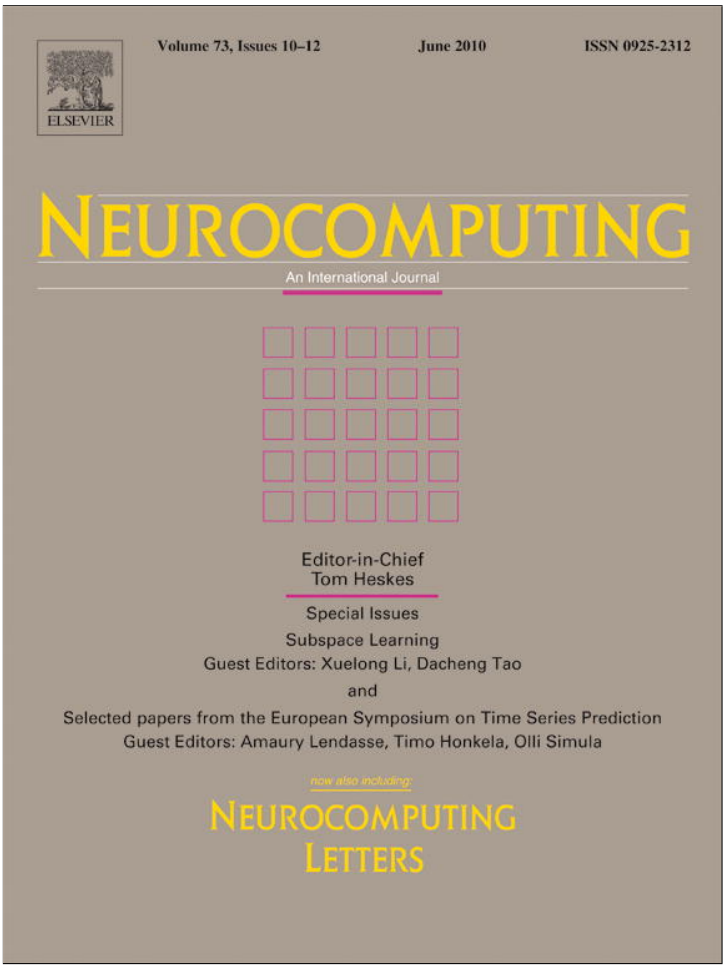


Neurocomputing 73 (2010) 2186–2195

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Semi-supervised Gaussian process latent variable model with pairwise constraints

Xiumei Wang [a,b], Xinbo Gao [b,*], Yuan Yuan [c], Dacheng Tao [d], Jie Li [b]

[a] School of Sciences, Xidian University, Xi'an 710071, China
[b] School of Electronic Engineering, Xidian University, Xi'an 710071, China
[c] School of Engineering and Applied Science, Aston University, Birmingham B4 7ET, United Kingdom
[d] School of Computer Engineering, Nanyang Technological University, 50 Nanyang Avenue, Blk N4, 639798, Singapore

ARTICLE INFO

Article history:
Received 30 April 2009
Received in revised form
18 January 2010
Accepted 30 January 2010
Communicated by J. Zhang
Available online 25 March 2010

Keywords:
Dimensionality reduction
Gaussian process latent variable model
Pairwise constraints
Semi-supervised learning

ABSTRACT

In machine learning, Gaussian process latent variable model (GP-LVM) has been extensively applied in the field of unsupervised dimensionality reduction. When some supervised information, e.g., pairwise constraints or labels of the data, is available, the traditional GP-LVM cannot directly utilize such supervised information to improve the performance of dimensionality reduction. In this case, it is necessary to modify the traditional GP-LVM to make it capable of handing the supervised or semi-supervised learning tasks. For this purpose, we propose a new semi-supervised GP-LVM framework under the pairwise constraints. Through transferring the pairwise constraints in the observed space to the latent space, the constrained priori information on the latent variables can be obtained. Under this constrained priori, the latent variables are optimized by the maximum a posteriori (MAP) algorithm. The effectiveness of the proposed algorithm is demonstrated with experiments on a variety of data sets.

1. Introduction

In machine learning and statistical pattern recognition, dimensionality reduction is of great importance and has been extensively studied. Existing methods for dimensionality reduction can be divided into two general categories, i.e., determined framework and probabilistic framework.

The determined methods can be further divided into two classes: linear and nonlinear methods. The linear methods, e.g., principal component analysis (PCA) [1,2] and multidimensional scaling (MDS) [3], try to seek a set of optimal bases for the subspace selection. However, they cannot catch the curvature and nonlinear structures embedded in the observed data. The nonlinear methods, e.g., the kernel extensions of PCA (kernel PCA) and MDS cannot adaptively deal with the different real data sets due to the fixed model parameters [4], and the cost functions in these extensions put great impact on the results of dimensionality reduction, and are hard to determine. These shortcomings will lead to some negative effect for their applications [5–7].

The probabilistic methods can also be divided into linear and nonlinear methods. Linear methods, e.g., probabilistic principal component analysis (PPCA) [8] and factor analysis (FA) [9,10], are conducted to tackle the problems of the small/sparse training sample sets. However, PPCA and FA are linear latent variable models and cannot catch nonlinear structure for its linear property. To this end, a nonlinear probabilistic model, the Gaussian process latent variable model (GP-LVM), is proposed recently to offer a projection from the latent space to the observed space [11,12].

The GP-LVM is a dual probabilistic interpretation to PPCA. It establishes a nonlinear mapping from the latent variable space, i.e., low-dimensional space, to the observed space by giving a Gaussian process prior to the mapping function [13]. A Gaussian process can be used as a priori probability distribution over functions in Bayesian inference and it has been widely used in machine learning for regression [14] and classification tasks [15]. Through modeling the joint probability density of the observed data, the GP-LVM can obtain low-dimensional manifolds with a little number of samples. However, the GP-LVM is an unsupervised learning method, which does not utilize the available supervised information for learning, thus it cannot capture the structure in the observed set preferably.

In general, there are two kinds of supervised information, that is, label sets and pairwise constraints. The former can provide the category label for each sample in the training set, while the latter only offers the constrained relationship for two samples, i.e., indicating whether two samples belong to the same class or not.

* Corresponding author. Tel.: +86 2988201838; fax: +86 2988201620.
E-mail addresses: wangxiumei@gmail.com (X. Wang), xbgao@mail.xidian.edu.cn (X. Gao), yuany1@aston.ac.uk (Y. Yuan), dctao@ntu.edu.sg (D. Tao), leejie@mail.xidian.edu.cn (J. Li).

0925-2312/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2010.01.021

So we can reach the conclusion that pairwise constraints are much weaker and more general priori information than the label information. When a great deal of labeled samples are available, supervised learning methods can work very well [16,17]. However, labeled data is often limited, and labeling samples requires much human expertise, so it is an expensive job to obtain supervised information. While, the pairwise constraints are easier to be obtained than the label information, so pairwise constraints have been widely used in the dimensionality reduction methods, such as in PCA [18,19], Fisher linear discriminant analysis (FLDA) [20] and locality preserving projection (LPP) [21,22]. Since pairwise constraints cannot offer the detail label sets, especially for multi-class tasks, they belong to semi-supervised information.

In this paper, we propose a new semi-supervised Gaussian process latent variable model which utilizes some supervised information as priori information, i.e., pairwise constraints mentioned above, for dimensionality reduction. Since the pairwise constraints are traditionally defined in the original observed space rather than in the latent variable space, we have to first transfer the pairwise constraints from the observed space to the latent space. Then we can obtain the priori information of the latent variables according to the transferred pairwise constraints. At the same time, the likelihood of the observed data can be calculated through the typical GP-LVM. So the posteriori probability of the latent variables will be obtained based on the Bayes theorem. Finally, the latent variables can be optimized through MAP algorithm.

The rest of this paper is organized as follows. In Section 2, some previous work is summarized, such as GP-LVM and the pairwise constraints. Section 3 describes how to establish the semi-supervised GP-LVM and its realization algorithm in detail. The experimental results and analysis are given in Section 4 to show the improved performance of the proposed method. The final section offers our conclusion.

2. Background

In this section we review some previous work on GP-LVM. More formally, let $Y=[y_1,\ldots,y_N]^T$ be the matrix denoting N observed examples, i.e., the high-dimensional data set to be processed. Each object y_i is described by a D-dimensional feature vector with $y_i \in R^D$. We use $X=[x_1,\ldots,x_N]^T$ to denote the low-dimensional set with x_i representing positions in latent space of the corresponding high-dimensional point, $x_i \in R^d$, $d<D$.

2.1. Gaussian process latent variable model

The GP-LVM is the dual representation of the PPCA. PPCA supposes that all variables are drawn from the same Gaussian distribution independently. PPCA determines the principal axes of a set of observed data vectors by maximum-likelihood estimation of projection. The dual representation of PPCA supposes that the linear mapping vectors in the projected matrix are independent identity distribution. By imposing a nonlinear Gaussian process priori to the transform f in each mapping direction, the GP-LVM can be established [13]. The positions of the data in the latent space can be obtained through integrating over f and maximizing likelihood function of the observed data set. The detail description will be given as follows.

The mapping function $f: X \to Y$ is a Gaussian process priori given by

$$f \sim N(0,K) \tag{1}$$

with the covariance between x_i and x_j, and the kernel function value is determined by a Mercer kernel function, for example, the radius basis function (RBF) kernel. The RBF kernel is employed as the nonlinear mapping function, which can be substituted with

$$k(x_i,x_j)=\theta_{rbf}\exp\left(-\frac{\gamma}{2}(x_i-x_j)^T(x_i-x_j)\right)+\theta_{white}\delta_{ij}, \tag{2}$$

where $k(x_i, x_j)$ is the element in the i-th row and the j-th column of the covariance matrix K and $\delta_{i,j}$ is the Kronecker delta function. $\theta=[\theta_{rbf}, \gamma, \theta_{white}]$ is a collector of the kernel parameters.

In the GP-LVM, a Gaussian process prior is imposed on the mapping function f_d in each dimension,

$$p(f)=\prod_{d=1}^{D} p(f_d)=\prod_{d=1}^{D} N(f_d|0,K), \tag{3}$$

where D is the dimension of the observed data and K can be computed through the RBF kernel defined in Eq. (2).The Gaussian processes are natural generalizations of multivariate Gaussian random variables to infinite index sets. It provides a promising non-parametric Bayesian approach to metric regression [23] and classification problems [24]. A Gaussian process priori over a function defines a flexible probabilistic distribution. Then the likelihood for every dimension can be obtained through marginalizing the mapping function.

$$p(y_{:,d}|X,\theta)=\int p(y_{:,d}|X,f_d,\theta)p(f_d)df_d=N(y_{:,d}|0,K). \tag{4}$$

The likelihood for the whole observed data can be viewed as a product of D number of independent Gaussian processes, and each process is related to a different dimension of the data set. So the observed data likelihood function can be obtained as follows:

$$P(Y|X,\theta)=\frac{1}{(2\pi)^{DN/2}|K|^{D/2}}\exp\left(-\frac{1}{2}tr(K^{-1}YY^T)\right). \tag{5}$$

The framework of the GP-LVM includes two modules: (a) initialization of the latent variables X and hyper-parameter θ; (b) optimizing algorithm with scale conjugate gradient (SCG) method. In the first module, the latent variables X are initialized with PCA, where the number of the projected vectors remaining in PCA is consistent with the dimension of the latent variable X. The RBF kernel hyper-parameters are initialized as $\theta=[1,1,1]$. The second module can be divided into two steps. In the first step, the hyper-parameters will be obtained through maximizing the likelihood which can be calculated by Eq. (5), that is to say, the process to obtain the optimal hyper-parameter is realized through maximizing the likelihood by SCG method. In the second step, the likelihood is updated with the new hyper-parameter, and then the latent variable can be obtained through maximizing the likelihood by SCG method. These two steps are executed iteratively in the optimizing process until convergence. The algorithmic flowchart of the GP-LVM is shown in Fig. 1.

2.2. Pairwise constraints

Unlike the class labels, the pairwise constraints do not give the labels of the training samples. This kind of constraints only offers the pairs information. That is, two samples belong to the same class or different classes. More specifically, we consider the following two types of pairwise constraints: *must-link* constraints and *cannot-link* constraints as follows:

- ***Must-link* constraints:** constraints specify that two samples should be assigned into one class. The constraints data set can be noted as
  $M=\{(y_i,y_j)|\ y_i$ and y_j belong to the same class$\}$.
- ***Cannot-link* constraints:** constraints specify that two samples should be assigned into different classes. The constraints data set can be denoted as
  $C=\{(y_i,y_j)|\ y_i$ and y_j belong to the different classes$\}$.

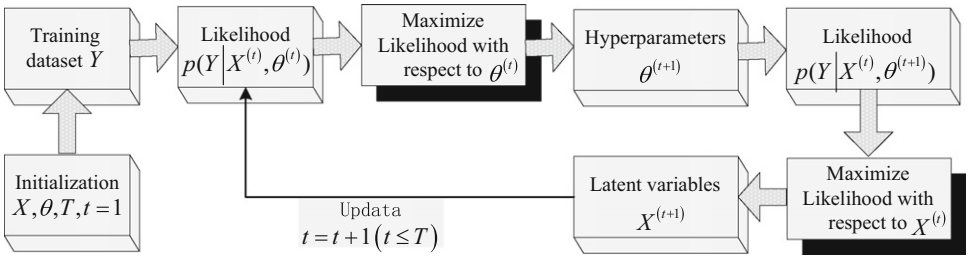


Fig. 1. The flowchart of the GP-LVM.

In this paper, we try to find a way to introduce the pairwise constraints into the probabilistic latent variable model. As well known, this kind of constraints is defined to the observed data, so the problem we have to deal with is how to introduce the constrained knowledge into the latent space and then use this knowledge to optimize the model. In the following section, we will focus on this problem and point out a way for transferring pairwise constraints to the latent space, and then utilize this knowledge to obtain some priori information for the latent variables.

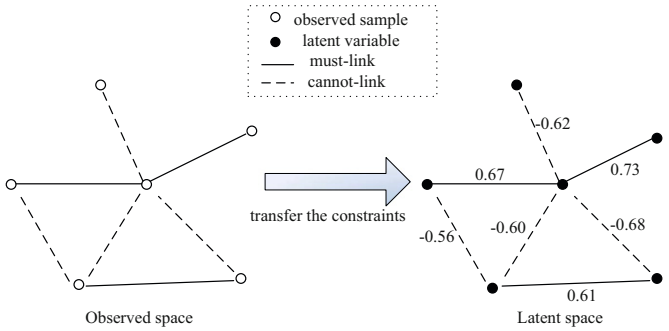


Fig. 2. The transfer scheme of the pairwise constraints.

3. Semi-supervised Gaussian process latent variable model

To resolve the aforementioned problem, a new semi-supervised learning framework for GP-LVM based on pairwise constraints will be established in this section. Since this paper mainly studies the embedding method of pairwise constraints in the latent variable model, this section puts the emphasis on how to transfer the constraints on the observed examples to the latent variables, and how to use this constrained information in the latent space. This section presents the process to establish the semi-supervised GP-LVM utilizing the constrained information.

3.1. Pairwise constraints in the latent variable model

As well known, the pairwise constraints are traditionally defined in the original observed space rather than in the latent variable space, we will give the steps on transferring this constrained information to the latent variables.

If the pair of samples $(y_i, y_j)\in M$, the latent variables (x_i, x_j) corresponding to (y_i, y_j) will belong to the same class. In the same way, if the pair of samples $(y_i, y_j)\in C$, the latent variables (x_i, x_j) corresponding to (y_i, y_j) will belong to the different classes. Then according to the relationship among the observed samples, we can infer the priori information of the latent variables. Besides the relationship of the pairwise constraints, the distances of the observed samples are also an important character in describing the whole dataset. Therefore a weight matrix should be obtained with considering not only pairwise constraints but also the distances of the samples. We define the weight matrix which offers a positive weight if the sample pair belongs to the same class and a negative weight if the sample pair belongs to different classes. At the same time, a large value will be assigned to the weight if the distance between two samples is small, otherwise a small value will be assigned. We define a weight matrix $W\in R^{N\times N}$ as

$$W_{i,j}=\begin{cases}\dfrac{e^t}{1+e^t}, & (y_i,y_j)\in M\\ -\dfrac{e^t}{1+e^t}, & (y_i,y_j)\in C\\ 0, & \text{otherwise}\end{cases} \tag{6}$$

where $t=\|x_i, x_j\|$ represents the Euclidean distance between two latent variables x_i and x_j. The value $W_{i,j}$ will be determined by both the distance and the pairwise constraints. If the two samples belong to M, i.e., the same class, the weight value is positive. If they belong to C, the weight value is negative. The weight values will change with the distance among the latent variables as shown in Fig. 2.

As shown in Fig. 2, real line represents the *must-link* relationship in two samples, and dashed line represents *cannot-link* relationship. The values will be positive if the sample pair belongs to same class and negative if the sample pair belongs to different classes. The values are also influenced by the distances of samples.

The priori probability of the latent variables can be defined as

$$P(X)=\frac{1}{Z}\exp\left(-\sum_{i,j=1}^{N}d(x_i,x_j)\right), \tag{7}$$

where $d(x_i, x_j)=W_{i,j}\|(x_i, x_j)\|$ and Z is a constant. So the Eq. (7) can be rewritten as

$$P(X)=\frac{1}{Z}\exp\left(-\sum_{i,j=1}^{N}d(x_i,x_j)\right)=\frac{1}{Z}\exp\left(-tr(X^TWX)\right)$$
$$=\frac{1}{Z}\exp\left(-tr(WXX^T)\right). \tag{8}$$

where W represents a weight matrix. Just as mentioned above, the weight matrix is defined according to the pairwise constraints and the distances of the samples. Then the weight matrix W can be obtained by Eq. (6).

3.2. The semi-supervised GP-LVM

Given the constrained priori information of the latent variables in Eq. (8), the detail description of the semi-supervised framework will be given. The GP-LVM is a latent variable model through defining a joint distribution over the observed variables Y and the latent variables X.The hyper-parameters and the latent variables

can be optimized through maximizing the likelihood function as Eq. (5).

$$P(Y|X,\theta)=\prod_{i=1}^{D}\frac{1}{(2\pi)^{N/2}|K|^{1/2}}\exp\left(-\frac{1}{2}y_{:,d}^{T}K^{-1}y_{:,d}\right). \tag{9}$$

According to the Bayes theorem

$$p(X|Y,\theta)=\frac{p(Y|X,\theta)p(X)}{p(Y)}. \tag{10}$$

It can be rewritten as

$$p(X|Y,\theta)\propto p(Y|X,\theta)p(X). \tag{11}$$

We have obtained the priori information of the latent variables X according to the Eq. (8). The posterior of X can be calculated based on the Bayes theorem. The maximum of posterior is equal to maximizing the right hand of Eq. (11). The log posterior is given by

$$L=\ln\, p(X|Y,\theta)=\ln p(Y|X,\theta)+\ln p(X). \tag{12}$$

The first part in the right hand of Eq. (12) is the same as log-likelihood of the GP-LVM,

$$\ln p(Y|X,\theta)=-\frac{DN}{2}\ln 2\pi-\frac{D}{2}\ln|K|-\frac{1}{2}tr(K^{-1}YY^{T}). \tag{13}$$

The second part in the right hand of Eq. (12) is a priori information for latent variables based on pairwise constraints.

$$\ln p(X)=-tr(WXX^{T})+\text{constant} \tag{14}$$

Therefore, finding the maximum posterior configuration of the GP-LVM equivalent to maximizing the posterior probability,

$$L=-\frac{D}{2}\ln|K|-\frac{1}{2}tr(K^{-1}YY^{T}-2WXX^{T})+\text{constant}. \tag{15}$$

In the above process, we substitute the priori with the constrained knowledge in Eq. (8), and then the framework of semi-supervised GP-LVM is established based on pairwise constraints.

In the following experiments, we apply the scaled conjugate gradient with regard to latent variables X and hyper-parameters for training. The detailed steps for training process will be given in Table 1. Firstly the latent variables will be initialized through PCA and all the values in the hyper-parameters are set to 1. Then the latent variables and the hyper-parameters will be optimized alternatively. Finally the semi-supervised GP-LVM can be established with the hyper-parameters obtained from the training process. The whole and detailed procedures for semi-supervised GP-LVM are shown in Table 1.

Table 1
An algorithm for learning semi- supervised GP-LVM.

Input: The high-dimensional data $Y\in R^{N\times D}$, and set of *must-link* constraints $M=(y_i, y_j)$, the set of *cannot-link* constraints $C=(y_i, y_j)$, the maximum iteration of training T.
Initialization: the latent variables $X\in R^{N\times d}$ through PCA, the hyper-parameters $\Theta=[1,1,1]$.
For $t=1$ to T

1. Calculate the kernel matrix $K_{\Theta}^{(t-1)}=K(X^{(t-1)},\Theta^{(t-1)})$ and the weight matrix $W^{(t-1)}$ according to the pairwise constraints M and C;
2. Calculate the log-likelihood:
$$L^{(t-1)}=-\frac{DN}{2}\ln 2\pi-\frac{D}{2}\ln|K^{(t-1)}|-\frac{1}{2}tr(WXX^{T});$$
3. Optimize the hyper-parameters $\Theta^{(t)}=\underset{\Theta}{\operatorname{argmin}}\{-L^{(t-1)}\}$ using scale conjugate gradient method;
4. Update the kernel matrix $K^{(t-1)}=K(X^{(t-1)}, \Theta^{t})$;
5. Calculate the log-likelihood
$$L^{(t-1)}=-\frac{DN}{2}\ln 2\pi-\frac{D}{2}\ln|K^{(t-1)}|-\frac{1}{2}tr(WXX^{T});$$
6. Optimize the latent variables $X^{(t)}=\underset{X}{\operatorname{argmin}}\{-L^{(t-1)}\}$ using scale conjugate gradient method;
Check convergence: the training stage of semi-supervised GP-LVM converges if
$\text{Error}(t)=\sum_{i=1}^{N}\|x_i^{t}-x_i^{t-1}\|^{2}\le\varepsilon$
7. Set $t\leftarrow t+1$, go back to 2, until convergence.

Output: the hyper-parameters Θ and X.

Table 1 summarizes the learning algorithm of the proposed semi-supervised GP-LVM. It reaches the optimal solution by the alternative iteration between the hyper-parameters Θ and latent variables X. After that, for the test point y^{*}, we can follow the GP-LVM algorithm to obtain its position x^{*} in the latent space.

We assume that the $p(y^{*}|x^{*})$ satisfies a Gaussian distribution, i.e.,

$$p(y^{*}|x^{*})=N(y^{*}|\mu^{*},\sigma_{*}^{2}I). \tag{16}$$

The mean μ^{*} and the variance σ_{*}^{2} can be represented respectively,

$$\begin{cases}\mu^{*}=Y^{T}K^{-1}k\\ \sigma_{*}^{2}=k(x^{*},x^{*})-k^{T}K^{-1}k\end{cases} \tag{17}$$

where K denotes the kernel matrix developed from the training set, and k is the vector of covariance between the test point y^{*} and N training points. The points x^{*} can be optimized by maximizing Eq. (16) with gradients-descent method.

3.3. Discussion and remarks on the semi-supervised GP-LVM

Eq. (15) suggests a general framework for incorporating constraints into the GP-LVM. Particular choices of the pairwise constraints would construct the different weight matrix and produce corresponding algorithms. That is, if the data set can be divided into three parts: *must-link* constrained data, *cannot-link* constrained data and *unlabeled data*, then the semi-supervised GP-LVM can be built in three ways as follows:

- SSGP-LVM-M: Only the *must-link* constraints is used in the model, then the weight matrix will be shown as,
$$W_{i,j}=\begin{cases}\dfrac{e^{t}}{1+e^{t}}, & (y_i,y_j)\in M\\ 0, & \text{otherwise}\end{cases} \tag{18}$$
- SSGP-LVM-CM: Both the *must-link* and *cannot-link* constraints are used in the model, then the weight matrix can be shown as,
$$W_{i,j}=\begin{cases}\dfrac{e^{t}}{1+e^{t}}, & (y_i,y_j)\in M\\ -\dfrac{e^{t}}{1+e^{t}}, & (y_i,y_j)\in C\\ 0, & \text{otherwise}\end{cases} \tag{19}$$
- SSGP-LVM-CMU: Both the constrained points and the unlabeled points are used for training process, the weight matrix will be
$$W_{i,j}=\begin{cases}\dfrac{e^{t}}{1+e^{t}}, & (y_i,y_j)\in M\\ -\dfrac{e^{t}}{1+e^{t}}, & (y_i,y_j)\in C\\ \dfrac{e^{t}}{N^{2}(1+e^{t})}, & \text{otherwise}\end{cases} \tag{20}$$

It is the same with Eq. (17). N is the total number of observed samples.

We can choose one of the weight matrix frameworks according to what we need in the experiments. If we only want to use the *must-link* constraints, we will choose the Eq. (18) to generate the weight matrix. When both of the *must-link* constraints and *cannot-link* constraints are required, we should use Eq. (19).

And the Eq. (20) should be demanded if both the constrained data and unlabeled data are chosen in the training algorithm.

4. Experiments and analysis

In this section, to validate the effectiveness of the proposed semi-supervised dimensionality reduction method, we conduct several experiments on some benchmark data sets, including *USPS handwritten digits*, *oil* data set, *ORL* face database, *YaleB* face database and five UCI data sets, i.e., *iris, wine, balance, sonar, letter*. The experiments consist of four parts. In the first part, we will visualize the latent variables in 2-D space, the *USPS Handwritten digits* data sets are used to verify the effectiveness of the proposed method. Then in the second part, we study the classification accuracy influenced by the constrained percents on the *oil* data set and three UCI data sets (*balance, iris and sonar*). In the third part, we address the classification accuracy changing with the dimension of the latent space, in which the *oil* data set, *USPS* data sets and *ORL* face data will be used. In the last part, the superiority of the proposed method is demonstrated by comparing with other semi-supervised method which is based on the pairwise constraints.

4.1. Data sets

We test the proposed method on a broad range of data sets, including *USPS handwritten digits* [11], *oil* data set [11], *ORL face* database [25], *YaleB face* database [25] and five UCI data sets[26], *i.e. iris, wine, balance, sonar, letter*. The detailed description on data sets is given in Table 2.

The first data set is *handwritten digits* which play an important role in pattern classification, and it is popular for testifying the performance of some algorithms in data visualization. The second data set is the multi-phase oil flow data. *Oil* data set has three classes with 1000 samples which have input dimensionality of 12. There are three phases (classes) of flow associated with the data: *stratified*, *annular* and *homogenous*. The following five data sets are all from the UCI data sets, and their attributes are not very high. The *ORL* face database and *YaleB* face database has much greater dimensionality than the data sets mentioned above. The *ORL* face database containing 400 images of 40 individuals is selected as test-bed. For each of individual, there are ten different images taken at different times, varying with the lighting and facial expressions, as shown in Fig. 3. The *YaleB* face database contains a total of 640 images of 10 individuals. The size of each cropped image for both databases is 32×32 pixels.

Table 2
The basic information of the data sets.

Data	Total number	Dimensionality	Class
USPS	7921	256	10
Oil	1000	12	3
Iris	150	4	3
Wine	178	13	3
Sonar	208	60	2
Balance	625	4	3
ORL face	400	1024	40
Letters	3680	16	5
YaleB face	2414	1024	38

4.2. Visualization in 2-D latent space

We use *USPS handwritten digits* data sets to verify the effectiveness of the proposed semi-supervised GP-LVM for data visualization. This database includes ten digits from '0', '1–9' in the 256-dimensional space. In the following experiments, we choose three digits '3', '5' and '8' as shown in Fig. 4.

For each digit, 300 samples are randomly selected. The *must-link* constraints and *cannot-link* constraints have been defined according to their label sets. The four sub-figures in Fig. 5 visualize results of four models: GP-LVM, BC-GP-LVM [27], SSGP-LVM-M and SSGP-LVM-CM. in each subfigure, '3' is represented by red crosses, '5' is denoted as blue circles and '8' is represented by magenta stars.

Fig. 5(a) shows the result of the traditional GP-LVM. Since it is an unsupervised learning method, the traditional GP-LVM does not utilize the label information or pairwise constraints, which leads to the dimensionality reduction result with less discriminative information. Therefore, it can be found in Fig. 5(a) that all the samples of the three digits are overlapped and cannot be distinguished from each other. As a nonlinear dimensionality reduction method, the traditional GP-LVM establishes a smooth mapping from the latent space to the data space, so it will preserve local distances in the latent space. That is, if two samples are adjacent to each other in the latent space, their corresponding samples mapped by the GP-LVM will be also adjacent in the observe space. However if their corresponding samples are

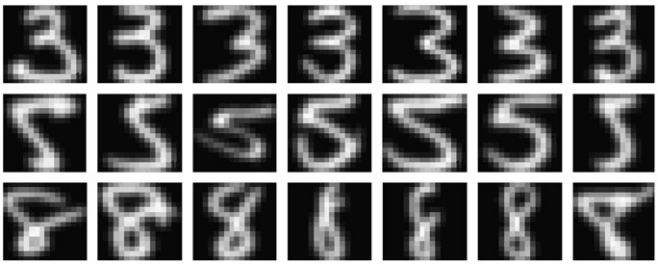


Fig. 4. Sample digit images from the handwritten digits database '3', '5'and '8'.

Fig. 3. Sample face images from the ORL database. For each subject, there are 10 face images with different facial expression and details.

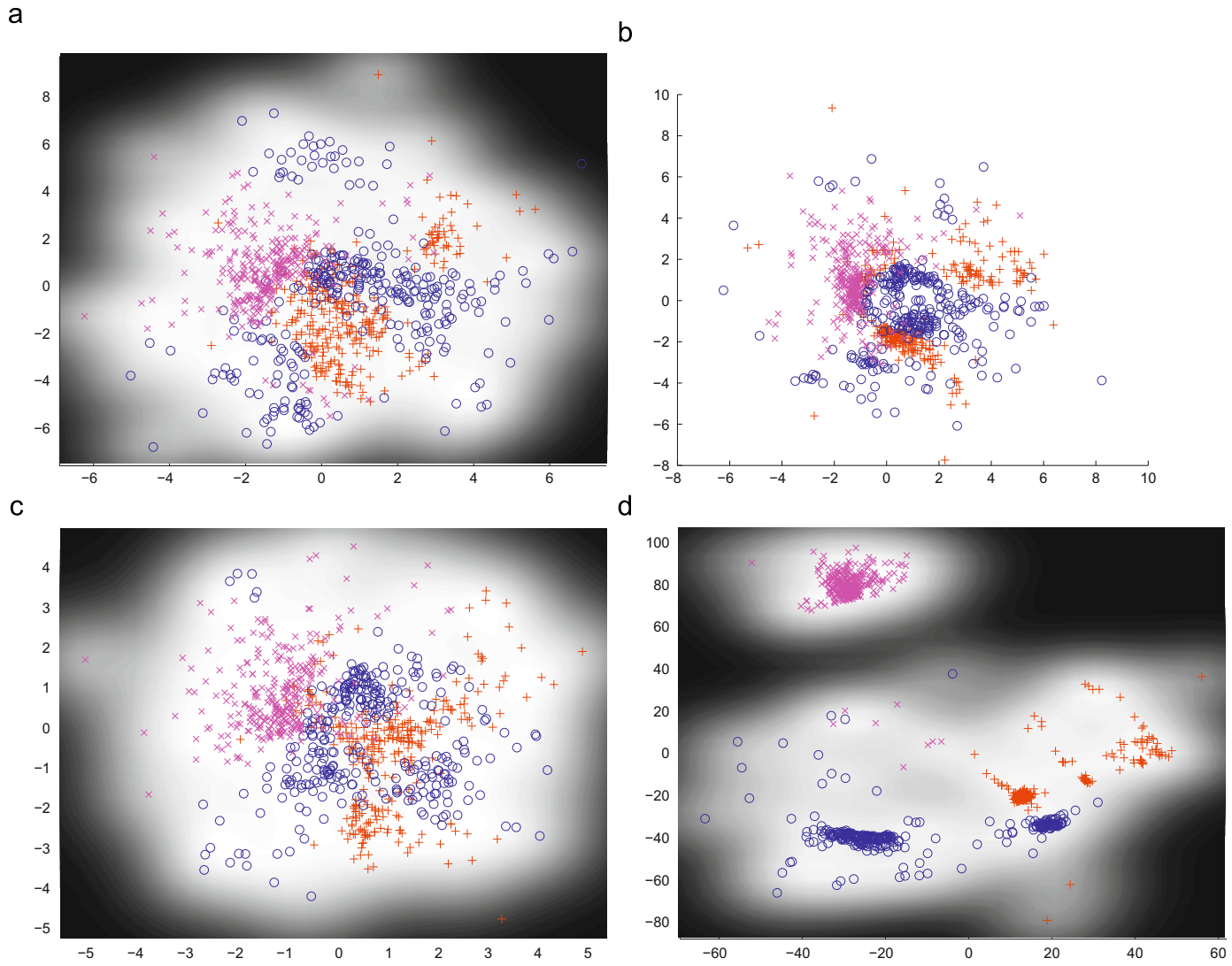


Fig. 5. The digit images visualized with (a) GP-LVM, (b) BC-GP-LVM, (c) SSGP-LVM-M and (d) SSGP-LVM-CM.

adjacent in the observed space, the GP-LVM cannot guarantee these two latent variables adjacent to each other. To this end, a back constrained GP-LVM (BC-GP-LVM) is proposed [27]. By adding distance constraints for pairs' observed samples, the BC-GP-LVM overcomes the shortcoming that the traditional GP-LVM cannot preserve the local neighborhood in the latent space. As shown in Fig. 5(b), the BC-GP-LVM can roughly take three digits '3', '5' and '8' apart. Since the BC-GP-LVM is still an unsupervised learning algorithm, the dimensionality reduction result cannot provide more discriminative information. It can be seen in Fig. 5(b) that some samples of digits '3' (with blue circle) and '5'(with magenta cross) are overlapped each other. Fig. 5(c) is the result of the SSGP-LVM-M in which only the *must-link* constraints have been utilized, and the weight matrix is computed with Eq. (16). Fig. 5(d) shows the result of the SSGP-LVM-CM which utilizes both the *must-link* constraints and the *cannot-link* constraints, and the weight matrix is computed through Eq. (20). It is obvious that the result of the SSGP-LVM-M has less discriminative information than that of the SSGP-LVM-CM. It is because that the SSGP-LVM-M does not use the *cannot-link* constraints. And the Fig. 5(d) reaches more discriminative result than other three methods, in which samples in same class are distributed tightly, and compact manifolds can be obtained for each class, while samples in different classes can be well separated.

4.3. *Classification accuracy influenced by number of constraints*

The proposed method is based on the pairwise constraints, so it may be influenced by the number of constraints. In this subsection, we evaluate the performance of the semi-supervised GP-LVM on the *oil* data set and three UCI data sets. The proposed semi-supervised GP-LVM (including SSGP-LVM-M, SSGP-LVM-CM and SSGP-LVM-CMU) will be compared with the GP-LVM under different level of constraints. The nearest neighborhood (NN) classifier is used for classification after dimensionality reduction.

In the following experiments, the pairwise constraints are randomly generated from the training set. If two samples belong to the same class, the relation between them is defined as *must-link* constraints; otherwise it is defined as *cannot-link* constraints. The others which have not been chosen are unlabeled samples. We repeat each experiment 50 times independently. The classification mean error versus constrained number by using the GP-LVM and the proposed methods are shown in Fig. 6.

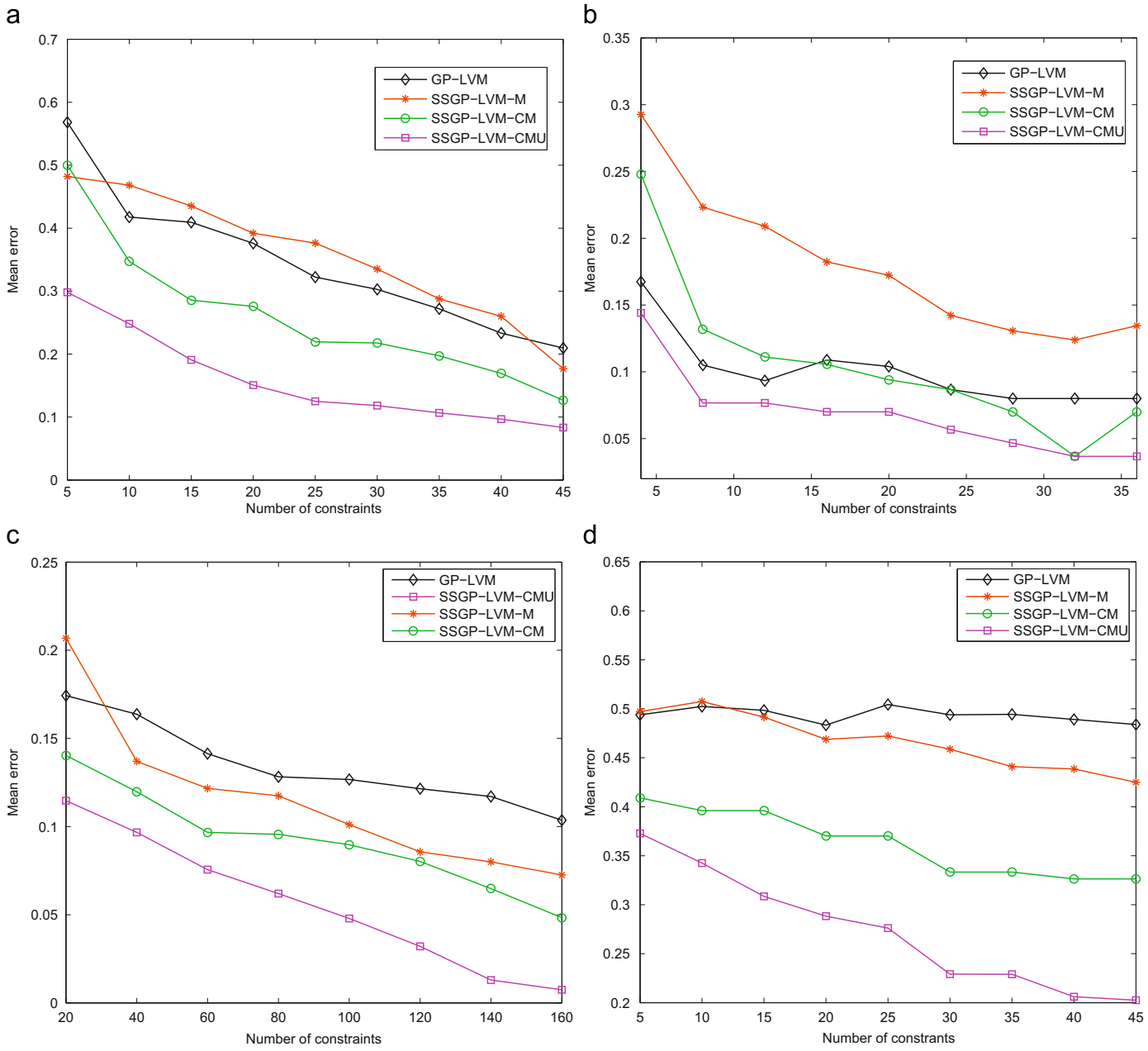


Fig. 6. The comparison of classification error rates between the proposed 3 methods and the GP-LVM on 4 data sets with different number of constraints. (a) balance data ($d=2$) (b) iris data ($d=2$) (c) oil data ($d=4$) (d) sonar ($d=4$)

The results in Fig. 6 demonstrate that the proposed semi-supervised models nearly always achieve the lower mean error than the traditional GP-LVM. The main reason is that the proposed models utilize the constrained information, and the GP-LVM does not use any constraints or label knowledge. In the three constrained models (SSGP-LVM-M, SSGP-LVM-CM and SSGP-LVM-CMU), the SSGP-LVM-M which only use the *must-link* constraints obtains poor performance compared with SSGP-LVM-CM and SSGP-LVM-CMU. Therefore only using the *must-link* constraints cannot achieve ideally results. By adding the *cannot-link* constraints, the SSGP-LVM-CM is superior to the GP-LVM and SSGP-LVM-M on the three of four data sets, except for *Iris* data set. The advantages of the SSGP-LVM-CMU over other methods can be seen in this experiment. For each dataset, the classification error rate of the SSGP-LVM-CMU is much lower than other three dimensionality reduction algorithms. Especially on *sonar* data set, the SSGP-LVM-CMU gives the mean error less than 30 percent; while the mean error which the GP-LVM can offers is always around 50 percent. That is because that the SSGP-LVM-CMU utilizes both the pairwise constraints and unlabeled samples.

4.4. *Classification accuracy influenced by dimension of constraints*

To observe the dependence of the semi-supervised GP-LVM method on the dimensions of the latent space, we test the algorithms in different dimensional latent space on three data sets, i.e., *ORL* face data, *oil* data set and *handwritten digits* data in this subsection. The pairwise constraints are randomly selected, and we repeat each experiment 50 times independently. The percent of constraints is given in the following Table 3.

Table 3
Statistics in percentiles of constraints.

Data	Total number	Class	Number of constraints
ORL	400	40	4*40
Oil	1000	3	100*3
USPS(3,5)	300*2	2	100*2

As shown in Table 3, the numbers of constrained samples in three data sets are different. Percent of the constrained samples in each data set is around 30%. The results for three data sets are shown in Fig. 7.

Fig. 7 shows the plots for mean error vs. number of dimensions. As can be seen from Fig. 7, the mean errors are decreasing with increasing dimensions of the latent space. That is, these curves have the same tendency. Although compared with the GP-LVM, the advantages of SSGP-LVM-M and SSGP-LVM-CM are not obvious in Fig. 7(a). The SSGP-LVM-CMU has much more advantage than other three models. Fig. 7(b) and (c) show the performance comparison on the *oil* data set and *USPS* data sets respectively. Compared with the GP-LVM, the proposed methods, especially the SSGP-LVM-CM and SSGP-LVM-CMU, significantly outperform the traditional GP-LVM for all the three data sets. As the number of dimension grows, the performance of the proposed method can keep the advantage consistently.

4.5. Classification accuracy comparison with constraints score

In order to evaluate the performance of the proposed methods against other semi-supervised method based on pairwise constraints, such as constraint *score* introduced in [28], we use *wine*, *YaleB* face and *letter* data sets in these experiments. For the *letter* data set, we choose the first 5 letters' samples: 'a', 'b', 'c', 'd' and 'e'. The pairwise constraints are randomly selected, and each selection is repeated 50 times independently. The experimental results on three data sets are shown in Fig. 8.

For the SSGP-LVM-CMU and Constraints Score, we choose the same number of constraints. For *wine* data set, the number of constraints is 20, that is, the numbers of the *must-link* and *cannot-link* are 10. For the second data set, *YaleB* face database, the number of constraints is 30 including 15 *must-link* and 15 *cannot-link*. There are 40 constraints selected randomly in *letter* data set. By comparing the results shown in Fig. 8, it can be concluded that the proposed method outperforms significantly, especially for the last two data sets. For the *wine* data set, the advantage of the proposed method is not obvious. This is because that the samples have a linear structure manifold, and constraints *score* is a linear method for feature selection, while the proposed method is a nonlinear dimensionality reduction model.

In these experiments, we validate the effectiveness of the proposed semi-supervised GP-LVM on a broad range of benchmark data sets, including *USPS handwritten digits, oil* data set, *ORL* face database, *YaleB* face database and five UCI data sets, i.e., *iris, wine, balance, sonar, letter.* The experiments are divided into four subsections. In the first subsection, the proposed method is validated in an intuitionist way, where the *oil* data is mapped into 2-dimensional space. Since the proposed method is based on the pairwise constraints, the performance may be influenced by the number of constraints. In the second subsection, we evaluate the performance of the semi-supervised GP-LVM on the *oil* data set and three UCI data sets, including *balance, sonar* and *iris*. To consider the dependence of the semi-supervised GP-LVM method on the dimensions of the latent space, we test the algorithms in different dimensional latent space on three data sets, i.e., *ORL* face data, *oil* data set and *handwritten digits* data in the third

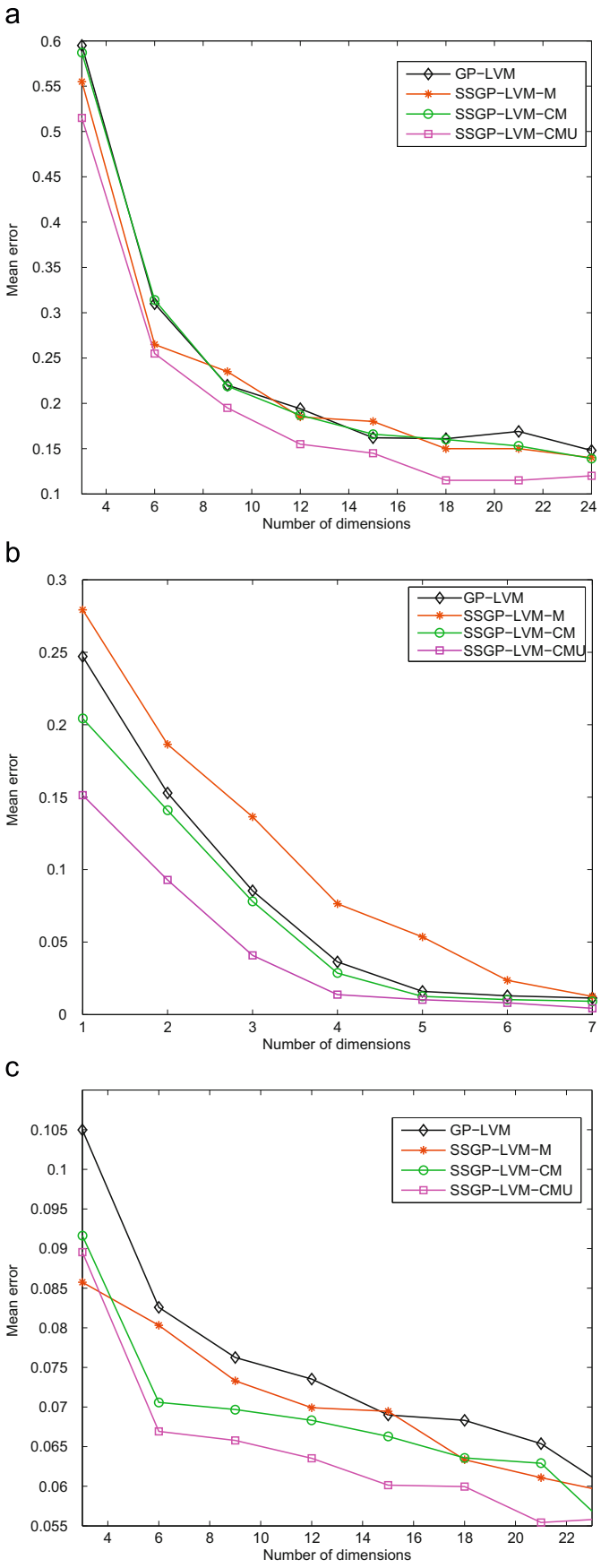


Fig. 7. The comparison of classification error rates between the proposed 3 methods and the GP-LVM on 3 data sets with different number of dimensions. (a) ORL data (b) oil data (c) USPS (3,5).

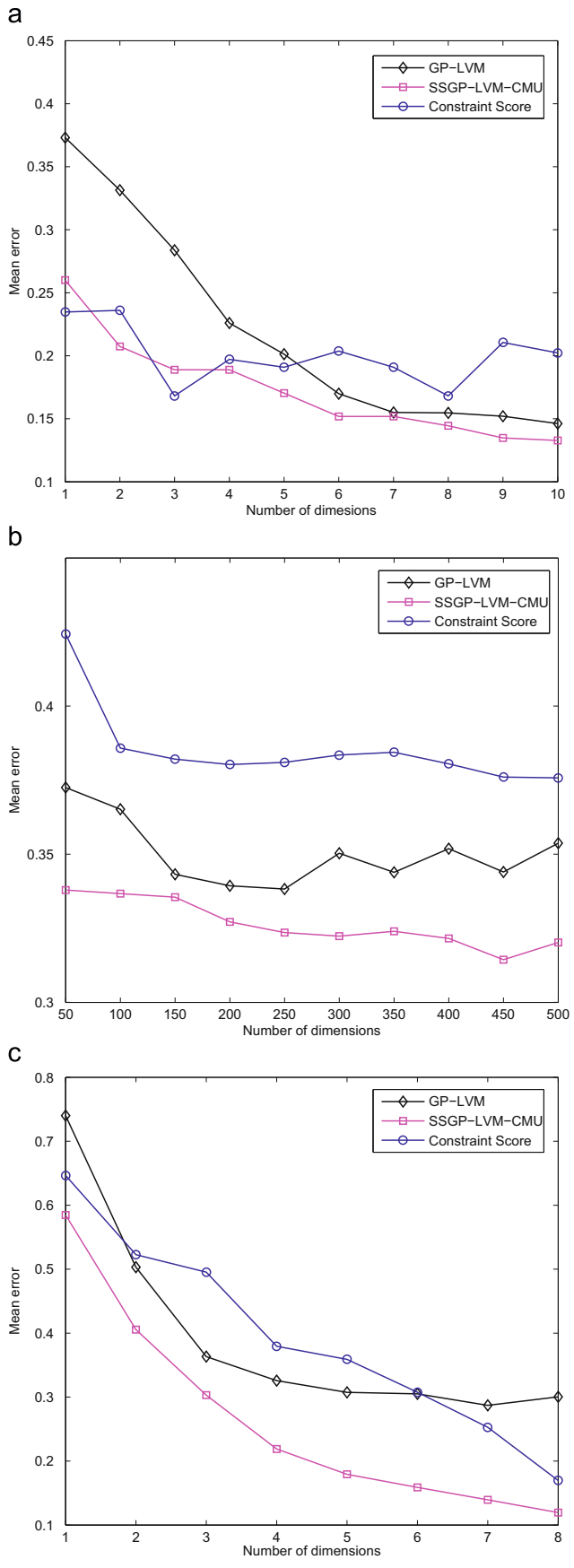


Fig. 8. The performance comparison of 3 dimensionality reduction methods on 3 data sets. (a) *wine* data (b) *YaleB* database (c) *letter* data.

subsection. In the forth subsection, the comparison experiment is conducted on *wine*, *YaleB* face and *letter* data sets to illustrate the superiority of the proposed method over other semi-supervised method related to pairwise constraints. The experiments on various datasets validate the superiority of the proposed method.

5. Conclusions

This paper proposes a novel semi-supervised dimensionality reduction model, i.e., the semi-supervised Gaussian process latent variable model. It discovers the discriminative structure of the high-dimensional data in the low-dimensional latent space through utilizing the pairwise constraints. We also detailedly describe how to constrain the latent variables with semi-supervised information. Compared with traditional latent variable models, the proposed semi-supervised model is much more discriminative. A great deal of experimental results is provided to testify the effectiveness of the proposed method. In future work, we will conduct the theoretical analysis of the proposed semi-supervised GP-LVM by combining with the graph model [29,30] with the latent variable model, and establish a general framework for the semi-supervised latent variable model for dimensionality reduction problems.

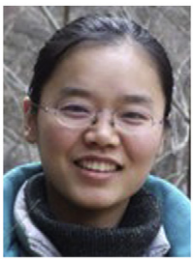
Acknowledgements

We want to thank the helpful comments and suggestions from the anonymous reviewers. This research was supported by the National Natural Science Foundation of China (60771068, 60702061, 60832005), the Ph.D Programs Foundation of Ministry of Education of China (No. 20090203110002), the Natural Science Basic Research Plane in Shaanxi Province of China (2009JM8004), the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) in China and the National Laboratory of Automatic Target Recognition, Shenzhen University, China.

References

[1] I. Joliffe, Principal Component Analysis, Springer, New York, 1986.
[2] Y. Pang, D. Tao, Y. Yuan, X. Li, Binary two-dimensional PCA, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 38 (4) (2008) 1176–1180.
[3] I. Borg, P. Groenen, Modern multidimensional scaling: theory and applications, 2nd ed., Springer, New York, 2005.
[4] B. Scholkopf, A. Smola, K.-R. Muller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Computation 10 (5) (1998) 1299–1319.
[5] Y. Pang, Y. Yuan, X. Li, Iterative subspace analysis based on feature line distance, IEEE Transactions on Image Processing, 18 (4) (2009) 903–907.
[6] Y. Pang, Y. Yuan, X. Li, Gabor-Based region covariance matrices for face recognition, IEEE Transactions on Circuits and Systems for Video Technology 18 (7) (2008) 989–993.
[7] T. Lin, H. Zha, S. Lee, Riemannian manifold learning for nonlinear dimensionality reduction, in: Proceeding of the European Conference on Computer Vision, Graz, Austria, May 2006, pp. 44–55.
[8] M.E. Tipping, C.M. Bishop, Probabilistic principal component analysis, Journal of the Royal Statistical Society, Series B 61 (3) (1999) 611–622.
[9] D.J. Bartholomew, Latent Variable Models and Factor Analysis, Charles Griffin & Co. Ltd, London, 1987.
[10] D.J. Bartholomew, Statistical Factor Analysis and Related Methods, Wiley, New York, 2004.
[11] N.D. Lawrence, Gaussian process models for visualization of high dimensional data, Advance in Neural Information Processing Systems 16 (2004) 329–336.
[12] N.D. Lawrence, Probabilistic non-linear principal component analysis with Gaussian process latent variable models, Journal of Machine Learning Research 6 (2005) 1783–1816.
[13] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning, the MIT press, Cambridge, 2006.
[14] M. Girolami, S. Rogers, Variational Bayesian multinomial probit regression with Gaussian process priors, Neural Computation 18 (8) (2006) 1790–1817.
[15] M. Opper, O. Winther, Gaussian processes for classification: mean-field algorithms, Neural Computation 12 (11) (2000) 2655–2684.

[16] F. Wang, C. Zhang, Robust self-tuning semi-supervised learning, Neurocomputing 71 (16–18) (2008) 2931–2939.
[17] J. Zhao, K. Lu, X. He, Locality sensitive semi-supervised feature selection, Neurocomputing 71 (10–12) (2008) 1842–1849.
[18] D. Zhang, Z. Zhou, S. Chen, Semi-supervised dimensionality reduction, in: Proceeding of the International Conference on Data Mining, Minneapolis, USA, Apr. 2007, pp. 629–634.
[19] S.C. H. Hoi, W. Liu, M.R. Lyu, W. Ma, Learning distance metrics with contextual constraints for image retrieval, in: Proceeding of the International Conference on Computer Vision and Pattern Recognition, NewYork, USA, Jun. 2006, pp. 2072–2078.
[20] A. Bar-Hillel, T. Hertz, N. Shental, D. Weinshall, Learning a mahalanobis metric from equivalence constraints, Journal of Machine Learning Research 6 (2005) 937–965.
[21] X. He, P. Niyogi, Locality preserving projections, Advance in Neural Information Processing Systems 16 (2003) 153–160.
[22] H. Cevikalp, J. Verbeek, F. Jurie, A. Kläser, Semi-supervised dimensionality reduction using pairwise equivalence constraints, in: Proceeding of the International Conference on Computer Vision Theory and Applications, Funchal, Portugal, Jan. 2008, pp. 489–496.
[23] C.K.I. Williams, C.E. Rasmussen, Gaussian processes for regression, Advance in Neural Information Processing Systems 8 (1995) 598–604.
[24] C.K.I. Williams, D. Barber, Bayesian classification with Gaussian processes, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (12) (1998) 1342–1351.
[25] S.P. Yu, K. Yu, V. Tresp et al., Supervised probabilistic principal component analysis, in: Proceedings of the Intinternational Conference on Knowledge Discovery and Data Mining, Philadelphia, USA, Aug 2006, pp. 464–473.
[26] http://archive.ics.uci.edu/ml.
[27] N.D. Lawrence, J. Quinonero-Candela, Local distance preservation in the GP-LVM through back constraints, in: Proceeding of the International Conference on Machine Learning, Pittsburgh, USA, Jun 2006, pp. 513–520.
[28] D. Zhang, S. Chen, Z. Zhou, Constraint Score: a new filter method for feature selection with pairwise constraints, Pattern Recognition 41 (2008) 1440–1451.
[29] Z. Li, J. Liu, X. Tang, Pairwise constraint propagation by semidefinite programming for semi-supervised classification, in: Proceedings of the International Conference on Machine Learning, Helsinki, Finland, Jun 2008, pp. 576–583.
[30] S. Basu, M. Bilenko, R.J. Mooney, A probabilistic framework for semi-supervised clustering, in: Proceeding of the International Conference on Knowledge Discovery and Data Mining, Seattle, USA, Aug. 2004 pp. 59–68.



Xiumei Wang received the BMath Degree from Shandong Normal University in 2002 and the MSc degree from Xidian University in 2005. She is currently a Ph.D. candidate at the School of Sciences at the Xidian University. Her research interests mainly involve nonparametric statistical models and machine learning.

Xinbo Gao received the BSc, MSc and Ph.D. degrees in signal and information processing from Xidian University, China, in 1994, 1997 and 1999, respectively. From 1997 to 1998, he was a research fellow in the Department of Computer Science at Shizuoka University, Japan. From 2000 to 2001, he was a postdoctoral research fellow in the Department of Information Engineering at the Chinese University of Hong Kong. Since 2001, he joined the School of Electronic Engineering at Xidian University. Currently, he is a Professor of Pattern Recognition and Intelligent System, and Director of the VIPS Lab, Xidian University. His research interests are computational intelligence, machine learning, computer vision, pattern recognition and artificial intelligence. In these areas, he has published 4 books and around 100 technical articles in refereed journals and proceedings including IEEE TIP, TCSVT, TNN, TSMC etc. He is on the editorial boards of journals including EURASIP Signal Processing (Elsevier), Neurocomputing (Elsevier). He served as general chair/co-chair or program committee chair/co-chair or PC member for around 30 major international conferences.

Yuan Yuan is currently a Lecturer with the School of Engineering and Applied Science, Aston University, United Kingdom. She received her BEng degree from the University of Science and Technology of China, China, and the Ph.D. degree from the University of Bath, United Kingdom. She has over sixty scientific publications in journals and conferences on visual information processing, compression, retrieval etc. She is an associate editor of International Journal of Image and Graphics (World Scientific), an editorial board member of Journal of Multimedia (Academy Publisher), a guest editor of Signal Processing (Elsevier), and a guest editor of Recent Patents on Electrical Engineering. She was a chair of some conference sessions, and a member of program committees of many conferences. She is a reviewer for several IEEE transactions, other international journals and conferences.

Dacheng Tao received the B.Eng. degree from the University of Science and Technology of China (USTC), the M.Phil degree from the Chinese University of Hong Kong (CUHK), and the Ph.D. degree from the University of London (Lon). Currently, he is a Nanyang Assistant Professor with the School of Computer Engineering in the Nanyang Technological University and holds a visiting post in Lon. He is a Visiting Professor in the Xi Dian University and a Guest Professor in the Wu Han University. His research is mainly on applying statistics and mathematics for data analysis problems in computer vision, multimedia, machine learning, data mining, and video surveillance. He has published more than 100 scientific papers including IEEE TPAMI, TIP, TKDE, CVPR, ECCV, NIPS, ICDM; ACM TKDD, Multimedia, KDD etc., with best paper runner up awards and finalists. One of his TPAMI papers received an interview with ScienceWatch.com (Thomson Scientific). His H-Index in google scholar is 14 and his Erdös number is 3. He holds the K. C. WONG Education Foundation Award.

Jie Li received the BSc, MSc and Ph.D. degrees in Circuit and System from Xidian University, China, in 1995, 1998 and 2005, respectively. Since 1998, she joined the School of Electronic Engineering at Xidian University. Currently, she is an Associate Professor of Xidian University. Her research interests include computational intelligence, machine learning, and image processing. In these areas, she has published over 30 technical articles in refered journals and proceedings including IEEE TCSVT, IJFS etc.