Latent Variable Models

CS3750
Xiaoting Li

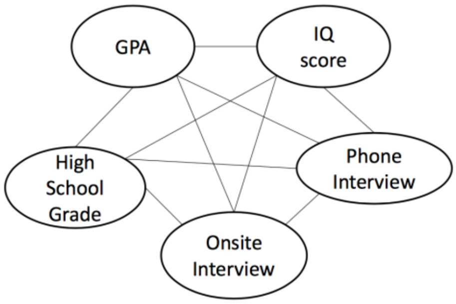
1

Outline

- **Latent Variable Models**
  - Expectation Maximization Algorithm (EM)
- **Factor Analysis**
- **Probabilistic Principal Component Analysis**
  - Model Formulation
  - Maximum Likelihood for PPCA
  - EM for PPCA
  - Examples
- **Sensible Principal Component Analysis**
  - Model Formulation
  - EM for SPCA
- **References**

2

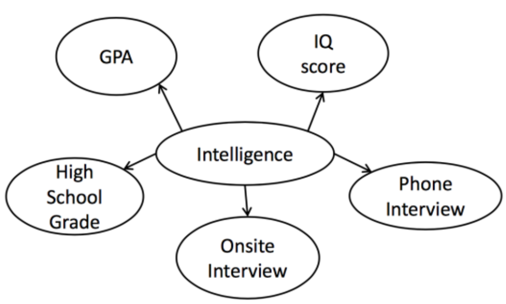
Latent Variable Models: Motivation



Name	High School Grade	University Grade	IQ score	Phone Interview	Onsite Interview
John	4.0	4.0	120	3/4	?
Helen	3.2	N/A	112	2/4	?
Sophia	3.5	3.6	N/A	4/4	85/100
Jack	3.6	N/A	N/A	3/4	?

3

Latent Variable Models: Motivation

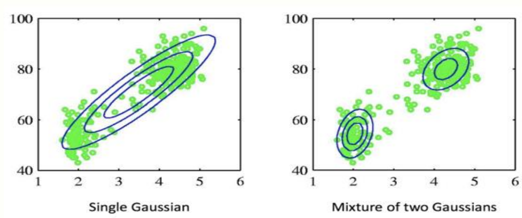


Name	High School Grade	University Grade	IQ score	Phone Interview	Onsite Interview
John	4.0	4.0	120	3/4	?
Helen	3.2	N/A	112	2/4	?
Sophia	3.5	3.6	N/A	4/4	85/100
Jack	3.6	N/A	N/A	3/4	?

4

Latent Variable Models: Motivation

- Gaussian mixture models
  - A single Gaussian is not a good fit to data
  - But two different Gaussians may do
  - True class of each point is unobservable



Example of a dataset that is best fit with a mixture of two Gaussians. Mixture models allow us to model clusters in the dataset.

5

Latent Variable Models

A latent variable model *p* is a probability distribution over two sets of variables *s,x*:

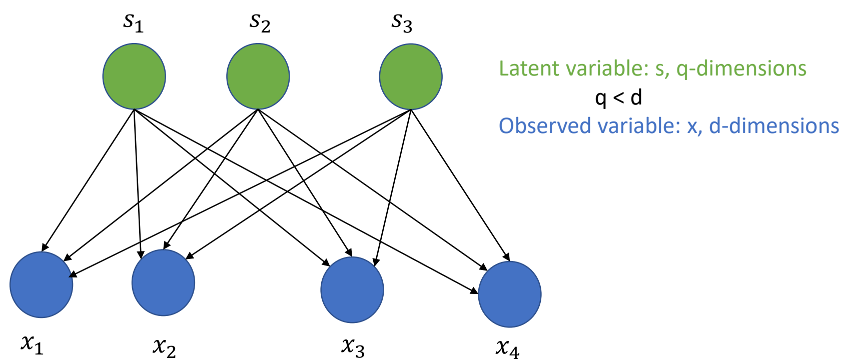
$$p(s, x; \theta)$$

where the *x* variables are observed at learning time in a dataset D and the *s* are never observed

6

Latent Variable Models

- The goal of a latent variable model is to express the distribution *p(x)* of the variables $x_1, \dots, x_d$ in terms of a smaller number of latent variables *s* = (s_1,…, s_q) where q < d



7

Expectation-Maximization (EM) algorithm

- EM algorithm is a hugely important and widely used algorithm for learning directed latent-variable graphical
- The key idea of the method:
  - Compute the parameter estimates iteratively by performing the following two steps:
    - **1. Expectation step**. For all hidden and missing variables (and their possible value assignments) calculate their expectations for the current set of parameters **Θ**'
    - **2. Maximization step**. Compute the new estimates of **Θ** by considering the expectations of the different value completions
    - **Stop when no improvement possible**

8

Factor Analysis

- Assumptions:
  - Underlying latent variable has a Gaussian distribution
    - s $\sim N(0, \mathrm{I})$, independent, Gaussian with unit variance
  - Linear relationship between latent and observed variables
  - *Diagonal* Gaussian noise in data dimensions
    - $\epsilon \sim N(0, \Psi)$, Gaussian noise

9

Factor Analysis

- A common latent variable where the relationship is linear:

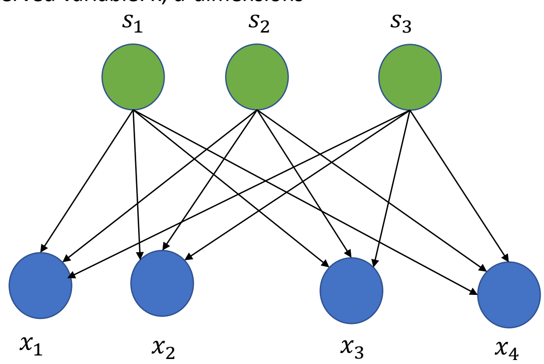
$$\mathrm{x} = Ws + \mu + \epsilon$$

  - $\mathrm{d-dimensional\ observation\ vector}\ x$
  - q-dimensional vector of latent variable s
  - $d \times q$ matrix W relates the two sets of variables, $q < d$
  - μ permits the model to have non-zero mean
  - s $\sim N(0, \mathrm{I})$, independent, Gaussian with unit variance
  - $\epsilon \sim N(0, \Psi)$, Gaussian noise
    - Then x $\sim N(\mu, WW^T + \Psi)$

10

Factor Analysis

Latent variable: s, q-dimensions
Observed variable: x, d-dimensions



$s \sim N(0, I)$

$Remapping$: Ws (weight matrix: w)

+

μ (location parameter)

+

$\epsilon \sim N(0, \Psi)$, Gaussian noise

$x = Ws + \mu + \epsilon$
$x \sim N(\mu, WW^T + \Psi)$

Parameters of interest: W (weight matrix),Ψ (variance of noise), $\boldsymbol{\mu}$

11

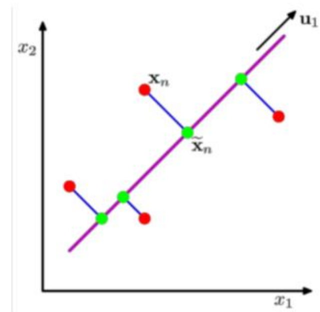
Factor Analysis: Optimization

- Use EM to solve parameters
- E-step:
  - compute posterior p(s|x)
- M-step:
  - Take derivatives of the expected complete log likelihood with respect to parameters

12

Principal Component Analysis

- General motivation is to transform the data into some reduced dimensionality representation
- Linear transformation of a *d* dimensional input **x** to q dimensional vector **s** such that $q < d$ under which the retained variance is maximal
- Limitation:
  - Absence of an associated probabilistic model for the observed data
  - Computational intensive for covariance matrix
  - Does not deal properly with missing data



13

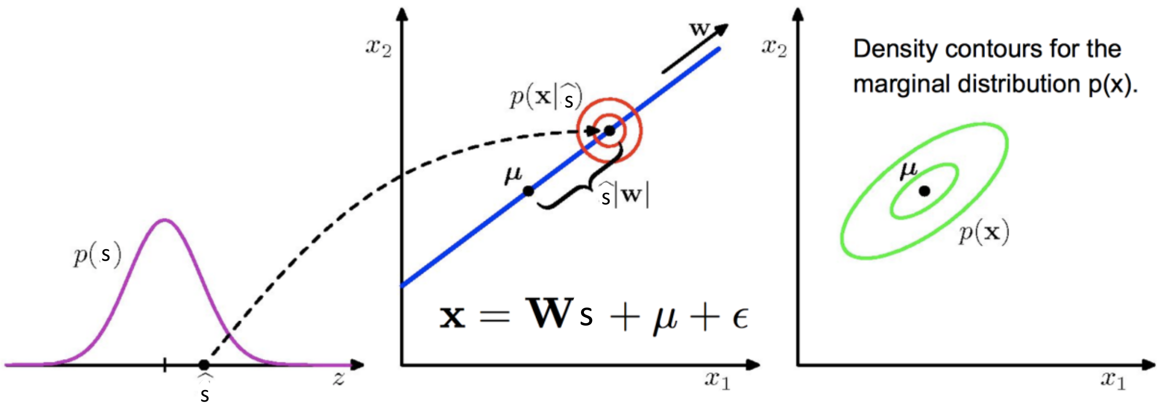
Probabilistic PCA

- Motivations:
  - The corresponding likelihood measure would permit comparison with other density–estimation techniques and would facilitate statistical testing.
    - Provides a natural framework for thinking about hypothesis testing
  - Offers the potential to extend the scope of conventional PCA.
  - Can be utilized as a constrained Gaussian density model.
    - Constrained covariance
  - Allows us to deal with missing values in the data set.
  - Can be used to model class conditional densities and hence it can be applied to classification problems.

14

Generative View of PPCA

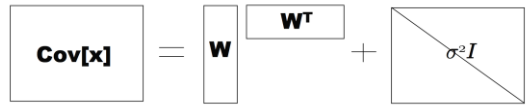
- Generative view of the PPCA for a 2-d data space and 1-d latent space



PPCA

- Assumptions:
  - Underlying latent variable $q-\dim s$ has a Gaussian distribution
  - Linear relationship between $q-\dim$ latent s and $d-\dim$ observed x variables
  - Isotropic Gaussian noise in observed dimensions
    - Noise variances constrained to be equal

PPCA

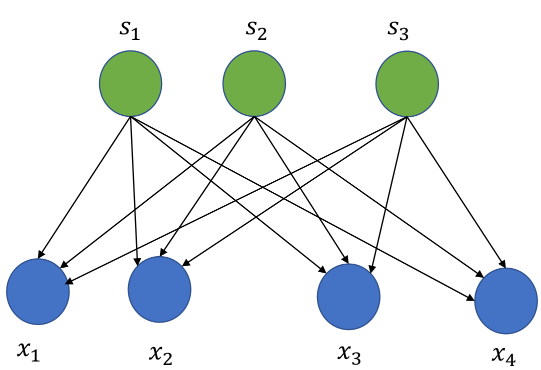


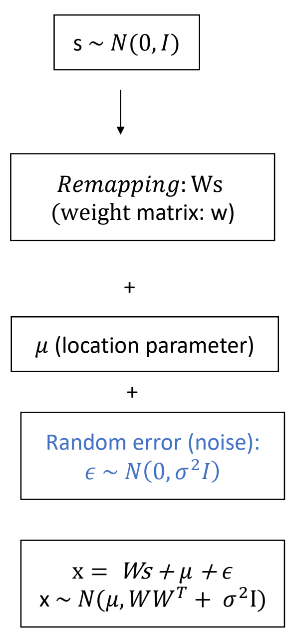
- A special case of factor analysis
  - noise variances constrained to be equal:
    - $\epsilon \sim N(0, \sigma^2 \mathrm{I})$
  - the s conditional probability distribution over x-space:
    - $\mathrm{x}|s \sim N(Ws + \mu, \sigma^2 \mathrm{I})$
  - latent variables:
    - $\mathrm{s} \sim N(0, I)$
  - observed data x be obtained by integrating out the latent variables:
    - $\mathrm{x} \sim N(\mu, C)$
      - $E[x] = E[\mu + Ws + \epsilon] = \mu + WE[s] + E[\epsilon] = \mu + W0 + 0 = \mu$
    - $C = WW^T + \sigma^2 \mathrm{I}$ (the observation covariance model)
      - $C = Cov[x] = E[(\mu + Ws + \epsilon - \mu)(\mu + Ws + \epsilon - \mu)^T] = E[(Ws + \epsilon)(Ws + \epsilon)^T] = WW^T + \sigma^2 \mathrm{I}$
  - The maximum likelihood estimator for μ is given by the mean of data, ***S*** is sample covariance matrix of the observations $\{x_n\}$
  - Estimates for W and σ^2 can be solved in two ways
    - Closed form
    - EM Algorithms

17

PPCA

Latent variable: s, q-dimensions
Observed variable: x, d-dimensions





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

18

Factor Analysis vs. PPCA

- PPCA
  - x $\sim N(\mu, WW^T + \sigma^2 \mathrm{I})$
  - Isotropic error
- Factor Analysis
  - x $\sim N(\mu, WW^T + \Psi)$
  - The error covariance is a diagonal matrix
  - FA doesn't change if you scale variables
    - FA looks for directions of large correlation in the data
    - FA doesn't chase large-noise features that are uncorrelated with other features
  - FA changes if you rotate data
    - can't interpret multiple factors as being unique

Maximum Likelihood for PPCA

- The log-likelihood for the observed data under this model is given by

$$\mathcal{L} = \sum_{n=1}^{N} \ln\{p(x_n)\} = -\frac{Nd}{2}\ln(2\pi) - \frac{N}{2}\ln|\mathrm{C}| - \frac{N}{2}Tr\{C^{-1}S\}$$

  - where S is the sample covariance matrix of the observations $\{(x_n)\}$

$$S = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)(x_n - \mu)^T$$

  - $C = WW^T + \sigma^2 \mathrm{I}$
- The log-likelihood is maximized when the columns of W span the principal subspace of the data.
  - Fit parameters $(\boldsymbol{W}, \mu, \sigma)$ to maximum likelihood: make the constrained model covariance as close as possible to the observed covariance

Maximum Likelihood for PPCA

- Consider the derivatives with respect to W
  - $\frac{\partial \mathcal{L}}{\partial W} = N(C^{-1}SC^{-1}W \; - C^{-1}\mathrm{W})$
- Maximizing with respect to W
  - $W_{ML} = \; U_q(\wedge_q - \sigma^2 I)^{1/2} R$
- Where
  - the q column vectors in Uq are eigenvectors of S, with corresponding eigenvalues in the diagonal matrix Λq
  - R is an arbitrary $q \; \times \; q$ orthogonal rotation matrix.
- For W = W_{ML}, the maximum-likelihood estimator for σ^2 is given by
  - $\sigma_{ML}^2 = \frac{1}{d-q} \sum_{j=q+1}^{d} \lambda_j$
  - the average variance associated with the discarded dimensions

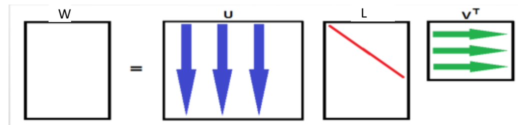
21

Maximum Likelihood for PPCA

- Consider the derivatives with respect to W
  - $\frac{\partial \mathcal{L}}{\partial W} = N(C^{-1}SC^{-1}W \; - C^{-1}\mathrm{W})$
  - At the stationary points $SC^{-1}W = W$, assuming that C^{-1} exists
- Three possible classes of solutions
  - W = 0, minimum of the log-likelihood
  - $C = S$
    - Covariance model is exact
    - $WW^T \; = S \; - \; \sigma^2 I$ has a known solution at W = $U(\wedge - \sigma^2 I)^{1/2} R$, where U is a square matrix whose columns are the eigenvectors of S, with ∧ is the corresponding diagonal matrix of eigenvalues, R is an arbitrary orthogonal matrix
  - $SC^{-1}W = W, but\ W \neq 0\ and\ C \neq S$

22

Maximum Likelihood for PPCA



- Consider the derivatives with respect to W
  - $\frac{\partial \mathcal{L}}{\partial W} = N(C^{-1}SC^{-1}W \; - C^{-1}\mathrm{W})$
  - At the stationary points $SC^{-1}W = W$, assuming that C^{-1} exists
- Case: $SC^{-1}W = W, but\ W \neq 0\ and\ C \neq S$
  - Express the parameter matrix W in terms of singular value decomposition (SVD):
    - $W = ULV^T$, $\boldsymbol{U}$: $d\ x\ q$ orthonormal vectors, $\boldsymbol{L}$: $q\ x\ q$ matrix of singular values, $\boldsymbol{V}$: $q\ x\ q$ orthogonal matrix
    - $C^{-1}W = \mathrm{W}\,(\sigma^2 I + W^T W)^{-1} = UL(\sigma^2 I + L^2)^{-1} V^T$
  - At the stationary points
    - $SUL(\sigma^2 I + L^2)^{-1} V^T = ULV^T$
    - $SUL = U(\sigma^2 I + L^2)L$

Maximum Likelihood for PPCA

- Column vectors of $\boldsymbol{U}, u_j$, are eigenvectors of $\boldsymbol{S}$, with eigenvalue λ_j, such that $\sigma^2 + l_j^2 = \lambda_j$
  - $Su_j = (\sigma^2 + l_j^2)u_j$
  - $l_j^2 = (\lambda_j - \sigma^2)^{1/2}$
- (substitute into SVD) , $W = U_q(\Lambda_q - \sigma^2 I)\,R$
  - U_q : *d x q* with *q* column eigenvectors $u_j$of $\boldsymbol{S}$
  - Λ_q : *q x q* diagonal matrix with elements: $\lambda_1 \ldots \lambda_q$, (eigenvalues to u_j), or σ^2 (equivalent to $l_j = 0$)
  - ***R***: arbitrary orthogonal matrix, equivalent to a rotation in principal subspace (or a re-parametrization)

EM for PPCA

- **Goal:** to estimate the model parameters W and σ^2, based on the observed dataset
- Rather than solve directly, can apply EM
- EM can be scaled to very large high-dimensional datasets.
- Consider the latent variables $\{s_n\}$ to be 'missing' data
- Need Complete-data log-likelihood:
  - $\mathcal{L}_C = \sum_{n=1}^{N} \ln\{p(x_n, s_n\}$
  - since
    - x$|s \sim N(Ws + \mu, \sigma^2 \mathrm{I})$ and s $\sim N(0, I)$
  - we have
    - $p(x_n, s_n)$ = $(2\pi\sigma^2)^{-d/2}\exp(-\frac{||x_n - Ws_n - \mu||^2}{2\sigma^2})(2\pi)^{-\frac{q}{2}}\exp(-\frac{||s_n||^2}{2})$

25

EM for PPCA

- **E-step**
  - Compute expectation of complete log likelihood with respect to posterior of latent variables
  - Take the expectation of $\mathcal{L}_C$ with respect to the distributions $p(s_n|x_n$, W, $\sigma^2)$
  - $\langle\mathcal{L}_C\rangle = -\sum_{n=1}^{N} \frac{d}{2}\ln(\sigma^2) + \frac{1}{2}tr(\langle s_n s_n^T\rangle) + \frac{1}{2\sigma^2}(x_n - \mu)^T(x_n -$

26

PPCA Examples

- Missing data
  - A natural approach to the estimation of the principal axes in cases where some or indeed all, of the data vectors exhibit one or more missing (at random) values
  - Fig. 1 (a): projection of 38 examples from the 18-dimensional Tobamovirus data (Ripley 1996) using standard PCA
  - Fig.1 (b): an equivalent PPCA projection obtained by using an EM algorithm
    - Simulated missing data by randomly removing each value in the data set with probability 20%

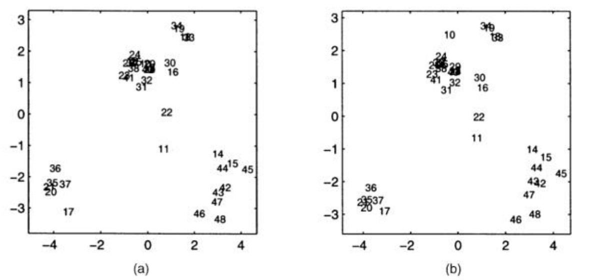


Fig. 1. Projections of the *Tobamovirus* data by using (a) PCA on the full data set and (b) PPCA with 136 missing values

PPCA Examples

- Mixtures of probabilistic principal component analysis models
  - Combining multiple PCA models, notably for image compression
  - Fig.2: three PCA projections of the virus data obtained from a three-component mixture model, optimized by using an EM algorithm
  - Effectively implements a simultaneous automated clustering and visualizing data

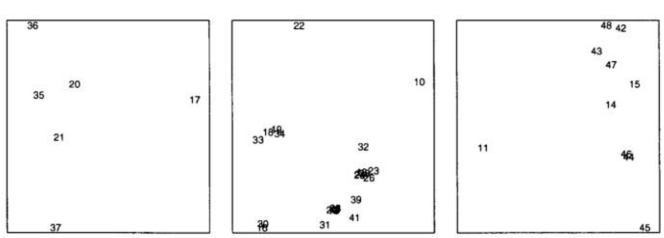


Fig. 2. Projections of the *Tobamovirus* data obtained from a three-component PPCA mixture model: the locations of these three projection planes can be superimposed on the single principal component projection plot (Fig. 1(a)) to aid the interpretation of the data structure further

PPCA Examples

- Controlling the degrees of freedom
  - Applied as a covariance model of data
  - Permits control of the model complexity through the choice of q
    - The covariance model in PPCA comprises $dq + 1 - q(q-1)/2$ free parameters
  - Table 1: estimated prediction error for various Gaussian models fitted to the Tobamovirus data
    - PPCA with q = 2 gives the lowest error

Table 1. Complexity and bootstrap estimate of the prediction error for various Gaussian models of the *Tobamovirus* data†

Covariance model	*q (equivalent)*	*Number of parameters*	*Prediction error*
Isotropic	(0)	1	18.6
Diagonal	(—)	18	19.6
PPCA	1	19	16.8
	2	36	14.8
	3	52	15.6
Full	(17)	171	3193.5

†The isotropic and full covariance models are equivalent to special cases of PPCA, with $q = 0$ and $q = d - 1$ respectively.

Sensible Principal Component Analysis (SPCA)

- SPCA
  - x = Ws + v
  - x ~ $N(0, WW^2 + \sigma^2 I)$
- Similar to PCA, the differences are:
  - Require noise covariance matrix to be a multiple $\sigma^2 I$ of the identity matrix, but do not take the limit as $\sigma^2 I \rightarrow 0$
  - During EM 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 for SPCA

- SPCA
  - x $\sim N(0, WW^T + \sigma^2 I)$
- E-step:
  - $\beta = W^T(WW^T + \sigma^2 I)^{-1}$
  - $\langle s_n | x_n \rangle = \beta(X - \mu)$
  - $\Sigma_s = nI - n\beta W + \langle s_n | x_n \rangle \langle s_n | x_n \rangle^T$
  - Log-likelihood in terms of weight matrix W, and a *centered* observed data matrix X $- \mu$, noise covariance $\sigma^2 I$, and conditional latent mean $\langle s_n | x_n \rangle$

31

EM for SPCA

- SPCA
  - x $\sim N(0, WW^T + \sigma^2 I)$
- M-step:
  - $W^{new} = (\mathrm{X} - \mu) \langle s_n | x_n \rangle^T \Sigma_s^{-1}$
  - $\sigma^{2\,new} = \mathrm{trace}[SS^T - W\langle s_n | x_n \rangle (\mathrm{X} - \mu)^T]/n^2$
  - Differentiate LL in terms of W and σ^2 and set to zero

32

EM for SPCA

- Since $\sigma^2 I$ is diagonal, the inversion in the e-step can be performed efficiently using the matrix inversion lemma:
  - $(WW^T + \sigma^2 I)^{-1} = (\frac{I}{\sigma^2} - W(I + \frac{W^T W}{\sigma^2})^{-1} W^T / (\sigma^2)^2)$
- Since we are only taking the trace of the matrix in the m–step, we do not need to compute the full sample covariance SS^T, but instead can compute only the variance along each coordinate
  - $\sigma^{2\,new} = \mathrm{trace}[SS^T - W\langle s_n | x_n \rangle (\mathrm{X} - \mu)^T]/n^2$
  - Shows that learning for SPCA enjoys a complexity limited by $O(dnq)$ and not worse
- Methods that explicitly compute the sample covariance matrix have complexities $O(nd^2)$
  - EM algorithm does not require computation of sample covariance matrix, $O(dnq)$
    - Huge advantage when *q* << *d* (# of principal components is much smaller than original # of variables)

33

Software

- Matlab
  - https://www.mathworks.com/help/stats/ppca.html
- R Programming
  - https://www.rdocumentation.org/packages/pcaMethods/versions/1.64.0/topics/ppca

34

Reference:

- Roweis, Sam T. "EM algorithms for PCA and SPCA." *Advances in neural information processing systems*. 1998.
- Tipping, Michael E., and Christopher M. Bishop. "Probabilistic principal component analysis." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.3 (1999): 611-622.
- Bishop, Christopher M. "Latent variable models." *Learning in graphical models*. Springer, Dordrecht, 1998. 371-403.
- https://www.cs.toronto.edu/~hinton/csc2515/notes/lec7middle.pdf
- https://www.cs.toronto.edu/~rsalakhu/STA4273_2015/notes/Lecture8_2015.pdf
- https://www.cs.ubc.ca/~schmidtm/Courses/540-W16/L12.pdf
- https://www.seas.upenn.edu/~cis520/lectures/PCA_PLS_CCA.pdf
- http://people.cs.pitt.edu/~milos/courses/cs3750/lectures/class8.pdf
- http://people.cs.pitt.edu/~milos/courses/cs2750-Spring2019/Lectures/Class19.pdf
- https://ermongroup.github.io/cs228-notes/learning/latent/
- https://people.cs.pitt.edu/~milos/courses/cs3750-Fall2007/lectures/class17.pdf
- https://people.cs.pitt.edu/~milos/courses/cs3750-Fall2014/lectures/class13.pdf
- https://liorpachter.wordpress.com/tag/ppca/