# Supplemental material to "Unbiased Regression Trees for Longitudinal and Clustered Data"

Wei Fu and Jeffrey S. Simonoff

December 8, 2014

This supplemental material provides information about additional simulation studies of the properties of the unbiased RE-EM tree method. In this material, we discuss the performance of the proposed method in regards to unbalanced data, multiple types of covariates, unevenly spaced time points, and linear dependence between the response variable and covariates.

## 1 Sensitivity of the unbiased RE-EM tree to unbalanced data structure

We generalize the simulation design in sections 3.2 and 3.3 of the paper in order to test how sensitive the proposed unbiased RE-EM tree method is to unequal sample sizes per object. Recall in sections 3.2 and 3.3 of the paper there are 100 individuals and 5 observations for each individual in the training set and 50 observations for each individual in the testing set. Here, we have four types of unbalanced/balanced data structure,

- Type 1, 5 observations for each individual for 100 individuals in the training data (balanced type, same as sections 3.2 and 3.3);
- Type 2, $(4, 5, 6)$ observations for each individual for $(33, 33, 34)$ individuals, respectively;
- Type 3, $(3, 5, 7)$ observations for each individual for $(33, 33, 34)$ individuals, respectively;
- Type 4, $(2, 5, 8)$ observations for each individual for $(33, 33, 34)$ individuals, respectively.

The testing data consist of 50 observations for each individual as usual. There are 100 replications and the measures of performance are the percentage of the time the correct tree structure is recovered, PMSE of the response variable $y$, and PMSE of the fixed effect. Figure 1 shows the performance of the proposed unbiased RE-EM tree algorithm.

By design, lack of balance is increasing from type 1 to type 4. From the figure we can see that in terms of recovering the correct tree structure and the PMSE of the fixed effect, the performance

of the proposed method is quite insensitive to the unbalanced data structure, since the performance for the most unbalanced type 4 is similar to that under the balanced type 1. That is to say, the estimation of the fixed effect is very insensitive to an unbalanced data structure. However, in the RIC cases the PMSE of the response $y$ degenerates with an increasing lack of balance in the training data. Closer inspection shows that the PMSE under type 4 has notably larger standard deviation than in other situations, which makes the difference of PMSEs not statistically significant between type 4 and the other types. That is, the proposed method is only affected by an unbalanced structure when the lack of balance is larger and the random effects structure is complex.

In the next section, we restrict ourselves to balanced data.

## 2 Linear term in the fixed effect

We consider one more variation in the simulation, a linear term in the fixed effect, by adding $0.5T$ to the fixed effect. Previously, the time index variable $T$ was linear in the random effect in RIC cases, and the fixed effect is generated based on a tree structure. Since any regression tree cannot perfectly recover the linear structure, we would expect predictive power to degrade compared to the earlier situation where the assumed tree structure of the fixed effect is correct. Figure 2 gives simulation results in terms of predictive performance (note that since the fixed effect does not take the form of a tree it is impossible to recover the "correct" tree structure).

Comparison with Figure 8 of the paper shows that for all cases, the PMSEs of the response variable $y$ are only slightly larger than the ones we obtained before, and the unbiased tree generally outperforms the CART-based tree, meaning that the existence of the linear term $0.5T$ in the fixed effect does not hurt the overall (PMSE of $y$) predictive performance very much. However, the PMSEs of the fixed effect degrade rapidly for both methods, particularly in the RIC cases. Thus, while the random effect can compensate somewhat for poor estimation of the fixed effect, as the random effect becomes more complex the poor estimation of the fixed effect becomes more pronounced. Still, the unbiased tree outperforms the CART-based tree even in the situation where the tree model is not the correct fixed effect.

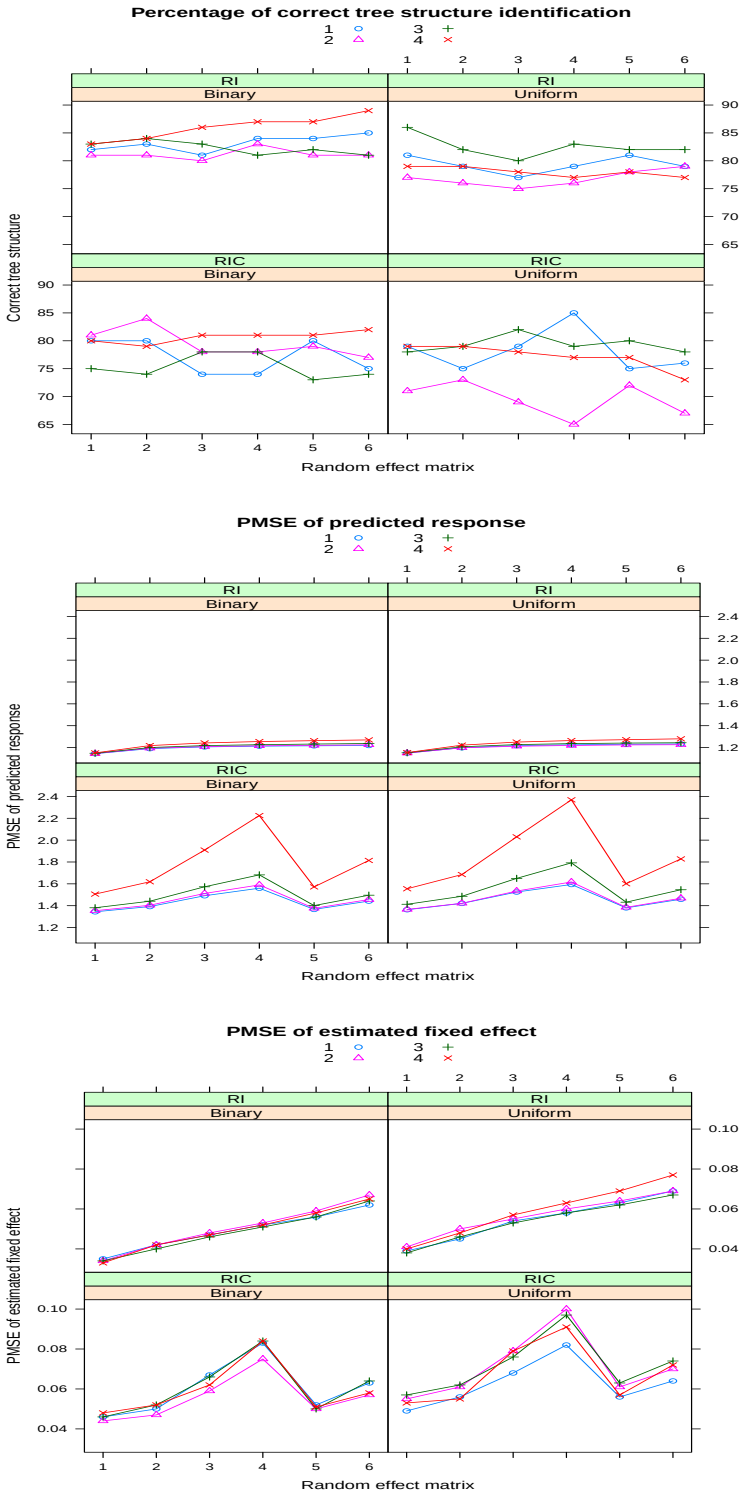


Figure 1: Proportion of the time that the correct tree structure is recovered, PMSE of the response $y$, and PMSE of the fixed effect, respectively, for each unbalanced type.

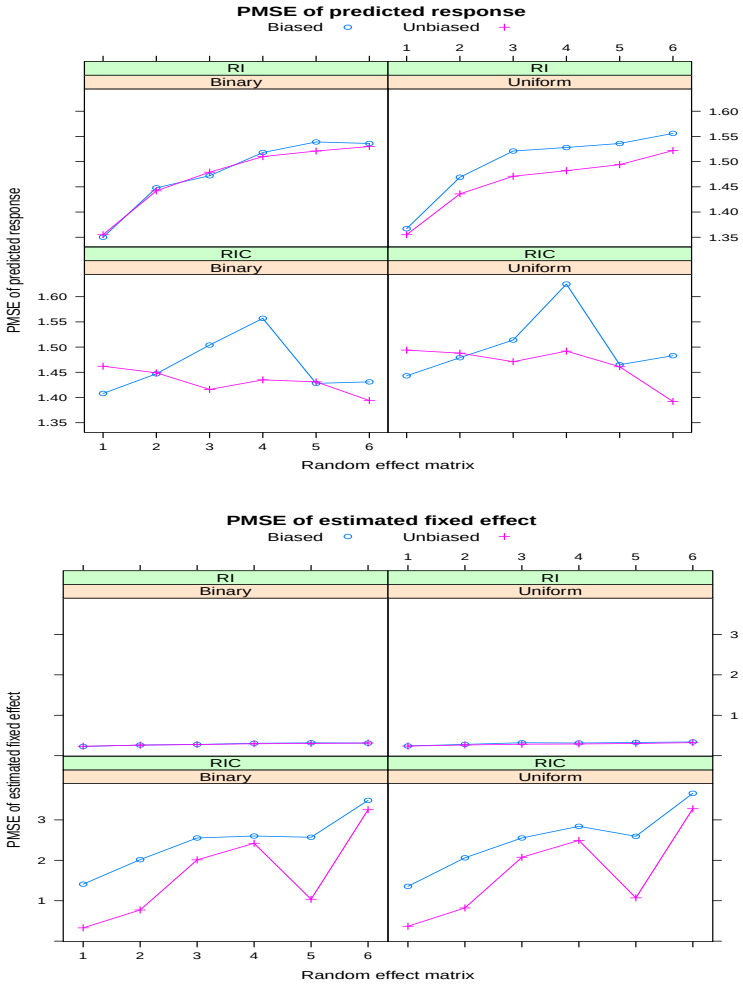


Figure 2: PMSE of the response $y$ and the fixed effect, respectively, with a linear effect term $0.5T$ in the fixed effect without $X_4$ included.