A Dynamic Model Selection Algorithm for Mixtures of Gaussian Processes

Longbo Zhao, Jinwen Ma
Department of Information Science, School of Mathematical Sciences & LMAM, Peking University
Beijing, 100871, China
Email: longbzhao@pku.edu.cn, jwma@math.pku.edu.cn

***Abstract*—The mixture of Gaussian processes (MGP) is a powerful and widely used model in machine learning. However, it remains a challenging problem to determine the actual number of GP components in the mixture, i.e., the model selection problem. Synchronously Balancing Criterion (SBC) has been recently proposed and shown to be effective for the model selection of MGPs, but it is rather time consuming to use SBC directly since we need to repeat the conventional learning process on a large number of candidate models. In this paper, based on the convexity of the negative SB Criterion objective function, we propose a dynamic model selection algorithm under the framework of the hard-cut EM algorithm with the GP number dynamically changed step by step according to the increase of SBC. It is demonstrated by the experiments on some typical synthetic datasets and an artificial toy dataset that our proposed algorithm is not only much more efficient on implementation time, but also more effective on model selection, in comparison with the conventional SBC based model selection method.**

Keywords: Mixture of Gaussian processes, parameter learning, model selection, hard-cut EM algorithm; synchronously balancing criterion

I. Introduction

The Gaussian Process (GP) model is powerful and widely used in regression and classification problems. However, there are two main limitations. Firstly, it cannot fit the multi-modal dataset well. Secondly, its parameter learning involves the inverse of covariance matrix, which has a large computational complexity $O(N^3)$ [1], where N is the number of training samples. In order to overcome these limitations, Tresp [2] proposed the mixture of Gaussian Processes (MGP) in 2000. From then on, many variants of the MGP model have been suggested and could be classified into two main forms: the generative models [1,6,7] and the conditional models [2-5]. Generally, the generative model is preferred since it can infer missing inputs from outputs [7]. Here, we adopt the most simplified and refined generative model so far. In fact, for the generative model of MGP, there have been many investigations on its parameter learning [8,9] and the precise hard-cut EM algorithm [10] is currently a good method for this aim.

However, since the performance of MGP for any learning task depends heavily on the number of GP components, the selection of number of GP components, being referred to as the model selection problem, is also important but rather difficult. The classical model selection criterions, such as AIC [11], BIC [12] and MML [13], have been demonstrated effective for Gaussian or finite mixtures, but cannot work so well for MGPs. Recently, we proposed an effective model selection criterion for MGPs, called Synchronously Balancing Criterion (SBC) [14]. In fact, it can be considered as an improved version of AIC and BIC and the experimental results demonstrate that SBC can detect the true component number with high probability when the penalty coefficient δ is in the feasible interval. However, in the conventional way, we need to train and check each candidate model separately for a large set of GP number, which involves heavy computation. In this paper, under the framework of the hard-cut EM algorithm we will establish a dynamic model selection algorithm with the GP number dynamically changed step by step according to the increase of SBC. Due to the convexity of the negative SBC, this dynamic model selection algorithm is effective and efficient, which can quickly detect the true number of GPs in a dataset.

The rest of the paper is organized as follows. Section 2 introduces the GP and MGP models. We present the dynamic model selection algorithm for MGPs in Section 3. The experimental results are contained in Section 4. Finally, we make a brief conclusion in Section 5.

II. The GP and MGP Models

A. The GP Model

Given a sample dataset $\mathbf{D}=\{\boldsymbol{X},\boldsymbol{Y}\}=\{(\boldsymbol{x}_i,y_i):i=1,2,...,N\}$, where $\boldsymbol{x}_i$ is a d-dimensional input vector, and y_i is an output. A GP model is defined as follows:

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

where $m(\boldsymbol{X})$ and $K(\boldsymbol{X},\boldsymbol{X})$ denote the mean vector and covariance matrix, respectively. Without loss of generality, we assume $m(\boldsymbol{X}) = \boldsymbol{0}$. For the covariance matrix, we adopt the squared exponential (SE) covariance function [15]:

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j;\theta) = \sigma_f^2 exp(-\frac{\sigma_l^2}{2}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2) + \sigma_n^2 I_{(i=j)}, \tag{2}$$

where $\boldsymbol{\theta}=\{\sigma_f^2,\sigma_l^2,\sigma_n^2\}$ denote the parameters of the GP model. So, the log likelihood of the GP model on the given dataset can be given as follows:

$$\log p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\theta}) = \log N(\boldsymbol{Y}|\boldsymbol{0}, K(\mathbf{X},\mathbf{X})). \tag{3}$$

In order to learn these parameters $\boldsymbol{\theta}$, we can perform the maximum likelihood estimation (MLE) procedure [15], that

is, we can get

$$\hat{\theta} = argmax_{\theta} \log N(\boldsymbol{Y}|\boldsymbol{0}, K(\boldsymbol{X},\boldsymbol{X})). \quad (4)$$

B. The MGP Model

On the basis of the GP model, we can combine a number of GPs together and form the MGP model. These GP components have different hyperparameters. We adopt the generative model and denote C and N as the numbers of GPs and traing samples in the MGP model, respectively.

Specifically, we define our generative MGP model as follows:

Step 1. Partition the samples into the GP components with the following probability distribution:

$$p(z_n = c) = \pi_c, \quad (5)$$

where c=1,...,C and n=1,...,N.

Step 2. Give the partition of the samples, each input $\boldsymbol{x}_i$ is distributed according to a corresponding Gaussian distribution:

$$p(\boldsymbol{x}_i|z_n = c) \sim N(\boldsymbol{\mu}_c, \boldsymbol{S}_c); \quad (6)$$

Denote $\boldsymbol{I}_c$=$\{n|z_n = c\}$, $\boldsymbol{X}_c$=$\{\boldsymbol{x}_n|z_n = c\}$, $\boldsymbol{Y}_c$=$\{y_n|z_n = c\}$ (c=1,...,C, n=1,...,N) as the sample indexes, inputs and outputs of the training samples in the c-th GP component, respectively.

Step 3. Given $\boldsymbol{X}_c$, the corresponding c-th GP component can be mathematically defined as follows:

$$\boldsymbol{Y}_c \sim N(\boldsymbol{0}, K(\boldsymbol{X}_c, \boldsymbol{X}_c)) \quad (7)$$

with the hyperparameters $\boldsymbol{\theta}_c$=$\{\sigma^2_{fc},\sigma^2_{lc},\sigma^2_{nc}\}$.

In summary, we mathematically define the MGP model by Eqs. (5),(6),(7). Based on these equations, the log likelihood function is derived as follows:

$$\log(p(\boldsymbol{Y}_c|\boldsymbol{X}_c, \boldsymbol{\Theta}, \boldsymbol{\Psi})) = \sum_{c=1}^{C} \Big(\sum_{n \in \boldsymbol{I}_c} (\log(\pi_c p(\boldsymbol{x}_n|\boldsymbol{\mu}_c, \boldsymbol{S}_c))) + \log(p(\boldsymbol{Y}_c|\boldsymbol{X}_c, \boldsymbol{\theta}_c)) \Big) \quad (8)$$

where $\boldsymbol{\Theta}$=$\{\boldsymbol{\theta}_c$: c=1, ..., $C\}$ and $\boldsymbol{\Psi}$=$\{\boldsymbol{\mu}_c$, $\boldsymbol{S}_c$, π_c: c=1, ..., $C\}$ denote the whole set of (hyper)parameters for outputs and inputs, respectively.

III. The Dynamic Model Selection Learning Procedure

We use the EM algorithm as the basic learning framework for the parameter learning and model selection of MGPs. Let z_{nc} be the hidden variables, where z_{nc} is a Kronecker delta. Denote z_{nc}=1, if the sample ($\boldsymbol{x}_n$,y_n) belongs to the c-th GP component. Derive the log likelihood function of the complete data from Eq.(8) to be

$$\log(p(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{\Psi})) = \sum_{c=1}^{C} \Big(\sum_{n=1}^{N} (z_{nc} \log(\pi_c p(\boldsymbol{x}_n|\boldsymbol{\mu}_c, \boldsymbol{S}_c))) + \log(p(\boldsymbol{Y}_c|\boldsymbol{X}_c, \boldsymbol{\theta}_c)) \Big) \quad (9)$$

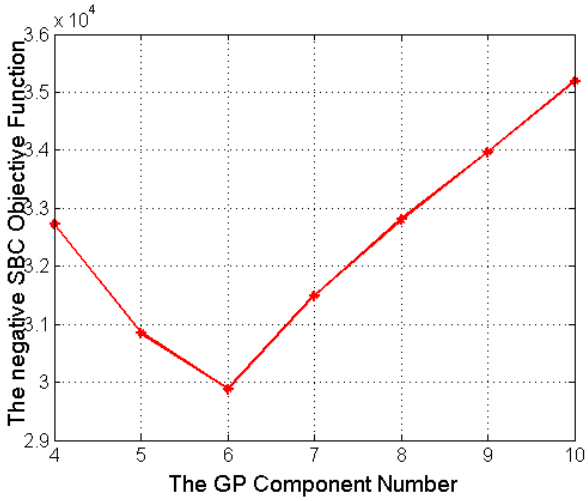

Fig. 1. The sketch of the negative SBC objective function value with the GP component number

Our dynamic model selection algorithm introduces the SBC based step by step model selection mechanism into the hard-cut EM algorithm. In order to do so, we begin to briefly introduce the hard-cut EM algorithm and SBC.

A. The Hard-Cut EM Algorithm

The main idea of the hard-cut EM algorithm is to assign the samples to the corresponding GP components according to the maximum a posterior (MAP) criterion in E-step:

$$k_n = argmax_{1 \le c \le C} \{\pi_c p(\boldsymbol{x}_n|\boldsymbol{\mu}_c, \boldsymbol{S}_c) p(y_n|\boldsymbol{\theta}_c)\} \quad (10)$$

that is, $z_{k_n n}$=1. With the known partition, we can learn the hyperparameters of each GP component via the MLE procedure in M-step.

For the parameters $\boldsymbol{\Psi}$, we have

$$\pi_c = \frac{1}{N} \sum_{n=1}^{N} z_{k_n n}; \quad (11)$$

$$\boldsymbol{\mu}_c = \frac{\sum_{n=1}^{N} (\boldsymbol{x}_n z_{k_n n})}{\sum_{n=1}^{N} z_{k_n n}}; \quad (12)$$

$$\boldsymbol{S}_c = \frac{\sum_{n=1}^{N} z_{k_n n} (\boldsymbol{x}_n - \boldsymbol{\mu}_c)^T (\boldsymbol{x}_n - \boldsymbol{\mu}_c)}{\sum_{n=1}^{N} z_{k_n n}}. \quad (13)$$

For the hyperparameters $\boldsymbol{\Theta}$, we perform the MLE procedure on each GP component as Eq.(4) does.

B. Synchronously Balancing Criterion (SBC)

SBC is a new and effective model selection criterion for mixtures of Gaussian processes. It takes the advantages of both AIC and BIC. The main idea of SBC is to make the changes of the log likelihood and the penalty term synchronously balanced with the component number C. Mathematically, the SBC is expressed as follows:

$$SBC(C) = \log likelihood - \delta N \log C, \quad (14)$$

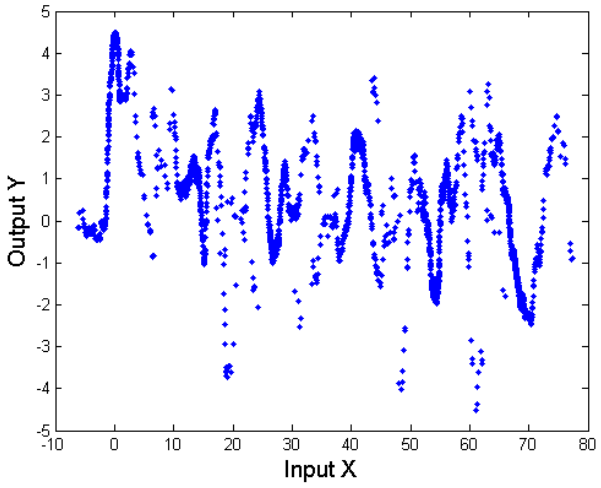

Fig. 2. The sketch of data points in the synthetic dataset of the first group with $= 1{:}7$.

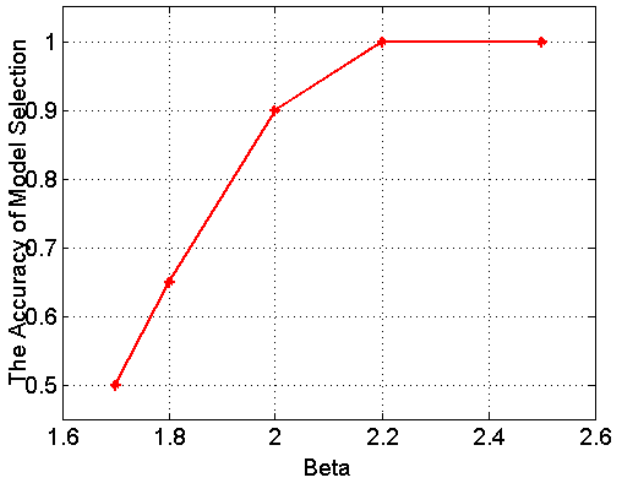

Fig. 3. The accuracy of model selection on the first group dataset with different value of .

where δ is the penalty coefficient and generally has a feasible interval (1.3, 1.7) for model selection. When δ is within this feasible interval, SBC can obtain the true number of GP components with high probability [14].

C. The Learning Procedure

In the MGP model, the sketch of the negative SBC objective function with the GP component number is shown in Fig.1, the true component number C =6. From Fig.1, we can observe that the negative SBC objective function obtains the minimum value with the true GP component number C , and the negative SBC objective function has the convexity property. Therefore, if $SBC(C)$=$max_{C_0 2 f C\text{-}1,\ C,\ C+1 g}\{SBC(C_0+1)\}$, we can set C as the final component number via the minimum selection and stable strategy.

With the help of the hard-cut EM algorithm, SBC, and the convexity property of negative SBC objective function, we now present the learning procedure of the dynamic model selection algorithm for MGPs as follows:

Step 1. Initialization: set an initial GP component number $C_0(C_0 \geq 2)$.

Step 2. Perform the hard-cut EM algorithm to train the MGP model with the GP component numbers C_0-1, C_0, C_0+1, respectively, and record the corresponding SBC objective function values $SBC(C_0-1)$, $SBC(C_0)$, $SBC(C_0+1)$.

Step 3. Obtain the maximum SBC objective function value: $BestC$=$argmax_{C 2 f C_0\text{-}1,\ C_0,\ C_0+1 g}\{SBC(C)\}$.

Step 4. Discuss $BestC$ and continue the iteration until stop.

(i). If $BestC$=C_0-1, set C_0=C_0-1. Perform the hard-cut EM algorithm to train the MGP model with the GP component number C_0-1 and go to step 3;

(ii). If $BestC$=C_0+1, set C_0=C_0+1. Perform the hard-cut EM algorithm to train the MGP model with the GP component number C_0+1 and go to step 3;

(iii). If $BestC$=C_0, stop the algorithm and set the current C_0 as the final result of our algorithm.

According to the above learning procedure, we do not need to know the candidate set in advance. Moreover, according to the convexity of the negative SBC objective function, the dynamic model selection algorithm can quickly converges and save much repeat computation.

IV. Experimental Results

In order to test the accuracy and effectiveness of our dynamic model selection algorithm for MGP model, we carry out simulation experiments on several typical synthetic datasets. Moreover, we perform our algorithm on an artificial toy data.

A. Simulation Experiments

In our simulation experiments, we generate three groups of synthetic datasets with one dimensional input: the first and second groups consist of datasets with six GP components in the mixture model, while the third group consists of datasets with eight GP components.

1) On the datasets of the first group: In the first group, these synthetic datasets are the same except the degree of overlap. For each component, the input variable is subject to a 1-dimensional Gaussian distribution. The means of six components are shown as follows:

$$means = [0, 8, 16, 32, 40, 48] * \beta, \tag{15}$$

where $1/\beta$ can be served as the degree of overlap. Here, we set $\beta = \{1.7, 1.8, 2.0, 2.2, 2.5\}$. Thus, this group has five data sets. Set $s = [576, 504, 432, 360, 504, 576]$ as the sample numbers of 6 components, respectively. According to the inputs, we generate the corresponding outputs with Eq.(7). Fig.2 shows the sketch of data points in the dataset with β=1.7.

Perform our algorithm on each of the five datasets with the SBC penalty coefficient δ = 1.6. During the initialization, we randomly initialize the hyperparameters and the component number C_0. Repeat the experiment 20 times on each dataset.

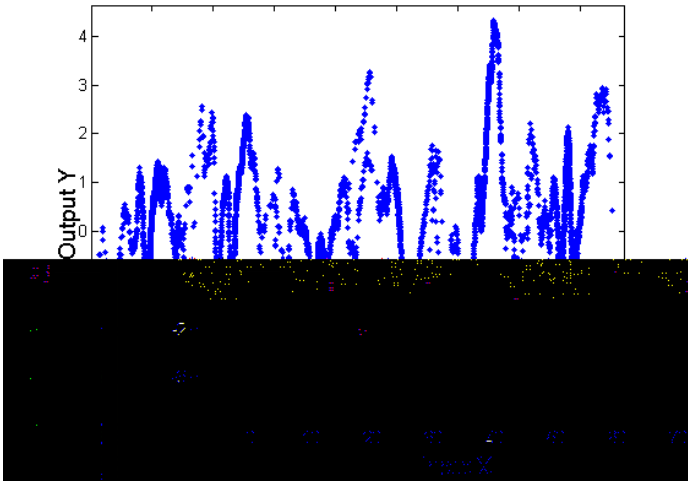

Fig. 4. The sketch of data points in the synthetic dataset of the second group with $= 1{:}7$.

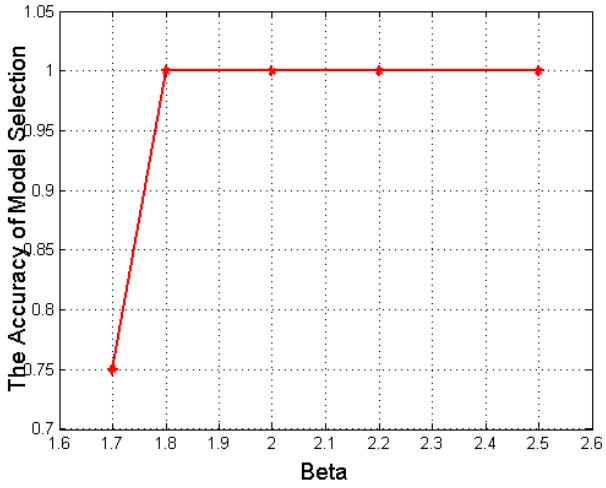

Fig. 5. The accuracy of model selection on the second group dataset with different value of .

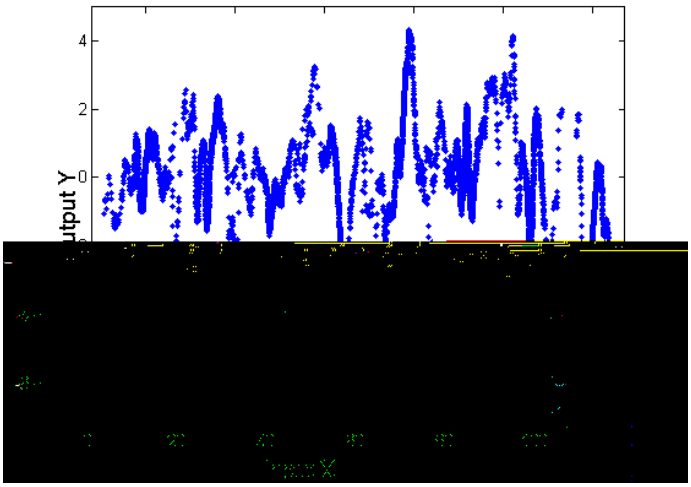

Fig. 6. The sketch of data points in the synthetic dataset of the third group with $= 1{:}8$.

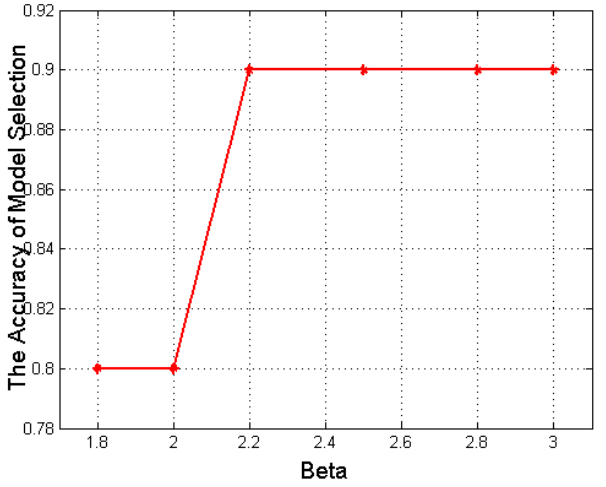

Fig. 7. The accuracy of model selection on the third group dataset with different value of .

The experimental results of the model selection accuracy with different β is shown in Fig.3

From Fig.3, we can observe that our algorithm selects the correct value of C = 6 with very high accuracy p, when the overlapping degree $1/\beta$ is small enough, whereas the accuracy decreases when the degree of overlap $1/\beta$ becomes larger, i.e., β=1.7, p=0.5. In this situation, the major reason may be that the sample number of each component is too small to interpret the large degree of overlap in the MGP model, which will be validated by the experiments on the datasets of the other two groups.

2) On the datasets of the second group: In this situation, we generate the synthetic datasets still with $\beta = \{1.7, 1.8, 2.0, 2.2, 2.5\}$, i.e., keep the same degrees of overlap, but increase the sample numbers by setting s = [1296, 1134, 972, 810, 1134, 1296]. Fig.4 shows the sketch of data points in the dataset with β=1.7.

Perform our algorithm on each of these datasets with the SBC penalty coefficient δ = 1.6. The initialization is randomly set. Repeat the experiment 20 times on each dataset. The experimental results is shown in Fig.5.

From Fig.5, we can observe that our algorithm also obtain the correct value of C = 6 with high accuracy p with different degree of overlap $1/\beta$. Especially, when the degree of overlap $1/\beta$ is large, our algorithm can still obtain a better experimental result, i.e., β=1.7, p=0.75, since the large degree of overlap can be better interpreted by this group of datasets with a larger sample number.

3) On the datasets of the third group: In the last situation, we generate a group of datasets with C =8 components. Set s = [1296, 1134, 972, 810, 1134, 1296, 1080, 944] as the sample number. For an input of GP component, it is subject

to a Gaussian distribution. The means of eight components are given as follows:

$$mean = [0, 8, 16, 32, 40, 48, 56, 64] * \beta \quad (16)$$

We set $\beta = \{1.8, 2.0, 2.2, 2.5, 2.8, 3.0\}$. So, there are six datasets in this group. Fig.6 shows the sketch of data points in the dataset with $\beta = 1.8$. Perform our algorithm on each of these six data sets with the SBC penalty coefficient $\delta = 1.6$. For initialization, we apply the randomization method to initialize the (hyper)parameters and initialize the component number $C_0 \geq 6$. Repeat 10 times on each dataset. The experimental results are shown in Fig.7. From Fig.7, we can observe that our algorithm works so well on these six datasets with different overlapping degrees. Especially, when the overlapping degree $1/\beta$ is so large like $\beta = 1.8$, our algorithm can still get the correct value of $C = 8$ with high accuracy p=0.8.

B. Experiments on an Artificial Toy Data

We finally perform our algorithm on an artificial toy dataset which is classical and often used as a beach mark dataset for the MGP modelling [6, 7, 16, 17]. It consists of four components, and each component is generated from a continuous function with Gaussian noise. In our experiment, each component has 500 samples, being shown in Fig.8.

We implement our algorithm on this toy dataset with the SBC penalty coefficient $\delta = 1.6$. We randomly initialize the hyperparameters as well as the component number C_0, and repeat the experiment 30 times. As a result, our algorithm obtains the true number of component $C = 4$ with high accuracy p=0.833.

V. Conclusion

With the help of the hard-cut EM algorithm, the effective model selection criterion SBC as well as the convexity property of the negative SBC objective function, we have established the dynamic model selection algorithm for MGPs which can quickly detect the true number of GPs through dynamically changing the component number step by step according to the increase of SBC objective function. At each iteration, we select the component number C with the optimal SBC objective function and output the result when it cannot be changed. The experiments are conducted on three groups of synthetic datasets with 6 or 8 typical and different GPs as well as an artificial dataset. It is demonstrated by the experimental results that our proposed dynamic model selection algorithm can achieve the true number of GP components number with high probability.

Acknowledgment

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

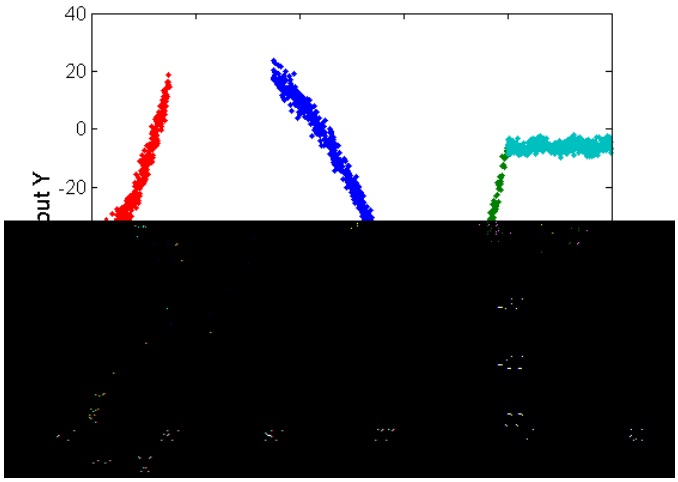

Fig. 8. The sketch of data pints in the artificial toy dataset.

References

[1] Yuan, C., Neubauer,C., Variational mixture of Gaussian process experts Advances in Neural Information Processing Systems, pp. 1897 - 1904 (2008)

[2] Tresp, V., Mixtures of Gaussian processes. Advances in Neural Information Processing Systems, vol. 13, pp. 654 - 660 (2000)

[3] Fergie, M.P.: Discriminative Pose Estimation Using Mixtures of Gaussian Processes. The University of Manchester (2013)

[4] Nguyen, T., Bonilla, E.: Fast Allocation of Gaussian Process Experts. In: Proceedings of The 31st International Conference on Machine Learning, pp. 145 - 153 (2014)

[5] Rasmussen, C.E., Ghahramani, Z.: Infinite mixtures of Gaussian process experts. In: Advances in Neural Information Processing Systems, vol. 14, pp. 881 - 888 (2001)

[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 (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] Di Wu, Ziyi Chen, and Jinwen MA: An MCMC Based EM Algorithm for Mixtures of Gaussian Processes. X. Hu et al.(Eds) ISNN 2015. LNCS 9377, pp.327 – 334, 2015.

[9] Yan Yang,Jinwen MA: An efficient EM approach to parameter learning of te mixture of Gaussian processes. Liu, D., Zhang, H., Polycarpou, M., Alippi, C., He, H.(Eds) ISNN 2011. LNCS 6676, pp.165 – 174, 2011.

[10] 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)

[11] Akaike, H.: A new look at the statistical identification model. IEEE Trans. on Automat. Control 19(6), 716 – 723 (1974)

[12] Liddle, A.R.: Information criterion for astrophysical model selection. Monthly Notices of the Royal Astronomical Society: Letters 377(1), pp.74 – 78, (2007)

[13] Neil J, Korb K B. The MML Evolution of Classification Graphs[J]. Genetic Programming, 1998.

[14] Longbo Zhao, Ziyi Chen, and Jinwen MA, An Effective Model Selection Criterion for Mixtures of Gaussian Processes. X. Hu et al.(Eds) ISNN 2015. LNCS 9377, pp.345 – 354, 2015.

[15] Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. The MIT Press, Cambridge (2006)

[16] 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)

[17] Fergie, M.P.: Discriminative Pose Estimation Using Mixture of Gaussian Processes. The University of Manchester (2013)