SI Appendix (for "Methods for causal inference from gene perturbation experiments and validation")

Nicolai Meinshausen [*], **Alain Hauser** [†], **Joris M. Mooij** [‡], **Jonas Peters** [§], **Philip Versteeg** [‡], **and Peter Bühlmann** [*]

[*] Seminar for Statistics, ETH Zurich, CH-8092 Zurich, [†] Department of Engineering and Information Technology, Bern University of Applied Sciences, CH-3400 Burgdorf, [‡] Informatics Institute, University of Amsterdam, 1090 GH Amsterdam, The Netherlands, and [§] Max 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 $\hat{S}(\mathcal{E})$ 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 invariance 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

$$S(\mathcal{E}) = \bigcap \{S;\ H_{0,S}(\mathcal{E}) \text{ holds}\}, \qquad \textbf{[S.1]}$$

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

$$S^* = \mathrm{pa}(Y) \supseteq S(\mathcal{E}), \qquad \textbf{[S.2]}$$

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

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

$$\begin{aligned} H_{0,S,\gamma}(\mathcal{E}): \qquad & \gamma_k = 0 \text{ if } k \notin S \text{ and } \exists F_\varepsilon \text{ such that } \forall\, e \in \mathcal{E}, \\ & Y^e = X^e\gamma + \varepsilon^e,\ \varepsilon^e \perp X^e_S,\ \varepsilon^e \sim F_\varepsilon, \end{aligned}$$

where F_ε 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 γ satisfy the Invariance Assumption above. For that reason, we say that S, γ are "plausible causal variables/predictors and coefficients". We relax the parameter γ in $H_{0,S,\gamma}(\mathcal{E})$ by considering

$$H_{0,S}(\mathcal{E}): \text{ there exists } \gamma \text{ such that } H_{0,S,\gamma}(\mathcal{E}) \text{ holds.} \qquad \textbf{[S.3]}$$

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].[1]

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

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

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_{regr}. Assuming faithfulness we obtain a screening property: the variables corresponding to $S_{\mathrm{regr}} \supseteq S^* = \mathrm{pa}(Y)$ must be a superset of the causal variables in $S^* = \mathrm{pa}(Y)$ [9]. In practice with finite sample data, we use an estimator $\hat{S}_{\mathrm{regr}}$ 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 $\hat{S}_{\mathrm{regr}}$, we compute $\hat{S}(\mathcal{E})$ in **[S.4]** as an intersection of subsets of $\hat{S}_{\mathrm{regr}}$ 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 $\hat{S}_{\mathrm{regr}}$.
   For the following steps, work with the reduced set of predictor variables from $\hat{S}_{\mathrm{regr}}$ 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 $\hat{S}_{\mathrm{regr}}$ based on the Lasso satisfies the asymptotic variable screening property if $\mathbb{P}[\hat{S}_{\mathrm{regr}} \supseteq S_{\mathrm{regr}}]$ converges to 1 as sample size and possibly the dimension increases.

1. Perform statistical tests for $H_{0,S}(\mathcal{E})$ in **[S.3]** at significance level α (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 $H_{0,\emptyset}$ 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*

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

Proof: We have that

$$\begin{aligned}\mathbb{P}[\hat{S}(\mathcal{E}) \subseteq S^*] &= \mathbb{P}[\cap_S \{H_{0,S}(\mathcal{E}) \text{ not rejected}\} \subseteq S^*] \\ &\geq \mathbb{P}[H_{0,S^*}(\mathcal{E}) \text{ not rejected}] \geq 1 - \alpha,\end{aligned}$$

where the first inequality is due to the fact that an intersection of sets is smaller than each set alone, and the last inequality follows by the fact that H_{0,S^*} holds (by (A1)) and that the test controls the type I error (by (A2)) and applying this fact to the complement of a type I error. □

The statement of the theorem also holds for the ICP method with pre-screening (step 0 above) if it satisfies $\hat{S}_{\mathrm{regr}} \supseteq S^*$. Such a property holds with high probability, as discussed above in the paragraph "Computational short-cut". In general, we do not have that $\mathbb{P}[\hat{S}(\mathcal{E}) = S^* = \mathrm{pa}(Y)]$ is large, saying that we can identify the causal variables with high probability. On the other hand, and important in practice, the procedure does not require to specify which of the causal variables are identifiable from the data generating distributions from the experimental settings $\mathcal{E}$.

Violation of linearity. If the data generating probability distribution would violate the linearity assumption in formula [3] (i.e., the model would be misspecified), the ICP method (based on the linearity assumption in formula [3]) is expected to yield less findings rather than producing false positives and thus, it will be conservative [7, Sec.6]. In order to have better power, one would need to extend the ICP procedure using nonlinear regression models.

Inferring a strong intervention effect (SIE) by predicted direct intervention effects. We argue here that a strong intervention effect (SIE), which is a total causal effect, can be inferred from predicted direct intervention effects as e.g. from the output of the invariant causal prediction (ICP) method.

Consider for simplicity a structural equation for Y which is linear as in formula [4], $Y \leftarrow \sum_{j \in \mathrm{pa}(Y)} \gamma_j^* X_j + \varepsilon_Y$. Leaving identifiability issues aside, we expect that with increasing sample size, those variables with a large coefficient $|\gamma_j^*|$ will be selected first by ICP. If the direct effect of variable X_j is strong, meaning that $|\gamma_j^*|$ is large, we should see a fairly strong total average effect $|\frac{d}{dx}\mathbb{E}[Y|\mathrm{do}(X_j = x)]|$; unless in the "unlikely case" (i.e., with a nearly non-faithful distribution) where the direct effect would be (nearly) canceled by indirect effects corresponding to directed paths from X_j to Y other than $X_j \to Y$ in the causal directed acyclic graph.

Therefore, the indices j (corresponding to variable X_j) with large estimates (from a method or algorithm) for $|\gamma_j^*|$ typically correspond to the indices (or variables) with large values of $|\frac{d}{dx}\mathbb{E}[Y|\mathrm{do}(X_j = x)]|$. Finally, variables X_j with a strong total causal effect are more likely to result in SIEs.

Adaptation of the ICP-method to Sachs et al. data [8]. The estimation of the causal graph works node-wise, that is we take each variable in turn as a target variable Y and treat all other variables as potentially causal predictors for Y. When applying the ICP method **[S.4]** to a given target variable Y, one needs to exclude interventions on the target itself. A straightforward adaption of the ICP-method in **[S.4]** is as follows. We have eight different environments which we denote for simplicity by $\mathcal{E} = \{1, \ldots, 8\}$. Let $\mathcal{E}_{ij} = \{i, j\}$ be the subset of two environments $\{i, j\} \in \mathcal{E}$. Let $\hat{S}_\alpha(\mathcal{E}_{ij})$ be the estimated causal parent set when using **[S.4]** for these two environments when controlling the error at level α. Let $\mathcal{P}$ be the set of all pairs of environments $\mathcal{E}_{ij}$ with $i < j$ where no intervention on the target Y occurs in environments $\mathcal{E}_{ij}$. A straightforward generalization of **[S.4]** is to estimate the causal parent set as

$$\tilde{S}(\mathcal{E}) = \bigcup_{\mathcal{E}_{ij} \in \mathcal{P}} \hat{S}_{\alpha/|\mathcal{P}|}(\mathcal{E}_{ij}), \qquad \textbf{[S.5]}$$

where $\hat{S}$ is 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 $\hat{S}(\mathcal{E}_{ij}) = \emptyset$, unless the intervention are (perhaps accidentally) fine-tuned to lead to invariance in a set other than the parental set $\mathrm{pa}(Y)$. For the data in [8] the graph remains indeed 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 parents 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".

Edge	[8]a	[8]b	[6]	[2]	ICP	hiddenICP
RAF→MEK	✓	✓				✓
MEK→RAF			✓	✓		✓
MEK→ERK	✓	✓	✓			
PLCg→PIP2	✓	✓		✓	✓	✓
PLCg→PIP3		✓		✓		
PLCg→PKC	✓			✓		
PIP2→PLCg			✓		✓	
PIP2→PIP3				✓		
PIP2→PKC	✓					
PIP3→PLCg	✓					
PIP3→PIP2	✓	✓	✓		✓	✓
PIP3→AKT	✓					
AKT→ERK			✓		✓	✓
ERK→AKT		✓		✓	✓	✓
ERK→PKA				✓		
PKA→RAF	✓	✓				
PKA→MEK	✓	✓	✓	✓		
PKA→ERK	✓	✓			✓	
PKA→AKT	✓	✓	✓	✓		✓
PKA→PKC				✓		
PKA→P38	✓	✓	✓			
PKA→JNK	✓	✓	✓	✓		
PKC→RAF	✓	✓	✓			
PKC→MEK	✓	✓	✓	✓		
PKC→PLCg			✓			
PKC→PIP2			✓			
PKC→AKT			✓			
PKC→PKA		✓	✓			
PKC→P38	✓	✓	✓	✓		✓
PKC→JNK	✓	✓	✓	✓	✓	✓
P38→JNK						✓
P38→PKC						✓
JNK→PKC						✓
JNK→P38				✓		✓

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 $P_j^{(1)}, \ldots, P_j^{(k)}$ the p-values that variable X_j is causal for a response Y in the datasets $r = 1, \ldots, k$, i.e., $P_j^{(r)}$ is the smallest level α such that $j \in \hat{S}(\mathcal{E})$ 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

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

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 σ_r^2 denotes the variance of the noise level in the rth dataset: it is often hard though to have knowledge about σ_r^2 and one might then simply use $\sigma_r^2 \equiv 1$ 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 $P_j^{(r)}$ 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.

Data and software. An R-script to compute the causal effects with invariant causal prediction can be found at http://stat.ethz.ch/∼nicolai/experimentKemmeren.R. The raw and processed gene knockout experimental data [5] can be found on the webpage http://deleteome.holstegelab.nl/ under 'Downloads/Causal Inference'.

1. P. Bühlmann and S. van de Geer. Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer, New York, NY, 2011.
2. D. Eaton and K. Murphy. Exact Bayesian structure learning from uncertain interventions. In Proceedings of the 10th International Conference on Artificial Intelligence and Statistics (AISTATS), pages 107–114, 2007.
3. L.V. Hedges and I. Olkin. Statistical method for meta-analysis. Academic press, 2014.
4. T.R. Hughes, M.J. Marton, A.R. Jones, C.J. Roberts, R. Stoughton, C.D. Armour, H.A. Bennett, E. Coffey, H. Dai, Y.D. He, M.J. Kidd, A.M. King, M.R. Meyer, D. Slade, P.Y. Lum, S.B. Stepaniants, D.D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S.H. Friend. Functional discovery via a compendium of expression profiles. Cell, 102:109–126, 2000.
5. P. Kemmeren, K. Sameith, L.A. van de Pasch, J.J. Benschop, T.L. Lenstra, T. Margaritis, E. O'Duibhir, E. Apweiler, S. van Wageningen, C.W. Ko, S. van Heesch, M.M.. Kashani, G. Ampatziadis-Michailidis, M.O.. Brok, N.A. Brabers, A.J. Miles, D. Bouwmeester, S.R. van Hooff, H. van Bakel, E. Sluiters, L.V. Bakker, B. Snel, P. Lijnzaad, D. van Leenen, M.J. Groot Koerkamp, and F.C. Holstege. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. Cell, 157:740–752, 2014.
6. J.M. Mooij and T. Heskes. Cyclic causal discovery from continuous equilibrium data. In Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI), pages 431–439, Corvallis, OR, 2013. AUAI Press.
7. J. Peters, P. Bühlmann and N. Meinshausen Causal inference using invariant prediction: identification and confidence intervals, 2016 (with discussion). Preprint arXiv:1501.01332.
8. K. Sachs, O. Perez, D. Pe'er, D. Lauffenburger, and G. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. Science, 308:523–529, 2005.
9. P. Spirtes, C. Glymour, and R. Scheines. Causation, Prediction, and Search. MIT Press, second edition, 2000.
10. R. Tibshirani. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B, 58:267–288, 1996.
11. S.A. Stouffer, E.A. Suchman, L.C. DeVinney, S.A. Star, and R.M. Williams. The american soldier: Adjustment during army life, vol. 1, 1949.