Variable selection: forward selection

2020-05-10

We do forward selection to choose a subset of the predictor variables for the final model.

$$Y_i = S_i(\beta + b_i + \Gamma(\alpha' X_i)) + \epsilon_i. \tag{1}$$

We have the set of biosignatures is $\mathbf{X} = \{X_i;\ i \in \mathbf{N}, 1 \leq i \leq 215\}$

Procedures to pick up n biosignatures:

- 1. We start with a model with only one biosignature X_i. The models are fitted with different X_is and the X_i with the highest *purity* value is chosen, which is the variable called "w0_1329". We mark it as Y_1
- 2. Remove the variable Y_1 from set X ($X^{-1} = X/Y_1$). Choose another covariate X_i from the set X^{-1}. Fit the models with $\{Y_1, X_i\}$. The variable X_i that achieves the *best performance* is chosen.
- 3. Repeat step 2 for $n-1$ times. Remove variable Y_j ($j \in \{1, .., n-1\}$) from the set X ($X^{-(n-1)} = X/Y_j$). Choose another covariate X_i from the set $X^{-(n-1)}$. Fit the models with $\{Y_j, X_i\}$. The variable X_i that achieves the *best performance* is chosen.

Definition of the *best performance*

As our definition, the *purity* measures how much the differences are between the two distributions. However, I did not use the purity as the performance measure directly, since sometimes the LME may estimate a covariance matrix with small values corresponding to a quite large inverse value and a large purity. Therefore, I also include Γ_1 and Γ_2 to measure the performance, since how could the two distributions separate from each other depends on:

- the angle θ between Γ_1 and Γ_2. (direction)
- the norms of Γ_1 and Γ_2. (speed)

Therefore, the best performance is the result that has:

- large *purity* in a reasonable range (< 10000)
- large norms of Γ_1 and Γ_2
- angle $\theta \in [60°, 120°]$

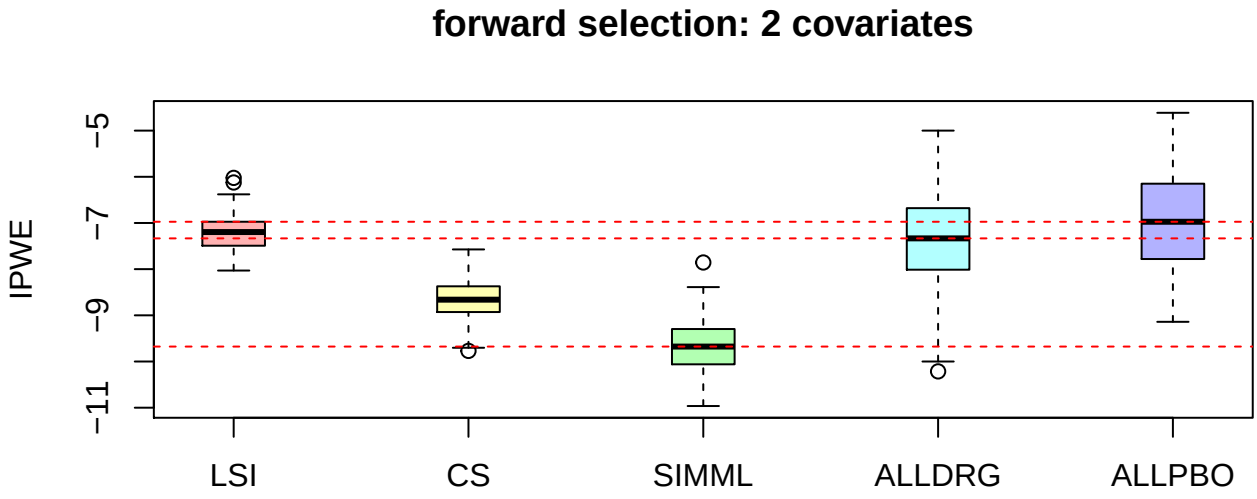
Based on the criteria and procedures, the covariates that are picked up from the first time, second time, to the 14th time are:

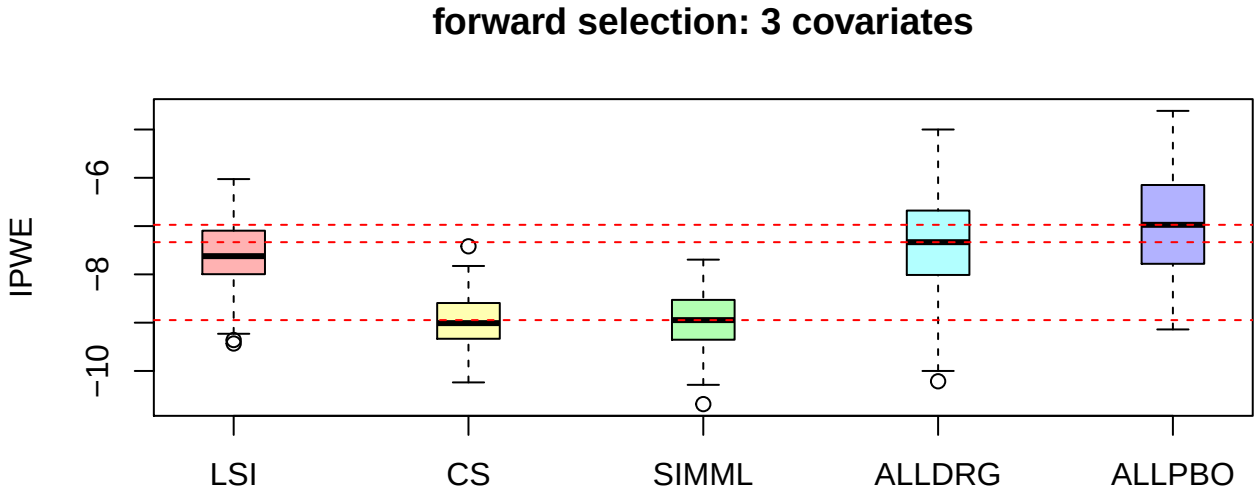
1. w0_1329;	2. decreaseRate;	3. w0_1413;	4. w0_1123;	5.w0_1235
6. w0_1271;	7. w0_1051	8. w0_1143;	9. w0_1113;	10. w0_1243
11. w0_1181;	12. w0_1171;	13. w0_1355;	14. w0_1237	15.w0_1021
16. w0_1147	17. w0_1407	18.w0_1309		

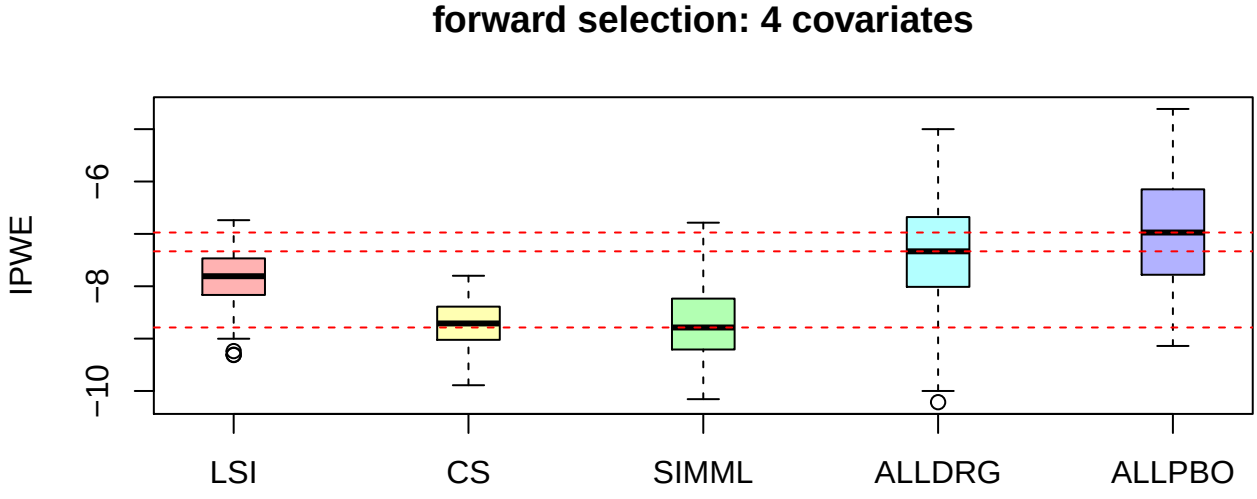
I then fit the model with 2, 3, to 14 of those covariates and calculate the IPWE (10 fold cross validation with 100 times repetitions). The boxplots are shown below.

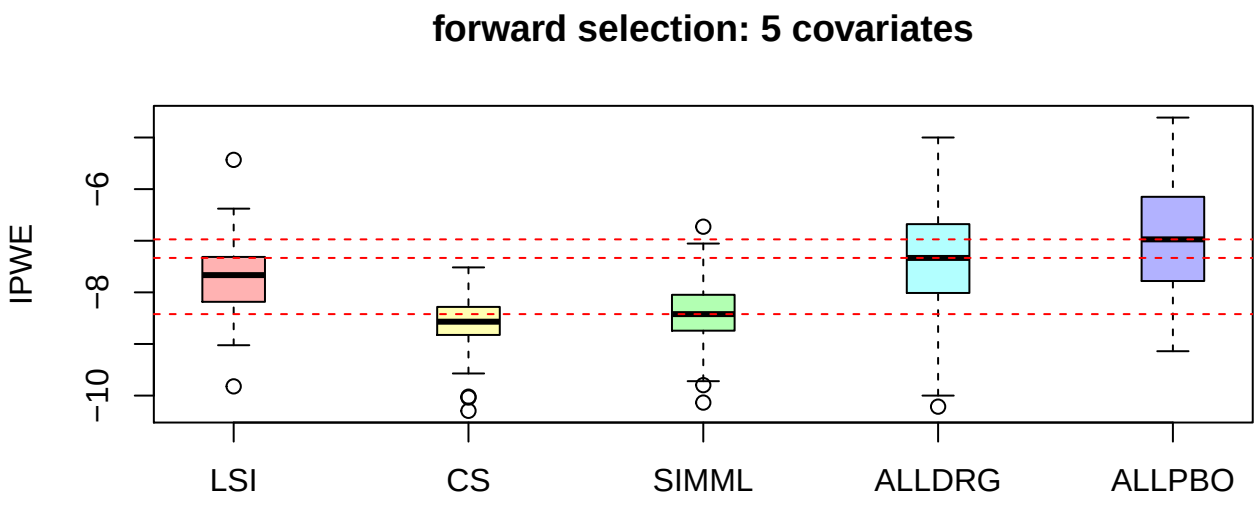
With more covariates selected, the preformance of our longituidnal model increases. 14 covariates might be not enough, I am choosing more covariates.

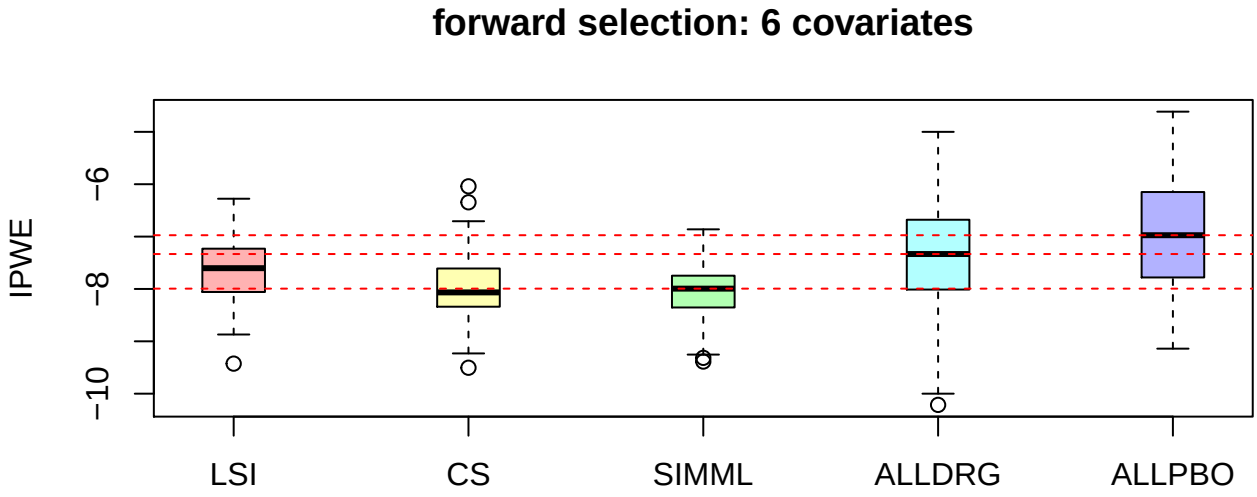
(In the box plots, the method from left to right are our longitudinal method, linear change score method, SIMML, IPWE called by subjects in drug group, IPWE called by subjects in placebo group)

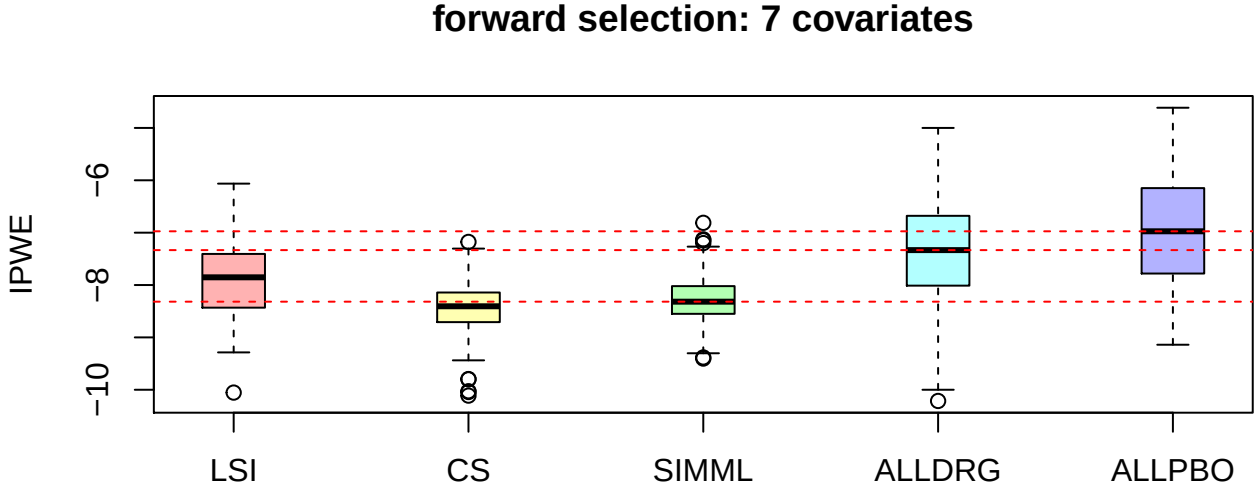

forward selection: 2 covariates
IPWE
−5
−7
−9
−11
LSI
CS
SIMML
ALLDRG
ALLPBO

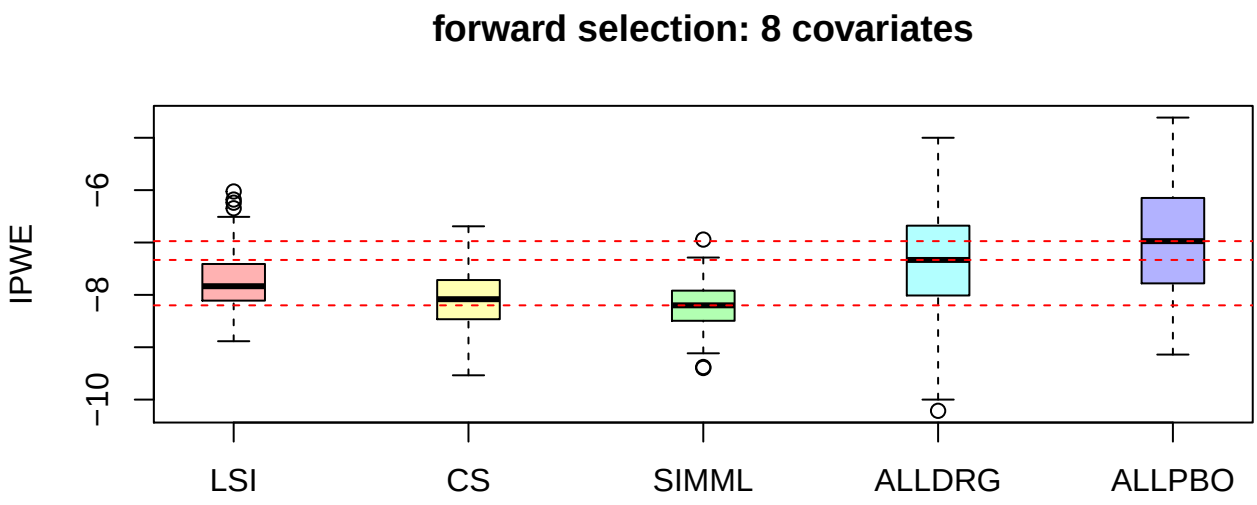

forward selection: 3 covariates
IPWE
−6
−8
−10
LSI
CS
SIMML
ALLDRG
ALLPBO

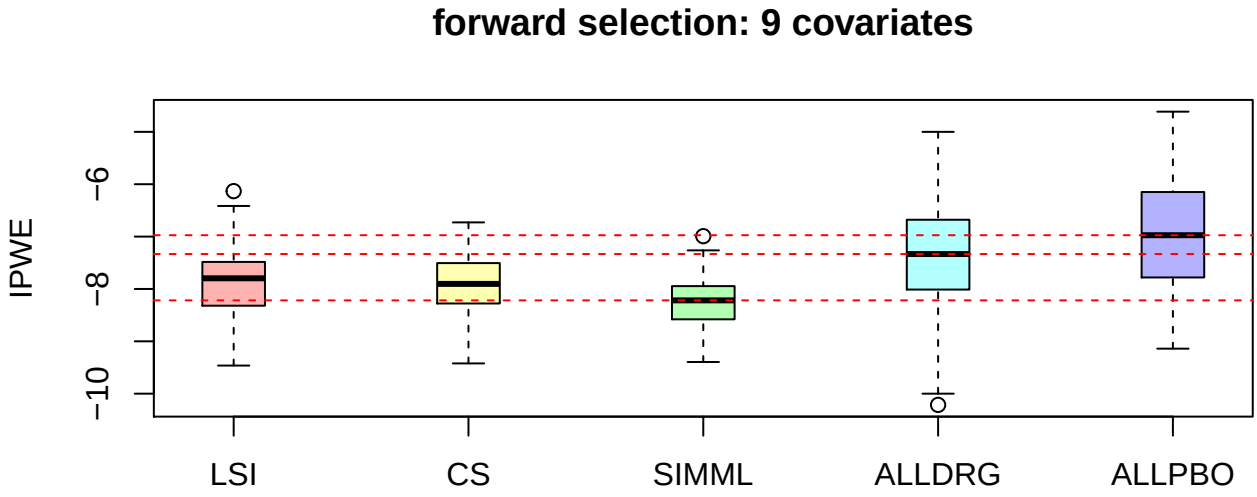

forward selection: 4 covariates
IPWE
−6
−8
−10
LSI
CS
SIMML
ALLDRG
ALLPBO

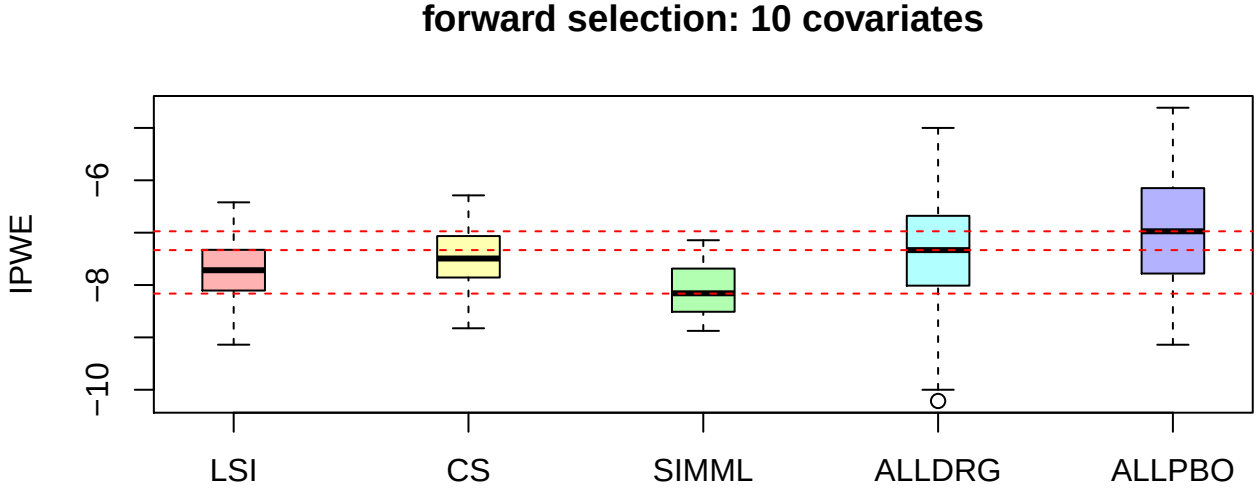

forward selection: 5 covariates
IPWE
−6
−8
−10
LSI
CS
SIMML
ALLDRG
ALLPBO

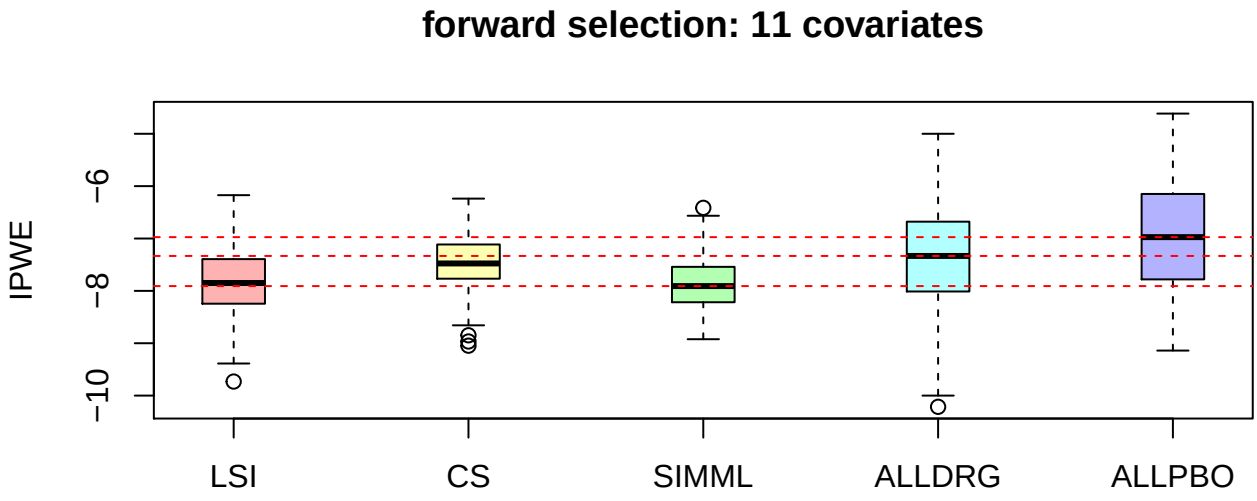

forward selection: 6 covariates
IPWE
−6
−8
−10
LSI
CS
SIMML
ALLDRG
ALLPBO

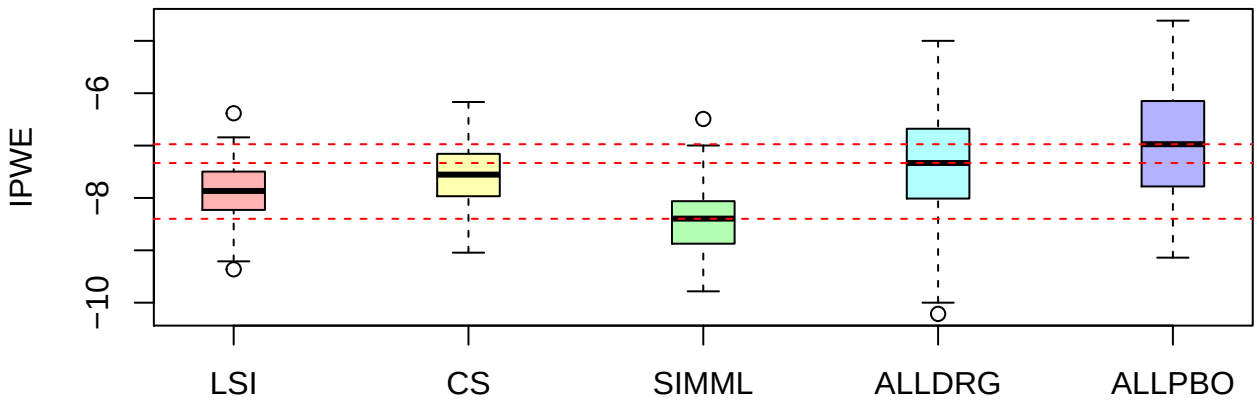

forward selection: 7 covariates
IPWE
−6
−8
−10
LSI
CS
SIMML
ALLDRG
ALLPBO

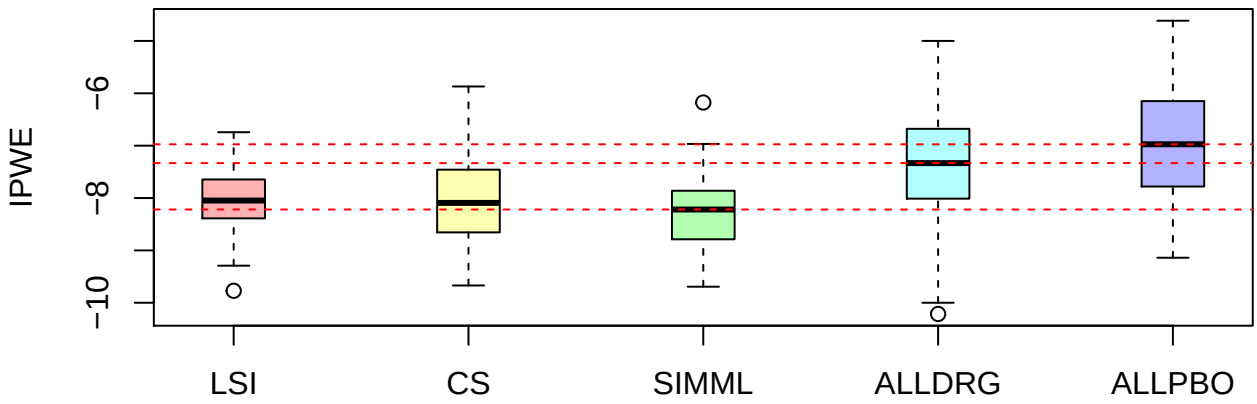

forward selection: 8 covariates
IPWE
−6
−8
−10
LSI
CS
SIMML
ALLDRG
ALLPBO

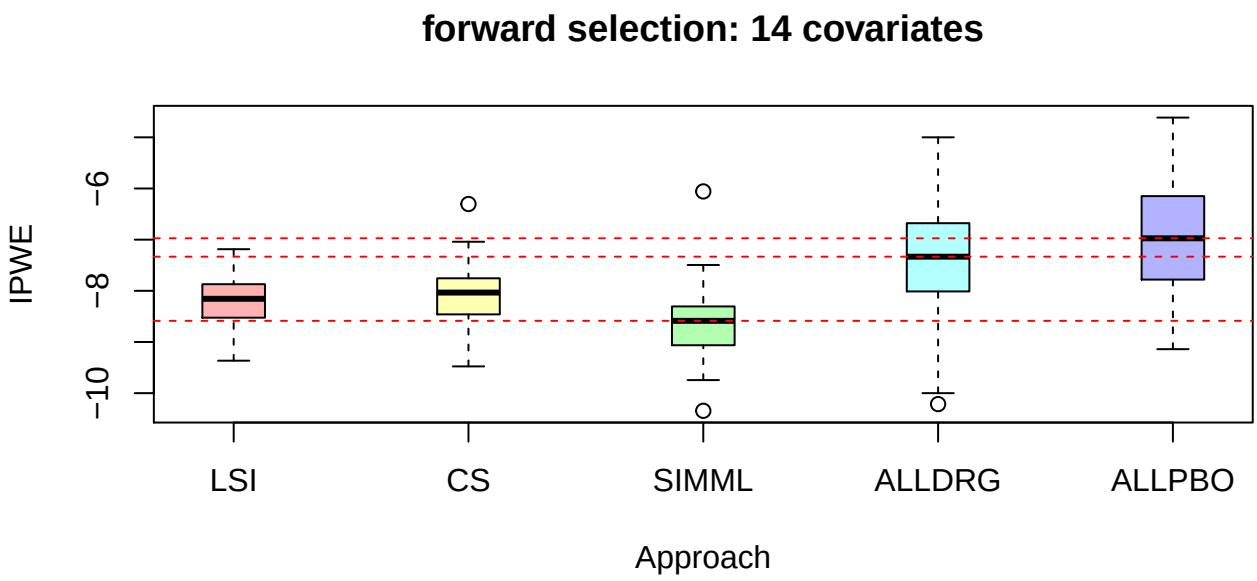

forward selection: 9 covariates
IPWE
−6
−8
−10
LSI
CS
SIMML
ALLDRG
ALLPBO

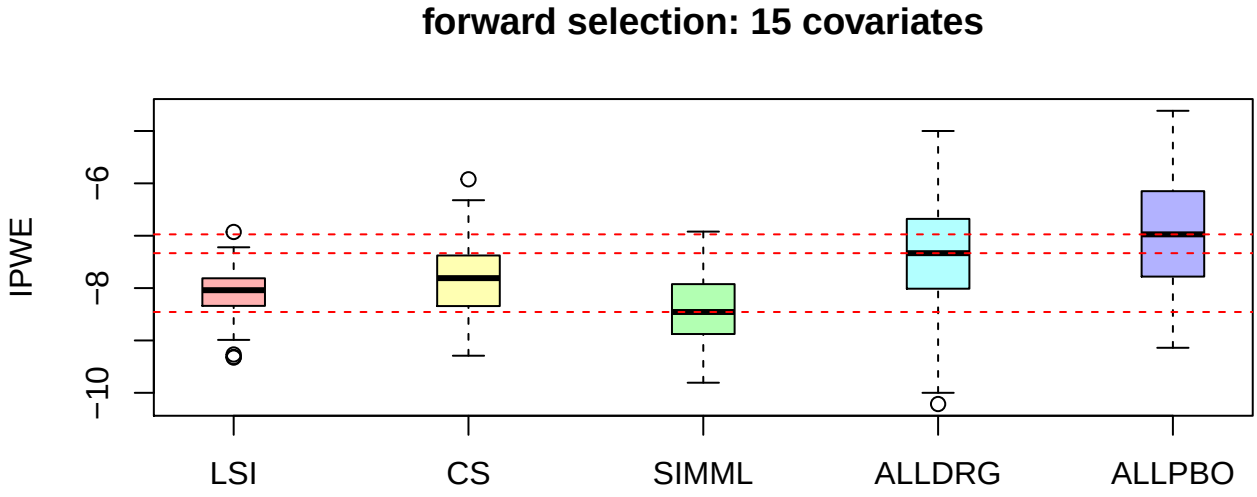

forward selection: 10 covariates
IPWE
−6
−8
−10
LSI
CS
SIMML
ALLDRG
ALLPBO

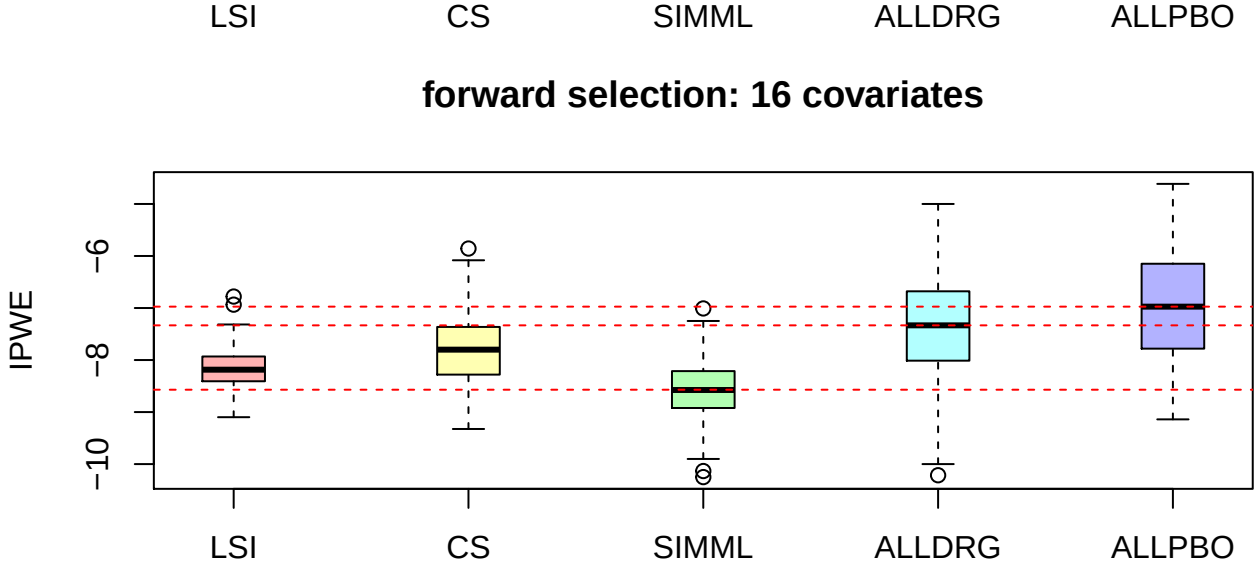

forward selection: 11 covariates
IPWE
−6
−8
−10
LSI
CS
SIMML
ALLDRG
ALLPBO

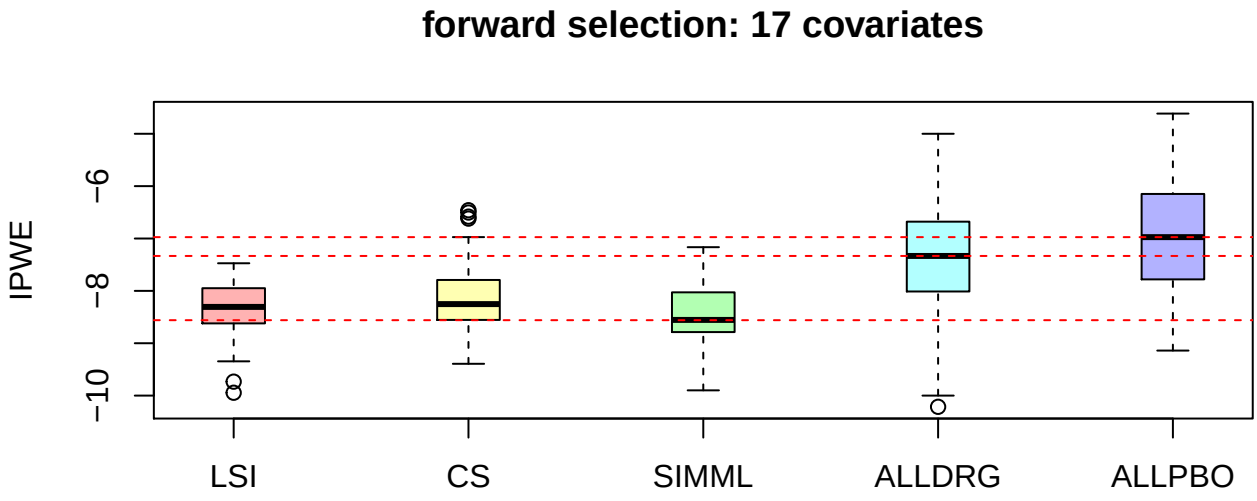

forward selection: 12 covariates
IPWE
−6
−8
−10
LSI
CS
SIMML
ALLDRG
ALLPBO

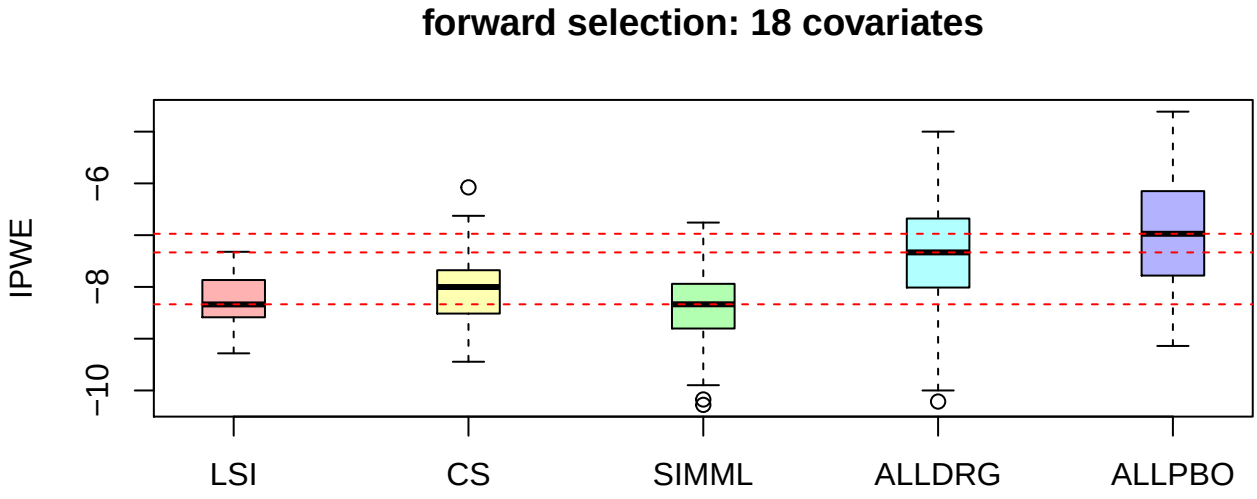

forward selection: 13 covariates
IPWE
−6
−8
−10
LSI
CS
SIMML
ALLDRG
ALLPBO

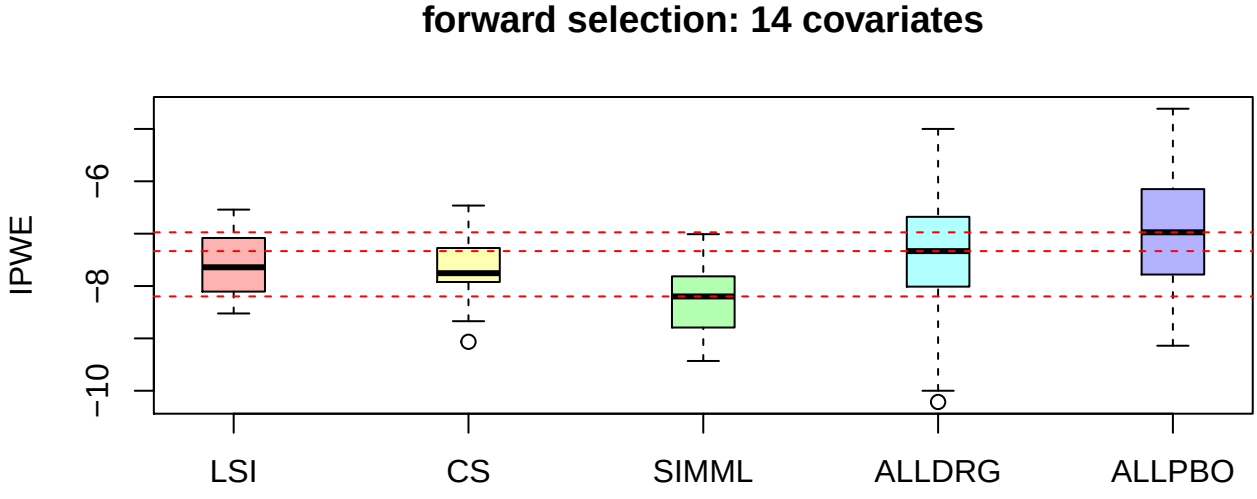

forward selection: 14 covariates
IPWE
−6
−8
−10
LSI
CS
SIMML
ALLDRG
ALLPBO
Approach

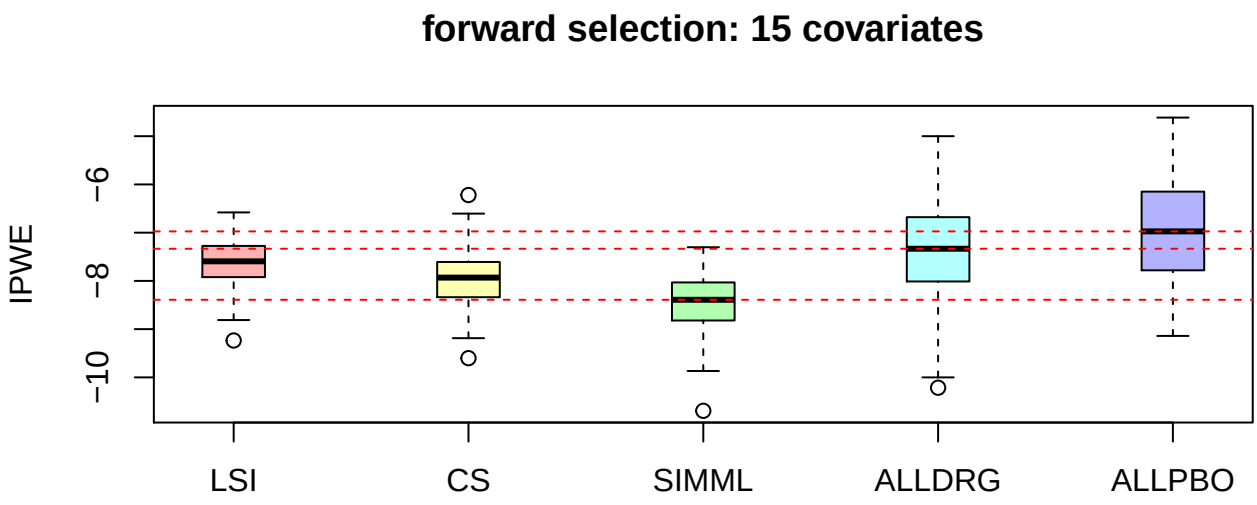

forward selection: 15 covariates
IPWE
−6
−8
−10
LSI
CS
SIMML
ALLDRG
ALLPBO

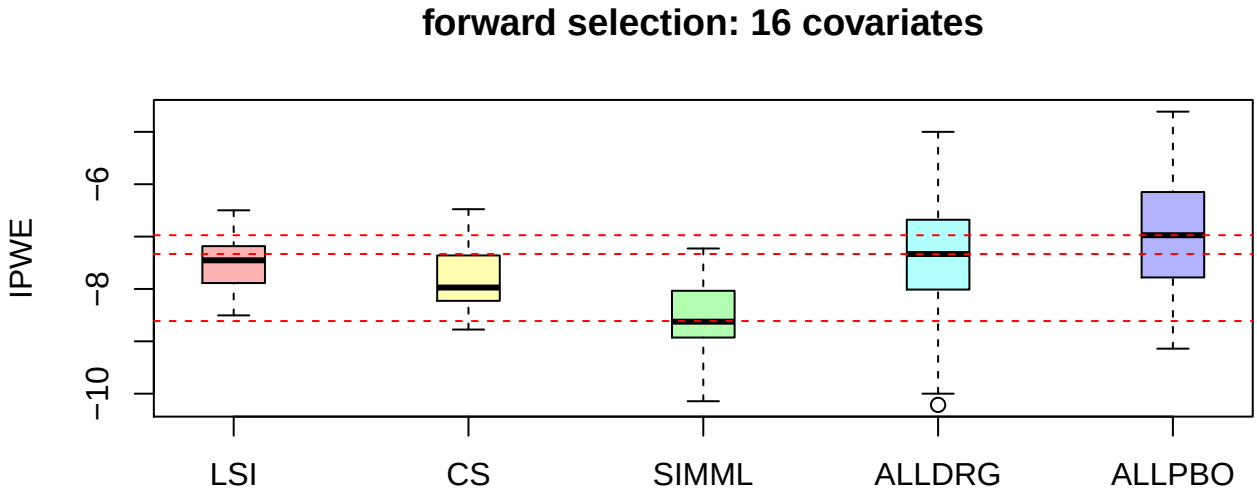

forward selection: 16 covariates
IPWE
−6
−8
−10
LSI
CS
SIMML
ALLDRG
ALLPBO

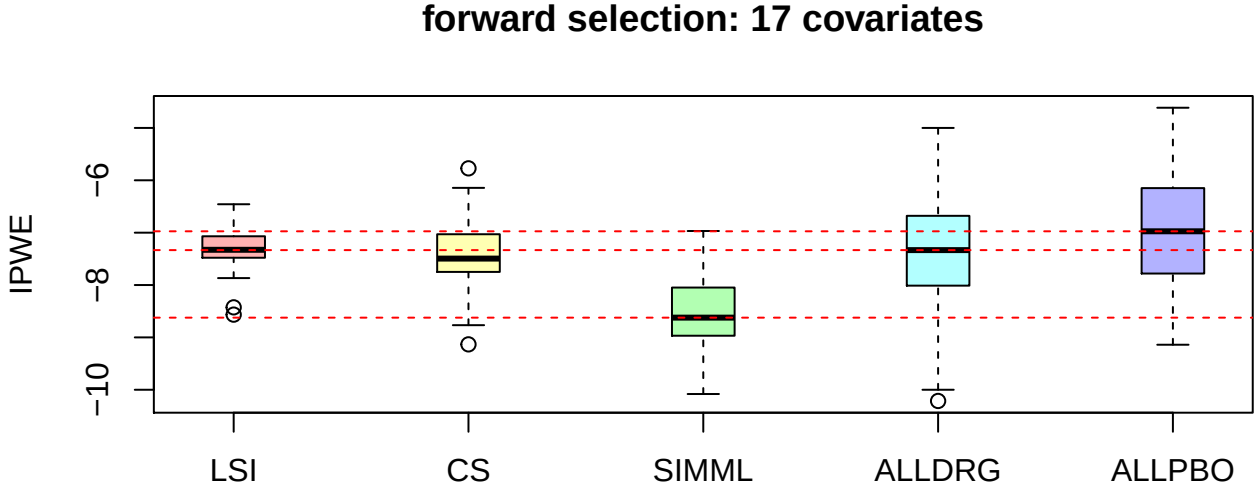

forward selection: 17 covariates
IPWE
−6
−8
−10
LSI
CS
SIMML
ALLDRG
ALLPBO

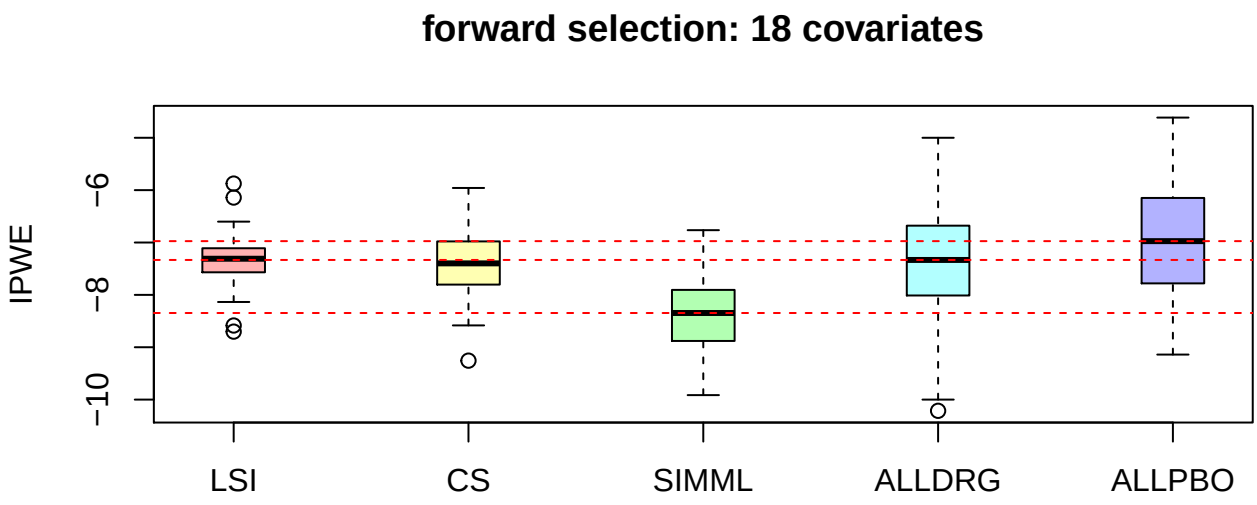

forward selection: 18 covariates
IPWE
−6
−8
−10
LSI
CS
SIMML
ALLDRG
ALLPBO

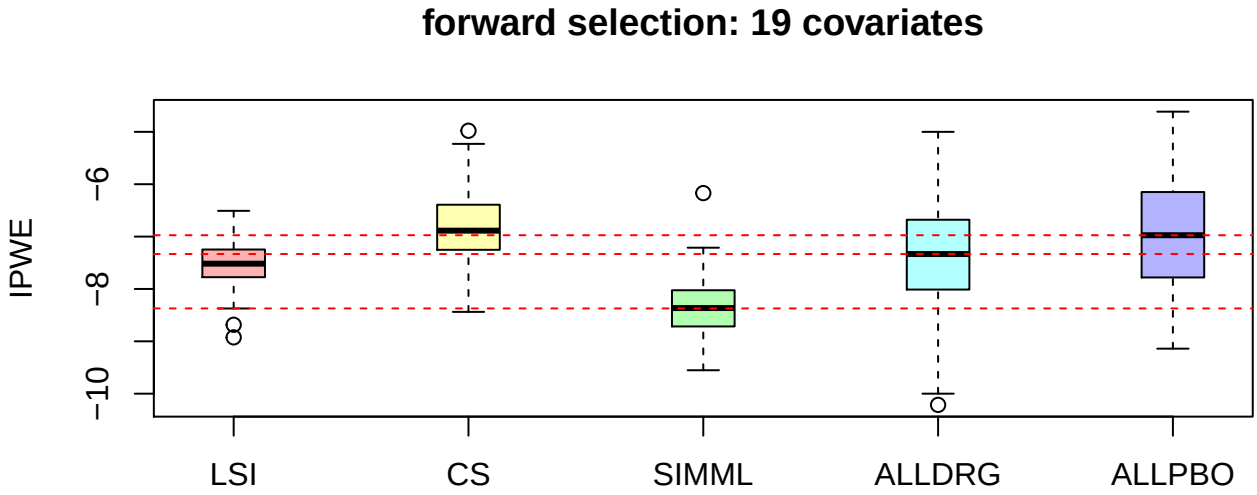

forward selection: 14 covariates
IPWE
−6
−8
−10
LSI
CS
SIMML
ALLDRG
ALLPBO

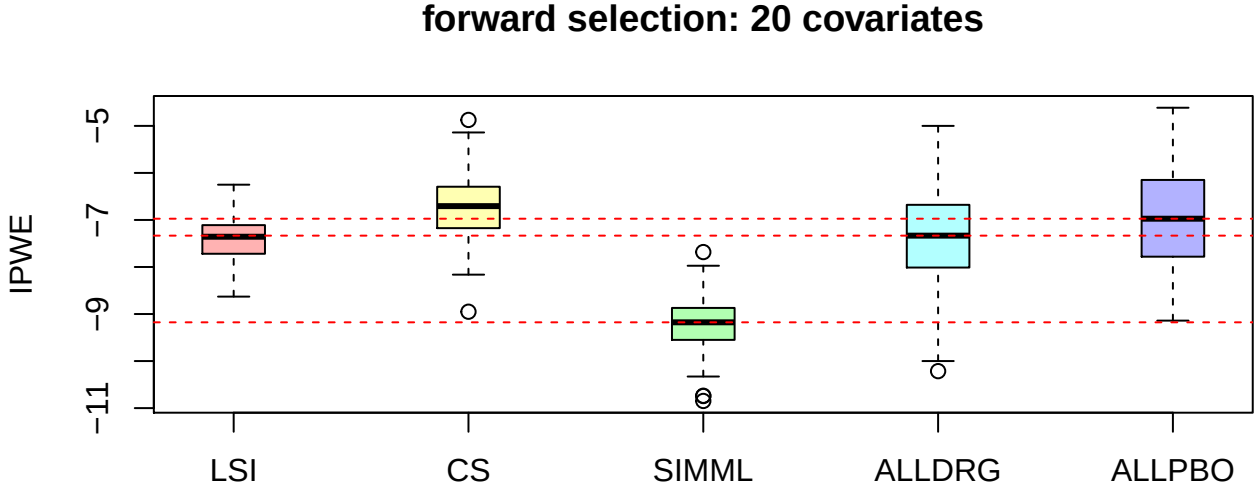

forward selection: 15 covariates
IPWE
−6
−8
−10
LSI
CS
SIMML
ALLDRG
ALLPBO

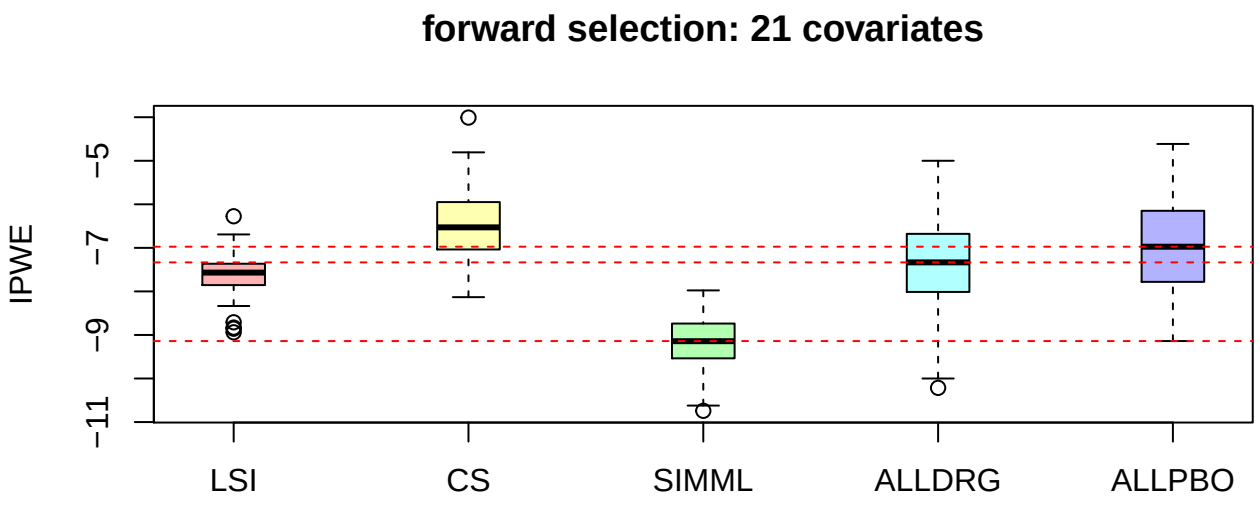

forward selection: 16 covariates
IPWE
−6
−8
−10
LSI
CS
SIMML
ALLDRG
ALLPBO


forward selection: 17 covariates
IPWE
−6
−8
−10
LSI
CS
SIMML
ALLDRG
ALLPBO


forward selection: 18 covariates
IPWE
−6
−8
−10
LSI
CS
SIMML
ALLDRG
ALLPBO


forward selection: 19 covariates
IPWE
−6
−8
−10
LSI
CS
SIMML
ALLDRG
ALLPBO


forward selection: 20 covariates
IPWE
−5
−7
−9
−11
LSI
CS
SIMML
ALLDRG
ALLPBO


forward selection: 21 covariates
IPWE
−5
−7
−9
−11
LSI
CS
SIMML
ALLDRG
ALLPBO