SELF-PACED MIXTURE OF T DISTRIBUTION MODEL

Yang Zhang[†,*], Qingtao Tang[†,*], Li Niu [‡],Tao Dai[†], Xi Xiao[†], Shu-Tao Xia[†]

[†] Department of Computer Science and Technology, Tsinghua University, China

[‡] 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.*, mixture 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 DModel (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 $\boldsymbol{x} \in \mathbb{R}^p$ follows the p-variate t distribution with degrees of freedom $\nu \in \mathbb{R}_+$, mean $\boldsymbol{\mu} \in \mathbb{R}^n$, and correlation matrix $\Sigma \in \Pi(p)$ if its joint probability density function (PDF) is given by

$$t\left(\boldsymbol{x}|\nu, \boldsymbol{\mu}, \Sigma\right) = \frac{\Gamma[(\nu+p)/2]}{\Gamma(\nu/2)\nu^{p/2}\pi^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \cdot \left[1 + \frac{1}{\nu}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]^{-\frac{\nu+p}{2}}.$$

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

$$\phi(\boldsymbol{x};\boldsymbol{\Psi}) = \sum_{j=1}^{g} \pi_j t(\boldsymbol{x};\nu_j,\boldsymbol{\mu}_j,\Sigma_j),$$

where π_j is the mixing coefficient of the j-th component and $\boldsymbol{\Psi} = \{\boldsymbol{\pi},\boldsymbol{\nu},\boldsymbol{\mu},\Sigma\}$, in which $\boldsymbol{\pi} = (\pi_1,\pi_2,\ldots,\pi_g)^{\mathrm{T}}$, $\boldsymbol{\nu} = (\nu_1,\nu_2,\ldots,\nu_g)^{\mathrm{T}}$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\ldots,\boldsymbol{\mu}_g)$, and $\Sigma = (\Sigma_1,\Sigma_2,\ldots,\Sigma_g)$.

Given the dataset $\mathcal{D} = \{\boldsymbol{x}_i|i=1,2,\ldots,n\}$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ denotes a p-dim sample, the model parameters of TMM $\boldsymbol{\Psi}$ can be estimated by minimum the negative log likelihood, *i.e.*,

$$\min_{\boldsymbol{\Psi}} -\sum_{i=1}^{n}\log\sum_{j=1}^{g}\pi_j t(\boldsymbol{x}_i;\nu_j,\boldsymbol{\mu}_j,\Sigma_j), \tag{1}$$

which can be solved by EM algorithm.

Self-paced Learning

The objective function can be written as

$$\sum_{i=1}^{n} v_i\ell_i + f(\boldsymbol{v},\lambda)$$

where v_i is learning weight of $\boldsymbol{x}_i$, ℓ_i is negative log likelihood of $\boldsymbol{x}_i$, $f(\boldsymbol{v},\lambda)$ is a regularizer, if

- $f(\boldsymbol{v},\lambda)$ is convex with respect to $v_i \in [0,1]$;
- $v^\star(\lambda,\ell)$ is monotonically decreasing with respect to ℓ, and it holds that $\log_(\ell \rightarrow 0)v^\star(\lambda,\ell)=1$,$\log_(\ell \rightarrow \inf)v^\star(\lambda,\ell)=0$;
- $v^\star(\lambda,\ell)$ is monotonically increasing with respect to λ, and it holds that $\log_(\lambda \rightarrow 0)v^\star(\lambda,\ell)=0$,$\log_(\lambda \rightarrow \inf)v^\star(\lambda,\ell)=1$, where $v^\star(\lambda,\ell) = \arg\min_{v\in[0,1]} v\ell + f(\boldsymbol{v},\lambda)$.

The Proposed Model

Objective Function

The objective function is given by

$$E(\boldsymbol{v};\boldsymbol{\Psi},\lambda) = -\sum_{i=1}^{n} v_i \boxed{\log\sum_{j=1}^{g}\pi_j t(\boldsymbol{x}_i;\nu_j,\boldsymbol{\mu}_j,\Sigma_j)} - \lambda||\boldsymbol{v}||_1$$

where $\boldsymbol{\Psi} = \{\boldsymbol{\pi},\boldsymbol{\nu},\boldsymbol{\mu},\Sigma\}$; λ is a hyper-parameter, as a threshold when fixing $\boldsymbol{\Psi}$; v is the learning weight, also the outlier indicator with $v \in \{0,1\}$; $\lambda||\boldsymbol{v}||_1$ enforces the sparsity of v since there exist only a few outliers in the training samples.

Optimization of $\boldsymbol{v}$

Fixing $\boldsymbol{\Psi}$, we estimate $\boldsymbol{v}$ by solving

$$\min_{\boldsymbol{v}\in\{0,1\}} E(\boldsymbol{v};\boldsymbol{\Psi},\lambda) = \sum_{i=1}^{n} v_i\ell_i - \lambda||\boldsymbol{v}||_1$$

Considering $v_i \in \{0,1\}$, the problem can be written as

$$\min_{\boldsymbol{v}\in\{0,1\}} \sum_{i=1}^{n} v_i(\ell_i - \lambda)$$

It is obvious that the solution is

$$v_i = \begin{cases} 0 & \ell_i > \lambda, \\ 1 & \ell_i \le \lambda. \end{cases}$$

Optimization of $\boldsymbol{\Psi}$

Fixing $\boldsymbol{v}$, we estimate $\boldsymbol{\Psi}$ by solving

$$\min_{\boldsymbol{\Psi}} E(\boldsymbol{\Psi};\lambda,\boldsymbol{v}) \Leftrightarrow \min_{\boldsymbol{\Psi}} \sum_{i=1}^{n} v_i\ell_i - \lambda||\boldsymbol{v}||_1 \Leftrightarrow \min_{\boldsymbol{\Psi}} \sum_{i=1}^{n} v_i\ell_i$$

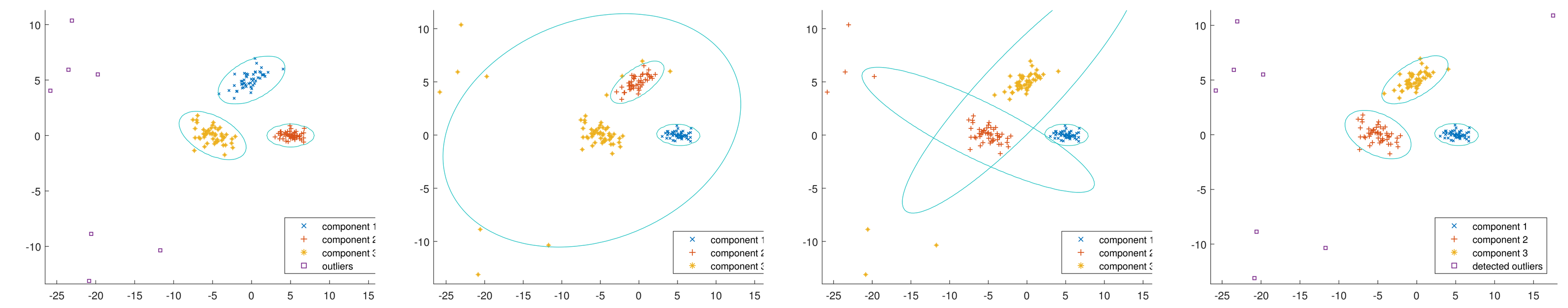
Use EM algorithm to optimize $\boldsymbol{\Psi}$ iteratively.

Summary of the algorithm

Input: dataset $\mathcal{D} = \{\boldsymbol{x}_i, i=1,2,\ldots,n\}$, learning pace $a > 1$, number of components g
Output: $\Psi_j (j=1,2,\ldots,g)$
Initialize Ψ by the result of k-means.
Initialize λ to the median of $\ell_i, i=1,2,\ldots,n$.
while $\triangle\Psi \simeq 0$ **do**
  while *not converged* **do**
    Update $\boldsymbol{v} = \arg\max_{\boldsymbol{v}\in\{0,1\}} E(\boldsymbol{v};\boldsymbol{\Psi},\lambda)$).
    while *not converged* **do**
      E-step
      Update $\hat{\tau}_{ij}$ and $\hat{u}_{ij}$, which are intermediate variables in TMM.
      M-step
      Update $\boldsymbol{\Psi}$.
    end
  end
  $\lambda \leftarrow a\lambda$.
end

Experiments

On the synthetic dataset



Synthetic dataset | GMM | TMM | SPTMM

On real dataset

	Clean				Noisy			
	MSE	W/B	DB	Dunn	MSE	W/B	DB	Dunn
k-means	2.80	**0.28**	**0.59**	2.35	3.55	0.69	0.72	2.06
GMM	2.87	**0.28**	0.60	**2.37**	4.67	1.16	1.81	0.63
TMM	2.93	0.29	0.61	**2.37**	4.66	0.81	1.41	0.89
SPTMM	**2.55**	0.29	0.61	**2.37**	**2.94**	**0.53**	**0.34**	**2.30**

Seeds dataset

	Clean				Noisy			
	MSE	W/B	DB	Dunn	MSE	W/B	DB	Dunn
k-means	3.44	2.23	0.69	2.66	3.64	2.50	0.72	2.61
GMM	3.54	2.48	0.59	3.34	3.75	2.79	1.14	1.35
TMM	3.62	2.74	0.57	3.33	3.80	3.02	0.58	3.46
SPTMM	**0.98**	**0.99**	**0.43**	**3.80**	**1.67**	**1.05**	**0.43**	**3.66**

Thyroid dataset

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.