Causal Inference in Machine Learning
in Computational Biology

ICB Retreat, Kloster Irsee

October 25, 2016

F. Alexander Wolf | falexwolf.de
Institute of Computational Biology
Helmholtz Zentrum München

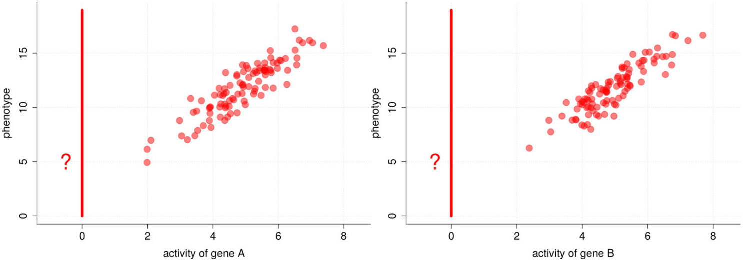
HelmholtzZentrum münchen
Deutsches Forschungszentrum für Gesundheit und Umwelt

HELMHOLTZ
| ASSOCIATION

Problem figures from Jonas Peters

Gene A and gene B both correlate with a phenotype.

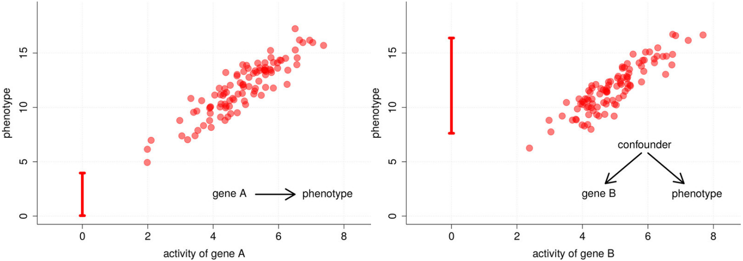
▷ What is the best prediction for the phenotype if we delete a gene?



Problem figures from Jonas Peters

Gene A and gene B both correlate with a phenotype.

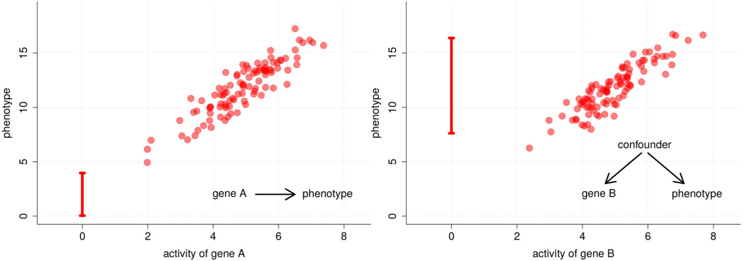
- ▷ What is the best prediction for the phenotype if we delete a gene?
- ▷ It certainly depends on the "causal structure" of the system.



Problem figures from Jonas Peters

Gene A and gene B both correlate with a phenotype.

- ▷ What is the best prediction for the phenotype if we delete a gene?
- ▷ It certainly depends on the “causal structure” of the system.

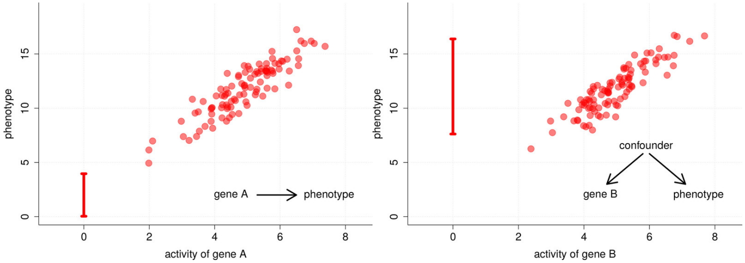


- ▷ To describe the *interventional* distribution, a **predictive** model needs to incorporate the causal structure of the system.

Problem figures from Jonas Peters

Gene A and gene B both correlate with a phenotype.

- What is the best prediction for the phenotype if we delete a gene?
- It certainly depends on the "causal structure" of the system.



- To describe the *interventional* distribution, a **predictive** model needs to incorporate the causal structure of the system.

How trustworthy is a given Machine Learning model? ▷ Ribeiro, Singh & Guestrin, arXiv:1602.04938 (2016)

Predictive models

- ▷ To fit the **observational data**, we need

$$Y = f(X_A, X_B) + N \;\mid\; \varnothing.$$

Predicts wrong interventional distribution.

Predictive models

- To fit the **observational data**, we need

$$Y = f(X_A, X_B) + N \;\mid\; \varnothing.$$

  Predicts wrong interventional distribution.

- To describe the **interventional data**, we'd rather set

$$Y = f(X_A) + N \;\mid\; \mathrm{do}(X_B = 0).$$

  Fails to describe observational distribution. Most likely, it's also terribly wrong in quantifying the effect of X_A on Y.

Predictive models

- To fit the **observational data**, we need

$$Y = f(X_A, X_B) + N \ \mid \ \varnothing.$$

Predicts wrong interventional distribution.

- To describe the **interventional data**, we'd rather set

$$Y = f(X_A) + N \ \mid \ \mathrm{do}(X_B = 0).$$

Fails to describe observational distribution. Most likely, it's also terribly wrong in quantifying the effect of X_A on Y.

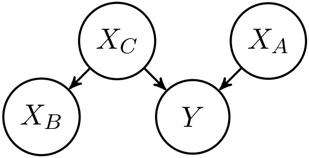
- Measure the confounder X_C, and assume there are no further confounders. Then,

$$Y = f(X_A, X_C) + N \ \mid \ \varnothing \quad \text{or} \quad \mathrm{do}(X_B = 0).$$

is a predictive model, which fits **both observational and interventional data**. Some people call it **"causal model"**.

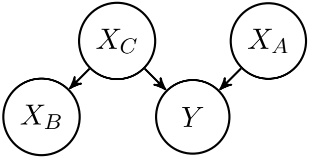
Graphical models

Visualize **cause-effect relations**.



Graphical models

Visualize **cause-effect relations**.

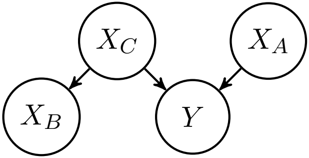


This “looks” like a **directed acyclic graphical (DAG) model**, which is a **conditional independence structure** that encodes

$$X_i \perp\!\!\!\perp \mathrm{NonDescendants}(X_i) \mid \mathrm{Parents}(X_i). \quad \text{(\textit{Markov property})}$$

Graphical models

Visualize **cause-effect relations**.



This "looks" like a **directed acyclic graphical (DAG) model**, which is a **conditional independence structure** that encodes

$$X_i \perp\!\!\!\perp \mathrm{NonDescendants}(X_i) \mid \mathrm{Parents}(X_i). \quad (\textit{Markov property})$$

If we specify the functional form that generates the distribution as

$$X_i = f_i(\mathrm{Pa}(X_i), N_i),$$

we call the DAG **structural equation model**.

Relation to causality

Relation to causality

Observational distribution (*Markov factorization*)

$$p(X_1, \ldots, X_d) = \prod_{i=1}^{d} p(X_i | \mathrm{Pa}(X_i)) \overset{\text{e.g.}}{=} \prod_{i=1}^{d} \mathcal{N}(X_i | f_i(\mathrm{Pa}(X_i)), \sigma^2)$$

Relation to causality

Observational distribution (*Markov factorization*)

$$p(X_1, \ldots, X_d) = \prod_{i=1}^{d} p(X_i | \mathrm{Pa}(X_i)) \overset{\text{e.g.}}{=} \prod_{i=1}^{d} \mathcal{N}(X_i | f_i(\mathrm{Pa}(X_i)), \sigma^2)$$

Interventional distribution ("surgery on the graph")

$$p(X_1, \ldots, X_d | \mathrm{do}(X_j = x_j)) = \prod_{i \neq j} p(X_i | \mathrm{Pa}(X_i), X_j = x_j)$$

Relation to causality

Observational distribution (*Markov factorization*)

$$p(X_1, \ldots, X_d) = \prod_{i=1}^{d} p(X_i | \mathrm{Pa}(X_i)) \overset{\text{e.g.}}{=} \prod_{i=1}^{d} \mathcal{N}(X_i | f_i(\mathrm{Pa}(X_i)), \sigma^2)$$

Interventional distribution ("surgery on the graph")

$$p(X_1, \ldots, X_d | \mathrm{do}(X_j = x_j)) = \prod_{i \neq j} p(X_i | \mathrm{Pa}(X_i), X_j = x_j)$$

- Correct interventional distributions are **only** obtained from the observational distribution, if **all edges** denote cause-effect relationships.

Relation to causality

Observational distribution (*Markov factorization*)

$$p(X_1, \ldots, X_d) = \prod_{i=1}^{d} p(X_i | \mathrm{Pa}(X_i)) \overset{\text{e.g.}}{=} \prod_{i=1}^{d} \mathcal{N}(X_i | f_i(\mathrm{Pa}(X_i)), \sigma^2)$$

Interventional distribution ("surgery on the graph")

$$p(X_1, \ldots, X_d | \mathrm{do}(X_j = x_j)) = \prod_{i \neq j} p(X_i | \mathrm{Pa}(X_i), X_j = x_j)$$

- Correct interventional distributions are **only** obtained from the observational distribution, if **all edges** denote cause-effect relationships.
  - ▷ The likelihood for interventional data is highly sensitive to non-causal edges.

Relation to causality

Observational distribution (*Markov factorization*)

$$p(X_1, \ldots, X_d) = \prod_{i=1}^{d} p(X_i | \mathrm{Pa}(X_i)) \overset{\text{e.g.}}{=} \prod_{i=1}^{d} \mathcal{N}(X_i | f_i(\mathrm{Pa}(X_i)), \sigma^2)$$

Interventional distribution ("surgery on the graph")

$$p(X_1, \ldots, X_d | \mathrm{do}(X_j = x_j)) = \prod_{i \neq j} p(X_i | \mathrm{Pa}(X_i), X_j = x_j)$$

- Correct interventional distributions are **only** obtained from the observational distribution, if **all edges** denote cause-effect relationships.
  - ▷ The likelihood for interventional data is highly sensitive to non-causal edges.
  - ▷ The model can efficiently be learned and easily falsified.

Structure Learning

How to learn conditional independence structure from data?

Structure Learning

How to learn conditional independence structure from data?

- Constraint-based methods. Pearl & Verma (1991) Spirtes, Glymour & Scheines (2000)
  Perform systematic conditional independence tests.

- Score-based methods. Chickering (2002)
  Maximize the likelihood or posterior of a graphical model.

Structure Learning

How to learn conditional independence structure from data?

- Constraint-based methods. Pearl & Verma (1991) Spirtes, Glymour & Scheines (2000)
  Perform systematic conditional independence tests.

  + PC algorithm scales well to large dimensions.

- Score-based methods. Chickering (2002)
  Maximize the likelihood or posterior of a graphical model.

  − Does not scale.

Structure Learning

How to learn conditional independence structure from data?

- Constraint-based methods. Pearl & Verma (1991) Spirtes, Glymour & Scheines (2000)
  Perform systematic conditional independence tests.
  - + PC algorithm scales well to large dimensions.
  - + Consistency results exist.

- Score-based methods. Chickering (2002)
  Maximize the likelihood or posterior of a graphical model.
  - − Does not scale.
  - − Consistency results only in low dimensions.

Structure Learning

How to learn conditional independence structure from data?

- Constraint-based methods. Pearl & Verma (1991) Spirtes, Glymour & Scheines (2000)
  Perform systematic conditional independence tests.
  - + PC algorithm scales well to large dimensions.
  - + Consistency results exist.
  - − "Not very reliable".

- Score-based methods. Chickering (2002)
  Maximize the likelihood or posterior of a graphical model.
  - − Does not scale.
  - − Consistency results only in low dimensions.
  - + "More reliable".

Structure Learning

How to learn conditional independence structure from data?

- Constraint-based methods. Pearl & Verma (1991) Spirtes, Glymour & Scheines (2000)
  Perform systematic conditional independence tests.
  - + PC algorithm scales well to large dimensions.
  - + Consistency results exist.
  - − "Not very reliable".
  - − Not a generative method.

- Score-based methods. Chickering (2002)
  Maximize the likelihood or posterior of a graphical model.
  - − Does not scale.
  - − Consistency results only in low dimensions.
  - + "More reliable".
  - + Generative method.

Structure Learning

How to learn conditional independence structure from data?

- Constraint-based methods. Pearl & Verma (1991) Spirtes, Glymour & Scheines (2000)
  Perform systematic conditional independence tests.
  - + PC algorithm scales well to large dimensions.
  - + Consistency results exist.
  - − "Not very reliable".
  - − Not a generative method.
  - − Problematic in the presence of hidden variables.
- Score-based methods. Chickering (2002)
  Maximize the likelihood or posterior of a graphical model.
  - − Does not scale.
  - − Consistency results only in low dimensions.
  - + "More reliable".
  - + Generative method.
  - + Bayesian ansatz allows to resolve hidden variables.

SGS and PC algorithm Spirtes, Glymour & Scheines (2000)

PC algorithm is most popular constraint-based method.

1. Start with a fully connected graph.

SGS and PC algorithm Spirtes, Glymour & Scheines (2000)

PC algorithm is most popular constraint-based method.

1. Start with a fully connected graph.
2. Reduce edges by conditional independence tests.

SGS and PC algorithm Spirtes, Glymour & Scheines (2000)

PC algorithm is most popular constraint-based method.

1. Start with a fully connected graph.
2. Reduce edges by conditional independence tests.

SGS Test all combinations and conditions $X_i \perp\!\!\!\perp X_j | S$.

SGS and PC algorithm Spirtes, Glymour & Scheines (2000)

PC algorithm is most popular constraint-based method.

1. Start with a fully connected graph.
2. Reduce edges by conditional independence tests.

SGS Test all combinations and conditions $X_i \perp\!\!\!\perp X_j | S$.

PC(a) Test $X_i \perp\!\!\!\perp X_j | \varnothing$.

SGS and PC algorithm Spirtes, Glymour & Scheines (2000)

PC algorithm is most popular constraint-based method.

1. Start with a fully connected graph.
2. Reduce edges by conditional independence tests.

SGS Test all combinations and conditions $X_i \perp\!\!\!\perp X_j | S$.

PC(a) Test $X_i \perp\!\!\!\perp X_j | \varnothing$.

(b) On remaining edges and connected components, test $X_i \perp\!\!\!\perp X_j | X_k$.

SGS and PC algorithm Spirtes, Glymour & Scheines (2000)

PC algorithm is most popular constraint-based method.

1. Start with a fully connected graph.
2. Reduce edges by conditional independence tests.

SGS Test all combinations and conditions $X_i \perp\!\!\!\perp X_j | S$.

PC(a) Test $X_i \perp\!\!\!\perp X_j | \varnothing$.

(b) On remaining edges and connected components, test $X_i \perp\!\!\!\perp X_j | X_k$.

(c) And so forth.

SGS and PC algorithm Spirtes, Glymour & Scheines (2000)

PC algorithm is most popular constraint-based method.

1. Start with a fully connected graph.
2. Reduce edges by conditional independence tests.

   SGS Test all combinations and conditions $X_i \perp\!\!\!\perp X_j | S$.

   PC(a) Test $X_i \perp\!\!\!\perp X_j | \varnothing$.

   (b) On remaining edges and connected components, test $X_i \perp\!\!\!\perp X_j | X_k$.

   (c) And so forth.
3. Orient edges, where possible: *colliders*.

SGS and PC algorithm Spirtes, Glymour & Scheines (2000)

PC algorithm is most popular constraint-based method.

1. Start with a fully connected graph.
2. Reduce edges by conditional independence tests.

   SGS Test all combinations and conditions $X_i \perp\!\!\!\perp X_j | S$.

   PC(a) Test $X_i \perp\!\!\!\perp X_j | \varnothing$.

   (b) On remaining edges and connected components, test $X_i \perp\!\!\!\perp X_j | X_k$.

   (c) And so forth.
3. Orient edges, where possible: *colliders*.

Greedy equivalence search Chickering (2002)

GES is most popular score-based method.

1. Start with an empty graph.

SGS and PC algorithm Spirtes, Glymour & Scheines (2000)

PC algorithm is most popular constraint-based method.

1. Start with a fully connected graph.
2. Reduce edges by conditional independence tests.

   SGS Test all combinations and conditions $X_i \perp\!\!\!\perp X_j | S$.

   PC(a) Test $X_i \perp\!\!\!\perp X_j | \varnothing$.

   (b) On remaining edges and connected components, test $X_i \perp\!\!\!\perp X_j | X_k$.

   (c) And so forth.
3. Orient edges, where possible: *colliders*.

Greedy equivalence search Chickering (2002)

GES is most popular score-based method.

1. Start with an empty graph.
2. Greedily add edges by computing a score, usually the likelihood.

Note: Faithfulness and Biological Networks

- A distribution is *faithful* to the graph $\mathcal{G}$, if there are no other independence relations than those encoded in the graph.
  - ▷ All variable couplings in the distribution lead to statistical association.

Note: Faithfulness and Biological Networks

- A distribution is *faithful* to the graph $\mathcal{G}$, if there are no other independence relations than those encoded in the graph.
  - ▷ All variable couplings in the distribution lead to statistical association.

One can easily construct distributions that do not show statistical associations between coupled variables. For example,

$$Y = (X_1 \wedge \overline{X}_2) \vee (\overline{X}_1 \wedge X_2), \quad X_1, X_2 \sim \mathrm{Ber}(0.5),$$

implies

$$Y \perp\!\!\!\perp X_1 \qquad Y \perp\!\!\!\perp X_2.$$

Note: Faithfulness and Biological Networks

- A distribution is *faithful* to the graph $\mathcal{G}$, if there are no other independence relations than those encoded in the graph.
  - ▷ All variable couplings in the distribution lead to statistical association.

One can easily construct distributions that do not show statistical associations between coupled variables. For example,

$$Y = (X_1 \wedge \overline{X}_2) \vee (\overline{X}_1 \wedge X_2), \quad X_1, X_2 \sim \mathrm{Ber}(0.5),$$

implies

$$Y \perp\!\!\!\perp X_1 \qquad Y \perp\!\!\!\perp X_2.$$

Then, only the interventional distribution shows association

$$Y = X_1 \mid \mathrm{do}(X_2 = 0), \quad X_1 \sim \mathrm{Ber}(0.5).$$

Note: Faithfulness and Biological Networks

- A distribution is *faithful* to the graph $\mathcal{G}$, if there are no other independence relations than those encoded in the graph.
  - ▷ All variable couplings in the distribution lead to statistical association.

One can easily construct distributions that do not show statistical associations between coupled variables. For example,

$$Y = (X_1 \wedge \overline{X}_2) \vee (\overline{X}_1 \wedge X_2), \quad X_1, X_2 \sim \mathrm{Ber}(0.5),$$

implies

$$Y \perp\!\!\!\perp X_1 \qquad Y \perp\!\!\!\perp X_2.$$

Then, only the interventional distribution shows association

$$Y = X_1 \mid \mathrm{do}(X_2 = 0), \quad X_1 \sim \mathrm{Ber}(0.5).$$

▷ Aside from **unmeasured confounders**, violated **faithfulness** poses the strongest limitation to causal conclusions in biology.

Time series data

Consider a d-dimensional time series X_{ti}, for example

$$X_{t1} = X_{(t-1)1} + N_{t1}$$

$$X_{t2} = X_{(t-1)2} + N_{t2}$$

$$X_{t3} = X_{(t-1)1} \wedge \overline{X}_{(t-1)2} + N_{t3}$$

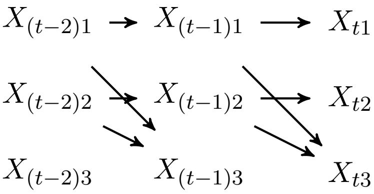
Time series data

Consider a d-dimensional time series X_{ti}, for example

$$X_{t1} = X_{(t-1)1} + N_{t1}$$

$$X_{t2} = X_{(t-1)2} + N_{t2}$$

$$X_{t3} = X_{(t-1)1} \wedge \overline{X}_{(t-1)2} + N_{t3}$$



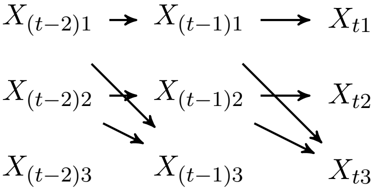
Time series data

Consider a d-dimensional time series X_{ti}, for example

$$X_{t1} = X_{(t-1)1} + N_{t1}$$

$$X_{t2} = X_{(t-1)2} + N_{t2}$$

$$X_{t3} = X_{(t-1)1} \wedge \overline{X}_{(t-1)2} + N_{t3}$$



- Time ordering **resolves directions** on the graph!
  - ▷ Here: $X_{t2} \perp\!\!\!\perp X_{(t-1)3} | X_{(t-1)2}$, but $X_{t3} \not\!\perp\!\!\!\perp X_{(t-1)2} | X_{(t-1)3}$.

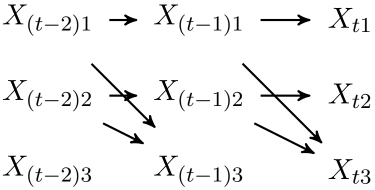
Time series data

Consider a d-dimensional time series X_{ti}, for example

$$X_{t1} = X_{(t-1)1} + N_{t1}$$

$$X_{t2} = X_{(t-1)2} + N_{t2}$$

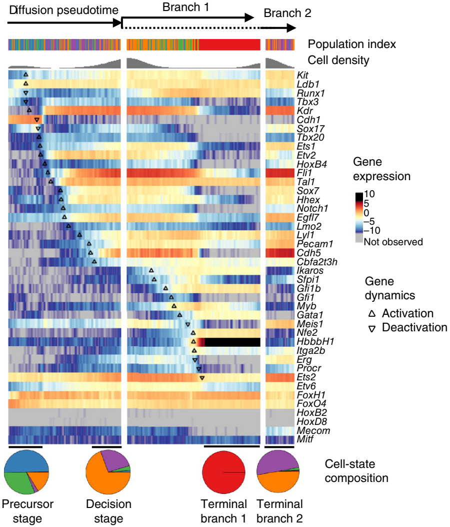
$$X_{t3} = X_{(t-1)1} \wedge \overline{X}_{(t-1)2} + N_{t3}$$



- Time ordering **resolves directions** on the graph!
  - ▷ Here: $X_{t2} \perp\!\!\!\perp X_{(t-1)3} | X_{(t-1)2}$, but $X_{t3} \not\!\perp\!\!\!\perp X_{(t-1)2} | X_{(t-1)3}$.

- **Granger Causality** and **Transfer Entropy** correspond to specific tests in the PC algorithm, but get the example above wrong.

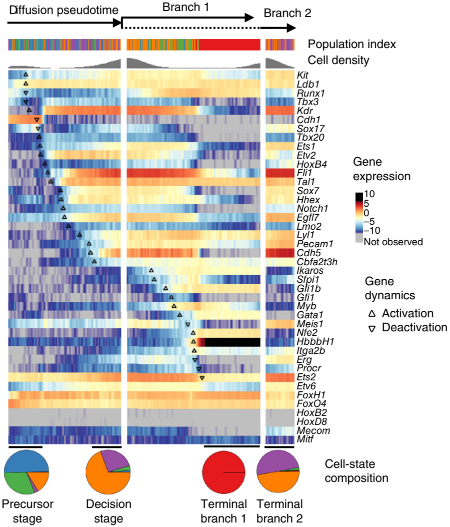
Inferring gene regulation from single-cell data



Haghverdi, Büttner, Wolf, Buettner & Theis, Nature Methods 13, 845 (2016)

Inferring gene regulation from single-cell data

Structure learning on gene expression pseudotime series is hard.

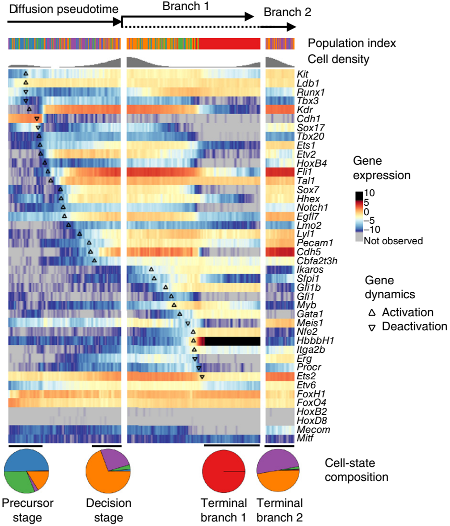


Haghverdi, Büttner, Wolf, Buettner & Theis, Nature Methods 13, 845 (2016)

Inferring gene regulation from single-cell data

Structure learning on gene expression pseudotime series is hard.

- Few dynamic noise. Relatively non-informative Hill kinetics.



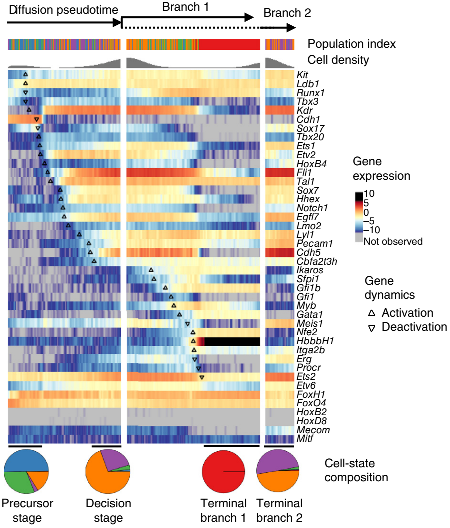
Haghverdi, Büttner, Wolf, Buettner & Theis, Nature Methods 13, 845 (2016)

Inferring gene regulation from single-cell data

Structure learning on gene expression pseudotime series is hard.

- Few dynamic noise. Relatively non-informative Hill kinetics.

▷ Use global geometric properties of the data.



Haghverdi, Büttner, Wolf, Buettner & Theis, Nature Methods 13, 845 (2016)

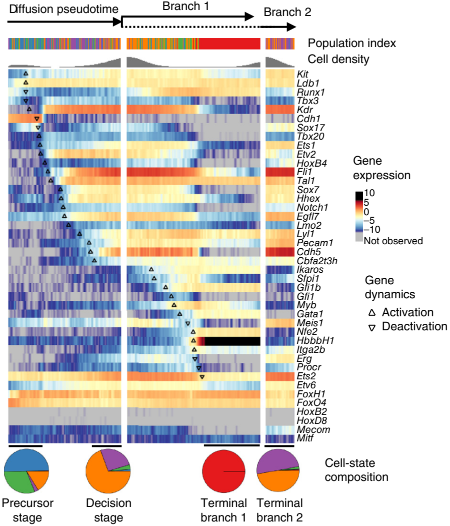
Inferring gene regulation from single-cell data

Structure learning on gene expression pseudotime series is hard.

- Few dynamic noise. Relatively non-informative Hill kinetics.

▷ Use global geometric properties of the data.

▷ Developed PC algorithm with tests of functional relations instead of statistical associations.



Haghverdi, Büttner, Wolf, Buettner & Theis, Nature Methods 13, 845 (2016)

Learning undirected Gaussian graphical models

- Learning the structure of undirected graphical models is easier than learning DAG structure because we don't need to worry about acyclicity.

Learning undirected Gaussian graphical models

- Learning the structure of undirected graphical models is easier than learning DAG structure because we don't need to worry about acyclicity.

- It is harder than learning DAG structure since the likelihood does not decompose, i.e. no greedy technique can be employed. Only in the Gaussian case, there is an immediate solution.

Learning undirected Gaussian graphical models

- Learning the structure of undirected graphical models is easier than learning DAG structure because we don't need to worry about acyclicity.

- It is harder than learning DAG structure since the likelihood does not decompose, i.e. no greedy technique can be employed. Only in the Gaussian case, there is an immediate solution.

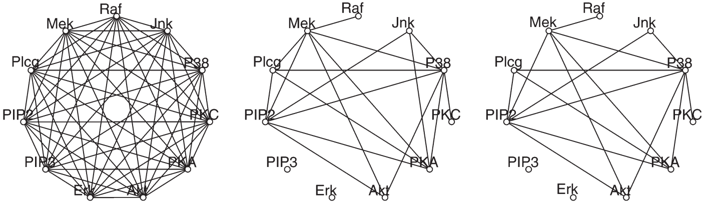
Graphical Lasso Friedman, Hastie & Tibshirani, Biostatistics 9, 432 (2008)

$$\mathrm{cost}(\mathbf{\Sigma}^{-1}) = \underbrace{-\log\det(\mathbf{\Sigma}^{-1}) + \mathrm{tr}(\mathbf{S}\mathbf{\Sigma})}_{-\text{loglikelihood}} + \underbrace{\lambda||\mathbf{\Sigma}^{-1}||_1}_{\text{sparsity prior}}$$

The precision matrix $\mathbf{\Sigma}^{-1}$ receives an L_1 prior.

▷ Limitations: Gaussian data. No causal interpretation.

Learning undirected Gaussian graphical models



L1 norm= 2.27182

L1 norm= 0.08915

L1 norm= 0.04251

data from Sachs, Perez, Pe'er, Lauffenburger & Nolan, Science 308, 523 (2005)

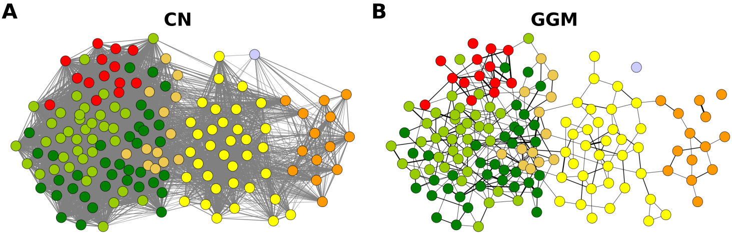
Graphical Lasso Friedman, Hastie & Tibshirani, Biostatistics 9, 432 (2008)

$$\mathrm{cost}(\mathbf{\Sigma}^{-1}) = \underbrace{-\log\det(\mathbf{\Sigma}^{-1}) + \mathrm{tr}(\mathbf{S}\mathbf{\Sigma})}_{-\text{loglikelihood}} + \underbrace{\lambda||\mathbf{\Sigma}^{-1}||_1}_{\text{sparsity prior}}$$

The precision matrix $\mathbf{\Sigma}^{-1}$ receives an L_1 prior.

▷ Limitations: Gaussian data. No causal interpretation.

Learning undirected Gaussian graphical models



Krumsiek, Suhre, Illig, Adamski & Theis, BMC Systems Biology 5, 21 (2011)

Graphical Lasso Friedman, Hastie & Tibshirani, Biostatistics 9, 432 (2008)

$$\mathrm{cost}(\mathbf{\Sigma}^{-1}) = \underbrace{-\log\det(\mathbf{\Sigma}^{-1}) + \mathrm{tr}(\mathbf{S}\mathbf{\Sigma})}_{-\text{loglikelihood}} + \underbrace{\lambda||\mathbf{\Sigma}^{-1}||_1}_{\text{sparsity prior}}$$

The precision matrix $\mathbf{\Sigma}^{-1}$ receives an L_1 prior.

▹ Limitations: Gaussian data. No causal interpretation.

Causal Inference

Causal Inference

There are two problems known as “causal inference”. Shalizi, Chap. 25 (2016)

Causal Inference

There are two problems known as “causal inference”. Shalizi, Chap. 25 (2016)

- Given data about a system, find its causal structure.

Causal Inference

There are two problems known as "causal inference". Shalizi, Chap. 25 (2016)

- Given data about a system, find its causal structure.
- Given the causal structure of a system, estimate effects variables have on each other.

Causal Inference

There are two problems known as “causal inference”. Shalizi, Chap. 25 (2016)

- Given data about a system, find its causal structure.
- Given the causal structure of a system, estimate effects variables have on each other.

We mostly talked about the first topic, because it's “more related to machine learning”.

Causal Inference

There are two problems known as “causal inference”. Shalizi, Chap. 25 (2016)

- Given data about a system, find its causal structure.
- Given the causal structure of a system, estimate effects variables have on each other.

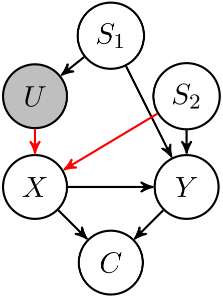
We mostly talked about the first topic, because it’s “more related to machine learning”.

Note: Very often, people estimate causal structure from subject knowledge.

Estimate effects variables have on each other

Backdoor criterion

How to compute a causal effect in this graph?

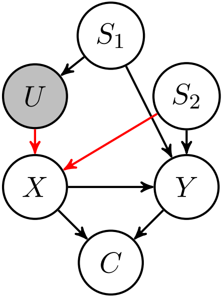


Estimate effects variables have on each other

Backdoor criterion

How to compute a causal effect in this graph? Block all causal pathways by conditioning on the right set of variables $S = \{S_1, S_2\}$.

$$p(Y|\mathrm{do}(X)) = \sum_s p(Y|X, S = s) p(S = s)$$



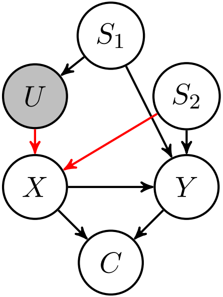
Estimate effects variables have on each other

Backdoor criterion

How to compute a causal effect in this graph? Block all causal pathways by conditioning on the right set of variables $S = \{S_1, S_2\}$.

$$p(Y|\mathrm{do}(X)) = \sum_s p(Y|X, S = s)p(S = s)$$

- ▷ Propensity scores.



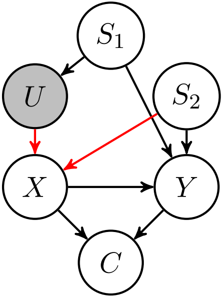
Estimate effects variables have on each other

Backdoor criterion

How to compute a causal effect in this graph? Block all causal pathways by conditioning on the right set of variables $S = \{S_1, S_2\}$.

$$p(Y|\mathrm{do}(X)) = \sum_s p(Y|X, S = s)p(S = s)$$

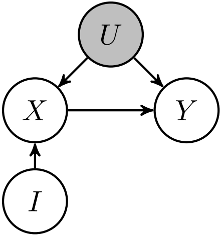
- ▷ Propensity scores.



Instrumental variables

You have no clue how to block all causal pathways, but you have some “external” way of varying X. Then

$$\beta = \frac{\mathrm{Cov}(I, Y)}{\mathrm{Cov}(I, X)}.$$



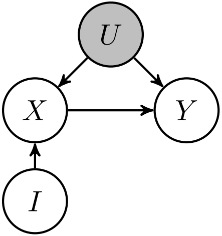
Estimate effects variables have on each other

- Randomization: I is coin toss that assigns treatment.

Instrumental variables

You have no clue how to block all causal pathways, but you have some "external" way of varying X. Then

$$\beta = \frac{\mathrm{Cov}(I, Y)}{\mathrm{Cov}(I, X)}.$$



Estimate effects variables have on each other

- Randomization: I is coin toss that assigns treatment.

- Mendelian randomization, e.g. to investigate causal effect of Gene Expression on Metabolite Level

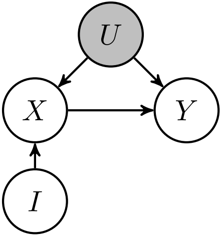
$$\beta = \frac{\mathrm{Cov}(\mathrm{SNP}, \mathrm{MetaboliteLevel})}{\mathrm{Cov}(\mathrm{SNP}, \mathrm{GeneExpression})}$$

Shin, Fauman, Petersen, Krumsiek & et al., Nature Genetics 46, 543 (2014)

Instrumental variables

You have no clue how to block all causal pathways, but you have some "external" way of varying X. Then

$$\beta = \frac{\mathrm{Cov}(I, Y)}{\mathrm{Cov}(I, X)}.$$



Summary

Directed graphical models can be used to “organize” causal reasoning.

Summary

Directed graphical models can be used to “organize” causal reasoning.

▷ Inference using constraint or score based methods.

Summary

Directed graphical models can be used to “organize” causal reasoning.

- Inference using constraint or score based methods.
- Time series data helps identifying causal directions.

Summary

Directed graphical models can be used to “organize” causal reasoning.

- ▷ Inference using constraint or score based methods.
- ▷ Time series data helps identifying causal directions.
- ▷ Have the potential to improve on inference of biological networks?

  Sachs, Perez, Pe’er, Lauffenburger & Nolan, Science 308, 523 (2005)

  Maathuis, Colombo, Kalisch & Bühlmann, Nature Methods 7, 247 (2010)

  Hill et al., Nature Methods 13, 310 (2016)

Summary

Directed graphical models can be used to “organize” causal reasoning.

▷ Inference using constraint or score based methods.

▷ Time series data helps identifying causal directions.

▷ Have the potential to improve on inference of biological networks?

Sachs, Perez, Pe’er, Lauffenburger & Nolan, Science 308, 523 (2005)

Maathuis, Colombo, Kalisch & Bühlmann, Nature Methods 7, 247 (2010)

Hill et al., Nature Methods 13, 310 (2016)

Thank you! Thanks to Fabian and all members of ICB-ML!

Transfer Entropy Schreiber (2000) and Granger Causality Granger (1969)

Transfer Entropy Schreiber (2000) and Granger Causality Granger (1969)

Consider a d-dimensional time series X_{ti}.

- Transfer Entropy is conditional mutual information

$$\begin{aligned}\mathrm{TE}_{i\to j} &= \mathrm{MI}_{X_{(t-1)i};X_{tj}|S} \\ &= H_{X_{tj}|S} - H_{X_{tj}|X_{(t-1)i},S}\end{aligned}$$

where originally, $S = X_{(t-1)j}$, and later $S = \{$all observed variables$\}$.

Transfer Entropy Schreiber (2000) and Granger Causality Granger (1969)

Consider a d-dimensional time series X_{ti}.

- Transfer Entropy is conditional mutual information

$$\begin{aligned} \mathrm{TE}_{i\to j} &= \mathrm{MI}_{X_{(t-1)i};X_{tj}|S} \\ &= H_{X_{tj}|S} - H_{X_{tj}|X_{(t-1)i},S} \end{aligned}$$

  where originally, $S = X_{(t-1)j}$, and later $S = \{$all observed variables$\}$.

- Granger Causality is "almost the same"

$$\mathrm{GC}_{i\to j} = \log(\Sigma_{X_{tj}|S}) - \log(\Sigma_{X_{tj}|X_{(t-1)i},S}),$$

  we just measure uncertainty by covariance instead of entropy. In the Gaussian case, GC is equivalent with TE. Barnett, Barrett & Seth, PRL 103, 238701 (2009)

Transfer Entropy Schreiber (2000) and Granger Causality Granger (1969)

Consider a d-dimensional time series X_{ti}.

- Transfer Entropy is conditional mutual information

$$\begin{aligned}\mathrm{TE}_{i\to j} &= \mathrm{MI}_{X_{(t-1)i};X_{tj}|S} \\ &= H_{X_{tj}|S} - H_{X_{tj}|X_{(t-1)i},S}\end{aligned}$$

  where originally, $S = X_{(t-1)j}$, and later $S = \{\text{all observed variables}\}$.

- Granger Causality is “almost the same”

$$\mathrm{GC}_{i\to j} = \log(\Sigma_{X_{tj}|S}) - \log(\Sigma_{X_{tj}|X_{(t-1)i},S}),$$

  we just measure uncertainty by covariance instead of entropy. In the Gaussian case, GC is equivalent with TE. Barnett, Barrett & Seth, PRL 103, 238701 (2009)

▷ Estimators for MI (in the Gaussian case, partial correlation) are popular for measuring conditional independence — their computation amounts to evaluating a single test in the PC algorithm.

Limitations of Transfer Entropy and Granger Causality

- Conditioning on all variables leads to a terrible *curse of dimensionality*.

Limitations of Transfer Entropy and Granger Causality

- Conditioning on all variables leads to a terrible *curse of dimensionality*.
- Say $X_1, X_2 \sim \mathrm{Ber}(0.5)$ describe the expression of two independent genes, and $X_3 = X_1 + X_2$ their sum. Then X_3 is a *collider* in the graph

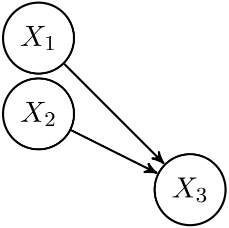
$$X_1 \not\perp\!\!\!\perp X_2 | X_3. \qquad \text{(compare “selection bias”)}$$

Limitations of Transfer Entropy and Granger Causality

- Conditioning on all variables leads to a terrible *curse of dimensionality*.
- Say $X_1, X_2 \sim \mathrm{Ber}(0.5)$ describe the expression of two independent genes, and $X_3 = X_1 + X_2$ their sum. Then X_3 is a *collider* in the graph

$$X_1 \not\perp\!\!\!\perp X_2 | X_3. \qquad \text{(compare ``selection bias'')}$$

- ▷ Granger Causality and Transfer Entropy yield an information flow $X_{(t-1)1} \to X_{t2}$. But it's non-causal, i.e. not helpful for prediction!

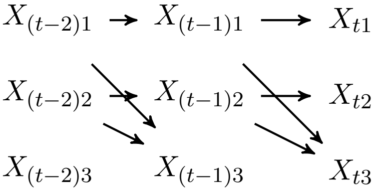


Limitations of Transfer Entropy and Granger Causality

- Conditioning on all variables leads to a terrible *curse of dimensionality*.
- Say $X_1, X_2 \sim \mathrm{Ber}(0.5)$ describe the expression of two independent genes, and $X_3 = X_1 + X_2$ their sum. Then X_3 is a *collider* in the graph

$$X_1 \not\perp\!\!\!\perp X_2 | X_3. \qquad \text{(compare “selection bias”)}$$

▷ Granger Causality and Transfer Entropy yield an information flow $X_{(t-1)1} \to X_{t2}$. But it's non-causal, i.e. not helpful for prediction!

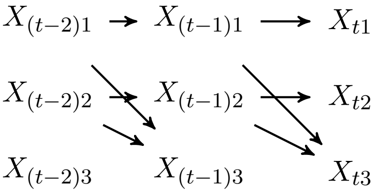


Limitations of Transfer Entropy and Granger Causality

- Conditioning on all variables leads to a terrible *curse of dimensionality*.
- Say $X_1, X_2 \sim \mathrm{Ber}(0.5)$ describe the expression of two independent genes, and $X_3 = X_1 + X_2$ their sum. Then X_3 is a *collider* in the graph

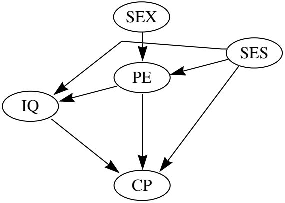
$$X_1 \not\perp\!\!\!\perp X_2 | X_3. \qquad \text{(compare "selection bias")}$$

  - ▷ Granger Causality and Transfer Entropy yield an information flow $X_{(t-1)1} \to X_{t2}$. But it's non-causal, i.e. not helpful for prediction!



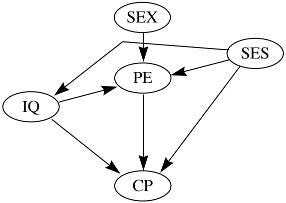
- General Note: Time Series data very helpful to resolve directions!

College admission example Heckerman, Meek & Cooper (1997)



$\log p(D \mid \mathbf{m}_1) \cong -45653$

$p(\mathbf{m}_1 \mid D) \cong 1.0$

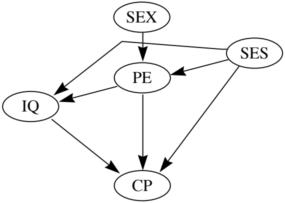


$\log p(D \mid \mathbf{m}_2) \cong -45699$

$p(\mathbf{m}_2 \mid D) \cong 1.2 \times 10^{-10}$

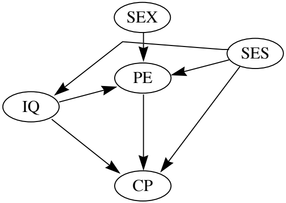
- PC algorithm chooses second most likely model! After it decides that SEX and IQ are marginally independent, it never considers the independence of SEX and IQ given PE.

College admission example Heckerman, Meek & Cooper (1997)



$\log p(D \mid \mathbf{m}_1) \cong -45653$

$p(\mathbf{m}_1 \mid D) \cong 1.0$

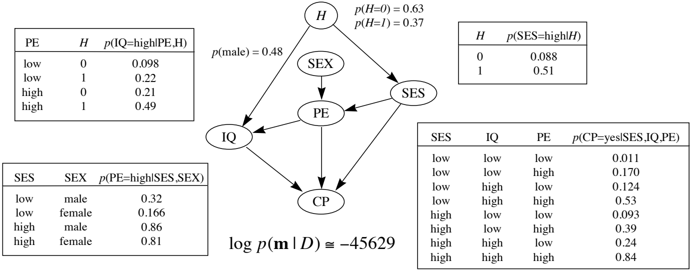


$\log p(D \mid \mathbf{m}_2) \cong -45699$

$p(\mathbf{m}_2 \mid D) \cong 1.2 \times 10^{-10}$

- PC algorithm chooses second most likely model! After it decides that SEX and IQ are marginally independent, it never considers the independence of SEX and IQ given PE.
- Most of the most likely model seems plausible in terms of a causal interpretation. The direct influence of SES on IQ though is likely to be due to a hidden common cause, e.g. IQ of parents.

College admission example Heckerman, Meek & Cooper (1997)



PE	*H*	*p*(IQ=high|PE,H)
low	0	0.098
low	1	0.22
high	0	0.21
high	1	0.49

H	*p*(SES=high|*H*)
0	0.088
1	0.51

SES	SEX	*p*(PE=high|SES,SEX)
low	male	0.32
low	female	0.166
high	male	0.86
high	female	0.81

SES	IQ	PE	*p*(CP=yes|SES,IQ,PE)
low	low	low	0.011
low	low	high	0.170
low	high	low	0.124
low	high	high	0.53
high	low	low	0.093
high	low	high	0.39
high	high	low	0.24
high	high	high	0.84

- PC algorithm chooses second most likely model! After it decides that SEX and IQ are marginally independent, it never considers the independence of SEX and IQ given PE.
- Most of the most likely model seems plausible in terms of a causal interpretation. The direct influence of SES on IQ though is likely to be due to a hidden common cause, e.g. IQ of parents.

Barnett, L., A. B. Barrett & A. K. Seth, 2009, Physical Review Letters **103**, 238701.

Chickering, D. M., 2002, The Journal of Machine Learning Research **2**, 445.

Friedman, J., T. Hastie & R. Tibshirani, 2008, Biostatistics **9**, 432.

Granger, C. W. J., 1969, Econometrica **37**, 424.

Haghverdi, L., M. Büttner, F. A. Wolf, F. Buettner & F. J. Theis, 2016, Nature Methods **13**, 845.

Heckerman, D., C. Meek & G. Cooper, 1997, Technical Report MSR-TR- 97-05, Microsoft Research .

Hill, S. M., L. M. Heiser, T. Cokelaer, M. Unger, N. K. Nesser, D. E. Carlin, Y. Zhang, A. Sokolov, E. O. Paull, C. K. Wong, K. Graim, A. Bivol et al., 2016, Nature Methods **13**, 310.

Krumsiek, J., K. Suhre, T. Illig, J. Adamski & F. J. Theis, 2011, BMC Syst. Biol. **5**, 21.

Maathuis, M. H., D. Colombo, M. Kalisch & P. Bühlmann, 2010, Nature Methods **7**, 247.

Pearl, J. & T. Verma, 1991, A Theory of Inferred Causation, in Principles of Knowledge Representation and Reasoning: Proceeding of the Second International Conference, pp. 441–452.

Ribeiro, M. T., S. Singh & C. Guestrin, 2016, 1602.04938.

Sachs, K., O. Perez, D. Pe'er, D. A. Lauffenburger & G. P. Nolan, 2005, Science **308**, 523.

Schreiber, T., 2000, Physical Review Letters **85**, 461.

Shalizi, C. R., 2016, Advanced Data Analysis from an Elementary Point of View (Cambridge University Press).

Shin, S.-Y., E. B. Fauman, A.-K. Petersen, J. Krumsiek & et al., 2014, Nature Genetics **46**, 543.

Spirtes, P., C. Glymour & R. Scheines, 2000, Causation, Prediction, and Search (MIT Press, Cambridge, MA, USA), 2nd edition.