Same example as last time, but start with R instead of S:

$$R = \begin{bmatrix} 1 & .4 & 0 \\ .4 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Note that R is the same as the covariance matrix of the standardized variables

$tr(R) = 3$

$\lambda_1 = 1.4$	$\lambda_2 = 1$	$\lambda_3 = .6$
$\vec{e}_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{bmatrix}$	$\vec{e}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$	$\vec{e}_3 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \\ 0 \end{bmatrix}$

$$Y_1 = .71\, Z_1 + .71\, Z_2 = .71\left(\frac{X_1-\mu_1}{16}\right) + .71\left(\frac{X_2-\mu_2}{10}\right)$$

$$= .044(X_1-\mu_1) + .071(X_2-\mu_2)$$

$$Y_2 = Z_3 = \frac{X_3-\mu_3}{5}$$

etc.

Structure matrix:

	Z_1	Z_2	Z_3
Y_1	.84	.84	0
Y_2	0	0	1
Y_3	.55	-.55	0

③

If we consider discarding some of the components, how many should go?

① Set a threshold for the cumulative percent variance.

Ex: Say we use 90% $\quad$ 279, 77, 25 out of 351

Just Y_1: $\frac{279}{351} = .79$

Y_1, Y_2: $\frac{279+77}{351} = .93$ $\quad$ * Stop

Standardized: 1.4, 1, .6 out of 3

Y_1: $1.4/3 = .47$ $\qquad$ Y_1, Y_2, Y_3: $3/3 = 1$ *

Y_1, Y_2: $2.4/3 = .8$

④

② Set a minimum eigenvalue (only useful in the standardized version)

A commonly-used minimum is 1.

(called Kaiser's rule)

This discards any "below average" components.

In our example, Y_3 would be discarded.

③ If the original $X_1, \dots, X_p$ are predictors in a multiple regression, use the p principal components as predictors instead, and keep the set of significant predictors.

Factor Analysis

Spearman & Pearson (1904)

		C	F	E	M	D	Mu
Classics	C	1	.83	.78	.70	.66	.63
French	F	.83	1	.67	.67	.65	.59
English	E	.78	.67	1	.64	.54	.51
Math	M	.70	.67	.64	1	.45	.51
Discern. Pitch	D	.66	.65	.54	.45	1	.40
Music	Mu	.63	.59	.51	.51	.40	1

Look at the C & E columns +

notice: $\frac{.83}{.67} \approx \frac{.70}{.64} \approx \frac{.66}{.54} \approx \frac{.63}{.51}$

1.24   1.09   1.22   1.24

They searched for a model that would explain this behavior.

They proposed: $Z_i = a_i F + e_i$

0 expectation, & indep. of F and each other

Z_i: ith standardized variable

F: a single underlying factor, also standardized

How does their model cause the constant ratio pattern that they observed? (7)

$$E[Z_i] = 0, \quad V[Z_i] = 1$$

$$E[F] = 0, \quad E[F] = 1$$

$$\underbrace{V[Z_i]}_{=1} = a_i^2 \underbrace{V[F]}_{=1} + V[e_i]$$

$$1 - a_i^2 = V[e_i]$$

$$Corr[Z_i, Z_j] = \frac{Cov(Z_i, Z_j)}{\sqrt{V(Z_i)V(Z_j)}} = Cov(Z_i, Z_j)$$

(8)

$$= E[Z_i Z_j] - E[Z_i]E[Z_j]$$

$$= E[Z_i Z_j]$$

$$= E\left[(a_i F + e_i)(a_j F + e_j)\right]$$

$$= E\left[a_i a_j F^2 + a_i F e_j + a_j F e_i + e_i e_j\right]$$

$$= a_i a_j$$

$$\frac{Cor(Z_i, Z_j)}{Cor(Z_i, Z_k)} = \frac{a_i a_j}{a_i a_k} = \frac{a_j}{a_k}$$, which is the pattern they saw.

$$z_i = a_i F + e_i$$

$$\frac{X_i - \mu_i}{\sigma_i} = a_i F + e_i$$

$$X_i = \mu_i + \sigma_i a_i F + \sigma_i e_i$$

$$X_i = \mu_i + l_i F + \varepsilon_i$$

Next: allow multiple factors

HW #6 due Nov 26 8.1, 8.2,
8.10 a,b ← do in a stat package