Probabilistic Principal Component Analysis and the E-M algorithm

The Minh Luong

CS 3750

October 23, 2007

Outline

- Probabilistic Principal Component Analysis
  - Latent variable models
  - Probabilistic PCA
    - Formulation of PCA model
    - Maximum likelihood estimation
      - Closed form solution
      - EM algorithm
        - » EM Algorithms for regular PCA
        - » Sensible PCA (E-M algorithm for probabilistic PCA)
  - Mixtures of Probabilistic Principal Component Analysers

Review of PCA

- Primary uses:
  - Analyze data and extract variables with similar concepts (principal components)
  - Project the data onto a lower dimensional space
  - Principal components which explain a greater amount of the variance are considered to be more important
- Accomplishes this by:
  - Maximizing variance of the projected data $\boldsymbol{x}$
  - Represent matrix $\boldsymbol{x}$ in a different (q-dimensional) space using a set of orthonormal vectors $\boldsymbol{W}$
    - Weight matrix $\boldsymbol{W}$ is a $d \times q$ matrix that represents a re-mapping of original data $\boldsymbol{y}$ into its "ideal" principal subspace, represented by $\boldsymbol{x}$
    - Each of q orthonormal columns of the weight matrix $\boldsymbol{W}$, $\boldsymbol{w_i}$, represents a separate principal component
  - Likelihood of a point in $\boldsymbol{y}$ is the distance^2 between it and its reconstruction, $\boldsymbol{Wx}$

Limitations of PCA

- Non-parametric
  - no probabilistic model for observed data
- The variance-covariance matrix needs to be calculated
  - Can be very computation-intensive for large datasets with a high # of dimensions
- Does not deal properly with missing data
  - Incomplete data must either be discarded or imputed using ad-hoc methods
- Outlying data observations can unduly affect the analysis

Motivation behind probabilistic PCA

- Addresses limitations of regular PCA
- PCA can be used as a general Gaussian density model in addition to reducing dimensions
- Maximum-likelihood estimates can be computed for elements associated with principal components
- Captures dominant correlations with few parameters
- Multiple PCA models can be combined as a probabilistic mixture
- Can be used as a base for Bayesian PCA

Latent variable models

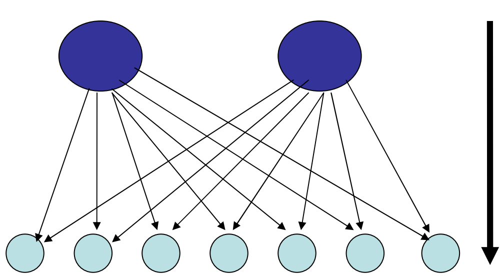
- Latent variable(s): unobserved variable (s)
  - Offer a lower dimensional representation of the data and their dependencies
- Latent variable model:
  - ***y:*** observed variables (*d*-dimensions)
  - ***x:*** latent variables (*q*-dimensions)
  - $q<d$
- Less dimensions results in more parsimonious models

Probabilistic PCA (PPCA)

- Latent variable model with linear relationship (factor analysis)
  - $\boldsymbol{y} \sim \boldsymbol{Wx} + \boldsymbol{\mu} + \boldsymbol{\varepsilon}$
  - Latent variables: $\boldsymbol{x} \sim N(0, I)$
  - Error (or noise): $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\psi})$
  - Location term (mean): $\boldsymbol{\mu}$
- Probabilistic PCA: Noise variances constrained to be equal ($\psi_i=\sigma^2$)
    - Error: $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I)$ (isotropic noise model)
  - $\boldsymbol{y}|\boldsymbol{x} \sim N(\boldsymbol{WX} + \boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$
  - $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{C_y})$, where $\boldsymbol{C_y}=\boldsymbol{WW^T} + \sigma^2\boldsymbol{I}$ (where $\boldsymbol{C_y}$ is the covariance matrix for the observed data $\boldsymbol{y}$)
- Normal PCA is a limiting case of probabilistic PCA, taken as the limit as the covariance of the noise becomes infinitesimally small ($\boldsymbol{\psi} = \lim_{\sigma^2 \to 0} \sigma^2 \boldsymbol{I}$)

Illustration of probabilistic PCA

Latent variables (x) q = 2
(hidden variables, underlying concepts)



Observed variables (y) d = 7
(data)

$\boldsymbol{x \sim N(0, I)}$

Remapping: Wx
(Weight matrix: W)

+

μ (location parameter)

+

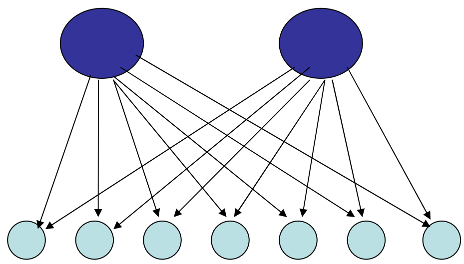
Random error (noise): ε
ε ~ N(0, σ²I)

$\boldsymbol{y = Wx + \mu + \varepsilon}$
$\boldsymbol{y \sim N(\mu, WW^T + \sigma^2 I)}$

Parameters of interest: W (weight matrix), σ^2 *(variance of noise)*

Illustration of probabilistic PCA

Latent variables (x) q = 2
(hidden variables, underlying concepts)



Note: Observed variables become independent of each other given latent factors

Observed variables (y) d = 7
(data)

PPCA (Maximum likelihood PCA)

- Log-likelihood for Gaussian noise model:
  - $LL = -N/2\ \{d \ln(2\pi) + \ln|\mathbf{C}_y| + \mathrm{tr}(\boldsymbol{C}^{-1}_y \boldsymbol{S})\}$
    - $\boldsymbol{C}_y = \boldsymbol{W}\boldsymbol{W}^T + \boldsymbol{\sigma}^2$
- Maximum likelihood estimates for above:
  - $\boldsymbol{\mu}$**:** mean of the data
  - $\boldsymbol{S}$ (sample covariance matrix of the observations $\boldsymbol{Y}$):
    - $\boldsymbol{S} = (1/N) \sum_{n=1}^{N} (Y_n - \mu)(Y_n - \mu)^T$
- MLE's for $\boldsymbol{W}$ and $\boldsymbol{\sigma}^2$ can be solved in two ways:
  - closed form (Tipping and Bishop)
  - EM algorithm (Roweis)

Tr(A) = sum of diagonal elements of A

MLE's for probabilistic PCA (closed form)

- Likelihood of LL is maximized with respect to $\boldsymbol{W}$ and $\boldsymbol{\sigma^2}$, MLE's can be obtained in closed form:
  - $$\sigma^2_{\mathrm{ML}} = \frac{1}{d-q} \sum_{j=q+1}^{d} \lambda_j$$
    - Represents the variance lost in the projection, averaged over the # dim decreased
  - $\boldsymbol{W_{ML} = U_q\, (\Lambda_q - \sigma^2 I)^{1/2}\, R}$
    - Represents the mapping of the latent space (containing $\boldsymbol{X}$) to that of the principal subspace (containing $\boldsymbol{Y}$)
    - Columns of $\boldsymbol{U_q}$ (d x q matrix): principal eigenvectors of **S**
    - $\boldsymbol{\Lambda}_q$ (q x q diagonal matrix): corresponding eigenvalues $\boldsymbol{\lambda_{1..q}}$
    - $\boldsymbol{R}$: q x q arbitrary rotation matrix (can be set to R=I)

Derivation of MLEs

- $LL = -N/2\ \{d \ln(2\pi) + \ln|C_y| + \mathrm{tr}(C^{-1}_y S)\}$

The 1st derivative of LL w/ respect to $\boldsymbol{W}$:

- dL/dW =N$\boldsymbol{(C^{-1}SC^{-1}W - C^{-1}W)}$, where $\boldsymbol{W = ULV^T} = \sigma^2 \boldsymbol{I} + \boldsymbol{WW^T}$
- The stationary points are $\boldsymbol{SC^{-1}W = W}$.
- Non-trivial case: $\boldsymbol{W \neq 0,\ C \neq S}$
- SVD: $\boldsymbol{W = ULV^T}$, $\boldsymbol{U}$: d x q orthonormal vectors, $\boldsymbol{L}$: q x q matrix of singular values, $\boldsymbol{V}$: q x q orthogonal matrix,
  - $\boldsymbol{C^{-1}W = W(\sigma^2 I + W^T W)^{-1} = UL(\sigma^2 I + L^2)^{-1} V^T}$
- At the stationary points:
  - $\boldsymbol{SUL(\sigma^2 I + L^2)V^T = ULV^T}$
  - $\boldsymbol{SUL = U(\sigma^2 I + L^2)L}$
- Column vectors of $\boldsymbol{U}$, $\boldsymbol{u_j}$, are eigenvectors of $\boldsymbol{S}$, with eigenvalue λ_j, such that $\sigma^2 + l_j^2 = \lambda_j$
  - $l_j^2 = (\lambda_j - \sigma^2)^{1/2}$
- (substitute into SVD) $\boldsymbol{W = U_q\, (\Lambda_q - \sigma^2 I)\, R}$
  - $\boldsymbol{U_q}$: d x q with q column eigenvectors $\boldsymbol{u_j}$ of $\boldsymbol{S}$
  - $\boldsymbol{\Lambda}_j$: $\lambda_1 \ldots \lambda_q$, ($q$ eigenvalues of $\boldsymbol{u_j}$), or $\boldsymbol{\sigma^2}$ (corresponding d-q "discarded" rows of $\boldsymbol{W}$)
  - $\boldsymbol{R}$: arbitrary orthogonal matrix, equivalent to a rotation in principal subspace (or a re-parametrization)

Derivation of MLEs (cont)

- Substitute above results into original *LL* expression
- $LL = -N/2\{d \ln(2\pi) + \sum_{j=1}^{q} \ln(\lambda_j) + \sum_{j=q+1}^{d} \lambda_j + (d - q)\ln \sigma^2 + q\}$
  - $\lambda_1 \ldots \lambda_q$, are q non-zero eigenvalues of $\boldsymbol{u_j}$ and $\lambda_{q+1} \ldots \lambda_d$, are zero
- Taking derivative of above with respect to σ^2 and solving for zero gives:

$$\sigma^2_{\mathrm{ML}} = \frac{1}{d-q} \sum_{j=q+1}^{d} \lambda_j$$

Differences between factor analysis and probabilistic PCA (PPCA)

- Covariance
  - PPCA (and standard PCA) is covariant under rotation of the original data axes
  - Factor analysis is covariant under component-wise rescaling
- Principal components (or factors)
  - In PPCA: different principal components (axes) can be found incrementally
  - Factor analysis: factors from a two-factor model may not correspond to those from a one-factor model

Dimensionality reduction and optimal reconstruction

- Using Bayes rule, we can obtain a posterior estimate of the latent variables
  - $x|y \sim N(M^{-1}W^T (y - \mu), \sigma^2 M^{-1})$,
    - where $M = W^T W + \sigma^2 I$, M is a $q \times q$ matrix
  - ***Cond. latent mean:*** $E[x|y] = \langle x_n|y_n \rangle = M^{-1}W^T (y_n - \mu)$
- Reconstruction of the observed data with respect to the new subspace:
  - The latent projection of regular PCA is skewed towards the origin (due to marginal distribution for x)
    - $y_n = W_{ML}\langle x_n|y_n \rangle + \mu$ is not orthogonal and thus not optimal
  - Optimal reconstruction of the observed data may still be obtained from conditional latent mean:
    - $y_n = W_{ML}(W_{ML}^T W_{ML})^{-1} M \langle x_n|y_n \rangle + \mu$

Motivation behind using E-M for PCA

- Naive PCA and MLE PCA computation-heavy for high dimensional data or large data sets
- PCA does not deal properly with missing data
  - E-M algorithm estimates ML values of missing data at each iteration
- Naïve PCA uses simplistic way (distance^2 from observed data) to access covariance
  - Sensible PCA (SPCA) defines a proper covariance structure whose parameters can be estimated through the E-M algorithm

E-M algorithm (review)

- Iterative process to estimate parameters consisting of two steps for each iteration
  - Expectation (data step): complete all hidden and missing variables Θ (or latent variables) from current set of parameters
  - Maximization (likelihood step): Update set of parameters $\Theta`$, using MLE, from complete set of data from previous step
- Likelihood obtained from MLEs guaranteed to improve in successive iterations
- Continue iterations until negligible improvement is found in likelihood

E-M algorithm for normal PCA

- Amounts to an iterative procedure for finding subspace spanned by the q leading eigenvectors without computing covariance
- E-step: $\boldsymbol{X} = \boldsymbol{(W^T W)^{-1} W^T Y}$
  - Fix subspace and project data, $\boldsymbol{y}$, into it to give values of hidden states $\boldsymbol{x}$
  - Known: $\boldsymbol{Y}$: d-dimensional observed data
  - Unknown (latent): $\boldsymbol{X}$: q-dimensional unknown states
- M-step: $W_{new} = \boldsymbol{YX^T(XX^T)^{-1}}$
  - Fix values of hidden states and choose subspace orientation that minimizes squared reconstruction errors

E-M algorithm and missing data

- Data with missing obs filled out: ***x***, Complete data (with blanks not filled out): ***y***

E-step (fill in missing variables):

- If data point ***y*** is complete, then ***y*=y*** and ***x**** is found as usual
- If the data point ***y*** is not complete, ***x**** and ***y**** are the solution to the least squares problem. Compute ***x*** by projecting the observed data ***y*** into the current subspace.
  - For each (possibly incomplete) point ***y***, find the unique pair of points **(x*,y*)** that minimize the norm $\|Wx^*-y^*\|$.
  - Constrain ***x**** to be in the current principal subspace and ***y**** in the subspace defined by known info about ***y***
    - If ***y*** can be completely solved in system of equations, set corresponding column of ***X*** to ***x**** and the corresponding column of ***Y*** to ***y****
    - Otherwise, QR factorization can be used on a particular constraint matrix to find least squares solution

E-M algorithm and missing data (E-step)

$$W = \begin{pmatrix} 1 & 1 \\ 1 & 0.5 \\ 2 & 1 \end{pmatrix} \quad X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad Y = \begin{pmatrix} 3 \\ 1 \\ ? \end{pmatrix}$$

Wx = y

$$\begin{aligned} x_1 + x_2 &= 3 \\ x_1 + 0.5\,x_2 &= 1 \\ 2x_1 + x_2 &= y \end{aligned} \xrightarrow{\textit{solve}} X^* = \begin{pmatrix} -1 \\ 4 \end{pmatrix} \quad Y^* = \begin{pmatrix} 3 \\ 1 \\ 2 \end{pmatrix}$$

Set ***x*** **=** *(-1, 4)****,*** ***y*** **=** *(3, 1, 2)****,*** proceed to M-step

If two elements are missing in ***Y***, then we use QR factorization to find the pair (x*, y*) with the least squares of the norm **||*Wx*-y**||**, according to the constraints specified in the set of equations ***Wx = y***.

EM for probabilistic PCA (Sensible Principal Component Analysis)

- Probabilistic PCA model:
  - $\boldsymbol{Y} \sim N(\mu, \boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I})$
- Similar to normal PCA model, the differences are:
  - We do not take the limit as σ^2 approaches 0
  - During E-M iterations, data can be directly generated from the SPCA model, and the likelihood estimated from the test data set
  - Likelihood much lower for data far away from the training set, even if they are near the principal subspace
- EM algorithm steps implemented as follows:
  - E: $\boldsymbol{\beta} = \boldsymbol{W}^T(\boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I})^{-1}$, $\langle x_n|y_n\rangle = \boldsymbol{\beta}(\boldsymbol{Y} - \mu)$, $\boldsymbol{\Sigma}_x = n\boldsymbol{I} - n\boldsymbol{\beta}\boldsymbol{W} + \langle x_n|y_n\rangle\langle x_n|y_n\rangle^T$
    - Log-likelihood in terms of weight matrix $\boldsymbol{W}$, and a *centered* observed data matrix $\boldsymbol{Y}- \mu$, noise covariance $\sigma^2\boldsymbol{I}$, and conditional latent mean $\langle x_n|y_n\rangle$
  - M: $\boldsymbol{W}^{new} = (\boldsymbol{Y} - \mu)\langle x_n|y_n\rangle^T \boldsymbol{\Sigma}_x^{-1}$, $\sigma^{2\ new} = trace[\boldsymbol{X}\boldsymbol{X}^T - \boldsymbol{W}\langle x_n|y_n\rangle(\boldsymbol{Y}- \mu)^T]/n^2$
    - Differentiate LL in terms of $\boldsymbol{W}$ and σ^2 and set to zero.

Advantages of using E-M algorithm in probabilistic PCA models

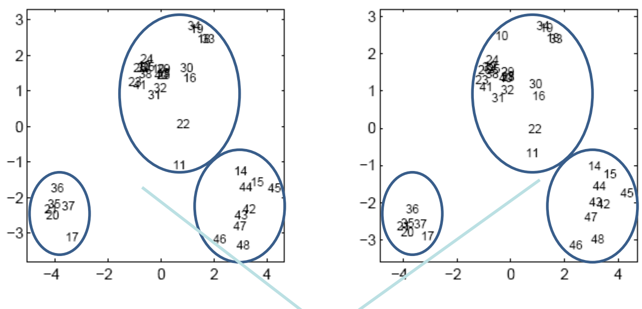
- Convergence:
  - Tipping and Bishop showed (1997) that the only stable local extremum is the *global maximum* at which the true principal subspace is found
- Complexity:
  - Methods that explicitly compute the sample covariance matrix have complexities $O(nd^2)$
  - E-M algorithm does not require computation of sample covariance matrix, $O(dnq)$
    - Huge advantage when $q \ll d$ (# of principal components is much smaller than original # of variabes)

E-M algorithm for PPCA (illustration)

Example: 38 observations (with 18 data points each) from *Tobamovirus* data set (Ripley, 1996)

Standard PCA (on complete data)

Probabilistic PCA (using EM algorithm) with 20% (136) missing values



3 clusters

Other methods for PCA

- Power iteration methods
  - Iteratively update eigenvector estimates through repeated multiplication by matrix to be diagonalized
  - Extremely inefficient to calculate explicitly ($O(nq^2)$)
  - E-M algorithm provides efficient way to obtain sample covariance matrix, without explicitly calculating it
  - Iterative methods to compute SVD are closely related to the E-M algorithm
- Learning methods for the principal subspace
  - Sanger's and Oja's rule
  - Typically require more iterations and the learning parameter to be set by hand

Mixtures of probabilistic PCAs

- A combination of local probabilistic PCA models
- Multiple plots may reveal more complex data structures than a PCA projection alone
- Applications:
  - Image compression (Dony and Haykin 1995)
  - Visualization (Bishop and Tipping, 1998)
- Clustering mechanisms of mixture PPCA:
  - Local linear dimensionality reduction
  - Semi-parametric density estimation

Mixtures of probabilistic PCAs

- $LL = \sum_{n=1}^{N} ln\{p(\boldsymbol{y_n})\} = \sum_{n=1}^{N} ln\ \{\sum_{i=1}^{M} \boldsymbol{\pi}_i\ p(\boldsymbol{y_n}|i)\}$
  - $p(\boldsymbol{y}|i)$ is a single PPCA model and $\boldsymbol{\pi}_i$ is the corresponding mixing proportion
  - Different mean vectors μ_i, weighting matrices $\boldsymbol{W}_i$, and noise error parameters σ_i^2 for each of M probabilistic PCA models
- An iterative E-M algorithm can be used to solve for parameters
- Guaranteed to find a *local* maximum of the log-likelihood