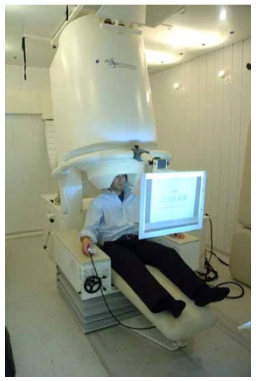
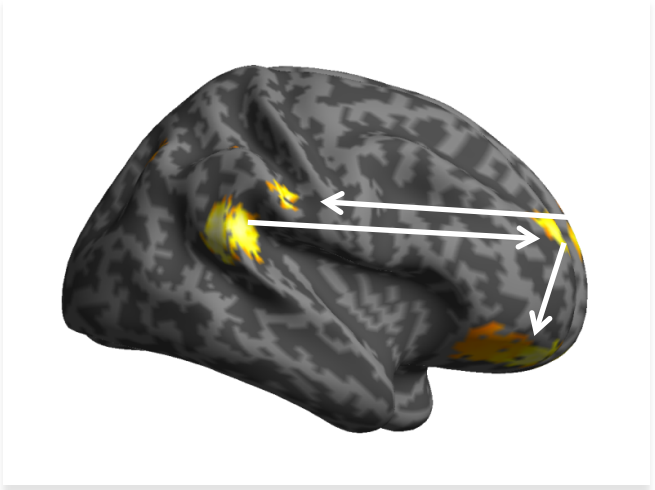
Variational Bayesian Inference

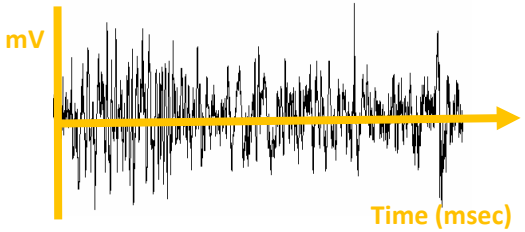
ECE 6504:
Advanced Topics in Machine Learning

Rosalyn Moran
rosalynj@vtc.vt.edu



Brain Responses



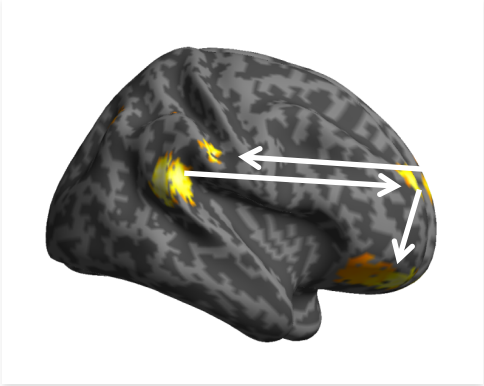


How were those data generated?

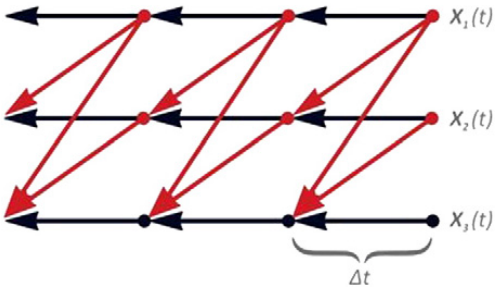
Overview

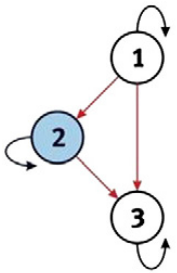
- **Problem Statement, Approach**
- KL Divergence
- Mean Field Partitioning & Structured Variational E-M
- Example: Gaussian Mixture Model

Problem Statement



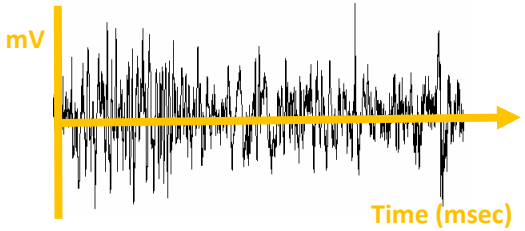
Brain Process, {x, θ}:





Valdes-Sosa et al, 2011

Imaging data, y:



Dynamic Causal Models:
A Probabilistic Graphical Model (DAG)
Including latent variables x, observed variables y, model parameters θ and priors over parameters
(more Thursday)

Today's Lecture

Main Issues in PGMs

VB: A procedure to do inference:
That implicitly 'does double duty' in Directed Graphs!

- **Representation**
  - How do we store $P(X_1, X_2, \ldots, X_N)$
  - What does my model mean/imply/assume? (Semantics)

- **Inference**
  - How do I answer questions/queries with my model, such as
  - Marginal Estimation: $P(X_5 \mid X_1, X_4)$
  - Most Probable Explanation: argmax $P(X_1, X_2, \ldots, X_N)$

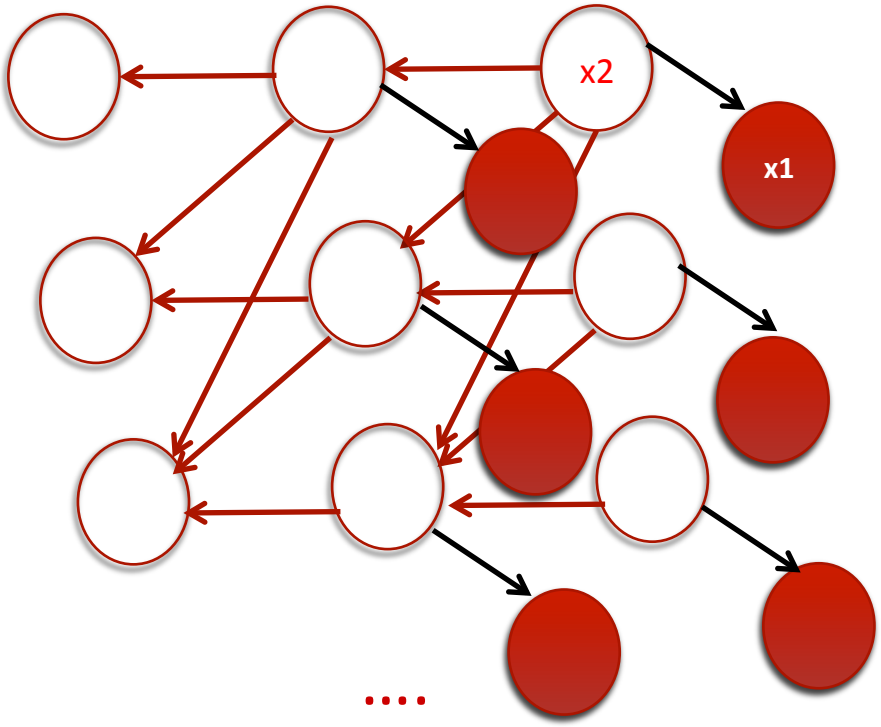
- **Learning**
  - How do we learn parameters and structure of $P(X_1, X_2, \ldots, X_N)$ from data
  - What is the right model for my data?

Today's Lecture

Motivation:
Inference

X : unobserved (latent variables)
& observed variables

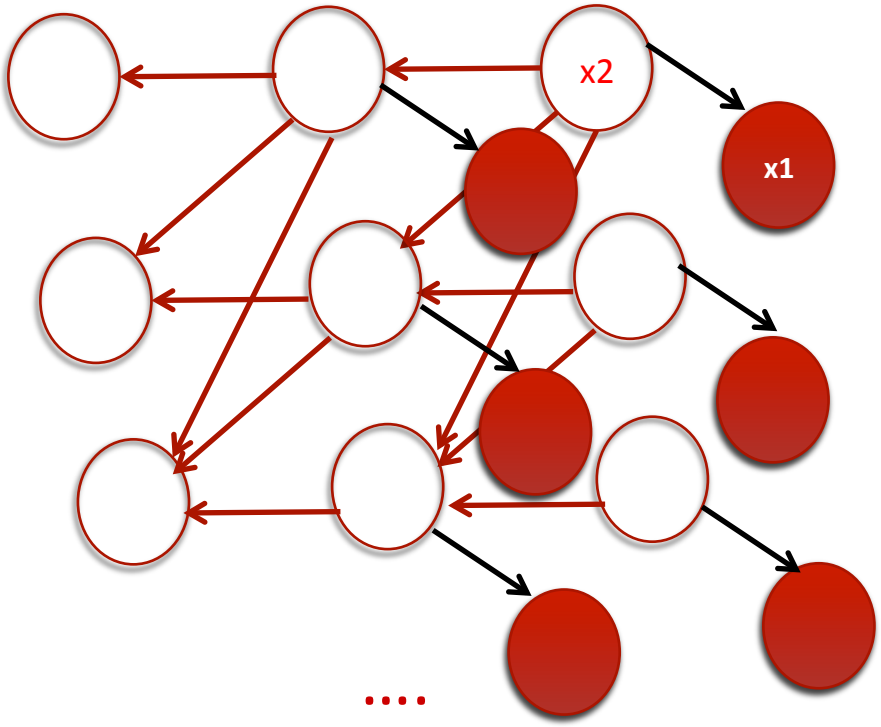
Query conditional X eg. $P(x_1 \mid x_2)$



Today's Lecture

Motivation: Inference

A large graph : computationally expensive using exact methods
(If Continuous RV's : Integrals may be intractable)



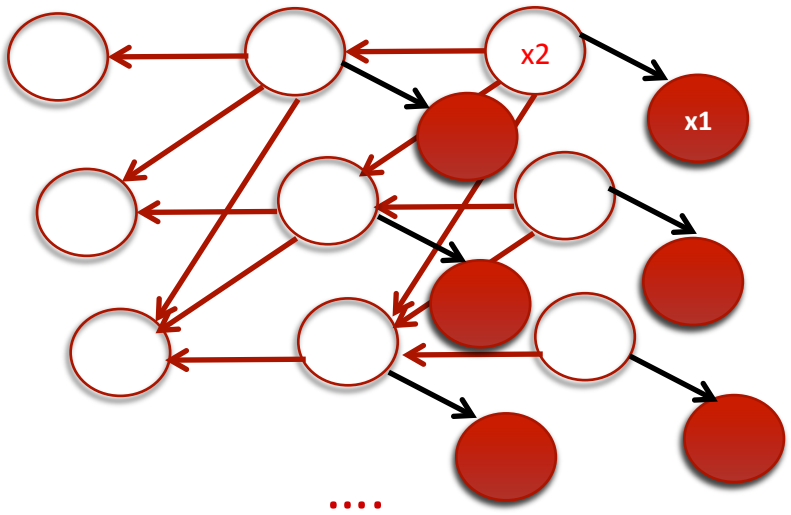
VB Procedure in a Nutshell

Solution: Approximate Inference using constrained optimization

Where: The approximation arises from

1) **constructing an approximating distribution over X: q(X). This distribution will be simpler than the true or 'target' distribution p(X)**

and 2) a factorization - exploited for q(X), mean field procedure

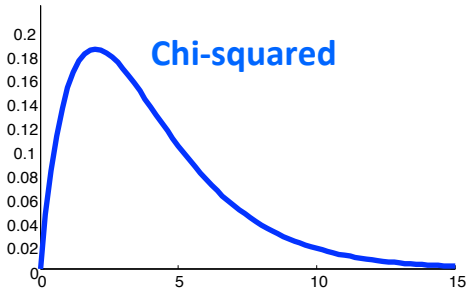


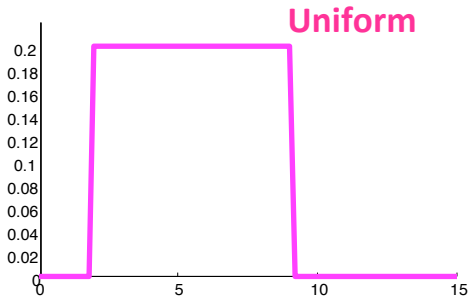
Overview

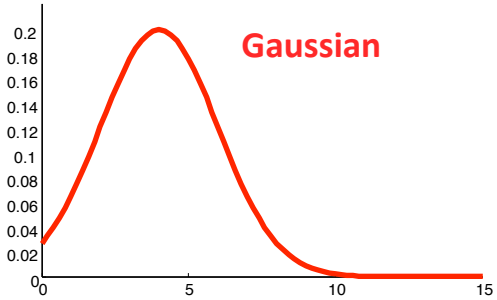
- Problem Statement, Approach
- **KL Divergence**
- Mean Field Partitioning & Structured Variational E-M
- Example(s): Gaussian Mixture Model
  Inferring Model Order in an AR process

Approach

Similarity:







Recall from Information Theory:

Shannon (1948): how much information is received when we observe a specific Value of the variable x?

$$h(x_i) = \log \frac{1}{p(x_i)}$$ Self Information/Surprise

$$h(x_i, y_i) = h(x_i) + h(y_i)$$

Total information sums when x_i and y_i are independent & a monotonic function

Potential Cost Functions

Deriving Cost Function, Motivation 1:

The average surprise of a random variable is the entropy. For discrete RV's:

$$H(x) = \sum_i p(x_i) \log \frac{1}{p(x_i)}$$

for continuous RV's:

$$H(x) = \int p(x) \log \frac{1}{p(x)} dx$$

Mutual Information:

$$I(x,y) = \sum_i \sum_j p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)\,p(y_j)}$$

Given two RVs, *x* and *y* with a joint probability mass/density function *p(x,y)* and marginal probability densities *p(x)* and *p(y)*; the mutual information *I(x,y)* is the relative entropy between the joint distribution and the product of the marginals

Potential Cost Functions

Deriving Cost Function, Motivation 1:

Searching for a distribution *q*, which is "close" to *p*.
Could employ Euclidean distances,
Often use a divergence term: The Kullback-Leibler Divergence or the

Relative Entropy between q and p: $$KL(q\|p) = \sum q(x) \log \frac{q(x)}{p(x)}$$

q: the *approximate or proposal* distribution

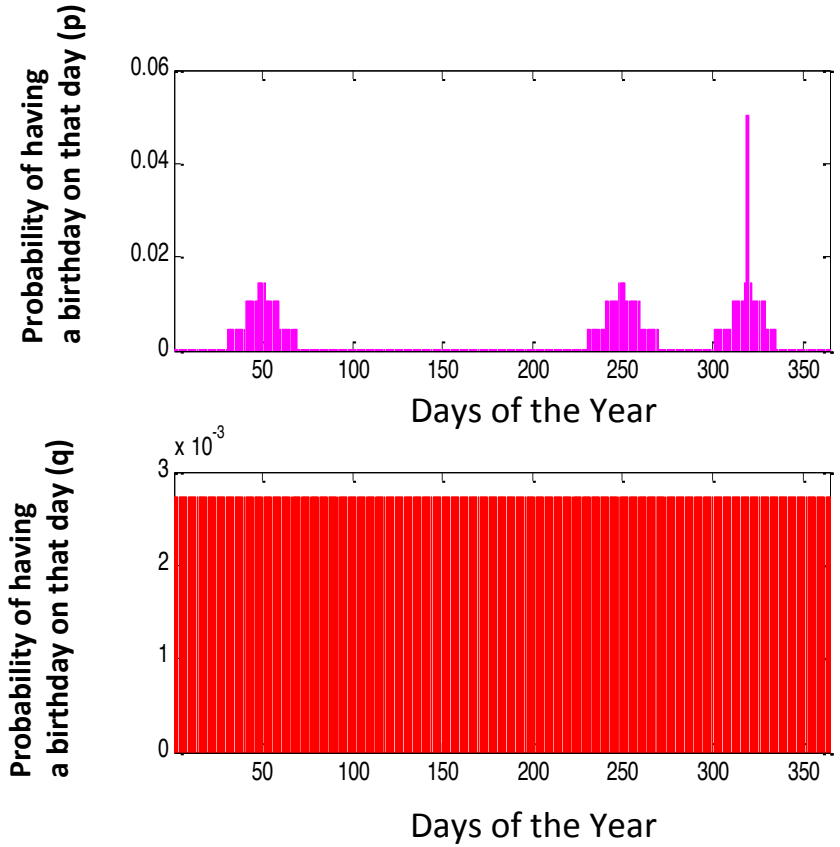
KL
- always nonnegative
- equals zero iff *p = q*
- Not symmetric

for continuous RV's:

$$KL(q\|p) = \int q(x) \ln \frac{q(x)}{p(x)} dx$$

$$KL(q\|p) = E_q \left\{ \ln \frac{q(x)}{p(x|y)} \right\}$$

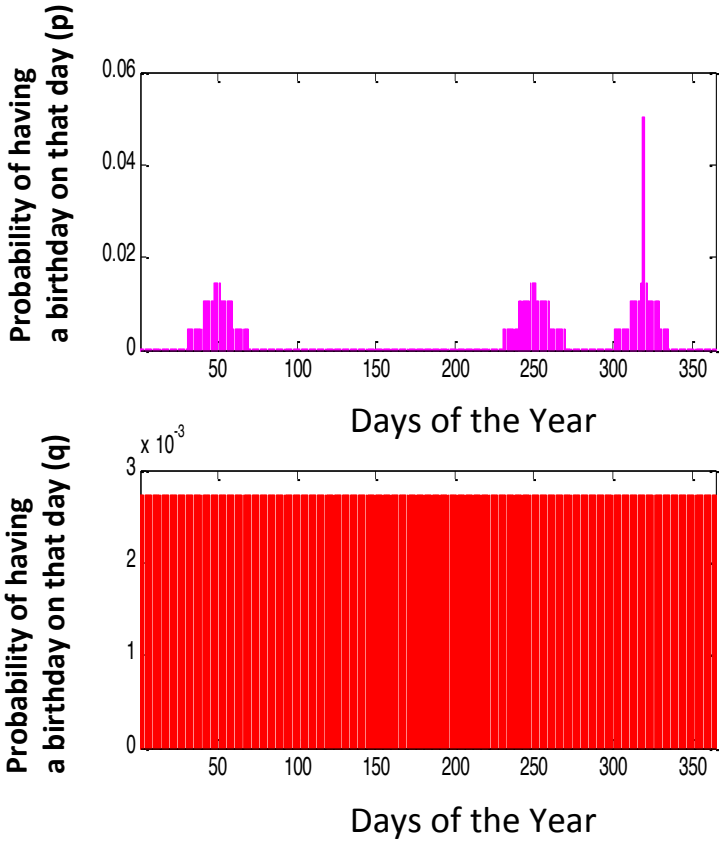
KL Divergence

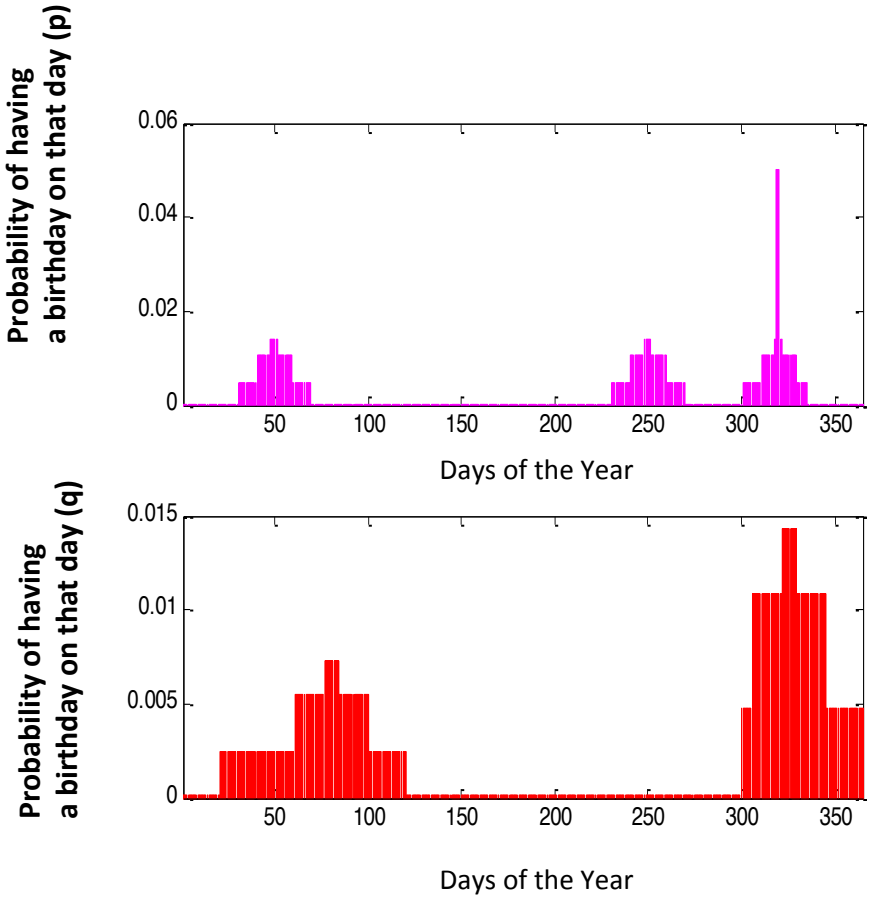


KL[q|p] = 3.53 nats

$$KL(q\|p) = \sum q(x) \log \frac{q(x)}{p(x)}$$

KL Divergence



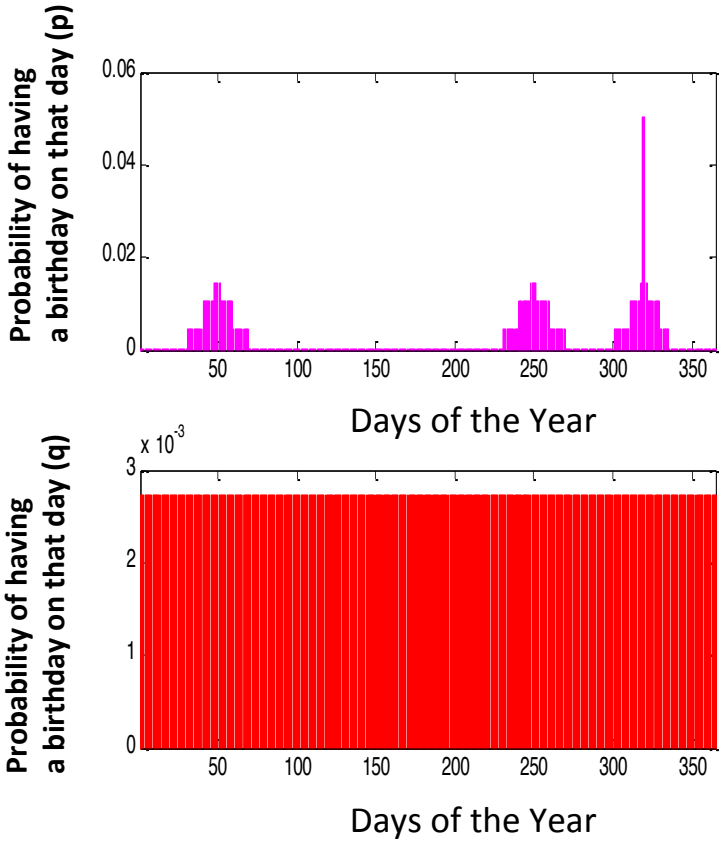


KL[q | p] = 3.53 nats

KL[q | p] = 3.21 nats

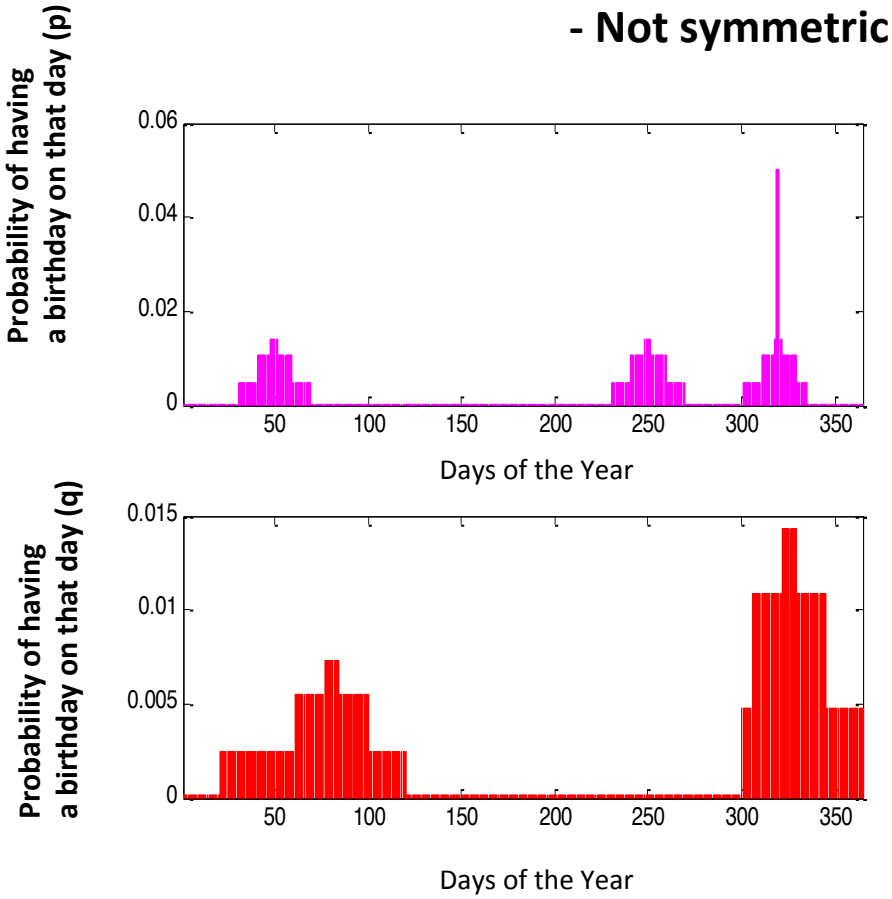
$$KL(q\|p) = \sum q(x) \log \frac{q(x)}{p(x)}$$

KL Divergence



KL[q | p] = 3.53 nats

- Not symmetric



KL[q | p] = 3.21 nats

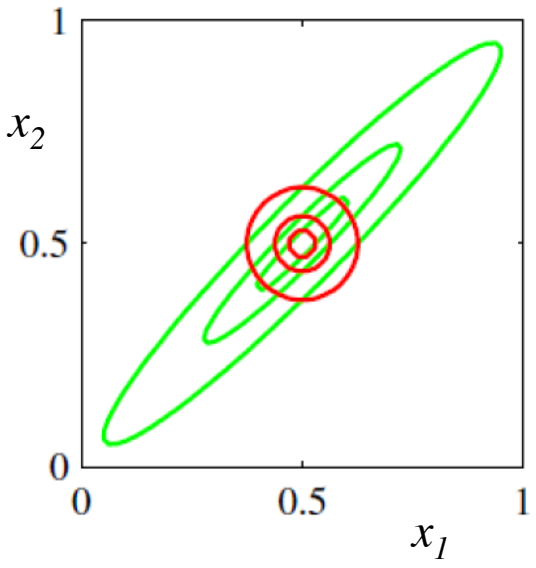
KL[p | q] = 2.41 nats

$$KL(p\|q) = \sum p(x) \log \frac{p(x)}{q(x)}$$

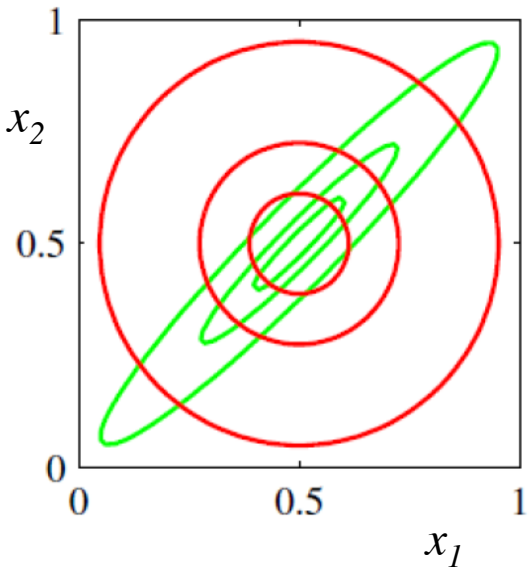
KL Divergence

The KL Divergence Criteria: $q(x) = q(x_1)q(x_2)$

Minimizing KL(q , p)



Minimizing *KL(p , q)*



Bishop, 10.2

Minimizing *KL(q,p)* tends to produce approximations where uncertainty is underestimated

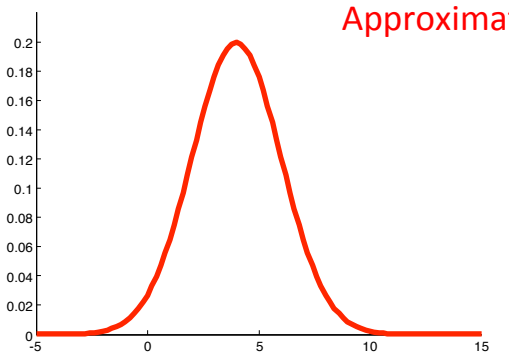
KL Divergence

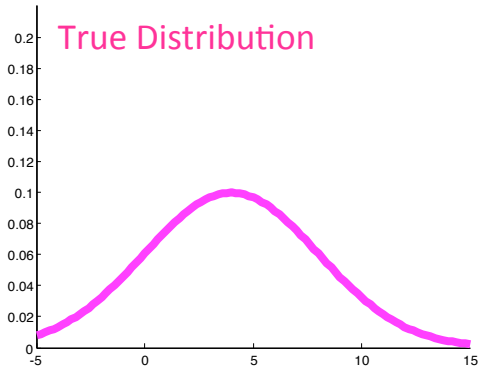
Search for a distribution q: that minimizes

$$KL(q\|p) = \sum q(x) \log \frac{q(x)}{p(x)}$$

- always nonnegative
- equals zero iff $p = q$
- Not symmetric

The Kullback Leibler Divergence *from* the approximate distribution *to* the true distribution





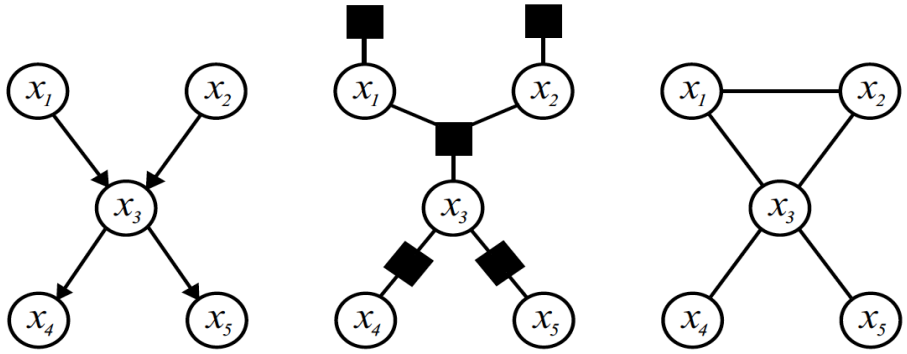
Why "q to p"?

KL Divergence & Energy Functional

Recall from previous lecture

: Joint Probabilities

- Factorization in Markov Networks
- Conditional Probabilities in Bayes Nets



$$p(X) = \frac{1}{Z}\prod_c \psi_c(x_c)$$

- *Joint distribution defined as the product of potentials: non-negative functions. In general chosen to represent preferred configurations of local variables (eg. correlations)*

$$\psi_c(x_c) = \exp(-E(x_c))$$

- *Represented as exponentials these are Boltzmann distributions and E is the energy function*

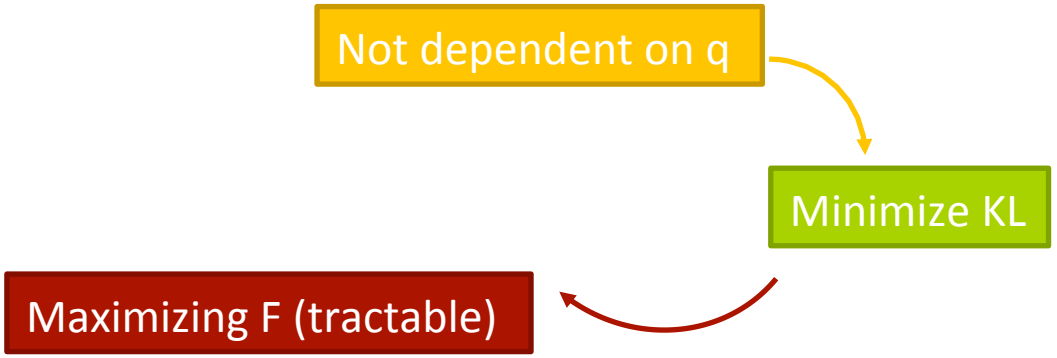
$$Z = \sum_x \prod_c \psi_c(x_c)$$

- *Partition Function; a normalization function equal to the probability of the evidence in directed graphs*

KL Divergence & Energy Functional

Theorem $F(p,q) = \ln Z - KL(q \mid p)$

The Energy Functional (Free Energy) = Log Partition Function - KL



Importantly:
Since KL is always positive F represents a lower bound on the log partition function,
A lower bound on the log evidence which can be used to score models

$$\ln Z \geq F(p,q)$$

KL Divergence & Energy Functional

Theorem $F(p,q) = \ln Z - KL(q \mid p)$

The Energy Functional (Free Energy) = Log Partition Function - KL

Proof $\ln Z = \ln \sum_x p(x)$

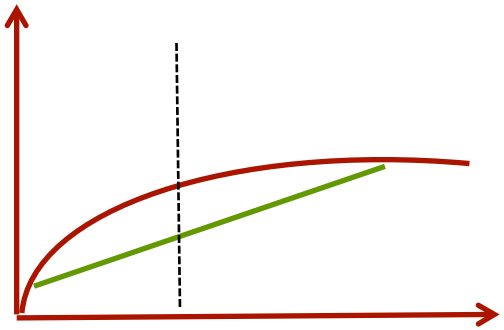
$$\ln Z = \ln \sum_x q(x) \frac{p(x)}{q(x)}$$

…Introduce *q*, algebraically

….By Jensen's Inequality:

if *X* is a random variable and ϕ is ***a concave function***, then

$\phi\{E(X)\} \geq E\{\phi(X)\}$



$$\ln \sum q(x) \frac{p(x)}{q(x)} \geq \underbrace{\sum q(x) \ln \frac{p(x)}{q(x)}}_{F}$$

KL Divergence & Energy Functional

Theorem $F(p,q) = \ln Z - KL(q \mid p)$

The Energy Functional (Free Energy) = Log Partition Function - KL

Proof $\ln Z = \ln \sum_x p(x)$

$$\ln Z = \ln \sum_x q(x) \frac{p(x)}{q(x)}$$

…Introduce *q*, algebraically

….By Jensen's Inequality:

$$\ln \sum q(x) \frac{p(x)}{q(x)} \geq \sum q(x) \ln \frac{p(x)}{q(x)}$$

….F: lower bound

$$\sum q(x) \ln \frac{p(x)}{q(x)} = \sum q(x) \ln p(x) - \sum q(x) \ln q(x)$$

…Rewriting RHS

KL Divergence & Energy Functional

Theorem $F(p,q) = \ln Z - KL(q \mid p)$

The Energy Functional (Free Energy) = Log Partition Function - KL

Proof $\ln Z = \ln \sum_x p(x)$

$$\ln Z = \ln \sum_x q(x) \frac{p(x)}{q(x)}$$

…Introduce *q*, algebraically

….By Jensen's Inequality:

$$\ln \sum q(x) \frac{p(x)}{q(x)} \geq \sum q(x) \ln \frac{p(x)}{q(x)}$$

….F: lower bound

$$\sum q(x) \ln \frac{p(x)}{q(x)} = \sum q(x) \ln p(x) - \sum q(x) \ln q(x)$$

…Rewriting RHS

$$F = \langle \ln p(x) \rangle_q + H[q] \qquad F = \sum_c \langle \ln \psi_c(x) \rangle_q + H[q]$$

…Intermediate Result
Used in Algorithmic Application

Internal Energy + Entropy

KL Divergence & Energy Functional

Theorem $$F = \ln Z - KL(q \mid p)$$

The Energy Functional (Free Energy) = Log Partition Function - KL

Proof

$$F = \langle \ln p(x) \rangle_q + H[q]$$

...But

$$KL(q \mid p) = \langle \ln q(x) \rangle_q - \langle \ln p(x) \rangle_q$$

$$\ln p(x) = \sum_{\psi_c} \ln \psi_c(x_c) - \ln Z$$

...And since by definition

$$KL(q \mid p) = -H[q] + \left\langle \sum_{x_c} \psi(x_c) \right\rangle_q + \langle \ln Z \rangle_q$$

...Z not dependent on q

KL Divergence & Energy Functional

Theorem $$F = \ln Z - KL(q \mid p)$$

The Energy Functional (Free Energy) = Log Partition Function - KL

Proof

$$F = \left\langle \ln p(x) \right\rangle_q + H\left[q\right]$$

...But

$$KL(q \mid p) = \left\langle \ln q(x) \right\rangle_q - \left\langle \ln p(x) \right\rangle_q$$

$$\ln p(x) = \sum_{\psi_c} \ln \psi_c(x_c) - \ln Z$$

...And since by definition

$$KL(q \mid p) = -H[q] - \left\langle \sum_{x_c} \psi(x_c) \right\rangle_q + \left\langle \ln Z \right\rangle_q$$

KL Divergence & Energy Functional

Theorem $$F = \ln Z - KL(q \mid p)$$

The Energy Functional (Free Energy) = Log Partition Function - KL

Proof

$$F = \left\langle \ln p(x) \right\rangle_q + H\left[q\right]$$

…But

$$KL(q \mid p) = \left\langle \ln q(x) \right\rangle_q - \left\langle \ln p(x) \right\rangle_q$$

$$\ln p(x) = \sum_{\psi_c} \ln \psi_c(x_c) - \ln Z$$

…And since by definition

$$KL(q \mid p) = -H[q] - \left\langle \sum_{x_c} \psi(x_c) \right\rangle_q + \left\langle \ln Z \right\rangle_q$$

…Simplify since
Z not dependent on q

KL Divergence & Energy Functional

Theorem

$$F(p,q) = \ln Z - KL(q \mid p)$$

The Energy Functional (Free Energy) = Log Partition Function - KL

Proof

$$F = \left\langle \ln p(x) \right\rangle_q + H\left[q\right]$$

...But

$$KL(q \mid p) = \left\langle \ln q(x) \right\rangle_q - \left\langle \ln p(x) \right\rangle_q$$

$$\ln p(x) = \sum_{\psi_c} \ln \psi_c(x_c) - \ln Z$$

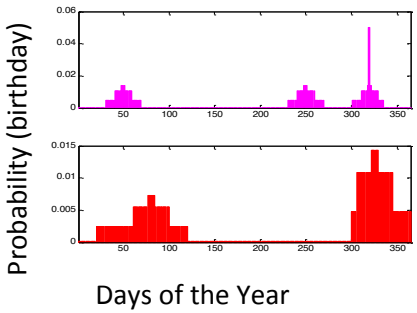
...And since by definition

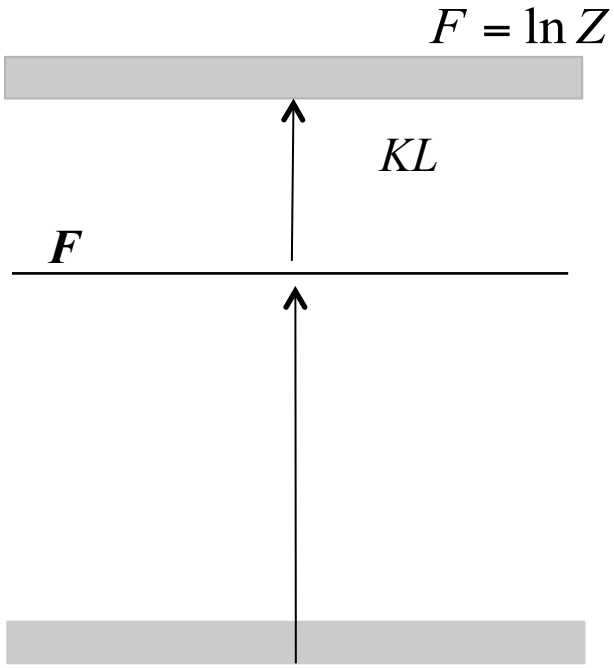
$$KL(q \mid p) = -H[q] - \left\langle \sum_{x_c} \psi(x_c) \right\rangle_q + \left\langle \ln Z \right\rangle_q$$

$$KL(q \mid p) = -F + \ln Z$$

Qed

Summary





$$F = \ln Z - KL(q \mid p)$$

$$F = \sum_c \left\langle \ln \psi_c(x) \right\rangle_q + H[q]$$

- Because KL is always positive, *F* provides a lower bound on the log model evidence. When KL is zero the densities are the same and *F* will become equal to the model's log evidence.
- *F* is a Functional, employ calculus of variations and maximize by considering different input functions, ie. explore different *q's*

Today's Lecture

Main Issues in PGMs

VB: A procedure to do inference:
That implicitly 'does double duty' in Directed Graphs!

- **Representation**
  - How do we store $P(X_1, X_2, \ldots, X_N)$
  - What does my model mean/imply/assume? (Semantics)

- **Inference**
  - How do I answer questions/queries with my model, such as
  - Marginal Estimation: $P(X_5 \mid X_1, X_4)$
  - Most Probable Explanation: argmax $P(X_1, X_2, \ldots, X_N)$

- **Learning**
  - How do we learn parameters and structure of $P(X_1, X_2, \ldots, X_N)$ from data
  - What is the right model for my data?

VB Approach in a Nutshell

Solution: Approximate Inference using constrained optimization

Where: The approximation arises from constructing an approximating distribution over X: q(X) which is closest in p(X) "in the KL sense" *(see K-F 11.1)*

So far have derived a cost function
Which can be maximized

$$F = \sum_c \left\langle \ln \psi_c(x) \right\rangle_q + H[q]$$

And is equivalent to minimizing KL(q|p)

$$F = \ln Z - KL(q \mid p)$$

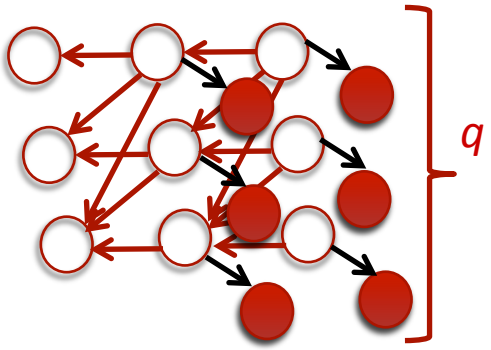
VB Approach in a Nutshell

Solution: Approximate Inference using constrained optimization

Where: The approximation arises from constructing an approximating distribution over X: q(X) which is closest in p(X) "in the KL sense" *(see K-F 11.1)*

So far have derived a cost function $F = \sum_c \langle \ln \psi_c(x) \rangle_q + H[q]$

This Energy functional contains: *an expectation over logarithms of potential functions/factors. If factors are small; the expectation can be performed easily. This depends on the choice of q*
And
The entropy of q, evaluation again dependent on choice of q



Overview

- Problem Statement, Approach
- KL Divergence
- **Mean Field Partitioning & Structured Variational E-M**
- Example: Gaussian Mixture Model

Approximate Inference

Maximize: $F = \sum_{\varphi} \left\langle \ln \varphi(x) \right\rangle_q + H\left[q\right]$

Consider a Restricted Family of Distributions &
Find the Member of this family for which the KL divergence is minimized

The Mean Field Approach: Assume the approximating or proposal density factorizes over groups of parameters - where this factorization is *a relaxation* (a superspace) of the space of true marginals.

$$q(X) = \prod_i q(x_i)$$

Where x_i is the *i*th group of parameters. Common notation: *"not i"*

$$q(X) = q(x_i)q(x_{\setminus i})$$

Fixed Point Equations

- This structured variational approach aims to optimize *F* over a *coherent* distribution *q* (ie. giving a proper joint distribution), at the expense of capturing all the information in *p*.

- Other optimization methods can trade-off information for coherency (eg. expectation propagation)

- Guaranteed convergence

- Guaranteed lower bound on *ln(Z)*

$$\max_q F = \max_q \sum_\phi E_q(\ln \phi) + \sum_i H_{q_i}(x_i)$$

$$\forall i$$

$$\sum_i q_i(x_i) = 1$$

$$q(x) = \prod_i q_i(x_i)$$

Fixed Point Equations

Constrained Optimization, solved via Lagrange multipliers, take derivative & set equal to 0

Stationary Points

Theorem: The distribution $q(x_i)$ is a local maximum of the mean-field iff

$$q(x_i) = \frac{\exp[I(x_i)]}{Z_i}$$

Where

$$I(x_i) = \sum_{\phi} E_{q_{\setminus i}} \left[\ln \phi | x_i \right]$$

Is the Variational Energy for the *i*'th partition

Fixed Point Equations

$$F = \sum_{\varphi} \langle \ln \phi(x) \rangle_q + H[q]$$

$$F_i = \sum q_i \Big[\sum_{\phi} q_{\setminus i} \ln(\phi) \Big] - \sum q_i \ln q_i + c$$

Consider terms dependent on *i*

The Variational Energy for the *i*'th partition

$$I(x_i)$$

The expectation of the log of the total joint probability *taken over parameters not* in the partition

Test: **Theorem**: The distribution $q(x_i)$ is a local maximum of the mean-field iff

$$q(x_i) = \frac{\exp[I(x_i)]}{Z_i}$$

Rewrite F_i above using result

$$F_i = \sum q_i \left[\ln \frac{\exp(I(x_i))}{q_i} \right] + c$$

Fixed Point Equations for Mean Field

Solution

$$q(x_i) = \frac{\exp[I(x_i)]}{Z}$$

Hence: The basis of variational methods using a mean field approximation is to obtain the optimal solution for a particular factor *i*, by considering the log of the joint distribution over all hidden and visible variables and then taking the Expectations with respect to all of the other factors not in *i*.

An iterative algorithm for doing inference

Expectations taken over "known" messages (ie from previous update on other factors)

Variational EM

Given the mean field approximation: $q(X) = q(x_i)q(x_{\setminus i})$

E-Step: Evaluate Expectations
of latent variables in set *i.*

M-Step: Compute Parameters set *i*
using current values of the expectations

Compute Objective Function Update
Repeat If ΔF > Threshold

Convergence to global maximum guaranteed when distributions are part of the exponential (convex) family

Ghahramani, Z., 1995

VB Approach + Algorithm in a Nutshell

Solution: Approximate Inference using constrained optimization

Where: The approximation arises from constructing an approximating distribution over X: q(X) which is closest in p(X) "in the KL sense" *(see K-F 11.1)*

So far have derived a cost function $$F = \sum_{\phi} \langle \ln \phi \rangle_q + H[q]$$

Approximated *q* using a factorization $$q(X) = \prod_i q(x_i)$$

Found iterative update equations for *q* using local stationary points

$$q(x_i) = \frac{\exp[I(x_i)]}{Z}$$

Today's Lecture

Main Issues in PGMs

VB: A procedure to do inference
- Could use $Q(X_5)$ (or a simple conditional)
- While finding *ln(Z)'s* lower bound

- **Representation**
  - How do we store $P(X_1, X_2, \ldots, X_N)$
  - What does my model mean/imply/assume? (Semantics)

- **Inference**
  - How do I answer questions/queries with my model, such as
  - Marginal Estimation: $P(X_5 \mid X_1, X_4)$
  - Most Probable Explanation: argmax $P(X_1, X_2, \ldots, X_N)$

- **Learning**
  - How do we learn parameters and structure of $P(X_1, X_2, \ldots, X_N)$ from data
  - What is the right model for my data?

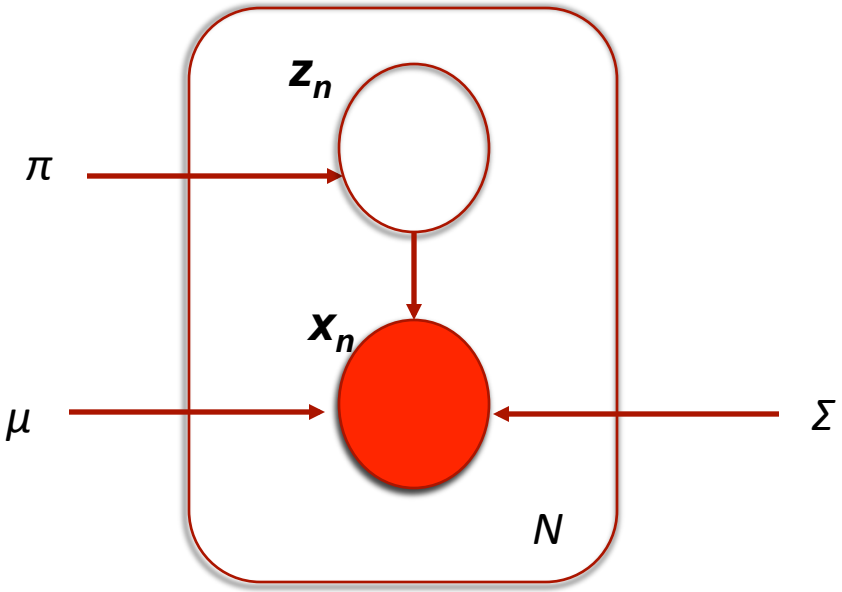
Overview

1. Problem Statement, Approach
2. KL Divergence
3. Mean Field Partitioning & Variational E-M
4. **Example: Gaussian Mixture Model**

Variational Gaussian Mixture Model (Bishop 10.2)

A Gaussian Mixture Model for a set of N i.i.d. data points,

- For each observation $\{\boldsymbol{x}_n\}$ have corresponding latent variable $\{\boldsymbol{z}_n\}$, where n = 1,…N.
  - And $\boldsymbol{z}_n$ is K-dimensional binary variable having a 1-of-K representation



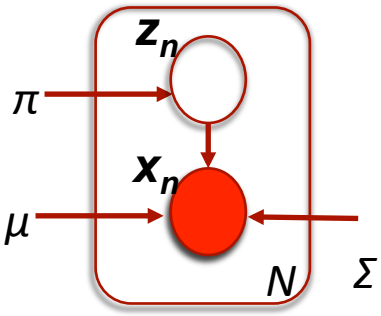
π_k: *mixing coefficients*

$$0 > \pi_k < 1$$

$$\sum_{k=1}^{K} \pi_k = 1$$

Inside the plate: variables grow with size of dataset
Outside parameters do not

Variational Gaussian Mixture Model (Bishop 10.2)



Conditional multinomial distribution over Latent Clusters given mixing coefficients

$$p(Z|\pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}}$$

Conditional Gaussian Distribution over x, given a particular value for z

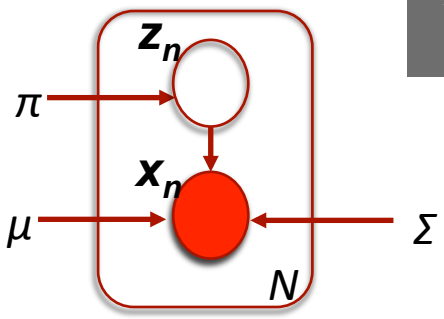
$$p(x|z_k = 1) = N(x|\mu_k, \Sigma_k)$$

Marginal Distribution of *x*

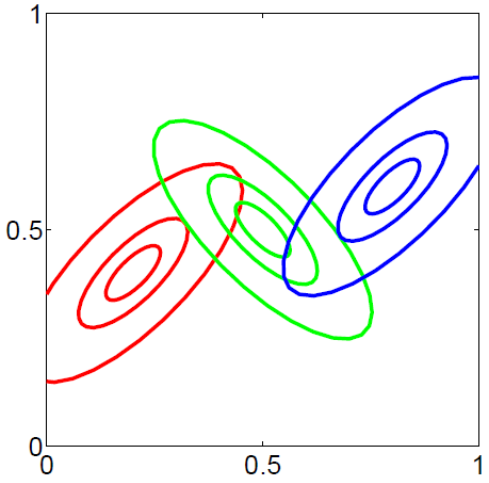
$$p(x) = \sum_z p(z)\, p(x \mid z) = \sum_{k=1}^{K} \pi_k N(x \mid \mu_k, \Sigma_k)$$

Ie. the marginal distribution of *x is a Gaussian mixture*

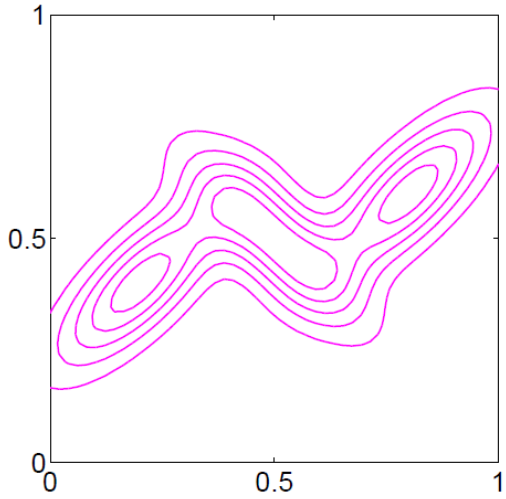
Variational Gaussian Mixture Model (Bishop 10.2)



$$p(x) = \sum_z p(z)p(x \mid z) = \sum_{k=1}^{K} \pi_k N(x \mid \mu_k, \Sigma_k)$$



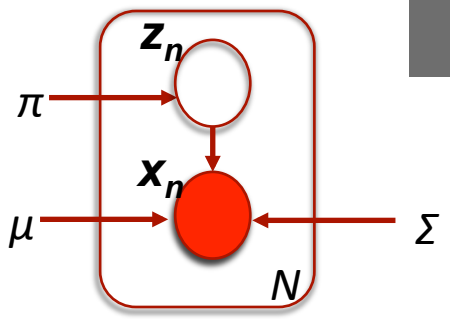
Mixture of 3 Gaussians:
Sampled from the joint distribution



Samples from the Marginal Distribution

Problem: the data set is unlabeled $p(z|x, \pi, \mu, \Lambda)$

Variational Densities



1 : Joint Distribution of all the random variables
Corresponding to the directed acyclic graph

$$p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{z}|\boldsymbol{\pi}) p(\boldsymbol{\pi}) p(\boldsymbol{\mu}|\boldsymbol{\Lambda})\, p(\boldsymbol{\Lambda})$$

2 : Distributions over parameters

Dirichlet prior for mixing coefficients $\quad p(\pi) = \mathrm{Dir}(\pi|\alpha_0) = C(\alpha_0) \prod_{k=1}^{K} \pi_k^{\alpha 0-1}$

Parameters, α_k

Gaussian-Wishart prior governing mean and precision of each Gaussian component

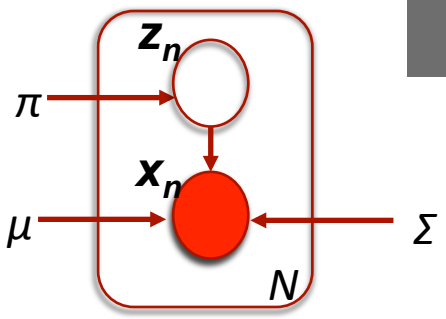
$$p(\mu, \Lambda) = \prod_{k=1}^{K} N(\mu_k \mid m_0, \beta_0^{-1}, \Lambda_k^{-1}) W(\Lambda_k \mid w_0, \nu_0)$$

Parameters, m, β, w and v

3 : Assume a variational distribution which factorizes latent variables and parameters where

$$q(z, \pi, \mu, \Lambda) = q(z|\pi, \mu, \Lambda) q(\pi, \mu, \Lambda)$$

Variational Densities



The expectation of the log of the total joint probability *taken over parameters not* in the partition

4 : From our general result

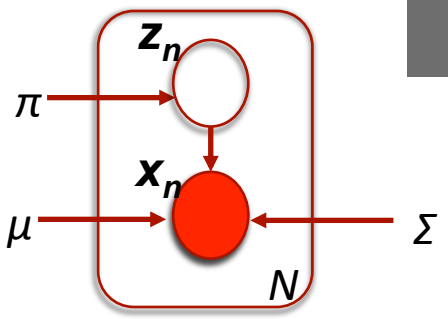
$$\ln q^*(Z) = E_{\pi,\mu,\Lambda}[\ln p(X, Z, \pi, \mu, \Lambda)] + const$$

Variational Energy for the z-partition

5 : And from the Graph

$$\ln q^*(Z) = E_{\pi}[\ln p(Z|\pi)] + E_{\mu,\Lambda}[\ln p(X|Z, \mu, \Lambda)] + const$$

Variational Densities



4 : From our general result

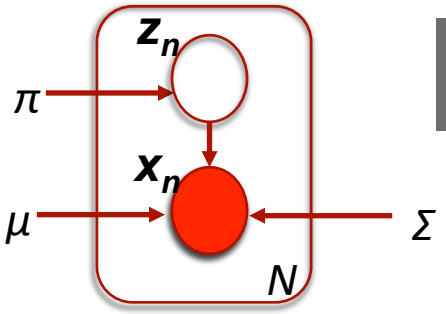
$$\ln q^*(Z) = E_{\pi,\mu,\Lambda}[\ln p(X, Z, \pi, \mu, \Lambda)] + const$$

Variational Energy for the z-partition

5 : And from the Graph

$$\ln q^*(Z) = E_{\pi}[\ln p(Z|\pi)] + E_{\mu,\Lambda}[\ln p(X|Z, \mu, \Lambda)] + const$$

"M-step": Initialise and/or "update first"



Variational Equations:

6 : Iterate Until Convergence

M-Step: Update Parameters

$$\beta_k = \beta_0 + N_k$$

$$m_k = \frac{1}{\beta_k}(\beta_0 m_0 + N_k \bar{x}_k)$$

$$w_k^{-1} = w_0^{-1} + N_k S_k + \frac{\beta_0 N_k}{\beta_0 + N_k}(\bar{x}_k - m_0)(\bar{x}_k - m_0)^T$$

$$\nu_k = \nu_0 + N_k + 1$$

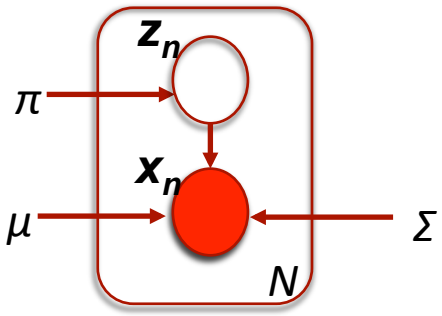
$$\alpha_k = \alpha_0 + N_k$$

E-Step: Evaluate Expectations of Labels (/Responsibilities)

$$E[z_{nk}] \propto \tilde{\pi}_k \tilde{\Lambda}_k^{1/2} \exp\left\{-\frac{D}{2\beta_k} - \frac{\nu_k}{2}(x_n - m_k)^T W_k (x_n - m_k)\right\}$$

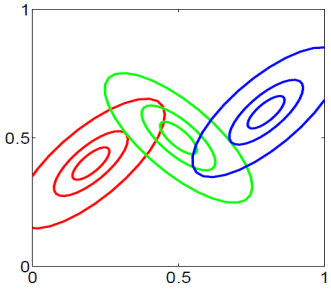
Compute Update on Objective Function
If > Threshold Repeat

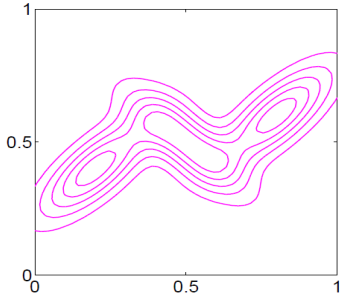
Variational Equations:



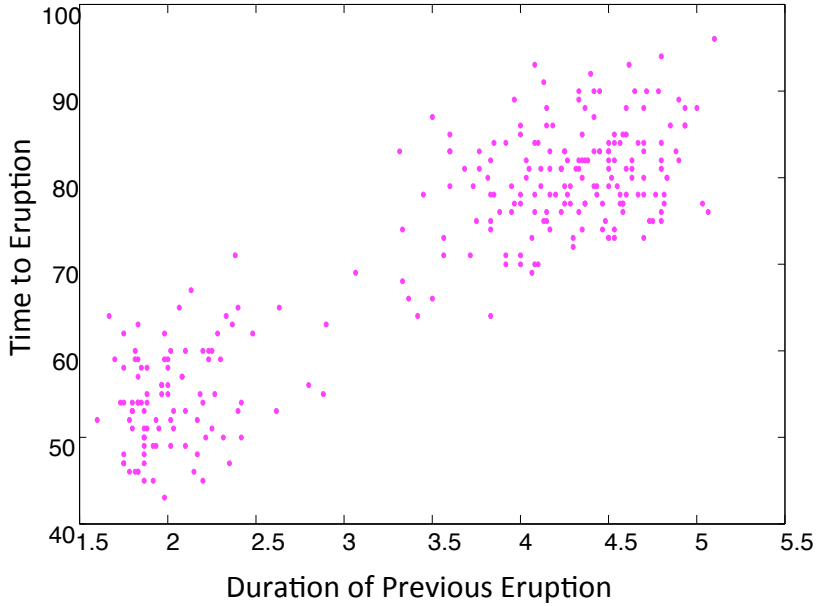
Variational EM: Demonstration

Problem: the data set is unlabeled

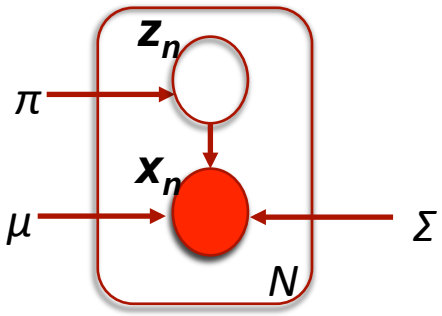




Eg. "Old Faithful" Geyser Eruption



Variational Equations:



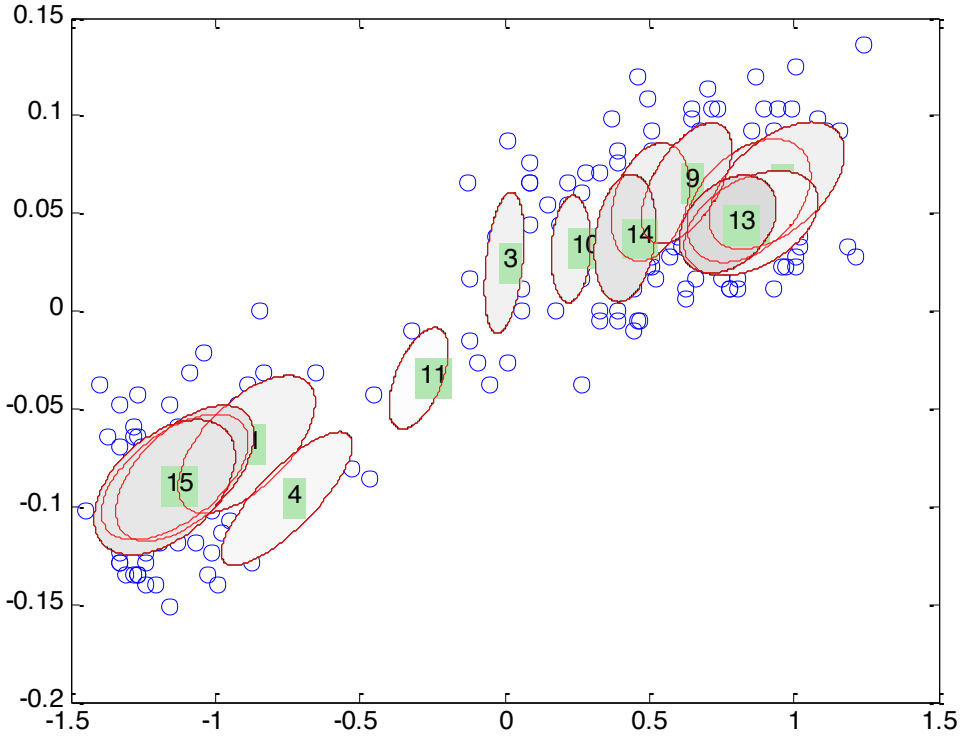
Variational EM: Demonstration

Problem: the data set is unlabeled

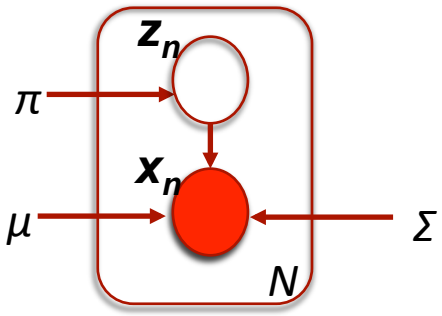
Iteration 0:

K = 15 Clusters

Alpha prior 0.001



Variational Equations:



Variational EM: Demonstration

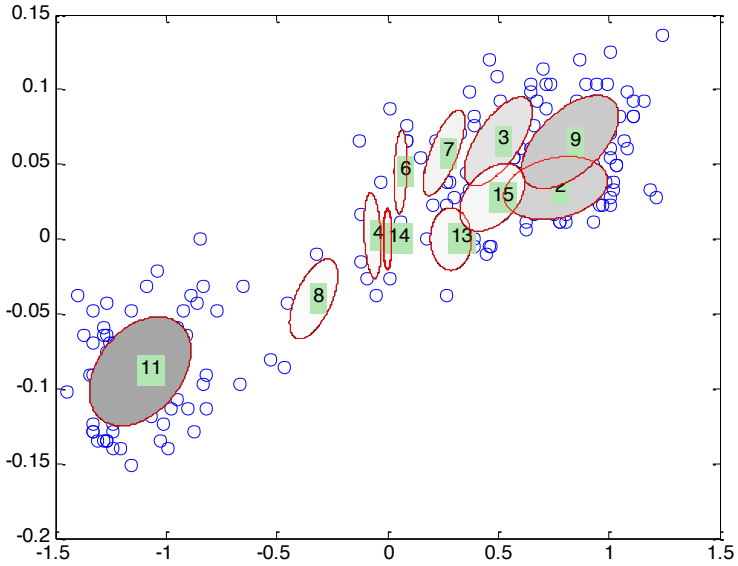
Problem: the data set is unlabeled

Iteration 10:

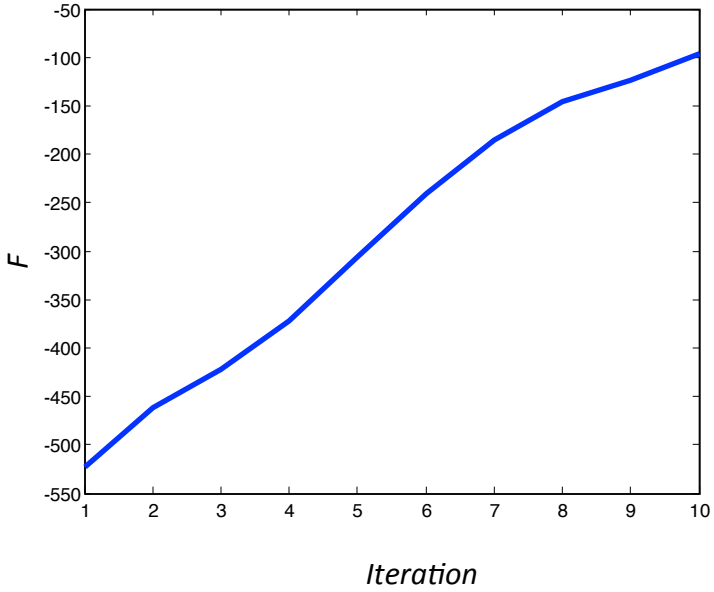
K = 15 Clusters

Alpha prior 0.001

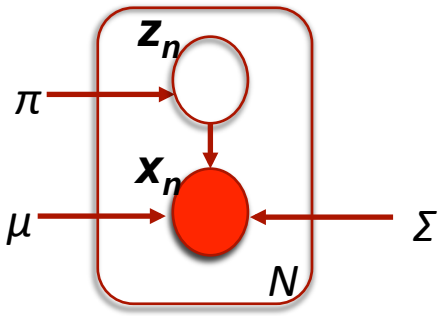
Posteriors Latent Variables:



Objective Function:



Variational Equations:



Variational EM: Demonstration

Problem: the data set is unlabeled

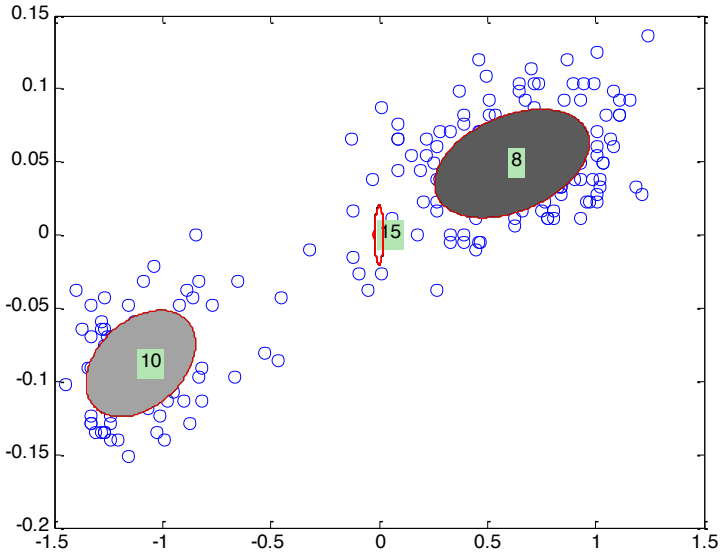
Iteration 100:

K = 15 Clusters

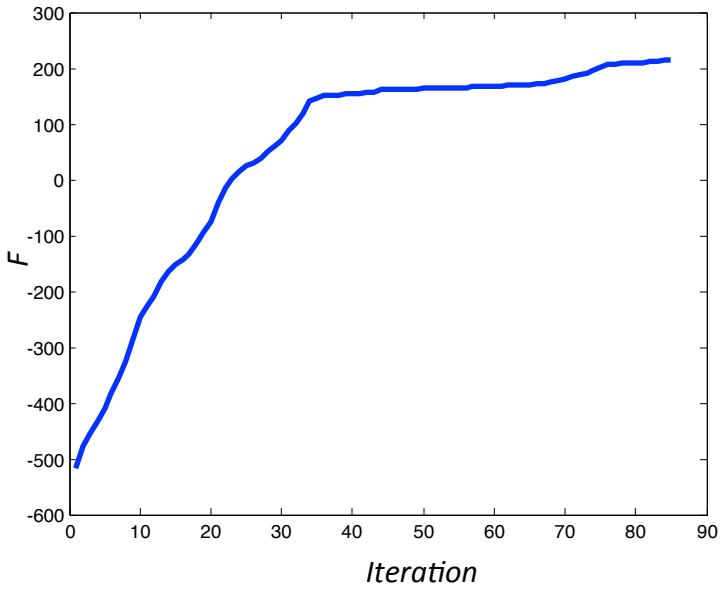
Alpha prior 0.001

Posteriors Latent Variables:

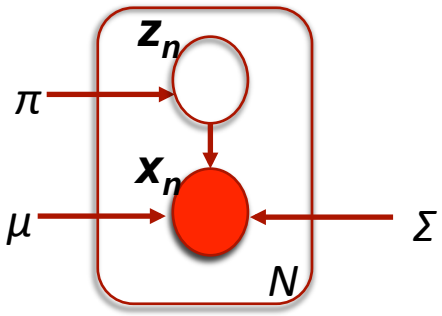
3 non-negligible components



Objective Function:



Variational Equations:



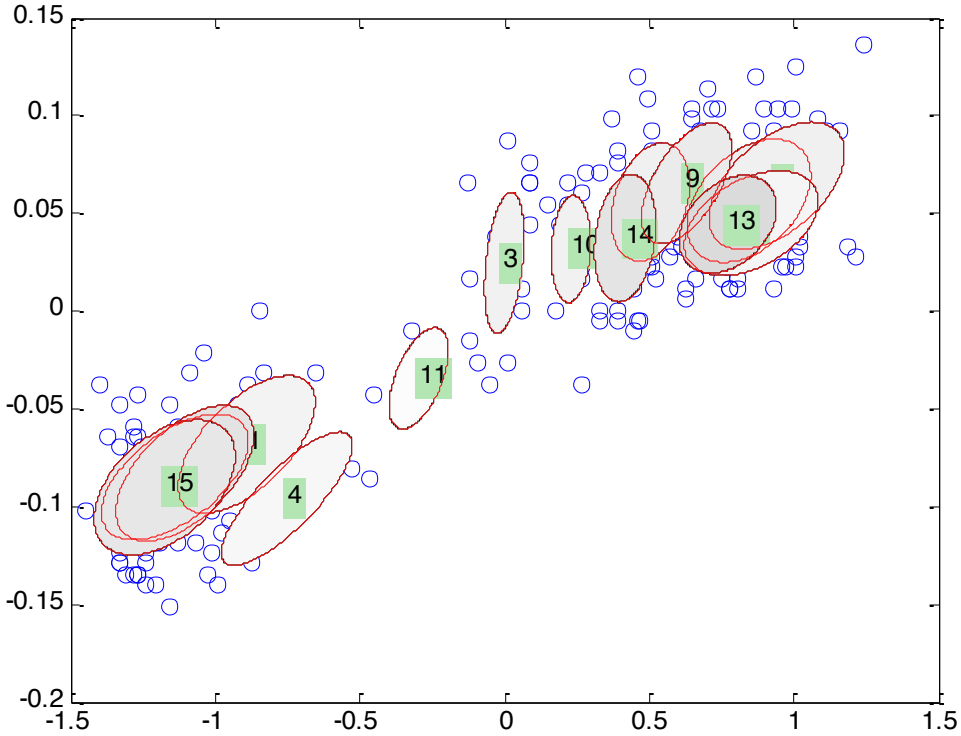
Variational EM: Demonstration

Problem: the data set is unlabeled

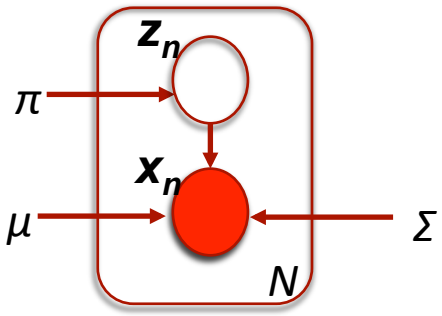
Iteration 0:

K = 15 Clusters

Alpha prior 1



Variational Equations:



Variational EM: Demonstration

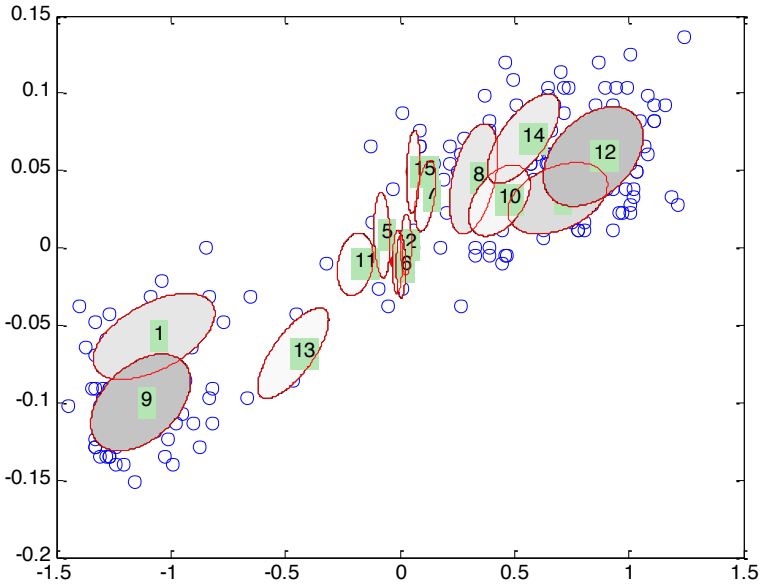
Problem: the data set is unlabeled

Iteration 10:

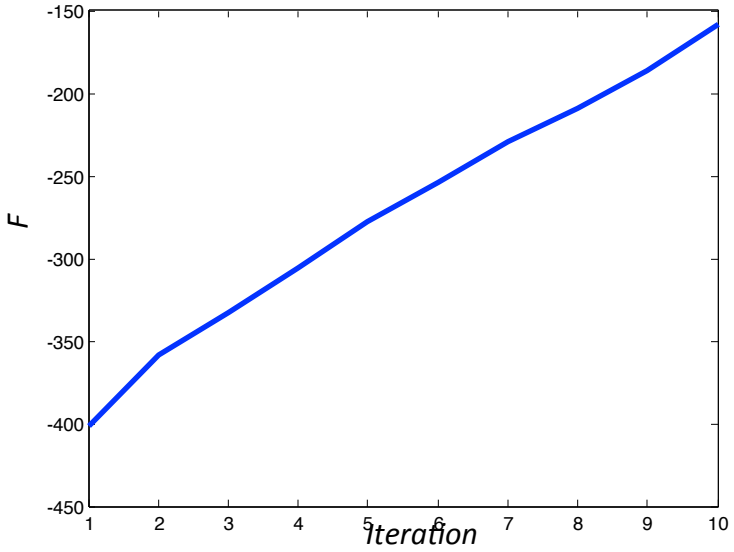
K = 15 Clusters

Alpha prior 1

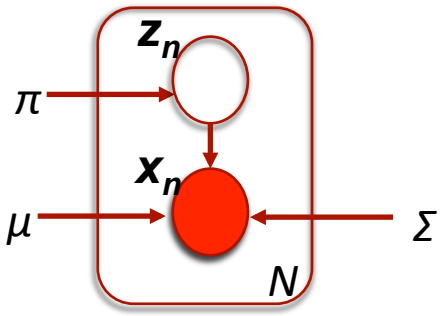
Posteriors Latent Variables:



Objective Function:



Variational Equations:



Variational EM: Demonstration

Problem: the data set is unlabeled

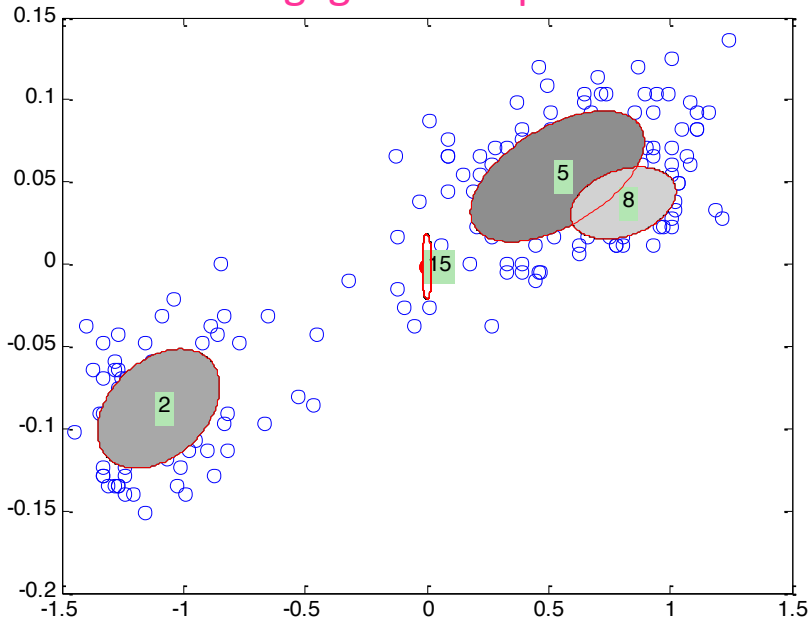
Iteration 100:

K = 15 Clusters

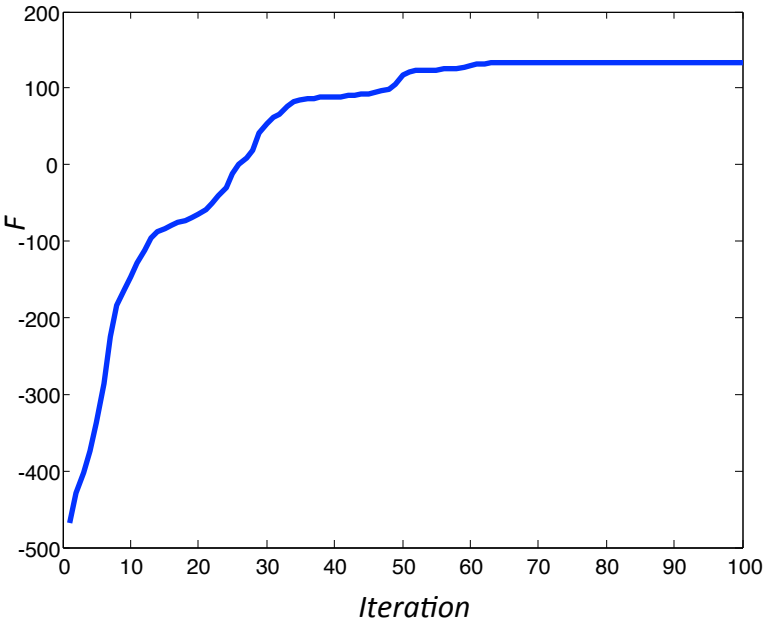
Alpha prior 1

Posteriors Latent Variables:

4 non-negligible components



Objective Function:



Summary

- Several Motivations to Arrive at Free Energy Cost Function
- Objective function's lower bound on log marginal likelihood implies utility in model comparison
- KL perspective given proposal density which should match true underlying posterior
- Mean field approach leads to EM-like update equations
- Can compute directly using partitions (shown), some algorithms employ gradient ascent
- Examples for dynamic models of brain activity next

References / Further Reading

- C. M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006.
- S.J. Roberts & W.D. Penny, Variational Bayes for Generalized Autoregressive Models, IEEE Trans Signal Proc., 2002
- D. J. C. MacKay, Information Theory, Inference and Learning Algorithms. Cambridge University Press, 2003.
- Z. Ghahramani and M. J. Beal. Propagation algorithms for variational Bayesian learning. In T. K. Leen, T. Dietterich, and V. Tresp, editors, Advances in Neural Information Processing Systems, volume 13, Cambridge MA, 2001. MIT Press.
- R. M. Neal and G. E. Hinton. A new view of the EM algorithm that justifies incremental and other variants. In M. I. Jordan, editor, Learning in Graphical Models.
- Code implemented in matlab :
- Old Faithful;: http://www1.aston.ac.uk/eas/research/groups/ncrg/resources/netlab/downloads/ and E. Khan at http://www.cs.ubc.ca/~murphyk
  AR: derived from spm: http:/fil.ion.ucl.ac.uk/spm
  rosalyn@vtc.vt.edu