Mixtures of Probabilistic Principal Component Analysers

Review

Ayoub Ghriss, Romain Girard, Alexis Thual

1 PCA and Latent Variable Models

1.1 Principal Component Analysis (PCA)

1.1.1 Short introduction to PCA

Principal Component Analysis (PCA) is a popular dimensionality reduction technique, that can be exploited in a number of applications, including density estimation, regression, and classification. PCA is based on the suspicion that a high-dimensional observed random vector could in fact be well represented by a lower-dimensional vector.

Let us formalize PCA in the following way. Given observed points $\{t_n\}, n \in \{1 \dots N\}$, of dimension d, it aims at finding a number $q < d$ of orthonormal axes (thus forming a linear subspace of dimension $q < d$) such that the variance of the projection of the observed vectors onto this subspace is maximal. The idea is that the directions along which the variance of the observed data is maximal are those which carry the most information about the individual observations, and should therefore be preserved as such to discriminate the observations. On the other hand, the directions along which the observed variance is minimal give little information about the individual observations: all observed vectors are ”roughly the same” along this direction, hence this axe rather gives information about the structure of the problem; this information can be stored in the form of the lower-dimensional linear subspace, and forgotten at the level of individual observations.

The q orthonormal axes retained are called *principal axes*, noted w_j (forming the matrix W), forming the *principal subspace*. Those principal axes correspond to the q largest eigenvectors (i.e. the q eigenvectors associated with the largest eigenvalues) of the sample covariance matrix $\mathrm{S} = N^{-1} \sum_{n=1}^{N} (t_n - \bar{t})(t_n - \bar{t})^T$. This yields a q-dimensional representation of t_n by $x_n = W^T(t_n - \bar{t})$ where $W = (w_1,\ w_2, \dots, w_q)$. x_n can be shown to be the optimal q-dimensional representation of t_n in the reconstruction error sense, that is $\hat{t}_n = Wx_n + \bar{t}$ minimises $\sum \|t_n - \hat{t}_n\|^2$.

1.1.2 Main drawbacks of such presentation

Despite its indisputable efficiency and its wide use and range of applications, standard PCA doesn't come with an associated generative model, or, in other words, with a probabilistic interpretation. A probabilistic view of PCA would indeed allow, for example, likelihood computation, Bayesian comparison of several models, and finer modeling than a strictly deterministic view. Such a model would also, thanks to its flexibility, lead to a possible mixture of Principle Component Analyzers, compensating for the linear (thus limited) decomposition offered by standard PCA. The idea would be to model nonlinear structures with a mixture of locally linear components (here PCA), which seems like a fair thing to aim at. The main result of the article is a derivation of a probabilistic model for Principal Component Analysis, which yields a natural extension to a mixture of local PCA components. Such representation would for instance allow using an EM algorithm presented during the course to find the associated parameters of our model.

1.2 Factor Analysis

1.2.1 Introduction to Factor Analysis

Factor Analysis (FA) can be seen to some extent as the probabilistic couterpart of standard PCA. FA naturally arises as a simple example of a latent variable model (that is a mapping $t = y(x; W)$) following the following hypothesis:

$$t = Wx + \mu + \epsilon$$

FA furthermore assumes the latent variables to be independent and Gaussian $x \sim \mathcal{N}(0, I)$, and the noise to be Gaussian as well $\epsilon \sim \mathcal{N}(0, \Psi)$, with Ψ diagonal. The matrix W, of size $d \times q$, is called the factor loadings matrix. Note that with this model, the vector t follows a Gaussian distribution $t \sim \mathcal{N}(\mu, \Psi + WW^T)$. A major property of this model is that, given the diagonality of Ψ, the observed variables t are conditionally independent given the latent variables x, which greatly simplifies the analysis.

1.2.2 Link with PCA

Although the link is not obvious, PCA can be shown to be a limiting case of Factor Analysis (in the same way that the k-means algorithm is a limiting case of the analysis of a Gaussian Mixture Model). Let us note first an important difference between FA and PCA: there is no closed form solution for W and Ψ, hence we need to estimate those with iterative algorithms.

In general, W_{FA} does not correspond to the subspace spanned by principal axes, whereas W_{PCA} does. However, the authors show in this paper that PCA arises from Factor Analysis (that is, the maximum likelihood estimator for W_{FA} is W_{PCA}) when $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, that is in the *homoscedastic residuals model.*

2 Probabilistic Principal Component Analysers

2.1 Introducing a generative model

Let's introduce the latent variable model which relates a d-dimensional data vector t to a corresponding q-dimensional latent variables x with $q < d$, for isotropic noise $\epsilon \sim \mathcal{N}(0,\ \sigma^2\mathrm{I})$:

$$\mathrm{t} = W\mathrm{x} + \mu + \epsilon \tag{1}$$

Such model indeed seems quite intuitive: W represents the linear surface one will project on, μ it's shift to the origin and ϵ allows some noise around the surface.

One can derive the distribution $p(\mathrm{t}|\mathrm{x})$ as shown in equation 2 from equation 1 and eventually obtains the marginal distribution $p(\mathrm{t})$ in equation 3 making the assumption that the prior x is gaussian.

$$p(\mathrm{t}|\mathrm{x}) = (2\pi\sigma^2)^{-d/2} \exp\Big(- \frac{1}{2\sigma^2} \left\| \mathrm{t} - W\mathrm{x} - \mu \right\| \Big) \tag{2}$$

$$\begin{aligned} p(\mathrm{t}) &= \int p(\mathrm{t}|\mathrm{x}) p(\mathrm{x}) d\mathrm{x} \\ &= (2\pi)^{-d/2} \det(\Sigma^{-1/2}) \exp\Big(- \frac{1}{2} (\mathrm{t} - \mu)^T \Sigma^{-1} (\mathrm{t} - \mu) \Big) \end{aligned} \tag{3}$$

where $\Sigma = \sigma^2 I + WW^T$

We then obtain $p(\mathrm{x}|\mathrm{t})$ using Bayes rule. Hence, the log-likelihood formulation :

$$\mathcal{L} = -\frac{N}{2} \{ d \ln(2\pi) + \ln|\Sigma| + \mathrm{tr}(\Sigma^{-1}\mathrm{S}) \} \tag{4}$$

Where where $M = \sigma^2 I + W^T W$

2.2 Algorithmic methods based on Probabilistic PCA

2.2.1 Parameters evaluation

Such representation allows one to see PCA as analogous to a simple likelihood maximisation. Indeed, given a sample $(\mathrm{t}_n)_n$, one can estimate the model parameters μ, W and σ using usual maximum-likelihood estimators.

The article proves that the ML estimator of W is a function of the first q dominant eigenvalues and eigenvectors of the covariance matrix of observed $(\mathrm{t}_n)_n$. Furthermore, the ML estimator of σ is a function of the remaining $d - q$ eigenvalues.

2.2.2 Implementing a mixture of PCAs

In the case of high dimensional data $(\mathrm{t}_n)_n$, it is only natural to think of the EM algorithm in order to try and estimate the parameters of our various linear components.

We can develop an iterative EM algorithm for optimisation of all of the model parameters $\pi_i, \mu_i, \mathrm{W}_i$ and σ_i^2. If $R_{n,i} = p(t_n, i)$ is the posterior *responsibility* of mixture i for generating data point t_n, given by

$$R_{n,i} = \frac{p(\mathrm{t}_n|i)\pi_i}{p(\mathrm{t}_n)}$$

The EM iteration, where the overlined notation is used for the updated parameters is as follows :

$$\overline{\pi_i} = \frac{1}{N} \sum_{n=1}^{N} R_{n,i},\ \overline{\mu}_i = \frac{\sum_{n--1}^{N} R_{n,i} \mathrm{t}_n}{\sum_{n=1}^{N} R_{n,i}},\ \overline{\mathrm{S}}_i = \frac{1}{\overline{\pi}_i N} \sum_{n=1}^{N} R_{n,i} (\mathrm{t}_n - \overline{\mu}_i)(\mathrm{t}_n - \overline{\mu}_i)^{\mathrm{T}}$$

Based on S_i we can obtain σ_i and W_i as in the standard PCA.

3 Applications and Results

3.1 VQPCA and Mixture of PPCA

Several examples illustrate the performance of PPCA. The algorithm is first tested on a synthetic dataset comprising 500 points uniformly sampled on an hemisphere with gaussian noise. 12 PCA are then fitted on the data: every point t_n is softly assigned to mixture components according to their responsibility $R_{n,i} \propto p(t_n|i)$. PPCA is compared to Vector-Quantization-PCA (VQPCA), which assigns points to the closer of 12 randomly-initiated clusters, performs a PCA on each of these clusters, recomputes the cluster centres and iterates until the allocations are constant.

Comparison between the two algorithms is made through the overall reconstruction error. Even though VQPCA is explicitly designed to minimise this quantity, PPCA outperforms VQPCA on some datasets and generally performs comparably to VQPCA. Moreover, it seems to overfit less than VQPCA.

3.2 Image Compression

PPCA was then tested on image compression: an image was partitioned in patches of 8x8 pixels and PCA, VQPCA and PPCA were applied on this set of patches. PPCA visually outperforms PCA and yields overall reconstruction error comparable to VQPCA. Finally, as q controls the number parameters in our fitted model, PPCA is particularly suited for problems which require a lot of parameters - such as noisy spiral data or handwritten digit recognition, where PPCA performs well.

3.3 Handwritten Digit Recognition

As detailed in the article, we implemented a mixture of Probabilistic PCA with $q = 10$: for each test observation t_j we calculate $p(t_j/i)$ the probability of being generated by the PPCA i. By assigning the digit to most likely PCA, we could reach a classification error of 7% using a training set of 500 digit and a test set of 10^4 digit.

Furthermore, combined with clustering for each digit, we can could achieve an error of 5% by introducing 2 clusters per digit, and assigning the test digit to the most likely cluster.

4 Conclusion

By developing a probabilistic view of Principle Component Analysis, the method described in the reviewed paper addresses the gap between Factor Analysis and standard PCA. Showing that PCA can be obtained as a limiting case of Factor Analysis, the authors enable the elaboration of Probabilistic PCA, and of mixture models of such Principle Component Analysers, thus bringing a more flexible estimator that has a wide range of applications.