ROBUST IMPORTANCE-WEIGHTED CROSS-VALIDATION UNDER SAMPLE SELECTION BIAS

Wouter M. Kouw [1,3*], *Jesse H. Krijthe* [2], *Marco Loog* [3]

[1] Department of Electrical Engineering, Eindhoven University of Technology
[2] Institute for Computing and Information Sciences, Radboud University Nijmegen
[3] Department of Intelligent Systems, Delft University of Technology

ABSTRACT

Cross-validation under sample selection bias can, in principle, be done by importance-weighting the empirical risk. However, the importance-weighted risk estimator produces suboptimal hyperparameter estimates in problem settings where large weights arise with high probability. We study its sampling variance as a function of the training data distribution and introduce a control variate to increase its robustness to problematically large weights.

Index Terms— Sample selection bias, cross-validation.

1. INTRODUCTION

Classification under sample selection bias refers to settings where training data is collected locally, but the goal is to generalize to a larger target population [1, 2, 3]. For example, data collected in one hospital with the aim of generalizing to a national population or data collected from a process that is assumed to be stationary, but in reality drifts slowly over time. To control for selection bias, each sample is weighted by a factor that matches its observation probability in the training data distribution to that in the target population, a process known as *importance-weighting* [4, 2]. However, importance-weighting underperforms and can even fail, if the weights are too large [2]. The weights directly scale the estimator's sampling variance (see Section 3.2) and sampling skewness [5]. Importance-weighting is therefore not suited to problem settings where large weights occur with high probability. It needs to be more robust to be considered a practical solution.

One could consider weight truncation as a means of avoiding large weights, but that introduces a substantial bias in the estimator [6]. Alternatively, variance reduction techniques incorporate additional information about the statistic of interest to avoid introducing a bias. Control variates are such a technique, requiring an additional function of the random variable that correlates well with the statistic of interest and whose expected value is known [7]. Because of

the correlation, the statistic is known to rise – or fall in case of negative correlation – above its expected value whenever the control variate does. The deviation of the control variate from its expected value can be subtracted from the statistic, thereby reducing the statistic's deviation from its expectation (i.e. sampling variance) [7, 8]. In importance-weighting, the importance weights themselves constitute an additional function that correlates with the importance-weighted loss and whose expected value is known [9]. An importance-weighted control variate can reduce the sampling variance of an importance-weighted risk estimator, thereby increasing its robustness to large weights.

We are interested in cross-validation under sample selection bias. In standard cross-validation, the risk is repeatedly estimated and averaged to obtain an estimate of the expected generalization error [10]. With this estimate, optimal hyperparameters can be selected. Cross-validation becomes difficult under sample selection bias, as the estimator essentially over-fits to the distribution from which the training data is generated [11]. Importance-weighted cross-validation has been explored as a solution [4], but is known to produce suboptimal hyperparameter estimates in problem settings with large weights [5]. We argue for including the weight-based control variate in the importance-weighted risk estimator such that importance-weighted cross-validation is more robust to large weights.

Our contributions can be summarized as follows:

- We apply a controlled importance-weighted risk estimator to a cross-validation procedure.

- We empirically show that the inclusion of the control variate increases robustness to large weights.

In the remainder of the paper we present the problem setting in more detail (Section 2), discuss the importance-weighted risk estimator (Section 3) and show how the control variate reduces sampling variance (Section 4). We then run a cross-validation experiment to demonstrate the effectiveness of the controlled estimator (Section 5). We briefly discuss conclusions and extensions in Section 6.

---

*Majority of work done while WK was at Delft University of Technology.

2. SAMPLE SELECTION BIAS

Consider an input space $\mathcal{X} \subseteq \mathbb{R}^D$, and an output space $\mathcal{Y}$, with $\mathcal{Y} = \{-1, +1\}$ for binary classification or $\mathcal{Y} \subseteq \mathbb{N}$ for multi-class. The population, or *target domain*, is a probability distribution $p_{\mathcal{T}}$ defined over this pair of spaces. The distribution of the training data $p_{\mathcal{S}}$, gathered under sample selection bias, is referred to as the *source domain*. Under sample selection bias, the data distributions differ $p_{\mathcal{S}}(x) \neq p_{\mathcal{T}}(x)$ with $p_{\mathcal{S}}(x)$ having a smaller variance than $p_{\mathcal{T}}(x)$, but the posterior distributions remain equal $p_{\mathcal{S}}(y \mid x) = p_{\mathcal{T}}(y \mid x)$. Data points from the source domain are referred to as x_i with labels y_i, while target domain data is referred to as z_j with labels u_j. The challenge is to use labeled source data $\{(x_i, y_i)\}_{i=1}^n$ and unlabeled target data $\{z_j\}_{j=1}^m$ to predict target labels u_j.

Consider a function h, parameterized by θ, that maps a data point to a real number, $h_\theta : \mathcal{X} \to \mathbb{R}$. The real-valued outputs are predictions of classes and we therefore refer to h as a *classifier*. Predictions are evaluated using a loss function $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$, which compares the prediction to the true label. The expected loss is called the *risk* of the classifier, $R(h_\theta)$ and the average loss is called the empirical risk $\hat{R}(h_\theta)$.

2.1. An example setting

Throughout the paper, we illustrate the behavior of estimators with a running example of a classification problem under sample selection bias. The posterior distribution in both domains is a unit cumulative normal distribution, $p_{\mathcal{T}}(y \mid x) = \Phi(yx \mid 0, 1)$. The target data distribution is taken to be a unit normal distribution $p_{\mathcal{T}}(x) = \mathcal{N}(0, 1)$, while the source data distribution is set to $\mathcal{N}(-1, \gamma)$. Figure 1 plots the data distributions, for $\gamma = 1/\sqrt{2}$. Note that $p_{\mathcal{S}}$ has higher probability mass between $x = -3$ and $x = -1/2$, implying these values will tend to be observed more often in the source data set than in the target set.
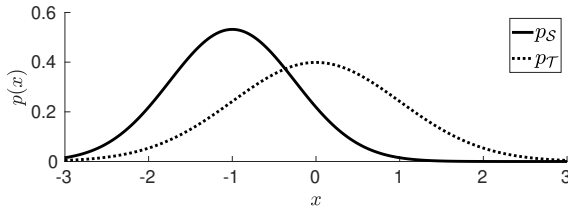

Fig. 1. Data distributions in the example setting.

In order to evaluate risk, a choice of loss function and classifier is needed. For this example, we use a quadratic loss, $\ell(h_\theta(x), y) = (x\theta - y)^2$, with $\theta = 1/\sqrt{\pi}$ as coefficient.

3. IMPORTANCE-WEIGHTING

Empirical risk minimization describes a classifier's performance by its expected loss. The risk function R integrates the loss ℓ over a joint distribution p, and is therefore domain-specific. We are interested in minimizing the target risk $R_{\mathcal{T}}$:

$$R_{\mathcal{T}}(h_\theta) = \int_{\mathcal{X}} \int_{\mathcal{Y}} \ell(h_\theta(x), y) p_{\mathcal{T}}(x, y) \, \mathrm{d}y \mathrm{d}x, \qquad (1)$$

which can be estimated with the sample average over data drawn from the target domain:

$$\hat{R}_{\mathcal{T}}(h_\theta) = \frac{1}{m} \sum_{j=1}^m \ell(h_\theta(z_j), u_j). \qquad (2)$$

However, this estimator cannot be used, since we lack target labels u.

We are interested in estimators that do *not* depend on u. One possibility is the importance-weighted risk estimator, which re-formulates the target risk to include the source distribution: $\iint \ell(h_\theta(x), y) p_{\mathcal{S}}(x, y) [p_{\mathcal{T}}(x, y)/p_{\mathcal{S}}(x, y)] \mathrm{d}y \mathrm{d}x$. Under sample selection bias, $p_{\mathcal{T}}(y \mid x) = p_{\mathcal{S}}(y \mid x)$, which simplifies the re-formulation to:

$$R_{\mathcal{W}}(h_\theta) = \int_{\mathcal{X}} \int_{\mathcal{Y}} \ell(h_\theta(x), y) \frac{p_{\mathcal{T}}(x)}{p_{\mathcal{S}}(x)} p_{\mathcal{S}}(x, y) \, \mathrm{d}y \mathrm{d}x. \quad (3)$$

Using samples from the source domain $\{(x_i, y_i)\}_{i=1}^n$, this risk can be estimated through:

$$\hat{R}_{\mathcal{W}}(h_\theta) = \frac{1}{n} \sum_{i=1}^n \ell(h_\theta(x_i), y_i) \frac{p_{\mathcal{T}}(x_i)}{p_{\mathcal{S}}(x_i)}. \qquad (4)$$

Note that it does not depend on target labels u.

3.1. Weights

The ratio of probability distributions is referred to as the importance weight function $w(x) = p_{\mathcal{T}}(x)/p_{\mathcal{S}}(x)$. Each weight corrects the probability of observing a single pair (x_i, y_i) under the source distribution to its probability under the target distribution. Higher values indicate higher importance to the target domain. In the example setting, the weight function is: $w(x) = \gamma \exp([\gamma^{-2}(-1-x)^2 - x^2]/2)$. Its variance has an analytical solution, $\mathbb{V}_{\mathcal{S}}[w(x)] = \gamma^2/(2\gamma^2 - 1) \cdot \exp(1/(2\gamma^2 - 1)) - 1$, provided $\gamma > 1/\sqrt{2}$, shown in Figure 2. Note that the weights' variance grows exponentially as γ shrinks. In other words, as the source domain becomes a more limited view of the target population, the spread over the importance of each sample grows sharply.

3.2. Sampling variance

In expectation, the estimators $\hat{R}_{\mathcal{T}}$ and $\hat{R}_{\mathcal{W}}$ are equivalent. However, they behave differently for finite sample sizes. The variance of an estimator with respect to data, is known as its *sampling variance*. In the following, we will compare the sampling variances of $\hat{R}_{\mathcal{W}}$ and $\hat{R}_{\mathcal{T}}$.
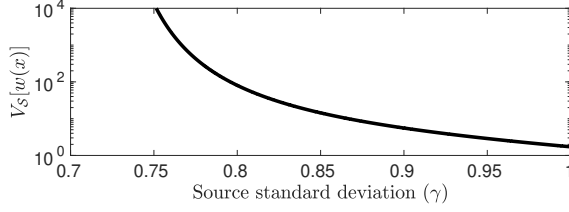
Fig. 2. Variance of the importance weights $w(x)$ as a function of the standard deviation γ of the source data distribution in the example setting.

Assuming data is drawn independently and is identically distributed (*iid*), the sampling variance of the target risk estimator is

$$\mathbb{V}_{\mathcal{T}}[\hat{R}_{\mathcal{T}}] = \mathbb{E}_{\mathcal{T}}\Big[\Big(\frac{1}{m}\sum_{j=1}^{m}\ell(h_\theta(z_j), u_j) - R_{\mathcal{T}}\Big)^2\Big]$$

$$= \frac{1}{m}\,\mathbb{E}_{\mathcal{T}}\big[\big(\ell(h_\theta(x), y) - R_{\mathcal{T}}\big)^2\big] = \frac{\sigma_{\mathcal{T}}^2}{m}, \quad (5)$$

where $\sigma_{\mathcal{T}}^2$ is the sampling variance of $\hat{R}_{\mathcal{T}}$ with respect to a single point. Assuming the source data is drawn *iid* as well, the sampling variance of the importance-weighted risk estimator follows similarly:

$$\mathbb{V}_{\mathcal{S}}[\hat{R}_{\mathcal{W}}] = \mathbb{E}_{\mathcal{S}}\Big[\Big(\frac{1}{n}\sum_{i=1}^{n}\ell(h_\theta(x_i), y_i)w(x_i) - R_{\mathcal{W}}\Big)^2\Big]$$

$$= \frac{1}{n}\,\mathbb{E}_{\mathcal{S}}\big[\big(\ell(h_\theta(x), y)w(x) - R_{\mathcal{W}}\big)^2\big] = \frac{\sigma_{\mathcal{W}}^2}{n}. \quad (6)$$

Since $R_{\mathcal{W}} = R_{\mathcal{T}}$, Equations (5) and (6) reveal that the sampling variances of $\hat{R}_{\mathcal{T}}$ and $\hat{R}_{\mathcal{W}}$ only differ by a scaling induced by the importance weights:

$$\sigma_{\mathcal{T}}^2 - \sigma_{\mathcal{W}}^2 = \mathbb{E}_{\mathcal{T}}\big[\,\ell(h_\theta(x), y)^2\big(1 - w(x)\big)\,\big]. \quad (7)$$

Hence, in problem settings where large weights occur often, the importance-weighted risk estimator has a higher sampling variance than the target risk estimator. This makes sense: as the training data distribution provides a more narrow view of the population, the risk estimate becomes more uncertain. Figure 4 plots $\sigma_{\mathcal{W}}^2$ and $\sigma_{\mathcal{T}}^2$ for the example setting (σ_β^2 will be discussed in Section 4). Note that the shape of the curve of $\hat{R}_{\mathcal{W}}$ reflects the influence of weight variance (see Figure 2).

3.3. Occurrence of large weights

A large sampling variance does not necessarily mean a given data set will contain a large weight. Drawing a point with a large weight occurs with some probability. For illustration, we plot the probability that an importance-weight exceeds a constant c in the example setting. Figure 3 shows curves for three choices of γ. For $\gamma = 0.7$, the probability of a weight exceeding the value 10 is about 0.2.
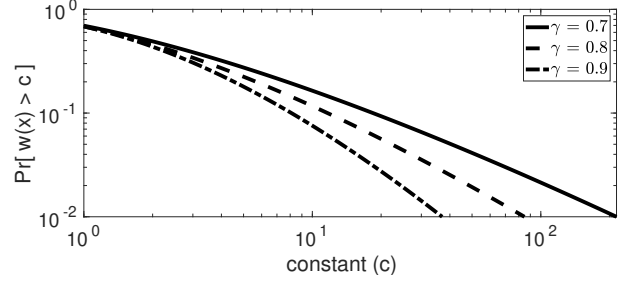
Fig. 3. Probability of an importance weight exceeding the constant c, for different values of source standard deviation γ.

4. REDUCING SAMPLING VARIANCE

4.1. Control variates

A control variate is a function g of a random variable x that correlates with a particular statistic $f(x)$ of that random variable and whose expected value is known [7, 8]. It reduces sampling variance as follows: suppose that $g(x)$ is positively correlated to the statistic $f(x)$. Then, whenever $g(x_i)$ rises above its expected value, $g(x_i) - \mathbb{E}[g(x)] > 0$, $f(x_i)$ will also rise above its expected value, $f(x_i) - \mathbb{E}[f(x)] > 0$. If g is negatively correlated, then f will fall below its expectation whenever g rises above its expectation. By subtracting the control variate from the statistic, $f(x_i) - (g(x_i) - \mathbb{E}[g(x)])$, the statistic's deviation from its expected value can be manipulated.

It is however important that the control variate is appropriately scaled. For this purpose, a parameter β is introduced. If β is fixed, then the estimator will be unbiased:

$$\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}f(x_i) - \beta\big(g(x_i) - \mathbb{E}[g(x)]\big)\Big]$$

$$= \mathbb{E}[f(x)] - \beta(\mathbb{E}[g(x)] - \mathbb{E}[g(x)]) = \mathbb{E}[f(x)] \quad (8)$$

Our statistic of interest is the weighted loss. For common choices of loss functions, such as quadratic, logistic or hinge, the importance weights will generally correlate well with the weighted loss. The expected value of the importance weights is always known, since

$$\mathbb{E}_{\mathcal{S}}[w(x)] = \int_{\mathcal{X}}\frac{p_{\mathcal{T}}(x)}{p_{\mathcal{S}}(x)}p_{\mathcal{S}}(x)\mathrm{d}x = \int_{\mathcal{X}}p_{\mathcal{T}}(x)\mathrm{d}x = 1. \quad (9)$$

This means the weights can *always* be introduced as a control variate to an importance-weighted estimator. The controlled importance-weighted risk estimator becomes:

$$\hat{R}_\beta(h_\theta) = \frac{1}{n}\sum_{i=1}^{n}\ell(h_\theta(x_i), y_i)w(x_i) - \beta(w(x_i) - 1). \quad (10)$$

Note that the weights $w(x_i)$ have already been computed, implying that additional computational cost is restricted to β.

The effect of the control variate on the sampling variance of the importance-weighted risk estimator can be seen via [7]:

$$
\begin{aligned}
\mathbb{V}_{\mathcal{S}}[\hat{R}_\beta] &= \mathbb{E}_{\mathcal{S}}\big[(\hat{R}_\beta - R_\beta)^2\big] \\
&= \frac{1}{n}\,\mathbb{E}_{\mathcal{S}}\big[\big(\ell_w(x,y) - \beta(w(x)-1) - R_{\mathcal{W}}\big)^2\big] \\
&= \frac{1}{n}\big(\sigma_{\mathcal{W}}^2 - 2\beta\mathbb{C}_{\mathcal{S}}[\ell_w(x,y),w(x)] + \beta^2\mathbb{V}_{\mathcal{S}}[w(x)]\big) = \frac{\sigma_\beta^2}{n}\,,
\end{aligned}
$$
(11)

where $\ell_w(x,y) = \ell(h_\theta(x),y)w(x)$ is shorthand for the weighted loss function given a fixed classifier h_θ, and $\mathbb{C}_{\mathcal{S}}$ refers to covariance. We can minimize this sampling variance for the scaling parameter β. Taking the derivative of (11) with respect to β and setting it to 0 yields:

$$
\beta^* = \mathbb{C}_{\mathcal{S}}\big[\ell_w(x,y),w(x)\big] / \mathbb{V}_{\mathcal{S}}\big[w(x)\big]\,.
$$
(12)

Plugging β^* back in simplifies the sampling variance to:

$$
\sigma_\beta^2 = \sigma_{\mathcal{W}}^2 - \mathbb{C}_{\mathcal{S}}\big[\ell_w(x,y),w(x)\big]^2 / \mathbb{V}_{\mathcal{S}}\big[w(x)\big]\,.
$$
(13)

Considering that both the squared covariance term and the variance term are non-negative, the sampling variance of $\hat{R}_\beta$ is never larger than that of $\hat{R}_{\mathcal{W}}$ [9]. In particular, σ_β^2 can be alternatively formulated as $\sigma_{\mathcal{W}}^2(1-\rho^2)$, where ρ denotes the correlation between the weighted loss and the weights [8]. Essentially, the more the weights correlate – *positively* or *negatively* – with the weighted loss, the larger the reduction in variance.

We computed σ_β^2 for the example setting and show it alongside $\sigma_{\mathcal{W}}^2$ and $\sigma_{\mathcal{T}}^2$ in Figure 4. Note that σ_β^2 also diverges as γ shrinks, like $\sigma_{\mathcal{W}}^2$, but does so at a slower rate.

4.2. Regression estimator

In practice, β will need to be estimated. Both the weight variance and the covariance between the weighted loss and the weights can be estimated from data. In that case, Equation 12 becomes the solution to a least-squares problem [8]:

$$
\hat{\beta} = \frac{\sum_{i=1}^n \big(\ell_w(x_i,y_i) - \sum_{i=1}^n \ell_w(x_i,y_i)\big)\big(w(x_i)-1\big)}{\sum_{i=1}^n \big(w(x_i)-1\big)^2}\,.
$$
(14)

However, β now depends on the same observed variable as the weighted loss and the weights, which is somewhat problematic. It can be shown that the deviation of $\hat{\beta}$ from β^* drops off on the order of $O(n^{-1/2})$ (see Theorem 1 from [9]). This estimation error in β causes a bias in the risk estimator on the order of $O(n^{-1})$ [9].

5. IMPORTANCE-WEIGHTED CROSS-VALIDATION

Standard cross-validation will not yield optimal regularization parameter estimates under sample selection bias.
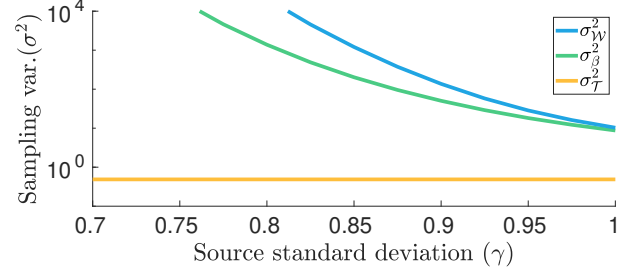
Fig. 4. Sampling variance of the target (yellow, $\sigma_{\mathcal{T}}^2$), the importance-weighted (blue, $\sigma_{\mathcal{W}}^2$) and the controlled importance-weighted (green, σ_β^2) risk estimators as a function of γ in the example setting.

Importance-weighted cross-validation is a potential solution [4], but is not robust to large weights. The following subsections describe experiments that compare risk estimators during k-fold cross-validation, and evaluate their ability to find appropriate regularization parameters.

5.1. Data

We consider a synthetic and a natural data set. In the synthetic setting, the target data distribution is a unit bivariate normal $\mathcal{N}([0,0] \mid I)$, while the source data distribution consists of a bivariate normal with a shifted mean and a scaled identity covariance matrix, $\mathcal{N}([-1,0] \mid \gamma I)$. The posterior distribution is of the form: $p(y = -1 \mid x_1, x_2) = 1 - \Phi(-[x_1, x_2])$ and $p(y = 1 \mid x_1, x_2) = \Phi(-[x_1, x_2])$. Figure 5 visualizes the class-conditional distributions $p(x \mid y)$ for $\gamma = 1/\sqrt{2}$. Note that the source domain is a local sampling of the larger target domain and that the nonlinear nature of the underlying decision boundary is not apparent. We draw data sets using rejection sampling, with 50 source samples and 1000 target samples.
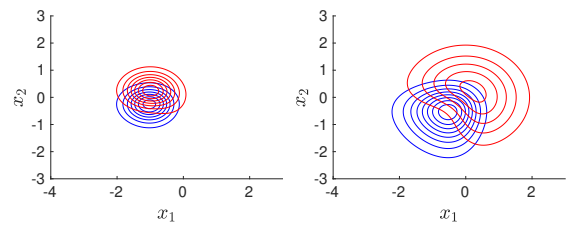
Fig. 5. Synthetic problem, with $\gamma = 1/\sqrt{2}$. Class-conditional distributions for source (left) and target domain (right).

The natural setting is derived from the ozone level detection data set from the UCI machine learning repository[1]. The task is to predict ozone days versus normal weather days, given various weather station measurements. All samples with missing values are removed, the data has been z-scored,

and it has been projected onto its first 10 principal components. The source domain is a local sampling in time: a Gaussian centered at the start date with a standard deviation that is a proportion γ of the total number of samples. For small values of γ, the source domain contains only samples close to the start date. For large values, the source domain is roughly equally spread over time. We draw 80 samples without replacement from both classes to form the source data set. These 80 samples are also included in the target data, so as not to form a 'hole' in feature space.

For each setting, we perform 100 000 repetitions of sampling a data set. The top 10% of repetitions with the largest weight variance are considered the data sets with problematically large weights.

5.2. Experimental setup

The source set is split into $k = 5$ folds, a classifier is trained using a particular value for the regularization parameter λ and its loss is computed on each data point in the held-out validation set. That loss is stored, along with those points' estimated weights. An importance-weighted L_2-regularized linear least-squares classifier is taken. Its parameters are estimated through $\theta_\lambda = (X_t W_t X_t^\top + \lambda I)^{-1}(X_t^\top W_t y_t)$ where t indicates training indices and W is a diagonal matrix with the estimated weights as its entries. For λ, we considered a range from 10^{-3} to 10^6 over 200 logarithmic steps.

Weights are estimated by fitting a normal distribution to data from each domain, and computing the ratio of the target probability over the source probability of each source data point $\hat{w}(x_i) = \hat{p}_\mathcal{T}(x_i)/\hat{p}_\mathcal{S}(x_i)$. We compare the importance-weighted risk estimator ($\hat{R}_{\hat{w}}$) with its control variate counterpart ($\hat{R}_{\hat{\beta}}$). We average their estimated risks over all data sets and specifically over the 10% of data sets with the largest weight variance (indicated with ">" in the legend of Figures 6, 7 and 8). We also include validation on the labeled target samples ($\hat{R}_\mathcal{T}$) as the oracle solution. After risk estimation, the λ is selected that minimized risk. The classifier is then retrained using all source data and the selected λ, and evaluated using the target risk based on the true target labels as the final measure. This process is repeated for each data set and we report the final average as $\bar{R}_\mathcal{T}$.

Repeating the above procedure for each data set allows us to perform non-parametric hypothesis tests for statistically significant differences between the final risks of the estimators. Since the data is paired and the estimators' sampling distributions are skewed beyond normality, we employ a Wilcoxon signed-rank test [12].

5.3. Results

Figure 6 presents the final target risks for each risk estimator on the synthetic problem setting. Note that the errorbars in the plot are too small to see. When looking at all data sets,

there is nearly no difference between $\hat{R}_{\hat{w}}$ and $\hat{R}_{\hat{\beta}}$. This indicates that the addition of the control variate does not deteriorate importance-weighted cross-validation in general. When looking at the sets with large weight variance, it can be seen that $\hat{R}_{\hat{w}}$ deteriorates strongly, while $\hat{R}_{\hat{\beta}}$ does not. This shows that the addition of the control variate makes the importance-weighted risk estimator more robust to large weights. The difference between $\hat{R}_{\hat{w}}$ and $\hat{R}_{\hat{\beta}}$ for the data sets with large weight variance (dotted lines) is statistically significant for each value of γ, with the largest p-value on the order of 10^{-30}.
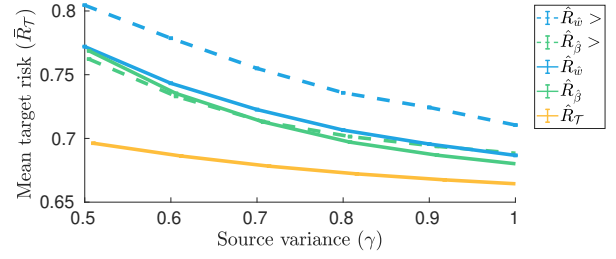
Fig. 6. Mean target risks for the synthetic problem setting, as a function of source variance γ.

Figure 7 presents the mean target risks in the Ozone problem setting. Note that the controlled estimator $\hat{R}_{\hat{\beta}}$ is essentially unaffected by the large weights since it produces the same final risks in both the case of all data sets and the top 10% of largest weight variance. Both sets of risks are below that of the uncontrolled estimator $\hat{R}_{\hat{w}}$, which still deteriorates in the data sets with large weight variance. All differences between $\hat{R}_{\hat{w}}$ and $\hat{R}_{\hat{\beta}}$ are statistically significant for all values of γ, with the largest p-value on the order of 10^{-50}. Lastly, all risks increase slightly as γ grows, which is due to less data outliers around the start date.
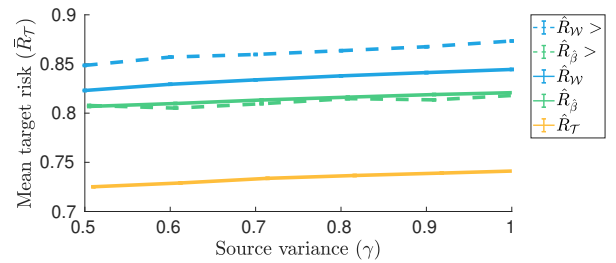

Fig. 7. Mean target risks for Ozone level detection setting, as a function of the relative scale of the source domain γ.

5.4. Weight estimators

The effect of the control variate is independent of the choice of importance-weight estimator. We illustrate this point by performing the same experiment on the synthetic setting using Kernel Mean Matching (KMM) [13] and the Kullback-

Leibler Importance Estimation Procedure (KLIEP) [14]. Figure 8 shows similar results as in Figure 6, except that the control variate even leads to better final target risks for the average over all data sets (green lines are below blue lines).
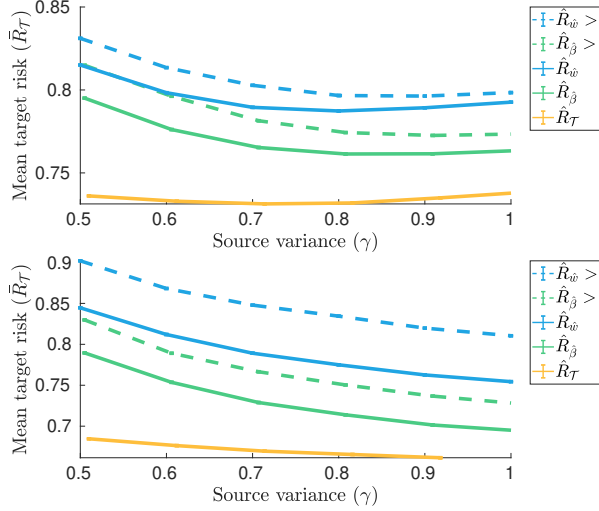
Fig. 8. Mean target risks for the synthetic problem setting, as a function of γ. (Top) KMM, (bottom) KLIEP.

Some non-parametric weight estimators are formulated such that they naturally avoid large weights. However, this behavior depends on one or more hyperparameters, such as a kernel bandwidth parameter. Unfortunately, finding an optimal bandwidth parameter would require cross-validation, and consequently, such estimators are not suited to cross-validation procedures. One has to resort to heuristics, such as Silverman's rule or the average distance to k-nearest-neighbours. In the above experiment, we set the kernel bandwidth in KMM and KLIEP to the average Euclidean distance from the source points to their five nearest neighbours of the target points.

6. CONCLUSION & DISCUSSION

We introduced a control variate to reduce the sampling variance of the importance-weighted risk estimator. With its inclusion, the importance-weighted risk estimator is more robust large weight variance. Consequently, during k-fold cross-validation, it selects better hyperparameters than the uncontrolled importance-weighted risk estimator.

We have studied an additive linear control variate. Alternatively, one could consider more complex control variates, such as higher-order moments of the weights, or multiplicative control variates [8]. If these have a stronger correlation with the weighted loss, they could lead to larger reductions in variance. However, there are a wide variety of possible alternatives, and it is unclear how to search over these.

7. REFERENCES

[1] B Zadrozny, "Learning and evaluating classifiers under sample selection bias," in *International Conference on Machine learning*, 2004, p. 114.

[2] C Cortes, M Mohri, M Riley, and A Rostamizadeh, "Sample selection bias correction theory," in *Algorithmic Learning Theory*, 2008, pp. 38–53.

[3] J Quionero-Candela, M Sugiyama, A Schwaighofer, and ND Lawrence, *Dataset shift in machine learning*, MIT Press, 2009.

[4] M Sugiyama, M Krauledat, and K-R Müller, "Covariate shift adaptation by importance weighted cross validation," *Journal of Machine Learning Research*, vol. 8, pp. 985–1005, 2007.

[5] WM Kouw and M Loog, "Effects of sampling skewness of the importance-weighted risk estimator on model selection," in *International Conference on Pattern Recognition*, 2018, pp. 1468–1473.

[6] EL Ionides, "Truncated importance sampling," *Journal of Computational and Graphical Statistics*, vol. 17, no. 2, pp. 295–311, 2008.

[7] BL Nelson, "On control variate estimators.," *Computers & OR*, vol. 14, no. 3, pp. 219–225, 1987.

[8] AB Owen, *Monte Carlo theory, methods and examples*, 2013.

[9] A Owen and Y Zhou, "Safe and effective importance sampling," *Journal of the American Statistical Association*, vol. 95, no. 449, pp. 135–143, 2000.

[10] R Kohavi et al., "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *International Joint Conference on Artificial Intelligence*, 1995, vol. 14, pp. 1137–1145.

[11] GC Cawley and NLC Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research*, vol. 11, pp. 2079–2107, 2010.

[12] WJ Conover, "Practical nonparametric statistics," 1980.

[13] J Huang, AJ Smola, A Gretton, KM Borgwardt, B Schölkopf, et al., "Correcting sample selection bias by unlabeled data," in *Advances in Neural Information Processing Systems*, 2007, p. 601.

[14] M Sugiyama, S Nakajima, H Kashima, PV Buenau, and M Kawanabe, "Direct importance estimation with model selection and its application to covariate shift adaptation," in *Advances in Neural Information Processing Systems*, 2008, pp. 1433–1440.