Outlier Analysis

Domenick Braccia

8/15/2019

Overview

Here, outlier features from each of the three pipelines considered - dada2, mothur and qiime - are examined. Features are labeled outliers based on 6 different error metrics: "median_error", "iqr_error", "rcov_error", "mean_error", "var_error" and "cov_error"

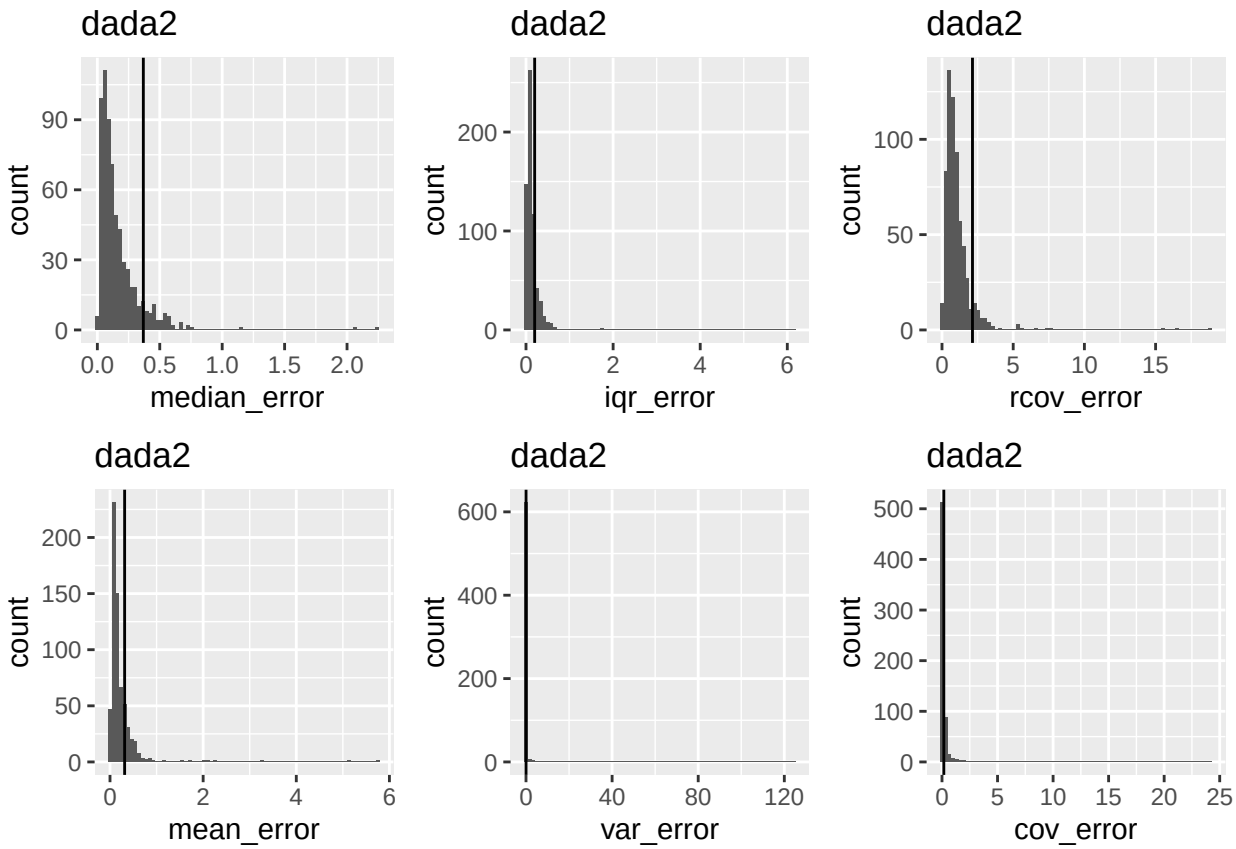
Retrieving saved data

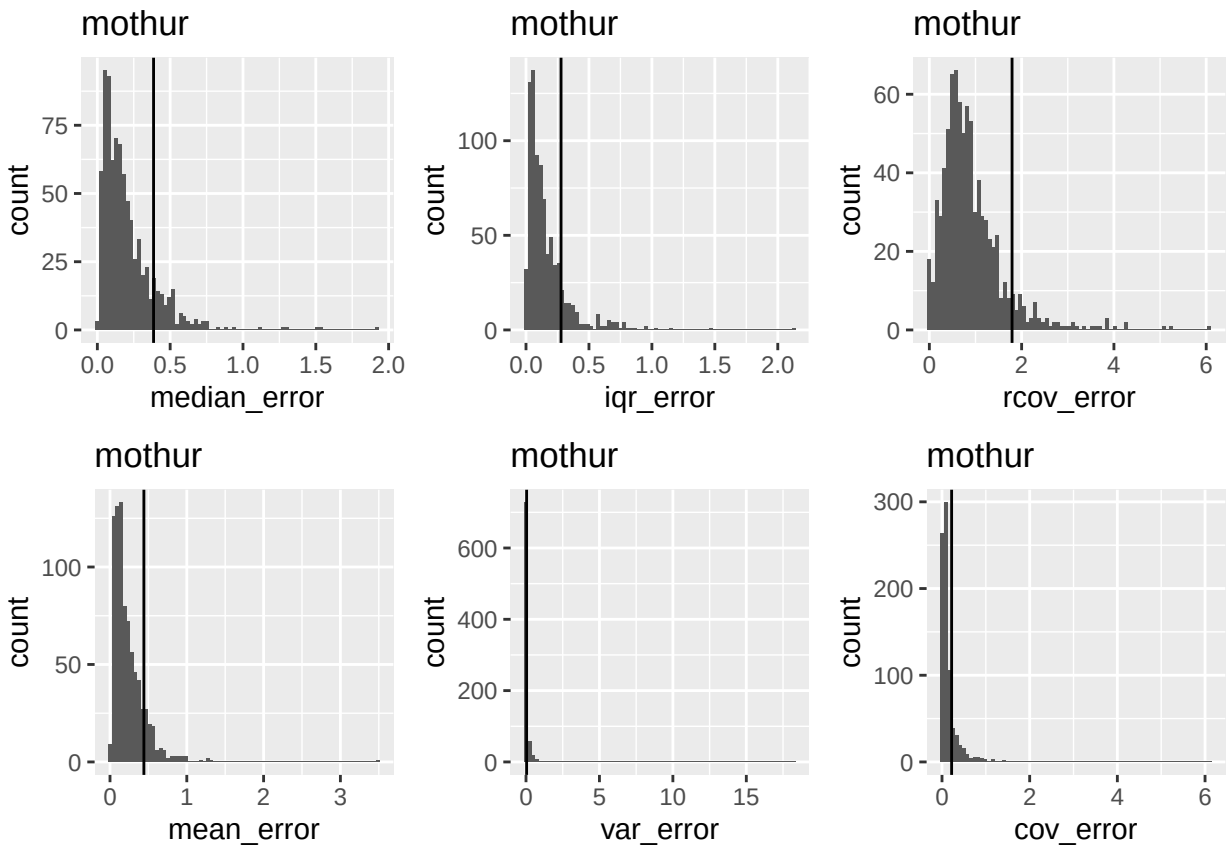
Data comes from relative abundance error metrics section of "relative_abundance_assessment_results.Rmd" file and is saved to the ~/data/ folder. Boxplots for each error metric are drawn from which outliers will be extracted and examined.

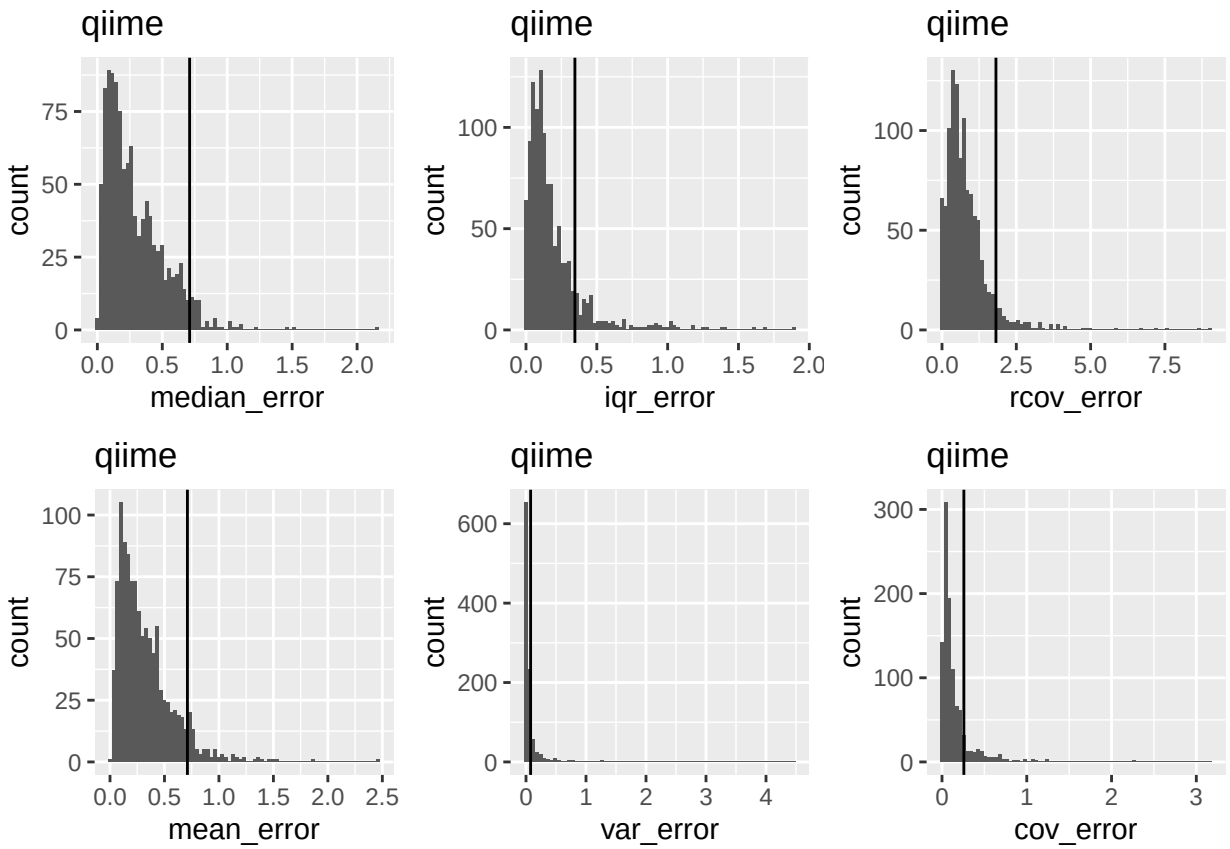
```
# loading data saved from rel_abundance_est Rmd file #
rel_abu_error <- readRDS(file = "../data/rel_abu_error.RDS")
rel_abu_error_summary <- readRDS(file = "../data/rel_abu_error_summary.RDS")
```

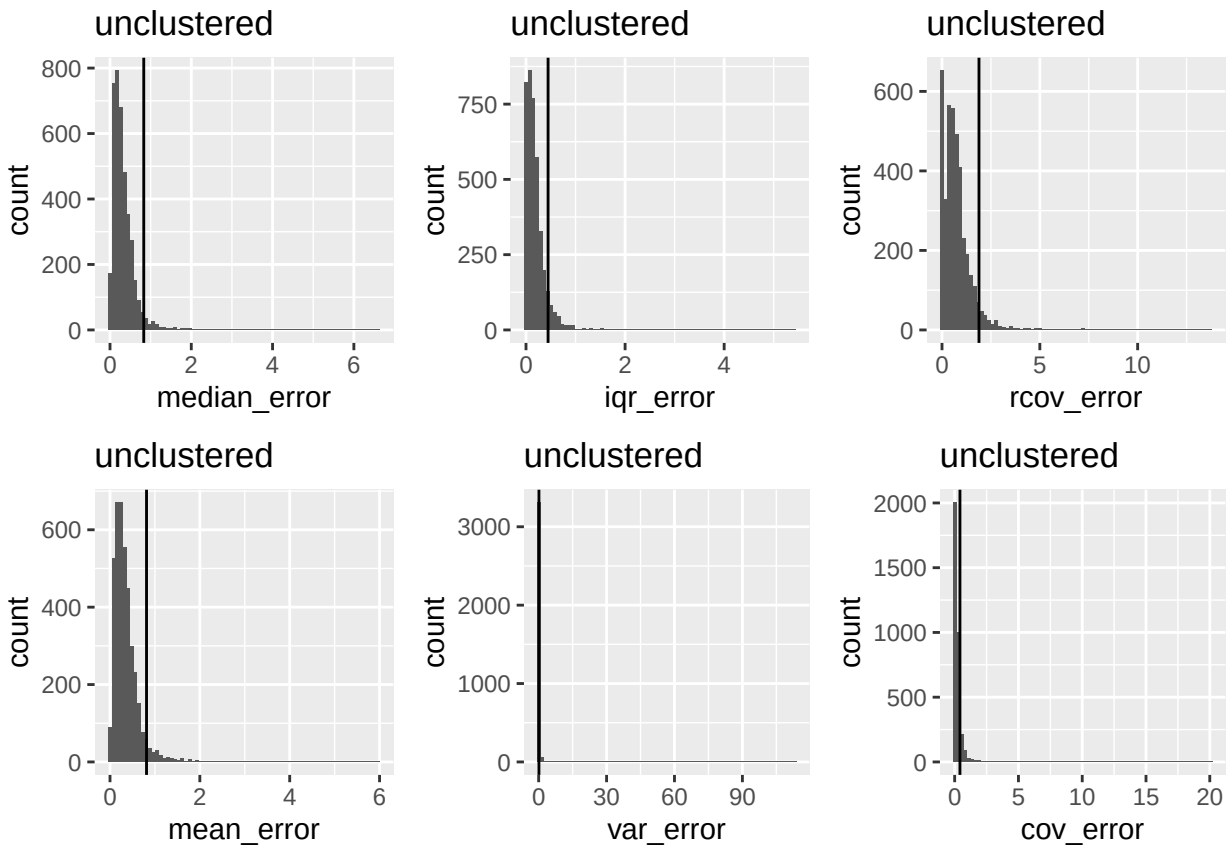
Plotting error distributions

Below are distributions of the feature level error metrics examined. The vertical line on each plot indicates the value at which features are labeled "outliers"


dada2
count
median_error
dada2
count
iqr_error
dada2
count
rcov_error
dada2
count
mean_error
dada2
count
var_error
dada2
count
cov_error


mothur
count
median_error
mothur
count
iqr_error
mothur
count
rcov_error
mothur
count
mean_error
mothur
count
var_error
mothur
count
cov_error


qiime
count
median_error
qiime
count
iqr_error
qiime
count
rcov_error
qiime
count
mean_error
qiime
count
var_error
qiime
count
cov_error



Examining "repeat" outliers

Below, two columns are added to the outlier summary tables.

`repeat_prop` contains the proportion of outlier features that appear as outliers in more than one sample. Overall, there tends to be very few - if any - repeat outliers for each of the three pipelines examined.

`prop_greater_2x` contains the proportion of outlier features that are greater than 2x the minumum outlier error metric value. These proportions seems to vary greatly by pipeline.

NOTE: for the variance and covariance error metrics, a handful of features were not identified as outliers or inlers. This seems to stem from the original definition of an outlier and inlier from earlier in the analysis, since most features have the same value for `var_error` or `cov_error` (see histograms for these metrics above).

```
#Q: do features within a pipeline show up as outliers for more than one sample?
#Q: how many outlier features have values 2X the smallest outlier value?

##### GENERIC #####
#repeat_outliers <- function(metric_cat) {
repeat_outliers_row <- vector(mode = "numeric")
repeat_outleirs <- matrix(0, nrow = 4, ncol = 6)
prop_greater_2x_row <- vector(mode = "numeric")
prop_greater_2x <- matrix(0, nrow = 4, ncol = 6)
for (i in seq(pipelines)) {
    for (j in seq(metrics)) {
```

```
        #calculating number of outliers > 2x the minumum outlier value
        current_outliers <- filter(outliers_cat[[metric_cat[j]]], pipe == pipelines[i], outlier_cat == 
        prop_greater_2x_row[j] <- sum(current_outliers > 2 * min(current_outliers)) / nrow(current_outli

        #making repeat outlier rows for summary tables
        current_metric <- outliers_cat[[metric_cat[j]]] %>%
            filter(pipe == pipelines[i], outlier_cat == "outlier") %>%
            group_by(feature_id) %>%
            summarise(num_samples = n())
        repeat_outliers_row[j] <- sum(current_metric$num_samples > 1) / length(current_metric$num_sample
    }
    #print(repeat_outliers_row)
    repeat_outleirs[i, ] <- repeat_outliers_row
    #print(prop_greater_2x_row)
    prop_greater_2x[i, ] <- prop_greater_2x_row
}

for (i in seq(metrics)) {
    outliers[[i]] %<>%
        add_column(repeat_prop = repeat_outleirs[, i]) %>%
        add_column(prop_greater_2x = prop_greater_2x[, i])
}
print(outliers, width = Inf)
```

```
## $median_error
## # A tibble: 4 x 7
## # Groups:   pipe [4]
##   metric pipe        inlier outlier outlier_prop repeat_prop
##   <chr>  <chr>        <int>   <int>        <dbl>       <dbl>
## 1 median dada2          595      45       0.0756      0.0227
## 2 median mothur         780      47       0.0603      0.0222
## 3 median qiime         1079      22       0.0204      0
## 4 median unclustered   3795     156       0.0411      0.0331
##   prop_greater_2x
##             <dbl>
## 1          0.0889
## 2          0.191
## 3          0.136
## 4          0.167
##
## $iqr_error
## # A tibble: 4 x 7
## # Groups:   pipe [4]
##   metric pipe        inlier outlier outlier_prop repeat_prop
##   <chr>  <chr>        <int>   <int>        <dbl>       <dbl>
## 1 iqr    dada2          600      40       0.0667      0
## 2 iqr    mothur         786      41       0.0522      0.025
## 3 iqr    qiime         1021      80       0.0784      0.0127
## 4 iqr    unclustered   3741     210       0.0561      0.0244
##   prop_greater_2x
##             <dbl>
## 1           0.575
## 2           0.561
```

```
## 3           0.438
## 4           0.233
##
## $rcov_error
## # A tibble: 4 x 7
## # Groups:   pipe [4]
##   metric pipe        inlier outlier outlier_prop repeat_prop
##   <chr>  <chr>        <int>   <int>        <dbl>       <dbl>
## 1 rcov   dada2          606      34       0.0561      0
## 2 rcov   mothur         789      38       0.0482      0.0556
## 3 rcov   qiime         1044      57       0.0546      0.0556
## 4 rcov   unclustered   3770     181       0.0480      0.0343
##   prop_greater_2x
##             <dbl>
## 1           0.294
## 2           0.289
## 3           0.298
## 4           0.221
##
## $mean_error
## # A tibble: 4 x 7
## # Groups:   pipe [4]
##   metric pipe        inlier outlier outlier_prop repeat_prop
##   <chr>  <chr>        <int>   <int>        <dbl>       <dbl>
## 1 mean   dada2          581      59       0.102       0.0172
## 2 mean   mothur         788      39       0.0495      0
## 3 mean   qiime         1071      30       0.0280      0
## 4 mean   unclustered   3775     176       0.0466      0.0353
##   prop_greater_2x
##             <dbl>
## 1           0.322
## 2           0.282
## 3           0.167
## 4           0.159
##
## $var_error
## # A tibble: 4 x 8
## # Groups:   pipe [4]
##   metric pipe        inlier outlier `<NA>` outlier_prop repeat_prop
##   <chr>  <chr>        <int>   <int>  <int>        <dbl>       <dbl>
## 1 var    dada2          553      83      4        0.150      0.0375
## 2 var    mothur         721      92     14        0.128      0.0952
## 3 var    qiime          921     131     49        0.142      0.0656
## 4 var    unclustered   2948     432    571        0.147      0.0485
##   prop_greater_2x
##             <dbl>
## 1           0.940
## 2           0.707
## 3           0.718
## 4           0.755
##
## $cov_error
## # A tibble: 4 x 8
## # Groups:   pipe [4]
```

```
##   metric pipe        inlier outlier `<NA>` outlier_prop repeat_prop
##   <chr>  <chr>        <int>   <int>  <int>        <dbl>       <dbl>
## 1 cov    dada2          571      65      4       0.114       0.0317
## 2 cov    mothur         750      63     14       0.084       0.125
## 3 cov    qiime          948     104     49       0.110       0.0505
## 4 cov    unclustered   3081     299    571       0.0970      0.0418
##   prop_greater_2x
##             <dbl>
## 1           0.708
## 2           0.444
## 3           0.587
## 4           0.361
```