For a more systematic comparison of the 4 proposed penalized Gini scores, we closely follow the simulations outlined in Li et al. (2019) involving discrete features with different number of distinct values, which poses a critical challenge for MDI. The data has 1000 samples with 50 features. All features are discrete, with the jth feature containing $j+1$ distinct values $0, 1, \ldots, j$. We randomly select a set S of 5 features from the first ten as relevant features. The remaining features are noisy features. All samples are i.i.d. and all features are independent. We generate the outcomes using the following rule:

$$P(Y=1|X) = \text{Logistic}(\frac{2}{5}\sum_{j\in S} x_j/j - 1)$$

Treating the noisy features as label 0 and the relevant features as label 1, we can evaluate a feature importance measure in terms of its area under the receiver operating characteristic curve (AUC). We grow 100 deep trees (minimum leaf size equals 1, $m_{try}=3$), repeat the whole process 100 times and report the average AUC scores for each method in Table 1. For this simulated setting, $\widehat{PG}^{(1)}_{oob}$ achieves the best AUC score under all cases, most likely because of the separation of the signal from noise mentioned above. We notice that the AUC score for the OOB-only $\widehat{PG}^{(0)}_{oob}$ is competitive to the permutation importance and the AIR score.