Supplementary Material for "A transfer learning approach based on random forest with application to breast cancer prediction in underrepresented populations"

Tian Gu[1], Yi Han[2], and Rui Duan[†1]

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA
[2]School of Mathematical Sciences, Shanghai Jiaotong University, Shanghai, China
[†]Corresponding author: rduan@hsph.harvard.edu

Detailed Simulation Settings

Setting 1

In Setting 1, we consider that the source and the target populations share a similar variable importance ranking, where the similarity between the two populations is measured by the correlation of their variable importance rankings. To generate $m_s(x)$ and $m_t(x)$, we apply some non-linear transformations on each feature in X and obtain the transformed features Z. We then combine the transformed features through a linear combination to obtain $m_s(x)$ and $m_t(x)$, i.e., $m_s(x) = Z\beta_s$ and $m_t(x) = Z\beta_t$, where β_s and β_t are p-dimensional vectors whose magnitude determines the feature importance. By changing the correlation between β_t and β_s, we change the similarity degree of the feature importance.

Setting 1, varying the variable importance rank between the target and the source population. For both sites, we generate $p = 20$ predictors from truncated normal distribution between $[-1, 1]$ with X of mean -0.5 and X of mean 0.5. we generate $Y_s = \beta_s m_s(x) + \epsilon_s$ and $Y = \beta m_t(x) + \epsilon$, where $\epsilon \sim N(0, 1)$ and $\epsilon_s \sim N(0, 1)$. Let $m_s(x)$ and $m_t(x)$ be the X-transformation functions for the source and the target data, respectively, where $m_s(x) = \sum_{i=1}^{10} X_i^2 - \sum_{i=11}^{20} 0.5X_i^2$ and $m_t(x) = \sum_{i=1}^{10} X_i^2 - \sum_{i=11}^{20} 0.5X_i^2$. Define τ as Kendall correlation between β and β_s. We change the relatedness of the variable importance rank in two populations by varying τ. Specifically, for a given β_s, we first sequentially swap the 1-st ranked term in β_s with the p'-th ranked term, $p' = 1, \ldots, p$, and then assign opposite sign of β_s to create β.

Setting 2

In Setting 2, we consider that the discrepancy between $m_s(X)$ and $m_t(X)$ is independent or weakly correlated with $m_s(X)$. To achieve this, we first generate $m_s(x)$ in the same way described in Setting 1. We then generate the function $\delta(x)$, a function of a random subset of all the features, on which we apply different feature transformations and linear combinations compared to $m_s(x)$. We obtain

$m_t(x) = m_s(x) + \delta(x)$. We vary the variance explained by the source model $m_s(x)$ to control the similarity between the source and the target populations.

Setting 2, $\delta(x)$ is weakly correlated with $m_s(x)$. For both sites, we generate $p = 20$ predictors from uniform distribution between $[-1, 1]$. Let $m_s(x) = \sum_{i=1}^{10} X_i^2 - \sum_{i=11}^{20} 0.5X_i^2$. Denote $\delta(x) = \delta(x; \Delta) = \sum_{i=1}^{15} \Delta I(X_i >= thres_i)$, where Δ is pre-specified to control the overall magnitude of $\delta(x)$ and $thres_i$ is a set of random threshold. We generate $Y_s = m_s(X_s) + \epsilon_s$, and $Y = m_t(X) = m_s(X) + \delta(X; \Delta) + \epsilon$, where $\epsilon \sim N(0, 1)$ and $\epsilon_s \sim N(0, 1)$. We vary the ratio between the source variance and the δ variance by varying Δ following uniform distribution between 0.6 to 50 (larger ratio represents smaller calibration term which indicating the source is more useful to the target).

Setting 3

In Setting 3, we consider that the discrepancy term is correlated with $m_s(X)$. We generate Y_s following the same data generating mechanism in Setting 2 except that we set $m_t(X) = Cm_s(X) + \delta(X)$, where C is a constant. In this case, the true discrepancy is $m_t(X) - m_s(X) = (C - 1) * m_s(X) + \delta(X)$. With $C \neq 1$, $m_s(X)$ is correlated with the discrepancy, and we vary C to alter the strength of the correlation.

Setting 3, $\delta(x)$ is highly correlated with $m_s(x)$. We generate outcomes Y_s following the same data generating mechanism in setting 2 above. We generate the target outcome through $Y = 5 \times m_s(X) + \delta(X; \Delta) + \epsilon$ so that $\delta(X; \Delta)$ is correlated with $m_s(X)$.