Effects of sampling skewness in importance-weighted cross-validation

WM Kouw M Loog

Abstract

Importance-weighting is a popular technique for dealing with sample selection bias and covariate shift, with characteristics such as unbiasedness and low computational complexity. However, the sampling distribution of an importance-weighted risk estimator can be skewed: for small sample sizes, it produces overestimates for the majority of data sets, and large underestimates for sets in the tail of the sampling distribution. These over- and underestimates lead to sub-optimal regularization parameters in importance-weighted cross-validation.

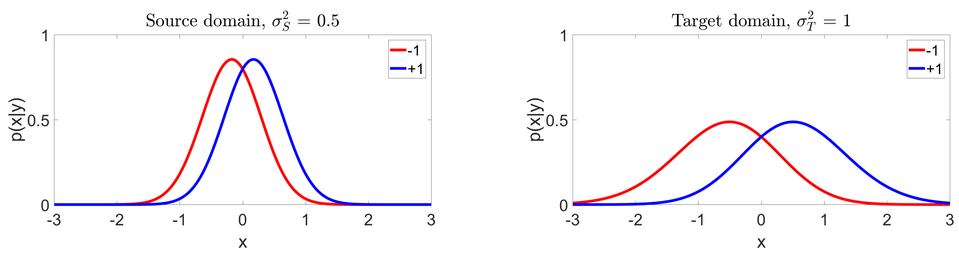
Covariate shift

In covariate shift settings, the training data comes from a source domain and the test data from a target domain. The data distributions are different:

$$p_{\mathcal{S}}(x) \neq p_{\mathcal{T}}(x)$$

But the posterior and prior distributions are equivalent:

$$p_{\mathcal{S}}(y \mid x) = p_{\mathcal{T}}(y \mid x) \ , \ p_{\mathcal{S}}(y) = p_{\mathcal{T}}(y)$$



Importance-weighting

We are interested in the target risk function:

$$R_{\mathcal{T}}(h) = \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} \ell(h(x), y) p_{\mathcal{T}}(x, y) \ \mathrm{d}x$$

But we don't have labeled target samples to estimate it. Instead, using the fact that the posteriors are equivalent, we re-write the target risk into a weighted source risk:

$$R_{\mathcal{W}}(h) = \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} \ell(h(x), y) p_{\mathcal{S}}(x, y) \frac{p_{\mathcal{T}}(x)}{p_{\mathcal{S}}(x)} \ \mathrm{d}x$$

Estimator

The importance-weighted risk can be estimated by the weighted source sample average:

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

where w(x) is the ratio of marginal data distributions. To get an idea of how this estimator behaves for finite sample sizes, we can have a look at the moments of its sampling distribution:

$$\mathbb{E}_{\mathcal{S}}\left[\hat{R}_{\mathcal{W}}(h)\right] = R_{\mathcal{T}}(h)$$

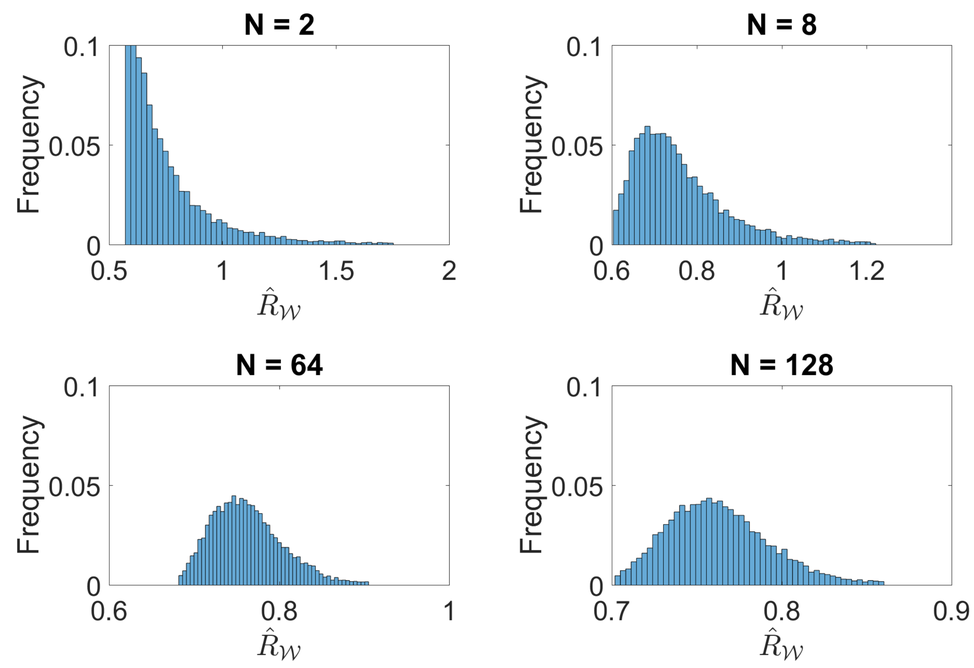
$$\mathbb{V}_{\mathcal{S}}\left[\hat{R}_{\mathcal{W}}(h)\right] = \frac{1}{n}\left(\mathbb{E}_{\mathcal{T}}[\ell(h(x), y)^2 \mathbf{w}(\mathbf{x})] - R_{\mathcal{T}}(h)^2\right)$$

$$\Gamma_{\mathcal{S}}\left[\hat{R}_{\mathcal{W}}(h)\right] = \frac{1}{n^2} \mathbb{V}_{\mathcal{S}}[\hat{R}_{\mathcal{W}}(h)]^{-2/3}$$
$$\left(\mathbb{E}_{\mathcal{T}}[\ell(h(x), y)^3 \mathbf{w}(\mathbf{x})^{\mathbf{2}}] - 3 \ R_{\mathcal{T}}(h) \mathbb{V}_{\mathcal{S}}[\hat{R}_{\mathcal{W}}(h)] - R_{\mathcal{T}}(h)^3\right)$$

Compared the target risk estimator's moments, these are scaled by the importance weights (bold). In other words, the larger the moments of the weight distribution, the less accurate the estimator.

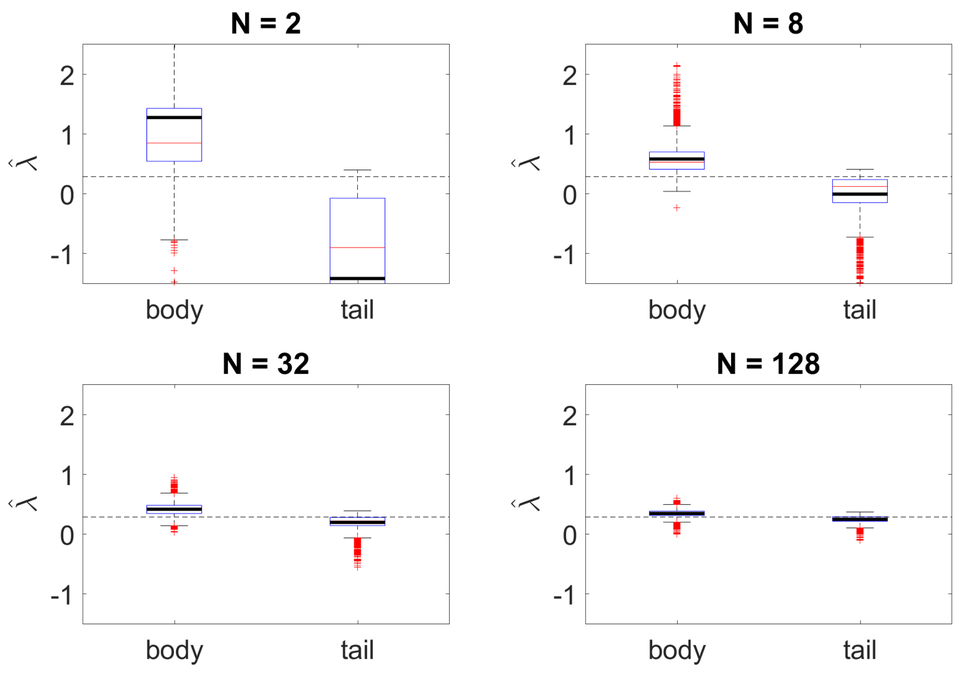
Experiment

We created a synthetic setting with $p_{\mathcal{S}}(x) = \mathcal{N}(x \mid 0, 0.5)$ and $p_{\mathcal{T}}(x) = \mathcal{N}(x \mid 0, 1)$, cumulative normal distributions as posteriors and class priors of 1/2. Using Bayes' rule, we derived the class-conditional distributions ($p_{\mathcal{S}}(x \mid y)$, $p_{\mathcal{T}}(x \mid y)$) and drew samples of varying size using rejection sampling, repeated 10000 times. For each data set, we computed the importance-weighted risk:

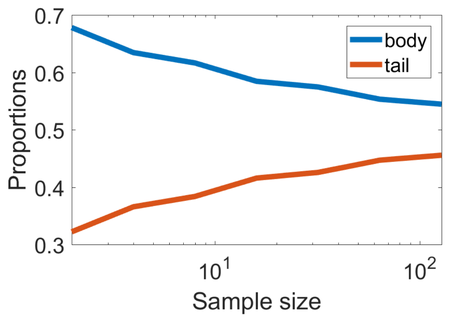


Cross-validation

Since the importance-weighted risk is used for evaluating different regularization parameters during cross-validation, it induces a skew in their estimates as well:



Although the estimator is unbiased, for small sample sizes, there are much more over- than underestimations.



Discussion

- For higher-dimensional settings, more validation data is required to reduce the skew.
- In the reverse setting where the source domain is wider than the target domain, the skew is negative instead of positive.
- We believe that the results will hold for all loss functions; the skewness is caused by the weights, not the loss function.

TU Delft

prlab.tudelft.nl
wmkouw@gmail.com
https://arxiv.org/abs/1804.07344