# The Un -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

***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 un -Hardcut EM algorithm for the parameter learning of TMGP. Specifically, the un -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-likSpecifica52utimpone8oo9717(by)-414.func7.99001(maximizi01231(by)-414.ral)BT9.963 TL1 0 0 1 48.964 46**

t7eoce.00).5450of

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

$$(y_1; y_2; \cdots; y_N) \sim GP(m(X); C(X; X') + \quad ^2 I); \quad (2)$$

where $^2$ 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 $= \{l; f; \ ^2\}$ 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

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

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

Here $y = [(y_1; y_2; \cdots; y_N)]'$ 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 $t_p( \ ; \ ; \ )$ with a center , a covariance matrix , and a degree $\in (0; +\infty]$ of freedom, if given the weight , $X$ has the multivariate normal distribution:

$$X \mid \ ; \ ; \ \sim \mathcal{N}_p( \ ; -): \quad (7)$$

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

$$\mid \ ; \ ; \ \sim Gamma(\frac{}{2}; \frac{}{2}); \quad (8)$$

where the $Gamma( \ ; \ )$ density function is

$$f( \ \mid \ ; \ ) = \frac{^{\alpha} \ ^{\alpha-1} \exp(-\ )}{( \ )}; \quad > 0; \quad > 0; \quad > 0:$$

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

$$\frac{(\frac{\nu+p}{2} \mid \ \mid^{-\frac{1}{2}})}{( \ )^{\frac{p}{2}} \ (\frac{\nu}{2})}[1 + \frac{_X( \ ; \ )}{}]^{-\frac{+p}{2}}; \quad (9)$$

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

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

$$\begin{aligned} &\mid \ ; \ ; \ ; X = \\ &\mid \ _X( \ ; \ ); \ \sim Gamma(\frac{+p}{2}; \frac{+ \ _X( \ ; \ )}{2}); \end{aligned} \quad (10)$$

whence

$$\mathbb{E}( \ \mid \ ; \ ; \ ; X) = \frac{+p}{+ \ _X( \ ; \ )}: \quad (11)$$

For the input $X = \{X_1; \cdots; X_N\}$ and the latent variable $= \{ \ _1; \cdots; \ _N\}$, we comprise the complete data $\{x_1; \cdots; x_N; \ _1; \cdots; \ _N\}$. Then the log-likelihood function of parameters , and , ignoring constants, is

$$\mathcal{L}( \ ; \ ; \ | X; \ ) = \mathcal{L}_N( \ ; \ | X; \ ) + \mathcal{L}_G( \ \mid \ ); \quad (12)$$

where

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

and

$$\mathcal{L}_G( \ \mid \ ) = -n \ln( \ (\frac{}{2})) + \frac{n}{2} \ln(\frac{}{2}) + \frac{}{2} \sum_{i=1}^{N} (\ln( \ _i) - \ _i): \quad (14)$$

Then the ML estimation of $\{ \ ; \ \}$ and the ML estimation of can be obtained from $\mathcal{L}_N( \ ; \ | X; \ )$ and $\mathcal{L}_G( \ \mid \ )$ respectively. Finally, we get the ML estimation of and from $\mathcal{L}_N( \ ; \ | X; \ )$ are

$$\hat{} = \frac{\sum_{i=1}^{N} \ _i X_i}{\sum_{i=1}^{N} \ _i}; \quad (15)$$

and

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

Therefore, the maximum likelihood estimation of the center , namely $\hat{}$ is the weighted mean of the observations $\{X_1; \cdots; X_N\}$, the maximum likelihood estimation of the covariance matrix , namely $\hat{}$ is the average weighted sum of observations squares $\{X_1; \cdots; X_N\}$ about $\hat{}$ with weights $\{ \ _1; \cdots; \ _N\}$. Maximum Likelihood estimation of obtained by maximizing $\mathcal{L}_G( \ \mid \ )$ given by (19), that is, by solving

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

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

## III. The TMGP Model and Its un -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

$$Pr(z_i = k) = \ _k; k = 1; : : : ; K: \tag{18}$$

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

$$x_i|(z_i = k) \sim t(\ _k;\ \ _k;\ \ _k); \tag{19}$$

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

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

where $C_k$ is the covariance matrix of $k$-th expert parameterized by $\ _k$.

### B. The Un -Hardcut EM Algorithm

We further propose the un -Hardcut EM algorithm to learn parameters in the TMGP model. In each component of the mixture model, there are two parameter vectors $\ _k = \{\ _k;\ \ _k;\ \ _k\}$ and $\ _k = \{l_k; f_k;\ \ ^2_k\}$ (fig.1). We use the EM algorithm to get the parameter vector $\ _k$ and MLE to get the parameter vector $\ _k$.

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 $\ _k = \{\ _k;\ \ _k;\ \ _k\}$ with the unknown variable $\ _i$. Lange, Little and Taylor (1989) [20] suggested how to use the EM algorithm to get parameters , and

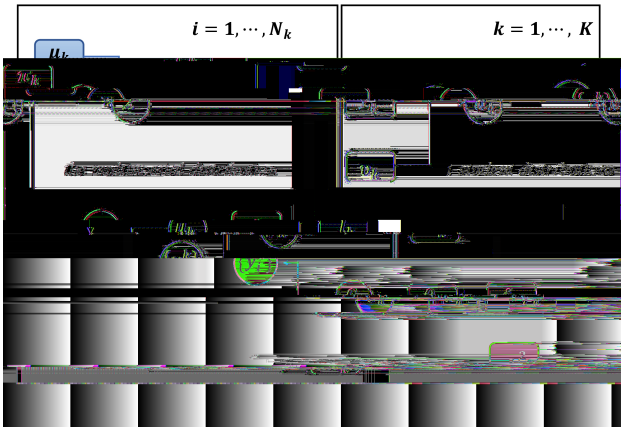


Fig. 1. The flowchart of data generation by the non-central student-$t$ mixture of Gaussian processes (TMGP) model. An input data $x_i^k$ of the $k$-th component is subject to a non-central student-$t$ distribution with a parameter vector$\alpha_k = \{\mu_k,\ \ _k, \nu_k\}$. The predictive output $y_i^k$ of the $k$-th component is subject to a Gaussian process. Suppose that $\mu_k$ 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 $\pi_k$.

of $t$ distribution. This method can extend to the TMGP model directly. Let $\ _k$ be a latent variable in the $k$-th component. We obtain its expectation according to the formula (11)

$$!^{(t+1)}_{i,k} = \mathbb{E}(\ _k|X;\ \ ^{(t)}_k) = \frac{\ ^{(t)}_k + p}{\ ^{(t)}_k + \ ^{(t)}_{i,X}(\ ^{(t)}_{i,k};\ \ ^{(t)}_{i,k})}: \tag{21}$$

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

$$\ ^{(t+1)}_k = \frac{\sum_{i=1}^N !^{(t+1)}_{i,k} X_i}{\sum_{i=1}^N !^{(t+1)}_{i,k}}; \tag{22}$$

and

$$\ ^{(t+1)}_k = \frac{1}{n}\sum_{i=1}^N !^{(t+1)}_{i,k}(X_i - \ ^{(t+1)}_k)(X_i - \ ^{(t+1)}_k)': \tag{23}$$

According to (14) and (17), we update $\ _k$ by solving the equation below for

$$\begin{aligned} -\ (\tfrac{\ }{2}) + \ln(\tfrac{\ }{2}) + \frac{1}{N}\sum_{i+1}^N[\ln(!^{(t+1)}_{i,k}) - !^{(t+1)}_{i,k}] \\ + 1 + [\ (\frac{p + \ ^{(t)}}{2}) - \ln(\frac{p + \ ^{(t)}}{2}] = 0: \end{aligned} \tag{24}$$

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

$$p(Z_t = k; X_t; y_t) = \ _k \cdot t(X_k|\ \ _k;\ \ _k;\ \ _k) \cdot \mathcal{GP}(y_t|0; l_k^2 + \ ^2_k): \tag{25}$$

We choose the most likely category as the input label.

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

**Algorithm 1** The Un -Hardcut EM algorithm for learning the paraneters 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 $\{ _k\}_{k=1}^K$, parameters of the input discussion $_k = \{ _k; _k; _k\}_{k=1}^K$, parameters of the Gaussian procession, $_k = \{l_k; f_k; ^2_k\}_{k=1}^K$;
**Initialization:** Initialize $\{Z_t\}_{t=1}^N$ via $k$-means.
**Iteration:**

**while** not converges **do**
  **for** $k = 1; : : : ; K$ **do**
    Update mixture parameters of $k$-th component

$$_k = p(Z_t = k|X_t; y_t) = \frac{_k \cdot t(X_k| _k; _k; _k) \cdot \mathcal{GP}(y_t|0; l_k^2 + ^2_k)}{\sum_{k=1}^K _k \cdot t(X_k| _k; _k; _k) \cdot \mathcal{GP}(y_t|0; l_k^2 + ^2_k)}$$

    (Update $^{(t)}_k$, $^{(t)}_k$ and $^{(t)}_k$ via MLE estimation)
    **while** not converges **do**
      (E-Step) Calculate $!^{(t+1)}_{i,k}$ for $i = 1; : : : ; n$ in (21),
      (M-Step) Calculate $^{(t+1)}_k$ in (22) and $^{(t+1)}_k$ in (23),
      Update $_k$ by solving (24).
    **end while**
    Obtain the GP parameters $_k$ by maximizing the likelihood function

$$p(y_{k,1}; : : : ; y_{k,N_k}) \mid X_{k,1}; : : : ; X_{k,N_k} = GP(0; K(X_{k,i}; X_{k,j} \mid _k) + ^2_k I)',$$

    Update $Z_i$ via hard-cut allocation,

$$Z_i = \arg \max_{k=1,\ldots,K} _k t(x_i;$$

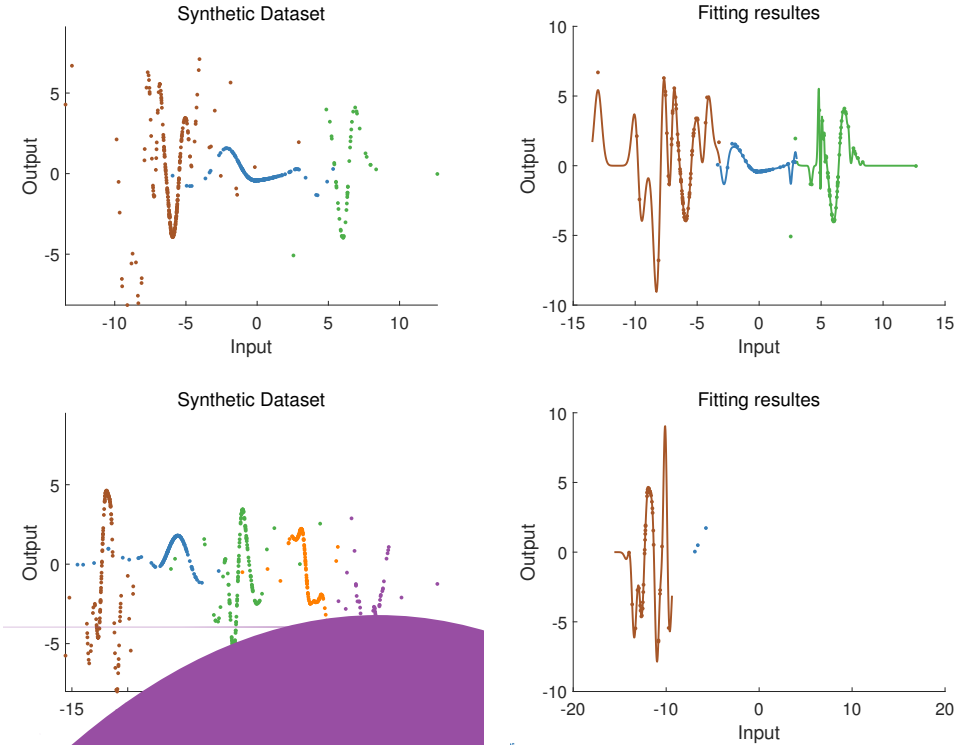


Fig. 2. The sketches of the synthetic data set with three and five Gaussian processes of the TMGP model (left column), and the fitting curves of the TMGP model using un$\nu$-Hardcut EM Algorithm(right column).

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(Un$\nu$-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(Un$\nu$-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 un -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 un -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 $_k$ and the solving process is time-consuming. As a result,the un -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.

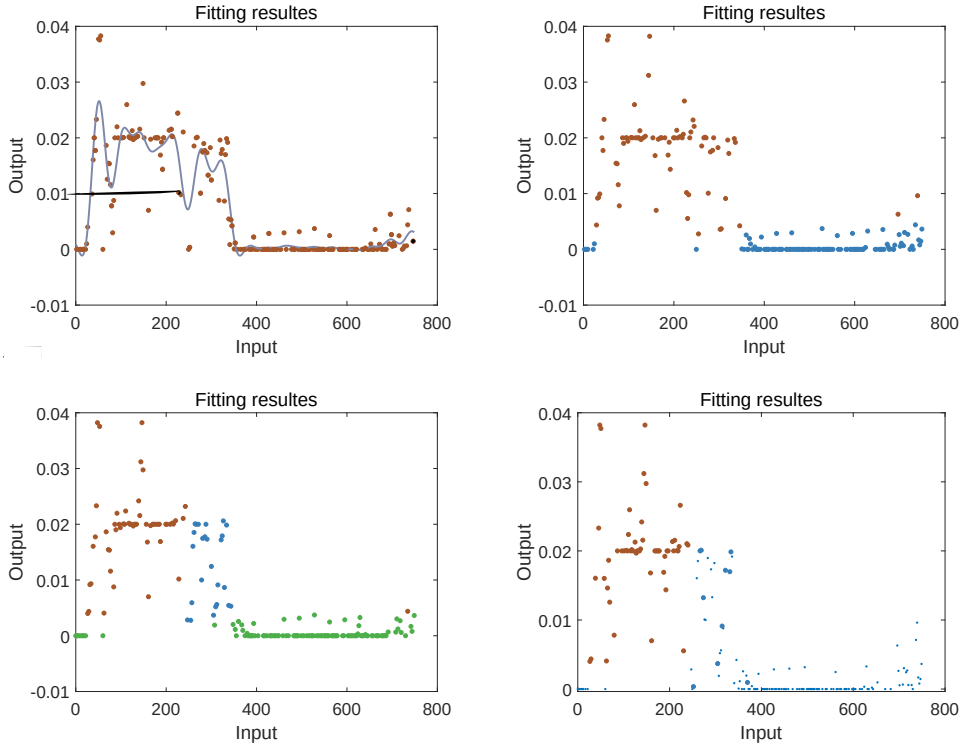


Fig. 3. The results of the TMGP model with different numbers of components on the coal gas concentration dataset.


[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.