Latent Variable Methods Course

Learning from data

Instructor: Kevin Dunn
kevin.dunn@connectmv.com
http://connectmv.com

Revision: 269:35e2 compiled on 15-12-2011

Copyright, sharing, and attribution notice

We appreciate:

- if you let us know about **any errors** in the slides
- **any suggestions to improve the notes**
- telling us if you use the slides, especially commercially, so we can inform you of major updates
- emailing us to ask about different licensing terms

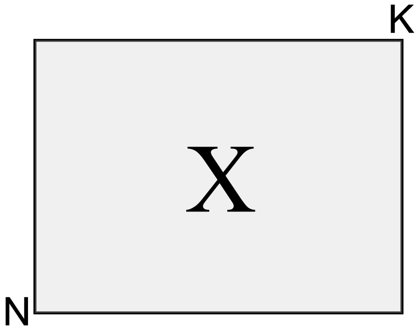
All of the above can be done by writing us at

`courses@connectmv.com`

If reporting errors/updates, please quote the current revision number: `269:35e2`

Data sources

- PCA considers a single data table (matrix)
- We will call it $\mathbf{X}$



- N observations
- K variables
- What goes in the columns of $\mathbf{X}$?
- What goes in the rows?

Visualization

How would you visualize such a data table?

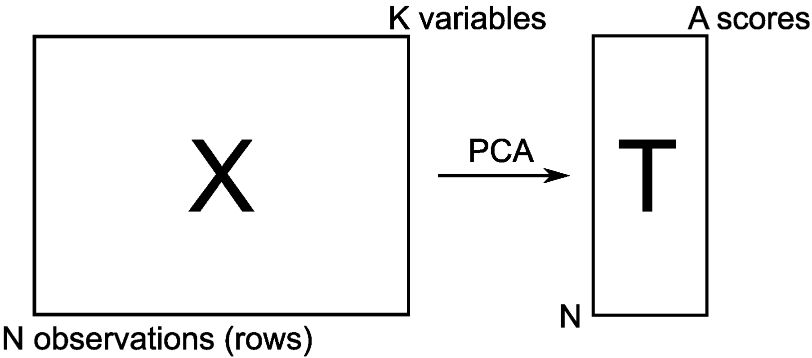
For example: assume $N = 300$ and $K = 50$

- One column at a time (time-series, histograms, boxplot)
- One row at a time (e.g. spectral data)
- Scatterplot matrix, requires $K(K-1)/2$ pairs of scatterplots

What is PCA (Principal Components Analysis)?

Mathematical objective

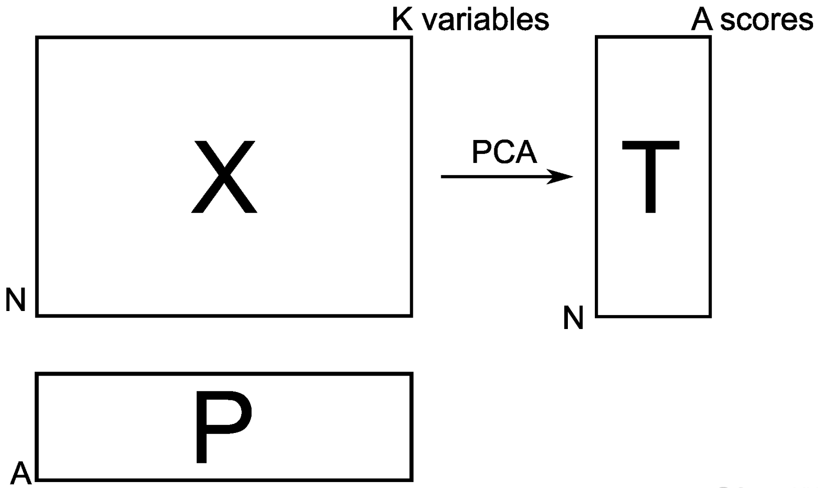
PCA: find me the best summary of my data, $\mathbf{X}$, with the fewest number of summary variables, called scores, $\mathbf{T}$.



Objectives for this class

PCA model will calculate from **X**:

- scores: **T**
- loadings: **P**



Objectives for this class

- Intuitive meaning of the scores, $\mathbf{T}$ and loadings, $\mathbf{P}$ and errors in a PCA model
- The interpretation of each of these
- How to start investigating a new data table

Time to break out the math

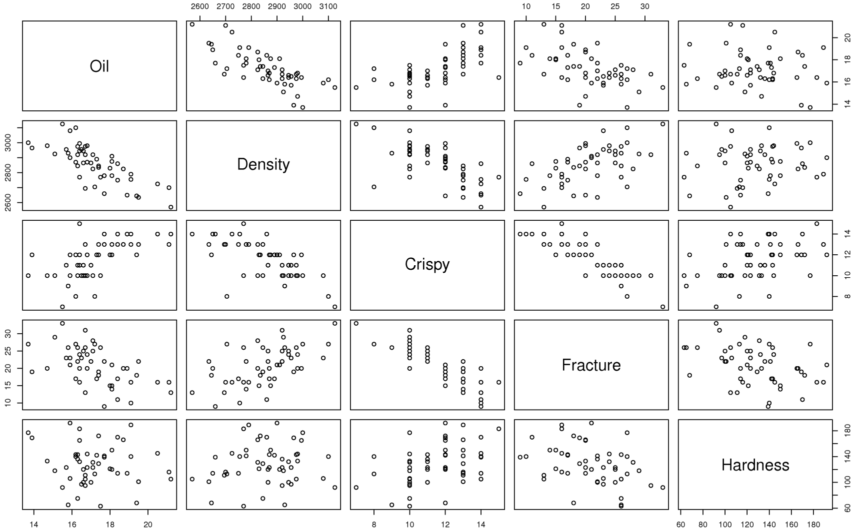
- Notation for scores: $t_{1,1}, t_{2,1}, \ldots, t_{n,1}, \ldots, t_{N,1}$
- Notation for loadings: $p_{1,1}, p_{2,1}, \ldots, p_{k,1}, \ldots, p_{K,1}$
- Length of a vector: $\|a\|$

Let's get started

1. Data preprocessing
2. Geometric interpretation (hand waving explanation)
3. Analytical geometry (to understand the hand waving)
4. Algebraic approach (to formalize the notation)
5. Look at applying this all in software

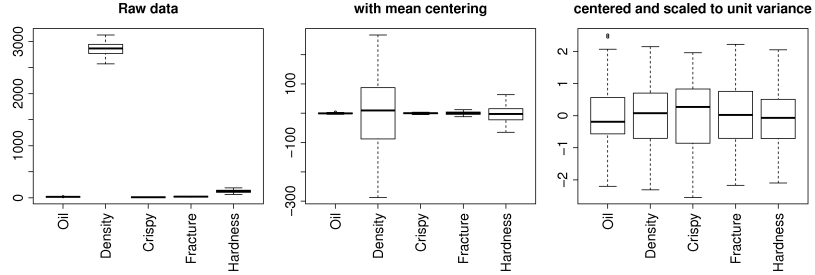
Preprocessing by example

Raw data:



Preprocessing by example

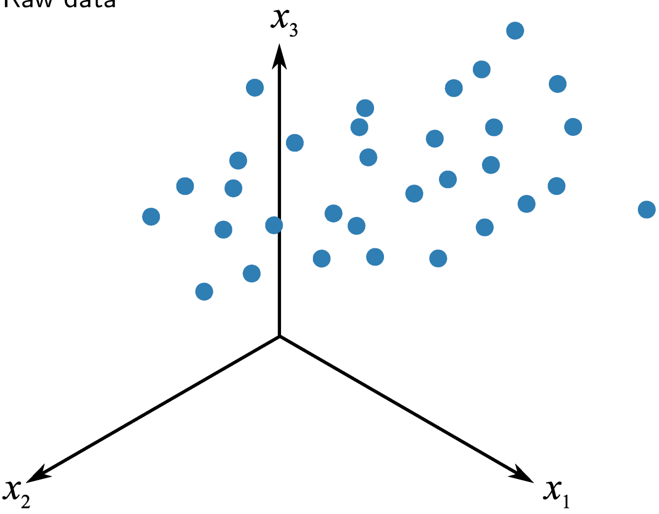
Center and scale the raw data



- Centering: $\mathbf{x}_{k,\text{center}} = \mathbf{x}_{k,\text{raw}} - \text{mean}\left(\mathbf{x}_{k,\text{raw}}\right)$
- Scaling: $\mathbf{x}_k = \dfrac{\mathbf{x}_{k,\text{center}}}{\text{standard deviation}\left(\mathbf{x}_{k,\text{center}}\right)}$
- Does *not change* relationships between variables

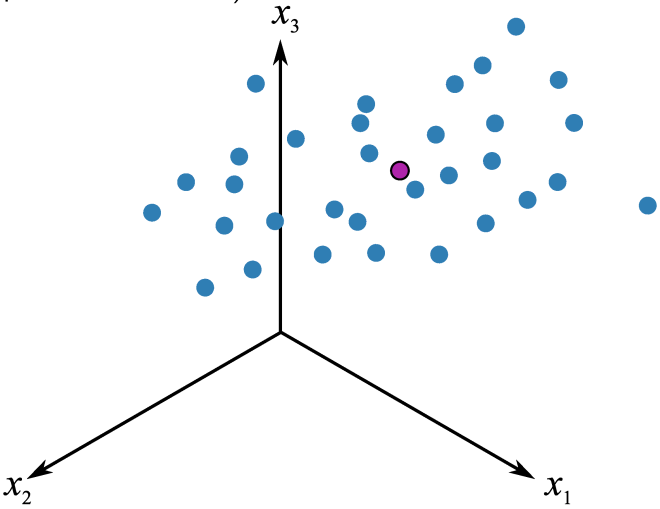
Geometric explanation of preprocessing

Raw data



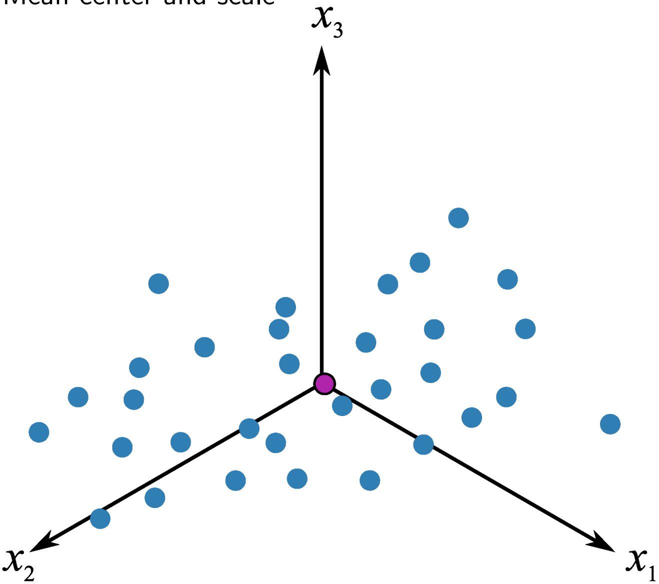
Geometric explanation of preprocessing

Calculate the mean of each variable (creates a “new” reference point in the swarm)

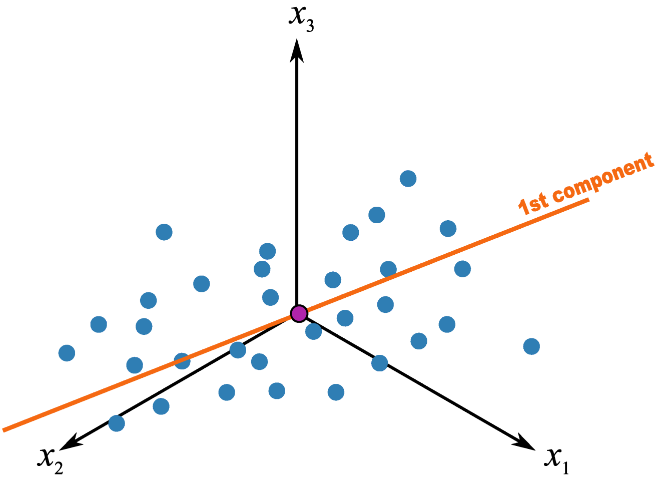


Geometric explanation of preprocessing

Mean center and scale

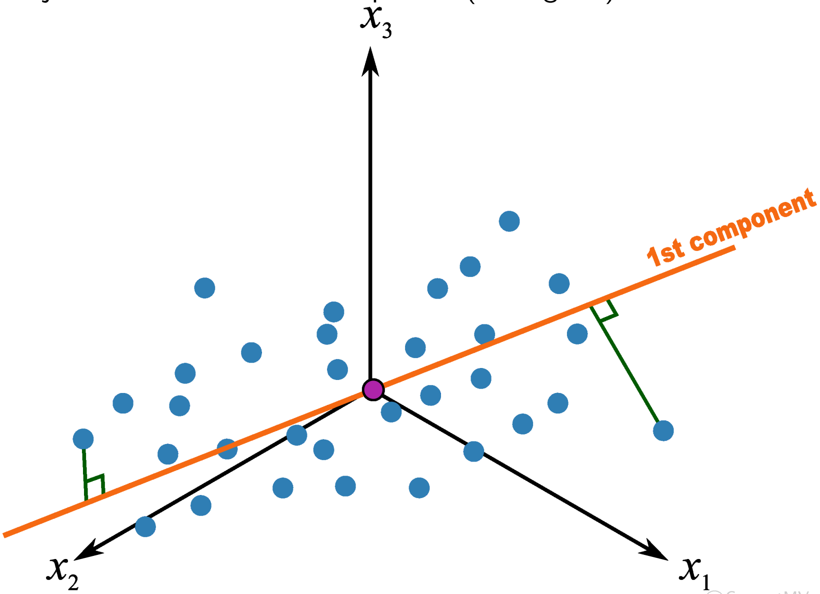


Geometric explanation of PCA



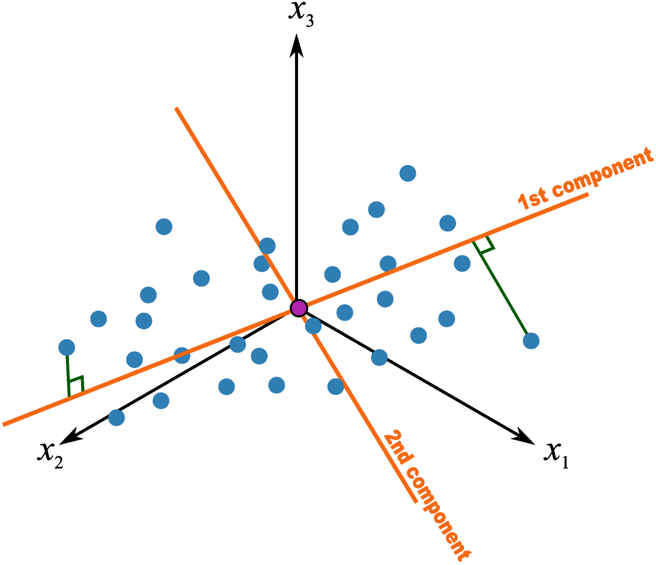
Geometric explanation of PCA

Project observations onto component (90 degrees)



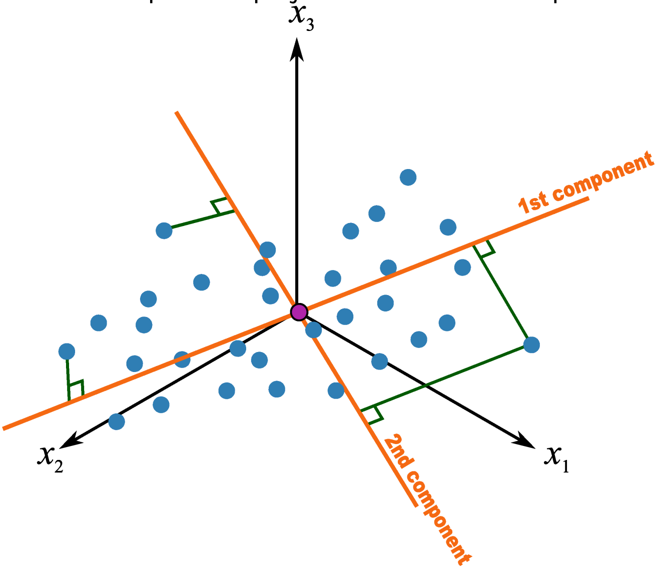
Geometric explanation of PCA

Second component: best-fit line; perpendicular to 1st component



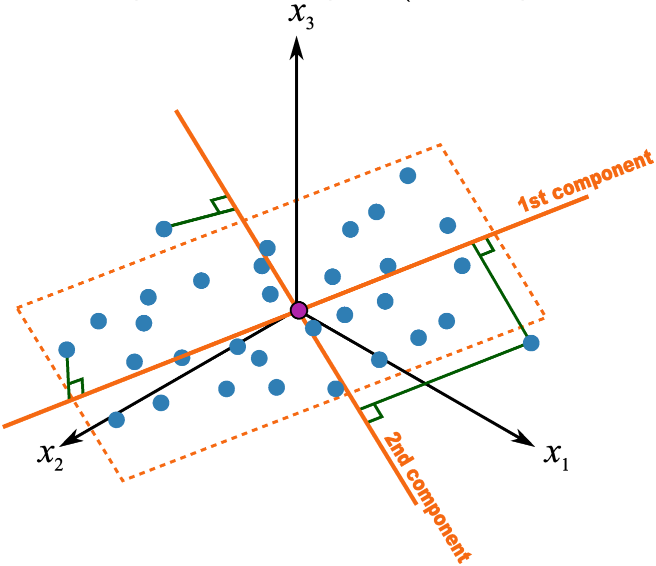
Geometric explanation of PCA

Second component: project onto second component



Geometric explanation of PCA

The 2 components form a plane (2-D subspace inside a 3D space)



Geometric explanation of PCA

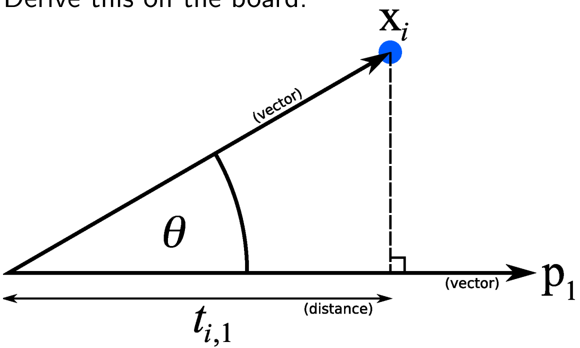
What have we done here?

Broken $\mathbf{X}$ down into 2 parts:

- projected points "*on the plane*"
- residual distance "*off the plane*"

Mathematical derivation for PCA

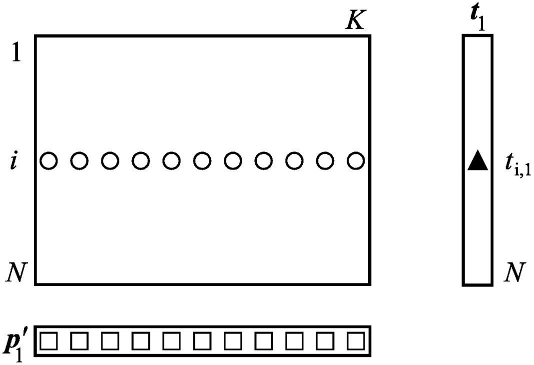
Derive this on the board:



$$\cos\theta = \frac{\text{adjacent length}}{\text{hypotenuse}} = \frac{t_{i,1}}{\|\mathbf{x}_i\|} \qquad \text{and also } \cos\theta = \frac{\mathbf{x}_i^T\mathbf{p}_1}{\|\mathbf{x}_i\|\|\mathbf{p}_1\|}$$

$$\begin{aligned} \frac{t_{i,1}}{\|\mathbf{x}_i\|} &= \frac{\mathbf{x}_i^T\mathbf{p}_1}{\|\mathbf{x}_i\|\|\mathbf{p}_1\|} \\ t_{i,1} &= \mathbf{x}_i^T\mathbf{p}_1 \\ (1\times 1) &= (1\times K)(K\times 1) \end{aligned}$$

Mathematical derivation for PCA



$$
\begin{aligned}
t_{i,1} &= \mathbf{x}_i^T \mathbf{p}_1 \\
&= x_{i,1}p_{1,1} + x_{i,2}p_{2,1} + \ldots + x_{i,k}p_{k,1} + \ldots + x_{i,K}p_{K,1}
\end{aligned}
$$

- K individual terms add up (i.e. linear combination) to give t_1
- Stack $t_{i,1}$ values from N rows: $\mathbf{T} = \mathbf{XP}$

Interpreting $t_i = \mathbf{x}_i^T \mathbf{p}_1$

Given that:

- values in $\mathbf{x}_i^T$ are centred and scaled, and
- entries in $\mathbf{p}$ are between -1 and $+1$

using

$$= \quad x_{i,1}p_{1,1} + x_{i,2}p_{2,1} + \ldots + x_{i,k}p_{k,1} + \ldots + x_{i,K}p_{K,1}$$

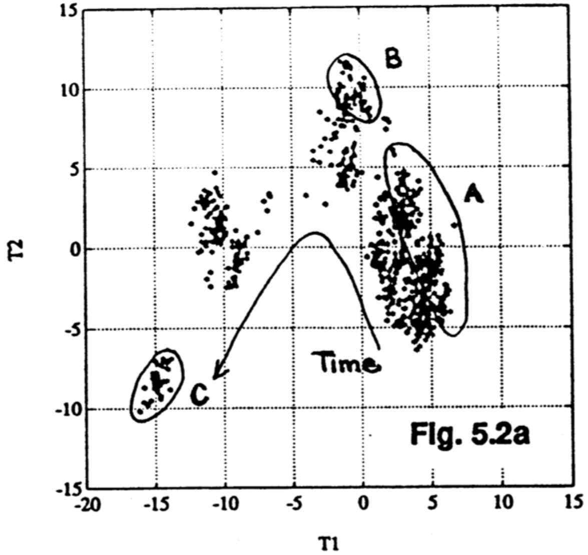
how would you

- get a large positive value of $t_{i,1}$?
- get a large negative value of $t_{i,1}$?
- get a value of $t_{i,1} \approx 0$?

- What can you say about observation (row) 13 and 22 if $t_{13,1} \approx t_{22,1}$?

Score plots: interpretation

Clustering



Score plots: interpretation

Also look for:

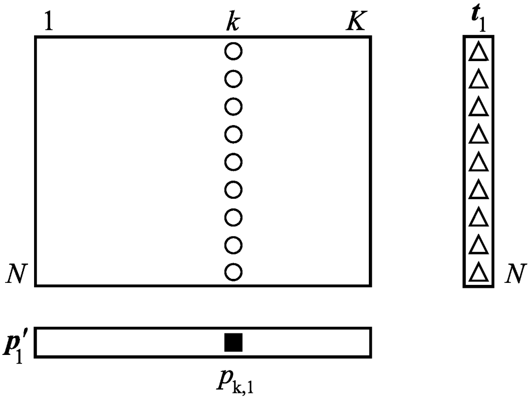
- outliers
- patterns in the sequence order (if time-based row order)
- colour-code score plots by another variable (good/bad)

We'll see more of these tips as we work with the software.

Key point: anything you would normally have done to visualize a column can be done with a score.

Mathematical derivation for PCA

We'll look at this in next class:



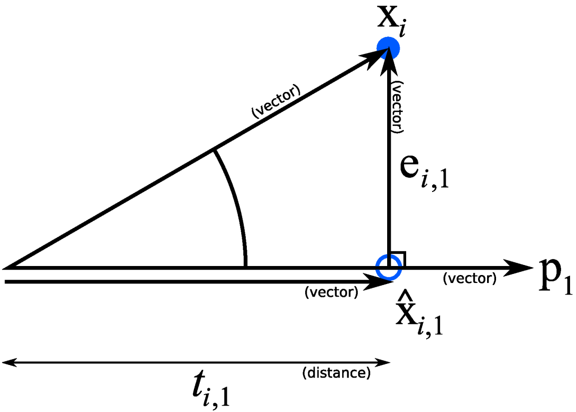
For now though:

- Columns that are related have similar loadings
- "Direction vectors" = "Loadings"
- Link between the real-world (K) and latent-variable world (A)

$$\begin{aligned} \mathbf{T} &= \mathbf{XP} \\ (N \times A) &= (N \times K)(K \times A) \end{aligned}$$

Predicted values for each observation

Take a look at this on the board as well:



$$
\begin{aligned}
\widehat{\mathbf{x}}_{i,1}^T &= t_{i,1}\mathbf{p}_1^T \\
(1 \times K) &= (1 \times 1)(1 \times K) = \text{best prediction of } \mathbf{x}_{i,1} \text{ with 1 component}
\end{aligned}
$$

In this case: "best" = "smallest error"

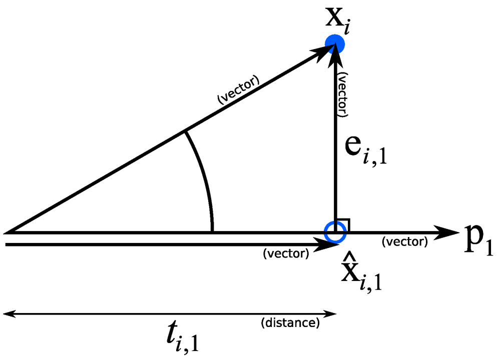
The residuals

Residuals for row i after extracting one component $= \mathbf{e}_{i,1}$

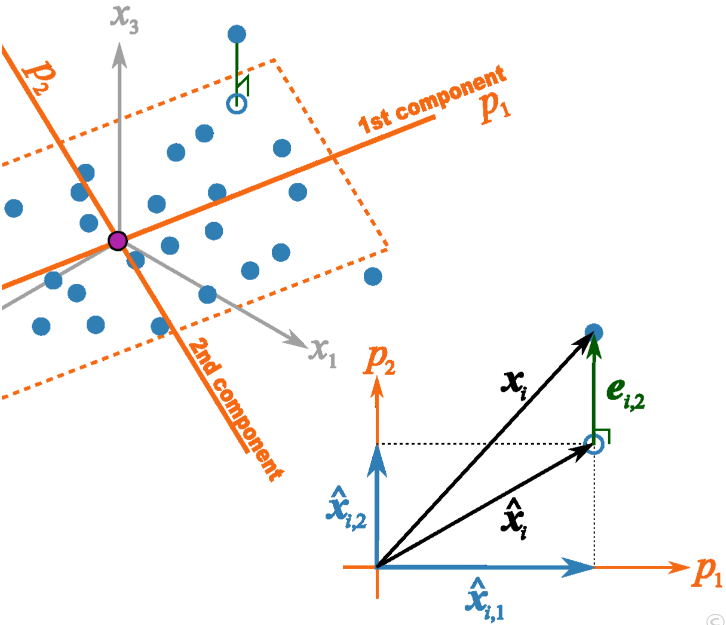
$$\mathbf{e}_{i,1}^T = \mathbf{x}_i^T - \hat{\mathbf{x}}_{i,1}^T \qquad \text{(each is a } 1 \times K \text{ vector)}$$

Another way of stating this:

$$\begin{aligned} \mathbf{x}_i^T &= \widehat{\mathbf{x}}_{i,1}^T + \mathbf{e}_{i,1}^T \\ \mathbf{x}_i^T &= t_{i,1}\mathbf{p}_1^T + \mathbf{e}_{i,1}^T \end{aligned}$$

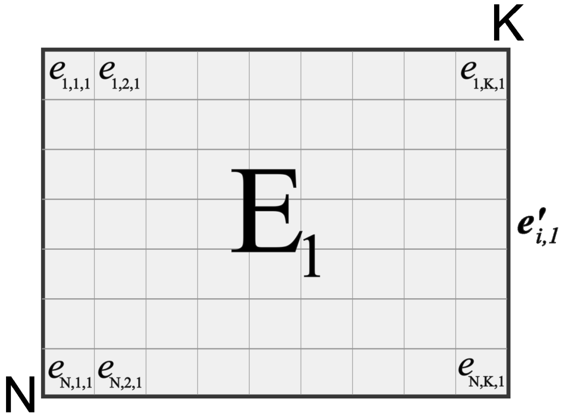


Predictions, residuals, vectors: explained



The residuals

Assemble the residuals for every row in a matrix, $\mathbf{E}_1$



The residuals

The next few slides discuss the residuals

- important part of fitting a model
- ideally, contains no information (just noise)

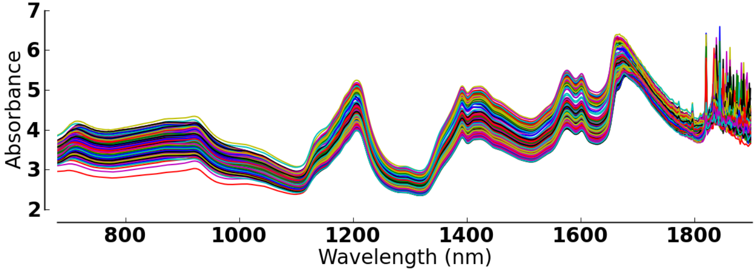
We will consider

- whole matrix residuals
- column residuals (per variable)
- row residuals (per observation)

Main way of quantifying residuals:

- calculate their sum of squares (ssq)
- in this case the ssq = variance
- and $R^2 = \dfrac{\text{variance explained by model}}{\text{initial variance}}$

Residuals: spectral example

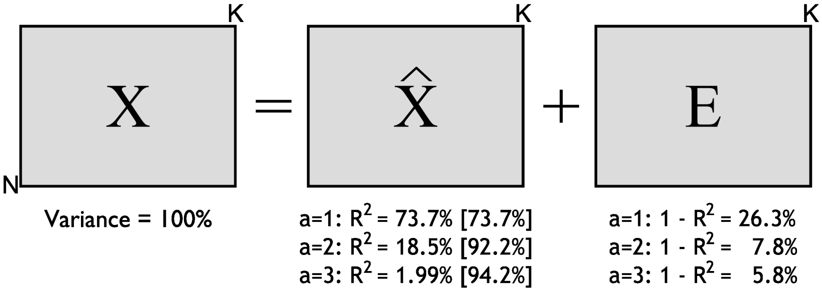


- Data on course website
- Try it yourself: http://datasets.connectmv.com/info/tablet-spectra
- N = 460
- K = 650

Whole matrix residuals

- $\mathbf{X} = \mathbf{TP}' + \mathbf{E} = \widehat{\mathbf{X}} + \mathbf{E}$
- Quantify how well the model ($\mathbf{TP}'$) fits the data
- $R_a^{2(\text{overall})} = 1 - \dfrac{\text{Var}(\mathbf{X} - \widehat{\mathbf{X}}_a)}{\text{Var}(\mathbf{X})} = 1 - \dfrac{\text{Var}(\mathbf{E}_a)}{\text{Var}(\mathbf{X})}$
  - $R^2_{a=0} = 0.0$ (no components, means no variance explained)
  - R^2 increases with every component added
  - $R_{a=1}^{2(\text{overall})} < R_{a=2}^{2(\text{overall})} < \ldots < R_{a=A}^{2(\text{overall})} = 1.0$

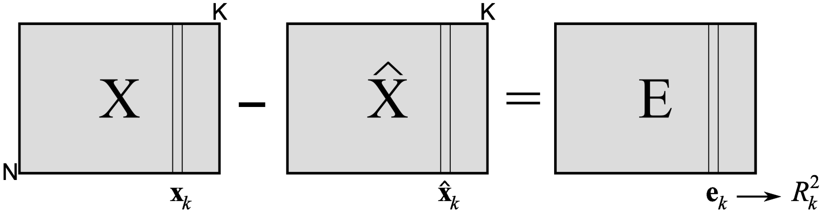
Matrix residuals: spectral example



- $R^2_{a=1} = 73.7\%$
- $R^2_{a=2} = 92.2\%$ (an additional 18.5%)
- $R^2_{a=3} = 94.2\%$ (an additional 2.00%)

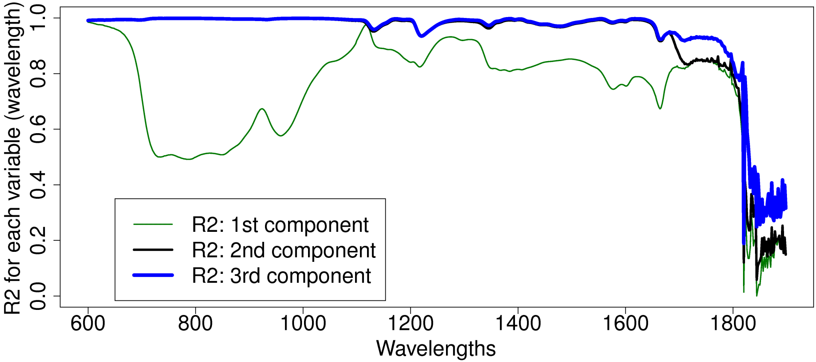
Column residuals

- R^2 can be calculated for each column

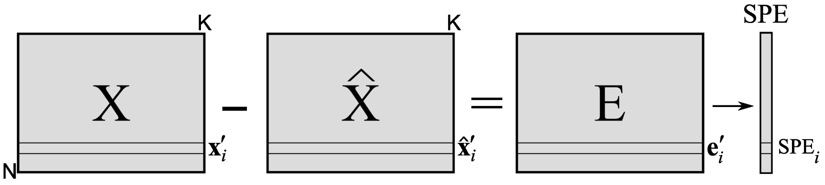


- $R^2_k = 1 - \dfrac{\text{Var}(\mathbf{x}_k - \widehat{\mathbf{x}}_k)}{\text{Var}(\mathbf{x}_k)}$
- indicates how well each column is explained by the model
- is 0.0 when there are no components
- increases for every every component added

Column residuals: spectral example



Row residuals



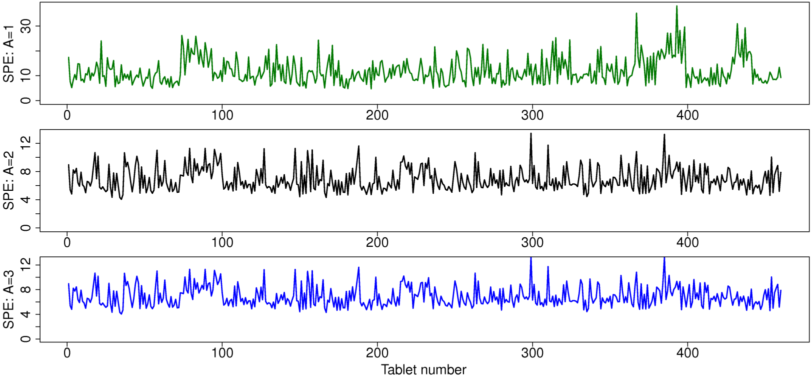
- $\mathbf{e}_i' = \mathbf{x}_i' - \widehat{\mathbf{x}}_i'$
  $\mathbf{e}_i' = \begin{bmatrix}(x_{i,1} - \hat{x}_{i,1}) & (x_{i,2} - \hat{x}_{i,2}) & \ldots & (x_{i,k} - \hat{x}_{i,k}) & \ldots & (x_{i,K} - \hat{x}_{i,K})\end{bmatrix}$
- Variance of residuals in a row $= e_{i,1}^2 + e_{i,2}^2 + \ldots + e_{i,K}^2$
- Call this SPE = squared prediction error $= \mathbf{e}_i^T \mathbf{e}_i$
- Square root of SPE = "distance to model's X-space"
- "DModX" (used in some software) is related to $\sqrt{\text{SPE}_i}$

Square prediction error

Distance from each observation to the model's plane:

- Smallest SPE ?
- Distribution of SPE values
- If SPE $>$ 95% limit:
  - poorly explained by the model
  - something new in this observation
  - new phenomenon?

Row residuals: spectral example



Data sets to look at in class

- **Website**: http://datasets.connectmv.com
- Click on the "Peas" link and download CSV file
- Click on the "Food consumption" link and download CSV file
- Click on the "Food texture" link and download CSV file

For next class

1. Assignment: instructions will be posted on course website
2. Read the paper by Wold (item 12)
   - http://literature.connectmv.com/item/12
   - This will help you understand the material in next class
   - No Q&A: but strongly recommend you are familiar with the concepts
3. Next class will cover
   - *how* we calculate the components
   - *how many* components should be calculated
   - *using* the PCA model on new data