



# Robust Mixture Modeling using the Pearson Type VII Distribution

J. Sun

Centre for Plant Integrative Biology & School of Computer Science,  
University of Nottingham

16 Feb. 2012

- 1 Introduction
- 2 Pearson Type VII Distribution Mixtures
  - Developing the EM Algorithm for the Pearson Type VII Mixtures
  - Outlier Detection Criterion
- 3 Experimental Results
  - The synthetic data and illustrative demonstration
  - Comparison of MoT and MoP on Outlier Detection
- 4 Conclusions

# Outline

- 1 Introduction**
- 2 Pearson Type VII Distribution Mixtures**
  - Developing the EM Algorithm for the Pearson Type VII Mixtures
  - Outlier Detection Criterion
- 3 Experimental Results**
  - The synthetic data and illustrative demonstration
  - Comparison of MoT and MoP on Outlier Detection
- 4 Conclusions**



## Clustering

- Clustering is the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often proximity according to some defined distance measure.
- Data clustering is considered as a form of unsupervised learning in machine learning.



- Finite mixture modelling (FMM) is a commonly used parametric data clustering method in the machine learning community.
- In FMM, data are assumed to be sampled from a mixture of probability distributions, i.e.

$$p(t) = \sum_{k=1}^K \pi_k p(t|k)$$

where  $p(t)$  is the likelihood or probability of the data point  $t$ ,  $\pi_k, 1 \leq k \leq K$  is the mixing proportions, and  $p(t|k)$  is the conditional probability, i.e. the probability of the data point in the  $k$ -th component.



- With different variable types, we can assume different conditional probability  $p(t|k)$ .
- Mixture of Gaussian (MoG) is the most popular clustering method for continuous variables.
- In MoG, each mixture component  $p(t|k)$  is a Gaussian distribution  $\mathcal{N}(t|\mu_k, \Lambda_k)$ :

$$\mathcal{N}(t|\mu_k, \Lambda_k) = \frac{|\Lambda_k|^{-\frac{1}{2}}}{(2\pi)^{-d/2}} \exp\left(-\frac{1}{2}(t - \mu_k)^T \Lambda_k^{-1}(t - \mu_k)\right)$$



- Given a data set  $\mathcal{Y} = \{t_1, \dots, t_N\}$ , we need to estimate the parameters of the mixture models. In MoG, the parameters are  $\Theta = \{\mu_k, \Lambda_k, \pi_k, 1 \leq k \leq K\}$ .
- To estimate the parameters, the maximum likelihood (ML) approach is usually applied. We need to maximize the log-likelihood function:

$$\mathcal{L}(\Theta|\mathcal{Y}) = \log \left( \prod_{n=1}^N p(t_n|\Theta) \right) = \sum_{n=1}^N \log \left( \sum_{k=1}^K \pi_k p(t_n|k) \right)$$



- To maximize  $\mathcal{L}(\Theta|\mathcal{Y})$ , usually the Expectation-Maximization (EM) algorithm is applied.
- Notations:
  - $\mathcal{Y}$ : the observed variables;
  - $\mathcal{Z}$ : the latent variables;
  - $p(\mathcal{Y}, \mathcal{Z})$  is called the *complete likelihood*;
  - $p(\mathcal{Z}|\mathcal{Y})$  is called the *posterior distribution*.

## The EM Algorithm

- E-step (Expectation):

$$Q(\Theta, \Theta^{(i-1)}) = E \left[ \log p(\mathcal{Y}, \mathcal{Z} | \mathcal{Y}, \Theta^{(i-1)}) \right]$$

- M-step (Maximization):

$$\Theta^{(i)} = \arg \max_{\Theta} Q(\Theta, \Theta^{(i-1)})$$

where

$$\begin{aligned}
 & E \left[ \log p(\mathcal{Y}, \mathcal{Z} | \mathcal{Y}, \Theta^{(i-1)}) \right] \\
 &= \int_{z \in \Upsilon} \underbrace{\log p(\mathcal{Y}, z | \Theta)}_{\text{Complete log-Likelihood}} \underbrace{p(z | \mathcal{Y}, \Theta^{i-1})}_{\text{Posterior}} dz
 \end{aligned}$$



- To apply the EM algorithm to the MoG, we can introduce a latent variable  $z_n$  for each data point.  $z_n = k$  means that the data point  $t_n$  belongs to the  $k$ -th cluster.
- Given the latent variable, we can write the complete log-likelihood of the MoG as follows:

$$\mathcal{L}_C(\mathcal{Y}, \mathcal{Z}|\Theta) = \sum_n \sum_k \delta(z_n = k) \log \left[ p(t_n|k)\pi_k \right]$$

where  $\delta(z_n = k)$  is the Kronecker delta, and  $\mathcal{Z} = \{z_1, \dots, z_N\}$ .

The EM algorithm for the MoG can be described as follows:

- E-step: Compute the posterior of  $z_n$ :

$$p(z_n = k | t_n) = \frac{p(t_n | k) \pi_k}{\sum_{k'} p(t_n | k') \pi_{k'}}$$

- M-step: Maximize  $\mathcal{L}_C$  for the parameters:

$$\begin{aligned}\mu_k &= \frac{\sum_n p(z_n = k | t_n) t_n}{\sum_n p(z_n = k | t_n)} \\ \Lambda_k &= \frac{\sum_n p(z_n = k | t_n) (t_n - \mu_k) (t_n - \mu_k)^T}{\sum_n p(z_n = k | t_n)} \\ \pi_k &= \frac{\sum_n p(z_n = k | t_n)}{N}.\end{aligned}$$



- MoG usually works well.
- But it is sensitive to outliers.  
**Definition**<sup>1</sup>: An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs.
- **Demo.**

<sup>1</sup>: Barnett, V. and Lewis, T.: 1994, Outliers in Statistical Data. John Wiley & Sons., 3rd edition.



- To deal with the outliers, usually the Student t-distribution is used as the building block, i.e.

$$p(t|k) = \mathcal{S}_t(t|\mu_k, \Lambda_k, \nu_k)$$

where

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

- $\nu$  is the degree of freedom. It can be proved that as  $\nu \rightarrow \infty$ ,  $\mathcal{S}_t \rightarrow \mathcal{N}$

# Outline

- 1 Introduction
- 2 **Pearson Type VII Distribution Mixtures**
  - Developing the EM Algorithm for the Pearson Type VII Mixtures
  - Outlier Detection Criterion
- 3 **Experimental Results**
  - The synthetic data and illustrative demonstration
  - Comparison of MoT and MoP on Outlier Detection
- 4 **Conclusions**



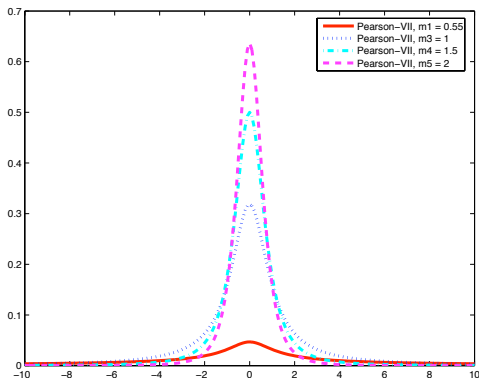
- We introduce a new distribution, called Pearson Type VII distribution, which subsumes the Student t-distribution.
- The PearsonVII distribution is of the following form:

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

where  $\mathbf{\Delta}^2 = (\mathbf{t} - \mu)^T \mathbf{\Lambda}^{-1} (\mathbf{t} - \mu)$ ,  $m > d/2$ ,  $\Gamma(\cdot)$  is the gamma function.

- The Student t-distribution is a special case of the Pearson type VII distribution if we set

$$m = \frac{\nu + d}{2}, \mathbf{\Lambda} = \mathbf{\Lambda}\nu.$$



**Figure:** The Pearson type VII distribution with different degrees of freedom.

- As before, we can introduce a discrete latent variable  $z_n$  for each data point. The E-step is the same as in the MoG:

$$p(z_n = k | t_n) = \frac{p(t_n | k) \pi_k}{\sum_{k'} p(t_n | k') \pi_{k'}}$$

except that  $p(t_n | k)$  is the Pearson Type VII distribution.

- In the maximization of the complete likelihood, we cannot obtain analytical solutions for  $\{\mu_k, \Lambda_k\}$  of the Pearson VII distributions.

$$\mathcal{L}_C(\mathcal{Y}, \mathcal{Z} | \Theta) \propto \sum_n \sum_k p(z_n = k | t_n) \log \left[ \pi_k |\Lambda_k|^{-\frac{1}{2}} [1 + \Delta_k^2]^{-m_k} \right]$$

- To obtain analytical solutions of the parameters, we can introduced a new latent variable  $u_n$  for each data point  $t_n$ .
- Precisely, we prove that:

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

where

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

is the gamma distribution with parameters  $a$  and  $b$ .



- Given the scale mixture representation of the Pearson VII distribution, the EM algorithm can be obtained.
- E-step: two latent variables  $z_n, u_n$  for each data point  $t_n$ . Their posteriors can be computed as follows:

$$p(z_n = k | t_n) = \frac{p(t_n | k) \pi_k}{\sum_{k'} p(t_n | k') \pi_{k'}}$$
$$p(u_n | t_n, k) = \mathcal{G}(u_n | a_{nk}, b_{nk})$$

where

$$a_{nk} = m_k$$
$$b_{nk} = \frac{(t_n - \mu_k)^T \Lambda_k^{-1} (t_n - \mu_k) + 1}{2}$$



- M-step. The parameters can be obtained:

$$\mu_k = \frac{\sum_n p(z_n = k|t_n) \langle u_n \rangle_k t_n}{\sum_n p(z_n = k|t_n) \langle u_n \rangle_k}$$

$$\Lambda_k = \frac{\sum_n p(z_n = k|t_n) \langle u_n \rangle_k (t_n - \mu_k)(t_n - \mu_k)^T}{\sum_n p(z_n = k|t_n)}$$

$$\pi_k = \frac{1}{N} \sum_n p(z_n = k|t_n)$$

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

## The M-Step Differences between MoG and MoP

$$\mu_k = \frac{\sum_n p(z_n = k|t_n) \langle u_n \rangle_k t_n}{\sum_n p(z_n = k|t_n) \langle u_n \rangle_k}$$
$$\Lambda_k = \frac{\sum_n p(z_n = k|t_n) \langle u_n \rangle_k (t_n - \mu_k)(t_n - \mu_k)^T}{\sum_n p(z_n = k|t_n)}$$



- The degree of freedom can be obtained by solving the following non-linear equation

$$\sum_n p(z_n = k|t_n) \left[ \langle \log u_n \rangle_k - \log(2) - \psi \left( m_k - \frac{d}{2} \right) \right] = 0$$

or specifically:

$$\psi \left( m_k - \frac{d}{2} \right) = \frac{\sum_n p(z_n = k|t_n) [\langle \log u_n \rangle_k - \log(2)]}{\sum_n p(z_n = k|t_n)}$$

where  $\psi(\cdot)$  is the digamma function, i.e.

$$\psi(x) = \frac{d}{dx} \log \Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}$$

- The posterior expectation of  $u$  can be used as an indicator of outliers.
- Given a data point  $t$ , if the value

$$e \equiv \sum_k p(z = k|t) \frac{2m_k}{(t - \mu_k)^T \mathbf{\Lambda}_k^{-1} (t - \mu_k) + 1}$$

is sufficiently small,  $t$  can be flagged as an outlier, or approximately, the value:

$$\kappa = \sum_k p(z = k|t) (\mathbf{t} - \mu_k)^T \mathbf{\Lambda}_k^{-1} (\mathbf{t} - \mu_k)$$

is sufficiently large.

# Outline

- 1 Introduction
- 2 Pearson Type VII Distribution Mixtures
  - Developing the EM Algorithm for the Pearson Type VII Mixtures
  - Outlier Detection Criterion
- 3 **Experimental Results**
  - The synthetic data and illustrative demonstration
  - Comparison of MoT and MoP on Outlier Detection
- 4 Conclusions

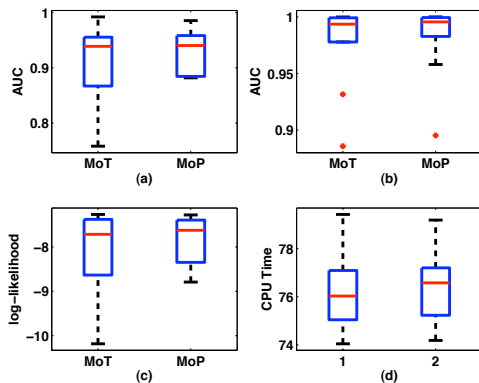
Demo



- The mixture of Student t-distribution (MoT) is capable of robust clustering and outlier detection.
- The comparison between the MoP and the MoT is in terms of
  - the rates of outlier detection accuracy on the training data sets and the test data set;
  - the out-of-sample log-likelihood (measuring the clustering capability); and
  - the CPU time used for training (measuring the convergence speed).



## Comparison of MoT and MoP on Outlier Detection



**Figure:** 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.

# Outline

- 1 Introduction
- 2 Pearson Type VII Distribution Mixtures
  - Developing the EM Algorithm for the Pearson Type VII Mixtures
  - Outlier Detection Criterion
- 3 Experimental Results
  - The synthetic data and illustrative demonstration
  - Comparison of MoT and MoP on Outlier Detection
- 4 Conclusions



- A new robust clustering algorithm based on Pearson Type VII mixtures.
- The scalar representation of the Pearson type VII distribution.
- An outlier detection criterion.
- Experimental results showed that the PearsonVII mixture is viable than the Student  $t$  mixtures.

Questions?