

# Causal Gene Ranking

Daniel Stekhoven

NORDSTAT 2010 - 17.06.10



# Design of experiments

## Given

- a **response** (e.g. phenotype of interest);
- and a set of **observational** data;

## we want

- a **stable ranking** for the causal effects of the variables on the response;

## with the intention to

- offer powerful designs for selecting important variables for future **interventional** experiments.

## The data

### *Arabidopsis thaliana* (thale cress) gene expression data

- **observational** microarray data with  $n = 47$  and  $p = 21'326$ ;
- samples are from 35 *A. thaliana* **oeotypes** (D. Weigel, Tübingen);
- samples were hand-picked according to the response (L. Hennig, Grussem Lab, ETH Zurich).

### The response

- point of time, when first flower occurs;
- a robust measure for time: **number of leaves** grown.



# Ranking

according to the stability of causal effects

## method outline

- 1 **Subsampling** the data (as part of stability selection);
- 2 estimating a **complete partially directed acyclic graph** (CPDAG) using the PC algorithm;
- 3 estimating **lower bounds** for the causal effects of the variables using intervention calculus;
- 4 **repeat** step 1. - 3. many times (stability selection);
- 5 take summary of rankings as final result.

# Assumptions

We assume that...

- the distribution of  $(X_1, \dots, X_p, Y)$  is multivariate Normal. Moreover; it is **Markovian** and **faithful** to an (unknown) directed acyclic graph (DAG);
- $X_1, \dots, X_p$  have equal variance, this allows to compare the causal effects of the different variables.

# Complete partially directed acyclic graph

CPDAG

## Markov equivalence class

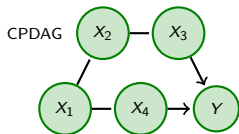
- the *true* DAG is typically **not** identifiable from observational data;
- but the **equivalence class** of a DAG is.

# Complete partially directed acyclic graph

CPDAG

## Markov equivalence class

- the *true* DAG is typically **not** identifiable from observational data;
- but the **equivalence class** of a DAG is.

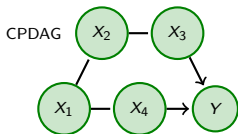


# Complete partially directed acyclic graph

CPDAG

## Markov equivalence class

- the *true* DAG is typically **not** identifiable from observational data;
- but the **equivalence class** of a DAG is.



## PC-algorithm

- based on low order partial correlations;
- computational feasible;
- asymptotically consistent;
- handles  $p \gg n$  problems;

if *true* DAG is **sparse**!

# Computing causal effects

given a DAG

The causal effect of  $X_i$  on  $Y$  is given by

$$\beta_{i|pa_i(G)} = \begin{cases} 0 & \text{if } Y \in pa_i(G), \\ \text{coef of } X_i \text{ in } Y \sim X_i + pa_i(G) & \text{if } Y \notin pa_i(G), \end{cases}$$

where  $pa_i(G)$  is the subset of parental nodes of  $X_i$  in the DAG  $G$  and  $Y \sim X_i + pa_i(G)$  is the linear regression of  $Y$  on  $X_i$  and  $pa_i(G)$ .

# Computing causal effects

given a DAG

effect of  $X_1$  on  $Y$

- in DAG  $G$  the parental set of  $X_1$  is

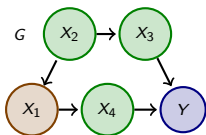
$$pa_1(G) = X_2;$$

- and  $Y \notin pa_1(G)$  thus the causal effect of  $X_1$  on  $Y$  is given by  $\beta_{1|X_2}$ , where

$$Y = \beta_{1|X_2} X_1 + \beta X_2 + \varepsilon.$$

But...

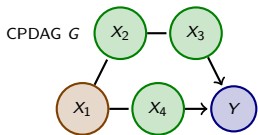
We do not have the true DAG - only a CPDAG.



## Lower bounds for causal effects

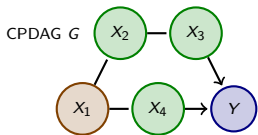
Example:

- 3 undirected edges;

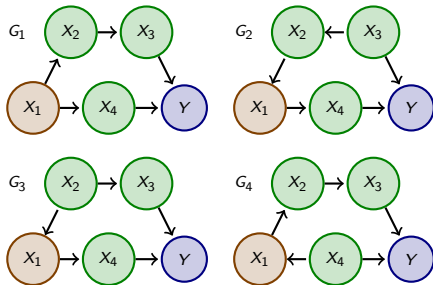


# Lower bounds for causal effects

## Example:

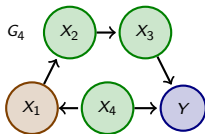
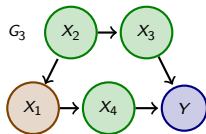
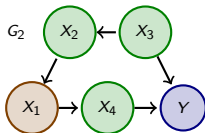
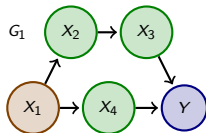
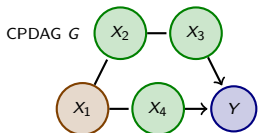


- 3 undirected edges;
- 8 possible graphs - but only 4 in the equivalence class;
  - no new v-structures;
  - no cycles.



# Lower bounds for causal effects

## Example:



- 3 undirected edges;
- 8 possible graphs - but only 4 in the equivalence class;
  - no new v-structures;
  - no cycles.
- compute causal effects for each DAG  $G_i$ ,  $i = 1, \dots, 4$

$$\{\beta_{1|\emptyset}, \beta_{1|X_2}, \beta_{1|X_3}, \beta_{1|X_4}\}$$

- the lower bound is then given by
- $$\min\{|\beta_{1|\emptyset}|, |\beta_{1|X_2}|, |\beta_{1|X_3}|, |\beta_{1|X_4}|\}.$$

# Lower bounds for causal effects

## proof of concept

M. Maathuis, D. Colombo, M. Kalisch & P. Bühlmann (2010).  
Predicting causal effects in large-scale systems from observational  
data. *Nature Methods* 7, 247 - 248

Yeast (*Saccharomyces cerevisiae*) single-gene deletion mutants.

# Stability selection

## procedure

- 1 draw a **subsample** of size  $\lfloor \frac{n}{2} \rfloor$  from the data;
- 2 estimate lower bounds for the causal effects;
- 3 repeat step 1 - 3, e.g. 100 times;
- 4 record the **relative selection frequencies**  $\pi$  of the **top  $q$**  variables.

N. Meinshausen & P. Bühlmann (2010). Stability Selection. *To appear as discussion paper in the Journal of the Royal Statistical Society, Series B*, arXiv: 0809.2932v2.

## Stability selection

### error bound

When applying stability selection, the following holds,

$$\mathbb{E}[V] \leq \frac{1}{2\pi_{thr} - 1} \frac{q^2}{p},$$

where  $\pi_{thr}$  the stability value at which a variable is considered stable and  $q$  the number of selected variables.

## Stability selection

### error bound

When applying stability selection, the following holds,

$$\mathbb{E}[V] \leq \frac{1}{2\pi_{thr} - 1} \frac{q^2}{p},$$

where  $\pi_{thr}$  the stability value at which a variable is considered stable and  $q$  the number of selected variables.

### therefore

- running stability selection we get the  $\pi$ 's;
- we can choose the  $q$ ;
- $p$  is given

$\Rightarrow$  an upper bound for the PFER of the stability for each variable.

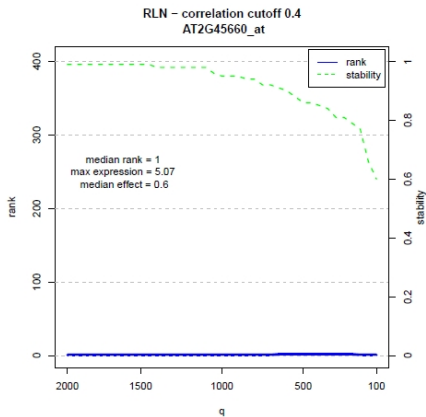
# Causal gene ranking

according to summary stability

# Causal gene ranking

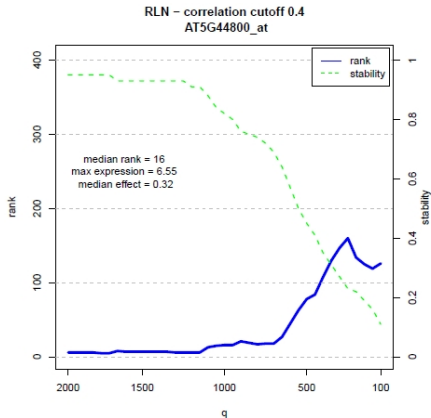
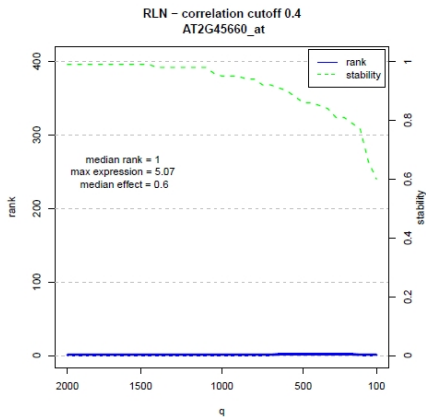
according to summary stability

## stable vs unstable variables



# Causal gene ranking according to summary stability

## stable vs unstable variables



# Causal gene ranking

	Gene	summary rank	median effect	expression	error (PCER)	name
1	AT2G45660	1	0.60	5.07	0.0017	AGL20 (SOC1)
2	AT4G24010	2	0.61	5.69	0.0021	ATCSLG1
3	AT1G15520	2	0.58	5.42	0.0017	PDR12
4	AT3G02920	5	0.58	7.44	0.0024	replication protein-related
5	AT5G43610	5	0.41	4.98	0.0101	ATSUC6
6	AT4G00650	7	0.48	5.56	0.0020	FRI
7	AT1G24070	8	0.57	6.13	0.0026	ATCSLA10
8	AT1G19940	9	0.53	5.13	0.0019	AtGH9B5
9	AT3G61170	9	0.51	5.12	0.0034	protein coding
10	AT1G32375	10	0.54	5.21	0.0031	protein coding
11	AT2G15320	10	0.50	5.57	0.0027	protein coding
12	AT2G28120	10	0.49	6.45	0.0026	protein coding
13	AT2G16510	13	0.50	10.7	0.0023	AVAP5
14	AT3G14630	13	0.48	4.87	0.0039	CYP72A9
15	AT1G11800	15	0.51	6.97	0.0028	protein coding
16	AT5G44800	16	0.32	6.55	0.0704	CHR4
17	AT3G50660	17	0.40	7.60	0.0059	DWF4
18	AT5G10140	19	0.30	10.3	0.0064	FLC
19	AT1G24110	20	0.49	4.66	0.0059	peroxidase, putative
20	AT1G27030	20	0.45	10.1	0.0059	unknown protein

- biological validation by gene knockout experiments in progress.

## Conclusion

### caveats

- a DAG is a very strong assumption, e.g. feedback loops;
- estimating a graph with almost 22'000 vertices remains difficult.

### however...

- a rough model, but yet still good (cf Maathuis et al., 2010);
- we score 3 of the 100 known important genes for flowering in the top 20's out of 22'000 genes in total, i.e.

$$\mathbb{P}[K = 3] \approx 10^{-6}.$$

# Outlook

in progress

- including **hidden variables** in causal graphs;
- perform real lab experiments with **mutant** plants.

# Acknowledgement

Thank you...








Prof Peter Bühlmann  
*Seminar for Statistics*



PD Dr. Lars Hennig  
*Institute for Plant Sciences*

**...and - thank you for Your attention.**

## References and further reading

-  M. Maathuis, D. Colombo, M. Kalisch & P. Bühlmann.  
*Predicting causal effects in large-scale systems from observational data.*  
Nature Methods 7, 247 - 248, 2010.
-  P. Spirtes, C. Glymour & R. Scheines.  
*Causation, Prediction, and Search.*  
Vol. 1, 2nd edn, The MIT Press, 2000.
-  M. Kalisch & P. Bühlmann.  
*Estimating high-dimensional directed acyclic graphs with the PC-algorithm.*  
Journal of Machine Learning Research 8, 2007.
-  M.H. Maathuis, M. Kalisch & P. Bühlmann.  
*Estimating high-dimensional intervention effects from observational data.*  
The Annals of Statistics, 2009.
-  N. Meinshausen & P. Bühlmann.  
*Stability Selection.*  
To appear, arXiv: 0809.2932v2, 2009.