Principal Components Analysis
with application to remote sensing image analysis

D G Rossiter
Cornell University, Section of Soil & Crop Sciences
April 12, 2020
Topic: Factor Analysis

A generic term for methods that consider the *inter-relations* between a set of variables.

- Often the set of *predictors* which might be used in a multiple linear regression.
  - Multivariate observations on same objects (e.g., soil samples)
  - **Remote sensing**: a set of *co-registered* images of a scene
    * all bands of one image
    * bands of multiple (co-registered) sensors
    * one band or band product (e.g., NDVI) of a *time-series* of images

- This is an analysis of the *structure* of the *multivariate feature space* covered by a set of variables.
Uses of factor analysis in remote sensing

1. Discover **relations between images**, and possible **groupings** of them

2. Discover **groupings of pixels** in a set of images (→ classification)

3. **Interpret** the resulting groupings in terms of processes

4. Diagnose **multi-collinearity**, since images are usually correlated
   - determine which images are most correlated
   - quantify **redundancy**, find the **most informative subset** of images

5. For **data reduction** for model inputs; two approaches:
   - Identify **representative** images for a **minimum data set**
   - Compute **synthetic images**
The simplest form of factor analysis; a data reduction technique.

Gives insight into the relation between a set of variables within a dataset
- This is completely data driven; different sets of observations from the same population will give different relations
- So, a data mining approach

Gives insight into the relation between a set of pixels in the multivariate space spanned by the set of images
What does PCA do?

1. The **vector space** made up of the original observations (e.g., stack of pixel values in a set of images) is **projected** onto another vector space;

2. The new space has the **same dimensionality** as the original\(^1\), i.e., there are as many variables in the new space as in the old;

3. In this space the new **synthetic images**, also called **principal components** are **orthogonal** to each other, i.e. **completely uncorrelated**;

4. The synthetic images are arranged in **decreasing order of variance explained**; and the total variance is unchanged;

5. The contribution of each original variable to each synthetic image is given;

6. Each observation can be **re-projected** into the new (PC) space, by its value of the synthetic images

\(^1\)unless the original was rank-deficient
Mathematics: the data matrix

PCA is a direct calculation from a matrix constructed from the multivariate dataset.

\[
X: \text{centred data matrix (i.e., difference from mean), where:}
\]

- **rows** are observations (e.g., pixels, observation locations, sampled individuals)
- **columns** are variables (e.g., reflectance in a band, pixel values) measured at each observation
- may **scale** by dividing values by the variable’s sample standard deviation
  - standardized vs. unstandardized, see below

This gives the location of each observation in multivariate attribute space.
Mathematics: the covariance or correlation matrix

The data matrix $X$ is used to build a matrix that shows the relation between data items:

- $C = X^T X$ : the **covariance** (unscaled) or **correlation** (scaled) matrix

- this is symmetric and positive (semi-)definite, so has all real roots
Why can the correlation matrix be misleading?

- The individual pairwise correlations do *not* take into account the degree to which *both* of the variables may be correlated to others.

- In the case where both are highly correlated to a several others this is *apparent* correlation, which may not reflect a real process.

- Solution: compute *partial correlations* - the bivariate correlation between the two *residuals* from linear regression of each variable on all the others, less the one with which to pair. This accounts for the "lurking" effect of other variables, and shows what correlation remains that can not be otherwise accounted for. (see example below)
Mathematics: Eigen decomposition

The key insight is that the Eigen decomposition\(^2\) of \(C\) orders the synthetic variables into descending amounts of variance, and ensures they are orthogonal (Hotelling 1933).

- Decompose a square, symmetric positive-definite matrix, e.g., the correlation matrix \(C\) formed from a data matrix such that \(AC = \lambda C\)

- **Eigenvalues**: a diagonal matrix \(\lambda\); off-diagonals 0, i.e., no covariances, so orthogonal; **Eigenvectors**: the transformation matrix \(A\)

- The eigenvectors provide a coördinate transformation such that the matrix multiplied by the diagonal eigenvalues matrix is the same as multiplication by a matrix made up of the eigenvectors

- Eigenvectors span an orthogonal vector space onto which we can project the original data.

\(^2\) (German *eigen* \(\approx\) English “own, belonging to oneself”)

---

D G Rossiter

Cornell University
College of Agriculture and Life Sciences
Computation

• $|C - \lambda I| = 0$: a determinant to find the **eigenvalues** of the correlation matrix
  - these are sometimes called the **characteristic values**
  - their relative magnitude is the proportion of the original covariance explained

• Then the axes of the new space, the **eigenvectors** $\gamma_j$ (one per dimension) are the solutions to $(C - \lambda_j I)\gamma_j = 0$

• Obtain **synthetic variables** by projection: $Y = PC$ where $P$ is the row-wise matrix of eigenvectors (rotations).
Details

In practice the system is solved by the Singular Value Decomposition (SVD) of the data matrix.

This is equivalent but more stable than directly extracting the eigenvectors of the correlation matrix.

Accessible explanations with worked examples:


Standardized vs. unstandardized – 1

**Standardized** each variable (e.g., reflectance in a band) has its mean subtracted (so $\bar{x}_j = 0$) and is divided by its sample standard deviation (so $\sigma(x_j) = 1$);

- All variables (e.g., bands) are **equally important**, no matter their absolute values or spreads;
- Gives **equal weight** to all variables;
- This is usually what we want if variables are measured on different scales.
  - e.g., multivariate measurements of soil constituents
  - e.g., co-registered images from different sensors
Unstandardized use the \textit{original variables}, in their original scales of measurement; generally the means are also subtracted to centre the variables

- Variables with \textit{wider spreads} (often due to measurement scale) are \textit{more important}, since they contribute more to the original variance
- This preserves the importance of variables with \textit{more variance = more information}
- E.g., sensor with different radiometric resolutions (so wider range of numeric values); higher resolution will have more weight
- Bands with more variability will have more weight – maybe we want this.
Potential difference between un/standardized!

Example: variables with three orders of magnitude difference in standard deviation (in original measurement scale):

First, **unstandardized** PCs:

<table>
<thead>
<tr>
<th></th>
<th>k.a</th>
<th>k.b</th>
<th>p.a</th>
<th>caco3.a</th>
<th>caco3.b</th>
<th>sand.b</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>206.1333028</td>
<td>172.1004094</td>
<td>53.2891699</td>
<td>25.5945396</td>
<td>24.9265695</td>
<td>20.3849822</td>
</tr>
<tr>
<td>cec.b</td>
<td>1.3351728</td>
<td>0.9138645</td>
<td>0.7702299</td>
<td>0.7265857</td>
<td>0.4260936</td>
<td>0.4153947</td>
</tr>
<tr>
<td>dens.b</td>
<td>0.2721703</td>
<td>0.2313593</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Importance of components:

<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
<th>PC5</th>
<th>PC6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard deviation</td>
<td>250.7955</td>
<td>100.7590</td>
<td>47.69829</td>
<td>34.18739</td>
<td>27.84061</td>
<td>16.31755</td>
</tr>
<tr>
<td>Proportion of Variance</td>
<td>0.8092</td>
<td>0.1306</td>
<td>0.02927</td>
<td>0.01504</td>
<td>0.00997</td>
<td>0.00343</td>
</tr>
<tr>
<td>Cumulative Proportion</td>
<td>0.8092</td>
<td>0.9398</td>
<td>0.96909</td>
<td>0.98413</td>
<td>0.99410</td>
<td>0.99753</td>
</tr>
</tbody>
</table>

PC1 is numerically very large; K values have much larger standard deviations (higher absolute values of all measurements).
Second, **standardized** PCs

<table>
<thead>
<tr>
<th></th>
<th>dens.a</th>
<th>silt.a</th>
<th>ph.a</th>
<th>cec.a</th>
<th>caco3.a</th>
<th>p.a</th>
<th>dens.b</th>
<th>sand.b</th>
<th>clay.b</th>
<th>oc.b</th>
<th>cec.b</th>
<th>caco3.b</th>
<th>p.b</th>
<th>sand.a</th>
<th>clay.a</th>
<th>oc.a</th>
<th>k.a</th>
<th>silt.b</th>
<th>ph.b</th>
<th>k.b</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Importance of components:

<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
<th>PC5</th>
<th>PC6</th>
<th>PC7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard deviation</td>
<td>2.4798</td>
<td>1.8233</td>
<td>1.4605</td>
<td>1.36289</td>
<td>1.16650</td>
<td>1.08224</td>
<td>0.85127</td>
</tr>
<tr>
<td>Proportion of Variance</td>
<td>0.3075</td>
<td>0.1663</td>
<td>0.1067</td>
<td>0.09289</td>
<td>0.06805</td>
<td>0.05857</td>
<td>0.03624</td>
</tr>
<tr>
<td>Cumulative Proportion</td>
<td>0.3075</td>
<td>0.4738</td>
<td>0.5805</td>
<td>0.67336</td>
<td>0.74141</td>
<td>0.79998</td>
<td>0.83622</td>
</tr>
</tbody>
</table>

Same standard deviation (by design), much less total variance explained by PC1.
Difference in variance represented by PCs

Standardized

Unstandardized
Standardized
More or less equal loadings (lengths of arrows)

(see below for explanation of biplots).

Unstandardized
K dominates loadings
Topic: Remote sensing examples

These will help visualize the transformation from original space into PC space.
Example: Time series of diurnal temperature differences

- 中国江苏沛县 Pei county in JiangSu province, PRC

- MODIS\(^3\) daily land-surface temperature product\(^4\)

- Images are **diurnal temperature differences** (day – night) between two MODIS products, units are $\Delta^\circ$C

- 30 Oct – 03 Nov 2000 (Julian days 304 ff.); Soil is drying after a heavy rain

- Objective: **try to relate DTD to soil texture and organic matter**


---

\(^3\) [http://modis.gsfc.nasa.gov](http://modis.gsfc.nasa.gov)


Original images

\[ \Delta \degree C \] (day - night)

Low DTD on first day after rain; increases overall then decreases (closer day/night air T?)

But: **similar overall pattern** of high vs. low DTD (**redundancy**)
## read images as a raster stack

```r
require(raster)
(list <- dir(pattern='CK_DTD_30[0-9]{1}.img$'))
stackDTD <- stack(list, RAT=FALSE)
## add a Z value to represent the time
stackDTD <- setZ(stackDTD, c(304,306,307,308))
## PCA works with points
stackDTD.pts <- rasterToPoints(stackDTD)
stackDTD.df <- as.data.frame(stackDTD.pts)
```

---

**Radney Gill**

Cornell University

College of Agriculture and Life Sciences
Simple and partial correlations

> with(stackDTD.df, cor(CK_DTD_307, CK_DTD_308)) # simple correlation of two days
[1] 0.9029475
> # compute residuals from linear model on other days
> r307 <- residuals(lm(CK_DTD_307 ~ CK_DTD_304+CK_DTD_306, data=stackDTD.df))
> r308 <- residuals(lm(CK_DTD_308 ~ CK_DTD_304+CK_DTD_306, data=stackDTD.df))
> cor(r307, r308) # partial correlation = simple correlation of residuals
[1] 0.5054682
> plot(CK_DTD_308 ~ CK_DTD_307, data=stackDTD.df); plot(r308 ~ r307)

Partial correlation accounts for correlations of each with the other days
PCA processing in R

In this example we use **unstandardized** PCA because the four DTD images are on the **same scale** from the **same sensor** and represent the **same phenomenon**.

```r
## PCA -- unstandardized, use original DTD, ignore coordinates
pca <- prcomp(stackDTD.pts[,3:dim(stackDTD.pts)[2]], scale = FALSE, retx=TRUE)
summary(pca) # show variance explained by each PC
pc$rotation # show contribution of each original band to each PC
screeplot(pca)

## extract synthetic bands, convert back to raster stack
stackDTD.scores <- cbind(stackDTD.pts[,1:2], pca$x)
stackDTD.scores <- data.frame(stackDTD.scores)
coordinates(stackDTD.scores) <- ~ x + y; gridded(stackDTD.scores) <- TRUE
stackDTD.scores <- stack(stackDTD.scores)
# name the synthetic bands in the raster stack
stackDTD.scores <- setZ(stackDTD.scores, paste("PC", 1:4, sep=""))
```

D G Rossiter
Structure of prcomp object

> str(pca)
List of 5
  $ sdev   : num [1:4] 1.823 0.404 0.323 0.208
  $ rotation: num [1:4, 1:4] -0.459 -0.590 -0.466 -0.474 -0.675 ...
     ..- attr(*, "dimnames")=List of 2
     .. ..$: chr [1:4] "CK_DTD_304" "CK_DTD_306" "CK_DTD_307" "CK_DTD_308"
     .. ..$: chr [1:4] "PC1" "PC2" "PC3" "PC4"
  $ center : Named num [1:4] 10.6 13.4 12.9 11.8
     ..- attr(*, "names")= chr [1:4] "CK_DTD_304" "CK_DTD_306" "CK_DTD_307" "CK_DTD_308"
  $ scale : logi FALSE
  $ x      : num [1:784, 1:4] -6.17 -5.71 -4.64 -4.83 -4.43 ...
     ..- attr(*, "dimnames")=List of 2
     .. ..$: NULL
     .. ..$: chr [1:4] "PC1" "PC2" "PC3" "PC4"
- attr(*, "class")= chr "prcomp"

rotation are the eigenvectors

x are the PC scores (location of each pixel in the PC space)

center are the image means (subtracted from all values)
PCA results – Importance of components

> summary(pca)

Importance of components:

<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Std.</td>
<td>1.8232</td>
<td>0.40406</td>
<td>0.3230</td>
<td>0.20810</td>
</tr>
<tr>
<td>Proportion</td>
<td>0.9145</td>
<td>0.04492</td>
<td>0.0287</td>
<td>0.01191</td>
</tr>
<tr>
<td>Cum.</td>
<td>0.9145</td>
<td>0.95938</td>
<td>0.9881</td>
<td>1.00000</td>
</tr>
</tbody>
</table>

The four DTD images are highly-correlated, 92% of the information is in common

I.e., over the four days the same areas tend to have narrow and wide DTD ranges
PCA results – rotations

> pc$rotation

<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
</tr>
</thead>
<tbody>
<tr>
<td>DTD304</td>
<td>-0.4447</td>
<td>0.5893</td>
<td>0.5870</td>
<td>0.3322</td>
</tr>
<tr>
<td>DTD306</td>
<td>-0.6084</td>
<td>0.3379</td>
<td>-0.5200</td>
<td>-0.4952</td>
</tr>
<tr>
<td>DTD307</td>
<td>-0.4717</td>
<td>-0.3857</td>
<td>-0.3677</td>
<td>0.7025</td>
</tr>
<tr>
<td>DTD308</td>
<td>-0.4578</td>
<td>-0.6243</td>
<td>0.4998</td>
<td>-0.3884</td>
</tr>
</tbody>
</table>

**PC1** “intensity” of the phenomenon over all days – all signs the same (arbitrary), similar magnitudes

**PC2** contrast between first two and second two days

**PC3** contrast between middle two and end two days

**PC4** Third and first days, contrasted with second and fourth days

Note PCs are *orthogonal* (independent)
Biplots

These show **scores** of the observations as synthetic variables in the 2-PC space (observation numbers) **Loadings** of each variable in the 2-PC space shown by the arrows (longer = higher loading).

First PC is **overall intensity** across all 4 dates; other PCs show **contrasts** between dates.
Note decreasing information content as PC number increases.

There is still some information in PC2, 3, 4 that contrasts with the overall pattern.
PCs (synthetic bands) – individual stretch

This allows to visualize contrasts within a PC
Another example of synthetic images

Cordillera de la Costa, W branch Río Aragua, Aragua state, Venezuela
Conclusion

• PCA is a powerful data reduction technique

• It is linear – so if variables are highly skewed or multi-model, apply a transformation

• It can also reveal the inter-relation between variables (e.g., reflectances in various spectral bands)

• It is completely data-driven and is scene-specific

• An important choice is between standardized and unstandardized
Topic: PCA of long time series of imagery to reveal seasonality

Groundbreaking paper, 200+ citations:


- Sensor was AVHRR

- Images were monthly maximum NDVI, 1986-1988 (three full years)
Synthetic images

Green = “+” score, Red = “-”

Sign is arbitrary, here chosen to show high NDVI in green.
Loadings

PC1 All bands contribute $\approx$ equally: (overall intensity)

PC2 shows annual movement of Intertropical Convergence Zone

PC3/4 show deviations from PC2 seasonality
Interpretation

**Synthetic bands** images summarizing all original images, according to the PCs

**Loadings** relation between original images (dates, x-axis) and contribution to the PC (y-axis).

- Note decreased absolute loadings at higher PCs, this is because they represent less of the total variability of the 36 images

- **PC1**: ≈ equal contribution of all dates (overall vegetation vigour averaged over the 3 years)

- **PC2**: + correlation in N. hemisphere summer, - in winter (seasonality) – Intertropical Convergence Zone

- **PC3/4**: deviations from PC2 seasonality – note time lag of greening/senescence

- **PC5**: detecting sensor drift
Topic: PCA for a multi- to hyper-variate dataset

A small dataset of 30 soil properties observed at 87 locations in the Lake Valencia basin, Venezuela

Also recorded coördinates (not used here) and geomorphic region

Main interest is to see the **inter-relation between variables** (grouping, contrasts)

- Could we only measure surface soil properties w.o. loss of information?

- Could we omit some (expensive?) measurements w.o. loss of information?

```r
> dim(dlv.r)
[1] 87 33
> names(dlv.r)
[1] "utm.n"  "utm.e"  "r.geo"  "dens.a"  "vcs.a"  "cs.a"  "ms.a"  "fs.a"
[9] "vfs.a"  "sand.a"  "silt.a"  "clay.a"  "oc.a"  "ph.a"  "cec.a"  "caco3.a"
[17] "p.a"    "k.a"    "dens.b"  "vcs.b"  "cs.b"  "ms.b"  "fs.b"   "vfs.b"
[25] "sand.b" "silt.b"  "clay.b"  "oc.b"   "ph.b"  "cec.b"  "caco3.b" "p.b"
[33] "k.b"
```
Geomorphic regions explain some variation

```r
> ## boxplot of bulk density of the subsoil, by geomorphic region
> require(ggplot2)
> qplot(x=as.factor(r.geo), y=dens.b, data=dlv.r, geom=c("boxplot", "point"), shape=I(3))
```

![Boxplot of bulk density by geomorphic region](image)
Computing standardized PCs

> pc <- prcomp(dlv.r[4:33], center=TRUE, scale.=TRUE)
> summary(pc)

Importance of components:

<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
<th>PC5</th>
<th>PC6</th>
<th>PC7</th>
<th>PC8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Values</td>
<td>3.1651</td>
<td>1.9973</td>
<td>1.8312</td>
<td>1.48272</td>
<td>1.33402</td>
<td>1.14400</td>
<td>1.06427</td>
<td>1.01107</td>
</tr>
<tr>
<td>Proportion of Variance</td>
<td>0.3339</td>
<td>0.1330</td>
<td>0.1118</td>
<td>0.07328</td>
<td>0.05932</td>
<td>0.04362</td>
<td>0.03776</td>
<td>0.03408</td>
</tr>
<tr>
<td>Cumulative Proportion</td>
<td>0.3339</td>
<td>0.4669</td>
<td>0.5787</td>
<td>0.65195</td>
<td>0.71127</td>
<td>0.75490</td>
<td>0.79265</td>
<td>0.82673</td>
</tr>
<tr>
<td></td>
<td>PC9</td>
<td>PC10</td>
<td>PC11</td>
<td>PC12</td>
<td>PC13</td>
<td>PC14</td>
<td>PC15</td>
<td>PC16</td>
</tr>
<tr>
<td>Values</td>
<td>0.87600</td>
<td>0.79861</td>
<td>0.76065</td>
<td>0.71565</td>
<td>0.68304</td>
<td>0.60664</td>
<td>0.5797</td>
<td>0.52760</td>
</tr>
<tr>
<td>Proportion of Variance</td>
<td>0.02558</td>
<td>0.02126</td>
<td>0.01929</td>
<td>0.01707</td>
<td>0.01555</td>
<td>0.01227</td>
<td>0.0112</td>
<td>0.00928</td>
</tr>
<tr>
<td>Cumulative Proportion</td>
<td>0.85231</td>
<td>0.87357</td>
<td>0.89285</td>
<td>0.90992</td>
<td>0.92548</td>
<td>0.93774</td>
<td>0.9489</td>
<td>0.95822</td>
</tr>
<tr>
<td></td>
<td>PC17</td>
<td>PC18</td>
<td>PC19</td>
<td>PC20</td>
<td>PC21</td>
<td>PC22</td>
<td>PC23</td>
<td>PC24</td>
</tr>
<tr>
<td>Values</td>
<td>0.50731</td>
<td>0.43862</td>
<td>0.42948</td>
<td>0.38868</td>
<td>0.33886</td>
<td>0.3146</td>
<td>0.28247</td>
<td>0.25723</td>
</tr>
<tr>
<td>Proportion of Variance</td>
<td>0.00858</td>
<td>0.00641</td>
<td>0.00615</td>
<td>0.00504</td>
<td>0.00383</td>
<td>0.0033</td>
<td>0.00266</td>
<td>0.00221</td>
</tr>
<tr>
<td>Cumulative Proportion</td>
<td>0.96680</td>
<td>0.97322</td>
<td>0.97936</td>
<td>0.98440</td>
<td>0.98823</td>
<td>0.9915</td>
<td>0.99418</td>
<td>0.99639</td>
</tr>
<tr>
<td></td>
<td>PC25</td>
<td>PC26</td>
<td>PC27</td>
<td>PC28</td>
<td>PC29</td>
<td>PC30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Values</td>
<td>0.20782</td>
<td>0.1647</td>
<td>0.14647</td>
<td>0.11458</td>
<td>0.04917</td>
<td>0.03113</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proportion of Variance</td>
<td>0.00144</td>
<td>0.0009</td>
<td>0.00072</td>
<td>0.00044</td>
<td>0.00008</td>
<td>0.00003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cumulative Proportion</td>
<td>0.99783</td>
<td>0.9987</td>
<td>0.99945</td>
<td>0.99989</td>
<td>0.99997</td>
<td>1.00000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

12 PCs (out of 30) explained 90% of the standardized variance of 30 standardized variables
> screeplot(pc,npcs=16)
Synthetic variables

The standardized variables are multiplied by these factors to make up the synthetic variables (PCs)

> print(pc$rotation[,1:4])

<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
</tr>
</thead>
<tbody>
<tr>
<td>dens.a</td>
<td>0.18958475</td>
<td>0.31398497</td>
<td>-0.09623236</td>
<td>0.08835655</td>
</tr>
<tr>
<td>vcs.a</td>
<td>0.07924278</td>
<td>0.04038316</td>
<td>0.23037009</td>
<td>-0.04841979</td>
</tr>
<tr>
<td>cs.a</td>
<td>0.16208849</td>
<td>-0.12027543</td>
<td>0.36370566</td>
<td>-0.07868327</td>
</tr>
<tr>
<td>ms.a</td>
<td>0.25194385</td>
<td>-0.12848505</td>
<td>0.21008011</td>
<td>0.00832416</td>
</tr>
<tr>
<td>fs.a</td>
<td>0.28293845</td>
<td>-0.13631074</td>
<td>-0.01492522</td>
<td>0.05422037</td>
</tr>
<tr>
<td>vfs.a</td>
<td>0.27325614</td>
<td>-0.09108109</td>
<td>-0.06474594</td>
<td>0.04104924</td>
</tr>
<tr>
<td>sand.a</td>
<td>0.22462929</td>
<td>-0.22823542</td>
<td>-0.12208377</td>
<td>0.00453792</td>
</tr>
<tr>
<td>silt.a</td>
<td>-0.04055455</td>
<td>-0.03393146</td>
<td>0.09453285</td>
<td>-0.12836263</td>
</tr>
<tr>
<td>clay.a</td>
<td>-0.19025200</td>
<td>0.24511807</td>
<td>0.03729835</td>
<td>0.09628884</td>
</tr>
<tr>
<td>oc.a</td>
<td>-0.22319928</td>
<td>-0.09812177</td>
<td>0.11735081</td>
<td>-0.10324012</td>
</tr>
<tr>
<td>ph.a</td>
<td>-0.02002915</td>
<td>-0.16078982</td>
<td>-0.34144997</td>
<td>-0.03262936</td>
</tr>
<tr>
<td>cec.a</td>
<td>-0.11061555</td>
<td>-0.31369218</td>
<td>0.08430074</td>
<td>0.07384959</td>
</tr>
<tr>
<td>caco3.a</td>
<td>-0.16458980</td>
<td>-0.33225111</td>
<td>-0.08797103</td>
<td>-0.21477108</td>
</tr>
<tr>
<td>p.a</td>
<td>-0.04422812</td>
<td>-0.17512030</td>
<td>0.04131721</td>
<td>0.49348626</td>
</tr>
<tr>
<td>k.a</td>
<td>-0.14509102</td>
<td>-0.11296314</td>
<td>0.06868405</td>
<td>0.38938453</td>
</tr>
<tr>
<td>dens.b</td>
<td>0.21263155</td>
<td>0.28604769</td>
<td>-0.08179030</td>
<td>0.11929324</td>
</tr>
<tr>
<td>vcs.b</td>
<td>0.09246395</td>
<td>0.01244499</td>
<td>0.41692142</td>
<td>-0.09615635</td>
</tr>
<tr>
<td>cs.b</td>
<td>0.14040571</td>
<td>-0.00038048</td>
<td>0.39081377</td>
<td>-0.05444895</td>
</tr>
<tr>
<td>ms.b</td>
<td>0.23184462</td>
<td>-0.08208393</td>
<td>0.21322777</td>
<td>-0.00210617</td>
</tr>
<tr>
<td>fs.b</td>
<td>0.26409667</td>
<td>-0.14157130</td>
<td>-0.04843741</td>
<td>0.07295430</td>
</tr>
<tr>
<td>vfs.b</td>
<td>0.27090087</td>
<td>-0.08083178</td>
<td>-0.11666254</td>
<td>0.07852025</td>
</tr>
<tr>
<td>sand.b</td>
<td>0.25024710</td>
<td>-0.18340556</td>
<td>-0.09229342</td>
<td>0.04529317</td>
</tr>
<tr>
<td>silt.b</td>
<td>-0.16691593</td>
<td>-0.03215390</td>
<td>0.14970295</td>
<td>-0.01619294</td>
</tr>
<tr>
<td>Variable</td>
<td>Component 1</td>
<td>Component 2</td>
<td>Component 3</td>
<td>Component 4</td>
</tr>
<tr>
<td>----------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
</tr>
<tr>
<td>clay</td>
<td>-0.1968</td>
<td>0.2769</td>
<td>-0.0003</td>
<td>-0.0466</td>
</tr>
<tr>
<td>oc</td>
<td>-0.1817</td>
<td>-0.1227</td>
<td>0.1911</td>
<td>0.0868</td>
</tr>
<tr>
<td>ph</td>
<td>0.0754</td>
<td>-0.1272</td>
<td>-0.3265</td>
<td>-0.0362</td>
</tr>
<tr>
<td>cec</td>
<td>-0.1593</td>
<td>-0.2679</td>
<td>-0.0194</td>
<td>-0.0503</td>
</tr>
<tr>
<td>caco3</td>
<td>-0.1435</td>
<td>-0.3046</td>
<td>-0.0261</td>
<td>-0.2908</td>
</tr>
<tr>
<td>p</td>
<td>-0.0707</td>
<td>-0.0569</td>
<td>0.1007</td>
<td>0.4387</td>
</tr>
<tr>
<td>k</td>
<td>-0.1522</td>
<td>-0.1233</td>
<td>-0.0062</td>
<td>0.4048</td>
</tr>
</tbody>
</table>
Biplot

> plot.it <- function(which) {
+   g <- ggbiplot(pc, choice=which, obs.scale = 1, var.scale = 1,
+                  groups = r.geo.f, ellipse = TRUE, circle = TRUE)
+   g <- g + scale_color_discrete(name = '') + theme(legend.direction = 'horizontal',
+                  legend.position = 'top')
+   print(g)}
> plot.it(1:2)
> plot.it(3:4)

(see next slides)
- normal ellipse of each group in PC space; 1 s.d. by default
- distance between the points \( \approx \) Mahalanobis distance; inner product between the variables \( \approx \) correlation
- graph produced with \texttt{ggbiplot} package
Interpretation of PC1 and 2

- strong correlations between top/sub for most properties
- sand fractions highly correlated
- clay opposite sand
- bulk density opposite CEC/OC/silt/CaCO3
- PC1 +dense, sandy, low silt, low OC
- PC2 +low fertility, base saturation, carbonates
- But these are not exactly aligned with the PCs
- Geomorphic regions cluster points in PC1/2 space
- especially 2 (piedmont) and 7 (recently-emerged lake sediments)
- PC3 contrasts top and subsoil sand
- PC4 contrasts K, P with CaCO3
Limitations of PCA:

• PCA is a **data reduction** technique

• It does not impose any **structure** on the PCs or synthetic variables, they come out directly from the eigen decomposition of the correlation or covariance matrix

• So, the resulting PCs or synthetic variables may not be easily **interpretable**
  
  - the loadings may not line up with any of the axes – see Lake Valencia example
  
  - Clear associations of variables but none lined up with a PC
Latent variable analysis – concept

• “Factor analysis” as used in social sciences

• Hypothesis: the set of observed variables is a measureable expression of some (smaller) number of latent variables

  - These can not be directly measured,
  - They influence a number of the observed variables.
  - Example: “math ability”, “ability to think abstractly” are assumed to exist (based on external evidence or theory) and measured with various tests (observed variables).

• The set of observed variables can be analyzed with PCA and then (1) reduced; (2) rotated into interpretable components
Latent variable analysis – computation

- from \( k \) observed variables hypothesize \( p < k \) latent variables ("factors")

- decompose \( k \times k \) variance-covariance matrix \( \Sigma \) of the original variables into:
  1. a \( p \times k \) loadings matrix \( \Lambda \) (the \( k \) columns are the original variables, the \( p \) rows are the latent variables);
  2. and a \( k \times k \) diagonal matrix of unexplained variances per original variable (its uniqueness) \( \Psi \), so \( \Sigma = \Lambda^\prime \Lambda + \Psi \)

- In PCA \( p = k \), there is no \( \Psi \), and all variance is explained by the synthetic variables; there is only one way to do this.

- In factor analysis, the loadings matrix \( \Lambda \) is not unique
  - it can be multiplied by any \( k \times k \) orthogonal matrix, known as rotations.
  - The factor analysis algorithm finds a rotation to satisfy user-specified conditions; e.g., varimax.
Latent variables – example

Lake Valencia, topsoils only; 15 observed variables

Hypothesize two latent variables
(1) particle-size distribution (texture), (2) reaction (pH, carbonates)

```r
> (fa <- factanal(dlv.r[4:18], 2))
```

Uniquenesses:

<table>
<thead>
<tr>
<th>dens.a</th>
<th>vcs.a</th>
<th>cs.a</th>
<th>ms.a</th>
<th>fs.a</th>
<th>vfs.a</th>
<th>sand.a</th>
<th>silt.a</th>
<th>clay.a</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.840</td>
<td>0.960</td>
<td>0.683</td>
<td>0.280</td>
<td>0.030</td>
<td>0.153</td>
<td>0.243</td>
<td>0.005</td>
<td>0.218</td>
</tr>
<tr>
<td>oc.a</td>
<td>ph.a</td>
<td>cec.a</td>
<td>caco3.a</td>
<td>p.a</td>
<td>k.a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.618</td>
<td>0.983</td>
<td>0.970</td>
<td>0.911</td>
<td>0.998</td>
<td>0.891</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Cumulative Var 0.325 0.414

Loadings:

<table>
<thead>
<tr>
<th>dens.a</th>
<th>vfs.a</th>
<th>ph.a</th>
<th>Factor1</th>
<th>Factor2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.401</td>
<td>0.919</td>
<td>-0.131</td>
<td></td>
<td></td>
</tr>
<tr>
<td>vcs.a</td>
<td>0.199</td>
<td>sand.a</td>
<td>0.843</td>
<td>-0.216</td>
</tr>
<tr>
<td>cs.a</td>
<td>0.547</td>
<td>silt.a</td>
<td>-0.136</td>
<td>0.988</td>
</tr>
<tr>
<td>ms.a</td>
<td>0.848</td>
<td>clay.a</td>
<td>-0.717</td>
<td>-0.519</td>
</tr>
<tr>
<td>fs.a</td>
<td>0.983</td>
<td>oc.a</td>
<td>-0.618</td>
<td>k.a</td>
</tr>
<tr>
<td>0.328</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Latent variables – interpretation

- **uniqueness**: \( \Psi \), noise left over after factors are fitted, i.e., variable is unexplained
  - here P, K, pH, CaCO3 most unique \( \rightarrow \) more factors are needed

- **variance explained** as in PCA. Note in PCA \( k = p \) and all variance is eventually explained

- **loadings** as in PCA
  - Factor 1 is associated with all sand fractions (coarse-textured soils) and bulk density, opposed to clay concentration and organic C
  - Factor 2 is associated with silt opposed to clay
  - So both latent variables are most associated with particle-size distribution; *not* according to our original hypothesis – this should be modified
Latent variables – plot

Factor 1 (most variance explained) was rotated to show the maximum contrast (varimax rotation)
Latent variables – DTD images example

Hypothesis: one factor, “intensity”

Uniquenesses:
CK_DTD_304  CK_DTD_306  CK_DTD_307  CK_DTD_308
     0.181       0.047       0.073       0.171

Loadings:
          Factor1
CK_DTD_304  0.905
CK_DTD_306  0.976
CK_DTD_307  0.963
CK_DTD_308  0.910

          Factor1
SS loadings  3.527
Proportion Var  0.882

Strong correlation with each image, little uniqueness, 88% of variance explained.

Compare with PCA results.
Single “best” image under the hypothesis of one process
Compare with **PCA synthetic band 1**.
Conclusion: PCA vs. latent variable analysis

**PCA** pure data reduction, maybe can interpret

**Latent variable analysis** aim is to produce interpretable variables representing the latent process