Model Selection Criterion For Multivariate Student's-t Mixture Model For Bounded Support Data

Ons Bouarada, Muhammad Azam, Nizar Bouguila
Concordia Institute for Information Systems Engineering
Concordia University, Montreal, QC, Canada

***Abstract*—We consider the task of modeling a multivariate data-set by a Student's t distribution mixture without knowing the prior number of clusters. Model Selection is the process of choosing the optimal number of mixture components (clusters) that ensures the best clustering accuracy. This paper proposes model selection using Minimum Message Length (MML) on bounded Student's t mixture model, along with applications on various data-sets and comparisons with other model selection procedures.**

***Index Terms*—Bounded Student's-t Mixture Model (BSMM), Minimum Message Length (MML), Model Selection, Data Clustering, Bayesian Estimation**

I. Introduction

Mixture models provide statistical inference on sub-populations of various random phenomena [1] and are commonly used for clustering tasks. Due to their high flexibility and their capability of modeling complex data distributions, mixture models are the subject of an increasing attention and have been applied on various distributions [2]. The mixture model based on the Student's-t distribution is our object of focus in this paper. By its definition, the Student's-t distribution is more heavily tailed than the Gaussian distribution, thus its mixture model (SMM) is more robust to outliers than the Gaussian mixture model (GMM) [3] In real world data, the values are usually concentrated within bounds. Therefore, it is suitable to introduce bounded support to every component of the mixture model, which allows more flexibility and better fitting to the different shapes of the data.

The multivariate Student's-t distribution has a small number of parameters: mean, covariance matrix, degrees of freedom and mixing parameter for each component of the mixture. These parameters can be estimated iteratively using the Expectation Maximization (EM) algorithm. Had we known the prior number of mixture components, the EM algorithm alone would have given a very good accuracy. However, this is not the case in real world data; the number of clusters is often unknown and needs to be initialized before proceeding to the EM algorithm. Choosing the number of components has a significant effect on the clustering performance, as too many components lead to overfitting the data and too few components reduce the model's flexibility.

To overcome this issue, various strategies have been presented to identify the optimal number of components: these can be stochastic, deterministic of resampling approaches [1]. More specifically, deterministic approaches include methods based on information theory concepts like the Minimum Message Length (MML) [4], Akaike's Information Criterion (AIC), the different versions of the Minimum Description Length (MDL) Criterion and the Mixture Minimum Description Length (MMDL).

This paper will focus on Minimum Message Length (MML) with the bounded multivariate Student's-t mixture model and compare it to other model selection algorithms like (...)

The remainder of this paper is organized as follows: section 2 describes the bounded Student's-t mixture model; section 3 explains the model selection with Minimum Message Length in detail; section 4 lays out the experimental results on 3 types of data-sets; and section 5 presents the conclusion.

II. Multivariate Bounded Student's-t Mixture Model

Let S be a multivariate Student's-t probability density function with the following parameters: a mean μ, a covariance matrix Σ and ν degrees of freedom. For a random vector x of dimension d, S can be written as follows:

$$S(x|\theta) = S(x|\mu,\Sigma,\nu) = \frac{1}{(\nu\pi)^{d/2}\Gamma(\nu/2)} \times \frac{\Gamma(\nu/2+d/2)|\Sigma|^{-1/2}}{[1+\nu^{-1}(x-\mu)^T\Sigma^{-1}(x-\mu)]^{(\nu+d)/2}} \tag{1}$$

Where $\theta = \{\mu, \Sigma, \nu\}$.

Let $X = x_1, \ldots, x_N$ denote an observed sample of N vectors of dimension d each. Modeling X as Student's-t mixture with K components implies that for every vector x_i, the ,marginal probability distribution of x_i is written as follows:

$$f(x_i|\Theta) = \sum_{k=1}^{K} \pi_k P(x_i|\theta_k) \tag{2}$$

Where π_k and θ_k are respectively the mixing proportion and the set of parameters for the k^{th} mixture component, and finally $\Theta = \{\theta_1, \ldots, \theta_K; \pi_1, \ldots, \pi_K\}$. π_k is the mixing proportion and represents the prior probability that x_i belongs to the k^{th} component, thus satisfies:

$$\pi_k \geq 0 \quad \text{and} \quad \sum_{k=1}^{K} \pi_k = 1 \tag{3}$$

$P(x_i|\theta_k)$ can be defined as a Student's-t probability density function with bounded support. We define a support region $\Omega_k \in \Re^d$ for each component k of the mixture model, and an indicator function as:

$$\mathrm{H}(x_i|\Omega_k) = \begin{cases} 1 & \text{if } x_i \in \Omega_k \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Then, $p(x_i)$ is defined as follows:

$$P(x_i|\theta_k) = \frac{S(x_i|\theta_k)H(x_i|\Omega_k)}{\int_{\Omega_k} S(x|\theta_k)\,dx} \tag{5}$$

The bounded aspect here is defined by multiplying the Student's-t PDF by the indicator function H. The division by the integral $\int_{\Omega_k} S(x|\theta_k)\,dx$ normalizes the bounded function P by the share of $S(x_i|\theta_k)$ that belongs to the support region Ω_k for the k^{th} component of the mixture [3].

A. Expectation Step

Now that we defined the PDF for this mixture model, we proceed to the expectation step of the EM algorithm. Here, at the iteration t of the EM algorithm, we define by τ_{ik} the posterior probability that the vector x_i belongs to the k^{th} component for $i \in 1, \ldots, N$ and $k \in 1, \ldots, K$.

$$\tau_{ik}^{(t)} = \frac{\pi_k^{(t} P(x_i|\theta_k^{(t)})}{\sum_{j=1}^{K} \pi_j^{(t)} P(x_i|\theta_j^{(t)})} \tag{6}$$

The estimation step includes also calculating the log-likelihood of the model $L(\Theta)$ at the current iteration:

$$L(\Theta) = \sum_{i=1}^{N} \log\Big(\sum_{k=1}^{K} \pi_k P(x_i|\theta_k)\Big) \tag{7}$$

B. Maximization Step

The goal of the Maximization step in the EM algorithm is to update the model parameters in a way that maximizes the previously calculated log-likelihood function [5]. As the logarithm is monotonically increasing, it is more suitable to minimize the negative log-likelihood function $J(\Theta) = -L(\Theta)$.

By applying the Jensen inequality, we find that at the t^{th} iteration:

$$J(\Theta) \leq -\sum_{i=1}^{N}\sum_{k=1}^{K} \tau_{ik}^{(t)} \Big\{ \log \pi_k + \log S(x_i|\theta_k) - \log \int_{\Omega_k} S(x|\theta_k)\,dx \Big\} \tag{8}$$

Thus minimizing $J(\Theta)$ becomes equivalent to minimizing $E(\Theta)$, where:

$$E(\Theta) = -\sum_{i=1}^{N}\sum_{k=1}^{K} \tau_{ik}^{(t)} \Big\{ \log \pi_k + \log S(x_i|\theta_k) - \log \int_{\Omega_k} S(x|\theta_k)\,dx \Big\} \tag{9}$$

In this case, we regard $E(\Theta)$ as an error function that needs to be minimized [3]. Therefore, we calculate the partial derivatives of $E(\Theta)$ with respect to every parameter in Θ, then we solve the equations:

$$\frac{\partial E(\Theta)}{\partial \mu_k} = 0 \quad ; \quad \frac{\partial E(\Theta)}{\partial \Sigma_k} = 0 \quad ; \quad \frac{\partial E(\Theta)}{\partial \nu_k} = 0$$

For the component k, the updates for the mean vector $\nu_k^{(t+1)}$ and the covariance matrix $\Sigma_k^{(t+1)}$ are solutions for the first and second equation respectively. After calculations, these solutions are the following:

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{N} z_{ik}^{(t)} (h(x_i|\theta_k^{(t)})x_i - R_k)}{\sum_{i=1}^{N} z_{ik}^{(t)} h(x_i|\mu_k^{(t)}, \Sigma_j^{(t)}, v_k^{(t)})} \tag{10}$$

$$\Sigma_j^{(t+1)} = \frac{\sum_{i=1}^{N} z_{ik}^{(t)} h(x_i|\theta_k^{(t)})(x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{N} z_{ik}^{(t)}} - G_k \tag{11}$$

Where h, R_k and G_k are respectively defined as follows:

$$h(x|\theta_k) = \frac{\nu_k + D}{\nu_k + (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)} \tag{12}$$

$$R_k = \frac{\sum_{m=1}^{M} (S_{mk} - \mu_k^{(t)}) h(S_{mk}; \theta_k^t) H(S_{mk}|\Omega_k)}{\sum_{m=1}^{M} H(S_{mk}|\Omega_k)} \tag{13}$$

$$G_k = \frac{1}{\sum_{m=1}^{M} H(S_{mk}|\Omega_k)} \times \sum_{m=1}^{M} \Big(\Sigma_k^{(t)} - (S_{mk} - \mu_k^{(t)}) \times (S_{mk} - \mu_k^{(t)})^T h(S_{mk}|\theta_k^{(t)})\Big) H(S_{mk}|\Omega_k) \tag{14}$$

S_{mk} is a generated sample of m vectors from the Student's-t distribution S with the parameters $\Theta_k = \{\mu_k, \Sigma_k, \nu_k\}$. As for the degrees of freedom, their update for the k^{th} component $\nu_k^{(t+1)}$ is a solution to the equation:

$$-\psi(v_j/2) + \log(v_j/2) + \psi(v_j/2 + D/2) - \log(v_j/2 + D/2) + 1 + \frac{\sum_{i=1}^{N} z_{ij}^{(t)} (\log(h(x_i|\mu_j, \Sigma_j, v_j) - h(x_i|\mu_j, \Sigma_j, v_j))}{\sum_{i=1}^{N} z_{ij}^{(t)}} - F_j = 0 \tag{15}$$

Where

$$\mathrm{F}_j = \frac{1}{\sum_{m=1}^{M} \mathrm{H}(\mathrm{s}_{mj}|\Omega_j)} \sum_{m=1}^{\mathrm{M}} \left(-\psi\left(v_j/2\right) + \log\left(v_j/2\right) + 1 + \log h\left(s_m \mid \mu_j^{(t)}, \Sigma_j^{(t)}, v_j^{(t)}\right) - h\left(s_{mj} \mid \mu_j^{(t)}, \Sigma_j^{(t)}, v_j^{(t)}\right) + \psi\left(v_j/2 + D/2\right) - \log\left(v_j/2 + D/2\right)\right) \mathrm{H}\left(\mathrm{s}_{mj} \mid \Omega_j\right) \tag{16}$$

III. Model Selection using Minimum Message Length

As its name suggests, the Minimum Message Length (MML) method is based on compressing a message that contains the data clustered by the evaluated mixture model [1], [6]. The better fit is the model, the greater is its capacity to compress. This approach suggests modeling the observed data

$X = x_1, \ldots, x_N$ by a mixture of distributions with different number of components K, and calculating the message length $MessLen$ (i.e., the amount of measured information after data compression) for each value of K. Then, the mixture that has the optimal number of clusters is the one that scores the minimum message length. $MessLen$ is defined as follows

$$MessLen(K) \simeq -\log\big(p(\Theta_K)\big) - \mathcal{L}(\Theta_K, X) + \frac{1}{2}\log F(\Theta_K) + \frac{N_p}{2}(1+\log(\frac{1}{12})) \tag{17}$$

Where $p(\Theta_K)$ is the prior probability, $\mathcal{L}(\Theta_K, X)$ is the log-likelihood and $F(\Theta_K)$ is the Fisher information matrix. The Fisher matrix is used to measure the amount of information contained in the evaluated mixture model.

A. Fisher information matrix calculation

For a random variable x that follows a distribution f around a parameter θ, the Fisher information $I_x(\theta)$ describes how sensitive f is to changes in the parameter θ [6]. It is defined as the following:

$$I_X(\theta) = \sum_{x \in X} \Big(\frac{\partial}{\partial \theta} \log f(x|\theta)\Big)^2 P_\theta(x) \tag{18}$$

For the case of our mixture model, the determinant of the Fisher information matrix is the product of all information matrices for all mixture components [1]. of the information matrix of

IV. Experiments and Results

The model is implemented with python and we have 3 main experiments on 3 different types of datasets: synthetic data, a commonly used dataset for clustering and a video dataset (for feature extraction)

A. Experiment1

B. Experiment2

... .

C. Discussions

V. Conclusion

The conclusion goes here.

Acknowledgment

The authors would like to thank...

References

[1] N. Bouguila and D. Ziou, "High-dimensional unsupervised selection and estimation of a finite generalized dirichlet mixture model based on minimum message length," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1716–1717, 2007.

[2] D. Peel and G. J. Mclachlan, "Robust mixture modelling using the t distribution," *Statistics and Computing*, no. 10, pp. 339–348, 2000.

[3] T. M. Nguyen and Q. J. Wu, "Multivariate student's-t mixture model for bounded support data," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Windsor, ON, Canada, 2013, pp. 5548–5552.

[4] M. A. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 381–396, 2002.

[5] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[6] M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain, "Simultaneous feature selection and clustering using mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1154–1166, 2004.