



Contents lists available at ScienceDirect

## Pattern Recognition Letters

journal homepage: [www.elsevier.com/locate/patrec](http://www.elsevier.com/locate/patrec)Robust mixture clustering using Pearson type VII distribution<sup>☆</sup>Jianyong Sun<sup>a,d,\*</sup>, Ata Kabán<sup>b</sup>, Jonathan M. Garibaldi<sup>c</sup><sup>a</sup> CPIX, School of Bioscience, The University of Nottingham, UK<sup>b</sup> School of Computer Science, University of Birmingham, Birmingham, UK<sup>c</sup> School of Computer Science, The University of Nottingham, Nottingham, UK<sup>d</sup> Tianjin Key Laboratory of Intelligence Computing and Novel Software Technology, Tianjin University of Technology, 300191 Tianjin, China

## ARTICLE INFO

## Article history:

Received 17 December 2008

Available online xxx

Communicated by W. Pedrycz

## Keywords:

Robust mixture modeling

Pearson type VII distribution

Outlier detection

Robust learning

## ABSTRACT

A mixture of Student  $t$ -distributions (MoT) has been widely used to model multivariate data sets with atypical observations, or outliers for robust clustering. In this paper, we developed a novel robust clustering approach by modeling the data sets using mixture of Pearson type VII distributions (MoP). An EM algorithm is developed for the maximum likelihood estimation of the model parameters. An outlier detection criterion is derived from the EM solution. Controlled experimental results on the synthetic datasets show that the MoP is more viable than the MoT. The MoP performs comparably if not better, on average, in terms of outlier detection accuracy and out-of-sample log-likelihood with the MoT. Furthermore, we compared the performances of the Pearson type VII and the student  $t$  mixtures on the classification of several real pattern recognition data sets. The comparison favours the developed Pearson type VII mixtures.

© 2010 Elsevier B.V. All rights reserved.

## 1. Introduction

It has been widely known that Gaussian mixture modeling is very sensitive to outliers. The performance of the Gaussian mixtures degrades on pattern recognition due to outliers. In the statistical machine learning community, the Student  $t$ -distribution is normally used as a building block for robust learning, such as clustering (Peel and McLachlan, 2000; Svendsen and Bishop, 2005), visualization (Vellido et al., 2006) and robust projections (Archambeau et al., 2006). This distribution has heavy tails, so that non-zero probabilities can be given for observations that are far away from the main clusters of the data set.

In this paper, we propose to use a broader family of distribution, called the Pearson type VII distribution (Pearson, 1916), for robust mixture modeling. Since the Pearson type VII distribution also has heavy tails, and subsumes the Student  $t$ -distribution, we would expect some advantages from using a mixture of Pearson distributions. To fit a mixture of Pearson type VII distributions with the multivariate data set by maximum likelihood (ML) estimation, the scale mixture representation of the Pearson type VII distribution is first derived. An EM algorithm is then developed based on the scale mixture representation. An outlier criterion is derived for the detection of atypical observations, or outliers from the data

sets. The advantage of using the mixture of Pearson (MoP) can be shown in the EM algorithm when estimating the degrees of freedom, which controls the degree of robustness. The degrees of freedom of the MoP can be more simply inferred than those of the mixture of  $t$ -distributions (MoT).

In the rest of the paper, Section 2 describes the Pearson type VII distribution and its scale mixtures form. Section 3 provides an EM algorithm for the MoP. Experimental results are presented in Section 4, firstly to show the comparison of the performances of the Pearson type VII, the Student  $t$  and the Gaussian mixtures, and an EM algorithm for the Pearson type VII mixtures without using the scale mixture representation. The performances of the developed algorithm and the Student  $t$  mixtures on classification of some benchmark datasets are presented also in this section. Section 5 concludes the paper.

## 2. The Pearson type VII distribution

The Pearson type VII distribution (Pearson, 1916) is defined as follows:

$$p(\mathbf{t}; \mu, \Lambda, m) = \frac{\Gamma(m)}{\pi^{d/2} \Gamma(m - \frac{d}{2})} |\Lambda|^{-\frac{1}{2}} [1 + \Delta^2]^{-m}$$

where  $\Delta^2 = (\mathbf{t} - \mu)^T \Lambda^{-1} (\mathbf{t} - \mu)$  is the Mahalanobis distance,  $m$  is the degree of freedom that controls the degree of robustness ( $2m > d$ ), and  $d$  is the dimensionality of  $\mathbf{t}$ .

The Pearson VII family distribution is an instance of elliptical symmetric distribution (Arellano-Valle et al., 2006), which is of the type  $p(\mathbf{t}) = |\Lambda|^{-\frac{1}{2}} \phi\left\{(\mathbf{t} - \mu)^T \Lambda^{-1} (\mathbf{t} - \mu)\right\}$  with  $\phi$  satisfying

<sup>☆</sup> Based on "Robust Mixture Modelling by using the Pearson Type VII Distribution", by (J. Sun, A. Kaban and J.M. Garibaldi) which appeared in Proceedings of the International Joint Conference on Neural Networks, © 2010IEEE.

\* Corresponding author at: CPIX, School of Bioscience, The University of Nottingham, UK. Tel.: +44 115951 6108; fax: +44 115951 6292.

E-mail address: [j.sun@cpib.ac.uk](mailto:j.sun@cpib.ac.uk) (J. Sun).

$$\int_0^\infty u^{\frac{d}{2}-1} \phi(cu) du = \frac{\Gamma(\frac{d}{2})}{(c\pi)^{\frac{d}{2}}}$$

where  $\Gamma(\cdot)$  is the gamma function. The Pearson type VII distribution has been widely used in many scientific areas, such as modeling the stock market (Nagahara, 1996), X-ray measurements (Prevéy, 1986) and many others. It can be shown that the normal distribution arises as a special case of Pearson type VII distribution if  $m$  approaches infinity (Pearson, 1916). The Student  $t$ -distribution  $S(\mathbf{t}; \mu, \Lambda, \nu)$  with degree of freedom  $\nu$ :

$$S(\mathbf{t}; \mu, \Lambda, \nu) = \frac{|\Lambda|^{-\frac{1}{2}} \Gamma(\frac{\nu+d}{2})}{(\pi\nu)^{\frac{d}{2}} \Gamma(\frac{\nu}{2}) \left\{ 1 + \frac{\Lambda^2}{\nu} \right\}^{\frac{\nu+d}{2}}}$$

is also a special case of the Pearson type VII distribution if we set

$$m = \frac{\nu+d}{2}, \quad \Lambda = \Lambda\nu \quad (1)$$

Fig. 1 shows the Pearson type VII distribution with  $\mu = 0, \Lambda = 1, d = 1$  and different degrees of freedom.

As stated in (Fernández and Steel, 2000), the Pearson type VII distribution and the Student  $t$ -distribution both belong to the class of scale mixtures of normals (West, 1987). Through some mathematical manipulation (see Appendix), it can be proven that the Pearson type VII distribution is of the following form:

$$p(\mathbf{t}; \mu, \Lambda, m) = \int_0^\infty \mathcal{N}\left(\mathbf{t}|\mu, \frac{\Lambda}{u}\right) \mathcal{G}\left(u|m - \frac{d}{2}, \frac{1}{2}\right) du \quad (2)$$

where

$$\mathcal{G}(u|a, b) = b^a u^{a-1} \frac{\exp\{-bu\}}{\Gamma(a)}$$

is the gamma distribution with parameters  $a$  and  $b$ . The Pearson type VII distribution can then be interpreted hierarchically as follows:

$$p(\mathbf{t}|u) = \mathcal{N}\left(\mathbf{t}|\mu, \frac{\Lambda}{u}\right)$$

$$p(u) = \mathcal{G}\left(u|m - \frac{d}{2}, \frac{1}{2}\right)$$

where  $u$  is a latent variable.

In the finite mixture modeling context, the log-likelihood of a data set  $\mathcal{Y} = \{\mathbf{t}_1, \dots, \mathbf{t}_N\}$  can be written as follows:

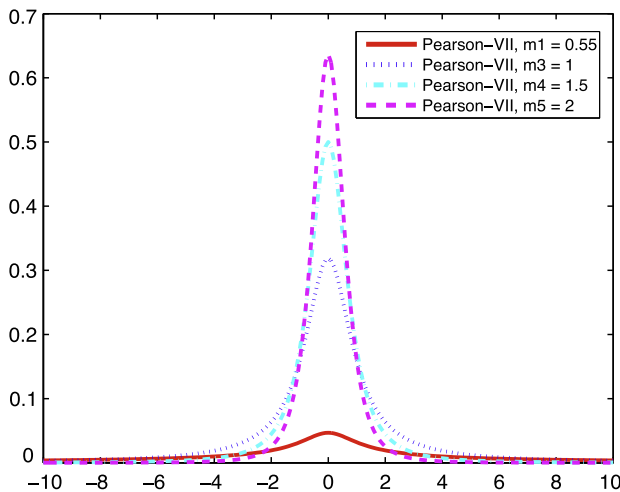


Fig. 1. The Pearson type VII distribution with different degrees of freedom.

$$\mathcal{L}(\Theta|\mathcal{Y}) = \sum_{n=1}^N \log p(\mathbf{t}_n) = \sum_{n=1}^N \log \left( \sum_{k=1}^K p(\mathbf{t}_n|k)p(k) \right)$$

with the model parameters  $\Theta = \{\mu_k, \Lambda_k, m_k, p(k)\}, 1 \leq k \leq K$ , where  $p(\mathbf{t}_n|k)$  is assumed as a Pearson type VII distribution.

### 3. The ML estimation of the mixtures of Pearson type VII distribution

In this section, we propose the use of the EM algorithm for the maximum likelihood estimation (MLE) of the model parameters. First of all, if we let

$$p(\mathbf{t}_n, u_n, k) = p(\mathbf{t}_n|u_n, k)p(u_n|k)p(k) = \mathcal{N}\left(\mathbf{t}_n|\mu_k, \frac{\Lambda_k}{u_n}\right) \mathcal{G}\left(u_n|m_k - \frac{d}{2}, \frac{1}{2}\right) p(k), \quad (3)$$

then the log-likelihood function of  $\mathcal{Y}$  can be written as

$$\mathcal{L}(\Theta|\mathcal{Y}) = \sum_n \sum_k \int p(\mathbf{t}_n, u_n, k) du_n \quad (4)$$

To apply the EM algorithm in the finite mixture modeling, discrete latent variables  $z_n, 1 \leq n \leq N$  are usually introduced, for the class assignment of each data point  $\mathbf{t}_n$ , while  $z_n = k$  indicates that  $\mathbf{t}_n$  belongs to the  $k$ th mixture component. If we further consider  $u_n$ 's as latent variables, the complete log-likelihood function can then be written as follows:

$$\mathcal{L}_c(\Theta|\mathcal{Y}) = \sum_n \sum_k \delta(z_n = k) \log [p(\mathbf{t}_n|u_n, k)p(u_n|k)p(k)] \quad (5)$$

where  $\delta(\cdot)$  is the Kronecker delta.

#### 3.1. E-step

In the E-step, we need to infer the posteriors  $p(z_n = k, u_n|\mathbf{t}_n)$  (denoted by  $p(k, u_n|\mathbf{t}_n)$  for convenience), which can be factorized as follows:

$$p(k, u_n|\mathbf{t}_n) = p(k|\mathbf{t}_n)p(u_n|\mathbf{t}_n, k)$$

According to Bayes' rule, the posteriors can be evaluated as follows:

$$p(k|\mathbf{t}_n) = \frac{p(\mathbf{t}_n|k)p(k)}{\sum_{k'} p(\mathbf{t}_n|k')p(k')} \quad (6)$$

$$p(u_n|\mathbf{t}_n, k) = \frac{p(\mathbf{t}_n|u_n, k)p(u_n|k)}{p(\mathbf{t}_n|k)} \quad (7)$$

where  $p(z_n = k|\mathbf{t}_n)$  is often called 'responsibilities' in literature: the value gives the probability that  $\mathbf{t}_n$  belongs to the  $k$ th cluster. Due to the conjugacy of the gamma prior for  $u_n$ , we obtain:

$$p(u_n|\mathbf{t}_n, k) = \mathcal{G}(u_n|a_{nk}, b_{nk}) \quad (8)$$

where

$$a_{nk} = m_k \quad (9)$$

$$b_{nk} = \frac{(\mathbf{t}_n - \mu_k)^T \Lambda_k^{-1} (\mathbf{t}_n - \mu_k) + 1}{2} \quad (10)$$

The expectation of the complete log-likelihood  $\langle \mathcal{L}_c \rangle$  w.r.t the posteriors can be evaluated as follows:

$$\begin{aligned} \langle \mathcal{L}_c \rangle &= \sum_n \sum_k p(k|\mathbf{t}_n) \langle \log [p(\mathbf{t}_n|u_n, k)p(u_n|k)p(k)] \rangle \\ &= \sum_{nk} p(k|\mathbf{t}_n) \langle \log [p(\mathbf{t}_n|u_n, k)] \rangle + \sum_{nk} p(k|\mathbf{t}_n) \langle \log p(u_n|k) \rangle \\ &\quad + \sum_{nk} p(k|\mathbf{t}_n) \log p(k) \\ &= \mathcal{F}_1 + \mathcal{F}_2 + \mathcal{F}_3 \end{aligned} \quad (11)$$

If we let  $\tilde{m}_k = m_k - \frac{d}{2}$ ,  $\mathcal{F}_1$  and  $\mathcal{F}_2$  can be evaluated as:

$$\mathcal{F}_1 = \sum_{nk} p(k|\mathbf{t}_n) \left[ \frac{d}{2} \langle u_n \rangle_k - \frac{1}{2} \log |\Lambda_k| - \frac{1}{2} \langle u_n \rangle_k (\mathbf{t}_n - \mu_k)^T \Lambda_k^{-1} (\mathbf{t}_n - \mu_k) - \frac{d}{2} \log(2\pi) \right]$$

$$\mathcal{F}_2 = \sum_{nk} p(k|\mathbf{t}_n) \left[ \tilde{m}_k \log \left( \frac{1}{2} \right) + (\tilde{m}_k - 1) \langle \log u_n \rangle_k - \frac{1}{2} \langle u_n \rangle_k - \log \Gamma(\tilde{m}_k) \right]$$

where

$$\langle u_n \rangle_k = \int u_n p(u_n | \mathbf{t}_n, k) du_n = \frac{a_{nk}}{b_{nk}}$$

$$\langle \log u_n \rangle_k = \int [\log u_n] p(u_n | \mathbf{t}_n, k) du_n = \psi(a_{nk}) - \log b_{nk}$$

and  $\psi$  is the digamma function.

### 3.2. M-step

In the M-step, we maximize  $\langle \mathcal{L}_C \rangle$  w.r.t.  $p(k)$ ,  $\mu_k$ ,  $\Lambda_k$  and  $m_k$ . Taking the derivatives of  $\langle \mathcal{L}_C \rangle$  w.r.t.  $\mu_k$ , and solving the stationary equation, we obtain the close form as follows:

$$\mu_k = \frac{\sum_n p(k|\mathbf{t}_n) \langle u_{nk} \rangle \mathbf{t}_n}{\sum_n p(k|\mathbf{t}_n) \langle u_{nk} \rangle} \quad (12)$$

Taking derivatives w.r.t.  $\Lambda_k$ , and zeroing the equation, we obtain:

$$\Lambda_k = \frac{\sum_n p(k|\mathbf{t}_n) \langle u_{nk} \rangle (\mathbf{t}_n - \mu_k) (\mathbf{t}_n - \mu_k)^T}{\sum_n \langle z_{nk} \rangle}$$

$p(k)$  can be updated as:

$$p(k) = \frac{1}{N} \sum_n p(k|\mathbf{t}_n) \quad (13)$$

It can be seen that the above update equations for  $\mu_k$ ,  $\Lambda_k$  and  $p(k)$  are the same as the update equations in the MLE for the mixture of Student  $t$ -distribution (see (Peel and McLachlan, 2000) for details). To update  $m_k$ , we need to solve the following equation:

$$\sum_n p(k|\mathbf{t}_n) \left[ \langle \log u_n \rangle_k - \log(2) - \psi \left( m_k - \frac{d}{2} \right) \right] = 0 \quad (14)$$

or specifically:

$$\psi \left( m_k - \frac{d}{2} \right) = \frac{\sum_n p(k|\mathbf{t}_n) [\langle \log u_n \rangle_k - \log(2)]}{\sum_n p(k|\mathbf{t}_n)} \quad (15)$$

Comparing to the non-linear equation used for estimating the degree of freedom  $v_k$  in MoT (Peel and McLachlan, 2000)

$$\psi \left( \frac{v_k}{2} \right) - \log \left( \frac{v_k}{2} \right) = \frac{\sum_n p(k|\mathbf{t}_n) [\langle u_n \rangle_k - \langle \log u_n \rangle_k - 1]}{\sum_n p(k|\mathbf{t}_n)} \quad (16)$$

We see that the estimation of  $m_k$  in the MoP could be easier than the estimation of  $v_k$  in MoT.<sup>1</sup> It was claimed in (Peel and McLachlan, 2000) that the search for  $v_k$ ,  $1 \leq k \leq K$  is time consuming. On the other hand, the search algorithm proposed in (Pike and Hill, 1966) is very efficient for solving the inverse digamma function.

### 3.3. Scaling

Considering the time complexity of the algorithm, in the E-step of each iteration, the computation of the parameters of  $p(u_n | \mathbf{t}_n, k)$ ,

<sup>1</sup> In our experiments, we used the Newton–Raphson method for solving Eqs. (16) and (15). The algorithm terminates when the difference between two consecutive solutions is less than  $1.0e-4$ . Note that a faster algorithm for the inverse of digamma as required in Eq. (15) is available as developed in (Miranda and Fackler, 2002) (the corresponding Matlab codes can be obtained from <http://www4.ncsu.edu/~pfackler>).

$1 \leq k \leq K$ ,  $a_{nk}$  and  $b_{nk}$  for each data point takes  $\mathcal{O}(d^3K)$ , and the responsibility  $p(k|\mathbf{t}_n)$  needs  $\mathcal{O}(d^3K)$  time as well. In the M-step, the computation of the model parameters  $\mu_k$ ,  $\Lambda_k$ ,  $p(k)$ ,  $1 \leq k \leq K$  takes  $\mathcal{O}(NK)$ . Therefore, in total, the time complexity of the MoP is  $\mathcal{O}(d^3KN)$ , which is exactly the same as that of the MoT except that we do not count the time used for the estimation of the degree of freedom.

The most expensive operation of the developed algorithm is the inversion of the covariance matrix for each component, which is  $\mathcal{O}(d^3)$ . One way to alleviate this problem is to use a diagonal form of  $\Sigma_k$ , in which case the cubic operation is no longer required. Alternatively, we can replace the full covariance matrix by a low rank approximation, such as factor analyzers, as resolved in (Archambeau et al., 2008). Apart from the expensive matrix inversion, the stability of the covariance matrix has also been widely acknowledged in the finite mixture models.

In our training, to avoid the instability problem caused by using the full covariance, we choose to put a Wishart distribution as a prior over the inverse covariance matrix  $\Upsilon_k = \Lambda_k^{-1}$ . That is:

$$p(\Upsilon_k) = C_{\mathcal{W}}(\mathbf{W}_0, \eta_0) |\Upsilon_k|^{(\eta_0 - d - 1)/2} \exp \left( -\frac{1}{2} \text{Tr}(\mathbf{W}_0^{-1} \Upsilon_k) \right)$$

where

$$C_{\mathcal{W}}(\mathbf{W}_0, \eta_0) = |\mathbf{W}_0|^{-\eta_0/2} \left( 2^{\eta_0 d} \pi^{d(d-1)/4} \prod_{i=1}^d \Gamma \left( \frac{\eta_0 + 1 - i}{2} \right) \right)^{-1} \quad (17)$$

where  $\eta_0 \geq d + 1$ . In our experiment, we set  $\eta_0 = d + 2$  and  $\mathbf{W}_0 = \mathbf{I}$  which is the  $d$ -dimensional identity matrix. Then the MAP (maximum a posteriori) estimation of the covariance matrix can be written as follows:

$$\Lambda_k = \frac{\sum_n p(k|\mathbf{t}_n) \langle u_{nk} \rangle (\mathbf{t}_n - \mu_k) (\mathbf{t}_n - \mu_k)^T + \mathbf{W}_0^{-1}}{\sum_n \langle z_{nk} \rangle + \eta_0 - d - 1}$$

### 3.4. Post-processing

The same as in generative models, the developed Pearson type VII model can also be applied to new, previously unseen data from the same source. For a given test data set, we need to calculate the posterior distributions of  $u_n$  associated with each test point  $\mathbf{t}_n$ . To compute, we fix the parameters  $\mu_k$ ,  $\Lambda_k$ ,  $m_k$  and  $p_k$ ,  $1 \leq k \leq K$  obtained from the training set and perform the E-step once, i.e. Eqs. (9) and (10). The log-likelihood of the test data set (called “out-of-sample log-likelihood”) can be computed by

$$\mathcal{L}_o = \sum_n p(k) \log p(\mathbf{t}_n | k)$$

As for the mixture of Student  $t$ -distribution (Peel and McLachlan, 2000), the posterior expectation of  $u$  can be used as an indicator of outliers. Given a data point  $\mathbf{t}$ , if we define

$$e \equiv \sum_k p(k|\mathbf{t}) \langle u \rangle_k = \sum_k p(k|\mathbf{t}) \frac{m_k}{\frac{(\mathbf{t} - \mu_k)^T \Lambda_k^{-1} (\mathbf{t} - \mu_k) + 1}{2}} \quad (18)$$

then,  $\mathbf{t}$  is flagged as an outlier if the value  $e$  is sufficiently small, or approximately, the value:

$$\kappa = \sum_k \langle z_k \rangle (\mathbf{t} - \mu_k)^T \Lambda_k^{-1} (\mathbf{t} - \mu_k) \quad (19)$$

is sufficiently large. To detect the outliers from a test data set, an appropriate value of  $\kappa$  or  $e$  can simply be adopted.

### 3.5. Model selection

To determine the optimal number of mixture components, we adopt the minimum message length (MML) criterion as developed in (Figueiredo and Jain, 2002) and applied in many literature such as in our previous work (Sun et al., 2007a; Sun and Kabán, 2010).

To adopt the MML criterion, a penalty term is added to the likelihood as follows:

$$-\frac{\hat{n}}{2} \sum_{k:\pi_k>0} \log \left( \frac{N\pi_k}{12} \right) - \frac{k_{nz}}{2} \log \left( \frac{N}{12} \right) - \frac{k_{nz}(\hat{n}+1)}{2} + \mathcal{L}(\Theta|\mathcal{Y}) \quad (20)$$

where  $\mathcal{L}(\Theta|\mathcal{Y})$  is the data log-likelihood,  $\hat{n}$  is the dimensionality of the parameters,  $k_{nz}$  is the number of non-zero-probability components. The free parameters involved in the proposed algorithm are the means and the full covariance matrices of  $\mathcal{N}(\mathbf{w}_n|u_n, z_n = k)$ . Thus the dimensionality of the  $k$ th parameter  $\theta_k = (\mu_k, \Sigma_k)$ , is  $d + d(d-1)/2$ .

The optimal number of components can be found by maximizing the penalised log-likelihood. The maximization of the penalised log-likelihood leads to the same update equations for  $\mu_k, \Sigma_k, \nu_k, 1 \leq k \leq K$ , only the mixing proportions  $\pi_k$  is updated as follows:

$$\pi_k = \frac{\max \left\{ 0, \sum_{n=1}^N p(k|\mathbf{t}_n) - \frac{\hat{n}}{2} \right\}}{\sum_{j=1}^K \max \left\{ 0, \sum_{n=1}^N p(j|\mathbf{t}_n) - \frac{\hat{n}}{2} \right\}} \quad (21)$$

At each training step, only the non-zero-probability components of the mixtures contribute to the update of the parameters in the variational M step.

### 3.6. The EM algorithm without using latent variable $u$

Besides the above EM-based algorithm for maximising the log-likelihood, we can also develop an EM algorithm without introducing the latent variable  $u$  for estimation of the model parameters. The EM algorithm can be easily derived. In the E-step, the posterior  $p(k|\mathbf{t}_n)$  is the same as in Eq. (6). The expectation of the complete log-likelihood can be written as:

$$\langle \mathcal{L}_t \rangle = \sum_{n=1}^N \sum_{k=1}^K p(k|\mathbf{t}_n) \log p(\mathbf{t}_n|k; \mu_k, \Lambda_k, m_k) p(k) \quad (22)$$

In the M-step, we maximize  $\langle \mathcal{L}_t \rangle$  w.r.t the model parameters. Here, we do not have analytical solutions to the model parameters. The derivative of  $\langle \mathcal{L}_t \rangle$  w.r.t  $\mu_k$  is:

$$\frac{\partial \langle \mathcal{L}_t \rangle}{\partial \mu_k} = \sum_n p(k|\mathbf{t}_n) \frac{2m_k \Lambda_k^{-1} (\mu_k - \mathbf{t}_n)}{1 + \Lambda_{nk}^2} \quad (23)$$

where  $\Lambda_{nk}^2 = (\mathbf{t}_n - \mu_k)^T \Lambda_k^{-1} (\mathbf{t}_n - \mu_k)$ . The derivatives of  $\langle \mathcal{L}_t \rangle$  w.r.t  $\mathbf{A}_k (\Lambda_k = \mathbf{A}_k \mathbf{A}_k^T)$  is:

$$\frac{\partial \langle \mathcal{L}_t \rangle}{\partial \mathbf{A}_k} = \sum_n p(k|\mathbf{t}_n) \left( \Lambda_k^{-1} - \frac{2m_k \Pi_{nk}}{1 + \Lambda_{nk}^2} \right) \mathbf{A}_k \quad (24)$$

where  $\Pi_{nk} = (\mathbf{t}_n - \mu_k)(\mathbf{t}_n - \mu_k)^T$ . The derivatives of  $\langle \mathcal{L}_t \rangle$  w.r.t  $m_k$  is:

$$\frac{\partial \langle \mathcal{L}_t \rangle}{\partial m_k} = \sum_n p(k|\mathbf{t}_n) \left( -\log(1 + \Lambda_{nk}^2) + \psi(m_k) - \psi\left(m_k - \frac{d}{2}\right) \right)$$

The mixing coefficient  $p(k)$  can be obtained the same as in Eq. (13). The disadvantage of this EM solution is that we do not have an outlier criterion for the purpose of outlier detection. The advantage is that the EM algorithm without  $u$  converges faster than that of the proposed algorithm. This EM algorithm can then be used as a 'ground truth' to compare with the proposed algorithm.

## 4. Experimental results

To test the performance of the MoP, firstly the performances of the mixture of Gaussian (MoG), the MoT and the MoP are demonstrated using some synthetic data sets. Secondly, we compare the proposed algorithm with the EM algorithm that is not based on the scale mixture representation of the Pearson type VII distribution. Thirdly, we carried out a set of controlled experiments on synthetic data sets to compare against the MoT, and the mixture of Gaussian. Finally, the performance of the Pearson type VII, the student  $t$  and the Pearson type II (MoPII) (Medasani and Krishnapuram, 1999) on classification of several benchmark datasets from the University of California-Irvine, data repository (Murphy and Aha, 1973).

### 4.1. The synthetic data and illustrative demonstration

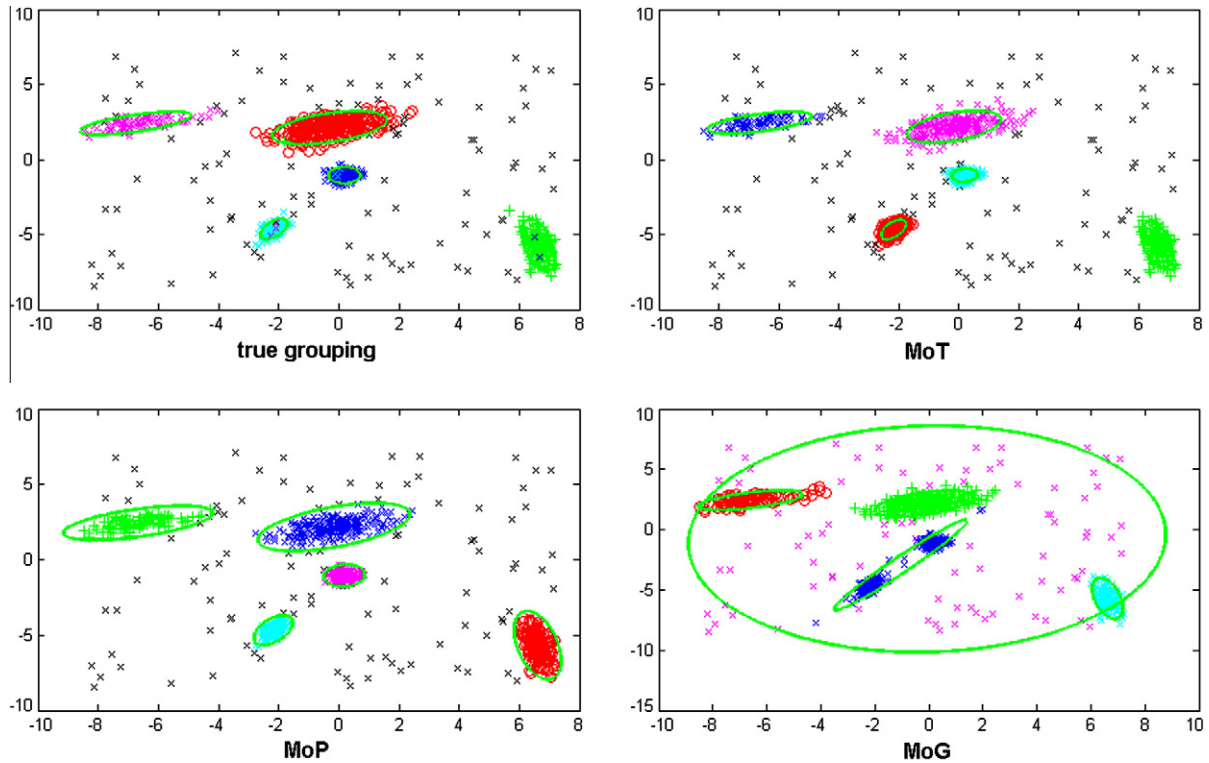
In our experiments, the synthetic data sets are created by sampling 1000 data points from a mixture of five Gaussian components. 100 uniformly distributed data points are generated to simulate the outliers. We demonstrate the performances of the the MoT, the mixture of Gaussian (MoG) and the MoP by using a 2-dimensional and a 5-dimensional dataset. The leftmost plots of Figs. 2, 3 (2-dimensional data) and Fig. 4 (5-dimensional data) show three example data sets with true grouping, while the remaining three plots in the three figures, show the estimated groupings by the MoT, the MoG and the MoP (clockwise), respectively. In Fig. 4, only the first two dimensions are displayed. From Fig. 2, it can be seen that the estimated clustering result by the MoP and the MoT compare well with the true clustering. But in Figs. 3 and 4, it can be observed that the MoT cannot recover the true grouping, while the MoP works better. In all the examples, as expected, the MoG is unable to find the right groupings due to the existence of outliers.

Fig. 5 shows the log-likelihoods obtained with varied number of mixture components ( $1 \leq K \leq 10$ ) by applying the MML criterion as described in Section 3.5. The same 2-dimensional and 5-dimensional datasets are used in Fig. 5(a) and (b), respectively. We run the MoP with MML 10 times for each  $K$ . Note that components will be pruned out during the training process if no enough data points support the components. The error bars in the figure show the standard deviation of the log-likelihood during the training of the runs. From the figure, we see that the maximum log-likelihoods are achieved at  $K=5$ , which indicates that the MoT with MML is able to find the optimal number of mixing components in the presence of outliers. Note that in both plots of Fig. 5, the maximum number of cluster  $K$  is no more than 8, this indicates that at least two mixture components were pruned out during the training.

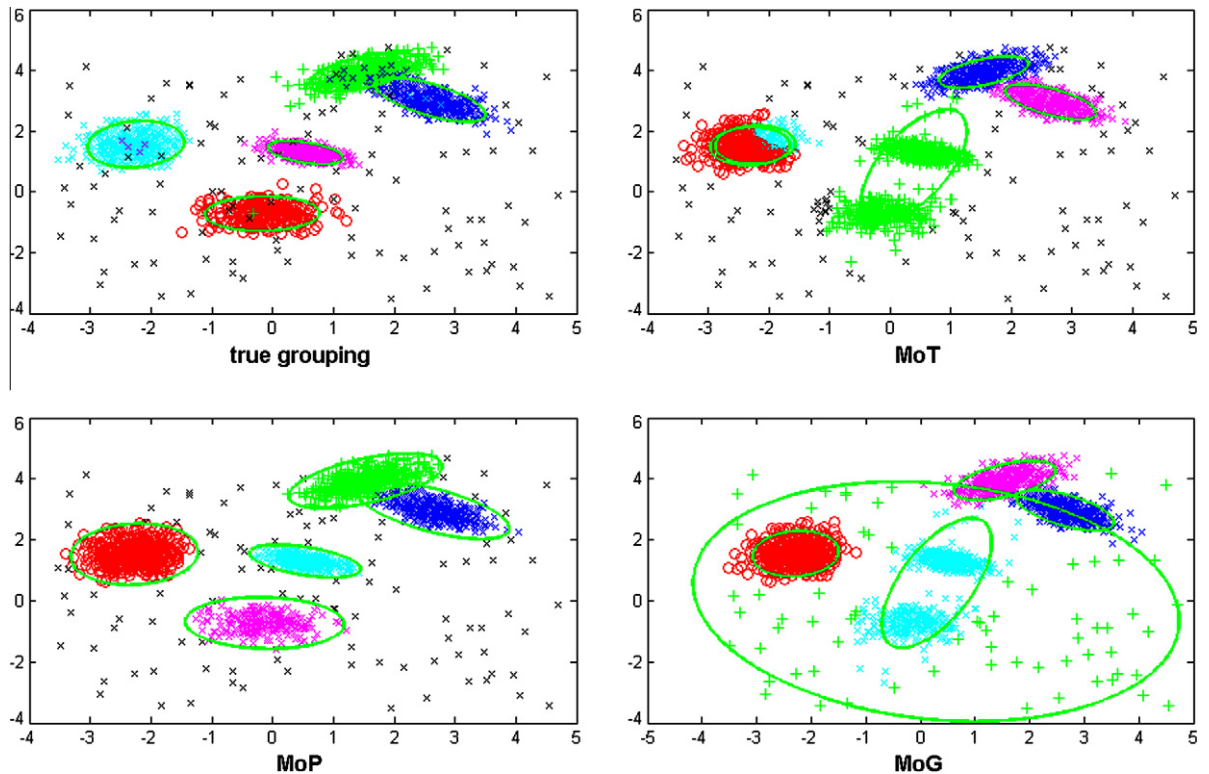
To compare the proposed EM algorithms with and without the latent variable  $u$ , the same 2-dimensional dataset is used. Fig. 6 shows the log-likelihoods obtained by using the EM algorithm with and without the latent variable  $u$  against iterations averaged over ten runs on the synthetic data set. From the figure, we see that the proposed algorithm is very close to the ground-truth EM algorithm.

### 4.2. Comparison of MoP and MoT on outlier detection

It is well known that the mixture of Student  $t$ -distribution (MoT) is capable of robust clustering and outlier detection (Peel and McLachlan, 2000). In this section, we compare the proposed algorithm (MoP) with the mixture of Student  $t$ -distribution (MoT) to explore the capability of the MoP method to detect outliers.



**Fig. 2.** An example illustrating the performance of MoT, MoG and MoP. The upper left plot shows the true grouping, while the remaining plots show the estimated groupings by MoT, MoG and MoP (clockwise), respectively.



**Fig. 3.** Another example illustrating the performance of MoT, MoG and MoP. The upper left plot shows the true grouping, while the remaining plots show the estimated groupings by MoT, MoG and MoP (clockwise), respectively.

We compare the MoP with the MoT in terms of (1) the rates of outlier detection accuracy on the training data sets and the test

data set; (2) the out-of-sample log-likelihood; and (3) the CPU time used for training. The out-of-sample log-likelihood will evaluate

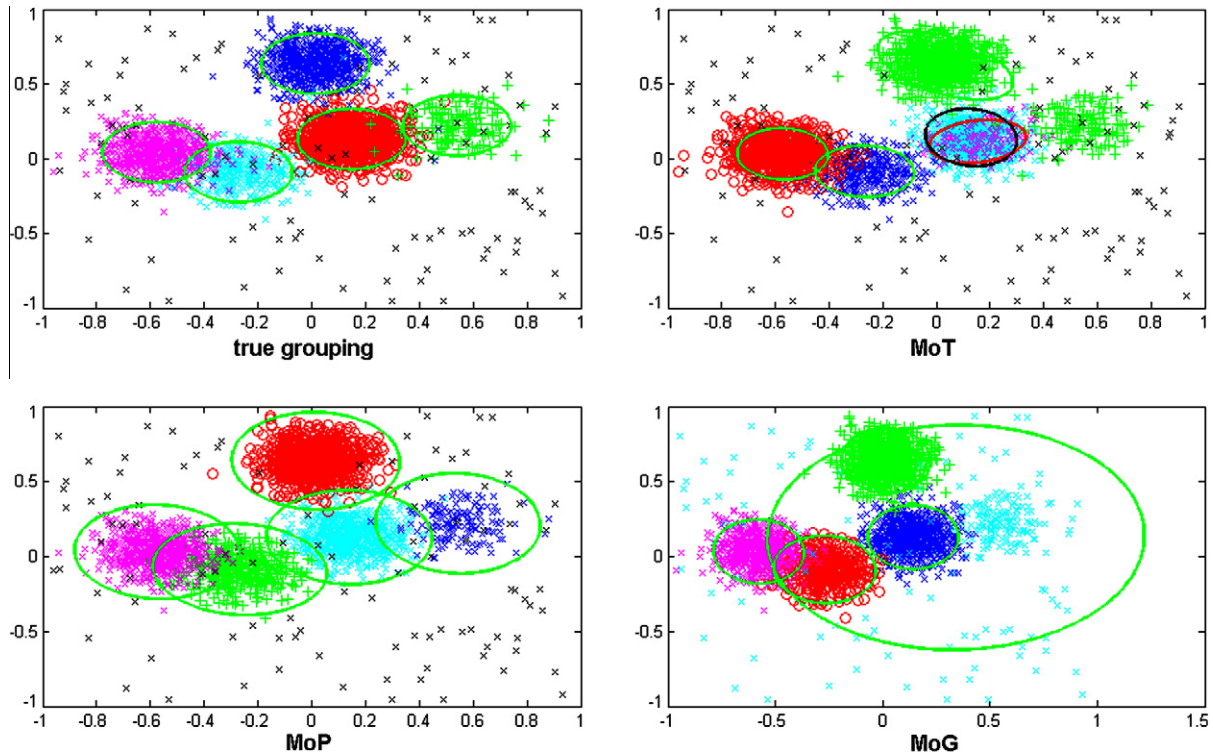


Fig. 4. A 5-dimensional data example illustrating the performance of MoT, MoG and MoP. The upper left plot shows the true grouping, while the remaining plots show the estimated groupings by MoT, MoG and MoP (clockwise), respectively.

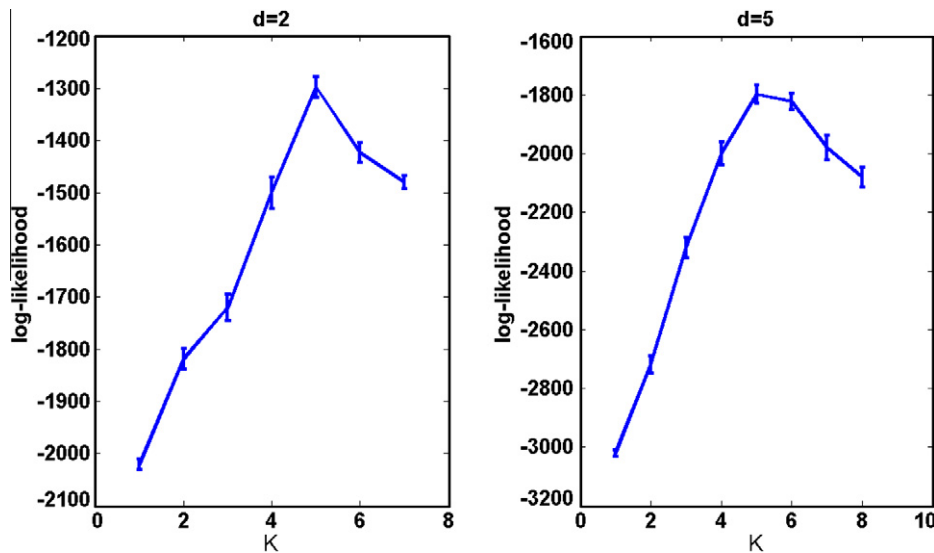


Fig. 5. The demonstration of the MoP with the minimum message length criterion to determine optimal number of mixture components. The data sets used in the plots are 2-dimensional and 5-dimensional, respectively.

the robust clustering capability of the MoP, and the CPU time will compare the convergence speeds of MoT and MoP.

Fig. 7 shows the results averaged over 50 runs on the synthetic datasets with 1000 data points and 5 clusters. In the experiments, we adopt the receiver operating characteristics (ROC) analysis (Fawcett, 2004) to measure the performance of outlier detection accuracy of the MoP and the MoT. The area under the ROC curve (AUC) gives us the probability that the outliers are correctly detected. In plots (a) and (b) of Fig. 7, the averaged AUC values are shown for the training and testing data sets. From the plots, we see that the average AUC values of the MoP is slightly better than

that of the MoP, but not significantly. In plot (d), the boxplot of the CPU time used by the MoT and the MoP are shown. It shows that the time used by the MoP is similar to that of the MoT, but with smaller variance. In the other words, the MoP appears more stable than the MoT.

From plot (c), it can be seen that the average out-of-sample log-likelihood of the MoP ( $-7.3101$ ) is slightly better than that of the MoT ( $-7.3584$ ). Though the improvement is not significant (the  $p$ -value obtained from the two-tailed  $t$ -test is 0.8914, which suggests that the null hypothesis cannot be rejected and so that the two means are equal), the study still demonstrates that the

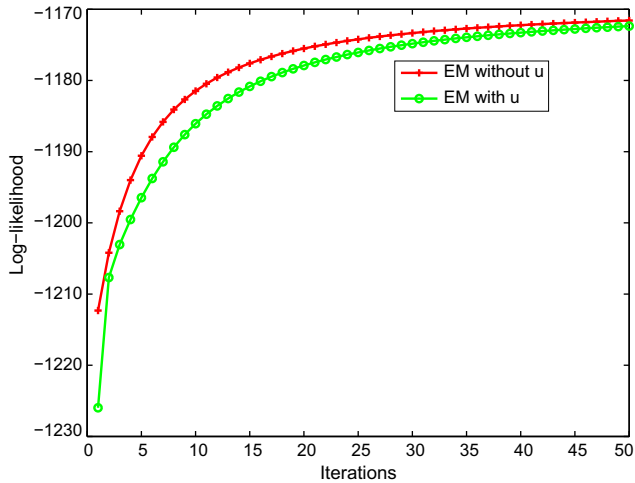


Fig. 6. The optimization process of the EM algorithms without latent variable  $u$  and the EM algorithm with latent variable  $u$  on the synthetic data set.

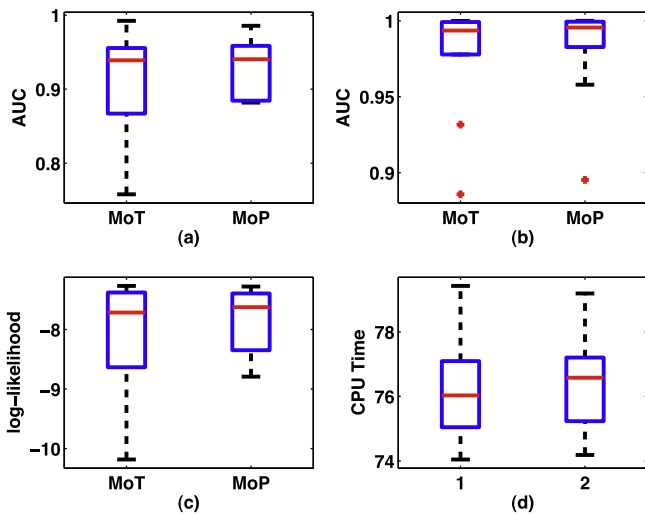


Fig. 7. The comparison of the MoP and the MoT in terms of (a) the outlier detection rates (measured by AUC) of the training data sets; (b) the AUC values of the test data sets; (c) the out-of-sample log-likelihood; (d) the CPU time spent on training.

new method may be used as an alternative to the Student  $t$ -distribution for robust learning. Moreover, as experienced in the experiments, we found that the solution to Eq. (16) sometimes is not available. That is, in the iteration of the Newton–Raphson algorithm, the intermediate solution becomes negative. In contrast, it is always easy to find the solution to Eq. (15) for the MoP. Thus, we conclude that the Pearson VII distribution is a better building block for robust learning than the Student  $t$ -distribution.

#### 4.3. Comparison of MoT, MoP and MoPII on classification

In this section, we compare the performance of the Student  $t$  mixtures, the Pearson type VII mixtures, and the Pearson type II mixtures on classification. Of the distributions of Pearson type, the Pearson-II has been demonstrated in the application of classification in (Medasani and Krishnapuram, 1999). Therefore, it is of interest to include it in our comparison. The Pearson type II distribution (Kotz, 1997) is written as

$$p(\mathbf{x}|\mu, A, \mathbf{K}) = \begin{cases} \frac{\Gamma(K+1+d/2)}{\Gamma(K+1)\pi^{d/2}} |\mathbf{W}|^{1/2} \mathbf{D}^K & \text{if } \mathbf{x} \in \text{Region } \mathcal{R} \\ 0 & \text{elsewhere} \end{cases} \quad (25)$$

where  $\mathbf{D} = [1 - (\mathbf{x} - \mu)^T \mathbf{W}(\mathbf{x} - \mu)]$ , and  $\mathbf{W} = (d + 2(K + 1))^{-1} \mathbf{A}^{-1}$ . The region  $\mathcal{R}$  denotes the interior of the hyper-ellipsoid, i.e.  $\mathcal{R} = \{\mathbf{x} | (\mathbf{x} - \mu)^T \mathbf{W}(\mathbf{x} - \mu) \leq 1\}$ . The parameter  $K$  determines the shape of the density function. In the algorithm developed in (Medasani and Krishnapuram, 1999), the estimation of the parameter  $K$  is not provided. In the following experiments, we tried the  $K$  values in the same way as suggested in (Medasani and Krishnapuram, 1999).

As in (Medasani and Krishnapuram, 1999), we applied Bayes' rule to carry out classification. The detailed procedure is as follows. First, we model each cluster  $\beta_k$  in the training data set as a mixture of multiple components. If we let  $p(x_j|\beta_k)$  be the conditional probability of selecting  $x_j$  given class  $\beta_k$  is given by  $p(x_j|\beta_k) = \sum_{i=1}^{N_k} p(\varpi_{ki}) p(x_j|\varpi_{ki})$  where  $N_k$  is the number of components in class  $\beta_k$ ,  $p(\varpi_{ki})$ ,  $1 \leq i \leq N_k$  is the mixing proportion of the class  $\beta_k$  and  $p(x_j|\varpi_{ki})$  is the conditional probability function. If we assume that  $P(\beta_k)$  is the priori probability for class  $\beta_k$ , then a data point  $x$  can be classified as class  $\beta_k$  if  $P(\beta_k)p(x|\beta_k) \geq P(\beta_\ell)p(x|\beta_\ell)$ ,  $\forall \ell \neq k$ .

To compare the different mixture modeling methods, five data sets, namely 'Breast Cancer', 'Pima Indian Diabetes', 'Heart Diseases', 'Bupa Liver', 'German Credit Card' and 'Wine' from the UCI data repository were used for this study. Table 1 lists the characteristics of the used data sets. In the table, 'NA' denotes 'not applicable' in case there are no data points in class 3.

In our experiments, each of the data sets was divided into training and testing sets, where 75% of the data are used as the training sets. In dividing the data set, we ensured that the data in the training set included all the classes. Twenty such divisions were generated randomly for each data set. To carry out classification of the data set, firstly the component parameters of the training data sets were estimated, then Bayes' rule was applied based on the resultant conditional probabilities for the classification of the training and testing data sets.

We use the classification rate, which is the accuracy of the classification, as the comparison criterion. The classification rates for the five data sets by using the three algorithms, MoP, MoT and MoPII are listed in Table 2. From the table, we see that considering the classification rate of the training data sets, the MoP performs better than the MoT except the Breast cancer data, and better than the MoPII except the Pima and the Wine data. Considering the classification rate of the test data sets, we see that the MoP is better

Table 1 Characteristics of the data sets.

Data sets	No. of classes	Data in			No. of features
		class 1	class 2	class 3	
Breast Cancer	2	444	239	NA	9
Pima Indian Diabetes	2	500	268	NA	8
Heart disease	2	150	120	NA	13
Bupa Liver Disorders	2	145	200	NA	6
German Credit Card	2	525	225	NA	24
Wine	3	59	71	48	13

Table 2

The comparison of the testing and training classification rate on five UCI benchmark data sets using MoP, MoT and MoPII.

Data sets	MoT		MoP		MoPII	
	Training	Testing	Training	Testing	Training	Testing
Breast Cancer	0.9841	0.9415	0.9836	0.9488	0.9529	0.9438
Pima Indian Diabetes	0.7773	0.6734	0.7806	0.6805	0.7758	0.7057
Heart disease	0.9700	0.5633	0.9957	0.5460	0.9114	0.5260
Bupa Liver Disorders	0.8469	0.6552	0.8510	0.6636	0.8180	0.6339
German Credit Card	0.7067	0.5924	0.8337	0.6752	0.7323	0.6024
Wine	0.9944	0.9111	0.9962	0.9222	1.0000	0.8500

than the MoT except the Heart disease data, and better than the MoPII except the breast cancer data.

From the results, again we may conclude that the proposed MoP is more viable than the MoT and the MoPII in classification.

## 5. Conclusion

We have developed a new robust mixture modeling approach based on the Pearson type VII distribution. The scale mixture representation of the Pearson type VII distribution is presented first. An EM algorithm is developed to fit the model with the maximum likelihood estimation. An outlier detection criterion is then derived from the EM formulation. Experimental results showed that the performance of mixture of Pearson VII distribution is similar to that of mixture of  $t$ -distribution in terms of outlier detection and likelihood maximization but more stable. On the other hand, the experiments on classification show that the Pearson type VII mixture is more stable than those of the Student- $t$  and the Pearson type II mixtures. In conclusion, the new method presented in this paper provides a more stable option for robust and sparse learning by using the Pearson type VII distribution as a building block.

We have already developed algorithms for robust clustering on data sets with measurement errors (Sun et al., 2007a; Sun and Kabán, 2010) and for visualising high-dimensional data sets with measurement errors (Sun et al., 2007b). These developed algorithms are all based on the mixtures of Student  $t$ -distributions. The newly developed Pearson Type VII distribution can be adopted in these studies to replace the Student  $t$ -distribution. It will be interesting to see the differences and this will be our future work.

## Appendix A

We show how the Pearson type VII distribution is represented as a scale mixture distribution. If we let

$$\tilde{p}(\mathbf{t}; \mu, \Lambda, m) = \int \mathcal{N}\left(\mathbf{t} | \mu, \frac{\Lambda}{u}\right) \mathcal{G}\left(u | m - \frac{d}{2}, \frac{1}{2}\right) du \quad (26)$$

We can evaluate the integral as follows:

$$\begin{aligned} \tilde{p}(\mathbf{t}; \mu, \Lambda, m) &= \frac{|\Lambda|^{-\frac{1}{2}}}{\pi^{\frac{d}{2}} 2^m \Gamma(\tilde{m})} \int u^{m-1} \exp\left\{-\left[\frac{1}{2}(1 + \Delta^2)\right]u\right\} du \\ &= \frac{|\Lambda|^{-\frac{1}{2}}}{\pi^{\frac{d}{2}} \Gamma(\tilde{m})} [1 + \Delta^2]^{-m} \int \tilde{u}^{m-1} \exp(-\tilde{u}) d\tilde{u} \end{aligned} \quad (27)$$

$$\begin{aligned} &= \frac{|\Lambda|^{-\frac{1}{2}} \Gamma(m)}{\pi^{\frac{d}{2}} \Gamma(\tilde{m})} [1 + \Delta^2]^{-m} \\ &= \frac{|\Lambda|^{-\frac{1}{2}} \Gamma(m)}{\pi^{\frac{d}{2}} \Gamma(\tilde{m})} [1 + \Delta^2]^{-m} \end{aligned} \quad (28)$$

(27) and (28), we re-parameterize

$$\tilde{u} = \frac{1}{2} [1 + \Delta^2] u$$

Therefore, we see that  $\tilde{p}(\mathbf{t}; \mu, \Lambda, m) = p(\mathbf{t}; \mu, \Lambda, m)$ .

## References

- Archambeau, C., Delannay, N., Verleysen, M., 2006. Robust probabilistic projections. In: Proc. 23rd Internat. Conf. on Machine Learning.
- Archambeau, C., Delannay, N., Verleysen, M., 2008. Mixtures of robust probabilistic principal component analyzers. *Neurocomputing* 71 (7–9), 1274–1282.
- Arellano-Valle, R., del Pino, G., Iglesias, P., 2006. Bayesian inference in spherical linear models: Robustness and conjugate analysis. *J. Multivar. Anal.* 97, 179–197.
- Fawcett, T., 2004. ROC graphs: Notes and practical considerations for researchers. *Machine Learn.*
- Fernández, C., Steel, M., 2000. Bayesian regression analysis with scale mixtures of normals. *Economet. Theor.* 16 (1), 80–101.
- Figueiredo, M., Jain, A., 2002. Unsupervised learning of finite mixture models. *IEEE Trans. Pattern Anal. Machine Intell.* 24 (3), 381–396.
- Kotz, S., 1997. Multivariate distributions at a cross-road. In: *Statistical Distributions in Scientific Work*. Reidel, Dordrecht.
- Medasani, S., Krishnapuram, R., 1999. A comparison of Gaussian and Pearson mixture modeling for pattern recognition and computer vision applications. *Pattern Recognition Lett.* 20, 305–313.
- Miranda, M., Fackler, P., 2002. *Applied Computational Economics and Finance*. The MIT Press.
- Murphy, P., Aha, D., 1973. UCI repository of machine learning databases. <<http://archive.ics.uci.edu/ml/>>.
- Nagahara, Y., 1996. Non-Gaussian distribution for stock returns and related stochastic differential equation. *Asia-Pacific Financ. Markets* 3 (2), 121–149.
- Pearson, K., 1916. Mathematical contributions to the theory of evolution, xix: Second supplement to a memoir on skew variation. *Philos. Trans. Roy. Soc. of London, Ser. A, Containing Papers of a Mathematical or Physical Character* 216, 429–457.
- Peel, D., McLachlan, G., 2000. Robust mixture modelling using the  $t$  distribution. *Statist. Comput.* 10, 339–348.
- Pike, M., Hill, I., 1966. Algorithm 291. *Comm. ACM* 9 (9), 684.
- Prevéy, P., 1986. The use of pearson VII distribution functions in X-ray diffraction residual stress measurement. *Adv. X-Ray Anal.* 29, 103–111.
- Sun, J., Kabán, A., 2010. A fast algorithm for robust mixtures in the presence of measurement errors. *IEEE Trans. Neural Network* 21 (8), 1206–1220.
- Sun, J., Kabán, A., Raychaudhury, S., 2007a. Robust mixtures in the presence of measurement errors. In: *Proc. 24th Internat. Conf. on Machine Learning*.
- Sun, J., Kabán, A., Raychaudhury, S., September 2007b. Robust visual mining of data with error information. In: *Proc. 11th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD07)*, Warsaw, Poland, pp. 573–580.
- Svensen, M., Bishop, C., 2005. Robust bayesian mixture modelling. *Neurocomputing* 64, 235–252.
- Vellido, A., Lisboa, P., Vicente, D., 2006. Handling outliers and missing data in brain tumor clinical assessment using t-GTM. *Comput Biol. Med.*
- West, M., 1987. On scale mixtures of normal distributions. *Biometrika* 74 (3), 646–648.