What is PCA?

- method of simplifying multivariate data
  - reduce number of variables
  - detect structure
- start with variables that correlate to varying degrees
- generate linear combinations that capture most variance
- the best ones are then factors which may better describe the data

What is PCA? continued

- deterministic algebraic trick
- unlike maximum likelihood factor analysis methods
- outcome of PCA and FA should be similar
  - PCA usually for reduction
  - FA usually for structure

geometrical view of PCA

if you have a 2- or 3- dimentional data set

ie 2 or 3 variables scatterploted

you could imagine rotating the axes

PCA as multivariate extension of regression

if you have 2 variables you can plot a scatterplot

if there is a correlation you can fit a line

line represents the what's common (covariance)

how do you get what's common out of multiple variables?

AN(c)OVA view

- both are ways of dividing up variance
- anova: test hypothesis that variables explain response
- PCA: divide up variance and then look for explanation of factors

related techniques

- exploration:
  - pcfa
  - heirarchical clustering
  - pls (more later)
- confirmation:
  - ML factor analysis
  - SEM (next week)
  - supervised clustering

PCA requirements/assumptions

- **linear relationships**
- **interval data (approximately)**
  - **appropriate scale**
- **untruncated variables**
- **proper specification**
  - **no irrelevant variables**
- **no high multicollinearity**

Some examples of calculating principal components

call	sqrt(eigenvalues)	eigenvalues	eigenvectors	rotated data
prcomp(ph,scale.=T)	sdev		rotation	scores
princomp(scale(ph))	sdev		loadings	x
eigen(cor(ph))		values	vectors	scale(ph) %*% eig$vectors
svd(cor(ph))		d	u	scale(ph) %*% sv$vectors

try PCA on Lin's open field data

```
pc1 <- prcomp(ph)
plot(pc1)
biplot(pc1)
image(t(as.matrix(ph)), xaxt='n')
mtext(colnames(ph),3,at=0:9/9)
heatmap(as.matrix(ph))
heatmap(scale(ph))
```

try PCA on Lin's open field data --correctly

```
pc2 <- prcomp(ph, scale.=T)
plot(pc2)
biplot(pc2)
```

Interpretation: eigenvalues

remember it’s the variance

how many components do you pay attention to?

2 popular rules-of-thumb

- **variance > 1**
- **scree inflection**

Interpretation: rotation aka loadings

pc2

str(pc2)

loadings(pc2)

varimax(pc2$rotation[,1:2])

pc2$rotation

heatmap(pc2$rot)

heatmap(pc2$x)

looking at loadings

plot pairs of loading columns (=eigenvectors)

plot(PC1 ~ PC2 , data=pc2$r)

plot(PC2 ~ PC3 , data=pc2$r)

another convenient way

with(data.frame(pc2$r), plot(PC1,PC2))

with(data.frame(pc2$r), text(PC1,PC2,rownames(pc2$r),pos=3, cex=0.6))

cloud(PC1~PC2+PC3, data=data.frame(pc2$r))

Partial least squares

Herve Abdi tutorial

PCA - akin to unsupervised clustering

PLS - supervised analogue

PLS resembles regression analysis with multiple responses

applicable where you have predictors and responses

pioneered by Herman Wold and his son Svante

http://www.utd.edu/~herve/Abdi-PLSR2007-pretty.pdf

toy example of PLS

```
install.packages('pls')

X <- read.table('http://rcourse.iop.kcl.ac.uk/S11/x')

Y <- read.table('http://rcourse.iop.kcl.ac.uk/S11/y')

pls3 <- plsr(Y ~ X, 3)

biplot(pls3)

biplot(pls2, which='y')
```

```
What is PCA?
        method of simplifying multivariate data
                reduce number of variables
                detect structure
        start with variables that correlate to varying degrees
        generate linear combinations that capture most variance
        the best ones are then factors which may better describe the data

What is PCA?  continued
        deterministic algebraic trick
        unlike maximum likelihood factor analysis methods
        outcome of PCA and FA should be similar
                PCA usually for reduction
                FA usually for structure

geometrical view of PCA
        if you have a 2- or 3- dimentional data set
                ie 2 or 3 variables scatterploted
        you could imagine rotating the axes

PCA as multivariate extension of regression
        if you have 2 variables you can plot a scatterplot
        if there is a correlation you can fit a line
        line represents the what's common (covariance)
        how do you get what's common out of multiple variables?

AN(c)OVA view
        both are ways of dividing up variance
        anova: test hypothesis that variables explain response
        PCA: divide up variance and then look for explanation of factors

related techniques
        exploration:
                pcfa
                heirarchical clustering
                pls  (more later)
        confirmation:
                ML factor analysis
                SEM (next week)
                supervised clustering

PCA requirements/assumptions
        linear relationships
        interval data (approximately)
                appropriate scale
        untruncated variables
        proper specification
                no irrelevant variables
        no high multicollinearity

*PCA.htm

try PCA on Lin's open field data
        pc1 <- prcomp(ph)
        plot(pc1)
        biplot(pc1)
        image(t(as.matrix(ph)), xaxt='n')
        mtext(colnames(ph),3,at=0:9/9)
        heatmap(as.matrix(ph))
        heatmap(scale(ph))
```