TECHNICAL REPORT
R-472
February 2018

What is Gained from Past Learning

Judea Pearl
Cognitive Systems Laboratory
Computer Science Department
University of California, Los Angeles, CA 90024 USA
judea@cs.ucla.edu

Abstract

We consider ways of enabling systems to apply previously learned information to novel situations so as to minimize the need for retraining. We show that theoretical limitations exist on the amount of information that can be transported from previous learning, and that robustness to changing environments depends on a delicate balance between the relations to be learned and the causal structure of the underlying model. We demonstrate by examples how this robustness can be quantified.

Keywords: Transfer Leaning, Domain Adaptation, Robustness, Compositional Regression, Causal Models.

1 Introduction

Assume that we have learned a certain relation R in environment that is governed by a probability function P. Now the environment changes and P turns into P^*. We would still like to estimate R, but in the new environment, P^*. The basis of much works on "Transfer Learning," "Robust Learning," "Domain Adaptation," and "Life Long Learning" (L2L) hinges on the intuition that it would be a great waste to start learning $R(P^*)$ from scratch, instead of amortizing what we learned in P.

This intuition assumes, of course, that the two environments share some features in common, and that the shared features are significant in determining R. Surely, if the two environments are totally different, then we might as well start learning things from scratch – there is simply no other choice. Similarly, if the target relation R is defined exclusively on the novelty part of P^*, no advantage would be realized by transferring what was learned in P.

To anchor this intuition in a formal setting[1] let us assume that the target relation

[1]The formal analysis provided by transportability theory (Pearl and Bareinboim, 2011, 2014) asks what variables must be observed in the new environment, so that R can be estimated consistently (in the limit of a large sample) despite environmental changes. In contrast, we now seek a finite-sample assessment of the gains that can be realized by borrowing information from P, regardless of whether asymptotic consistency is feasible.

R can be decomposed into a set S of sub-relations, and that these sub-relations fall into two categories:

S_A - sub-relations on which P and P^* agree, and

S_D - sub-relations on which P and P^* disagree.

One obvious saving that can be realized from knowing S_A is in learning time. If we have trained the learner on 100 cases from each distribution, we can estimate S_A using all 200 samples, and S_D using the 100 samples of P^*. The net result being that some portions of R receive extra samples, which render them more precise, thus making the estimate of R more precise (i.e., less susceptible to sampling bias). Conversely, if we aspire to achieve a given precision in R, less samples, or shorter learning time, would be realized overall.

A simple example can illustrate this logic.

Example 1. *Let X and Y be two sets of variables governed by a joined distribution $P = P(x, y)$. X could represent class labels and Y a set of measurements, or features. If our task is to infer X on the basis of measurements of Y, then the relation of interest is $R = P(x|y)$, which can be learned by drawing samples from P.*

Let us assume that P changes into P^* such that the prior probability remains the same, $P(x) = P^*(x)$, but the conditional probability $P(y|x)$ changes. (This would be the case, for example, when the instruments for measuring Y undergo changes.) We can either learn $P^*(x|y)$ from scratch, by drawing samples from P^*, or we can borrow samples drawn previously from P, pool them with what we observe in P^* and obtain an improved estimate of $R = P^*(x|y)$. This can be done by decomposing R into a product of prior and conditional probabilities, then capitalizing on the equality $P(x) = P^*(x)$.

We have:

$$\begin{aligned} R &= P^*(x|y) \\ &= P^*(y|x)P^*(x)/P^*(y) \\ &= P^*(y|x)P(x)/P^*(y) \end{aligned} \tag{1}$$

The last expression permits us to use the more precise estimated $P(x)$ rather than rely solely on the small-sample estimate of $P^*(x|y)$.

This simple example raises a fundamental question: Is it always beneficial to decompose a relation into components, estimate each component individually, some with improved precision, then recombine the results?

A competing intuition might claim that the exercise of decomposing, estimating, and combining introduces new sources of noise, compared to, say, estimating the relation in one shot.

The question is further complicated by the fact that decompositions are not unique. Eq. (1), for example can also be written as:

$$R = P^*(x|y) = P^*(y|x)P(x)/\sum_{x'} P^*(y|x')P(x') \tag{2}$$

This calls for refraining from learning $P^*(y)$ directly in the new environment, but estimating $P^*(y|x)$ and $P^*(y|x')$ for all x' at the new environment, then averaging the results to get a composite estimate of $P^*(y)$ as shown in the denominator.

It is not at all clear that the refinement offered by the denominator of (2) would improve precision over the estimator defined in (1). Assume for example that Y is a single binary variable, whereas X is a vector of continuous variables. Decomposing the $P(y)$ as in the denominator of Eq. 1 would entail estimating all factors $P(y|x')$ and averaging the estimates. Estimating $P(y)$ from scratch, in contrast, may offer definite advantages, despite the fact that we have not borrowed any information from P.

We thus ask the following questions:

1. Given a relation R, which of its decompositions gain by borrowing and which does not?
2. Which relations R have a beneficial decomposition and which do not?
3. Given that borrowing is beneficial, can we quantify the benifit?

2 The Transfer Benefit Ratio (TBR)

To get a theoretical handle on the problem, let us take the simple problem of estimating the regression coefficient τ of Y on X in the chain model of Fig. 1.

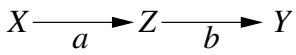

Figure 1: A chain model where b changes and a remains the same.

Here b is the regression coefficient of Y on Z,

$$b = \partial/dz E[Y|Z = z],$$

a is the regression coefficient of Z on X

$$a = \partial/\partial x E[Z|X = x],$$

and, based on the chain structure:

$$R = \tau = \partial/\partial x E[Y|X = x] = a * b.$$

Let us assume that we estimate a and b using Ordinary Least Square (OLS) on a large number (N_1) of cases from P. Now b changes to b^*, while a remains the same. How are we to estimate τ in P^*, if we can draw only a small number of samples (N_2) in the new environment?

We have two options:

1. We ignore the estimates obtained in the training environment and estimate τ from scratch, obtaining
$$\hat{\tau} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

2. We estimate a and b separately, and multiply their estimates, with a receiving samples from both environments and b from P^* only.

Let $\hat{a}, \hat{b}$, be respectively the OLS estimators of a, b. To measure the benefit of borrowing the estimate $\hat{a}$ from the training environment, we need to compare the efficiency of $\hat{a}\hat{b}$ to that of $\hat{\tau}$, recalling that $\hat{a}$ is estimated using N_1 training cases from P, and $\hat{\tau}$ and $\hat{b}$ are estimated using N_2 training cases from P^*.

The ratio of the asymptotic variances of these two estimators will measure the merit of transferring knowledge from one environment to another, and will be called here the Transfer Benefit Ratio (TBR).

This measure translates directly to improvement in the learning speed. When TBR is high, a small number of cases (N_2) in the novel environment would be sufficient to achieve a given precision, whereas a low TBR would require a high number of cases to achieve such precision.

Intuitively, the benefit of transfer would be more pronounced when the part shared by the two environments is noisy and the novel part is noiseless. Under such conditions, assessments of the target quantity τ are highly vulnerable to inaccuracies in estimating the relation between X and Z, and it is here that the training conducted in P can be most beneficial.

Exact analysis (see Appendix I) reveals that, for $N_2 \ll N_1$, the TBR is given by the following formula

$$\underset{N_2/N_1 \to 0}{TBR} = \frac{1 - \rho_b^2 \, \rho_a^2}{\rho_a^2(1 - \rho_b^2)}, \tag{3}$$

where ρ_a^2 and ρ_b^2 are the squared correlation coefficients

$$\rho_a^2 = \frac{cov^2(XZ)}{var(X)var(Z)} \quad \rho_b^2 = \frac{cov^2(YZ)}{var(Y)var(Z)}. \tag{4}$$

Equation (3) quantifies the intuition that transfer learning is more beneficial when the novelty between the two environments is almost deterministic (ρ_b approaches 1) so that the few observations conducted in the new environment would suffice to complete the adaptation.

Appendix I generalizes this result to any N_1/N_2 ratio and presents 3-dimensional charts of how the TBR varies with both the N_1/N_2 ratio and the statistics of X, Y, and Z. Remarkably, it shows that TBR is greater than unity even for $N_1 = N_2$. This means that there is benefit to the two-step estimation of τ (using the product $\hat{a}\hat{b}$ over the single step estimator $\hat{\tau}$, even when the environment does not change and we are faced with the problem of estimating τ given the chain model of Fig. 1. This phenomenon reflects a more general pattern in estimation: proper utilization of

modeling assumptions can improve estimation efficiency, provided those assumptions are valid (Cox, 1960; Pearl, 2012).

Clearly, this exercise is oversimplified in that it assumes just two linear relationships $X \to Z$ and $Z \to Y$ one invariant and one novel. Yet, such rudimentary analysis must be conducted to understand the speed-up provided by prior learning, the factors that determine this speed-up, and how to optimize those factors.

In more realistic situations, it is not at all clear that a speedup would be achieved regardless of the problem structure. In our example, we capitalized on the chain structure, which rendered X and Y conditional independent given Z. Under such conditions, the product estimator is superior to the one-shot estimator even when no environmental change takes place (i.e., $N_1 = N_2$). On the other hand, when $N_1 \gg N_2$, the benefit of transfer learning is realized even in the absence of independence constraints.

We have so far not considered the possibility of minimizing the number of variables needed to be measured in the new environment. Cases exist where, despite differences between P and P^*, R can be estimated entirely in the source environment, without taking any measurements in P^*. In other cases, some measurements in the new environments are needed, but the number of variables involved can be minimized by proper design (Pearl and Bareinboim, 2011).

3 Conclusions

We have demonstrated by simple examples that it is possible to quantify the benefit of borrowing information from previous learning, and that this benefit depends on the structure of the data generating model. This leaves open the general question of deciding, for any given relation R, how can it best benefit from previous learning, and how robust can it be to changes in the target environment? We conjecture that the understanding of such theoretical questions is necessary for designing algorithms that take maximum advantage of previous learning and spend minimum resources on re-learning that which could be borrowed.

Acknowledgement

I am indebted to Professor Jinyong Hahn for teaching me the secrets of asymptotic variance analysis. The 3-dimensional plots of Fig. 3 were produced by Elias Bareinboin.

This research was supported in parts by grants from Defense Advanced Research Projects Agency #W911NF-16-057, National Science Foundation #IIS-1302448, #IIS-1527490, and #IIS-1704932, and Office of Naval Research #N00014-17-S-B001.

Appendix I – Composition and Transfer in a Two-Stage Process

In experiments involving a two-stage process as in Fig. 1, Cox (1960) has shown that the estimated regression coefficient between treatment and response has a reduced variance if computed as a product of two estimates, one for each stage of the process. Below we summarize Cox's analysis and adapt it to the problem of information transfer across populations.

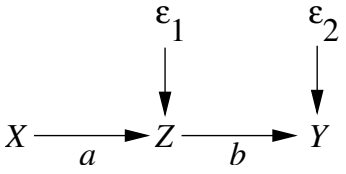

Figure 2: A two-stage process with intermediate variable Z.

The linear model depicted in Fig. 1 can be representd by the following structural equations:

$$z = ax + \epsilon_1,\ y = bz + \epsilon_2 \quad \text{with} \quad cov(x, \epsilon_1) = cov(x, \epsilon_2) = cov(\epsilon_1, \epsilon_2) = 0. \tag{5}$$

The process is depicted in Fig. 2. Our target of analysis is the regression coefficient of Y on X, i.e., the coefficient of x in the equation

$$y = \tau x + \epsilon_3 \quad \text{with} \quad cov(x, \epsilon_3) = 0. \tag{6}$$

As before, let $\hat{a}, \hat{b}$, and $\hat{\tau}$ be respectively the OLS estimators of a, b, τ. Cox showed that the asymptotic variance of $\hat{\tau}$ is greater than that of the product $\hat{a}\hat{b}$, or

$$var(\hat{\tau})/var(\hat{a}\hat{b}) \geq 1,$$

with equality holding only in pathological cases of perfect determinism. Specifically, he computed the n-sample variances to be:

$$var(\hat{\tau}) = [var(\epsilon_2) + b^2 var(\epsilon_1)]/nvar(X) \tag{7}$$

$$var(\hat{b}) = var(\epsilon_2)/n[a^2 var(X) + var(\epsilon_1)] \tag{8}$$

$$var(\hat{a}) = var(\epsilon_1)/nvar(X) \tag{9}$$

$$\begin{aligned} var(\hat{a}\hat{b}) &= a^2 var(\hat{b}) + b^2 var(\hat{a}) \\ &= \frac{a^2 var(X)(var(\epsilon_2) + b^2 var(\epsilon_1)) + b^2 var^2(\epsilon_1)}{nvar(X)[a^2 var(X) + var(\epsilon_2)]}. \end{aligned} \tag{10}$$

Thus,

$$\begin{aligned} \frac{var(\hat{\tau})}{var(\hat{a}\hat{b})} &= \frac{a^2 var(X) + var(\epsilon_1)}{a^2 var(X) + var(\epsilon_1) b^2 var(\epsilon_1)/[var(\epsilon_2) + b^2 var(\epsilon_1]} \\ &= \frac{a^2 var(X) + var(\epsilon_1)}{a^2 var(X) + var(\epsilon_1) F} \end{aligned} \tag{11}$$

which is greater than 1 because $F = b^2 var(\epsilon_1)/[var(\epsilon_2) + b^2 var(\epsilon_1)] \geq 1$.

The relation to transfer learning surfaces when a and b are estimated from two diverse populations, Π and Π^*. Let us assume that a is the same in the two populations, and is estimated by $\hat{a}$ using N_1 samples, pooled from both. b is presumed to be different, and is estimated by $\hat{b}$ using N_2 samples form Π^* alone. We need to compare the efficiency of estimating τ using the product $(\hat{a}\hat{b})$, to that of estimating τ directly, using N_2 samples from Π^*. The TBR, or the ratio of the asymptotic variances of these two estimators, can now be calculated as follows:

Keeping track of the number of samples entering each estimator, we have

$$var(\hat{\tau}; N_2) = var(\epsilon_2) + b^2 var(\epsilon_2)/N_2 var(X) \tag{12}$$

$$var(\hat{b}; N_2) = var(\epsilon_2)/N_2[a^2 var(X) + var(\epsilon_1)] \tag{13}$$

$$var(\hat{a}; N_1) = var(\epsilon_1)/N_1 var(X) \tag{14}$$

$$\begin{aligned} var(\hat{a}\hat{b}; N_1, N_2) &= a^2 var(\hat{b}) + b^2 var(\hat{a}) \\ &= \frac{N_1 a^2 var(X) var(\epsilon_2) + b^2 var(\epsilon_1)[a^2 N_2 var(x) + N_2 var^2(\epsilon_1)}{N_1 N_2 var(X)[a^2 var(X) + var(\epsilon_2)]}. \end{aligned} \tag{15}$$

Taking the ratio, we have

$$TBR = \frac{var(\hat{\tau}; N_2)}{var(\hat{a}\hat{b}; N_1, N_2)} \tag{16}$$

$$= \frac{N_1[a^2 var(X) + var(\epsilon_1)][var(\epsilon_2 + b^2 var(\epsilon_1)]}{a^2 var(X)[N_1 var(\epsilon_2) + N_2 b^2 var(\epsilon_1)] + N_2 b^2 var(\epsilon_1)} \tag{17}$$

$$= \frac{a^2 var(X) + var(\epsilon_1)}{a^2 var(X) F_1 + var(\epsilon_1) F_2}, \tag{18}$$

where

$$F_1 = \frac{var(\epsilon_2) + b^2 var(\epsilon_1) N_2/N_1}{var(\epsilon_2) + b^2 var(\epsilon_1)} \tag{19}$$

and

$$F_2 = N_2 b^2 var(\epsilon_1)/N_1[var(\epsilon_2) + b^2 var(\epsilon_1)]. \tag{20}$$

Since both F_1 and F_2 are smaller than 1 for $N_2 < N_1$, we conclude that the TBR is greater than one for $N_2 < N_1$, which means that it is beneficial to decompose the estimation task into two stages and use a higher number of samples, N_1, to estimate the shared component: $cov(X, Z)$.

Expression (17) can be simplified using correlation coefficients, as defined in Eq. (4) and gives:

$$TBR = \frac{1 - \rho_b^2 \rho_a^2}{\rho_a^2(1 - \rho_b^2) + \rho_b^2(1 - \rho_a^2) N_1/N_2} \tag{21}$$

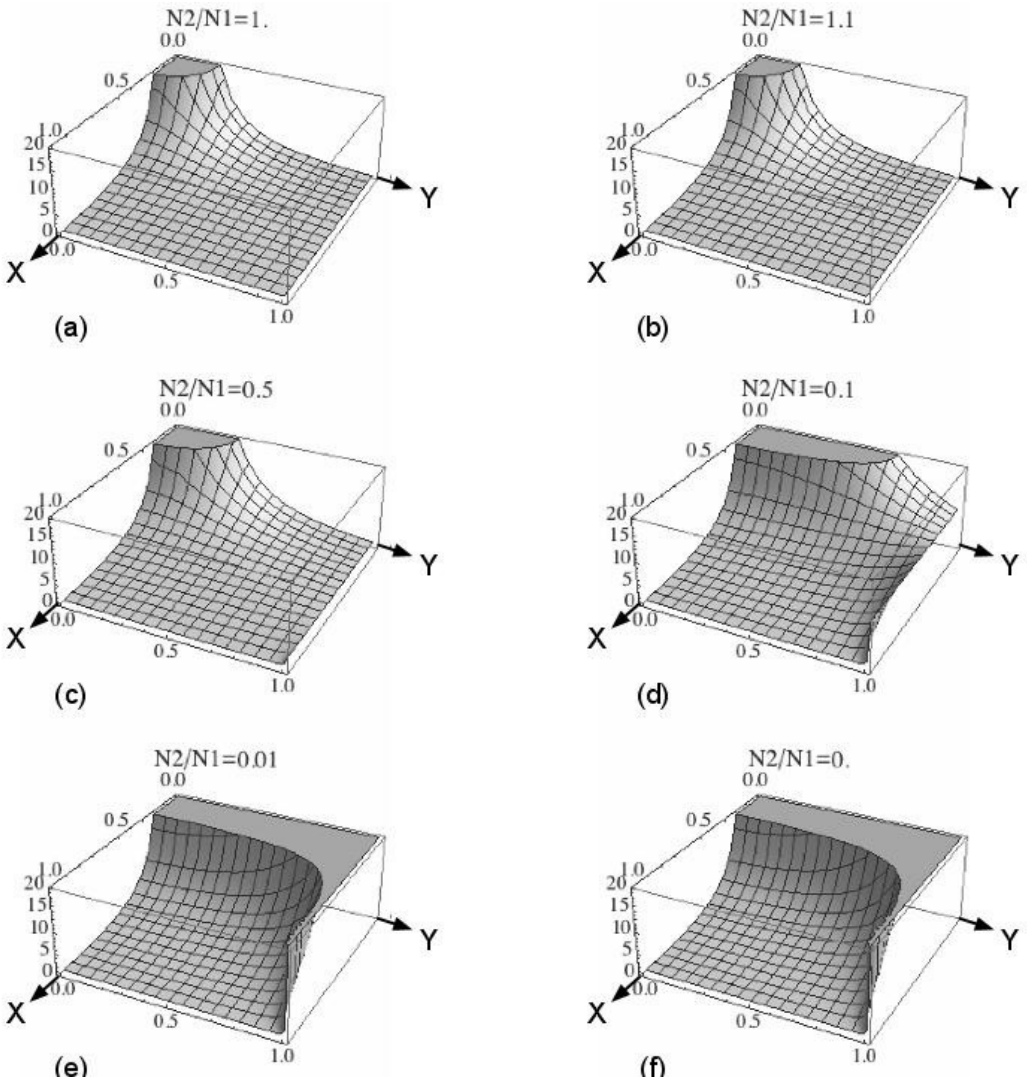

Figure 3: Illustrating the behavior of the Transfer Benefit Ratio (Eq. (21)) for different values of N_2/N_1 with X and Y axes representing ρ_a and ρ_b respectively. (a) $N_2/N_1 = 1$ (no transfer) TBR represents the benefit of decomposition alone. (c) $N_2/N_1 = 0.5$ represents data sharing between two equi-sampled studies. (d) $N_2/N_1 = 0.1$ showing a more pronounced benefit near the $\rho_b = 1$ region, where the $Z \to Y$ process becomes noiseless. (f) the limit case when $N_2/N_1 \to 0$, sharing marked benefit throughout the $\rho_b = 1$ and $\rho_a = 0$ regions, and no benefit near the $\rho_b = 0, \rho_a = 1$ corner.

The behavior of Eq. (21) for different values of N_2/N_1 is illustrated in Fig. 3(a, b, cd).

For $N_2 = N_1$ we obtain Cox's ratio (11) which quantifies the benefit of decomposition alone, without transfer. The ratio greatly exceeds one when both ρ_a^2 and ρ_b^2 are small, and approaches one when either or both of ρ_a^2 and ρ_b^2 are near one. This means that the benefit of decomposition is substantial if and only if both processes are noisy, whereas if either one of them comes close to being deterministic, decomposition has

no benefit.

This is reasonable; there is no benefit to decomposition unless Z brings new information which is not already in X or Y.

For $N_2 < N_1$, however, the TBR ratio represents the benefit of both decomposition and transfer. For the ratio to greatly exceed one we now need that both ρ_a^2 and ρ_b^2 be small. However, the TBR becomes unity (useless transfer) only when ρ_a is unity; $\rho_b = 1$ does not render it useless. It means that transfer is useless only when the process in agreement ($X \to Z$) is deterministic. Having disagreement on a deterministic mechanism does not make the transfer useless, as long as the process in agreement is corrupted by noise and can benefit from the extra samples from Π.

Indeed, taking the extreme case of deterministic $Z \to Y$ process ($\rho_b = 1$), there is a definite advantage to borrowing N_1 samples from the source population to estimate a and multiply it by b, rather than estimating c directly with the N_2 samples available at the target population. Two such samples can determine b precisely, and can hardly aid in the estimation of a.

The limit of TBR as N_1/N_2 increases indefinitely and represents transfer between a highly explored environment (large N_1) and one highly novel (low N_2). The limit of (21) reads:

$$\underset{N_2/N_1 \to 0}{TBR} = \frac{1 - \rho_b^2\ \rho_a^2}{\rho_a^2(1 - \rho_b^2)},$$

which establishes Eq. (3). It reveals that the Transfer Benefit Ratio will be most significant when the populations share noisy components (e.g., low correlation between X and Z) and differ in noiseless components (high correlation between Y and Z). Under such conditions, accurate assessment of the target quantity τ is highly vulnerable to inaccuracies in estimating a, and it is here that the large sample taken from Π can be most beneficial.

Appendix II – Extension to Saturated Models

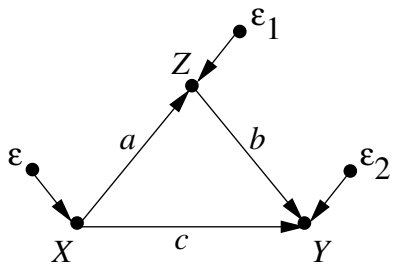

Figure 4: Saturated model in which Y depends on both X and Z.

In Appendix I, the benefit of transfer learning was demonstrated using an "over-identified" model (Fig. 2) which embodied the conditional independence $X \perp\!\!\!\perp Y|Z$, and for which the product estimator $\hat{a}\hat{b}$ was consistent. The question we analyze in

this Appendix is whether benefit can be demonstrated in "saturated" models as well (also called "just identified"), such as the one depicted in Fig. 4.

This model represents the following regression equations

$$Y = bz + cx + \epsilon_1$$
$$Z = ax + \epsilon_2$$

and the target quantity is again the total regression coefficient τ in the equation

$$y = \tau x + \epsilon \qquad \text{with } cov(x, \epsilon) = 0,$$

which is given by $\tau = cov(X, Y)/var(X) = c + ab$.

Again, τ can be estimated in two ways:

1. A one-shot way: compute the OLS regression of Y on X, call this estimator $\hat{\tau}$.
2. A two-shot way: compute the sum: $\hat{\theta} = \hat{c} + \hat{a}\hat{b}$ where $\hat{a}, \hat{b}$, and $\hat{c}$ are the OLS estimators of a, b, c respectively.

We now ask whether the variance of the composite estimator $\hat{\theta}$ will be smaller than the one-shot estimator $\hat{\tau}$, as we have seen in the over-identified model of Fig. 1. We further ask whether data sharing would be beneficial in case a is the same in both population while b and c are different.

Using an analysis similar to that of Appendix I, one can show that the answer to the first question is negative, while that of the second question is positive. In other words, we lose the intrinsic advantage of decomposition, but we can still draw advantage from data sharing if a is the same in the two populations. Formally, while the efficiency of the composite estimator $\hat{\theta} = \hat{a}\hat{b} + \hat{c}$ is identical to that of the one-shot estimator $\hat{\tau}$,[2] the variance of the former can be reduced if a is estimated using a larger sample than would be available to the one-shot estimator. In particular, assuming that $\hat{a}$ is estimated using N_1 samples and $\hat{b}, \hat{c}$, and $\hat{\tau}$ using N_2 samples, the asymptotic variances of $\hat{\theta}$ and $\hat{\tau}$, can be obtained by the delta method, and read:

$$var(\hat{c} + \hat{a}\hat{b}) = var(\epsilon_2)/N_2 var(X) + b^2 var(\epsilon_1)/N_1 var(X) \tag{22}$$
$$var(\hat{\tau}) = b^2 var(\epsilon_1) + var(\epsilon_2)]/N_2 var(X) \tag{23}$$

Consequently, the TBR is given by

$$\begin{aligned} TBR &= var(\hat{\tau})/var(\hat{c} + \hat{a}\hat{b}) \\ &= [1 - (1 - N_2/N_1) b^2 var(\epsilon_1)/(b^2 var(\epsilon_1) + var(\epsilon_2))]^{-1}. \end{aligned} \tag{24}$$

We see that for a single population and $N_1 = N_2$ decomposition in itself carries no benefit, ($TBR = 1$); the one-shot estimator is as good as the two-shot estimator. This

[2]The equality $\hat{\tau} = \hat{a}\hat{b} + \hat{c}$ is a mathematical identity, which holds for all sample sizes, not merely for asymptotic variance. I am indebted to Prof. Jinyong Hahn for demonstrating this fact. (See Hahn and Pearl, 2011.)

stands in contrast to the over-identified model of Fig. 1, for which the TBR was greater than unity (Eq. (21)) except in pathological cases. Moreover, the loss of benefit is not due to the disappearance of over-identification conditions from the model, but due to the composite estimator's failure to detect and utilize such conditions when they are valid. This can be seen from the fact that Eq. (24) (as well as the equality $\hat{\tau} = \hat{c} + \hat{a}\hat{b}$) remains unaltered even when $c = 0$. In other words, it is not the actual value of c that counts but the structure of the estimator we postulate. If we are ignorant of the fact that $c = 0$ in the actual model and go through the trouble of estimating τ by the sum $\hat{c} + \hat{a}\hat{b}$, instead of $\hat{a}\hat{b}$, the variance will be greater than what we would have gotten had we detected the model structure correctly and used the estimator $\hat{\tau} = \hat{a}\hat{b}$ to reflect our knowledge.

For $N_2/N_1 < 1$ however, the picture changes dramatically; Eq. (24) demonstrates a definite benefit to composite estimation ($TBR > 1$) which increases with $var(\epsilon_2)$. The intuition is similar to that given in Appendix I. When the $Z \to Y$ process was almost deterministic. we obtained $TBR > 1$. Here too, if the Y equation is deterministic, we can estimate it precisely with just a few samples (N_2) from P^* and use additional ($N_1 - N_2$) samples for estimating the noisy $X \to Z$ process which is common to both populations. The one-shot estimator will suffer from this noise if allowed only N_2 sample from P^*.

References

Cox, D. (1960). Regression analysis when there is prior information about supplementary variables. *The Journal of the Royal Statistical Society, Series B* **22** 172–176.

Hahn, J. and Pearl, J. (2011). Precision of composite estimators. Tech. Rep. R-388, <http://ftp.cs.ucla.edu/pub/stat_ser/r388.pdf>, Department of Computer Science, University of California, Los Angeles, CA. In preparation.

Pearl, J. (2012). Some thoughts concerning transfer learning, with applications to meta-analysis and data-sharing estimation. Tech. Rep. R-387, ⟨http://ftp.cs.ucla.edu/pub/stat_ser/r387.pdf⟩, Department of Computer Science, University of California, Los Angeles, CA. Working paper.

Pearl, J. and Bareinboim, E. (2011). Transportability of causal and statistical relations: A formal approach. In *Proceedings of the Twenty-Fifth Conference on Artificial Intelligence (AAAI-11)* (W. Burgard and D. Roth, eds.). AAAI Press, Menlo Park, CA, 247–254. Available at: ⟨http://ftp.cs.ucla.edu/pub/stat_ser/r372a.pdf⟩.

Pearl, J. and Bareinboim, E. (2014). External validity: From *do*-calculus to transportability across populations. *Statistical Science* **29** 579–595.