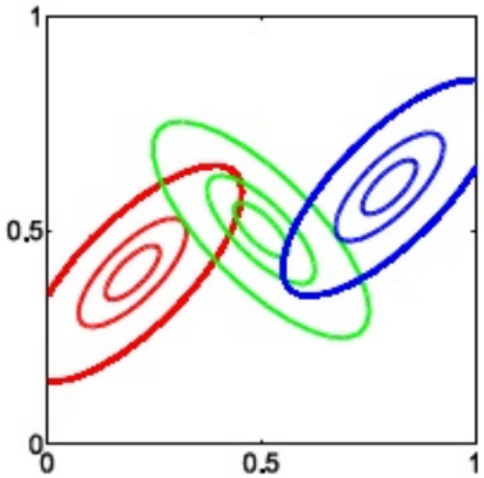
Introduction to Graphical Models

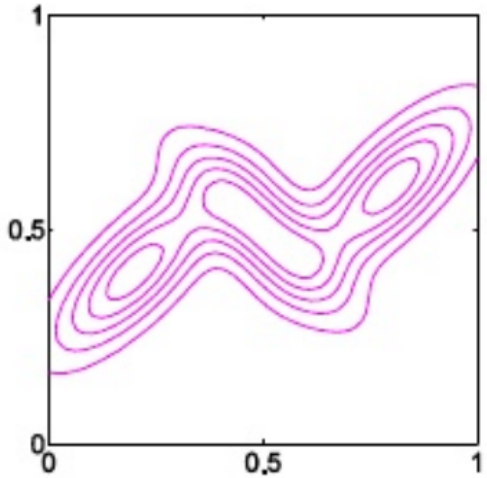
Samrachana Adhikari

December 16, 2019

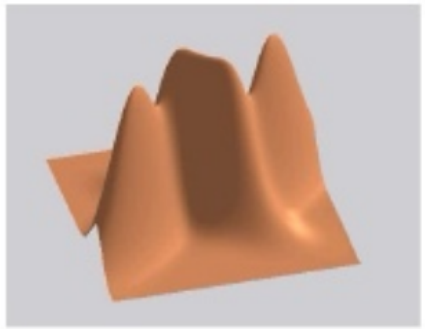
Motivation: Mixture of Distributions

$p(\mathbf{x}|k)$





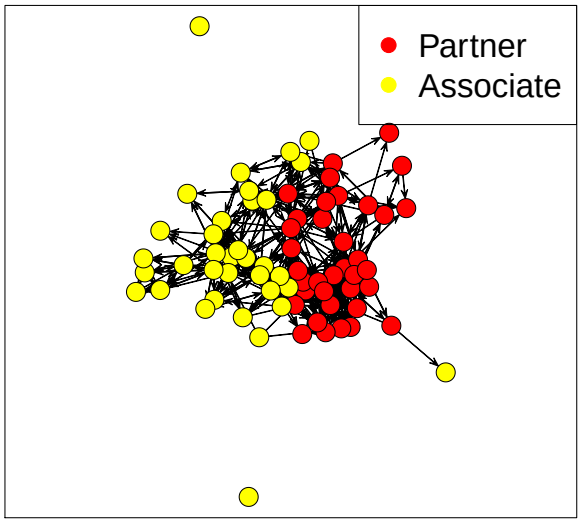
$p(\mathbf{x})$

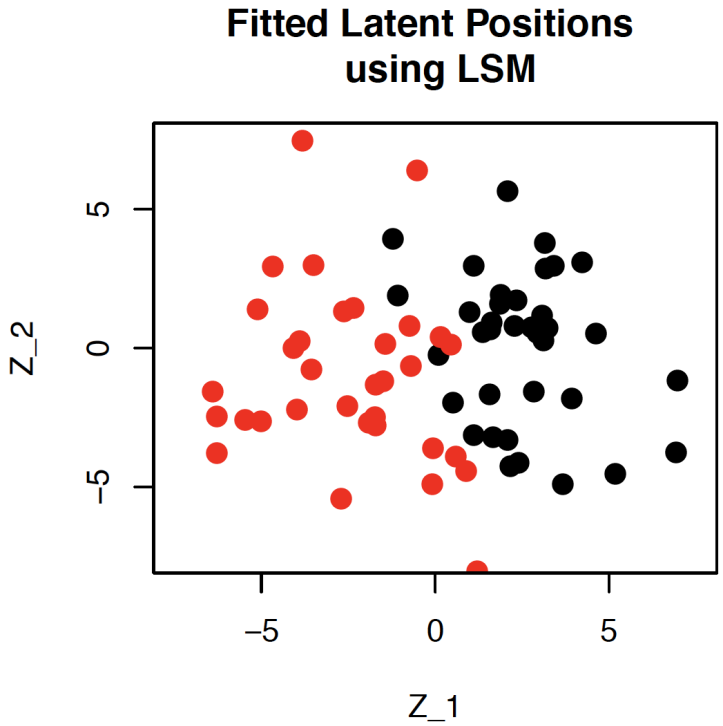


$p(\mathbf{x})$

Source: Bishop (2006).

Motivation: Social Network Analysis/ Community Detection





Source Adhikari and Dabbs (2018)

Motivation: Topic Modeling

computer	chemistry	cortex	orbit	infection
methods	synthesis	stimulus	dust	immune
number	oxidation	fig	jupiter	aids
two	reaction	vision	line	infected
principle	product	neuron	system	viral
design	organic	recordings	solar	cells
access	conditions	visual	gas	vaccine
processing	cluster	stimuli	atmospheric	antibodies
advantage	molecule	recorded	mars	hiv
important	studies	motor	field	parasite

Source Blei and Lafferty (2009)

Mixture Distribution

A distribution f is a mixture of K component distributions $f_1, f_2, \ldots, f_K$ if

$$f(x) = \sum_{k1}^{K} \lambda_k f_k(x)$$

- λ_k represent mixing weights
- $\lambda_k > 0$
- $\sum_k \lambda_k = 1$

Parametric Mixture Models

Assuming f_k are all from the same parametric distributions the mixture model becomes:

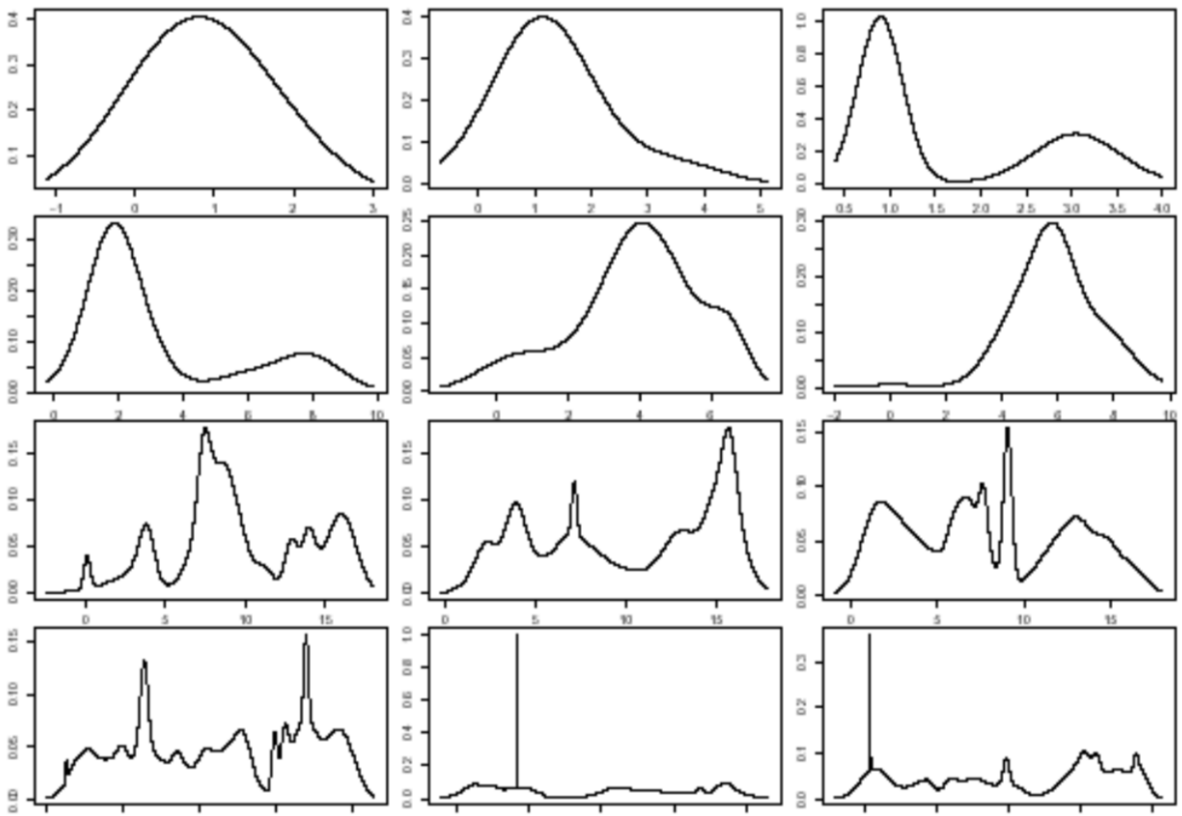
$$f(x) = \sum_{k=1}^{K} \lambda_k f(x; \Theta_k)$$

- Θ_k is the parameter vector of the k^{th} component f_k.
- Vector of parameters of the mixture model is $\Theta = (\lambda_1, \ldots, \lambda_k, \Theta_1, \ldots, \Theta_k)$

Why Mixture Models

- A flexible approach to model complex distributions
- When $K = 1$ we are back to parametric estimation
- As number of components K increases, the distribution tends to move from parametric to non-parametric representation
- A reasonable K which is usually small than the number of data points provides a good approximation of a semi-parametric density
- An approach for probabilistic clustering, with K denoting the number of clusters

The following figure shows various density functions of Gaussian mixtures with $k = 2$ components (first row), $k = 5$ components (second row), $k = 25$ components (third row) and $k = 50$ components (fourth row):

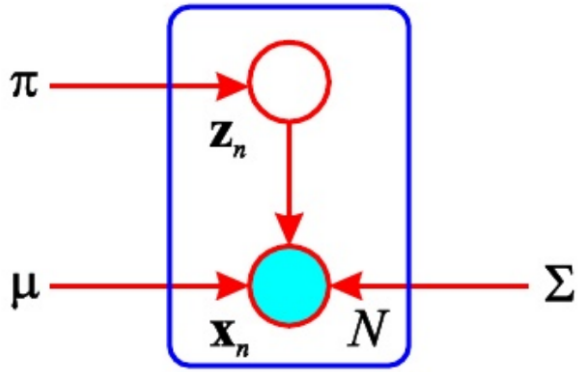


Latent Variable Point of View

- Introduce a latent variable Z
- Z represents the class membership of the random variable X. Then,

$$\begin{aligned} Z &\sim \text{Multinomial}(\lambda_1, \lambda_2, \ldots, \lambda_k) \\ f(X|Z=k) &= f(x; \Theta_k) \end{aligned}$$

Graphical Model Representation of Mixture of Gaussians



- f_ks have Gaussian or Normal distribution with different means and variances
- Nodes represent random variables
- Colored node is an observed random variable
- Symbols outside the box are parameters
- Edges represent direction of dependence

Maximum Likelihood Estimation

$$l = logp(X|\lambda, \mu, \Sigma) = \sum_{n=1}^{N} log \sum_{k=1}^{K} \lambda_k \text{Normal}(x_n|\mu_k, \Sigma_k)$$

- No closed form solution maximizing the likelihood!

Maximum Likelihood Estimation

$$\frac{\delta l}{\delta \Theta_i} = \sum_{n=1}^{N} \frac{\lambda_j f(x_i;\Theta_j)}{\sum_k . \lambda_k f(x_i;\Theta_k)} \frac{\delta log f(x_i;\Theta_j)}{\delta \Theta_j}$$

Define: $w_{ij} = \frac{\lambda_j f(x_i;\Theta_j)}{\sum_k . \lambda_k f(x_i;\Theta_k)} = P(Z_i = j | X_i, \Theta_i)$

Parameter Estimation by Iterative. Algorithms

1. Expectation Maximization
2. Sampling from posterior distribution

Expectation Maximization

- Initialization: Start with guesses about the mixture components $\Theta_1, \Theta_2, \ldots, \Theta_K$ and the mixing weights $\lambda_1, \ldots, \lambda_K$
- Until convergence, iteratively
  1. E-step: Using the current parameter guesses, calculate the weights $w_{ij} = \frac{\lambda_j f(x;\Theta)}{\sum_{k=1}^{K} \lambda_k f(x_i;\Theta_k)}$
  2. M-step: Using the current weights maximize the weighted likelihood to get new parameter estimates
- Return the final parameter estimates and cluster probabilities

Sampling from Posterior Distribution

$$
\begin{aligned}
f(X|Z=k) &= \text{Normal}(\mu_k, \Sigma_K) \\
Z &\sim \text{Multinomial}(\lambda_1, \lambda_2, \ldots, \lambda_K) \\
\mu_k &\sim \text{Normal}(0, \sigma_0^2) \\
\Sigma_k &\sim \text{Inverse-Wishart}(a, b) \\
\lambda_1, \ldots, \lambda_K &\sim \text{Dirichlet}(\alpha, K)
\end{aligned}
$$

Goal is to sample from the posterior distribution $p(\lambda, \mu, \Sigma|X)$.

Markov Chain Monte Carlo Algorithm

1. Initialize $\mu^0, \Sigma^0, \lambda^0$
2. Update Z sampling from $Z^{(j+1)} \sim Z|x, Z^{(j)}, \lambda^{(j)}, \mu^{(j)}, \Sigma^{(j)}$
3. Update μ sampling from $\mu^{(j+1)} \sim \mu|x, Z^{(j+1)}, \lambda^{(j)}, \mu^{(j)}, \Sigma^{(j)}$
4. ...
5. j = j+1. Go to step 2.

By iteratively updating and sampling from the conditional distribution for each parameter, we can sample from the joint posterior distribution of the parameter given data.

Sampling from the Posterior

- Gibbs sampling
- Metropolis-Hastings sampling
- Metropolis with Gibbs sampling
- Computationally more expensive than analytical approaches

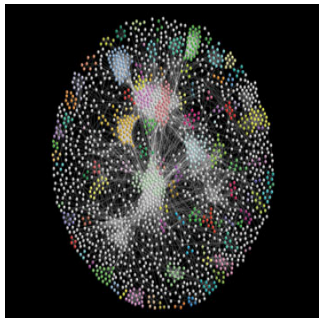
How to Choose K

- K or number of components in the mixture model needs to be pre-specified
- Classic problem in clustering or unsupervised learning
- Few ways to choose K
- Fit mixture models with a range of values for K
  1. Cross-validation: Train on a subset of data and compute likelihood on test data
  2. Choose K that maximizes information criteria
  3. Examples are: AIC, BIC or DIC

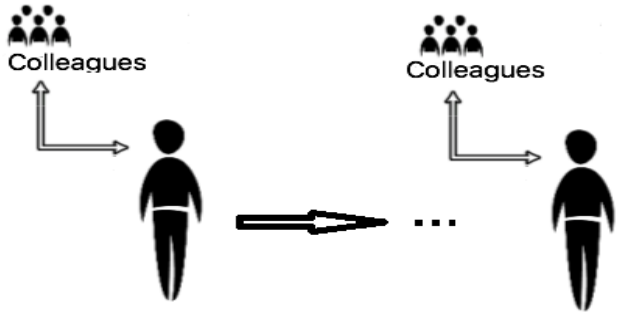
Social Network Analysis

Facebook friendship network

Network of protein interaction[1]



Advice seeking network of teachers[2]



[1] Created by Artavanis-Tsakonas Lab at HMS

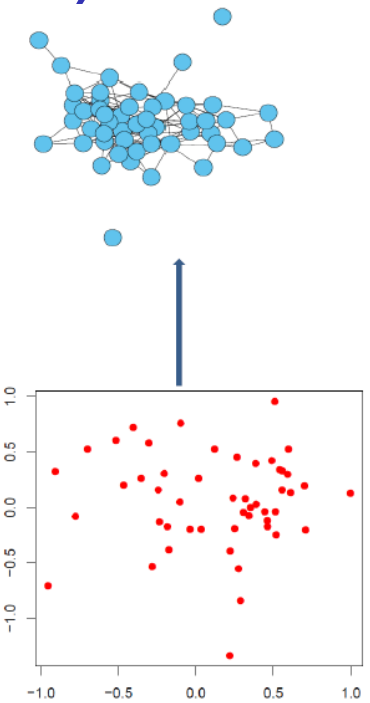
[2] (?, ?)

Latent Space Network Model (LSM) (?, ?)

Observed network Y

Z,β

Latent positions and intercept



Model

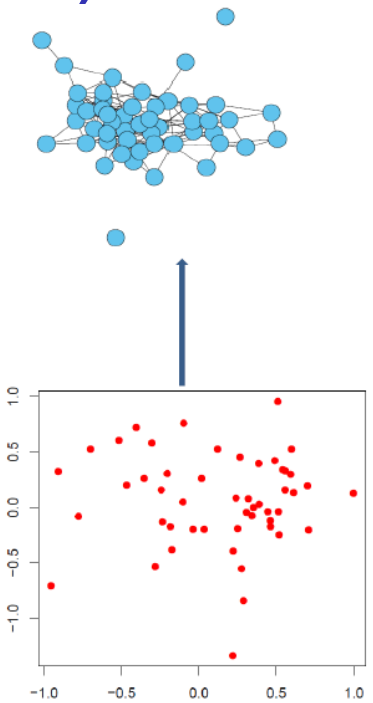
$$
\begin{aligned}
y_{ij} &\sim \text{Bernoulli}(p_{ij}) \\
\text{logit}(p_{ij}) &= \beta_0 - ||Z_i - Z_j|| \\
Z_i &\sim \text{MVN}(\mu, \Sigma)
\end{aligned}
$$

Latent Space Network Model (LSM) (?, ?)

Observed network Y

Z,β

Latent positions and intercept



Model

$$
\begin{aligned}
y_{ij} &\sim \text{Bernoulli}(p_{ij}) \\
\text{logit}(p_{ij}) &= \beta_0 - ||Z_i - Z_j|| \\
Z_i &\sim \text{MVN}(\mu, \Sigma)
\end{aligned}
$$

Topic Modeling

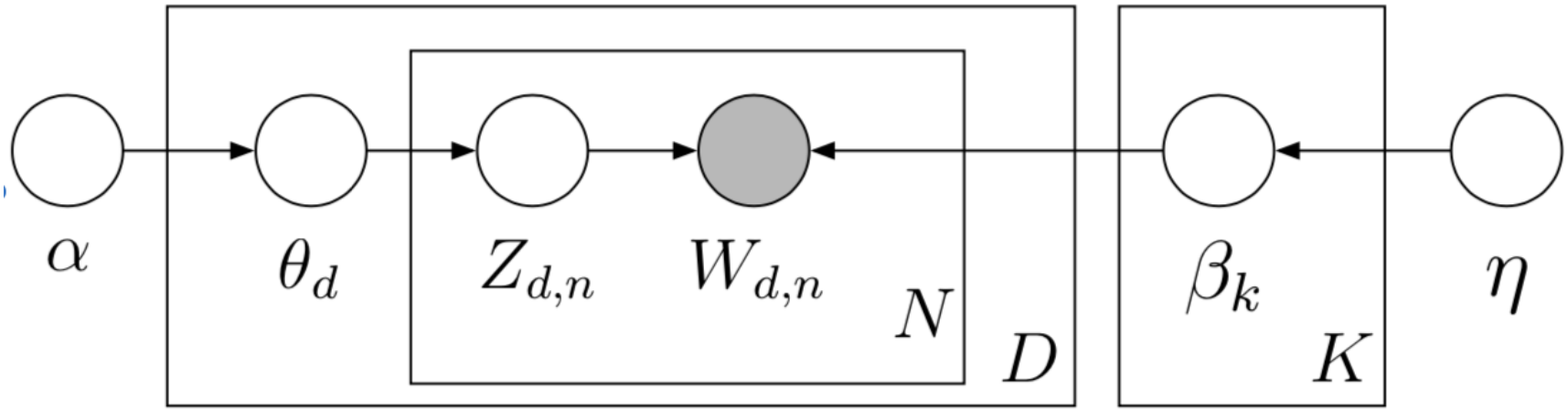
computer	chemistry	cortex	orbit	infection
methods	synthesis	stimulus	dust	immune
number	oxidation	fig	jupiter	aids
two	reaction	vision	line	infected
principle	product	neuron	system	viral
design	organic	recordings	solar	cells
access	conditions	visual	gas	vaccine
processing	cluster	stimuli	atmospheric	antibodies
advantage	molecule	recorded	mars	hiv
important	studies	motor	field	parasite

Source Blei and Lafferty (2009)

Topic Modeling

- $K =$ specified number of topics
- $V =$ size of the vocabulary, $D =$ number of document and $N =$ words per document
- $\alpha =$ a positive vector of length K
- ν is a scalar
- Topic proportions θ are distributions over topic indices
- Topics β are distributions over the vocabulary

Latent Dirichlet Allocation



LDA

1. For each topic, draw a distribution over words $\beta_k \sim \text{Dirichlet}(\nu)$
2. For each document,
   - Draw a vector of topic proportions $\theta_d \sim \text{Dirichlet}(\alpha)$
   - For each word,
     - Draw a topic assignment
       $Z_{d,n} \sim \text{Multinomial}(\theta_d), Z_{d,n} \in \{1, \ldots, K\}$
     - Draw a word $W_{d,n} \sim \text{Multinomial}(\beta_{Zd,n}, W_{d,n} \in \{1, \ldots, V\}$

Latent Dirichlet Allocation

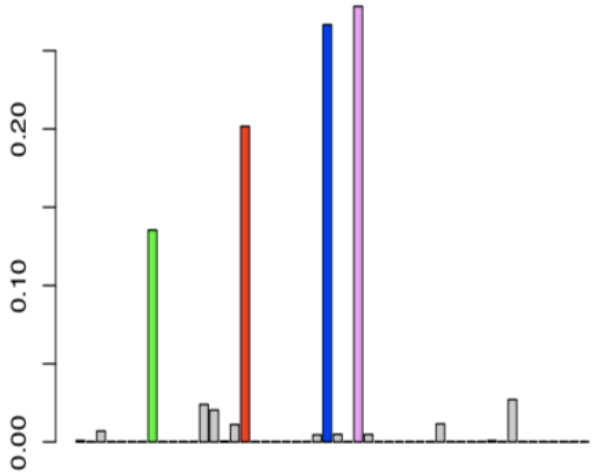
Chance and Statistical Significance in Protein and DNA Sequence Analysis

Samuel Karlin and Volker Brendel

Top words from the top topics (by term score)

sequence	measured	residues	computer
region	average	binding	methods
pcr	range	domains	number
identified	values	helix	two
fragments	different	cys	principle
two	size	regions	design
genes	three	structure	access
three	calculated	terminus	processing
cdna	two	terminal	advantage
analysis	low	site	important

Expected topic proportions



References

Bishop's book