



Automatic Model Selection Algorithm Based on BYY Harmonic Learning for Mixture of Gaussian Process Functional Regressions Models

Xiaogang Guo¹, Taohong Guo¹ and Jiewen Ma² (✉)

¹Department of Information and Computational Science, School of Mathematical Science and LMAM, Peking University, Beijing 100871, China
jwma@math.pku.edu.cn

Abstract. For the mixture model, determining the best functional regression model is a difficult task. This article proposes a data-driven automatic model selection algorithm based on Bayesian Ying-Yang (BYY) harmonic learning for Gaussian process functional regression (GPFR) model. BYY harmonic learning has been successfully applied to the model selection of Gaussian mixture model (GMM), but it cannot be directly used for GPFR model. We deal with the case of the single and multiple regression functional regression model based on Gaussian process (GP) model, through which GPFR model is transformed into GPFR model into a GMM. Then, we can make model selection for GPFR model via BYY harmonic learning. Experiments show that the proposed automatic model selection algorithm can find the best functional regression model for the simulated data set.

Keywords: Mixture of Gaussian Process Functional Regression Model, Model Selection, Bayesian Ying-Yang Harmonic Learning, Cross-Entropy

1 Introduction

Gaussian process (GP) model is an effective tool for Bayesian linear and nonlinear classification and regression, e.g., classifying the image of handwritten digits and detecting the existence of a face [1]. However, the current algorithm for model selection on the simulated data set is not effective. Therefore, in this paper, we transform the GPFR model into a GMM [2, 3] and the existing research has been devoted to finding the best model, although the effectiveness of the algorithm is still not clear [4–8].

Like the mixture model, GPFR model also faces the problem of model selection, and determining the best Gaussian process functional regression (GPFR) model is a difficult task. Since a linear combination of GPFR models is still a GPFR model, we can use the GPFR model to solve the problem of model selection. In addition, the algorithm for model selection is still a difficult task, and we can

al deignatatic del electi alg cith . The traditi al eth di t ch e the ti al . ber f GPFR c et thr gh cetati tatical electi critei . For exa le, Qia g et al. [6] c sed the liti g ectati - xi i ati (SEM) alg cith ba ed the Ba e ia i f c ai critei (BIC) [9]. H e e , all the i t- i g tatical electi criteia fle ca e a i c e . ber f GPFR c et a d the e f a tatical electi critei i c c a high ti e c l e , i c e e eed t e eal the h le a a e e t i ati g c c e f c d i f f e e t . ber f GPFR c et . M e e , t cha tic i l a t i e t h d , c h a c e e i b l e j . M a k chai M te Carl [10] a d Dirichlet c c e e [11], ha e al bee e d t deal ith the del electi c ble f i -GPFR del [5, 7, 8]. H e e , the e e t h d c e i e c l l e c t i g a l a g e . ber f a l e , h i c h e l l i a h i g c t a t i a l c t .

F c Ga ia i t c e del (GMM), the a t a t i c del electi alg cith ba ed Ba e ia Yi g- Ya g (BYY) ha e l e a i g [12, 13] ha e a c i e d b e t t e r c e l l a d h i g h e c t a t i e e d t h a t h e b a e d t a t i t a c l e c t i c r i t e i a a d t cha tic i l a t i e t h d [14 20]. c d i i g a

ba ed Bed ha e l e a i g c i -GPFR de. al Bed

i -GPFR f i a h i, c i a d e, b d e d
8

del i -GPFR f c aalg cith electi e t c d e d a t a t i c
c e l l i

Here, i is the indicator variable $i \in \{1, \dots, G\}$ and z is the latent variable. For the GMM, we establish the full i -BYB theorem: $q(z = g) = \pi_g$; $q(\mathbf{z} | z = g) = \mathcal{N}(\mathbf{z} | \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$; $p(\mathbf{z}) = \frac{1}{I} \sum_{i=1}^I \delta(\mathbf{z} - \mathbf{z}_i)$, i.e. the empirical distribution;

$$p(z = g | \mathbf{z}) = \frac{\pi_g \mathcal{N}(\mathbf{z} | \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)}{\sum_{s=1}^G \pi_s \mathcal{N}(\mathbf{z} | \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)}. \quad (3)$$

Moreover, π_g is the i -BYB parameter, i.e. $\pi_g \geq 0$ and $\sum_{g=1}^G \pi_g = 1$. The i -BYB has

$$H(p||q) = J(\Theta) = \frac{1}{I} \sum_{i=1}^I \sum_{g=1}^G \frac{\pi_g \mathcal{N}(\mathbf{z}_i | \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)}{\sum_{s=1}^G \pi_s \mathcal{N}(\mathbf{z}_i | \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)} \ln \left(\pi_g \mathcal{N}(\mathbf{z}_i | \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \right), \quad (4)$$

where $J(\Theta)$ is called harmonic function and $\Theta = \{\pi_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\}_{g=1}^G$.

According to BYB harmonic learning, the aim of $J(\Theta)$ is to find the optimal parameters Θ that minimize the harmonic function $J(\Theta)$. Here, we can make the selection of the parameters Θ by minimizing $J(\Theta)$. In the case of minimizing $J(\Theta)$, the i -BYB is the i -BYB of the i -BYB harmonic function. Compared with the automatic model selection algorithm based on variational Bayes, the i -BYB is a data-driven method, the i -BYB harmonic learning has a better performance and higher computational efficiency [14–20].

3 Automatic Model Selection Algorithm Based on BYB Harmonic Learning

First, we briefly introduce the i -GPFR model. A GP is a collection of random variables, a finite or infinite set of i -BYB which is subject to a Gaussian distribution [1]. To specify a GP $\{f(\mathbf{x}) | \mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^D\}$, we need to determine the mean function $m(\mathbf{x})$ and the covariance function $c(\mathbf{x}, \mathbf{x}')$, where

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] \text{ and } c(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]. \quad (5)$$

where the GP is denoted as

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), c(\mathbf{x}, \mathbf{x}')). \quad (6)$$

In the i -GPFR model, since $D = 1$, we denote the input x instead of \mathbf{x} . The i -GPFR model with G GPFR components can be established through the full i -BYB framework:

$$q(z = g) = \pi_g, \text{ where } \pi_g \geq 0 \text{ and } \sum_{g=1}^G \pi_g = 1; \quad (7)$$

$$q(y(x)|z = g) = \mathcal{GPFR}(x | \mathbf{b}_g, \boldsymbol{\theta}_g, r_g) = \mathcal{GP}(m(x | \mathbf{b}_g), c(x, x' | \boldsymbol{\theta}_g) + r_g^{-1} \delta(x, x')). \quad (8)$$

In E. (8), $\delta(x, x')$ is the Kronecker delta function,

$$m(x|\mathbf{b}_g) = \varphi(x)^T \mathbf{b}_g \text{ and } c(x, x'|\boldsymbol{\theta}_g) = \theta_{g0}^2 \exp \left\{ -\frac{(x - x')^2}{2\theta_{g1}^2} \right\}, \quad (9)$$

where $\varphi(x) = [\varphi_1(x), \varphi_2(x), \dots, \varphi_P(x)]^T$ is a column vector of B -like [21] and $c(x, x'|\boldsymbol{\theta}_g)$ is defined as the scaled exponential covariance function. θ_{g0}, θ_{g1} , and r_g are free parameters.

The Yingchi of the \mathbf{x} -GPFR model

$$q(z = g, y(x)) = q(z = g)q(y(x)|z = g) = \pi_g \mathcal{GPFR}(x|\mathbf{b}_g, \boldsymbol{\theta}_g, r_g) \quad (10)$$

and Yingchi is

$$p(z = g, y(x)) = p(y(x))p(z = g|y(x)) = p(y(x)) \frac{\pi_g \mathcal{GPFR}(x|\mathbf{b}_g, \boldsymbol{\theta}_g, r_g)}{\sum_{s=1}^G \pi_s \mathcal{GPFR}(x|\mathbf{b}_s, \boldsymbol{\theta}_s, r_s)}. \quad (11)$$

We denote a training dataset $\mathcal{D} = \{\mathcal{C}_i\}_{i=1}^I$, where $\mathcal{C}_i = \{(x_{in}, y_{in})\}_{n=1}^{N_i}$ is a set of training samples of length N_i . It is generally assumed that x_{i1}, \dots, x_{iN_i} are independent and identically distributed i.i.d. $[x_{i1}, x_{iN_i}]$ ($i = 1, \dots, I$). Let $\mathbf{x}_i = [x_{i1}, \dots, x_{iN_i}]^T$, $\mathbf{y}_i = [y_{i1}, \dots, y_{iN_i}]^T$, and $\Theta = \{\pi_g, \mathbf{b}_g, \boldsymbol{\theta}_g, r_g\}_{g=1}^G$. For the \mathbf{x} -GPFR model,

$$H(p||q) = \sum_{g=1}^G \int p(y(x))p(z = g|y(x)) \ln(q(z = g)q(y(x)|z = g)) dy(x) \quad (12)$$

can be accurately estimated by

$$J(\Theta) = \frac{1}{I} \sum_{i=1}^I \sum_{g=1}^G \frac{\pi_g \mathcal{N}(\mathbf{x}_i|\mathbf{m}_{ig}, \mathbf{C}_{ig})}{\sum_{s=1}^G \pi_s \mathcal{N}(\mathbf{x}_i|\mathbf{m}_{is}, \mathbf{C}_{is})} \ln(\pi_g \mathcal{N}(\mathbf{x}_i|\mathbf{m}_{ig}, \mathbf{C}_{ig})) \quad (13)$$

with $\mathbf{m}_{ig} = m(\mathbf{x}_i|\mathbf{b}_g)$ and $\mathbf{C}_{ig} = c(\mathbf{x}_i, \mathbf{x}_i|\boldsymbol{\theta}_g) + r_g^{-1} \mathbf{I}_{N_i}$, where \mathbf{I}_{N_i}

$\hat{C}_i = \{(x_n, \hat{y}_{in})\}_{n=1}^N$ & $\hat{f}_i(x)$ with $x_n = x_{in} + (n-1)\Delta$. Denoting σ_1^2 as the variance of Gaussian noise

$$\sigma_1^2 = \frac{1}{N_i} \sum_{n=1}^{N_i} (y_{in} - \hat{f}_i(x_{in}))^2. \quad (14)$$

It is clear that

$$\sigma_2^2 = \frac{1}{N_i} \sum_{n=1}^{N_i} (y_{in} - f_i(x_{in}))^2 \quad (15)$$

is a unbiased estimate of the variance of Gaussian noise in the case of a single C_i & $\hat{f}_i(x)$. Hence, σ_1^2 is a good estimate of the variance. The advantage is that there are no significant differences between $\hat{f}_i(x)$ and $\hat{f}_i(x)$. In this case, \hat{C}_i is a good data set for C_i . That is to say, the difference between the two sets of data is very small. \hat{C}_i is a good estimate of C_i in all cases, which will be validated through experiments in Sect. 4.

Let $\hat{D} = \{\hat{C}_1, \dots, \hat{C}_I\} = [x_1, x_2, \dots, x_N]^T$, and $\hat{y}_i = [\hat{y}_{i1}, \hat{y}_{i2}, \dots, \hat{y}_{iN}]^T$. \hat{D} can be regarded as a data set of the full Bayesian GMM:

$$q(z = g) = \pi_{g|V}, \text{ where } \pi_g \geq 0 \text{ and } \sum_{g=1}^G \pi_g = 1; q(\mathbf{y}_i | z = g) = \mathcal{N}(\mathbf{y}_i | \mathbf{m}_g, \mathbf{C}_g), \quad (16)$$

where $\mathbf{m}_g = m(\mathbf{b}_g)$ and $\mathbf{C}_g = c(\mathbf{b}_g, \mathbf{r}_g^{-1} \mathbf{I}_N)$. In this case, the

$$q(z = g, \mathbf{y}_i) = q(z = g)q(\mathbf{y}_i | z = g) = \pi_g \mathcal{N}(\mathbf{y}_i | \mathbf{m}_g, \mathbf{C}_g) \quad (17)$$

and the Yang's

$$p(z = g, \mathbf{y}_i) = p(\mathbf{y}_i)p(z = g | \mathbf{y}_i) = p(\mathbf{y}_i) \frac{\pi_g \mathcal{N}(\mathbf{y}_i | \mathbf{m}_g, \mathbf{C}_g)}{\sum_{s=1}^G \pi_s \mathcal{N}(\mathbf{y}_i | \mathbf{m}_s, \mathbf{C}_s)}. \quad (18)$$

Then, the corresponding function is

$$J(\Theta) = \frac{1}{I} \sum_{i=1}^I \sum_{g=1}^G \frac{\pi_g \mathcal{N}(\mathbf{y}_i | \mathbf{m}_g, \mathbf{C}_g)}{\sum_{s=1}^G \pi_s \mathcal{N}(\mathbf{y}_i | \mathbf{m}_s, \mathbf{C}_s)} \ln(\pi_g \mathcal{N}(\mathbf{y}_i | \mathbf{m}_g, \mathbf{C}_g)). \quad (19)$$

As in the case of GMM, the maximum likelihood estimate of the parameters of GPFR can be obtained by maximizing $J(\Theta)$ through an iterative method.

After the training process, we can determine the classification accuracy of the model, i.e. let

$$z_i = \arg \max_{g \in \{1, 2, \dots, G\}} \frac{\pi_g \mathcal{N}(\mathbf{y}_i | \mathbf{m}_g, \mathbf{C}_g)}{\sum_{s=1}^G \pi_s \mathcal{N}(\mathbf{y}_i | \mathbf{m}_s, \mathbf{C}_s)} \quad (i = 1, 2, \dots, I). \quad (20)$$

The test data GPFR can be used to get a training set and their corresponding labels. The classification accuracy can be determined by this way. Besides, we can predict the test data by calculating their conditional distribution given the labels. The details are referred to [4-8].

4 Experimental Results

In this section, we evaluate the synthetic data and the real-world data to test if the effective effect is considered a statistical model selection algorithm. We compare the GPFR model trained via the proposed algorithm with the GP model, the GPFR model, and the GPFR model trained through the traditional EM algorithm [2, 3] and the SEM algorithm [6].

Since we are mainly concerned with the prediction ability of the GPFR model, the standard evaluation index (RMSE) is chosen as the evaluation metric. It is assumed that there are T test cases and the test data of the t th ($t = 1, 2, \dots, T$) test case are $y_{t1}, y_{t2}, \dots, y_{tM}$, which are the corresponding predicted value $\hat{y}_{t1}, \hat{y}_{t2}, \dots, \hat{y}_{tM}$, respectively. If $\|_{\mathcal{V}}$ that

$$\text{RMSE} = \sqrt{\frac{1}{TM} \sum_{t=1}^T \sum_{m=1}^M (y_{tm} - \hat{y}_{tm})^2}. \quad (21)$$

As a result, a smaller RMSE indicates a better prediction result.

4.1 On Synthetic Datasets

The synthetic data are defined as S_2, S_3, \dots, S_{10} , respectively, where the subscripts represent the number of features. For each feature, we have 20 training cases and 10 test cases for a GP with a linear effect. The effect and parameters of the Gaussian process model to generate the synthetic data are listed in Table 1, where S_l ($l = 2, 3, \dots, 10$) are generated by the l th GP. Each case consists of 100 input, which is a standard distributed in $[-3, 3]$. The 60 input on the left side of the test cases are $k_{\mathcal{V}}$ and the 40 on the right side are used for testing.

First, we demonstrate the effective effect of the constructed GP model through α -effect. A training case is a standard choice for each case S_9 . Figure 1 shows the reconstruction effect of the training case. Figure 1 is composed of 9 sub-graphs, each of which shows a training case, its reconstruction, and their corresponding effect. As can be seen from the graphs, although there are significant differences between training cases and their reconstruction, their corresponding effects are similar, which indicates that the constructed GP model is effective.

When testing is considered algorithm, G is initialized as $l+3$ for S_l . To illustrate the bad effect of a global GPFR compared to the prediction ability, we train the GPFR model consisting of $l-1$ and $l+1$ GPFR compared to the EM algorithm [2, 3], which are denoted as $\mathcal{G}_{\mathcal{V}}(\text{GPFR}(-1))$ and $\mathcal{G}_{\mathcal{V}}(\text{GPFR}(+1))$, respectively. Similarly, the $\mathcal{G}_{\mathcal{V}}(\text{GP}(-1))$ and $\mathcal{G}_{\mathcal{V}}(\text{GP}(+1))$ are denoted as $\mathcal{G}_{\mathcal{V}}(\text{GP}(-1))$ and $\mathcal{G}_{\mathcal{V}}(\text{GP}(+1))$, respectively. Besides, P is set to be 20. Table 2 shows the α -effectual results.

From Table 2, we see that the RMSE of the GPFR model and the $\mathcal{G}_{\mathcal{V}}(\text{GPFR}(-1))$ model are smaller than those of the GP model and the $\mathcal{G}_{\mathcal{V}}(\text{GP}(-1))$ model, respectively, which

Table 1. Mean function and parameter of the Gaussian process used to generate the synthetic data set.

Mean function	θ^T	$\sqrt{r^{-1}}$
x^2	[0.5, 0.5]	0.15
$\left(-4(x+1.5)^2+9\right)1_{\{x<0\}}+\left(4(x-1.5)^2-9\right)1_{\{x\geq 0\}}$	[0.528, 0.4]	0.144
$8 \sin (1.5 x)-1$	[0.556, 0.3]	0.139
$\sin (1.5 x)+2 x-5$	[0.583, 0.2]	0.133
$\sin (4 x)-0.5 x^2-2 x$	[0.611, 0.1]	0.128
$-x^2$	[0.639, 0.1]	0.122
$\left(4(x+1.5)^2-9\right)1_{\{x<0\}}+\left(-4(x-1.5)^2+9\right)1_{\{x\geq 0\}}$	[0.667, 0.2]	0.117
$5 \cos (3 x+2)$	[0.694, 0.3]	0.111
$\cos (1.5 x)-2 x+5$	[0.722, 0.4]	0.106
$\cos (4 x)+0.5 x^2+2 x$	[0.75, 0.5]	0.1

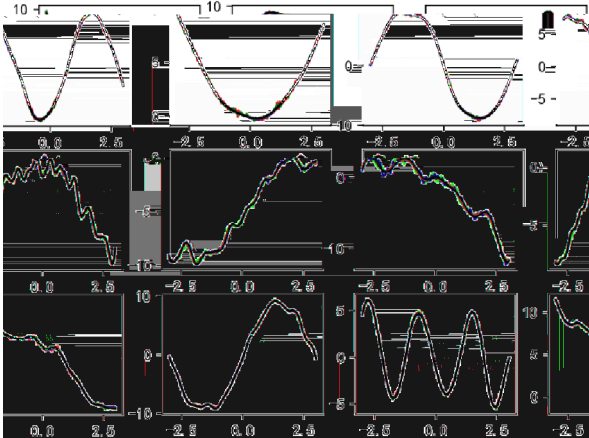


Fig. 1. The result of reconstruction of the synthetic data set. The red, green, blue, and black color represent the original case, the reconstructed case, the GP reconstruction of the original case, and the GP reconstruction of the reconstructed case, respectively.

to evaluate the effectiveness of modeling the mean function as a linear combination of B-spline. Because of the \mathbf{K} -GP (\mathbf{K} -GPFR) model and the GP (GPFR) model, the need for fitting the \mathbf{K} -GP model is not considered. Furthermore, we can see that a \mathbf{K} -GP model based on GPFR can not affect the prediction result badly. For S_2, S_3, \dots, S_9 , both the SEM algorithm and the proposed algorithm find the correct model based on GPFR can not affect the RMSE much. However, the prediction error of the SEM algorithm is higher than that of the proposed algorithm. On the other hand,

the SEM algorithm need to repeat the whole parameter learning process for different values of GPFR coefficient. On the other hand, since different training sets have different distributions, we have to select the best model for each data set. This is the aim of the SEM algorithm. However, the SEM algorithm has a high time complexity. For S_{10} , since the SEM algorithm failed to find the best GPFR coefficient, its RMSE is larger than that for the selected algorithm.

Taking S_9 for example, we select the clustering result for the selected algorithm in Fig. 2, where different colors represent different clusters. On the left and right side of Fig. 2 are the clustering result for the selected algorithm on the training and test data sets, respectively. It is clear that the selected algorithm correctly found all the clusters.

Table 2. RMSE and clustering time for all the methods on the synthetic data sets.

	S_2		S_3		S_4	
	RMSE	Time (s)	RMSE	Time (s)	RMSE	Time (s)
GP	5.5831	6.87	4.7878	9.88	4.7798	13.09
$\hat{\mathbf{K}}$ -GP (-1)	5.5239	6.12	4.6125	8.90	4.3580	15.84
$\hat{\mathbf{K}}$ -GP (+1)	4.8240	7.39	4.6035	17.40	4.3488	23.31
GPFR	5.0759	6.68	4.6864	12.42	4.3051	17.72
$\hat{\mathbf{K}}$ -GPFR (-1)	5.0214	8.07	0.9416	15.95	0.9510	18.93
$\hat{\mathbf{K}}$ -GPFR (+1)	1.6846	14.20	1.0680	22.66	0.9319	25.79
$\hat{\mathbf{K}}$ -GPFR (SEM)	0.4312	20.63	0.4856	41.59	0.5469	58.97
$\hat{\mathbf{K}}$ -GPFR (BYY)	0.4401	9.46	0.4746	15.03	0.5403	18.64
	S_5		S_6		S_7	
	RMSE	Time (s)	RMSE	Time (s)	RMSE	Time (s)
GP	4.9213	17.85	4.9897	15.07	5.3096	20.47
$\hat{\mathbf{K}}$ -GP (-1)	4.4775	15.03	4.3834	20.14	4.3025	30.66
$\hat{\mathbf{K}}$ -GP (+1)	4.5205	28.59	4.3813	29.19	4.3082	30.53
GPFR	4.8079	26.73	4.8649	24.25	4.9871	21.45
$\hat{\mathbf{K}}$ -GPFR (-1)	0.8756	31.66	1.0776	29.67	1.3295	32.09
$\hat{\mathbf{K}}$ -GPFR (+1)	0.9252	30.58	1.0270	35.37	1.0281	38.44
$\hat{\mathbf{K}}$ -GPFR (SEM)	0.5638	81.34	0.6057	87.52	0.6540	92.78
$\hat{\mathbf{K}}$ -GPFR (BYY)	0.5573	25.49	0.6137	23.66	0.6571	27.82
	S_8		S_9		S_{10}	
	RMSE	Time (s)	RMSE	Time (s)	RMSE	Time (s)
GP	4.8180	19.67	4.4758	17.82	4.8438	21.67
$\hat{\mathbf{K}}$ -GP (-1)	4.4904	24.13	4.1223	32.36	4.5730	32.78

(continued)

Table 2. (continued)

	\mathcal{S}_2		\mathcal{S}_3		\mathcal{S}_4	
	RMSE	Time (s)	RMSE	Time (s)	RMSE	Time (s)
$\mathbf{\hat{x}}_k$ -GP (+1)	4.4818	23.70	4.1214	30.95	4.5878	28.14
GPFR	4.6871	26.73	4.3686	21.67	4.6865	20.97
$\mathbf{\hat{x}}_k$ -GPFR (-1)	1.5325	33.78	1.0585	40.27	1.5279	41.56
$\mathbf{\hat{x}}_k$ -GPFR (+1)	1.1891	35.96	0.9789	39.38	1.0343	49.78
$\mathbf{\hat{x}}_k$ -GPFR (SEM)	0.6448	99.49	0.6233	116.85	1.4379	130.65
$\mathbf{\hat{x}}_k$ -GPFR (BYY)	0.6421	28.91	0.6199	28.62	0.6317	32.46

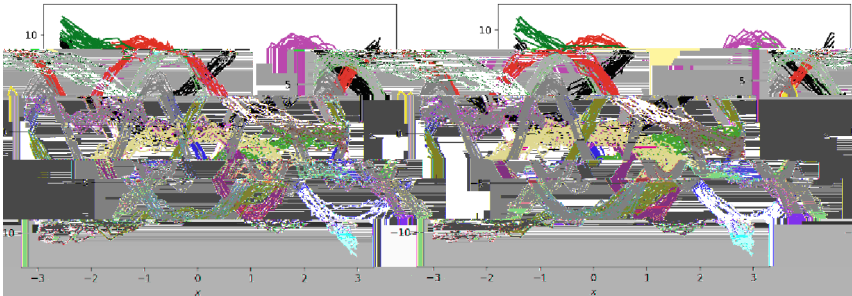


Fig. 2. Clustering results for synthetic data sets \mathcal{S}_7 and \mathcal{S}_9 .

4.2 On Real-World Datasets

Here, we utilize the electricity load data collected by the North China Grid Company [8], which records electricity load every 15 min in 2009 and 2010. Hence, daily electricity load can be regarded as a sequence with 96 points. We divide the data into training and testing data according to the year, which are respectively \mathcal{R}_1 and \mathcal{R}_2 , sequentially. Each sub-dataset consists of 200 training sequences and 165 test sequences. Moreover, the 56 points in the left side of test sequences are known and the 40 points in the right side are used for testing.

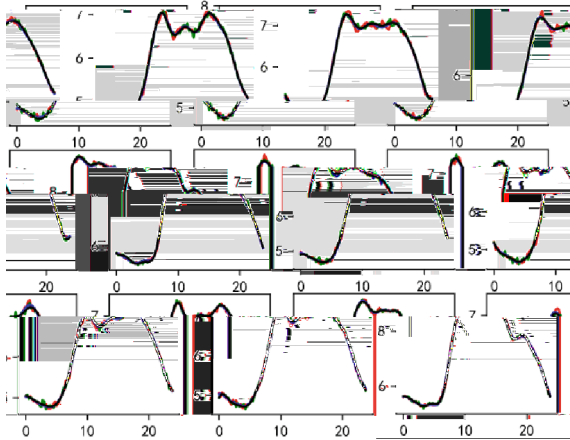


Fig. 3. The result of forecasting the electricity load data set. The red, green, blue, and black color represent the original set, the reconstructed set, the test set, and the forecasted set, respectively.

Although all the cases have the same result, we can see that the data has the same pattern. Like the synthetic data set, we can find the 9 training cases of \mathcal{R}_1 , which forecasting cases are presented in Fig. 3. A can be seen from the figure, the forecasting results based on GP model are effective for the electricity load data set.

Since the number of cases in \mathcal{R}_1 and \mathcal{R}_2 are less than G , we set $G = 3, 6, 9, 12, 15$ for the \mathbf{x} -GP and \mathbf{x} -GPFR model training using the EM algorithm. For each case, we used the SEM algorithm, G is set to be 15. Besides, P is set to be 30. The experimental results are described in Table 3. For \mathcal{R}_1 and \mathcal{R}_2 , the RMSE of the SEM algorithm is smaller than that of the SEM algorithm in the case of forecasting the data set. On the left and right side of Fig. 4 are the clustering results of the SEM algorithm for the training and test data set, respectively. For \mathcal{R}_1 and \mathcal{R}_2 , the number of clusters of the SEM algorithm are 13 and 11, respectively. As can be seen from Fig. 4, the clustering results of the SEM algorithm are different from the clustering results of the GP algorithm, that is to say, the clustering results of the GP algorithm are reasonable.

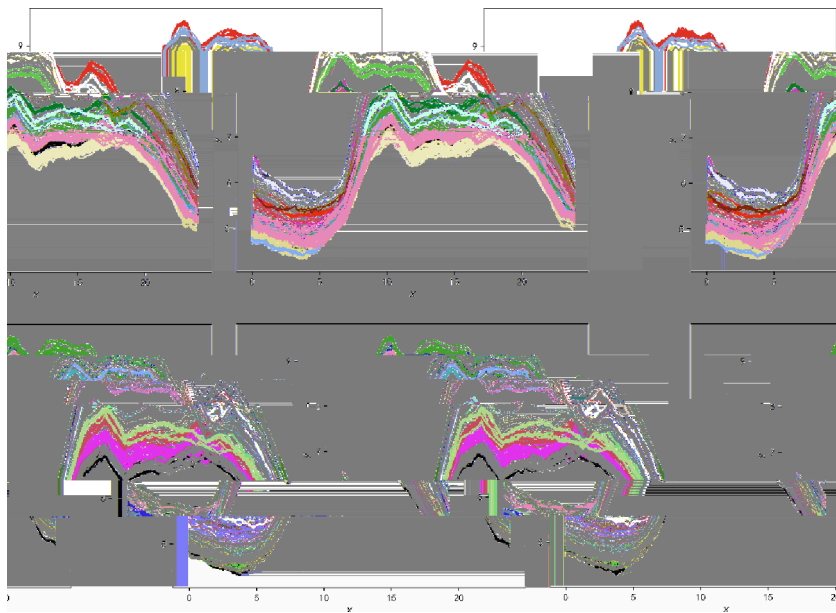


Fig. 4. Classification results for selected automatic selection algorithm \mathcal{R}_1 and \mathcal{R}_2 .

Table 3. RMSE and fitting error for all the methods \mathcal{R}_1 and \mathcal{R}_2 .

	\mathcal{R}_1		\mathcal{R}_2	
	RMSE	Fitting error (%)	RMSE	Fitting error (%)
GP	0.9599	19.43	0.8977	20.39
\mathbf{k} -GP (3)	0.9390	20.33	0.8846	21.42
\mathbf{k} -GP (6)	0.9387	22.54	0.8854	22.66
\mathbf{k} -GP (9)	0.9380	25.99	0.8853	26.09
\mathbf{k} -GP (12)	0.9395	29.83	0.8847	31.23
\mathbf{k} -GP (15)	0.9401	34.76	0.8872	36.91
GPFR	0.5584	21.30	0.5499	21.59
\mathbf{k} -GPFR (3)	0.2089	24.45	0.2133	20.64
\mathbf{k} -GPFR (6)	0.1701	25.76	0.1731	24.77
\mathbf{k} -GPFR (9)	0.1356	28.93	0.1455	29.45
\mathbf{k} -GPFR (12)	0.1248	34.65	0.1314	33.63
\mathbf{k} -GPFR (15)	0.1178	35.88	0.1301	36.78
\mathbf{k} -GPFR (SEM)	0.1323	150.76	0.1377	170.17
\mathbf{k} -GPFR (BYY)	0.1109	33.97	0.1201	34.58

14. Cheng, G., Li, L., Ma, J.: A gradient BYY harmonic learning algorithm for straight line detection. In: Song, F., Zhang, J., Tang, Y., Cao, J., Yang, W. (ed.) ISSN 2008. LNCS, vol. 5263, pp. 618–626. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-87732-5_69
15. Li, L., Ma, J.: A BYY calculus-based EM algorithm for Gaussian mixture learning. *Appl. Math. Comput.* **205**(2), 832–840 (2008)
16. Li, L., Ma, J.: A BYY likelihood-based EM algorithm for Gaussian mixture learning. In: Song, F., Zhang, J., Tang, Y., Cao, J., Yang, W. (ed.) ISSN 2008. LNCS, vol. 5263, pp. 600–609. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-87732-5_67
17. Ma, J., Wang, T., Xu, L.: A gradient BYY harmonic learning-based Gaussian mixture learning algorithm for the detection of the selected model. *Neurocomputing* **56**, 481–487 (2004)
18. Ma, J., Gao, B., Wang, Y., Cheng, Q.: Comparison of the algorithm for the BYY harmonic learning-based Gaussian mixture learning algorithm for the detection of the selected model. *Int. J. Pattern Recogn. Artif. Intell.* **19**(5), 701–713 (2005)
19. Ma, J., Li, J.: The BYY algorithm for Gaussian mixture learning algorithm for the detection of the selected model. *Pattern Recogn.* **40**(7), 2029–2037 (2007)
20. Ma, J., He, X.: A fast reduced BYY harmonic learning algorithm for Gaussian mixture learning algorithm for the detection of the selected model. *Pattern Recogn. Lett.* **29**(6), 701–711 (2008)
21. Bucci, C.D.: On calculating the binary J. *Acoust. Soc. Am.* **6**, 50–62 (1972)