Toward Tidy Principles for Matrix-Decomposed Data

New Developments in Graphing Multivariate Data
Section on Statistical Graphics
Joint Statistical Meetings 2022

Jason Cory Brunson

Laboratory for Systems Medicine
Division of Pulmonary, Critcial Care, and Sleep Medicine
University of Florida

July 28, 2022

Acknowledgments

Development

- Emily Paul (UPenn)
- Joyce Robbins (Columbia)

Experiment

- Tom Agresta (UConn)
- Ritchie Vaughan (UVA)
- Martinna Bertolini (UFRJ)
- Carol Mathews (UF)

Support

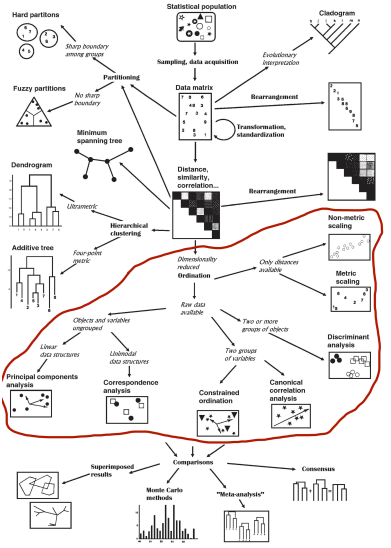
UCONN HEALTH

UF UNIVERSITY of FLORIDA

Ordination

"[A]ny technique that extracts artificial variables in order to reduce the dimensionality of the data is referred to as ***ordination****."*

		model: unsupervised	model: supervised
data	discrete	clustering	classification
data	continuous	dimension reduction	regression



Podani J • 2000 • *Introduction to the Exploration of Multivariate Biological Data* • https://web.archive.org/web/20200221000313/http://ramet.elte.hu/~podani/books.html

Principal components analysis

Derivation

- $X_{n\times p}$ data
- $\overline{x}_{n\times 1}$ data centroid
- $Y = X - 1\overline{x}^\top$ centered data
- $Y = U_{n\times q} D_{q\times q} V^\top_{q\times p}$ singular value decomposition

Interpretation

- D inertia
- U_r, V_r standard coordinates (orthonormal)
- $U_r D_r$, $V_r D_r$ principal coordinates
- V_r variable loadings
- $U_r D_r$ case scores

Application

- X' new data
- $(X' - 1\overline{x}^\top)V_r$ scores (supplementary)

Linear discriminant analysis

Derivation

- $G_{n\times k}$ groups
- $N = \text{diag}(n_1, \ldots, n_k)$ group counts
- $\overline{X}_{k\times p} = N^{-1}G^\top X$ group centroids
- $C = \frac{1}{n}X^\top X$ covariance matrix
- $\overline{Y} = \overline{X} - 1\overline{x}^\top$ centered group centroids
- $\overline{Y}C^{-1/2} = U_{k\times q}D_{q\times q}V^\top_{q\times p}$

Interpretation

- V_r variable loadings
- $U_rD_r = \overline{Y}C^{-1/2}V$ group centroid scores

Application

- $YC^{-1/2}V_r$ case scores (supplementary)
- X' new data
- $X'C^{-1/2}V_r$ scores (supplementary)

General Multidimensional Analysis

1. Preprocess data $X \rightsquigarrow Y$
   - centering
   - double-centering
2. Generalized SVD $Y = NDM^\top = (A^{-1/2}U)D(B^{-1/2}V)^\top$, where A, B are positive semi-definite and $N^\top AN = M^\top BM = I$ (orthonormalization)
   - weights
   - sphering

   Low-rank approximation $Y \approx N_r D M_r^\top$
3. Biplot of $F = U_r D^a$ and $G = V_r D^b$, with $a + b = 1$
   - row-principal
   - column-principal
   - symmetric

Greenacre M • 1984 • *Theory and applications of correspondence analysis* • http://www.carme-n.org/

Need

R is replete with ordination methods!

CRAN Task View: Multivariate Statistics

Maintainer: Paul Hewson
Contact: Paul.Hewson at plymouth.ac.uk
Version: 2014-09-19

CRAN Task View: Analysis of Ecological and Environmental Data

Maintainer: Gavin Simpson
Contact: ucfagls at gmail.com
Version: 2014-05-31

… but they are

- **specialized**: unweildy & uninformative inspection methods
- **heterogeneous**: diverse, dissimilar, domain-specific conventions
- **standalone**: not easily interoperable with other tools or integrable into external workflows

Design

Typical implementations:

General implementation:

Tidy management:

Inspiration

*[T]he **tidyverse** is a collection of R packages that share a high-level design philosophy and low-level grammar and data structures, so that learning one package makes it easier to learn the next.*

The tidyverse strives to be

- **human-centered**: supports data analysis conducted by humans
- **consistent**: ensures learning transfers between packages
- **composable**: enables modular thinking and doing
- **inclusive**: developed and informed by a broad community

Wickham &al • 2019 • *The Journal of Open Source Software* • doi:10.21105/joss.01686
Tidyverse team • *Tidyverse design guide* • https://design.tidyverse.org/

Inspiration

lazy, surly, & pithy data frames

relational algebra for data sets

convenient summarization
of statistical models

grammatical production
of statistical graphics

Implementation

Engine

Recovery methods for (your!) S3 model classes:

- left & right matrix factors (singular vectors)
  $U_{n\times k}$, $V_{p\times k}$
- transformations of coordinate spaces
  $A_{n\times n}$, $B_{p\times p}$
- inertia and its distribution unto the factors
  $D = \text{diag}(d_1, \ldots, d_k)$, (a, b)
- active & supplementary elements
  $U_r D_r = XV_r$, $X'V_r$

Dashboard

Class 'tbl_ord':

- wrapper for ordination models
- clear & consistent formatting

Functions:

- augment with model metadata
- redistribution of inertia
- tidily inspect & summarize
- annotate rows and columns
- build biplots grammatically
- add ordination plot layers

Example workflow

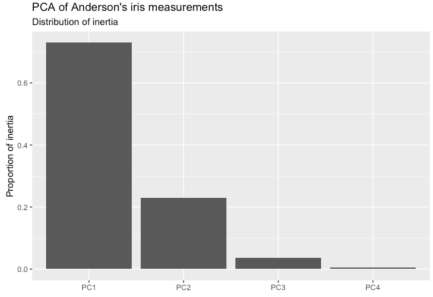
```
head(iris)
#>   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
#> 1          5.1         3.5          1.4         0.2  setosa
#> 2          4.9         3.0          1.4         0.2  setosa
#> 3          4.7         3.2          1.3         0.2  setosa
#> 4          4.6         3.1          1.5         0.2  setosa
#> 5          5.0         3.6          1.4         0.2  setosa
#> 6          5.4         3.9          1.7         0.4  setosa
summary(iris)
#>   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
#>  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
#>  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
#>  Median :5.800   Median :3.000   Median :4.350   Median :1.300
#>  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
#>  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
#>  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
#>        Species
#>  setosa    :50
#>  versicolor:50
#>  virginica :50
#>
#>
#>
```

```
(iris_pca <- ordinate(iris, cols = 1:4, model = ~ prcomp(., scale. = TRUE)))
#> # A tbl_ord of class 'prcomp': (150 x 4) x (4 x 4)'
#> # 4 coordinates: PC1, PC2, ..., PC4
#> #
#> # Rows (principal): [ 150 x 4 | 1 ]
#>     PC1    PC2     PC3 ... |   Species
#>                            |   <fct>
#> 1 -2.26 -0.478  0.127      | 1 setosa
#> 2 -2.07  0.672  0.234  ... | 2 setosa
#> 3 -2.36  0.341 -0.0441     | 3 setosa
#> 4 -2.29  0.595 -0.0910     | 4 setosa
#> 5 -2.38 -0.645 -0.0157     | 5 setosa
#> # … with 145 more rows
#> # ℹ Use `print(n = ...)` to see more rows
#> #
#> # Columns (standard): [ 4 x 4 | 3 ]
#>      PC1     PC2     PC3 ... |   .name        .center .scale
#>                              |   <chr>         <dbl>  <dbl>
#> 1  0.521 -0.377   0.720     | 1 Sepal.Length   5.84  0.828
#> 2 -0.269 -0.923  -0.244 ... | 2 Sepal.Width    3.06  0.436
#> 3  0.580 -0.0245 -0.142     | 3 Petal.Length   3.76  1.77
#> 4  0.565 -0.0669 -0.634     | 4 Petal.Width    1.20  0.762
```

Example workflow

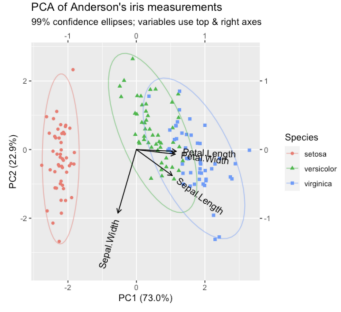
```
iris_meta <- data.frame(
  Species = c("setosa", "versicolor", "virginica"),
  Colony = c(1L, 1L, 2L),
  Cytotype = c("diploid", "hexaploid", "tetraploid"),
  Ploidy = c(2L, 6L, 4L)
)
(iris_pca <- left_join_rows(iris_pca, iris_meta, by = "Species"))
#> # A tbl_ord of class 'prcomp': (150 x 4) x (4 x 4)'
#> # 4 coordinates: PC1, PC2, ..., PC4
#> #
#> # Rows (principal): [ 150 x 4 | 4 ]
#>      PC1    PC2     PC3 ... |   Species Colony Cytotype Ploidy
#>                             |   <chr>    <int> <chr>     <int>
#> 1 -2.26 -0.478  0.127      | 1 setosa       1 diploid       2
#> 2 -2.07  0.672  0.234  ... | 2 setosa       1 diploid       2
#> 3 -2.36  0.341 -0.0441     | 3 setosa       1 diploid       2
#> 4 -2.29  0.595 -0.0910     | 4 setosa       1 diploid       2
#> 5 -2.38 -0.645 -0.0157     | 5 setosa       1 diploid       2
#> # … with 145 more rows
#> # ℹ Use `print(n = ...)` to see more rows
#> #
#> # Columns (standard): [ 4 x 4 | 3 ]
#>      PC1     PC2     PC3 ... |   .name        .center .scale
#>                              |   <chr>          <dbl>  <dbl>
#> 1  0.521 -0.377   0.720     | 1 Sepal.Length    5.84  0.828
#> 2 -0.269 -0.923  -0.244 ... | 2 Sepal.Width     3.06  0.436
#> 3  0.580 -0.0245 -0.142     | 3 Petal.Length    3.76  1.77
#> 4  0.565 -0.0669 -0.634     | 4 Petal.Width     1.20  0.762
```

```
tidy(iris_pca) %T>% print() %>%
  ggplot(aes(x = .name, y = .prop_var)) +
  geom_col() +
  labs(x = "", y = "Proportion of inertia") +
  ggtitle("PCA of Anderson's iris measurements",
          "Distribution of inertia")
#> # A tibble: 4 × 4
#>   .name .sdev .inertia .prop_var
#>   <fct> <dbl>    <dbl>     <dbl>
#> 1 PC1   1.71     435.      0.730
#> 2 PC2   0.956    136.      0.229
#> 3 PC3   0.383     21.9     0.0367
#> 4 PC4   0.144      3.09    0.00518
```

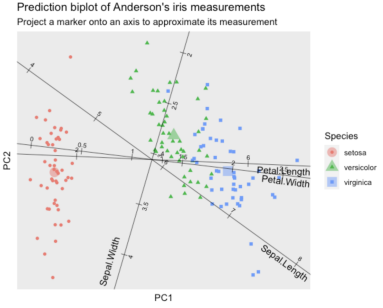


Example workflow

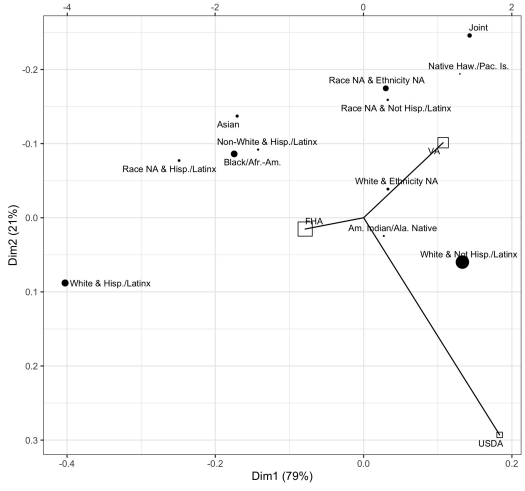
```
ggbiplot(iris_pca, sec.axes = "cols", scale.factor = 2) +
  geom_rows_point(aes(color = Species, shape = Species)) +
  stat_rows_ellipse(aes(color = Species), alpha = .5, level = .99) +
  geom_cols_vector() +
  geom_cols_text_radiate(aes(label = .name)) +
  expand_limits(y = c(-3.5, NA)) +
  ggtitle("PCA of Anderson's iris measurements",
          "99% confidence ellipses; variables use top & right axes")
```



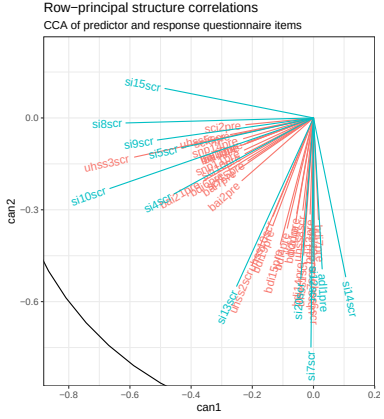

```
ggbiplot(iris_pca, axis.type = "predictive", axis.percents = FALSE) +
  theme_biplot() +
  geom_rows_point(aes(color = Species, shape = Species)) +
  stat_rows_center(
    aes(color = Species, shape = Species),
    size = 5, alpha = .5, fun.data = mean_se
  ) +
  geom_cols_axis(aes(label = .name, center = .center, scale = .scale)) +
  ggtitle("Prediction biplot of Anderson's iris measurements",
          "Project a marker onto an axis to approximate its measurement")
```

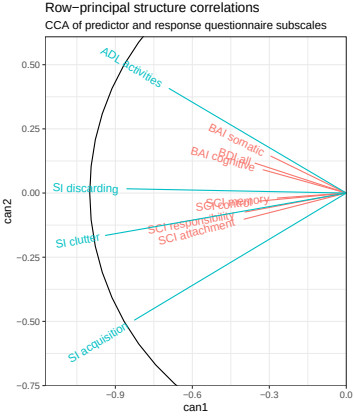


Origination of home loans by program and racial-ethnic group



Associations between hoarding and other mental health disorders





Use case 3

Limitations & needs

S3 class methods
- quality measures
- interpolation
- prediction
- predictive biplot elements

Biplot functionality
- predictive biplots
- joint row-and-column layers
  - interpolative vector sum
  - predictive projection

Involvement
- accessibility
- issues
- contributions

Fin

This is the end
Beautiful friend