EFA vs. PCA

Mengqian LU

Rule of Thumb

- It’s not the data’s own job to tell you what it measures, you should at least have a clue.
- It’s important to know the differences between methods, their assumptions, their limitations, their applications and advantages.
- Exploratory analysis results can sometimes be misleading – Check model/method’s assumptions; whether constraints are implausible; sample specific (still remember this?)

EFA vs. PCA

- Different Objectives:
  1. EFA is to determine both the nature and the number of latent variables (intelligence such as spatial reasoning and verbal intelligence), accounting for the variance-covariance of observed factor indicators (observable variables, 6 intelligence tests);
  2. PCA is to reduce high dimensional data space to fewer components, summarizing their variance

EFA vs. PCA (Cont'd)

- EFA and PCA are different, however, to a certain level, PCA is a special kind of EFA – especially in terms of "Extraction Method"

How to find unique solution?

- Options 1: Have $\mathbf{M} = \mathbf{\Lambda\Psi}^{-1}\mathbf{\Lambda}$ be diagonal, with diagonal elements in descending order of magnitude → ordered orthogonal factors with descending contributions (same logic as PCA)
- Potential problems:
  1. Variables may have substantial loadings on >1 factor
  2. From the 2nd factor, they often turn to be bipolar, i.e., a mixture of positive and negative loadings -> hard to interpret
- Possible solution: rotate the loadings instead, but there has been debate over this, in my opinion, there is nothing wrong to introduce domain knowledge

- Different assumptions on the sample covariance:

1. PCA: $$\boxed{\begin{aligned} Cov(x) &= \frac{1}{n}\sum_i x_i x_i^T = C_x \\ C_x &= U\Lambda U^T \end{aligned}}$$

PCA analyze ALL variance, C_x

2. EFA: $$\boxed{\mathrm{cov}(x) = \Sigma = \Lambda\Lambda^T + \Psi}$$

EFA analyze COMMON variance, $\Lambda\Lambda^T$

EFA vs. PCA (Cont'd)

- PCA extracts COMPONENTS
  - Keep all components → full component solution;
  - Keep fewer components (principal components) → truncated solution.
- PCA perfectly reproduces original correlation matrix
  - With unique mathematical solution
  - With uncorrelated, orthogonal components, ordered in descending variances
- PCA yields component loadings that explain relations between observation and extracted components

$$PC_1 = L_{11}O_1 + L_{12}O_2 + L_{13}O_3 + L_{14}O_4$$
$$PC_2 = L_{21}O_1 + L_{22}O_2 + L_{23}O_3 + L_{24}O_4$$
$$PC_3 = L_{31}O_1 + L_{32}O_2 + L_{33}O_3 + L_{34}O_4$$
$$PC_4 = L_{41}O_1 + L_{42}O_2 + L_{43}O_3 + L_{44}O_4$$

EFA vs. PCA Extraction Methods

- Given this correlation matrix:
- There are two structures in this 4 dimensional space: **O_1 & O_2**; **O_3 & O_4**
- Your PCs should be formed as $PC_1 = L_{11}O_1 + L_{12}O_2$ to $PC_2 = L_{23}O_3 + L_{24}O_4$ capture and separate the two groups, but

PCA doesn't group variables, it reproduces variations

	O_1	**O_2**	**O_3**	**O_4**
O_1	**1.0**			
O_2	**0.7**	**1.0**		
O_3	**0.3**	**0.3**	**1.0**	
O_4	**0. 3**	**0.3**	**0.5**	**1.0**

EFA vs. PCA Extraction Methods

- We can't ignore any cross correlation (those 0.3's)

	O_1	O_2	O_3	O_4
O_1	**1.0**			
O_2	**0.7**	**1.0**		
O_3	**0.3**	**0.3**	**1.0**	
O_4	**0. 3**	**0.3**	**0.5**	**1.0**

- In order to maximize variance reproduced by PC_1 and PC_2, We must have:

$$PC_1 = L_{11}O_1 + L_{12}O_2 + L_{13}O_3 + L_{14}O_4$$
$$PC_2 = L_{21}O_1 + L_{22}O_2 + L_{23}O_3 + L_{24}O_4$$

- ALL variables contribute to the extracted PC_1 and PC_2, but $|L_{11}|$ and $|L_{12}|$ will be larger for PC_1, $|L_{23}|$ and $|L_{24}|$ will be larger for PC_2.

$$PC_1 = 0.8O_1 + 0.8O_2 + 0.68O_3 + 0.68O_4$$
$$PC_2 = -0.46O_1 - 0.46O_2 + 0.54O_3 + 0.54O_4$$

PCA maximizes variance, it doesn't find groups of indicators that measure the same thing

EFA vs. PCA Component Matrix

$$PC_1 = 0.8O_1 + 0.8O_2 + 0.68O_3 + 0.68O_4$$
$$PC_2 = -0.46O_1 - 0.46O_2 + 0.54O_3 + 0.54O_4$$

	PC_1	PC_2	PC_3	PC_4
O_1	0.8	-0.46	*	*
O_2	0.8	-0.46	*	*
O_3	0.68	0.54	*	*
O_4	0.68	0.54	*	*

- Row = Observed Variables
- Column = Components
- ΣValue^2 by column = Eigenvalue for that component, eg Eigenvalue for PC_1 = $.8^2 + .8^2 + 0.68^2 + 0.68^2 = 2.2048$
- Eigenvalue/#Observed Values = Variance explained by that component, eg PC_1 →2.2048/4 = 55%
- Σ Value^2 by row = extracted communality for that variable, eg R^2 for O_1 = $.8^2+0.46^2$
- This only stands when the solution is **orthogonal**
- Same extraction applies to EFA, but it gives “Factor Matrix”

EFA vs. PCA Extraction Methods

$$\operatorname{cov}(x) = \Sigma = \Lambda\Lambda^T + \Psi$$

- Maximum Likelihood (MLE)
  - Assume normality – estimation, assessment of fit
  - Focuses on finding “best guesses” for loadings and error variances

EFA vs. PCA
Outcomes and what you are looking for

- PCA, outputs are linear combinations of the observed variables $\{O_1, \ldots O_4\}$ and $\{PC_1, \ldots PC_4\}$ are the outcomes.

$$PC_1 = L_{11}O_1 + L_{12}O_2 + L_{13}O_3 + L_{14}O_4$$
$$PC_2 = L_{21}O_1 + L_{22}O_2 + L_{23}O_3 + L_{24}O_4$$
$$PC_3 = L_{31}O_1 + L_{32}O_2 + L_{33}O_3 + L_{34}O_4$$
$$PC_4 = L_{41}O_1 + L_{42}O_2 + L_{43}O_3 + L_{44}O_4$$

The type of construct measured by a component is an 'emergent' construct, from the observed, formative variables.

- EFA, factors are considered as the cause of the observed variables, but factors $\{F_1, F_2\}$ are the predictors for $\{O_1, \ldots O_4\}$

$$O_1 = L_{11}F_1 + L_{12}F_2 + u_1$$
$$O_2 = L_{21}F_1 + L_{22}F_2 + u_2$$
$$O_3 = L_{31}F_1 + L_{32}F_2 + u_3$$
$$O_4 = L_{41}F_1 + L_{42}F_2 + u_4$$

The type of construct measured by a factor is a 'reflective' construct, the manifest variables are a reflection of the latent variables

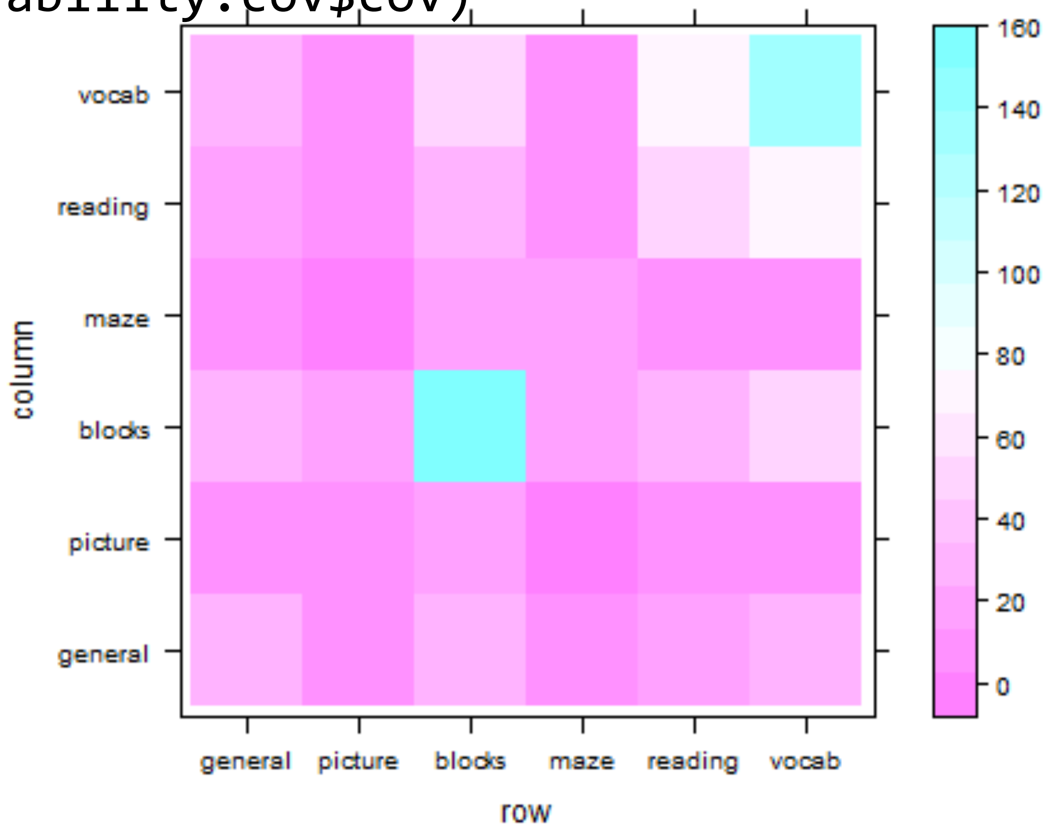
Exploratory Factor Analysis [Summary]

- “Exploratory” means trying the following alternatives
  - Number of factors, rotations, threshold for loadings, factor scores
- Although we said that we take advantage of the non-unique solutions of EFA, we still hope for the best scenario that: we get the same answer (not so different to have completely opposite interpretations) regardless of choices of the above alternatives; in reality, you need to pick one and defend it.
- PCA and EFA are both exploratory techniques for multivariates with different research questions and assumptions on variance-covariance matrix.

EFA VISUALIZATION

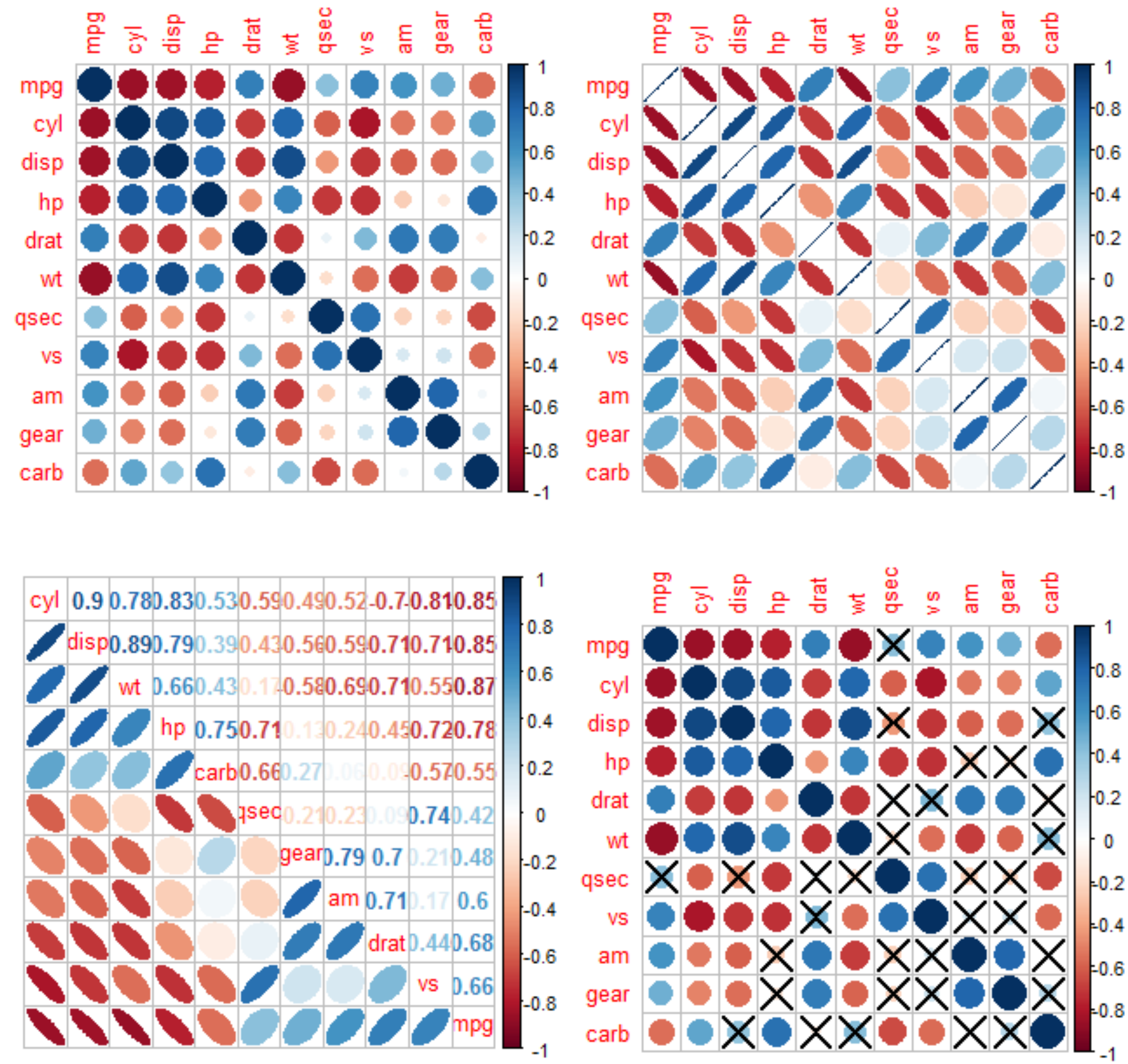
Many ways to visualize cov/cor max

```
library(lattice)
levelplot(ability.cov$cov)
```

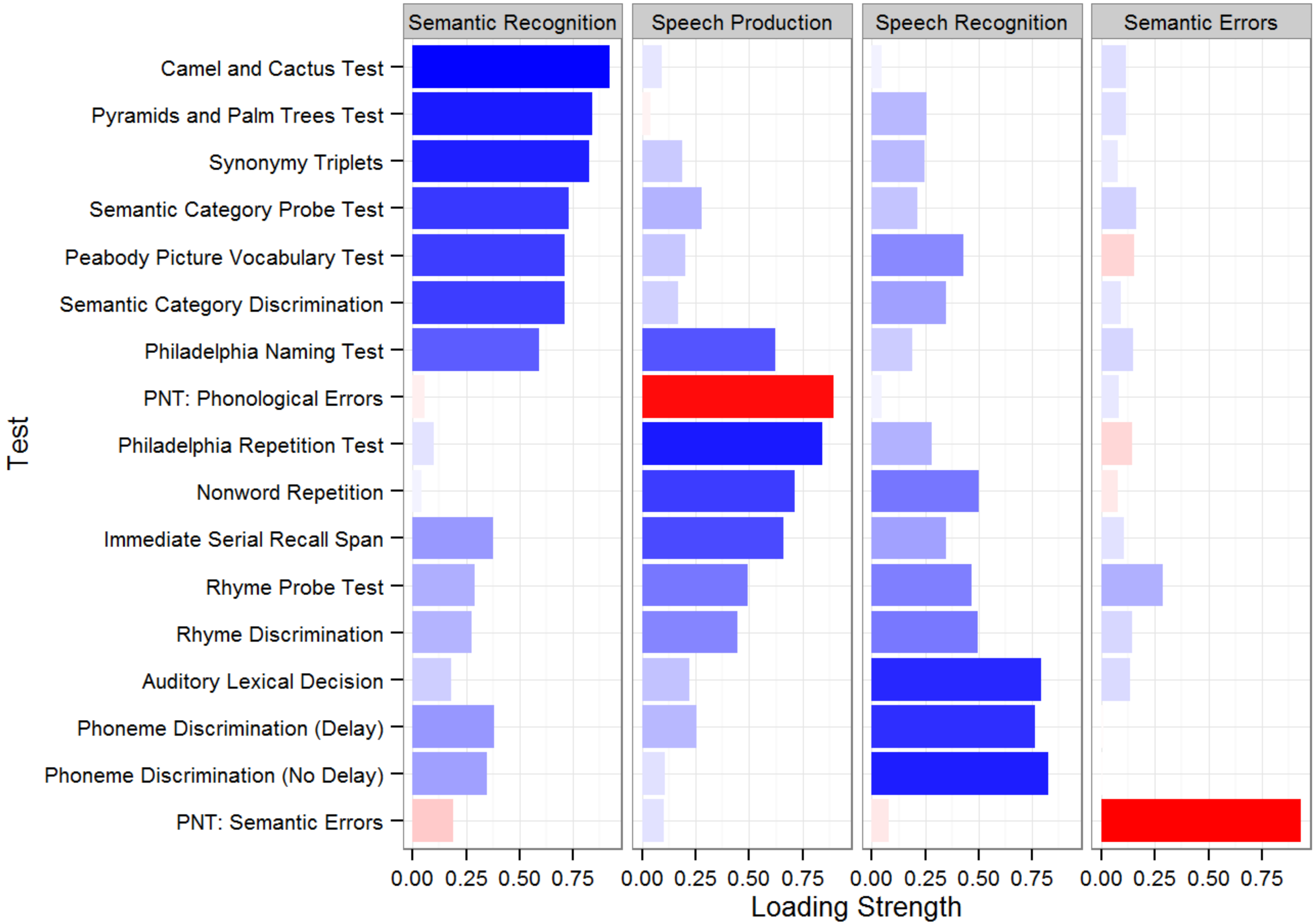


Many ways to visualize cov/cor max (cont’d)

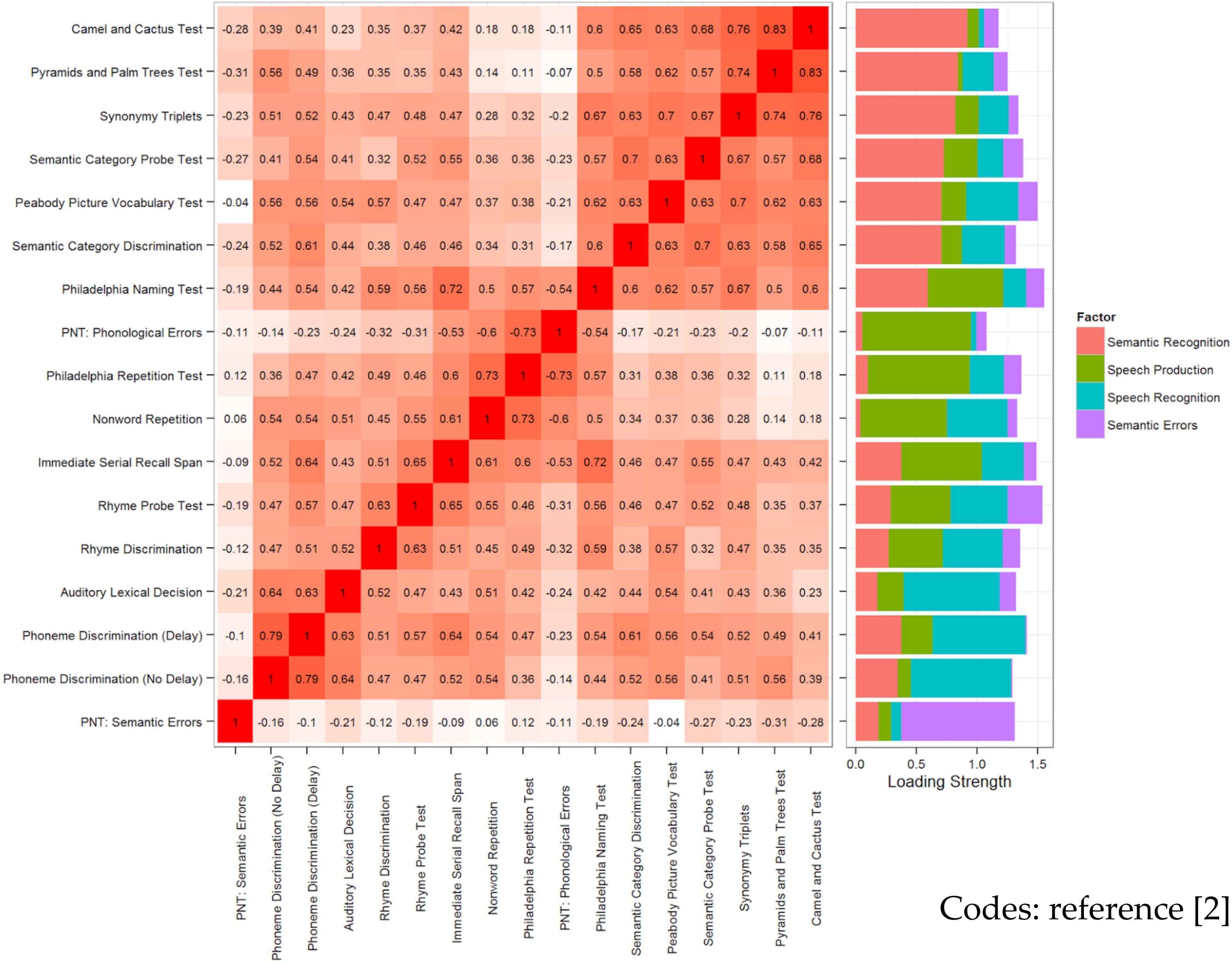
```
# Correlation matrix
library(corrplot)
M = cor(mtcars)
corrplot(M, method = "circle")
corrplot(M, method = "ellipse")
corrplot.mixed(M,lower="ellipse",upper='number',order="FPC")
# p.mat is the calculated p.value matrix
corrplot(M, p.mat = res1[[1]], sig.level = 0.01)
```



mpg
cyl
disp
hp
drat
wt
qsec
vs
am
gear
carb
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1



Semantic Recognition
Speech Production
Speech Recognition
Semantic Errors
Test
Camel and Cactus Test
Pyramids and Palm Trees Test
Synonymy Triplets
Semantic Category Probe Test
Peabody Picture Vocabulary Test
Semantic Category Discrimination
Philadelphia Naming Test
PNT: Phonological Errors
Philadelphia Repetition Test
Nonword Repetition
Immediate Serial Recall Span
Rhyme Probe Test
Rhyme Discrimination
Auditory Lexical Decision
Phoneme Discrimination (Delay)
Phoneme Discrimination (No Delay)
PNT: Semantic Errors
0.00 0.25 0.50 0.75
Loading Strength



Codes: reference [2]