

Challenges (& Some Solutions) and Making Connections

Real-life Search

- All search algorithm theorems have form:
 - “If the world behaves like ..., then (probability 1) the algorithm will recover the true structure”
- Two types of problems:
 - Probability 1 proofs don't help in the short run
 - World might not fit our assumptions

Search in the Short Run

- Bayes net learning algorithms can give the wrong answer if the data fail to reflect the “true” associations and independencies
 - Of course, this is a problem for all inference: we might just be really unlucky
 - Note: This is not (really) the problem of unrepresentative samples (e.g., black swans)

Convergence in Search

- In search, we would like to bound our possible error as we acquire data
 - I.e., we want search procedures that have uniform convergence
- Without uniform convergence,
 - Cannot set confidence intervals for inference
 - Not every Bayesian, regardless of priors over hypotheses, agrees on probable bounds, no matter how loose

Pointwise Convergence

- Assume hypothesis H is true
- Then
 - For any standard of “closeness” to H , and
 - For any standard of “successful refutation”,
 - Then for every hypothesis that is not “close” to H , there is a sample size for which that hypothesis is refuted

Uniform Convergence

- Assume hypothesis H is true
- Then
 - For any standard of “closeness” to H , and
 - For any standard of “successful refutation”,
 - There is a sample size such that for *all* hypotheses H^* that are not “close” to H , H^* is refuted.

Theorems about Convergence

- There are procedures that, for every model, pointwise converge to the OME class containing the true causal model. (Spirtes, Glymour, & Scheines, 1993)
- There is no procedure that, for every model, uniformly converges to the OME class containing the true causal model. (Robins, Scheines, Spirtes, & Wasserman, 1999; 2003)

Uncooperative World

- Mixed populations
- Variable definition problem
- Structure mis-specification
- Selection bias

Mixed Populations

- A mixture of populations #1 and #2 can have different statistics than either population by itself
 - The different statistics can be *quite* different
 - E.g., association where the sub-populations all have independencies
 - Or positive association when they all have negative associations
 - And so on...

Mixed Populations

- Simple example
 - Population #1: $P(X) = 0.2$; $P(Y) = 0.2$
 - Population #2: $P(X) = 0.8$; $P(Y) = 0.8$
 - And mix the populations 50% / 50%
- In this particular mixture,
 - $P(X) = 0.5$; $P(X | Y) = 0.68$
 - $\Rightarrow X$ and Y are associated in the mixture

Mixed Populations

- More complicated:
 - Consider the following data:

Men			Women		
	Treated	Untreated		Treated	Untreated
Alive	3	20	Alive	16	3
Dead	3	24	Dead	25	6

$$P(A | T) = 0.5$$

$$P(A | U) = 0.45\dots$$

$$0.333$$

$$P(A | T) = 0.39$$

$$P(A | U) =$$

Treatment is superior in both groups!

Mixed Populations

- More complicated:
 - Consider the following data:

Pooled

	Treated	Untreated
Alive	19	23
Dead	28	30

$$P(A | T) = 0.404$$

$$P(A | U) = 0.434$$

Better off *not*
being *Treated!*



Variable Definition

- Illustrative example:
 - Suppose you want to know the impact of *Total Cholesterol* on *Heart Disease*
- Problem:
 - $TC = \text{Low Density Lipid} + \text{High Density Lipid}$
 - $LDL \xrightarrow{+} HD$ & $HDL \xrightarrow{-} HD$
 - \Rightarrow There is no fact of matter about the effect of TC on HD

Variable Definition

- In particular, consider predicting the effect of an intervention on *Total Cholesterol*
 - Cannot make a determinate prediction, since it depends on how the intervention occurs
 - Increase *TC* by raising *LDL*? Or raising *HDL*?
- And search is also difficult in this condition
- Similar situation arises with other variables

Variable Definition

- Idea: Just divide up the world into the finest-grained structure available to you
 - Won't work in all cases, but will for many
- Problem: But lots of fine-grained structure doesn't make a causal difference
 - We don't need to know molecular structure to do lots of causal inference
- So which differences matter?
 - Depends on causal structure \Rightarrow Cycle...

Model Mis-specification

- If the model contains parameters that do not occur in the true structure (or *vice versa*), then the estimated parameters can be deeply misleading about the nature of the world (though formally correct)
 - That is, the “best” estimated model might still be quite different from the truth

Model Mis-specification

- Example:

True model: $X \leftarrow Y \rightarrow Z \leftarrow W$

Estimated model: $X \rightarrow W \rightarrow Y \leftarrow Z$

– Some of the estimation errors:

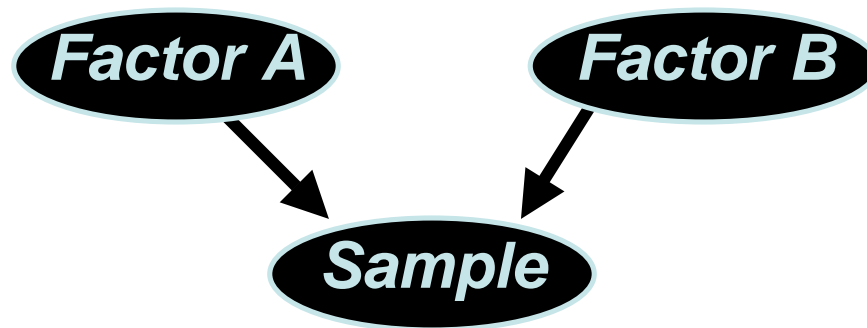
- $P(W | X)$ parameters will all be (close-to-) identical
- $P(Y | Z, W)$ parameters depend on W , even though W and Y are (in the data) independent
- And so on...

Selection Bias

- Sometimes, a variable of interest is a cause of whether people get in the sample
 - E.g., measure various skills or knowledge in college students
 - Or measuring joblessness by a phone survey during the middle of the day
- Simple problem: You might get a skewed picture of the population

Selection Bias

- If two variables matter, then we have:



- *Sample* = 1 for everyone we measure
- That is equivalent to conditioning on *Sample*
- \Rightarrow Induces an association between *A* and *B*!

Connections

- Bayes nets can help us understand many other techniques
 - Regression
 - Factor Analysis, PCA, ICA
 - Naïve Bayes classification
- And give novel inspiration for other types of problems
 - Markov blanket classifiers

Regression

- Given data D , find the coefficients for:

$$Y = \sum_i a_i X_i + \varepsilon_Y$$

that minimize squared error in prediction

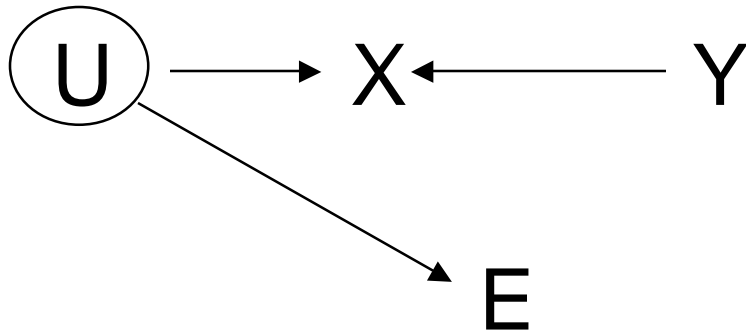
- Relatively simple computation
- There are also nonlinear versions
- Non-zero coefficient iff

$$X_i \not\perp\!\!\!\perp Y \mid \{X_1, \dots, X_n\} \setminus X_i$$

Regression as Search

- Regression is a (classical) parameter estimation method that assumes a particular causal structure
 - But many people use regression for *search* (i.e., non-zero coefficient \Leftrightarrow causal link)
- Regression is a reliable parameter estimator, but unreliable for search

Regression as Search



(and knowledge that X and Y precede E in time)

- Regression \Rightarrow X and Y both cause E
- Constraint-based Bayes net search \Rightarrow Neither X nor Y cause E

Factor Analysis

- Assume linear equations
- Given some set of (observed) features, determine the coefficients for (a fixed number of) unobserved variables that minimize the error

Factor Analysis

- If we have one factor, then we find coefficients to minimize error in:

$$F_i = a_i + b_i \times U$$

where U is the unobserved variable (with fixed mean and variance)

- Two factors \Rightarrow Minimize error in:

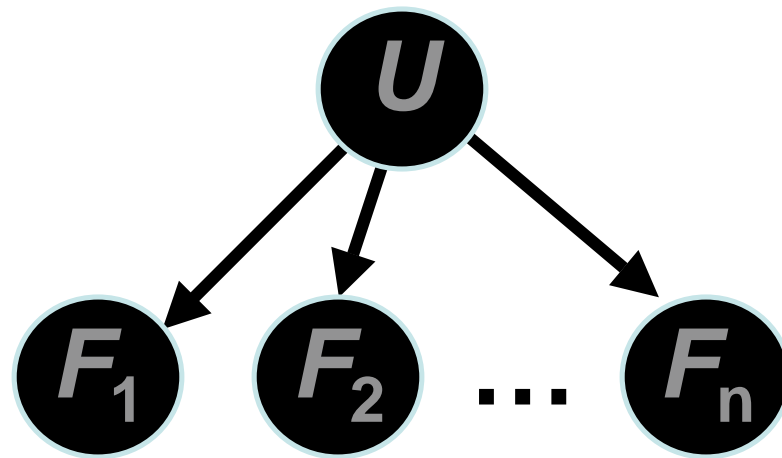
$$F_i = a_i + b_{i,1} \times U_1 + b_{i,2} \times U_2$$

Factor Analysis

- The decision about exactly how many factors to use is typically based on some “simplicity vs. fit” tradeoff
- Also, the interpretation of the unobserved factors must be provided by the scientist
 - The data do not dictate the meaning of the unobserved factors (though it can sometimes be “obvious”)

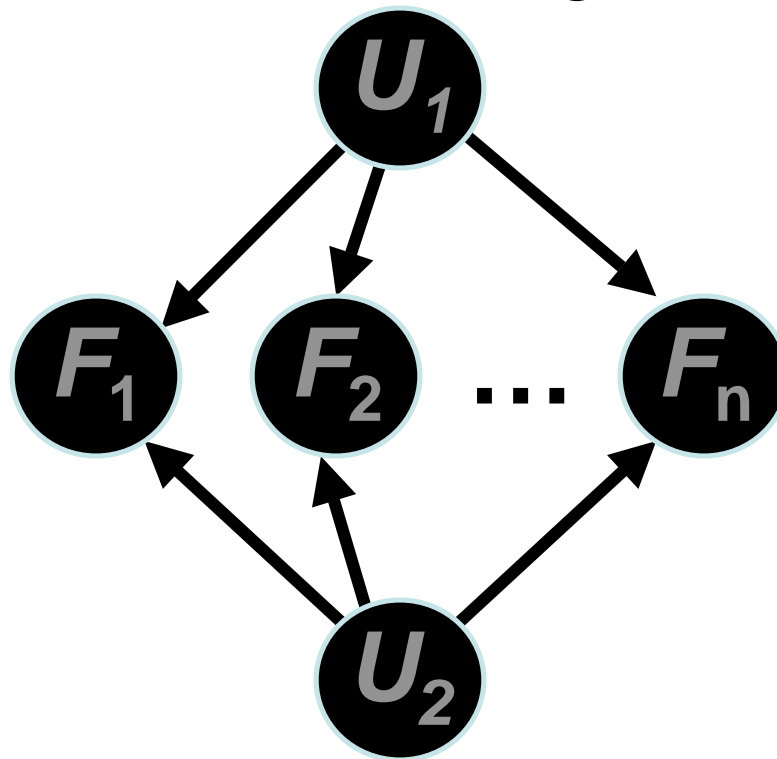
Factor Analysis & Graphs

- One-variable factor analysis is equivalent to finding the ML parameter estimates for the SEM with graph:



Factor Analysis & Graphs

- Two-variable factor analysis is equivalent to finding the ML parameter estimates for the SEM with graph:



PCA and ICA

- Two widely-used analysis techniques:
 - Principal Components Analysis (PCA)
 - Independent Components Analysis (ICA)
- Both are generalizations of factor analysis
- Both have natural reconstructions / explanations in terms of Bayes nets

Naïve Bayes Classifiers

- General classification problem:
Suppose we have classes \mathbf{C} defined on a set of features F_1, \dots, F_n ,
Classify a novel instance by computing $P(C_i | F_1, \dots, F_n)$
- In general, this computation is *hard!*

Naïve Bayes Classifiers

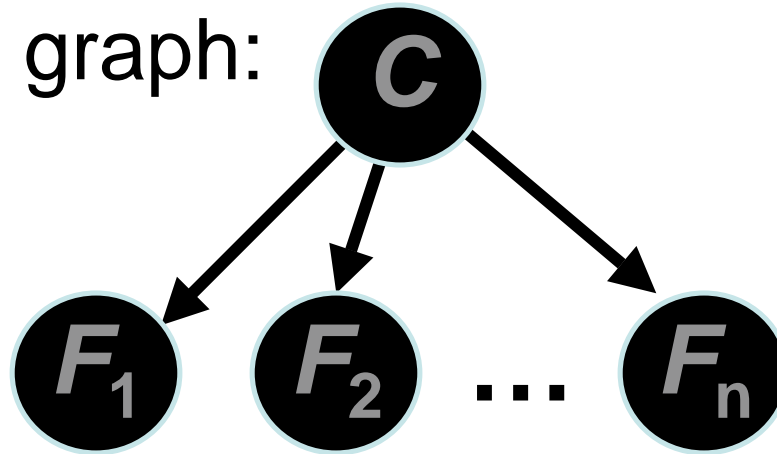
- Naïve Bayes assumption: Each feature is independent of every other feature conditional on the class
 - For all i, j, k , $F_i \perp\!\!\!\perp F_j \mid C_k$
- $\Rightarrow P(C_k \mid F_1, \dots, F_n) \propto$
 $P(C_k) \times P(F_1 \mid C_k) \times \dots \times P(F_n \mid C_k)$

Naïve Bayes Classifiers

- Advantages
 - Fast classification of novel cases
 - Large sample sizes for parameter estimation
 - Relatively small number of parameters
- Disadvantages
 - The fundamental assumption almost never holds in the real world
 - I.e., the class almost never screens off features
 - But sometimes it works “well enough”

Naïve Bayes as a Graph

- Consider the graph:



- Markov + faithfulness \Rightarrow Naïve Bayes assumption

Markov Blanket Classification

- Generally, doing classification requires using only the variables that matter
- The *Markov Blanket* of X is the subset of variables such that X is independent of all of the other variables in the system
 - I.e., the Markov blanket is the set that screens off X from the rest of the system

Markov Blanket Classification

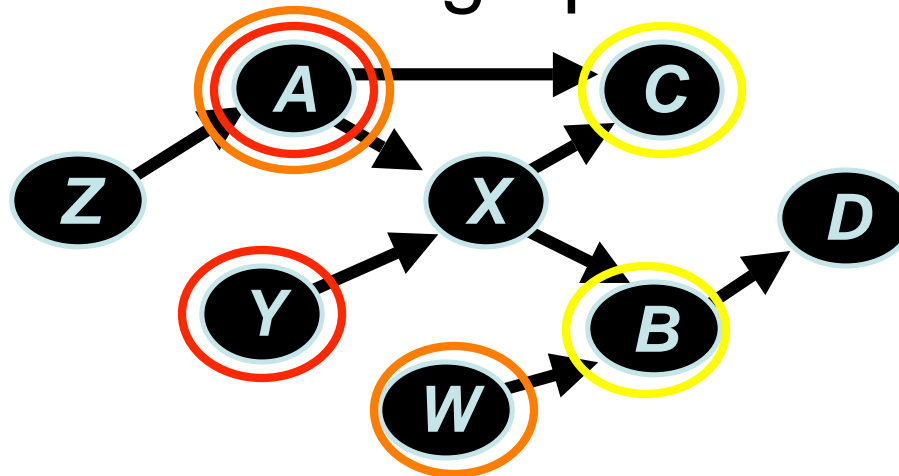
- Classification is computing conditional probabilities, so the Markov blanket suffices for optimal classification
 - Given the Markov blanket, no other variables in the system carry information about X
- Graphical (causal) structure defines the inputs for the optimal classifier

Markov Blanket Classification

- The Markov blanket of X contains:
 - The children of X
 - The parents of X
 - The parents of the children of X
- Justification:
 - First two are obvious
 - Knowing the children of X induces an association between X and the children's parents
 - We're conditioning on a collider, as in $Gas \rightarrow Car \leftarrow Battery$

Markov Blanket Classification

- Consider X in this graph:



- Markov blanket of X is:
 - Parents of X ; Children of X ; and Parents of Children of X

Markov Blanket Classification

- Given some graph, we thus have an efficient method for calculating the inputs to an optimal classifier for all X
 - And in lots of cases, the quantitative side of the classifier is also easy to assemble
- Of course, if you have the wrong variables, or bad measurements, then...