Methods for causal inference from gene perturbation experiments and validation

Meinshausen, N.; Hauser, A.; Mooij, J.M.; Peters, J.; Versteeg, P.; Bühlmann, P.

DOI
10.1073/pnas.1510493113

Publication date
2016

Document Version
Other version

Published in
Proceedings of the National Academy of Sciences of the United States of America

Citation for published version (APA):
SI Appendix  (for “Methods for causal inference from gene perturbation experiments and validation”)  

Nicolai Meinshausen 1, Alain Hauser 1, Joris M. Mooij 1, Jonas Peters 3, Philip Versteeg 1, and Peter Bühlmann 1

1Seminar for Statistics, ETH Zurich, CH-8092 Zurich. 2Department of Engineering and Information Technology, Bern University of Applied Sciences, CH-3400 Burgdorf. 3Informatics Institute, University of Amsterdam, 1090 GH Amsterdam, The Netherlands, and 4Max Planck Institute for Intelligent Systems, D-72076 Tübingen

Details on the method of invariant causal prediction (ICP).  
We outline here the main idea of the ICP method which estimates a set of variables \( S(\mathcal{E}) \) satisfying formula [5]. Before doing so, we emphasize the underlying main assumptions: 

(A1) the Invariance Assumption as in the section on “Causal inference based on invariant across experiments”;  

(A2) construction of a statistical test for the null-hypothesis \([S.3]\) below which controls the type I error for any set \( S \subseteq \{1, \ldots, p\} \).

The Invariance Assumption in (A1) is a requirement regarding the space of experimental settings \( \mathcal{E} \). It holds under the following condition.  

(B) The space of experimental settings \( \mathcal{E} \) is such that the experimental conditions \( e \in \mathcal{E} \) do not affect the structural equation for \( Y \) in formula [4]. For example, \( \mathcal{E} \) consists of do-interventions at variables \( X_j \) for some \( j \in \{1, \ldots, p\} \) (but there is no do-intervention at variable \( Y \)).

The ICP method. We describe here the method in 3 steps.  

Identifiability for population version. As a starting point to explain the ICP method, we note that there is an identifiability issue: there might be many sets \( S^* \) and corresponding parameters and error distributions which fulfill the Invariance Assumption (and such other sets, say \( S \), fulfill the hypothesis \( H_{0, S}(\mathcal{E}) \) in \([S.3]\) below). Therefore, as quantity which is identifiable from the probability distribution generating the data, we consider

\[
S(\mathcal{E}) = \bigcap \{S; H_{0, S}(\mathcal{E}) \text{ holds}\},
\]

\([S.1]\)

where \( H_{0, S}(\mathcal{E}) \) is defined in \([S.3]\) below. Under assumption (A1) we then have that

\[
S^* = \text{pa}(Y) \supseteq S(\mathcal{E}),
\]

\([S.2]\)

because (A1) says that \( S^* = \text{pa}(Y) \) fulfills the Invariance assumption and hence \( H_{0, S^*}(\mathcal{E}) \) holds.

The hypothesis for the statistical test mentioned in (A2) is as follows:

\[
H_{0, S, \gamma}(\mathcal{E}) : \quad \gamma_k = 0 \text{ if } k \notin S \text{ and } \exists F_e \text{ such that } \forall e \in \mathcal{E},
\]

\[
Y^e = X^e \gamma + \varepsilon^e, \quad \varepsilon^e \perp X^e_S, \quad \varepsilon^e \sim F_e,
\]

where \( F_e \) is the same for all \( e \) and “\( \perp \)” denotes independence. That is, the null-hypothesis \( H_{0, S, \gamma}(\mathcal{E}) \) is a scenario where a particular set of variables indexed by \( S \subseteq \{1, \ldots, p\} \) and a particular regression vector \( \gamma \) satisfy the Invariance Assumption above. For that reason, we say that \( S, \gamma \) are “plausible causal variables/predictors and coefficients”. We relax the parameter \( \gamma \) in \( H_{0, S, \gamma}(\mathcal{E}) \) by considering

\[
H_{0, S}(\mathcal{E}) : \text{there exists } \gamma \text{ such that } H_{0, S, \gamma}(\mathcal{E}) \text{ holds}. \]

\([S.3]\)

Statistical testing for finite sample data. The null-hypothesis in \([S.3]\) can be statistically tested by using a test which incorporates constancy of the regression parameter of \( Y^e \) against \( X^e_S \) across all \( e \in \mathcal{E} \), and of constancy of the corresponding residual variances across \( e \in \mathcal{E} \), see [7].

The ICP method then proceeds by using the empirical version of \([S.1]\). It considers the intersection of all sets of plausible variables:

\[
\hat{S}(\mathcal{E}) = \bigcap \{S; H_{0, S}(\mathcal{E}) \text{ is not rejected by the statistical test at significance level } \alpha\}[S.4]
\]

Computational short-cut. The construction in \([S.4]\) shows that in principle, we would have to go through all possible subsets \( S \subseteq \{1, \ldots, p\} \). This would become very quickly computationally infeasible.

The strategy is to start testing \( H_{0, S}(\mathcal{E}) \) with small subsets \( S \) and subsequently move to larger subsets if all previous small subsets lead to rejection of the corresponding \( H_{0, S}(\mathcal{E}) \). If two disjoint subsets are not rejected, we can stop the search since then the estimate \( \hat{S}(\mathcal{E}) \) would be empty. The same holds if the empty set is not rejected.\(^2\) Such a strategy often markedly improves the computational speed.

Furthermore, an additional effective way to improve computational speed is given by estimating first the set of variables with non-zero regression parameters in a linear model of \( Y \) versus \( X \) in the pooled data among all environments, denoted by \( S_{\text{regr}} \). Assuming faithfulness we obtain a screening property: the variables corresponding to \( S_{\text{regr}} \supseteq S^* = \text{pa}(Y) \) must be a superset of the causal variables in \( S^* = \text{pa}(Y) \) [9]. In practice with finite sample data, we use an estimator \( \hat{S}_{\text{regr}} \) containing the non-zero estimated regression coefficients from the Lasso [10], see e.g. [1] for sufficient conditions ensuring that the Lasso estimator satisfies the variable screening property.\(^3\) Based on the set of variables \( \hat{S}_{\text{regr}} \), we compute \( \hat{S}(\mathcal{E}) \) in \([S.4]\) as an intersection of subsets of \( \hat{S}_{\text{regr}} \) only. This, combined with the strategy mentioned above leads to substantial gains in computational efficiency.

We summarize the ICP method as follows.

0. (optional pre-screening if \( p \) is large) Consider the pooled data \((Y, X) = \{(Y^e, X^e); e \in \mathcal{E}\}\). Compute a Lasso regression of \( Y \) versus \( X \) and denote the set of selected variables with non-zero Lasso-estimated regression coefficients as \( \hat{S}_{\text{regr}} \).

For the following steps, work with the reduced set of predictor variables from \( \hat{S}_{\text{regr}} \) only (and denote it as the original data as \( X^e (e \in \mathcal{E})\)).

\(^1\)Non-constancy of the error variances implies that \( H_{0, S}(\mathcal{E}) \) must be false, and we only need to control the probability of a false rejection. This doesn’t imply that the error distributions are characterized by the error variances only.

\(^2\)For a detailed description, see the summary of the ICP method below.

\(^3\)The estimator \( \hat{S}_{\text{regr}} \) based on the Lasso satisfies the asymptotic variable screening property if \( P[S_{\text{regr}} \supseteq \hat{S}_{\text{regr}}] \) converges to 1 as sample size and possibly the dimension increases.
Perform statistical tests for $H_{0,S}(\mathcal{E})$ in [S.3] at significance level $\alpha$ (e.g. $\alpha = 0.05$) (by testing constancy of the regression parameters and of the residual variances; see [7] for details).

Compute $\hat{S}(\mathcal{E})$ in [S.4]: start by testing small sets $S$ and subsequently move to larger sets $S$ as follows:

(i) If $H_{0,S}$ is not rejected: set $\hat{S}(\mathcal{E}) = \emptyset$ and stop; otherwise go to step (ii).

(ii) Consider sets $S$ of cardinality 1 and consider consecutively the corresponding intersections of the sets where $H_{0,S}(\mathcal{E})$ is not rejected; as soon as the intersection becomes $\emptyset$, output $\hat{S}(\mathcal{E}) = \emptyset$ and stop; otherwise, denote by $\hat{S}_1(\mathcal{E})$ the current intersection, set $m = 1$ and go to the next step.

(iii) Discard all supersets of $\hat{S}_m(\mathcal{E})$ and consider among the remaining sets the ones with cardinality $m + 1$. Construct consecutively the corresponding intersections of $\hat{S}_m(\mathcal{E})$ with the sets where $H_{0,S}(\mathcal{E})$ is not rejected; as soon as the intersection becomes $\emptyset$, output $\hat{S}(\mathcal{E}) = \emptyset$ and stop; otherwise, denote by $\hat{S}_{m+1}(\mathcal{E})$ the current intersection and increase $m$ by one ($m \leftarrow m + 1$).

(iv) Repeat step (iii) until there are no further sets to consider anymore.

Theorem 1. [7] Consider a linear model for each experimental setting and assume (A1)-(A2). Then, the estimated set $\hat{S}(\mathcal{E})$ from the ICP method (as described in step 1 above) satisfies

$$\mathbb{P}[\hat{S}(\mathcal{E}) \subseteq S^* = \text{pa}(Y)] \geq 1 - \alpha,$$

where $S^*$ denotes the causal parent set as defined in [S.4]. Taking the union over the results from pairs of environments as in [S.5] allows to exclude pairs of environments where an intervention on $Y$ occurred. For some interventions the location of the intervention is precisely known (so-called abundance interventions). We can thus remove pairs of environments from the union in [S.5] for which an abundance intervention on the target $Y$ occurs. However, excluding these pairs of environments has in general no effect on the estimated graph. The reason is that if interventions occur on $Y$ in environments $\mathcal{E}_{ij}$, no set of variables is invariant any longer and $S(\mathcal{E}_{ij}) = \emptyset$, unless the intervention is (perhaps accidentally) fine-tuned to lead to invariance in a set other than the parental set $\text{pa}(Y)$. For the data in [8] the graph remains unchanged whether we remove pairs of environments where an abundance intervention occurred or not.

There are also interventions of “activity-type” [6] in the data for which the target of the intervention is unknown. The precise location of these interventions are in general not known a-priori, see [6] for details. Leaving pairs of environments $\mathcal{E}_{ij}$ with abundance interventions on $Y$ out of the union in [S.5] made no difference to the estimated graph, as expected. By the same reasoning, we do not reduce the set of pairs of environments in [S.5] based on the activity interventions, as interventions on $Y$ in a given pair of environments $\mathcal{E}_{ij}$ will just yield an empty set $\hat{S}(\mathcal{E}_{ij})$ for the estimated set of causal parent of $Y$ and leave the union in [S.5] unchanged.

It is still essential, however, to split the environments into pairs (as in [S.5]) in the presence of (unknown) interventions.
Table S1: Direct causal relationships between the biochemical agents in the flow cytometry data of [8], according to different causal discovery methods. The consensus network according to [8] is denoted here by “[8]a” and their reconstructed network by “[8]b”.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RAF→MEK</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>MEK→RAF</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>MEK→ERK</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PLCg→PIP2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PLCg→PIP3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PLCg→PKC</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PIP2→PLCg</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PIP2→PIP3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PIP2→PKC</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PIP3→PLCg</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PIP3→PIP2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PIP3→PIP3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PIP3→PKC</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PIP3→AKT</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>AKT→ERK</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ERK→AKT</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ERK→PKA</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PKA→RAF</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PKA→MEK</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PKA→ERK</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PKA→AKT</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PKA→PKC</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PKA→P38</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PKA→JNK</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PKC→RAF</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PKC→MEK</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PKC→PLCg</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PKC→PIP2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PKC→AKT</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PKC→PKA</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PKC→P38</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PKC→JNK</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>P38→JNK</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>JNK→PKC</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>JNK→P38</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
on $Y$ in some environments, as the parental set is then not invariant across all environments simultaneously. We would thus have no power to detect causal effects when applying [S.4] over all environments simultaneously without splitting it into a union over pairs as in [S.5].

Results on Sachs et al. data [8]. Table S1 shows the direct causal relations reported in the literature and the ones found by our invariant causal prediction methods. Note that the actual ground truth for these data is not known. Nevertheless, there is quite some overlap the different results. In particular, only 34 out of 110 possible edges have been reported.

Using several datasets. We add here a remark that it is sometimes feasible to have access to different datasets of the same kind of gene perturbations. For example, the datasets in [5] and [4] have an overlap of 5225 genes (6170 and 5361 in [5] and [4], respectively) whose expressions were measured. Due to the fact that the invariant causal prediction method provides confidence statement for causal variables, we can aggregate the results from different datasets with the methods from meta analysis [3]. To be more precise, denote by $P^{(1)}_j, \ldots, P^{(k)}_j$ the p-values that variable $X_j$ is causal for a response $Y$ in the datasets $r = 1, \ldots, k$, i.e., $P^{(r)}_i$ is the smallest level such that $j \in \hat{S}(E)$ in dataset $r$ (see also formula [5]). Assuming independence among the $k$ different datasets, we can then use Stouffer’s method [11]: the aggregated p-value is

$$P^{aggr}_j = \Phi \left( \frac{\sum_{r=1}^{k} w_r \Phi^{-1} (P^{(r)}_j)}{\sqrt{\sum_{r=1}^{k} w_r^2}} \right),$$

with large positive weight $w_r$ if the dataset $r$ has large sample size $n_r$ or is of high quality. Typically we would choose $w_r = n_r/\sigma_r^2$ where $\sigma_r^2$ denotes the variance of the noise level in the $r$th dataset: it is often hard though to have knowledge about $\sigma_r^2$ and one might then simply use $\sigma^2 \equiv 1$ for all $r$. We note that this aggregation method, for any choice of non-negative weights, controls the probability for a type I error of claiming a false positive whenever the p-values $P^{(r)}_j$ are valid or conservative as in formula [5]. In cases where different datasets would describe different environments or interventions, we could in principle combine them into our framework from the Invariance Assumption without meta analysis aggregation, leading to potentially improved identifiability as indicated in formula [6]. But then, the issue of standardization to the same scale of the error variance across datasets needs to be addressed.
