Introduction to Machine Learning

Lecture 11 - Unsupervised Learning: Principal Component Analysis

Guang Bing Yang, PhD

Introduction to Continuous Latent Variables

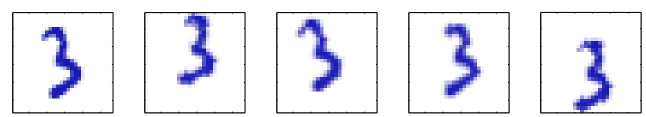
- Continuous Latent Variables
- Principal Component Analysis
- Factor Analysis

Continuous Latent Variables

- Previously discussed discrete latent variables, such as mixture of Gaussians.
- Sometimes, it is more appropriate to think in terms of continuous factors which control the data we observe.
- This motivation for such models is:
  - for many datasets, data points lie close to a manifold of much lower dimensionality compared to that of the original data space.
  - Training continuous latent variable models often called **dimensionality reduction,** since there are typically many fewer latent dimensions.
- Examples include:
  - Principal Components Analysis,
  - Factor Analysis,
  - Independent Components Analysis

Intrinsic Latent Dimensions

- In this dataset, there is only 3 degrees of freedom of variability, vertical and horizontal translations, and the rotations.



Copied from the book: Pattern Recognition and Machine Learning by Christopher M. Bishop

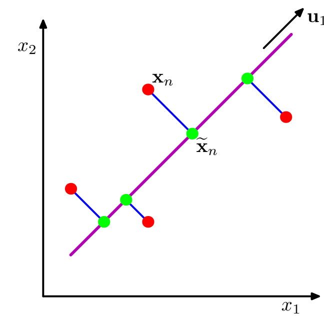
- Each image undergoes a random displacement and rotation within some larger image field.
- The resulting images have 100 x 100 = 10,000 pixels.
- However, the data points live on a subspace of the data space whose intrinsic dimensionality is three.

Generative View

- To generate the data points, first selecting a point from a distribution in the latent space, then sampling a point from the conditional distribution in the input space
- Taking Gaussian distribution for both latent and observed variables is the simplest latent variable model.
- This generative approach leads to probabilistic formulation of the **Principal Component Analysis** and **Factor Analysis**.
- Take look at the standard PCA, and then consider its probabilistic formation
- Mixture of PCAs, Bayesian PCA are advantages of using EM for parameters.

Principal Component Analysis

- Widely used for data compression, visualization, feature extraction, and dimensionality reduction
- The goal is find M principal components underlying D-dimensional data
  - Select the top M eigenvectors of S (data covariance matrix): $\{u_1, \dots u_M\}$
  - project each input vector X into this subspace, e.g., $z_{n1} = X_n^T u_1$
  - The full projection into M dimensions:
  - $\begin{bmatrix} \mathbf{u}_1^\top \\ \cdots \\ \mathbf{u}_M^\top \end{bmatrix} [\mathbf{x}_1 \cdots \mathbf{x}_N] = [\mathbf{z}_1 \cdots \mathbf{z}_N]$



Maximum Variance

- Consider a dataset $\{x_1, \dots, x_N\}, x_n \in R^D$. Our goal is to project data onto a space having dimensionality M < D.
- Consider the projection into M=1 dimensional space
- Define the direction of this space using a D-dimensional unit vector u1, so that $u_1^T u_1 = 1$
- Objective is to maximize the variance of the projected data w.r.t. u_1
  - $\frac{1}{N}\sum_{n=1}^{N}\{u_1^T x_n - u_1^T \bar{x}\}^2 = u_1^T S u_1,$
  - where sample mean and data covariance:
    - $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$
    - $S = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})(x_n - \bar{x})^T$

Maximum Variance

- To maximize the variance of the projected data, use Language multiplier with constrain $||u_1|| = 1$
- maximum: $u_1^T S u_1 + \lambda(1 - u_1^T u_1)$
- setting the derivative w.r.t. u_1 to zero:
- $Su_1 = \lambda_1 u_1$
- Hence u_1 must be an eigenvector of S, and
- the maximum variance of the projected data is given by: $u_1^T S u_1 = \lambda_1,$
- Optimal u_1 is the principal component (eigenvector with maximal eigenvalue)

Minimum Error

- Introduce a complete orthonormal set of D-dimensional basis vectors:
  - $\{u_1, \ldots, u_D\}, \ u_i^T u_j = \delta_{ij},$
  - let $x_n = \sum_{i=1}^{D} \alpha_{ni} u_i, \ \alpha_{ni} = x_n^T u_i$, rotation of the coordinate system to a new system defined by u_i
- This enables the data to be represented by the projection into M-dimensional subspace as (represent M-dimensional linear subspace by the first M of the basis vectors):
  - $\tilde{x}_n = \sum_{i=1}^{M} z_{ni} u_i + \sum_{i=M+1}^{D} b_i u_i$

Minimum Error

- For the $\tilde{x}_n = \sum_{i=1}^{M} z_{ni} u_i + \sum_{i=M+1}^{D} b_i u_i$, where z_{ni} depend on the particular data point and b_i are constants.
- To minimize the distortion w.r.t. u_i, z_{ni}, and b_i
  - $J = \frac{1}{N} \sum_{n=1}^{N} ||x_n - \tilde{x}_n||^2$
  - minimize J w.r.t. z_{ni} and b_i: $z_{nj} = x_n^T u_j, \ \ b_j = \bar{x}^T u_j$
  - Then the objective J reduces to:, $J = \frac{1}{N} \sum_{n=1}^{N} \sum_{i=M+1}^{D} (x_n^T u_i - \bar{x}^T u_i)^2 = \sum_{i=M+1}^{D} u_i^T S u_i$

Minimum Error

- The general solution is obtained by choosing u_i to be eigenvectors of the covariance matrix:
  - $Su_i = \lambda_i u_i$
- The distortion is then given by: $J = \sum_{i=M+1}^{D} \lambda_i$
- The M components are the eigenvectors of S with lowest eigenvalues when objective J is minimized.

Applications of PCA

- Apply PCA on 2429 19 x 19 grayscale images from CBCL database

- with only 3 components
- For pre-processing, PCA with 3 components obtains 79% accuracy on face/non-face discrimination test vs. 76.8% for mixture of Gaussian with 84 components.

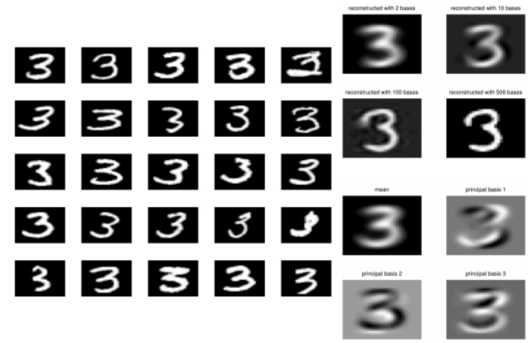
Applications of PCA

- Apply PCA on 2429 19 x 19 grayscale images from CBCL database
- good for “eigenfaces”

Eigenfaces — PCA on 3 components

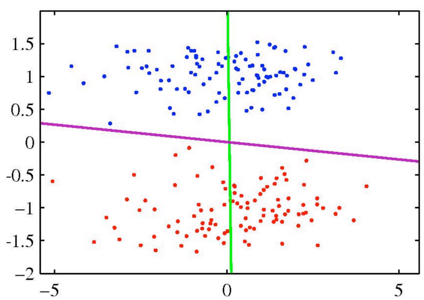
Applications of PCA

- Apply PCA on digit images



PCA vs. Fisher’s LDA

- Both PCA and Fisher’s LDA can be apply for linear dimensionality reduction.
- PCA chooses direction of maximum variance using unsupervised approach.
- The magenta curve shows a strong class overlap
- LDA uses supervised way with the class labels.
- The green curve shows a projection into it.



yguangbing@gmail.com, Guang.B@chula.ac.th Apr 9th, 2021

PCA for High-Dimensional Data

- In some applications of PCA, the number of data points is smaller than the dimensionality of the data space, i.e. N<D.
- To find the eigenvectors of the D x D data covariance matrix S, the computation expense is $O(D^3)$.
- Thus, direct application of PCA is often computationally infeasible.
- To solve this disadvantage, here is a solution:
  - Let X be the N x D centered data matrix. The corresponding eigenvector equation becomes: $\frac{1}{N}X^TXu_i = \lambda_i u_i$

yguangbing@gmail.com, Guang.B@chula.ac.th Apr 9th, 2021

PCA for High-Dimensional Data

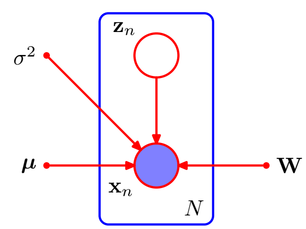
- Pre-multiply by X: $\frac{1}{N}XX^TXu_i = \lambda_i(Xu_i)$
- Let $v_i = Xu_i$, hence $\frac{1}{N}XX^Tv_i = \lambda_i v_i$
- This is an eigenvector equation for the N x N matrix.
  - $O(N^3) << O(D^3)$.
- To determine eigenvectors, multiply by X^T: $(\frac{1}{N}XX^T)(X^Tv_i) = \lambda_i X^T v_i$
- Hence, X^Tv_i is an eigenvector of S with eigenvalue λ_i

Probabilistic PCA

- Many advantages of Probabilistic PCA (PPCA):
  - It represents a constrained form of the Gaussian distribution.
  - Able to derive EM algorithm for PCA which is computationally efficient.
  - PPCA can deal with missing values in the data set.
  - Mixture of PPCAs can be formulated in a principled way.
  - PPCA forms the basis for a Bayesian PCA, in which the dimensionality of the principal subspace can be determined from the data.
  - The existence of a likelihood function allows direct comparisons with other probabilistic density models
  - PPCA can be used to model class conditional densities and hence it can be applied to classification problems.

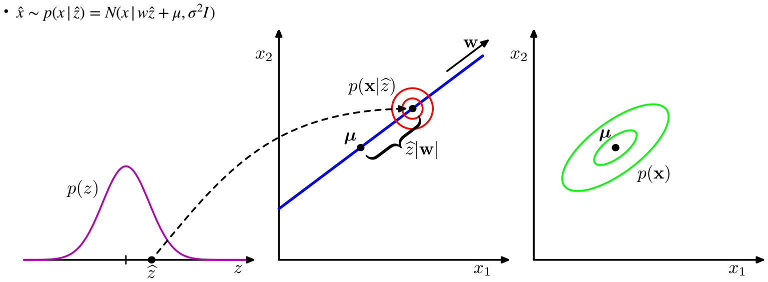
Probabilistic PCA

- Assumptions to formulate the PPCA:
  - underlying latent M-dim variable z has a Gaussian distribution.
  - linear relationship between M-dim latent z and D-dim observed x variables.
  - isotropic Gaussian noise in observed
  - $p(z) = N(z\,|\,0, I)$
  - $p(x\,|\,z) = N(x\,|\,Wz + \mu, \sigma^2 I)$
- The mean of x is a linear function of z governed by the D x M matrix W and the D-dim vector μ.
- The columns of W span the principal subspace of the data space (Columns of W are the principal components, σ^2 is sensor noise).



Probabilistic PCA

- Draw a value of the latent variable from its prior distribution:
  - $\hat{z} \sim p(z)$
- Draw a value for x from from an isotropic Gaussian distribution:
  - $\hat{x} \sim p(x\,|\,\hat{z}) = N(x\,|\,w\hat{z} + \mu, \sigma^2 I)$



Generative view of the PPCA for a 2-d data space and 1-d latent space

Marginal Data Density

- The joint p(z, x), the marginal data distribution p(x) and the posterior p(z|x) are Gaussians
- $p(X) = \int_z p(z)p(x|z)dz = N(x|\mu, WW^T + \sigma^2 I), \;\; x = Wz + \mu + \epsilon$
- This is the marginal data density, also known as predictive distribution.
- By computing mean and covariance of Gaussian:
- $E[x] = E[\mu + Wz + \epsilon] = \mu + WE[z] + E[\epsilon] = \mu + W0 + 0 = \mu$
- $$\begin{aligned} C = \mathrm{Cov}[x] & \\ &= E[(x-\mu)(x-\mu)^T] \\ &= E[(\mu + Wz + \epsilon - \mu)(\mu + Wz + \epsilon - \mu)^T] \\ &= E[(Wz+\epsilon)(Wz+\epsilon)^T] \\ &= WW^T + \sigma^2 I \end{aligned}$$

yguangbing@gmail.com, Guang.B@chula.ac.th Apr 9th, 2021

Redundancy in Parameterization

- Given the marginal distribution:
  - $p(X) = \int_z p(z)p(x|z)dz = N(x|\mu, WW^T + \sigma^2 I), \;\; x = Wz + \mu + \epsilon$
- Let R be an orthogonal matrix, then define a new matrix:
  - $\hat{W} = WR, RR^T = I$
  - $\hat{W}\hat{W}^T = WRR^TW^T = WW^T,$
  - the redundancy in parameterization as if rotating the latent space coordinates.
- Hence, there is a whole family of matrices that all of which give rise to the same marginal distribution when rotating within the latent space.

yguangbing@gmail.com, Guang.B@chula.ac.th Apr 9th, 2021

Joint Density for PPCA

- The joint density for PPCA is given as:

$$p(\begin{bmatrix} \mathbf{z} \\ \mathbf{x} \end{bmatrix}) = \mathcal{N}(\begin{bmatrix} \mathbf{z} \\ \mathbf{x} \end{bmatrix} \mid \begin{bmatrix} 0 \\ \mu \end{bmatrix}, \begin{bmatrix} I & \mathbf{W}^\top \\ \mathbf{W} & \mathbf{W}\mathbf{W}^\top + \sigma^2 \mathbf{I} \end{bmatrix})$$

  - and covariance:

$$\begin{aligned} Cov[\mathbf{z}, \mathbf{x}] &= E[(\mathbf{z} - 0)(\mathbf{x} - \mu)^T] = E[\mathbf{z}(\mu + \mathbf{W}\mathbf{z} + \epsilon - \mu)^T] \\ &= E[\mathbf{z}(\mathbf{W}\mathbf{z} + \epsilon)^T] = \mathbf{W}^T \end{aligned}$$

Reduce O(D^3) to O(M^3) by applying matrix inversion lemma:

$$C^{-1} = \sigma^{-1}I - \sigma^{-2}W(W^TW + \sigma^2 I)^{-1}W^T$$

yguangbing@gmail.com, Guang.B@chula.ac.th Apr 9th, 2021

Posterior Distribution for PPCA

- The posterior distribution is about the inference problem in PPCA:
  - $p(z|x) = N(z|m, V)$
  - $m = M^{-1}W^T(x - \mu)$
  - $V = \sigma^2 M^{-1}$
  - $M = W^TW + \sigma^2 I$
- Mean of inferred z is projection of centred x, a linear operation.
- The posterior variance does not depend on the input x at all
- Since $C^{-1} = \sigma^{-1}I - \sigma^{-2}W(W^TW + \sigma^2 I)^{-1}W^T$, and $C = WW^T + \sigma^2 I$.

yguangbing@gmail.com, Guang.B@chula.ac.th Apr 9th, 2021

Maximum Likelihood

- Using maximum likelihood (by integrating out latent variables) to determine the model parameters:

$$\begin{aligned} L(\theta;\mathbf{X}) &= \log p(\mathbf{X}|\theta) = \sum_n \log p(\mathbf{x}_n|\theta) \\ &= -\frac{N}{2}\log|\mathbf{C}| - \frac{1}{2}\sum_n (\mathbf{x}_n-\mu)\mathbf{C}^{-1}(\mathbf{x}_n-\mu)^T \\ &= -\frac{N}{2}\log|\mathbf{C}| - \frac{1}{2}Tr[\mathbf{C}^{-1}\sum_n (\mathbf{x}_n-\mu)(\mathbf{x}_n-\mu)^T] + \text{const} \end{aligned}$$

- Let $\mu_{ML} = \bar{x}$, then
- $\log p(X|\theta) = -\frac{N}{2}\log|C| - \frac{1}{2}Tr[C^{-1}S] + \text{const}$

yguangbing@gmail.com, Guang.B@chula.ac.th Apr 9th, 2021

Maximum Likelihood

- Maximizing w.r.t. W: $W_{ML} = U_M(L_M - \sigma^2 I)^{1/2}R$
- where
  - $U_M \in R^{D\times M}$ matrix whose columns are given by the M principal eigenvectors of the data covariance matrix S.
  - $L_M \in R^{M\times M}$ diagonal matrix containing M largest eigenvalues.
  - $R \rightarrow M \times M$ an arbitrary orthogonal matrix.

If the eigenvectors have been arranged in the order of decreasing values of the corresponding eigenvalues, then the columns of W define the principal subspace of standard PCA.

yguangbing@gmail.com, Guang.B@chula.ac.th Apr 9th, 2021

Maximum Likelihood

- Maximizing w.r.t. σ^2: $\sigma^2_{ML} = \frac{1}{D-M}\sum_{i=M+1}^{D}\lambda_i$
- which is the average variance associated with the discarded dimensions

yguangbing@gmail.com, Guang.B@chula.ac.th Apr 9th, 2021

EM for PPCA

- Instead of solving directly, we can use EM. The EM can be scaled to very large high-dimensional datasets.
- The complete-data log-likelihood takes form:
  - $\log p(X, Z|\mu, W, \sigma^2) = \sum_n [\log p(x_n|z_n) + \log p(z_n)]$
- In E-step:
  - compute expectation of complete log likelihood w.r.t. z, using the current parameters.
  - Need to derive $E[z_n]$, $E[z_n z_n^T]$ w.r.t. the true posterior: $p(z|x)$

yguangbing@gmail.com, Guang.B@chula.ac.th Apr 9th, 2021

EM for PPCA

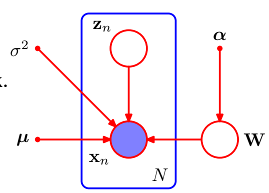
- In M-step:
  - Maximize w.r.t. parameters W and σ^2
  - EM avoids direct $O(ND^2)$ construction of covariance matrix.
  - Instead EM involves sums over data cases: O(NDM). It can also be implemented online, without storing data.

EM for PPCA

- It is able to derive standard PCA as a limit of probabilistic PCA as the noise term goes to zero: $\sigma^2 \to 0$
- Maximum likelihood parameters are the same.
- Inferring the distribution over latent variables is easier: The posterior mean reduces to:
  - $\lim_{\sigma^2 \to 0} (W^T W + \sigma^2 I)^{-1} W^T (x - \mu) = (W^T W)^{-1} W^T (x - \mu)$
- which represents an orthogonal projection of the data point onto the latent space – standard PCA.
- Posterior covariance goes to zero

Bayesian PCA

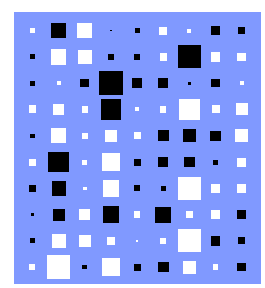
- A Bayesian viewpoint and place priors over model parameters is given as:
  - define an independent Gaussian prior over each column of W.
  - employ the evidence approximation (empirical Bayes) framework.
- Such Gaussian has an independent variance
  - $p(W|\alpha) = \prod_{i=1}^{M} (\frac{\alpha_i}{a\pi}) \exp[\frac{1}{2}\alpha_i W_i^T W_i]$
  - where w_i is the with column of W.

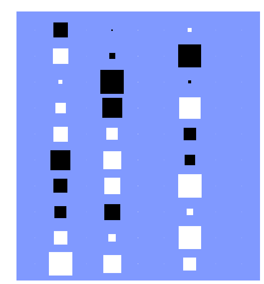


Bayesian PCA

- The values of α_i are re-estimated during training by maximizing the marginal likelihood.
  - $p(X|\alpha, \mu, \sigma^2) = \int p(X|W, \mu, \sigma^2) p(W|\alpha) dW$

1. Hinton diagram of the matrix W: each element of W is depicted as a square (white for positive and black for negative)
2. Bayesian PCA discovers appropriate dimensionality


PPCA


Bayesian PCA

Factor Analysis

- Use a linear Gaussian latent variable model which is related to PPCA.
- Assume:
  - underlying latent M-dim variable z has a Gaussian distribution
  - linear relationship between M-dim latent z and D-dim observed X variables
  - diagonal Gaussian noise in observed dimensions given as:
    - $p(z) = N(z \mid 0, I)$
    - $p(x \mid z) = N(x \mid Wz + \mu, \Psi)$
    - W is a D x M factor loading matrix
    - $\Psi \rightarrow M \times M$ diagonal matrix
- The only difference between PPCA and FA is that in Factor Analysis the 20 conditional distribution of the observed variable x has diagonal rather than isotropic covariance.

Factor Analysis

- Same as the PPCA, the joint p(z, x), the marginal p(x) and the posterior p(z|x) are Gaussians.
- Marginal distribution: $p(X) = \int_z p(z)p(x \mid z)dz = N(x \mid \mu, WW^T + \Psi)$
- The joint distribution:

$$p(\begin{bmatrix} \mathbf{z} \\ \mathbf{x} \end{bmatrix}) = \mathcal{N}(\begin{bmatrix} \mathbf{z} \\ \mathbf{x} \end{bmatrix} \mid \begin{bmatrix} 0 \\ \mu \end{bmatrix}, \begin{bmatrix} I & \mathbf{W}^\top \\ \mathbf{W} & \mathbf{W}\mathbf{W}^\top + \Psi \end{bmatrix})$$

Use EM to solve the parameters.

PCA and FA

Recap

- Introduced continuous latent variables and the most important applications: the principal component analysis, or PCA
- PCA is widely applied in dimensionality reduction, lossy data compression, feature extraction, and data visualization, and more.
- PCA can be defined as the orthogonal projection of the data onto a lower dimensional linear space, known as the principal subspace, such that the variance of the projected data is maximized, Or,
- It can be defined as the linear projection that minimizes the average project cost, defined as the mean squared distance between the data points and their projections.

PCA and FA

Recap

- By maximize the projected variance, the data have reduced dimensionality M < D
- Where the eigenvector with the highest eigenvalue is the first principal component, and hence the second one having the second largest eigenvalue, and so on so forth
- These eigenvectors with highest values of eigenvalues are representation of the observed data X in D- dimension, the representatives have M- dimension
- Hence, PCA can reduce the data dimensionality, and compress the data with loss.
- Factor analysis is a linear-Gaussian latent variable model that is closed related to probabilistic PCA.

Questions and lab