Working Paper

TECHNICAL REPORT
R-387
January 2012

Some Thoughts Concerning Transfer Learning, with Applications to Meta-analysis and Data-sharing Estimation

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

January 19, 2012

1 Introduction

A deeply entrenched axiom in the theory of learning states that the more one learns the easier it is to learn. In other words, the more proficient one becomes in performing familiar tasks, the easier it is to learn new tasks. This phenomenon, long recognized by psychologists and educators, has also been demonstrated in machine learning, especially in self-taught classification tasks, where unlabeled samples are used during training (Thrun, 1996; Ando and Zhang, 2005; Caruana, 1997). It has been shown, for example, that the performance of an image classifier can be improved by initially giving it unlimited access to unlabeled, randomly chosen images downloaded from the Internet (Raina et al., 2007). Evidently, at this initial stage of the process the classifier learns generic relationships, applicable to all visual patterns and, subsequently, when the classifier is trained with limited samples from a specific target domain (say, to distinguish elephants from rhinos), the classifier makes use of its previously learned knowledge to perform the target task more efficiently.

But what precisely is learned in that initial phase of the task, and why is it helpful in performing the final phase? What improvement, if any, can we expect from the first phase, and what needs to be assumed before we can quantify the improvement?

This paper aims to capture these and related questions using extremely simple, yet quantifiable models that may serve as building blocks for understanding the more complex phenomenon of transfer learning as it occurs in practice.

Whereas the justifications usually given to transfer learning are tailored to specific classification algorithms,[1] this paper aims to gain an understanding of this phenomenon

[1] See Daumé, III and Marcu, 2006; Blitzer et al., 2006; Ben-David et al., 2007; Daumé, III, 2007; Jiang and Zhai, 2007a; Satpal and Sarawagi, 2007; Jiang and Zhai, 2007b; Blitzer et al., 2008; Japkowicz and Stephen, 2002; Shimodaira, 2000; Heckman, 1979; Zadrozny, 2004; Zhu, 2005; Chapelle et al., 2006; Baxter,

from a general framework, starting from traditional estimation theory. In that way, the findings of the analysis will also be applicable to a broader class of tasks that could benefit from data sharing. These include, for example, Meta Analysis and ordinary statistical estimation.

In meta analysis, one attempts to combine results from many experimental and observational studies, usually conducted under different conditions and on diverse populations, so as to construct an aggregate measure of effect size that is "better," in some sense, than any one study in isolation. However, meta-analytical techniques treat each of the combined studies as a "black box" sensory device that provides noisy measurements of some unknown quantity; rarely do they explicate or utilize commonalities and differences among the combined studies to ensure that generalizations to the target population reflect the common and suppress the dissimilar. (See Rosenthal, 1995; Raudenbush and Bryk, 1985, for exceptions using moderator variables). The examples analyzed in this paper promise to enrich meta analysis with principled ways of handling differences and commonalities among pooled datasets.

But the issue of decomposition impacts many other areas of statistical estimation, especially those concerned with multi-component systems or those subject to time-varying changes. Such systems. invariably require local estimations of their components through studies that involve different sample sizes, taken under different conditions. Whenever we collect data from diverse environments we are presented with an opportunity of benefiting from their commonalities because, almost by definition, the common components of these environments receive more samples than any of their disparate parts and this increased sample size, if properly utilized, can lead to an improved overall performance. But what constitute a "shared" component and how "relevant" is that component to the overall task are questions that require formal definitions and careful mathematical analysis. The simple examples treated in the following sections should shed light on these general issues.

In Section 2 we address questions that emanates from attempts to formally define the notion of "commonalities" and "relevance." In Section 3 we map these questions onto the problem of improving precision in regression analysis and demonstrate the benefit of both decomposition and data sharing using a two-phase process previously analyzed by Cox (1960). In Section 4, we quantify these benefits of in saturated models, where decomposition in itself offers no benefit, yet decomposition combined with data sharing does. Finally, in Section 5, we speculate on the generality of these results and their applicability in transfer leaning, meta analysis and other data-sharing estimation tasks.

2 The Benefit of Transfer Learning

We have two populations, Π and Π^*, governed by two probability functions, P and P^*. Assume that we have taken data from both, and we wish to estimate some probabilistic relation R on Π^*. We ask whether there is an advantage to mixing the two datasets, as opposed to taking samples from P^* alone?[2]

1997; Ben-David and Schuller, 2003.

[2] This question is related to the problem of observational transportability (Pearl and Bareinboim, 2011) with the difference that we now pose no restriction on the variables that can be measured in Π^*. In

Intuitively data fusion can be justified when the two populations share some features in common, and when the shared features are significant in determining R. Assume that the target relation R can be decomposed into a set S of subrelations, which fall into two categories:

S_A - subrelations on which P and P^* agree, and

S_D - subrelations on which P and P^* disagree.

If we have 100 samples 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 accurate, and that should make the estimate of R less susceptible to sampling bias.

In this paper we explore the generality of this intuition, and formalize it in several models.

Example 1. *Let $R = P^*(x|y)$ and let $P(x)$ be in the agreement set, S_A. We have:*

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

which permits us to use the more accurately estimated $P(x)$ rather than rely solely on the noisy estimate of $P^*(x|y)$.

This simple example raises a fundamental question: Is it alway 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.

A related question arises from the fact that decompositions are not unique. Eq. 1, for example can also be written as:

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

which calls for estimating $P^*(y|x)$ and $P^*(y|x')$ for all x' at the target population, then average the estimates to get $P^*(y)$ in the denominator.

We ask whether it is always the case that estimating R^* through the r.h.s. of (1) or (2) will result in improved precision, compared to estimating $P^*(x|y)$ in one shot. Moreover, would the refinement offered by the denominator of (2) improve precision over the estimator defined in (1)?

That over-refinement may be risky can be seen in the following example.

Pearl and Bareinboim (2011), we needed to take data from Π because the cost of measuring V^* in Π^* was prohibitive. Now we assume that measurement cost is zero, and we ask: is there still an advantage to borrowing information from P, as opposed to taking samples from P^* alone?

Example 2. *Let $R = P^*(x|y)$ and let P and P^* be defined on three variables, X, Y and Z, such that $P^*(z) \neq P(z)$, and*[3]

$$P^*(x, y, z) = \frac{P^*(z)}{P(z)} P(x, y, z) = P^*(z)P(x, y|z) \tag{3}$$

Writing:

$$\begin{aligned} P^*(x, y) &= \sum_z P^*(z)P^*(x|z)P^*(y|x, z) \\ &= \sum_z P^*(z)P(x|z)P(y|x, z) \end{aligned} \tag{4}$$

we have:

$$P^*(x|y) = \frac{\sum_z P^*(z)P(x|z)P(y|x, z)}{\sum_{z,x'} P^*(z)P(x'|z)P(y|x', z)} \tag{5}$$

The question is: would we do better letting only $P^*(z)$ be estimated in Π^* as compared to estimating $R = P^*(x|y)$ directly in Π^*, borrowing no information from Π?

This is not a trivial question. Assume Z is a multidimensional vector, while X and Y are binary. The estimation of every term involving Z would be subject to a large sampling error, while skipping Z altogether amounts to estimating a simple conditional probability table between two binary variables. On the other hand, what if we have only 10 samples in Π^* and 10,000 in Π, it would seem reasonable to use Eq. (4) above and mix samples from both populations.

The question boils down to whether the difference between $P^*(x|y)$ and $P(x|y)$ is significant and whether a noisy correction of the source of this difference, $P^*(z)$, can reduce that difference.

To obtain a theoretical handle on this problem we take an extreme case and imagine that Π offers us infinite samples. We now ask: should we estimate R from scratch (i.e., in Π^*) or use some knowledge from Π? If the latter, is it clear what part should be borrowed from Π? We further assume that differences between Π and Π^* are encoded in the form of a selection diagram G (Pearl and Bareinboim, 2011) with the help of which we ask the following questions:

1. Given a selection diagram G, is it the case that no matter how we decompose R we always gain by borrowing sub-relations from Π?

2. If so, how do we find out a decomposition that permits us to benefit from Π.

3. If not, which decompositions gain by borrowing and which do not. Moreover, which relations have a beneficial decomposition and which do not.

4. Given that borrowing is beneficial, can we quantify the improvement?

[3] The problem may emerge in classification tasks, where x is the class label, y is a set of sensory inputs, and Z a set of features whose distribution $P(z)$ varies from environment to environment.

3 The Transfer Benefit Ratio (TBR)

A way of quantifying the benefit of estimator decomposition was outlined by Cox (1960). In experiments involving a two-stage process, Cox 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. He characterized the role of the intermediate variable as that of reducing "the effect of random variation entering the system after the treatments have exerted their effect."

Below we summarize Cox analysis of two-stage estimation and adapt it to the problem of information transfer across populations.

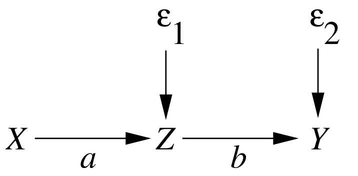

Figure 1: A two-stage process with intermediate variable Z.

Consider the linear model depicted in Fig. 1, representing the following structural equations:

$$z = ax + \epsilon_1,\ \ y = bz + \epsilon_2 \ \ \text{with} \ \ cov(x, \epsilon_1) = cov(x, \epsilon_2) = cov(\epsilon_1, \epsilon_2) = 0 \tag{6}$$

Our target of analysis is the regression coefficient of Y on X, i.e., the coefficient of x in the equation

$$y = \tau x + \epsilon_3 \ \ \text{with} \ \ cor(x, \epsilon_3) = 0 \tag{7}$$

Clearly, $\tau = ab = cov(XY)/var(X)$, and can be estimated by OLS, using

$$\hat{\tau} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

Let $\hat{a}, \hat{b}$, be respectively the OLS estimators of a, b. 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:

$$\begin{aligned}
var(\hat{\tau}) &= [var(\epsilon_2) + b^2 var(\epsilon_1)]/nvar(X) && (8)\\
var(\hat{b}) &= var(\epsilon_2)/n[a^2 var(X) + var(\epsilon_1)] && (9)\\
var(\hat{a}) &= var(\epsilon_1)/nvar(X) && (10)\\
var(\hat{a}\hat{b}) &= a^2 var(\hat{b}) + b^2 var(\hat{a}) &&\\
&= \frac{a^2 var(X)(var(\epsilon_2) + b^2 var(\epsilon_1)) + b^2 var^2(\epsilon_1)}{nvar(X)[a^2 var(X) + var(\epsilon_2)]} && (11)
\end{aligned}$$

Thus,

$$\begin{aligned}\frac{var(\hat{\tau})}{var(\hat{a}\hat{b})} &= \frac{a^2var(X) + var(\epsilon_1)}{a^2var(X) + var(\epsilon_1)b^2var(\epsilon_1)/[var(\epsilon_2) + b^2var(\epsilon_1]} \\ &= \frac{a^2var(X) + var(\epsilon_1)}{a^2var(X) + var(\epsilon_1)F} \qquad (12)\end{aligned}$$

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

The relation to transfer learning surfaces when a and b are estimated from two diverse populations. 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 ratio of the asymptotic variances of these two estimators will measure the merit of transferring knowledge from one population to another, and will be called here the Transfer Benefit Ratio (TBR).

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

$$\begin{aligned}var(\hat{\tau}; N_2) &= var(\epsilon_2) + b^2var(\epsilon_2)/N_2var(X) \qquad (13)\\ var(\hat{b}; N_2) &= var(\epsilon_2)/N_2[a^2var(X) + var(\epsilon_1)] \qquad (14)\\ var(\hat{a}; N_1) &= var(\epsilon_1)/N_1var(X) \qquad (15)\\ var(\hat{a}\hat{b}; N_1, N_2) &= a^2var(\hat{b}) + b^2var(\hat{a}) \\ &= \frac{N_1a^2var(X)var(\epsilon_2) + b^2var(\epsilon_1)[a^2N_2var(x) + N_2var^2(\epsilon_1)}{N_1N_2var(X)[a^2var(X) + var(\epsilon_2)]} \qquad (16)\end{aligned}$$

Taking the ratio, we have

$$\begin{aligned}TBR &= \frac{var(\hat{\tau}; N_2)}{var(\hat{a}\hat{b}; N_1, N_2)} \qquad (17)\\ &= \frac{N_1[a^2var(X) + var(\epsilon_1)][var(\epsilon_2 + b^2var(\epsilon_1)]}{a^2var(X)[N_1var(\epsilon_2) + N_2b^2var(\epsilon_1)] + N_2b^2var(\epsilon_1)} \qquad (18)\\ &= \frac{a^2var(X) + var(\epsilon_1)}{a^2var(X)F_1 + var(\epsilon_1)F_2} \qquad (19)\end{aligned}$$

Where

$$F_1 = \frac{var(\epsilon_2) + b^2var(\epsilon_1)N_2/N_1}{var(\epsilon_2) + b^2var(\epsilon_1)} \qquad (20)$$

and

$$F_2 = N_2b^2var(\epsilon_1)/N_1[var(\epsilon_2) + b^2var(\epsilon_1] \qquad (21)$$

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 (18) can be simplified using correlation coefficients, and gives:

$$TBR = \frac{1 - \rho_b^2 \rho_a^2}{\rho_a^2(1 - \rho_b^2) + \rho_b^2(1 - \rho_a^2)N_1/N_2} \tag{22}$$

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

$$\rho_a^2 = \frac{cov^2(XZ)}{var(X)var(Z)} \quad \rho_b^2 = \frac{cov^2(YZ)}{var(Y)var(Z)} \tag{23}$$

For $N_2 = N_1$ we obtain Cox's ratio (18) 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 ($X \rightarrow Z$) 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 $Z \rightarrow Y$ 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.

It is interesting to examine the limit of this ratio as N_1/N_2 increases indefinitely, representing a mixture of two environments, one highly familiar (large N_1) and one highly novel (low N_2). The limit of (22) reads:

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

and 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 c is highly vulnerable to inaccuracies in estimating the relation between X and Z, and it is here that the many samples taken from Π can be most beneficial.

These behaviors are illustrated in Fig. 2(abcd) for different values of N_2/N_1.

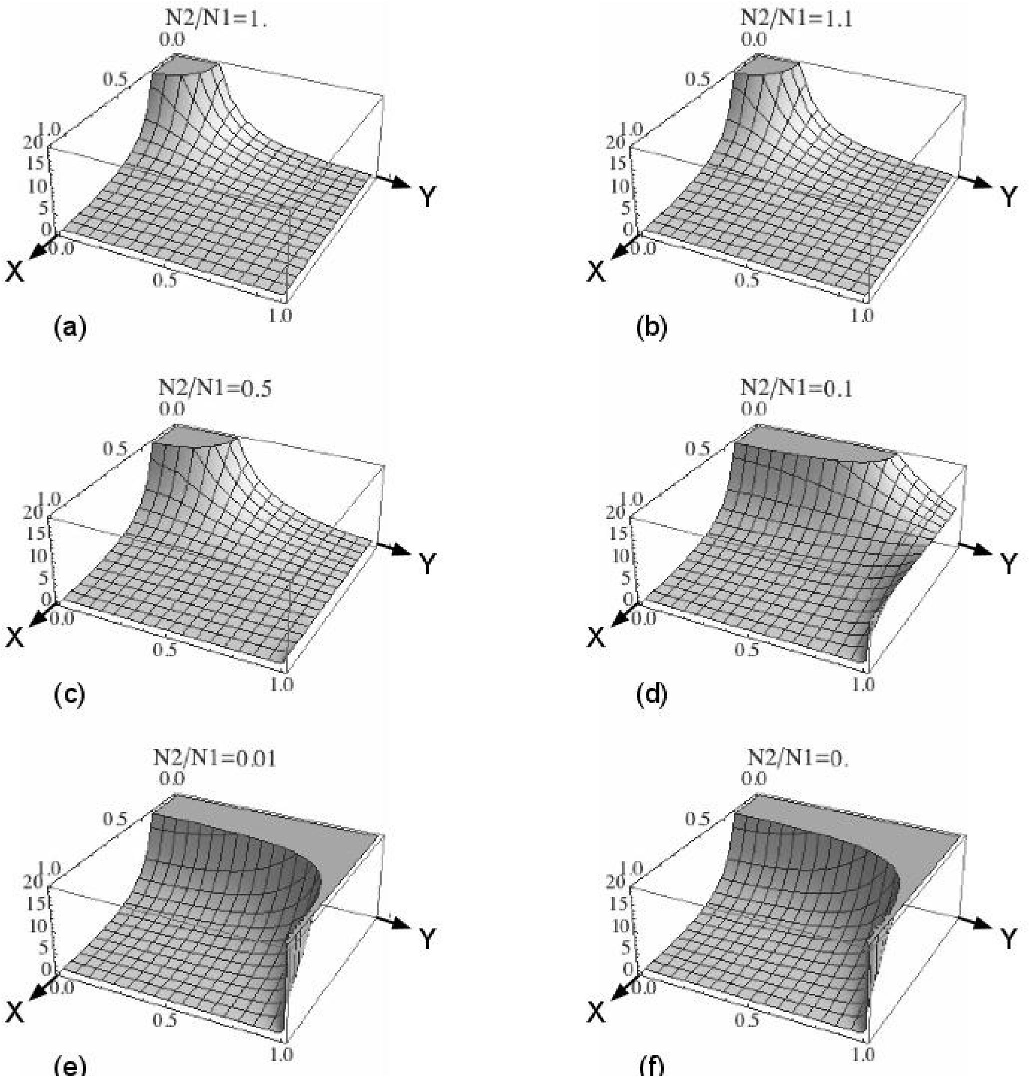

Figure 2: Illustrating the behavior of the Transfer Benefit Ratio (Eq. (22)) for different values of N_2/N_1 with X and Y axes representing ρ_a and ρ_b respectively. (a) $N_2/N_1 = 1$ (no transfer) TBR represents the benefit of decomposition alone. (c) $N_2/N_1 = 0.5$ represents data sharing between 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.

4 Extension to Saturated Models

In Section 3, the benefit of transfer learning was demonstrated using an "over-identified" model (Fig. 1) 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 explore in this section is whether benefit can be demonstrated in "saturated" models as well (also called "just identified"), such as the one depicted in Fig. 3.

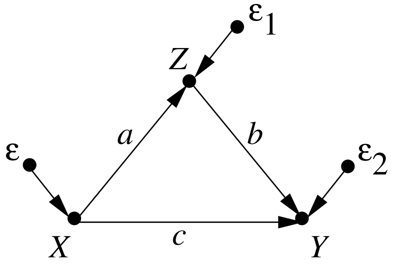

Figure 3: Saturated model in which Y depends on both X and Z.

This model represents the following regression equations

$$
\begin{aligned}
Y &= bz + cx + \epsilon_1 \\
Z &= ax + \epsilon_2
\end{aligned}
$$

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 Section 3, 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 $\hat{\tau}$,[4] 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:

$$var(\hat{c} + \hat{a}\hat{b}) = var(\epsilon_2)/N_2 var(X) + b^2 var(\epsilon_1)/N_1 var(X) \tag{24}$$

$$var(\hat{\tau}) = b^2 var(\epsilon_1) + var(\epsilon_2)]/N_2 var(X) \tag{25}$$

[4] 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.)

Consequently, the Transfer Benefit Ratio is given by

$$\begin{aligned} TBR &= var(\hat{\tau})/var(\hat{c}+\hat{a}\hat{b}) \\ &= [1-(1-N_2/N_1)b^2 var(\epsilon_1)/(b^2 var(\epsilon_1)+var(\epsilon_2)]^{-1} \end{aligned} \tag{26}$$

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 stands in contrast to the over-identified model of Fig. 1, for which the TBR was greater than unity (Eq. (22)) 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. (26) (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. (26) demonstrates a definite benefit to composite estimation ($TBR > 1$) which increases with $var(\epsilon_2)$. The intuition is similar to that given in Section 3. When the $Z \rightarrow Y$ process was almost deterministic. we obtained $TBR > 1$ (even when Cox ratio was one). 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 $X \rightarrow Z$ process which is common to both populations. The one-shot estimator will suffer from this noise if allowed only N_2 sample from P^*.

5 Discussion and applications

In this section we summarize the preceding observations and speculate on their implications in several areas of statistical estimation.

5.1 The benefit of decomposition

We note that the explanation given by Cox (1960) for the improved precision of the composite estimator $\hat{a}\hat{b}$ (Fig. 1) is not complete. The role of the intermediate variable in this model is not to reduce "the effect of random variation entering the system" but to enable us to exploit the conditional independence $Y \perp\!\!\!\perp X | Z$ that constrains the data. This can be seen by examining the composite estimator $\hat{c} + \hat{a}\hat{b}$ of the saturated model of Fig. 3, in which the intermediate variable also reduces random variation entering the system and, yet, it does not result in improved precision over the single-shot estimator. The two are in fact identical (see footnote 4) even when $c = 0$ in the data-generating model.

Hahn and Pearl (2011) show that the superiority of composite estimators persists in non-parametric models, for example, taking X, Y, and Z to be binary variables in the model of Fig. 1. The superiority disappears however in the binary version of Fig. 3, which is saturated. It is tempting therefore to generalize and conjecture that composite estimators will be found superior whenever they exploit structural information that is glossed over by their non-composite counterparts.

5.2 Application to machine learning

The model described in Fig. 3 can serve as a skeleton for analyzing transfer learning in classification tasks such as those described in the introduction section. Assume that X_1 and X_2 are two sets of features and Y is a class label. Consider the following 3-phase learning task:

1. An unsupervised "incubation" phase, where the classifier is given N_1 unlabeled samples from $P(x_1, x_2)$.
2. A supervised training phase, where N_2 labeled samples are drawn from $P(y|x_1, x_2)$, and
3. An execution phase, where the classifier is presented with instances of X_1 and is asked to classify Y accordingly.

In this task, the results of Section 3 are directly applicable, and Eq. 26 provides us, in the case of linear models, an explanation and a quantification of the benefit expected from the incubation phase. To make these results applicable to practical learning tasks, further generalizations are required from linear to multi-variate semi-parametric models, as well as a translation from variances to probabilities of misclassification.

The setup outlined above is not standard in machine learning, where it is usually assumed that all features are presented to the classifier at the execution phase. However, if we regard X_2 to be low-level sensory inputs and X_1 to be high-level abstract features (i.e., stochastic functions of X_2), it is not unreasonable to expect that, at the execution phase, the classifier will use only the features in X_1 for its decision function, while delegating the $P(x_1|x_2)$ computation to auxiliary, special purpose machinery. In such settings, having an accurate estimate of the relation $P(x_2|x_1)$ serves to construct a more informed decision function d: $X_1 \rightarrow Y$.

5.3 Applications to surrogate-endpoint analysis

In the health sciences, a problem area where transfer learning is both critical and implicit is the analysis of *surrogate endpoint*. At its core, the problem concerns a randomized trial where one seeks "endpoint surrogate," namely, a variable Z that would allow good predictability of an outcome Y for both treatment ($X = 1$) and control ($X = 0$). In the words of Ellenberg and Hamilton (1989) "investigators use surrogate endpoints when the endpoint of interest (e.g., life expectancy or survival) is too difficult and/or expensive to measure routinely and when they can define some other, more readily measurable, endpoint (a biological marker), which is sufficiently well correlated with the first to justify its use as a substitute."

Putting aside debates on whether good prediction, void of causal mediation is sufficient for the task (see Prentice, 1989; Joffe and Green, 2009; Pearl and Bareinboim, 2011; Pearl, 2011), the problem fits well into the model of Figs. 1 and 3. The task involves three phases:

1. A small number (N_2) of samples are drawn gfrom $P(y, z|x)$ (with X randomized) to determine if Z promises to be a good predictor of Y.
2. A larger number (N_1) of samples are drawn from $P(z|x)$ to determine if X has a sufficiently strong influence on Z.
3. Given the two tests above and their associated estimates, $P(z|x)$ and $P(y, z|x)$, an estimate is then sought of $P(y|x)$ which gives the causal effect of X on Y, since X is randomized throughout.

In view of the relative sizes of the two samples involved, it is clear that a composite estimate of $P(y|x)$ would be superior to the one-shot estimator which can be obtained from the N_2 samples of phase-1. The analysis of Section 4 also teaches us that this superiority increases when the output process is less noisy than the $X \rightarrow Z$ process. This unfortunately is not the case in most practical applications.

5.4 Towards a compositional meta-analysis

Consider the diagrams depicted in Fig. 4, each representing a study conducted on a different population and under a different set of conditions.[5] Solid circles represent variables that were measured in the respective study and hollow circles variables that remained unmeasured. An arrow ■→ represents an external influence affecting a mechanism by which the study population is assumed to differ from the target population Π^*, shown in Fig. 4(a). For example, Fig. 1(c) represents an observational study on population Π_c in which variables X, Z and Y were measured, W was not measured and the prior probability $P_c(Z)$ differs from that of the target population $P^*(Z)$. Diagrams (b)–(f) represent observational studies while (g)–(j) stand for experimental studies with X randomized (hence the missing arrows into X).

Despite differences in populations, measurements and conditions, each of the studies may provide information that bears on the target relation $R(\Pi^*)$ which, in this example, we take to be the causal effect of X on Y, $P^*(y|do(x))$ or; given the structure of Fig. 4(a),

$$R(\Pi^*) = P^*(y|do(x)) = \sum_z P^*(y|x, z)P^*(z).$$

While $R(\Pi^*)$ can be estimated directly from some of the studies, (e.g., (g)) and indirectly from others (e.g., (b) and (d)), it cannot be estimated from those studies in which the population differs substantially from Π^*, (e.g., (c), (e), (f)). The estimates of R provided by the former studies may differ from each other due to sampling variations and measurement errors, and can be aggregated in the standard tradition of meta analysis. The latter studies, however, should not be averaged with the former, since they do not provide unbiased

[5] These diagrams were defined in Pearl and Bareinboim (2011), where they are called *selection diagrams.*

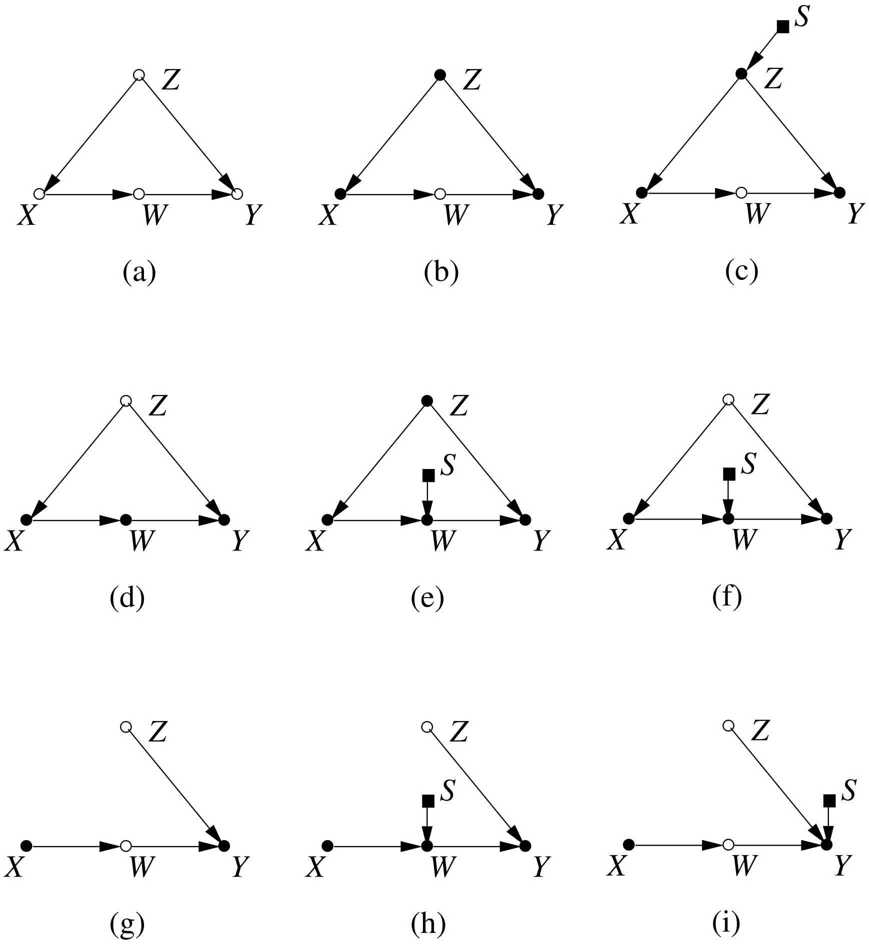

Figure 4: Diagrams representing 8 studies ((b)–(i)) conducted under different conditions on different populations, aiming to estimate the causal effect of X on Y in the target population, shown in 4(a).

estimates of R. They are not totally useless, though, for they can provide information that renders the former estimates more accurate. For example, although we cannot identify R from study 4(c), since $P_c(z)$ differs from $P^*(z)$, we can nevertheless use the estimates of $P_c(x|z), P_c(y|z,x)$ that 4(c) provides to improve the accuracy of $P^*(x|z)$ and $P^*(y|z,x)$[6] which may be needed for estimating R by indirect methods. For example, $P^*(y|z,x)$ is needed in study 4(b) if we use the estimator $R = \sum_z P^*(y|x,z)P^*(z)$, while $P^*(x|z)$ is needed if we use the inverse probability estimator $R = \sum_z P^*(x,y,z)/P^*(x|z)$.

Similarly, consider the randomized studies depicted in 4(h) and 4(i). None is sufficient for identifying R in isolation, yet taken together, they permit us to borrow $P_i(w|do(x))$

[6] The absence of boxed arrows into X and Y in Fig. 4(c) implies the equalities

$$P_c(x|z) = P^*(x|z) \text{ and } P_c(y|z,x) = P^*(y|z,x).$$

from 4(i) and $P_h(y|w, do(x))$ from 4(h) and synthesize a bias-free estimator:

$$\begin{aligned} R &= \sum_w P^*(y|w, do(x))P^*(w|do(x)) \\ &= \sum_w P_h(y|w, do(x)P_i(w|do(x)) \end{aligned}$$

The challenge of synthetic meta analysis is to take a collection of studies, annotated with their respective selection diagrams (as in Fig. 4), and construct an estimator of a specified relation $R(\Pi^*)$ that makes maximum use of the samples available, by exploiting the commonalities among the populations studied and the target population Π^*. As the relation $R(\Pi^*)$ changes, the synthesis strategy will change as well.

I speculate that data-pooling strategies based on the principles outlined in this paper will one day replace the blind methods currently used in meta analysis.

References

Ando, R. and Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research* **6** 1817–1853.

Baxter, J. (1997). A bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine Learning* **28** 7–39.

Ben-David, S., Blitzer, J., Crammer, K. and Pereira, F. (2007). Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems 19* (B. Schölkopf, J. Platt and T. Hoffman, eds.). MIT Press, Cambridge, Massachusetts, USA, 137–144.

Ben-David, S. and Schuller, R. (2003). Exploiting task relatedness for mulitple task learning. In *Proceedings of Computational Learning Theory (COLT)*.

Blitzer, J., Crammer, K., Kulesza, A., Pereira, F. and Wortman, J. (2008). Learning bounds for domain adaptation. In *Advances in Neural Information Processing Systems 20* (J. Platt, D. Koller, Y. Singer and S. Roweis, eds.). MIT Press, Cambridge, Massachusetts, USA, 105–134.

Blitzer, J., McDonald, R. and Pereira, F. (2006). Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Sydney, Australia.

Caruana, R. (1997). Multitask learning. *Machine Learning* **28** 41–75.

Chapelle, O., Schölkopf, B. and Zien, A. (eds.) (2006). *Semi-Supervised Learning*. MIT Press, New York, NY.

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.

Daumé, III, H. (2007). Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. Omnipress, Madison, WI, 256–263.

Daumé, III, H. and Marcu, D. (2006). Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence* **26** 101–126.

Ellenberg, S. and Hamilton, J. (1989). Surrogate endpoints in clinical trials: Cancer. *Statistics in Medicine* **8** 405–413.

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.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica* **47** 153–161.

Japkowicz, N. and Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis* **6** 429–450.

Jiang, J. and Zhai, C. (2007a). Ainstance weighting for domain adaptation in NLP. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. Prague, Czech Republic.

Jiang, J. and Zhai, C. (2007b). A two-stage approach to domain adaptation for statistical classifiers. In *Proceedings of the ACM 16th Conference on Information and Knowledge Management*. ACM, Lisbon, Portugal.

Joffe, M. and Green, T. (2009). Related causal frameworks for surrogate outcomes. *Biometrics* **65** 530–538.

Pearl, J. (2011). Principal stratification a goal or a tool? *The International Journal of Biostatistics* **7**. Article 20, DOI: 10.2202/1557-4679.1322. Available at: http://www.bepress.com/ijb/vol7/iss1/20.

Pearl, J. and Bareinboim, E. (2011). Transportability of causal and statistical relations: A formal approach. Tech. Rep. R-372-A, <http://ftp.cs.ucla.edu/pub/stat_ser/r372a.pdf>, University of California Los Angeles, Computer Science Department, CA. Forthcoming, 2011 AAAI Proceedings.

Prentice, R. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine* **8** 431–440.

Raina, R., Battle, A., Lee, H., Packer, B. and Ng, A. Y. (2007). Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th international conference on Machine learning*. ICML '07, ACM, New York, NY, USA.
URL `http://doi.acm.org/10.1145/1273496.1273592`

Raudenbush, S. W. and Bryk, A. S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics* **10** 75–98.

Rosenthal, R. (1995). Writing meta-analysic reviews. *Psycho. Bull.* **118** 183–192.

Satpal, S. and Sarawagi, S. (2007). Domain adaptation of conditional probability models via feature subsetting. In *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases*. Warsaw, Poland.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference* **90** 227–244.

Thrun, S. (1996). Is learning the n-th thing any easier than learning the first? In *Advances in Neural Information Processing Systems*. The MIT Press, 640–646.

Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. In *Proceedings of the 21st Annual International Conference on Machine Learning*. Banff, Canada.

Zhu, X. (2005). Semi-supervised learning literature survey. Tech. Rep. 1530, University of Wisconsin-Madison, WI.