CSC2515    Fall 2015
Introduction to Machine Learning

Lecture 7B: Probabilistic PCA &
Factor Analysis

Some of the figures are provided by Chris Bishop
from his textbook: ”Pattern Recognition and Machine Learning”
Return to Graphical Model View

• Earlier we discussed PCA – considered as projection from inputs to reduced dimensional representation

• Now: as latent variable model

• Latent variables in mixture models are multinomials (referring to cluster identity).

• Today we’ll consider continuous latent variables
Generative View

- Each data example generated by first selecting a point from a distribution in the latent space, then generating a point from the conditional distribution in the input space.

- Mixture models have multinomial latents.

- For continuous latents, and inputs, now looking at simple models: Gaussian distributions in both latent and data space, linear relationship betwixt.

- This view underlies Probabilistic PCA, Factor Analysis.
Probabilistic PCA

- Probabilistic, generative view of data
- Assumptions:
  - underlying latent variable has a Gaussian distribution
  - linear relationship between latent and observed variables
  - isotropic Gaussian noise in observed dimensions

\[
p(z) = \mathcal{N}(z|0, I)
\]
\[
p(x|z) = \mathcal{N}(x|Wz + \mu, \sigma^2 I)
\]
\[
x = Wz + \mu + \epsilon
\]
Probabilistic PCA: Constrained covariance

- Marginal density for PPCA ($x$ is D-dim., $z$ is M-dim):
  \[ p(x|\theta) = \mathcal{N}(x|\mu, WW^T + \sigma^2 I) \]
  - where $\theta = W, \mu, \sigma$

- Effective covariance is low-rank outer product of two long skinny matrices plus a constant diagonal matrix

\[
\text{Cov}[x] = W \quad W^T + \sigma^2 I
\]

- So PPCA is just a constrained Gaussian model:
  - Standard Gaussian has $D + D(D+1)/2$ effective parameters
  - Diagonal-covariance Gaussian has $D + D$, but cannot capture correlations
  - PPCA: $DM + 1 - M(M-1)/2$, can represent $M$ most significant correlations
Factor Analysis

Can be viewed as generalization of PPCA

Historical aside – controversial method, based on attempts to interpret factors: e.g., analysis of IQ data identified factors related to race

Assumptions:
- underlying latent variable has a Gaussian distribution
- linear relationship between latent and observed variables
- diagonal Gaussian noise in data dimensions

\[
p(z) = \mathcal{N}(z|0, I)
\]
\[
p(x|z) = \mathcal{N}(x|Wz + \mu, \Psi)
\]

W: factor loading matrix (\(D \times M\))

\(\Psi\): data covariance (diagonal, or axis-aligned; vs. PCA’s spherical)
Summary of Latent Factor Methods

- Aim to find low-dimensional subspace that captures essential properties of data

- Assumes that even though data is high-dimensional, there are some small number of continuous underlying (latent) factors, whose variability accounts for variations in observations

- Example: latent factors underlying images are lighting, object identities, pose, etc.

- Different methods vary in terms of their assumptions about these factors, and how the observations relate to the factors