

## LDA training on random subsets of different sizes

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
##
```

```
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
```

```
##
```

```
##      filter, lag
```

```
## The following objects are masked from 'package:base':
```

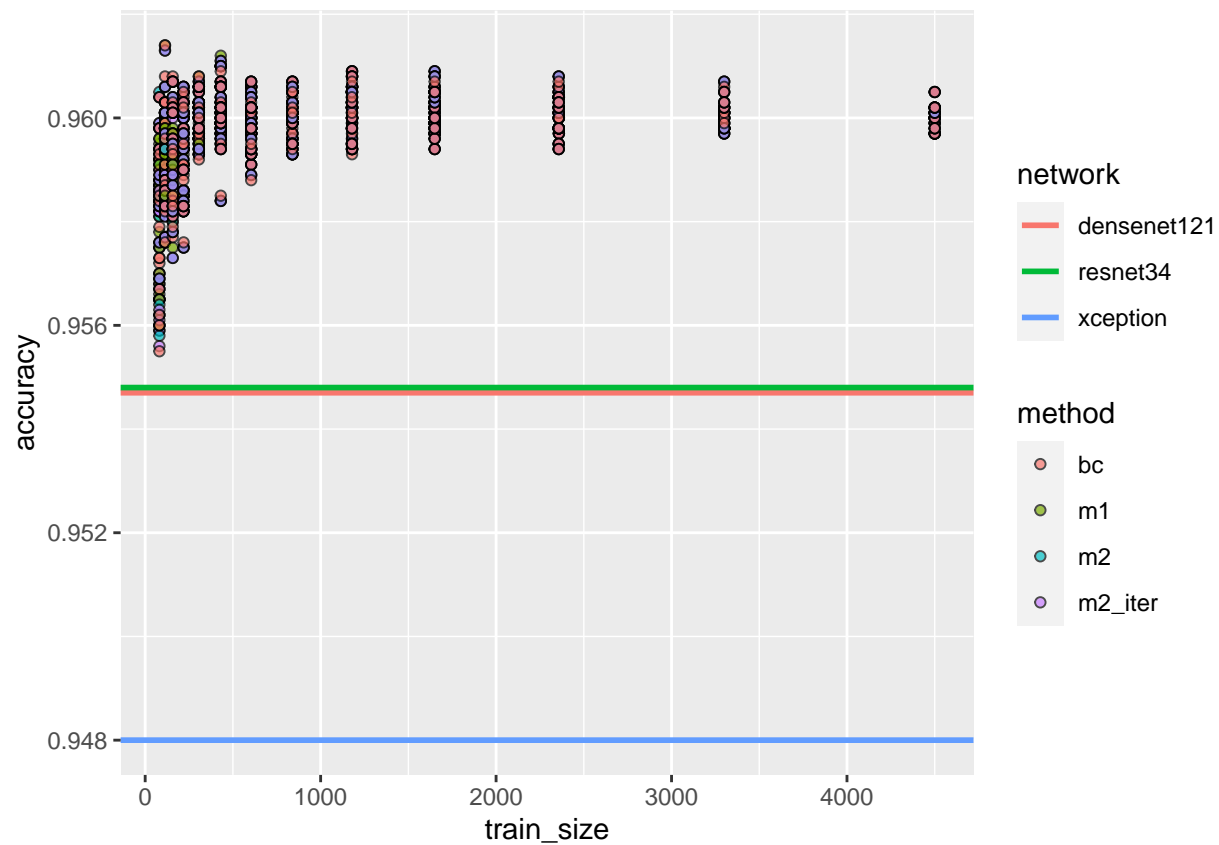
```
##
```

```
##      intersect, setdiff, setequal, union
```

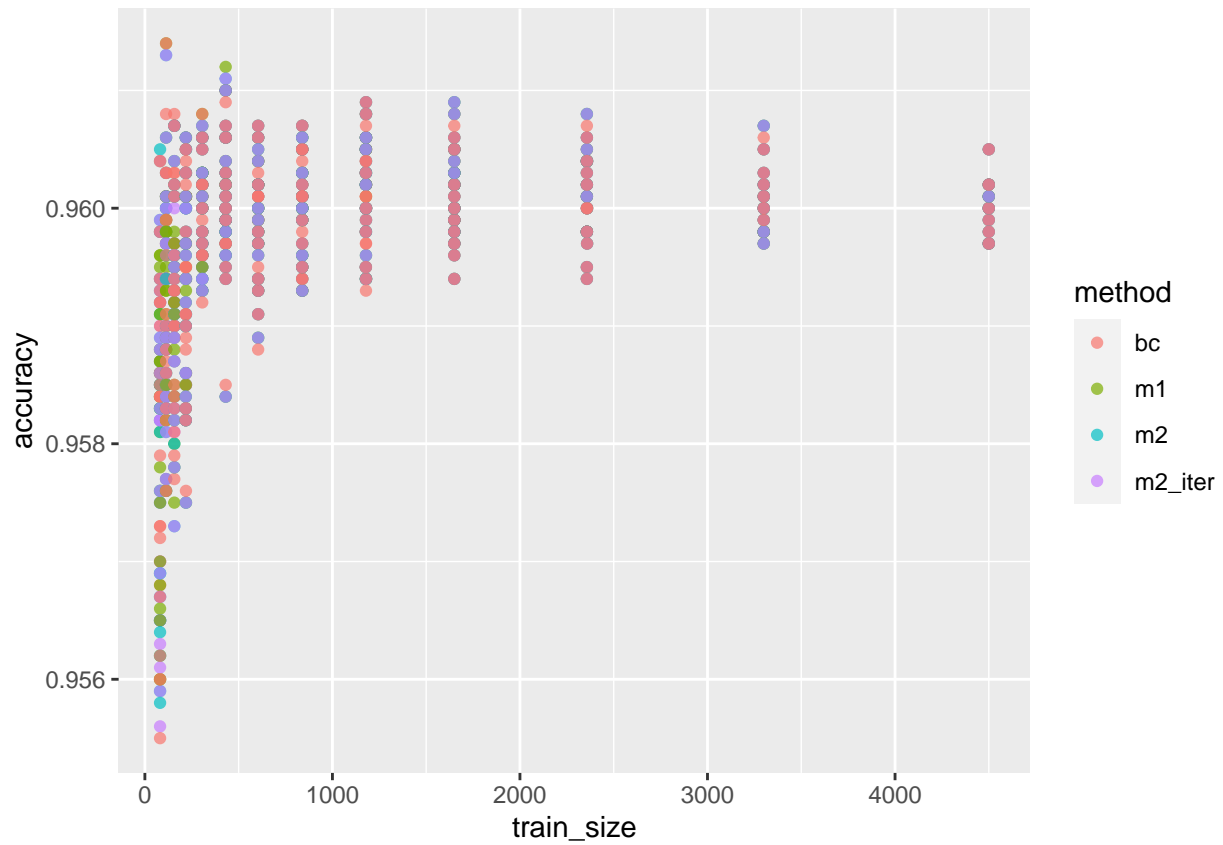
Experiment code is in the file `training_subsets_sizes_experiment.py`. Experiment on CIFAR10 dataset. This experiment trains WeightedLDAEnsemble on various subsets of different sizes of data on which neural networks were trained. Subsets of the same size are disjoint. Three different coupling methods are used: method one and two from (Wu, Lin, and Weng 2004) and Bayes covariant method (Šuch and Barreda 2016). Goal of this experiment is to determine, for which size of the LDA training set, the ensemble achieves the best performance.

```
acc_ens_subsets <- read.csv("../data/data_train_val_c10/1/exp_subsets_sizes_train_outputs/accuracies.csv")
acc_nets <- read.csv("../data/data_train_val_c10/1/exp_subsets_sizes_train_outputs/net_accuracies.csv",
```

```
scatter <- ggplot() + geom_point(data=acc_ens_subsets, mapping=aes(x=train_size, y=accuracy, fill=method))
scatter
```

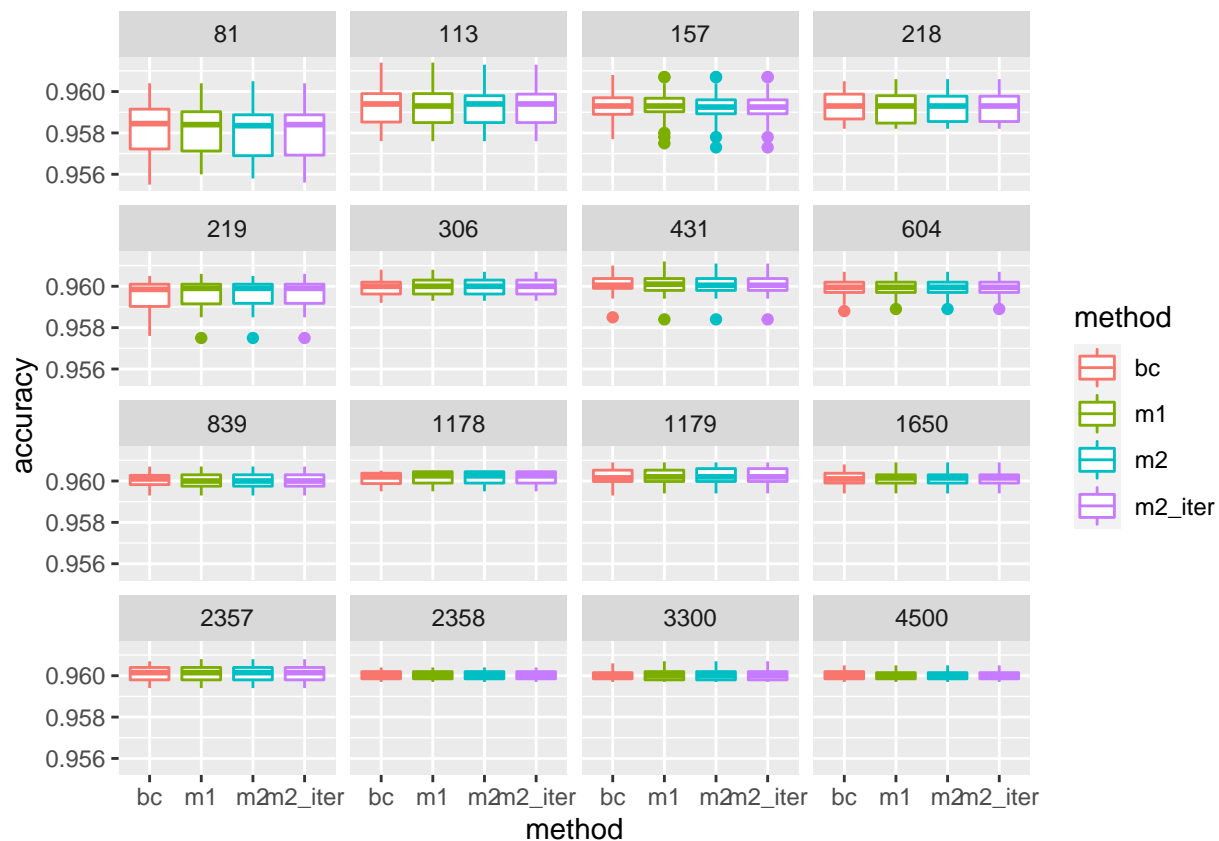


```
scatter2 <- ggplot() + geom_point(data=acc_ens_subsets, mapping=aes(x=train_size, y=accuracy, color=method))
scatter2
```



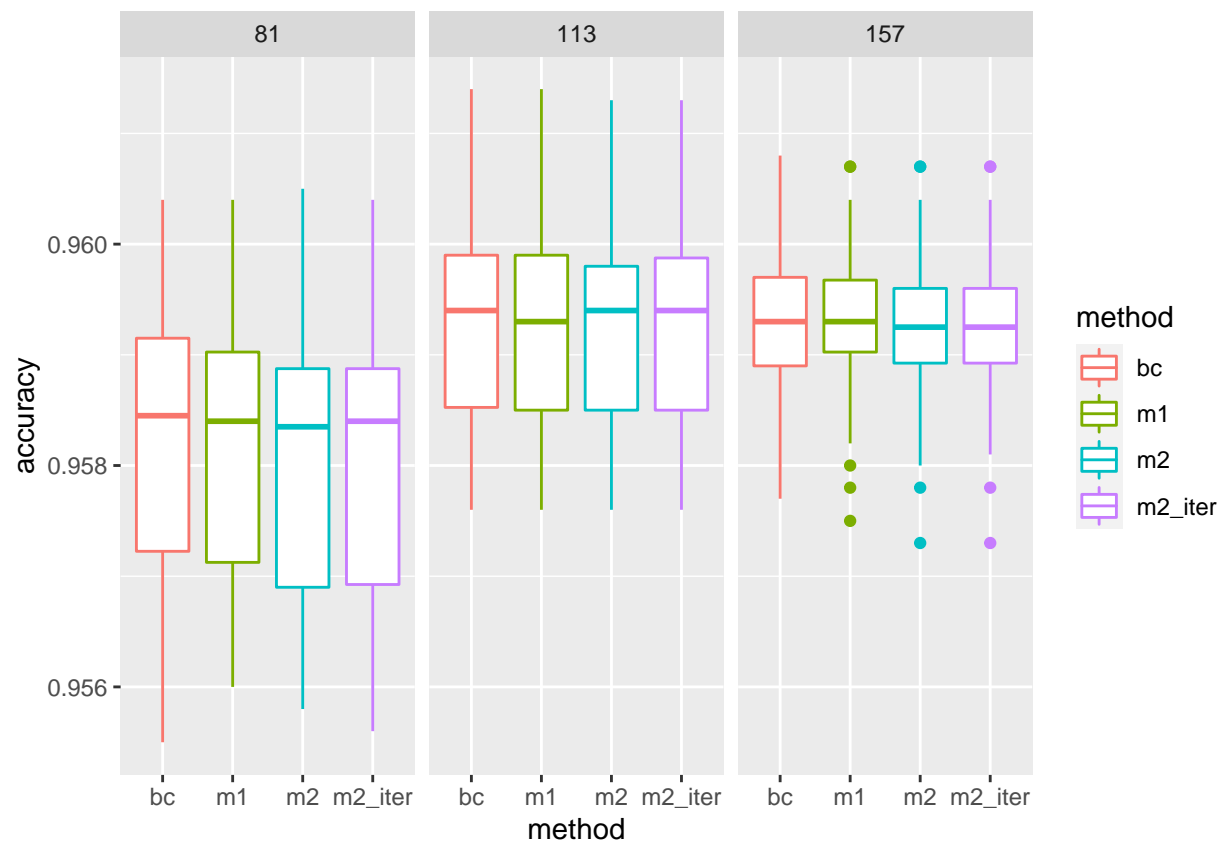
Scatter plots are quite messy, so we will use boxplots.

```
box_acc <- ggplot() + geom_boxplot(data=acc_ens_subsets, mapping=aes(x=method, y = accuracy, color=method))
box_acc
```

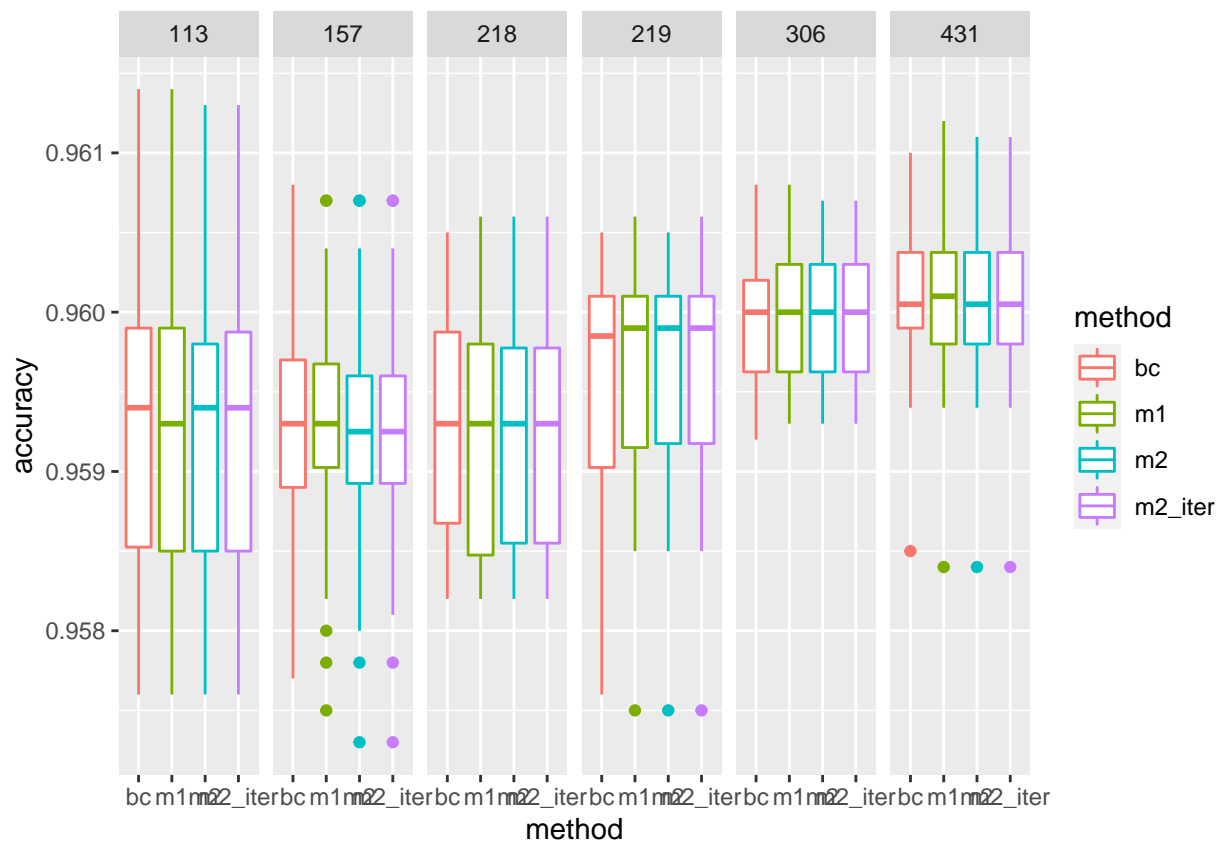


Accuracy seems to be slowly increasing with increasing training set size. We will inspect this closer by plotting for smaller subset sizes range.

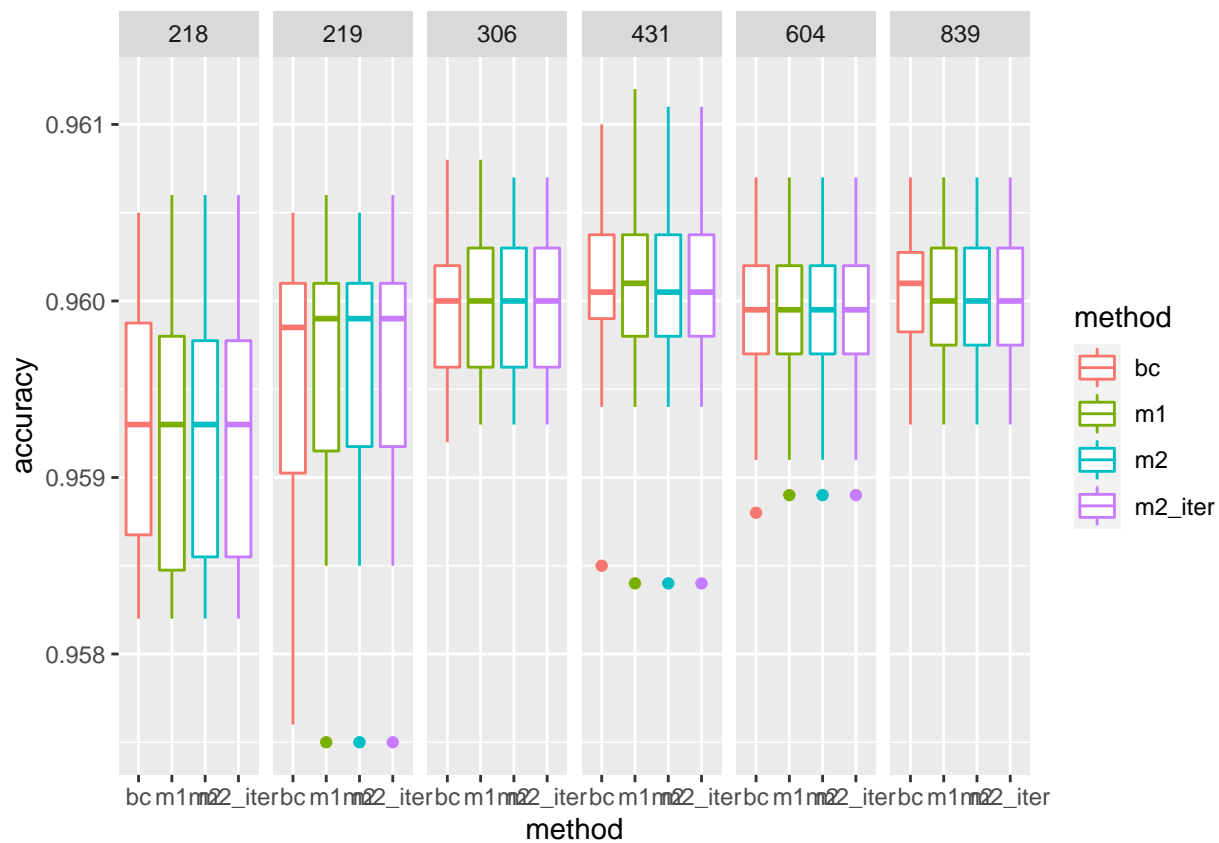
```
acc_ens_subsets %>% filter(0 < train_size & train_size < 200) %>% ggplot() + geom_boxplot(mapping=aes(x=
```



```
acc_ens_subsets %>% filter(100 < train_size & train_size < 500) %>% ggplot() + geom_boxplot(mapping=aes
```

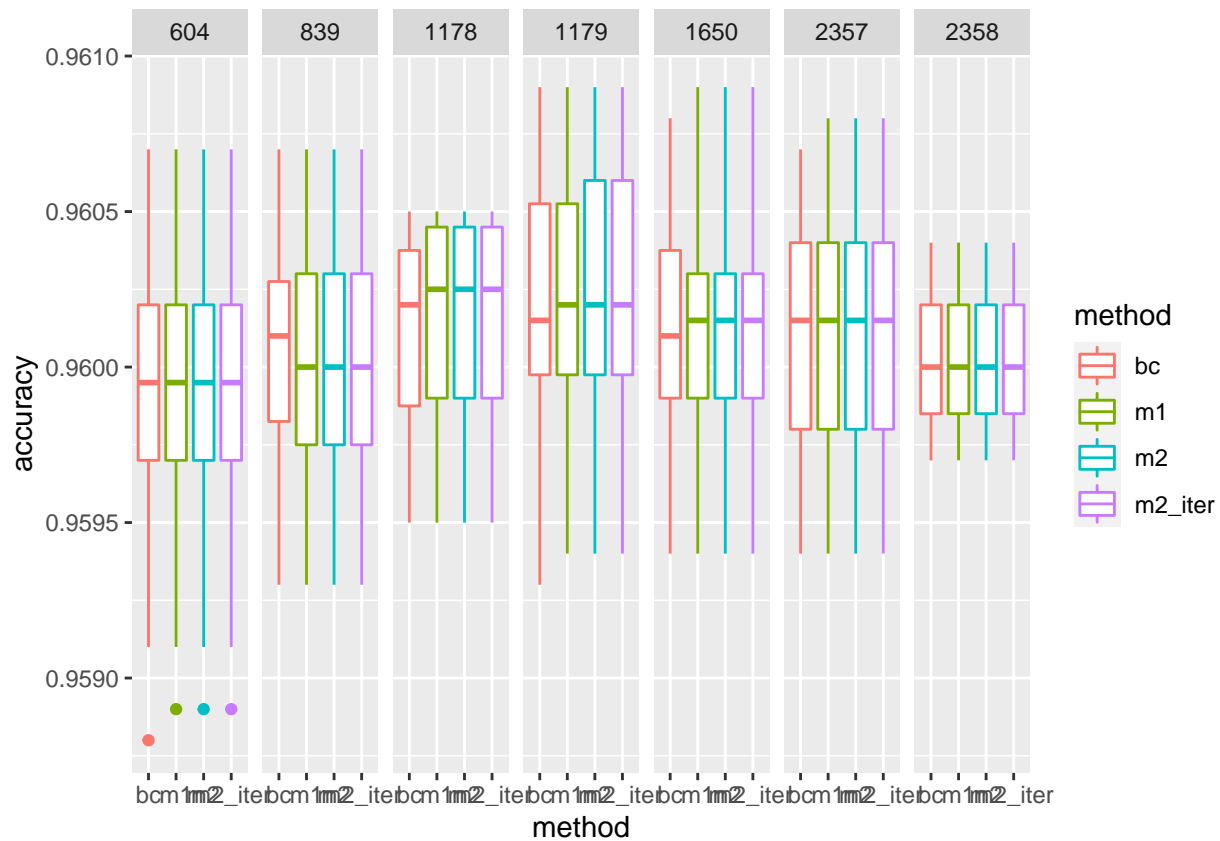


```
acc_ens_subsets %>% filter(200 < train_size & train_size < 1000) %>% ggplot() + geom_boxplot(mapping=aes(method, accuracy))
```



Here the accuracy increase seems to be stopping at train set size 430.

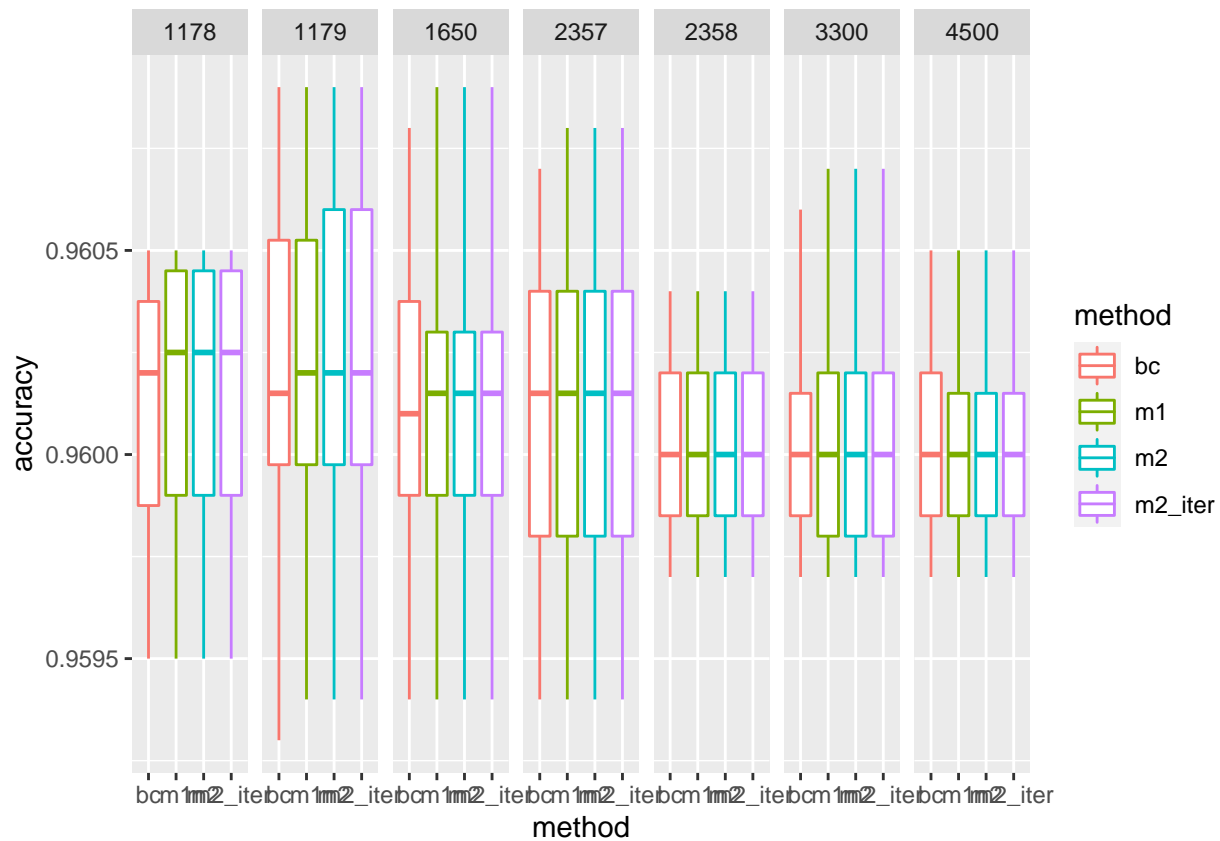
```
acc_ens_subsets %>% filter(500 < train_size & train_size < 3000) %>% ggplot() + geom_boxplot(mapping=aes(method, accuracy))
```



Here is visible some further increase, but it is not very stable and decreases again.

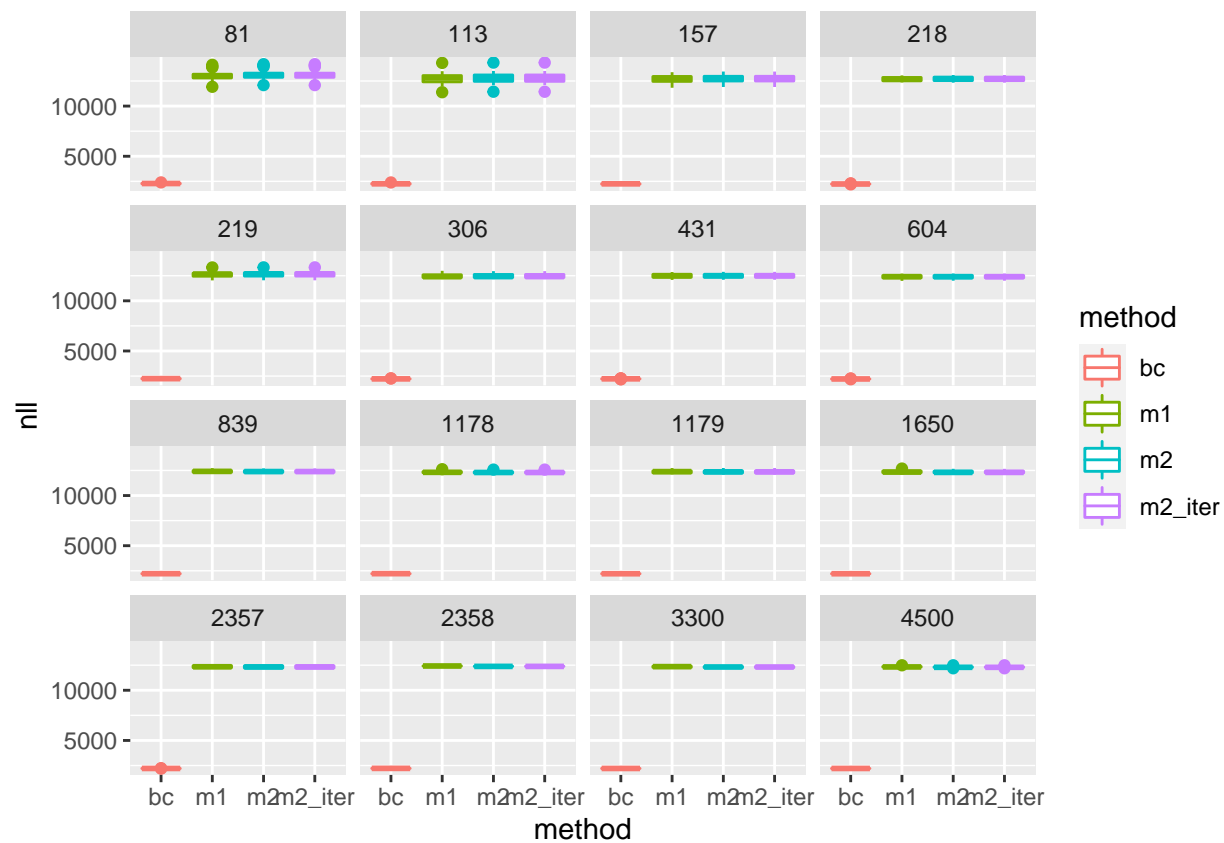
```
acc_ens_subsets %>% filter(1000 < train_size & train_size < 5000) %>% ggplot() + geom_boxplot(mapping=a
```





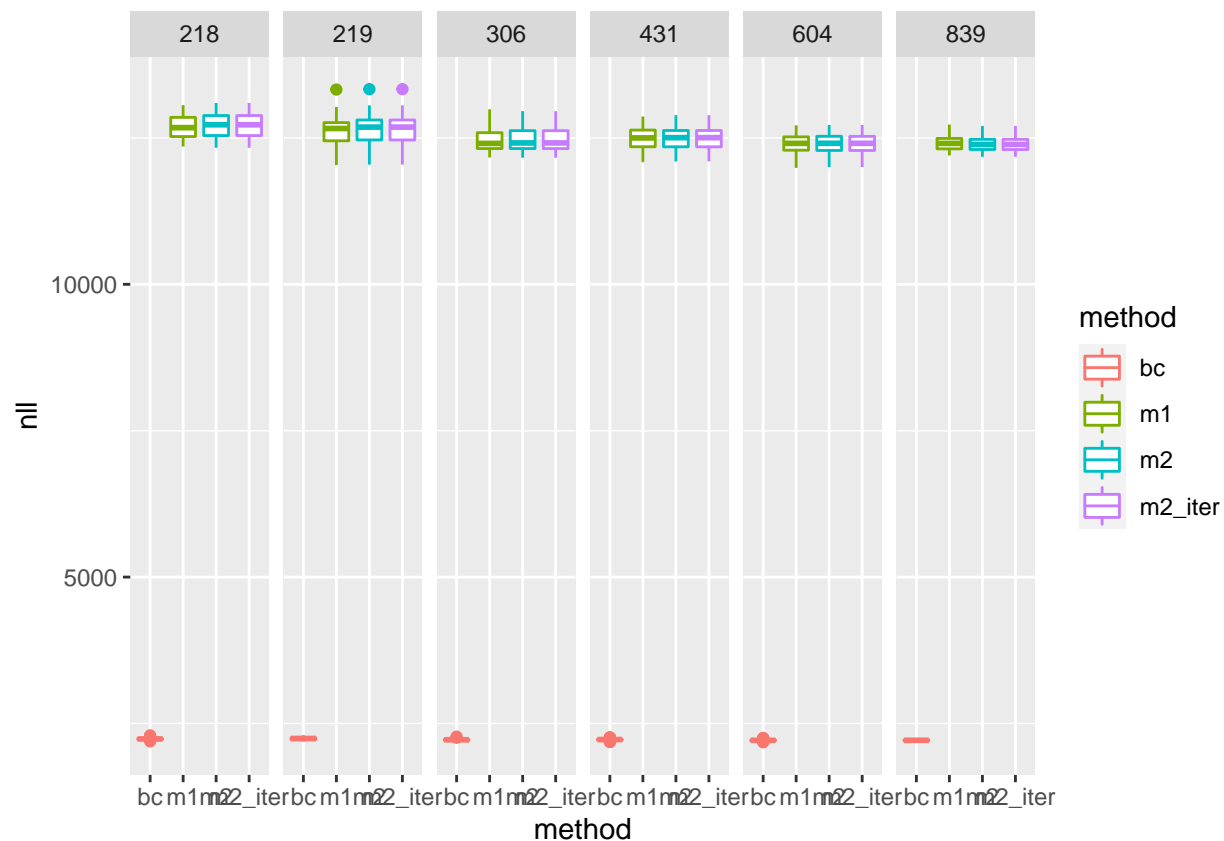
It seems, that from train set size around 1000 accuracy decreases again.

```
box_nll <- ggplot() + geom_boxplot(data=acc_ens_subsets, mapping=aes(x=method, y = nll, color=method))
box_nll
```



NLI doesn't seem to be doing much with changing LDA train size. Still, this metric needs further attention because of the zero probabilities produced by ensemble.

```
acc_ens_subsets %>% filter(200 < train_size & train_size < 1000) %>% ggplot() + geom_boxplot(mapping=aes(method))
```



Šuch, Ondrej, and Santiago Barreda. 2016. “Bayes Covariant Multi-Class Classification.” *Pattern Recognition Letters* 84: 99–106.

Wu, Ting-Fan, Chih-Jen Lin, and Ruby C Weng. 2004. “Probability Estimates for Multi-Class Classification by Pairwise Coupling.” *Journal of Machine Learning Research* 5 (Aug): 975–1005.