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 through Kendall correlation. 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 apply different non-linear transformation on each feature in X and X_s to generate the transformed features Z_s and Z_t. We then combine the transformed features through a linear combination to obtain $m_s(x)$ and $m_t(x)$, i.e., $m_s(x) = Z_s\beta_s$ and $m_t(x) = Z_t\beta_t$, where β_s and β_t are p-dimensional vectors whose magnitude determines the feature importance. Specifically,

- $m_s(X) = \beta_s \times (\sum_{i=1}^{5} \sqrt{|X_i|} - 0.5\sum_{i=6}^{9} \sqrt{|X_i|} - 0.8\sum_{i=10}^{15} \sqrt{|X_i|} + 0.5\sum_{i=16}^{20} \sqrt{|X_i|})$
- $m_t(X) = \beta_t \times (\sum_{i=1}^{10}(X_i)^2 - 2\sum_{i=11}^{20}(X_i)^2)$
- $\beta_t = \mathbf{a}(1, \ldots, 20)^T$, where $\mathbf{a} = (a_1, \ldots, a_{20})$ is a set of random number to determine the sign of β_t, i.e., $a_i \sim Bern(0.5), i \in \{1, \ldots, 20\}$

Define τ as Kendall correlation between β_t and β_s. We change the relatedness of the variable importance rank in two populations by varying τ. Specifically, for a given β_t described above, we first sequentially swap the 1-st ranked term in β_t (20) with the p'-th ranked term, $p' = 1, \ldots, p$ to create β_s', and then assign opposite sign to create $\beta_s = -\beta_s'$. In such way, we change the correlation between β_t and β_s, and thus change the similarity degree of the feature importance.

Lastly, we generate $Y_s = \beta_s m_s(X_s) + \epsilon_s$ and $Y = \beta_t m_t(X) + \epsilon_t$, where $\epsilon_t \epsilon_s \sim N(0, 1)$ are random noises.

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 (e.g. we use $X_1, \ldots, X_{15}$ as an example to illustrate below), on which we apply different feature transformations and linear combinations compared to $m_s(x)$:

- $m_s(X) = \sum_{i=1}^{10} X_i^2 - 0.5 \sum_{i=11}^{20} X_i^2$
- $\delta(x) = \delta(x; \Delta) = \sum_{i=1}^{15} \Delta I(X_i >= thres_i)$, where Δ is pre-specified to control the overall magnitude of $\delta(x)$ and $thres_i$ is a set of random threshold. We randomly choose $thres = (-0.5, -0.5, -0.5, -0.5, 4, 4, 4, 4, 0.3, 0.3, 0.3, 0.3, 0.7, 0.7, 0.7)$.

We obtain $m_t(x) = m_s(x) + \delta(x)$. Specifically, We generate $Y_s = m_s(X_s) + \epsilon_s$ and $Y = m_t(X) = m_s(X) + \delta(X; \Delta) + \epsilon_t$, where $\epsilon_t, \epsilon_s \sim N(0, 1)$ are random noises. We control Δ to vary the variance explained by the source model $m_s(x)$ to control the similarity between the source and the target populations. We let Δ follows uniform distribution between 0.6 to 50 (larger Δ represents larger discrepancy term which indicates smaller variance explained by the source model).

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 between 1 and 10 to alter the strength of the correlation (C closer to 1 represents larger variance explained by the source model).