Comparison of purity and likelihood criteria in non constrained GEM models

2020-06-28

In our previous setting, we consider the constrained "GEM" model in the longitudinal setting

$$\mathbf{y_{ki}} = \mathbf{X_i}(\beta_\mathbf{k} + \mathbf{b_{ki}} + \mathbf{\Gamma_k}(\alpha' \mathbf{x_i})) + \epsilon_\mathbf{ki} \tag{1}$$

where

- $\mathbf{y_{ki}} = (y_{ki1}, ..., y_{kim_i})'$ is a vector of outcomes for subject i at treatment k, measured repeatedly over m_i time points.
- $\beta_\mathbf{k} + \mathbf{\Gamma_k}(\alpha' \mathbf{x_i})$ presents the fixed effect in the longitudinal model, while $\mathbf{x_i} = (x_{i1}, ...x_{ip})'$ are p predictor variables.
- $\mathbf{b_{ki}}$ is the random effect for subject i at the treatment k
- $\epsilon_\mathbf{ki}$ is the random error for subject i at the treatment k

To choose α, we consider two criteria

- maximize the purity
- maximize the log-likelihood function

Generally, in practice we do not expect the GEM model t obe the true dta generating model. Without the constrain, the model can be generalized as

$$\mathbf{y_{ki}} = \mathbf{X_i}(\beta_\mathbf{k} + \mathbf{b_{ki}} + \sum_{\mathbf{j=1}}^{\mathbf{p}} \mathbf{\Gamma_{kj}} \mathbf{x_{ij}}) + \epsilon_\mathbf{ki} \tag{2}$$

that is, instead of sharing a same coefficient $\mathbf{\Gamma_k}\alpha'$, each predictor x_{ij} has its own coefficient Γ_{kj}. The generazlied model is not a single index model.

In this study, we wanted to check the performance of those two crietria when the single index model structure is misspecified.

Simulation 1

The data sets are generated following the below parameter setting:

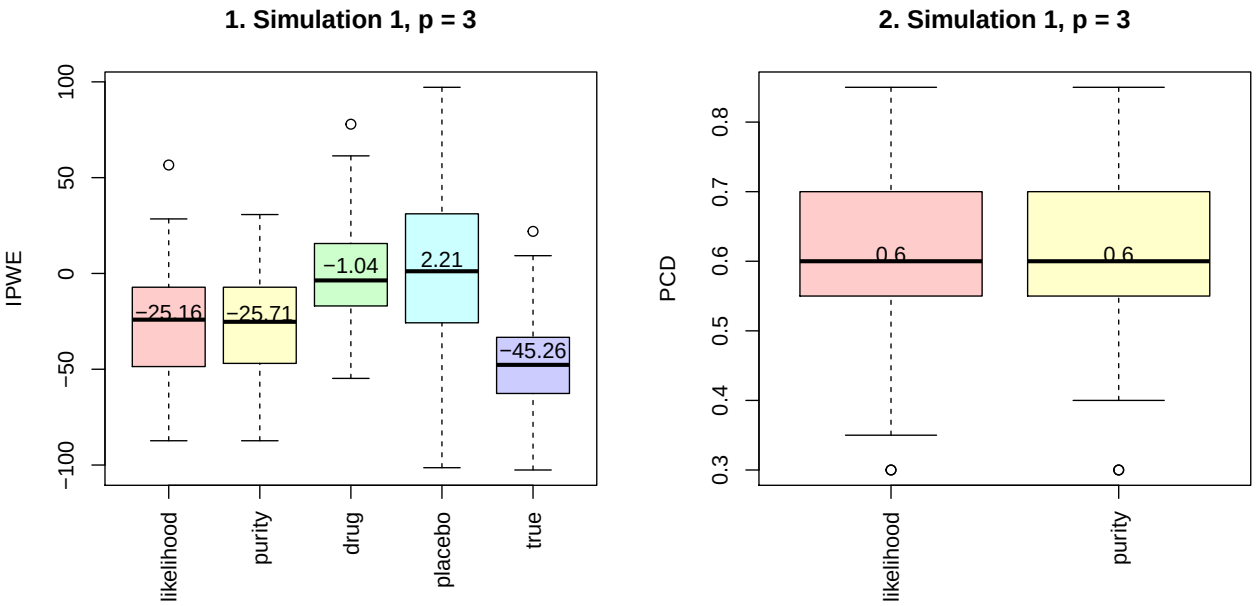
$$\mathbf{y_{ki}} = \mathbf{X_i}(\beta_\mathbf{k} + \mathbf{b_{ki}} + \sum_{j=1}^{p} \Gamma_{kj} x_{ij}) + \epsilon_\mathbf{ki}$$

- Suppose we have $k = \{1, 2\}$ treatments. $k = 1$ represents the placebo group while $k = 2$ represents the drug group.
- $\mathbf{X_i} = [1, t, t^2]$, and $t = [0, 1, 2, 3, 4, 6, 8]$, which is the design matrix for fixed effect and random effect
- $\beta_\mathbf{1} = \beta_\mathbf{2} = [1, -0.05, -0.02]$

- $\mathbf{b_{1i}} \sim N(0, \mathbf{D_1}), \mathbf{b_{2i}} \sim N(0, \mathbf{D_2})$, $\mathbf{D_1} = \mathbf{D_2} = \begin{pmatrix} 1 & 0.3 & 0.1 \\ 0.3 & 1 & 0.5 \\ 0.1 & 0.5 & 1 \end{pmatrix}$
- $\mathbf{\Gamma_{1j}} \sim MVN((0.3, -0.4, -0.3)', \mathbf{\Sigma_{\Gamma_1}})$, $j \in \{1, 2, .., p\}$, $\mathbf{\Sigma_{\Gamma_1}} = \begin{pmatrix} 0.5 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.5 \end{pmatrix}$
- $\mathbf{\Gamma_{2j}} \sim MVN((-0.1, 0, -0.3)', \mathbf{\Sigma_{\Gamma_2}})$, $j \in \{1, 2, .., p\}$, $\mathbf{\Sigma_{\Gamma_1}} = \begin{pmatrix} 0.5 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.5 \end{pmatrix}$
- $\mathbf{x_i} \sim MVN(\mu_x, \Sigma_x)$, $\mu_x = \mathbf{0}_p$, Σ_x has diagonal equals to 1 and 0.5 everywhere else.
- $\epsilon_\mathbf{1}, \epsilon_\mathbf{2} \sim N(0, 1^2)$
- $p = \{3, 10\}$, which is the dimension of the predictors x_i.

The IPWE are calculated with the estimated treatment assignment by using the purity criterion and by using the log-likelihood criterion. The proportions of correct assignment are also calculated. The whole procedures are repeated for 100 times.

Results



When there are $p = 3$ predictors, the IPWE and PCD are shown in figure 1 and figure 2.

The IPWEs estimated by log-likelihood criterion, purity criterion, only drug group, only placebo group, and true group assignment are shown in the figure 1.

Although the performance of purity criterion and log-likelihood criterion are quite similar, purity criterion has a slightly better IPWE than log-likelihood criterion. Their proportions of correct assignment, shown in figure 2 are also very similar.

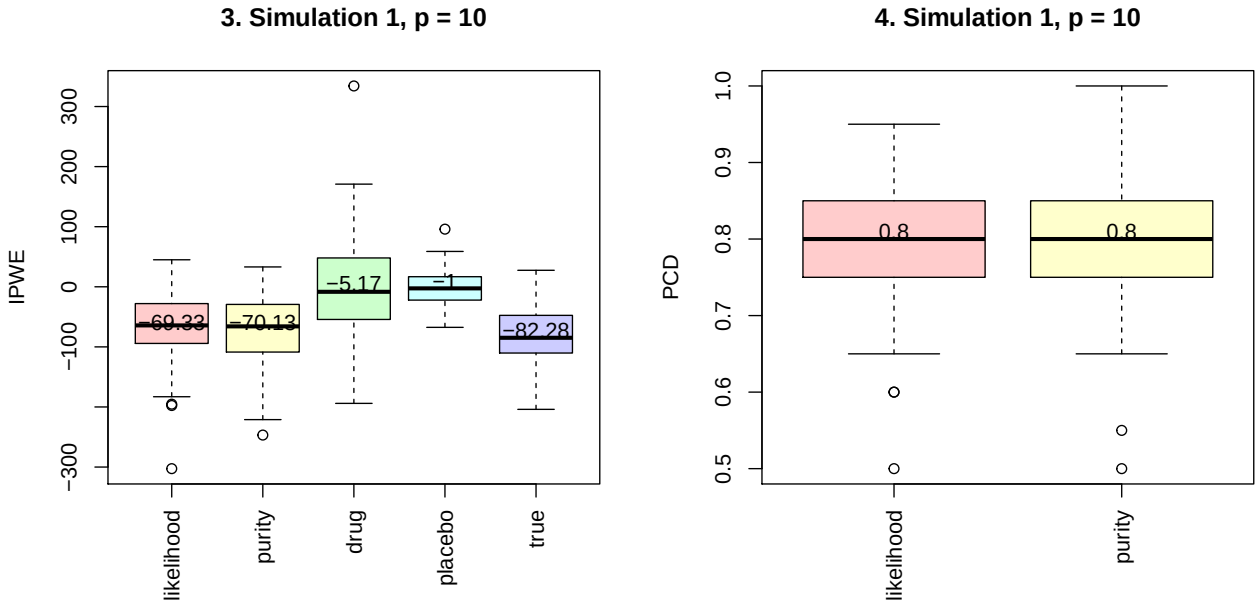


Figure 3 ane Figure 4 are the results under the same setting as Figure 1 and Figure 2 while have a higher dimension of predictors, $p = 10$. The results are very close to each other.

Simulation 2

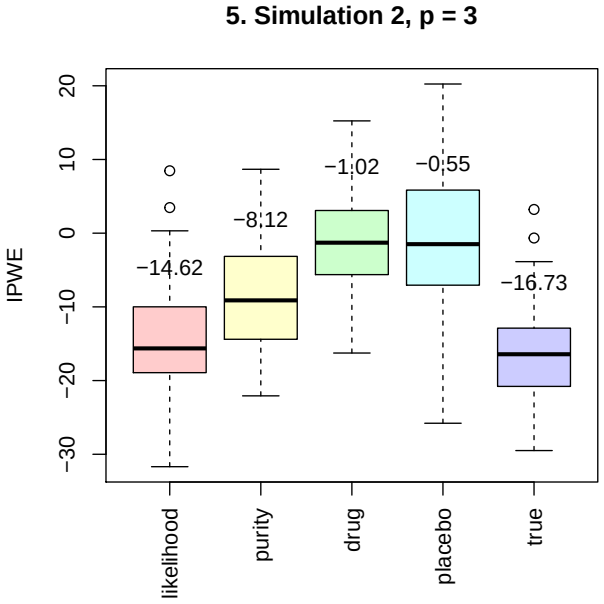
Since in simulation 1, we did not observe big difference between the two crietria, I then changed the simulation setting. (the parameters are more close to the EMBARC dataset)

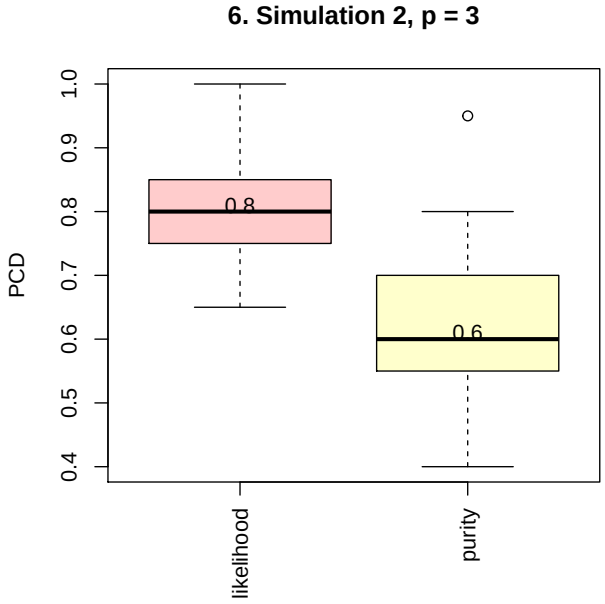
The data sets are generated following the below parameter setting:

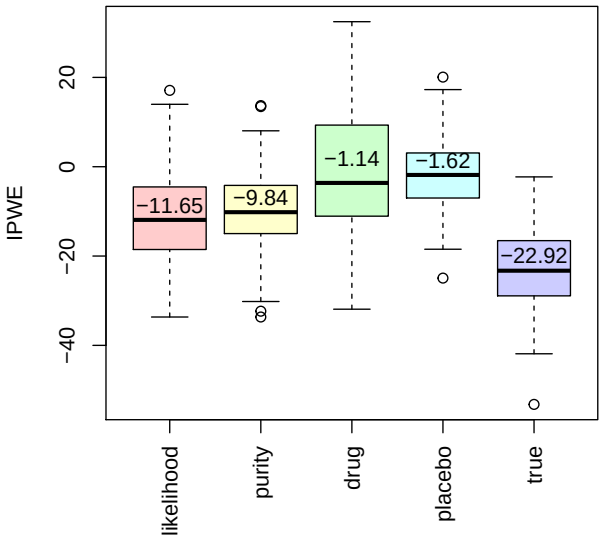
$$\mathbf{y_{ki}} = \mathbf{X_i}(\beta_\mathbf{k} + \mathbf{b_{ki}} + \sum_{j=1}^{p} \Gamma_{kj} x_{ij}) + \epsilon_\mathbf{ki}$$

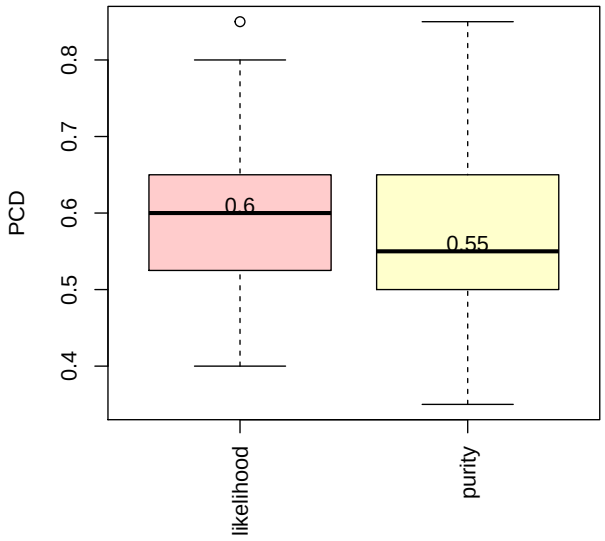
- Suppose we have $k = \{1, 2\}$ treatments. $k = 1$ represents the placebo group while $k = 2$ represents the drug group.
- $X_i = [1, t, t^2]$, and $t = [0, 1, 2, 3, 4, 6, 8]$, which is the design matrix for fixed effect and random effect
- $\beta_1 = \beta_2 = [0, -0.1, -0.01]$
- $b_1 \sim N(0, D_1), b_2 \sim N(0, D_2)$, $D_1 = D_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.01 & 0.005 \\ 0 & 0.005 & 0.02 \end{pmatrix}$
- $\Gamma_{1j} \sim MVN((0, -0.04, -0.01)', \Sigma_{\Gamma_1})$, $j \in \{1, 2, .., p\}$, $\Sigma_{\Gamma_1} = \begin{pmatrix} 0.1 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0.1 \end{pmatrix}$
- $\Gamma_{2j} \sim MVN((0, -0.01, -0.04)', \Sigma_{\Gamma_2})$, $j \in \{1, 2, .., p\}$, $\Sigma_{\Gamma_1} = \begin{pmatrix} 0.05 & 0 & 0 \\ 0 & 0.05 & 0 \\ 0 & 0 & 0.05 \end{pmatrix}$
- $x_i \sim MVN(\mu_x, \Sigma_x)$, $\mu_x = \mathbf{0}_p$, Σ_x has diagonal equals to 1 and 0.5 everywhere else.
- $\epsilon_1, \epsilon_2 \sim N(0, 1^2)$
- $p = \{3, 10\}$, which is the dimension of the predictors x_i.

Results









However, in this setting, the performance of log-likelihood is much better than the performance of purity criterion.

I think this may be caused that the estimation of purity is not stable while the random effect of concavity term is relatively small.

The quantiles of the estimated purities:

```
##           0%          25%          50%          75%         100%
##  -134.2603  2535535.8569  5636568.1028 10001369.9353 40051000.8994
```

We can see that there are many quite large purities, caused by a small estimation of $\hat{D}_1$ and $\hat{D}_2$ and as a consequence, large $(\hat{D}_1)^{-1}, (\hat{D}_2)^{-1}$

The quantiles of the estimated loglikelihood:

```
##        0%       25%       50%       75%      100%
## -2902.705 -2861.054 -2847.620 -2824.181 -2785.795
```

It is relatively stable, since we calculated the inverse of $D_k + \sigma_k^2 I$ instead of the inverse of D_k

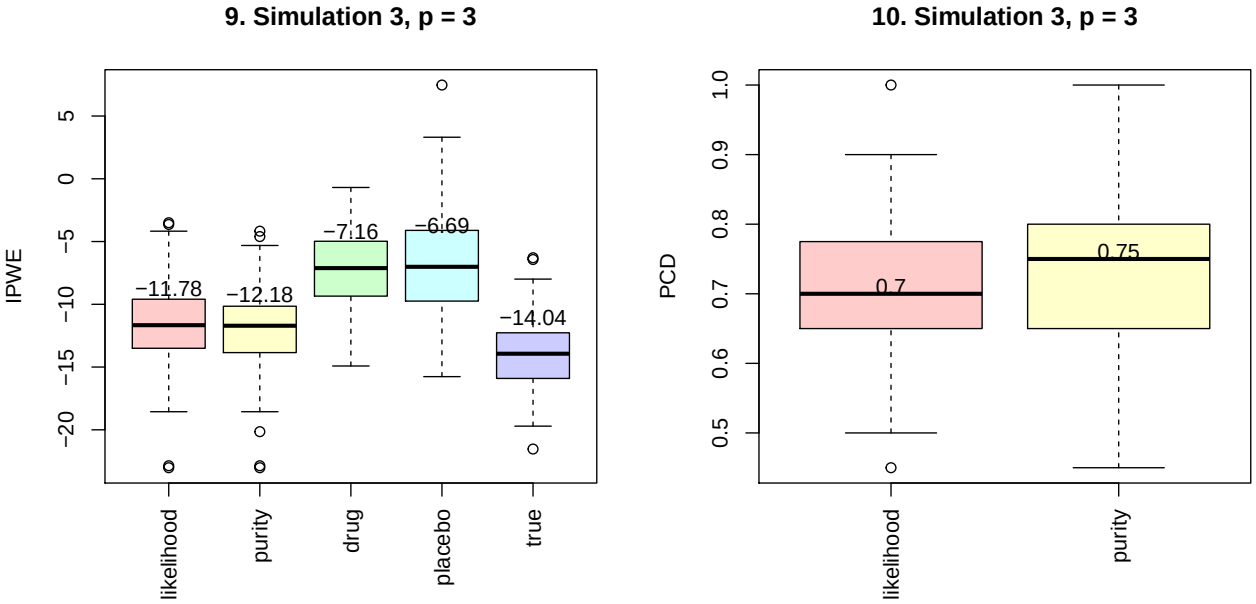
Simulation 3

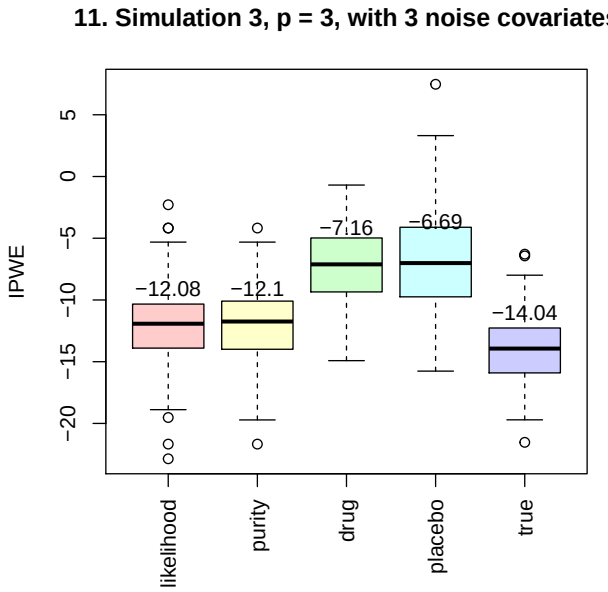
Since the random effect for concavity term is quite small and affects the estimation of purity, in the simulation 3, I used the same data generation setting as simulation 2, while the purity is calculated by outcomes' distributions instead of coefficients' distributions.

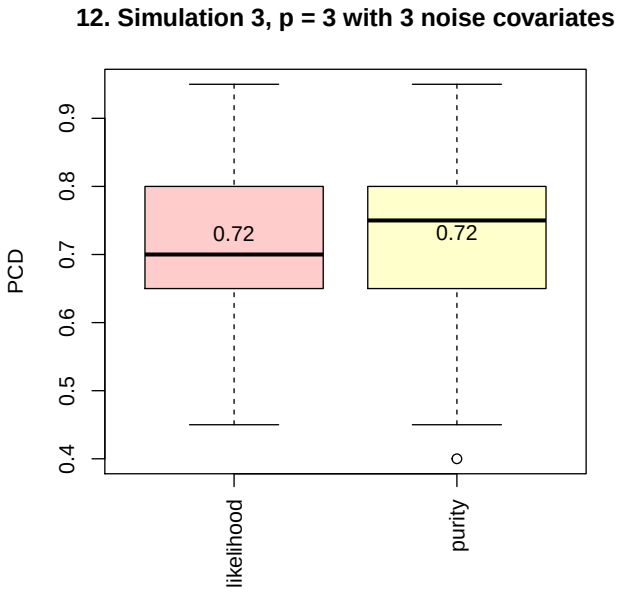
For the dimension of predictors, I set

- scenario 1. $p = 3$.
- scenario 2. $p = p_1 + p_2$, $p_1 = 3$, which is the dimension of covariates that are used to generate the outcome; $p_2 = 3$, which is the dimension of noises. The noises are generated from standard normal distributions.

Results







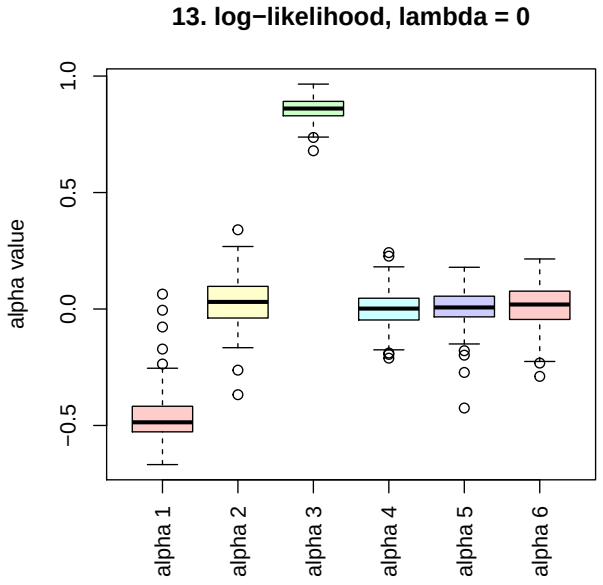
By using the random effects of intercept and slope for purity calculation, the results of purity method get better. The estimation gets more stable. And the results of purity method is slightly better than the log-likelihood method.

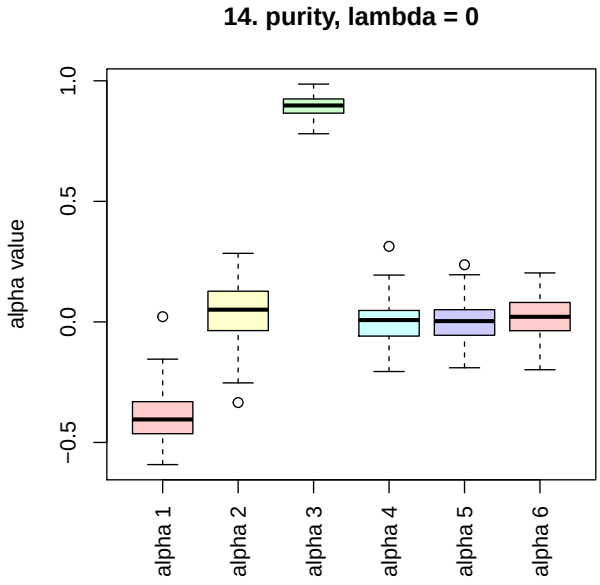
Penalization

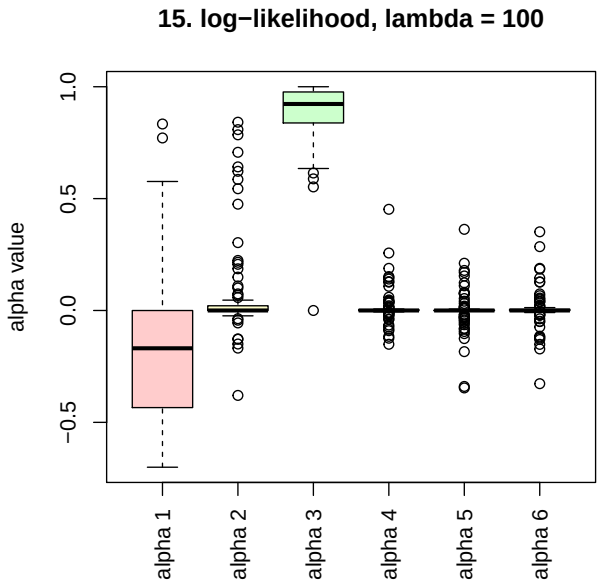
If we consider the L1 penalty for variable selection, the estimation of $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6)'$ vector ($p = 3 + 3 = 6$ dimensions) are shown in the following plots.

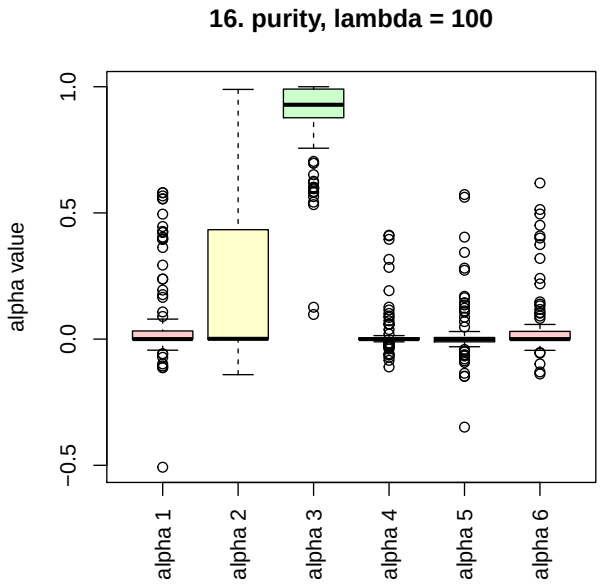
If the $\lambda = 0$, which means that we do not have penalty in the model, the estimated $\hat{\alpha}$ are just the $\hat{\alpha}$ for the calculation of IPWE in Figure 11. The esitmations of α by using the log-likelihood crietrion and purity crietrion are quite similar.

When λ is large, $\lambda = 100$, the estimation of the 6 elements are shown in the Figure 15 and 16. The elements $\alpha_4, \alpha_5, \alpha_6$ are estimated to be close to 0.







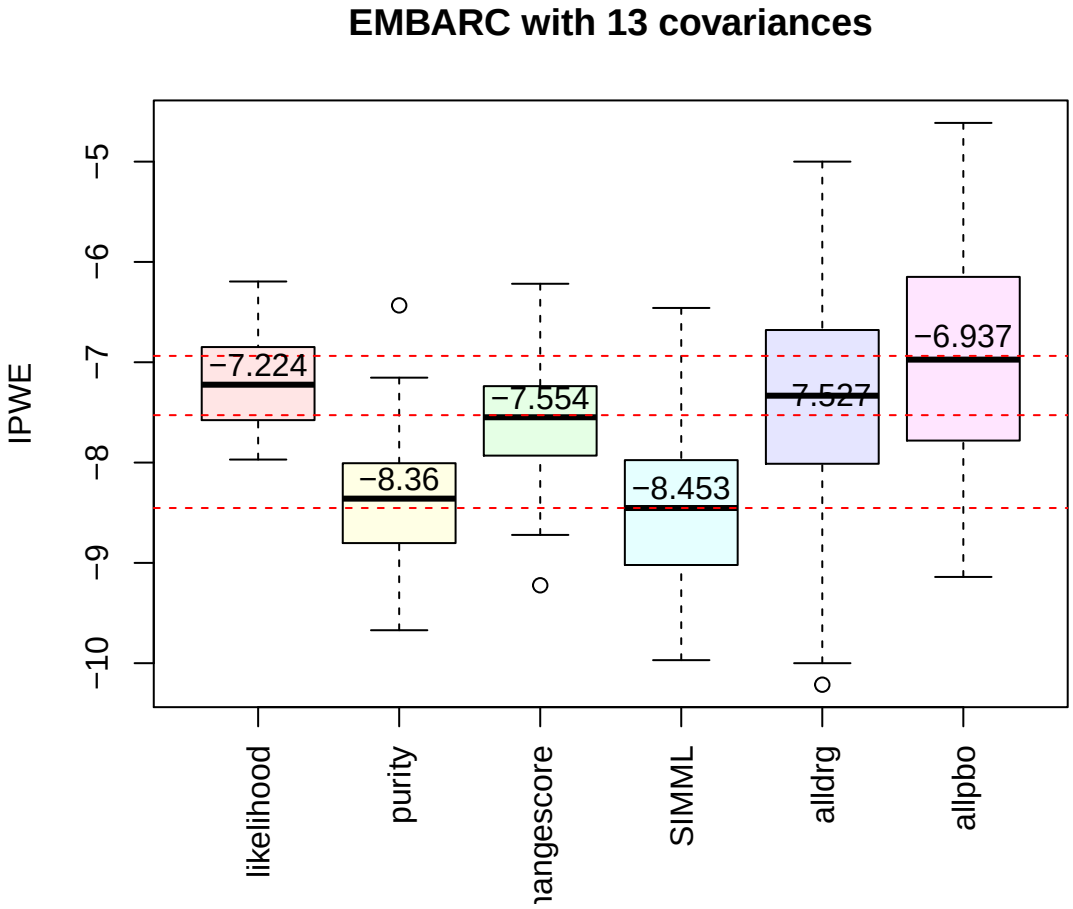


EMBARC

In previous results, by conducting forward and backward selection, we selected below 13 covariates. Fit the longtiudual single index model with those covariates can have a relatively good estimation of IPWE.

Now we estimation the α of those 13 covariates with log-likelihood criterion. The IPWE plot is shown below.

w0_1329	w0_1431	w0_1203	w0_1273	w0_1409	w0_1401
w0_1149	w0_1235	w0_1307	w0_1011	w0_1329	w0_1431
w0_1401					



From left to right, the boxplots show the IPWE estimated by log-likelihood criterion, purity criterion, linear change score method, SIMML, and all drug and all placebo.

The purity crietrion and SIMML have similar results and they have good estimations of IPWE. However, the loglikelihood method has a bad performance, the estimation is worse than the linear change score method.