ROBUST IMPORTANCE-WEIGHTED CROSS-VALIDATION UNDER SAMPLE SELECTION BIAS

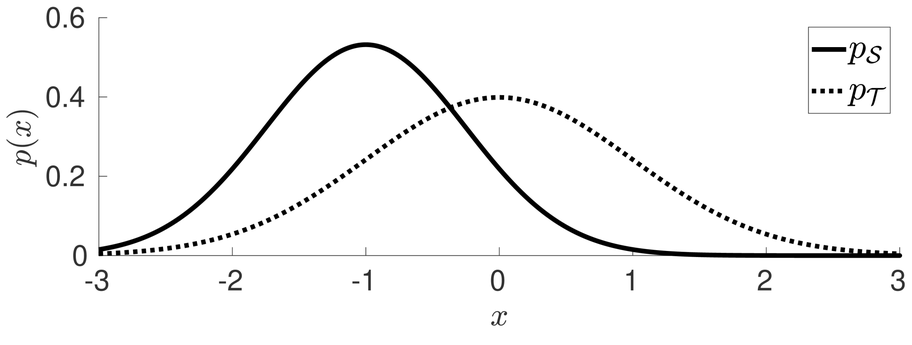
WM Kouw[13] JH Krijthe[2] M Loog[3]

[1] Eindhoven University of Technology [2] Radboud University Nijmegen [3] Delft University of Technology

Cross-validation under sample selection bias can, in principle, be done by importance-weighting the empirical risk. However, the importance-weighted risk estimator produces sub-optimal 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.

SAMPLE SELECTION BIAS

Sampling bias can occur spatially, when you collect data from one location but expect to generalize a wider target population, or temporally, when you collect for a short period of time but expect to generalize to a larger horizon. As a consequence the training data is differently distributed than test data. Below, we refer to the distribution of the training data - collected under sampling bias - as the *source* distribution (p_S) and the test data distribution as the *target* distribution (p_T).



TARGET RISK

We are interested in minimizing the target risk function, i.e. the expected loss with respect to the target distribution. This function is estimated with a sample average over labeled target samples:

$$R_{\mathcal{T}}(h_\theta) = \int \ell\big(h_\theta(x), y\big) p_{\mathcal{T}}\big(x, y\big) \; \mathrm{d}y\mathrm{d}x$$

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

IMPORTANCE-WEIGHTED RISK

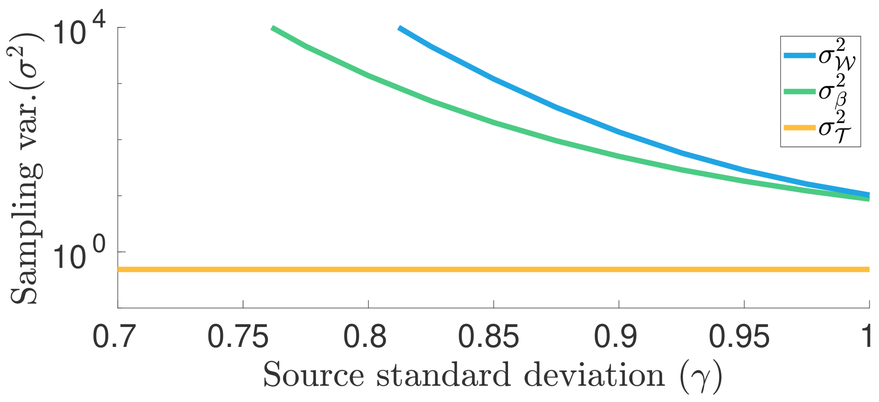
Under sample selection bias, we do not have target labels. Instead, we use the fact that the posteriors are equal, pS(y|x) = pT(y|x), to re-cast the target risk as the source risk weighted by the ratio of data distributions:

$$R_{\mathcal{W}}(h_\theta) = \int \ell\big(h_\theta(x), y\big) p_{\mathcal{T}}\big(x, y\big) \frac{p_{\mathcal{T}}(x)}{p_{\mathcal{S}}(x)} \; \mathrm{d}y\mathrm{d}x$$

$$\hat{R}_{\mathcal{W}}(h_\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell\big(h_\theta(x_i), y_i\big) w(x_i)$$

SAMPLING VARIANCE

It turns out that the importance-weights scale the sampling variance of the weighted estimator. High sampling variance means inaccurate estimates. In the figure below, we plot the sampling variance of the oracle target risk estimator (yellow) and the importance-weighted source risk estimator (blue).



CONTROL VARIATE

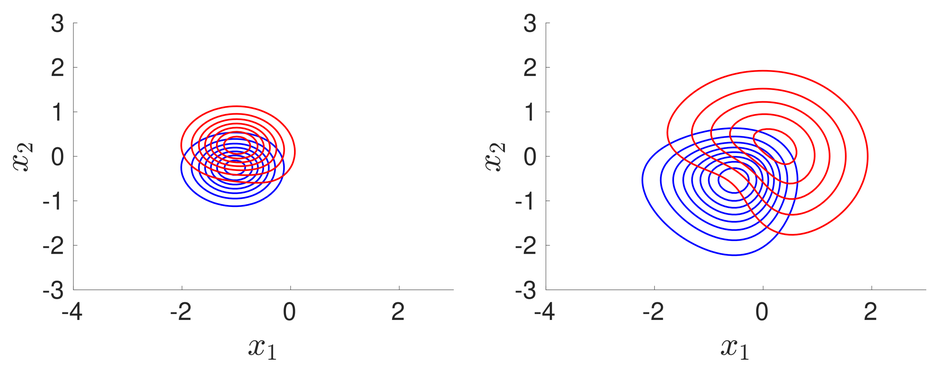
Variance reduction methods employ additional knowledge of an estimator to reduce the sampling variance of that estimator, and thereby produce a more accurate estimate given the same sample size. We use the importance-weights themselves as the control variate:

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

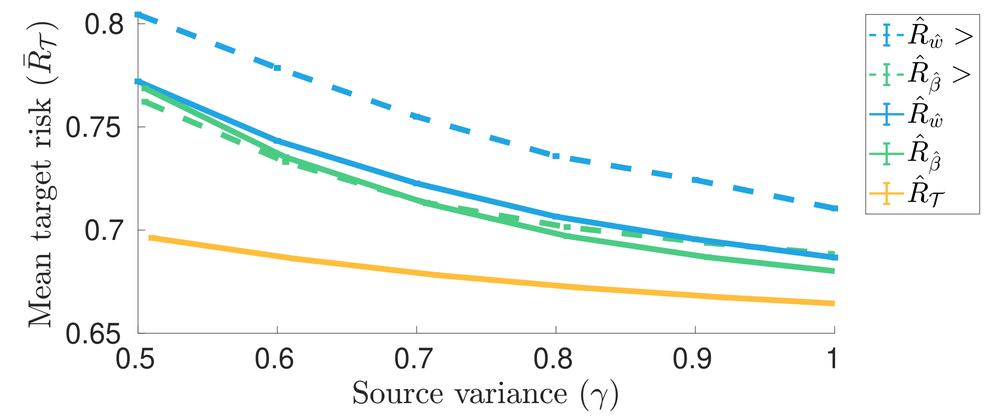
We know the expected value of the weights, namely *1*, and we also know that the weights are correlated to the weighted loss. If a weight rises above 1, then the weighted loss for that point also rises above its expected value (or falls below, in case of negative correlation). Subtracting the weight's deviation from the weighted loss, reduces the estimator's sampling variance. The β parameter is estimated separately and ensures appropriate scaling. The green line in the left-bottom Figure shows the new sampling variance. Note that it grows much more slowly.

EXPERIMENT

Consider a unit 2D Gaussian for the target distribution and a narrower 2D Gaussian for the source distribution, with a nonlinear decision boundary between them. Below are shown the two class-conditional distributions of the source (left) and target (right) distribution, for source std. dev. of 1/√2.



We draw 10^6 source and target data sets from these distributions. For each set, we do 5-fold importance-weighted cross-validation where we train an importance-weighted classifier on remaining folds and evaluate the risk using the weighted, controlled and oracle target risk estimators shown above. A regularization parameter is then selected, we train the importance-weighted classifier on all source data with the selected parameter and evaluate its risk using the labeled target data. Shown below are the target risks for the three estimators as a function of the standard deviation of the source distribution (i.e. the severity of sampling bias).



The dotted lines show the performance of the risk estimators for the top 10% largest weight variance data sets. We can conclude from this that, while the weighted risk estimator deteriorates as the weight variance increases, the controlled estimator performs at the same level.

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. The effect is independent of the weight estimator employed (see paper for experiments and results).

https://github.com/wmkouw/covshift-ctrlvar
https://arxiv.org/abs/1710.06514
https://wmkouw.github.io/