Proceedings of ICSP2020

The Unν-Hardcut EM Algorithm for Non-central Student-t Mixtures of Gaussian Processes

Xiaoyan Li, Tao Li, Jinwen Ma
Department of Information and Computational Sciences, School of Mathematical Sciences and LMAM
Peking university, Beijing, China
Email:1801110047@pku.edu.cn, li_tao@pku.edu.cn, jwma@math.pku.edu.cn

2020 15th IEEE International Conference on Signal Processing (ICSP) | 978-1-7281-4480-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICSP48669.2020.9320906

***Abstract*—The mixture of Gaussian processes (MGP) is capable of learning any general stochastic process with a given set of samples for regression. However, there are two drawbacks on the learning of the MGP model. First, it is sensitive to outliers. Second, it is hard to learn the samples with heavy tails. The non-central student-t mixture of Gaussian processes (TMGP) is an effective regression model to deal with these problems. But this TMGP model has more parameters than the MGP model. In order to overcome this difficulty, we propose a new kind of Hardcut EM algorithm referred to as the unν-Hardcut EM algorithm for the parameter learning of TMGP. Specifically, the unν-Hardcut EM algorithm is based on the general framework of the Hard-cut EM algorithm while the expectation maximization for the parameters of input distribution of each component is implemented by maximizing the log-likelihood function of non-central student-t distribution on the input data set, with the degree ν of freedom unestimated but predicted in a special iterative procedure. It is demonstrated by the experimental results on synthetic data sets that the unν-Hardcut EM algorithm of the TMGP model is effective for regression. Moreover, this new appraoch obtains good prediction performance on a coal gas concentration data set.**

***Keywords*—Mixture of Gaussian Processes (MGP), student-t Mixture, EM algorithm, Regression, Parameter Learning.**

I. Introduction

Gaussian process (GP) is a powerful machine learning model for time series regression and classification [1]–[4]. However, it cannot fit the multimodal data well and also has large computational complexity $O(N^3)$ [5], where N is the number of training samples [6]. In order to overcome these limitations, Tresp [7] proposed the mixture of Gaussian Processes (MGP). However, the MGP model is sensitive to samples with potential outliers [8]. Besides that, for many applied problems, the tails of the assumed probability distributions in the MGP model are often shorter than required [9]. There are two reasons for the above problems. Firstly, Gaussian Process (GP) is not robust to outliers. Secondly, certain Gaussian distribution is selected as the gating function to combine these predictive Gaussian processes together. Generally, the Gaussian distribution has the aforementioned disadvantages. In this paper, we propose non-central student-t distribution as the gating distribution to combine these predictive Gaussian processes in the TMGP model to release the infulence of input outliers to regression.

Many methods can be utilized to solve these problems. As it is well known, student-t distribution provides a longer tail and robust with outliers [10]. In the mixture model, student-t mixture model (TMM) is longer tailed and more robust to potential outliers [11]–[13]. However, it cannot model time series or process data. In order to extend these characteristics to the mixture of Gaussian processes along the input region, we establish the non-central student-t mixture of Gaussian processes (TMGP) model. That is, we utlize a non-central student-t as the input distribution of each component Gaussian process. By extending the tail of the input distribution and enhancing the robustness of the input distribution to outliers, the same effect is achieved on the output distribution. So, each component has one more parameter than that of the Gaussian or normal mixture of Gaussian processes model. For parameter learning, we propose the unν-Hardcut EM algorithm of the TMGP model, which is based on the general framework of the Hard-cut EM algorithm. Moreover, the expectation maximization for the parameters of input distribution of each component is implemented by maximizing the log-likelihood function of non-central student-t distribution on the input data set, with the degree ν of freedom unestimated but predicted in a special iterative procedure. It is further demonstrated by the experimental result on synthetic data sets and a coal gas concentration data set that the TMGP model with the unν-Hardcut EM algorithm is less sensitive to outliers and more robust to the heavy tails of data distribution.

The rest of this paper is organized as follows. Section II describes the related models and learning methods. Section III presents the details of the TMGP model and the unν-Hardcut EM algorithm. The experimental results on synthetic data sets and real-world data sets are contained in Section IV and Section V, respectively. Finally,we make a breif conclusion in Section VI.

II. Related Models and Learning Methods

A. Gaussian Process

Gaussian Process is a typical stoachastic process in which any group of the states are subject to a Gaussian distribution. As a powerful machine learning model, it is easy to learn the parameters and make the inference [14]. We can define a Gaussian process of y with respect to an input x with a mean function $m(x)$ and a covariance matrix $C(x,x^{'})$ by

$$y = g(x) \sim GP(m(x), C(x,x^{'})). \quad (1)$$

Suppose that we have dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ in which x_i and y_i are a pair of input and output variables at sampling time i. As a statistical learning model, Gaussian process can be mathematically defined by

$$(y_1, y_2, \ldots, y_N) \sim GP(m(X), C(X, X^{'}) + \sigma^2 I), \quad (2)$$

where σ^2 dominates the noise golbally. For simplity, we generally set $m(X) = 0$. The covariance matrix with covariance function is $C(X, X^{'}) = [c(x_i, x_j)]_{N\times N}$. The most commonly used covariance function is the squared exponential function [15]–[17], which is defined by

$$c(x_i, x_j) = l^2 exp(-\frac{1}{2} f^2 \parallel x_i - x_j \parallel^2). \quad (3)$$

We can obtain the hyperparameter $\Theta = \{l, f, \sigma^2\}$ through the Maximum Likelihood Estimation (MLE) method. Actually, the predictive output of Gaussian process regression is given by

$$y_* | X, y, x_* \sim \mathcal{N}(\hat{y}_*, cov(f_*)), \quad (4)$$

where

$$\hat{y}_* = \mathbb{E}[y_* \mid X, y, x_*] = C(x_*, X)[C(X, X) + \sigma^2 I]^{-1} y, \quad (5)$$

$$cov(f_*) = C(x_*, x_*) - C(x_*, X)[C(X, X) + \sigma^2 I]^{-1}[C(x_*, X)]^{'}. \quad (6)$$

Here $y = [(y_1, y_2, \ldots, y_N)]^{'}$ is the output vector, $C(X, X) = [c(x_i, x_j)]_{N\times N}$ and $C(x_*, X) = [c(x_*, x_j)]_{1\times N}$ denotes the covariance relationship vector of the training inputs to the test input.

B. ML Estimation for Non-central Student-t Distribution

A p-dimensional random variable X is subject to a non-central t distribution $t_p(\mu, \Sigma, \nu)$ with a center μ, a covariance matrix Σ, and a degree $\nu \in (0, +\infty]$ of freedom, if given the weight τ, X has the multivariate normal distribution:

$$X \mid \mu, \Sigma, \tau \sim \mathcal{N}_p(\mu, \frac{\Sigma}{\tau}). \quad (7)$$

Furthermore, weight τ is subject to a $Gamma$ distribution [18], i.e.,

$$\tau \mid \mu, \Sigma, \nu \sim Gamma(\frac{\nu}{2}, \frac{\nu}{2}), \quad (8)$$

where the $Gamma(\alpha, \beta)$ density function is

$$f(\tau \mid \alpha, \beta) = \frac{\beta^\alpha \tau^{\alpha-1} \exp(-\beta\tau)}{\Gamma(\alpha)}, \tau > 0, \alpha > 0, \beta > 0.$$

By integrating τ from the joint density of (X, τ), we can get the density function of the marginal distribution of X, namely, $t_p(\mu, \Sigma, \nu)$,

$$\frac{\Gamma(\frac{\nu+p}{2}|\Sigma|^{-\frac{1}{2}})}{(\pi\nu)^{\frac{p}{2}}\Gamma(\frac{\nu}{2})}[1 + \frac{\delta_X(\mu, \Sigma)}{\nu}]^{-\frac{\nu+p}{2}}, \quad (9)$$

where $\delta_X = (X - \mu)^{'}\Sigma^{-1}(X - \mu)$ that is the Mahalanobis distance from X to the center μ concerning Σ. The density function (9) depends on X through $\delta_X(\mu, \Sigma)$. Thus, the distribution is ellipsoidal symmetric about μ.

We further derive the parameter learning function to $\{\mu, \Sigma, \nu\}$ through the ML estimation method. From the multivariance normal distribution (7), a p-dimensional random variable X with the given indicator τ being subject to $Gamma$ distribution is subject to a non-central student-t distribution. Thus, given $\{\mu, \Sigma, \tau, \nu\}$, the random variable $\tau\delta_X(\mu, \Sigma)$ is subject to χ_p^2 distribution, that is as $\Gamma(p/2, 1/2)$. On the other hand, from (8), the indicator τ is subject to a $Gamma$ distribution. So, taking X as samples are subject to (9), the conditional posterior distribution of τ, i.e., its distribution with be given $\{\mu, \Sigma, \nu, X\}$ is,

$$\begin{aligned} &\tau|\mu, \Sigma, \nu, X = \\ &\tau|\delta_X(\mu, \Sigma), \nu \sim \mathit{Gamma}(\frac{\nu + p}{2}, \frac{\nu + \delta_X(\mu, \Sigma)}{2}), \end{aligned} \quad (10)$$

whence

$$\mathbb{E}(\tau|\mu, \Sigma, \nu, X) = \frac{\nu + p}{\nu + \delta_X(\mu, \Sigma)}. \quad (11)$$

For the input $X = \{X_1, \ldots, X_N\}$ and the latent variable $\tau = \{\tau_1, \ldots, \tau_N\}$, we comprise the complete data $\{x_1, \ldots, x_N, \tau_1, \ldots, \tau_N\}$. Then the log-likelihood function of parameters μ, Σ and ν, ignoring constants, is

$$\mathcal{L}(\mu, \Sigma, \nu|X, \tau) = \mathcal{L}_N(\mu, \Sigma|X, \tau) + \mathcal{L}_G(\nu|\tau), \quad (12)$$

where

$$\begin{aligned} \mathcal{L}_N(\mu, \Sigma|X, \tau) = &-\frac{n}{2}\ln|\Sigma| - \frac{1}{2}\mathit{tr}(\Sigma^{-1})\sum_{i=1}^{N}\tau_i X_i X_i^{'} \\ &+ \mu^{'}\Sigma^{-1}\sum_{i=1}^{N}\tau_i X_i - \frac{1}{2}\mu^{'}\Sigma^{-1}\sum_{i=1}^{N}\tau_i, \end{aligned} \quad (13)$$

and

$$\mathcal{L}_G(\nu|\tau) = -n\ln(\Gamma(\frac{\nu}{2})) + \frac{n\nu}{2}\ln(\frac{\nu}{2}) + \frac{\nu}{2}\sum_{i=1}^{N}(\ln(\tau_i) - \tau_i). \quad (14)$$

Then the ML estimation of $\{\mu, \Sigma\}$ and the ML estimation of ν can be obtained from $\mathcal{L}_N(\mu, \Sigma|X, \tau)$ and $\mathcal{L}_G(\nu|\tau)$ respectively. Finally, we get the ML estimation of μ and Σ from $\mathcal{L}_N(\mu, \Sigma|X, \tau)$ are

$$\hat{\mu} = \frac{\sum_{i=1}^{N}\tau_i X_i}{\sum_{i=1}^{N}\tau_i}, \quad (15)$$

and

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{N}\tau i(X_i - \hat{\mu})(X_i - \hat{\mu})^{'}. \quad (16)$$

Therefore, the maximum likelihood estimation of the center μ, namely $\hat{\mu}$ is the weighted mean of the observations $\{X_1, \ldots, X_N\}$, the maximum likelihood estimation of the covariance matrix Σ, namely $\hat{\Sigma}$ is the average weighted sum of observations squares $\{X_1, \ldots, X_N\}$ about $\hat{\mu}$ with weights $\{\tau_1, \ldots, \tau_N\}$. Maximum Likelihood estimation of τ obtained by maximizing $\mathcal{L}_G(\nu|\tau)$ given by (19), that is, by solving

$$-\phi(\frac{\nu}{2}) + \ln(\frac{\nu}{2}) + \frac{1}{n}\sum_{i=1}^{N}(\ln(\tau_i) - \tau_i) + 1 = 0 \quad (17)$$

for ν, where $\phi(x) = d\ln(\Gamma(x))/dx$ is the digamma function. Equation (9) is discussed in the reference essay [18].

III. The TMGP Model and Its unν-Hardcut EM Algorithm

A. The TMGP model

A single GP cannot characterize a multimodal data set along the input regression because the structure of the GP model is rather simple. However, there are many multimodal data sets avaliable in practical applications. To tackle this problem, we extend the single GP model to th TMGP model in which different components are involved along the input region and each component is subject to a GP model independently. The gating network combines these predictive Gaussian processes together along the input region and we selecte the gating function as the non-central student-t mixture distribution. Let non-central student-t mixture distribution be the input distribution such that it can enhance the robustness of the model to outliers and extend the tail of the input distribution. For simplity, we still denote the data set of the TMGP model by $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, and describe the detail of the TMGP model as follows.

We describe the TMGP model mathematically as follows. We assume that there are K components involved along with the input region. Let $\{x_i\}_{i=1}^N$ be the set of p-dimensional inputs, $\{y_i\}_{i=1}^N$ be the set of outputs, and $\{z_i\}_{i=1}^N$ be indicators [19]. The set of indicators $\{z_i\}_{i=1}^N$ is subject to the multinomial distribution, which can be defined by

$$Pr(z_i = k) = \pi_k; k = 1, \ldots, K. \tag{18}$$

The input x_i is subject to a non-central student-t distribution, which can be defined by

$$x_i|(z_i = k) \sim t(\mu_k, \Sigma_k, \nu_k), \tag{19}$$

where the center is μ_k, the covariance matrix is Σ_k and the degree is $\nu_k \in (0, +\infty]$ of freedom. Finally, the predictive output of the K-th Gaussian process regression with certain covariance matrix by leaned hyperparameter vector $\Theta_k = \{l_k, f_k, \theta_k^2\}$,

$$y_i \sim \mathcal{GP}(0, C_k), \tag{20}$$

where C_k is the covariance matrix of k-th expert parameterized by θ_k.

B. The Unν-Hardcut EM Algorithm

We further propose the unν-Hardcut EM algorithm to learn parameters in the TMGP model. In each component of the mixture model, there are two parameter vectors $\alpha_k = \{\mu_k, \Sigma_k, \nu_k\}$ and $\theta_k = \{l_k, f_k, \sigma_k^2\}$ (fig.1). We use the EM algorithm to get the parameter vector α_k and MLE to get the parameter vector θ_k.

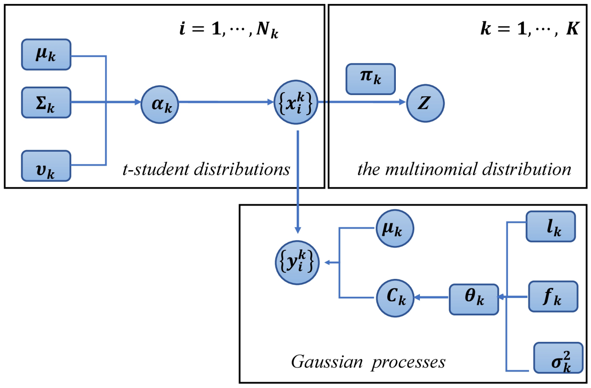

Fig. 1. The flowchart of data generation by the non-central student-t mixture of Gaussian processes (TMGP) model. An input data x_i^k of the k-th component is subject to a non-central student-t distribution with a parameter vector$\alpha_k = \{\mu_k, \Sigma_k, \nu_k\}$. The predictive output y_i^k of the k-th component is subject to a Gaussian process. Suppose that μ_k is the mean function and C_k the covariance matrix with hyper-parameters $\theta_k = \{l_k, f_k, \sigma_k^2\}$. The indicator Z is generated by the multinomial distribution with π_k.

We have discussed in Section II about the ML estimation of the non-central student-t distribution, but it is not easy to get the parameter vector $\alpha_k = \{\mu_k, \Sigma_k, \nu_k\}$ with the unknown variable τ_i. Lange, Little and Taylor (1989) [20] suggested how to use the EM algorithm to get parameters μ, Σ and ν of t distribution. This method can extend to the TMGP model directly. Let τ_k be a latent variable in the k-th component. We obtain its expectation according to the formula (11)

$$\omega_{i,k}^{(t+1)} = \mathbb{E}(\tau_k|X, \alpha_k^{(t)}) = \frac{\nu_k^{(t)} + p}{\nu_k^{(t)} + \delta_{i,X}^{(t)}(\mu_{i,k}^{(t)}, \Sigma_{i,k}^{(t)})}. \tag{21}$$

Then we maximize the likelihood function (13) to obtaine $\mu_k^{(t+1)}$ and $\Sigma_k^{(t+1)}$

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^N \omega_{i,k}^{(t+1)} X_i}{\sum_{i=1}^N \omega_{i,k}^{(t+1)}}, \tag{22}$$

and

$$\Sigma_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^N \omega_{i,k}^{(t+1)} (X_i - \mu_k^{(t+1)})(X_i - \mu_k^{(t+1)})'. \tag{23}$$

According to (14) and (17), we update ν_k by solving the equation below for ν

$$\begin{aligned} -\phi(\frac{\nu}{2}) + \ln(\frac{\nu}{2}) + \frac{1}{N}\sum_{i+1}^N [\ln(\omega_{i,k}^{(t+1)}) - \omega_{i,k}^{(t+1)}] \\ + 1 + [\phi(\frac{p + \nu^{(t)}}{2}) - \ln(\frac{p + \nu^{(t)}}{2}] = 0. \end{aligned} \tag{24}$$

The indicator Z in the TMGP model update via Hard-cut allocation. Based on the TMGP model, we obtain the probability

$$p(Z_t = k, X_t, y_t) = \pi_k \cdot t(X_k|\mu_k, \Sigma_k, \nu_k) \cdot \mathcal{GP}(y_t|0, l_k^2 + \sigma_k^2). \tag{25}$$

We choose the most likely category as the input label.

Based on the above analysis, we can design the unν-Hardcut EM algorithm as shown in Algorithm 1.

Algorithm 1 The Unν-Hardcut EM algorithm for learning the parameters of the TMGP model.

```
Input: the set of data is $\mathcal{D} = \{X_i, y_i\}_{i=1}^{N}$, the number of experts $K$;
Indicator: the set of indicators is $\{Z_t\}_{t=1}^{N}$;
Output: Mixing proportions $\{\pi_k\}_{k=1}^{K}$, parameters of the input discussion $\alpha_k = \{\mu_k, \Sigma_k, \nu_k\}_{k=1}^{K}$, parameters of the Gaussian procession, $\theta_k = \{l_k, f_k, \sigma_k^2\}_{k=1}^{K}$;
Initialization: Initialize $\{Z_t\}_{t=1}^{N}$ via $k$-means.
Iteration:
while not converges do
    for $k = 1, \ldots, K$ do
        Update mixture parameters of $k$-th component
            $\pi_k = p(Z_t = k|X_t, y_t) = \frac{\pi_k \cdot t(X_k|\mu_k, \Sigma_k, \nu_k) \cdot \mathcal{GP}(y_t|0, l_k^2 + \sigma_k^2)}{\sum_{k=1}^{K} \pi_k \cdot t(X_k|\mu_k, \Sigma_k, \nu_k) \cdot \mathcal{GP}(y_t|0, l_k^2 + \sigma_k^2)}$
        (Update $\mu_k^{(t)}$, $\Sigma_k^{(t)}$ and $\nu_k^{(t)}$ via MLE estimation)
        while not converges do
            (E-Step) Calculate $\omega_{i,k}^{(t+1)}$ for $i = 1, \ldots, n$ in (21),
            (M-Step) Calculate $\mu_k^{(t+1)}$ in (22) and $\Sigma_k^{(t+1)}$ in (23),
            Update $\nu_k$ by solving (24).
        end while
        Obtain the GP parameters $\theta_k$ by maximizing the likelihood function
            $p(y_{k,1}, \ldots, y_{k,N_k}) \mid X_{k,1}, \ldots, X_{k,N_k} = GP(0, K(X_{k,i}, X_{k,j} \mid \theta_k) + \sigma_k^2 I)$'
        Update $Z_i$ via hard-cut allocation,
            $Z_i = \arg \max_{k=1,\ldots,K} \pi_k t(x_i, \alpha_k) p(y_i \mid X_i, \theta_k)$
            .
    end for
end while
```

IV. Experimental Results

In this section, we conduct experiments on synthetic datasets which are composed of three and five Gaussian processes named $\mathcal{S}_3$ and $\mathcal{S}_5$. For each Gaussian process, we generate 200 training samples as well as 500 test samples. The training samples and test samples are selected from non-central student-t distributions randomly. In this experiment, we have $\theta_k^1 = \ln l_k^2$, $\theta_k^2 = \ln f_k^2$ and $\theta_k^3 = \ln \sigma_k^2$. We set $\theta^1 = \{0.3, 0.1, 0.8, 0.4, 0.2\}$, $\theta^2 = \{2, 0.75, 1.5, 0.5, 1.5\}$ and $\theta^3 = \{0.1, 0.1, 0.1, 0.1, 0.1\}$ to generate the samples. The implementation is based on the GPML toolbox [1] and all experiments are conducted on a personal computer (Inter(R) Core (TM) i7-8550U CPU 1.8GHz, 8G RAM).

A. Experimental Results on Synthetic Datasets

In this subsection, we give the results of the TMGP model using the unν-Hardcut EM algorithm. Firstly, we visualize the results of the algorithm on $\mathcal{S}_3$ and $\mathcal{S}_5$ in Fig. 2.

It can be seen from Fig.2 that there are the results of different components in the TMGP model. The left column of Fig.2 shows the data set with three Gaussian processes of TMGP (up) and five Gaussian processes of TMGP (down). The right column of Fig.2 shows the model result using the unν-Hardcut EM algorithm correspondingly. In these synthetic datasets, each component data have long tails so that the mixing proportion is higher. The reason is that its input distribution has heavy tails. In this case, it is clear that the model is suitable to make significant predictions despite the numbers of GPs of the TMGP model.

The TMGP model can handle the multimodal data with heavy tails and outliers because the student-t distribution can handle outliers effectively. But it is not so easy to learn the degree ν of freedom. We need to solve the non-linear equation (24) for updating the parameter, which increases the computations of the algorithm.

B. Comparison on Synthetic Datasets

To further investigate the TMGP model, we compare it with some typical models as well as their corresponding learning algorithms on synthetic data sets. In fact, these compararive models can be summarized as follows.

- LR(SVM): Linear Regression using support vector regression with Gaussian kernel;
- GP(GL): a single Gaussian process using Gaussian likelihood function;
- GP(TL): a single Gaussian process using central t distribution likelihood function;
- MGP (Hard-cut EM): Mixtures of Gaussian processes using the Hard-cut EM algorithm;
- MGP (LOOCV): Mixtures of Gaussian processes using the LOOCV algorithm;
- TMGP (Unν-Hardcut EM): Non-central t distribution mixtures of Gaussian processes using the unν-Hardcut EM algorithm.

We contain all the experimental results of the proposed and comarative models in Table I, where the Rooted Mean Square Error (RMSE) is used to measure the fitting performance. In experiment, we utilize the SVM toolbox in MATLAB. Moreover, we adopt the GPML toolbox for learning the GP model.

In Table I, the average predicted RMSEs of the TMGP model are always the smallest. In fact, the linear regression is not suitable to process data no matter if the input distributions have the same degree ν of freedom. A single GP is rather simple to characterize multimodal data sets. But its performance is better than the linear regression model. The TMGP model performs better than MGP for its heavy tails and robust to outliers. The table illustrates that the regression results of the TMGP model and the MGP model are close.

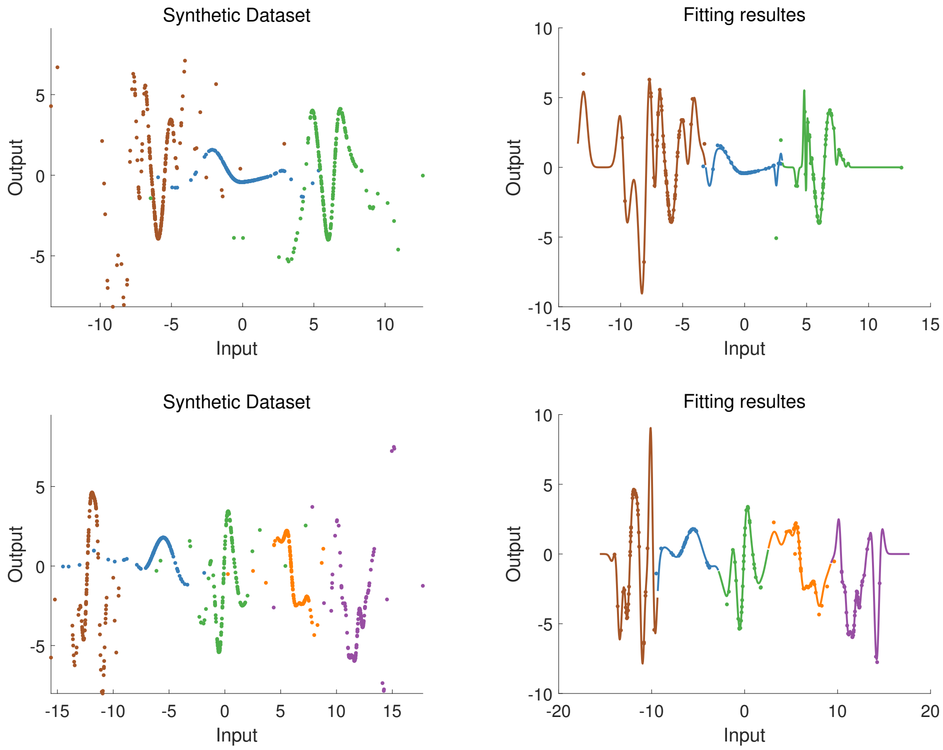

Fig. 2. The sketches of the synthetic data set with three and five Gaussian processes of the TMGP model (left column), and the fitting curves of the TMGP model using unν-Hardcut EM Algorithm(right column).

TABLE I

THE RESULTS OF THE PROPOSED AND COMPARATIVE ALGORITHMS.

	Method	$\nu = 3 * ones(K, 1)$	$\nu = [4, 6, 3]$
K=3	LR(SVM)	0.4855	0.5579
	GP(GL)	0.4783	0.3642
	GP(TL)	0.459	0.3482
	MGP(Hard-cut)	0.4179	0.3027
	MGP(LOOCV)	0.4282	0.3302
	TMGP(Unν-Hardcut)	**0.3387**	**0.2876**
	Method	$\nu = 3 * ones(K, 1)$	$\nu = [4, 6, 3, 5, 2]$
K=5	LR(SVM)	0.7387	2.3042
	GP(GL)	0.6565	1.1196
	GP(TL)	0.5363	1.0985
	MGP(Hard-cut)	0.4829	1.3533
	MGP(LOOCV)	0.5108	1.1301
	TMGP(Unν-Hardcut)	**0.441**	**1.0754**

V. Application to the Modeling of Coal Gas Concentration Data

In this section, we apply the TMGP model with the unν-Hardcut EM algorithm to the modeling of the gas data set which is recorded by coal mine detections in 2018. Actually, this real-world dataset consists of the observations of gas consentration per five seconds in a specific coal mine face. We firstly calculate the means of gas concentration data every day as our experiment samples. In this case, we use the TMGP model with 1-4 components respectively.

In Fig.3, we set $K = 1$, $K = 2$, $K = 3$ and $K = 4$ to illustrate the results of the TMGP model. It can be seen from Fig.3 that the TMGP model with four components is better than the others. It should be noted that different colors are used to distuiguish different classification results of the components. Fig.3(4) shows the change of coal gas consentration with season.

VI. Conclusion

We have established the non-central student-t mixture of Gaussian processes (TMGP) model with the proposed unν-Hardcut EM algorithm for learning the time series data with potential outliers and distribution heavy tails. It is demonstrated by the experimental results on synthetic and coal gas concentration datasets that this TMGP approach is less sensitive to outliers and more robust to the heavy tails of data distribution. However, according to our experiments, there are still two problems. Firstly, it is still difficult to determine the number of Gaussian processes in the mixture. Secondly, we need to solve a non-linear equation to get the parameter ν_k and the solving process is time-consuming. As a result,the unν-Hardcut EM algorithm converges slowly. Therefore, in the future, we will improve the TMGP approach by investigating these two problems.

Acknowledgment

This work was supported by the National Key R & D Program of China (2018YFC0808305).

References

[1] C. E. Rasmussen and H. Nickisch, *Gaussian Processes for Machine Learning (GPML) Toolbox*. JMLR.org, 2010.

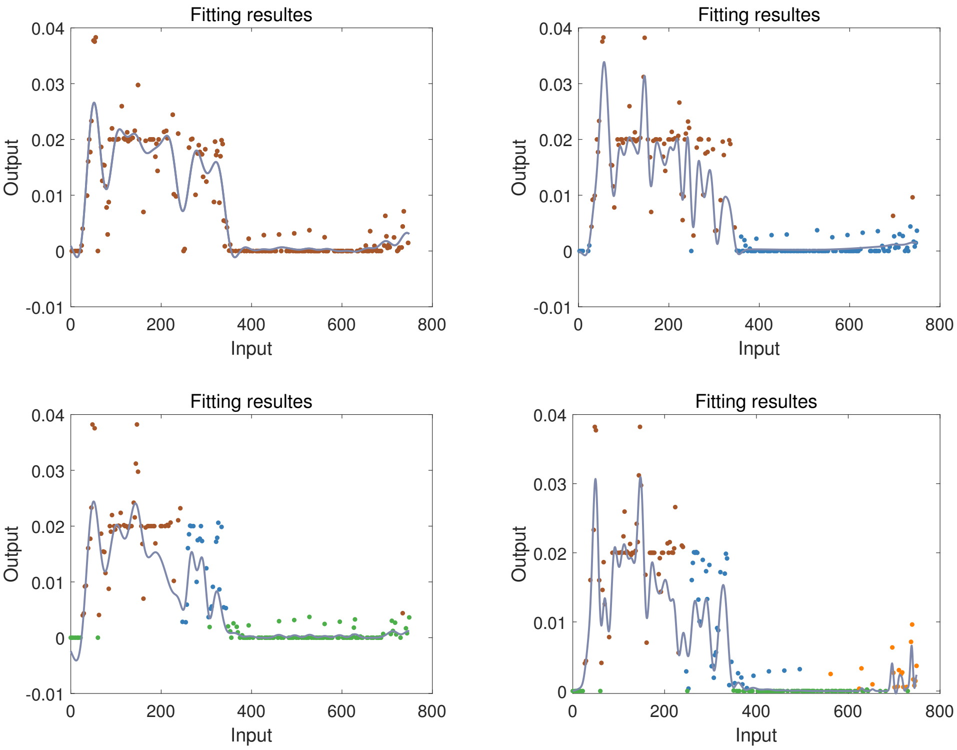

Fig. 3. The results of the TMGP model with different numbers of components on the coal gas concentration dataset.

[2] R. C. Grande, T. J. Walsh, G. Chowdhary, S. Ferguson, and J. P. How, "Online regression for data with changepoints using gaussian processes and reusable models," *IEEE Transactions on Neural Networks & Learning Systems*, vol. 28, no. 9, pp. 2115–2128, 2017.

[3] M. Lazaro-Gredilla and S. Van Vaerenbergh, "A gaussian process model for data association and a semidefinite programming solution," *IEEE Transactions on Neural Networks & Learning Systems*, vol. 25, no. 11, pp. 1967–1979, 2014.

[4] D. Wu and J. Ma, "A two-layer mixture model of gaussian process functional regressions and its mcmc em algorithm," *IEEE Transactions on Neural Networks & Learning Systems*, pp. 1–11, 2018.

[5] Y. Chao and C. Neubauer, "Variational mixture of gaussian process experts," in *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008*, 2008.

[6] L. Zhao and J. Ma, "A dynamic model selection algorithm for mixtures of gaussian processes," in *2016 IEEE 13th International Conference on Signal Processing (ICSP)*, 2016, pp. 1095–1099.

[7] V. Tresp, "Mixtures of gaussian processes," *Advances in Neural Information Processing Systems*, vol. 13, pp. 654–660, 2001.

[8] X. Pan, H. Zhu, and Q. Xie, "A robust nonsymmetric student's-t finite mixture model for mr image segmentation," in *2015 IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 1830–1834.

[9] L. KL, R. Little, and J. Taylor, "Robust statistical modeling using the t distribution," *Journal of the American Statistical Association*, vol. 84, 01 1989.

[10] R. E. Schafer, "Interval estimation of product reliability by use of the noncentral t distribution," *IRE Transactions on Reliability and Quality Control*, vol. RQC-9, no. 1, pp. 77–81, 1960.

[11] J. Lai and H. Zhu, "A fusion algorithm: Fully convolutional networks and student's-t mixture model for brain magnetic resonance imaging segmentation," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 1598–1602.

[12] T. M. Nguyen and Q. M. J. Wu, "Bounded asymmetrical student's-t mixture model," *IEEE Transactions on Cybernetics*, vol. 44, no. 6, pp. 857–869, 2014.

[13] Y. Zhou, H. Zhu, and X. Tao, "Robust mr image segmentation using the trimmed likelihood estimator in asymmetric student's-t mixture model," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2017, pp. 644–647.

[14] Y. Lei and H. Yang, "A gaussian process ensemble modeling method based on boosting algorithm," in *Proceedings of the 32nd Chinese Control Conference*, 2013, pp. 1704–1707.

[15] A. Solin and S. Särkkä, "Gaussian quadratures for state space approximation of scale mixtures of squared exponential covariance functions," in *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2014, pp. 1–6.

[16] S. Särkkä and R. Piché, "On convergence and accuracy of state-space approximations of squared exponential covariance functions," in *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2014, pp. 1–6.

[17] A. Abusnina, D. Kudenko, and R. Roth, "Selection of covariance functions in gaussian process-based soft sensors," in *2014 IEEE International Conference on Industrial Technology (ICIT)*, 2014, pp. 371–378.

[18] C. Liu and D. B. Rubin, "Ml estimation of the t distribution using em and its extensions, ecm amd ecme," *Statistica Sinica*, vol. 5, no. 1, pp. 19–39, 1995.

[19] Z. Chen, J. Ma, and Y. Zhou, "A precise hard-cut em algorithm for mixtures of gaussian processes," 2014.

[20] K. L. Lange, R. J. A. Little, and J. M. G. Taylor, "Robust statistical modeling using the t distribution," *Publications of the American Statistical Association*, vol. 84, no. 408, pp. 881–896, 1989.