

Abstract

Discovering structure in high-dimensional, observational data, as for example in microarray gene expression experiments, is an elaborate and crucial task. We introduce a method to stably infer the causal influence of predictor variables on a response. Combining the estimation of Markov equivalence classes of directed acyclic graphs using the PC-algorithm and causal intervention calculus for the effects of the variables on the response, and putting these two parts in a stability selection environment, we are not only able to rank variables according to their stable, causal-type influence, but also to assign the per-comparison error rate to each of them. Our causal inference method takes the cumulative nature of effects through a cascaded pathway into account. Furthermore, assigning ranks using stability makes the approach less prone to sampling variability and allows to choose the amount of regularization. We apply our method to real data from observational gene expression experiments of *Arabidopsis thaliana* with floral development as response of main interest.

Causal Graphs

A **graph** consists of

- a set of nodes N , e.g. variables;
- a set of edges E , e.g. relations of nodes.

A directed edge denotes the causal influence of a parent node onto its child node.

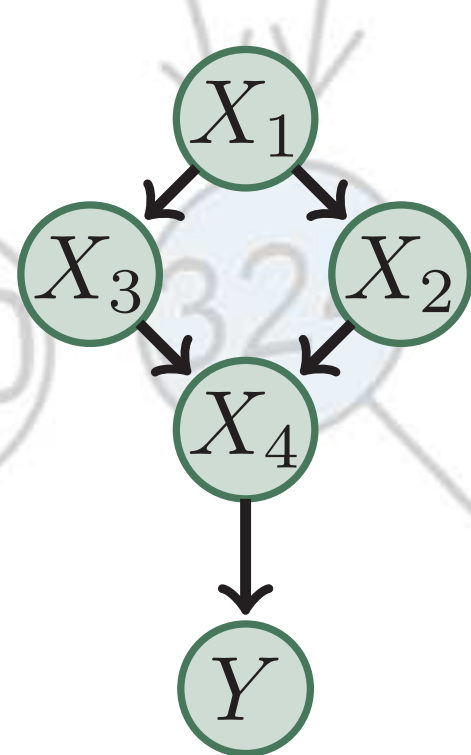


fig. 1

On the left an example for a directed causal graph is given:

- X_1 , the season of year;
- X_2 , is there precipitation;
- X_3 , is the sprinkler on;
- X_4 , is the floor wet;
- Y , is the floor slippery.

A **Directed Acyclic Graph** (DAG, see fig. 1) has

- no *undirected* edges;
- no directed *cycles*.

If not all edges can be directed, we have a **completed partially directed acyclic graph** (CPDAG, see fig. 2). A CPDAG describes a Markov equivalence class of DAGs.

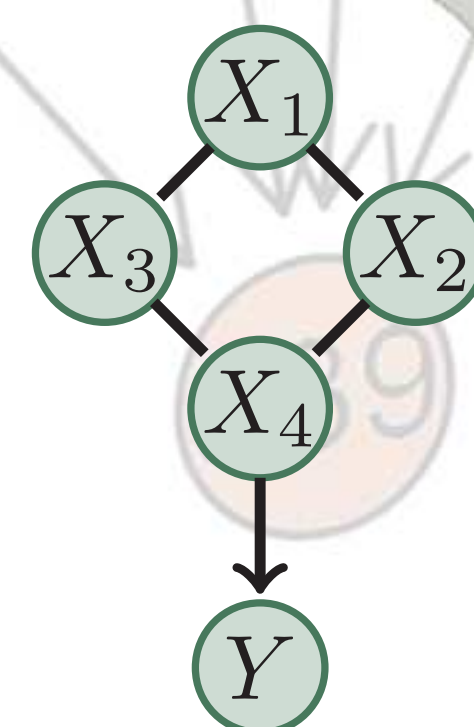


fig. 2

A CPDAG can be estimated from observational data using the PC-algorithm [Spirtes et al., 2000].

Data

We are investigating a data set from *Arabidopsis thaliana* with the following properties:

- 55 samples of observational microarray gene expression experiments;
- 21'325 genes measured per experiment;
- samples are from 35 different ecotypes (D. Weigel, MPI, Tübingen);
- sampled according to a strict protocol, therefore offering high homogeneity;
- response is the mean number of leaves of a genotype before the first blossom develops;
- variables standardized using robust multi-chip average (RMA).

Stability Selection

Since we have a high-dimensional setting, where $p \gg n$, variable selection is notoriously difficult. Stability selection [Meinshausen et al., 2009] offers finite sample control for error rates and hence a transparent way to choose a feasible amount of regularization.

It is performed in the following way:

1. draw a subsample of size $\lfloor \frac{n}{2} \rfloor$ from the data;
2. apply the method to score the genes;
3. repeat step 1 - 3 several times;
4. record the relative frequencies of the top q scoring genes.

Under certain assumptions the expected number of false positives V is bounded by:

$$\mathbb{E}(V) \leq \frac{1}{2\pi_{thr} - 1} \frac{q^2}{p}$$

Using this bound we can assign the **per-comparison error rate** (PCER, defined as $\mathbb{E}(V)/p$) to each gene and rank them accordingly.

Causal Effects

We estimate the causal effect of a given gene X_j on a response variable Y as the **regression coefficient** β_j of the linear regression

$$Y = \beta_j X_j + \beta_{pa(X_j)} pa(X_j) + \varepsilon$$

where $pa(X_j)$ is the set of all parent nodes of X_j and $\beta_{pa(X_j)}$ the corresponding coefficients.

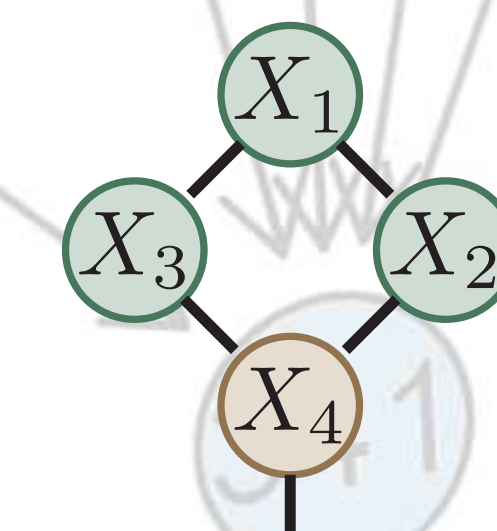


fig. 3

In fig. 3 the parental sets of X_4 are not uniquely defined:

- $pa_1(X_4) = \emptyset$;
- $pa_2(X_4) = \{X_2\}$;
- $pa_3(X_4) = \{X_3\}$ and;
- $pa_4(X_4) = \{X_2, X_3\}$.

Fit on each of these a regression

$$Y \sim X_4 + pa_i(X_4)$$

where $i \in \{1, \dots, 4\}$ the index of the parental sets. A **lower bound** for the causal effect of X_4 would then be

$$\min\{|\beta_{4,1}, \dots, \beta_{4,4}|\}$$

where $\beta_{j,i}$ is the regression coefficient of variable j in the i -th parental set [Maathuis et al., 2009].

Feasibility & parameter choice

Since from a computational point of view it is not feasible to use all genes at once when estimating the CPDAG, we perform a targeted approach using correlation learning, i.e. in each stability step we only consider a fixed number p_{feas} of highly correlated genes with respect to the response.

To estimate the CPDAG, only a single tuning parameter α is necessary, which describes the test le-

vel for the edges to be drawn in the PC algorithm (see [Kalisch et al., 2007]). We set α such that the number of edges to the response node are above the average neighbourhood size of the whole graph, i.e.

$$|adj(Y)| \geq \frac{|E|}{|N|}$$

where $adj(Y)$ is the set of adjacent nodes to Y .

Results

When using number of leaves as response together with the data described in the lower left box, subsample dimension fixed to $p_{feas} = 2000$ genes and a tuning parameter $\alpha = 0.1$, we top rank two genes directly involved in floral development;

- **SOC1**;
- **Flowering Locus A (FLA)**.

This favourable result illustrates the capability of the causal intervention approach to select important features out of a large set of available variables and respecting the cumulative increase of causal effects generated by extensive gene pathways. In the table to the right the top 20 genes are listed, sorted according to the PCER when the number of selected variables in each stability step is $q = 800$. One can observe that half of the genes belong to a certain gene family, but have unknown biological function. Validating these functional unknown genes in lab

experiments using mutant plants will be subject of future research.

Gene Locus	other name	Stability			PCER	Causal Intervention Effect				
		q = 'all'	rank q = 800	rank		average	rank	full sample	rank	
AT1G15520	PDR12	0.99	3	0.92	1	0.0017	0.59	3	0.50	51
AT1G19940	ATGH9B5	0.97	7	0.87	2	0.0019	0.56	5	0.42	169
AT2G45660	SOC1	1.00	1	0.97	2	0.0019	0.63	1	0.47	80
AT1G24070	CSLA10	0.93	20	0.83	4	0.0021	0.53	7	0.49	58
AT4G24010	CSLG1	0.98	6	0.83	4	0.0021	0.61	2	0.00	1999
AT2G15320	F27010.3	0.96	11	0.82	6	0.0022	0.51	14	0.17	1349
AT4G00650	FLA	1.00	1	0.80	7	0.0023	0.59	4	0.33	501
AT1G76640	F28016.1	0.96	11	0.77	8	0.0026	0.52	9	0.50	46
AT2G28120	F24013.9	0.95	14	0.75	9	0.0028	0.51	13	0.36	346
AT2G16510	F1P15.11	0.95	14	0.74	10	0.0029	0.50	15	0.49	60
AT3G02920	RPA32B	0.89	33	0.74	10	0.0029	0.56	6	0.43	129
AT1G32375	-	0.89	33	0.73	12	0.0031	0.52	11	0.52	41
AT1G11800	F25C20.3	0.92	23	0.72	13	0.0032	0.50	17	0.30	726
AT1G30120	PDH-E1 BETA	0.89	33	0.72	13	0.0032	0.49	18	0.45	109
AT1G33070	T9L6.11	0.88	44	0.72	13	0.0032	0.47	25	0.35	450
AT3G44440	F14L2.2	0.91	26	0.72	13	0.0032	0.53	8	0.48	65
AT1G08720	EDR1	0.89	33	0.71	17	0.0034	0.47	26	0.14	1486
AT2G27350	F12K2.7	0.85	68	0.71	17	0.0034	0.52	10	0.56	28
AT2G14560	LURP1	0.84	81	0.70	19	0.0035	0.50	16	0.39	257
AT3G09160	F3L24.2	0.94	17	0.70	19	0.0035	0.52	12	0.38	270

Table: 20 top ranked genes, values and ranks for stability with $q = 800$ and 2'000, PCER and values and ranks for causal intervention effects averaged over the 100 subsamples and for the whole (targeted) sample.

In summary our method finds genes which are very **stable** and have agreeably small PCER. It offers a tool for experimental biologists to choose candidate genes for future gene expression experiments in a model-based way using readily available observational data.

References

- [Spirtes et al., 2000] P. Spirtes, C. Glymour & R. Scheines. *Causation, Prediction, and Search*. Vol. 1, 2nd edn, The MIT Press, 2000.
- [Kalisch et al., 2007] M. Kalisch & P. Bühlmann. *Estimating high-dimensional directed acyclic graphs with the PC-algorithm*. Journal of Machine Learning Research 8, 2007.
- [Maathuis et al., 2009] M.H. Maathuis, M. Kalisch & P. Bühlmann. *Estimating high-dimensional intervention effects from observational data*. The Annals of Statistics, 2009.
- [Maathuis et al., 2010] M.H. Maathuis, D. Colombo, M. Kalisch & P. Bühlmann. *Predicting causal effects in large-scale systems from observational data*. Nature Methods, to appear, 2010.
- [Meinshausen et al., 2009] N. Meinshausen & P. Bühlmann. *Stability Selection*. Journal of the Royal Statistical Society, Series B (Discussion paper), 2010.

