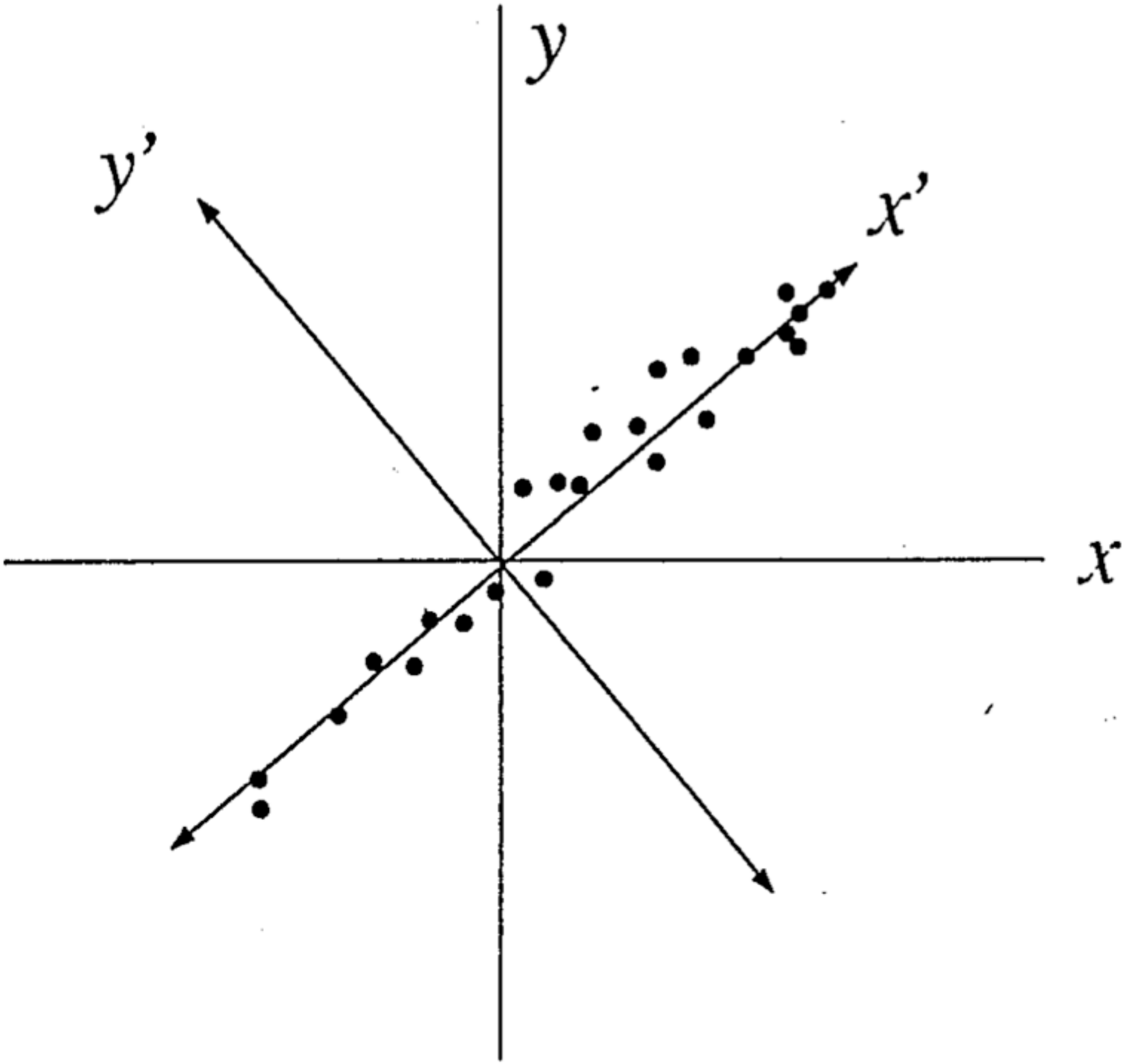
# March 10th
# Principal components analysis


y
y'
x'
x

Let **obs** be a vector-valued observation of dimensionality D
i.e. $\langle \text{measurement}_1, \text{measurement}_2, \cdots, \text{measurement}_D \rangle$
assume we have $n$ such observations

Let **obs** be a vector-valued observation of dimensionality D
i.e. $\langle \text{measurement}_1, \text{measurement}_2, \cdots, \text{measurement}_D \rangle$
assume we have $n$ such observations

$$\mathbf{A} = \sum_{i=1}^{n} \left( \mathbf{obs}_i \mathbf{obs}_i^T \right)$$

# outer product

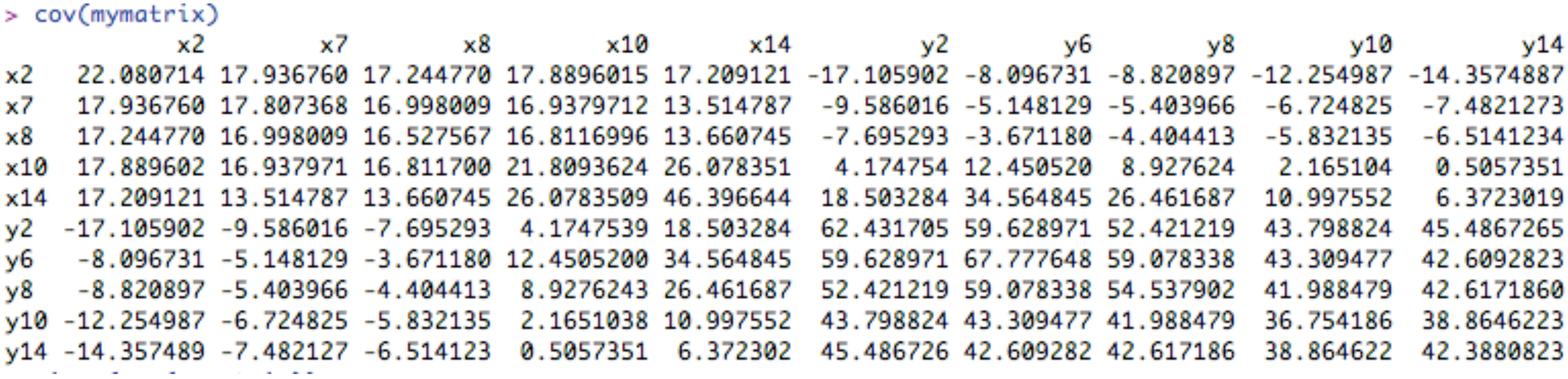
$$\mathbf{u} \otimes \mathbf{v} = \mathbf{u}\mathbf{v}^T = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix} = \begin{bmatrix} u_1v_1 & u_1v_2 & u_1v_3 \\ u_2v_1 & u_2v_2 & u_2v_3 \\ u_3v_1 & u_3v_2 & u_3v_3 \\ u_4v_1 & u_4v_2 & u_4v_3 \end{bmatrix}$$

# Eigenvector, eigenvalue

$$A\vec{x} = \lambda\vec{x}$$

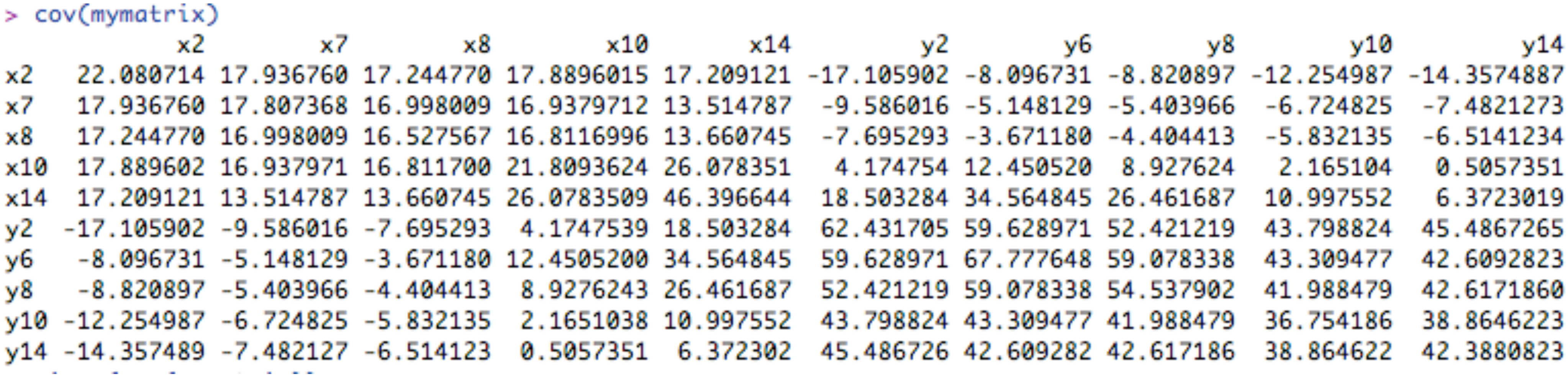
**i.e.**

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1\,k} \\ a_{21} & a_{22} & \cdots & a_{2\,k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k\,1} & a_{k\,2} & \cdots & a_{k\,k} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix},$$



```
> cov(mymatrix)
            x2        x7        x8        x10       x14         y2        y6        y8        y10         y14
x2   22.080714 17.936760 17.244770 17.8896015 17.209121 -17.105902 -8.096731 -8.820897 -12.254987 -14.3574887
x7   17.936760 17.807368 16.998009 16.9379712 13.514787  -9.586016 -5.148129 -5.403966  -6.724825  -7.4821273
x8   17.244770 16.998009 16.527567 16.8116996 13.660745  -7.695293 -3.671180 -4.404413  -5.832135  -6.5141234
x10  17.889602 16.937971 16.811700 21.8093624 26.078351   4.174754 12.450520  8.927624   2.165104   0.5057351
x14  17.209121 13.514787 13.660745 26.0783509 46.396644  18.503284 34.564845 26.461687  10.997552   6.3723019
y2  -17.105902 -9.586016 -7.695293  4.1747539 18.503284  62.431705 59.628971 52.421219  43.798824  45.4867265
y6   -8.096731 -5.148129 -3.671180 12.4505200 34.564845  59.628971 67.777648 59.078338  43.309477  42.6092823
y8   -8.820897 -5.403966 -4.404413  8.9276243 26.461687  52.421219 59.078338 54.537902  41.988479  42.6171860
y10 -12.254987 -6.724825 -5.832135  2.1651038 10.997552  43.798824 43.309477 41.988479  36.754186  38.8646223
y14 -14.357489 -7.482127 -6.514123  0.5057351  6.372302  45.486726 42.609282 42.617186  38.864622  42.3880823
```

# eigenvalues



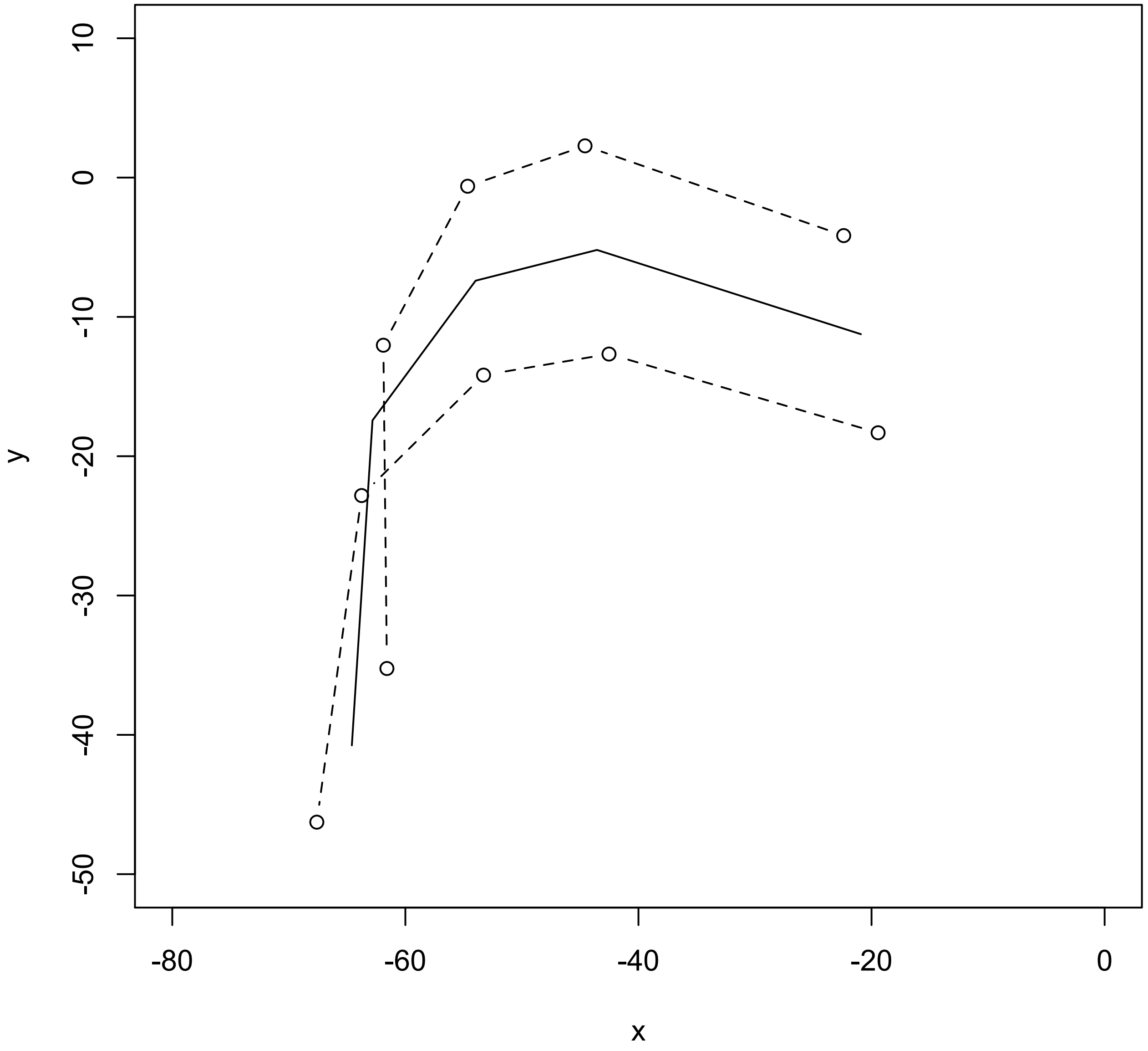
```
> cov(mymatrix)
            x2        x7        x8        x10       x14         y2        y6        y8        y10         y14
x2   22.080714 17.936760 17.244770 17.8896015 17.209121 -17.105902 -8.096731 -8.820897 -12.254987 -14.3574887
x7   17.936760 17.807368 16.998009 16.9379712 13.514787  -9.586016 -5.148129 -5.403966  -6.724825  -7.4821273
x8   17.244770 16.998009 16.527567 16.8116996 13.660745  -7.695293 -3.671180 -4.404413  -5.832135  -6.5141234
x10  17.889602 16.937971 16.811700 21.8093624 26.078351   4.174754 12.450520  8.927624   2.165104   0.5057351
x14  17.209121 13.514787 13.660745 26.0783509 46.396644  18.503284 34.564845 26.461687  10.997552   6.3723019
y2  -17.105902 -9.586016 -7.695293  4.1747539 18.503284  62.431705 59.628971 52.421219  43.798824  45.4867265
y6   -8.096731 -5.148129 -3.671180 12.4505200 34.564845  59.628971 67.777648 59.078338  43.309477  42.6092823
y8   -8.820897 -5.403966 -4.404413  8.9276243 26.461687  52.421219 59.078338 54.537902  41.988479  42.6171860
y10 -12.254987 -6.724825 -5.832135  2.1651038 10.997552  43.798824 43.309477 41.988479  36.754186  38.8646223
y14 -14.357489 -7.482127 -6.514123  0.5057351  6.372302  45.486726 42.609282 42.617186  38.864622  42.3880823
```

# eigenvalues

$$\mathrm{cov(mymatrix)} \overset{\textbf{first eigenvector}}{\begin{pmatrix} 0.0970286 \\ 0.0556497 \\ 0.0443798 \\ -0.063532 \\ -0.202193 \\ -0.471818 \\ -0.498884 \\ -0.452402 \\ -0.359686 \\ -0.367508 \end{pmatrix}} = \underset{\textbf{first eigenvalue}}{\underset{\uparrow}{258.49}} \overset{\textbf{first eigenvector}}{\begin{pmatrix} 0.0970286 \\ 0.0556497 \\ 0.0443798 \\ -0.063532 \\ -0.202193 \\ -0.471818 \\ -0.498884 \\ -0.452402 \\ -0.359686 \\ -0.367508 \end{pmatrix}}$$

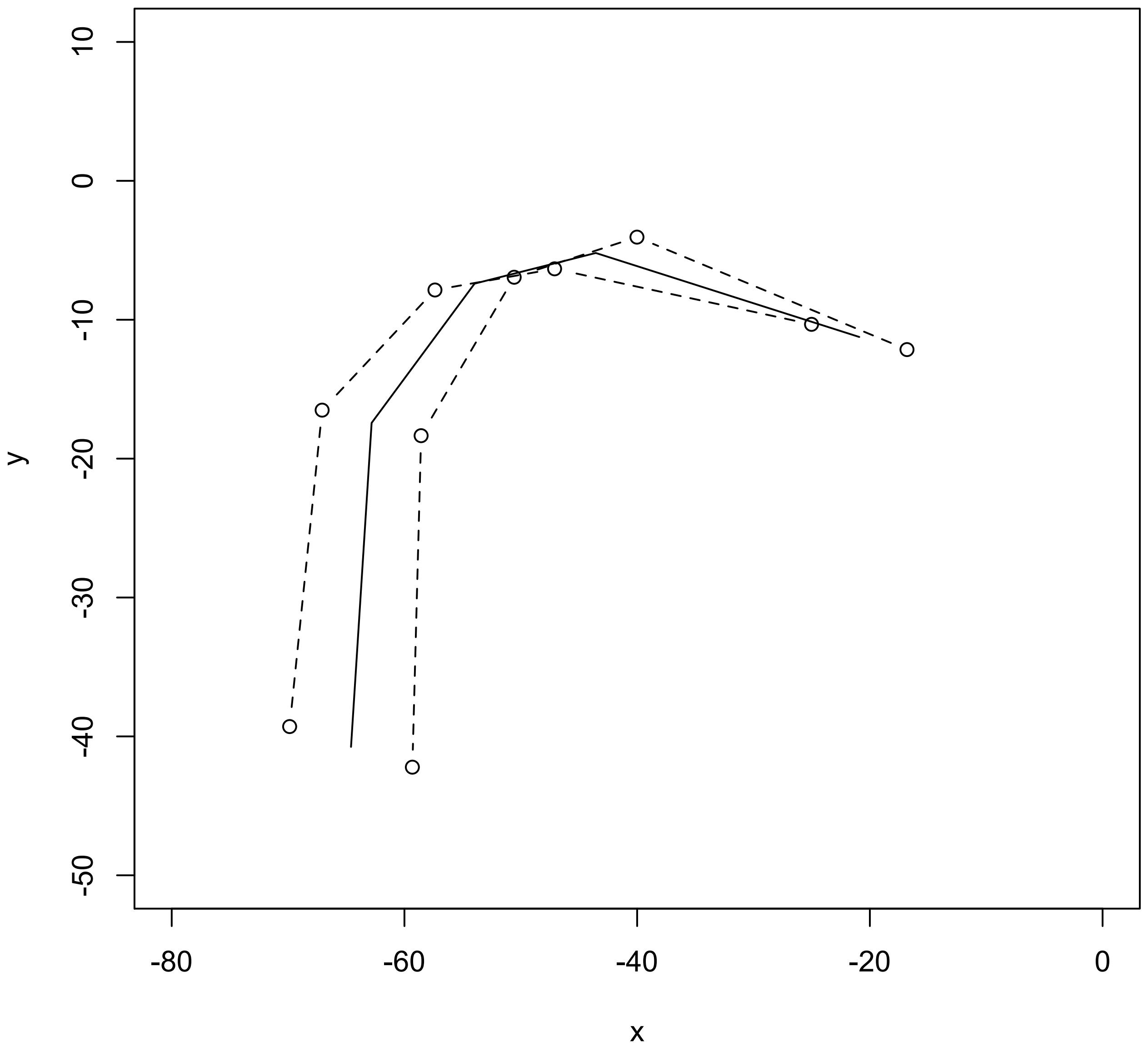
principal component #1

has a lot to do with whether the tongue is high or low



principal component #2

has a lot to do with whether the tongue is front or back



# Eigendecomposition

$$\mathbf{A} = \mathbf{P}\ \mathbf{D}\ (\mathbf{P}^{-1})$$

$$P \equiv \left[\vec{X}_1 \vec{X}_1 \cdots \vec{X}_k\right]$$

$$= \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{k\,1} \\ x_{12} & x_{22} & \cdots & x_{k\,2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1\,k} & x_{2\,k} & \cdots & x_{k\,k} \end{bmatrix}$$

$$\mathsf{D} \equiv \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix},$$

# Singular value decomposition

$$\mathbf{A} = \mathbf{U}\ \mathbf{D}\ (\mathbf{V}^T)$$

# Singular value decomposition

$$\mathbf{A} = \mathbf{U}\ \mathbf{D}\ (\mathbf{V}^T)$$

just take the top k eigenvalues

$$\mathbf{A} = \mathbf{U}_{\mathrm{k}}\ \mathbf{D}_{\mathrm{k}}\ (\mathbf{V}_{\mathbf{k}}^T)$$

# Term x document

$$\begin{array}{cc} & \mathbf{d}_j \\ & \downarrow \\ \mathbf{t}_i^T \rightarrow & \begin{bmatrix} x_{1,1} & \dots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \dots & x_{m,n} \end{bmatrix} \end{array}$$

# Term x document

$$\begin{array}{cc} & \begin{array}{ccc} & \mathbf{d}_j & \\ & \downarrow & \end{array} \\ \mathbf{t}_i^T \rightarrow & \begin{bmatrix} x_{1,1} & \dots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \dots & x_{m,n} \end{bmatrix} \end{array}$$

$$d_j = \langle \underset{\text{chicken}}{8}, \underset{\text{fried}}{2}, \underset{\text{oil}}{7}, \underset{\text{pepper}}{4} \rangle$$

# Term x document

$$\begin{array}{cc} & \begin{array}{ccc} & \mathbf{d}_j & \\ & \downarrow & \end{array} \\ \mathbf{t}_i^T \rightarrow & \begin{bmatrix} x_{1,1} & \dots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \dots & x_{m,n} \end{bmatrix} \end{array}$$

$$d_j = \langle \underset{\text{chicken}}{8}, \underset{\text{fried}}{2}, \underset{\text{oil}}{7}, \underset{\text{pepper}}{4} \rangle$$

**find me a document about "fried chicken":**

$$q = \langle 1, 1, 0, 0 \rangle$$

# Term x document

$$\begin{array}{cc} & \begin{matrix} & \mathbf{d}_j & \\ & \downarrow & \end{matrix} \\ \mathbf{t}_i^T \rightarrow & \begin{bmatrix} x_{1,1} & \dots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \dots & x_{m,n} \end{bmatrix} \end{array}$$

$$d_j = \langle \underset{\text{chicken}}{8}, \underset{\text{fried}}{2}, \underset{\text{oil}}{7}, \underset{\text{pepper}}{4} \rangle$$

**find me a document about “fried chicken”:**

$$q = \langle 1, 1, 0, 0 \rangle$$

➠ **the $j$ that maximizes $d \cdot q$**

# Higher weights to terms that are more telling

$$w_{i,j} = tf_{i,j} \times idf_i$$

$$idf_i = \log \left( \frac{\boxed{N}}{\boxed{n_i}} \right)$$

total # of docs

# docs in which this term appears

# LSA

Example of text data: Titles of Some Technical Memos

c1: *Human* machine *interface* for ABC *computer* applications
c2: A *survey* of *user* opinion of *computer system response time*
c3: The *EPS user interface* management *system*
c4: *System* and *human system* engineering testing of *EPS*
c5: Relation of *user* perceived *response time* to error measurement

m1: The generation of random, binary, ordered *trees*
m2: The intersection *graph* of paths in *trees*
m3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4: *Graph minors*: A *survey*

$$\{X\} =$$

| | **c1** | **c2** | **c3** | **c4** | **c5** | **m1** | **m2** | **m3** | **m4** |
|---|---|---|---|---|---|---|---|---|---|
| **human** | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| **interface** | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| **computer** | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **user** | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| **system** | 0 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | 0 |
| **response** | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| **time** | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| **EPS** | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| **survey** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **trees** | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| **graph** | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| **minors** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |

r (human.user) = -.38

r (human.minors) = -.29

Figure 1. A word by context matrix, X, formed from the titles of five articles about human-computer interaction and four about graph theory. Cell entries are the number of times that a word (rows) appeared in a title (columns) for words that appeared in at least two titles.

$$\{X\} = \{W\}\{S\}\{P\}'$$

$\{W\} =$

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 0.22 | -0.11 | 0.29 | -0.41 | -0.11 | -0.34 | 0.52 | -0.06 | -0.41 |
| 0.20 | -0.07 | 0.14 | -0.55 | 0.28 | 0.50 | -0.07 | -0.01 | -0.11 |
| 0.24 | 0.04 | -0.16 | -0.59 | -0.11 | -0.25 | -0.30 | 0.06 | 0.49 |
| 0.40 | 0.06 | -0.34 | 0.10 | 0.33 | 0.38 | 0.00 | 0.00 | 0.01 |
| 0.64 | -0.17 | 0.36 | 0.33 | -0.16 | -0.21 | -0.17 | 0.03 | 0.27 |
| 0.27 | 0.11 | -0.43 | 0.07 | 0.08 | -0.17 | 0.28 | -0.02 | -0.05 |
| 0.27 | 0.11 | -0.43 | 0.07 | 0.08 | -0.17 | 0.28 | -0.02 | -0.05 |
| 0.30 | -0.14 | 0.33 | 0.19 | 0.11 | 0.27 | 0.03 | -0.02 | -0.17 |
| 0.21 | 0.27 | -0.18 | -0.03 | -0.54 | 0.08 | -0.47 | -0.04 | -0.58 |
| 0.01 | 0.49 | 0.23 | 0.03 | 0.59 | -0.39 | -0.29 | 0.25 | -0.23 |
| 0.04 | 0.62 | 0.22 | 0.00 | -0.07 | 0.11 | 0.16 | -0.68 | 0.23 |
| 0.03 | 0.45 | 0.14 | -0.01 | -0.30 | 0.28 | 0.34 | 0.68 | 0.18 |

$\{S\} =$

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 3.34 | | | | | | | | |
| | 2.54 | | | | | | | |
| | | 2.35 | | | | | | |
| | | | 1.64 | | | | | |
| | | | | 1.50 | | | | |
| | | | | | 1.31 | | | |
| | | | | | | 0.85 | | |
| | | | | | | | 0.56 | |
| | | | | | | | | 0.36 |

**just two concepts**

$\{P\} =$

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 0.20 | 0.61 | 0.46 | 0.54 | 0.28 | 0.00 | 0.01 | 0.02 | 0.08 |
| -0.06 | 0.17 | -0.13 | -0.23 | 0.11 | 0.19 | 0.44 | 0.62 | 0.53 |
| 0.11 | -0.50 | 0.21 | 0.57 | -0.51 | 0.10 | 0.19 | 0.25 | 0.08 |
| -0.95 | -0.03 | 0.04 | 0.27 | 0.15 | 0.02 | 0.02 | 0.01 | -0.03 |
| 0.05 | -0.21 | 0.38 | -0.21 | 0.33 | 0.39 | 0.35 | 0.15 | -0.60 |
| -0.08 | -0.26 | 0.72 | -0.37 | 0.03 | -0.30 | -0.21 | 0.00 | 0.36 |
| 0.18 | -0.43 | -0.24 | 0.26 | 0.67 | -0.34 | -0.15 | 0.25 | 0.04 |
| -0.01 | 0.05 | 0.01 | -0.02 | -0.06 | 0.45 | -0.76 | 0.45 | -0.07 |
| -0.06 | 0.24 | 0.02 | -0.08 | -0.26 | -0.62 | 0.02 | 0.52 | -0.45 |

Figure 2. Complete SVD of matrix in Figure 1.

$$\left\{\hat{X}\right\} =$$

| | c1 | c2 | c3 | c4 | c5 | m1 | m2 | m3 | m4 |
|---|---|---|---|---|---|---|---|---|---|
| human | 0.16 | 0.40 | 0.38 | 0.47 | 0.18 | -0.05 | -0.12 | -0.16 | -0.09 |
| interface | 0.14 | 0.37 | 0.33 | 0.40 | 0.16 | -0.03 | -0.07 | -0.10 | -0.04 |
| computer | 0.15 | 0.51 | 0.36 | 0.41 | 0.24 | 0.02 | 0.06 | 0.09 | 0.12 |
| user | 0.26 | 0.84 | 0.61 | 0.70 | 0.39 | 0.03 | 0.08 | 0.12 | 0.19 |
| system | 0.45 | 1.23 | 1.05 | 1.27 | 0.56 | -0.07 | -0.15 | -0.21 | -0.05 |
| response | 0.16 | 0.58 | 0.38 | 0.42 | 0.28 | 0.06 | 0.13 | 0.19 | 0.22 |
| time | 0.16 | 0.58 | 0.38 | 0.42 | 0.28 | 0.06 | 0.13 | 0.19 | 0.22 |
| EPS | 0.22 | 0.55 | 0.51 | 0.63 | 0.24 | -0.07 | -0.14 | -0.20 | -0.11 |
| survey | 0.10 | 0.53 | 0.23 | 0.21 | 0.27 | 0.14 | 0.31 | 0.44 | 0.42 |
| trees | -0.06 | 0.23 | -0.14 | -0.27 | 0.14 | 0.24 | 0.55 | 0.77 | 0.66 |
| graph | -0.06 | 0.34 | -0.15 | -0.30 | 0.20 | 0.31 | 0.69 | 0.98 | 0.85 |
| minors | -0.04 | 0.25 | -0.10 | -0.21 | 0.15 | 0.22 | 0.50 | 0.71 | 0.62 |

$\underline{r}$ (human.user) = .94

$\underline{r}$ (human.minors) = -.83

Figure 3. Two dimensional reconstruction of original matrix shown in Fig. 1 based on shaded columns and rows from SVD as shown in Fig. 2. Comparing shaded and boxed rows and cells of Figs. 1 and 3 illustrates how LSA induces similarity relations by changing estimated entries up or down to accommodate mutual constraints in the data.

# LSA takes the TOEFL

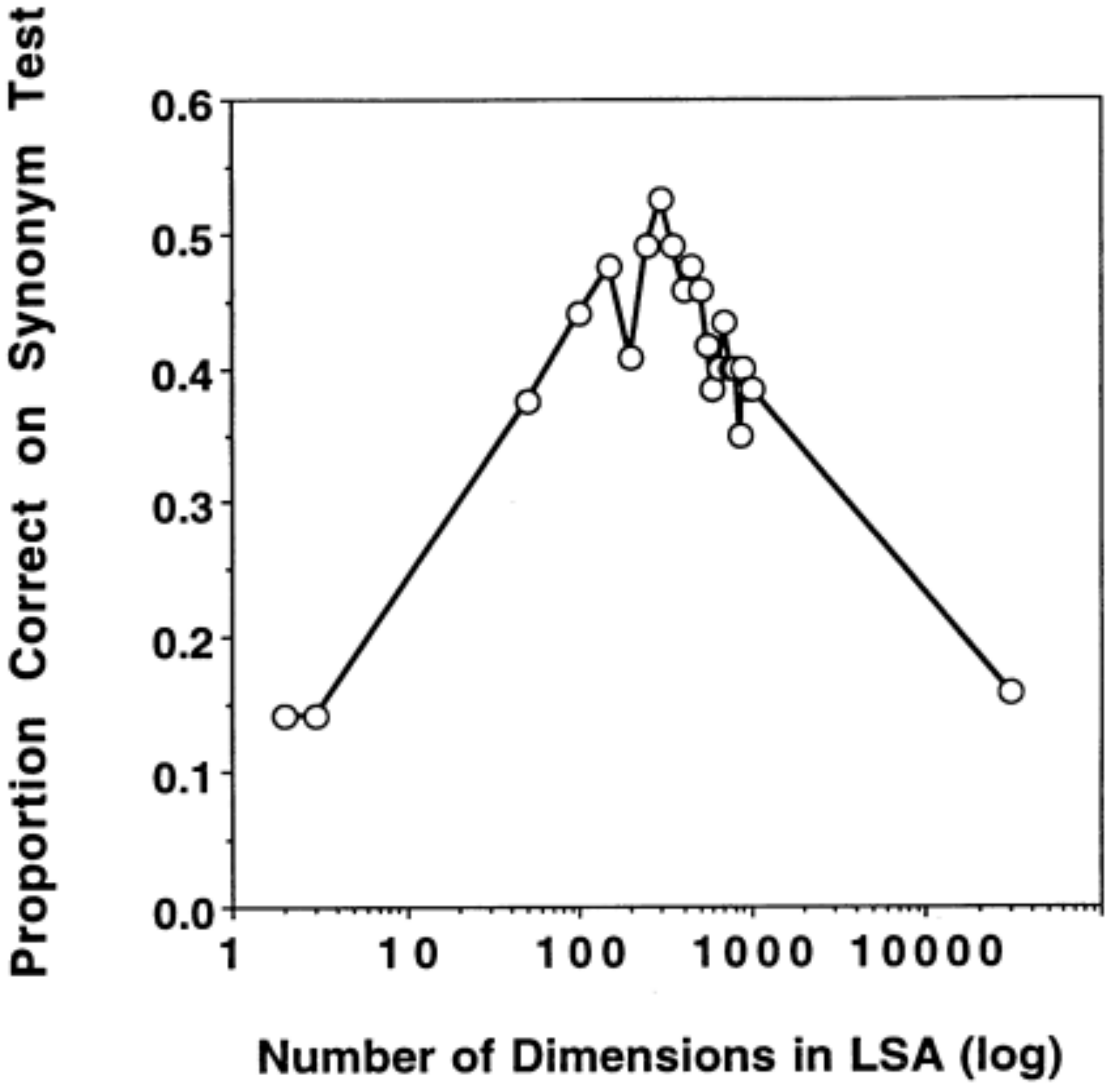
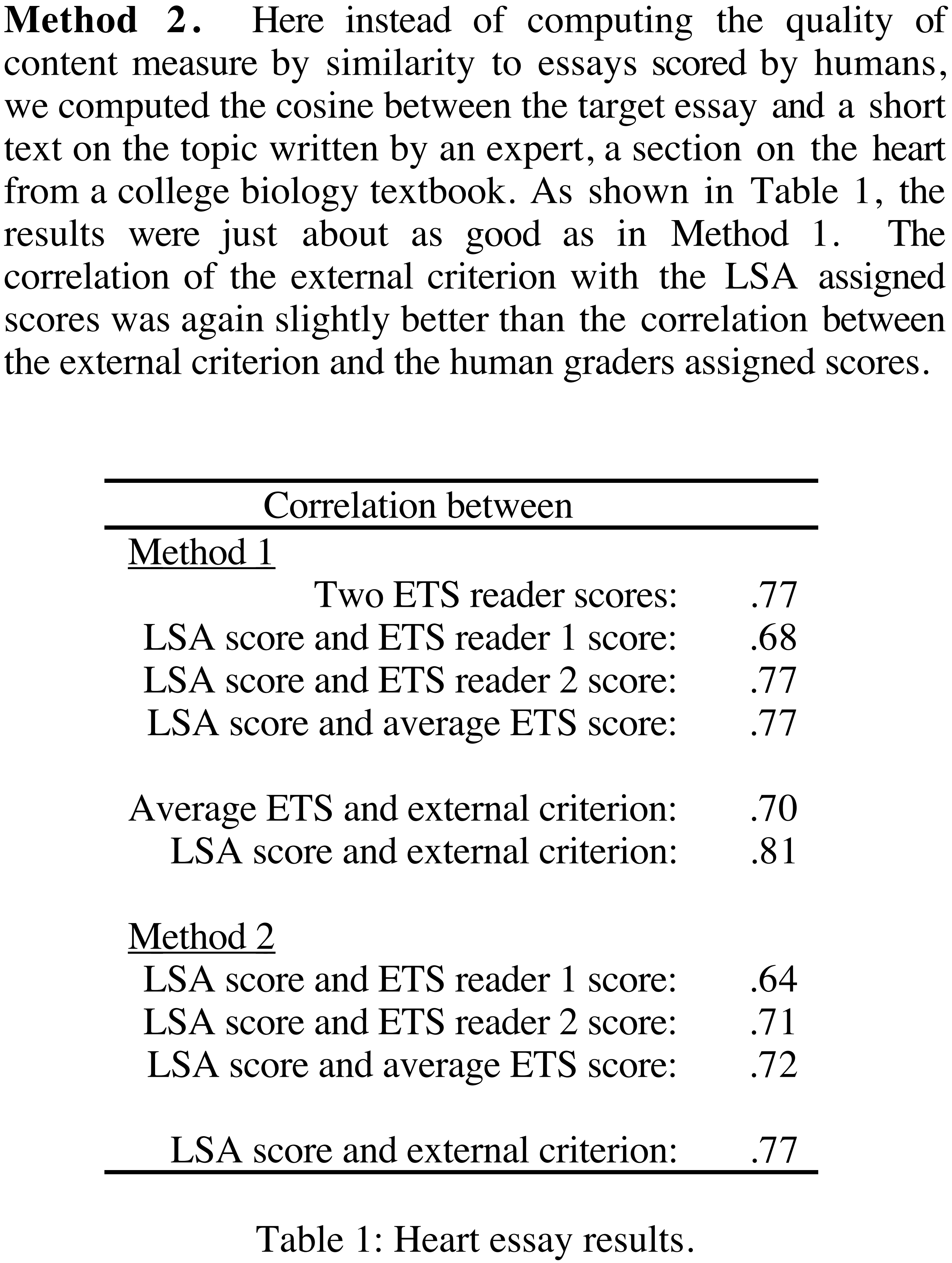


Figure 5. The effect of number of dimensions in an LSA corpus-based representation of meaning on performance on a synonym test (from ETS Test of English as a Foreign Language). The measure is the proportion of 80 multiple-choice items after standard correction for guessing. The point for the highest dimensionality is equivalent to a first-order co-occurrence correlation.

# LSA grades your essay



**Method 2.** Here instead of computing the quality of content measure by similarity to essays scored by humans, we computed the cosine between the target essay and a short text on the topic written by an expert, a section on the heart from a college biology textbook. As shown in Table 1, the results were just about as good as in Method 1. The correlation of the external criterion with the LSA assigned scores was again slightly better than the correlation between the external criterion and the human graders assigned scores.

| Correlation between | |
|---|---|
| Method 1 | |
| Two ETS reader scores: | .77 |
| LSA score and ETS reader 1 score: | .68 |
| LSA score and ETS reader 2 score: | .77 |
| LSA score and average ETS score: | .77 |
| Average ETS and external criterion: | .70 |
| LSA score and external criterion: | .81 |
| Method 2 | |
| LSA score and ETS reader 1 score: | .64 |
| LSA score and ETS reader 2 score: | .71 |
| LSA score and average ETS score: | .72 |
| LSA score and external criterion: | .77 |

Table 1: Heart essay results.