Unsupervised Learning II

Roberta De Vito

BROWN
Public Health

Factor Analysis: the idea

The theoretical model

$$
\begin{aligned}
x_1 &= \mu_1 + \lambda_{11} f_1 + \cdots + \lambda_{1k} f_k + \epsilon_1 \\
&\vdots \\
x_p &= \mu_p + \lambda_{p1} f_1 + \cdots + \lambda_{pk} f_k + \epsilon_p
\end{aligned}
$$

Questions on prismia (Q1 and Q2)

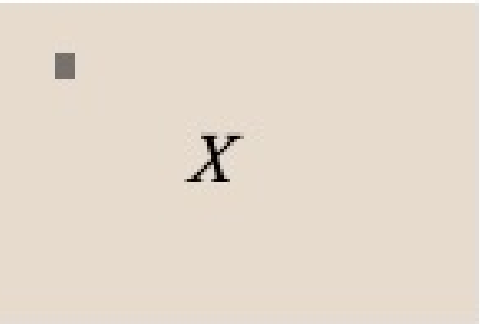
The matrix form

- the model can be written in matrix form

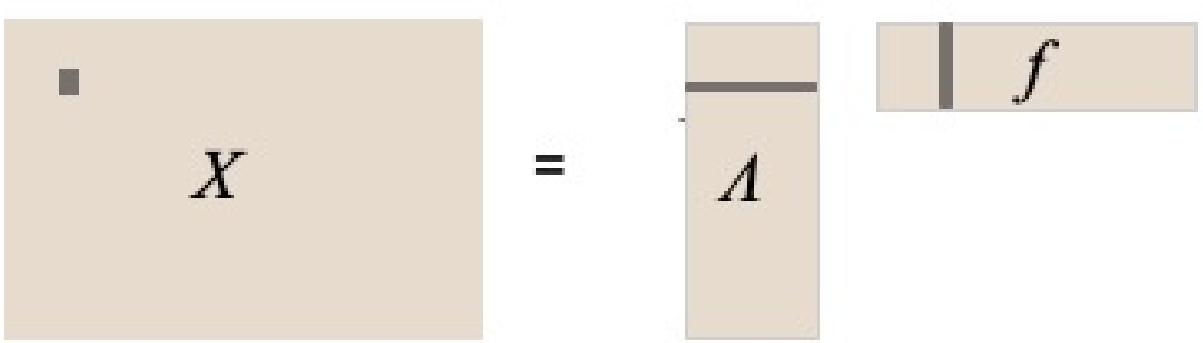
$$X = \mu + \Lambda f + \epsilon$$

- the error is represented by ϵ
- the factor loading matrix is Λ that is a $P \times K$ matrix
- we want to find K dimension suche that $K \ll P$

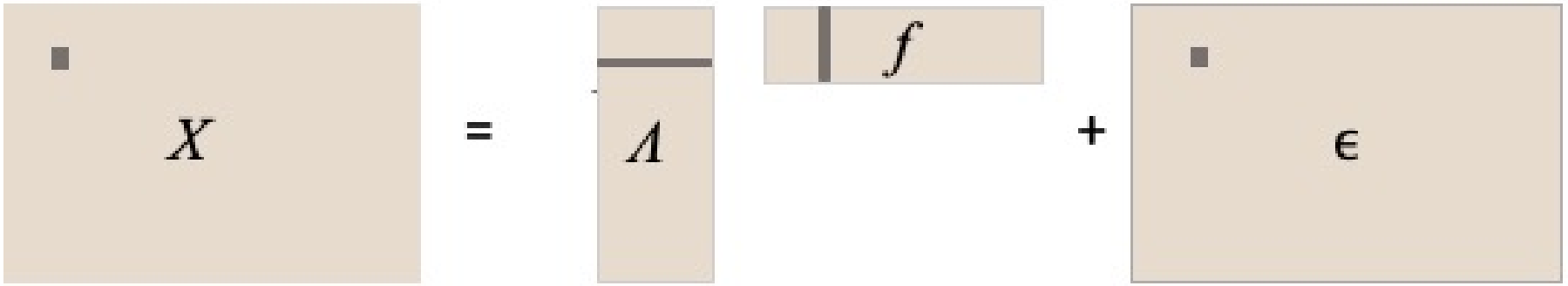
The matrix form: visualization



The matrix form: visualization



The matrix form: visualization



Model Assumptions

- $F \sim N(0, \mathcal{I}_k)$
- $\epsilon \sim N(0, \Psi)$ where $\Psi = diag(\psi_{11}, \ldots \psi_{pp})$ Q3 in prismia
- $Cov(F, \epsilon) = 0$
- $Var(X_j) = \lambda_{j1}^2 + \lambda_{jp}^2 + \psi_{jj} =$ communalities $+$ residual part
- In general: $\Sigma_X = Cov(X) = \Lambda\Lambda^\top + \Psi$

Model Properties

Invariance of scale

If we change the scale of X to $Y = CX$ with $C = diag(c_1, \ldots, c_p)$:

$$Var(Y) = C\Sigma C^\top = C\Lambda\Lambda^\top C^\top + C\Psi C^\top$$

Model Properties

Identifiability or Non-Uniqueness of Factor Loadings

If we take Q that is any orthogonal matrix ($QQ^\top = I$) then:

$$\Lambda^\star {\Lambda^\star}^\top = \Lambda Q Q^\top \Lambda^\top = \Lambda \Lambda^\top$$

$$\Sigma_X = Cov(X) = \Lambda Q Q^\top \Lambda^\top + \Psi = \Lambda \Lambda^\top + \Psi$$

Model Properties

This issue of Identifiability of the factor loadings has been solved by using specific rotations

This rotation (Varimax rotation) puts some loadings to zero and enhances other loadings (so close to 1)

Data set: ability.cov

general	general intelligence
picture	to complete a picture
blocks	drawing blocks
maze	orientation
reading	comprehension
vocab	dictionary

A first look to the data

```
> ability.cov$cov
        general picture  blocks   maze reading   vocab
general  24.641   5.991  33.520  6.023  20.755  29.701
picture   5.991   6.700  18.137  1.782   4.936   7.204
blocks   33.520  18.137 149.831 19.424  31.430  50.753
maze      6.023   1.782  19.424 12.711   4.757   9.075
reading  20.755   4.936  31.430  4.757  52.604  66.762
vocab    29.701   7.204  50.753  9.075  66.762 135.292
```

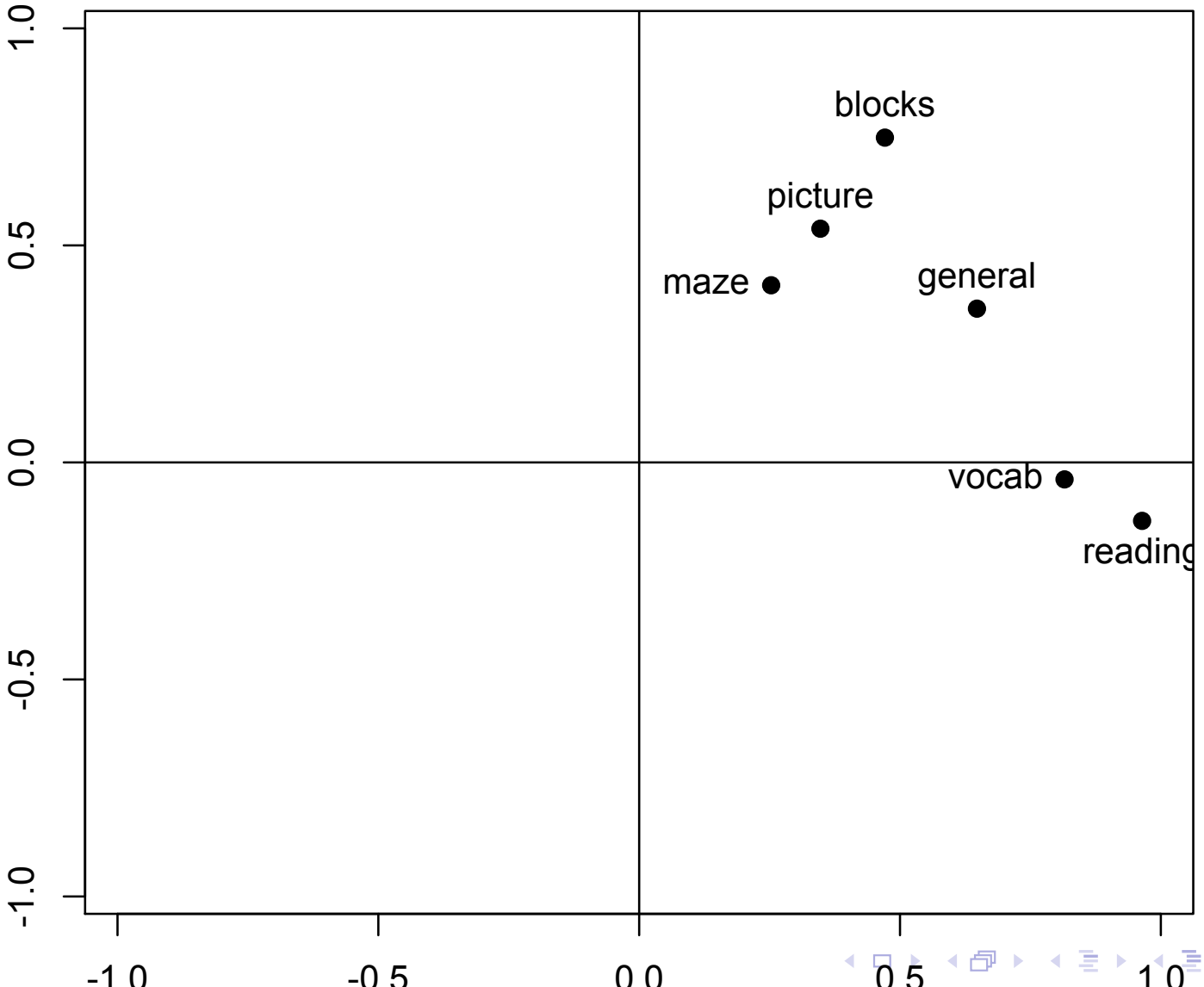
Model with one factor

```
Call:
factanal(factors = 1, covmat = ability.cov, rotation = "none")

Uniquenesses:
general picture  blocks    maze reading   vocab
  0.535   0.853   0.748   0.910   0.232   0.280

Loadings:
        Factor1
general 0.682
picture 0.384
blocks  0.502
maze    0.300
reading 0.877
vocab   0.849

               Factor1
SS loadings      2.443
Proportion Var   0.407

Test of the hypothesis that 1 factor is sufficient.
The chi square statistic is 75.18 on 9 degrees of freedom.
The p-value is 1.46e-12
```

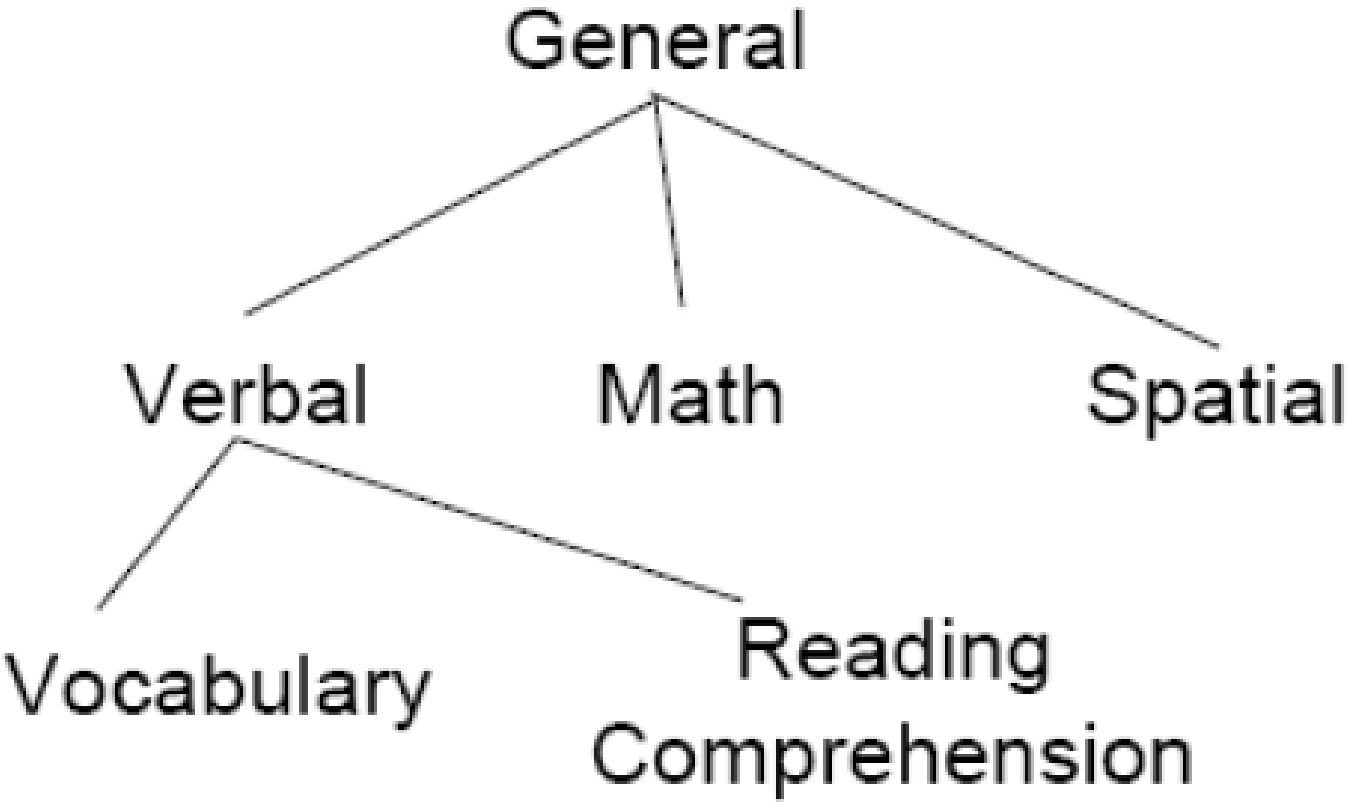
Model with one factor

```
Call:
factanal(factors = 2, covmat = ability.cov, rotation = "none")

Uniquenesses:
general picture  blocks    maze reading   vocab
  0.455   0.589   0.218   0.769   0.052   0.334

Loadings:
        Factor1 Factor2
general  0.648   0.354
picture  0.347   0.538
blocks   0.471   0.748
maze     0.253   0.408
reading  0.964  -0.135
vocab    0.815

               Factor1 Factor2
SS loadings      2.420   1.162
Proportion Var   0.403   0.194
Cumulative Var   0.403   0.597

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 6.11 on 4 degrees of freedom.
The p-value is 0.191
```

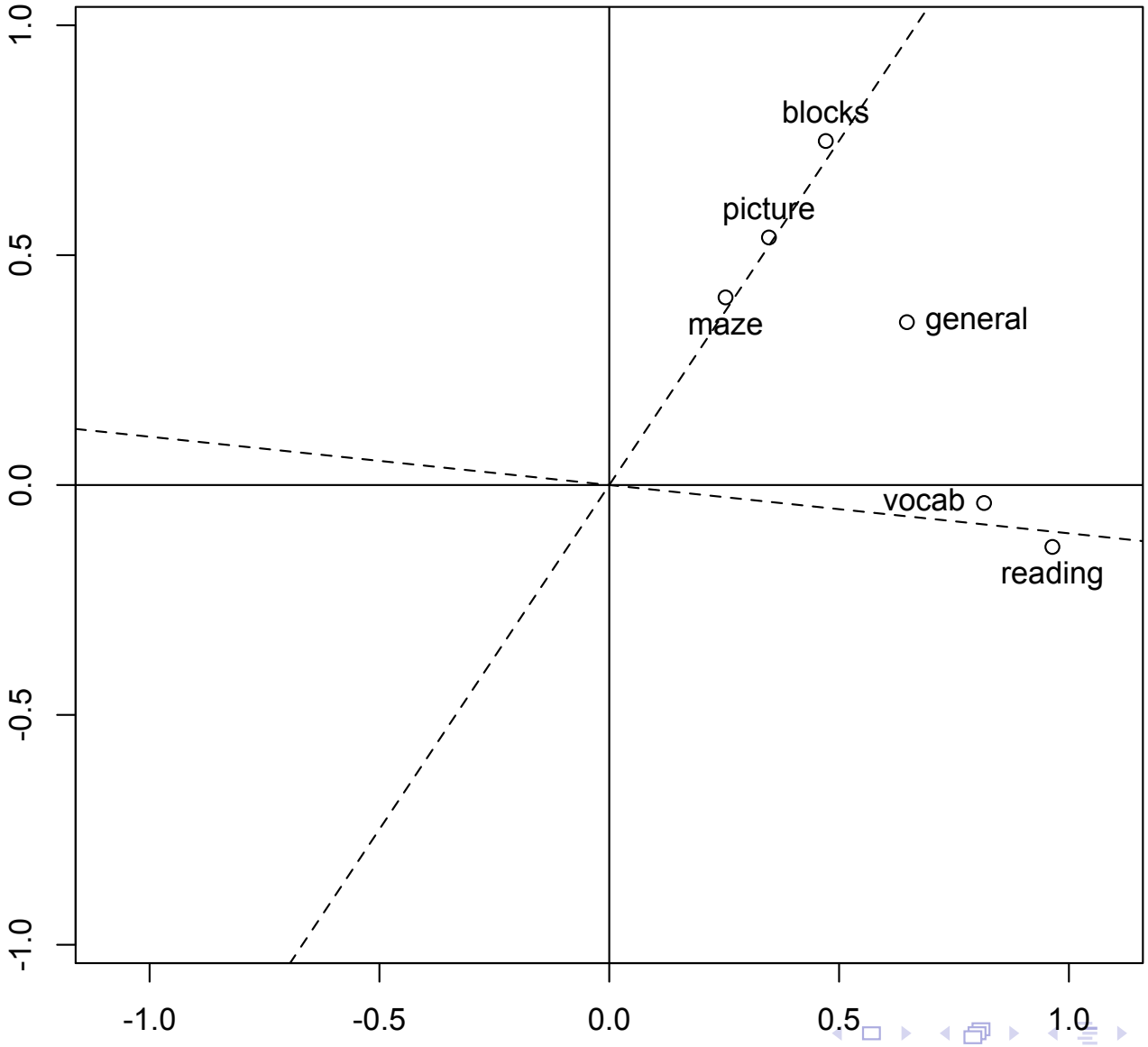
Plotting the factor model



Plotting the factor model



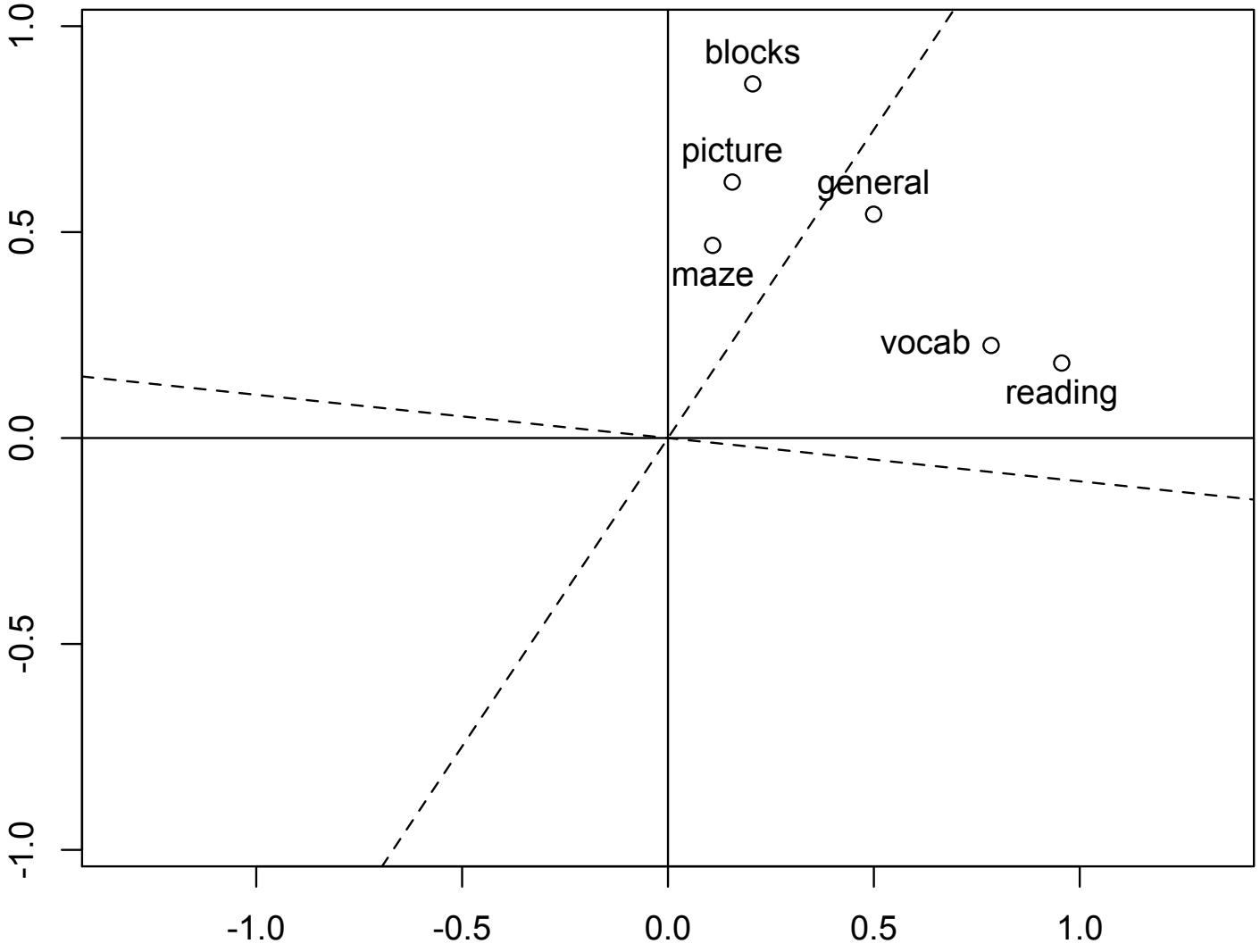
Plotting the factor model



Factor Model with Rotation

```
Call:
factanal(factors = 2, covmat = ability.cov, rotation = "varimax")

Uniquenesses:
general picture  blocks    maze reading   vocab
  0.455   0.589   0.218   0.769   0.052   0.334

Loadings:
        Factor1 Factor2
general 0.499   0.543
picture 0.156   0.622
blocks  0.206   0.860
maze    0.109   0.468
reading 0.956   0.182
vocab   0.785   0.225

               Factor1 Factor2
SS loadings      1.858   1.724
Proportion Var   0.310   0.287
Cumulative Var   0.310   0.597

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 6.11 on 4 degrees of freedom.
The p-value is 0.191
```

Plotting the factor model with rotation



The Boston housing Example

X_1: per capita crime rate,
X_2: proportion of residential land zoned for large lots,
X_3: proportion of nonretail business acres,
X_4: Charles River (1 if tract bounds river, 0 otherwise),
X_5: nitric oxides concentration,
X_6: average number of rooms per dwelling,
X_7: proportion of owner-occupied units built prior to 1940,
X_8: weighted distances to five Boston employment centers,
X_9: index of accessibility to radial highways,
X_{10}: full-value property tax rate per \$10,000,
X_{11}: pupil/teacher ratio ,
X_{12}: $1000(B-0.63)^2\,\boldsymbol{I}(B<0.63)$ where B is the proportion of African American,
X_{13}: % lower status of the population,
X_{14}: median value of owner-occupied homes in \$1000.

The Boston housing: 2 Factors (Q4)

```
Loadings:
               Factor1 Factor2
crime           0.233   0.569
large-lots     -0.699  -0.114
nonretail       0.674   0.532
nitric-oxides   0.744   0.466
room           -0.334  -0.205
prior1940       0.791   0.278
dist-Boston    -0.815  -0.299
highway         0.253   0.894
tax-rate        0.315   0.928
pupil/teacher   0.184   0.440
af-american    -0.222  -0.413
lower-status    0.571   0.403
owner          -0.376  -0.382

               Factor1 Factor2
SS loadings      3.666   3.379
Proportion Var   0.282   0.260
Cumulative Var   0.282   0.542
```

The Boston housing Factor Model Plot