Inference and Representation

DS-GA-1005, CSCI-GA.2569

Joan Bruna
Courant Institute of Mathematical Sciences
Center for Data Science
NYU

Lecture 4 Announcements

- Midterm:
  - Up until Gibbs sampling (today's lecture)
  - Long algorithmic question + conceptual questions (no long proofs) + extra (proof-based) question.

- Today:
  - BP
  - Gibbs Sampling
  - PCA and Factor Analysis

BP and Free Energy

- Let $p(x)$ be the joint distribution defined by our pairwise MRF. Consider another joint distribution $q(x)$ defined over the same domain.
- Assuming positive densities, we define a divergence

$$D_{KL}(q \mid\mid p) = \sum_x q(x) \log \frac{q(x)}{p(x)}$$

  - Kullback-Lieblier is not a distance (not symmetric and no triangle ineq.).
  - but non-negative:

$$\begin{aligned} D_{KL}(q \mid\mid p) &= \mathbb{E}_{x\sim q} \log \frac{q}{p}(x) \\ &= -\mathbb{E}_{x\sim q} \log \frac{p}{q}(x) \\ &\geq -\log \mathbb{E}_{x\sim q} \frac{p}{q}(x) \\ &= 0 \ . \end{aligned}$$

BP and Free Energy

- If we write $p(x)$ as a Gibbs distribution with energy $E(x)$

$$p(x) = \frac{1}{Z} e^{-E(x)}$$

the Kullback-Liebler divergence becomes

$$D_{KL}(q||p) = \sum_x q(x)E(x) + \sum_x q(x)\log q(x) + \log Z \ \ (\ \ \geq 0 \ ) \ .$$

- Zero divergence when

$$\sum_x q(x)E(x) + \sum_x q(x)\log q(x) := U(q) - S(q)$$

avg.energy entropy

reaches *free energy value* $F = -\log Z$.

$G(q) = U(q) - S(q)$: Gibbs free energy

Mean-Field Free Energy

- Consider an approximation $q(x)$ with separable form:

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

  - It is called *mean-field*. it does not explicitly model pairwise interactions.

Mean-Field Free Energy

- Consider an approximation $q(x)$ with separable form:

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

  - It is called *mean-field* because it does not explicitly model pairwise interactions.
  - What is the Gibbs free energy of this model when $E(x)$ is a pair-wise MRF?

$$E(x) = -\sum_{i,j} \log \psi_{ij}(x_i, x_j) - \sum_i \log \phi_i(x_i) \ .$$

Mean-Field Free Energy

- Consider an approximation $q(x)$ with separable form:

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

  - It is called *mean-field* because it does not explicitly model pairwise interactions.
  - What is the Gibbs free energy of this model when $E(x)$ is a pair-wise MRF?

$$E(x) = -\sum_{i,j} \log \psi_{ij}(x_i, x_j) - \sum_i \log \phi_i(x_i) \ .$$

  - Mean-field average Energy:

$$U(q) = -\sum_{(ij)} \sum_{x_i, x_j} q_i(x_i) q_j(x_j) \log \psi_{ij}(x_i, x_j) - \sum_i \sum_{x_i} q_i(x_i) \log \phi_i(x_i) \ .$$

$$S(q) = -\sum_i \sum_{x_i} q_i(x_i) \log q_i(x_i) \ .$$

Mean Field Free Energy

- Mean-field approximation: Minimize Gibbs Free Energy $q(x)$.
- *Variational Inference* (later in course) exploits such mean-field approximations over specific parametric families.
- The mean field model corresponds to one-node beliefs

$$q_i(x_i) \;\leftrightarrow\; b_i(x_i)$$

Mean Field Free Energy

- Mean-field approximation: Minimize Gibbs Free Energy $q(x)$.
- *Variational Inference* (later in course) exploits such mean-field approximations over specific parametric families.
- The mean field model corresponds to one-node beliefs

$$q_i(x_i) \;\leftrightarrow\; b_i(x_i)$$

- What about a two-node belief model?

Bethe Free Energy

- Let us construct a mean-field approximation that contains unary and pair-wise beliefs: b_i, b_{ij}

$$\forall\ i,j\ ,\ \sum_{x_i} b_i(x_i) = \sum_{x_i,x_j} b_{ij}(x_i,x_j) = 1\ .$$

$$\forall\ i,j\ ,\ \sum_{x_j} b_{ij}(x_i,x_j) = b_i(x_i)\ .$$

Bethe Free Energy

- Let us construct a mean-field approximation that contains unary and pair-wise beliefs: b_i, b_{ij}

$$\forall\ i,j\ ,\ \sum_{x_i} b_i(x_i) = \sum_{x_i,x_j} b_{ij}(x_i,x_j) = 1\ .$$

$$\forall\ i,j\ ,\ \sum_{x_j} b_{ij}(x_i,x_j) = b_i(x_i)\ .$$

- Under this approximation, the average energy is

$$U = -\sum_{ij}\sum_{x_i,x_j} b_{ij}(x_i,x_j)\log\psi_{ij}(x_i,x_j) - \sum_{i}\sum_{x_i} b_i(x_i)\log\phi_i(x_i)\ .$$

- Important observation: since $p(x)$ is a pair-wise MRF, its average energy has the previous form, and is exact (reaches global minima of free energy).

Bethe Free Energy

- The Entropy of a pairwise MRF does not have closed-form expression for general graphs, but for simply connected graphs we have

$$b(x) = \frac{\prod_{(ij)} b_{ij}(x_i, x_j)}{\prod_i b_i(x_i)^{d_i - 1}} \; .$$

d_i: degree of node i

Bethe Free Energy

- The Entropy of a pairwise MRF does not have closed-form expression for general graphs, but for simply connected graphs we have

$$b(x) = \frac{\prod_{(ij)} b_{ij}(x_i, x_j)}{\prod_i b_i(x_i)^{d_i - 1}} \ .$$

d_i: degree of node i

- It follows that

$$H_{\text{Bethe}} = -\sum_{(ij)} \sum_{x_i, x_j} b_{ij}(x_i, x_j) \log b_{ij}(x_i, x_j) + \sum_i (d_i - 1) \sum_{x_i} b_i(x_i) \log b_i(x_i) \ .$$

Bethe Free Energy

- The Entropy of a pairwise MRF does not have closed-form expression for general graphs, but for simply connected graphs we have

$$b(x) = \frac{\prod_{(ij)} b_{ij}(x_i, x_j)}{\prod_i b_i(x_i)^{d_i - 1}} \ .$$

d_i: degree of node i

- It follows that

$$H_{\text{Bethe}} = -\sum_{(ij)} \sum_{x_i, x_j} b_{ij}(x_i, x_j) \log b_{ij}(x_i, x_j) + \sum_i (d_i - 1) \sum_{x_i} b_i(x_i) \log b_i(x_i) \ .$$

- Thus minimizer of Bethe free energy $G_{\text{Bethe}} = U - H_{\text{Bethe}}$ contains the true Gibbs distribution $p(x)$ (recall

$$D_{KL}(q||p) = 0 \Leftrightarrow q = p \ .$$

Bethe Free Energy

- Bethe free energy: $G_{\text{Bethe}} = U - H_{\text{Bethe}}$

$$U = -\sum_{ij}\sum_{x_i,x_j} b_{ij}(x_i,x_j)\log\psi_{ij}(x_i,x_j) - \sum_{i}\sum_{x_i} b_i(x_i)\log\phi_i(x_i)\ .$$

$$H_{\text{Bethe}} = -\sum_{(ij)}\sum_{x_i,x_j} b_{ij}(x_i,x_j)\log b_{ij}(x_i,x_j) + \sum_{i}(d_i-1)\sum_{x_i} b_i(x_i)\log b_i(x_i)\ .$$

- On simply connected graphs, BP beliefs are global minima of the Bethe free energy.

Bethe Free Energy

- Bethe free energy: $G_{\text{Bethe}} = U - H_{\text{Bethe}}$

$$U = -\sum_{ij}\sum_{x_i,x_j} b_{ij}(x_i,x_j)\log\psi_{ij}(x_i,x_j) - \sum_{i}\sum_{x_i} b_i(x_i)\log\phi_i(x_i)\ .$$

$$H_{\text{Bethe}} = -\sum_{(ij)}\sum_{x_i,x_j} b_{ij}(x_i,x_j)\log b_{ij}(x_i,x_j) + \sum_{i}(d_i-1)\sum_{x_i} b_i(x_i)\log b_i(x_i)\ .$$

- On simply connected graphs, BP beliefs are global minima of the Bethe free energy.

- On general graphs, the Bethe Free Energy does not satisfy $G_{\text{Bethe}} \geq -\log Z$

Bethe Free Energy

- Bethe free energy: $G_{\text{Bethe}} = U - H_{\text{Bethe}}$

$$U = -\sum_{ij}\sum_{x_i,x_j} b_{ij}(x_i,x_j)\log\psi_{ij}(x_i,x_j) - \sum_{i}\sum_{x_i} b_i(x_i)\log\phi_i(x_i)\ .$$

$$H_{\text{Bethe}} = -\sum_{(ij)}\sum_{x_i,x_j} b_{ij}(x_i,x_j)\log b_{ij}(x_i,x_j) + \sum_{i}(d_i-1)\sum_{x_i} b_i(x_i)\log b_i(x_i)\ .$$

- On simply connected graphs, BP beliefs are global minima of the Bethe free energy.

- On general graphs, the Bethe Free Energy does not satisfy $G_{\text{Bethe}} \geq -\log Z$
- However, they provide a powerful characterization of BP solutions:

  A set of beliefs gives BP a fixed point in any graph G if and only if they are stationary points of the Bethe free energy.

Bethe Free Energy

- We construct a Lagrangian $\mathcal{L}(b)$ corresponding to the constraints

$$\forall\ i,j,x_i\ ,\ b_i(x_i) = \sum_{x_j} b_{ij}(x_i,x_j) \rightarrow\ \lambda_{ij}(x_i)$$

$$\forall\ i,j\ ,\ \sum_{x_i}\sum_{x_j} b_{ij}(x_i,x_j) = 1 \rightarrow\ \gamma_{ij}$$

$$\forall\ i,\ \sum_{x_i} b_i(x_i) = 1 \rightarrow\ \gamma_i$$

Bethe Free Energy

- We construct a Lagrangian $\mathcal{L}(b)$ corresponding to the constraints

$$\forall\ i, j, x_i\ ,\ b_i(x_i) = \sum_{x_j} b_{ij}(x_i, x_j) \rightarrow\ \lambda_{ij}(x_i)$$

$$\forall\ i, j\ ,\ \sum_{x_i}\sum_{x_j} b_{ij}(x_i, x_j) = 1 \rightarrow\ \gamma_{ij}$$

$$\forall\ i,\ \sum_{x_i} b_i(x_i) = 1 \rightarrow\ \gamma_i$$

- From satisfy $\dfrac{\partial \mathcal{L}(b)}{\partial b_{ij}(x_i, x_j)} = 0$ $\dfrac{\partial \mathcal{L}(b)}{\partial b_i(x_i)} = 0$, stationary points of BFE

$$\log b_{ij}(x_i, x_j) = \log \psi_{ij}(x_i, x_j) + \log \phi_i(x_i) + \log \phi_j(x_j) + \lambda_{ij}(x_j) + \lambda_{ji}(x_i) + \gamma_{ij} - 1$$

$$(d_i - 1)(\log b_i(x_i) + 1) = -(1 - d_i) \log \phi_i(x_i) + \sum_{j \in N(i)} \lambda_{ji}(x_i) + \gamma_i$$

Bethe Free Energy and BP

$$\log b_{ij}(x_i, x_j) = \log \psi_{ij}(x_i, x_j) + \log \phi_i(x_i) + \log \phi_j(x_j) + \lambda_{ij}(x_j) + \lambda_{ji}(x_i) + \gamma_{ij} - 1$$

$$(d_i - 1)(\log b_i(x_i) + 1) = -(1 - d_i) \log \phi_i(x_i) + \sum_{j \in N(i)} \lambda_{ji}(x_i) + \gamma_i$$

- Now, if we suppose messages/beliefs that are fixed point of BP, we define $\lambda_{ij}(x_j) = \log \prod_{k \in N(j) \setminus i} m_{kj}(x_j)$

Bethe Free Energy and BP

$$\log b_{ij}(x_i, x_j) = \log \psi_{ij}(x_i, x_j) + \log \phi_i(x_i) + \log \phi_j(x_j) + \lambda_{ij}(x_j) + \lambda_{ji}(x_i) + \gamma_{ij} - 1$$

$$(d_i - 1)(\log b_i(x_i) + 1) = -(1 - d_i) \log \phi_i(x_i) + \sum_{j \in N(i)} \lambda_{ji}(x_i) + \gamma_i$$

- Now, if we suppose messages/beliefs that are fixed point of BP, we define $\lambda_{ij}(x_j) = \log \prod_{k \in N(j) \setminus i} m_{kj}(x_j)$
- These multipliers satisfy the optimality KKT conditions of Lagrange multipliers, so

$$\text{Lagrange multipliers } \lambda_{ij}(x_j) \text{ of Bethe Free energy}$$

$$\updownarrow$$

$$\text{Messages } m_{ij}(x_j) \text{ of BP algorithm}$$

- This is a first hint of a major tool: characterize inference as solutions of optimization problems: **variational inference.**

Max-Product

- We have described an algorithm to estimate marginal (and conditional) distributions.
- How about inference tasks of the form $\arg\max_x p(x \mid y)$?
  - I.e. Maximum-a-posteriori inference.

Max-Product

- We have described an algorithm to estimate marginal (and conditional) distributions.
- How about inference tasks of the form $\arg\max_x p(x \mid y)$?
  - I.e. Maximum-a-posteriori inference.

- A simple variant is the *max-product algorithm,* used to estimate the state configuration with maximum probability.
- Marginalization:

$$m_{ij}^{(n+1)}(x_j) \leftarrow \sum_{x_i} \left(\tilde{\phi}_i(x_i; y)\psi_{ij}(x_i, x_j) \prod_{k \in N(i) \setminus j} m_{ki}^{(n)}(x_i) \right) .$$

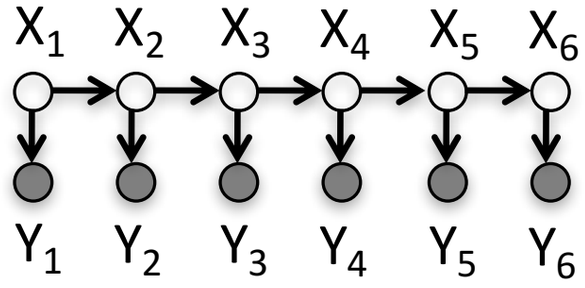
- Maximization:

$$m_{ij}^{(n+1)}(x_j) \leftarrow \max_{x_i} \left(\tilde{\phi}_i(x_i; y)\psi_{ij}(x_i, x_j) \prod_{k \in N(i) \setminus j} m_{ki}^{(n)}(x_i) \right) .$$

Marginal inference in HMMs

- "Filtering" problem is to do marginal inference to find:

$$\Pr(x_n \mid y_1, \ldots, y_n)$$



- How does one **compute** this?

- Applying rule of conditional probability, we have:

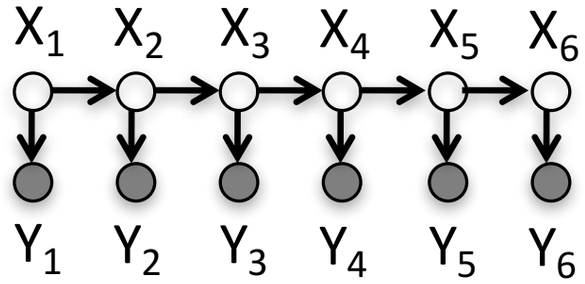
$$\Pr(x_n \mid y_1, \ldots, y_n) = \frac{\Pr(x_n, y_1, \ldots, y_n)}{\Pr(y_1, \ldots, y_n)}$$

- Naively, would seem to require k^{n-1} summations,

$$\Pr(x_n, y_1, \ldots, y_n) = \sum_{x_1, \ldots, x_{n-1}} \Pr(x_1, \ldots, x_n, y_1, \ldots, y_n)$$

Is there a more efficient algorithm?

Marginal inference in HMMs:



- Use **dynamic programming**

$$\Pr(x_n, y_1, \ldots, y_n) = \sum_{x_{n-1}} \Pr(x_{n-1}, x_n, y_1, \ldots, y_n)$$

$$\Pr(A = a) = \sum_b \Pr(B = b, A = a)$$

$$\Pr(\vec{A} = \vec{a}, \vec{B} = \vec{b}) = \Pr(\vec{A} = \vec{a}) \Pr(\vec{B} = \vec{b} \mid \vec{A} = \vec{a})$$

$$= \sum_{x_{n-1}} \Pr(x_{n-1}, y_1, \ldots, y_{n-1}) \Pr(x_n, y_n \mid x_{n-1}, y_1, \ldots, y_{n-1})$$

Conditional independence in HMMs

$$= \sum_{x_{n-1}} \Pr(x_{n-1}, y_1, \ldots, y_{n-1}) \Pr(x_n, y_n \mid x_{n-1})$$

$$\Pr(A = a, B = b) = \Pr(A = a) \Pr(B = b \mid A = a)$$

$$= \sum_{x_{n-1}} \Pr(x_{n-1}, y_1, \ldots, y_{n-1}) \Pr(x_n \mid x_{n-1}) \Pr(y_n \mid x_n, x_{n-1})$$

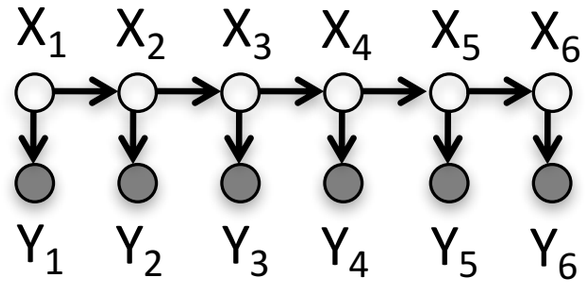
Conditional independence in HMMs

$$= \sum_{x_{n-1}} \Pr(x_{n-1}, y_1, \ldots, y_{n-1}) \Pr(x_n \mid x_{n-1}) \Pr(y_n \mid x_n)$$

- For n=1, initialize $\Pr(x_1, y_1) = \Pr(x_1) \Pr(y_1 \mid x_1)$
- Total running time is $O(nk^2)$ – linear time! Easy to do **filtering**

Marginal Inference in MRF

- This is a simply connected graph:



- Thus we can apply the BP algorithm:

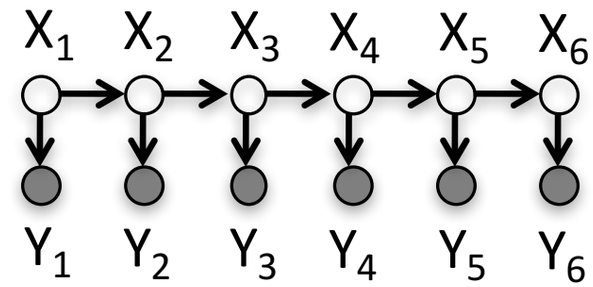
$$\Pr(x_n \ , \ y) = b_n(x_n)$$

$$b_n(x_n) = \frac{1}{Z_n} \Pr(y_n \mid x_n) m_{n-1,n}(x_n) \ .$$

$$m_{n-1,n}(x_n) = \sum_{x_{n-1}} \Pr(y_{n-1} \mid x_{n-1}) \Pr(x_n \mid x_{n-1}) m_{n-2,n-1}(x_{n-1}) \ .$$

$\phi_{n-1}(x_{n-1}, y_{n-1})$ $\psi_{n,n-1}(x_n, x_{n-1})$

MAP inference in HMMs:



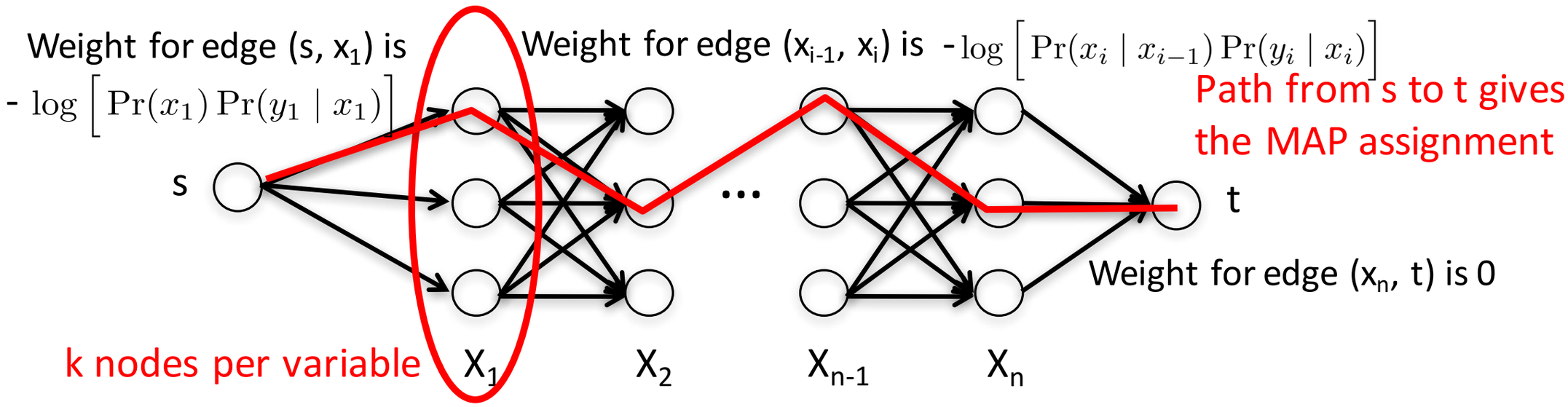
- MAP inference in HMMs can be solved in linear time!

$$\arg\max_{\mathbf{x}} \Pr(x_1, \ldots x_n \mid y_1, \ldots, y_n) = \arg\max_{\mathbf{x}} \Pr(x_1, \ldots x_n, y_1, \ldots, y_n)$$

$$= \arg\max_{\mathbf{x}} \log \Pr(x_1, \ldots x_n, y_1, \ldots, y_n)$$

$$= \arg\max_{\mathbf{x}} \log \Big[\Pr(x_1) \Pr(y_1 \mid x_1) \Big] + \sum_{i=2}^{n} \log \Big[\Pr(x_i \mid x_{i-1}) \Pr(y_i \mid x_i) \Big]$$

- Formulate as a shortest paths problem



Called the Viterbi algorithm

Monte-Carlo Estimation

- BP is an instance of optimization-based inference.
- Let's focus on marginal inference:

$$p(x_i) = \sum_{j \neq i} \sum_{x_j} p(x_1, \ldots, x_n) \ .$$

Monte-Carlo Estimation

- BP is an instance of optimization-based inference.
- Let's focus on marginal inference:

$$p(x_i) = \sum_{j \neq i} \sum_{x_j} p(x_1, \ldots, x_n) \ .$$

- This object can be written as an expectation:

$$p(x_i) = \mathbb{E}_{X \sim p} f_{i,x_i}(X) \ , \ \ f_{i,x_i}(X) = \mathbf{1}(X_i = x_i) \ .$$

Monte-Carlo Estimation

- BP is an instance of optimization-based inference.
- Let's focus on marginal inference:

$$p(x_i) = \sum_{j \neq i} \sum_{x_j} p(x_1, \ldots, x_n) \ .$$

- This object can be written as an expectation:

$$p(x_i) = \mathbb{E}_{X \sim p} f_{i,x_i}(X) \ , \ \ f_{i,x_i}(X) = \mathbf{1}(X_i = x_i) \ .$$

- Thus, another route to approximate inference is by replacing this expectation with *iid* samples:

$$x^1, \ldots, x^M \sim p(X) \text{ iid}$$

$$\hat{p}(x_i) = \frac{1}{M} \sum_{m=1}^{M} f_{i,x_i}(x^m) \ .$$

Monte-Carlo Estimation

- Thus, provided we can (efficiently) sample from the model, we can estimate any quantity that depends smoothly on the density.
- What is the quality of such estimate?

- Bias?

$$\mathbb{E}_{x^1 \ldots x^M \sim p}\ [\hat{p}(x_i)] = \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_{x^m \sim p} f_{i,x_i}(x^m)\ . = \mathbb{E} f_i(x) = p(x_i)$$

- Variance?
  - Law of large numbers: $\hat{p}(x_i) \overset{a.s.}{\to} p(x_i)\ ,\ (m \to \infty)\ .$
  - CLT:Under mild assumptions, $\sqrt{m}(\hat{p}(x_i) - p(x_i)) \overset{d}{\to} \mathcal{N}(0,1)\ .$

Monte-Carlo Estimation

- But, how do we sample from a graphical model?
  - If it is a BN, we saw in the first lecture that it lends itself to sampling by following topological order.
  - But how about undirected graphical models?

Gibbs Sampling

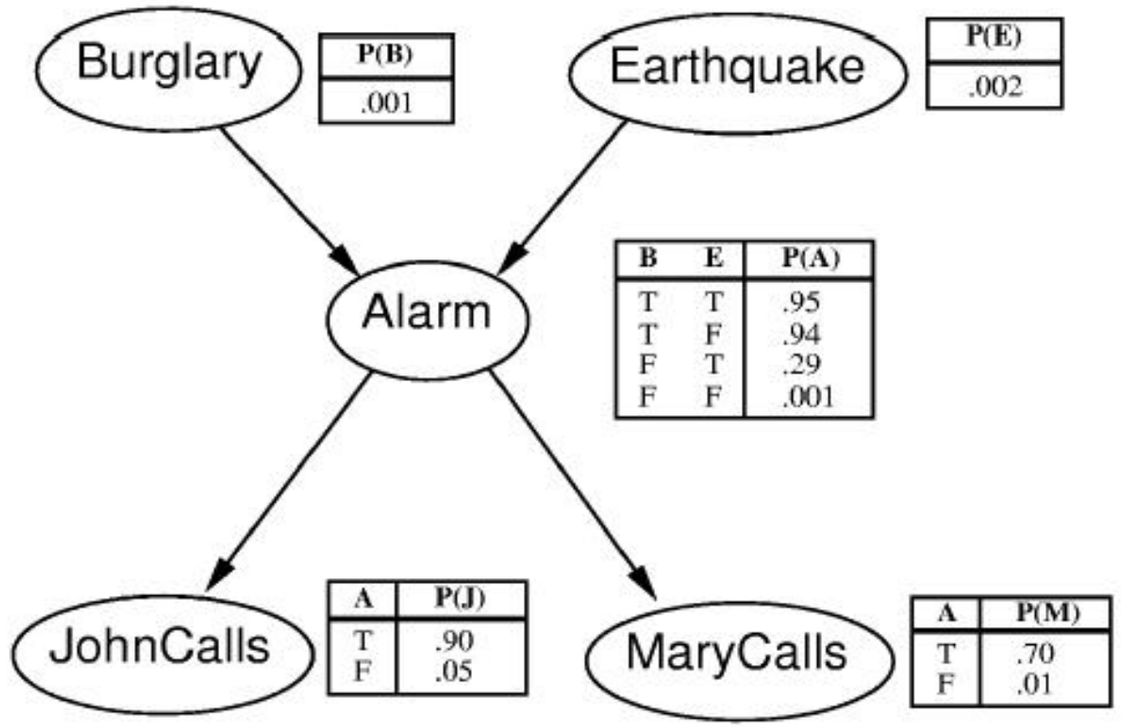
- Gibbs Sampling is an iterative algorithm that produces samples from undirected models.

- Suppose the model contains variables $x_1 \ldots x_n$
- Initialize starting values (e.g from uniform distribution)

- Do until (convergence):
  - Pick an ordering of the variables
  - For each x_i,
    - Sample $p(x_i \mid X_j = x_j)\,,\, j \neq i\ .$
    - update x_i

- Recall that we only need to condition on the Markov Blanket.

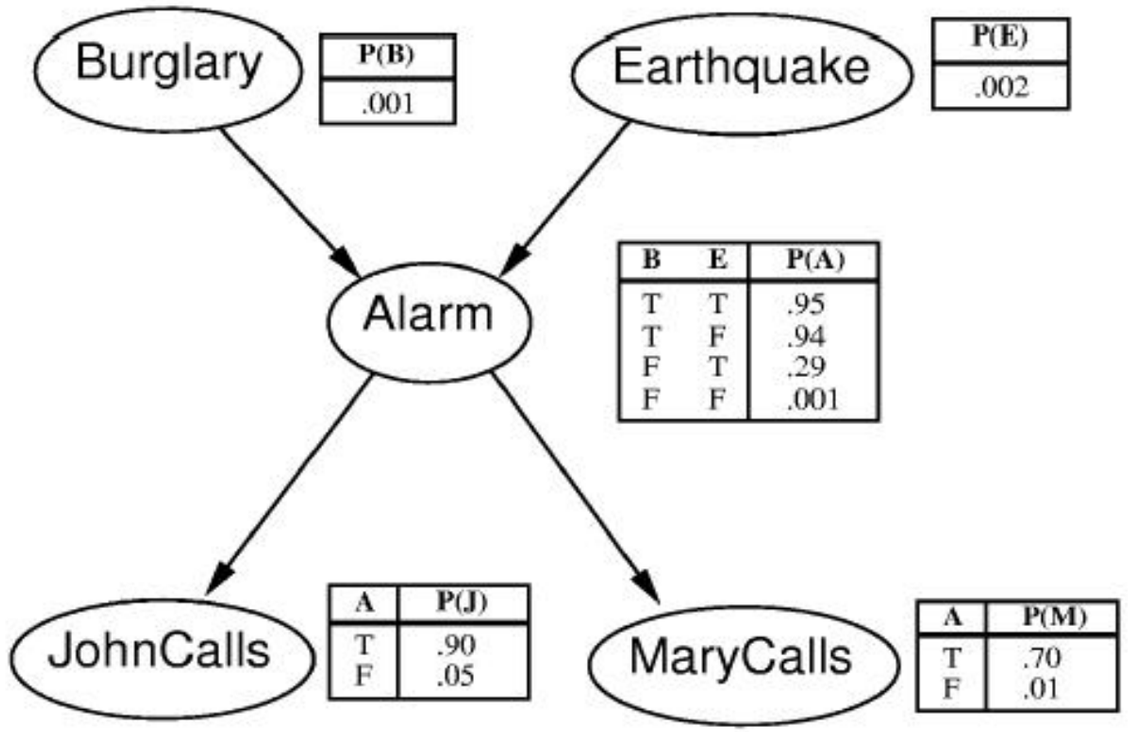
Gibbs Sampling: An Example



t	B	E	A	J	M
0	F	F	F	F	F
1					
2					
3					
4					

- Consider the alarm network
  - Assume we sample variables in the order B,E,A,J,M
  - Initialize all variables at t = 0 to False

Gibbs Sampling: An Example



t	B	E	A	J	M
0	F	F	F	F	F
1	F				
2					
3					
4					

- Sampling P(B|A,E) at t = 1: Using Bayes Rule,

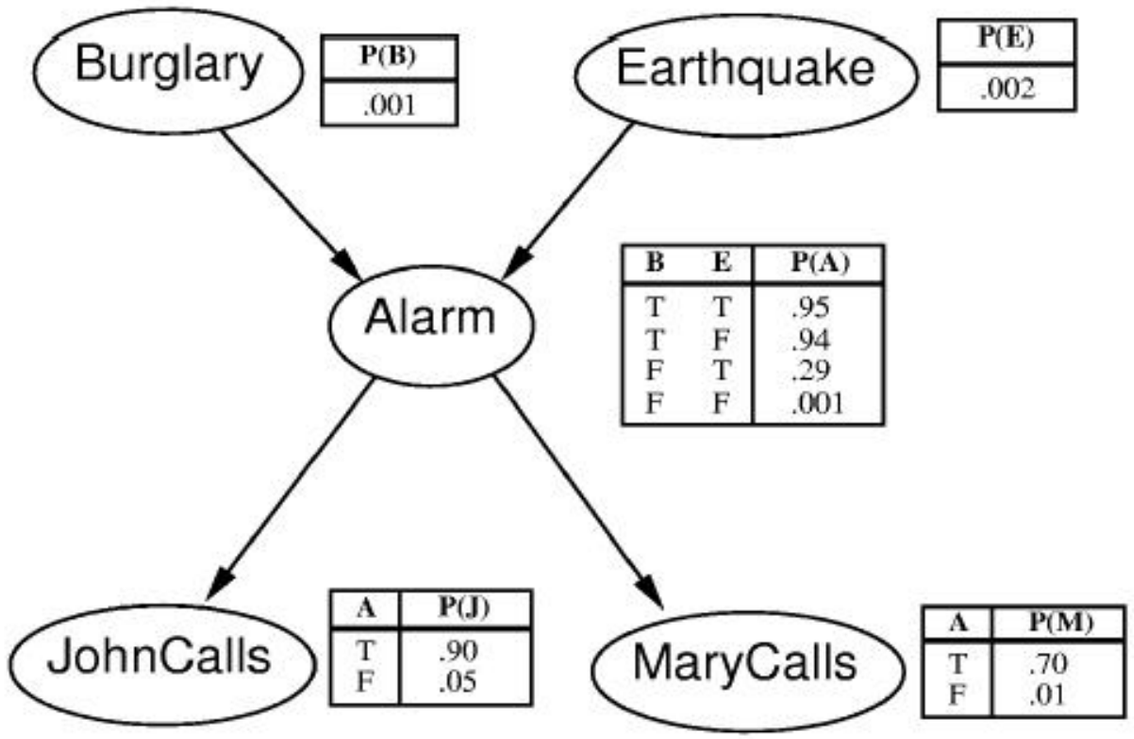
$$P(B \mid A, E) \propto P(A \mid B, E)P(B)$$

- A=false, E=false, so we compute:

$$P(B = T \mid A = F, E = F) \propto (0.06)(0.01) = 0.0006$$

$$P(B = F \mid A = F, E = F) \propto (0.999)(0.999) = 0.9980$$

Gibbs Sampling: An Example



t	B	E	A	J	M
0	F	F	F	F	F
1	F	T			
2					
3					
4					

- Sampling P(E|A,B): Using Bayes Rule,

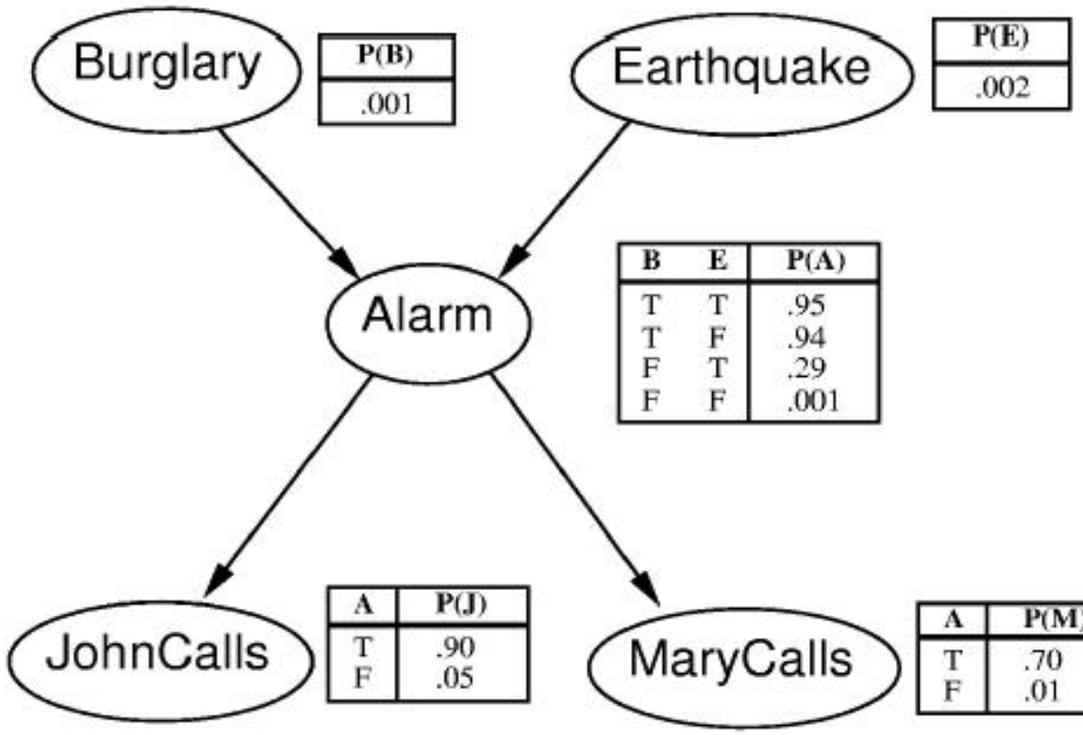
$$P(E \mid A, B) \propto P(A \mid B, E)P(E)$$

- (A,B) = (F,F), so we compute the following,

$$P(E = T \mid A = F, B = F) \propto (0.71)(0.02) = 0.0142$$

$$P(E = F \mid A = F, B = F) \propto (0.999)(0.998) = 0.9970$$

Gibbs Sampling: An Example



t	B	E	A	J	M
0	F	F	F	F	F
1	F	T	F		
2					
3					
4					

- Sampling P(A|B,E,J,M): Using Bayes Rule,

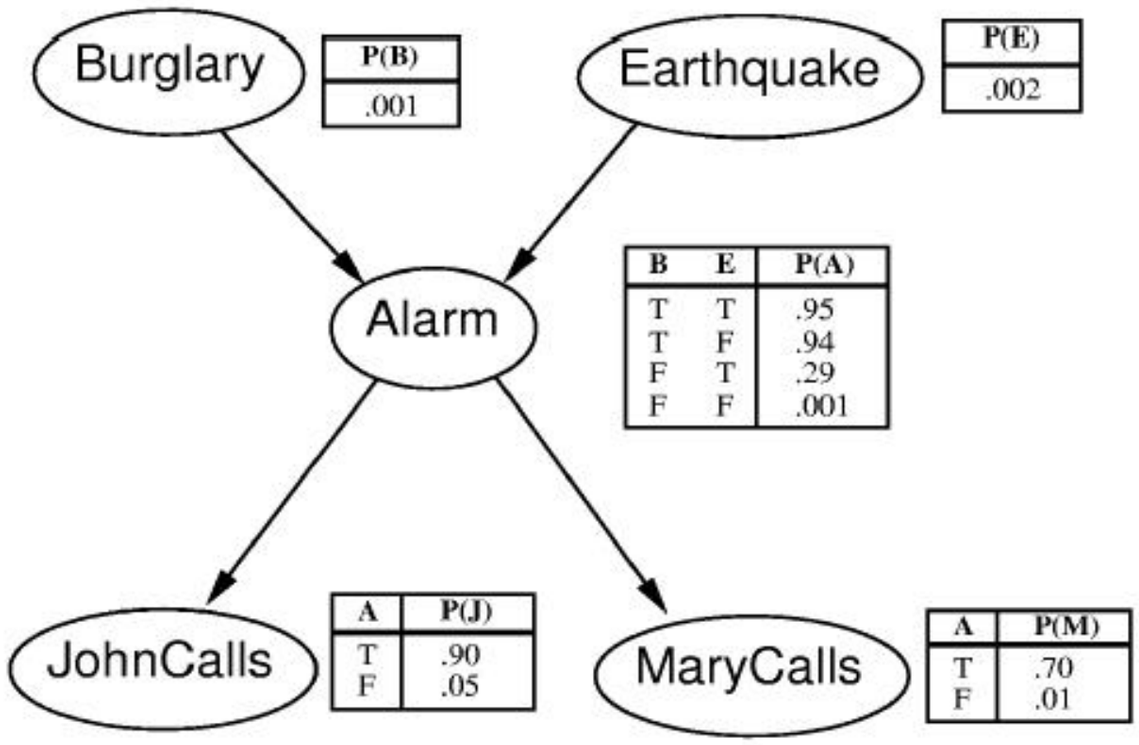
$$P(A \mid B, E, J, M) \propto P(J \mid A)P(M \mid A)P(A \mid B, E)$$

- (B,E,J,M) = (F,T,F,F), so we compute:

$$P(A = T \mid B = F, E = T, J = F, M = F) \propto (0.1)(0.3)(0.29) = 0.0087$$

$$P(A = F \mid B = F, E = T, J = F, M = F) \propto (0.95)(0.99)(0.71) = 0.6678$$

Gibbs Sampling: An Example



t	B	E	A	J	M
0	F	F	F	F	F
1	F	T	F	T	
2					
3					
4					

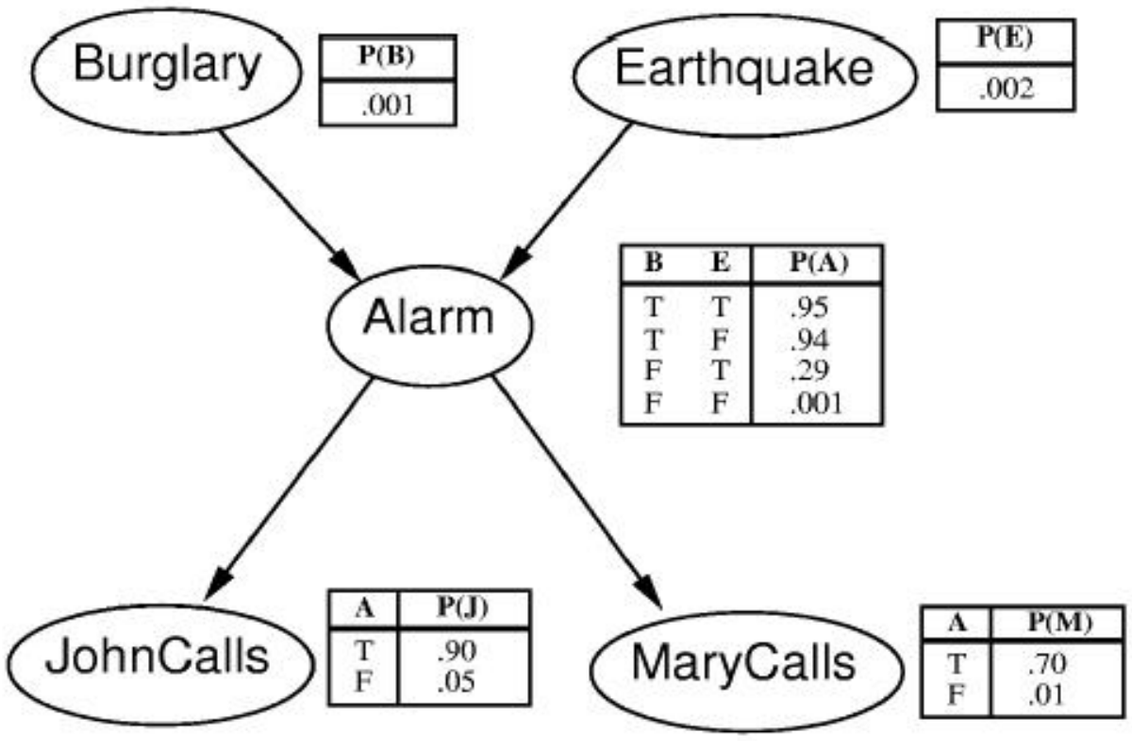
- Sampling P(J|A): No need to apply Bayes Rule

- A = F, so we compute the following, and sample

$$P(J = T \mid A = F) \propto 0.05$$

$$P(J = F \mid A = F) \propto 0.95$$

Gibbs Sampling: An Example



t	B	E	A	J	M
0	F	F	F	F	F
1	F	T	F	T	F
2					
3					
4					

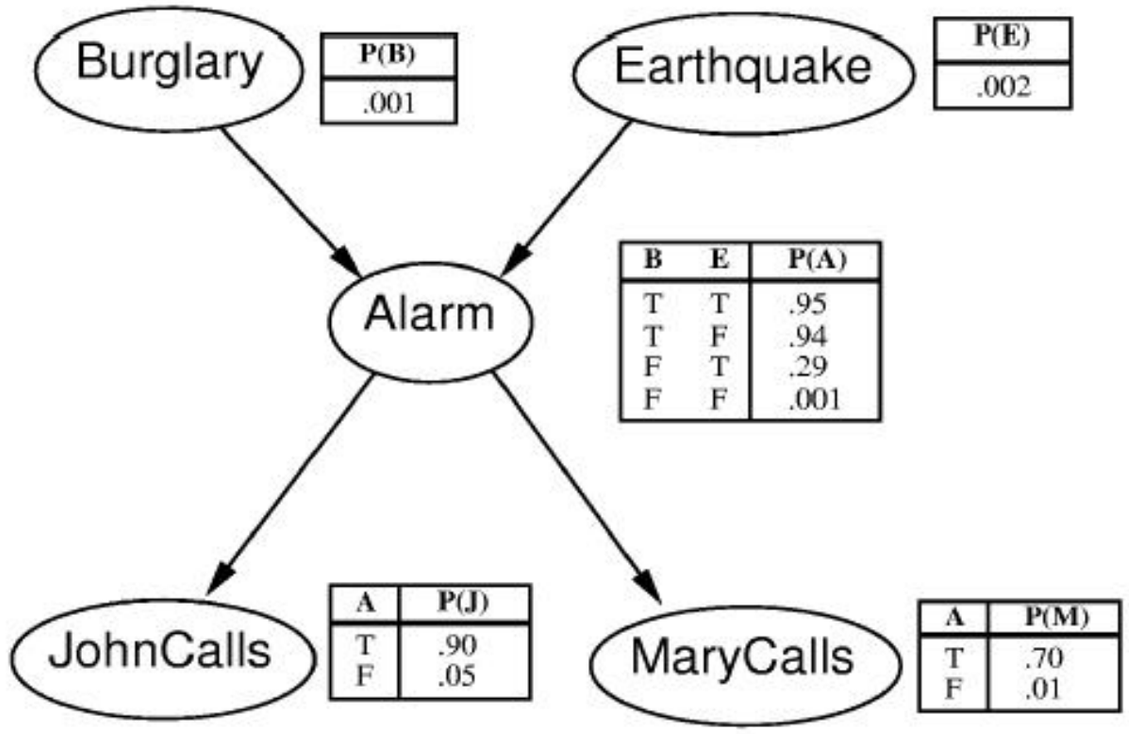
- Sampling P(M|A): No need to apply Bayes Rule

- A = F, so we compute the following, and sample

$$P(M = T \mid A = F) \propto 0.01$$

$$P(M = F \mid A = F) \propto 0.99$$

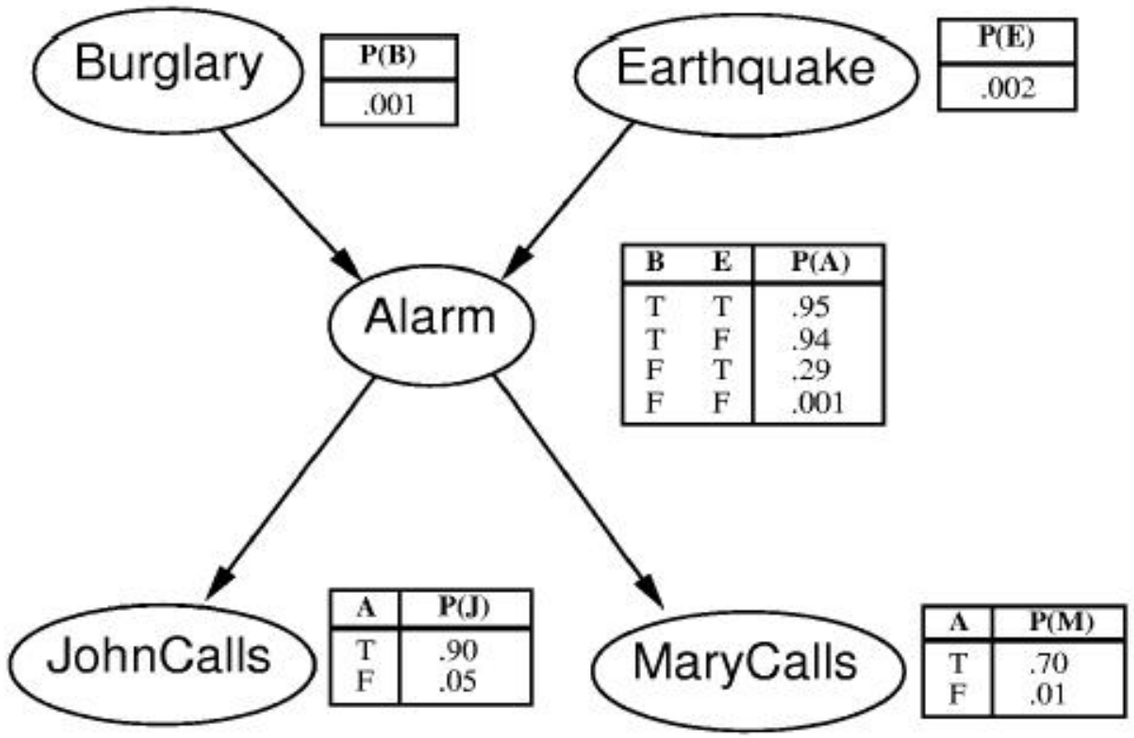
Gibbs Sampling: An Example



t	B	E	A	J	M
0	F	F	F	F	F
1	F	T	F	T	F
2	F	T	T	T	T
3					
4					

- Now t = 2, and we repeat the procedure to sample new values of B,E,A,J,M …

Gibbs Sampling: An Example



t	B	E	A	J	M
0	F	F	F	F	F
1	F	T	F	T	F
2	F	T	T	T	T
3	T	F	T	F	T
4	T	F	T	F	F

- Now t = 2, and we repeat the procedure to sample new values of B,E,A,J,M …

- And similarly for t = 3, 4, etc.

Gibbs Sampling and Markov Chains

- This algorithm is an instance of a broad family of tools: MCMC
- We will study in future lecture the main properties and uses of general MCMC methods.

Survey Data

- How are these statements inferred?

FiveThirtyEight

Politics Sports Science & Health Economics Culture

MAY 3, 2016 AT 2:45 PM

The Mythology Of Trump's 'Working Class' Support

His voters are better off economically compared with most Americans.

By Nate Silver

Filed under 2016 Election

FiveThirtyEight

Politics Sports Science & Health Economics Culture

MAR 14, 2016 AT 6:02 AM

What Trump Supporters Were Doing Before Trump

By Dan Hopkins

Filed under 2016 Election

JUNE 2, 2016

Election 2016

More 'warmth' for Trump among GOP voters concerned by immigrants, diversity

BY BRADLEY JONES AND JOCELYN KILEY | 10 COMMENTS

Survey Data

- We typically see these sort of data

How the GOP candidates' supporters differed on issues

POSITION	YEAR	STANCE OF SUPPORTERS TRUMP	CRUZ	RUBIO
Raise taxes on rich	2007	0.25	0.24	0.27
Pro-gay marriage	2007	0.31	0.20	0.28
Conservative ideology	2007	0.64	0.76	0.67
Pro-choice	2007	0.63	0.39	0.42
Stay in Iraq	2007	0.62	0.81	0.67
Hawk (vs. dove)	2008	0.68	0.67	0.52
No special help for blacks	2012	0.86	0.82	0.80
Obama rating	2012	0.21	0.17	0.20
Anti-Obamacare	2012	0.76	0.82	0.80
Very critical of system	2012	0.70	0.64	0.62
Pro-government spending	2012	0.20	0.15	0.15
Create pathway to citizenship	2012	0.21	0.29	0.37
Anti-Hispanic prejudice	2012	0.53	0.50	0.49
Anti-black prejudice	2012	0.58	0.54	0.55
Pro-NAFTA	2012	0.40	0.50	0.52

Stance on a position ranges from 0-1, with 0 being totally against and 1 being totally in agreement with

SOURCE: HOPKINS/MUTZ

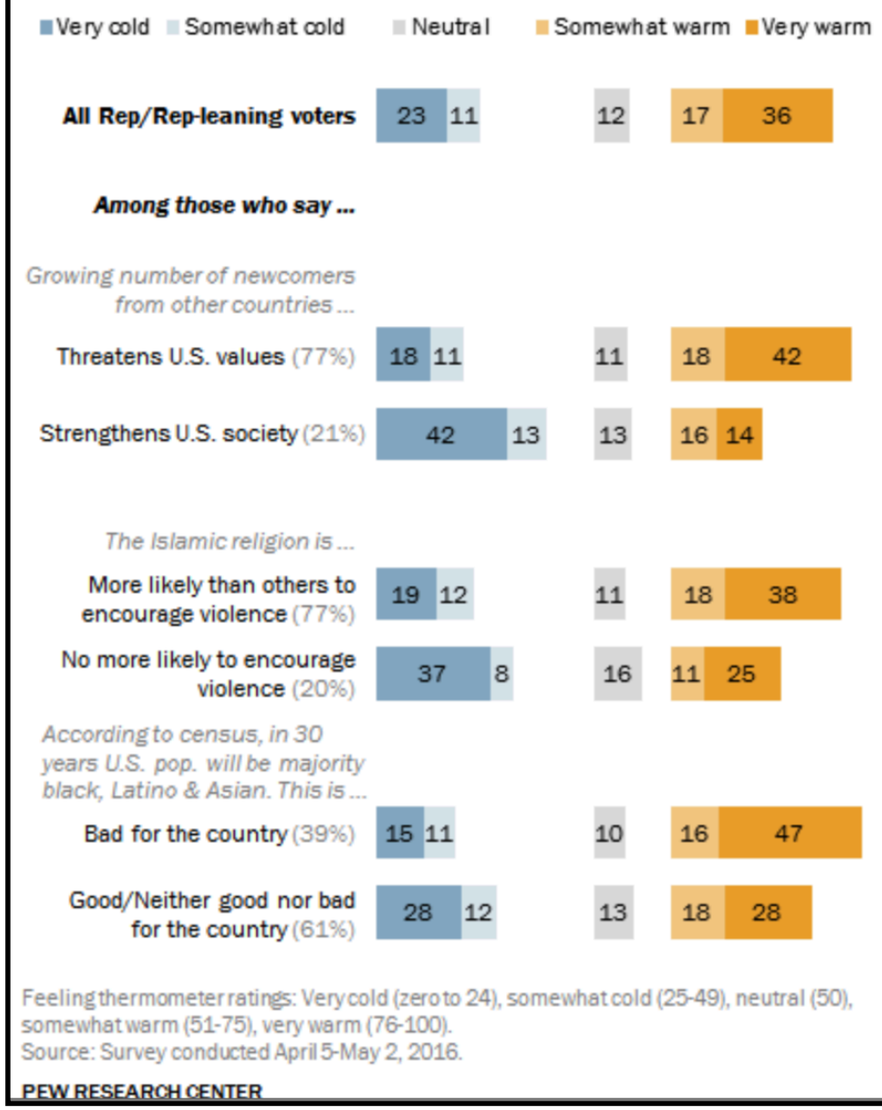
Share of Republican electorate with household income below $50,000

STATE	2012	2016
Alabama	37%	41%
Florida	34	33
Georgia	24	26
Illinois	28	23
Maryland	19	19
Massachusetts	24	20
Michigan	35	37
Mississippi	36	37
New Hampshire	26	27
Ohio	32	30
Oklahoma	41	30
South Carolina	36	27
Tennessee	35	33
Vermont	37	30
Virginia	25	19
Wisconsin	32	28
Average	**31**	**29**

SOURCE: EDISON RESEARCH EXIT POLLS

Within GOP, views of immigration, Islam, diversity strongly associated with ratings of Trump

% of Republican and Republican-leaning registered voters who rate Trump on a 'feeling thermometer' from 0 (coldest rating) to 100 (warmest rating) ...



Feeling thermometer ratings: Very cold (zero to 24), somewhat cold (25-49), neutral (50), somewhat warm (51-75), very warm (76-100).
Source: Survey conducted April 5-May 2, 2016.

PEW RESEARCH CENTER

Survey Data: other prime example

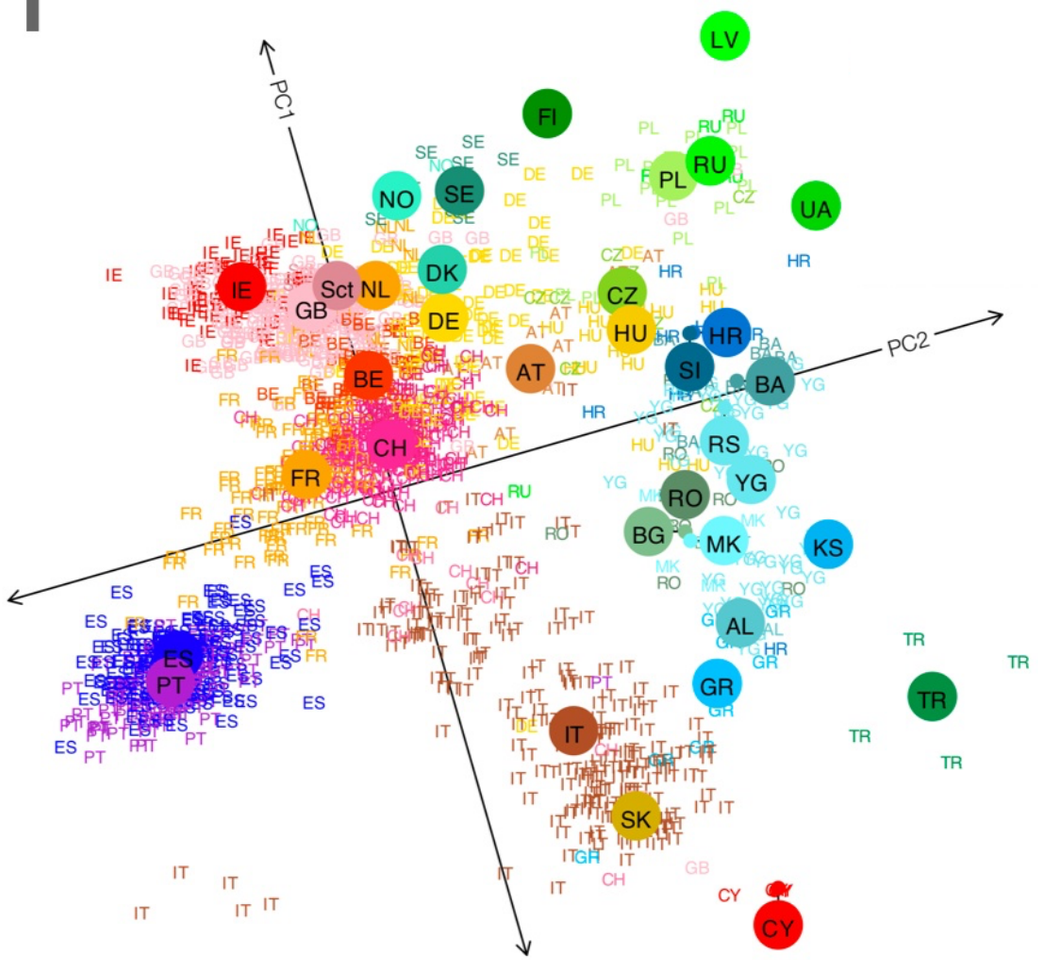
- Medical surveys
- Financial Markets
- EEG recordings

- Children's Depression Inventory [3]
  - 27 items scored 0,1,2 assessing aspects of depressive symptoms for children and adolescents
  - 1 total scale
    - sum of the 27 items after reverse coding 13 of them
    - higher scores indicate higher depressive symptom levels
  - 5 subscales measuring different aspects of depressive symptoms
    - negative mood, interpretation problems, ineffectiveness, anhedonia, and negative self-esteem
    - the total scale equals the sum of the subscales
  - total scale used in practice rather than subscales

(credit: G. Knafl, ohsu)

Gene Expression

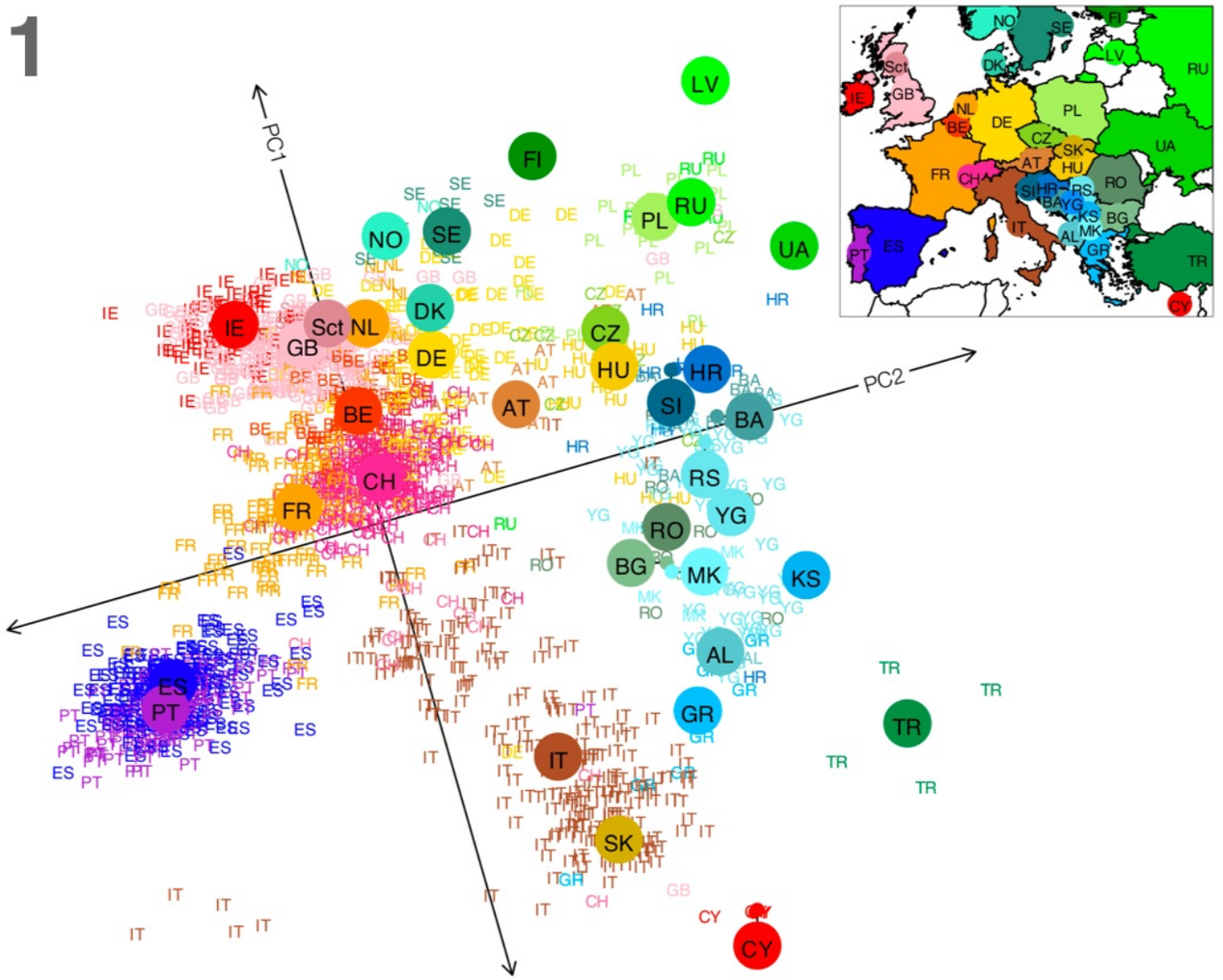
1



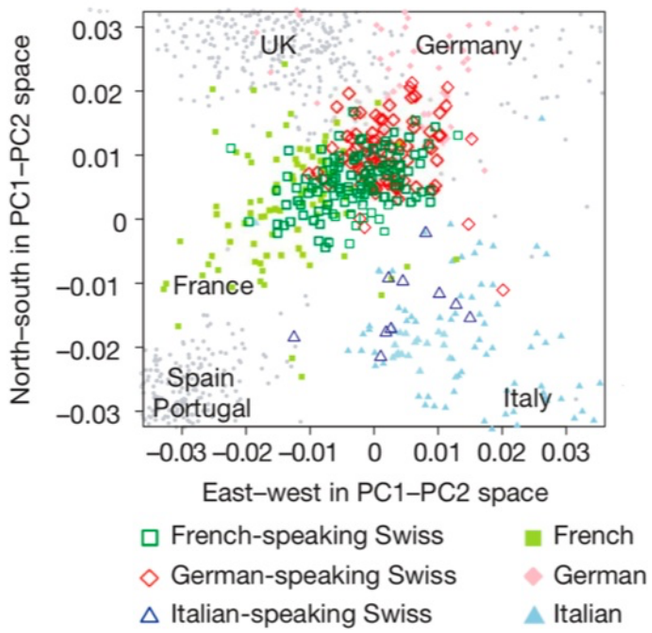
courtesy: John Novembre, UCLA

Gene Expression

1



2



courtesy: John Novembre, UCLA

Factor Analysis for Survey Data

- **Goal**: extract interpretable, summary information out of a series of correlated survey responses.

- Factor Analysis refers to a series of statistical techniques to achieve that.

- That is, given answers $x_1, \ldots x_L$ to L questions, infer latent variables (=factors) that explain the underlying phenomena under study.

Principal Component Analysis

- We start with the simplest setting: suppose that

$$X_l = \sum_{j=1}^{J} \alpha_{j,l} Y_j \ , \ l = 1 \ldots L \ .$$

with $Y_{1\ldots J}$ iid and $X_{1\ldots L}$ jointly Gaussian.

- Q: How to discover the 'latent' factors Y?

Principal Component Analysis

- We start with the simplest setting: suppose that

$$X_l = \sum_{j=1}^{J} \alpha_{j,l} Y_j \ , \ l = 1 \dots L \ .$$

with $Y_{1 \dots J}$ iid and $X_{1 \dots L}$ jointly Gaussian.

- Q: How to discover the 'latent' factors Y?
- Observation 1:

$$\mathbf{X} \text{ Gaussian} \Rightarrow \mathbf{Y} = A\mathbf{X} \text{ also Gaussian.}$$

Principal Component Analysis

- We start with the simplest setting: suppose that

$$X_l = \sum_{j=1}^{J} \alpha_{j,l} Y_j \ , \ l = 1 \dots L \ .$$

with $Y_{1\dots J}$ iid and $X_{1\dots L}$ jointly Gaussian.

- Q: How to discover the 'latent' factors Y?

- Observation 1:

$$\mathbf{X} \text{ Gaussian} \Rightarrow \mathbf{Y} = A\mathbf{X} \text{ also Gaussian.}$$

- Observation 2:

$$\text{If } \mathbf{Y} = (Y_1, \dots, Y_J) \text{ is jointly Gaussian,}$$
$$Y_i, Y_j \text{ independent} \Leftrightarrow Y_i, Y_j \text{ decorrelated.}$$

Principal Component Analysis

- We define $\mu_X = \mathbb{E}(X)\ ,\ \Sigma_X = \mathbb{E}\{(X - \mu_X)(X - \mu_X)^T\}\ .$
- **Reminder:** Let $\mathbf{Y} = A\mathbf{X} + b$. Then
  - $\mu_Y = A\mu_X + b.$
  - $\Sigma_Y = A\Sigma_X A^T.$

Principal Component Analysis

- We define $\mu_X = \mathbb{E}(X)\ ,\ \Sigma_X = \mathbb{E}\{(X-\mu_X)(X-\mu_X)^T\}\ .$
- **Reminder:** Let $\mathbf{Y} = A\mathbf{X} + b$. Then
  - $\mu_Y = A\mu_X + b.$
  - $\Sigma_Y = A\Sigma_X A^T.$

- In our previous model, if $A_{l,j} = \alpha_{l,j}$, we have $X = AY$.
- Hence $\Sigma_X = A\Sigma_Y A^T$
- Q: How to find A such that $A^{-1}\Sigma_X A^{-T}$ defines an uncorrelated random vector?

Principal Component Analysis

Reminder: [Real Spectral Theorem]: If $A \in \mathbb{R}^{n \times n}$ is symmetric and positive semidefinite ($A \succeq 0$), then A diagonalizes
in a real orthonormal basis with non-negative eigenvalues.

Principal Component Analysis

Reminder: [Real Spectral Theorem]: If $A \in \mathbb{R}^{n \times n}$ is symmetric and positive semidefinite ($A \succeq 0$), then A diagonalizes
in a real orthonormal basis with non-negative eigenvalues.

Fact: The covariance operator $\Sigma_X = \mathbb{E}((X - \mu_X)(X - \mu_X)^T)$ is **symmetric** and **positive semidefinite**.

Principal Component Analysis

Reminder: [Real Spectral Theorem]: If $A \in \mathbb{R}^{n \times n}$ is symmetric and positive semidefinite ($A \succeq 0$), then A diagonalizes in a real orthonormal basis with non-negative eigenvalues.

Fact: The covariance operator $\Sigma_X = \mathbb{E}((X - \mu_X)(X - \mu_X)^T)$ is **symmetric** and **positive semidefinite**.

Therefore, Σ_X admits a real eigenbasis:

$$\Sigma_X = U \Lambda U^T \ , \ \Lambda = \mathrm{diag}(\lambda_i) \succeq 0 \ .$$

Thus $U^T \Sigma_X U = \Lambda$.

Principal Component Analysis

Reminder: [Real Spectral Theorem]: If $A \in \mathbb{R}^{n \times n}$ is symmetric and positive semidefinite ($A \succeq 0$), then A diagonalizes in a real orthonormal basis with non-negative eigenvalues.

Fact: The covariance operator $\Sigma_X = \mathbb{E}((X - \mu_X)(X - \mu_X)^T)$ is **symmetric** and **positive semidefinite**.

Therefore, Σ_X admits a real eigenbasis:

$$\Sigma_X = U \Lambda U^T \ , \ \Lambda = \mathrm{diag}(\lambda_i) \succeq 0 \ .$$

Thus $U^T \Sigma_X U = \Lambda$.

Moreover, we can write $\Lambda = S \cdot S$, with $s_{i,i} = \sqrt{\lambda_i}$.
If $\min_i \lambda_i > 0$, it results that $\tilde{U} = U S^{-1}$ satisfies $\tilde{U}^T \Sigma_X \tilde{U} = \mathbf{1}$.

Principal Component Analysis

- We have just shown that $\mathbf{Y} = \tilde{U}(\mathbf{X} - \mu_X)$ provides variables that are uncorrelated and jointly Gaussian, thus independent.

Principal Component Analysis

- We have just shown that $\mathbf{Y} = \tilde{U}(\mathbf{X} - \mu_X)$ provides variables that are uncorrelated and jointly Gaussian, thus independent.

- Remarks
  - The decomposition is not unique: Any orthogonal transformation of Y also satisfies the same property.
  - PCA provides linear compression: if $J < \mathbf{rank}(\Sigma_X)$, what is the best linear approximation of X with J independent components?

$$\min_{A \in \mathbb{R}^{L \times J}} \mathbb{E}(\|X - AX\|^2) \ .$$

Principal Component Analysis

- We have just shown that $\mathbf{Y} = \tilde{U}(\mathbf{X} - \mu_X)$ provides variables that are uncorrelated and jointly Gaussian, thus independent.

- Remarks
  - The decomposition is not unique: Any orthogonal transformation of Y also satisfies the same property.
  - PCA provides linear compression: if $J < \mathbf{rank}(\Sigma_X)$, what is the best linear approximation of X with J independent components?

$$\min_{A \in \mathbb{R}^{L \times J}} \mathbb{E}(\|X - AX\|^2) \ .$$

$$A = \{ \text{ eigenvectors of } \Sigma_X \text{ corresponding to } J \text{ largest eigenvalues.} \}$$

(again, A is determined up to an orthogonal transformation)

Estimating the Principal Components

- So far, we have seen how to extract information from the covariance of X.

Estimating the Principal Components

- So far, we have seen how to extract information from the covariance of X.
- In practice, we will observe $x_1, \ldots, x_N$ iid samples of X
- Empirical Covariance:

$$\hat{\Sigma}_N = \frac{1}{N} \sum_{n<N} (x_n - \hat{\mu})(x_n - \hat{\mu})^T \ . \quad \in \mathbb{R}^{L \times L} \ .$$

Estimating the Principal Components

- So far, we have seen how to extract information from the covariance of X.
- In practice, we will observe $x_1, \ldots, x_N$ iid samples of X
- Empirical Covariance:

$$\widehat{\Sigma}_N = \frac{1}{N} \sum_{n<N} (x_n - \hat{\mu})(x_n - \hat{\mu})^T \ . \quad \in \mathbb{R}^{L \times L} \ .$$

- $\widehat{\Sigma}_N$ is symmetric, positive definite. (why?)

- Estimated Principal Components:

$$\widehat{\Sigma}_N = \widehat{U} \widehat{\Lambda} \widehat{U}^T \ .$$

Estimating the Principal Components

- Q: How good are these estimates?
  - i.e. for a desired accuracy ϵ, how many samples N=N(L) are required?

Estimating the Principal Components

- Q: How good are these estimates?
  - i.e. for a desired accuracy ϵ, how many samples N=N(L) are required?

- **Theorem** [Vershynin]: For distributions with bounded order-q moment, the empirical covariance satisfies

$$\|\widehat{\Sigma}_N - \Sigma\| \lesssim O(\log\log L)^2 \left(\frac{L}{N}\right)^{1/2-2/q} \qquad (q > 4)$$

Estimating the Principal Components

- Q: How good are these estimates?
  - i.e. for a desired accuracy ϵ, how many samples N=N(L) are required?

- **Theorem** [Vershynin]: For distributions with bounded order-q moment, the empirical covariance satisfies

$$\|\widehat{\Sigma}_N - \Sigma\| \lesssim O(\log\log L)^2 \left(\frac{L}{N}\right)^{1/2-2/q} \qquad (q > 4)$$

It results that for a desired approximation $\|\widehat{\Sigma}_N - \Sigma\| \leq \epsilon$ we need $O((\log\log L)^\alpha L) \approx O(L)$ samples.

Estimating the Principal Components

- Q: How good are these estimates?
  - i.e. for a desired accuracy ϵ, how many samples N=N(L) are required?

- **Theorem** [Vershynin]: For distributions with bounded order-q moment, the empirical covariance satisfies $(q > 4)$

$$\|\widehat{\Sigma}_N - \Sigma\| \lesssim O(\log\log L)^2 \left(\frac{L}{N}\right)^{1/2-2/q}.$$

- It results that for a desired approximation $\|\widehat{\Sigma}_N - \Sigma\| \leq \epsilon$ we need $O((\log\log L)^\alpha L) \approx O(L)$ samples. $\left(\frac{1}{\alpha} + \frac{1}{q} = \frac{1}{4}\right)$

- **Very Important Consequence: PCA does not suffer from the curse of dimensionality!**

Estimating Principal Components

- Q: What is the computational complexity of computing PCA?
  - Naïve estimation of covariance: $O(NL^2)$
  - Diagonalizing covariance: $O(L^3)$
  - So, in big data applications we are doomed.

Estimating Principal Components

- Q: What is the computational complexity of computing PCA?
  - Naïve estimation of covariance: $O(NL^2)$
  - Diagonalizing covariance: $O(L^3)$
  - So, in big data applications we are doomed.

- Reminder: Given a data matrix, the principal components are directly given by

The **Singular Value Decomposition** (SVD) of $B \in \mathbb{R}^{n \times p}$ is defined as $B = U \Lambda V^T$, with

$$U \in \mathbb{R}^{n \times n}, V \in \mathbb{R}^{p \times p}, \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_{\min(n,p)}),$$
$$UU^T = \mathbf{1}, V^T V = \mathbf{1} \ .$$

Computing the first p principal components costs $O(pNL)$.

Estimating Principal Components

- Q: What is the computational complexity of computing PCA?
  - Naïve estimation of covariance: $O(NL^2)$
  - Diagonalizing covariance: $O(L^3)$
  - So, in big data applications we are doomed.

- Reminder: Given a data matrix, the principal components are directly given by

The **Singular Value Decomposition** (SVD) of $B \in \mathbb{R}^{n \times p}$ is defined as $B = U \Lambda V^T$, with

$$U \in \mathbb{R}^{n \times n}, V \in \mathbb{R}^{p \times p}, \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_{\min(n,p)}),$$
$$UU^T = \mathbf{1}, V^T V = \mathbf{1} \ .$$

Computing the first p principal components costs $O(pNL)$.

- Alternatives?

Interlude: Randomized PCA

- Suppose we suspect that the data matrix $X \in \mathbb{R}^{N \times L}$ has rank $p \ll \min(N, L)$.
- Q: Can we leverage that prior into a faster algorithm?

Interlude: Randomized PCA

- Suppose we suspect that the data matrix $X \in \mathbb{R}^{N \times L}$ has rank $p \ll \min(N, L)$.
- Q: Can we leverage that prior into a faster algorithm?

- **Idea**: Break the estimation into two steps:
  1. Compute an approximate basis for the range of X:

$$X \approx QQ^T X \ , \ \text{with } Q \in \mathbb{R}^{N \times n} \ , \ n \ll N, \ Q^T Q = \mathbf{1} \ .$$

  2. Form $B = Q^T X \in \mathbb{R}^{n \times L}$ and compute its SVD: $B = \tilde{U} \Lambda V^T$.
  3. Set $U = Q\tilde{U}$.

- How to solve stage 1?

Interlude: Randomized PCA

- Suppose we suspect that the data matrix $X \in \mathbb{R}^{N \times L}$ has rank $p \ll \min(N, L)$.
- Q: Can we leverage that prior into a faster algorithm?

- **Idea**: Break the estimation into two steps:
  1. Compute an approximate basis for the range of X:

$$X \approx QQ^T X \ , \ \text{with } Q \in \mathbb{R}^{N \times n} \ , \ n \ll N, \ Q^T Q = \mathbf{1} \ .$$

  2. Form $B = Q^T X \in \mathbb{R}^{n \times L}$ and compute its SVD: $B = \tilde{U} \Lambda V^T$.
  3. Set $U = Q\tilde{U}$.

- How to solve stage 1? **Randomize!!**

Interlude: Randomized PCA

- If we target a rank-k approximation, we simply draw a random matrix $\Omega \in \mathbb{R}^{L \times (k+p)}$, form the matrix $Y = X\Omega \in \mathbb{R}^{N \times (k+p)}$, and perform a SVD of Y: $Y = V \Lambda Q^T$.

Interlude: Randomized PCA

- If we target a rank-k approximation, we simply draw a random matrix $\Omega \in \mathbb{R}^{L \times (k+p)}$, form the matrix $Y = X\Omega \in \mathbb{R}^{N \times (k+p)}$, and perform a SVD of Y: $Y = V \Lambda Q^T$.

- Strong guarantees from concentration of measure:

Theorem: [Halko, Martinsson, Tropp] Given data matrix $X \in \mathbb{R}^{N \times L}$ and $\Omega \in \mathbb{R}^{L \times (k+p)}$ drawn from iid standard Gaussian, the resulting Q satisfies

$$\|X - QQ^T X\| \leq \left(1 + C\sqrt{(k+p)\min(N,L)}\right) \lambda_{k+1} \ ,$$

whp, where λ_{k+1} is the $k+1$-th singular value of X.

$(1 - cp^{-p})$

Interlude: Randomized PCA

- If we target a rank-k approximation, we simply draw a random matrix $\Omega \in \mathbb{R}^{L\times(k+p)}$, form the matrix $Y = X\Omega \in \mathbb{R}^{N\times(k+p)}$, and perform a SVD of Y: $Y = V\Lambda Q^T$.

- Strong guarantees from concentration of measure:

Theorem: [Halko, Martinsson, Tropp] Given data matrix $X \in \mathbb{R}^{N\times L}$ and $\Omega \in \mathbb{R}^{L\times(k+p)}$ drawn from iid standard Gaussian, the resulting Q satisfies

$$\|X - QQ^T X\| \leq \left(1 + C\sqrt{(k+p)\min(N,L)}\right)\lambda_{k+1} \ ,$$

whp, where λ_{k+1} is the $k+1$-th singular value of X.

$(1 - cp^{-p})$

- Resulting computational gains:

from $O(NLk)$ to $O(NL\log(k))$ for k ppal components.

Factor Analysis

- We have seen that PCA can be interpreted as a linear latent model:

$$X_i = \sum_j \alpha_{i,j} Y_j + \mu_i \ , \ (i = 1, \ldots L) \ ,$$

with Y_j uncorrelated, unit variance.

Factor Analysis

- We have seen that PCA can be interpreted as a linear latent model:

$$X_i = \sum_j \alpha_{i,j} Y_j + \mu_i \ , \ (i = 1, \dots L) \ ,$$

with Y_j uncorrelated, unit variance.

- Lack of unicity: given an $L \times L$ orthogonal matrix R, we also have

$$X = \mu + AR^T RY = \mu + \tilde{A}\tilde{Y} \ ,$$

where $\tilde{Y}_j$ are also uncorrelated and unit variance.

Factor Analysis

- We have seen that PCA can be interpreted as a linear latent model:

$$X_i = \sum_j \alpha_{i,j} Y_j + \mu_i \ , \ (i = 1, \dots L) \ ,$$

with Y_j uncorrelated, unit variance.

- Lack of unicity: given an $L \times L$ orthogonal matrix R, we also have

$$X = \mu + AR^T RY = \mu + \tilde{A}\tilde{Y} \ ,$$

where $\tilde{Y}_j$ are also uncorrelated and unit variance.

- Also, an underlying assumption is that data has *low-rank,* i.e. covariance directly reveals dependencies in data.

Factor Analysis

- An alternative to PCA is *Factor Analysis*. Suppose a generative model of the form

$$X_i = \sum_{j \leq J} \alpha_{i,j} Y_j + \mu_i \ + \epsilon_i \ , \ (i = 1, \ldots L) \ ,$$

with $J < L$ and ϵ_i uncorrelated, zero-mean.

Factor Analysis

- An alternative to PCA is *Factor Analysis*. Suppose a generative model of the form

$$X_i = \sum_{j \leq J} \alpha_{i,j} Y_j + \mu_i \ + \epsilon_i \ , \ (i = 1, \ldots L) \ ,$$

with $J < L$ and ϵ_i uncorrelated, zero-mean.

- Interpretation:
  - Latent variables Y_j are common factors of variability.
  - Latent variables ϵ_i explain the remaining individual variability, uncorrelated from the rest.
- Example:
  - Factor analysis on the topics of your final project.

Factor Analysis

- Gaussian joint likelihood model:

$$X \sim \mathcal{N}(\mu, AA^T + \mathrm{diag}(\beta))$$

  - with $\beta_i = \mathbf{Var}(\epsilon_i)$.
  - Parameter Estimation? The covariance is a sufficient statistic:

$$\Sigma_X = AA^T + \mathrm{diag}(\beta) \ .$$

$$\uparrow$$

$$\text{low rank}$$

  - SVD is still useful, but does not automatically yield the solution.
  - We will soon see an alternative estimation algorithm (EM).

Independent Component Analysis

- We have seen that asking latent variables only to be decorrelated leads to lack of unicity (any orthogonal transformation is also decorrelated).

Independent Component Analysis

- We have seen that asking latent variables only to be decorrelated leads to lack of unicity (any orthogonal transformation is also decorrelated).
- What happens if we tighten our requirements, e.g. asking that

$$Y_i \text{ and } Y_j \text{ independent}$$

Independent Component Analysis

- We have seen that asking latent variables only to be decorrelated leads to lack of unicity (any orthogonal transformation is also decorrelated).
- What happens if we tighten our requirements, e.g. asking that

$$Y_i \text{ and } Y_j \text{ independent}$$

- We have seen that in the Gaussian case, this does not alleviate the problem. (why?)

Independent Component Analysis

- We have seen that asking latent variables only to be decorrelated leads to lack of unicity (any orthogonal transformation is also decorrelated).
- What happens if we tighten our requirements, e.g. asking that

$$Y_i \text{ and } Y_j \text{ independent}$$

- We have seen that in the Gaussian case, this does not alleviate the problem. (why?)

- But, as it turns out, it is the exception. The model becomes uniquely identifiable if

$$Y_i \text{ and } Y_j \text{ independent and non-Gaussian.}$$

Independent Component Analysis

- Without loss of generality, we can assume X has trivial covariance: $\Sigma_X = \mathbf{1}$.
- Since $\Sigma_Y = \mathbf{1}$ as well, it results that $\mathbf{1} = A\Sigma_Y A^T = AA^T$

Independent Component Analysis

- Without loss of generality, we can assume X has trivial covariance: $\Sigma_X = \mathbf{1}$.
- Since $\Sigma_Y = \mathbf{1}$ as well, it results that $\mathbf{1} = A\Sigma_Y A^T = AA^T$

- so we are reduced to finding an orthogonal transformation such that $Y = A^T X$ becomes independent and non-Gaussian.

Independent Component Analysis

- Without loss of generality, we can assume X has trivial covariance: $\Sigma_X = \mathbf{1}$.
- Since $\Sigma_Y = \mathbf{1}$ as well, it results that $\mathbf{1} = A\Sigma_Y A^T = AA^T$

- so we are reduced to finding an orthogonal transformation such that $Y = A^T X$ becomes independent and non-Gaussian.

- It is a form of "inverse" Central Limit Theorem method.

- Q: How to measure/estimate statistical independence?

Entropy

- For a continuous random variable X with density p(x), the *differential entropy* is

$$H(X) = -\int p(x) \log p(x) dx = -\mathbb{E}(\log p(X)) \ .$$

  – it extends the notion of uncertainty/information carried by X.

Entropy

- For a continuous random variable X with density p(x), the *differential entropy* is

$$H(X) = -\int p(x) \log p(x) dx = -\mathbb{E}(\log p(X)) \ .$$

  – it extends the notion of uncertainty/information carried by X.

- Entropy measures independence through the *mutual information*:

Given $X_1, \ldots, X_n$, the mutual information is

$$I(Y) = \sum_{i \leq n} H(Y_i) - H(Y) \ \geq 0 \ .$$

Entropy

- For a continuous random variable X with density p(x), the *differential entropy* is

$$H(X) = -\int p(x) \log p(x) dx = -\mathbb{E}(\log p(X)) \ .$$

  – it extends the notion of uncertainty/information carried by X.

- Entropy measures independence through the *mutual information*:

Given $X_1, \ldots, X_n$, the mutual information is

$$I(Y) = \sum_{i \leq n} H(Y_i) - H(Y) \ \geq 0 \ .$$

- **Fact:** $I(Y) = 0$ iff Y_i and Y_j are mutually independent.
- **Fact:** If A is unitary and $Y = A^T X$, then $H(Y) = H(X)$.

Independent Component Analysis

- So ICA attempts to solve the following problem:

$$\arg \min_{A^T A = \mathbf{1}} \sum_{i \leq L} H(\langle A_i, X \rangle) - H(X) = \arg \min_{A^T A = \mathbf{1}} \sum_{i \leq L} H(\langle A_i, X \rangle) .$$

- Ex from ESLL:

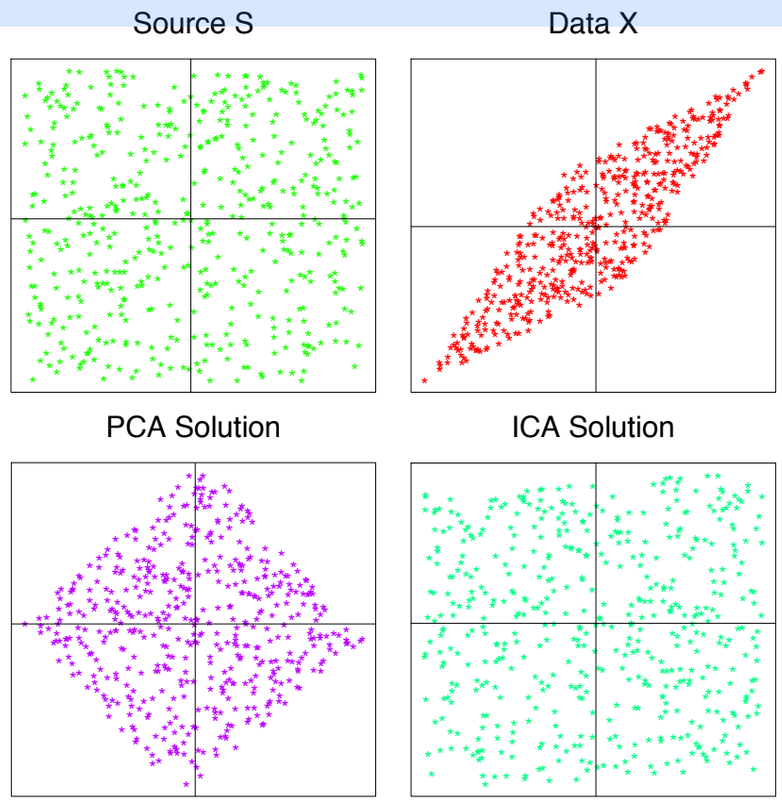


FIGURE 14.38. *Mixtures of independent uniform random variables. The upper left panel shows* 500 *realizations from the two independent uniform sources, the upper right panel their mixed versions. The lower two panels show the PCA and ICA solutions, respectively.*

- Challenge: computing entropy requires estimating the density: exposed to curse of dimensionality!