Probabilistic Nonlinear Dimensionality Reduction through Gaussian Process Latent Variable Models: an Overview

Matteo Bodini∗

Department of Computer Science
Università degli Studi di Milano
Via Celoria 18, 20133 Milano (Italy)

Email: ∗matteo.bodini@unimi.it

Abstract—From an algorithmic complexity point of view, machine learning methods scale and generalize better when using a few key features: using lots is computationally expensive, and overfitting can occur. High dimensional data is often counter-intuitive to perceive and process, but unfortunately it is common for observed data to be in a representation of greater dimensionality than it requires. This gives rise to the notion of dimensionality reduction, a sub-field of machine learning that is motivated to find a descriptive low-dimensional representation of data.

In this review it is explored a way to perform dimensionality reduction, provided by a class of Latent Variable Models (LVMs). In particular, the aim is to establish the technical foundations required for understanding the Gaussian Process Latent Variable Model (GP-LVM), a probabilistic nonlinear dimensionality reduction model. The review is organized as follows: after an introduction to the problem of dimensionality reduction and LVMs, Principal Component Analysis (PCA) is recalled and it is reviewed its probabilistic equivalent that contributes to the derivation of GP-LVM. Then, GP-LVM is introduced, and briefly a remarkable extension of the latter, the Bayesian Gaussian Process Latent Variable Model (BGP-LVM) is described. Eventually, and the main advantages of using GP-LVM are summarized.

Index Terms—Dimensionality reduction, Gaussian processes, latent variable models

I. INTRODUCTION

When we resort to machine learning methodologies to face a particular problem, it is frequent to deal with high dimensional data. Many common tasks can be considered: in a classification one the aim consists of determine the class of some new observations through a classifier, learned from a set of training data; in clustering, the goal is to group a set of data so that those in the same group, a cluster, are more similar (considering some notion of similarity) to each other than to data in other groups [1], [15]. Considering in particular the latter, we can notice that its purpose is to get the inherent low dimensional structure of a given data set. Either way, the main goal of machine learning techniques is to learn the low dimensional underlying patterns from a set of data, even of high dimensionality [1].

To face the above problems, several machine learning methods are available [18], [1], [15], [12]. Beyond these, LVMs [1], [2], [22], [4] have been widely used in image recognition [24], [5], facial expression analysis [3], [7], information retrieval [8], and speech recognition [9], [6]. In principle, a LVM refers to a statistical model which relates a set of variables, the so-called manifest variables, to a set of latent variables under the assumption that the responses on the manifest variables are controlled by latent variables. A remarkable aspect of these models lies in the fact that different meanings can be provided to the latent variables, depending on the specific task. This flexible definition has made LVMs largely be used in several machine learning tasks: for instance, in clustering latent variables can represent the clustering membership of samples [21]. In the main topic of this review, dimensionality reduction, it is assumed that latent variables are the low dimensional representations of high dimensional samples.

Many machine learning models can be considered as specific cases or variants of LVMs, e.g. neural networks [13], and PCA [22]. Among these, Gaussian Process Latent Variable Models (GP-LVMs) represent a large class of LVMs and have been applied in many scenarios. Such models can be seen as a mix of LVMs and Bayesian non parametric Gaussian Process (GP). A GP is a probabilistic model used in regression, classification and clustering [25], [17]. At first sight, GP based models are composed by a set of LVMs where each observed variable is the sum of the corresponding latent variable, and noise. The main difference from the other LVMs is that latent variables are seen as functional variables: they are the noise free form of observed variables. With LVMs the aim is to learn latent variables or the underlying pattern of high dimensional data, and GPLVMs try to infer the target variable of new sample by integrating out the latent variables. This stands for the major difference between the latter and other LVMs. For the purpose of infer latent variables, GP-LVM assumes that the functional variables are generated by GP from some low dimensional latent variables: these should be inferred from data. Hence, latent variables are learned by
integrating out the functional variables and maximizing the log marginal likelihood. Even if GP-LVM was first proposed for dimensionality reduction, it has been widely used in many scenarios, such as Facial Expression Recognition [11], Facial Expression Analysis [3], and Human pose estimation [10].

LVMs have been widely studied and applied, and here it is presented a review and an analysis especially of GP-LVM, focusing on the problem of dimensionality reduction. In the first paragraphs the foundations necessary to understand the model are described: first, PCA, and multivariate Gaussian distribution are recalled. Then, it is introduced a probabilistic version of PCA from which GP-LVM is derived [19]. Eventually, GP-LVM and one of its remarkable extension, the Bayesian GP-LVM, are introduced, and are summarized which are the main advantages of using GPLVM.

II. PRINCIPAL COMPONENT ANALYSIS

Within the set of multivariate analysis methods, Principal Component Analysis (PCA) is a statistical technique used to perform a dimensionality reduction which is linear. It was originally introduced in Pearson K. [23], and independently developed in Hotelling H. [14], where it is presented the standard algebraic derivation of PCA. For such reason it is also known as Hotelling Transform.

Just to recall the idea of PCA, consider a set of observable variables that can be measured. It is common that the collected data shows that some variables change together: this relation is formally measured as the covariance. PCA searches for an orthogonal set of principal components that retain maximal variance. A principal component can be seen as a combination of the observed variables. If two observed variables strongly covary, then it makes sense to describe the data with a single variable instead. The more linearly observed variables covary, the less information we lose from choosing a smaller set of principal components, thus effectively reducing dimensionality.

Let us now give a more formal definition of principal component. Given a set of observed d-dimensional data represented as a design matrix, $X = \{x_1, ..., x_n\}^T$, the q principal components $w_j$, $j \in \{1, ..., q\}$, are the orthonormal axes with maximal variance. The first principal component is a linear function $\alpha_1^T X$ that retains most variance of $X$, where $\alpha_1 = [\alpha_{11}, \alpha_{12}, ..., \alpha_{1n}]$ is a vector of n constants such that [16]:

$$\alpha_1^T X = \alpha_{11} x_1 + \alpha_{12} x_2 + ... + \alpha_{1n} x_n = \sum_{i=1}^n \alpha_{1i} x_i.$$  

The next principal components are found by looking for a linear function that is orthogonal to the selected principal components and retain maximum variance.

PCA can be performed by singular value decomposition (SVD) of design matrix $X$ [16]. SVD is a factorization by which we can write the design matrix as

$$X = U \Sigma V^T.$$  

Denoting $r = r(X)$ the rank of $X$, then $U \in \mathbb{R}^{n \times r}$ is a matrix of orthonormal columns that are the left singular vectors, $\Sigma \in \mathbb{R}^{r \times r}$ is a diagonal matrix of the singular values of $X$, and $V \in \mathbb{R}^{d \times r}$ is a matrix of orthonormal columns that are the right singular vectors.

A limitation of the standard well known PCA is the lack of a probabilistic solution. One way to formulate a probabilistic model is to introduce the notion of noise among observed variables, where it is often the case that the noise is assumed to follow a Gaussian distribution. In the remainder of this review the multivariate Gaussian distribution will take a central role, and some of its useful properties in the derivation of GP-LVM will be shortly recalled in the next paragraph.

III. MULTIVARIATE GAUSSIAN DISTRIBUTION

The Gaussian (normal) distribution is a reasonable prior assumption for data that is subject to the central limit theorem, which states that as the sample size of a population tends to infinity, the distribution becomes normally distributed. A random variable $X$ that is normally distributed with mean $\mu$ and variance $\sigma^2$ is usually denoted as

$$X \sim \mathcal{N}(\mu, \sigma^2).$$

The Gaussian density for a single variable $y$ is expressed as:

$$\mathcal{N}(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right),$$

$$\mathcal{N}(y|\mu, \sigma^2) \equiv p(y|\mu, \sigma^2).$$

For a d-dimensional vector $y$, the multivariate Gaussian distribution with mean vector d-dimensional $u$, $d \times d$ covariance matrix $\Sigma$, and the determinant of $\Sigma$ as $|\Sigma|$, is

$$\mathcal{N}(y|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2} \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(y - \mu)^T \Sigma^{-1}(y - \mu)\right).$$

The Gaussian distribution is closed under addition [1], scaling [2], and multiplication [3] - all of which yields a result that is also a Gaussian distribution [25]. These noticeable properties will be useful in the following sections as they allow the analytical integration of multivariate Gaussian distributions:

$$\sum_{i=1}^n y_i \sim \mathcal{N}\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right),$$

$$wy \sim \mathcal{N}(w\mu, w^2\sigma^2),$$

$$\mathcal{N}(x|a, A)\mathcal{N}(x|b, B) = Z^{-1} \mathcal{N}(x|c, C),$$

$$Z^{-1} = \frac{1}{(2\pi)^{d/2} |A + B|} \exp\left(-\frac{1}{2} (a - b)^T (A + B)^{-1} (a - b)\right),$$

$$c = C(A^{-1}a + B^{-1}b),$$

$$C = (A^{-1} + B^{-1})^{-1},$$

for means $\mu_i \in \mathbb{R}$ and variances $\sigma_i^2 \in \mathbb{R}$, $i = 1, 2, ..., n$, $w \in \mathbb{R}$, two d-dimensional mean vectors $a, b$, and $d \times d$ covariance matrices $A, B$. Let $l \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $h \sim \mathcal{N}(\mu_2, \sigma_2^2)$.
be jointly Gaussian distributed variables, if the variables are independent, then \( p(l, h) = p(l)p(h) \). The joint probability density is thus,

\[
p(l, h) = \frac{1}{\sqrt{2\pi\sigma_1^2\sigma_2^2}} \exp\left( -\frac{1}{2} \left( \frac{(l - \mu_1)^2}{\sigma_1^2} + \frac{(h - \mu_2)^2}{\sigma_2^2} \right) \right).
\]

Assuming independence, the joint probability density for a \( n \)-dimensional vector \( y \) is expressed as

\[
p(y) = \frac{1}{2\pi\sqrt{|D|}} \exp\left( -\frac{1}{2} (y - \mu)^T D^{-1} (y - \mu) \right),
\]

where \( D \in \mathbb{R}^{n \times n} \) is the diagonal matrix of the variances.

IV. A Probabilistic Model for PCA

Tipping and Bishop introduced a probabilistic framework for principal component analysis by constraining the noise distribution of a LVM [25].

A LVM transforms a set of \( n \) \( d \)-dimensional observed variables encoded as a design matrix, \( Y = [y_1, ..., y_n]^T \), to a set of \( n \) \( q \)-dimensional latent (unobserved) variables, \( X = [x_1, ..., x_n]^T \). Latent variables are parsimonious, it is generally the case that \( q \ll d \), explaining the original data with fewer variables. A notable latent variable model is that of factor analysis, one that assumes linearity in relation of the observed data set. For each observed data point, \( y_i \in Y \), \( 1 \leq i \leq n \), there is an associated latent variable \( x_i \). The original data can be represented in terms of the corresponding latent variable as

\[
y_i = Wx_i + \mu_i + \epsilon_i.
\]

The matrix \( W \) represents the linear relationship between the latent space with the data-space. Figure 1 shows how the input latent variable and the matrix that model the relation contributes to predicting an output observed variables. The parameter \( \mu_i \) allows for non-zero mean, and the \( \epsilon_i \) parameter represents noise within the model. We notice that standard PCA can be viewed as a variant of factor analysis where the noise parameter is not accounted for and furthermore, PCA reduce dimensionality without explaining the relation between the original data and its principal components. Regression is possible with a model that explains the relation between observed variables and latent variables.

The latent variable model developed by Tipping and Bishop performs principal component analysis by modelling the noise parameter of equation 5 as an isotropic, spherical Gaussian distribution. The noise values, \( \epsilon_i \in \mathbb{R}^{d \times 1} \), are sampled from an independent spherical Gaussian distribution

\[
\epsilon_i \sim N(0, \beta^{-1}I).
\]

The conditional probability distribution of a observed variables \( y_i \) given input variables \( x_i \) is thus Gaussian distributed as

\[
p(y_i|x_i) = N(Wx_i + \mu_i, \beta^{-1}I).
\]

The prior of latent variables is assumed to be standard Gaussian with zero mean and unit covariance, defined as \( x_i \sim N(0, I) \). The marginal distribution for the observed data \( y_i \) is obtained by integrating out the latent variables. From equation 4, an arbitrary rotation matrix \( R^T \) can be applied to the basis, forming the correlated Gaussian,

\[
p(y_i) = \frac{1}{2\pi\sqrt{|D|}} \exp\left( -\frac{1}{2} (Ry_i - R\mu_1)^T D^{-1} R^T (y_i - \mu_i) \right).
\]

This gives an eigenvalue decomposition of the inverse covariance matrix, and thus the covariance matrix,

\[
C^{-1} = RD^{-1}R^T,
\]

\[
C = RRD^TR^T.
\]

As a consequence, we can derive that given \( x_i \sim N(\mu_i, \beta^{-1}) \) and \( y_i = Wx_i \), then the distribution of the observed variables can be denoted as \( y_i \sim N(W\mu_i, WW^T) \). Thus, with a prior as standard Gaussian, \( N(0, I) \),

\[
Wx_i \sim N(0, WW^T),
\]

\[
y_i \sim N(0, C),
\]

where the observation covariance model is \( C = WW^T \beta^{-1}I \), with corresponding log-likelihood [19]

\[
\mathcal{L} = \frac{n}{2}(d\ln(2\pi) + \ln|C| + tr(C^{-1}S)).
\]

\[
S = \frac{1}{n} \sum_{i=1}^{n}(y_i - \mu)(y_i - \mu)^T.
\]

We can write the likelihood for a data point as

\[
p(y_i|x_i, W, \beta) = N(y_i|Wx_i, \beta^{-1}I).
\]

Integrating over the latent variables gives the marginal likelihood,

\[
p(y_i|W, \beta) = \int p(y_i|x_i, W, \beta)p(x_i)dx_i.
\]

As the prior of probabilistic PCA is modelled as a standard Gaussian distribution, \( p(x_i) = N(x_i|0, I) \), marginalisation of the integral obtains the marginal likelihood of each data point as

\[
p(y_i|W, \beta) = N(y_i|0, WW^T + \beta^{-1}I).
\]

Assuming that the data points are independent, the likelihood of the full data set is the product of each marginal likelihood,

\[
p(Y|W, \beta) = \prod_{i=1}^{n} p(y_i|W, \beta).
\]
A. The Principal Subspace of PPCA

Tipping and Bishop showed that all potential solutions for $W$, the likelihood [4], is of the form [20]
\[
W = U_q(K_q - \sigma^2 I)^{\frac{1}{2}} R.
\]
One particular case of interest is when the likelihood is maximised,
\[
W_{ML} = U_q LR, \tag{8}
\]
\[
L = (\Lambda_q - \sigma^2 I)^{\frac{1}{2}}
\]
The matrix $U_q$ contains the column vectors that are the principal eigenvectors, $\Lambda_q = [\lambda_1, ..., \lambda_q]$ represents the diagonal matrix of the corresponding eigenvalues, and $R$ represent an arbitrary orthogonal rotation matrix. Maximising the likelihood of $W$ by equation [8] on the latent variable model defined by equation [5] maps the latent space to the principal subspace of the observed data. Satisfying $W_{ML}$, the latent variable model, is effectively equivalent to standard principal component analysis.

V. GAUSSIAN PROCESS LATENT VARIABLE MODEL

The PCA formulation by Tipping and Bishop allows a probabilistic model, however, it assumes the relation of data is linear. The GP-LVM is a nonlinear latent variable model derived from a dual of the probabilistic PCA by replacing the inner product kernel with Gaussian processes [19]. A nonlinear embedding is a more suitable model for capturing high dimensional data.

A. Dual Probabilistic PCA

The dual probabilistic PCA introduced by Lawrence allow for latent mappings to be non-linearised through the kernel trick by performing computations implicitly in the high dimensional space [25, p.8]. Dual probabilistic PCA marginalises the parameters, $W$, and optimises latent variables, $X$. It is the dual approach of the standard probabilistic PCA where the parameters are optimised and marginalises the latent variables.

First, a conjugate prior to the likelihood of probabilistic PCA [7] is taken to be a spherical Gaussian distribution,
\[
p(W) = \prod_{i=1}^{d} \mathcal{N}(w_i|0, I).
\]
As marginalisation of both $W$ and $X$ is often intractable in practice, $W$ is selected for marginalisation as the conjugate prior is Gaussian distributed, thus, it can be integrated analytically. The marginalised likelihood of $W$ is
\[
p(Y|X, \beta) = \prod_{i=1}^{d} p(y_{i,i}|X, \beta),
\]
The $y_{i,i}$ parameter represents the $i^{th}$ column of $Y$, where
\[
p(y_{i,i}|X, \beta) = \mathcal{N}(y_{i,i}|0, XX^T + \beta^{-1} I). \tag{9}
\]
The objective function is the log-likelihood
\[
L = -\frac{dn}{2} \ln 2\pi - \frac{d}{2} \ln |K| - \frac{1}{2} tr(K^{-1}YY^T), \tag{10}
\]
In the original paper, Lawrence found the gradients of the log-likelihood [10] with respect to $X$ as
\[
\frac{\sigma L}{\sigma X} = K^{-1}YY^TK^{-1}X - dK^{-1}X.
\]
A stationary point where the gradients are zero is given by
\[
\frac{1}{d}YY^TK^{-1}X = X.
\]
The values for $X$ which maximise the likelihood are given by singular value decomposition of $X$.
\[
X = USV^T.
\]
$U$ is a matrix whose orthonormal column vectors are the first eigenvectors of $YY^T$. $\Sigma$ is a diagonal matrix of singular values, whose $i^{th}$ element is $\lambda_i = (\lambda_i - \frac{1}{2})^{-\frac{1}{2}}$, where $\lambda_i$ is the eigenvalue associated with the $i^{th}$ eigenvector $\frac{1}{\beta}YY^T$. $V$ is an arbitrary rotation matrix. Lawrence showed that the eigenvalue problem developed here is equivalent to the eigenvalue problem solved in probabilistic PCA, and thus, dual probabilistic PCA is also effectively equal to standard PCA when the likelihood is maximised. Dual probabilistic PCA assumes that the output dimensions are linear, independent, and identically distributed. Infringing upon these assumptions derive new probabilistic models.

B. Gaussian Processes

A Gaussian process (GP) is a non-parametric statistical model that is a distribution over functions. It treats each observed variable as an independent distribution. A prior probability distribution is an assumption of belief before taking into account of evidence. Observing the output of a continuous function provides information regarding its behaviour around that specific point. For a noiseless model, we can be certain that the input must intersect the point at an observed output, as shown in Figure 2. An observation refines our belief to obtain a posterior. A model with noise can infer that the mapping is nearby the observation, thus treating each observed variable as an independent distribution allows us to take into account of the noise model for prediction.

Formally, a Gaussian process is the infinite generalisation of the Gaussian distribution, specified by a mean function $m(x)$ and covariance function $k(x, x')$ of a real process $f(x)$ [25, p.13],
\[
f(x) \sim \mathcal{GP}(m(x), k(x, x')),
\]
where
\[
m(x) = \mathbb{E}[f(x)],
\]
\[
k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))].
\]
The mean function of a GP is generally assumed to be zero, unless stated otherwise. Polynomial regression models yield best results when the behaviour of the observed data resembles the selected polynomial function, but determining a suitable function is challenging. As a non-parametric model, Gaussian
In Bayesian reasoning, we update our prior knowledge with observed evidence to obtain a posterior. This methodology is expressed in Bayes’ Theorem,

$$p(w|y, X) = \frac{p(y|X, w)p(w)}{p(y|X)}.$$ 

The marginal likelihood is given by

$$p(y|X) = \int p(y|X, w)p(w)dw.$$ 

The posterior expresses what we know about the parameters using the likelihood and the prior. We can then use the posterior to make an informed prediction for test inputs.

The dual probabilistic PCA model uses a Gaussian process prior that is corrupted by Gaussian noise [19], \( \epsilon \sim N(0, \beta^{-1}I) \). The covariance function (kernel) is thus,

$$k(x_i, x_j) = x_i^T x_j + \beta^{-1}\delta_{ij}. \quad (11)$$

Parameters \( x_i \) and \( x_j \) are vectors from the space of inputs to the function and \( \sigma_{ij} \) represents the Kronecker delta, defined by

$$\delta_{ij} = \begin{cases} 1, & \text{if } i = j; \\ 0, & \text{if } i \neq j. \end{cases}$$

Taking inputs from matrix \( X \) and evaluating the covariance function at each observed variable gives the covariance matrix,

$$K = XX^T + \beta^{-1}I.$$ 

The element at \( i^{th} \) row and \( j^{th} \) column of \( K \) is given by the prior distribution \([11]\). Thus, the marginal likelihood of dual probabilistic PCA is a product of \( d \) independent Gaussian processes. The covariance function of a Gaussian process describes the properties of functions, such as variability. Learning in Gaussian processes is to determine hyperparameters of a covariance function that is suitable for the problem being modelled.

VI. BAYESIAN GAUSSIAN PROCESS LATENT VARIABLE MODEL

The Bayesian Gaussian Process Latent Variable Model (Bayesian GP-LVM) [27] extends GP-LVM. The standard GP-LVM method trains by finding the maximum a posteriori (MAP) estimate of \( X \) and jointly maximizing with respect to the hyperparameters. Bayesian GP-LVM performs variational inference to marginalise the latent variables. This method enables optimisation of the resulting lower bound on the marginal likelihood with respect to the hyperparameters.

The marginal likelihood of the observed data is obtained by integrating out the latent variables:

$$p(Y) = \int p(Y|X)p(X)dX.$$ 

Computationally, this integration is intractable in practice. A variational distribution, \( q(X) \), can instead be used to
approximate the posterior distribution over the latent variables, \( p(X|Y) \).

\[
q(X) = \prod_{i=1}^{n} \mathcal{N}(x_i|\mu_i, S_i).
\]

\( \mu_n \) and \( S_i \) are the variational parameters. \( S_i \) is taken as a diagonal covariance matrix. The variational distribution can then be used to obtain a Jensen’s lower bound on \( \log p(Y) \):

\[
F(q) = \int q(X) \log \frac{p(Y|X)p(X)}{q(X)} dX
= \int q(X) \log p(Y|X)p(X) dX - \int q(X) \log \frac{q(X)}{p(X)} dX
= \tilde{F}(q) - KL(q||p).
\]

The \( KL(q||p) \) term is the negative KL divergence between the variational posterior distribution \( q(X) \) and the prior distribution \( p(X) \) over the latent variables. As the KL divergence is uncertain observed data.

A fully marginalised GP-LVM approximated by variational inference establishes a Bayesian perspective that is robust to overfitting, thus applicable to data sets even with missing or uncertain observed data.

VII. SUMMARY

We have explained that dimensionality reduction with LVMs aspires to explain the underlying behavior of observed data. Dimensionality reduction was introduced through PCA and was formulated a probabilistic framework. We looked at the derivation of nonlinear latent variable models such as GP-LVM and its variational Bayesian counterpart.

Now that we have established a foundation for probabilistic latent variable models, we will look at the advantages of use GP-LVM and why BGP-LVM was introduced. First, 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. Further, besides widely used in dimensionality reduction, GP-LVM 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, we can also exert classical kernel methods for a further expansion of GPLVM [20].

REFERENCES