+ MODEL
ARTICLE IN PRESS
+ MODEL
Available online at www.sciencedirect.com
ScienceDirect
CAAI Transactions on Intelligence Technology xx (2016) 1—11
http://www.journals.elsevier.com/caai-transactions-on-intelligence-technology/

Review Article

A review on Gaussian Process Latent Variable Models

Ping Li, Songcan Chen*

College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

Available online ■ ■ ■

Abstract

Gaussian Process Latent Variable Model (GPLVM), as a flexible bayesian non-parametric modeling method, has been extensively studied and applied in many learning tasks such as Intrusion Detection, Image Reconstruction, Facial Expression Recognition, Human pose estimation and so on. In this paper, we give a review and analysis for GPLVM and its extensions. Firstly, we formulate basic GPLVM and discuss its relation to *Kernel Principal Components Analysis*. Secondly, we summarize its improvements or variants and propose a taxonomy of GPLVM related models in terms of the various strategies that be used. Thirdly, we provide the detailed formulations of the main GPLVMs that extensively developed based on the strategies described in the paper. Finally, we further give some challenges in next researches of GPLVM.

Keywords: GPLVM; Non-parametric method; Gaussian process

1. Introduction

In many machine learning tasks, we are often faced with various complex, particularly high dimensional, data/observations [1—3] for which our goal is to learn the low dimensional underlying patterns from those observations [4,5]. For example, in classification task [6—9], we want to identify a category of a new observation by a classifier learned from a set of training data. In clustering task, the goal is to group a set of observations in such a way that observations in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups [10,11], achieving to understand inherent (low dimensional) structure of a given data set.

Recently, many machine learning models have been proposed to address the above problems [1,3,12,13]. Among those models, latent variable models (LVMs) [3,14,15] as a kind of underlying patterns extraction methods, have been widely used in image recognition [16], information retrieval [17], speech recognition [18] and recommender systems [19]. A latent variable model generally refers to a statistical model that relates a set of variables (so-called manifest variables) to a set of latent variables under the assumption that the responses on the manifest variables are controlled by the latent variables. Furthermore, we can provide latent variables with various meanings for specific tasks. In *Dimension Reduction* (DR), we assume that the latent variables are the low dimensional representations of high dimensional samples. In clustering, the latent variables can be defined to represent the clustering membership of samples [20]. This flexible definition of latent variables has made LVMs widely be used in many machine learning tasks.

LVMs have a history of several decades and many machine learning models can be actually considered as its special cases or variants, e.g., neural networks [18], PCA [21], latent graphical models [3] and so on. Among these models, *Gaussian Process Variable Models* (GPLVMs) [15], as a large class of LVMs, have been explored and applied in many machine scenarios. They can be considered as the combination of LVM and a Bayesian non-parametric *Gaussian Process* (GP) [22] model. GP is a probabilistic model and has been extensively applied in many machine learning tasks such as regression [23—25], classification [6—8] and clustering [26].

* Corresponding author.
E-mail addresses: ping.li.nj@nuaa.edu.cn (P. Li), s.chen@nuaa.edu.cn (S. Chen).
Peer review under responsibility of Chongqing University of Technology.

http://dx.doi.org/10.1016/j.trit.2016.11.004

In general, we can consider these GP-based models to be a set of LVMs where each observed variable is the sum of the corresponding latent variable and noise. Different from other LVMs, these latent variables can be thought of as functional variables which are the noise-free form of observed variables. In LVMs, our goal is to learn the latent variables or the underlying pattern of data. While the above GP-based models try to infer the target variable of new sample by integrating out the latent variables. This is a major difference between GP-based model and other LVMs.

In order to infer the latent variables, GPLVM assumes that the functional variables are generated by GP from some low dimensional latent variables. It is these latent variables that we should inferred from data. In model inference, we can learn the latent variables by integrating out the functional variables and maximizing the log marginal likelihood. Although originally proposed for dimension reduction, GPLVM has been extended and widely used in many machine learning scenarios, such as Intrusion Detection [27], Image Reconstruction [28], Facial Expression Recognition [29], Human pose estimation [30], Age Estimation [31] and Image-Text Retrieval [32].

We can analyze the advantages of GPLVM from two aspects. Firstly, GPLVM can greatly benefit from the non-linear learning characteristic of GP which uses a non-linear kernel to replace the covariance matrix. Moreover, as a non-linear DR model, GPLVM has a strong link with *Kernel Principal Components Analysis* (KPCA) [33] (a popular DR method) and can be considered as a *Probabilistic Kernel Principal Components Analysis* (PKPCA). For such a link, we will discuss in next section. Secondly, most of the existing LVMs are parametric models in which there is a strong assumption on the projection function or data distribution. Such a parametric construction form partly loses flexibility in modeling. Therefore, in the past decades, many non-parametric machine learning methods have successively been proposed, such as Nearest Neighbor methods [34,35] and kernel estimates of probability densities [36,37]. GP and GPLVM can be treated as a class of Bayesian non-parametric model whose distribution-free form makes the models have a flexible structure which can grow in size to accommodate the complexity of the data.

Besides widely used in DR, GPLVM can also be extended to other machine learning tasks due to its characteristics below. Firstly, its distribution-free assumption on prior of latent variables provides us a lot of opportunities to improve it. Secondly, its generation process can be amenable to different tasks. Thirdly, we can also exert classical kernel methods for a further expansion of GPLVM, such as enhancing the scalability of the model, automatic selection of the feature dimension and so on. Despite GPLVM has been widely studied and extended, to our best knowledge, there has actually had no survey for those related models. So in this paper, we try to present a review and analysis of both GPLVM and its extensions.

The rest of this paper is organized as follows: In Section 2, we formulate the GPLVMs and discuss its relation to *Kernel Principal Components Analysis* (KPCA). In Section 3, we summarize its improvements or variants and propose a taxonomy of GPLVM related models. A specific review of GPLVM that extensively developed in the past decade is given in Section 4. Finally, in Section 5, we further give some challenges in next researches of GPLVM.

2. Gaussian process and Gaussian Process Latent Variable Model

2.1. Gaussian process

GP, as the a flexible Bayesian nonparametric model and the building block for GPLVM, has been widely used in many machine learning applications [38—40] for data analysis. In GP, we model a finite set of random function variables $\boldsymbol{f} = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_N)]^T$ as a joint Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{K}$, where $\boldsymbol{x}_i$ is the ith input. If the function f has a GP prior, we can write it as

$$f \sim \mathcal{GP}(\boldsymbol{\mu}, \boldsymbol{K}) \tag{1}$$

where in many cases we can specify a zero mean ($\boldsymbol{\mu} = \mathbf{0}$) and a kernel matrix $\boldsymbol{K}$ (with hyper-parameter $\boldsymbol{\theta}$) as covariance matrix. GP has been widely used in various machine learning scenarios such as regression, classification, clustering. In this section, we detailed the formulation of *Gaussian Process Regression* (GPR) to demonstrate the use of GP.

In GPR, our goal is to predict the response y^* of a new input $\boldsymbol{x}^*$, given a training dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ of N training samples, where $\boldsymbol{x}_i$ is the input variable and y_i is the corresponding continuous response variable. We model the response variable y_i as a noise-version of the function value $f(\boldsymbol{x}_i)$

$$y_i \sim \mathcal{N}\left(f(\boldsymbol{x}_i), \sigma^2\right) \tag{2}$$

where the distribution of noise is Gaussian $\mathcal{N}(0, \sigma^2)$ with variance σ^2. From the above definition, we can get the joint probability of the response variables and latent function variables $p(\boldsymbol{y}, \boldsymbol{f}) = p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f})$. Then we can know that the distribution of the latent function value f^* is a Gaussian distribution with mean $\mu(\boldsymbol{x}^*)$ and variance $var(\boldsymbol{x}^*)$:

$$\begin{aligned} \mu(\boldsymbol{x}^*) &= \boldsymbol{k}_{\boldsymbol{x}^*X}\left(\sigma^2 I + \boldsymbol{K}_{XX}\right)^{-1}\boldsymbol{y} \\ var(\boldsymbol{x}^*) &= k_{\boldsymbol{x}^*\boldsymbol{x}^*} - \boldsymbol{k}_{\boldsymbol{x}^*X}\left(\sigma^2 I + \boldsymbol{K}_{XX}\right)^{-1}\boldsymbol{k}_{X\boldsymbol{x}^*} \end{aligned} \tag{3}$$

where $\boldsymbol{k}_{\boldsymbol{x}^*X} = k(\boldsymbol{x}^*, \boldsymbol{X})$ is a n-dimensional row vector of the covariance between $\boldsymbol{x}^*$ and the N training samples $(\boldsymbol{k}_{\boldsymbol{x}^*X} = \boldsymbol{k}_{X\boldsymbol{x}^*}^T)$, $\boldsymbol{K}_{XX} = k(\boldsymbol{X}, \boldsymbol{X})$ denotes the kernel matrix of the N training samples.

2.2. Gaussian Process Latent Variable Model

GPLVM [15] is originally proposed for dimension reduction of high dimensional data. Its goal is to learn the low dimensional representation $\boldsymbol{X}^{N\times Q}$ of the data matrix $\boldsymbol{Y} \in \mathbb{R}^{N\times D}$, where N and D are the number and dimensionality of training

Please cite this article in press as: P. Li, S. Chen, A review on Gaussian Process Latent Variable Models, CAAI Transactions on Intelligence Technology (2016), http://dx.doi.org/10.1016/j.trit.2016.11.004