Results: forward selection

2020-05-21

Outline

If we consider the random effect for the quadratic term,

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

$$z = \beta + b_i + \Gamma(\alpha' X_i) \sim N\big(\beta + \Gamma(\alpha' X_i), D\big)$$

If we don't consider the random effect of the quadratic term,

$$Y_i = S_i\big(\beta + +\Gamma(\alpha' X_i)\big) + Z_i b_i + \epsilon_i. \tag{2}$$

$$Y \sim N\big(S(\beta + \Gamma(\alpha' X_i)), ZDZ' + \sigma I\big)$$

Variable selection:

Embarc

- 1. Forward selection:
  - Criteria: purity
  - Purity calculation: treat coefficient z as normal distribution
- 2. Forward selection:
  - Criteria: purity
  - Purity calculation: use the outcome Y as normal distribution
- 3. Forward selection:
  - Criteria: purity
  - Purity calculation: treat coefficient z as normal distribution, add penalty to the D matrix, $D^* = D + \lambda I$, set $\lambda = 0.1$.
- 4. Forward selection:
  - Criteria: IPWE (10 fold CV) (slow)
  - Purity calculation: treat coefficient z as normal distribution

Simulation

- Forward selection:
  - Criteria: purity
  - Purity calculation: treat coefficient z as normal distribution
- PCD, IPWE are calculated by using the selected coefficients
- PCD, IPWE are also calculated by using the true coefficients

EMBARC

I have 3 EMBARC dataset:

- one contains the longformat HDRS score at each week and demographic (287 subjects)
- one contains the behavior measure (166 subjects)
- one contains cortical thickness (158 subjects)

However only 103 subjects are in all of these three datasets. And there are more 215 available covariates in total.

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

Procedures to pick up n biosignatures:

- 1. Let M_0 denote the null model, which contains 0 covariates
- 2. For $k = 0, ..., p-1$ (p = 215 covariates):
  - (a) consider all $p - k$ models that augment the covariates in M_k with one additional covariate.
  - (b) choose the best among these $p - k$ models and call it M_{k+1}.
  - the best is defined as whether:
    - largest purity
    - smallest IPWE
- 3. For the selected covariates, conduct 10 fold CV with methods (longitudinal single index, linear change score method, SIMML). This procedure is repeated for 100 times.

Results

Forward selection 1

- Criteria: purity
- Purity calculation: treat coefficient z as normal distribution

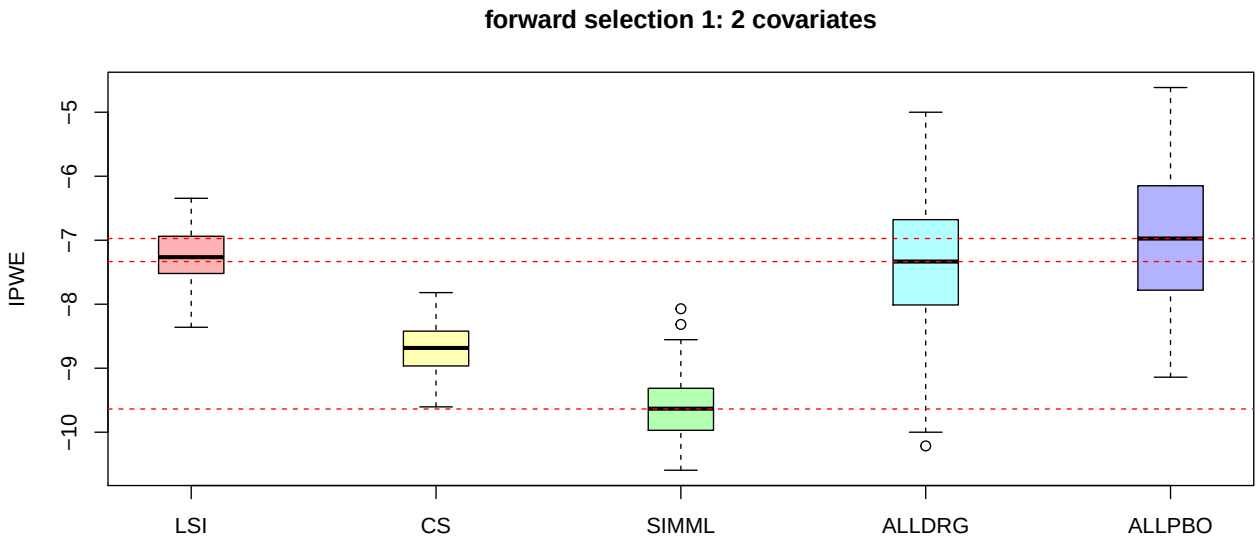
The selected covariates are:

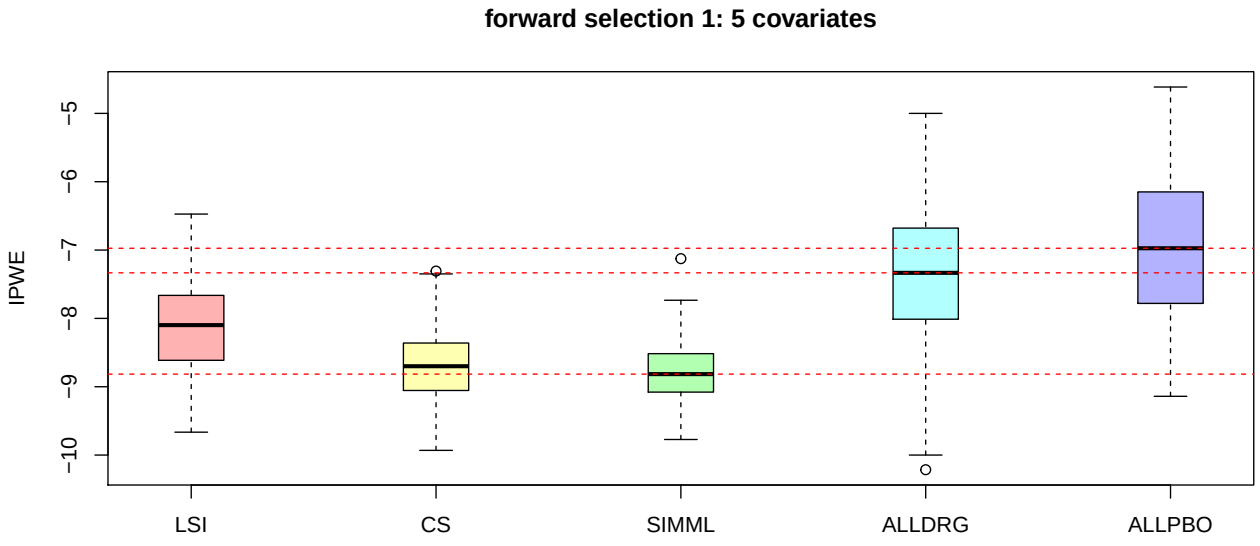
w0_1329	w0_1187	w0_1011	w0_1181	w0_1171
decreaseRate	w0_1029	w0_1437	w0_1285	w0_1183
w0_1295	w0_1179	w0_1431	w0_1241	w0_1261
w0_1235	w0_1339	w0_1143	w0_1287	w0_1163
w0_1127	w0_1379	w0_1293	w0_1435	w0_1407
w0_1087	w0_1051	w0_1311	w0_1213	w0_1351
w0_1065	w0_1227	w0_1101	w0_1145	w0_1045
w0_1343	w0_1373	w0_1251	w0_1385	w0_1073

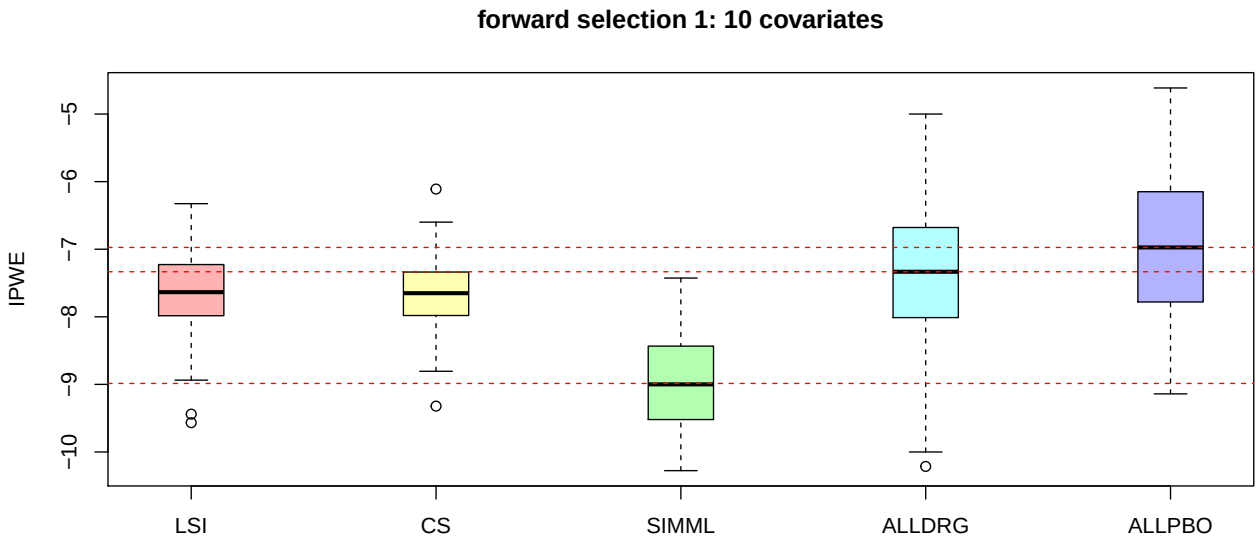
covarname	theta	purity	normgamma1	normgamma2
w0_1073	2.4802584	41.913281	1.3901240	0.0087236
w0_1075	2.8928182	38.906675	0.8872195	0.0198000
w0_1077	0.5300800	726.870345	1.0171346	0.0080150
w0_1079	0.2908866	42.047154	0.8020257	0.0384993
w0_1081	3.1302543	15.487358	2.1713318	0.7242384
w0_1083	3.1313148	5.320124	0.3271038	0.2481159
w0_1085	3.0120143	30.850141	1.2251240	0.1445782

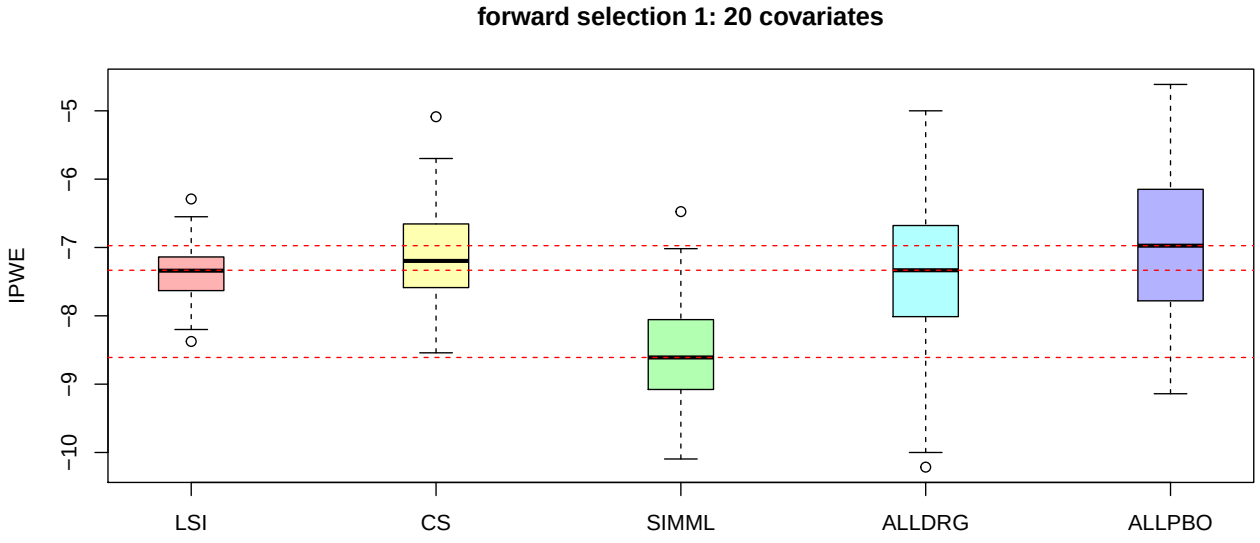
w0_1087	2.5638959	1539.958912	1.1535516	0.0120626
w0_1089	3.0808423	58.436855	1.8761604	0.4818834
w0_1091	0.1730876	41.750188	0.8132775	0.0649562

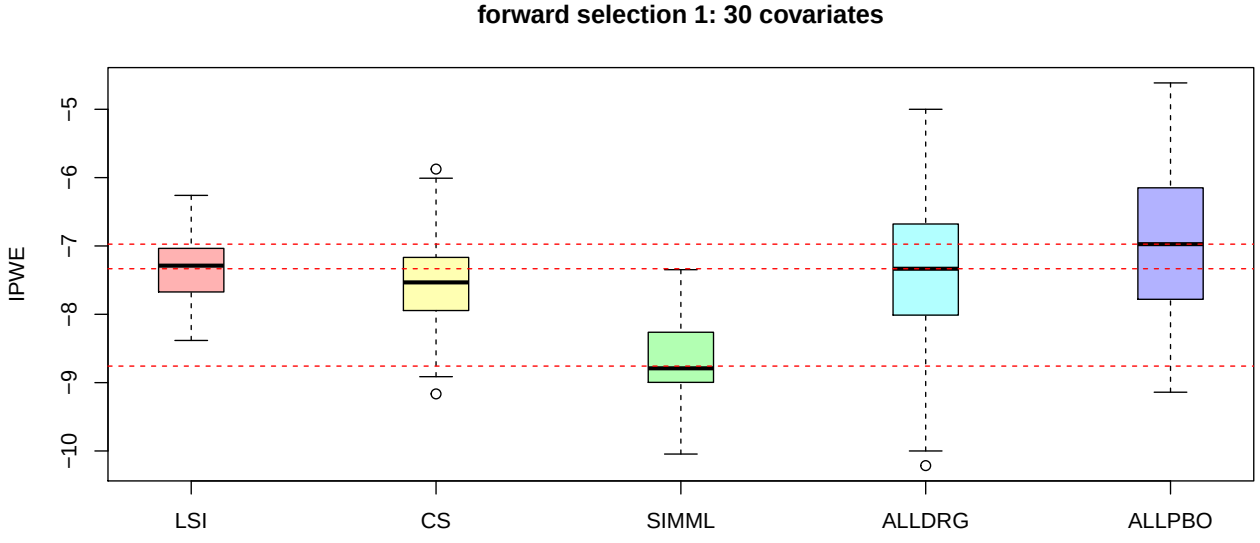
The box plots

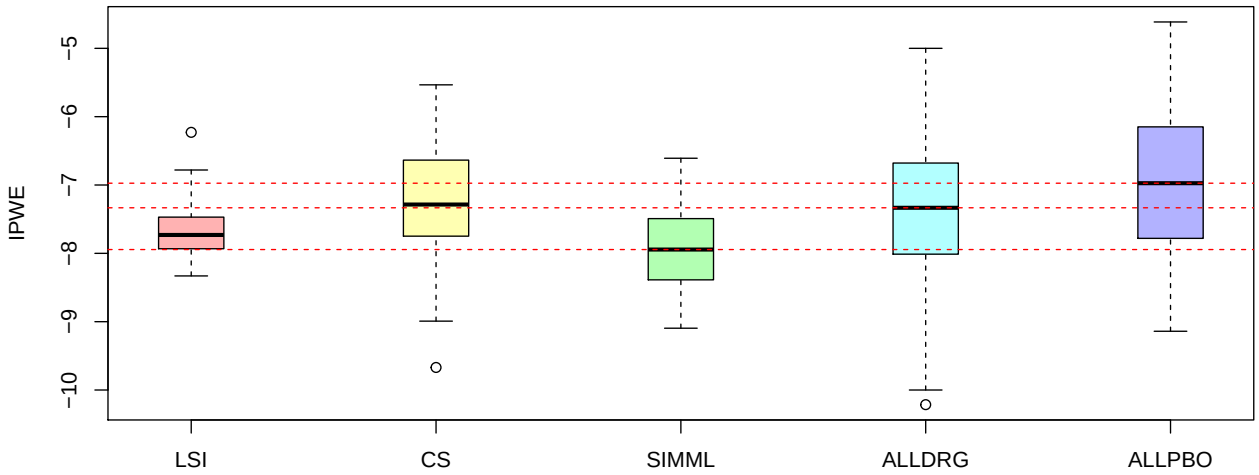





forward selection 1: 10 covariates
IPWE
−5
−6
−7
−8
−9
−10
LSI
CS
SIMML
ALLDRG
ALLPBO


forward selection 1: 20 covariates
IPWE
−5
−6
−7
−8
−9
−10
LSI
CS
SIMML
ALLDRG
ALLPBO


forward selection 1: 30 covariates
IPWE
−5
−6
−7
−8
−9
−10
LSI
CS
SIMML
ALLDRG
ALLPBO



Summary table

	LSI		CS		SIMML		ALLPBO		ALLDRG	
ncov	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
2	-7.25	0.40	-8.70	0.38	-9.62	0.50	-6.95	1.03	-7.35	1.09
5	-8.13	0.68	-8.67	0.53	-8.74	0.47	-6.95	1.03	-7.35	1.09
10	-7.62	0.65	-7.68	0.55	-8.97	0.69	-6.95	1.03	-7.35	1.09
15	-7.26	0.51	-7.70	0.73	-9.06	0.68	-6.95	1.03	-7.35	1.09
20	-7.36	0.39	-7.11	0.69	-8.54	0.73	-6.95	1.03	-7.35	1.09
25	-7.46	0.43	-6.36	0.66	-8.17	0.84	-6.95	1.03	-7.35	1.09
30	-7.33	0.42	-7.54	0.69	-8.67	0.55	-6.95	1.03	-7.35	1.09
35	-7.42	0.41	-7.43	0.83	-8.08	0.59	-6.95	1.03	-7.35	1.09
40	-7.69	0.40	-7.25	0.85	-7.90	0.58	-6.95	1.03	-7.35	1.09

Forward selection 2

- Criteria: purity
- Purity calculation: use the outcome Y as normal distribution

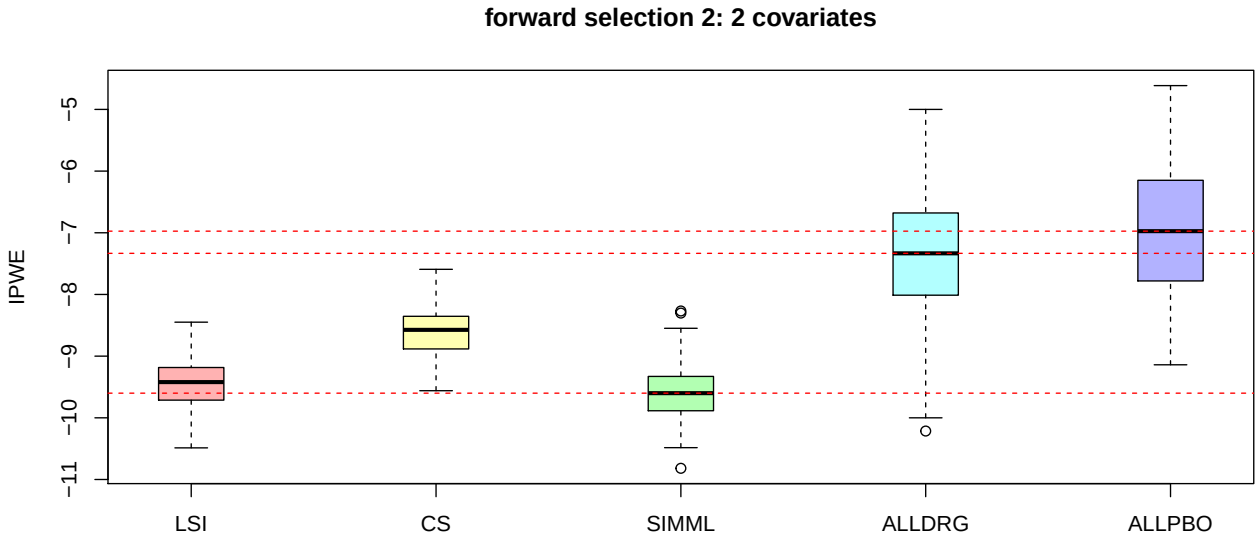
covarname	theta	purity	normgamma1	normgamma2
w0_1213	3.136531	9.157856	2.750641	1.0349410
w0_1215	3.136538	9.249266	2.246515	1.4302695
w0_1217	3.125252	9.517897	3.788619	1.4591685
w0_1219	3.125928	9.147530	2.310265	1.4878516
w0_1221	3.131540	9.154944	2.964951	0.9526455
w0_1223	3.124186	9.050182	2.219308	0.7468090
w0_1225	3.139726	9.235857	2.894701	0.6176171
w0_1227	3.131681	9.592079	4.134425	1.5361126
w0_1229	3.134735	9.016094	1.770590	0.6197603
w0_1231	3.131832	8.995163	1.913513	0.7749481

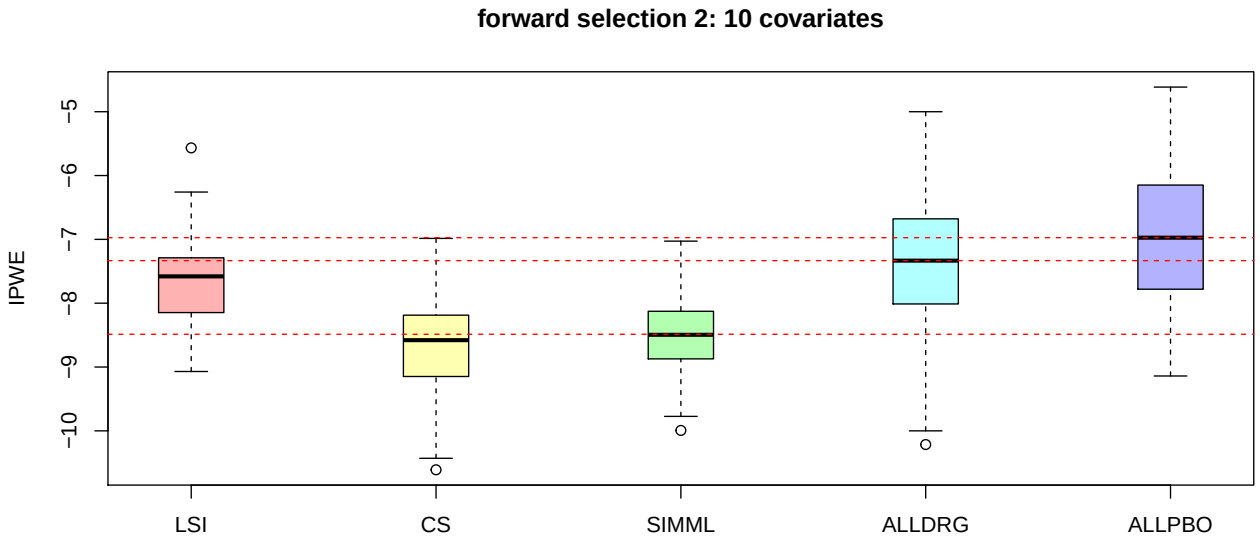
The selected covariates are:

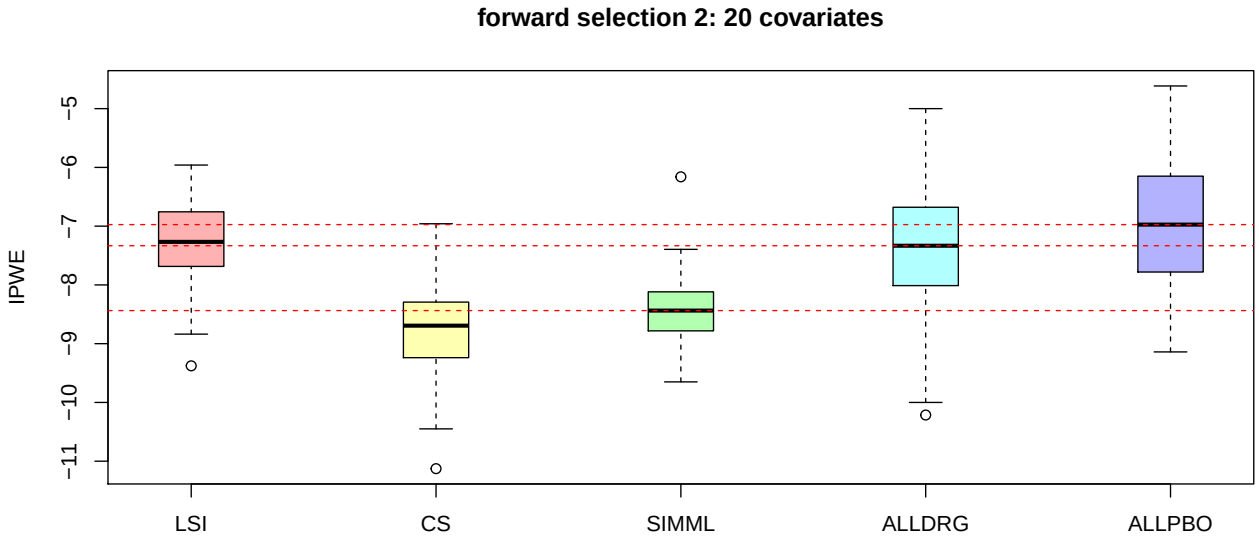
w0_1329	w0_1431	w0_1425	w0_1069	w0_1357
decreaseRate	w0_1235	w0_1337	w0_1177	w0_1071

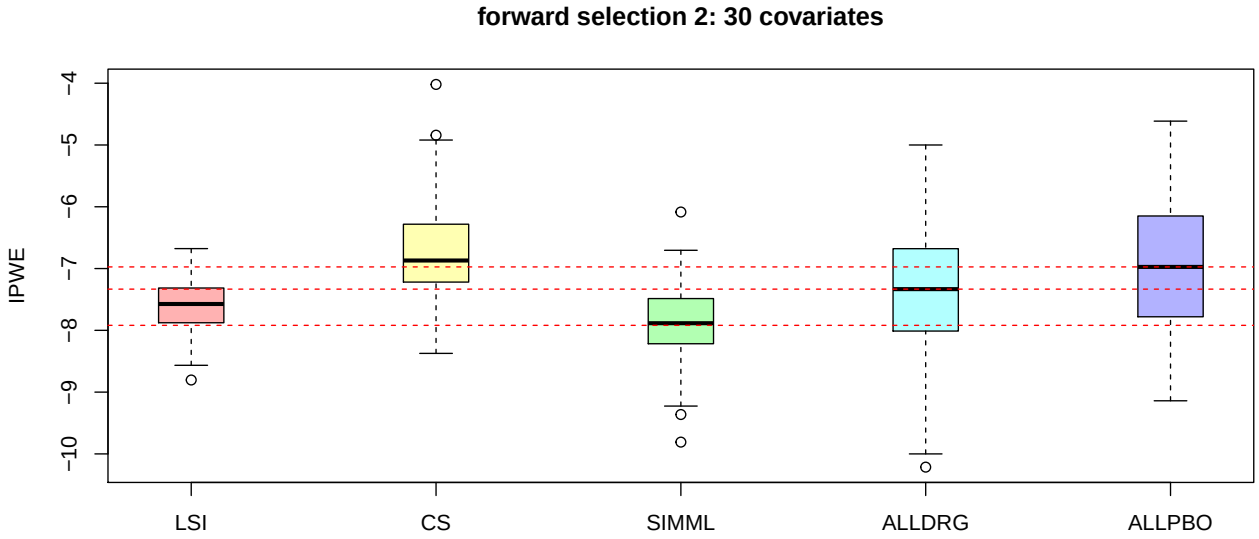
w0_1149	w0_1393	w0_1437	w0_1331	w0_1049
w0_1227	w0_1297	w0_1307	w0_1273	w0_1409
w0_1295	w0_1203	w0_1265	w0_1293	w0_1127
w0_1401	w0_1077	w0_1021	w0_1011	w0_1209

The box plots









Summary table

	LSI		CS		SIMML		ALLPBO		ALLDRG	
ncov	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
2	-9.44	0.39	-8.61	0.41	-9.59	0.44	-6.95	1.03	-7.35	1.09
5	-8.60	0.64	-8.60	0.48	-9.26	0.51	-6.95	1.03	-7.35	1.09
10	-7.68	0.65	-8.65	0.81	-8.48	0.60	-6.95	1.03	-7.35	1.09
15	-7.41	0.69	-8.21	0.87	-7.76	0.66	-6.95	1.03	-7.35	1.09
20	-7.31	0.68	-8.78	0.78	-8.44	0.52	-6.95	1.03	-7.35	1.09
25	-7.56	0.46	-7.96	0.89	-8.04	0.68	-6.95	1.03	-7.35	1.09
30	-7.57	0.43	-6.76	0.78	-7.87	0.63	-6.95	1.03	-7.35	1.09

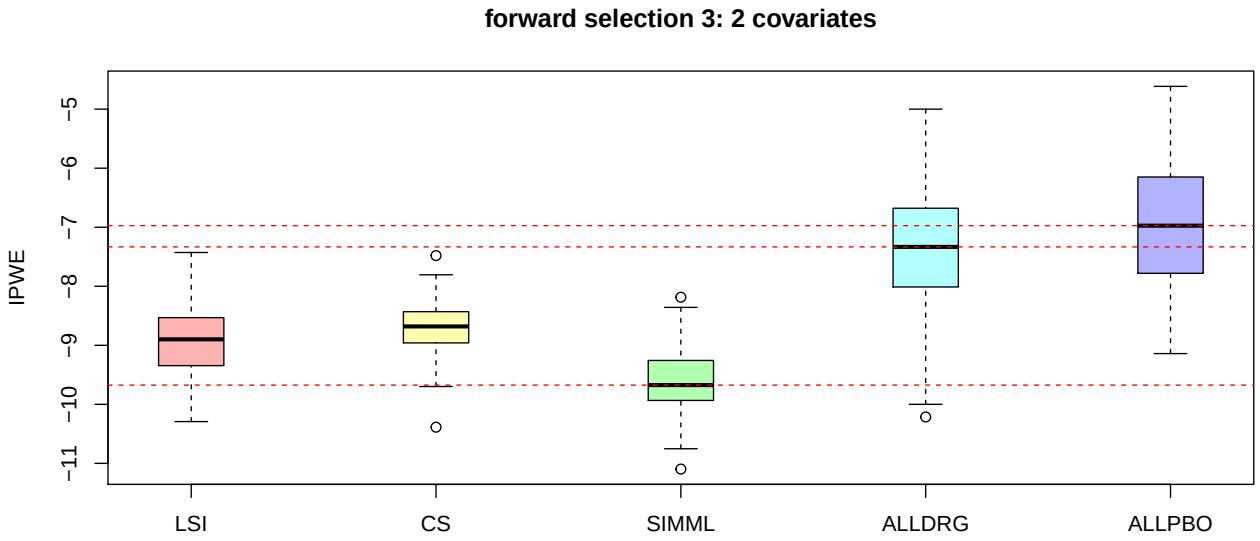
Forward selection 3:

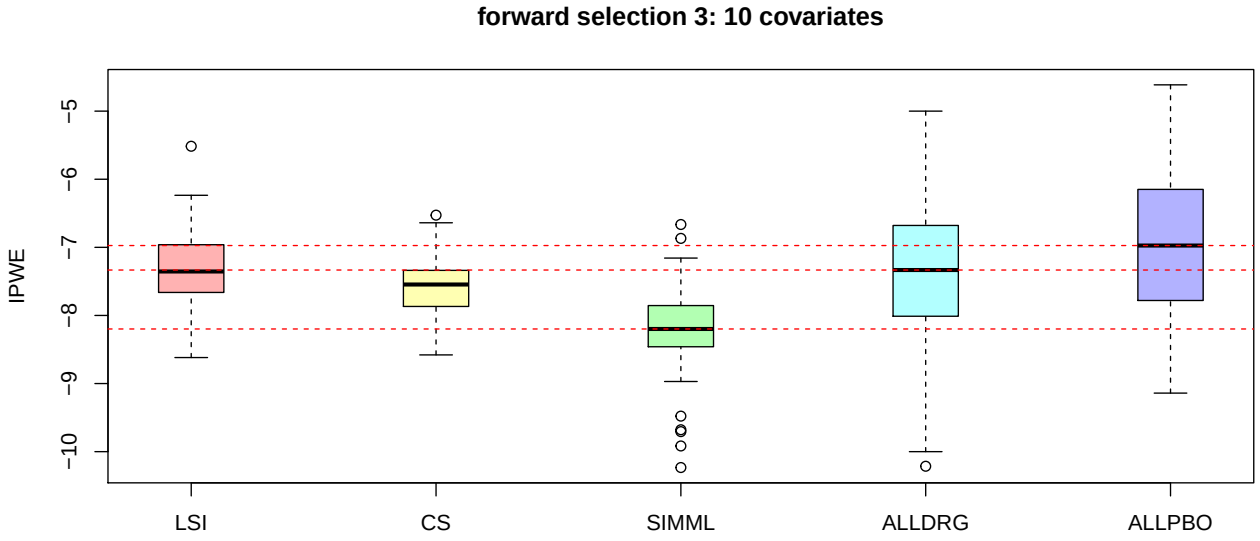
- Criteria: purity
- Purity calculation: treat coefficient z as normal distribution, add penalty to the D matrix, $D^* = D + \lambda I$, set $\lambda = 0.1$.

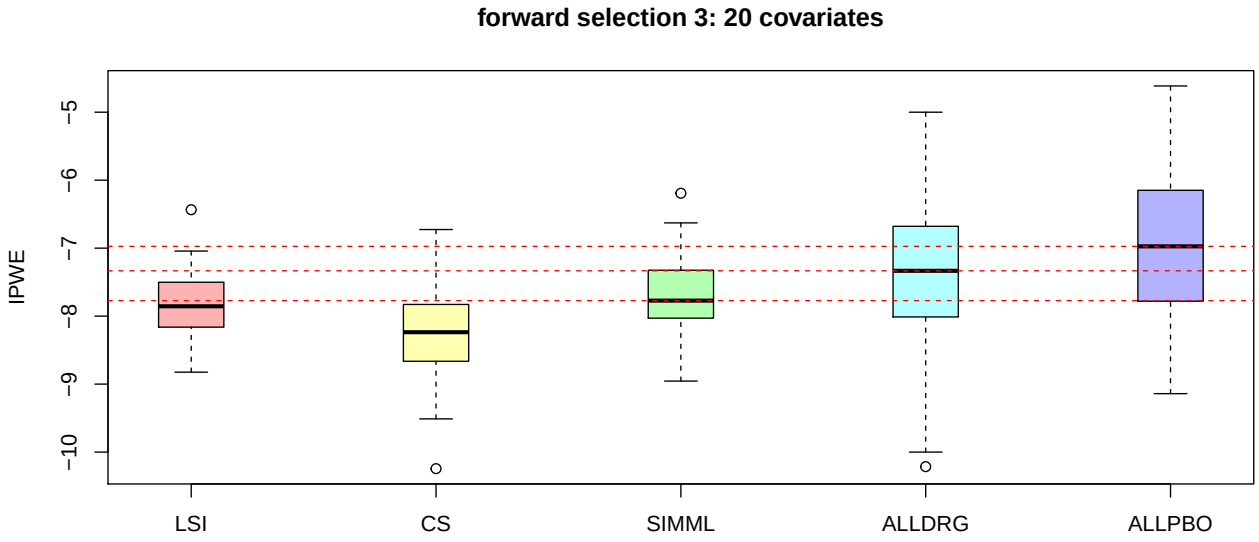
The selected covariates are:

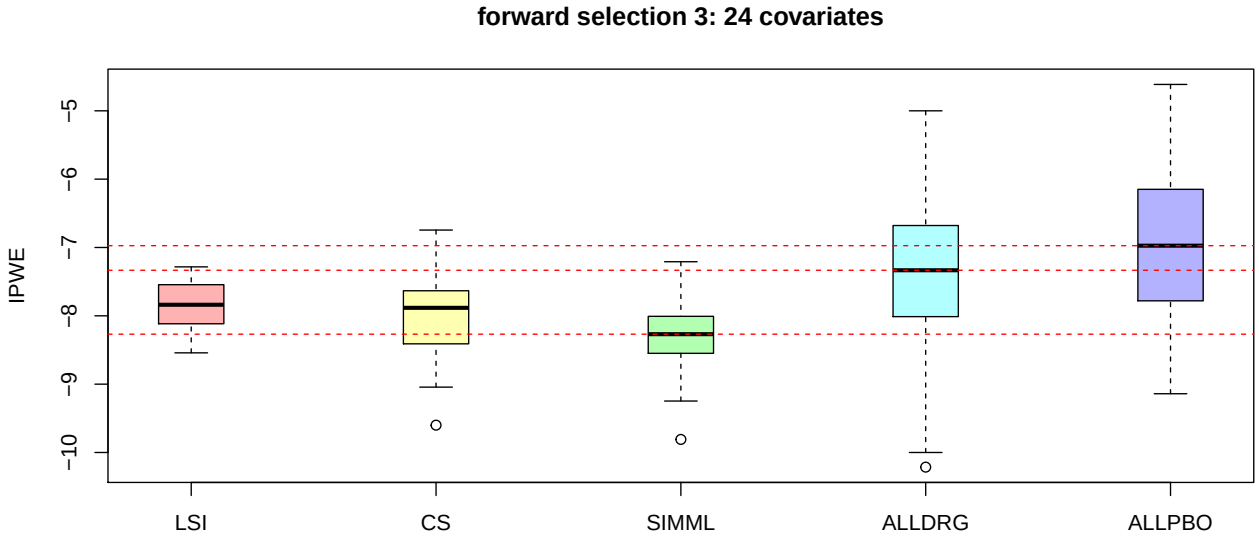
w0_1329	w0_1225	w0_1029	w0_1065	w0_1013	w0_1441
decreaseRate	w0_1089	w0_1407	w0_1213	w0_1257	w0_1263
w0_1149	w0_1137	w0_1111	w0_1395	w0_1357	w0_1103
w0_1375	w0_1277	w0_1175	w0_1041	w0_1181	w0_1385

The box plots









Summary table

	LSI		CS		SIMML		ALLPBO		ALLDRG	
ncov	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
2	-8.90	0.58	-8.68	0.43	-9.63	0.51	-6.95	1.03	-7.35	1.09
5	-7.80	0.58	-8.62	0.42	-8.47	0.57	-6.95	1.03	-7.35	1.09
10	-7.28	0.61	-7.60	0.49	-8.21	0.72	-6.95	1.03	-7.35	1.09
15	-7.27	0.50	-7.46	0.49	-7.80	0.53	-6.95	1.03	-7.35	1.09
20	-7.82	0.50	-8.27	0.68	-7.68	0.53	-6.95	1.03	-7.35	1.09
24	-7.83	0.35	-7.98	0.67	-8.35	0.52	-6.95	1.03	-7.35	1.09

Forward selection 4:

- Criteria: IPWE (10 fold CV) (slow)
- Purity calculation: treat coefficient z as normal distribution

For each subset of covariates, i.e. combination of biosignatures, the data is splitted into 10 parts, 9 as the training set and 1 as the test set, and 10 folds CV is conducted. The estimated IPWE is calculated in the

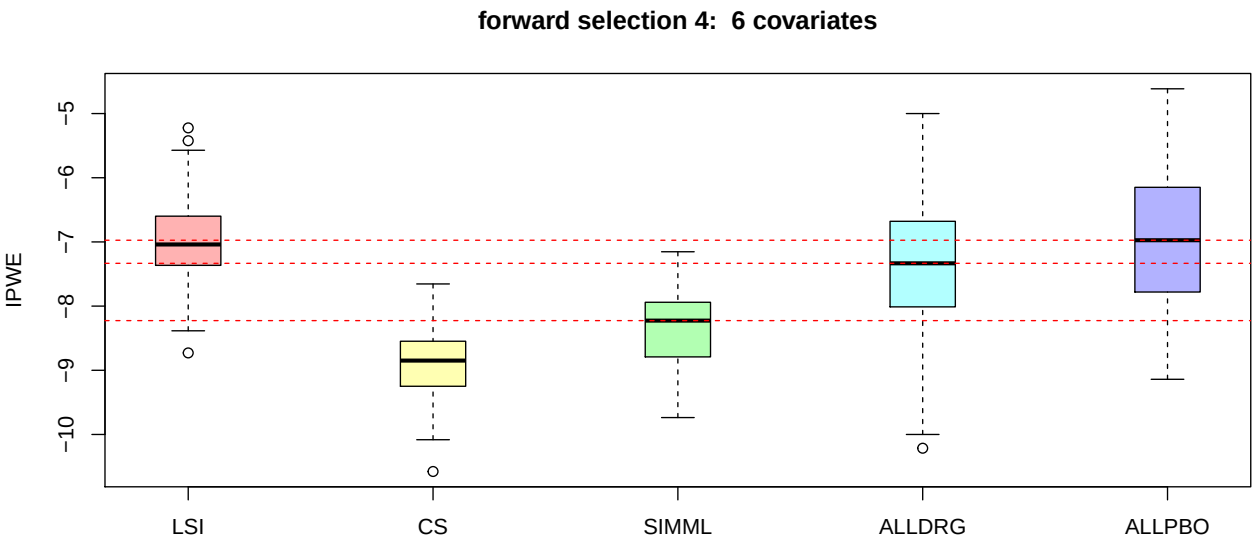
test set for each fold. The mean IPWE across the 10 CV is used as the criteria for selection.

covarname	ipwe
w0_1313	-7.080000
w0_1315	-6.914894
w0_1317	-8.148936
w0_1319	-8.782609
w0_1321	-7.016949
w0_1323	-7.452830
w0_1325	-7.580000
w0_1327	-7.425926
w0_1331	-8.260870

The selected covariates are:

w0_1329	decreaseRate	w0_1191	w0_1407	w0_1319	w0_1163

The boxplots



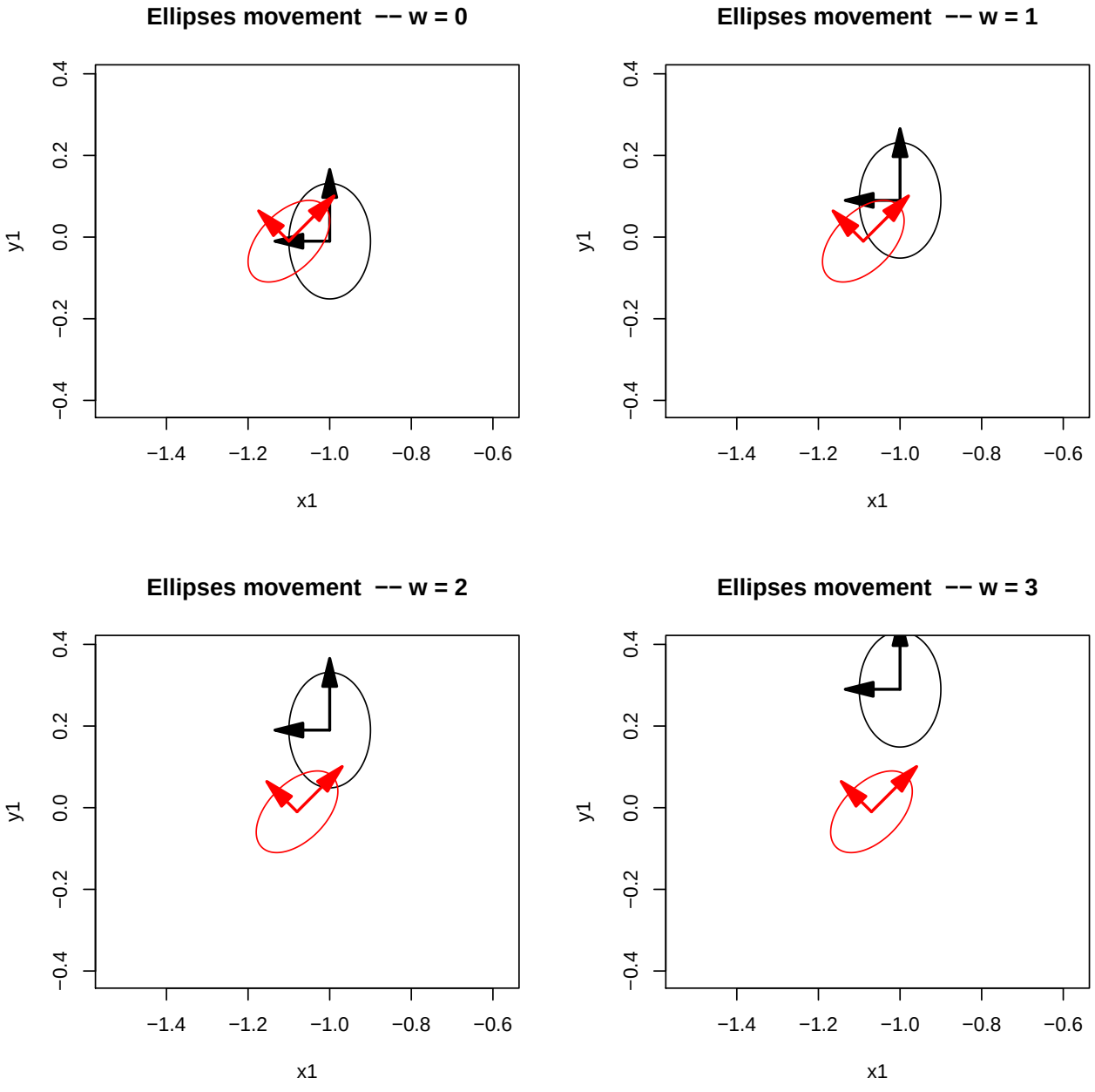
Simulation

Data parameter:

- $n = 100$ in each treatment group (drg, pbo)
- $\beta_{drg} = [40, -1, -0.01], \beta_{pbo} = [40, -1.1, -0.01]$
- $\Gamma_{drg} = [0, 0, 0.1], \Gamma_{pbo} = [0, 0.1, 0], \theta = \frac{\pi}{2}$
- $b_{drg} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.01 & 0 \\ 0 & 0 & 0.02 \end{bmatrix}, b_{pbo} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.01 & 0.005 \\ 0 & 0.005 & 0.01 \end{bmatrix}$
- $\epsilon_{drg}, \epsilon_{pbo} \sim N(0, 3^2)$
- $S = [1, t, t^2], t = [0, 1, 2, 3, 4, 6, 8]$

Covariates:

- 20 covariates used for data generation, $X \sim MVN(0, \Psi)$, Ψ has 1 in the diagonal and 0.5 anywhere else. $X_i, i \in [1, 20]$
- add 80 covariates as the noise (do not used for data generation). each covariate is independently generated from a standard normal distribution. $X_i, i \in [21, 100]$



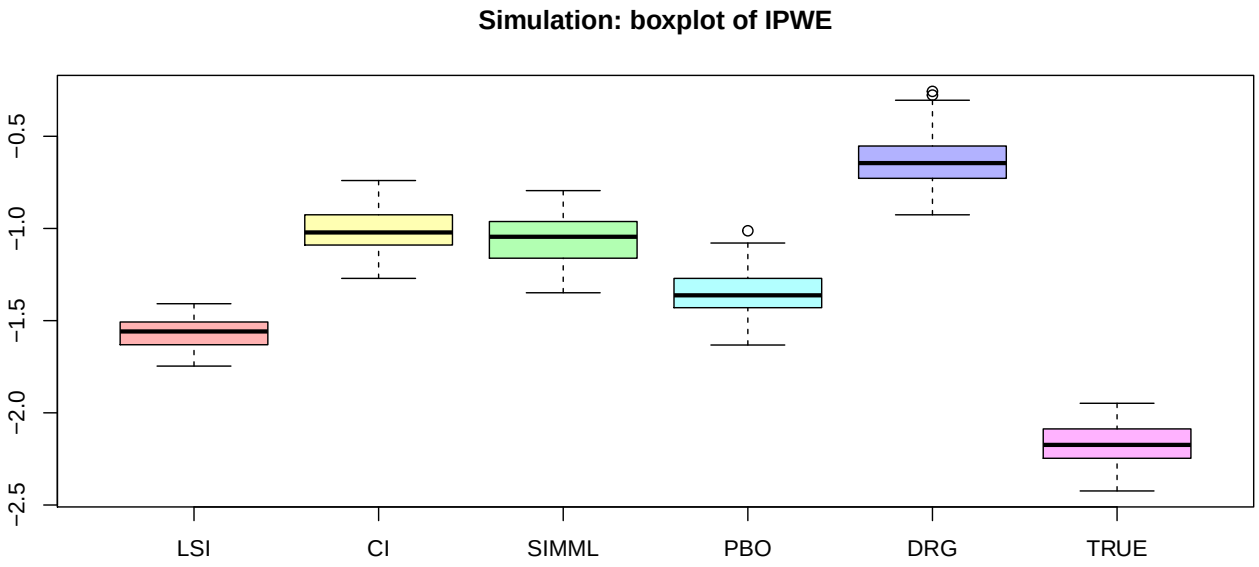
True covariates

If we esitmate the preformances of methods by using the true covariates, i.e., the covariates that used to generate the data set, the proportion of correct decision and IPWE are:

Table 11: Simulation: true covariates: Proportion of correct decision

longitudinal	change.score	simml
0.669	0.600	0.625
0.015	0.027	0.024

IPWE:

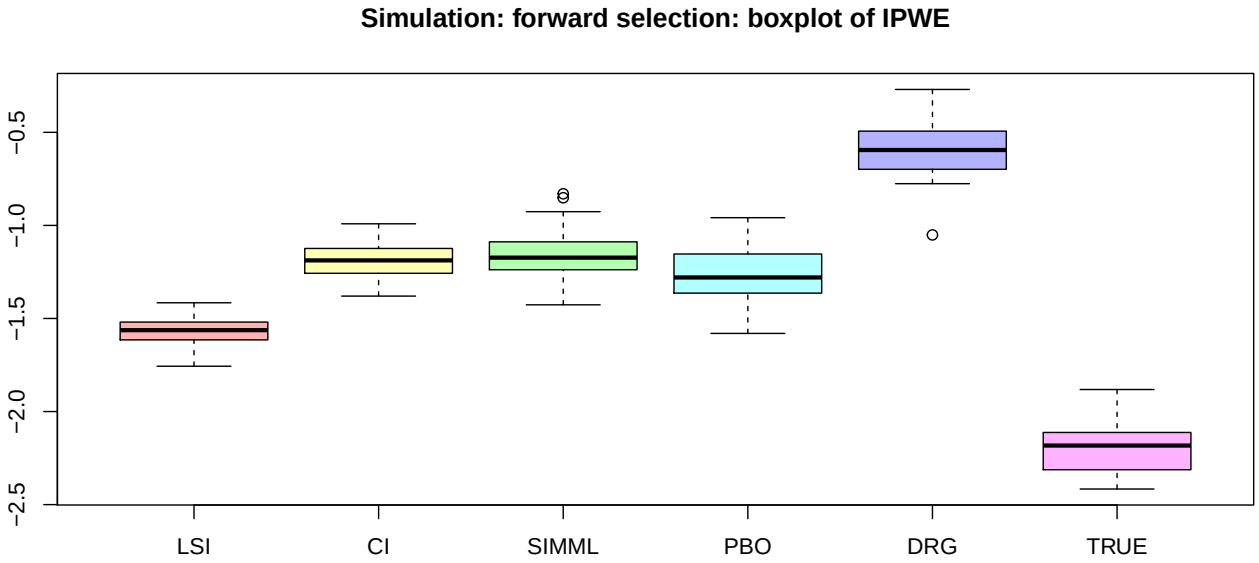


Forward selection: by using the purity

Table 12: PCD

longitudinal	change.score	simml
0.660	0.633	0.618
0.284	0.290	0.363

IPWE:



Forward selection: by using the IPWE

Table 13: PCD

longitudinal	change.score	simml
0.670	0.600	0.558
0.299	0.309	0.524

IPWE: