Predicting new MCIA scores

Max Mattesich, Joaquin Reyna, Anna Konstorum

2022-09-01

Predicting MCIA global (factor) scores for new test samples

It may be of interest to use the embedding that is calculated on a training sample set to predict scores on a test set (or, equivalently, on new data).

After loading the nipalsMCIA library (see the X vignette for instructions on how to install), we randomly split the NCI60 cancer cell line data into training and test sets.

```
library(nipalsMCIA)
library(ggplot2)
```

```
data(NCI60)

set.seed(8)
num_samples = dim(data_blocks[[1]])[1]
num_train = round(num_samples*0.7,0)
train_samples = sample.int(num_samples,num_train)

data_blocks_train<-data_blocks
data_blocks_test<-data_blocks
for (i in 1:length(data_blocks)){
  data_blocks_train[[i]]<-data_blocks_train[[i]][train_samples,]
  data_blocks_test[[i]]<-data_blocks_test[[i]][-train_samples,]
}
```

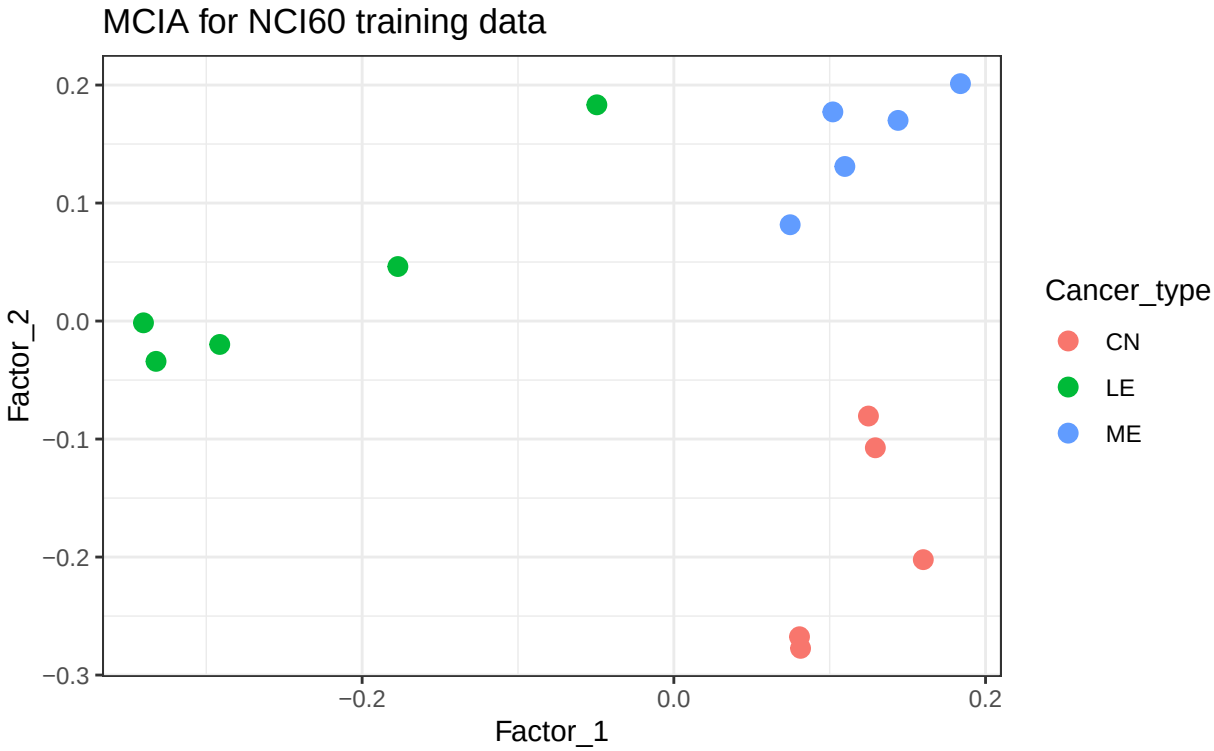
Run NIPALS-MCIA on training data

```
MCIA_train <- nipals_multiblock(data_blocks_train, preprocMethod='colprofile',num_PCs = 10,
                                plots='all', tol=1e-12)
```

Visualize model on training data using metadata on cancer type

The metadata for cancer type is contained in the sample name (before the period)
e.g. 'ME.UACC_62' is type 'ME'

```
cancer_type <- substr(rownames(data_blocks_train$mrna), 1, 2)
MCIA_out<-data.frame(MCIA_train$global_scores[,1:2])
MCIA_out$Cancer_type<-cancer_type
colnames(MCIA_out)<-c("Factor_1", "Factor_2", "Cancer_type")
ggplot(data = MCIA_out, aes(x=Factor_1, y=Factor_2, color=Cancer_type))+
  geom_point(size=3) +
  theme_bw() +
  ggtitle("MCIA for NCI60 training data")
```



Generate factor scores for test data using the MCIA_train model

We use the `predict_gs` function to generate new factor scores on the test data set using the MCIA_train model above

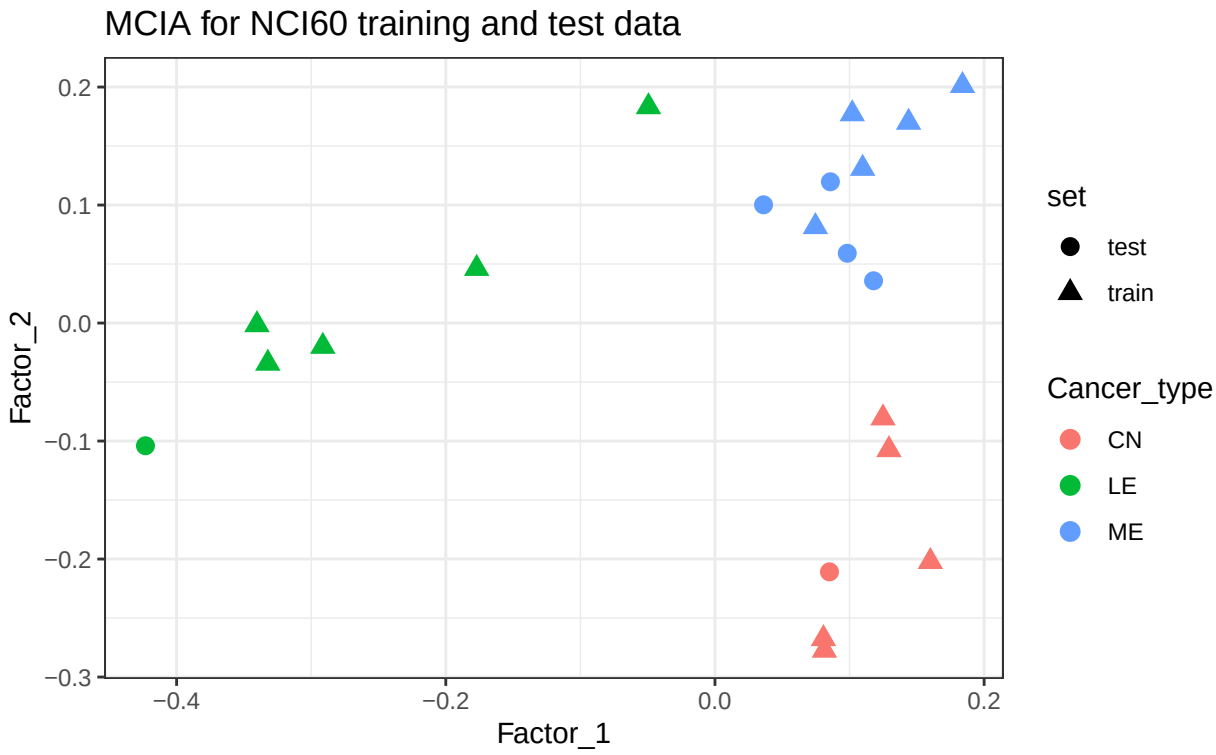
```
MCIA_test_scores <- predict_gs(MCIA_train,data_blocks_test)
```

Visualize new scores with old

We once again plot the top two factor scores for both the training and test datasets

```
cancer_type <- substr(rownames(data_blocks_test$mrna), 1, 2)
MCIA_out_test<-data.frame(MCIA_test_scores[,1:2])
MCIA_out_test$Cancer_type<-cancer_type
colnames(MCIA_out_test)<-c("Factor_1", "Factor_2", "Cancer_type")
MCIA_out_test$set<-"test"
MCIA_out$set<-"train"
MCIA_out_full<-rbind(MCIA_out,MCIA_out_test)
rownames(MCIA_out_full)<-NULL

ggplot(data = MCIA_out_full, aes(x=Factor_1, y=Factor_2, color=Cancer_type, shape=set))+
  geom_point(size=3) +
  theme_bw() +
  ggtitle("MCIA for NCI60 training and test data")
```


MCIA for NCI60 training and test data
Factor_2
Factor_1
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.2
0.0
0.2
set
test
train
Cancer_type
CN
LE
ME