Estimating complex causal effects from observational data

Juha Karvanen
Department of Mathematics and Statistics,
University of Jyväskylä,
Jyväskylä, Finland
juha.t.karvanen@jyu.fi

March 6, 2014

Abstract

Causal calculus is a tool to express causal effects in the terms of observational probability distributions. The application of causal calculus in the non-parametric form requires only the knowledge of the causal structure. However, some kind of explicit modeling is needed when numeric estimates of the causal effect are to be calculated. In this paper, the estimation of complicated nonlinear causal relationships from observational data is studied. It is demonstrated that the estimation of causal effects does not necessarily require the causal model to be specified parametrically but it suffices to model directly the observational probability distributions. The conditions when this approach produces valid estimates are discussed. Generalized additive models, random forests and neural networks are applied to the estimation of causal effects in examples featuring the backdoor and the frontdoor adjustment.

Keywords: causal estimation, data analysis, nonlinearity, structural equation model
1 Introduction

During the past two decades major advances have been taken in causal modeling. Structural causal models offer a mathematically well-defined concept to convey causal assumptions and causal calculus provides a systematic way to express the causal effects in the terms of the observational probabilities (Pearl, 1995, 2009). The benefits of this framework are numerous: First, the philosophically entangled concept of causality has a clear and practically useful definition. Second, in order to tell whether a causal effect can be identified in general it is sufficient to specify the causal model non-parametrically, i.e. specify only the causal structure of the variables. Third, the completeness of causal calculus has been proved (Huang and Valtorta, 2006; Bareinboim and Pearl, 2012a) and algorithms for the identification of the causal effects have been proposed (Tian and Pearl, 2002; Shpitser and Pearl, 2006a; Bareinboim and Pearl, 2012a). Fourth, the framework can be extended to deal with study design, selection bias and missing data (Karvanen, 2013; Bareinboim and Pearl, 2012b; Mohan et al., 2013).

Estimation of causal effects from the data has in general received only a little attention in the research on structural causal models. Notable exceptions are the estimation of the causal effect a binary treatment and the modeling using linear Gaussian models in the structural equations modeling community. With linear Gaussian models causal effects can be often directly identified in closed-form using path analysis as recently reminded by Pearl (2013). Linear models nicely illustrate many key concepts of causal analysis. However, in practical situations, the assumption of the linearity of all causal effects is often too restrictive.

This paper aims to narrow the gap between the theory of causal models and practical data analysis. As a starting point it is assumed that only the causal structure can be specified, i.e. the causal effects can be described qualitatively in the nonparametric form but not in the parametric form. Causal relationships are expected to have a complex nonlinear forms which cannot be a priori modeled but are assumed to be continuous, differentiable and sufficiently smooth. The key idea is to first use causal calculus to find out the observational probabilities needed to calculate the causal effects and then estimate these observational probabilities from the data using flexible models from the arsenal of modern data analysis. The estimated causal effects cannot be expressed in a closed-form but can be numerically calculated for a given input. The approach works if the required observational probabilities can be
reliably interpolated or extrapolated from the model fitted to the data.

The estimation procedure is given in detail in Section 2. Simulation examples demonstrating the estimation of complex nonlinear causal effects are presented in Section 3. Conclusions are given in Section 4.

2 Estimation of causal effects

An important building block of structural causal models is the do-operator. The notation \( p(Y \mid \text{do}(X = x)) \), or shortly \( p(Y \mid \text{do}(x)) \), refers to the distribution of random variable \( Y \) when an action that sets the value of variable \( X \) to \( x \) is applied. The rules of causal calculus, also known as do-calculus, give graph theoretical conditions when it is allowed to insert or delete observations, to exchange actions and observation and to insert or delete actions (Pearl, 2009). These rules and the standard probability calculus are sequentially applied to present a causal effect in the terms of observational probabilities.

A straightforward approach for the estimation of causal effects is to first to define the causal model explicitly in a parametric form up to some unknown parameters and then use data to estimate these parameters. For instance, in linear Gaussian models, the causal effects are specified by the linear models with unknown regression coefficients that can be estimated on the basis of the sample covariance matrix calculated from the data. This approach may turn out to be difficult to implement when the functional forms of the causal relationships are complex or unknown because it is not clear how the parametrized causal effects can be estimated from the data.

The estimation procedure proposed here utilizes implicit modeling of causal effects. The key principles of the estimation procedure are similar to those used for causal models with design (Karvanen, 2013). Here it is assumed that the causal structure is known, the causal effects are complex and there are data collected by simple random sampling from the underlying population. The estimation procedure can be presented as follows:

1. Specify the causal structure.
2. Check that the causal effect(s) of interest can be identified.
3. Apply causal calculus to present the causal effect(s) using observational distributions.
4. Fit flexible models for the observational distributions needed to calculate the causal effect.

5. Combine the fitted observational models and use them to predict the causal effect.

Step 1 is easy to carry out because the causal structure is assumed to be known. In Steps 2 and 3, the rules of causal calculus are applied either intuitively or systematically via the identification algorithms (Tian and Pearl, 2002; Shpitser and Pearl, 2006a,b; Bareinboim and Pearl, 2012a). In Step 4, the fitting of a model to the data is a statistical problem for which various data analysis methods in statistics and machine learning are available. Naturally, validation is needed to avoid over-fitting. In Step 5, the fitted observational models are combined as the result derived in Step 3 indicates. This usually requires integration over some variables, which can be often carried out by summation with the actual data as demonstrated in the examples in Section 3.

Steps 4 and 5 are the critical steps of the procedure. The availability of the data limits the causal effects that can be estimated. For instance, if the calculation of the causal effect \( p(Y \mid \text{do}(X = x)) \) requires an estimate of the observational distribution \( p(Y \mid x, z) \), the range of \( x \) and \( z \) in the data limits the values of \( X \) for which the causal effect can be calculated. Moreover, some combinations of \( x \) and \( z \) may be absent in the data. In this case, it must be assumed that the estimated \( p(Y \mid x, z) \) can be interpolated or extrapolated for the values absent. In other words, it is assumed that the functional forms of the causal effects are continuous, differentiable and smooth. If the causal effect is to be calculated only for a specific value \( X = x_1 \), then the success of the estimation depends only on the quality of estimate \( \hat{p}(Y \mid X = x_1, z) \).

3 Examples

The causal structures for the examples are presented in Figure 1. In both examples, the aim is to estimate the average causal effect \( E(Y \mid \text{do}(X = x)) \).
Figure 1: Causal structures for the examples (a) and (b). The solid circle indicates that the variable is observed and the open circle indicates that the variable is unobserved.

3.1 Example (a): backdoor adjustment

For a causal model with the causal structure shown in Figure 1(a), the causal effect of $X$ to $Y$ can be expressed as

$$E(Y \mid do(X = x)) = \int E(Y \mid x, z)p(z)dz.$$  

(1)

In other words, the observed association of $X$ and $Y$ is confounded by $Z$. The result (1) can be derived by applying the rules of causal calculus and is also known as the backdoor adjustment (Pearl, 1993). Mediator $W$ is not needed to calculate $E(Y \mid do(X = x))$.

In Example (a) the data generating mechanism is defined by the following structural equations

$$Z \sim \text{NPoly4Q}(\lambda_1 = 0, \lambda_2 = 0.1, \tau_3 = 0.2, \tau_4 = 0.05),$$
$$X = 4\phi(Z - 1) - 0.5 + \epsilon_X,$$
$$W = \phi(Z - 1) + \phi(Z + 1) + 2\phi(X - 0.5) - X,$$
$$Y = Z + W + \epsilon_Y,$$
$$\epsilon_X \sim N(0, 0.1^2),$$
$$\epsilon_Y \sim N(0, 0.1^2),$$
where the nonlinear structural relationships are defined using the density function of standard normal distribution

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{x^2}{2} \right),$$  \hspace{1cm} (2)

NPoly4Q($\lambda_1 = 0, \lambda_2 = 0.1, \tau_3 = 0.2, \tau_4 = 0.05$) stands for a normal polynomial quantile mixture (Karvanen, 2006) with mean 0, L-scale 0.1, L-skewness 0.2 and L-kurtosis 0.05 and the disturbance terms $\epsilon_X$ and $\epsilon_Y$ are independent from all other variables. The data generating mechanism is completely unknown in the estimation.

The data contain $n = 10000$ observations of variables $Z, X, W$ and $Y$. The joint distribution of $Z, X$ and $Y$ is illustrated in scatterplots in Figure 2. It can be seen that the observed dependence structure is complicated and cannot be modeled with simple regression models. There are also combinations, such as $x = 0$ and $z = 1$, for which the observed data do not directly provide any information on the possible values of $y$. Some assumptions are needed to overcome the lack of the data. A parametric model for the joint distribution may be assumed, or as it is done in this paper, the joint distribution of $Z, X$ and $Y$ is assumed to be sufficiently smooth to allow the estimation of $E(Y \mid x, z)$ for the whole observed range of $x$ and $z$.

In Figure 3 the observed conditional average $E(Y \mid x)$ and the average causal effect $E(Y \mid do(x))$ are compared. It can be seen that while $E(Y \mid do(x))$ is a decreasing function of $x$, the observed $E(Y \mid x)$ is an increasing function of $x$ on the interval $[0.0, 1.0]$ due to confounder $Z$. The lower panel of Figure 3 shows data generated from a hypothetical experiment where the value of $X$ is controlled (The data are for illustration only and are not used in the estimation). The differences compared the observational data in upper panel are evident.

Application of the estimation procedure from Section 2 with the result (1) requires $E(Y \mid x, z)$ to be estimated and integrated over the values of $Z$. For the estimation, three alternative methods are applied: generalized additive models (GAM) (Hastie and Tibshirani, 1990; Hastie et al., 2009) with spline smoothing, random forests (Breiman, 2001) and neural networks with a single hidden layer (Haykin, 1994; Hastie et al., 2009). When an estimator $\hat{E}(Y \mid x, z)$ is available, the integration can be carried out by summation over the values $z_i$ in the data

$$\hat{E}(Y \mid do(X = x)) = \frac{1}{n} \sum_{i=1}^{n} \hat{E}(Y \mid x, z_i).$$  \hspace{1cm} (3)
Figure 2: Scatterplot matrix of a subsample of 500 observations in Example (a).
Figure 3: Average causal effect compared to average conditional mean in Example (a). In the upper panel the curves are presented with observed data (subsample of 500 observations) and in the lower panel with data generated from the hypothetical experiment where the values of $X$ are controlled (8 values of $Y$ for each $x$).
The GAM estimation is carried out using R (R Core Team, 2013) and the package `gam` (Hastie, 2013). The random forest estimation is carried out with the R package `randomForests` (Liaw and Wiener, 2002) and the neural networks estimation is carried out with the R package `nnet` (Venables and Ripley, 2002). In addition to variables $X$ and $Z$, their product is also given as the input for all three estimation methods. The R package `caret` (Kuhn et al., 2014) is utilized to select the tuning parameters via ten-fold cross-validation. The tuning parameters are the degrees of freedom in the spline smoothing in GAM and the size of the hidden layer and the weight decay in neural networks.

The estimated average causal effects are presented in Figure 4. It can be seen that GAM and neural networks performed very well but the random forests failed to estimate the causal effect. In fact, random forests were included here as a method that is not really suitable for the estimation task. First, random forests are designed to work with high number of potential predictors. Second, the binary decision rules do not provide a smooth estimator that can extrapolated to the data points not observed. Despite these restrictions random forests may be a good method for the estimation of causal effects in some other scenarios.

### 3.2 Example (b): frontdoor adjustment

For a causal model with the causal structure shown in Figure 1(b), the causal effect of $X$ to $Y$ can be expressed as

$$
E(Y \mid \text{do}(X = x)) = \int p(z \mid X = x) \int E(Y \mid X = x', Z = z)p(X = x')dx'dz.
$$

This result is known as the frontdoor adjustment (Pearl, 1995). Here $U$ is an observed confounder that has an effect both to $X$ and to $Y$. Despite this, the causal effect of $X$ to $Y$ can be still identified thanks to mediator $Z$ that is independent on $U$ on the condition of $X$.

In Example (b) the 10000 observations are generated from the following
Figure 4: Estimated average causal effects compared to the true average causal effect in Example (a).
structural equations

\[ U \sim N(0, 1), \]
\[ X' \sim \text{Unif}(-2, 2), \]
\[ X = X' + U, \]
\[ Z = 4\phi(X) + \epsilon_Z, \]
\[ Y = \phi(Z - 0.5) + 0.3 \ast Z - 0.1U, \]
\[ \epsilon_Z \sim N(0, 0.1^2). \]

The data contain only variables \( Z, X \) and \( Y \) whose joint distribution is illustrated in scatterplots in Figure 5. Again there combinations, such as \( x = 0 \) and \( z = 0 \), not present in the data but in this case the situation is different from Example (a) because variable \( Z \) depends only on \( X \) and therefore the observed distribution \( p(Z \mid x) \) equals the causal distribution \( p(Z \mid \text{do}(x)) \). From the observational distribution \( p(Y \mid x) \), one might to misjudge that \( Y \) decreases as the function \( X \) but in reality this is not the case as it can be seen from Figure 6. The highest expected values for \( Y \) are obtained when \( X \) is set to have values around zero. The observed decreasing trend is due to the unobserved confounder \( U \).

The frontdoor adjustment (4) leads to estimator

\[
\hat{E}(Y \mid \text{do}(X = x)) = \frac{1}{n} \sum_{i=1}^{n} \hat{E}(Y \mid X = x_i, Z = \tilde{z}_i)
\]

where \( \tilde{z}_i \) is a random variable generated from the estimated distribution \( \hat{p}(Z \mid X = x) \). Often reliable approximation of (5) is obtained using the estimator

\[
\hat{E}(Y \mid \text{do}(X = x)) = \frac{1}{n} \sum_{i=1}^{n} \hat{E}\left(Y \mid X = x_i, Z = \hat{E}(Z \mid X = x)\right),
\]

where \( \tilde{z}_i \) is replaced by its expectation \( \hat{E}(Z \mid X = x) \). The same methods as in Example (a) are used for the estimation and estimator (6) is applied for the computational convenience.

The estimated average causal effects are shown in Figure 7. GAM and neural networks performed again well but as expected, random forests did not produce a smooth estimate.
Figure 5: Scatterplot matrix of a subsample of 500 observations in Example (b).
Figure 6: Average causal effect compared to average conditional mean in Example (b). In the upper panel the curves are presented with observed data (subsample of 500 observations) and in the lower panel with data generated from the hypothetical experiment where the values of $X$ are controlled (8 values of $Y$ for each $x$).
Figure 7: Estimated average causal effects compared to the true average causal effect in Example (b).
4 Conclusions

It was demonstrated that the estimation of causal effects does not necessarily require the causal model to be specified parametrically but it suffices to model the observational probability distributions. The key message can be summarized in the form of the equation

\[ \text{Causal estimation} = \text{causal calculus} + \text{data analysis}. \]

This result has practical importance because it allows modern data analysis tools to be applied in the estimation of causal effects. The causal structure can be often specified even if it is very difficult to forge a suitable parametric model for the causal effects. The data can modeled with flexible models for which ready-made software implementations exist.

The proposed approach assumes that the causal effects are sufficiently smooth to be estimated from the data. Accurate estimation of complicated effects requires large datasets. The high noise level and unobserved confounders further increase the amount of data needed. In the examples, the variables were continuous but the approach works similarly for discrete variables.

In addition to the data analysis methods applied in the examples, other methods are expected to be potentially useful. The estimation methods should provide smooth estimates to avoid problems similar to those encountered with random forests in the examples. It is recommended to try alternative estimation methods to understand the sensitivity of the results to the choice of the estimation method.

In the presented examples, the data was obtained by simple random sampling but the proposed estimation procedure can be applied with other study designs as well. The study design can be described using causal models with design and flexible models can applied also for the selection processes.

Estimation of complex causal effects from observational data is a challenging problem and the obtained estimates should be always interpreted cautiously. The results should be evaluated with respect to the existing scientific knowledge before any conclusions can be made.
References

Bareinboim, E. and Pearl, J. (2012a). Causal inference by surrogate experiments: $z$-identifiability. In de Freitas, N. and Murphy, K., editors, Proceedings of the Twenty-Eight Conference on Uncertainty in Artificial Intelligence, pages 113–120. AUAI Press.

Bareinboim, E. and Pearl, J. (2012b). Controlling selection bias in causal inference. In JMLR Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS), volume 22, pages 100–108.

Breiman, L. (2001). Random forests. Machine learning, 45(1):5–32.

Hastie, T. (2013). gam: Generalized Additive Models. R package version 1.09.

Hastie, T. and Tibshirani, R. (1990). Generalized additive models. CRC Press.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning. Springer, 2nd edition.

Haykin, S. (1994). Neural networks: a comprehensive foundation. Prentice Hall PTR.

Huang, Y. and Valtorta, M. (2006). Pearl's calculus of intervention is complete. In Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence, pages 217–224. AUAI Press.

Karvanen, J. (2006). Estimation of quantile mixtures via L-moments and trimmed L-moments. Computational Statistics & Data Analysis, 51(2):947–959.

Karvanen, J. (2013). Study design in causal models. Submitted for publication, arXiv:1211.2958.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., and the R Core Team (2014). caret: Classification and Regression Training. R package version 6.0-24.
Liaw, A. and Wiener, M. (2002). Classification and regression by random-forest. *R News*, 2(3):18–22.

Mohan, K., Pearl, J., and Tian, J. (2013). Graphical models for inference with missing data. In *Proceedings of Neural Information Processing Systems Conference (NIPS-2013)*.

Pearl, J. (1993). Comment: Graphical models, causality and intervention. *Statistical Science*, 8(3):266–269.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–710.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, second edition.

Pearl, J. (2013). Linear models: a useful microscope for causal analysis. *Journal of Causal Inference*, 1(1):155–170.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Shpitser, I. and Pearl, J. (2006a). Identification of conditional interventional distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI2006)*, pages 437–444. AUAI Press.

Shpitser, I. and Pearl, J. (2006b). Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, pages 1219–1226. AAAI Press.

Tian, J. and Pearl, J. (2002). A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, pages 567–573. AAAI Press/The MIT Press.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition.