Received: 22 January 2018 | Accepted: 30 April 2018

DOI: 10.1111/2041-210X.13029

APPLICATION

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

RRPP: An R package for fitting linear models to high-dimensional data using residual randomization

Michael L. Collyer[1] | Dean C. Adams[2,3]

[1]Department of Science-Biology, Chatham University, Pittsburgh, Pennsylvania

[2]Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa

[3]Department of Statistics, Iowa State University, Ames, Iowa

Correspondence
Michael L. Collyer, Department of Science-Biology, Chatham University, Pittsburgh, PA.
Email: m.collyer@chatham.edu

Funding information
Division of Environmental Biology, Grant/Award Number: 1556379 and 1737895; NSF DEB Awards, Grant/Award Number: 1737895 and 1556379

Handling Editor: Robert Freckleton

Abstract

1. Residual randomization in permutation procedures (RRPP) is an appropriate means of generating empirical sampling distributions for ANOVA statistics and linear model coefficients, using ordinary or generalized least-squares estimation. This is an especially useful approach for high-dimensional (multivariate) data.
2. Here, we present an R package that provides a comprehensive suite of tools for applying RRPP to linear models. Important available features include choices for OLS or GLS coefficient estimation, data or dissimilarity matrix analysis capability, choice among types I, II, or III sums of squares and cross-products, various effect size estimation methods, and an ability to perform mixed-model ANOVA.
3. The `lm.rrpp` function is similar to the `lm` function in many regards, but provides coefficient and ANOVA statistics estimates over many random permutations. The S3 generic functions commonly used with `lm` also work with `lm.rrpp`. Additionally, a `pairwise` function provides statistical tests for comparisons of least-squares means or slopes, among designated groups. Users have many options for varying random permutations. Compared to similar available packages and functions, RRPP is extremely fast and yields comprehensive results for downstream analyses and graphics, following model fits with `lm.rrpp`.
4. The RRPP package facilitates analysis of both univariate and multivariate response data, even when the number of variables exceeds the number of observations.

KEYWORDS
dissimilarity, generalized least-squares, high-dimensional data, multivariate

1 | INTRODUCTION

The ability to analyse multidimensional traits and other multivariate data has become a requisite skill for evolutionary biologists and ecologists over the last few decades. This reality is important, as high-dimensional data are not just encumbrance, but often necessary for addressing important research questions. Advanced computing technology has spurred the quick collection of "high-dimensional" data (e.g., genomics, imaging, remote sensing) and some characteristics of research subjects just cannot be described with scant information. High-dimensional (continuous) data are multivariate-response data that require strict attention to using all variables (p) in analyses, even if they exceed the number of observations (n), as they are needed to describe complex multidimensional traits of research subjects. For example, a comparative morphometric analysis might include many anatomical landmarks (large p, relative to n) whose two- or three-dimensional Cartesian coordinates comprise a single trait, organism "shape" (Adams, Rohlf, & Slice, 2013). It is not uncommon to have anatomical landmarks per specimen exceed the number of specimens, especially if such information is obtained by 3D laser scanners.

The general analytical problem high-dimensional data pose is that when p exceeds the error degrees of freedom of linear models (n minus the number of model parameters), parametric analyses

like multivariate analysis of variance (M-ANOVA) are not available. Parametric M-ANOVA involves approximating *F*-statistics from multivariate statistics (like Wilks' λ) to obtain approximate probability distributions. Because *F*-distributions require positive degrees of freedom, *n* must sufficiently exceed *p* for an appropriate *F*-distribution to be used. (We provide an example of this limitation in the Analytical Demonstrations, below).

In the last few decades, with better computing power, resampling procedures have largely offered solutions to this parametric limitation (e.g., Adams & Collyer, 2018a; Anderson, 2001a; Collyer, Sekora, & Adams, 2015). Resampling procedures offer two advantages. First, *F*-statistics can be estimated with respect to interobservation distances in multivariate data spaces, eliminating the need to concern oneself with variable number in calculating degrees of freedom (Anderson, 2001a). *F*-statistics calculated this way do not have appropriate *F*-distributions as probability distributions, but second, empirical probability distributions can be generated via the resampling procedure in lieu of parametric distributions. These distributions closely match parametric distributions for ideal conditions (large *n*:*p* ratio, multivariate normal error, homoscedasticity; Adams & Collyer, 2018b).

There have been two chief approaches for resampling procedures applied to high-dimensional data, largely based on the type of data involved. The first involves resampling the row vectors of (potentially but not necessarily large) matrices of data or residuals to generate random data outcomes, from which linear model coefficients, sums of squares (*SS*), and other statistics can be estimated many times (Adams & Collyer, 2015; Anderson & ter Braak, 2003; Collyer et al., 2015; Gonzalez & Manly, 1998). The second involves randomization of values in (potentially large) "distance" matrices that describe interobservation dissimilarities (e.g., Euclidean distance, Canberra distance; see Legendre & Legendre, 2012 for many examples). Joint randomization of matrix rows and columns produces random distance matrices, from which linear model coefficients and statistics can be estimated many times (Anderson, 2001a; McArdle & Anderson, 2001; Zapala & Schork, 2006). The latter offers the additional advantage of choosing among alternative dissimilarity indices (distances), which is beneficial if either only distances are available or the type of data collected influences how distances should be measured (Anderson, 2001a; but see Warton, Wright, & Wang, 2012).

Through many lines of research focusing on type I error rates, effect size estimation, and statistical power, it can be generally concluded that for linear models, randomization of residuals (vectors or distances) in a permutation procedure (RRPP) is a sound method for generating empirical distributions of linear model statistics (Adams, 2014; Adams & Collyer, 2015, 2018a,b; Adams & Felice, 2014; Anderson & ter Braak, 2003; Collyer et al., 2015; Freedman & Lane, 1983; Gonzalez & Manly, 1998). Recent evidence has shown that this method is effective for both ordinary least-squares (OLS) and generalized least-squares (GLS) estimation of linear model coefficients and *SS* (Adams & Collyer, 2018b). Additionally, not only can RRPP handle large *p*:*n* ratios (focusing on the largeness of *p*, not the smallness of *n*) but large *p*:*n* ratios can offer enhanced statistical power, using RRPP (Adams & Collyer, 2018b).

The purpose of this application is to introduce a powerful `R` (R Core Team, 2018) package, `RRPP`, which performs ANOVA and other statistical tests via RRPP for any *p*:*n* data scenario or any *n* × *n* dissimilarity matrix, with either OLS or GLS estimation, ability to choose among types I, II, or III *SS*, and several sampling distribution methods for estimating effect sizes. Prediction and plotting capabilities round out a suite of features that should be recognized as standard components of the general analytical toolkit for any researcher who has used linear models and ANOVA to analyse data. Although designed for high-dimensional data, `RRPP` can be used for data of any dimension and will produce fast, reliable results.

2 | DESCRIPTION

This description of RRPP is brief, but a detailed conceptual development is provided in Supporting Information Appendix S1. RRPP is the process of randomizing residuals from null linear models that when added to fitted values, produce random pseudovalues (Anderson, 2001b; Anderson & ter Braak, 2003; Collyer et al., 2015). Doing this many times and estimating linear model coefficients and *SS* for full models (containing effects of null models, plus an effect to test) over random permutations yield empirical sampling distributions of ANOVA statistics. Importantly, RRPP holds constant the effects of null models rather than conflate them with the effects that are tested (Collyer et al., 2015). This nonparametric procedure can be applied in a systematic way to calculate *P*-values for various linear model designs (Anderson, 2001b; Anderson & ter Braak, 2003; Collyer et al., 2015). Both OLS and GLS methods of estimation are possible (Adams & Collyer, 2018b), and different types of *SS* can be calculated. The statistical properties (type I error rates and statistical power) of RRPP have largely been validated (Adams & Collyer, 2018a,b; Anderson, 2001b). The `RRPP` package is the most comprehensive package to feature RRPP methodology for any linear model analysis, and performs similarly to the widely used `lm` function in the `R` package, `stats` (Table 1). However, with GLS capability, the RRPP package generalizes the purpose of the strictly univariate `gls` function in the `nlme` `R` package (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2018) to analysis of multivariate data. (Note that `gls` should not be confused with `glm`, the latter referring to generalized linear models rather than generalized least-squares.) The option to choose among *SS* types generalizes a purpose of the strictly univariate `R` package, `car` (Companion to Applied Regression; Fox & Weisberg, 2011) to analysis of multivariate data. The `RRPP` package is, to the best of our knowledge, the only R package with such capabilities for multivariate data.

3 | ANALYTICAL DEMONSTRATIONS

The analytical demonstrations presented here are also presented in a vignette for R (Supporting Information Appendix S2). The

TABLE 1 Comparison of common functions using `lm` and `lm.rrpp` functions

Function	`lm`	`lm.rrpp`
main function	• Estimates coefficients, fitted values, and residuals. • Provides model design matrix and its QR decomposition.	• Estimates coefficients, fitted values, and residuals for many permutations, and for every null-model/full-model comparison. • Provides model design matrices and their QR decompositions for every null-model/full-model comparison. • Provides ANOVA statistics and effect size estimates for many permutations. • Provides additional permutation information for downstream analysis. • Offers choice of SS type (I, II, III) in calculations. • Has a parallel processing option for large datasets (limited to Unix systems).
`summary`	• Provides *t*-test statistics for coefficients. • Provides full model ANOVA.	• Provides full model ANOVA. • Provides a summary of data, *SS* type (I, II, III), and the permutation procedure used. • Coefficients tests found instead in `coef` S3 generic function • Returns SSCP matrices for model effects and residuals
`coef`	• Returns linear model coefficients only.	• Returns linear model coefficients. • Provides analogous statistics to `summary.lm` but based on distributions of vector Euclidean distances for the many random coefficients generated by lm.rrpp.
`anova`	• Performs parametric ANOVA on univariate data only, with type I *SS*. (`drop1` can be used to perform type III SS.) Does not allow choice of random effects for *MS* in *F*-value calculations (although the aov function has such options).	• Performs nonparametric ANOVA with RRPP on either univariate or multivariate data, or dissimilarity matrices. Choice of *SS* type, random effects for *MS* in *F*-value calculations, and effect size type are available. • Data dimensionality is not an issue.
`manova`	• Performs parametric ANOVA on multivariate data only. • *F*-statistics can be approximated multiple ways • Data variables must be less in number than residual degrees of freedom. • `summary.manova.lm` returns SSCP matrices for model effects and residuals	• No such function is needed. • However, SSCP matrices are output in `summary.lm.fit`
`predict`	• Estimates parametric confidence or prediction intervals for points within or beyond data ranges. • Flexible options.	• Generates a list of nonparametric confidence intervals for all points requested, based on coefficients estimated many times in lm.rrpp. • plot.predict.lm.rrpp is an additional function that can plot predicted values and either confidence interval bars or ellipses in univariate or multivariate plots.
`plot`	• Returns multiple diagnostic plots of residuals.	• Can choose from among: o Diagnostic plots of residual Euclidean distances or PC scores o Regression plots for predictor variables o Principal component plots (from covariance matrices of model fitted values)
`pairwise`	• Not available for `lm`. A comparable function (in limited capacity) is `tukeyHSD` for `aov` objects (which could be obtained from `lm` objects)	• A novel function for performing pairwise tests for `lm.rrpp` model fits, for each least-squares means or slopes. Alternative null models can also be used.

vignette has more examples, provides R output, and facilitates concurrent demonstration in R. Three datasets are provided in `RRPP` (with examples in help pages) and are used here as examples of the package: `Pupfish` (Collyer et al., 2015), `PupfishHeads` (Gilbert, 2016), and `PlethMorph` (Adams & Collyer, 2018a). The first two datasets contain landmark-based geometric morphometric data collected from museum samples of Pecos pupfish (*Cyprinodon pecosensis*), representing body shape and cranial morphology, respectively. Within both datasets, the `$coords` objects are matrices of Procrustes residuals obtained from generalized Procrustes analysis (GPA) of configurations of anatomical landmarks. For the purposes of this example, it is sufficient to recognize that Procrustes residuals embody a highly multivariate dataset representing shape (see the `R` package, `geomorph`; Adams, Collyer, Kaliontzopoulou, & Sherratt, 2017; for applications of GPA in the context of shape analysis). The third dataset contains averaged linear measurements of 37 species of Plethodontid salamanders, plus a covariance matrix based on a Brownian model of evolution, given the phylogenetic relationship among the species (see Adams & Collyer, 2018a, for details).

TABLE 2 ANOVA statistics for fixed-effects ($_F$) and mixed-effects($_M$) models, for the `PupfishHeads` (head size) example. Italicized values indicate where *F*-values were calculated with respect to random effects (Locality:year) rather than residuals. *Z* and *p* are based on 1,000 random permutations, using RRPP

Source	*df*	*SS*	*MS*	R^2	F_F	Z_F	p_F	F_M	Z_M	p_M
Sex	1	0.609	0.609	0.045	14.409	1.713	0.001	14.409	1.713	0.001
Locality	1	0.471	0.471	0.034	11.147	1.667	0.001	*1.726*	*0.697*	*0.248*
Locality:year	6	1.638	0.273	0.120	6.457	3.230	0.001	6.457	3.230	0.001
Model	8	2.718	0.340	0.199	8.037	12.687	0.001			
Residuals	259	10.949	0.042	0.801						
Total	267	13.667								

3.1 | Example: Pupfish cranial morphology and mixed-model ANOVA

In the first example, we use a univariate-dependent variable (head size, measured as the centroid size of the cranial landmark configuration; Bookstein, 1991) with a mixed-effects model design. The following code highlights the analytical steps:

```
> data("PupfishHeads")
> PupfishHeads$logHeadSize <- log(PupfishHeads$headSize)
> fit <- lm.rrpp(logHeadSize ~ sex + locality/year, SS.type = "I",
  data = PupfishHeads)
> summary(fit)
> anova(fit, effect.type = "F")
```

Of important note, we choose to log-transform head size and include it as a separate variable in the `RRPP` data frame, `PupfishHeads`. We accomplished this via the code above—rather than using `log(headSize)`—because downstream functions like `predict.lm.rrpp` work better without functions in the formula. ANOVA was performed using random distributions of *F*-statistics to calculate *z*-scores and *p*-values (but one could use alternative distributions—see the `RRPP` help page). The S3 Generic functions (`summary`, `anova`) return summaries that remind the user how random data were generated, the type of *SS*, and how *z*-scores were calculated. This particular ANOVA summary is a default that fails to consider the year fish were sampled as a random effect (Table 2). A mixed-model ANOVA update can be performed by changing the expected mean-square (*MS*) error estimates in each *F* calculation:

```
> anova(fit, effect.type = "F",
  error = c("Residuals", "locality:year", "Residuals"))
```

This adjustment illustrates that the head size variation does not significantly differ between localities, with respect to the variation among sampling events (Table 2). Although not apparent in this example, the `anova.lm.rrpp` function can also perform multimodel inferential tests (as demonstrated in the help page and Supporting Information Appendix S2). One might wish to also look at individual model coefficients, and ascertain which have the largest effect:

```
> coef(fit, test = TRUE)
```

TABLE 3 Tests of coefficient vector lengths (distance, *d*), with observed values, upper 95% confidence limit, *Z*-scores, and *p*-values provided. *Z* and *p* are based on 1,000 random permutations, using RRPP. The standard is the mean for females, from the lake habitat, in 1987. In each case, *d* is amount of change in head size for the coefficient indicated

Coefficient	*d*	*95% UCL*	*Z*	*p*
Male	0.096	0.058	4.202	0.002
SH	0.092	0.058	3.790	0.001
SH:1998	0.064	0.135	0.237	0.344
Lake:1999	0.357	0.136	6.937	0.001
SH:1999	0.013	0.139	−0.964	0.846
SH:2000	0.072	0.102	0.992	0.167
Lake:2001	0.230	0.131	4.180	0.002
Lake:2002	0.310	0.138	5.901	0.001

This function produces a table much like `summary.lm` output, but with bootstrap-generated confidence intervals of coefficients (Table 3).

It might be of interest to visualize model predictions for certain effects, holding constant other effects. For example, if we want to look at confidence intervals to compare male and female head sizes, holding constant the effects of locality and sampling period, we could do the following:

```
> sizeDF <- data.frame(sex = c("Female", "Male"))
> rownames(sizeDF) <- c("Female", "Male")
> sizePreds <- predict(fit, sizeDF)
> plot(sizePreds)
```

The plots (e.g., Figure 1) are perfectly amenable (e.g., point type and colour, line thickness, alternative labels, and additional text can be added or adjusted with typical `par` arguments). Supporting Information Appendix S3 also demonstrates how to use different types of *SS*.

3.2 | Example: Pupfish body shape and high-dimensional data

In the second example, we highlight the `RRPP` ability to efficiently handle large data computations. For this demonstration, a

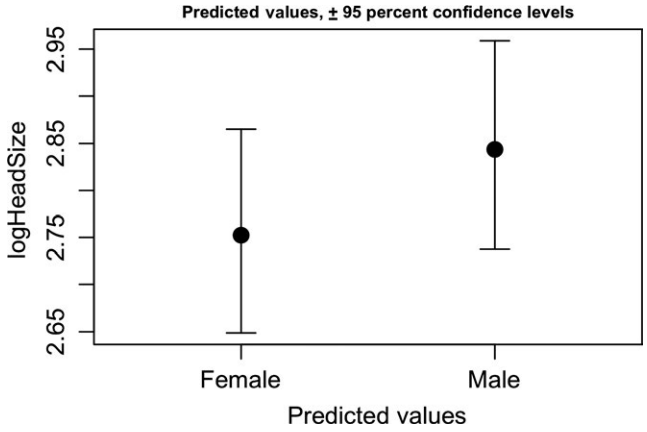

FIGURE 1 Example plot with the `predict` S3 generic function. All S3 generic examples use the `lm.rrpp` suffix. The example data, `PupfishHeads`, and the example, `lm.rrpp` fit were used for this plot

54(n) × 112(p) matrix of Procrustes residuals are the data. In every one of the 1,000 random permutations, RRPP shuffles residual vectors the same way for four different null models, estimates coefficients for four different full models, estimates the *SS* as the difference between residual *SS* (*RSS*) for four null-full model comparisons, and calculates the total *SS*, before calculating *MS*, R^2, *F*, Cohen's f^2, and Euclidean distances of coefficient vectors across all 1,000 permutations. This process, plus packaging of results, took approximately 0.5 seconds on a notebook computer, without any parallel processing.

In the second example, the initial steps are quite the same as the first example:

```
> data(Pupfish)
> Pupfish$logSize <- log(Pupfish$CS)
> fit <- lm.rrpp(coords ~ logSize + Sex*Pop,
  SS.type = "I", data = Pupfish)
> summary(fit, formula = FALSE)
> anova(fit)
> coef(fit, test = TRUE)
```

ANOVA results (Supporting Information Appendix S2) reveal that after accounting for body size allometry, not only are there significant interpopulation differences in body shape and sexual dimorphism in body shape, but sexual dimorphism also significantly varies between the two populations. A fuller evaluation of these results and why parametric M-ANOVA is not possible is provided in Collyer et al. (2015). In Supporting Information Appendix S2, we also provide additional script for attempting a parametric M-ANOVA and demonstrate the inability to perform such an analysis with high-dimensional data.

We can look at the precision of group mean estimation, accounting for allometric shape variation, by doing the following:

```
> shapeDF <- expand.grid(Sex = levels(Pupfish$Sex), Pop =
  levels(Pupfish$Pop))
> rownames(shapeDF) <- paste(shapeDF$Sex, shapeDF$Pop, sep = ".")
> shapePreds <- predict(fit, shapeDF, confidence = 0.95)
> plot(shapePreds, PC = TRUE, ellipse = TRUE)
```

This produces a plot (Figure 2) that puts nonparametric confidence ellipses on the least-squares means for the four groups analysed (two populations by two sexes). This can be compared to a principal component plot performed on the covariance matrix of the model fitted values with the following:

```
> plot(fit, type = "PC")
```

These plots differ as there is a rotational difference between the covariance matrices estimated with 4 predicted and 54 fitted values. Additionally, the former illustrates prediction precision and the latter sample dispersion. Additional plotting examples are provided in the examples in the `lm.rrpp` help page and Supporting Information Appendix S2. Both functions allow passing `par` arguments to the plot as well as saving plot data for more advanced plotting.

The function, `pairwise`, can be used to test pairwise differences between least-squares means with:

```
> PWT <- pairwise(fit, groups = interaction(Pupfish$Sex,
Pupfish$Pop))
> summary(PWT, confidence = 0.95)
```

Much like the `tukeyHSD` function in the `R stats` package (R Core Team, 2018), `pairwise` will generate tables with confidence intervals and *p*-values for the pairwise statistic, Euclidean distance between least-squares means. This function could also be used for pairwise comparison of slopes in analysis of covariance (ANCOVA) designs, as the help page and Supporting Information Appendix S2 demonstrate.

Because the Procrustes residuals are projected into a Euclidean tangent space (see `geomorph` function, `gpagen`; Adams et al., 2017), this analysis could be performed with an object of class `dist` (values from lower half of a distance matrix), representing the interspecimen shape (Euclidean) distances (as described in the Introduction), using the following code:

```
> D <- dist(Pupfish$coords) # inter-observation Euclidean distances
> Pupfish$D <- D
> fitD <- lm.rrpp(D ~ logSize + Sex*Pop, SS.type = "I", data =
Pupfish)
> anova(fitD)
> anova(fit)
```

The ANOVA results with either method are exactly the same (Supporting Information Appendix S2).

3.3 | Example: Plethodontid morphology, phylogenetics, and GLS estimation

In the third example, we highlight GLS estimation. The following code creates two `lm.rrpp` fits using OLS and GLS, respectively, and evaluates them as in previous examples:

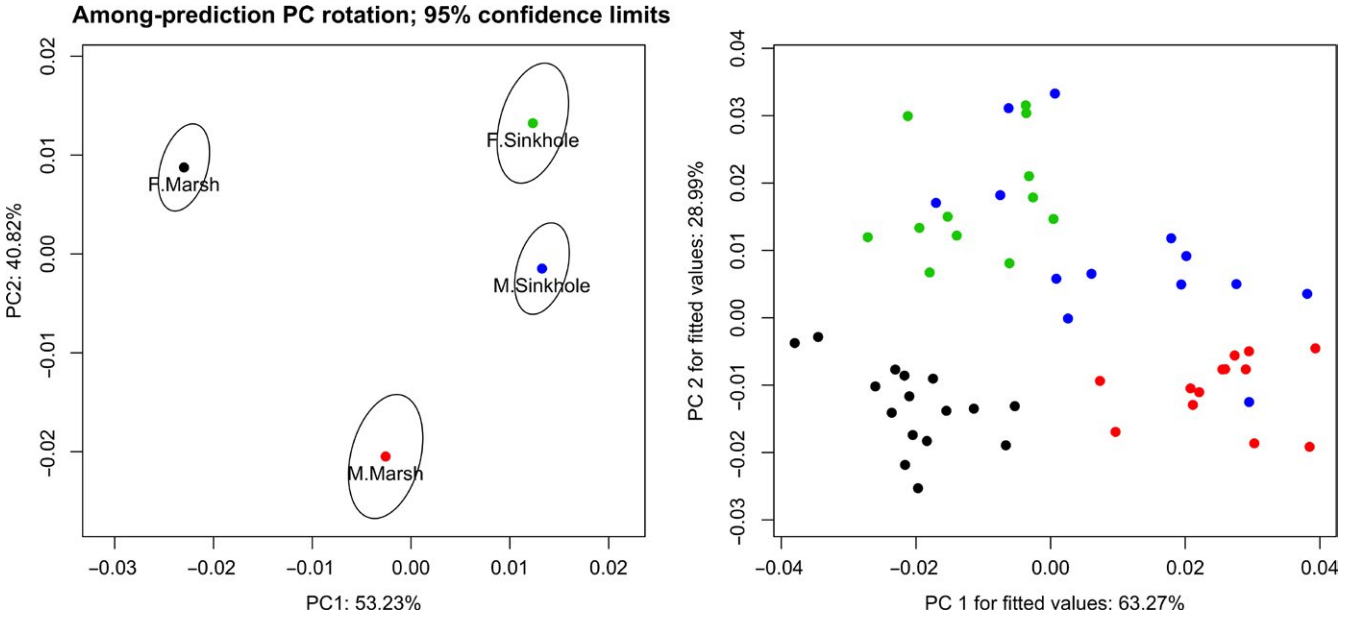

FIGURE 2 A PC plot with confidence ellipses for predicted values (left) and a PC plot with all data projected, based on the covariance matrix of fitted values (right). The data, `Pupfish`, were used for these examples

```
> data(PlethMorph)
> fitOLS <- lm.rrpp(TailLength ~ SVL, data = PlethMorph)
> fitGLS <- lm.rrpp(TailLength ~ SVL, data = PlethMorph, Cov =
  PlethMorph$PhyCov)
> anova(fitOLS)
> anova(fitGLS)
> coef(fitOLS, test = TRUE)
> coef(fitGLS, test = TRUE)
```

Although analyses on either model fit indicate a significant relationship between tail length and snout-to-vent length (SVL), the GLS coefficients test and ANOVA show how phylogenetic autocorrelation among species augments the OLS-estimated relationship (Supporting Information Appendix S2). The `lm.rrpp` help page and Supporting Information Appendix S2 has further illustration for a multivariate example, plus plotting of the results. Parallel processing is also available (for Unix systems) for any `lm.rrpp` analysis.

3.4 | Analytical summary

On the surface, these three examples and their analyses should seem intuitive to any user of R who has used the `lm` function plus its associated S3 generics (`coef`, `predict`, `resid`, `fitted`, `summary`, and `anova`), all of which can be used on `lm.rrpp` model fits (Table 1). The functions, `pairwise` (not an S3 generic) and `anova`, also allow pairwise comparisons of least-squares means or slopes and multimodel inferences, respectively. Advanced users will recognize, however, much more extensive useable results for adaptive programming. The output from a `lm.rrpp` fit is arranged hierarchically as follows:

```
> attributes(fit)
$names
[1] "call"     "LM"       "ANOVA"    "PermInfo"
$class
[1] "lm.rrpp"
```

Within the `$LM` partition, all attributes of the `lm` function are found, in addition to coefficients for every random permutation. Within the `$ANOVA` partition, the *SS* type, plus *SS*, *MS*, R^2, *F*, and Cohen's f^2 for all permutations, as well as effect sizes estimated for each of these are provided. Within the `$PermInfo` partition, the number of permutations, type (RRPP or randomization of "full" data values, FRPP), and sampling frame in every permutation (schedule) are provided. Thus, `lm.rrpp` is the workhorse that makes all downstream analysis efficient.

4 | RRPP COMPARED TO SIMILAR R PACKAGES

The `lm.rrpp` and `anova.lm.rrpp` have some overlap in purpose of high-dimensional data analysis with the `adonis` and `adonis2` functions of the `vegan` package (Oksanen et al., 2017) and the `manylm` and `anova.manylm` functions of the `mvabund` package (Wang, Naumann, Wright, Eddelbuettel, & Warton, 2017; Wang, Naumann, Wright, & Warton, 2012). All of these packages use resampling procedures to perform ANOVA on data of any dimension (although `adonis` and `adonis2` require estimating dissimilarity matrices first). `RRPP` is unique, however, in offering GLS estimation, pairwise comparison tools, choice of three *SS* types, and the capability to work with either data

or distance matrices. In Appendix S3 of the Supporting Information, we provide a comprehensive comparison of the three packages in terms of results provided and computation efficiency. For cases that functions overlap in purpose, they all provided similar results, appropriate type I error rates, and similar statistical power (Supporting Information Figure S2). However, `lm.rrpp` is much more flexible with options (Supporting Information Table S1), produces more results as output, and has much faster computation speeds (Supporting Information Figure S1). The `lm.rrpp` function accomplishes all objectives of the `adonis` and `adonis2` functions but differs from `manylm` chiefly with the method of addressing singular error covariance matrices (as a result of $p > n$). The `manylm` and `anova.manylm` functions can optionally use ridge regularization (Warton, 2008) for obtaining appropriately ranked covariance matrices in a penalized-likelihood framework. The `lm.rrpp` function avoids determinants of singular error covariance matrices by relying on statistics based on traces (Adams & Collyer, 2018b). These alternative approaches produce remarkably similar results for our limited comparison (see Appendix S3 and associated R scripts in the Supporting Information).

5 | CURRENT LIMITATIONS

The `RRPP` package works with continuous data or distance matrices (that can be based on frequency or binary data). It currently does not extend to generalized linear models, like the `manyglm` function of `mvabund`. A method of RRPP for generalized linear models has not yet been developed, so its application is restricted to general (not generalized) linear models. The functions of `RRPP` are based on least-squares estimation of coefficients. Although mixed-model ANOVA is possible (via error specification in the `anova` function), multivariate random effects use expected mean squares (EMS) estimation, which is less ideal than restricted maximum likelihood (REML), if model designs are imbalanced. The purpose of this package is to work with high-dimensional data, for which REML is currently not possible (because it would require inverting singular covariance matrices). Users should exercise caution and investigate multiple approaches for designs that are both high-dimensional and highly imbalanced. Additionally, although a covariance matrix input is an option for GLS estimation of coefficients, currently only one matrix can be input and we assume no philosophy for multiple sources of nonindependence among observations (e.g., phylogenetic history and repeated measures). However, a vector of weights can be passed onto `lm.rrpp`, as with `lm`. Doing this has the effect of reweighting the covariance matrix, as **CW**, where **W** is a diagonal matrix of the square root of the weights. If **W** and **C** are inherently related (maybe because both are estimated, empirically), one might wish to guard against including both in the model set up, preferentially opting to calculate a new covariance matrix to use in `lm.rrpp`; e.g., **V** = **CW**.

ACKNOWLEDGEMENTS

This R package was developed with funding from NSF DEB Awards 1737895 (M.L.C.) and 1556379 (D.C.A.). We thank M. C. Gilbert for example data. We thank E. Baken and B. Juarez for testing RRPP functions on novel data, and helping us to refine functions, and we thank T. Turner (curator of fishes) and A. Snyder (collections manager) at the Museum of Southwestern Biology for assistance with pupfish collections. This manuscript benefitted from the conscientious reviews of two anonymous reviewers on an earlier version. The authors have no conflicts of interest.

AUTHORS' CONTRIBUTIONS

M.L.C. conceived and developed this R package based on recent research and scripts designed by D.C.A. and M.L.C. Both M.L.C. and D.C.A. contributed data examples, developed this article, and tested `RRPP` functions.

DATA AND PACKAGE ACCESSIBILITY

Documentation and source code are freely available on CRAN (https://cran.r-project.org/web/packages/RRPP) and GITHUB (https://github.com/mlcollyer/RRPP). For updates, we recommend direct installation using the R package, `devtools` (Wickham & Chang, 2017):

```
> devtools::install_github("mlcollyer/RRPP")
```

ORCID

Michael L. Collyer http://orcid.org/0000-0003-0238-2201

REFERENCES

Adams, D. C. (2014). A method for assessing phylogenetic least-squares models for shape and other high-dimensional multivariate data. *Evolution*, *68*, 2675–2688. https://doi.org/10.1111/evo.12463

Adams, D. C., & Collyer, M. L. (2015). Permutation tests for phylogenetic comparative analyses of high-dimensional shape data: What you shuffle matters. *Evolution*, *69*, 823–829. https://doi.org/10.1111/evo.12596

Adams, D. C., & Collyer, M. L. (2018a). Multivariate phylogenetic comparative methods: Evalutions, comparisons, and recommendations. *Systematic Biology*, *67*, 14–31. https://doi.org/10.1093/sysbio/syx055

Adams, D. C., & Collyer, M. L. (2018b). Phylogenetic ANOVA: Group-clade aggregation, biological challenges, and a refined permutation procedure. *Evolution*, https://doi.org/10.1111/evo.13492. in press.

Adams, D. C., Collyer, M. L., Kaliontzopoulou, A., & Sherratt, E. (2017). Geomorph: Software for geometric morphometric analyses. R package version 3.0.6. http://CRAN.R-project.org/package=geomorph.

Adams, D. C., & Felice, R. (2014). Assessing phylogenetic morphological integration and trait covariation in morphometric data using evolutionary covariance matrices. *PLoS ONE*, *9*(4), e94335. https://doi.org/10.1371/journal.pone.0094335

Adams, D. C., Rohlf, F. J., & Slice, D. E. (2013). A field comes of age: Geometric morphometrics in the 21st century. *Hystrix*, *24*, 7–14.

Anderson, M. J. (2001a). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, *26*, 32–46.

Anderson, M. J. (2001b). Permutation tests for univariate or multivariate analysis of variance and regression. *Canadian Journal of Fisheries and Aquatic Sciences*, *58*(3), 626–639. https://doi.org/10.1139/f01-004

Anderson, M. J., & ter Braak, C. J. F. (2003). Permutation tests for multifactorial analysis of variance. *Journal of Statistical Computation and Simulation*, *73*, 85–113. https://doi.org/10.1080/00949650215733

Bookstein, F. L. (1991). *Morphometric tools for landmark data: Geometry and biology*. Cambridge: Cambridge University Press.

Collyer, M. L., Sekora, D. J., & Adams, D. C. (2015). A method for analysis of phenotypic change for phenotypes described by high-dimensional data. *Heredity*, *115*, 357–365. https://doi.org/10.1038/hdy.2014.75

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression*. Thousand Oaks, CA: Sage Publications.

Freedman, D., & Lane, D. (1983). A nonstochastic interpretation of reported significance levels. *Journal of Business and Economic Statistics*, *1*, 292–298.

Gilbert, M. C. (2016). *Impacts of habitat fragmentation on the cranial morphology of a threatened desert fish (Cyprinodon pecosensis)*. Masters thesis, Western Kentucky University, Bowling Green, KY, USA.

Gonzalez, L., & Manly, B. F. J. (1998). Analysis of variance by randomization with small data sets. *Environmetrics*, *9*, 53–65. https://doi.org/10.1002/(ISSN)1099-095X

Legendre, P., & Legendre, L. F. (2012). *Numerical ecology*, Vol. 24. Amsterdam: Elsevier.

McArdle, B. H., & Anderson, M. J. (2001). Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology*, *82*, 290–297. https://doi.org/10.1890/0012-9658(2001)082[0290:FMMTCD]2.0.CO;2

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., … Wagner, H. (2017). VEGAN: Community ecology package. Retrieved from https://cran.r-project.org/package=vegan

Pinheiro, J., Bates, D., DebRoy, S., & Sarkar, D., & R Core Team. (2018). _NLME: Linear and nonlinear mixed effects models. Retrieved from https://cran.r-project.org/package=nlme

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://cran.r-project.org/

Wang, Y., Naumann, U., Wright, S., Eddelbuettel, D., & Warton, D. (2017). `mvabund`: Statistical methods for analysing multivariate abundance data. Retrieved from https://cran.r-project.org/package=mvabund

Wang, Y. I., Naumann, U., Wright, S. T., & Warton, D. I. (2012). `mvabund` – An R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution*, *3*, 471–474. https://doi.org/10.1111/j.2041-210X.2012.00190.x

Warton, D. I. (2008). Penalized normal likelihood and ridge regularization of correlation and covariance matrices. *Journal of the American Statistical Association*, *103*(481), 340–349. https://doi.org/10.1198/016214508000000021

Warton, D. I., Wright, S. T., & Wang, Y. (2012). Distance-based multivariate analyses confound location and dispersion effects. *Methods in Ecology and Evolution*, *3*(1), 89–101. https://doi.org/10.1111/j.2041-210X.2011.00127.x

Wickham, H., & Chang, W. (2017). `devtools`: Tools to make developing R packages easier. Retrieved from https://cran.r-project.org/package=devtools

Zapala, M. A., & Schork, N. J. (2006). Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and relative variables. *Proceedings of the National Academy of Sciences, U.S.A.*, *103*, 19430–19435. https://doi.org/10.1073/pnas.0609333103

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Collyer ML, Adams DC. `RRPP`: An R package for fitting linear models to high-dimensional data using residual randomization. *Methods Ecol Evol*. 2018;9:1772–1779. https://doi.org/10.1111/2041-210X.13029