15 Semi-supervised & transfer learning

We have seen examples of both supervised learning (regression, classification) and unsupervised learning (density estimation, clustering) so far this semester. Here we study semi-supervised learning, a mixed setting where we have access to two datasets, one labeled and one unlabeled. Specifically, we observe labeled data

$$(X_1, Y_1,), ..., (X_{N_\ell}, Y_{N_\ell})$$

but also additional unlabeled data

$$(X_{N_\ell+1}, ..., X_n),$$

and we want to estimate a regression function or mean or other quantity. A common underlying motivation is that labels can be costly to obtain, while unlabeled data is cheap – so that N_ℓ might be relatively small, whereas $n - N_\ell$ is large. It is then natural to ask: can the additional unlabeled data help, and if so for what tasks? We also discuss transfer learning, a related paradigm where we have data from a source distribution that differs from the target distribution (and data from the target may be labeled or unlabeled or even missing entirely); there our goal is to construct procedures that perform well in the target distribution.

Remark 1. An important fact about the semi-supervised data structure is that it can also be represented as observations $(Z_1, ..., Z_n)$ for

$$Z_i = (X_i, R_i, R_i Y_i)$$

where R_i is an indicator for whether we see a labeled outcome Y_i for observation i. Note when $R_i = 0$ we do not get to observe Y_i and instead only see the arbitrary placeholder $R_i Y_i = 0$. This framing helps connect more clearly to missing data problems, where there is a long history of work on both identification and estimation under missing at random and other assumptions [Robins et al., 1994, Tsiatis, 2006, van der Laan and Robins, 2003]. Connections to missing data are pervasive in semi-supervised inference and transfer learning, but are relatively rarely emphasized in the machine learning literature.

1 Identification

Perhaps the simplest and most fundamental problem in semi-supervised learning is to estimate the mean $\mathbb{E}(Y)$. Note when the observations $Z = (X, R, RY)$ are iid from some overall distribution $\mathbb{P}$, this is the mean across both the labeled and unlabeled data

$$\mathbb{E}(Y) = \mathbb{E}(Y \mid R = 1)\mathbb{P}(R = 1) + \mathbb{E}(Y \mid R = 0)\mathbb{P}(R = 0).$$

We can of course estimate $\mathbb{E}(Y \mid R = 1)$ unbiasedly with the sample average of the outcomes in the labeled data, but without further assumptions we have no information at all about the mean in the unlabeled data, $\mathbb{E}(Y \mid R = 0)$. Therefore in general the parameter $\mathbb{E}(Y)$ is not identified, i.e., we could not know what it was even if we had an infinite sample, and were given the distribution $\mathbb{P}$ exactly (since it tells us nothing about $\mathbb{E}(Y \mid R = 0)$). Relatedly, if we do not make any further assumptions, the extra unlabeled data is useless to us even in estimating the mean in the labeled data $\mathbb{E}(Y \mid R = 1)$. These fundamental limits point to the crucial role of adding extra structural assumptions, which is sometimes underappreciated in the literature.

1.1 Missing at random assumptions

Since the mean $\mathbb{E}(Y)$ is not identified if we do not assume anything at all about how the unlabeled data relate to the labeled data, we necessarily need additional structure to make use of the unlabeled data. The most common assumption is the *missing completely at random (MCAR)* assumption that

$$R \perp\!\!\!\perp (X, Y), \tag{MCAR}$$

i.e., that whether you are labeled/unlabeled is completely independent of your features and underlying outcomes. This really encodes two separate assumptions: first, that $R \perp\!\!\!\perp X$, so that $d\mathbb{P}(x \mid R = 1) = d\mathbb{P}(x \mid R = 0)$, i.e., the covariate distribution is exactly the same in the labeled and unlabeled groups, and second that $R \perp\!\!\!\perp Y \mid X$, so that labeling is independent of the underlying outcome given features (i.e., the conditional outcome distribution given features is the same whether you are labeled or not).

The $R \perp\!\!\!\perp Y \mid X$ part of Assumption MCAR is untestable, since the outcomes are unobserved whenever $R = 0$. However the $R \perp\!\!\!\perp X$ part is testable; we could simply estimate the feature distributions in the labeled and unlabeled data and compare. In fact, it turns out the mean $\mathbb{E}(Y)$ is identified even if we drop the testable $R \perp\!\!\!\perp X$ part of Assumption MCAR. Thus we have some hope at reliably estimating $\mathbb{E}(Y)$ even under the strictly weaker *missing at random (MAR)* assumption

$$R \perp\!\!\!\perp Y \mid X. \tag{MAR}$$

However, there are some extra difficulties related to estimation that come with this weaker Assumption MAR, due to the fact that the covariate distributions can be different in the labeled and unlabeled data, since $R \not\perp\!\!\!\perp X$, and $\mathbb{P}(R = 1 \mid X)$ can be challenging to estimate.

Remark 2. MCAR/MAR terminology was coined by Rubin [1976].

2 Mean estimation

Here we consider estimating the mean $\psi = \mathbb{E}(Y)$ under Assumption MCAR. Note that in this case we have

$$\mathbb{E}(Y) = \mathbb{E}(Y \mid R = 1) = \mathbb{E}(Y \mid R = 0)$$

so the means in the labeled and unlabeled data are equal, since Assumption MCAR also implies $R \perp\!\!\!\perp Y$.

Naive estimator

The simplest estimator just omits the unlabeled data and averages the labeled outcomes, i.e.,

$$\widehat{\psi}_\ell = \frac{1}{N_\ell} \sum_{i:R_i=1} Y_i = \frac{1}{n} \sum_{i=1}^{n} \frac{R_i Y_i}{N_\ell / n}.$$

In the first result below we verify this estimator $\widehat{\psi}_\ell$ is unbiased and characterize its variance.

Remark 3. In what follows we assume observations Z_i are iid with $\mathbb{P}(R_i = 1) = \pi$, so that the number of labeled observations $N_\ell = \sum_i R_i \sim \text{Bin}(n, \pi)$ is random. In a slight abuse of notation we sometimes write $\mathbb{P}(R = 1) = \pi = n_\ell/n$ to keep in mind the ratio scaling, i.e., simply defining $n_\ell = n\pi$. An alternative option would be to assume $\mathbb{P}\{(R_1, ..., R_n) = (r_1, ..., r_n)\} = 1/\binom{n}{N_\ell}$ so that $\sum_i R_i = N_\ell$ would be a fixed sample size, but then the R_i indicators would have some dependence. Similar results can be obtained in either case.

Remark 4. Although we assume $\mathbb{P}(R_i = 1) = \pi$, nothing in our analyses requires this to be bounded below by a constant, e.g., we could have $\pi = n_\ell/n = 1/\sqrt{n}$ accommodating much more unlabeled data.

Theorem 1. *Assume MCAR. Then $\widehat{\psi}_\ell$ is conditionally unbiased given $R^n = (R_1, ..., R_n)$ and*

$$var(\widehat{\psi}_\ell \mid R^n) = \frac{var(Y)}{N_\ell}.$$

Proof. Let $R^n = (R_1, ..., R_n)$ as in the theorem statement. For the bias we have

$$\mathbb{E}(\widehat{\psi}_\ell \mid R^n) = \frac{1}{n}\sum_{i=1}^n \frac{R_i \mathbb{E}(Y_i \mid R_i = 1)}{N_\ell/n} = \frac{1}{n}\sum_{i=1}^n \frac{R_i \psi}{N_\ell/n} = \psi \frac{\sum_i R_i}{N_\ell} = \psi$$

using iterated expectation and Assumption MCAR. For the variance we have

$$\text{var}(\widehat{\psi}_\ell \mid R^n) = \frac{1}{N_\ell^2}\sum_{i=1}^n R_i \text{var}(Y_i \mid R_i = 1) = \frac{\text{var}(Y \mid R = 1)}{N_\ell}$$

by the iid assumption. □

Therefore the variance of $\widehat{\psi}_\ell$ is what we might expect: the marginal outcome variance scaled by the number of labeled samples. Note we can also write this as

$$\frac{\text{var}(Y)}{N_\ell} = \frac{\mathbb{E}\{\text{var}(Y \mid X)\} + \text{var}\{\mathbb{E}(Y \mid X)\}}{N_\ell}$$

by the law of total variance.

Remark 5. Of course conditional unbiasedness implies marginal unbiasedness as well. For the variance, note we have

$$\begin{aligned}\text{var}(\widehat{\psi}_\ell) &= \mathbb{E}\{\text{var}(\widehat{\psi}_\ell \mid R^n)\} + \text{var}\{\mathbb{E}(\widehat{\psi}_\ell \mid R^n)\} = \mathbb{E}\{\text{var}(\widehat{\psi}_\ell \mid R^n)\} \\ &= \text{var}(Y)\mathbb{E}\{\mathbb{1}(N_\ell > 0)/N_\ell\} \leq \frac{2\text{var}(Y)}{n_\ell}\end{aligned}$$

using the law of total variance, and Lemma A.2 of Devroye et al. (1996) in the last step.

Remark 6. Note that $(R_1, ..., R_n) \mid N_\ell$ is uniform over all binary n-vectors with N_ℓ ones, and $\mathbb{E}(R_i \mid N_\ell) = N_\ell/n$. Conditionally given N_ℓ the labels R_i, $i = 1, ..., n$, have some dependence.

Doubly robust estimator

Now the question is: can this simple estimator be improved? In fact it can. Suppose we have an estimator $\widehat{\mu}(x)$ of the regression function

$$\mu(x) = \mathbb{E}(Y \mid X = x, R = 1) = \mathbb{E}(Y \mid X = x)$$

built from a separate independent training sample (e.g., using sample splitting). Then we can construct a semiparametric / doubly robust / debiased / one-step / double machine learning estimator

$$\widehat{\psi} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{R_i\{Y_i - \widehat{\mu}(X_i)\}}{N_\ell/n} + \widehat{\mu}(X_i) \right]$$

(see van der Laan and Robins [2003], Tsiatis [2006], Chernozhukov et al. [2018], Kennedy [2022], etc. for more history and details). The precise form of this estimator comes from semiparametric theory involving influence functions. The following result shows how it retains the unbiasedness of the naive estimator while typically improving variance.

Theorem 2. *Assume MCAR and let $\widehat{\mu}$ be constructed on a separate independent sample. Then the estimator $\widehat{\psi}$ is conditionally unbiased given $N_\ell = \sum_i R_i$ and*

$$var(\widehat{\psi} \mid N_\ell) \leq \frac{\mathbb{E}\{var(Y \mid X)\}}{N_\ell} + \frac{var\{\mu(X)\}}{n} + \left(1 - \frac{N_\ell}{n}\right) \frac{\|\widehat{\mu} - \mu\|^2}{N_\ell}$$

so that

$$N_\ell \Big\{ var(\widehat{\psi}_\ell \mid N_\ell) - var(\widehat{\psi} \mid N_\ell) \Big\} \;\geq\; \left(1 - \frac{N_\ell}{n}\right) \Big[var\{\mu(X)\} - \|\widehat{\mu} - \mu\|^2 \Big].$$

Proof. In this proof all expectations and variances are conditional given the training sample used to construct $\widehat{\mu}$. For unbiasedness note that we have

$$\begin{aligned}
\mathbb{E}(\widehat{\psi} \mid N_\ell) &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\frac{R_i\{Y_i - \widehat{\mu}(X_i)\}}{N_\ell/n} + \widehat{\mu}(X_i) \mid N_\ell \right] \\
&= \frac{\mathbb{E}(R_i \mid N_\ell)\mathbb{E}\{\mu(X) - \widehat{\mu}(X)\}}{N_\ell/n} + \mathbb{E}\{\widehat{\mu}(X)\} = \mathbb{E}\{\mu(X)\} = \psi
\end{aligned}$$

where in the second line we used iterated expectation and $\mathbb{E}(R_i \mid N_\ell) = N_\ell/n$, and in the last equality that $\mathbb{E}\{\mathbb{E}(Y \mid X, R = 1)\} = \mathbb{E}\{\mathbb{E}(Y \mid X)\} = \mathbb{E}(Y) = \psi$. Now consider the variance. Let $\widehat{\phi} = \frac{R(Y - \widehat{\mu})}{N_\ell/n} + \widehat{\mu}$ so that $\widehat{\psi} = \mathbb{P}_n(\widehat{\phi})$. Then we have

$$\text{var}(\widehat{\phi} \mid N_\ell) = \text{var}(\widehat{\phi} - \phi + \phi \mid N_\ell) = \text{var}(\phi \mid N_\ell) + \text{var}(\widehat{\phi} - \phi \mid N_\ell)$$

since the covariance is zero, noting

$$\begin{aligned}
\text{cov}(\widehat{\phi} - \phi, \phi \mid N_\ell) &= \mathbb{E}\left[\left(\frac{R}{N_\ell/n} - 1 \right) \left(\mu - \widehat{\mu} \right) \left\{ \frac{R(Y - \mu)}{N_\ell/n} + \mu - \psi \right\} \,\middle|\, N_\ell \right] \\
&= \mathbb{E}\left[\left(\frac{1}{N_\ell/n} - 1 \right) (\mu - \widehat{\mu}) \frac{R(Y - \mu)}{N_\ell} + 0 \,\middle|\, N_\ell \right] = 0
\end{aligned}$$

by iterated expectation and that $\mathbb{E}(R_i \mid N_\ell) = N_\ell/n$. Now the first term in the variance is

$$\begin{aligned}\operatorname{var}\left\{\frac{R(Y-\mu)}{N_\ell/n}+\mu \;\middle|\; N_\ell\right\} &= \mathbb{E}\left[\left\{\frac{R(Y-\mu)}{N_\ell/n}+(\mu-\psi)\right\}^2 \;\middle|\; N_\ell\right] = \mathbb{E}\left\{\frac{R(Y-\mu)^2}{(N_\ell/n)^2}+(\mu-\psi)^2 \;\middle|\; N_\ell\right\} \\ &= \mathbb{E}\left\{\frac{R\operatorname{var}(Y \mid X)}{(N_\ell/n)^2}+(\mu-\psi)^2 \;\middle|\; N_\ell\right\} = \frac{\mathbb{E}\{\operatorname{var}(Y \mid X)\}}{N_\ell/n}+\operatorname{var}\{\mu(X)\}.\end{aligned}$$

using iterated expectation and MCAR so that $\operatorname{var}(Y \mid X, R=1) = \operatorname{var}(Y \mid X)$. For the second term, since $\widehat{\phi}-\phi = \left(\frac{R}{n_\ell/n}-1\right)(\mu-\widehat{\mu})$, we have

$$\begin{aligned}\operatorname{var}(\widehat{\phi}-\phi \mid N_\ell) &\le \mathbb{E}\{(\widehat{\phi}-\phi)^2 \mid N_\ell\} = \mathbb{E}\left\{\left(\frac{R}{N_\ell/n}-1\right)^2\left(\mu-\widehat{\mu}\right)^2 \;\middle|\; N_\ell\right\} \\ &= \left\{N_\ell/n\left(\frac{1-N_\ell/n}{N_\ell/n}\right)^2+(1-N_\ell/n)\right\}\|\widehat{\mu}-\mu\|^2 = \left(\frac{n-N_\ell}{N_\ell}\right)\|\widehat{\mu}-\mu\|^2.\end{aligned}$$

Now by iterated expectation, MCAR, and the iid assumption we have

$$\operatorname{cov}(\widehat{\phi}_i, \widehat{\phi}_j \mid N_\ell) = \mathbb{E}\left[\left\{\frac{R_i(Y_i-\widehat{\mu}_i)}{N_\ell/n}+\widehat{\mu}_i-\psi\right\}\left\{\frac{R_j(Y_j-\widehat{\mu}_i)}{N_\ell/n}+\widehat{\mu}_j\right\} \;\middle|\; N_\ell\right] = 0$$

so that

$$\begin{aligned}\operatorname{var}(\widehat{\psi} \mid N_\ell) &= \frac{1}{n^2}\sum_{i=1}^n \operatorname{var}(\widehat{\phi}_i \mid N_\ell)+\frac{1}{n^2}\sum_{i\neq j}\operatorname{cov}(\widehat{\phi}_i, \widehat{\phi}_j \mid N_\ell) = \frac{1}{n^2}\sum_{i=1}^n \operatorname{var}(\widehat{\phi}_i \mid N_\ell) \\ &\le \frac{1}{n}\left[\frac{\mathbb{E}\{\operatorname{var}(Y \mid X)\}}{N_\ell/n}+\operatorname{var}\{\mu(X)\}+\left(\frac{n-N_\ell}{N_\ell}\right)\|\widehat{\mu}-\mu\|^2\right]\end{aligned}$$

which gives the variance bound result. The comparison of the variances is obtained by simply subtracting this variance bound from the variance result in the previous theorem. □

Note that the doubly robust estimator $\widehat{\psi}$ is guaranteed to have smaller conditional variance than $\widehat{\psi}_\ell$ whenever

$$\|\widehat{\mu}-\mu\|^2 \le \operatorname{var}\{\mu(X)\},$$

a mild condition. Even for the naive estimator $\widetilde{\mu} = \mathbb{E}(Y)$ that does not use any covariate information we have $\|\widetilde{\mu}-\mu\|^2 = \operatorname{var}\{\mu(X)\}$, so this condition would typically hold as long as the covariates are even a little predictive of the outcome. Also recall there are universally consistent estimators for which $\mathbb{E}\|\widehat{\mu}-\mu\|^2 \to 0$ for any distribution with bounded second moment.

2.1 Missing at random

Under the weaker MAR assumption, which allows $R \not\perp\!\!\!\perp X$ so the covariate distributions can be different in the labeled and unlabeled data, the estimator from the previous section would be biased. A similar doubly robust estimator can be used, though, of the form

$$\widehat{\psi} = \frac{1}{n}\sum_{i=1}^n\left[\frac{R_i\{Y_i-\widehat{\mu}(X_i)\}}{\widehat{\pi}(X_i)}+\widehat{\mu}(X_i)\right]$$

where $\widehat{\pi}(x)$ is an estimate of $\pi(x) = \mathbb{P}(R=1 \mid X=x)$, called the *propensity score*. Zhang et al. [2023] study the MAR semi-supervised learning setup.

3 Related topics

3.1 Prediction-powered inference

"Prediction-powered inference" [Angelopoulos et al., 2023a,b] is a recent take on semi-supervised learning with a slightly different emphasis, including focusing on:

1. M-estimation rather than mean estimation, i.e., the goal is to estimate some $\theta_0 = \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}\{\ell_\theta(X, Y)\}$ for a loss function $\ell_\theta(X, Y)$, instead of say $\mathbb{E}(Y)$,
2. inference rather than estimation, e.g., confidence sets for θ,
3. access to a black-box predictor f that can be evaluated on the labels.

For M-estimation one first estimates $\mathbb{E}\{\ell_\theta(X, Y)\}$ (which is a mean estimation problem) and then minimizes it over θ. For inference one can use central limit theorem arguments to prove asymptotic normality. Although access to a black-box f is interesting motivation, one can also just view $f(X_i)$ as another feature, so the covariates would be $\{X_i, f(X_i)\}$ instead of just X_i. It is also important to point out that others have also considered these issues, e.g., Chakrabortty and Cai [2018], Zhang et al. [2019], etc.

The proposed PPI estimator of $\mathbb{E}\{\ell_\theta(X, Y)\}$ is

$$\frac{1}{n}\sum_{i=1}^{n} \frac{R_i}{N_\ell/n}\Big\{\ell_\theta(X_i, Y_i) - \ell_\theta(X_i, f(X_i))\Big\} + \frac{1}{n}\sum_{i=1}^{n} \frac{(1 - R_i)\ell_\theta(X_i, f(X_i))}{(n - N_\ell)/n}$$

which is similar to the doubly robust estimator from the previous section (albeit omitting the labeled $R = 1$ observations in the second term, which will result in a loss of efficiency).

3.2 Regression & classification

In mean estimation-style problems, the target parameter explicitly depends on the marginal distribution of features X, so it is intuitive that improvement could be possible with access to unlabeled data (under MAR). However, for regression and classification problems (where target quantities are conditional on X) one typically needs to add even more structure beyond MAR or MCAR, to somehow link the target to the covariate distribution.

Wasserman and Lafferty [2007] consider semi-supervised nonparametric regression under the MAR assumption. They consider manifold assumptions, i.e., that the distribution of X lives on a lower dimensional manifold, and semi-supervised smoothness assumptions, i.e., that there is some relationship between $\mu(x)$ and the covariate density $p(x)$ (e.g., that their Hessians are the same or a simple function of one another). They show that faster rates are possible with the latter smoothness, but not manifold assumptions.

Rigollet [2007] considered a clustering-based assumption, motivated by the heuristic that points in the same cluster should have the same label. His formalization is that if a density level set $L_\lambda = \{x : p(x) > \lambda\}$ consists of k disjoint connected components C_j, then the Bayes classifier $\mathbb{1}\{\mu(x) > 1/2\}$ should be constant on each C_j, i.e., the decision boundary can only lie in low-density regions. In the semi-supervised setting Rigollet [2007] went on to show that one can achieve faster rates under this assumption, using a classifier that relies on estimating the level set and its connected components.

4 Transfer learning

In transfer learning problems, one observes two datasets from two different distributions, say a source P_s and target P_t, and the goal is to construct procedures that perform well in the target distribution P_t. Of course one could always just only use P_t (and without any assumptions, this is all one could do), but the hope is to gain some extra benefit from having access to P_s (either merely to improve performance, or because the observations from P_t may be unlabeled or missing). Just as in semi-supervised learning, in order to make use of the source sample from P_s we need to link P_s and the target P_t in some way. There are three main approaches, detailed below: posterior drift, covariate shift, and label shift.

Remark 7. Consider the simple setting where the samples from P_s and P_t are both labeled. We can use the same missing data formalization as before and write the data as an iid sample $(Z_1, ..., Z_n) \sim \mathbb{P}$ with

$$Z_i = (R_i, X_i, Y_i)$$

where $R_i \in \{s, t\}$ denotes whether the observation is in the source or the target, i.e., we can write $d\mathbb{P}(X, Y \mid R = r) = dP_r(X, Y)$ for $r \in \{s, t\}$.

4.1 Posterior drift

One way to relate P_s and P_t is through the *posterior drift* model. This assumes the marginal covariate distributions are the same, i.e.,

$$R \perp\!\!\!\perp X \iff d\mathbb{P}(X \mid R = s) = d\mathbb{P}(X \mid R = t)$$

but the conditional outcome distributions can differ, i.e., $d\mathbb{P}(Y \mid X, R = s) \neq d\mathbb{P}(Y \mid X, R = t)$. However there needs to be some connection between the conditional outcome distributions to get any benefit for regression, for example.

Cai and Pu [2024] recently studied nonparametric regression under posterior drift, with the goal of using the $R = s$ source data to help estimate the target regression $\mu_t(x) = \mathbb{E}(Y \mid X = x, R = t)$. They linked μ_s to μ_t by assuming μ_t is ϵ-close in L_1 distance to μ_s plus a polynomial of order k, i.e., there exists a polynomial $f(x)$ of order k such that

$$\int \left| \mu_t(x) - \Big\{ \mu_s(x) + f(x) \Big\} \right| \, dx \leq \epsilon.$$

The gap ϵ is called the *bias strength*; when $\epsilon = 0$ they call μ_s a *perfect reference* for μ_t, otherwise an imperfect reference. They also assume μ_s and μ_t are Holder-smooth with their own indices. Under this model $\mathcal{P}$ they show the minimax risk is lower bounded as

$$\inf_{\widehat{\mu}_t} \sup_{P \in \mathcal{P}} \mathbb{E}_P \|\widehat{\mu}_t - \mu_{t,P}\|_P \gtrsim n_{\max}^{-1/(2+d/\beta_{\max})} + \Big\{ \min(\epsilon, n_t^{-1/(2+d/\beta_t)}) n_t^{-1/(2+d/\beta_t)} \Big\}^{1/2} + \frac{1}{\sqrt{n_t}}$$

where β_t is the smoothness of μ_t, n_t is the number of target samples, $n_{\max} = \max(n_s, n_t)$, and $\beta_{\max} = \max(\beta_t, \beta_s)$. For this rate to be faster than the $n_t^{-1/(2+d/\beta_t)}$ rate that one would get only using the target sample, one needs small bias strength $\epsilon < n_t^{-1/(2+d/\beta_t)}$ and either more source smoothness $\beta_s > \beta_t$ or more source samples $n_s > n_t$. Cai and Pu [2024] also give a local polynomial-based "confidence thresholding" estimator that matches the minimax lower bound up to log factors.

4.2 Covariate shift

Another way to relate P_s and P_t is through the *covariate shift* model. This assumes the conditional outcome distributions are the same under the source and target, i.e.,

$$Y \perp\!\!\!\perp R \mid X \iff d\mathbb{P}(Y \mid X, R = s) = d\mathbb{P}(Y \mid X, R = t),$$

but the covariate distributions can differ, i.e., $d\mathbb{P}(X \mid R = s) \neq d\mathbb{P}(X \mid R = t)$. When the target distribution is unlabeled, we can reframe the data structure as iid observations of

$$Z_i = (X_i, S_i, S_i Y_i)$$

where now $S_i = \mathbb{1}(R_i = s)$ is an indicator for whether the observation comes from the source distribution, and in the target where $S_i = 0$ we do not get to see outcomes, we only see the placeholder $S_i Y_i = 0$. Covariate shift is now just precisely a missing at random assumption on the missing indicator S, i.e.,

$$S \perp\!\!\!\perp Y \mid X$$

so we can use tools from missing data to identify and estimate quantities defined via the target distribution. For example, suppose we want to estimate the mean in the target distribution

$$\psi = \mathbb{E}(Y \mid S = 0).$$

Without any assumptions, we have no information about Y in the $S = 0$ group because the $S = 0$ group is unlabeled. However, under covariate shift this mean is identified as

$$\begin{aligned}\psi &= \mathbb{E}(Y \mid S = 0) = \mathbb{E}\{\mathbb{E}(Y \mid X, S = 0) \mid S = 0\} \\ &= \mathbb{E}\{\mathbb{E}(Y \mid X, S = 1) \mid S = 0\}\end{aligned}$$

where the first line uses iterated expectation and the second the covariate shift assumption (as long as $0 < \mathbb{P}(S = 1 \mid X) < 1$). This quantity could be estimated by regressing Y on X in the source data to get $\widehat{\mu}_1(x)$, an estimate of $\mu_1(x) = \mathbb{E}(Y \mid X = x, S = 1)$, and then imputing it for the target observations and averaging, i.e.,

$$\widehat{\psi}_{pi} = \sum_{i:S_i=0} \widehat{\mu}(X_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{(1 - S_i)\widehat{\mu}(X_i)}{N_t/n}$$

where $N_t = \sum_i (1 - S_i)$ is the number of observations in the target sample. This is a plug-in-style estimator, which could be improved with a doubly robust estimator of the form

$$\frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{S_i}{N_t/n}\right)\frac{1 - \widehat{\pi}(X_i)}{\widehat{\pi}(X_i)}\Big\{Y_i - \widehat{\mu}_1(X_i)\Big\} + \frac{1 - S_i}{N_t/n}\widehat{\mu}_1(X_i)\right]$$

for $\widehat{\pi}(x)$ an estimate of the source propensity score $\pi(x) = \mathbb{P}(S = 1 \mid X = x)$. This kind of estimator also appears in the causal inference literature, when estimating treatment effects on the treated [Kennedy et al., 2015] or transporting causal effects [Zeng et al., 2023].

4.3 Label shift

The label shift model assumes conditional covariate distributions are the same, i.e.,

$$R \perp\!\!\!\perp X \mid Y \iff d\mathbb{P}(X \mid Y, R = s) = d\mathbb{P}(X \mid Y, R = t)$$

but marginal outcome distributions can differ. See Maity et al. [2020] for details.

References

A. N. Angelopoulos, S. Bates, C. Fannjiang, M. I. Jordan, and T. Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023a.

A. N. Angelopoulos, J. C. Duchi, and T. Zrnic. Ppi++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023b.

T. T. Cai and H. Pu. Transfer learning for nonparametric regression: Non-asymptotic minimax analysis and adaptive procedure. *arXiv preprint arXiv:2401.12272*, 2024.

A. Chakrabortty and T. Cai. Efficient and adaptive linear regression in semi-supervised settings. *The Annals of Statistics*, 46(4):1541–1572, 2018.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. M. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

E. H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. 2022.

E. H. Kennedy, A. Sjölander, and D. Small. Semiparametric causal inference in matched cohort studies. *Biometrika*, 102(3):739–746, 2015.

S. Maity, Y. Sun, and M. Banerjee. Minimax optimal approaches to the label shift problem. *arXiv preprint arXiv:2003.10443*, 1, 2020.

P. Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research*, 8(7), 2007.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427): 846–866, 1994.

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

A. A. Tsiatis. *Semiparametric Theory and Missing Data*. New York: Springer, 2006.

M. J. van der Laan and J. M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer, 2003.

L. Wasserman and J. Lafferty. Statistical analysis of semi-supervised regression. *Advances in Neural Information Processing Systems*, 20, 2007.

Z. Zeng, E. H. Kennedy, L. M. Bodnar, and A. I. Naimi. Efficient generalization and transportation. *arXiv preprint arXiv:2302.00092*, 2023.

A. Zhang, L. D. Brown, and T. T. Cai. Semi-supervised inference: General theory and estimation of means. *The Annals of Statistics*, 47(5):2538–2566, 2019.

Y. Zhang, A. Chakrabortty, and J. Bradic. Double robust semi-supervised inference for the mean: selection bias under mar labeling with decaying overlap. *Information and Inference: A Journal of the IMA*, 12(3):2066–2159, 2023.