Causal Inference and Invariance
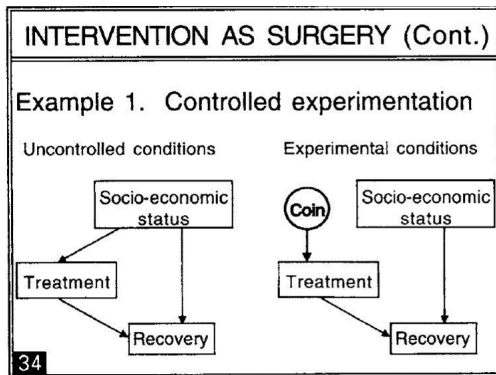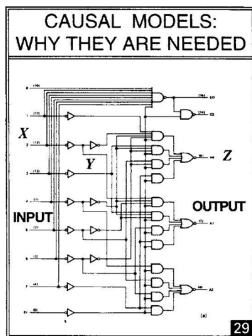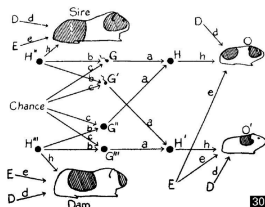
Charles Zheng and Qingyuan Zhao

Stanford University
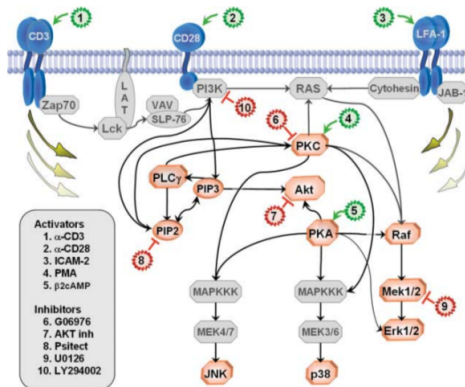
February 16, 2016

(Part 1/2)

CAUSAL MODELS: WHY THEY ARE NEEDED

INTERVENTION AS SURGERY (Cont.)
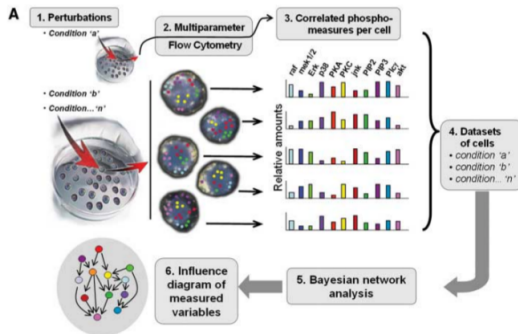
Example 1. Controlled experimentation

A hot application: systems biology

- Causal relationships $=$ *chemical interactions*.
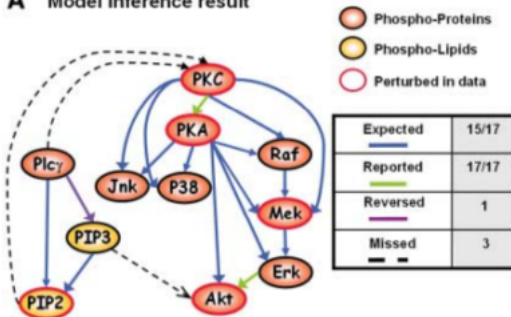- Experimenters *intervene* by injecting *activators* and *inhibitors*.

Protein signalling data

- Flow cytometry data from Sachs et al. *Science*, 2005.
- 1 observational data set + 9 interventions.

A **Model inference result**

- Causal inference applied to observational + interventional data.
- Recovered most of the known interactions.

The many facets of causality

- *Philosophy.* What is causality? How do we learn about cause and effect? *Aristotle, Hume.*
- *Computer science.* Can we build an artificial intelligence which reasons like humans? *Judea Pearl.*
- *Social science.*
- *Biology.*
- *Policy.* How will increasing interest rates influence unemployment?
- *Law.* Whose "fault" is it??
- *Statistics.* Answering the above questions using data!

Statistics and causality

- *Estimating causal effects from data.* Can we predict a causal effect based on observational or experimental data? E.g. effect of a medical treatment based on clinical trial data? Motivation for potential outcomes approach developed by Rubin, etc.

- *Bayesian networks/structure learning from data.* Can we model multivariate relationships using a network structure? Networks *can be* given causal interpretation, but causal inference is not the only motivation. Motivation for graphical lasso.

Section 1

Introduction

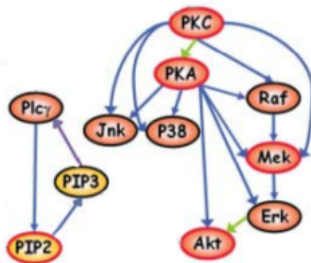*Graphical approach* pioneered by Judea Pearl.

TYPICAL DERIVATION IN CAUSAL CALCULUS

$P(c \mid do\{s\}) = \Sigma_t P(c \mid do\{s\}, t) P(t \mid do\{s\})$ — Probability Axioms

$= \Sigma_t P(c \mid do\{s\}, do\{t\}) P(t \mid do\{s\})$ — Rule 2

$= \Sigma_t P(c \mid do\{s\}, do\{t\}) P(t \mid s)$ — Rule 2

$= \Sigma_t P(c \mid do\{t\}) P(t \mid s)$ — Rule 3

$= \Sigma_{s'} \Sigma_t P(c \mid do\{t\}, s') P(s' \mid do\{t\}) P(t \mid s)$ — Probability Axioms

$= \Sigma_{s'} \Sigma_t P(c \mid t, s') P(s' \mid do\{t\}) P(t \mid s)$ — Rule 2

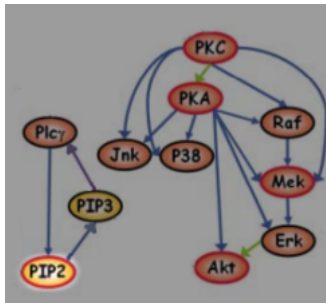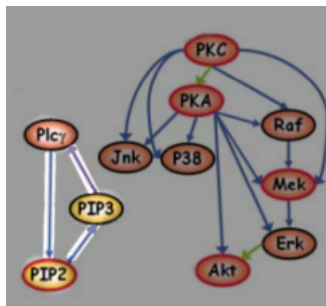$= \Sigma_{s'} \Sigma_t P(c \mid t, s') P(s') P(t \mid s)$ — Rule 3

Graphs: nodes and vertices

- Each variable in the dataset is given a *node*.
- Arrows indicate which variables *cause* which other variables. (Parents → children).
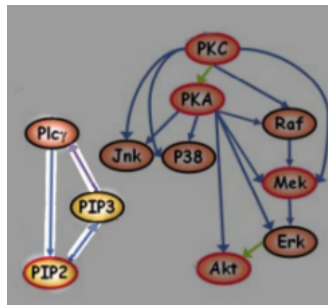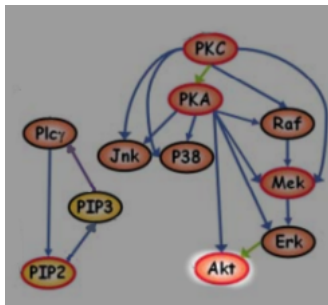
Intervening on variables in the system causes the distribution to change.

Causality and experiments

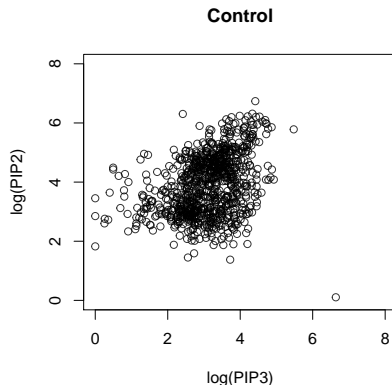- Not every variable will be affected by the intervention!
- Following the arrows tells you *which* variables which are affected.

- If we *inhibit* Akt, no other variables should be affected.
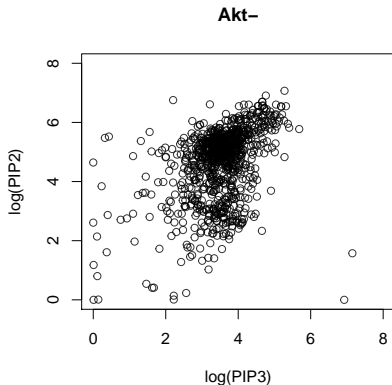- If we *inhibit* PIP2, then we may not only change the distribution of PIP2, but also PIP3.

Principle I: Which variables are affected.

Looking at Sachs data.
Joint distribution of PIP2 and PIP3 in the "control" case.

Akt−

Joint distribution of PIP2 and PIP3 when we intervene on Akt.

Principle I: Which variables are affected.
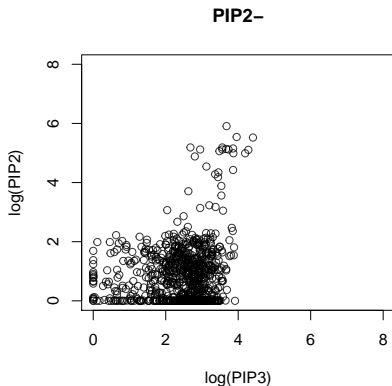
PIP2−

Joint distribution of PIP2 and PIP3 when we intervene on PIP2.

Principle I: Which variables are affected.
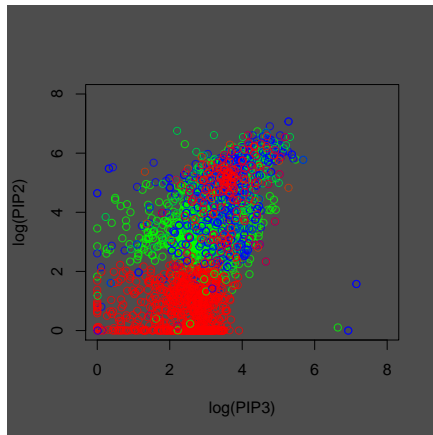
Control , **PIP2-** , **Akt-**

Intervening on PIP2 also affects the distribution of PIP3, while intervening on Akt does not (drastically) change the distribution.

Principle II: Conditional independence.
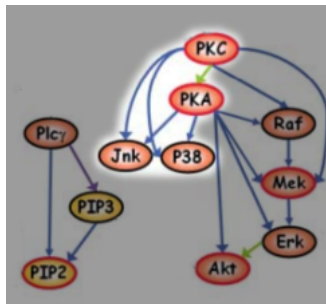
- Surprisingly, the structure of the causal graph implies certain *conditional independence* relationships.
- This allows the potential to infer causal relationships from observational data.

Principle II: Conditional independence.
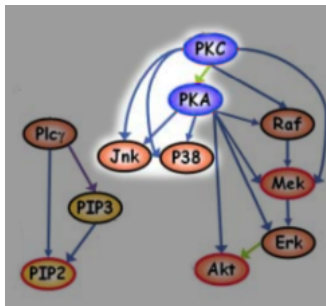
- Two variables are independent conditional on their common parents.
- Conditioning on PKC and PKA, Jnk and p38 should be independent.

Principle II: Conditional independence.
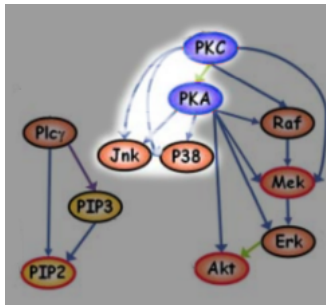
- "Once you and I condition on common factors, we are left with nothing in common."

Principle II: Conditional independence.

R = 0.608

log(Jnk)
n = 8319
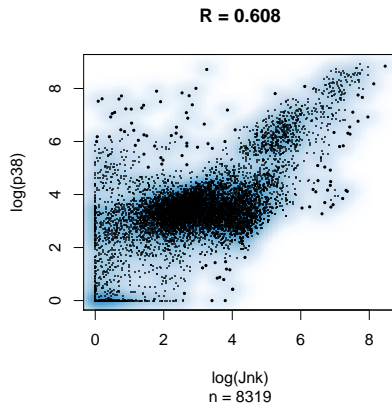
Marginally, p38 and Jnk are correlated.

Principle II: Conditional independence.
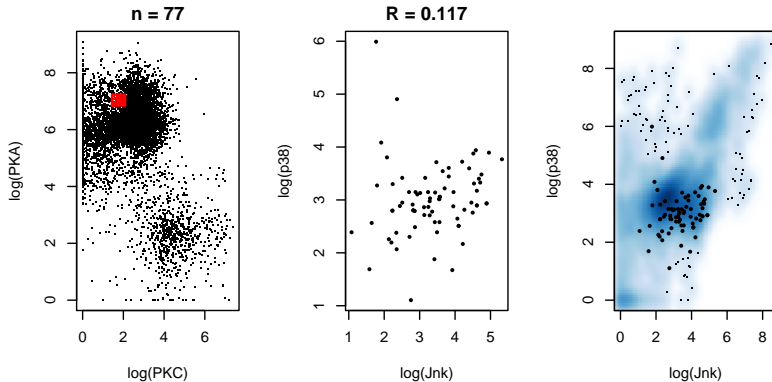
We can't condition on PKA and PKC since the data is continuous. But, conditioning on small windows seems to reduce association.

Principle II: Conditional independence.
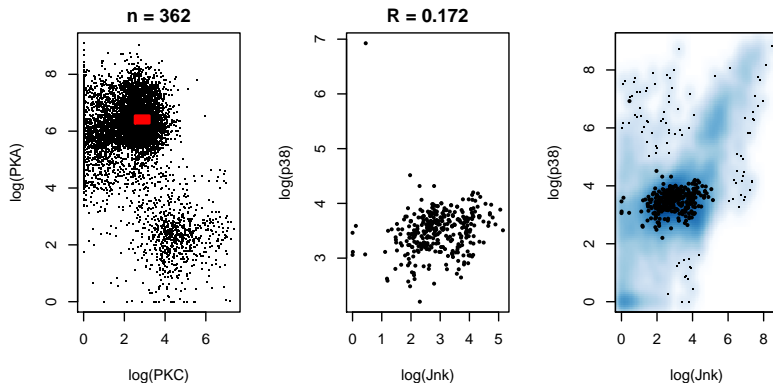
Left: We condition on (PKA, PKC) to lie within the indicated window.
Center: Conditional joint distribution of (Jnk, p38). *Right:* Conditional
join distribution, overlaid on marginal distribution.

Principle II: Conditional independence.
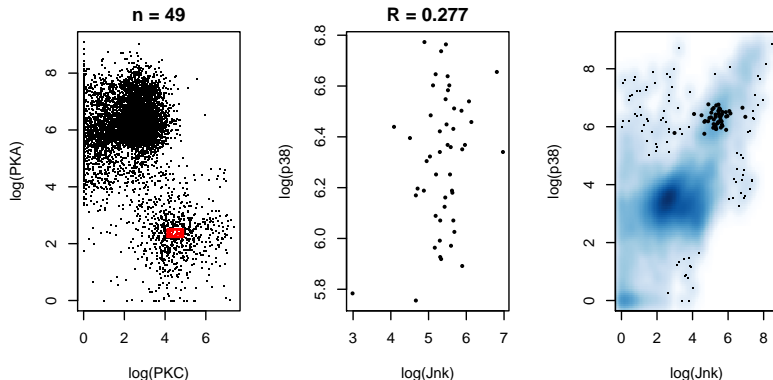
PKA and PKC *explain away* some (if not all) of the association between Jnk and p38. (Recall that R = 0.608 marginally.)

Principle III: Predictive invariance
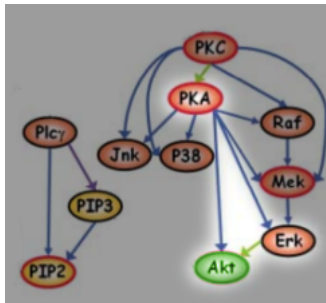
- The conditional distribution $\Pr[Akt|PKA, Erk]$ is invariant to interventions applied to other variables.
- Therefore, the optimal rule for predicting $\hat{Akt}(PKA, Erk)$ is invariant as well.

Principle III: Predictive invariance
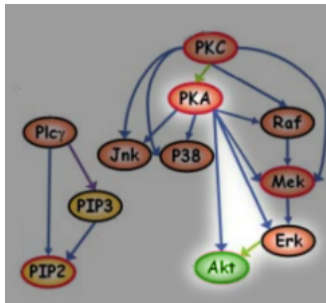
$\{PKA, Erk\}$ is an "invariant set" for *Akt* since:

- It includes all of the "direct" causes of *Akt* in the graph.
- It doesn't include any variables caused by *Akt*.
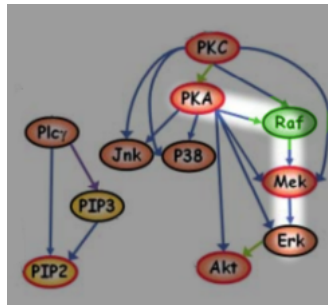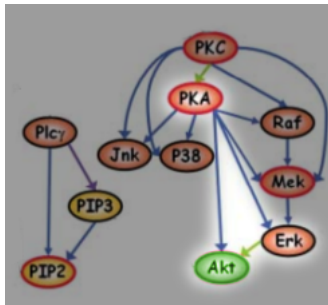
Principle III: Predictive invariance

In contrast, $\{PKA, Mek, Erf\}$ is *not* an invariant set for *Raf* since:

- .

- .
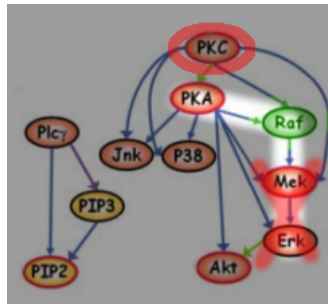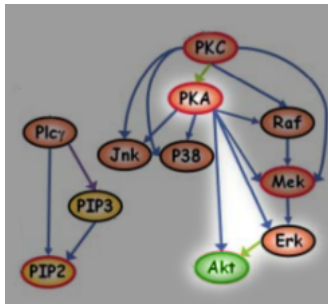
In contrast, $\{PKA, Mek, Erf\}$ is *not* an invariant set for Raf since:

- It is missing a direct cause of Raf.
- It contains variables which are caused by Raf.

$\{PKA, Erk\}$ is an invariant set for Akt.
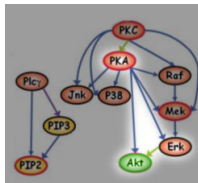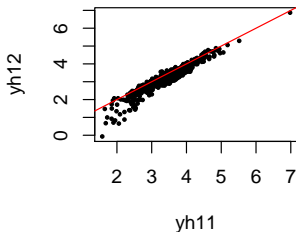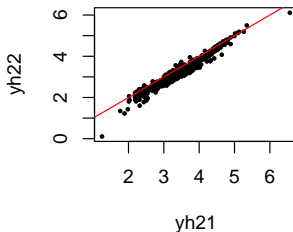
$\{PKA, Mek, Erf\}$ is *not* an invariant set for *Raf*.
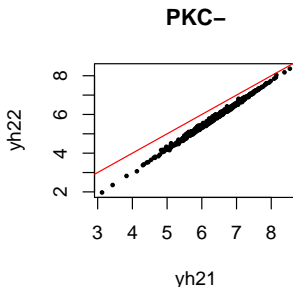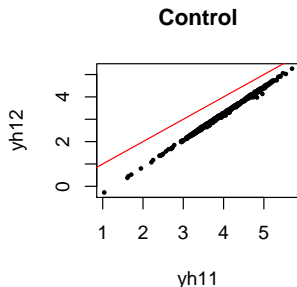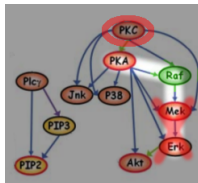
Predictive Invariance: Example
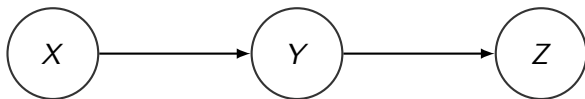
- Suppose we are trying to predict Y.
- $X \sim N(0, a)$.
- $Y | X \sim N(X, b)$.
- $Z | Y \sim N(Y, c)$.

Predictive Invariance: Example
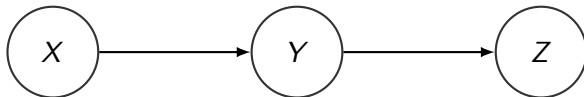
$X \sim N(0, a), \ Y|X \sim N(X, b), \ Z|Y \sim N(Y, c).$

- We can intervene by adding noise to X = changing $a \to a'$.
- Intervene by injecting noise to Z = changing $c \to c'$.
- Consider a linear model which predicts Y given X and Z.
- *Is the optimal prediction rule invariant under intervention?*

Predictive Invariance: Example

$$X \to Y \to Z$$

$$X \sim N(0, a), \ Y|X \sim N(X, b), \ Z|Y \sim N(Y, c).$$

The joint distribution is

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} a & a & a \\ a & a+b & a+b \\ a & a+b & a+b+c \end{bmatrix} \right)$$

The optimal prediction rule is given by

$$\mathbf{E}[Y|X, Z] = \mu_Y + \Sigma_{Y,XZ} \Sigma_{XZ}^{-1}(X - \mu_X, Z - \mu_Z) = \frac{c}{b+c}X + \frac{b}{b+c}Z.$$

Predictive Invariance: Example

$$X \sim N(0, a), \ Y|X \sim N(X, b), \ Z|Y \sim N(Y, c).$$

Optimal prediction rule:

$$\mathbf{E}[Y|X, Z] = \underbrace{\frac{c}{b+c}}_{\beta_X} X + \underbrace{\frac{b}{b+c}}_{\beta_Z} Z.$$

i.e. Y is a weighted average of X and Z ($\beta_X + \beta_Z = 1$).

- Imagine c is very small, i.e. $Z = Y + $ tiny noise. Then Z is a great predictor of Y! $\beta_Z \approx 1$.
- Conversely, if b is small, that means $Y = X + $ tiny noise. $\beta_X \approx 1$.
- If $b = c$, then $\beta_X = \beta_Z = 1/2$.

Predictive Invariance: Example

But is the OLS predictive rule invariant?

$$\mathbf{E}[Y|X, Z] = \frac{c}{b+c}X + \frac{b}{b+c}Z.$$

If we intervene on Z, changing c to c', the OLS coefficients change too. The model is not invariant.
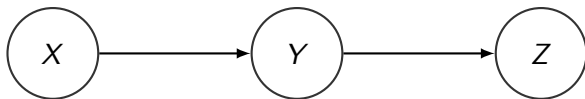
"Real-life" example. $X =$ lifetime total number of bagels eaten, $Y =$ Body Mass Index (BMI), $Z =$ how many pull-ups you can do? Z is a good predictor of Y, unless you "intervene" by offering a \$100 prize for doing 10 pull-ups.

Predictive Invariance: Example
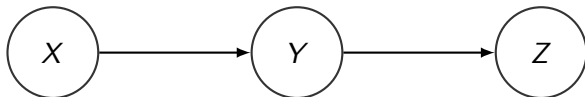
- In contrast, consider predicting Y using *only* X.
- $\{X\}$ is an invariant set for Y because it contains all direct parents and no children of Y.
- Indeed,

$$\mathbf{E}[Y|X] = \frac{\text{Cov}(Y, X)}{\text{Cov}(X)} X = \frac{a}{a} X = X.$$

  The OLS coefficient, 1, does not depend on a or c, and hence is *invariant* under interventions.

- *Exercise.* Is $\{Z\}$ an invariant set for Y?

Predictive Invariance: General case
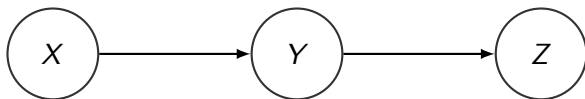
$$p(x, y, z) = p(x)p(y|x)p(z|y)$$

Interventions:

$$p(x) \rightarrow \tilde{p}(x) \text{ or } p(z|y) \rightarrow \tilde{p}(z|y)$$

Non-invariant rule:

$$p(y|x, z) = \frac{p(x, y, z)}{p(x, z)} = \frac{p(x)p(y|x)p(z|y)}{p(x)p(z|x)} = \frac{p(y|x)p(z|y)}{p(z|x)}$$

affected by intervention $p(z|y) \rightarrow \tilde{p}(z|y)$.

Overview: Principles of Causal Inference

Causal relationships in a system represented by a graph. The graph tells you:

I. which variables are affected by an intervention.

II. what conditional independence relationships exist in the joint distribution.

III. which sets of predictors and responses will have "invariant" optimal predictive rules.

These three intuitive principles are equivalent to *factorization* definition of graphical model:

$$p(x_1, \ldots, x_k) = \prod_{i=1}^{k} p(x_i | \text{Parents}(x_i)).$$

Section 2

Statistical Methods
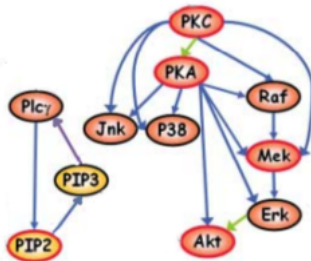
- Suppose we want to reduce the expression level of PKC in the cell. We have a treatment (an enzyme) which can inhibit PIP2– what would be the *treatment effect*

$$\mathbf{E}[PKC|do(PIP2)] - \mathbf{E}[PKC] =?$$

- *Controlled experiment.* Do an experiment where we randomize the treatment, estimate the treatment effect using the difference

  mean of the treated − mean of the controls

- *Observational data.* We observe that the enzyme we are considering is sometimes expressed in the cell naturally. Can we estimate the treatment effect even without having done a controlled experiment?
  - *Potential outcomes approach.* (By Rubin et al.) Match treated and untreated observations using *propensity scores.* Optional: sensitivity analyses.
  - *Graphical approach.* (Pearl et al.) Supposing we know the structure of the graph (or we can try to learn it), apply *calculus of interventions.*

Observational data
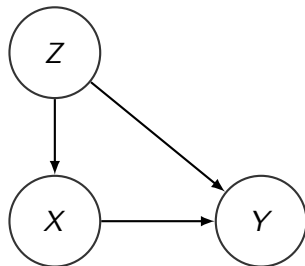
- $X = \{0, 1\}$ is the treatment variable, Y is the outcome of interest, Z are confounders.
- Want to estimate effect of treatment.
- No confounders?? Use $\mathbf{E}[Y|X=1] - \mathbf{E}[Y|X=0]$, done!
- *Unobserved* confounders?! We'll discuss next time..
- For now, assume all confounders are observed.

Calculus of interventions

- Our goal is to infer the average treatment effect,

$$\mathbf{E}[Y|do(X = 1)] - \mathbf{E}[Y|do(X = 0)]$$

- The 'do' notation refers to interventions.
- We have observational data,

$$p(x, y, z) = p(y|x, z)p(x|z)p(z).$$
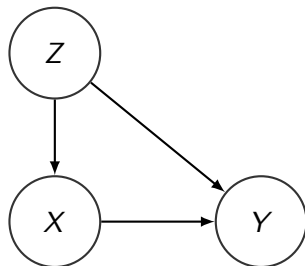
Adjustment for confounders

To compute $\mathbf{E}[y|do(X=1)] - \mathbf{E}[y|do(X=0)]$, it suffices to estimate $p(y|do(x))$.

- $p(x, y, z) = p(y|x, z)p(x|z)p(z)$.
- Step one: convert between 'do' notation and probability

$$p(y|x, z) = p(y|do(x), z).$$

- Step two: apply law of total probability

$$p(y|do(x)) = \sum_z p(y, z|do(x)) = \sum_z p(y|do(x), z)p(z|do(x))$$

$$= \sum_z p(y|x, z)p(z).$$

- Step three: plug in empirical $\hat{p}(y|x, z)$ obtained by conditioning on z.

Potential outcomes formulation

Potential outcomes uses notation $Y^{(0)}$ for $Y|do(X = 0)$ and $Y^{(1)}$ for $Y|do(X = 1)$.

- First, determine the nature of the "assignment mechanism," X. In this case, the probability that $X = 1$ is determined by Z, $f(z) = p(X = 1|z)$. The function $f(z)$ is called the *propensity score*.
- Estimate $f(z)$ using regression (e.g. logistic regression X on Z).
- Match 'cases' (which have $X = 1$) to 'controls' (with $X = 0$) that have similar propensity scores.
- Compute average treatment effect by a weighted average of:

$$\Pr[X = 1] \underbrace{(\mathbf{E}[Y|X = 1] - \mathbf{E}[Y|do(X = 0), X = 1])}_{\text{treatment effect on treated}}$$

$$+ \Pr[X = 0] \underbrace{(\mathbf{E}[Y|X = 0, do(X = 1)] - \mathbf{E}[Y|X = 0])}_{\text{treatment effect on untreated}}$$

Comparison of approaches

- Approaches are *almost* equivalent in this example. Matching based on $f(z) = p(x|z)$ and weighting based on probability of treatment produces the same result as conditioning on Z.

- Practically, there are still subtle differences. Easier to get confidence intervals in one approach than another, etc.

- The real devil is in the assumptions of the approaches. Which assumptions are needed, and how can they be checked? How much knowledge about the graph do we need for each approach?

Section 3

Conclusions

- Causal models imply statements about which variables get affected by interventions, which conditional independencies exist, and which sets of variables lead to invariant prediction rules.
- Real data may reveal a causal model to be more of a useful approximation than a literal description.
- Estimation of causal effects from experiments with imperfect randomization or observational data is a common goal in causal inference, and can be addressed using the graphical approach or potential outcomes framework.

Next time...

In part II of the talk, we'll go into more detail about the limitations of causal inference, and criticism of many common practices.

- Does causal inference require us to make too many strong assumptions?
- Do instrumental variable approaches give sensible results?
- How robust are causal inferences to hidden confounders?
- Is it realistic to be able to learn structure from completely observational data?

We'll also introduce the causal invariance approach by Peters, Meinshausen, and Bühlmann–could this new approach extend the applicability of causal inference?

References

- Sachs, Karen, et al. "Causal protein-signaling networks derived from multiparameter single-cell data." *Science* 308.5721 (2005): 523-529.
- Nagarajan, Radhakrishnan, Marco Scutari, and Sophie Lèbre. "Bayesian networks in R." Springer 122 (2013): 125-127.
- Pearl, Judea. *Causality.* Cambridge university press, 2009.
- Morgan, Stephen L., and Christopher Winship. *Counterfactuals and causal inference.* Cambridge University Press, 2014.
- Peters, Jonas, Peter Bühlmann, and Nicolai Meinshausen. "Causal inference using invariant prediction: identification and confidence intervals." arXiv preprint arXiv:1501.01332 (2015).