Multivariate Miscellany

Lecture 11.1 Dimension Reduction Applications

Lauren Sullivan

Module: Multivariate Models

Readings

Required for class:

- NA

Optional:

- Groves, A. M., Bauer, J. T., and Brudvig, L. A. (2020) Lasting signature of planting year weather on restored grasslands. *Scientific Reports*.

Dimension Reduction and its Application

We have talked in depth about how multivariate ordination can be a great way to reduce dimensions for multivariate, correlated data. But we have not yet put into practice this idea.

We are going to focus on how to reduce dimensions and then test hypotheses with these reduced dimensions.

We will start with creating a reduced dimension X variable.

Data Example

How does a restoration site's age and planting-year weather influence its total cover of desireable (sown species) and non-desireable (non-sown species, or likely weeds) species?

SCIENTIFIC REPORTS

natureresearch

OPEN

Lasting signature of planting year weather on restored grasslands

Anna M. Groves[1,2,3*], Jonathan T. Bauer[1,4] & Lars A. Brudvig[1,2]

The Data

```
restoration[1:15, 1:5]
```

```
## # A tibble: 15 x 5
##    Site           Age_2016 Biomass Jun1.dd.accum Jun1.precip.accum
##    <chr>             <dbl>   <dbl>         <dbl>             <dbl>
##  1 A2                   14    690.          266.              316.
##  2 A302                 19    653.          417.              338.
##  3 A4                   19    699.          414.              329.
##  4 ARR                  14    547.          266.              318.
##  5 B1                   16    449.          372.              180.
##  6 B2                   16    391.          372.              180.
##  7 B3                   13    287.          301.              247.
##  8 B4                   13    350.          301.              247.
##  9 BoudemanKappy        12    491.          341.              397.
## 10 BoudemanMain         12    474.          334.              399.
## 11 Brookdale             6    408.          339.              255.
## 12 BruceWillis           5    593.          254.              354.
## 13 ButlerWest           10    604.          275.              257.
## 14 C1                    8    340.          233.              205.
## 15 C2                   10    488.          330.              341.
```

Reducing Climate Dimension

```r
rest.pca <- prcomp(restoration[, c(4:14)], center = TRUE, scale = TRUE,
                   na.rm = TRUE)
summary(rest.pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     2.0826 1.5515 1.1395 1.0788 0.87243 0.74723 0.54375
## Proportion of Variance 0.3943 0.2188 0.1180 0.1058 0.06919 0.05076 0.02688
## Cumulative Proportion  0.3943 0.6131 0.7312 0.8370 0.90618 0.95694 0.98382
##                            PC8     PC9    PC10    PC11
## Standard deviation     0.34052 0.23931 0.06721 0.01667
## Proportion of Variance 0.01054 0.00521 0.00041 0.00003
## Cumulative Proportion  0.99436 0.99956 0.99997 1.00000
```

Here, PC1 explains 39% of the variance, and PC2 explains 22% of the variance.

Reducing Climate Dimension

```r
rest.pca$rotation[,1:2]
```

```
##                                PC1         PC2
## Jun1.dd.accum           0.25090229  0.37308177
## Jun1.precip.accum      -0.18631915  0.28574897
## Sep1.dd.accum           0.44438229  0.22338160
## Sep1.precip.accum      -0.26079668  0.49864516
## summer.dd.accum         0.42627484  0.06806152
## summer.precip.accum    -0.21809796  0.43920722
## max.month.dd.accum      0.44047904  0.01500657
## max.mean.month.precip -0.27287180  0.25669759
## avg.low.temp            0.35993837  0.37267878
## avg.mon.rain.days      -0.09149120  0.21439946
## max.days.no.precip     -0.03403896 -0.17916581
```

It looks like degree day (dd) tends to be correlating more with PC1, and preciptitation tends to be correlating more with PC2.

Reducing Climate Dimension

Let's pull out our values for each plot for both PC1 and PC2.

```
clim <- rest.pca$x[, 1:2]
clim[1:15, ]
```

```
##              PC1         PC2
##  [1,]  1.8125414 -0.04942355
##  [2,]  2.0271115  3.57509201
##  [3,]  1.9969256  3.34571551
##  [4,]  1.8528529 -0.08810994
##  [5,]  1.9733218 -0.17462123
##  [6,]  1.9733218 -0.17462123
##  [7,] -1.5542528  1.74385183
##  [8,] -1.5542528  1.74385183
##  [9,] -2.1168497  1.90154808
## [10,] -2.3962185  1.53356989
## [11,]  0.6954783  1.82267167
## [12,]  0.6386134  0.05734789
## [13,] -0.2683007  0.23327476
## [14,] -0.6728123 -0.34482529
## [15,]  1.2460297  1.28352702
```

Reducing Climate Dimension

And `cbind()` our climate dimensions to our original dataset since they are in the same order.

```
rest.m <- cbind(restoration, clim)
rest.m[1:15, c(1:4, 46:47)]
```

```
##               Site Age_2016  Biomass Jun1.dd.accum        PC1         PC2
## 1               A2       14 689.5750      266.2307  1.8125414 -0.04942355
## 2             A302       19 652.8725      417.2026  2.0271115  3.57509201
## 3               A4       19 698.8575      413.5930  1.9969256  3.34571551
## 4              ARR       14 547.0475      266.3968  1.8528529 -0.08810994
## 5               B1       16 449.2875      372.1416  1.9733218 -0.17462123
## 6               B2       16 390.6600      372.1416  1.9733218 -0.17462123
## 7               B3       13 287.2500      301.0465 -1.5542528  1.74385183
## 8               B4       13 349.7050      301.0465 -1.5542528  1.74385183
## 9   BoudemanKappy        12 491.1700      340.6284 -2.1168497  1.90154808
## 10   BoudemanMain        12 474.2325      333.8917 -2.3962185  1.53356989
## 11      Brookdale         6 407.6700      339.4839  0.6954783  1.82267167
## 12    BruceWillis         5 593.0175      253.5168  0.6386134  0.05734789
## 13     ButlerWest        10 604.4075      275.2122 -0.2683007  0.23327476
## 14             C1         8 339.6300      233.1260 -0.6728123 -0.34482529
## 15             C2        10 487.9013      330.4094  1.2460297  1.28352702
```

Linear Models with Reduced Dimensional Data

Let's look at how the average cover of sown species (the desireable ones) is a function of restoration age, and the climate variables.

```
sown.cover.lm <- lm(log(Mean.Sown.Cover) ~ Age_2016 + PC1 + PC2,
                    data = rest.m, na.rm = TRUE)
Anova(sown.cover.lm,  type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: log(Mean.Sown.Cover)
##              Sum Sq Df F value    Pr(>F)
## (Intercept) 176.567  1 82.6744 6.417e-14 ***
## Age_2016      0.097  1  0.0452    0.8321
## PC1           1.115  1  0.5223    0.4720
## PC2           2.044  1  0.9572    0.3309
## Residuals   168.720 79
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Linear Models with Reduced Dimensional Data

Let's look at how the average cover of non-sown species (the weedy ones) is a function of restoration age, and the climate variables.

```
nonsown.cover.lm <- lm(log(Mean.Nonsown.Cover) ~ Age_2016 + PC1 + PC2,
                       data = rest.m)
Anova(nonsown.cover.lm,  type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: log(Mean.Nonsown.Cover)
##             Sum Sq Df  F value    Pr(>F)
## (Intercept) 93.091  1 115.0220 < 2.2e-16 ***
## Age_2016     5.737  1   7.0881  0.009399 **
## PC1          0.071  1   0.0875  0.768220
## PC2          0.385  1   0.4752  0.492641
## Residuals   63.937 79
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Linear Models with Reduced Dimensional Data

What happens if you make your PC1 variable categorical based on how different the years are from average? We will make an Average category where the climate variable is .75 standard deviations away from the average climate PC1, and a Warmest cagtegory for all values above this range, and Coolest for all values below this range.

```
x <- rest.m$PC1
rest.m$group <- case_when(x > mean(x)+0.75*sd(x) ~ "Warmest",
                  x < mean(x)+0.75*sd(x) & x > mean(x)-0.75*sd(x) ~ "Average",
                  x < mean(x)-0.75*sd(x) ~ "Coolest")
rest.m$group <- factor(rest.m$group,
                       levels = c("Coolest", "Average", "Warmest"))
rest.m[1:5, c(1:3, 46:48)]
```

```
##   Site Age_2016  Biomass      PC1         PC2   group
## 1   A2       14 689.5750 1.812541 -0.04942355 Warmest
## 2 A302       19 652.8725 2.027111  3.57509201 Warmest
## 3   A4       19 698.8575 1.996926  3.34571551 Warmest
## 4  ARR       14 547.0475 1.852853 -0.08810994 Warmest
## 5   B1       16 449.2875 1.973322 -0.17462123 Warmest
```

Linear Models with Reduced Dimensional Data

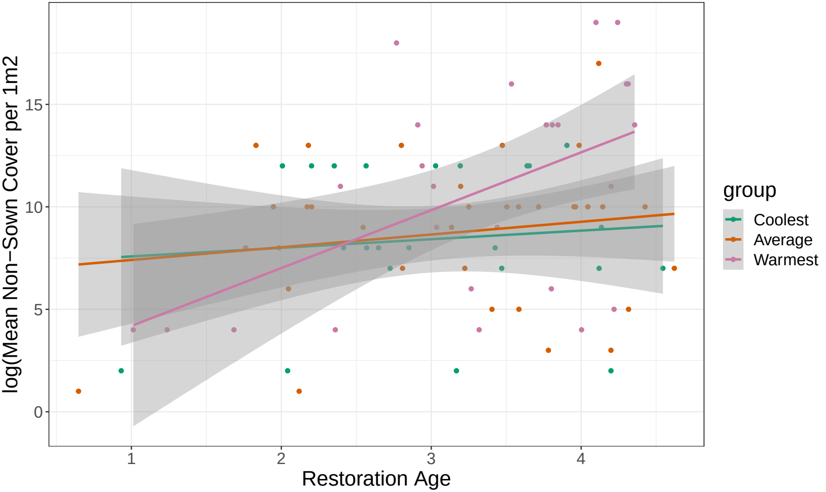
Let's look at how the average cover of non-sown species (the weedy ones) is a function of restoration age, and the climate variables.

```
nonsown.cover.lm2 <- lm(log(Mean.Nonsown.Cover) ~ Age_2016 + group,
                    data = rest.m)
Anova(nonsown.cover.lm2,  type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: log(Mean.Nonsown.Cover)
##             Sum Sq Df F value    Pr(>F)
## (Intercept) 66.491  1 81.8221 7.926e-14 ***
## Age_2016     5.680  1  6.9896  0.009885 **
## group        0.191  2  0.1173  0.889514
## Residuals   64.197 79
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Graphing This Model

$$\log(\text{mean.nonsown.cover}) \sim \text{Age} + \text{group}$$



Reduced Dimensional Y Variables

You can do the same thing for response variables. Say instead of wanting to look at plant cover, you want to look at how restoration age and climate affect the soil properties of each site.

```
restoration[, c(28:45)]
```

```
## # A tibble: 85 x 18
##       pH Organic~1 S.ppm P.mg.~2 Ca.mg~3 Mg.mg~4 K.mg.~5 Na.mg~6 B.hal~7 Fe.mg~8
##    <dbl>     <dbl> <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1   5.8      2.97     9      29    1194     178      84      21    0.35     559
##  2   6.1      2.64    13     122     886     144      76      21    0.23     230
##  3   6        2.72    11      35    1194     167      68      42    0.22     291
##  4   5.8      4.92    17     120    1402     231     107      28    0.4      153
##  5   6.5      0.68     7      40     336      54      18      23    0.2      223
##  6   6.4      0.72     8      24     279      40      21      24    0.1      151
##  7   6.1      1.16    10     182     475      83      61      27    0.29     245
##  8   5.9      3.44    11      87     893     125      76      22    0.24     235
##  9   6.2      1.98    11      25    1364     121     111      28    0.23     156
## 10   6.3      3.36    12      54    1367     200      99      24    0.42     132
## # ... with 75 more rows, 8 more variables: Mn.mg.per.kg <dbl>,
## #   Cu.half.detection <dbl>, Zn.mg.per.kg <dbl>, Al.mg.kg <dbl>,
## #   Clay.percent <dbl>, Silt.percent <dbl>, Sand.percent <dbl>,
## #   Water.Holding.Capacity <dbl>, and abbreviated variable names
## #   1: Organic.Matter.percent, 2: P.mg.per.kg, 3: Ca.mg.per.kg,
## #   4: Mg.mg.per.kg, 5: K.mg.per.kg, 6: Na.mg.per.kg, 7: B.half.detection,
## #   8: Fe.mg.per.kg
```

Reduced Dimensional Y Variables

Create a PCA for soil variables.

```
soil.pca <- prcomp(restoration[, c(28:45)], center = TRUE, scale = TRUE,
                   na.rm = TRUE)
summary(soil.pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4    PC5     PC6     PC7
## Standard deviation     2.6949 1.6734 1.4651 1.09292 0.9721 0.87011 0.80462
## Proportion of Variance 0.4035 0.1556 0.1193 0.06636 0.0525 0.04206 0.03597
## Cumulative Proportion  0.4035 0.5590 0.6783 0.74465 0.7972 0.83921 0.87517
##                            PC8     PC9    PC10    PC11    PC12   PC13   PC14
## Standard deviation     0.69395 0.64391 0.59302 0.53775 0.49874 0.3956 0.3422
## Proportion of Variance 0.02675 0.02303 0.01954 0.01607 0.01382 0.0087 0.0065
## Cumulative Proportion  0.90193 0.92496 0.94450 0.96057 0.97438 0.9831 0.9896
##                           PC15    PC16    PC17      PC18
## Standard deviation     0.29489 0.25243 0.19185 4.987e-16
## Proportion of Variance 0.00483 0.00354 0.00204 0.000e+00
## Cumulative Proportion  0.99442 0.99796 1.00000 1.000e+00
```

PC1 explains 40% of the variation, so let's use that for the Y.

Linear Model with Reduced Dimensional Y

```r
soil <- soil.pca$x[, 1:2]
colnames(soil) <- c("soilPC1", "soilPC2")
rest.all <- cbind(rest.m, soil)

soil.lm <- lm(soilPC1 ~ Age_2016 + PC1 + PC2, data = rest.all)
Anova(soil.lm,  type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: soilPC1
##             Sum Sq Df F value   Pr(>F)
## (Intercept)   1.15  1  0.1796 0.672847
## Age_2016      1.32  1  0.2062 0.651026
## PC1          46.34  1  7.2434 0.008683 **
## PC2          49.54  1  7.7432 0.006742 **
## Residuals   505.42 79
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```