A Comparative Analysis with Oxalis, continued

Abigail Mazie and Cécile Ané

April 26, 2011

Goals

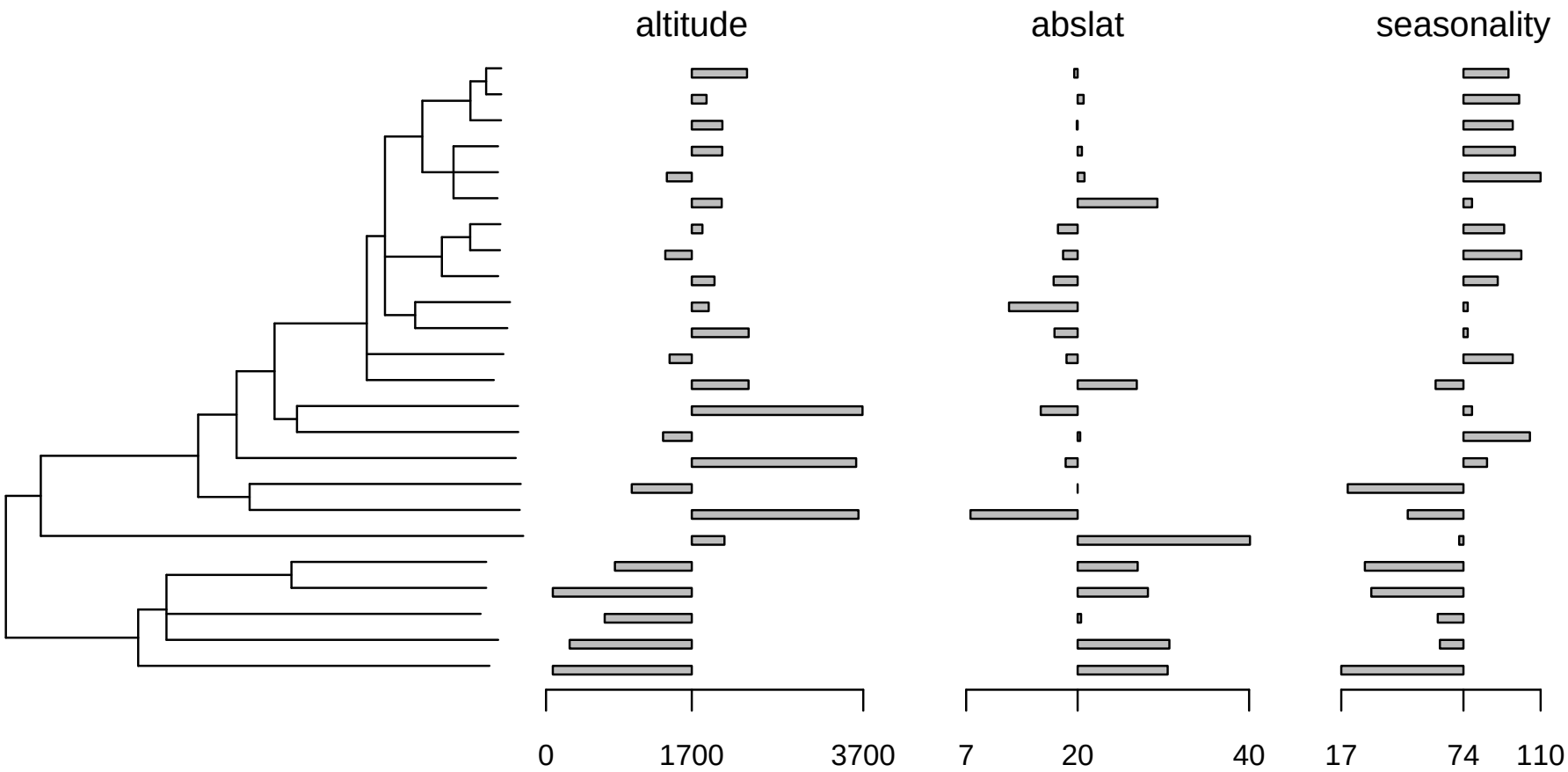
- Comparative analysis of *Oxalis* data, continued;
- Using R for comparative analysis;
- Illustrating multiple regression and principal component analysis (PCA).

Reading the Data

```
> library(ape)
> library(picante)
> library(phylobase)

> oxalis.dat = read.csv("oxalis-data.csv")
> rownames(oxalis.dat) = oxalis.dat$species
> oxalis.dat = oxalis.dat[, -1]

# add column with (natural) log of # scales/mm
> oxalis.dat$log.scales.mm = log(oxalis.dat$scales.mm)
```

```
> oxalis.dat
                     alt precip seasonality scales.mm   lat abslat          region log.scales.mm
OadenophyllaADEPH1  2081    785          72        40 -40.1   40.1           Basal      3.688879
ObrasiliensisBRAS2    79   1097          17         7 -30.5   30.5 SESouthAmerican      1.945910
OperdicariaMV79      275   1262          63        64 -30.7   30.7 SESouthAmerican      4.158883
OdebilisEE171        684   1613          62        10 -20.4   20.4 SESouthAmerican      2.302585
OhispidulaMV44MV342   79   1211          31         8 -28.2   28.2 SESouthAmerican      2.079442
ObipartitaMV59MV320  803   1510          28         5 -27.0   27.0 SESouthAmerican      1.609438
...
```

Visualizing the data



Possible Variables

- Temperature (minimum, maximum, mean)
- Precipitation
- Latitude
- Altitude
- Bioclimatic Variables

Bioclimatic Variables

- BIO1 = Annual Mean Temperature
- BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp))
- BIO3 = Isothermality (BIO2/BIO7) (* 100)
- BIO4 = Temperature Seasonality (standard deviation *100)
- BIO5 = Max Temperature of Warmest Month
- BIO6 = Min Temperature of Coldest Month
- BIO7 = Temperature Annual Range (BIO5-BIO6)
- BIO8 = Mean Temperature of Wettest Quarter
- BIO9 = Mean Temperature of Driest Quarter
- BIO10 = Mean Temperature of Warmest Quarter
- BIO11 = Mean Temperature of Coldest Quarter
- BIO12 = Annual Precipitation
- BIO13 = Precipitation of Wettest Month
- BIO14 = Precipitation of Driest Month
- BIO15 = Precipitation Seasonality (Coefficient of Variation)
- BIO16 = Precipitation of Wettest Quarter
- BIO17 = Precipitation of Driest Quarter
- BIO18 = Precipitation of Warmest Quarter
- BIO19 = Precipitation of Coldest Quarter

Why use principal component analysis?

- Too many variables for direct multiple regression (little power)
- PCA will allow us to reduce number of predictor variables to a few components, while retaining most information from all predictor variables
- Can use PCA on groups of predictor variables to reduce to one component

Bioclimatic Variables: Temperature Group

- BIO1 = Annual Mean Temperature
- BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp))
- BIO3 = Isothermality (BIO2/BIO7) (* 100)
- BIO4 = Temperature Seasonality (standard deviation *100)
- BIO5 = Max Temperature of Warmest Month
- BIO6 = Min Temperature of Coldest Month
- BIO7 = Temperature Annual Range (BIO5-BIO6)
- BIO8 = Mean Temperature of Wettest Quarter
- BIO9 = Mean Temperature of Driest Quarter
- BIO10 = Mean Temperature of Warmest Quarter
- BIO11 = Mean Temperature of Coldest Quarter

- BIO12 = Annual Precipitation
- BIO13 = Precipitation of Wettest Month
- BIO14 = Precipitation of Driest Month
- BIO15 = Precipitation Seasonality (Coefficient of Variation)
- BIO16 = Precipitation of Wettest Quarter
- BIO17 = Precipitation of Driest Quarter
- BIO18 = Precipitation of Warmest Quarter
- BIO19 = Precipitation of Coldest Quarter

Bioclimatic Variables: Precipitation Group

- BIO1 = Annual Mean Temperature
- BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp))
- BIO3 = Isothermality (BIO2/BIO7) (* 100)
- BIO4 = Temperature Seasonality (standard deviation *100)
- BIO5 = Max Temperature of Warmest Month
- BIO6 = Min Temperature of Coldest Month
- BIO7 = Temperature Annual Range (BIO5-BIO6)
- BIO8 = Mean Temperature of Wettest Quarter
- BIO9 = Mean Temperature of Driest Quarter
- BIO10 = Mean Temperature of Warmest Quarter
- BIO11 = Mean Temperature of Coldest Quarter

- BIO12 = Annual Precipitation
- BIO13 = Precipitation of Wettest Month
- BIO14 = Precipitation of Driest Month
- BIO15 = Precipitation Seasonality (Coefficient of Variation)
- BIO16 = Precipitation of Wettest Quarter
- BIO17 = Precipitation of Driest Quarter
- BIO18 = Precipitation of Warmest Quarter
- BIO19 = Precipitation of Coldest Quarter

Question

- With PCA, the author said that sometimes we would like to perform a PCA on the evolutionary **correlation** matrix, rather than the evolutionary **variance-covariance** matrix.

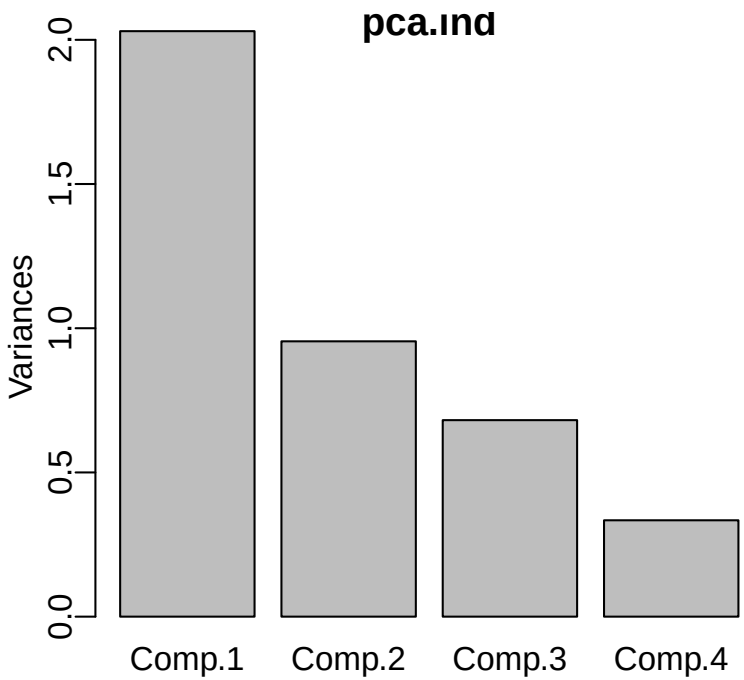
  When should we use one or another?

Note: PCA on correlation matrix = PCA on re-scaled variables, so that each one is centered and has variance 1.

Principal Component Analysis in R

```
# 1st column = atl, 2nd=precip, 3d=season, 6th=abslat
> head(oxalis.dat[,c(1,2,3,6)])
                     alt precip seasonality abslat
OadenophyllaADEPH1  2081    785          72   40.1
ObrasiliensisBRAS2    79   1097          17   30.5
OperdicariaMV79      275   1262          63   30.7
OdebilisEE171        684   1613          62   20.4
OhispidulaMV44MV342   79   1211          31   28.2
ObipartitaMV59MV320  803   1510          28   27.0
...

> pca.ind = prcomp(oxalis.dat[,c(1,2,3,6)], scale=T)

> summary(pca.ind)
Importance of components:
                        Comp.1 Comp.2 Comp.3  Comp.4
Standard deviation     1.4248 0.9770 0.8254 0.57803
Proportion of Variance 0.5075 0.2386 0.1703 0.08353
Cumulative Proportion  0.5075 0.7461 0.9165 1.00000

# plot variances explained by each axis
> plot(pca.ind)
```

How much is explained by the PC axes



We could keep 3 components for future multiple regression. Not very useful here, then.

4 original variables with clear interpretation versus 3 PC variables with less clear interpretation.

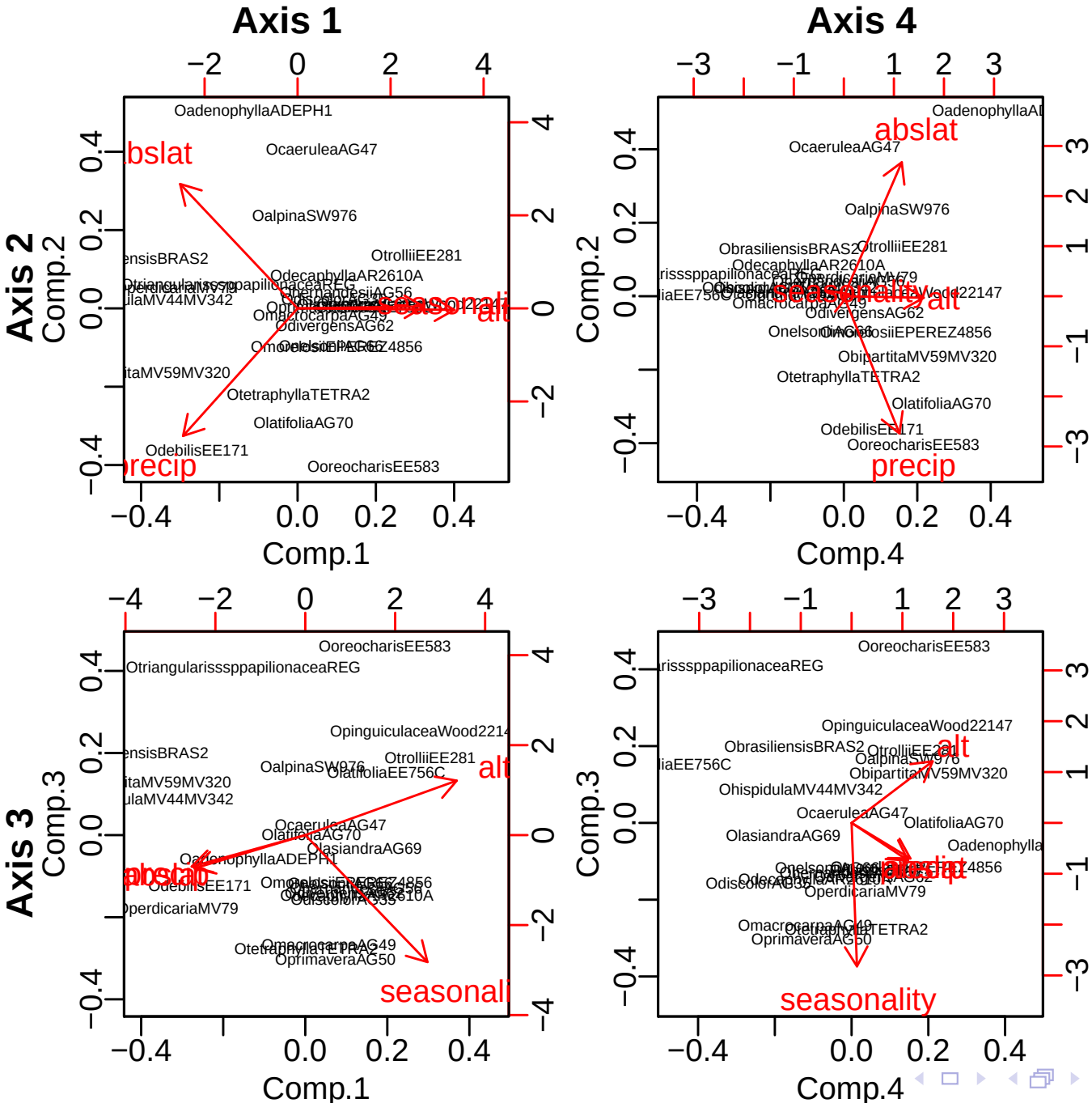
Interpretation of PC axes

```
> pca.ind$roration  # loadings
Loadings:
            Comp.1 Comp.2 Comp.3 Comp.4
alt          0.603  0.008  0.374  0.704
precip      -0.441 -0.716 -0.229  0.491
seasonality  0.487 -0.011 -0.872 -0.047
abslat      -0.452  0.698 -0.217  0.511
```

Interpretation:

- axis 1 = contrast between average of altitude & seasonality and average of precip & abslat.
- axis 2 = contrast between abs.latitude and precipitation
- axis 3 = - seasonality mostly
- axis 4 = $\sim$ average of alt, precip and abslat.

Interpretation of PC axes



How to get these “biplots”

```
layout(matrix(1:4,2,2))
par(mar=c(2.5,2,1.5,1.5), mgp=c(1.4,.4,0),oma=c(0,1.5,1.5,0))

biplot(pca.ind, cex=c(.5,1))
biplot(pca.ind, choices=c(1,3),cex=c(.5,1))
biplot(pca.ind, choices=c(4,2),cex=c(.5,1))
biplot(pca.ind, choices=c(4,3),cex=c(.5,1))

mtext("Axis 1", side=3, adj=.2, outer=T, font=2)
mtext("Axis 2", side=2, adj=.8, outer=T, font=2)
mtext("Axis 3", side=2, adj=.2, outer=T, font=2)
mtext("Axis 4", side=3, adj=.8, outer=T, font=2)
```

How to use PCA for Multiple Regression

```
> pca.ind$x   # scores
> newpred = as.data.frame(pca.ind$x[,1:3])
> newpred
                                    PC1   PC2   PC3
OadenophyllaADEPH1                 0.77 -2.36  0.24
ObrasiliensisBRAS2                 2.78 -0.60 -0.79
OperdicariaMV79                    2.12 -0.23  0.71
OdebilisEE171                      1.75  1.69  0.48
OhispidulaMV44MV342                2.56 -0.09 -0.34
ObipartitaMV59MV320                2.56  0.78 -0.50
...

> round(cor(pca.ind$x, subset(oxalis.dat,select=8)),2)
   log.scales.mm
PC1          0.26
PC2          0.08
PC3         -0.66
PC4          0.36
```

How to use PCA for Multiple Regression

```
> fit.ind = lm(oxalis.dat$log.scales.mm ~ PC1+PC2+PC3, data=newpred)
> summary(fit.ind)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.78621    0.15064  18.496 4.78e-14 ***
PC1         -0.18406    0.10800  -1.704 0.103841
PC2         -0.08198    0.15750  -0.521 0.608425
PC3          0.79530    0.18643   4.266 0.000378 ***

> summary(lm(oxalis.dat$log.scales.mm ~ PC1+PC3, data=newpred))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.7862     0.1480  18.825 1.25e-14 ***
PC1          -0.1841     0.1061  -1.735 0.097481 .
PC3           0.7953     0.1832   4.342 0.000287 ***

> summary(lm(oxalis.dat$log.scales.mm ~ PC2, data=newpred))
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.78621    0.20590  13.532 3.83e-12 ***
PC2         -0.08198    0.21528  -0.381    0.707
```

The coefficients are unchanged if we include or exclude other components in the list of predictors.

Multiple Regression

- first: standard, non-phylogenetic multiple regression.
- response: (natural) log of # scales per mm
- 4 predictors: seasonality, absolute latitude, altitude, precipitation.
- why not latitude?

$$\begin{aligned}\log(\text{scales/mm}) = \; & b_0 + b_1 * \text{seasonality} + b_2 * \text{abs.latitude} \\ & + b_3 * \text{altitude} + b_4 * \text{precip} \\ & + \text{residual variation}\end{aligned}$$

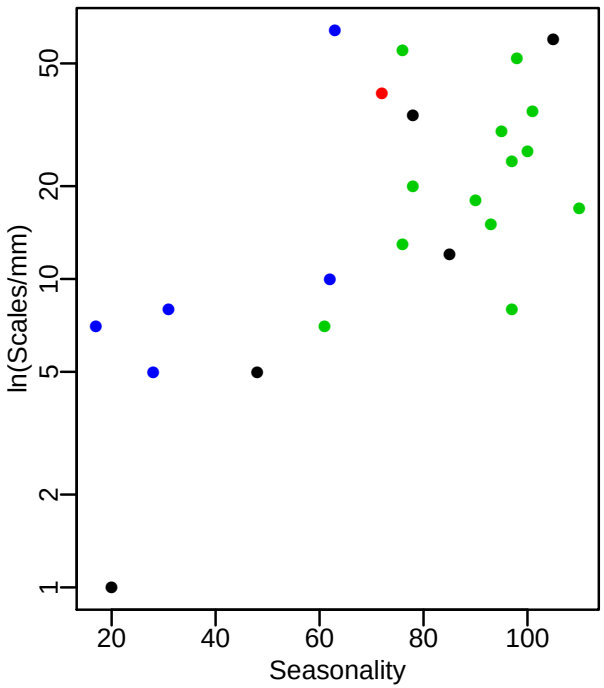
Multiple Regression

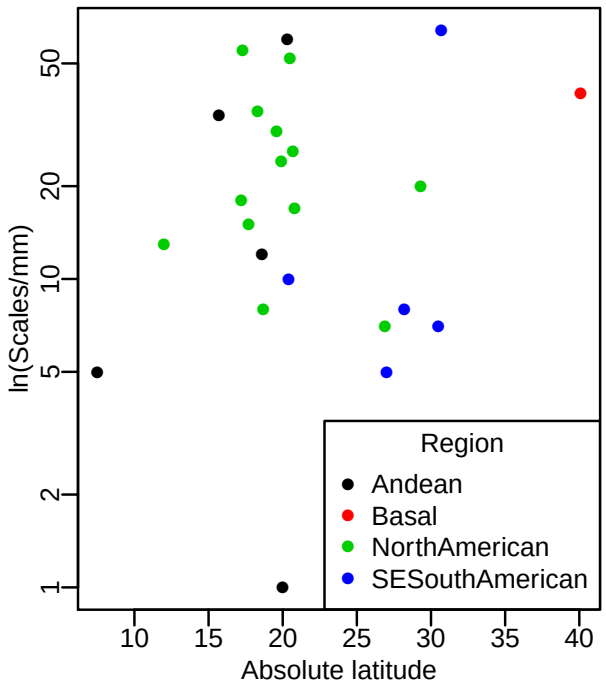
```
> fit1 = lm(log.scales.mm ~ seasonality+abslat+alt+precip, data=oxalis.dat)
> summary(fit1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.4030243  1.2105430  -1.985   0.0618 .
seasonality  0.0292465  0.0054350   5.381 3.42e-05 ***
abslat       0.0665638  0.0238482   2.791   0.0116 *
alt          0.0002407  0.0001844   1.306   0.2073
precip       0.0011831  0.0005624   2.104   0.0490 *

> fit2 = update(fit1, .~.-alt)
> summary(fit2)
...
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.3830970  0.9408541  -1.470   0.1571
seasonality  0.0303468  0.0054629   5.555 1.95e-05 ***
abslat       0.0511692  0.0210909   2.426   0.0248 *
precip       0.0008309  0.0005021   1.655   0.1136
```

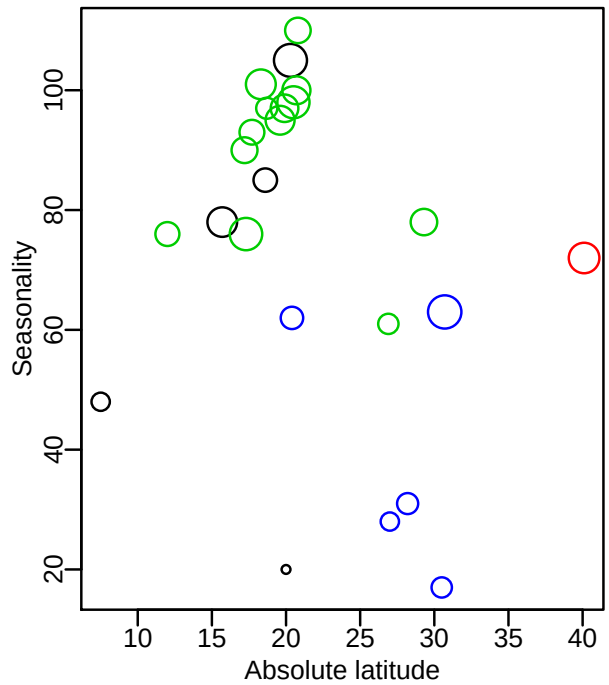
Multiple Regression

```
> fit3 = update(fit2, .~.-precip)
> summary(fit3)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.330993   0.721669  -0.459   0.6512
seasonality  0.027633   0.005422   5.096 4.78e-05 ***
abslat       0.049428   0.021919   2.255   0.0349 *

> plot(fit3) # diagnostic plots look okay

> fit4=update(fit3, .~.-seasonality)
> summary(fit4)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.45041    0.68987   3.552  0.00179 **
abslat       0.01556    0.03052   0.510  0.61522
```

Scatterplots







Scatterplots

```
pdf("scatter.pdf", height=3, width=8)
layout(matrix(1:3,1,3))
par(mar=c(3.1,3.1,.5,.5), mgp=c(1.4,.4,0))

plot(scales.mm ~ seasonality,  col=region, log="y",
 data=oxalis.dat, pch=16, xlab="Seasonality", ylab="ln(Scales/mm)")

plot(scales.mm ~ abslat,  col=region, log="y",
 data=oxalis.dat, pch=16, xlab="Absolute latitude", ylab="ln(Scales/mm)")
legend("bottomright", pch=16, col=1:4, legend=levels(oxalis.dat$region),
 title="Region")

plot(seasonality ~ abslat, data=oxalis.dat, pch=1, cex=(log(scales.mm)+1.5)/2,
 ylab="Seasonality", xlab="Absolute latitude", col=region)
dev.off()
```

Phylogenetic Multiple Regression, BM model

```
> fit1p = gls(log.scales.mm ~ seasonality+abslat+alt+precip,
              data=happy.consensus$data,
              correlation=corBrownian(1,happy.consensus$phy))
> summary(fit1p)
Coefficients:
                 Value Std.Error   t-value p-value
(Intercept) -1.8386655 1.4671421 -1.253229  0.2253
seasonality  0.0258897 0.0089130  2.904725  0.0091
abslat       0.0533496 0.0309965  1.721150  0.1015
alt          0.0003798 0.0002169  1.751202  0.0960
precip       0.0009950 0.0004984  1.996443  0.0604

> fit2p = update(fit1p, .~.-alt)
> summary(fit2p)
Coefficients:
                 Value Std.Error    t-value p-value
(Intercept) -0.6510246 1.3665549 -0.4763984  0.6390
seasonality  0.0221185 0.0090848  2.4346799  0.0244
abslat       0.0394230 0.0314688  1.2527653  0.2247
precip       0.0009390 0.0005224  1.7973132  0.0874
```

Phylogenetic Multiple Regression, BM model

```
> fit3p = update(fit2p, .~.-precip)
> summary(fit3p)
Coefficients:
                Value Std.Error   t-value p-value
(Intercept) 0.5678728 1.2478134 0.4550943  0.6537
seasonality 0.0248235 0.0094230 2.6343513  0.0155
abslat      0.0246305 0.0319457 0.7710130  0.4493

> fit3p = update(fit2p, .~.-abslat)
> summary(fit3p)
Coefficients:
                Value Std.Error   t-value p-value
(Intercept) 0.6847540 0.8662346 0.7904949  0.4381
seasonality 0.0195747 0.0089742 2.1812198  0.0407
precip      0.0007678 0.0005110 1.5024324  0.1479

> fit4p = update(fit3p, .~.-precip)
> summary(fit4p)
Coefficients:
                Value Std.Error  t-value p-value
(Intercept) 1.3146924 0.7793614 1.686884  0.1058
seasonality 0.0227868 0.0089614 2.542770  0.0185
```

Questions

- Is the matrix C the same for all size-dependent traits?
- If the evolutionary changes (i.e. bulb morphology), are being driven by the availability of ecological niches (i.e. environmental variables), can we assume Brownian Motion as the model for evolutionary process when it describes the evolution of the trait as random and it is suppose to result in a normal distribution of the trait value across the phylogeny? Shouldn't we be testing first for BM?

Notes:

- Recall log(scales/mm) $= b_0 + b_1 *$ seasonality$+$ $b_2 *$ abs.latitude $+ b_3 *$ altitude$+$ residual variation. BM assumption is on the residual variation.

Questions

- Is the matrix C the same for all size-dependent traits?
- If the evolutionary changes (i.e. bulb morphology), are being driven by the availability of ecological niches (i.e. environmental variables), can we assume Brownian Motion as the model for evolutionary process when it describes the evolution of the trait as random and it is suppose to result in a normal distribution of the trait value across the phylogeny? Shouldn't we be testing first for BM?

Notes:

- Recall log(scales/mm) $= b_0 + b_1 *$ seasonality$+$ $b_2 *$ abs.latitude $+ b_3 *$ altitude$+$ residual variation. BM assumption is on the residual variation.
- Maybe BM not perfect: adjust branch lengths to the data?

Questions

- Is the matrix C the same for all size-dependent traits?
- If the evolutionary changes (i.e. bulb morphology), are being driven by the availability of ecological niches (i.e. environmental variables), can we assume Brownian Motion as the model for evolutionary process when it describes the evolution of the trait as random and it is suppose to result in a normal distribution of the trait value across the phylogeny? Shouldn't we be testing first for BM?

Notes:

- Recall log(scales/mm) $= b_0 + b_1 *$ seasonality$+$ $b_2 *$ abs.latitude $+ b_3 *$ altitude$+$ residual variation. BM assumption is on the residual variation.
- Maybe BM not perfect: adjust branch lengths to the data?
- OU models with multiple optima, like multiple "niches": next week?

Phylogenetic Multiple Regression, BM model with λ

```
> fit1p = gls(log.scales.mm ~ seasonality+abslat, data=happy.consensus$data,
           correlation=corPagel(1,happy.consensus$phy))
> summary(fit1p)
Correlation Structure: corPagel
  lambda
0.2970
Coefficients:
                  Value Std.Error   t-value p-value
(Intercept) -0.19157210 0.7822855 -0.244888  0.8089
seasonality  0.02974626 0.0064620  4.603227  0.0002
abslat       0.04110774 0.0237137  1.733505  0.0977

> fit2p = update(fit1p, .~.-abslat)
> summary(fit2p)
Correlation Structure: corPagel
   lambda
0.6378635
Coefficients:
                Value Std.Error  t-value p-value
(Intercept) 0.9311904 0.5587248 1.666635  0.1098
seasonality 0.0287868 0.0074809 3.848046  0.0009
```

Phylogenetics Principal Component Analysis

Using function written by Liam Revell, download file here:

```
http://anolis.oeb.harvard.edu/~liam/R-phylogenetics/phyl.pca/

> source("phyl.pca.R")
> pca.phy2 = phyl.pca(happy.consensus$phy,
                      happy.consensus$dat[,c(1,2,3,6)],
                      mode="corr")
> pca.phy2$Eval
          PC1      PC2       PC3       PC4
PC1 1.566031 0.000000 0.0000000 0.0000000
PC2 0.000000 1.153042 0.0000000 0.0000000
PC3 0.000000 0.000000 0.7385365 0.0000000
PC4 0.000000 0.000000 0.0000000 0.5423898
```

Phylogenetics Principal Component Analysis

Alternatively: Do PCA on phylogenetically independent contrasts.

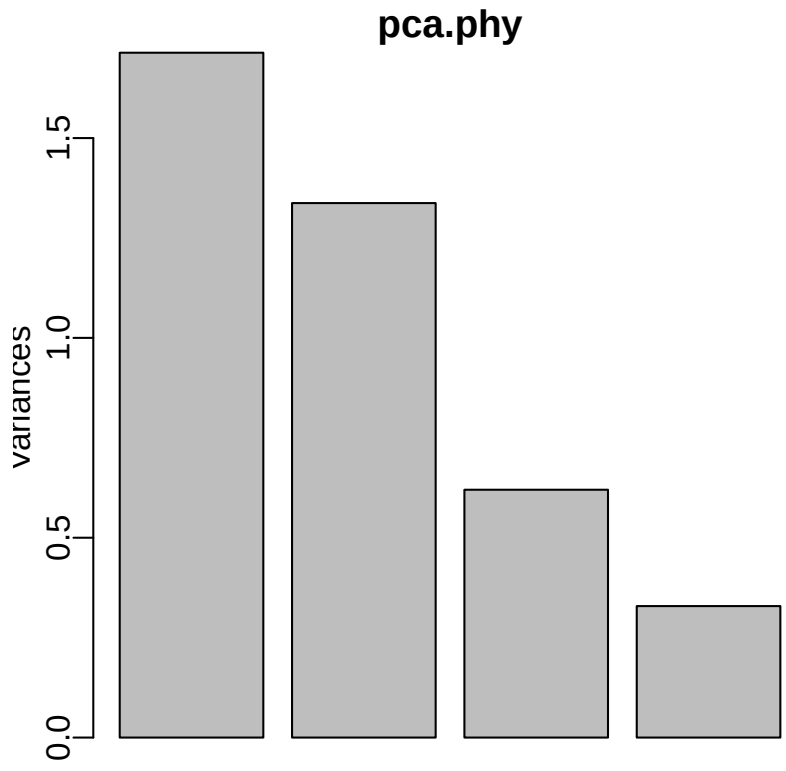
```
> happy = match.phylo.data(trees[[1]],oxalis.dat)

> pic.alt         = pic(happy$data["alt"],         happy$phy)
> pic.precip      = pic(happy$data["precip"],      happy$phy)
> pic.seasonality = pic(happy$data["seasonality"],happy$phy)
> pic.abslat      = pic(happy$data["abslat"],      happy$phy)

> pic.data = cbind(pic.alt, pic.precip, pic.seasonality, pic.abslat)
> pic.data
       pic.alt  pic.precip pic.seasonality  pic.abslat
25   73.881403  -53.457875      32.7993779 -118.240561
26   30.365122  -66.536279     115.0349694   30.696753
27   69.460995  -60.091738     -21.9336346  -16.850701
28  -11.709317   13.274186     -33.2314109   43.069406
29   12.080398  -62.594111     -61.8705167   69.819194
30  125.176153  -62.588076    -125.1761529  112.658538
31    3.756950  -40.454788      75.3165198  214.436457
...
```

Phylogenetics Principal Component Analysis

Important: *not* center the phylogenetic independent contrasts for PCA, i.e. do PCA “through the origin”.

```
> pca.phy = prcomp(pic.data, center=F, scale=T)
> summary(pca.phy)
Importance of components:
                          PC1   PC2   PC3    PC4
Standard deviation     1.309 1.156 0.787 0.5736
Proportion of Variance 0.428 0.334 0.155 0.0823
Cumulative Proportion  0.428 0.763 0.918 1.0000

> plot(pca.phy)
```

Phylogenetics Principal Component Analysis

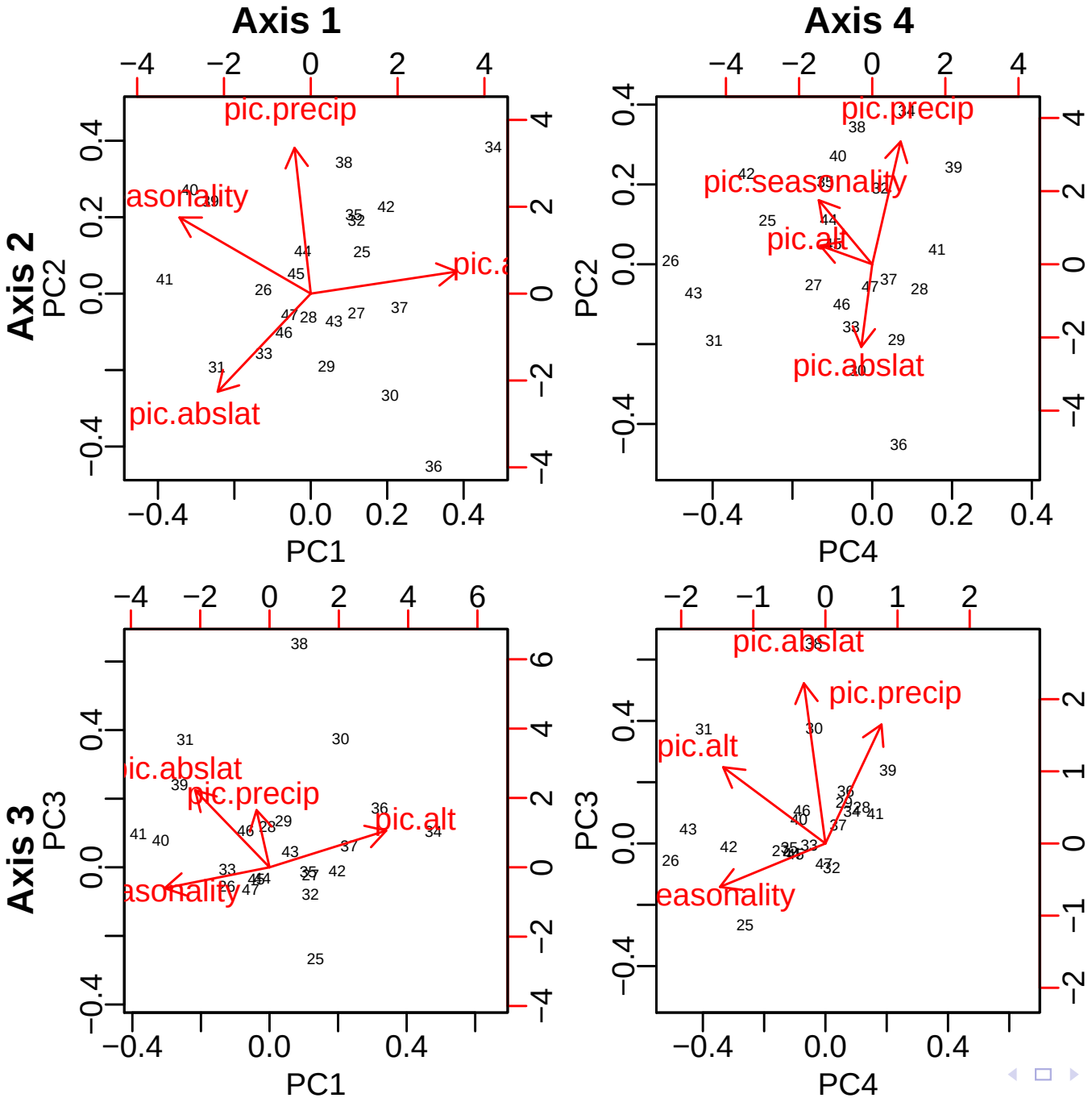


Phylogenetics PCA: interpreting axes

```
> pca.phy$rotation # loadings
                     PC1    PC2    PC3    PC4
pic.alt           0.6703  0.115  0.350 -0.644
pic.precip       -0.0745  0.756  0.546  0.353
pic.seasonality -0.6023  0.395 -0.199 -0.665
pic.abslat       -0.4270 -0.509  0.735 -0.136

> biplot(pca.phy)
```

axis 1: contrast between altitude and seasonality
axis 2: precipitation mostly
axis 3: absolute latitude mostly
axis 4: - average altitude and seasonality

Phylogenetics PCA: interpreting axes



Conclusions

- Multiple regression: seasonality still showed an association with bulb morphology.

  The association between absolute latitude and bulb morphology was statitically significant only when seasonality was included, and when phylogenetic correlation was ignored.
- PCA is useful to reduce the number of predictors prior to multiple regression, when there are many potential predictors compared to sample size.
- PC axes are harder to interpret than original variables. Useful to apply PCA on predefined groups of variables, for easier interpretation.
- Phylogenetic PCA has yet to be thoroughly tested.