Validation set vs training set LDA training - Approach one

```
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
```

```
library(tidyr)
```

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

```
library("ggpubr")
```

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

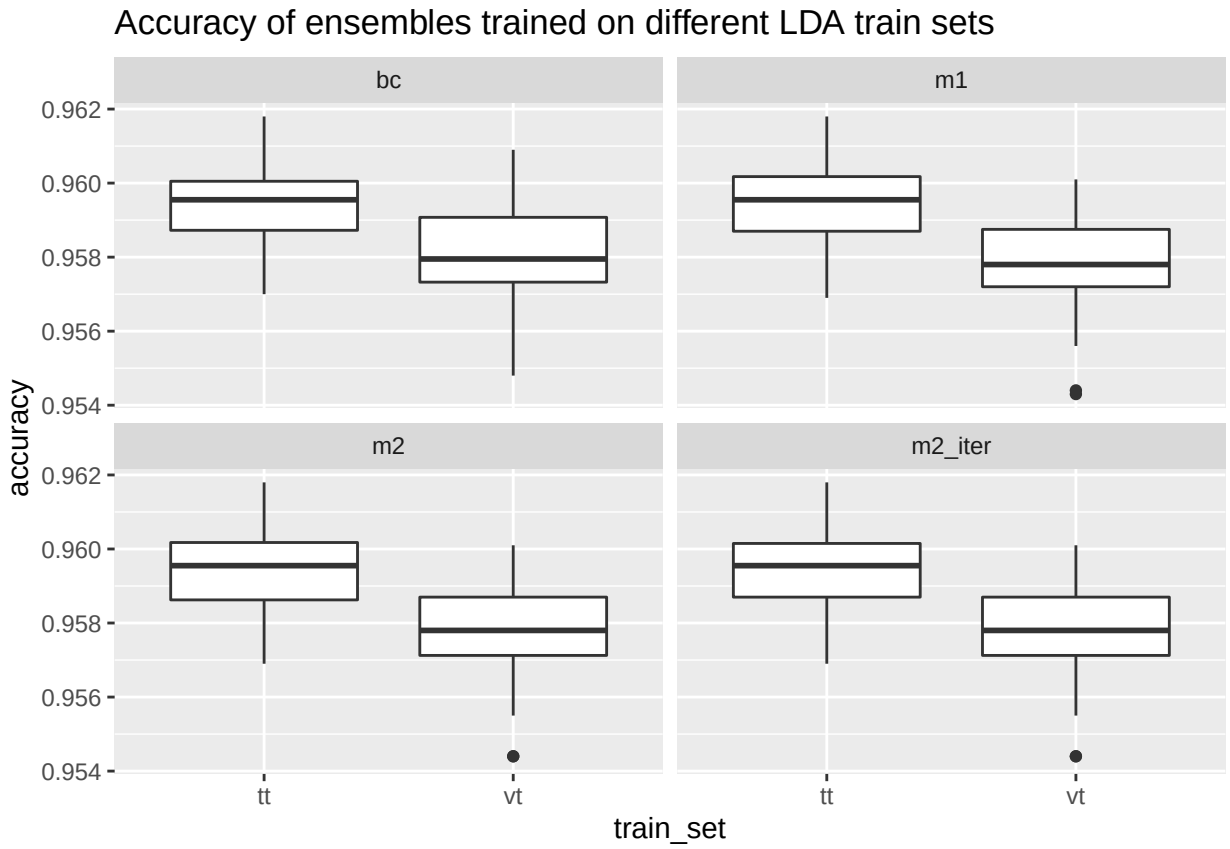
This experiment focuses on the question, whether training LDA on the same set of data as the neural networks were trained on has any adverse effects to the performance of the ensemble as opposed to training on a different set, not presented to the networks during the training. Experimment was performed with two slightly different approaches.

Aproach one Experiment code is in the file base_ensembling_experiment.py. Experiment on CIFAR10 dataset. This experiment was performed in 30 replications. In each replication a set of 500 samples from CIFAR10 training set was randomly chosen, with each class represented equally, this set is reffered to as validation set. This set was extracted from the CIFAR10 training set. Three neural networks were trained on the reduced training set. These networks were then combined using WeightedLDAEnsemble. For each replication, the ensemble was built twice. First LDA was trained on the extracted validation set and second on a randomly chosen set of 500 samples from the neural networks training set. These LDA training sets are referred to as vt and tt respectively.

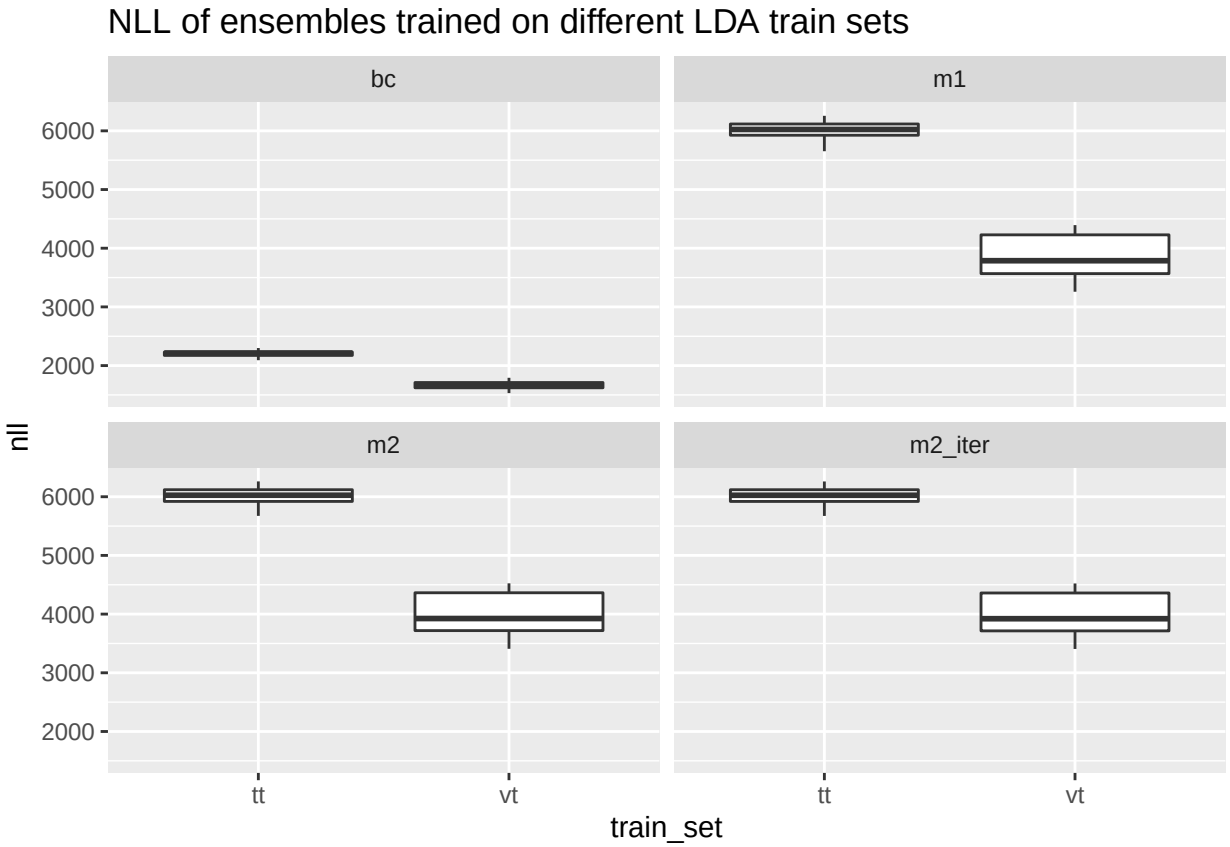
Approach two is described and visualized in files visualization_half_train_base_experiment_CIF10 and visualization_half_train_base_experiment_CIF100.

```
net_results <- read.csv("../data/data_train_val_c10/net_accuracies.csv")
ens_results <- read.csv("../data/data_train_val_c10/ensemble_accuracies.csv")
```

```
box_acc <- ggplot() + geom_boxplot(data=ens_results, mapping = aes(x=train_set, y=accuracy)) + facet_wra
  ggtitle("Accuracy of ensembles trained on different LDA train sets")
box_acc
```

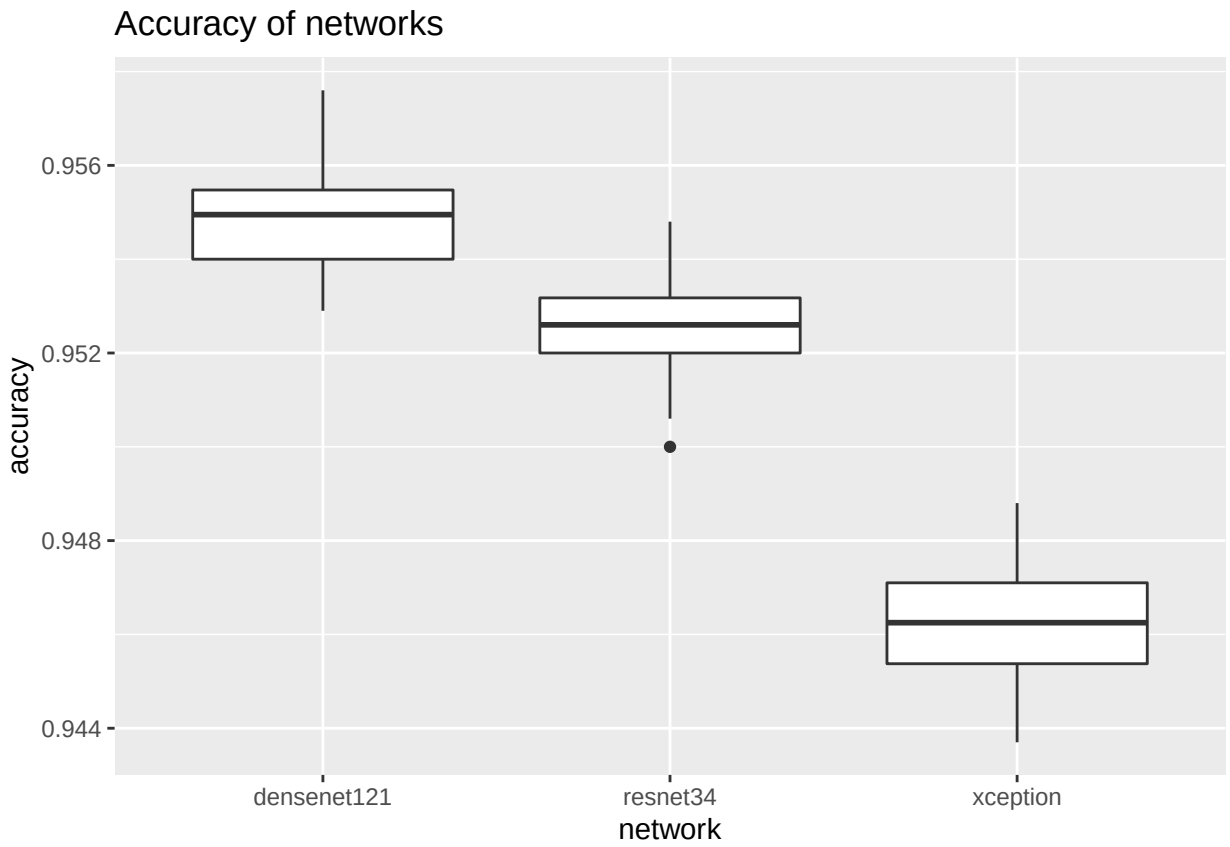



```
box_nll <- ggplot() + geom_boxplot(data=ens_results, mapping = aes(x=train_set, y=nll)) + facet_wrap(~m
  ggtitle("NLL of ensembles trained on different LDA train sets")
box_nll
```

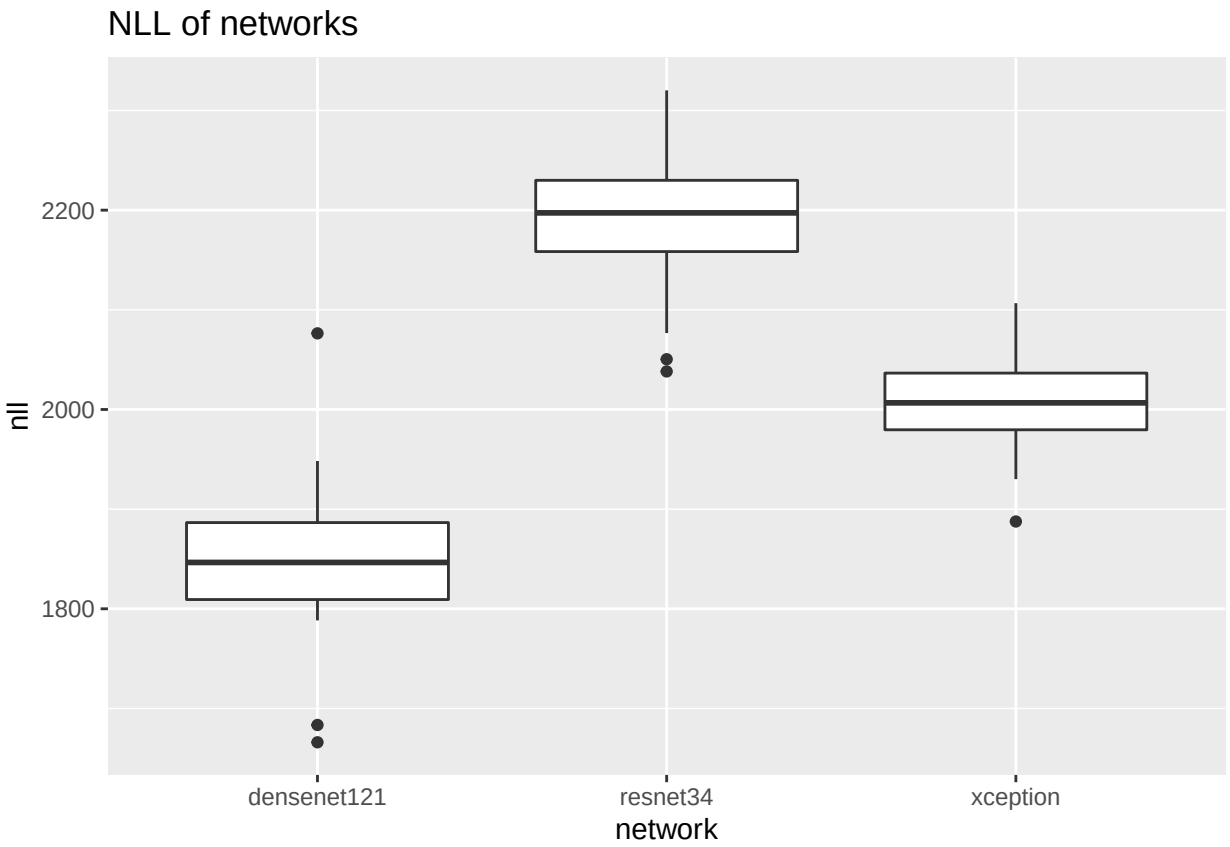


For some reason NLL has opposite trend to accuracy - train set with better accuracy has worse (larger) NLL. This needs some further attention.

```
box_net_acc <- ggplot() + geom_boxplot(data=net_results, mapping=aes(x=network, y=accuracy)) +
  ggtitle("Accuracy of networks")
box_net_acc
```




```
box_net_nll <- ggplot() + geom_boxplot(data=net_results, mapping=aes(x=network, y=nll)) +
  ggtitle("NLL of networks")
box_net_nll
```



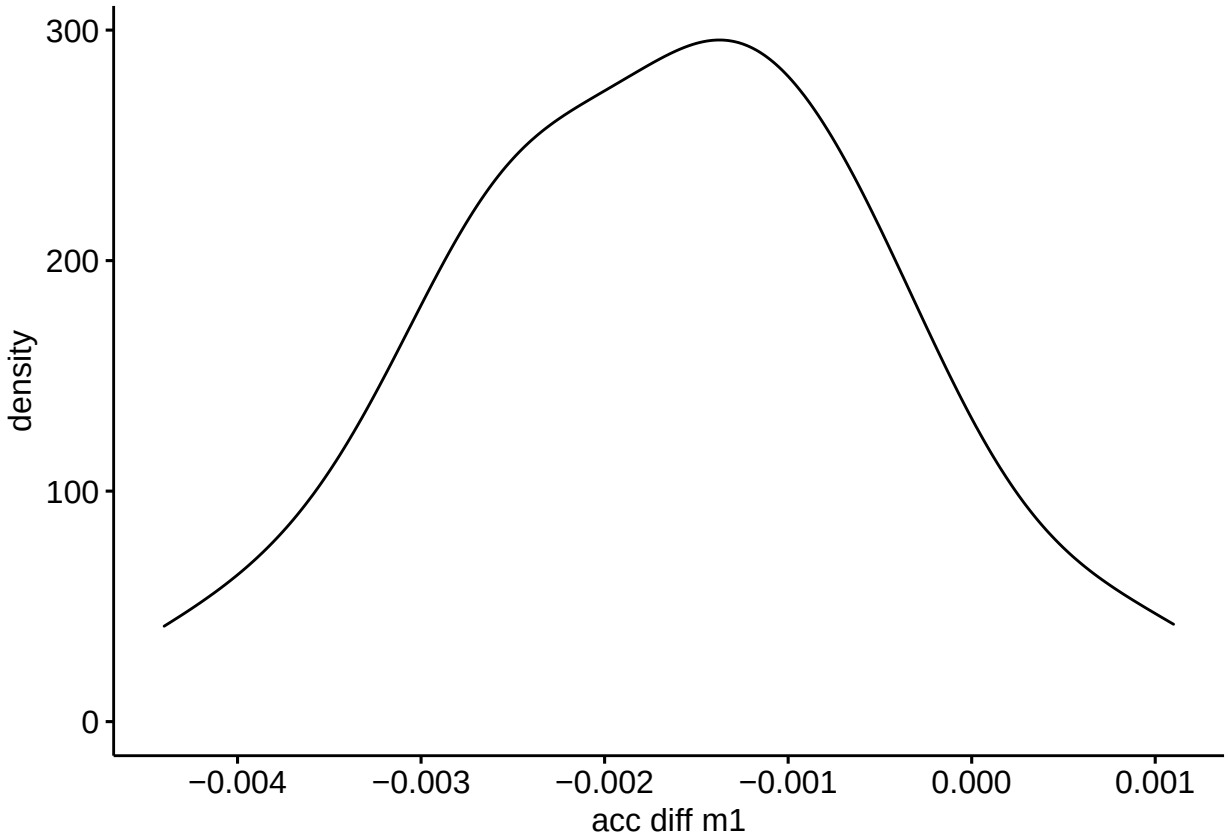

```
ens_results_wide <- pivot_wider(ens_results, names_from = c(train_set, method), values_from = c(accuracy
```

Testing statistical significance of difference for training on validation and train data Since ensembles in each replication are trained on the same set of networks, we use paired t-test.

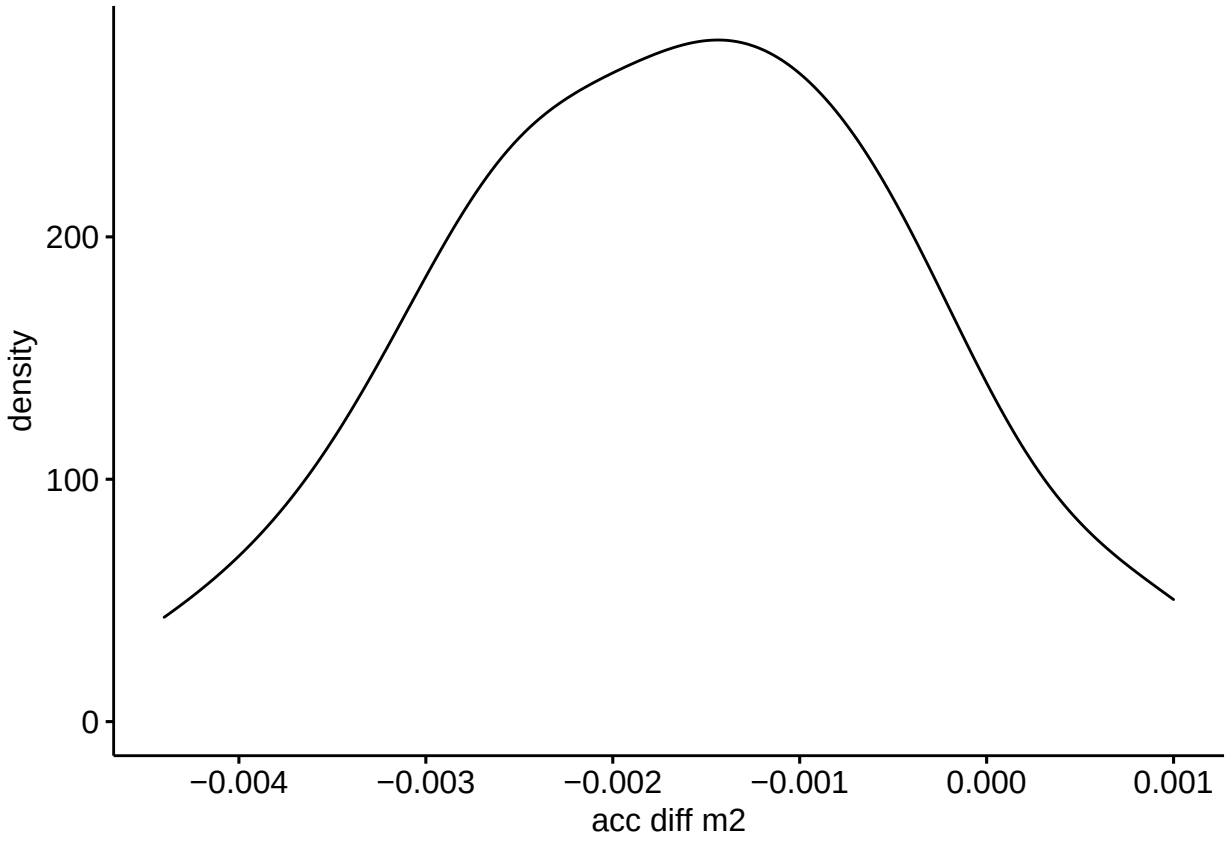
```
differences <- data.frame(ens_results_wide$accuracy_vt_m1 - ens_results_wide$accuracy_tt_m1, ens_results
                          ens_results_wide$accuracy_vt_m2_iter - ens_results_wide$accuracy_tt_m2_iter,
                          ens_results_wide$accuracy_vt_bc - ens_results_wide$accuracy_tt_bc,
                          ens_results_wide$nll_vt_m1 - ens_results_wide$nll_tt_m1, ens_results_wide$nll
                          ens_results_wide$nll_vt_m2_iter - ens_results_wide$nll_tt_m2_iter, ens_results
names(differences) <- c("acc_diff_m1", "acc_diff_m2", "acc_diff_m2_iter", "acc_diff_bc", "nll_diff_m1",
```

Ascertaining normality of differences

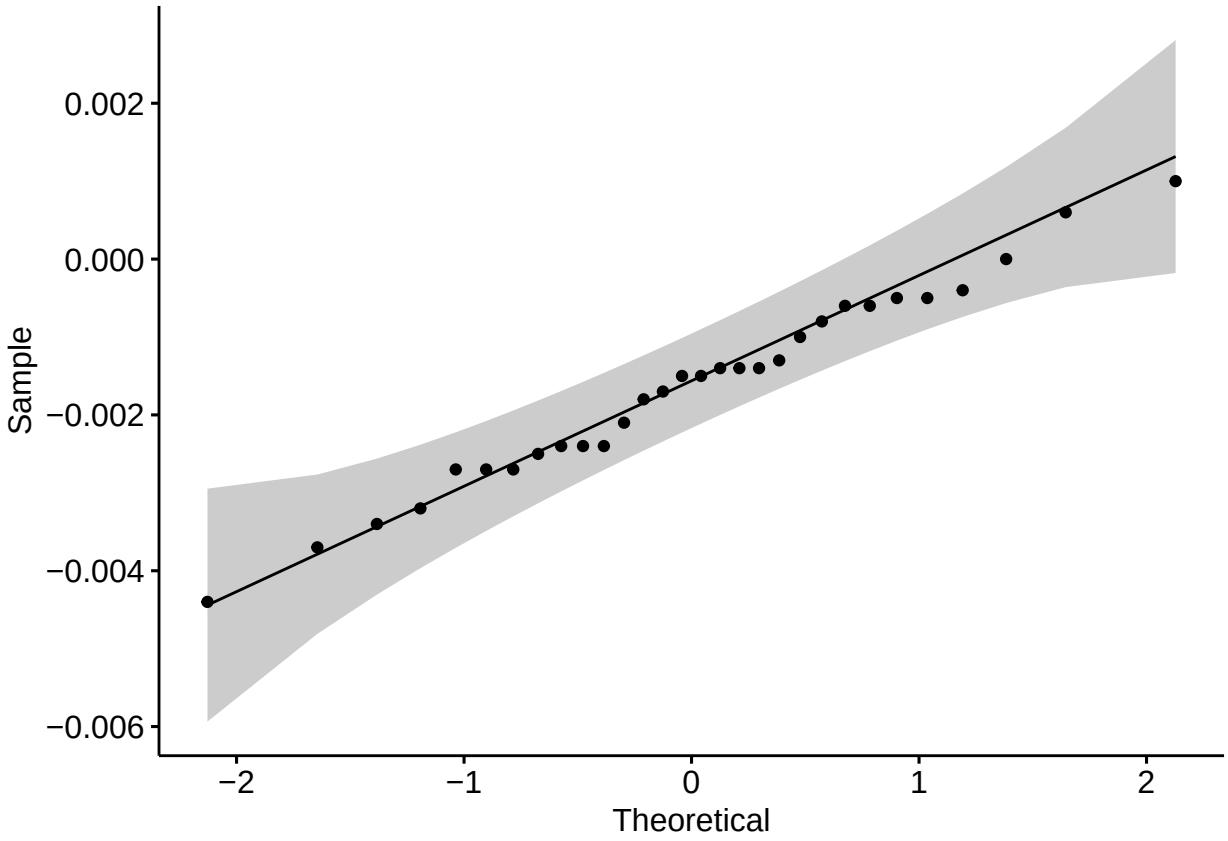
```
ggdensity(differences$acc_diff_m1, xlab="acc diff m1")
```



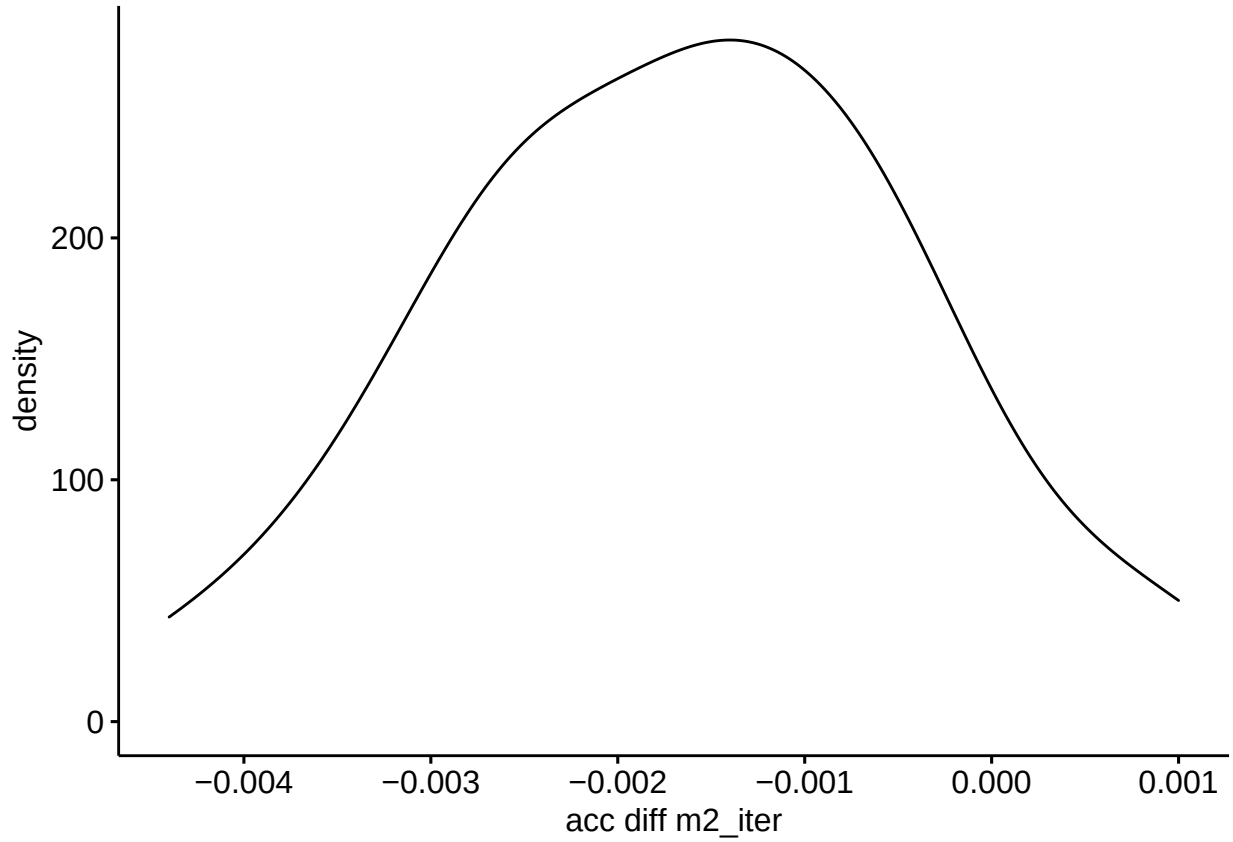

```
ggdensity(differences$acc_diff_m2, xlab="acc diff m2")
```



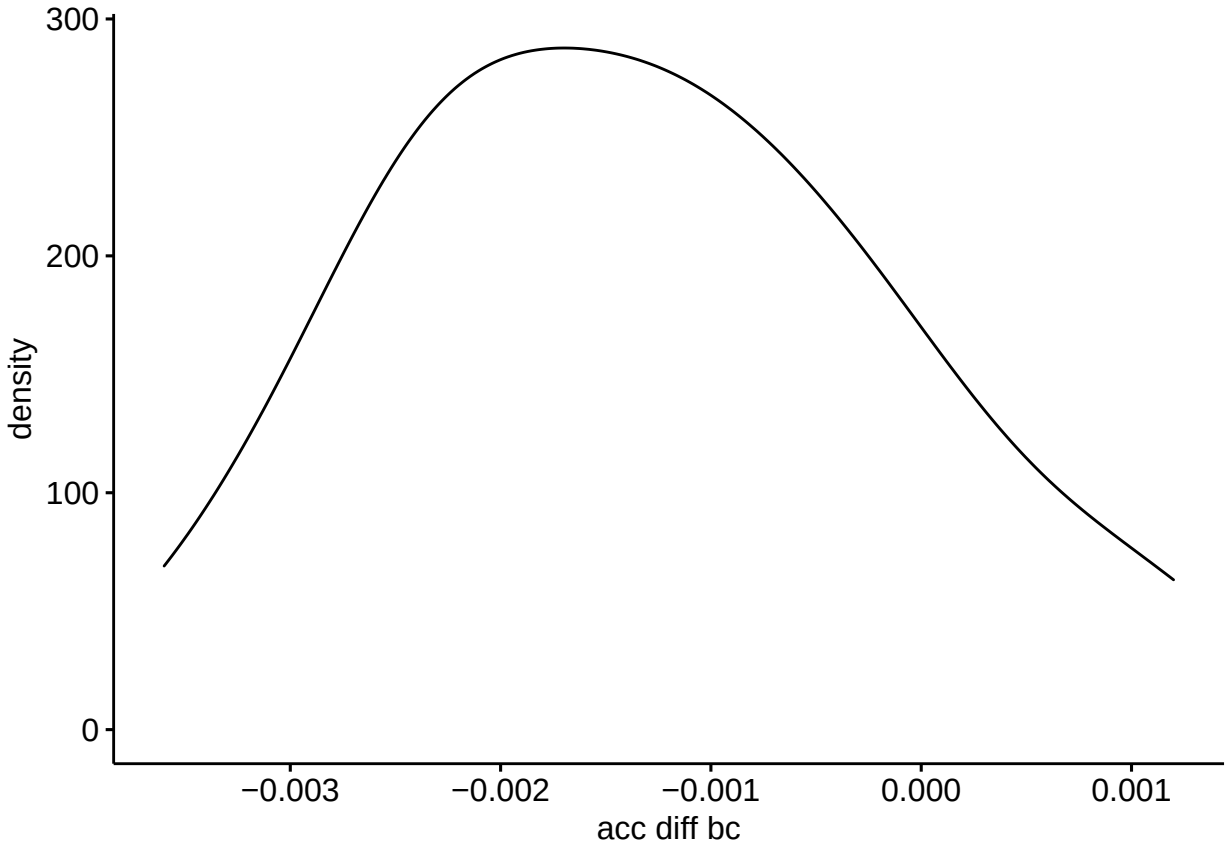

```
ggqqplot(differences$acc_diff_m2)
```



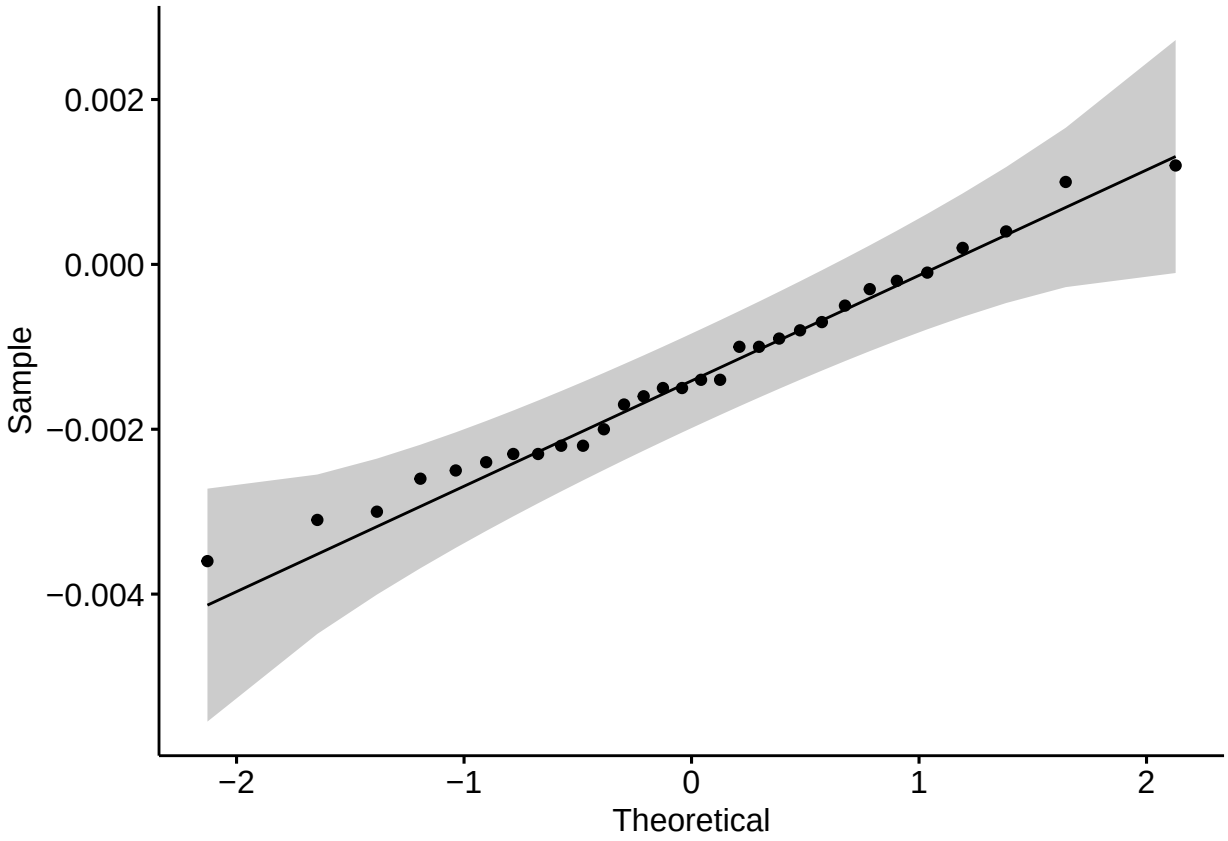

```
ggdensity(differences$acc_diff_m2_iter, xlab="acc diff m2_iter")
```




```
ggdensity(differences$acc_diff_bc, xlab="acc diff bc")
```




```
ggqqplot(differences$acc_diff_bc)
```




```
shapiro.test(differences$acc_diff_bc)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  differences$acc_diff_bc
## W = 0.98211, p-value = 0.8784
```

Accuracy differences seem to be normal, so we can proceed with the t-tests.

```
t.test(ens_results_wide$accuracy_vt_m1, ens_results_wide$accuracy_tt_m1, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  ens_results_wide$accuracy_vt_m1 and ens_results_wide$accuracy_tt_m1
## t = -7.2749, df = 29, p-value = 5.186e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.002105334 -0.001181332
## sample estimates:
## mean of the differences
##            -0.001643333
```

```r
t.test(ens_results_wide$accuracy_vt_m1, ens_results_wide$accuracy_tt_m1, paired=TRUE, alternative="less
```

```
##
##  Paired t-test
##
## data:  ens_results_wide$accuracy_vt_m1 and ens_results_wide$accuracy_tt_m1
## t = -7.2749, df = 29, p-value = 2.593e-08
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##          -Inf -0.001259514
## sample estimates:
## mean of the differences
##            -0.001643333
```

```r
t.test(ens_results_wide$accuracy_vt_m2, ens_results_wide$accuracy_tt_m2, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  ens_results_wide$accuracy_vt_m2 and ens_results_wide$accuracy_tt_m2
## t = -7.1694, df = 29, p-value = 6.84e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.002116413 -0.001176920
## sample estimates:
## mean of the differences
##            -0.001646667
```

```r
t.test(ens_results_wide$accuracy_vt_m2, ens_results_wide$accuracy_tt_m2, paired=TRUE, alternative="less
```

```
##
##  Paired t-test
##
## data:  ens_results_wide$accuracy_vt_m2 and ens_results_wide$accuracy_tt_m2
## t = -7.1694, df = 29, p-value = 3.42e-08
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##          -Inf -0.001256413
## sample estimates:
## mean of the differences
##            -0.001646667
```

```r
t.test(ens_results_wide$accuracy_vt_m2_iter, ens_results_wide$accuracy_tt_m2_iter, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  ens_results_wide$accuracy_vt_m2_iter and ens_results_wide$accuracy_tt_m2_iter
## t = -7.1969, df = 29, p-value = 6.363e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -0.002123182 -0.001183484
## sample estimates:
## mean of the differences
##            -0.001653333
```

```
t.test(ens_results_wide$accuracy_vt_m2_iter, ens_results_wide$accuracy_tt_m2_iter, paired=TRUE, alterna
```

```
##
##  Paired t-test
##
## data:  ens_results_wide$accuracy_vt_m2_iter and ens_results_wide$accuracy_tt_m2_iter
## t = -7.1969, df = 29, p-value = 3.182e-08
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##          -Inf -0.001262994
## sample estimates:
## mean of the differences
##            -0.001653333
```

```
t.test(ens_results_wide$accuracy_vt_bc, ens_results_wide$accuracy_tt_bc, paired=TRUE)
```

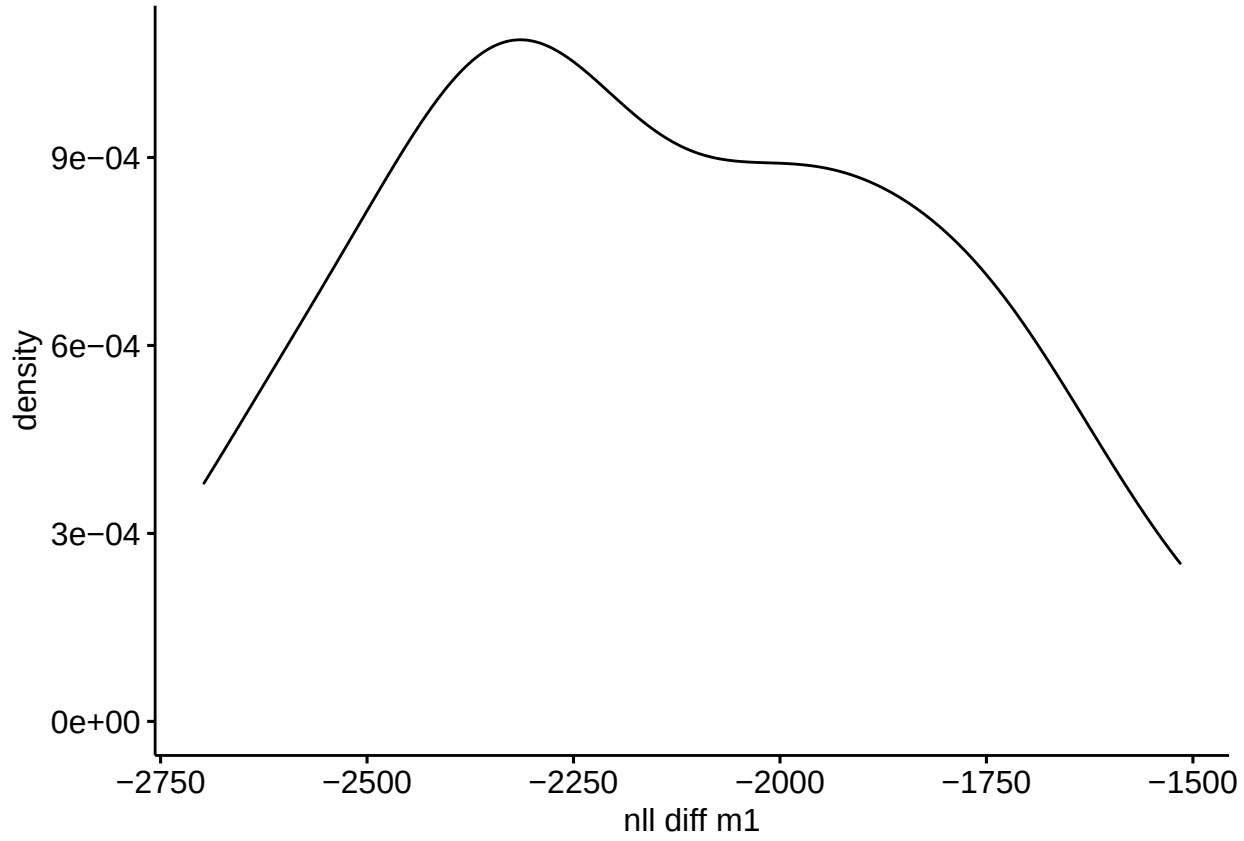
```
##
##  Paired t-test
##
## data:  ens_results_wide$accuracy_vt_bc and ens_results_wide$accuracy_tt_bc
## t = -6.0607, df = 29, p-value = 1.349e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.0017832767 -0.0008833899
## sample estimates:
## mean of the differences
##            -0.001333333
```

```
t.test(ens_results_wide$accuracy_vt_bc, ens_results_wide$accuracy_tt_bc, paired=TRUE, alternative="less
```

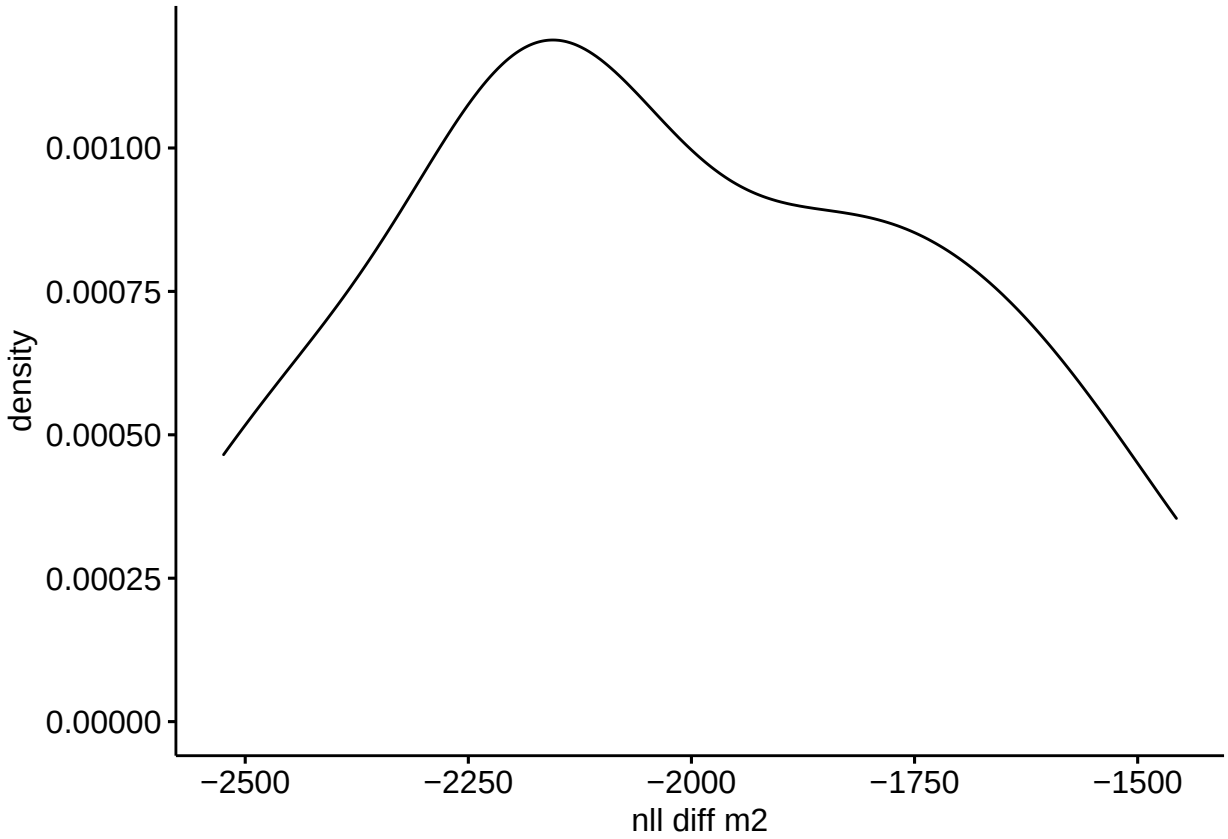
```
##
##  Paired t-test
##
## data:  ens_results_wide$accuracy_vt_bc and ens_results_wide$accuracy_tt_bc
## t = -6.0607, df = 29, p-value = 6.743e-07
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##           -Inf -0.0009595313
## sample estimates:
## mean of the differences
##            -0.001333333
```

For all coupling methods, we found, that ensemble models trained on subset (of size 500) of the neural networks training set have statistically significantly better accuracy than ensemble models trained on separate validation set (of the same size).

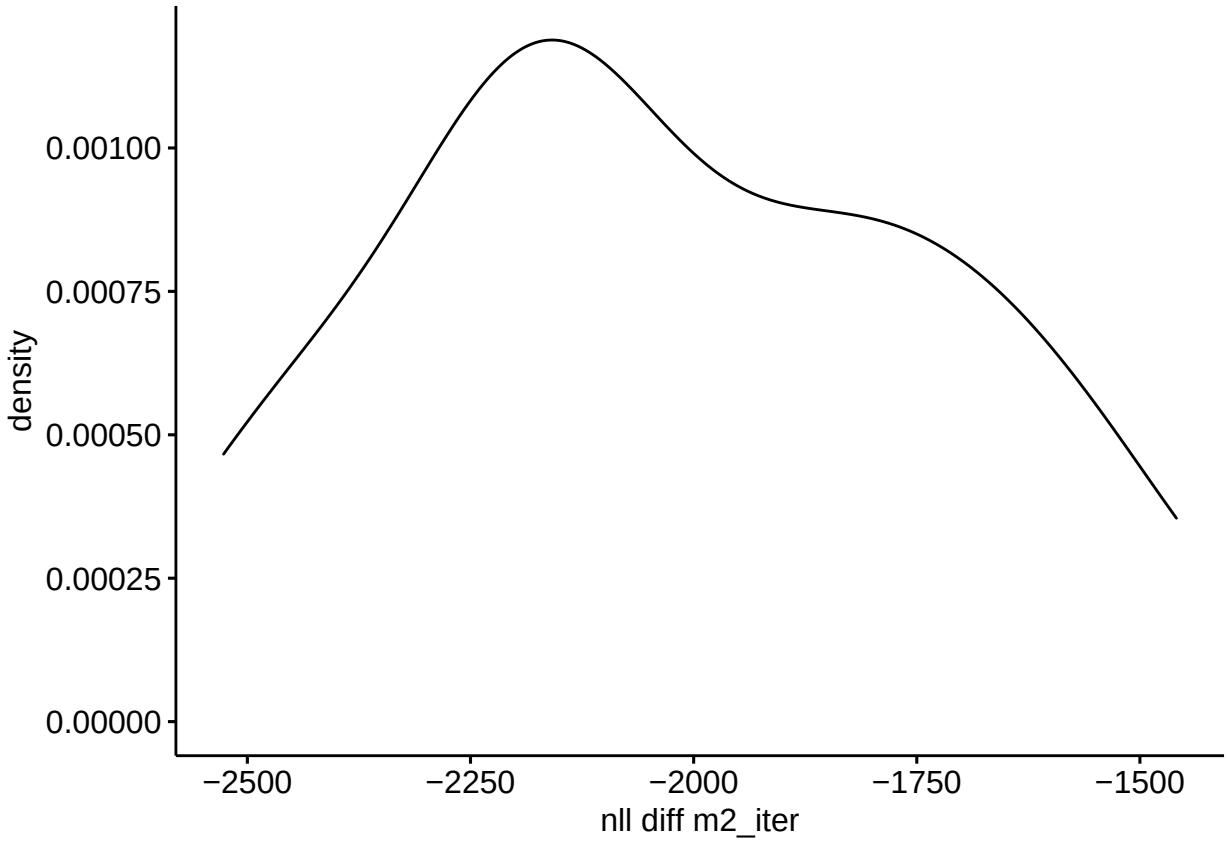
```
ggdensity(differences$nll_diff_m1, xlab="nll diff m1")
```



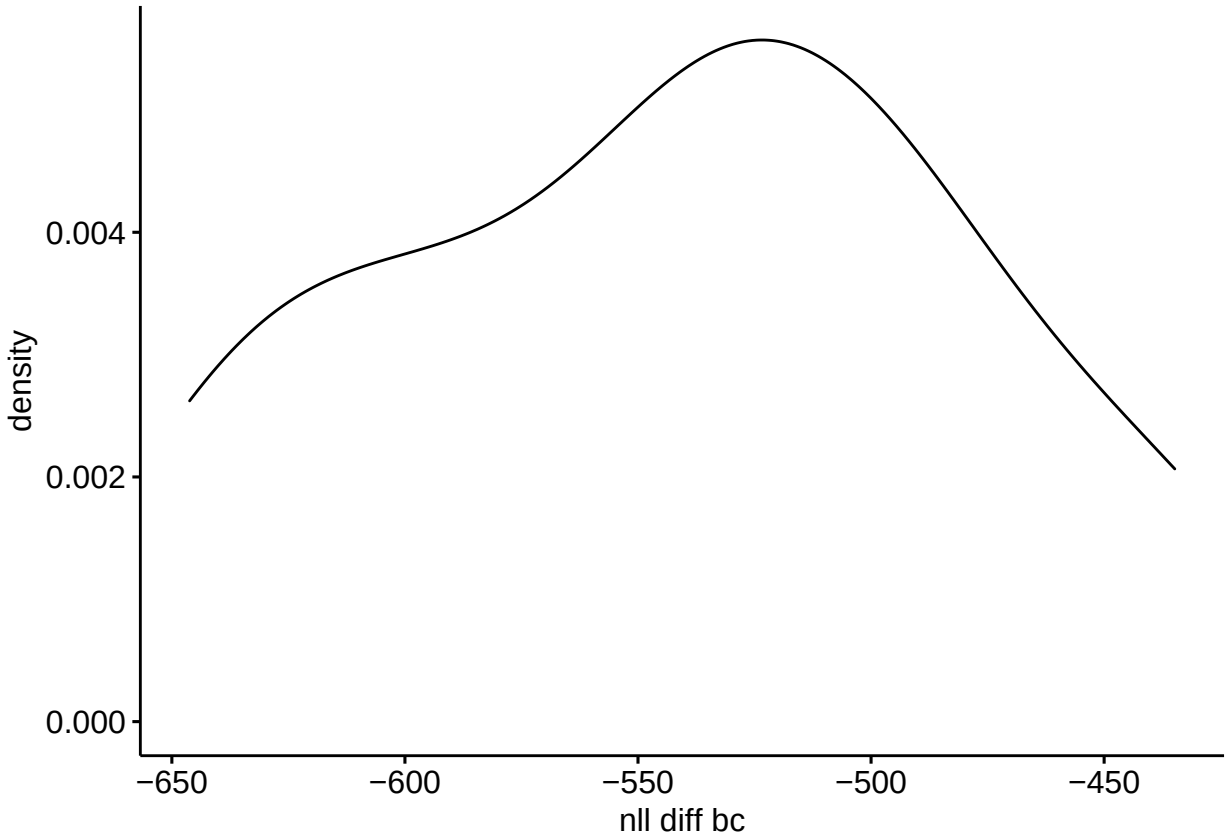

```
ggdensity(differences$nll_diff_m2, xlab="nll diff m2")
```




```
ggdensity(differences$nll_diff_m2_iter, xlab="nll diff m2_iter")
```




```
ggdensity(differences$nll_diff_bc, xlab="nll diff bc")
```



We postpone tests on nll due to strange properties of this metric on our ensembles. First, further research into this behavior will be needed.