The Journal of Open Source Software

MAGPA: A R package for multivariate analysis of genotype–phenotype association and visualization of 3D image

Yin Huang[1]

1 CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, University of Chinese Academy of Sciences

DOI: 10.21105/joss.02012

Software

- Review
- Repository
- Archive

Editor: Pending Editor

Submitted: 13 January 2020
Published: 13 January 2020

Summary

The **MAGPA** (multivariate analysis of genotype–phenotype association) is a package of multivariate correlation analysis and an interactive visualization tool for 3D image. This package was implemented for genetic association analysis of facial phenotypes and visualization related results. In addition, it can also be used in genome-wide association analysis of other multivariate phenotypes, especially three-dimensional image data. It can not only receive the prepared features, but also preprocess the features with principal component analysis and automatically select the number of variables. The genotype should be a SnpMatrix, which is a special object holding large arrays of single nucleotide polymorphism (SNP). Then canonical correlation analysis (CCA)(Benesty & Cohen, 2018) is used to extract the linear combination of variables to maximize the correlation with each SNP. For the interactive visualization, the function *visual3d* is required to provide at least a reference of 3D image object and a vector, such as the phenotypic changes under different genotypes. It can draw a 3D object with different style and gradient colors.

Statistical Methodology

CCA is a multivariate statistical method that reflects the overall correlation between two groups of variables by using the correlation between the pairs of comprehensive indicators, which is implemented in PLINK and has demonstrated its advantages in multivariable analysis of genotype-phenotype association. Here, briefly, X is the sample phenotypic matrix, the principle components of facial variations from each segment, and Y is the genotype of the sample. (X,Y) is the canonical correlation (Formula 1).

$$\rho(X,Y) = \frac{cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} \quad (1)$$

The target function of CCA is to maximize the (X ,Y) by optimizing the corresponding projection vector and , called the canonical correlation coefficients between X and Y, respectively (Formula 2).

$$\underbrace{arg\ max}_{a,b} \frac{cov(X',Y')}{\sqrt{D(X')}\sqrt{D(Y')}}, \quad X' = a^T X, Y' = b^T Y \quad (2)$$

Examples

In this example, the genotype data and phenotype data are used to demonstrate how to use the function *magpa* and show the one of the possible input to the function. The variable `geno` is a list with a SnpMatrix `genotypes` (2000 rows, 50 columns) and a data frame `map`. The variable `pheno` is a matrix (2000 rows, 300 columns) expanded by 2000 samples with 100 three-dimensional coordinate in each sample.

If your phenotype is pre-prepared, you can follow the step 3. Otherwise, you can follow the step 1, which set `pca` argument to `TRUE`. Or you can follow the step 2 to select features, and then follow the step 3.

1.Processing phenotype and multivariate association

The function *magpa* is performed multivariate analysis of genotype–phenotype association based on CCA, which calls the function "cca" to carry out the canonical correlation analysis and the function "F.test.cca" to test the statistical significance by employing Rao's statistic from the R package **yacca**(Butts, 2009). The first argument of *magpa* can be a file prefix name of plink(Purcell et al., 2007) output (.bed, .bim, .fam), an object read by *read.plink* of the **snpStats**(Solé, Guinó, Valls, Iniesta, & Moreno, 2006) package, or a SnpMatrix. The second argument is the phenotypic matrix (rows are the number of samples, and columns are the number of features).

```
data(geno);data(pheno)
gpa <- magpa(geno,pheno,pca = TRUE)
head(gpa)
```

2.Automatic extraction of the principal components

The function *pcapheno* is implemented to automatically extract the principal components from high-dimension data, which calls **paran** to performs Horn's parallel analysis(Dinno, 2009) for evaluating the components retained in a principle component analysis.

```
paral<- pcapheno(pheno)
new_pheno<- paral$pheno
head(new_pheno)
```

3.Multivariate association with prepared phenotype

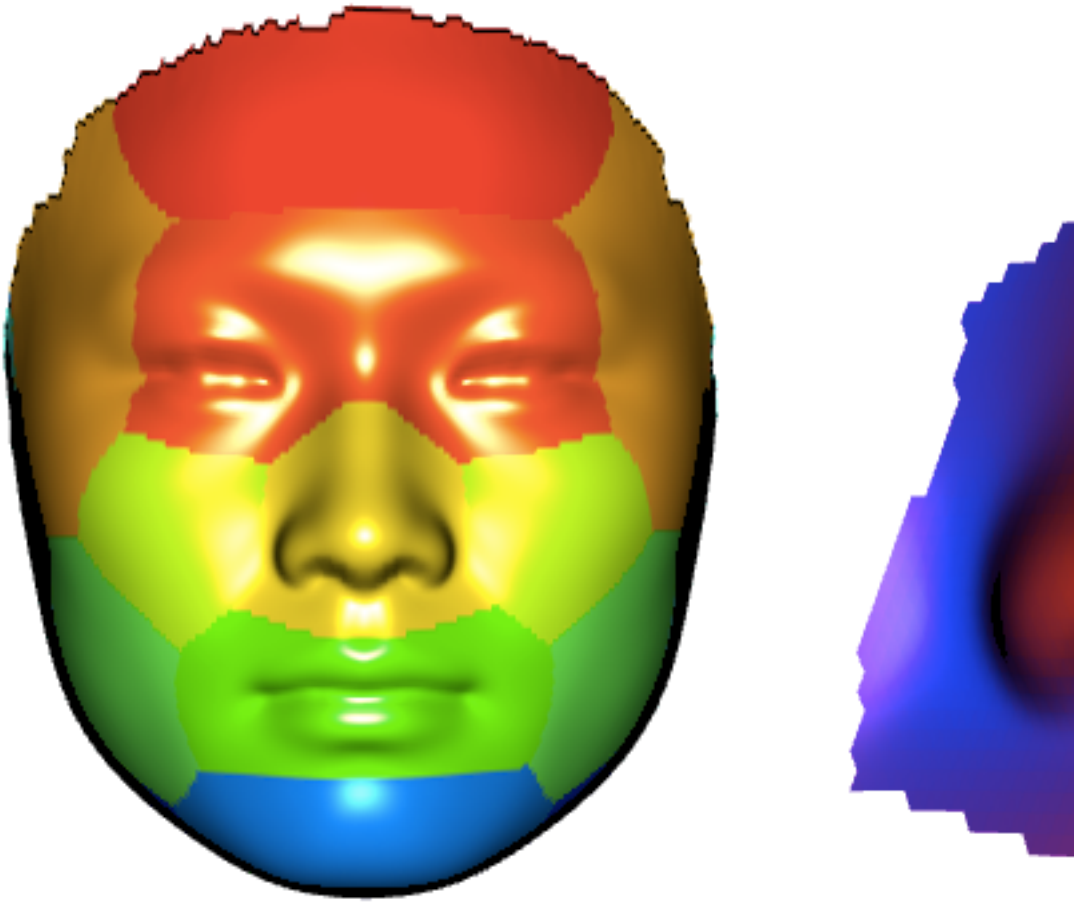
The result in a matrix included SNP, CHR, position, MAF, canonical correlation, chisq, and pvalue. It can be used to draw QQ plot and Manhattan plot.

```
gpa<- magpa(geno,new_pheno)
head(gpa)
```

Visualization for 3D Face

The function *visual3d* is an interactive graphing function based on the **rgl**(Adler, Nenadic, & Zucchini, 2003) package. This function is very flexible. The first argument is a reference of 3D image object, whose file (with `.obj` suffix) can be read by the *readobj* function. The

second argument can be a vector or list. And the thrid argument is a vector of colors or a color palette, if it is not given, the default color palette will be used. Two examples of using *visual3d* are as follows.You can see more in the help documentation.



References

Adler, D., Nenadic, O., & Zucchini, W. (2003). RGL: A R-library for 3D visualization with OpenGL. *Proceedings of the 35th Symposium of the Interface: Computing Science and Statistics*, 1–11.

Benesty, J., & Cohen, I. (2018). Canonical Correlation Analysis. In *Applied multivariate statistical analysis* (pp. 5–14). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-319-67020-1_2

Butts, C. T. (2009). Yet Another Canonical Correlation Analysis Package. Retrieved from https://rdrr.io/cran/yacca/man/yacca-package.html

Dinno, A. (2009). Implementing horn's parallel analysis for principal component analysis and factor analysis. *Stata Journal*, *9*(2), 291–298. doi:10.1177/1536867x0900900207

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., et al. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. doi:10.1086/519795

Solé, X., Guinó, E., Valls, J., Iniesta, R., & Moreno, V. (2006). SNPStats: A web tool for the analysis of association studies. *Bioinformatics*, *22*(15), 1928–1929. doi:10.1093/bioinformatics/btl268