SNS COLLEGE OF TECHNOLOGY

Coimbatore – 35

An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A++' Grade

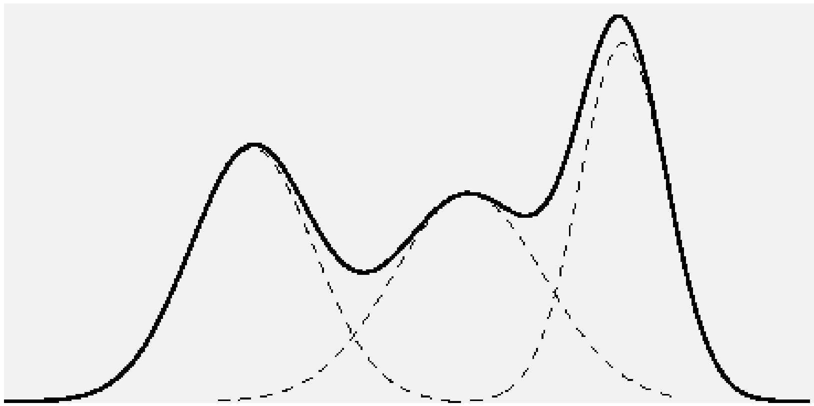
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

Probabilistic Models with Latent Variables

Density Estimation Problem

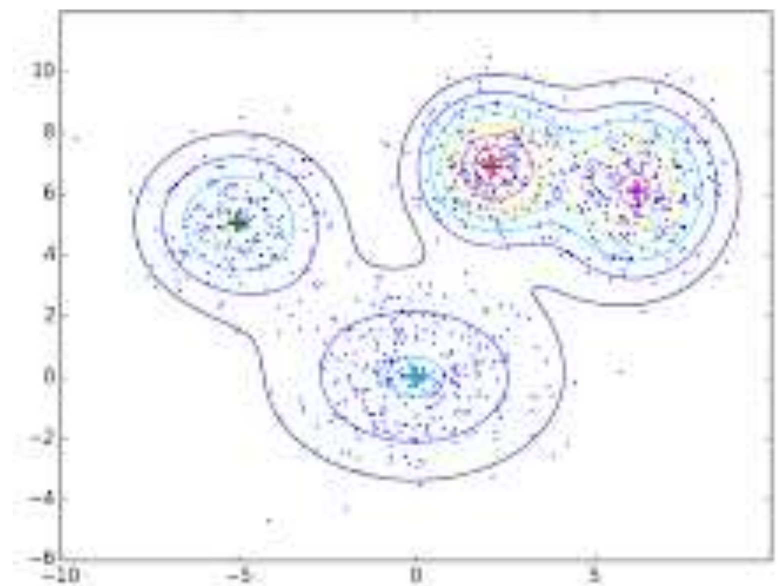
- Learning from unlabeled data $\{x_1, x_2, \ldots, x_N\}$
  - Unsupervised learning, density estimation
- Empirical distribution typically has multiple modes

Density Estimation Problem



From http://courses.ee.sun.ac.za/Pattern_Recognition_813

From http://yulearning.blogspot.co.uk



Density Estimation Problem

- Conv. composition of unimodal pdf's: multimodal pdf

$$f(x) = \sum_k w_k f_k(x) \text{ where } \sum_k w_k = 1$$

- Physical interpretation
  - Sub populations

Latent Variables

- Introduce new variable Z_i for each X_i
- Latent / hidden: not observed in the data

- Probabilistic interpretation
  - Mixing weights: $w_k \equiv p(z_i = k)$
  - Mixture densities: $f_k(x) \equiv p(x|z_i = k)$

Generative Mixture Model

For $i = 1, \ldots, N$

$Z_i \sim iid\ Mult$

$X_i \sim iid\ p(x|z_i)$

- $P(x_i, z_i) = p(z_i)p(x_i|z_i)$
- $P(x_i) = \sum_k p(x_i, z_i)$ recovers mixture distribution

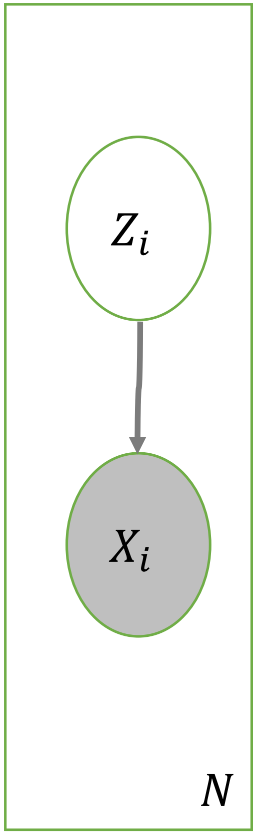


Plate Notation

Tasks in a Mixture Model

- Inference

$$P(z|x) = \prod_i P(z_i|x_i)$$

- Parameter Estimation
  - Find parameters that e.g. maximize likelihood
  - Does not decouple according to classes
  - Non convex, many local minima

Example: Gaussian Mixture Model

- Model

$$\text{For } i = 1, \dots, N$$
$$Z_i \sim iid\ Mult(\pi)$$
$$X_i \mid Z_i = k \sim iid\ N(x|\mu_k, \Sigma)$$

- Inference

$$P(z_i = k|x_i; \mu, \Sigma)$$

  - Soft-max function

Example: Gaussian Mixture Model

- Loglikelihood
  - Which training instance comes from which component?

$$l(\theta) = \sum_i \log p(x_i) = \sum_i \log \sum_k p(z_i = k) p(x_i | z_i = k)$$

- No closed form solution for maximizing $l(\theta)$

- Possibility 1: Gradient descent etc
- Possibility 2: Expectation Maximization

Expectation Maximization Algorithm

- Observation: Know values of $Z_i \Rightarrow$ easy to maximize

- Key idea: iterative updates
  - Given parameter estimates, “infer” all Z_i variables
  - Given inferred Z_i variables, maximize wrt parameters

- Questions
  - Does this converge?
  - What does this maximize?

Expectation Maximization Algorithm

- Complete loglikelihood

$$l_c(\theta) = \sum_i \log p(x_i, z_i) = \sum_i \log p(z_i) p(x_i|z_i)$$

- Problem: z_i not known
- Possible solution: Replace w/ conditional expectation

- Expected complete loglikelihood

$$Q(\theta, \theta_{old}) = E[\sum_i \log p(x_i, z_i)\,]$$

Wrt $p(z|x, \theta_{old})$ where θ_{old} are the current parameters

Expectation Maximization Algorithm

$$Q(\theta, \theta_{old}) = E[\sum_i \log p(x_i, z_i)]$$

$$= \sum_i \sum_k E[I(z_i = k)] \log[\pi_k p(x_i|\theta_k)]$$

$$= \sum_i \sum_k p(z_i = k|x_i, \theta_{old}) \log[\pi_k p(x_i|\theta_k)]$$

$$= \sum_i \sum_k \gamma_{ik} \log \pi_k + \sum_i \sum_k \gamma_{ik} \log p(x_i|\theta_k)$$

Where $\gamma_{ik} = p(z_i = k|x_i, \theta_{old})$

- Compare with likelihood for generative classifier

Expectation Maximization Algorithm

- Expectation Step
  - Update γ_{ik} based on current parameters

$$\gamma_{ik} = \frac{\pi_k p(x_i | \theta_{old,k})}{\sum_k \pi_k p(x_i | \theta_{old,k})}$$

- Maximization Step
  - Maximize $Q(\theta, \theta_{old})$ wrt parameters

- Overall algorithm
  - Initialize all latent variables
  - Iterate until convergence
    - M Step
    - E Step

Example: EM for GMM

- E Step remains the step for all mixture models

- M Step
  - $\pi_k = \frac{\sum_i \gamma_{ik}}{N} = \frac{\gamma_k}{N}$
  - $\mu_k = \frac{\sum_i \gamma_{ik} x_i}{\gamma_k}$
  - $\Sigma = ?$

- Compare with generative classifier

Analysis of EM Algorithm

- Expected complete LL is a lower bound on LL
- EM iteratively maximizes this lower bound

- Converges to a local maximum of the loglikelihood

Bayesian / MAP Estimation

- EM overfits
- Possible to perform MAP instead of MLE in M-step

- EM is partially Bayesian
  - Posterior distribution over latent variables
  - Point estimate over parameters

- Fully Bayesian approach is called Variational Bayes

(Lloyd's) K Means Algorithm

- Hard EM for Gaussian Mixture Model
  - Point estimate of parameters (as usual)
  - Point estimate of latent variables
  - Spherical Gaussian mixture components

$$z_i^* = \arg\max_{\mathrm{k}} p(z_i = k|x_i, \theta) = \arg\min_k ||x_i - \mu_k||_2^2$$

$$\text{Where } \mu_k = \frac{\sum_{i:z_i=k} x_i}{N}$$

- Most popular "hard" clustering algorithm

K Means Problem

- Given $\{x_i\}$, find k "means" $(\mu_1^*, \dots, \mu_k^*)$ and data assignments $(z_1^*, \dots, z_N^*)$ such that

$$(\mu^*, z^*) = \arg\min_{\mu, z} \sum_i \left|\left|x_i - \mu z_i\right|\right|_2^2$$

- Note: z_i is k-dimensional binary vector

Model selection: Choosing K for GMM

- Cross validation
  - Plot likelihood on training set and validation set for increasing values of k
  - Likelihood on training set keeps improving
  - Likelihood on validation set drops after "optimal" k

- Does not work for k-means! Why?

Principal Component Analysis: Motivation

- Dimensionality reduction
  - Reduces #parameters to estimate
  - Data often resides in much lower dimension, e.g., on a line in a 3D space
  - Provides "understanding"

- Mixture models very restricted
  - Latent variables restricted to small discrete set
  - Can we "relax" the latent variable?

Classical PCA: Motivation

- Revisit K-means

$$\min_{W,Z} J(W, Z) = |X - WZ^T|^2{}_F$$

  - W: matrix containing means
  - Z: matrix containing cluster membership vectors

- How can we relax Z and W?

Classical PCA: Problem

$$\min_{W,Z} J(W,Z) = ||X - WZ^T||^2{}_F$$

- X : $D \times N$
- Arbitrary Z of size $N \times L$,
- Orthonormal W of size $D \times L$

Classical PCA: Optimal Solution

- Empirical covariance matrix $\hat{\Sigma} = \frac{1}{N}\sum_i x_i x_i^T$
  - Scaled and centered data
- $\widehat{W} = V_L$ where V_L contains L Eigen vectors for the L largest Eigen values of $\hat{\Sigma}$
- $\widehat{z_i} = \widehat{W}^T x_i$
- Alternative solution via Singular Value Decomposition (SVD)
- W contains the "principal components" that capture the largest variance in the data

Probabilistic PCA

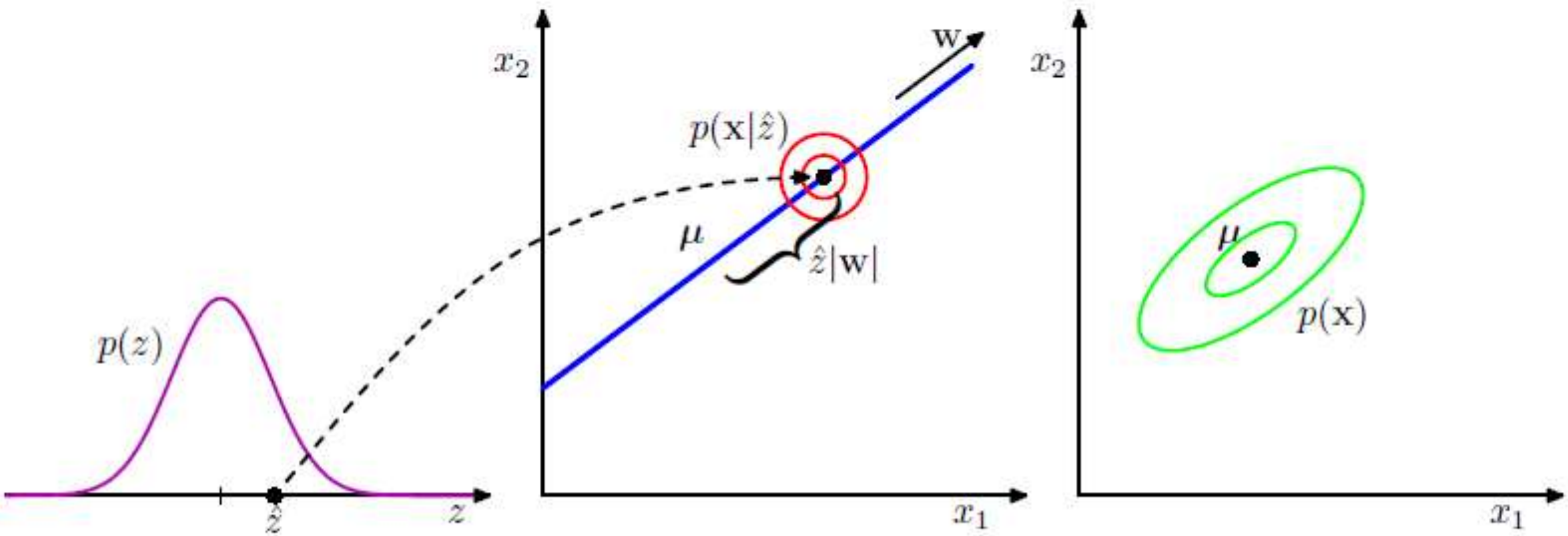
- Generative model

$$P(z_i) = N(z_i|\mu_0, \Sigma_0)$$
$$P(x_i|z_i) = N(x_i|Wz_i + \mu, \Psi)$$

Ψ forced to be diagonal

- Latent linear models
  - Factor Analysis
  - Special Case: PCA with $\Psi = \sigma^2 I$

Visualization of Generative Process

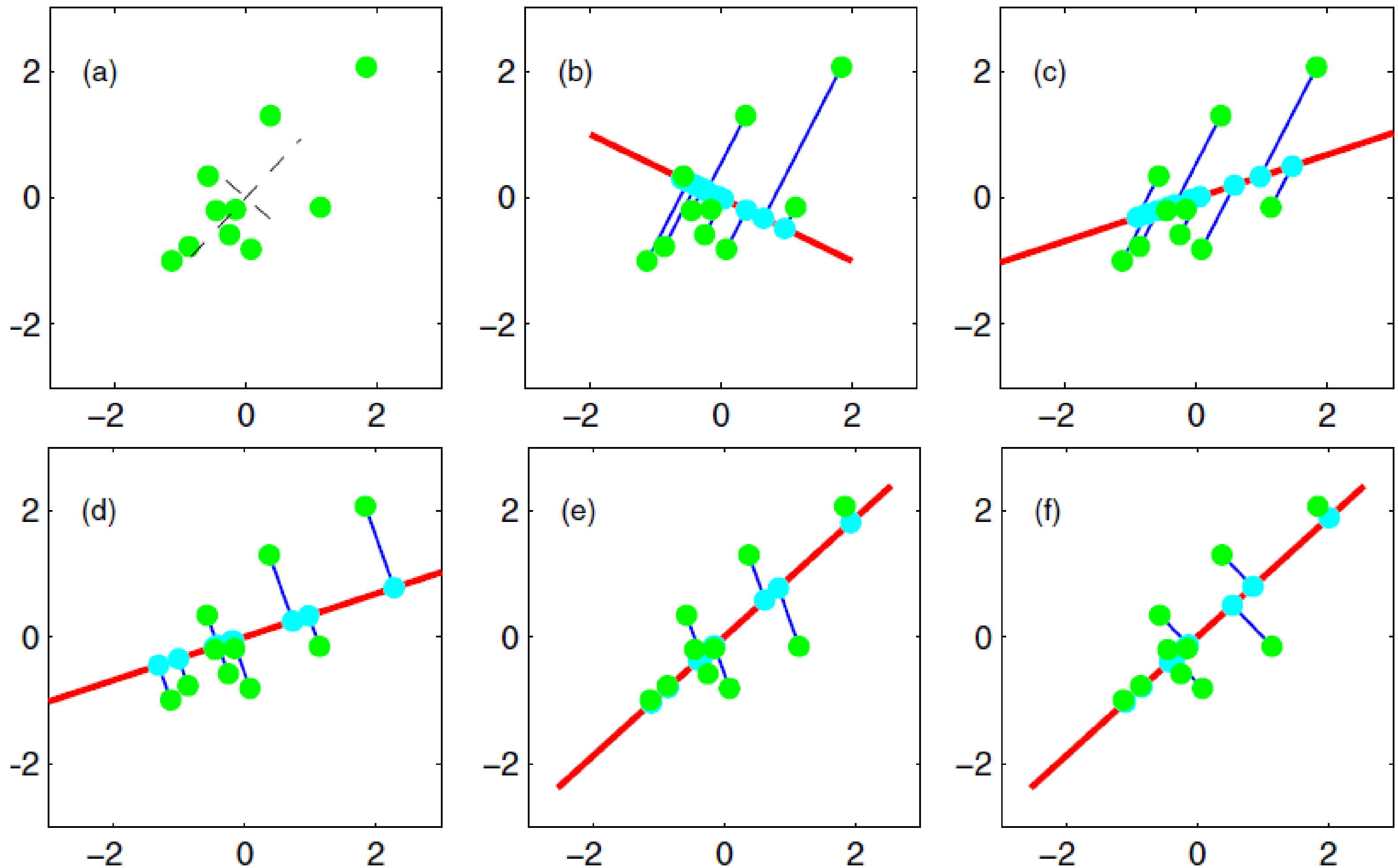


From Bishop, PRML

Relationship with Gaussian Density

- $Cov[x] = WW^T + \Psi$
- Why does Ψ need to be restricted?
- Intermediate low rank parameterization of Gaussian covariance matrix between full rank and diagonal
  - Compare #parameters

EM for PCA: Rod and Springs



From Bishop, PRML

Advantages of EM

- Simpler than gradient methods w/ constraints
- Handles missing data
- Easy path for handling more complex models
- Not always the fastest method

Summary of Latent Variable Models

- Learning from unlabeled data
- Latent variables
  - Discrete: Clustering / Mixture models ; GMM
  - Continuous: Dimensionality reduction ; PCA
- Summary / "Understanding" of data
- Expectation Maximization Algorithm