# Package ‘ConQuR’

May 9, 2022

**Type** Package

**Title** Batch Effects Removal for Microbiome Data in Large-Scale Epidemiology Studies via Conditional Quantile Regression

**Version** 2.0

**Author** Wodan Ling and Michael C. Wu

**Maintainer** Wodan Ling <wling@fredhutch.org>

**Description** This package conducts batch effects removal from a taxa read count table by a conditional quantile regression method. The distributional attributes of microbiome data - zero-inflation and over-dispersion, are simultaneously considered.

**License** GPL (>=2)

**Encoding** UTF-8

**LazyData** true

**NeedsCompilation** no

**Imports** quantreg, cqrReg, glmnet, dplyr, doParallel, gplots, vegan, ade4, compositions, randomForest, ROCR, ape, GUniFrac, fastDummies

**RoxygenNote** 7.1.1

**Suggests** knitr,
rmarkdown

**VignetteBuilder** knitr

**Depends** R (>= 3.5.0)

## R topics documented:

ConQuR *Remove batch effects from a taxa read count table*

## Description

Remove batch effects from a taxa read count table

## Usage

```
ConQuR(
  tax_tab,
  batchid,
  covariates,
  batch_ref,
  logistic_lasso = F,
  quantile_type = "standard",
  simple_match = F,
  lambda_quantile = "2p/n",
  interplt = F,
  delta = 0.4999,
  taus = seq(0.005, 0.995, by = 0.005),
  num_core = 2
)
```

## Arguments

`tax_tab` The taxa read count table, samples (row) by taxa (col).

`batchid` The batch indicator, must be a factor.

`covariates` The data.frame contains the key variable of interest and other covariates, e.g., data.frame(key, x1, x2).

`batch_ref` A character, the name of the reference batch, e.g.,“2”.

`logistic_lasso` A logical value, TRUE for L1-penalized logistic regression, FALSE for standard logistic regression; default is FALSE.

`quantile_type` A character, “standard” for standard quantile regression, “lasso” for L1-penalized quantile regression, “composite” for composite quantile regression; default is “standard”.

`simple_match` A logical value, TRUE for using the simple quantile-quantile matching, FALSE for not; default is FALSE.

`lambda_quantile`

A character, the penalization parameter in quantile regression if `quantile_type`=“lasso” or “composite”; only two choices “2p/n” or “2p/logn”, where p is the number of expanded covariates and n is the number of non-zero read count; default is “2p/n”.

`interplt` A logical value, TRUE for using the data-driven linear interpolation between zero and non-zero quantiles to stablize border estimates, FALSE for not; default is FALSE.

`delta` A real constant in (0, 0.5), determing the size of the interpolation window if interplt=TRUE, a larger delta leads to a narrower interpolation window; default is 0.4999.

`taus` A sequence of quantile levels, determing the “precision” of estimating conditional quantile functions; default is seq(0.005, 0.995, by=0.005).

`num_core` A real constant, the number of cores used for computing; default is 2.

## Details

- Choose `batch_ref` based on prior knowledge, or try several options, there is no default.
- The option “composite” of `quantile_type` is aggressive, use with caution.
- If choose `simple_match`=TRUE, `logistic_lasso`, `quantile_type`, `lambda_quantile`, `interplt` and `delta` won’t take effect.
- Always use a fine grid of `taus` if the size of data is adequate.

## Value

The corrected taxa read count table, samples (row) by taxa (col).

## References

- Ling, W. et al. (2021+). ConQuR: batch effects removal for microbiome data in large-scale epidemiology studies via conditional quantile regression.
- Ling, W. et al. (2020+). Statistical inference in quantile regression for zero-inflated outcomes. Statistica Sinica.
- Machado, J.A.F., Silva, J.S. (2005). Quantiles for counts. Journal of the American Statistical Association 100(472), 1226–1237.
- Koenker, R. & Bassett Jr, G. (1978). Regression quantiles. Econometrica: journal of the Econometric Society, 33-50.
- Koenker, R. (2005). Econometric Society Monographs: Quantile Regression. New York: Cambridge University.
- Zou, H. & Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. The Annals of Statistics 36, 1108-1126.

---

# `ConQuR_libsize` *Remove batch effects from a taxa read count table (library size is included in the model)*

---

## Description

Remove batch effects from a taxa read count table (library size is included in the model)

## Usage

```
ConQuR_libsize(
  tax_tab,
  batchid,
  covariates,
  libsize_tune = NULL,
  batch_ref,
  logistic_lasso = F,
  quantile_type = "standard",
  simple_match = F,
  lambda_quantile = "2p/n",
  interplt = F,
  delta = 0.4999,
  taus = seq(0.005, 0.995, by = 0.005),
  num_core = 2
)
```

## Arguments

`tax_tab` The taxa read count table, samples (row) by taxa (col).

`batchid` The batch indicator, must be a factor.

`covariates` The data.frame contains the key variable of interest and other covariates, e.g., data.frame(key, x1, x2).

`libsize_tune` The samples' library sizes, which is designed to pass values from `Tune_ConQuR_libsize`; do not assign values, keep it as NULL.

`batch_ref` A character, the name of the reference batch, e.g.,"2".

`logistic_lasso` A logical value, TRUE for L1-penalized logistic regression, FALSE for standard logistic regression; default is FALSE.

`quantile_type` A character, "standard" for standard quantile regression, "lasso" for L1-penalized quantile regression, "composite" for composite quantile regression; default is "standard".

`simple_match` A logical value, TRUE for using the simple quantile-quantile matching, FALSE for not; default is FALSE.

`lambda_quantile` A character, the penalization parameter in quantile regression if `quantile_type`="lasso" or "composite"; only two choices "2p/n" or "2p/logn", where p is the number of expanded covariates and n is the number of non-zero read count; default is "2p/n".

`interplt` A logical value, TRUE for using the data-driven linear interpolation between zero and non-zero quantiles to stablize border estimates, FALSE for not; default is FALSE.

`delta` A real constant in (0, 0.5), determing the size of the interpolation window if interplt=TRUE, a larger delta leads to a narrower interpolation window; default is 0.4999.

`taus` A sequence of quantile levels, determing the "precision" of estimating conditional quantile functions; default is seq(0.005, 0.995, by=0.005).

`num_core` A real constant, the number of cores used for computing; default is 2.

## Details

- Use this version when between-batch library size variability is not considered nuisance and maintained in the corrected taxa read count table.
- Choose `batch_ref` based on prior knowledge, or try several options, there is no default.
- The option "composite" of `quantile_type` is aggressive, use with caution.
- If choose `simple_match`=TRUE, `logistic_lasso`, `quantile_type`, `lambda_quantile`, `interplt` and `delta` won't take effect.
- Always use a fine grid of `taus` if the size of data is adequate.

### Value

The corrected taxa read count table, samples (row) by taxa (col).

### References

- Ling, W. et al. (2021+). ConQuR: batch effects removal for microbiome data in large-scale epidemiology studies via conditional quantile regression.
- Ling, W. et al. (2020+). Statistical inference in quantile regression for zero-inflated outcomes. Statistica Sinica.
- Machado, J.A.F., Silva, J.S. (2005). Quantiles for counts. Journal of the American Statistical Association 100(472), 1226–1237.
- Koenker, R. & Bassett Jr, G. (1978). Regression quantiles. Econometrica: journal of the Econometric Society, 33-50.
- Koenker, R. (2005). Econometric Society Monographs: Quantile Regression. New York: Cambridge University.
- Zou, H. & Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. The Annals of Statistics 36, 1108-1126.

---

## `PERMANOVA_R2` *PERMANOVA R2 of batch and variable of interest*

---

### Description

PERMANOVA R2 of batch and variable of interest

### Usage

```
PERMANOVA_R2(TAX, batchid, covariates, key_index)
```

### Arguments

| | |
|---|---|
| `TAX` | The taxa read count table, samples (row) by taxa (col). |
| `batchid` | The batch indicator, must be a factor. |
| `covariates` | The data.frame contains the key variable of interest and other covariates. |
| `key_index` | An integer, location of the variable of interest in `covariates`. |

### Details

Three PERMANOVA R2 will be computed: (1) the standard one (adnois), (2) on euclidified dissimilarities (adonis2, sqrt.dist=T), and (3) with a constant added to the non-diagonal dissimilarities such that all eigenvalues are non-negative in the underlying PCoA (adonis2, add=T).

### Value

A list

- tab_count - A table summarizing PERMANOVA R2 computed on the original taxa read count table in Bray-Curtis dissimilarity.
- tab_rel - A table summarizing PERMANOVA R2 computed on the corresponding relative abundance table in Euclidean dissimilarity (Aitchison dissimilarity).

### References

- Anderson, M. J. (2014). Permutational multivariate analysis of variance (PERMANOVA). Wiley statsref: statistics reference online, 1-15.

## `Plot_PCoA` *Stratified PCoA plots*

### Description

Stratified PCoA plots

### Usage

```
Plot_PCoA(
  TAX,
  factor,
  sub_index = NULL,
  dissimilarity = "Bray",
  GUniFrac_type = "d_0.5",
  tree = NULL,
  main = NULL,
  aa = 1.5
)
```

### Arguments

`TAX` The taxa read count table, samples (row) by taxa (col).

`factor` The variable for stratification, e.g., batchid or the variable of interest, must be a factor.

`sub_index` A vector of sample indices, to restrict the analysis to a subgroup of samples, e.g., c(1:5, 15:20); default is NULL.

`dissimilarity` The dissimilarity type, “Bray” for Bray-Curtis dissimilarity, “Aitch” for Aitchison dissimilarity, “GUniFrac” for generalized UniFrac dissimilarity; default is “Bray”.

`GUniFrac_type` The generalized UniFrac type, “d_1” for weighted UniFrac, “d_UW” for unweighted UniFrac, “d_VAW” for variance adjusted weighted UniFrac, “d_0” for generalized UniFrac with alpha 0, “d_0.5” for generalized UniFrac with alpha 0.5; default is “d_0.5”.

`tree` The rooted phylogenetic tree of R class “phylo”, must be provided when `dissimilarity`=“GUniFrac” default is NULL.

`main` The title of plot; default is NULL.

`aa` A real number, the character size for the title.

### Value

Print a PCoA plot.

### References

- Chen, J., & Chen, M. J. (2018). Package ‘GUniFrac’. The Comprehensive R Archive Network (CRAN).

## RF_Pred *Predict binary variables based on a taxa read count table by random forest*

### Description

Predict binary variables based on a taxa read count table by random forest

### Usage

```
RF_Pred(TAX, factor, fold = 5, seed = 2020)
```

### Arguments

| | |
|---|---|
| `TAX` | The taxa read count table, samples (row) by taxa (col). |
| `factor` | The binary variable to predict, e.g., the key variable, case/control, must be a factor. |
| `fold` | The number of folds; default is 5. |
| `seed` | The seed to generate fold indices for samples; default is 2020. |

### Value

A list

- pred - A table summarizing the predicted probabilities and true labels for all samples.
- auc_across_fold - AUC of the ROC curves across folds.
- auc_on_all - AUC of the ROC curve on all samples.

## RF_Pred_Multiclass *Predict multiclass variables based on a taxa read count table by random forest*

### Description

Predict multiclass variables based on a taxa read count table by random forest

### Usage

```
RF_Pred_Multiclass(TAX, factor, fold = 5, seed = 2020)
```

### Arguments

| | |
|---|---|
| `TAX` | The taxa read count table, samples (row) by taxa (col). |
| `factor` | The multiclass variable to predict, e.g., the key variable, never smoker/former smoker/current smoker, must be a factor. |
| `fold` | The number of folds; default is 5. |
| `seed` | The seed to generate fold indices for samples; default is 2020. |

### Value

A list

- pred - A table summarizing the predicted probabilities and true labels for all samples.
- cross_entropy_across_fold - mean cross-entropy across folds.

## RF_Pred_Regression *Predict continuous variables based on a taxa read count table by random forest*

### Description

Predict continuous variables based on a taxa read count table by random forest

### Usage

```
RF_Pred_Regression(TAX, variable, fold = 5, seed = 2020)
```

### Arguments

| | |
|---|---|
| `TAX` | The taxa read count table, samples (row) by taxa (col). |
| `variable` | The continuous variable to predict. |
| `fold` | The number of folds; default is 5. |
| `seed` | The seed to generate fold indices for samples; default is 2020. |

### Value

A list

- pred - A table summarizing the predicted and true values for all samples.
- rmse_across_fold - RMSEs across folds.

## Sample_Data *Example data, a taxa read count table, with batchid, key variable and covariates*

### Description

A dataset containing 100 taxa from 3 batches, key variable is sbp, with covariates, sex, race and age

### Usage

```
Sample_Data
```

## Format

A taxa read count (273 samples by 100 taxa), batchid and the metadata:

**batchid** factor, with levels 0, 1, 2

**sbp** key variable, systolic blood pressure, continuous variable

**sex** covariate 1, binary variable

**race** covariate 2, binary variable

**age** covariate 3, continuous variable

---

# `Tune_ConQuR` *Tune over variations of ConQuR*

---

## Description

Tune over variations of ConQuR

## Usage

```
Tune_ConQuR(
  tax_tab,
  batchid,
  covariates,
  batch_ref_pool,
  logistic_lasso_pool,
  quantile_type_pool,
  simple_match_pool,
  lambda_quantile_pool,
  interplt_pool,
  frequencyL,
  frequencyU,
  cutoff = 0.1,
  delta = 0.4999,
  taus = seq(0.005, 0.995, by = 0.005),
  num_core = 2
)
```

## Arguments

`tax_tab` The taxa read count table, samples (row) by taxa (col).

`batchid` The batch indicator, must be a factor.

`covariates` The data.frame contains the key variable of interest and other covariates, e.g., data.frame(key, x1, x2).

`batch_ref_pool` A vector of characters, the candidates for reference batch, e.g., c(“0”, “2”).

`logistic_lasso_pool`
A vector of logical values, whether or not using the L1-penalized logistic regression, e.g., c(T, F).

`quantile_type_pool`
A vector of characters, the candidates for quantile regression type, e.g., c(“standard”, “lasso”).

`simple_match_pool`
A vector of logical values, whether or not using the simple quantile-quantile matching, e.g., c(T, F).

`lambda_quantile_pool`
A vector of characters, the candidates for the penalization parameter in quantile regression (“lasso” or “composite”), e.g., c(NA, “2p/n”, “2p/logn”).

`interplt_pool` A vector of logical values, whether or not using the data-driven linear interpolation between zero and non-zero quantiles, e.g., c(T, F).

`frequencyL` A real constant between 0 and 1, the lower bound of prevalence that needs tuning.

`frequencyU` A real constant between 0 and 1, the upper bound of prevalence that needs tuning.

`cutoff` A real constant, the grid size of prevalence for tuning; default is 0.1.

`delta` A real constant in (0, 0.5), determing the size of the interpolation window if interplt=TRUE, a larger delta leads to a narrower interpolation window; default is 0.4999.

`taus` A sequence of quantile levels, determing the “precision” of estimating conditional quantile functions; default is seq(0.005, 0.995, by=0.005).

`num_core` A real constant, the number of cores used for computing; default is 2.

## Details

- “original”, i.e., the original data without correction is always a default candidate.
- If “standard” is one candidate for `quantile_type_pool`, always include NA as one candidate for `lambda_quantile_pool`.
- Be cautious with candidate “composite” for `quantile_type_pool`, the underlying assumption is strong and the computation might be slow.
- The tuning procedure finds the local optimal in each cutoff. If `frequencyL`=0.2, `frequencyU`=0.5 and `cutoff`=0.1, the functions determines the combination achieving maximum removal of batch variations on taxa present in 20%-30%, ..., 40%-50% of the samples, respectively.
- The same reference batch is used across taxa in the final optimal corrected table.

## Value

A list

- tax_final - The optimal corrected taxa read count table, samples (row) by taxa (col).
- method_final - A table summarizing variations of ConQuR chosen for each prevalence cutoff.

## References

- Ling, W. et al. (2021+). ConQuR: batch effects removal for microbiome data in large-scale epidemiology studies via conditional quantile regression
- Ling, W. et al. (2020+). Statistical inference in quantile regression for zero-inflated outcomes. Statistica Sinica.
- Machado, J.A.F., Silva, J.S. (2005). Quantiles for counts. Journal of the American Statistical Association 100(472), 1226–1237.
- Koenker, R. & Bassett Jr, G. (1978). Regression quantiles. Econometrica: journal of the Econometric Society, 33-50.

- Koenker, R. (2005). Econometric Society Monographs: Quantile Regression. New York: Cambridge University.
- Zou, H. & Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. The Annals of Statistics 36, 1108-1126.
- Anderson, M. J. (2014). Permutational multivariate analysis of variance (PERMANOVA). Wiley statsref: statistics reference online, 1-15.

## `Tune_ConQuR_libsize` *Tune over variations of ConQuR_libsize*

### Description

Tune over variations of ConQuR_libsize

### Usage

```
Tune_ConQuR_libsize(
  tax_tab,
  batchid,
  covariates,
  batch_ref_pool,
  logistic_lasso_pool,
  quantile_type_pool,
  simple_match_pool,
  lambda_quantile_pool,
  interplt_pool,
  frequencyL,
  frequencyU,
  cutoff = 0.1,
  delta = 0.4999,
  taus = seq(0.005, 0.995, by = 0.005),
  num_core = 2
)
```

### Arguments

`tax_tab` The taxa read count table, samples (row) by taxa (col).

`batchid` The batch indicator, must be a factor.

`covariates` The data.frame contains the key variable of interest and other covariates, e.g., data.frame(key, x1, x2).

`batch_ref_pool` A vector of characters, the candidates for reference batch, e.g., c(“0”, “2”).

`logistic_lasso_pool`
A vector of logical values, whether or not using the L1-penalized logistic regression, e.g., c(T, F).

`quantile_type_pool`
A vector of characters, the candidates for quantile regression type, e.g., c(“standard”, “lasso”).

`simple_match_pool`
A vector of logical values, whether or not using the simple quantile-quantile matching, e.g., c(T, F).

`lambda_quantile_pool`
: A vector of characters, the candidates for the penalization parameter in quantile regression (“lasso” or “composite”), e.g., c(NA, “2p/n”, “2p/logn”).

`interplt_pool`
: A vector of logical values, whether or not using the data-driven linear interpolation between zero and non-zero quantiles, e.g., c(T, F).

`frequencyL`
: A real constant between 0 and 1, the lower bound of prevalence that needs tuning.

`frequencyU`
: A real constant between 0 and 1, the upper bound of prevalence that needs tuning.

`cutoff`
: A real constant, the grid size of prevalence for tuning; default is 0.1.

`delta`
: A real constant in (0, 0.5), determing the size of the interpolation window if interplt=TRUE, a larger delta leads to a narrower interpolation window; default is 0.4999.

`taus`
: A sequence of quantile levels, determing the “precision” of estimating conditional quantile functions; default is seq(0.005, 0.995, by=0.005).

`num_core`
: A real constant, the number of cores used for computing; default is 2.

## Details

- Use this version when between-batch library size variability is not considered nuisance and maintained in the corrected taxa read count table.
- “original”, i.e., the original data without correction is always a default candidate.
- If “standard” is one candidate for `quantile_type_pool`, always include NA as one candidate for `lambda_quantile_pool`.
- Be cautious with candidate “composite” for `quantile_type_pool`, the underlying assumption is strong and the computation might be slow.
- The tuning procedure finds the local optimal in each cutoff. If `frequencyL`=0.2, `frequencyU`=0.5 and `cutoff`=0.1, the functions determines the combination achieving maximum removal of batch variations on taxa present in 20%-30%, ..., 40%-50% of the samples, respectively.
- The same reference batch is used across taxa in the final optimal corrected table.

## Value

A list

- tax_final - The optimal corrected taxa read count table, samples (row) by taxa (col).
- method_final - A table summarizing variations of ConQuR_libsize chosen for each prevalence cutoff.

## References

- Ling, W. et al. (2021+). ConQuR: batch effects removal for microbiome data in large-scale epidemiology studies via conditional quantile regression
- Ling, W. et al. (2020+). Statistical inference in quantile regression for zero-inflated outcomes. Statistica Sinica.
- Machado, J.A.F., Silva, J.S. (2005). Quantiles for counts. Journal of the American Statistical Association 100(472), 1226–1237.
- Koenker, R. & Bassett Jr, G. (1978). Regression quantiles. Econometrica: journal of the Econometric Society, 33-50.

- Koenker, R. (2005). Econometric Society Monographs: Quantile Regression. New York: Cambridge University.
- Zou, H. & Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. The Annals of Statistics 36, 1108-1126.
- Anderson, M. J. (2014). Permutational multivariate analysis of variance (PERMANOVA). Wiley statsref: statistics reference online, 1-15.

# Index