10-418/10-618 Machine Learning for Structured Data

Machine Learning Department
School of Computer Science
Carnegie Mellon University

Causal Inference
+
Bayesian Nonparametrics

Matt Gormley
Lecture 24
Nov. 30, 2022

Reminders

- **Homework 6: VAE + Structured SVM**
  - Out: Wed, Nov 16
  - Due: Wed, Nov 30 at 11:59pm

- **10-618 Mini-Project**
  - Team Formation Due: Tue, Nov 29
  - Proposal Due: Thu, Dec 1
  - Summary & Code Due: Fri, Dec 9

CAUSAL INFERENCE

Causal Hierarchy

Figure 1. The causal hierarchy. Questions at level 1 can be answered only if information from level i or higher is available.

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y \mid x)$	Seeing	What is? How would seeing *X* change my belief inY?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y \mid do(x), z)$	Doing, Intervening	What if? What if I do *X*?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x \mid x', y')$	Imagining, Retrospection	Why? Was it *X* that caused *Y*? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past two years?

Table from Pearl (2018)

Causal Models

Whiteboard:

- Structural Causal Models
  - Example: Linear SCM (structural equation model)
  - Example: Nonparametric SCM
  - Intervention
  - Graphical model induced by SCM
- Post-Intervention Distribution vs. Conditional Distribution
- Treatment Efficacy
  - average difference
  - experimental risk ratio

Identification

Identification:

- whether the causal effects are **identifiable**
- **the central question** in analysis of causal effects

Can the post-intervention distribution $p(y \mid do(x_0))$ be estimated by data sampled from the pre-intervention distribution $p(x, y, z)$?

Yes! (Sometimes.)

Case 1: when the model M is acyclic with all error terms (U_X, U_Y, U_Z) jointly independent, <u>all causal effects are identifiable</u>.

Case 2: when we can <u>marginalize out the causal effects</u>

Causal Markov Theorem

Theorem 1 (The Causal Markov Condition). *Any distribution generated by a Markovian model M can be factorized as:*

$$P(v_1, v_2, \ldots, v_n) = \prod_i P(v_i|pa_i) \tag{15}$$

where $V_1, V_2, \ldots, V_n$ are the endogenous variables in M, and pa_i are (values of) the endogenous "parents" of V_i in the causal diagram associated with M.

Corollary 1 (Truncated factorization). *For any Markovian model, the distribution generated by an intervention $do(X = x_0)$ on a set X of endogenous variables is given by the truncated factorization*

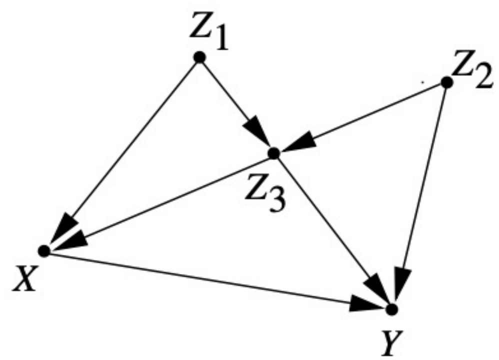
$$P(v_1, v_2, \ldots, v_k|do(x_0)) = \prod_{i|V_i \notin X} P(v_i|pa_i)\,|_{x=x_0} \tag{17}$$

where $P(v_i|pa_i)$ are the pre-intervention conditional probabilities.[8]

Figures from Pearl (2009)

Identification

Example: Model M
(error terms not shown)



1. All of the terms in the post-intervention distribution are from the pre-intervention distribution
2. Those terms could be learned from observational data

Pre-intervention distribution:

$$P(x, z_1, z_2, z_3, y) = P(z_1)P(z_2)P(z_3|z_1, z_2)P(x|z_1, z_3)P(y|z_2, z_3, x)$$

Post-intervention distribution:

$$P(z_1, z_2, z_3, y|do(x_0)) = P(z_1)P(z_2)P(z_3|z_1, z_2)P(y|z_2, z_3, x_0)$$

Causal effect of X on Y:

$$P(y|do(x_0)) = \sum_{z_1, z_2, z_3} P(z_1)P(z_2)P(z_3|z_1, z_2)P(y|z_2, z_3, x_0)$$

Figures from Pearl (2009)

Identification

Identification:

- whether the causal effects are **identifiable**
- **the central question** in analysis of causal effects

Can the post-intervention distribution $p(y \mid do(x_0))$ be estimated by data sampled from the pre-intervention distribution $p(x, y, z)$?

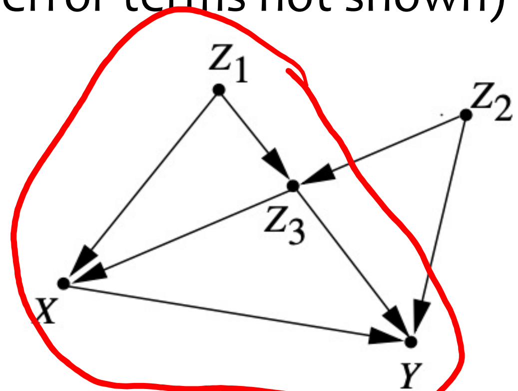
Yes! (Sometimes.)

Case 1: when the model M is acyclic with all error terms (U_X, U_Y, U_Z) jointly independent, <u>all causal effects are identifiable</u>.

Case 2: when we can <u>marginalize out the causal effects</u>

Unmeasured Confounders

Example: Model M
(error terms not shown)



Suppose in our previous identifiability example, we didn't observe z_2 in our data. Can we still estimate p(y | do(x_0))?

Pre-intervention distribution:

$$P(x, z_1, z_2, z_3, y) = P(z_1)P(z_2)P(z_3|z_1, z_2)P(x|z_1, z_3)P(y|z_2, z_3, x)$$

no data to learn these!!

Post-intervention distribution:

$$P(z_1, z_2, z_3, y|do(x_0)) = P(z_1)P(z_2)P(z_3|z_1, z_2)P(y|z_2, z_3, x_0)$$

Causal effect of X on Y:

$$P(y|do(x_0)) = \sum_{z_1, z_2, z_3} P(z_1)P(z_2)P(z_3|z_1, z_2)P(y|z_2, z_3, x_0)$$

$$\boxed{P(y|do(x_0)) = \sum_{z_1, z_3} P(z_1)P(z_3|z_1)P(y|z_1, z_3, x_0)}$$

Yes! Just marginalize over z_2

Figures from Pearl (2009)

Unmeasured Confounders

- Suppose we wish to measure causal effect of X on Y
- But some *confounding* variables are **unmeasurable** (e.g. genetic trait) and some are **measureable** (e.g. height)
- How to pick an **admissible set** of confounders which, if measured, would enable inference?

Definition 3 (Admissible sets – the back-door criterion). A set S is admissible (or "sufficient") for adjustment if two conditions hold:

1. No element of S is a descendant of X
2. The elements of S "block" all "back-door" paths from X to Y, namely all paths that end with an arrow pointing to X.

Definition 1 (d-separation). A set S of nodes is said to block a path p if either (i) p contains at least one arrow-emitting node that is in S, or (ii) p contains at least one collision node that is outside S and has no descendant in S. If S blocks *all* paths from X to Y, it is said to "d-separate X and Y," and then, X and Y are independent given S, written $X \perp\!\!\!\perp Y | S$.

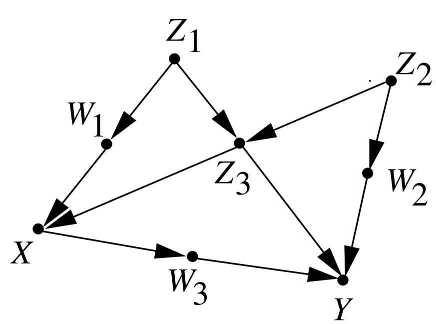
Figures from Pearl (2009)

Unmeasured Confounders

- Suppose we wish to measure causal effect of X on Y
- But some *confounding* variables are **unmeasurable** (e.g. genetic trait) and some are **measureable** (e.g. height)
- How to pick an **admissible set** of confounders which, if measured, would enable inference?

Definition 3 (Admissible sets – the back-door criterion)**.** A set S is admissible (or "sufficient") for adjustment if two conditions hold:

1. No element of S is a descendant of X
2. The elements of S "block" all "back-door" paths from X to Y, namely all paths that end with an arrow pointing to X.



Based on this criterion we see, for example, that the sets $\{Z_1, Z_2, Z_3\}$, $\{Z_1, Z_3\}$, $\{W_1, Z_3\}$, and $\{W_2, Z_3\}$, each is sufficient for adjustment, because each blocks all back-door paths between X and Y. The set $\{Z_3\}$, however, is not sufficient for adjustment because, as explained above, it does not block the path $X \leftarrow W_1 \leftarrow Z_1 \rightarrow Z_3 \leftarrow Z_2 \rightarrow W_2 \rightarrow Y$.

Figures from Pearl (2009)

EXAMPLE: IDENTIFYING CAUSAL EFFECT

Simpson's Paradox

	Treatment A	Treatment B
Small Stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large Stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

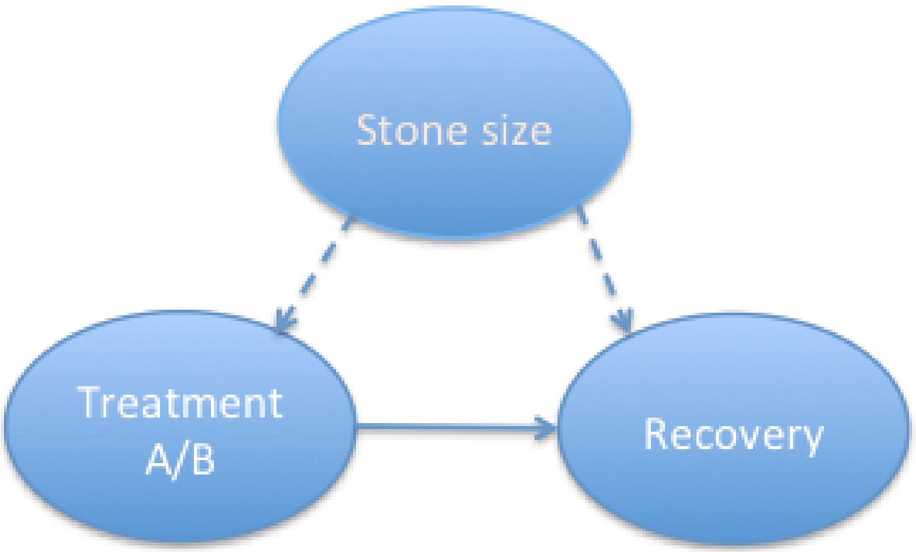


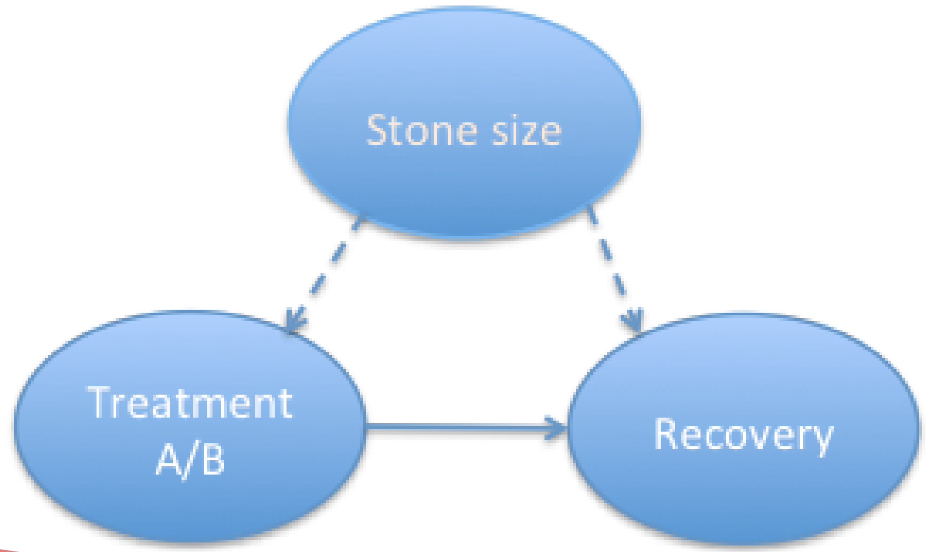
Figure from Kun Zhang's Spring 2019 10-708 Guest Lectures

For people with Small Stones, 93% of those who received Treatment A recovered; but only 87% of those who received Treatment B recovered.

pson's Paradox

So Treatment A is better than Treatment B right?

	Treatment A	Treatment B
Small Stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large Stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**



For people with Large Stones, 73% of those who received Treatment A recovered; but only 69% of those who received Treatment B recovered.

Not quite! Because if you look at both groups, 83% of those who received Treatment B recovered vs only 78% of those with Treatment A.

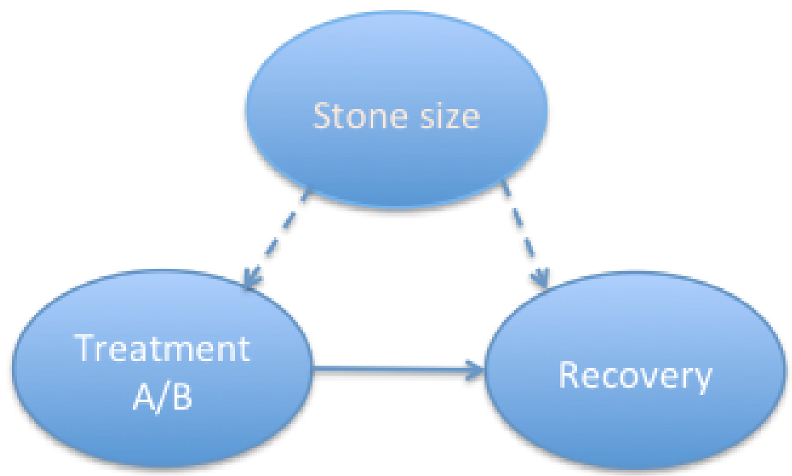
The problem is HOW the data was collected: i.e. the doctor's looked at stone size when selecting Treatment A or B

Figure from Kun Zhang's Spring 2019 10-708 Guest Lectures

Identification of Causal Effects

$P(X3 \mid \mathrm{do}\ (X2=1))$

- "Golden standard": randomized controlled experiments
  - **All the other factors** that influence the outcome variable are either fixed or vary at random, so any changes in the outcome variable must be due to the controlled variable



  - Usually expensive or impossible to do!

Identification of Causal Effects

Whiteboard:

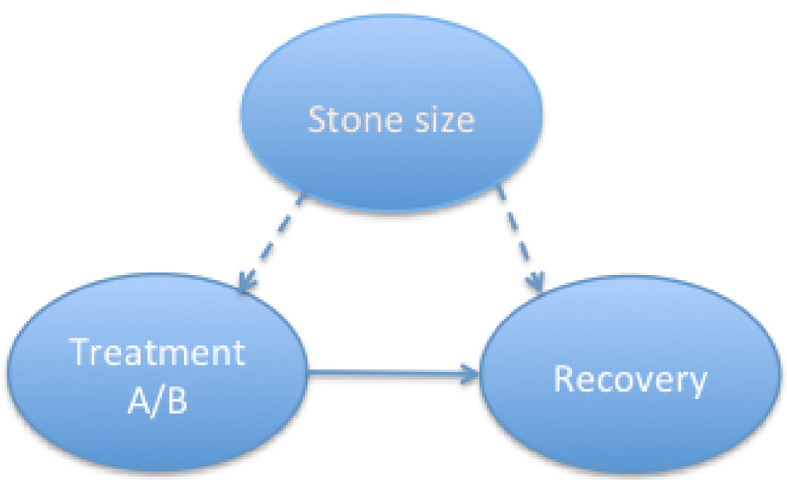
- Stone-size example:
  - Model 1: path diagram for randomized control trial
  - Model 2: path diagram for observational data
  - Model 3: path diagram for intervention

Identification of Causal Effects: Example

	Treatment A	Treatment B
Small Stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large Stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

$$P(R|T) = \sum_S P(R|T,S)P(S|T)$$

$$P(R\,|\,do(T)) = \sum_S P(R\,|\,T,S)P(S)$$

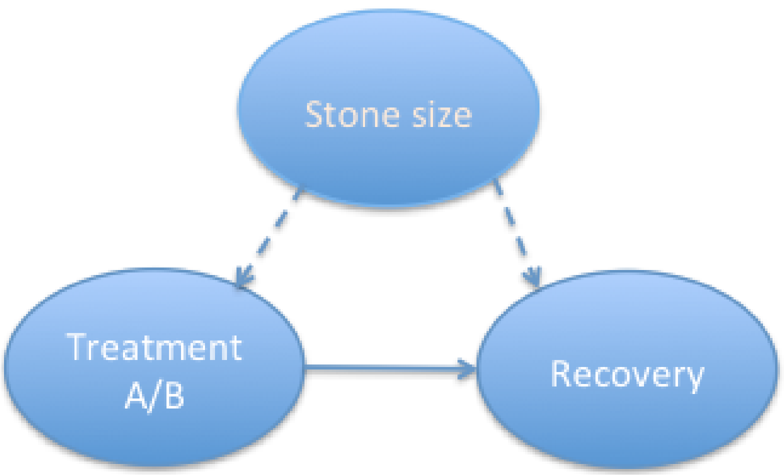


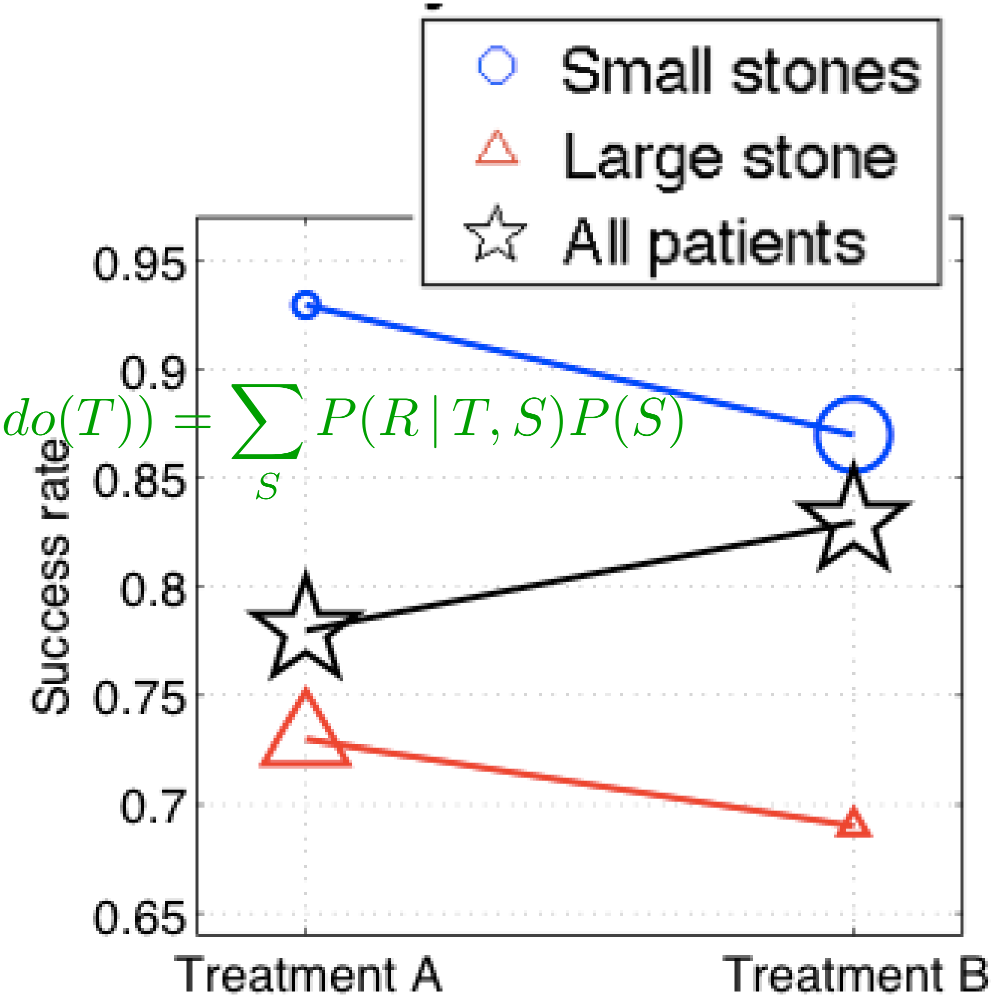
conditioning vs. manipulating

Identification of Causal Effects: Example

	Treatment A	Treatment B
Small Stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large Stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

$$P(R \mid do(T)) = \sum_S P(R \mid T, S) P(S)$$

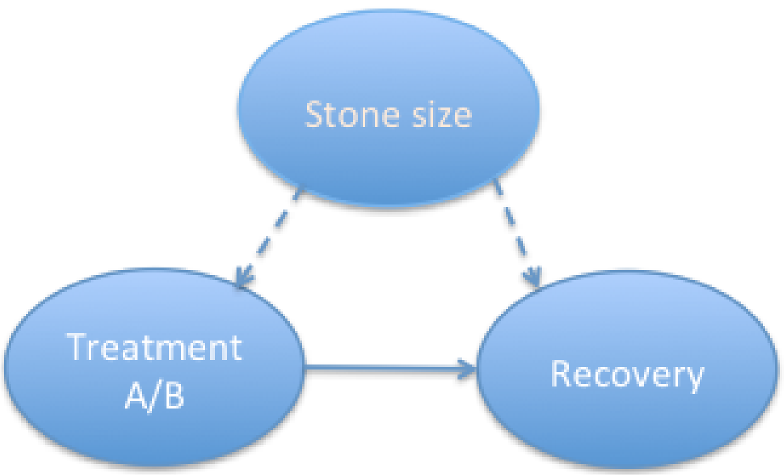


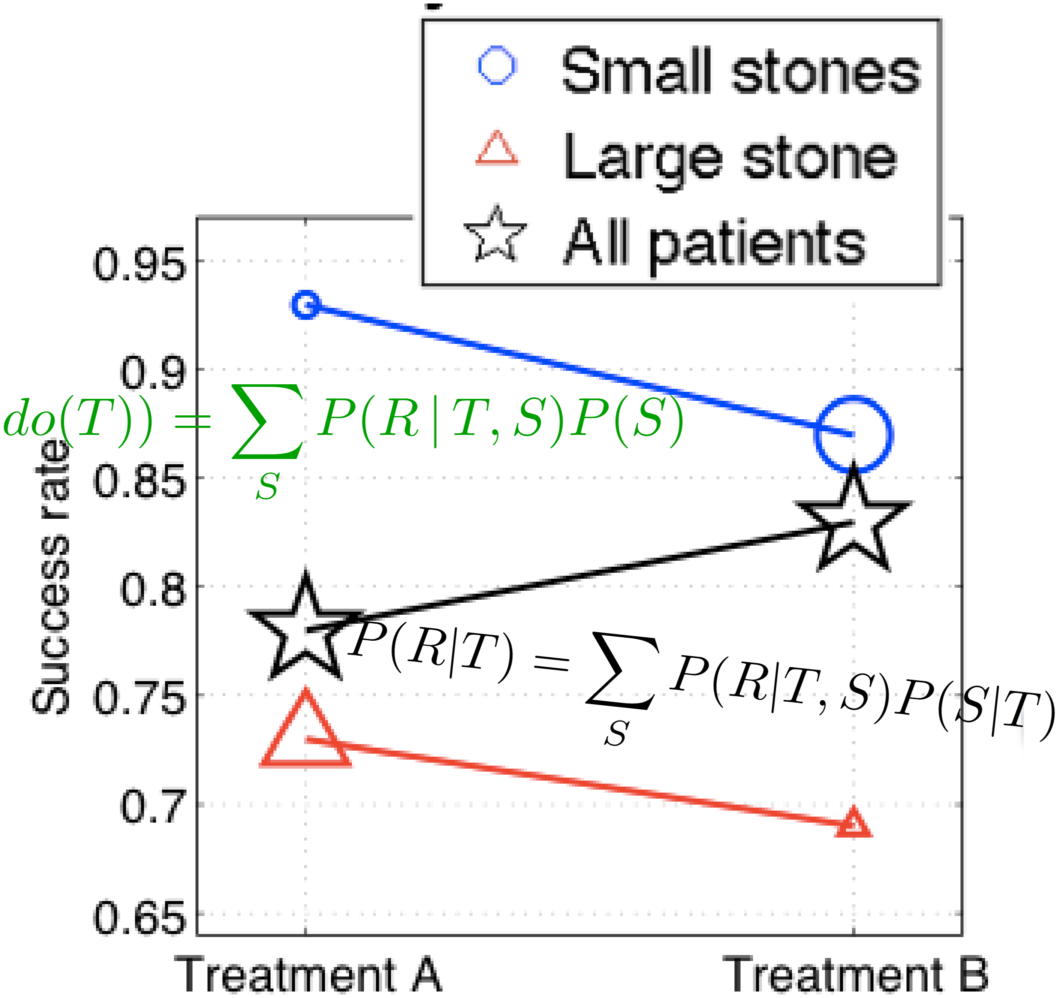


conditioning vs. manipulating

Identification of Causal Effects: Example

	Treatment A	Treatment B
Small Stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large Stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**





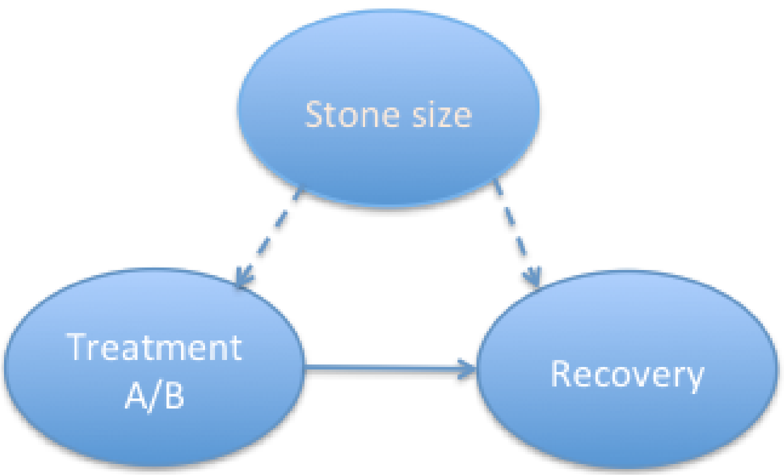
$$P(R \mid do(T)) = \sum_S P(R \mid T, S) P(S)$$

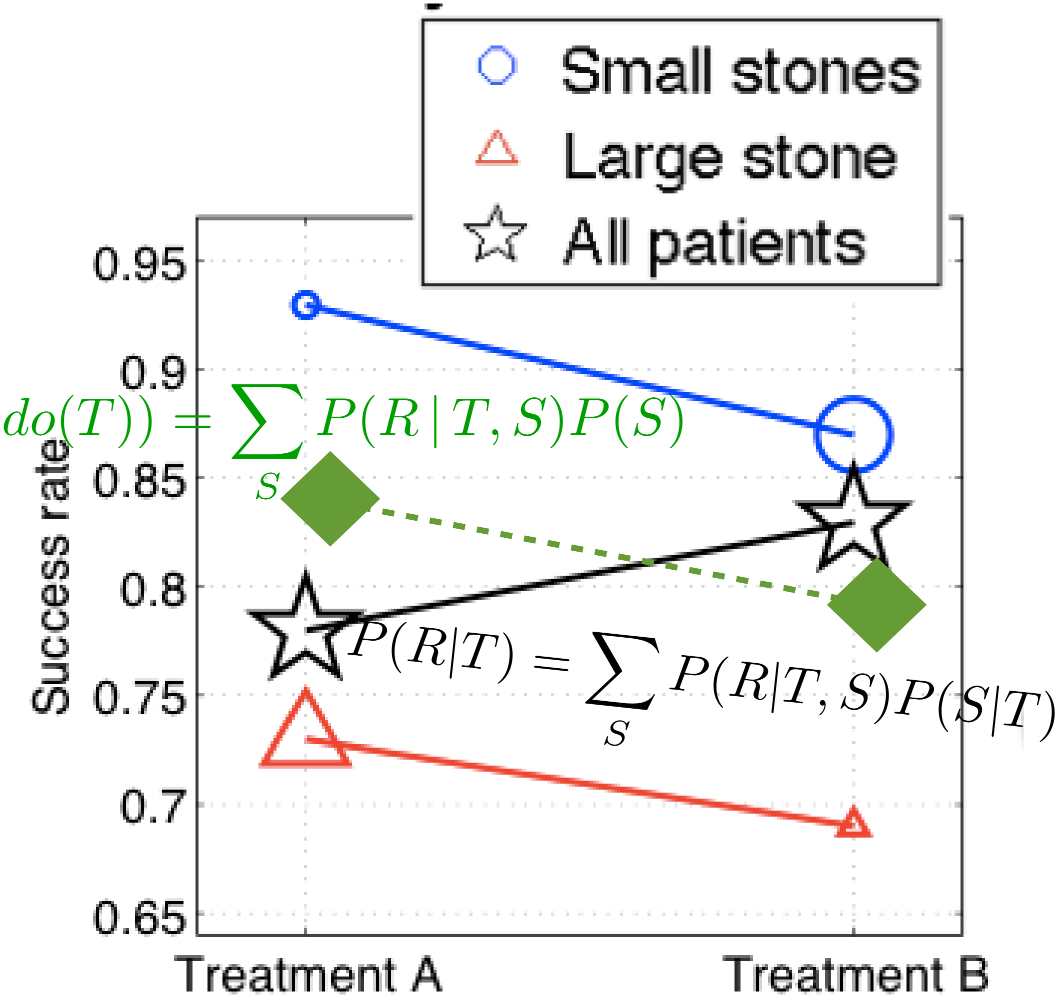
$$P(R|T) = \sum_S P(R|T,S)P(S|T)$$

conditioning vs. manipulating

Identification of Causal Effects: Example

	Treatment A	Treatment B
Small Stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large Stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**





$$P(R \mid do(T)) = \sum_S P(R \mid T, S) P(S)$$

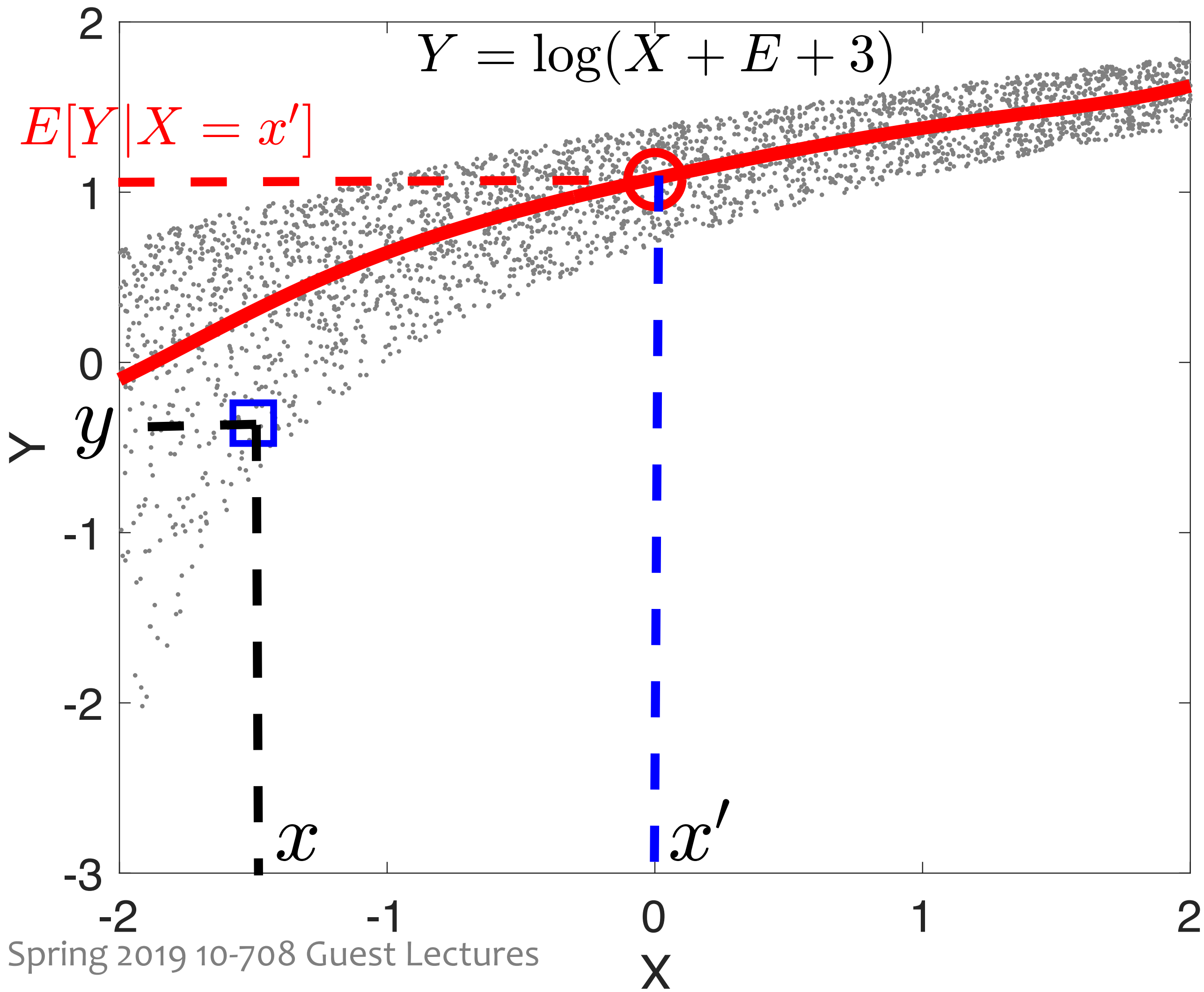
$$P(R|T) = \sum_S P(R|T,S)P(S|T)$$

conditioning vs. manipulating

COUNTERFACTUAL INFERENCE

Counterfactual Inference vs. Prediction

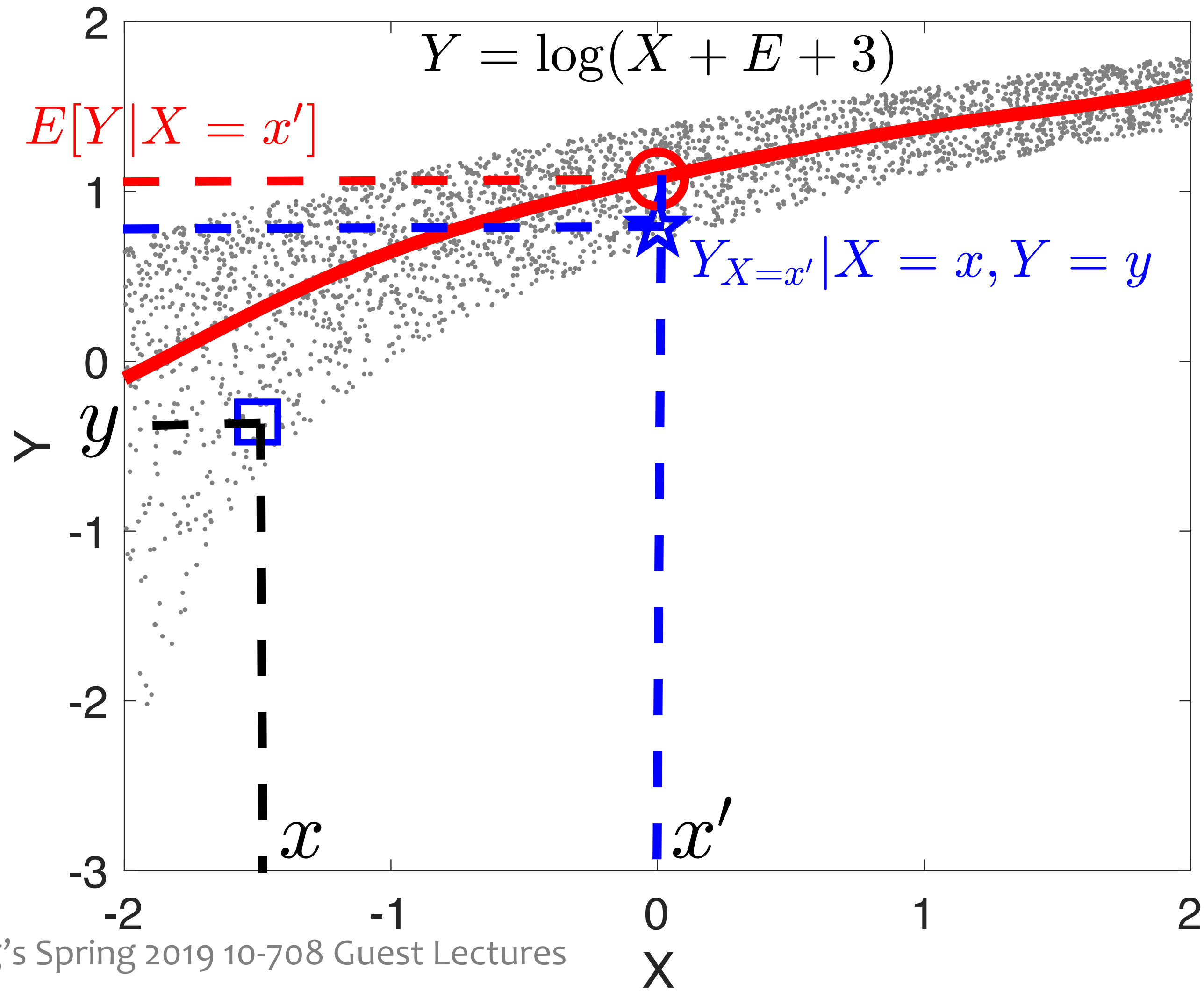
- Suppose X→Y with Y = log(X + E + 3). For an individual with (x,y), what would Y be if X had been x'?



Slide from Kun Zhang's Spring 2019 10-708 Guest Lectures

Counterfactual Inference vs. Prediction

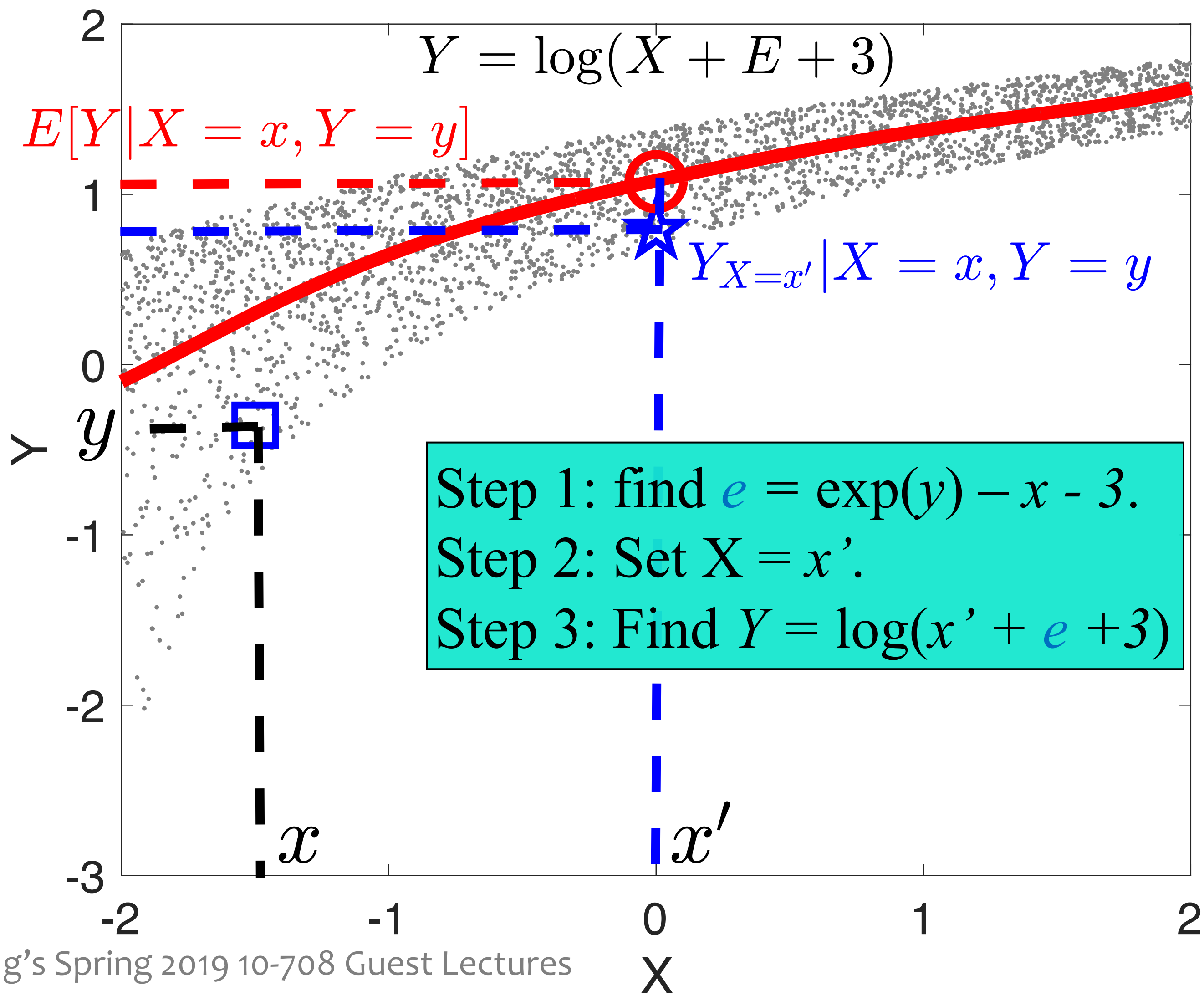
- Suppose X→Y with Y = log(X + E + 3). For an individual with (x,y), what would Y be if X had been x' ?



Slide from Kun Zhang's Spring 2019 10-708 Guest Lectures

Counterfactual Inference vs. Prediction

- Suppose X→Y with Y = log(X + E + 3). For an individual with (x,y), what would Y be if X had been x' ?



Slide from Kun Zhang's Spring 2019 10-708 Guest Lectures

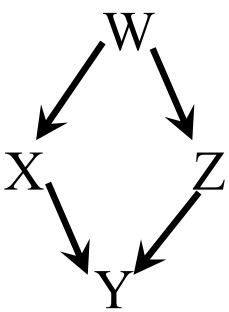
Standard Counterfactual Questions

- We talk about a particular situation (or unit) $U = u$, in which $X = x$ and $Y = y$
- What value would Y be had X been x' in situation u? I.e., we want to know $Y_{X=x'}(u)$, the value of Y in situation u if we do($X=x'$)
- u is not directly observable, so $P(Y_{X=x'} \mid X = x, Y = y)$ instead

For identification of causal effects, U is randomized. It is fixed for counterfactual inference.

Slide from Kun Zhang's Spring 2019 10-708 Guest Lectures

Counterfactual Inference



$$\boxed{\begin{array}{l} W = U_W \\ X = f_X(W, U_X) \\ Z = f_Z(W, U_Z) \\ Y = f_Y(X, Z, U_Z) \end{array}}$$

$$P(Y_{X=x'} \mid \underbrace{X = x,\ Y = y,\ W = w}_{\textit{evidence}})$$

- Three steps
  - Abduction: find $P(U \mid \textit{evidence})$
  - Action: Replace the equation for X by $X = x'$
  - Prediction: Use the modified model to predict Y

Slide from Kun Zhang's Spring 2019 10-708 Guest Lectures

CAUSAL DISCOVERY

Causal Discovery

- Goal:
  - Find a path diagram (i.e. causal model) that is best supported by the data
- Key Idea:
  - find causal structures that are consistent (in a d-separation sense) with the set of conditional independencies supported by the data
- Where to learn more?
  - Kun Zhang (CMU, Philosophy / ML) guest lectures from Spring 2020 10-708: http://www.cs.cmu.edu/~epxing/Class/10708-20/lectures.html

Causal Structure vs. Statistical Independence
(SGS, et al.)

Causal Markov condition: each variable is ind. of its non-descendants (non-effects) conditional on its parents (direct causes)

causal structure (causal graph)

$Y \rightarrow X \rightarrow Z$

Y -- X -- Z **?**

Statistical independence(s)

$Y \perp\!\!\!\perp Z \mid X$

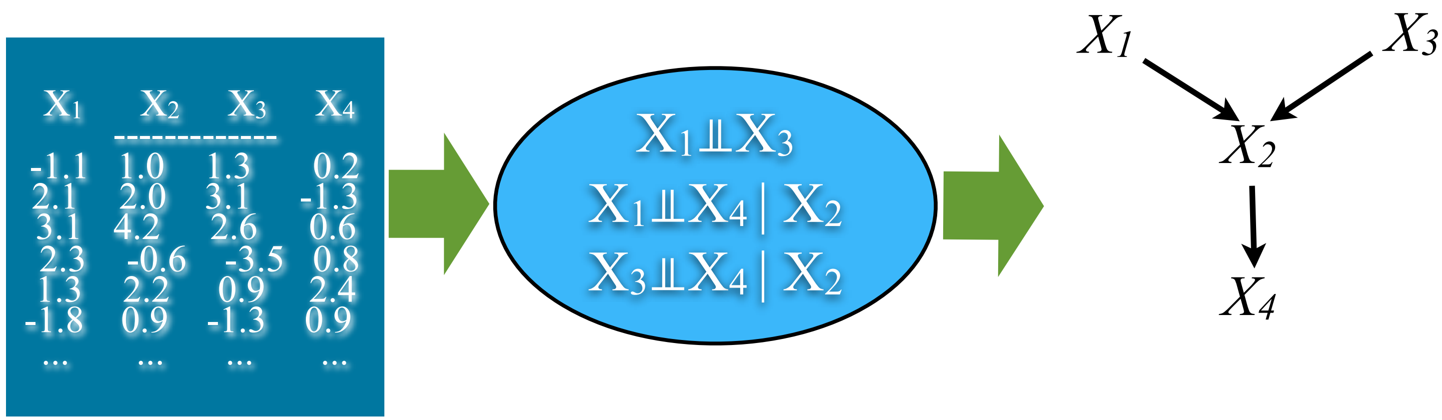
Faithfulness: all observed (conditional) independencies are entailed by Markov condition in the causal graph

Recall: $Y \perp\!\!\!\perp Z \Leftrightarrow P(Y|Z)=P(Y)$; $Y \perp\!\!\!\perp Z|X \Leftrightarrow P(Y|Z,X)=P(Y|X)$

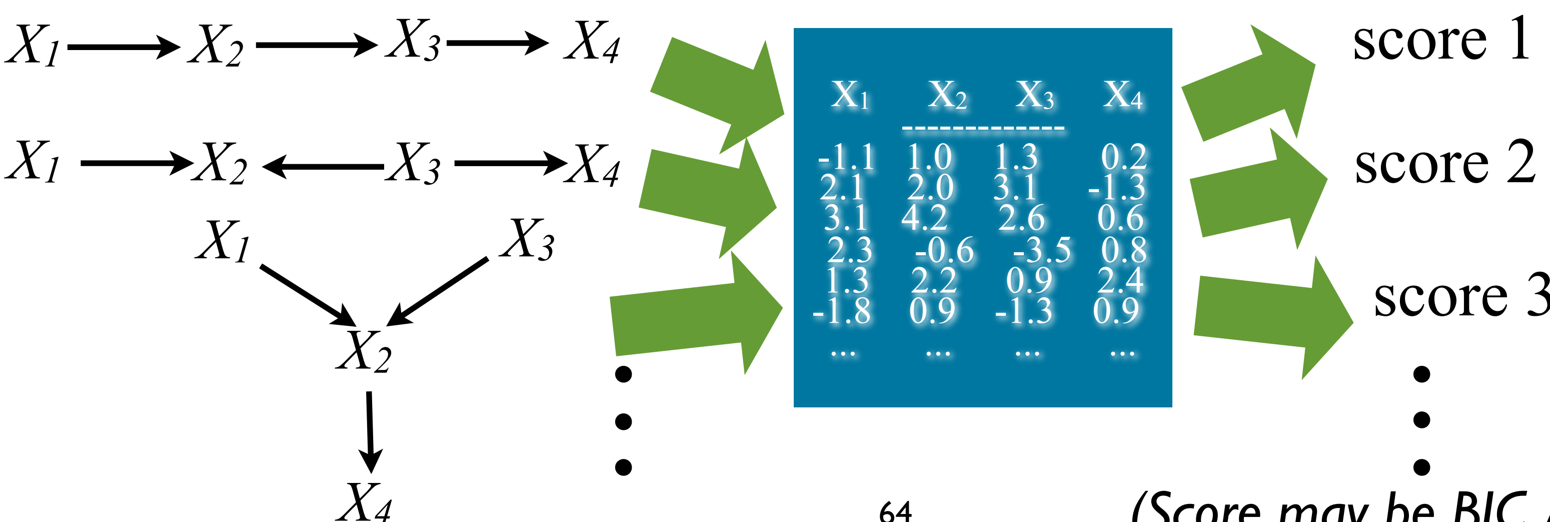
Slide from Kun Zhang's Spring 2019 10-708 Guest Lectures

Constraint-Based vs. Score-Based

- Constraint-based methods



- Score-based methods



Which one is the best?

64

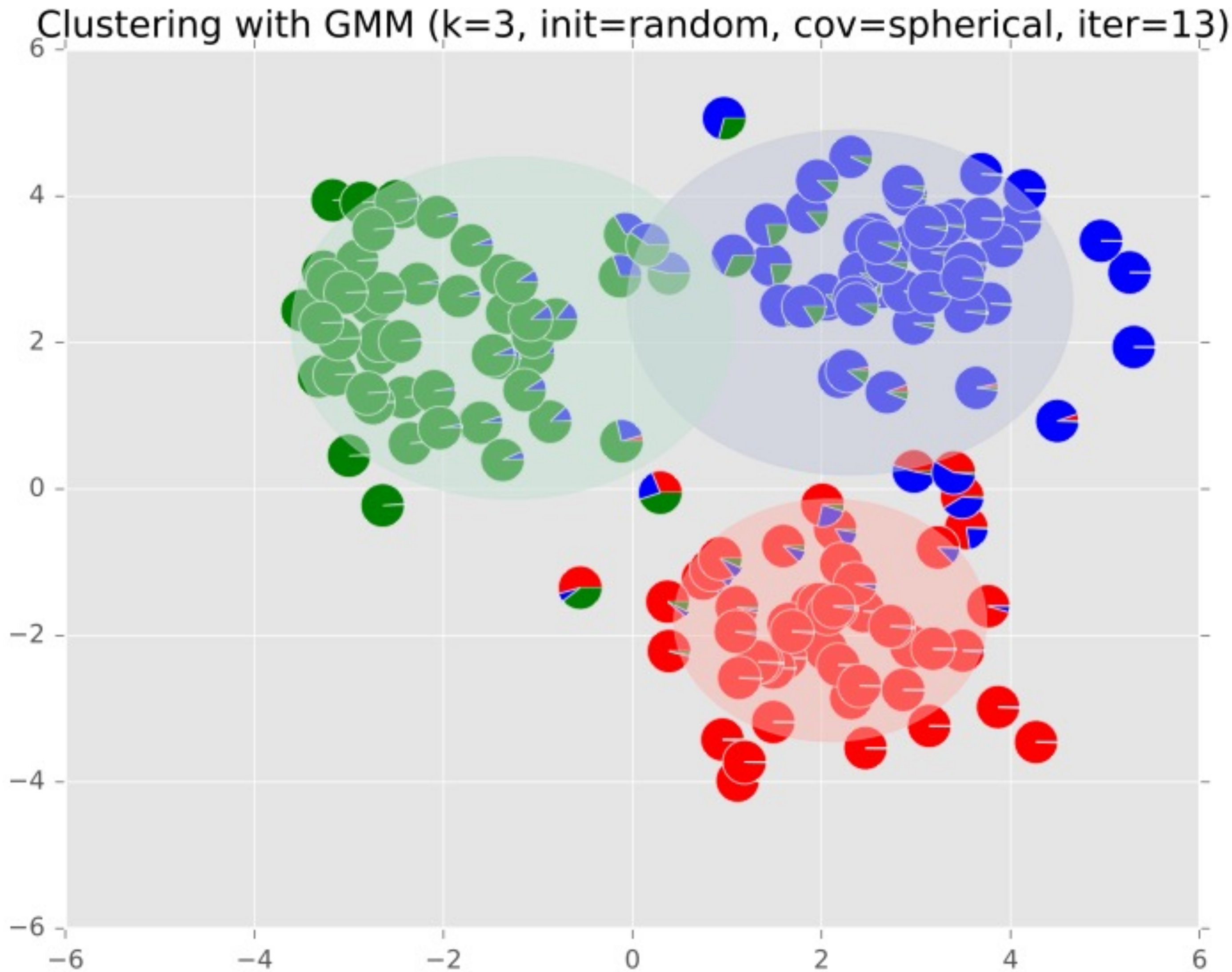
(Score may be BIC, AIC, etc.)

A CONUNDRUM: HOW TO PICK THE NUMBER OF LATENT CLUSTERS?

K-Means Algorithm

- **Given** unlabeled feature vectors
  $D = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)},\ldots, \mathbf{x}^{(N)}\}$
- **Initialize** cluster centers $c = \{\mathbf{c}^{(1)},\ldots, \mathbf{c}^{(K)}\}$
  and cluster assignments $z = \{z^{(1)}, z^{(2)},\ldots, z^{(N)}\}$
- **Repeat** until convergence:
  - for j in $\{1,\ldots,K\}$
    $\mathbf{c}^{(j)}$ = **mean** of **all** points assigned to cluster j
  - for i in $\{1,\ldots, N\}$
    $z^{(i)}$ = **index** j of cluster center **nearest** to $\mathbf{x}^{(i)}$

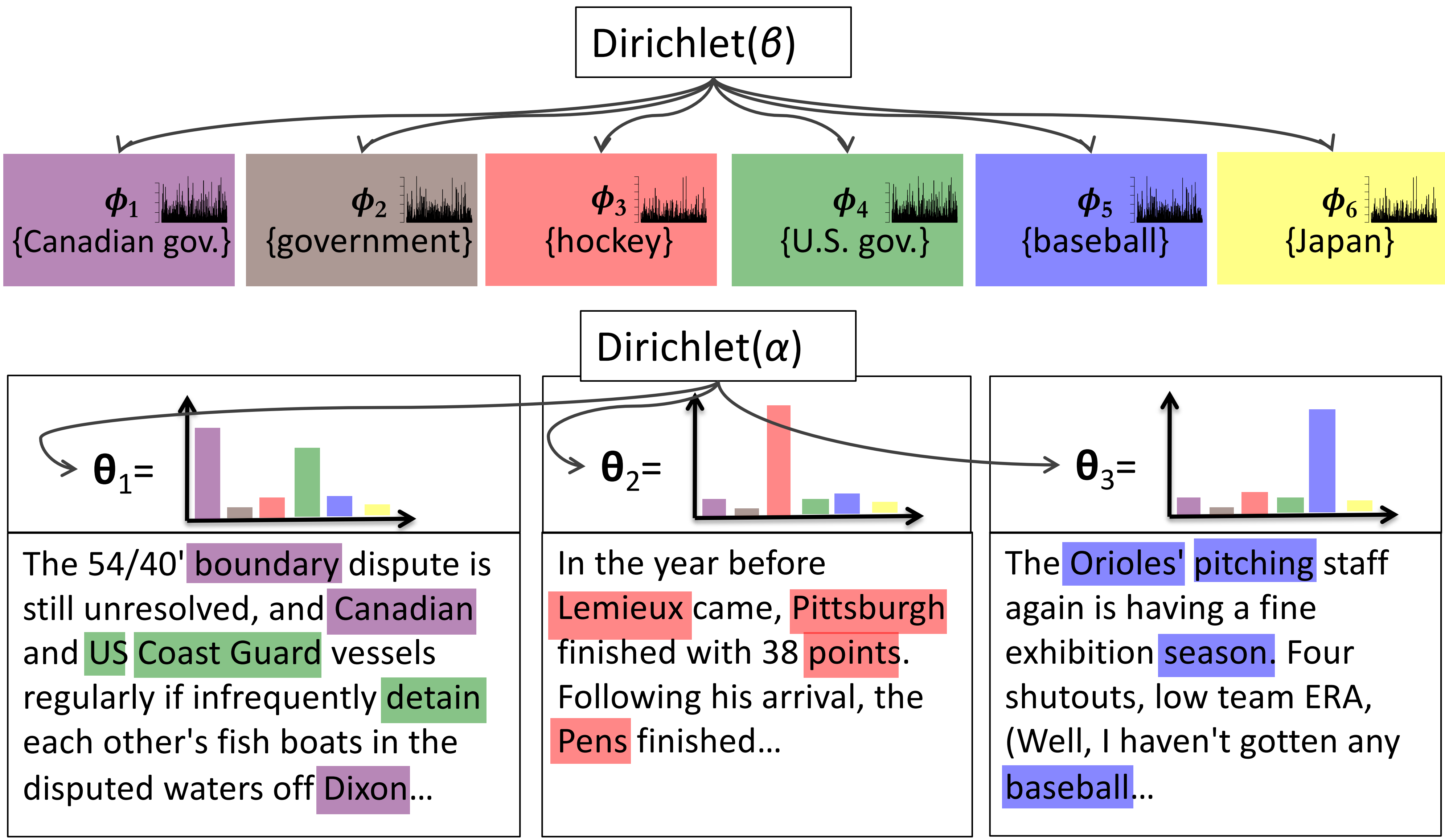
Example: GMM



Clustering with GMM (k=3, init=random, cov=spherical, iter=13)
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6

(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



Familiar models for unsupervised learning:

1. K-Means
2. Gaussian Mixture Model (GMM)
3. Latent Dirichlet Allocation (LDA)

But without labeled data, how do we know the right number of clusters / topics?

Outline

- **Motivation / Applications**
- **Background**
  - de Finetti Theorem
  - Exchangeability
  - Aglommerative and decimative properties of Dirichlet distribution
- **CRP and CRP Mixture Model**
  - Chinese Restaurant Process (CRP) definition
  - Gibbs sampling for CRP-MM
  - Expected number of clusters
- **DP and DP Mixture Model**
  - Ferguson definition of Dirichlet process (DP)
  - Stick breaking construction of DP
  - Uncollapsed blocked Gibbs sampler for DP-MM
  - Truncated variational inference for DP-MM
- **DP Properties**
- **Related Models**
  - Hierarchical Dirichlet process Mixture Models (HDP-MM)
  - Infinite HMM
  - Infinite PCFG

analogy to GMM

analogy LDA

BAYESIAN NONPARAMETRICS

Parametric vs. Nonparametric

- **Parametric models:**
  - **Finite** and **fixed** number of parameters
  - Number of parameters is **independent of the dataset**
- **Nonparametric models:**
  - **Have** parameters ("**infinite dimensional**" would be a better name)
  - Can be understood as having an **infinite** number of parameters
  - Can be understood as having a **random** number of parameters
  - Number of parameters can **grow with the dataset**
- **Semiparametric models:**
  - Have a **parametric** component and a **nonparametric** component

Table adapted from Jordan ICML 2005

Parametric vs. Nonparametric

	Frequentist	Bayesian
Parametric	Logistic regression, ANOVA, Fisher discriminant analysis, ARMA, etc.	Conjugate analysis, hierarchical models, conditional random fields
Semiparametric	Independent component analysis, Cox model, nonmetric MDS, etc.	[Hybrids of the above and below cells]
Nonparametric	Nearest neighbor, kernel methods, boostrap, decision trees, etc.	Gaussian processes, Dirichlet processes, Pitman-Yor processes, etc.

Table adapted from Ghahramani 2015

Parametric vs. Nonparametric

Application	Parametric	Nonparametric
function approximation	polynomial regression	Gaussian processes
classification	logistic regression	Gaussian process classifiers
clustering	mixture model, k-means	Dirichlet process mixture model
time series	hidden Markov model	infinite HMM
feature discovery	factor analysis, pPCA, PMF	infinite latent factor models

Parametric vs. Nonparametric

- **Def:** a *model* is a collection of distributions

$$\{p_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$$

- *parametric model*: the parameter vector is finite dimensional

$$\Theta \subset \mathcal{R}^k$$

- *nonparametric model*: the parameters are from a possibly infinite dimensional space, $\mathcal{F}$

$$\Theta \subset \mathcal{F}$$

Motivation #1

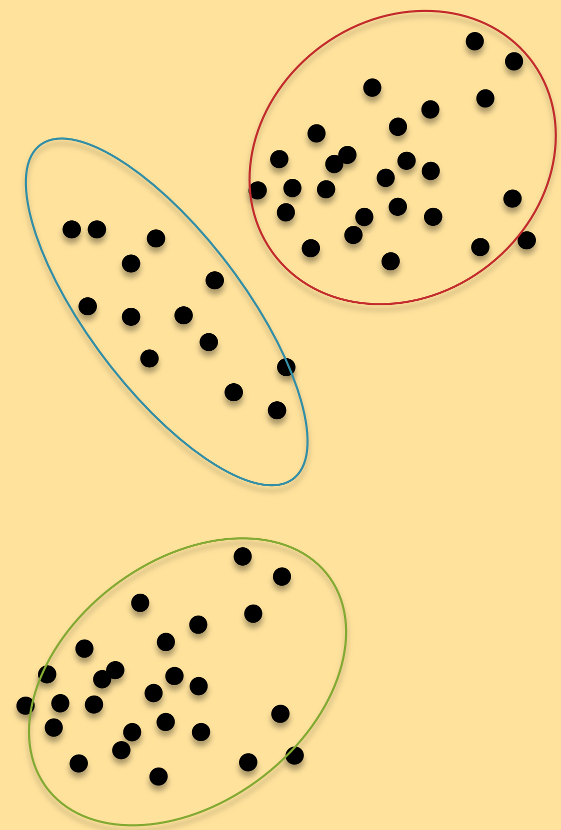
Model Selection

- For clustering:
  How many clusters in a **mixture model**?
- For topic modeling:
  How many topics in **LDA**?
- For grammar induction:
  How many non-terminals in a **PCFG**?
- For visual scene analysis:
  How many objects, parts, features?

Motivation #1

Model Selection

- For clustering:
  How many clusters in a **mixture model**?
- For topic modeling:
  How many topics in **LDA**?
- For grammar induction:
  How many non-terminals in a **PCFG**?
- For visual scene analysis:
  How many objects, parts, features?



Motivation #1

Model Selection

- For clustering:
  How many clusters in a **mixture model**?
- For topic modeling:
  How many topics in **LDA**?
- For grammar induction:
  How many non-terminals in a **PCFG**?
- For visual scene analysis:
  How many objects, parts, features?

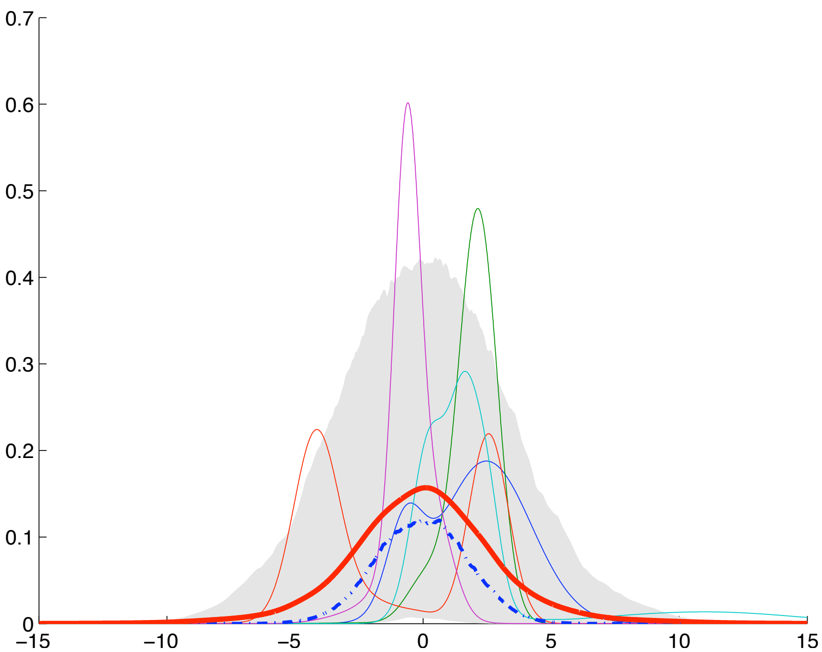
1. **Parametric approaches:** cross-validation, bootstrap, AIC, BIC, DIC, MDL, Laplace, bridge sampling, etc.
2. **Nonparametric approach:** average of an infinite set of models

Motivation #2

Density Estimation

- Given data, estimate a probability density function that best explains it
- A nonparametric prior can be placed over an infinite set of distributions

Prior:



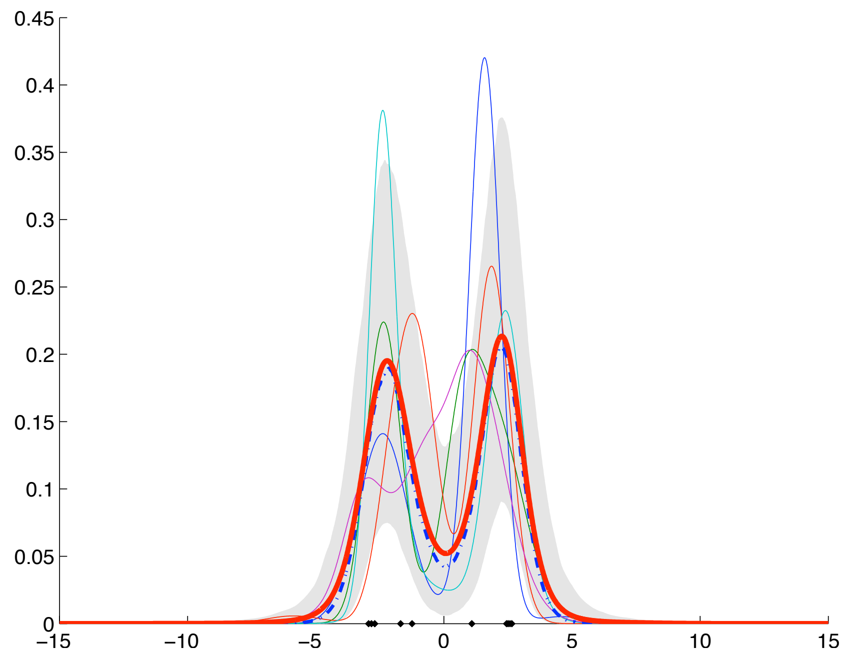
Red: mean density. Blue: median density. Grey: 5-95 quantile.
Others: draws.

Motivation #2

Density Estimation

- Given data, estimate a probability density function that best explains it
- A nonparametric prior can be placed over an infinite set of distributions

Posterior:



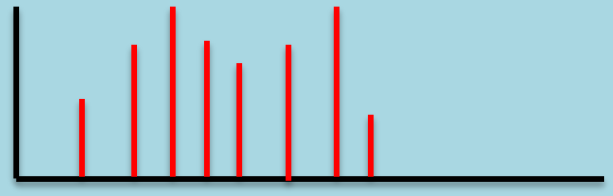
Red: mean density. Blue: median density. Grey: 5-95 quantile.
Black: data. Others: draws.

EXCHANGEABILITY AND DE FINETTI'S THEOREM

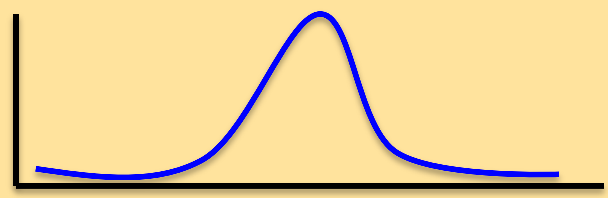
Background: Mixed Distribution

Suppose we have a random variable X drawn from some distribution $P_\theta(X)$ and X ranges over a set $\mathcal{S}$.

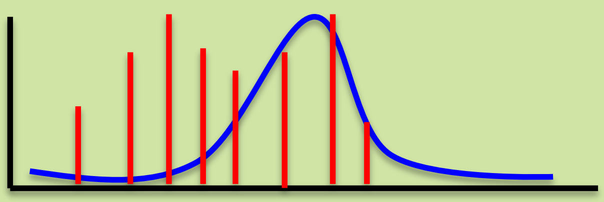
- Discrete distribution:
  $\mathcal{S}$ is a countable set.



- Continuous distribution:
  $P_\theta(X = x) = 0$ for all $x \in \mathcal{S}$

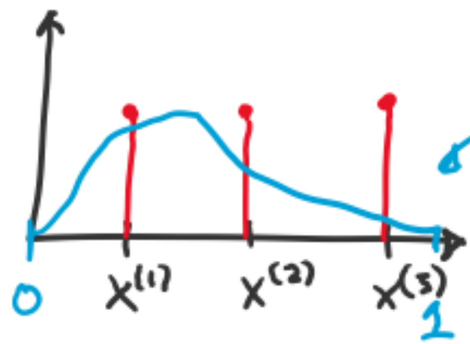


- Mixed distribution:
  $\mathcal{S}$ can be partitioned into two disjoint sets $\mathcal{D}$ and $\mathcal{C}$ s.t.
  1. $\mathcal{D}$ is countable and $0 < P_\theta(X \in D) < 1$
  2. $P_\theta(X = x) = 0$ for all $x \in \mathcal{C}$



Background: Mixed Distribution

Example:



$Beta(\alpha, \beta) = H$

$$x' \sim p_{mixed} = \left[\frac{1}{4} \sum_{i=1}^{3} \underbrace{\delta_{x^{(i)}}}_{\text{point mass distribution where all prob. mass is placed on subscript value}}\right] + \frac{1}{4} H$$

$$x' = \begin{cases} x^{(1)} & \text{w/prob } 1/4 \\ x^{(2)} & \text{w/prob } 1/4 \\ x^{(3)} & \text{w/prob } 1/4 \\ x'' \sim Beta(\alpha,\beta) = H & \text{w/prob } 1/4 \end{cases}$$

$$x \sim \delta_{x^{(1)}}$$

$$\Rightarrow \quad x = \begin{cases} x^{(1)} & \text{w/prob } 1.0 \\ \text{any other } x'' \neq x^{(1)} & \text{w/prob } 0.0 \end{cases}$$

Exchangability and de Finetti's Theorem

Exchangeability:

- **Def #1**: a joint probability distribution is **exchangeable** if it is invariant to permutation
- **Def #2:** The possibly infinite sequence of random variables $(X_1, X_2, X_3, \ldots)$ is **exchangeable** if for any finite permutation s of the indices $(1, 2, \ldots n)$:

$$P(X_1, X_2, \ldots, X_n) = P(X_{s(1)}, X_{s(2)}, \ldots, X_{s(n)})$$

Notes:

- *i.i.d.* and *exchangeable* are not the same!
- the latter says that if our data are reordered it doesn't matter

Exchangability and de Finetti's Theorem

Slide from Jordan ICML 2005

Theorem (De Finetti, 1935). *If* $(x_1, x_2, \ldots)$ *are infinitely exchangeable, then the joint probability* $p(x_1, x_2, \ldots, x_N)$ *has a representation as a mixture:*

$$\underbrace{p(x_1, x_2, \ldots, x_N)} = \int \left(\prod_{i=1}^{N} p(x_i \mid \theta) \right) \underline{dP(\theta)}$$

for some random variable θ.

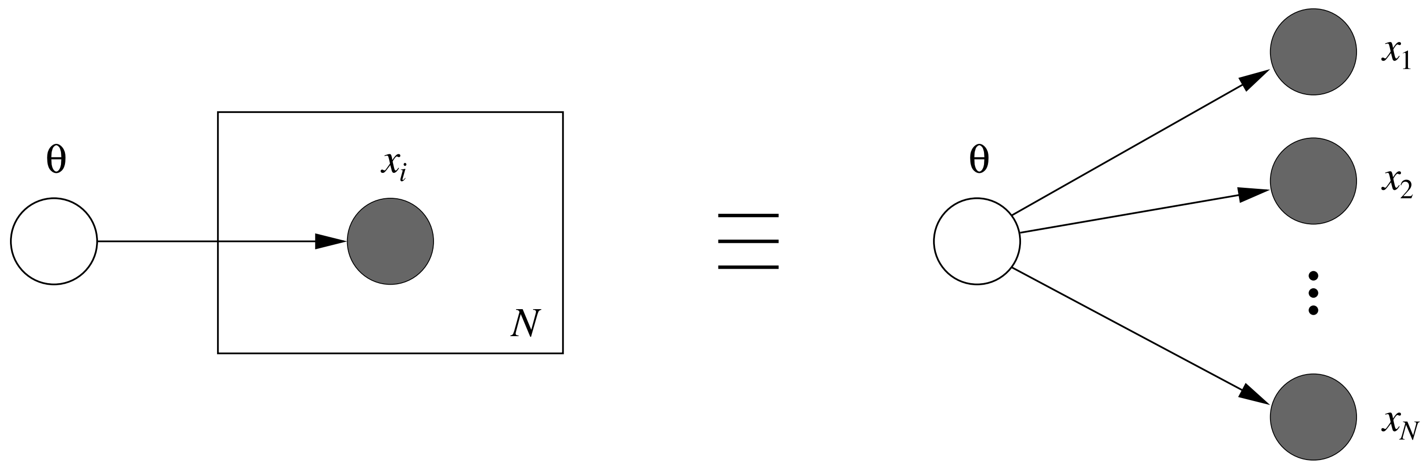
- The theorem wouldn't be true if we limited ourselves to parameters θ ranging over Euclidean vector spaces
- In particular, we need to allow θ to range over measures, in which case $P(\theta)$ is a measure on measures
  - the Dirichlet process is an example of a measure on measures...

Actually, this is the Hewitt-Savage generalization of the de Finetti theorem.
The original version was given for the Bernoulli distribution

Exchangability and de Finetti's Theorem

Slide from Jordan ICML 2005

- A *plate* is a "macro" that allows subgraphs to be replicated:



- Note that this is a graphical representation of the De Finetti theorem

$$p(x_1, x_2, \ldots, x_N) = \int p(\theta) \left(\prod_{i=1}^{N} p(x_i \,|\, \theta) \right) d\theta$$

Parametric vs. Nonparametric

Type of Model	Parametric Example	Nonparametric Example	
		Construction #1	Construction #2
distribution over counts	Dirichlet-Multinomial Model	Dirichlet Process (DP)	
		Chinese Restaurant Process (CRP)	Stick-breaking construction
mixture	Gaussian Mixture Model (GMM)	Dirichlet Process Mixture Model (DPMM)	
		CRP Mixture Model	Stick-breaking construction
admixture	Latent Dirichlet Allocation (LDA)	Hierarchical Dirichlet Process Mixture Model (HDPMM)	
		Chinese Restaurant Franchise	Stick-breaking construction

Chinese Restaurant Process & Stick-breaking Constructions

DIRICHLET PROCESS

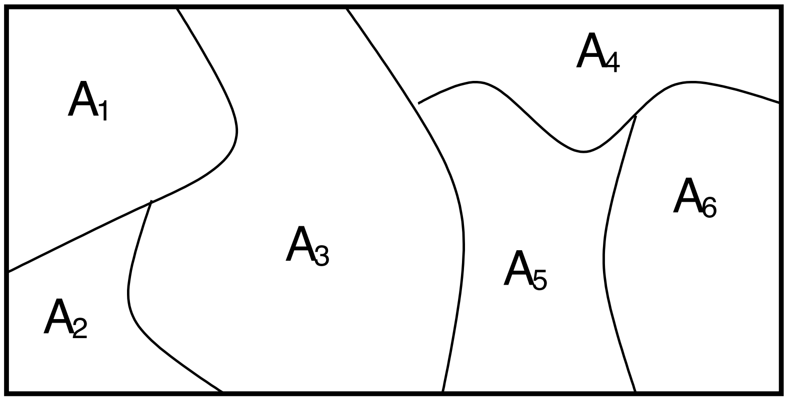
Dirichlet Process

Ferguson Definition

- Parameters of a DP:
  1. Base distribution, H, is a probability distribution over Θ
  2. Strength parameter, $\alpha \in \mathcal{R}$
- We say $G \sim \mathrm{DP}(\alpha, H)$
  if for any partition $A_1 \cup A_2 \cup \ldots \cup A_K = \Theta$
  we have:
  $(G(A_1), \ldots, G(A_K)) \sim \mathrm{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_K))$

In English: the DP is a distribution over probability measures s.t. marginals on finite partitions are Dirichlet distributed

A partition of the space Θ

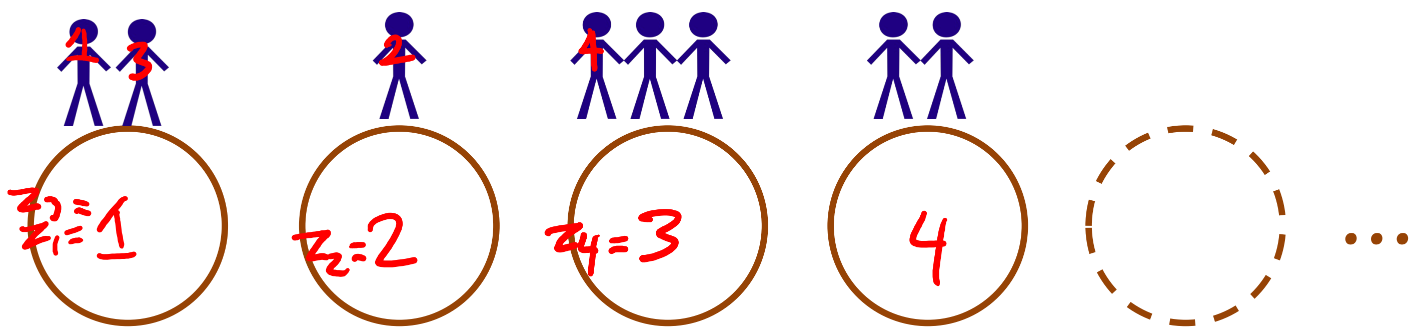


Chinese Restaurant Process

- Imagine a Chinese restaurant with an infinite number of tables
- Each customer enters and sits down at a table
  - The first customer sits at the first unoccupied table
  - Each subsequent customer chooses a table according to the following probability distribution:

    $p(\textit{kth occupied table}) \propto n_k$
    $p(\textit{next unoccupied table}) \propto \alpha$

    where n_k is the number of people sitting at the table k



$\frac{2}{8+\alpha}$ $\frac{1}{8+\alpha}$ $\frac{3}{8+\alpha}$ $\frac{2}{8+\alpha}$ $\frac{\alpha}{8+\alpha}$

Chinese Restaurant Process

Properties:

1. CRP defines a **distribution over clusterings** (i.e. partitions) of the indices $1,\ldots,n$
   - customer = index
   - table = cluster
2. We write $z_1, z_2, \ldots, z_n \sim CRP(\alpha)$ to denote a **sequence of cluster indices** drawn from a Chinese Restaurant Process
3. The CRP is an **exchangeable process**
4. **Expected number of clusters** given n customers (i.e. observations) is $O(\alpha \log(n))$
   - *rich-get-richer* effect on clusters: popular tables tend to get more crowded
5. Behavior of CRP with α:
   - As α goes to 0, the number of clusters goes to 1
   - As α goes to $+\infty$, the number of clusters goes to n

CRP vs. DP

Dirichlet Process: For both the CRP and stick-breaking constructions, if we marginalize out G, we have the following predictive distribution:

$$\theta_{n+1}|\theta_1, \ldots, \theta_n \sim \frac{1}{\alpha + n} \left(\alpha H + \sum_{i=1}^{n} \delta_{\theta_i} \right)$$

(Blackwell-MacQueen Urn Scheme)

The Chinese Restaurant Process is just a different construction of the Dirichlet Process where we have marginalized out G

Properties of the DP

1. Base distribution is the "mean" of the DP:

$$\mathbb{E}[G(A)] = H(A) \text{ for any } A \subset \Theta$$

2. Strength parameter is like "inverse variance"

$$V[G(A)] = H(A)(1 - H(A))/(\alpha + 1)$$

3. Samples from a DP are discrete distributions (stick-breaking construction of $G \sim \mathrm{DP}(\alpha, H)$ makes this clear)

4. Posterior distribution of $G \sim \mathrm{DP}(\alpha, H)$ given samples $\theta_1, \ldots, \theta_n$ from G is a DP

$$G|\theta_1, \ldots, \theta_n \sim \mathrm{DP}\left(\alpha + n, \frac{\alpha}{\alpha+n} H + \frac{n}{\alpha+n} \frac{\sum_{i=1}^n \delta_{\theta_i}}{n}\right)$$

Exchangability

Question: Q 1

Select All: Which of the following properties of an infinite sequence of random variables $X_1, X_2, X_3, \ldots$ ensure that they are infinitely exchangeable?

✓ A. For any pair of orderings $(i_1, i_2, \ldots, i_n)$ and $(j_1, j_2, \ldots, j_n)$ of the indices $(1, \ldots, n)$ the joint probability of the two orderings is the same 85

✓ B. The joint distribution is invariant to permutation 85

C. The joint distribution of the first n random variables can be represented as a mixture 25

✓ D. The random variables are independent and identically distributed 45

E - toxic

Answer: