Robust Cluster Analysis Via Mixtures Of Multivariate t-Distributions

Geoffrey J. McLachlan and David Peel

Department of Mathematics, University of Queensland,
St. Lucia, Queensland 4072, AUSTRALIA

Abstract. Normal mixture models are being increasingly used as a way of clustering sets of continuous multivariate data. They provide a probabilistic (soft) clustering of the data in terms of their fitted posterior probabilities of membership of the mixture components corresponding to the clusters. An outright (hard) clustering can be subsequently obtained by assigning each observation to the component to which it has the highest fitted posterior probability of belonging. However, outliers in the data can affect the estimates of the parameters in the normal component densities, and hence the implied clustering. A more robust approach is to fit mixtures of multivariate t-distributions, which have longer tails than the normal components. The expectation-maximization (EM) algorithm can be used to fit mixtures of t-distributions by maximum likelihood. The application of this model to provide a robust approach to clustering is illustrated on a real data set. It is demonstrated how the use of t-components provides less extreme estimates of the posterior probabilities of cluster membership.

1 Introduction

Finite mixtures of distributions have provided a mathematical-based approach to the statistical modeling of a wide variety of random phenomena (McLachlan and Basford (1988), McLachlan (1999)). Because of their usefulness as an extremely flexible method of modelling, finite mixture models have continued to receive increasing attention over the years, both from a practical and theoretical point of view. 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 (1977), McLachlan and Krishnan (1997)), as the iterates on the M-step are given in closed form.

However, for many applied problems, the tails of the normal distribution are often shorter than required. Also, the estimates of the component means and covariance matrices can be affected by observations that are atypical of the components in the normal mixture model being fitted. Hence we consider the fitting of mixtures of multivariate t-distributions. This 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.

Also, as the t-distribution provides a longer tailed alternative to the normal distribution, it tends to give less extreme estimates of the posterior probabilities of component membership.

One useful application of normal mixture models has been in the important field of cluster analysis. Besides having a sound mathematical basis, this approach is not confined to the production of spherical clusters, such as with k-means type algorithms that use Euclidean distance rather than the Mahalanobis distance metric which allows for within-cluster correlations between the variables in the feature vector $\boldsymbol{X}$. Moreover, unlike clustering methods defined solely in term so of the Mahalanobis distance, the normal mixture-based clustering takes into account the normalizing term $\mid \boldsymbol{\Sigma}_i \mid^{-1/2}$ in the estimate of the multivariate normal density adopted for the component distribution of $\boldsymbol{X}$ corresponding to the ith cluster. This term can make an important contribution in the case of disparate group-covariance matrices (McLachlan (1992, Chapter 2)).

Although even a crude estimate of the within-cluster covariance matrix $\boldsymbol{\Sigma}_i$ often suffices for clustering purposes, it can be severely affected by outliers. Hence it is highly desirable for methods of cluster analysis to be robust. By robustness, it is meant that the method is not affected significantly by small departures from the assumed model, such as the presence of outliers. The problem of providing protection against outliers in multivariate data is a very difficult problem and increases with the difficulty of the dimension of the data (Rocke and Woodruff (1997)). The related problem of making clustering algorithms more robust has received much attention recently as, for example, in McLachlan and Basford (1988, Chapter 3), De Veaux and Kreiger (1990), Campbell (1994), Davé and Krishnapuram (1996), Frigui and Krishnapuram (1996), Kharin (1996), and Rousseeuw, Kaufman, and Trauwaert (1996), and Zhuang et al. (1996), among others. In the past, there have been many attempts at modifying existing methods of cluster analysis to provide robust clustering procedures. Some of these have been of a rather *ad hoc* nature. The t-mixture model provides a sound mathematical basis for a robust method of clustering. We shall illustrate its usefulness in this context by a cluster analysis of a real data set.

2 Normal Mixture Model

We let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ denote an observed p-dimensional random 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 $\boldsymbol{X}$ with the g-component normal mixture probability density function (p.d.f.),

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

where the mixing proportions π_i are nonnegative and sum to one and where $\phi(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ 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 $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_g^T)^T$

and where $\boldsymbol{\theta}_i$ contains the elements of $\boldsymbol{\mu}_i$ and the distinct elements of $\boldsymbol{\Sigma}_i$ ($i = 1, \ldots, g$).

3 Multivariate *t*-Distribution

Consider the multivariate normal p.d.f. $\phi(\boldsymbol{x};\, \boldsymbol{\mu},\, \boldsymbol{\Sigma})$. One way to broaden this parametric family for potential outliers is to adopt the two-component normal mixture p.d.f.

$$(1-\epsilon)\phi(\boldsymbol{x};\, \boldsymbol{\mu},\, \boldsymbol{\Sigma}) + \epsilon\phi(\boldsymbol{x};\, \boldsymbol{\mu},\, c\boldsymbol{\Sigma}), \tag{2}$$

where c is large and ϵ is small, representing the small proportion of observations that have a relatively large variance. Huber (1964) subsequently considered more general forms of contamination of the normal distribution in the development of his robust M-estimators of a location parameter, as discussed further in Section 7. The normal scale mixture model (2) can be written as

$$\int \phi(\boldsymbol{x};\boldsymbol{\mu},\, \boldsymbol{\Sigma}/u)\, dH(u), \tag{3}$$

where H is the probability distribution that places mass $(1-\epsilon)$ at the point $u = 1$ and mass ϵ at the point $u = 1/c$. Suppose we now take H to be the distribution function of a chi-squared random variable on its degrees of freedom ν; that is, we now take the random variable U to be distributed as

$$U \sim \text{ gamma}(\tfrac{1}{2}\nu,\, \tfrac{1}{2}\nu), \tag{4}$$

where the gamma$(\alpha,\, \beta)$ density function $f(u;\, \alpha,\, \beta)$ is given by

$$f(u;\, \alpha,\, \beta) = \{\beta^\alpha u^{\alpha-1}/\Gamma(\alpha)\} \exp(-\beta u) I_{(0,\infty)}(u)\,; \qquad (\alpha,\, \beta > 0),$$

and the indicator function $I_{(0,\,\infty)}(u) = 1$ for $u > 0$ and is zero elsewhere. We then obtain the t-distribution with location parameter $\boldsymbol{\mu}$, positive definite inner product matrix $\boldsymbol{\Sigma}$, and ν degrees of freedom,

$$f(\boldsymbol{x};\, \boldsymbol{\mu},\, \boldsymbol{\Sigma},\, \nu) = \frac{\Gamma(\frac{\nu+p}{2})\, |\boldsymbol{\Sigma}|^{-1/2}}{(\pi\nu)^{\frac{1}{2}p}\Gamma(\frac{\nu}{2})\{1 + \delta(\boldsymbol{x},\, \boldsymbol{\mu};\, \boldsymbol{\Sigma})/\nu\}^{\frac{1}{2}(\nu+p)}}, \tag{5}$$

where

$$\delta(\boldsymbol{x},\, \boldsymbol{\mu};\, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$$

denotes the Mahalanobis squared distance between $\boldsymbol{x}$ and $\boldsymbol{\mu}$ (with $\boldsymbol{\Sigma}$ as the covariance matrix). As ν tends to infinity, U converges to one with probability one, and so $\boldsymbol{X}$ becomes marginally multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The family of t-distributions provides a heavy-tailed alternative to the normal family with mean $\boldsymbol{\mu}$ and covariance matrix that is equal to a scalar multiple of $\boldsymbol{\Sigma}$ (if $\nu > 2$). In the above and sequel, we are using f as a generic symbol for a p.d.f.

4 ML Estimation of Mixtures of *t*-Distributions

We consider now ML estimation for a g-component mixture of t-distributions, given by

$$f(\boldsymbol{x};\,\boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i f(\boldsymbol{x};\,\boldsymbol{\mu}_i,\,\boldsymbol{\Sigma}_i,\,\nu_i), \tag{6}$$

where now $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{g-1}, \boldsymbol{\theta}^T, \boldsymbol{\nu}^T)^T$ and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_g)^T$. The application of the EM algorithm for ML estimation in the case of a single component t-distribution has been described in McLachlan and Krishnan (1997, Sections 2.6 and 5.8). The results there can be extended to cover the present case of a g-component mixture of multivariate t-distributions.

In the EM-framework, the observed data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, are augmented by $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, where $\boldsymbol{z}_j$ is the component-label vector defining the component of origin of $\boldsymbol{x}_j$, and $z_{ij} = (\boldsymbol{z}_j)_i$ is 1 or zero, according as $\boldsymbol{x}_j$ belongs or does not belong to the ith component. In the light of the above characterization of the t-distribution, it is convenient to view the observed data augmented by the $\boldsymbol{z}_j$ as still being incomplete and introduce into the complete-data vector the additional missing data, $u_1, \ldots, u_n$, which are defined so that given $z_{ij} = 1$,

$$\boldsymbol{X}_j \mid u_j,\, z_{ij} = 1 \;\sim N(\boldsymbol{\mu}_i,\, \boldsymbol{\Sigma}_i/u_j), \tag{7}$$

independently for $j = 1, \ldots, n$, and

$$U_j \mid z_{ij} = 1 \sim \text{gamma}\left(\tfrac{1}{2}\nu_i,\, \tfrac{1}{2}\nu_i\right). \tag{8}$$

Given $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, the $U_1, \ldots, U_n$ are independently distributed according to (8).

It follows that on the $(k+1)$th iteration of the M-step, the $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are updated as

$$\boldsymbol{\mu}_i^{(k+1)} = \sum_{j=1}^{n} \tau_{ij}^{(k)} u_{ij}^{(k)} \boldsymbol{x}_j \Big/ \sum_{j=1}^{n} \tau_{ij}^{(k)} u_{ij}^{(k)} \tag{9}$$

and

$$\boldsymbol{\Sigma}_i^{(k+1)} = \sum_{j=1}^{n} \tau_{ij}^{(k)} u_{ij}^{(k)} (\boldsymbol{x}_j - \boldsymbol{\mu}_i^{(k+1)})(\boldsymbol{x}_j - \boldsymbol{\mu}_i^{(k+1)})^T \Big/ \sum_{j=1}^{n} \tau_{ij}^{(k)}, \tag{10}$$

where

$$u_{ij}^{(k)} = \frac{\nu_i^{(k)} + p}{\nu_i^{(k)} + \delta(\boldsymbol{x}_j,\, \boldsymbol{\mu}_i^{(k)};\, \boldsymbol{\Sigma}_i^{(k)})} \tag{11}$$

and where

$$\tau_{ij}^{(k)} = \frac{\pi_i^{(k)} f(\boldsymbol{x}_j;\, \boldsymbol{\mu}_i^{(k)},\, \boldsymbol{\Sigma}_i^{(k)},\, \nu_i^{(k)})}{f(\boldsymbol{x}_j;\, \boldsymbol{\Psi}^{(k)})}$$

is the current estimate of the posterior probability that $\boldsymbol{x}_j$ belongs to the ith component of the mixture $(i = 1, \ldots, g;\ j = 1, \ldots, n)$. It can be seen that the $\boldsymbol{\mu}_i^{(k+1)}$ and $\boldsymbol{\Sigma}_i^{(k+1)}$ are effectively chosen by weighted least-squares estimation.

The E-step updates the weights $\tau_i^{(k)} u_{ij}^{(k)}$, while the M-step effectively chooses $\boldsymbol{\mu}_i^{(k+1)}$ and $\boldsymbol{\Sigma}_i^{(k+1)}$ by weighted least-squares estimation. Thus the updates of the parameters are available in closed form if the degrees of freedom ν_i are specified in advance. Otherwise, the updated $\nu_i^{(k+1)}$ have to be computed iteratively. The process can be speeded up using the multicycle ECM and ECME algorithms (Liu and Rubin (1995), McLachlan and Krishnan (1997, Section 5.8)). The MIXFIT algorithm of McLachlan et al. (1997) for the fitting of normal mixture models has an option for the fitting of mixtures of t-components.

5 Clustering Applications of *t*-Mixture Models

The ML estimation of the component means $\boldsymbol{\mu}_i$ is robust in the sense that observations with large Mahalanobis distances are downweighted. This can be clearly seen from the form of the equation (9) for the MLE of $\boldsymbol{\mu}_i$. As ν_i (or its estimate if not specified) decreases, the degree of downweighting of an outlier increases. For finite ν_i as $\| \boldsymbol{x}_j \| \rightarrow \infty$, its effect on the ith component-location parameter estimate goes to zero, whereas its effect on the ith component-scale estimate remains bounded but does not vanish. Of course there is always the option of manually excluding observations considered to be grossly atypical of the bulk of the data, using the minimum covariance determinant criterion; see, for example, Hawkins and McLachlan (1997).

It can be therefore seen that the use of mixtures of t-distributions provides a sound statistical basis for formalizing and implementing the somewhat *ad hoc* approaches that have been proposed in the past. It also provides a framework for assessing the degree of robustness to be incorporated into the fitting of the mixture model through either the specification or the estimation of the degrees of freedom ν_i in the t-component p.d.f.'s.

6 Example

To illustrate the t-mixture model-based approach to clustering, we consider the crab data set of Campbell and Mahon (1974) on the genus *Leptograpsus*, which has been analysed further in Ripley (1996). 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 Fig. 1, we give the scatter plot of the second and third variates with their group of origin noted. Hawkins' (1981) 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 observations (from G_1).

We now cluster these data, ignoring the known classification of the data. We first fit a mixture of two normal components with equal covariance matrices. The implied clusters consist of one cluster containing 31 observations from G_1 and another containing all 50 observations from G_2 plus the remaining 19 observations from G_1. Hence this clustering misallocates 19 observations from G_1 but no observations from G_2. We next fitted a mixture of two t-components with common inner product matrix and degrees of freedom. The latter was estimated along with the other parameters from the data. The inferred value for ν was $\hat{\nu} = 22.5$. This t-mixture model-based solution produces a slightly improved outright clustering in that one fewer observation from G_1 is misallocated with the smaller sized cluster now containing 32 observations from G_1 and with the larger one containing 18 observations from G_1 and all 50 from G_2.

Although the use of the t-mixture model has only slightly improved the outright clustering, it does produce a less extreme probabilistic clustering of the observations. To demonstrate this point, we have listed the estimates of the posterior probabilities of membership of group G_1 under both normal and t-mixture models for those 19 observations from G_1 misclassified under the normal mixture model in Table 1. It can be seen that the use of the t-mixture model results in observation 14 having an estimated posterior probability of belonging to the first component of the mixture (corresponding to group G_1) of greater than 0.5. Hence this observation is no longer misclassified in an outright clustering of the data. The remaining 18 observations would still be misclassified, but all have increased estimated posterior probabilities of belonging to G_1; in particular, these estimated probabilities for observations 11, 18, and 26 are much closer to the threshold value of 0.5.

7 Previous Work on M-Estimation of Components

A common way in which robust fitting of normal mixture models has been undertaken, is by using M-estimates to update the component estimates on the M-step of the EM algorithm, as in McLachlan and Basford (1988) and Campbell (1994). In this case, the updated component means are $\boldsymbol{\mu}_i^{(k+1)}$ are given by (9), but where now the weights $u_{ij}^{(k)}$ are defined as

$$u_{ij}^{(k)} = \psi(d_{ij}^{(k)})/d_{ij}^{(k)}, \tag{12}$$

where

$$d_{ij}^{(k)} = \{(\boldsymbol{x}_j - \boldsymbol{\mu}_i^{(k)})^T \boldsymbol{\Sigma}_i^{(k)^{-1}} (\boldsymbol{x}_j - \boldsymbol{\mu}_i^{(k)})\}^{1/2}$$

and $\psi(s) = -\psi(-s)$ is Huber's (1964) ψ-function defined by

$$\begin{aligned} \psi(s) &= s, && \mid s \mid \leq c, \\ &= \operatorname{sign}(s)\, c, && \mid s \mid > c, \end{aligned} \tag{13}$$

for an appropriate choice of the tuning constant c. The ith component-covariance matrix $\boldsymbol{\Sigma}_i^{(k+1)}$ can be updated as (10), where $u_{ij}^{(k)}$ is replaced by $\{\psi(d_{ij}^{(k)})/d_{ij}^{(k)}\}^2$.

An alternative to Huber's ψ-function is a redescending ψ-function, for example, Hampel's (1973) piecewise linear function. However, there can be problems in forming the posterior probabilities of component membership. as there is the question as to which parametric family to use for the component p.d.f.'s (McLachlan and Basford, Section 2.8, 1988). One possibility is to use the form of the p.d.f. corresponding to the ψ-function adopted. However, in the case of any redescending ψ-function with finite rejection points, there is no corresponding p.d.f. In Campbell (1994), the normal p.d.f. was used, while in the related univariate work in De Veaux and Kreiger (1990), the t-density with three degrees of freedom was used, with the location and scale component parameters estimated by the (weighted) median and mean absolute deviation, respectively. It can be therefore seen that the use of mixtures of t-distributions provides a sound statistical basis for formalizing and implementing the somewhat *ad hoc* approaches that have been proposed in the past. It also provides a framework for assessing the degree of robustness to be incorporated into the fitting of the mixture model through the specification or estimation of the degrees of freedom ν_i in the t-component p.d.f.'s.

References

Campbell, N.A. (1994). Mixture models and atypical values. *Mathematical Geology* **16**, 465–477.

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.

Davé, R.N. and Krishnapuram, R. (1995). Robust clustering methods: a unified view. *IEEE Transactions on Fuzzy Systems* **5**, 270–293.

Dempster, A.P., Laird, N.M., and and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B* **39**, 1–38.

De Veaux, R.D. and Kreiger, A.M. (1990). Robust estimation of a normal mixture. *Statistics & Probability Letters* **10**, 1–7.

Frigui, H. and Krishnapuram, R. (1996). A robust algorithm for automatic extraction of an unknown number of clusters from noisy data. *Pattern Recognition Letters* **17**, 1223–1232.

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

Hampel, F.R. (1973). Robust estimation: a condensed partial survey. *Z. Wahrscheinlickeitstheorie verw. Gebiete* **27**, 87–104.

Hawkins, D.M. and McLachlan, G.J. (1997). High-breakdown linear discriminant analysis. *Journal of the American Statistical Association* **92**, 136–143.

Huber, P.J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics* **35**, 73–101.

Kharin, Y. (1996). *Robustness in Statistical Pattern Recognition*. Dordrecht: Kluwer.

Liu, C. and Rubin, D.B. (1995). ML estimation of the t distribution using EM and its extensions, ECM and ECME. *Statistica Sinica*, **5**, 19–39.

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

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

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

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

McLachlan, G.J., Peel, D., Basford, K.E., and Adams, P. (1997). MIXFIT: an algorithm for the automatic fitting and testing of normal mixture models. Unpublished manuscript.

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

Rocke, D.M. and Woodruff, D.L. (1997). Robust estimation of multivariate location and shape. *Journal of Statistical Planning and Inference* **57**, 245–255.

Rousseeuw, P.J., Kaufman, L., and Trauwaert, E. (1996). Fuzzy clustering using scatter matrices. *Computational Statistics and Data Analysis* **23**, 135–151.

Zhuang, X., Huang, Y., Palaniappan, K., and Zhao, Y. (1996). Gaussian density mixture modeling, decomposition and applications. *IEEE Transactions on Image Processing* **5**, 1293–1302.

Table 1. Estimated Posterior Probability of Membership of G_1

No.	Normal Mixture	t-Mixture $\hat{\nu} = 22.5$	No.	Normal Mixture	t-Mixture $\hat{\nu} = 22.5$
1	0.0000	0.0004	11	0.1610	0.3237
2	0.0000	0.0001	12	0.0042	0.0098
3	0.0003	0.0010	14	0.4932	0.6359
4	0.0016	0.0036	15	0.0116	0.0189
5	0.0007	0.0020	16	0.0002	0.0003
6	0.0056	0.0093	18	0.1702	0.2971
7	0.0002	0.0005	19	0.0047	0.0068
8	0.1450	0.1889	20	0.0733	0.0930
9	0.0011	0.0022	26	0.4163	0.4643
10	0.0004	0.0008			

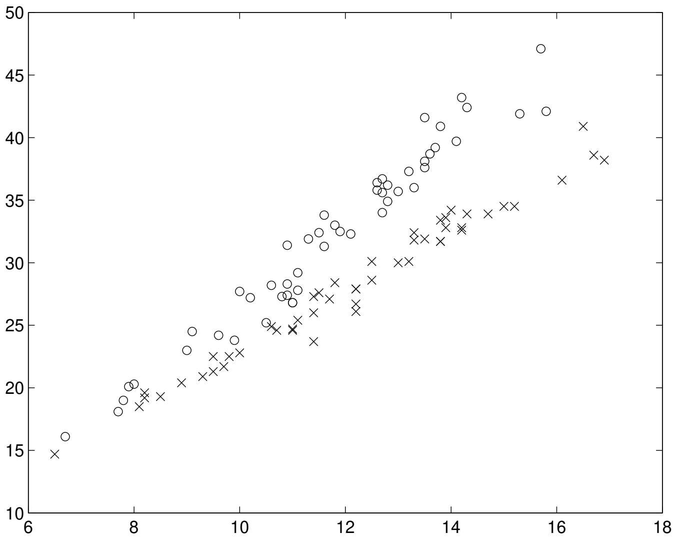

Fig. 1. Plot of third versus second variate for $n_1 = 50$ male and $n_2 = 50$ female "blue" crabs (∘ denotes male and $\times$ female)