MULTIVARIATE STUDENT'S-T MIXTURE MODEL FOR BOUNDED SUPPORT DATA

Thanh Minh Nguyen, Q. M. Jonathan Wu, Senior Member, IEEE

Department of Electrical and Computer Engineering, University of Windsor
401 Sunset Avenue, Windsor, ON, N9B3P4, Canada
{nguyen1j, jwu}@uwindsor.ca

ABSTRACT

The finite mixture model based on the Student's-t distribution, which is heavily tailed and more robust than the Gaussian mixture model (GMM), is a flexible and powerful tool to address many pattern recognition problems. However, the Student's-t distribution is unbounded. In many applications, the observed data are digitalized and have bounded support. A new finite multivariate Student's-t mixture model for bounded support data, which includes the GMM and the Student's-t mixture model (SMM) as special cases, is presented in this paper. We propose an extension of the Student's-t distribution in this paper. This new distribution is sufficiently flexible to fit different shapes of observed data, such as non-Gaussian, non-symmetric, and bounded support data. Another advantage of the proposed model is that each of its components can model the observed data with different bounded support regions. In order to estimate the model parameters, previous models represent the Student's-t distributions as an infinite mixture of scaled Gaussians. We propose an alternate approach in order to minimize the higher bound on the data negative log-likelihood function, and directly deal with the Student's-t distribution.

Index Terms— Bayesian estimation, bounded support regions, Density estimation.

1. INTRODUCTION

The finite mixture model is widely used in machine learning areas. The main advantage of this technique is in its capability to use prior knowledge to model the uncertainty in a probabilistic manner. In this technique, the Gaussian mixture model (GMM) [1–3] is a well-known method. The major advantage of the GMM is that the log-likelihood function used to estimate the parameters is inherently simple. Another advantage of the GMM is that it is easy to implement and requires a small number of parameters. These parameters can be efficiently estimated by adopting the expectation maximization (EM) algorithm [4–6]. However, the GMM is sensitive to outliers and may lead to excessive sensitivity to small numbers of data points [7, 8]. Also, for many applied problems, the tail of the Gaussian distribution is shorter than required.

Another way to fit different shapes of observed data is to use the generalized Gaussian mixture model (GGMM) [9,10]. In this model, each component is represented by a multi-dimensional generalized Gaussian distribution $\mathrm{T}(\mathrm{x}_i|\mu_j, \Sigma_j, \lambda_j)$. This distribution has one parameter more than the Gaussian distribution $\Phi(\mathrm{x}_i|\mu_j, \Sigma_j)$. The parameter λ_j controls the tails of the distribution and determines whether the latter is peaked or flat. However, the major disadvantage of GGMM is that this model assumes that the dimensions of the observed data are independent. Therefore, it is not suitable for analyzing correlated data.

Use of the finite Student's-t mixture model (SMM) has been proposed in order to improve the robustness of the algorithm [7,11]. In this model, each of the mixture's components is a Student's-t distribution $S(\mathrm{x}_i|\mu_j, \Sigma_j, v_j)$. The main advantage of the Student's-t distribution is that it is more heavily tailed than the Gaussian distribution. Unlike the GMM, each component of the SMM has an additional paramete–the degrees of freedom (v_j)–which is which is a robustness tuning parameter. Hence, the SMM of the longer tailed multivariate Student's-t distribution provides a much more robust approach than the GMM. In order to estimate the model parameters by adopting the EM algorithm, the Student's-t distribution in the previous model is represented as an infinite mixture of scaled Gaussians [11], which corresponds to an increase in the algorithm's complexity.

One drawback of the above-mentioned mixture models is that their distributions are unbounded with a support range of $(-\infty, +\infty)$. We observe in many real applications that the observed data always fall within the bounded support regions, and that the dimensions of the observed data are correlated. For example, in the area of signal processing, the power spectrum is semi-bounded. In the area of image computer vision, the pixels are usually in the limited range. Motivated by the aforementioned observations, we introduce in this paper a multivariate Student's-t mixture model for bounded support data, based on modeling of the probability density function. Differing from the above-mentioned mixture models, each component density in our model can model

This research has been supported in part by the Canada Research Chair Program, AUTO21 NCE, and the NSERC Discovery grant.

the observed data with different bounded support regions. We propose an extension of the Student's-t distribution that has a flexibility to fit different shapes of observed data, such as non-Gaussian, non-symmetric, and bounded support data. The proposed model can be used to analyze both univariate and multivariate data. We directly deal with the Student's-t distribution by proposing an alternate approach to estimate the model parameters, whereas the Student's-t distributions in previous models are represented as infinite mixtures of scaled Gaussians. We demonstrate through extensive simulations that the proposed model is superior to other methods based on the modeling of the probability density function of the data via finite mixture model. The remainder of this paper is organized as follows: section 2 describes the proposed method in detail; section 3 presents the parameter estimation; section 4 sets out the experimental results; and section 5 presents our conclusions.

2. PROPOSED METHOD

Given a finite mixture model with K components, the marginal distribution of the random variable x_i is

$$f(\mathrm{x}_i|\Theta) = \sum_{j=1}^{K} \pi_j p(\mathrm{x}_i|\Omega_j) \tag{1}$$

where Θ represents the model parameters. The prior probability π_i that pixel x_i is in label Ω_j satisfies the constraints

$$\pi_j \geq 0 \text{ and } \sum_{j=1}^{K} \pi_j = 1 \tag{2}$$

As shown in (1), the mixture models have relied on $p(\mathrm{x}_i|\Omega_j)$ to model the underlying distributions. Note that, $p(\mathrm{x}_i|\Omega_j)$ can be any kind of distribution. In GMM [1], GGMM [9, 10], and SMM [7, 11], $p(\mathrm{x}_i|\Omega_j))$ is the Gaussian distribution $\Phi(\mathrm{x}_i|\mu_j, \Sigma_j)$, generalized Gaussian distribution $\mathrm{T}(\mathrm{x}_i|\mu_j, \Sigma_j, \lambda_j)$, and the Student's-t distribution $S(\mathrm{x}_i|\mu_j, \Sigma_j, v_j)$, respectively. These distributions are all unbounded with support range $(-\infty, +\infty)$. In order to overcome this problem, we propose a new finite mixture model with bounded support region, non-Gaussian, non-symmetric distribution. First, for each label Ω_j, we define ∂_{Ω_j} to be the bounded support region in $\mathbb{R}^D$, and the indicator function as

$$\mathrm{H}(\mathrm{x}_i|\Omega_j) = \begin{cases} 1 \text{ IF } \mathrm{x}_i \in \partial_{\Omega_j} \\ 0 \text{ Otherwise} \end{cases} \tag{3}$$

And the multivariate Student's-t distribution $S(\mathrm{x}_i|\mu_j, \Sigma_j, v_j)$ as follows:

$$\begin{aligned} S(\mathrm{x}_i|\mu_j, \Sigma_j, v_j) &= \frac{1}{(v_j\pi)^{D/2}\Gamma(v_j/2)} \\ &\times \frac{\Gamma(v_j/2 + D/2)|\Sigma_j|^{-1/2}}{\left[1 + v_j^{-1}(\mathrm{x}_i - \mu_j)^{\mathrm{T}}\Sigma_j^{-1}(\mathrm{x}_i - \mu_j)\right]^{(v_j+D)/2}} \end{aligned} \tag{4}$$

In (4), i=(1,2,...,N), j=(1,2,...,K). The D-dimensional vector μ_j is the mean. The $D\mathrm{x}D$ matrix Σ_j is the covariance, $|\Sigma_j|$ denotes the determinant of Σ_j, and v_j is the degree of freedom. With the indicator function $\mathrm{H}(\mathrm{x}_i|\Omega_j)$ in (3) and the distribution $S(\mathrm{x}_i|\mu_j, \Sigma_j, v_j)$ in (4), we define a bounded multivariate Student's-t distribution :

$$\Psi(\mathrm{x}_i|\mu_j, \Sigma_j, v_j) = \frac{S(\mathrm{x}_i|\mu_j, \Sigma_j, v_j)\mathrm{H}(\mathrm{x}_i|\Omega_j)}{\int_{\partial_{\Omega_j}} S(\mathrm{x}|\mu_j, \Sigma_j, v_j)d\mathrm{x}} \tag{5}$$

In (5), $\int_{\partial_{\Omega_j}} S(\mathrm{x}|\mu_j, \Sigma_j, v_j)d\mathrm{x}$ is the normalization constant, and is identified as the share of $S(\mathrm{x}_i|\mu_j, \Sigma_j, v_j)$ that belongs to the support region ∂_{Ω_j}. The idea to define the distribution $\Psi(\mathrm{x}_i|\mu_j, \Sigma_j, v_j)$ in (5) is based on the fact that the observed data are digitalized and have bounded support. We assign $\Psi(\mathrm{x}_i|\mu_j, \Sigma_j, v_j)$ as equal to $S(\mathrm{x}_i|\mu_j, \Sigma_j, v_j)$ in the support region ∂_{Ω_j}, and as zero outside. It is worth mentioning that the proposed distribution in (5) will always satisfy the conditions of the probability density [2]:

$$\Psi(\mathrm{x}_i|\mu_j, \Sigma_j, v_j) \geq 0 \text{ and } \int_{-\infty}^{\infty} \Psi(\mathrm{x}|\mu_j, \Sigma_j, v_j)d\mathrm{x} = 1 \tag{6}$$

Given the distribution $\Psi(\mathrm{x}_i|\mu_j, \Sigma_j, v_j)$ in (5), the log-likelihood function is written in the form

$$L(\Theta) = \sum_{i=1}^{N} \log \sum_{j=1}^{K} \pi_j \frac{S(\mathrm{x}_i|\mu_j, \Sigma_j, v_j)\mathrm{H}(\mathrm{x}_i|\Omega_j)}{\int_{\partial_{\Omega_j}} S(\mathrm{x}|\mu_j, \Sigma_j, v_j)d\mathrm{x}} \tag{7}$$

From (7), we can see that each component of the proposed model has the ability to model the observed data with different bounded support regions ∂_{Ω_j}. We can define any shape of ∂_{Ω_j} based on prior knowledge about the observed data. Given the log-likelihood function in (7), our next objective is to optimize the parameter set in order to maximize this log-likelihood function.

3. PARAMETER LEARNING

In this section, rather than using the EM algorithm to estimate the model parameters and to maximize the log-likelihood function in (7), we propose an alternate approach to minimize the negative log-likelihood function. Another difference between the proposed model and the above-mentioned models is the approach used to estimate the parameters of the Student's-t distribution. In order to estimate the model parameters, we deal directly with the Student's-t distribution, whereas the Student's-t distribution in existing models is represented as an infinite mixture of scaled Gaussians. In order to determine the label Ω_j to which the observed data x_i should be assigned, we need to adjust the parameters $\Theta = \{\pi_j, \eta_j, \mu_j, \Sigma_j, v_j\}$ in order to maximize the likelihood function in (7). Since the logarithm is a monotonically increasing function, it is

more convenient to consider the negative logarithm of the likelihood function in (7) as an error function

$$J(\Theta) = -L(\Theta) = -\sum_{i=1}^{N} \log \sum_{j=1}^{K} \pi_j \Psi(\mathrm{x}_i|\mu_j, \Sigma_j, v_j) \quad (8)$$

Therefore, maximizing the likelihood $L(\Theta)$ is equivalent to minimizing $J(\Theta)$. In order to minimize the error function $J(\Theta)$, we define variable $z_{ij}^{(t)}$ as

$$z_{ij}^{(t)} = \frac{\pi_j^{(t)} \Psi(\mathrm{x}_i|\mu_j^{(t)}, \Sigma_j^{(t)}, v_j^{(t)})}{\sum_{m=1}^{K} \pi_m^{(t)} \Psi(\mathrm{x}_i|\mu_m^{(t)}, \Sigma_m^{(t)}, v_m^{(t)})} \quad (9)$$

where t indicates the iteration step. Because the values $z_{ij}^{(t)}$ in (9) always satisfy the conditions $\sum_{j=1}^{K} z_{ij}^{(t)} = 1$, we can apply Jensen's inequality [3], in the form $\log(\sum_{j=1}^{K} z_{ij}^{(t)} s) \geq \sum_{j=1}^{K} z_{ij}^{(t)} \log(s)$ to the error function in (8). This gives

$$J(\Theta) \leq -\sum_{i=1}^{N} \sum_{j=1}^{K} z_{ij}^{(t)} \left\{\log \pi_j + \log \Psi(\mathrm{x}_i|\mu_j, \Sigma_j, v_j)\right\} \quad (10)$$

Minimizing the negative log-likelihood function in (8), is equivalent to minimizing the error function $E(\Theta)$

$$E(\Theta) = -\sum_{i=1}^{N} \sum_{j=1}^{K} z_{ij}^{(t)} \left\{\log \pi_j + \log S(\mathrm{x}_i|\mu_j, \Sigma_j, v_j) - \log \int_{\partial\Omega_j} S(\mathrm{x}|\mu_j, \Sigma_j, v_j) d\mathrm{x}\right\} \quad (11)$$

To minimize this function, we consider the derivation of the error function $E(\Theta)$ with the means μ_j, Σ_j^{-1} and v_j at the (t+1) iteration step. According to the theory of robust statistics [12], any estimate T is defined by an implicit equation:

$$\sum_i \Upsilon(\mathrm{x}_i - \mathrm{T}) = 0 \quad (12)$$

This gives a numerical solution of the location of T as a weighted mean:

$$\mathrm{T} = \frac{\sum_i w_i \mathrm{x}_i}{\sum_i w_i}; \text{ where } w_i = \frac{\Upsilon(\mathrm{x}_i - \mathrm{T})}{\mathrm{x}_i - \mathrm{T}} \quad (13)$$

Now we can apply (12) to the $\partial E(\Theta)/\partial \mu_j = 0$, $\partial E(\Theta)/\partial \Sigma_j^{-1} = 0$, and $\partial E(\Theta)/\partial v_j = 0$. After some manipulation [13], we have the estimates of μ_j, Σ_j, and v_j at the (t+1) step

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^{N} z_{ij}^{(t)} \left(h(\mathrm{x}_i|\mu_j^{(t)}, \Sigma_j^{(t)}, v_j^{(t)}) \mathrm{x}_i - \mathrm{R}_j\right)}{\sum_{i=1}^{N} z_{ij}^{(t)} h(\mathrm{x}_i|\mu_j^{(t)}, \Sigma_j^{(t)}, v_j^{(t)})} \quad (14)$$

$$\Sigma_j^{(t+1)} = \frac{\sum_{i=1}^{N} z_{ij}^{(t)} h(\mathrm{x}_i|\mu_j^{(t)}, \Sigma_j^{(t)}, v_j^{(t)}) (\mathrm{x}_i - \mu_j)(\mathrm{x}_i - \mu_j)^{\mathrm{T}}}{\sum_{i=1}^{N} z_{ij}^{(t)}} - \mathrm{G}_j \quad (15)$$

The estimates of the degrees of freedom v_j are given by the solution of the equation

$$-\psi(v_j/2) + \log(v_j/2) + \psi(v_j/2 + D/2) - \log(v_j/2 + D/2) + 1 + \frac{\sum_{i=1}^{N} z_{ij}^{(t)} (\log h(\mathrm{x}_i|\mu_j, \Sigma_j, v_j) - h(\mathrm{x}_i|\mu_j, \Sigma_j, v_j))}{\sum_{i=1}^{N} z_{ij}^{(t)}} - \mathrm{F}_j = 0 \quad (16)$$

In (14), (15), (16), the $h(\mathrm{x}|\mu_j, \Sigma_j, v_j)$, R_j, G_j, and F_j are given by

$$h(\mathrm{x}|\mu_j, \Sigma_j, v_j) = \frac{v_j + D}{v_j + (\mathrm{x} - \mu_j)^T \Sigma_j^{-1} (\mathrm{x} - \mu_j)} \quad (17)$$

$$\mathrm{R}_j = \frac{\sum_{m=1}^{M} (\mathrm{s}_{mj} - \mu_j^{(t)}) h(\mathrm{s}_{mj}|\mu_j^{(t)}, \Sigma_j^{(t)}, v_j^{(t)}) \mathrm{H}(\mathrm{s}_{mj}|\Omega_j)}{\sum_{m=1}^{M} \mathrm{H}(\mathrm{s}_{mj}|\Omega_j)} \quad (18)$$

$$\mathrm{G}_j = \frac{1}{\sum_{m=1}^{M} \mathrm{H}(\mathrm{s}_{mj}|\Omega_j)} \sum_{m=1}^{M} (\Sigma_j^{(t)} - (\mathrm{s}_{mj} - \mu_j^{(t)}) \times (\mathrm{s}_{mj} - \mu_j^{(t)})^T h(\mathrm{s}_{mj}|\mu_j^{(t)}, \Sigma_j^{(t)}, v_j^{(t)})) \mathrm{H}(\mathrm{s}_{mj}|\Omega_j) \quad (19)$$

$$\mathrm{F}_j = \frac{1}{\sum_{m=1}^{M} \mathrm{H}(\mathrm{s}_{mj}|\Omega_j)} \sum_{m=1}^{M} (-\psi(v_j/2) + \log(v_j/2) + 1 + \log h(\mathrm{s}_{mj}|\mu_j^{(t)}, \Sigma_j^{(t)}, v_j^{(t)}) - h(\mathrm{s}_{mj}|\mu_j^{(t)}, \Sigma_j^{(t)}, v_j^{(t)}) + \psi(v_j/2 + D/2) - \log(v_j/2 + D/2)) \mathrm{H}(\mathrm{s}_{mj}|\Omega_j) \quad (20)$$

where $\mathrm{s}_{mj} \sim S(\mathrm{x}|\mu_j^{(t)}, \Sigma_j^{(t)}, v_j^{(t)})$ denotes the random vector that is drawn from the probability distribution S, and M is the number of random vectors s_{mj}. Note that, M is a large integer value. We use $M = 10^6$ for our experiments. The next step is to update the estimate of the prior probability π_j. The constraint $\sum_{j=1}^{K} \pi_j = 1$ enables

$$\pi_j = \frac{1}{N} \sum_{i=1}^{N} z_{ij}^{(t)} \quad (21)$$

So far, the discussion has focused on estimating $\Theta = \{\pi_j, \mu_j, \Sigma_j, v_j\}$ of the model. In the next section, we demonstrate the robustness, accuracy and effectiveness of the proposed model compared to other approaches.

4. EXPERIMENTS

We begin with one experiment on simulated data in Fig. 1 to demonstrate the robustness of the proposed distribution. As shown, the observed data is in the interval (-0.5, 0.5). In this experiment, the number of components for all compared methods is assigned a value of 2 (K=2). In Fig. 1(b), the values of the statistic $\mathcal{X}^2$ [10] obtained by GMM [1, 2] method is very poor , with $\mathcal{X}^2$=84854.08. The GGMM [9, 10] and SMM [7, 11] methods slightly improve the result, with $\mathcal{X}^2$=30874.85 and $\mathcal{X}^2$=21043.41, respectively. However, the estimated histogram remains poor. Compared to GMM, GGMM, and SMM, we find that our method is the most robust and has the lowest $\mathcal{X}^2$=10186.47.

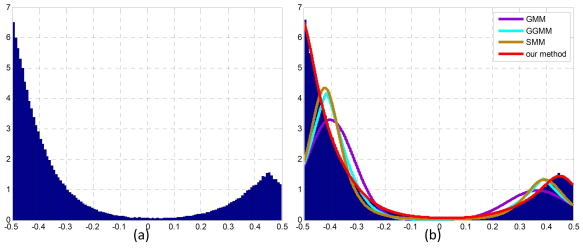

Fig. 1. The estimated histogram, (a): The histogram of the observed data, (b): The estimated histogram of GMM ($\mathcal{X}^2$=84854.08),GGMM ($\mathcal{X}^2$=30874.85), SMM ($\mathcal{X}^2$=21043.41), and our method ($\mathcal{X}^2$=10186.47).

In the second experiment , we shows a sample with 20,000 simulated points from four labels. Each label has 5,000 points. The ground truth distributions (two-dimensional view) of four labels are shown in Fig. 2(b). As shown in this figure, the intensity distribution of each label type is in the bounded support region, does not exhibit an exact Gaussian shape. Fig. 2(c)-(f), show the results of GMM, GGMM, SMM, and our method, respectively. In this experiment, the dimensions of the data are not independent. Therefore, the accuracy of GGMM in Fig. 2(d) is quite poor. GMM (Fig. 2(c)) and SMM (Fig. 2(e)) slightly improve the result. However, the error of the estimated distributions compared to the ground truth distributions in Fig. 3(b) remains quite high. The proposed method, as shown in Fig. 2(f), is more accurate compared to other methods.

Wavelet approximation coefficient is an important problem in computer vision as it plays a major role in a wide range of applications. In the next experiment, in order to give some implementation details about log-likelihood function, an image from Brodatz (www.ux.uis.no/ tranden/brodatz.html), as shown in Fig. 3(a), is used. In Fig. 3(b)-(e), we show the wavelet coefficients of the high-pass subband (CH). In Fig. 3(f) we plot the log-likelihood function versus the number of iteration. From that plot we see that our method reaches higher maximum values, which implies better performance.

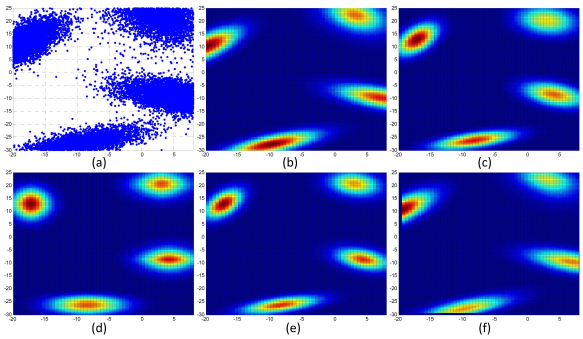

Fig. 2. The simulated point set experiment, (a): Original data, (b): Ground truth distribution, (c): GMM, (d): GGMM, (e): SMM, (f): our method.

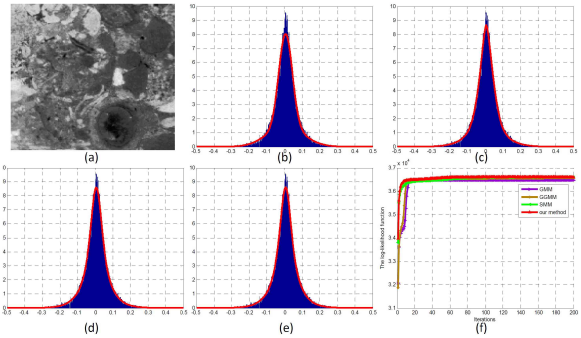

Fig. 3. Fig. 4. Approximation of the wavelet coefficients (K=2), (a): The original image from Brodatz datasets, (b): GMM ($\mathcal{X}^2$=53.82), (c): GGMM ($\mathcal{X}^2$=14.50), (d): SMM ($\mathcal{X}^2$=5.67), (e): BGGMM ($\mathcal{X}^2$=4.88), (f): Comparison of the log-likelihood function.

5. RELATION TO PRIOR WORK AND CONCLUSIONS

We have presented a bounded asymmetrical Student's-t mixture model to analyze both univariate and multivariate data. The proposed distribution, based on the Student's-t distribution, is heavily tailed and more robust than the Gaussian Mixture Model, and has the flexibility to fit different shapes of observed data such as non-Gaussian, non-symmetric, and bounded support data. Each component of the proposed model has the ability to model the observed data with different bounded support regions. In order to estimate model parameters, existing models represent the Student's-t distribution as an infinite mixture of scaled Gaussians. However, we propose an alternate approach in order to minimize the higher bound on the data negative log-likelihood function, and directly deal with the Student's-t distribution. Through extensive simulations, we demonstrate that the proposed model is superior to other clustering methods, based on the modeling of the probability density function of the data via finite mixture model.

6. REFERENCES

[1] McLachlan G., and Peel D., "*Finite Mixture Models*", New York, Wiley, 2000.

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

[3] Thanh M. N., and Wu Q. M. J., "Gaussian Mixture Model Based Spatial Neighborhood Relationships for Pixel Labeling Problem," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 42, no. 1, pp. 193–202, 2012.

[4] Dempster P., Laird N. M., and Rubin D. B., "Maximum likelihood from incomplete data via EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[5] McLachlan G., and Krishnan T., "*The EM Algorithm and Extensions*", Wiley, 1997.

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

[7] Peel D., and McLachlan G., "Robust Mixture Modeling Using the t Distribution," *Statistics and Computing*, vol. 10, no. 4, pp. 339–348, 2000.

[8] Thanh M. N., and Wu Q. M. J., "Robust Student's-t Mixture Model with Spatial Constraints and Its Application in Medical Image Segmentation," *IEEE Transactions on Medical Imaging*, vol. 31, no. 1, pp. 103–116, 2012.

[9] Allili M. S., Ziou D., Bouguila N., and Boutemedjet S., "Image and Video Segmentation by Combining Unsupervised Generalized Gaussian Mixture Modeling and Feature Selection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 10, pp. 1373–1377, 2010.

[10] Allili M., "Wavelet Modeling Using Finite Mixtures of Generalized Gaussian Distributions: Application to Texture Discrimination and Retrieval," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1452–1464, 2012.

[11] Liu C., and Rubin D., "ML estimation of the t distribution using EM and its extensions, ECM and ECME," *Statistica Sinica*, vol. 5, no. 1, pp. 19–39, 1995.

[12] Huber P. J., "Robust Statistics", Wiley, 1981, pp. 43–44.

[13] Hedelin P., and Skoglund J., "Vector Quantization Based on Gaussian Mixture Models," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 385–401, 2000.