JMLR Workshop and Conference Proceedings 6:147–156 NIPS 2008 workshop on causality

Distinguishing between cause and effect

Joris Mooij JORIS.MOOIJ@TUEBINGEN.MPG.DE
Max Planck Institute for Biological Cybernetics,
72076 Tübingen, Germany

Dominik Janzing DOMINIK.JANZING@TUEBINGEN.MPG.DE
Max Planck Institute for Biological Cybernetics
72076 Tübingen, Germany

Editor: Isabelle Guyon, Dominik Janzing and Bernhard Schölkopf

Abstract

We describe eight data sets that together formed the `CauseEffectPairs` task in the *Causality Challenge #2: Pot-Luck* competition. Each set consists of a sample of a pair of statistically dependent random variables. One variable is known to cause the other one, but this information was hidden from the participants; the task was to identify which of the two variables was the cause and which one the effect, based upon the observed sample. The data sets were chosen such that we expect common agreement on the ground truth. Even though part of the statistical dependences may also be due to hidden common causes, common sense tells us that there is a significant cause-effect relation between the two variables in each pair. We also present baseline results using three different causal inference methods.

Keywords: causal inference, benchmarks

1. Introduction

Arguably, the most elementary problem in causal inference is to decide whether statistical dependences between two random variables X, Y are due to (a) a causal influence from X to Y, (b) an influence from Y to X, or (c) a possibly unobserved common cause Z influencing X and Y. Most of the state-of-the-art causal inference algorithms address this problem only if X and Y are part of a larger set of random variables influencing each other. In that case, conditional statistical dependences rule out some causal directed acyclic graphs (DAGs) and prefer others (Spirtes et al., 1993; Pearl, 2000).

Recent work (Kano and Shimizu, 2003; Sun et al., 2006; Shimizu et al., 2006; Sun et al., 2008; Hoyer et al., 2009; Janzing and Schölkopf, 2008) suggests that the shape of the joint distribution shows asymmetries between cause and effect, which often indicates the causal direction with some reliability, i.e., one can distinguish between cases (a) and (b).

To enable more objective evaluations of these and other (future) proposals for identifying cause and effect, we have tried to select real-world data sets with pairs of variables where the causal direction is known. The best way to obtain the ground truth of the causal relationships in the systems that generated the data would be by performing interventions on one of the variables and observing whether the intervention changes the distribution of the other variable.

Data set	Number of samples	Variable 1	Variable 2	Causal relationship
pairs01	349	Altitude	Temperature	$1 \to 2$
pairs02	349	Altitude	Precipitation	$1 \to 2$
pairs03	349	Longitude	Temperature	$1 \to 2$
pairs04	349	Sunshine hours	Altitude	$1 \leftarrow 2$
pairs05	4177	Length	Age	$1 \leftarrow 2$
pairs06	4177	Age	Shell weight	$1 \to 2$
pairs07	4177	Diameter	Age	$1 \leftarrow 2$
pairs08	5000	Age	Wage per hour	$1 \to 2$

Table 1: Data sets in the CauseEffectPairs task.

Unfortunately, these interventions cannot be made in practice for many of the existing data sets because the original data-generating system is no longer available, or because of other practical reasons. Therefore, we have selected some data sets in which the causal direction should be clear by common sense.

In selecting the data sets for the CauseEffectPairs task, we applied the following selection criteria:

- the minimum number of data points should be a few hundred;
- the variables should have continuous values;
- there should be a significant cause–effect relationship between the two variables;
- the direction of the causal relationship should be known or obvious from the meaning of the variables;

We collected eight data sets satisfying these criteria, which we refer to as pairs01, ..., pairs08. They can be downloaded from Mooij et al. (2008). Some properties of the data sets are given in Table 1.

In this article, we describe the various data sets in the task and provide our "common sense" interpretation of the causal relationships present in the variables. We also present baseline results of all previously existing applicable causal inference methods that we know of.

2. Climate data

The first four pairs were obtained from climate data provided by the *Deutscher Wetterdienst* (DWD) and are available online at Deutscher Wetterdienst (2008). We merged several of the original data sets to obtain data for 349 weather stations in Germany, selecting only those weather stations with no missing data. After merging the data sets, we selected the following six variables: altitude, latitude, longitude, and annual mean values (over the years 1961–1990) of sunshine duration, temperature and precipitation. We converted the latitude and longitude variables from sexagesimal to decimal notation. Out of these six variables, we selected four different pairs with "obvious" causal relationships: altitude–temperature, altitude–precipitation, longitude–temperature and sunshine–altitude. We will now discuss each pair in more detail.

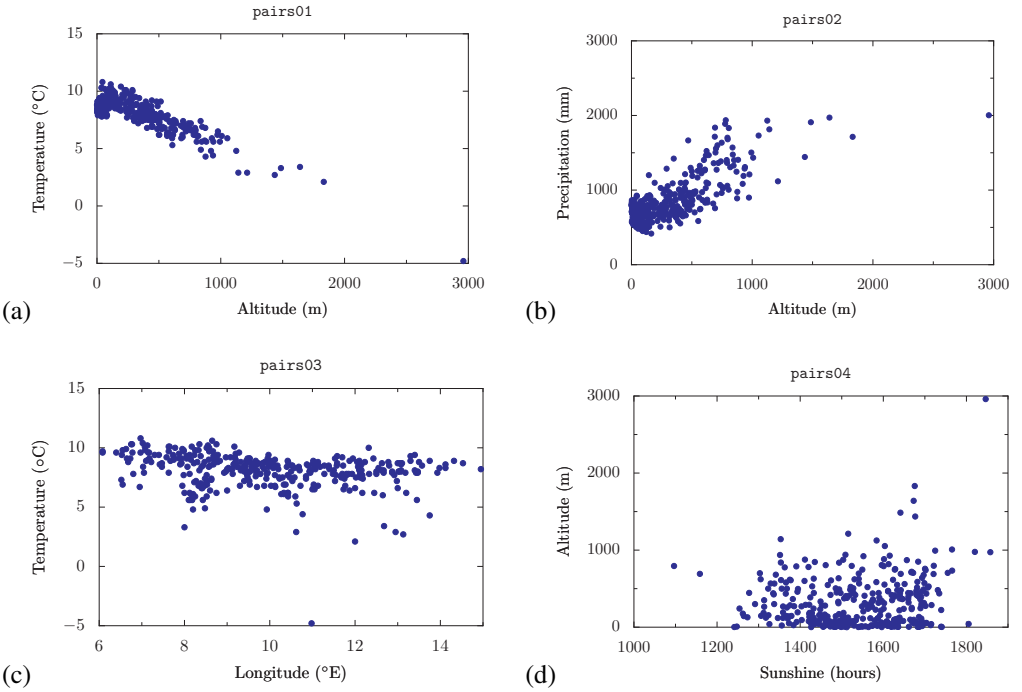

Figure 1: Scatter plots of the German climate data: (a) altitude–temperature, (b) altitude–precipiation, (c) longitude–temperature, (d) altitude–sunshine hours.

2.1 Altitude and temperature

As an elementary fact of meteorology, places with higher altitude tend to be colder than those that are closer to sea level (roughly 1 centigrade per 100 meter). There is no doubt that altitude is the cause and temperature the effect: one could easily think of an intervention where the thermometer is lifted by a balloon to measure the temperature at a higher point of the same longitude and latitude. On the other hand, heating or cooling a location does not change its altitude.

The altitudes in the DWD data set range from 0 m to 2960 m, which is sufficiently large to detect significant statistical dependences. The data is plotted in Figure 1(a).

One potential confounder is latitude, since all mountains are in the south and far from the sea, which is also an important factor for the local climate. The places with the highest average temperatures are therefore those with low altitude but lying far in the south (Upper Rhine Valley). Hence this confounder should induce positive correlations between altitude and temperature as opposed to the negative correlation between altitude and temperature which is already evident from the scatter plot. This suggests that the direct causal relation between altitude and temperature dominates over the confounder.

2.2 Altitude and precipitation

Altitude and precipitation form the second pair of variables that we selected from the DWD data; their relation is plotted in Figure 1(b).

It is known that altitude is also an important factor for precipitation since rain often occurs when air is forced to rise over a mountain range and the air becomes oversaturated with water

due to the lower temperature (orographic rainfall). This effect defines an indirect causal influence of altitude on precipitation via temperature. These causal relations are, however, less simple than the causal influence from altitude to temperature because gradients of the altitude with respect to the main direction of the wind are more relevant than the altitude itself. The hypothetical intervention that defines a causal relation could be to build artificial mountains and observe orographic rainfall.

2.3 Longitude and temperature

For the dependence between longitude and temperature, shown in Figure 1(c), a hypothetical intervention could be to move a thermometer between west and east. Even if one could adjust for altitude and latitude, it is unlikely that temperature would remain the same since the climate in the west is more oceanic and less continental than in the east of Germany. Therefore, longitude causes temperature.

2.4 Sunshine hours and altitude

The fourth and final pair of DWD variables are sunshine duration and altitude, shown in Figure 1(d). Linear regression between both quantities shows a slight increase of sunshine duration with altitude. Possible explanations are that higher cities are sometimes above low-hanging clouds. Cities in valleys, especially if they are close to rivers or lakes, typically have more misty days. Moving a sunshine sensor above the clouds clearly increases the sunshine duration whereas installing an artificial sun would not change the altitude. The causal influence from altitude to sunshine duration can be confounded, for instance, by the fact that there is a simple statistical dependence between altitude and longitude in Germany as explained in Subsection 2.1.

3. Abalone data

Another three pairs of variables were selected from the `Abalone` data set (Nash et al., 1994) in the UCI Machine Learning Repository (Asuncion and Newman, 2007). The data set contains 4177 measurements of several variables concerning the sea snail *Abalone*. The original data set contains the nine variables sex, length, diameter, height, whole weight, shucked weight, viscera weight, shell weight and number of rings. The number of rings in the shell is directly related to the age of the snail: adding 1.5 to the number of rings gives the age in years. Of these variables, we selected three pairs with obvious cause-effect relationships, which we now discuss in more detail.

3.1 Length and age

The data for the first Abalone pair, length and age, is plotted in Figure 2(a). For the variable "age" it is not obvious what a reasonable intervention would be since there is no possibility to change the time. However, waiting and observing how the length changes or how it changed from the past to the present can be considered as equivalent to the hypothetical intervention (provided that the relevant background conditions do not change too much). Clearly, this "intervention" would change the probability distributions of the length, whereas changing the length of snails (by a complicated surgery) would not change the distribution of age. Regardless of the difficulties of defining interventions, we expect common agreement on the ground truth (age causes length).

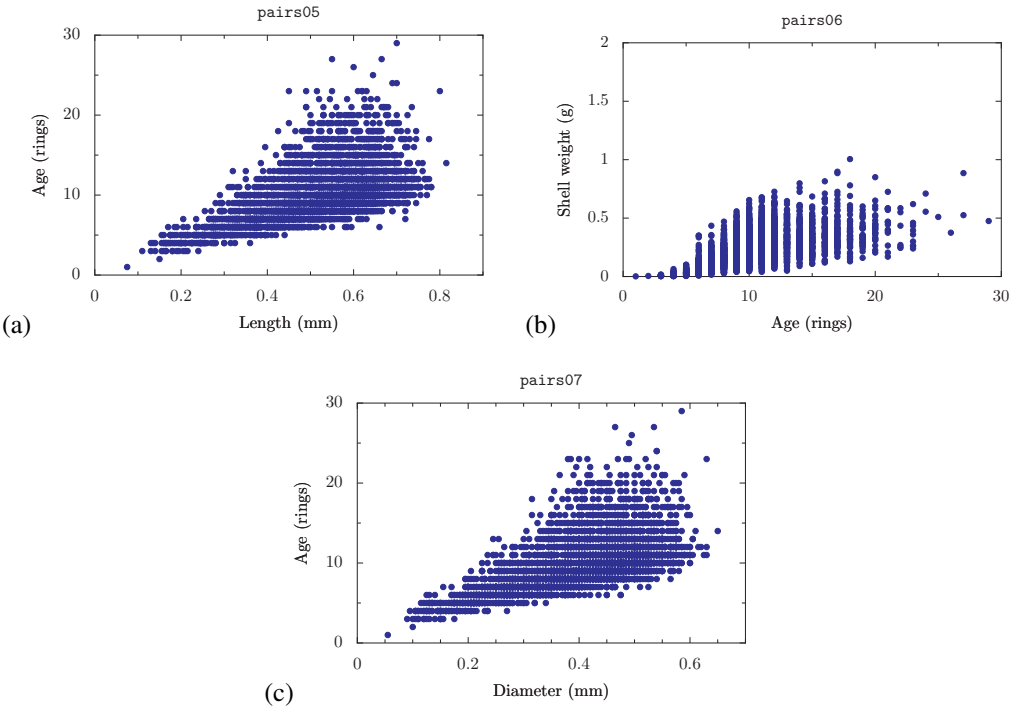

Figure 2: Scatter plots of the Abalone data: (a) length–age, (b) age–shell weight, (c) diameter–age.

3.2 Age and shell weight

The data are plotted in Figure 2(b). Similar considerations as in Subsection 3.1 hold for the ground truth: age causes shell weight but not vice versa.

3.3 Diameter and age

For the final pair, shell diameter and age, the data are plotted in Figure 2(c). Again, age causes diameter and not the other way around.

4. Age and wage per hour of employees in the USA

Our final data source was the `Census Income` data set (Kohavi, 1996) in the UCI Machine Learning Repository (Asuncion and Newman, 2007). We have selected the following variables: `1 AAGE` (age), and `7 AHRSPAY` (wage per hour) and selected the first 5000 instances for which wage per hour was not equal to zero. The scatter plot for this pair is shown in Figure 3. It clearly shows an increase of wage up to about 45 and decrease for higher age.

As already argued in the Abalone case, interventions on the variable "age" are difficult to define. Compared to the discussion in the context of the Abalone data set, it seems more problematic to consider waiting as a reasonable "intervention" since the relevant (economical) background conditions change rapidly compared to the length of the human life: If someone's salary is higher than the salary of a 20 year younger colleague *because* of his/her longer job experience, we cannot conclude that the younger colleague 20 years later will earn the same

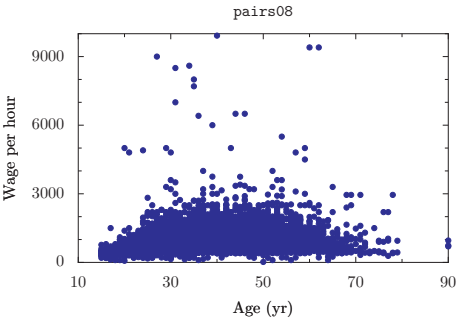

Figure 3: Scatter plot of Census data: age–wage per hour.

money as the colleague earns now. Possibly, the factory or even the branch of industry he/she was working in does not exist any more and his/her job experience is no longer appreciated. However, we know that employees sometimes indeed do get a higher income because of their longer job experience. Pretending longer job experience by a fake certificate of employment would be a possible intervention. On the other hand, changing the wage per hour is an intervention that is easy to imagine (though difficult for us to perform) and this would certainly not change the age.

5. Baseline results

At the time the challenge was held, only three methods existed for deciding upon the causal direction between two real-valued variables, to the best of our knowledge: the method proposed by Friedman and Nachman (2000), LiNGAM (Shimizu et al., 2006) and the causal inference method of Hoyer et al. (2009). In this Section, we report the results of applying these three methods to the data sets of the challenge task. These results may serve as baseline results for future evaluations.

5.1 Comparing marginal likelihood of Gaussian Process regression fits

The basic idea behind the method of Friedman and Nachman (2000) (when applied to the special case of only two variables X and Y) is fitting a Gaussian Process (Rasmussen and Williams, 2006) to the data twice: once with X as input and Y as output, and once with the roles of X and Y reversed. If the former fit has a larger marginal likelihood, this indicates that X causes Y, and otherwise, one concludes that Y causes X. We adopted a squared exponential covariance function and used the GPML code (Rasmussen and Williams, 2007).

The results are shown in Table 2. Only three out of eight causal direction inferences are correct.

5.2 LiNGAM

The causal inference method LiNGAM (an acronym for Linear, Non-Gaussian, Acyclic causal Models) assumes that effects are linear functions of their causes, plus independent additive noise. Shimizu et al. (2006) showed that if all (or all except one of the) noise distributions are non-Gaussian, the correct causal (data-generating) structure can be identified asymptotically using Independent Component Analysis. We have applied the implementation provided by the authors at (Hoyer et al., 2006) on the data sets of our challenge task.

Dataset	$S_{1\to 2}$	$S_{1\leftarrow 2}$	Decision	Ground truth	Correct?
pairs01	2.183×10^{02}	2.171×10^{02}	$1 \to 2$	$1 \to 2$	+
pairs02	3.355×10^{02}	3.385×10^{02}	$1 \leftarrow 2$	$1 \to 2$	-
pairs03	4.858×10^{02}	4.603×10^{02}	$1 \to 2$	$1 \to 2$	+
pairs04	4.821×10^{02}	4.889×10^{02}	$1 \leftarrow 2$	$1 \leftarrow 2$	+
pairs05	5.141×10^{03}	4.291×10^{03}	$1 \to 2$	$1 \leftarrow 2$	-
pairs06	4.568×10^{03}	4.801×10^{03}	$1 \leftarrow 2$	$1 \to 2$	-
pairs07	5.086×10^{03}	4.243×10^{03}	$1 \to 2$	$1 \leftarrow 2$	-
pairs08	6.842×10^{03}	6.869×10^{03}	$1 \leftarrow 2$	$1 \to 2$	-

Table 2: Baseline results for distinguishing the cause from the effect, using the method of Friedman and Nachman (2000); S denotes the logarithm of the marginal likelihood of the Gaussian Process fit.

Dataset	Diagnostic	Decision	Ground truth	Correct?
pairs01	OK	$1 \leftarrow 2$	$1 \to 2$	-
pairs02	Not really triangular at all	$1 \leftarrow 2$	$1 \to 2$	-
pairs03	Not really triangular at all	$1 \to 2$	$1 \to 2$	+
pairs04	Only somewhat triangular	$1 \to 2$	$1 \leftarrow 2$	-
pairs05	OK	$1 \to 2$	$1 \leftarrow 2$	-
pairs06	OK	$1 \leftarrow 2$	$1 \to 2$	-
pairs07	OK	$1 \to 2$	$1 \leftarrow 2$	-
pairs08	OK	$1 \to 2$	$1 \to 2$	+

Table 3: Baseline results for distinguishing the cause from the effect, using LiNGAM Shimizu et al. (2006).

Dataset	$p_{1\to 2}$	$p_{1\leftarrow 2}$	Decision	Ground truth	Correct?
pairs01	1.64×10^{-02}	9.43×10^{-15}	$1 \to 2$	$1 \to 2$	+
pairs02	1.50×10^{-13}	2.88×10^{-16}	$1 \to 2$	$1 \to 2$	+
pairs03	7.89×10^{-03}	7.02×10^{-04}	$1 \to 2$	$1 \to 2$	+
pairs04	5.50×10^{-05}	1.08×10^{-02}	$1 \leftarrow 2$	$1 \leftarrow 2$	+
pairs05	1.13×10^{-70}	7.79×10^{-23}	$1 \leftarrow 2$	$1 \leftarrow 2$	+
pairs06	1.56×10^{-210}	1.98×10^{-113}	$1 \leftarrow 2$	$1 \to 2$	-
pairs07	2.66×10^{-82}	5.85×10^{-26}	$1 \leftarrow 2$	$1 \leftarrow 2$	+
pairs08	$0.00 \times 10^{+00}$	1.60×10^{-80}	$1 \leftarrow 2$	$1 \to 2$	-

Table 4: Baseline results for distinguishing the cause from the effect, using the method of Hoyer et al. (2009).

The results are shown in Table 3. Only two out of eight causal direction inferences are correct.

5.3 Additive noise models

The basic idea of the recent method by Hoyer et al. (2009) is to assume that the effect can be written as some (not necessarily linear) function of the cause, plus additive noise, which is independent of the cause. In practice, one tests the causal model "X causes Y" as follows:

- perform regression of Y on X in order to estimate the function $f : \mathbb{R} \longrightarrow \mathbb{R}$ that best approximates the functional relationship between X and Y, i.e., such that $Y \approx f(X)$,
- calculate the residuals $Y - f(X)$ for all data points,
- check whether these residuals are independent of X, i.e., whether $(Y - f(X)) \perp\!\!\!\perp X$.

For the regression, we used standard Gaussian Process Regression (Rasmussen and Williams, 2006) using the GPML code (Rasmussen and Williams, 2007), with a squared exponential covariance function. For the independence test, we used the independence test based on the Hilbert Schmidt Independence Criterion (also known as HSIC) (Gretton et al., 2005), using the gamma approximation and Gaussian kernels with heuristically chosen kernel widths. The statistical test assumes independence as a null hypothesis and calculates corresponding p-values. Now in order to decide whether "X causes Y" or, alternatively, "Y causes X", one simply takes the model with the highest p-value for independence between residuals and regressor.

We report the results in Table 4. By using this method, we correctly classify six out of eight data sets. The small p-values may indicate that the assumption of additive noise is violated in these data sets, even in the correct causal direction. Still, by comparing the p-values in both directions, the correct decision is made in most cases.[1]

6. Discussion and remarks on submitted solutions

Finding data sets satisfying the criteria mentioned in Section 1 turned out to be challenging, which explains why the number of data sets in our task is relatively small (another reason is that we only decided to submit a task to the challenge just shortly before the deadline). For future evaluations, the number of data sets should be increased in order to obtain more significant conclusions when used as benchmarks for comparing causal inference algorithms.

We received 6 submissions as suggested solutions of this task. The number of correctly identified pairs were $2, 8, 5, 3, 5, 7$, while the submission with 7 correct solutions was (unfortunately) later changed to 5 correct ones. The winner team (Zhang and Hyvärinen) correctly identified 8 out of 8 causal directions. Their method will be described in the paper *Distinguishing causes from effects using nonlinear acyclic causal models*, published elsewhere in this workshop proceedings. One group (not the winning group) used the fact that the pairs contained common variables and used conventional methods in addition to a new method. Since the goal of our task was to consider only pairs of variables at a time, it was a weakness of our task to allow for such a solution strategy (the submission was accepted nevertheless, of course).

An additional desideratum for data sets used in similar future challenges would therefore be that all variable pairs should be disjoint. On the other hand, the constraint that the variables should have continuous values could be removed, which would make the task more challenging for the participants (and would also make it easier to find suitable data).

1. Meanwhile, we have improved the method by replacing the regression step by a dependence minimization procedure, which yields similar qualitative results, but with more plausible p-values (Mooij et al., 2009).

Acknowledgments

We would like to thank Bernhard Schölkopf for suggesting the DWD climate data. This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886.

References

A. Asuncion and D.J. Newman. UCI machine learning repository, 2007. URL http://www.ics.uci.edu/~mlearn/MLRepository.html.

Deutscher Wetterdienst. Website of the German weather service, 2008. URL http://www.dwd.de/.

N. Friedman and I. Nachman. Gaussian process networks. In *Proceedings of the 16th Annual Conference on Uncertainty in Artificial Intelligence*, pages 211–219, 2000.

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory: 16th International Conference (ALT 2005)*, pages 63–78, August 2005.

P. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems (NIPS*2008)*. MIT Press, 2009.

Patrik O. Hoyer, Antti Kerminen, and Shohei Shimizu. LiNGAM v. 1.4.2, December 2006. URL http://www.cs.helsinki.fi/group/neuroinf/lingam/.

D. Janzing and B. Schölkopf. Causal inference using the algorithmic Markov condition, 2008. URL http://arxiv.org/abs/0804.3678.

Y. Kano and S. Shimizu. Causal inference using nonnormality. In *Proceedings of the International Symposium on Science of Modeling, the 30th Anniversary of the Information Criterion*, pages 261–270, Tokyo, Japan, 2003.

Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996.

Joris Mooij, Dominik Janzing, and Bernhard Schölkopf. Distinguishing between cause and effect, 2008. URL http://www.causality.inf.ethz.ch/repository.php?id=14.

Joris Mooij, Dominik Janzing, Jonas Peters, and Bernhard Schölkopf. Regression by dependence minimization and its application to causal inference in additive noise models. To appear at the 26th International Conference on Machine Learning (ICML 2009), 2009.

W. Nash, T. Sellers, S. Talbot, A. Cawthorn, and W. Ford. The Population Biology of Abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (H. rubra) from the North Coast and Islands of Bass Strait. Sea Fisheries Division, Technical Report No. 48 (ISSN 1034-3288), 1994.

J. Pearl. *Causality: Models, reasoning, and inference*. Cambridge University Press, 2000.

C. E. Rasmussen and C. Williams. GPML code, 2007. URL http://www.gaussianprocess.org/gpml/code.

C. E. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. J. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Lecture Notes in Statistics. Springer, New York, 1993.

X. Sun, D. Janzing, and B. Schölkopf. Causal inference by choosing graphs with most plausible Markov kernels. In *Proceedings of the 9th International Symposium on Artificial Intelligence and Mathematics*, pages 1–11, Fort Lauderdale, FL, 2006.

X. Sun, D. Janzing, and B. Schölkopf. Causal reasoning by evaluating the complexity of conditional densities with kernel methods. *Neurocomputing*, 71:1248–1256, 2008.