 misallocations overall with the latter (4 from G_1 and 2 from G_2).

We now cluster these data, ignoring the known classification of the data, by fitting a mixture of two normal components with and without restrictions on the component covariance matrices. Rather surprisingly, it is found that the assumption of equal-component covariance matrices has a marked impact on the implied clustering of the data. For without any restrictions on the component-covariance matrices, the normal mixture model-based clustering results in one cluster containing 39 observations from G_1 and another containing all 50 observations from G_2, along with the remaining 11 observations from G_1. Identifying the larger-sized cluster with G_2 (And the smaller-sized cluster with G_1), the overall misallocation rate is 11. However, if we impose the restriction of equal component-covariance matrices, then a further 8 observations from G_1 are misallocated, as these 8 observations are moved from the smaller-sized cluster to the larger-sized cluster. This is an interesting example as it is a situation where, although it is reasonable to assume that the true group-structure is homoscedastic, the assumption of homoscedasticity in fitting the normal mixture model leads to a much inferior clustering of the data. This is brought about by there being several observations from G_1 (male crabs) that are near the boundary of the feature space on which the group-conditional densities weighted by their prior probabilities are apparently equal, and when knowledge of their true classification is not used, these points have an adverse effect on the group-density estimates (and the consequent clustering) produced under the assumption of homoscedasticity.

To illustrate just how sensitive the process is to the location of some of these observations, we considered the implied clustering of the 17th and 18th observations from G_1,

$$\boldsymbol{x}_{17} = (13.1, 10.6, 28.2, 32.3, 11.0)^T$$

and

$$\boldsymbol{x}_{18} = (13.1, 10.9, 28.3, 32.4, 11.2)^T.$$

Although these two points are very close in terms of Euclidean and also Mahalanobis distance, their estimated posterior probabilites of belonging to G_1 are quite different under the normal mixture model with equal covariance matrices, being equal to 0.564 and 0.170, respectively. To see why there should be this somewhat unexpected difference between these estimated posterior probabilities, we have plotted in Figure 2 the contours of the estimated component p.d.f.'s. for the 2nd and 3rd feature variables. It can be

seen that the 18th observation lies in the feature space where the contours of $f_2(\boldsymbol{x}; \hat{\boldsymbol{\Psi}})$ are very steep, so that even slight changes in the values of the feature variables can result in relatively large changes in the values of the component densities. Without the restriction of equality on the component-covariance matrices, the estimates of these two posterior probabilities are very similar, being 1.000 and 0.996, and now observation 18 is correctly allocated to G_1.

7. References

1. McLachlan, G.J. (1999). *Finite Mixture Models*. New York: Wiley.

2. Dempster, A.P., Laird, N.M., and and Rubin, D.B. (1997). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B* **39**, 1–38.

3. McLachlan, G.J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. New York: Wiley.

4. Cheeseman, P., and Stutz, J. (1995). Bayesian classification (AutoClass): theory and results. In *Advances in Knowledge Discovery and Data Mining*, U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (Eds.). Menlo Park, California: The AAAI Press, pp. 61–83.

5. Haughton, D. (1997). Packages for estimating finite mixtures: a review. *The American Statistician* **51**, 194–205.

6. Wallace, C.S. and Dowe D.L. (1994). Intrinsic classification by MML - the Snob program. In 7th Australian Joint Conf. on Art. Intell., 37–44.

7. McLachlan, G.J. and Basford, K.E. (1988). *Mixture Models: Inference and Applications to Clustering.* New York: Marcel Dekker.

8. McLachlan, G.J. and Peel, D. (1996). An algorithm for unsupervised learning via normal mixture models. In *ISIS: Information, Statistics and Induction in Science*, D.L. Dowe, K.B. Korb, and J.J. Oliver (Eds.), pp. 354–363. Singapore: World Scientific.

9. McLachlan, G.J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics* **36**, 318–324.

10. McLachlan, G.J., Peel D., Adams, P., and Basford, K.E. (1995). An algorithm for assessing by resampling the P-value of the likelihood ratio test on the number of components in normal mixture models. *Research Report No. 31*. Brisbane: Centre for Statistics, The University of Queensland.

11. McLachlan, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. New York: Wiley.

12. Hastie, T. and Tibshirani, R.J. (1996). Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society B* **58**, 155–176.

13. Campbell, N.A. and Mahon, R.J. (1974). A multivariate study of variation in two species of rock crab of genus *Leptograpsus*. *Australian Journal of Zoology* **22**, 417–425.

14. Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

15. Hawkins, D.M. (1981). A new test for multivariate normality and homoscedasticity. *Technometrics* **23**, 105–110.

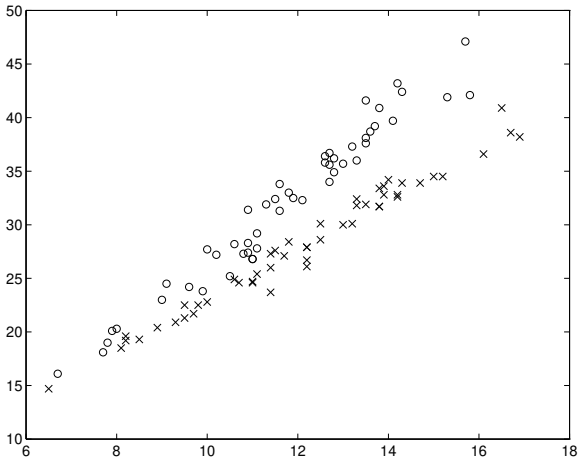

Figure 1. Plot of third versus second variate for $n_1 = 50$ male and $n_2 = 50$ female "blue" crabs ($\circ$ denotes male and $\times$ female)

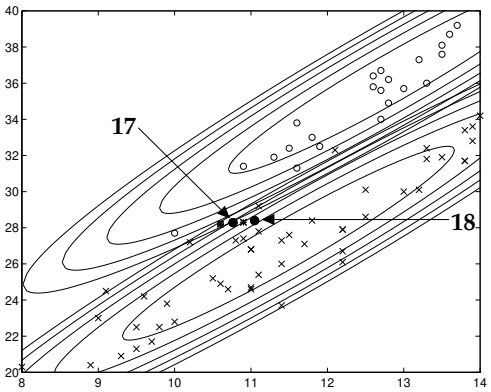

Figure 2. Contours of the fitted component densities on the second and third variates in the region of points 17 and 18 for the $n_1 = 50$ male and $n_2 = 50$ female "blue" crabs ($\circ$ denotes male and $\times$ female)