An Effective Model Selection Criterion for Mixtures of Gaussian Processes

Longbo Zhao, Ziyi Chen, and Jinwen Ma*

Department of Information Science, School of Mathematical Sciences and LMAM,
Peking University, Beijing, 100871, China
jwma@math.pku.edu.cn

Abstract. The Mixture of Gaussian Processes (MGP) is a powerful statistical learning framework in machine learning. For the learning of MGP on a given dataset, it is necessary to solve the model selection problem, i.e., to determine the number C of actual GP components in the mixture. However, the current learning algorithms for MGPs cannot solve this problem effectively. In this paper, we propose an effective model selection criterion, called the Synchronously Balancing or SB criterion for MGPs. It is demonstrated by the experimental results that this SB criterion is feasible and even outperforms two classical criterions: AIC and BIC, for model selection on MGPs. Moreover, it is found that there exists a feasible interval of the penalty coefficient for correct model selection.

Keywords: Mixture of Gaussian processes, Model selection, EM algorithm, Parameter learning, Likelihood.

1 Introduction

The Gaussian Process (GP) model is a powerful tool for machine learning. However, it has two limitations. Firstly, it can only fit a single modality dataset. Secondly, for the GP model, the learning algorithm has a large computational complexity $O(N^3)$[1], where N is the number of training samples. In order to solve these issues, Tresp [2] proposed the mixture of Gaussian processes (MGP) in 2000. From then on, various MGP models have been proposed and can be classified into two main forms: the conditional models [2-5] and the generative models [1, 6]. Here, we adopt the generative model since it can infer missing inputs from outputs [7]. In fact, with different number of GP components, the MGP model may lead to quite different experimental results for regression and classification. So, it is critical to know the true number of GP components in the mixture or dataset and thus to get the reasonable result. That is, we must determine the number C of GP components in the mixture for the parameter learning, which is referred to as the model selection problem for the learning of the mixture.

* Corresponding author.

X. Hu et al. (Eds.): ISNN 2015, LNCS 9377, pp. 345–354, 2015.
DOI: 10.1007/978-3-319-25393-0_38

For model selection, there are some classical criterions like AIC [8], BIC [9], etc., which have been demonstrated effectively for Gaussian Mixtures. However, for MGPs, these criterions do not fit well. In order to solve this model selection problem, we try to improve AIC, BIC criterion and propose a new and effective model selection criterion for model selection on MGPs, called the Synchronously Balancing or SB criterion.

For parameter learning, EM algorithm is an effective way for finite mixtures [10]. However, for the MGP model, the approximations in the implementation of E-step or M-step must be made since it cannot be computed efficiently yet. Among these approximation versions of the EM algorithm for MGPs, we adopt the recently proposed hard-cut EM algorithm [11]. However, the EM algorithm has the local maxima problem. To solve this problem, we further implement the SMEM algorithm [12] after the convergence of the hard-cut EM algorithm.

The rest of the paper is organized as follows. Section 2 introduces the GP and MGP models. Section 3 presents the SB criterion and gives the model selection framework. In Section 4, we test the SB criterion on three synthetic datasets and compare it with AIC, BIC. Moreover, we apply our SB criterion on an artificial toy dataset to select the number of actual components. Finally, we make a brief conclusion in Section 5.

2 The GP and MGP Models

2.1 The GP Model

Given a dataset consisting of N samples $\boldsymbol{D} = \{\boldsymbol{X}, \boldsymbol{Y}\} = \{(\boldsymbol{x}_i, y_i): i = 1,2,\cdots,N\}$, where $\boldsymbol{x}_i$ is a Q-dimensional input vector, and y_i is an output, a GP model is mathematically defined as follows:

$$\boldsymbol{Y} \sim N\big(m(\boldsymbol{X}), K(\boldsymbol{X}, \boldsymbol{X})\big) \tag{1}$$

where

$$m(\boldsymbol{X}) = [m(\boldsymbol{x}_1), m(\boldsymbol{x}_2), \cdots, m(\boldsymbol{x}_N)]^T \tag{2}$$

$$K(\boldsymbol{X}, \boldsymbol{X}) = \big[K(\boldsymbol{x}_i, \boldsymbol{x}_j)\big]_{N \times N} \tag{3}$$

denote the mean vector and covariance matrix, respectively. As in most cases, we can set $m(\boldsymbol{X}) = \boldsymbol{0}$, and adopt the squared exponential (SE) covariance function [13]:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j | \boldsymbol{\theta}) = f^2 exp\left(-\frac{l^2}{2}\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2\right) + \sigma^2 I(i = j) \tag{4}$$

where $\boldsymbol{\theta} = \{f, l, \sigma\}$ denotes the parameters of the GP model. Therefore, the log-likelihood function of the outputs can be derived as follows:

$$\log p(\boldsymbol{Y} | \boldsymbol{X}, \boldsymbol{\theta}) = \log\left[N\big(\boldsymbol{Y} | \boldsymbol{0}, K(\boldsymbol{X}, \boldsymbol{X} | \boldsymbol{\theta})\big)\right] \tag{5}$$

and we can obtain the estimation of these parameters via maximum likelihood estimation (MLE), that is

$$\widehat{\boldsymbol{\theta}} = argmax_{\boldsymbol{\theta}} log[N(\boldsymbol{Y}|\mathbf{0}, K(\boldsymbol{X}, \boldsymbol{X}|\boldsymbol{\theta}))] \tag{6}$$

2.2 The MGP Model

An MGP model is comprised of multiple Gaussian Process components, and in each component, the corresponding outputs are subject to a certain Gaussian Process. These Gaussian Processes have different parameters and are independent.

For our generative MGP model, the samples are partitioned into the GP components with the following probability

$$p(z_i = c) = \pi_c; \quad c = 1,2,\cdots,C \; i.i.d \; for \; i = 1,2,\cdots,N \tag{7}$$

where $z_i = c$ means that the i-th sample belongs to the c-th GP component.

Given the partition of the samples, each input $\boldsymbol{x}_i$ is subject to a Gaussian distribution, that is

$$p(\boldsymbol{x_i}|\, z_i = c) \sim \boldsymbol{N}(\boldsymbol{\mu}_c, \boldsymbol{S}_c); \quad c = 1,2,\cdots,C \; i.i.d \; for \; i = 1,2,\cdots,N \tag{8}$$

Denote $\boldsymbol{I}_c = \{i|z_i = c\}$, $\boldsymbol{X}_c = \{\boldsymbol{x}_i|z_i = c\}$ and $\boldsymbol{Y}_c = \{y_i|z_i = c\}$ as the indexes, inputs and outputs of the samples in the c-th GP component, respectively. Given $\boldsymbol{X}_c$, the corresponding outputs $\boldsymbol{Y}_c$ is subject to the GP given by Eq.(2) with the parameters $\boldsymbol{\theta}_c = \{f_c, l_c, \sigma_c\}$, and these GP components are independent.

work effectively for Gaussian Mixture Model. However, for the MGP model, the change of $penalty$ with C is too small in comparison with that of $\log likelihood$, so that the selected value of C tends to be large. In order to solve this problem, we try to improve these two criterions to make the changes of the log-likelihood and the penalty term synchronously balanced and construct the following effective criterion:

$$F = \log likelihood - \delta N \log C \tag{11}$$

Compared with AIC and BIC, such a penalty has a much larger variation with C so that the log-likelihood and penalty are more balanced.

3.2 Model Selection with SB Criterion

Our proposed model selection framework combines the advantages of the SB criterion, the hard-cut EM algorithm [11], and the SMEM algorithm [12]. More specifically, for some values of C, we train the MGP model with hard-cut EM algorithm, update the estimated parameters via SMEM algorithm to avoid local maxima, and then select the best value of C according to the SB criterion.

Before establishing our framework for model selection, we first introduce the hard-cut EM algorithm as well as the SMEM algorithm used in this framework.

The Hard-Cut EM Algorithm and the SMEM Algorithm. The main idea of the hard-cut EM algorithm is to partition the samples into the corresponding GP components according to the maximum a posterior (MAP) criterion in E-step, that is

$$z_i = argmax_{1 \le c \le C} \pi_c N(\boldsymbol{x}_i | \boldsymbol{\mu}_c, \boldsymbol{S}_c) N(y_i | 0, l_c^2 + \sigma_c^2) \tag{12}$$

With the known partition, the parameters of each GP component are estimated via MLE respectively in M-step, i.e.

$$\pi_c = \frac{1}{N} \sum_{i=1}^{N} I(z_i = c), \boldsymbol{\mu}_c = \frac{\sum_{i=1}^{N} I(z_i = c) \boldsymbol{x}_t}{\sum_{i=1}^{N} I(z_i = c)}, \boldsymbol{S}_c = \frac{\sum_{i=1}^{N} I(z_i = c)(\boldsymbol{x}_i - \boldsymbol{\mu}_c)(\boldsymbol{x}_i - \boldsymbol{\mu}_c)^T}{\sum_{i=1}^{N} I(z_i = c)} \tag{13}$$

and $\boldsymbol{\theta}_c$ is learnt by Eq. (6).

In the SMEM algorithm, merge and split operations are implemented after the convergence of EM iterations in order to avoid local maxima.

For convenience, we denote

$$post_{ci} = p(z_i = c | \boldsymbol{x}_i, y_i) = \frac{\pi_c N(\boldsymbol{x}_i | \boldsymbol{\mu}_c, \boldsymbol{S}_c) N(y_i | 0, l_c^2 + \sigma_c^2)}{\sum_{c=1}^{C} \pi_c N(\boldsymbol{x}_i | \boldsymbol{\mu}_c, \boldsymbol{S}_c) N(y_i | 0, l_c^2 + \sigma_c^2)} \tag{14}$$

as the posterior probability of the t-th sample belonging to the c-th GP component obtained from EM iterations, and denote $post_c = (post_{c1}, post_{c2}, \cdots, post_{cN})$. When $post_u$ and $post_v$ are almost equal, we can merge the u-th and the v-th GP components into one component. So, we define the merge criterion as the similarity between $post_u$ and $post_v$:

$$O_{merge}(u, v) = \frac{post_u post_v^T}{\|post_u\| \|post_v\|}, u \neq v \tag{15}$$

where $||\cdot||$ denotes the Euclidean vector norm, and we can merge the two GP components with the largest $O_{merge}(u, v)$.

After the merge operation, we attempt to split each GP component into two GP components, called the k_1-th and the k_2-th components, and estimate $post_{k_1}$and $post_{k_2}$ by minimizing $O_{merge}(k_1, k_2)$. Then, we only accept the split of the k^*-th GP component which leads to the smallest minimum $O_{merge}(k_1, k_2)$.

The Model Selection Framework. Denote the set of candidate values of C as

$$S = \{C | l \leq C \leq L\} \tag{16}$$

For each element C from the set S, we learn the MGP model with C components via the hard-cut EM algorithm and SMEM algorithm in turn to get the maximum likelihood:

Step 1 Initialization: Set $s = 1, BestL_C = -Inf$ and initialize the parameters $\{\boldsymbol{\Theta}^0, \boldsymbol{\Psi}^0\}$ in the MGP model.
Step 2 Parameter Learning:

At phase s, we perform the hard-cut EM algorithm with the initial parameters $\{\boldsymbol{\Theta}^{s-1}, \boldsymbol{\Psi}^{s-1}\}$. After convergence, we obtain the estimated parameters $\{\widetilde{\boldsymbol{\Theta}}^s, \widetilde{\boldsymbol{\Psi}}^s\}$, and the corresponding log-likelihood function L_C. If $L_C > BestL_C$, then set $BestL_C = L_C$.

Then, implement the SMEM algorithm [12] with the initial parameters $\{\widetilde{\boldsymbol{\Theta}}^s, \widetilde{\boldsymbol{\Psi}}^s\}$, and we can obtain the updated parameters $\{\boldsymbol{\Theta}^s, \boldsymbol{\Psi}^s\}$, and the corresponding log-likelihood L_C. If $L_C > BestL_C$, then set $BestL_C = L_C$.
Step 3 Set $s = s + 1$, if $s = MaxTime$, terminate and output $BestL_C$; otherwise, return to Step2.

After the learning process above, we have obtained the maximum log-likelihood $BestL_C$, for each C from the candidate set S. Then according to the SB criterion, we obtain the appropriate number of GP components as follows:

$$C^* = argmax_{C \in S}\{BestL_C - \delta N \log C\} \tag{17}$$

4 Simulation Experiments

In order to test the effectiveness and accuracy of our proposed SB criterion for model selection, we generate three typical synthetic datasets from the MGP model, and then apply the SB criterion to these datasets with various values of penalty coefficient δ and compare the SB criterion with two classical model selection criterions, AIC and BIC, on the large synthetic dataset. Moreover, we carry out the same experiment on an artificial toy dataset. Finally, by summarizing these experimental results, we obtain an appropriate empirical interval for δ that leads to reliable model selection.

4.1 On Three Typical Synthetic Datasets of MGP

Three synthetic datasets are generated from the MGP models with different sizes. For the small synthetic dataset, there are 939 samples and 5 GP components, as shown in Fig.1. The medium synthetic dataset has 2400 samples and 8 GP components, as plotted in Fig.2. The large synthetic dataset has 10000 samples and 10 GP components, as plotted in Fig.3.

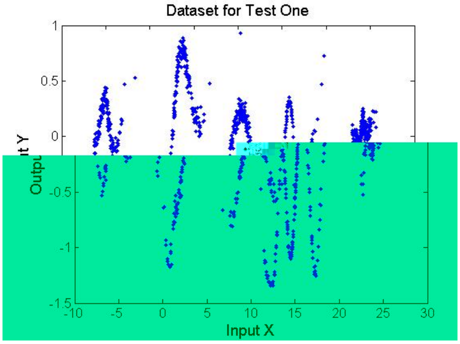


Fig. 1. The small synthetic dataset

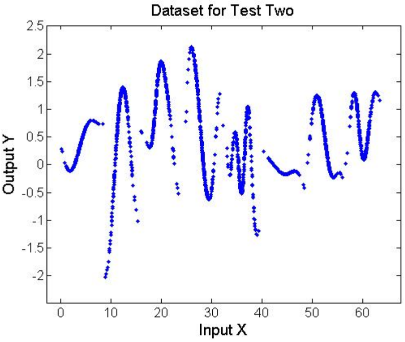


Fig. 2. The medium synthetic dataset

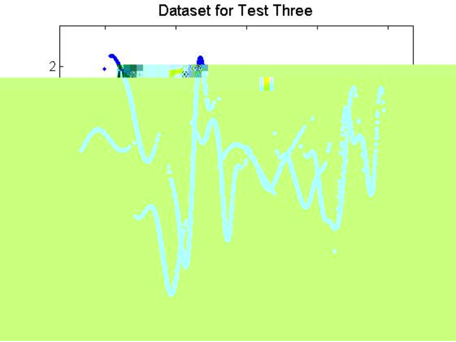


Fig. 3. The large synthetic dataset

Then, we apply the model selection framework above to these synthetic datasets with some $\delta \in (0,3)$. The candidate sets for the small, medium and large datasets are $S = \{2,3,\cdots,10\}$, $S = \{3,4,\cdots,13\}$ and $S = \{5,6,\cdots,15\}$, respectively. We repeat the experiment 18 times on the small dataset and 15 times on the medium and large datasets. For each value of δ, the number of experiments where the estimated value of C does not equal to the true value is shown in Figs.4-6 for the three datasets, respectively.

It can be seen from Figs. 4-6 that our proposed model selection framework selects the correct value of C with very high probability when the penalty coefficient δ lies in a suitable interval, whereas the error increases when δ gets away from this interval, since appropriate value of δ ensures the balance between the log-likelihood and the penalty. The suitable intervals are $(1.0,1.8)$, $(1.3,2.2)$ and $(1.10,1.75)$ for the small, medium and large datasets, respectively. Particularly, with the best value of δ, our proposed model selection framework based on the SB criterion gives correct result for all the 15 times on both the medium and the large synthetic dataset, whereas the large dataset has heavy overlaps among the GP components that makes model selection even more difficult, which firmly demonstrates the strong ability of our proposed model selection framework.

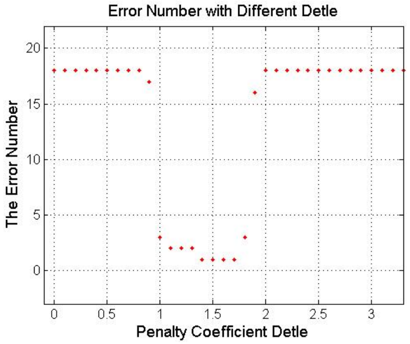


Fig. 4. Model selection result on the small synthetic dataset

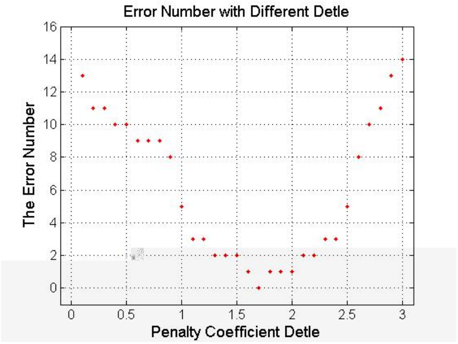


Fig. 5. Model selection result on the medium synthetic dataset

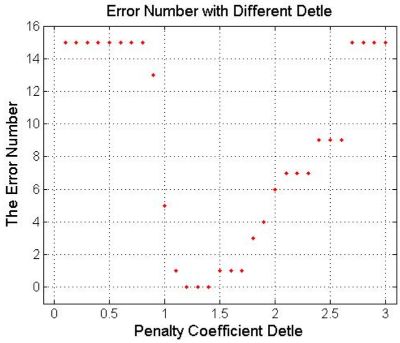


Fig. 6. Model selection result on the large synthetic dataset

4.2 Experimental Results with AIC and BIC Criterion

To compare our proposed SB criterion with two classical criterions, AIC and BIC, we also apply AIC and BIC criterions to the three synthetic datasets above. Figs.7 & 8 show the objective functions of AIC and BIC criterions against the value of C on the large synthetic dataset, respectively. From Figs.7 & 8, it can be seen that these two criterions prefer to select the maximum value of C from the candidate set, since the log-likelihood is dramatically increasing with C for MGP models whereas the penalty is relatively stable, so that the objective functions also increase with C. In contrast, due to the synchronous balance between the log-likelihood and the penalty, our proposed SB criterion significantly outperforms the AIC and BIC criterion on the synthetic dataset.

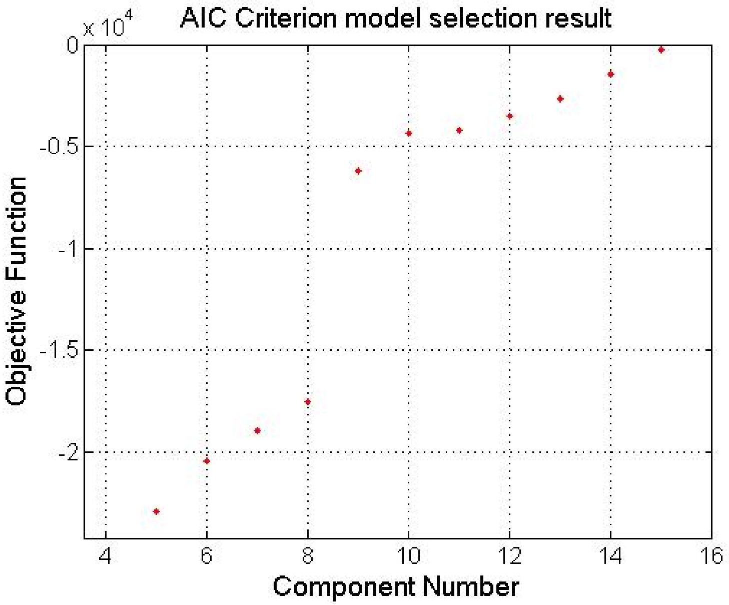


Fig. 7. The model selection result with AIC criterion on the large synthetic dataset

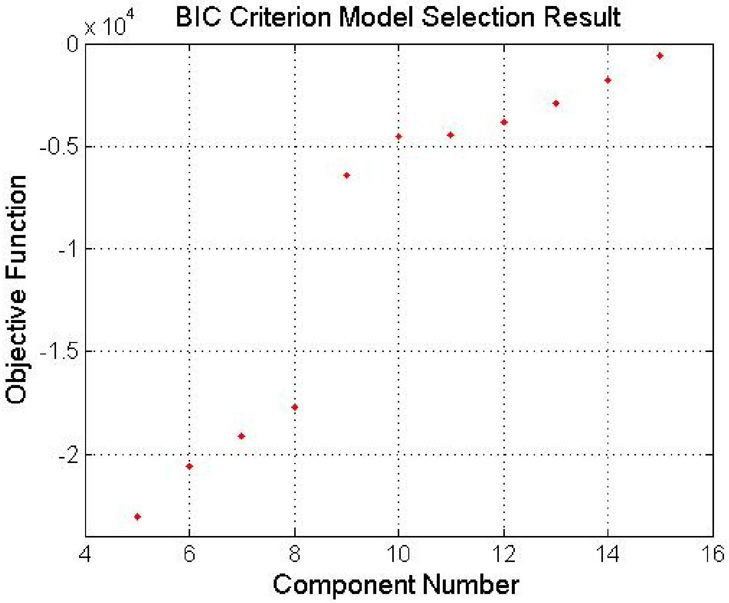
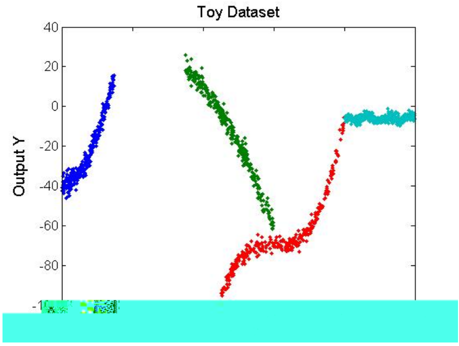


Fig. 8. The model selection result with BIC criterion on the large synthetic dataset

4.3 On an Artificial Toy Dataset

The artificial toy dataset is used to test some MGP models since it is highly multimodal [6, 7, 11, 14]. The dataset consists of four groups, and each group is generated from a continuous function with different levels of Gaussian noise. In our experiment, we generate 200 samples for each group, as shown in Fig.9. Then, we apply the SB criterion and repeat the experiment 25 times with the candidate set $S = \{2,3,\cdots,10\}$. For some values of δ, the number of experiments where the estimated value of $C \neq 4$(the true number of components) is shown in Fig.10. It can be observed from Fig. 10. that the SB criterion makes mistakes only twice among the 25 times, which means it can select the true number of components with very high probability, when δ comes from $(1.25,2.15)$, whereas the performance becomes poorer when δ gets too large or too small, as also shown in the experiments on the synthetic datasets above. Since the Toy dataset does not come from MGP models and is more similar to a real dataset, our proposed model selection framework also demonstrates potential applicability.



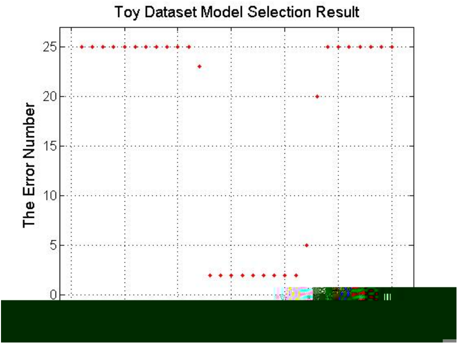


Fig. 9. The toy dataset

Fig. 10. Model selection result on toy dataset

4.4 Experimental Conclusion and Penalty Coefficient Choice

It can be summarized from the experimental results above that the performance of our proposed model selection framework heavily relies on the penalty coefficient δ. With a suitable value of δ, our model selection framework works well on both the synthetic datasets and the toy dataset. Besides, the appropriate intervals for δ in these experiments are close to each other and the intersection of these intervals is $(1.3,1.7)$, which leads to the correct value of C with very high probability. Therefore, $(1.3,1.7)$ can be an empirical interval of δ for model selection on a new dataset.

5 Conclusion

We have established an effective criterion for model selection of the MGP model, where the log-likelihood and the penalty are much more synchronously balanced in comparison with classical criterions like AIC and BIC. From the experimental results, it can be demonstrated that when the penalty coefficient is within a certain feasible interval, like $(1.3, 1.7)$, our proposed SB criterion can obtain the true number of GP components with very high probability, and significantly outperforms AIC and BIC.

Acknowledgement. This work was supported by the National Science Foundation of China under Grant 61171138

References

1. Yuan, C., Neubauer, C.: Variational mixture of Gaussian process experts. In: Advances in Neural Information Processing Systems, pp. 1897–1904 (2008)
2. Tresp, V.: Mixtures of Gaussian processes. In: Advances in Neural Information Processing Systems, vol. 13, pp. 654–660 (2000)
3. Nguyen, T., Bonilla, E.: Fast Allocation of Gaussian Process Experts. In: Proceedings of The 31st International Conference on Machine Learning, pp. 145–153 (2014)
4. Rasmussen, C.E., Ghahramani, Z.: Infinite mixtures of Gaussian process experts. In: Advances in Neural Information Processing Systems, vol. 14, pp. 881–888 (2001)
5. Fergie, M.P.: Discriminative Pose Estimation Using Mixtures of Gaussian Processes. The University of Manchester (2013)
6. Yang, Y., Ma, J.: An efficient EM approach to parameter learning of the mixture of Gaussian processes. In: Liu, D., Zhang, H., Polycarpou, M., Alippi, C., He, H. (eds.) ISNN 2011, Part II. LNCS, vol. 6676, pp. 165–174. Springer, Heidelberg (2011)
7. Meeds, E., Osindero, S.: An alternative infinite mixture of Gaussian process experts. In: Advances in Neural Information Processing Systems, vol. 18, pp. 883–890 (2005)
8. Akaike, H.: A new look at the statistical identification model. IEEE Trans. on Automat. Control 19(6), 716–723 (1974)
9. Liddle, A.R.: Information criterion for astrophysical model selection. Monthly Notices of the Royal Astronomical Society: Letters 377(1), L74–L78 (2007)
10. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of Royal Statistical Society, Series B (Methodological), 1–38 (1977)
11. Chen, Z., Ma, J., Zhou, Y.: A Precise Hard-Cut EM Algorithm for Mixtures of Gaussian Processes. In: Huang, D.-S., Jo, K.-H., Wang, L. (eds.) ICIC 2014. LNCS, vol. 8589, pp. 68–75. Springer, Heidelberg (2014)
12. Ueda, N., Nakano, R., Ghahramani, Y.Z., Hiton, G.E.: SMEM algorithm for mixture models. Neural Computation 12(9), 2109–2128 (2000)
13. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. The MIT Press, Cambridge (2006)
14. Fergie, M.P.: Discriminative Pose Estimation Using Mixture of Gaussian Processes. The University of Manchester (2013)