SELF-PACED MIXTURE OF T DISTRIBUTION MODEL
Yang Zhang1,*, Qingtao Tang1,*, Li Niu1, Tao Dai1, Xi Xiao1, Shu-Tao Xia1
1 Department of Computer Science and Technology, Tsinghua University, China
2 Department of Electrical and Computer Engineering, Rice University, U.S.
*Equal Contribution
Email: yangzhan16@mails.tsinghua.edu.cn

Abstract
Gaussian mixture model (GMM) is a powerful probabilistic model for representing the probability distribution of observations in the population. However, the fitness of Gaussian mixture model can be significantly degraded when the data contain a certain amount of outliers. Although there are certain variants of GMM (e.g., multi-component of Laplace, mixture of t distribution) attempting to handle outliers, none of them can sufficiently mitigate the effect of outliers if the outliers are far from the centroids. Aiming to remove the effect of outliers further, this paper introduces a Self-Paced Learning mechanism into mixture of t distribution, which leads to Self-Paced Mixture of t Distribution (SPTMM). We derive an Expectation-Maximization based algorithm to train SPTMM and show SPTMM is able to screen the outliers. To demonstrate the effectiveness of SPTMM, we apply the model to density estimation and clustering. Finally, the results indicate that SPTMM outperforms other methods, especially on the data with outliers.

Main Contributions
- This is the first work of employing Self-Paced Learning (SPL) to mixture model, with the aim to effectively remove the influence of outliers.
- We propose our SPTMM method which integrates SPL with TMM, and develop an EM based algorithm to solve the corresponding optimization problem.
- Extensive experiments demonstrate the superiority of our SPTMM method for density estimation and clustering.

Related Works
Mixture of t Distribution
The t distribution is defined as follows. A p-dim random vector $x \in \mathbb{R}^p$ follows the p-variate t distribution with degrees of freedom $ν \in \mathbb{R}_+$, mean $μ \in \mathbb{R}^p$ and correlation matrix $Σ \in \mathbb{R}^{p \times p}$ if its joint probability density function (PDF) is given by

$$t(x | ν, μ, Σ) = \frac{Γ\left(\frac{ν + p}{2}\right)}{ν^{\frac{p+1}{2}} |Σ|^\frac{1}{2} \Gamma\left(\frac{ν}{2}\right)} \left[1 + \frac{1}{ν} (x - μ)^T Σ^{-1} (x - μ)\right]^{-(ν+p)/2}.$$

The mixture of t distribution model (TMM) is a linear superposition of g-component t distribution, i.e.,

$$p(x | Ψ, ν, μ, Σ) = \sum_{i=1}^g π_i t(x | ν, μ_i, Σ_i),$$

where $π_i$ is the mixing coefficient of the $i$-th component and $Ψ = \{ν, μ, Σ\}$, in which $ν = (ν_1, ν_2, ..., ν_g)$, $μ = (μ_1, μ_2, ..., μ_g)$, and $Σ = (Σ_1, Σ_2, ..., Σ_g)$.

Given the dataset $D = \{x_1, x_2, ..., x_n\}$, where $x_i \in \mathbb{R}^p$ denotes a p-dim sample, the model parameters of TMM can be estimated by minimum the negative log likelihood, i.e.,

$$\min_{Ψ} \sum_{i=1}^n log \left(\sum_{j=1}^g π_j t(x_i | ν_j, μ_j, Σ_j)\right),$$

which can be solved by EM algorithm.

Self-paced Learning
The objective function can be written as

$$\sum_{i=1}^n \frac{1}{f(v, λ)} f(v, λ),$$

where $v_i$ is learning weight of $x_i$, $f(\cdot)$ is negative log likelihood of $x_i$, $f(v, λ)$ is a regularizer, if

- $f(\cdot)$ is convex with respect to $v_i \in [0, 1]$;
- $\nu(λ, f)$ is monotonically decreasing with respect to $f$ and it holds that $log(λ, f) \rightarrow 0$ or $\nu(λ, f) \rightarrow 0$ if $\nu(λ, \cdot)$ is monotonically increasing with respect to $λ$, and it holds that $log(λ, f) \rightarrow 0$ or $\nu(λ, f) \rightarrow 0$ if $\nu(λ, f)$ $→ 0$ or $\nu(λ, f) \rightarrow 1$, where $\nu(λ, f) = arg min_{v < 1} (f(v, λ) + f(\cdot)).$

The Proposed Model
Objective Function
The objective function is given by

$$E(Ψ | v, λ) = \max_{ν, μ, Σ} \min_v \sum_{i=1}^n v_i (f_i - λ |v|);$$

where $Ψ = \{ν, μ, Σ\}; λ$ is a hyper-parameter, as a threshold when fixing $Ψ, v$ is the learning weight, also the outlier indicator with $v \in [0, 1]; |v|_{\|v\|}$, enforces the sparsity of $v$ since there exist only a few outliers in the training samples.

Optimization of $v$
Fixing $Ψ$, we estimate $v$ by solving

$$\min_{v \geq 0} E(Ψ | v, λ) = \min_{v \geq 0} \sum_{i=1}^n v_i (f_i - λ |v|);$$

Considering $v_i \in [0, 1]$, the problem can be written as

$$\min_{v \geq 0} \sum_{i=1}^n v_i (f_i - λ |v|);$$

It is obvious that the solution is

$$v_i = \left\{ \begin{array}{ll} 0, & t_i > λ, \\ 1, & t_i \leq λ. \end{array} \right.$$

Optimization of $Ψ$
Fixing $v$, we estimate $Ψ$ by solving

$$\min_{Ψ} E(Ψ | v, λ) \Rightarrow \min_{Ψ} \sum_{i=1}^n v_i (f_i - λ |v|) \Rightarrow \min_{Ψ} \sum_{i=1}^n f_i,$$

Use EM algorithm to optimize $Ψ$ iteratively.

Conclusions
In this paper, we depicted a novel model SPTMM which integrates the Self-Paced Learning mechanism into mixture of t distribution, in order to improve the mixture models’ ability of handling outliers. Given the model, we developed an EM based algorithm that can solve the optimization problem in SPTMM efficiently. In addition to the mathematical justification, the experiments also display the value of the model. The results demonstrated that SPTMM clearly outperforms K-means, GMM and TMM for estimating the covariance matrix in the distributions. With respect to clustering, SPTMM is shown to be the best performer in most cases, in particular for the data with outliers. In the future, we would like to assess if SPTMM can be improved to perform better in a clean environment.