See Härdle and Simar, chapter 12.

6 FACTOR ANALYSIS

6.1 ORTHOGONAL FACTOR MODEL

- Let $X = (X_1, \ldots, X_p)^T \sim (\mu, \Sigma)$ be a random vector in $\mathbb{R}^p$.
- In PCA, we've seen that if Σ has only $q < p$ non zero eigenvalues,

$$X - \mu = \Gamma_{(1)} Y_{(1)} \tag{1}$$

where

$$\Gamma_{(1)} = (\gamma_1 | \ldots | \gamma_q),$$

and

$$Y_{(1)} = (Y_1, \ldots, Y_q)^T,$$

the latter being the first q PCs in $Y = \Gamma^T (X - \mu)$.

☛ Recall that $Y_{(1)} \sim (0, \Lambda_1)$, where $\Lambda_1 = diag(\lambda_1, \dots \lambda_q)$.

☛ Letting $Q = \underbrace{\Gamma_{(1)}\Lambda_{(1)}^{1/2}}_{p\times q}$ and $F = \underbrace{\Lambda_{(1)}^{-1/2}Y_{(1)}}_{q\times 1}$, rewrite (1) as

$$X = \mu + QF.$$

☛ Now we have

$$\begin{aligned}
&E(F) = 0\\
&var(F) = \Lambda_{(1)}^{-1/2} var(Y_{(1)}) \Lambda_{(1)}^{-1/2} = I_q\\
&\Sigma = var(X) = Q\ var(F)\ Q^T = QQ^T = \sum_{j=1}^{q} \lambda_j \gamma_j \gamma_j^T.
\end{aligned}$$

☛ X is a weighted sum of the uncorrelated factors in $F = (F_1, \dots, F_q)^T$, i.e.

$$X = \mu + F_1 Q_1 + \cdots + F_q Q_q,$$

where $Q_1, \dots, Q_q$ are the columns of

$$Q = (Q_1 | \dots | Q_q).$$

- Σ, originally with $p + \binom{p}{2} = \frac{p(p+1)}{2}$ free parameters (the diagonal and the unique off-diagonal entries), is completely explained by the orthogonal factors via the loading matrix Q, which has qp entries.

- When $q << p$,

  $\Rightarrow \frac{p(p+1)}{2}$ is roughly of order $O(p^2)$.

  $\Rightarrow qp$ is roughly of order $O(p)$.

  $\Rightarrow$ Achieved **dimension reduction**! (b/c # parameters reduced)

- Actually, the # of free parameters in Q is a subtle issue, <u>not just qp</u>....

- More on that later.

☛ Orthogonal factor model: A similar but more nuanced model for X.

☛ For a $p \times q$ non-random matrix

$$Q = (q_{j\ell})_{\substack{1 \le j \le p \\ 1 \le \ell \le q}} = (Q_1 | \dots | Q_q),$$

there is a random vector

$$F = (F_1, \dots, F_q)^T \text{ of } q \text{ common factors}$$

and a random vector

$$U = (U_1, \dots, U_p)^T \text{ of } p \text{ specific factors},$$

such that

$$\begin{aligned} X &= \mu + QF + U \\ &= \mu + F_1 Q_1 + \cdots + F_q Q_q + U. \end{aligned}$$

☛ F and U: assumed to be **latent** (i.e. unobserved).

☛ The components $q_{j\ell}$ of Q are called **loadings**.

☛ It is assumed:

- $E(F) = E(U) = 0$
- $var(F) = I_q$
- $var(U) \equiv \Psi \equiv diag(\psi_1, \ldots, \psi_p)$ (i.e., a diagonal matrix)
- $cov(U_i, U_j) = 0$ if $i \neq j$
- $cov(F, U) = 0$.

☛ Hence,

$$\begin{aligned}\Sigma \equiv var(X) &= var(QF + U)\\ &= var(QF) + var(U)\\ &= QQ^T + \Psi.\end{aligned}$$

☛ Think of it as a covariance matrix model of the form

$$QQ^T + \Psi.$$

☛ Inferential goal: estimate Q and Ψ.

☛ Dimensionality reduction? Yes, only $qp + p$ many entries in (Q, Ψ), much less than $p(p+1)/2$ when $q << p$.

☛ Correlations among $X_1, \ldots, X_p$: entirely explained by the common factors F. (via the loadings Q)

☛ U_j adds extra noise specific to the component X_j, $j = 1, \ldots, p$.

☛ QF and U: the unobserved systematic and error parts of X.

☛ Well-known *psychology* application by Charles Spearman: When $q = 1$, F is a latent **general intelligence factor**, where X_j's are scores for different cognitive tasks.

☛ Component by component, for each $j = 1, \ldots, p$,

$$X_j = \mu_j + \sum_{\ell=1}^{q} q_{j\ell} F_\ell + U_j,$$

where $q_{j\ell}$ is the (j, ℓ)-th element of Q.

☛ Since $cov(U, F) = 0$, $var(F) = I_q$ and $var(U) = diag(\psi_1, \ldots, \psi_p)$,

$$var(X_j) = \sum_{\ell=1}^{q} q_{j\ell}^2 + \psi_j,$$

where

- $\sum_{\ell=1}^{q} q_{j\ell}^2$ is called the communality
- ψ_j is called the specific variance (or *uniqueness* in some text) .

☛ The proportion of $var(X_j)$ explained by the q common factors is

$$\frac{\sum_{\ell=1}^{q} q_{j\ell}^2}{var(X_j)}. \tag{2}$$

(The closer to 1, the better $var(X_j)$ is explained by Q.)

☛ Correlation between X and F: Since $X = QF + U + \mu$,

$$cov(X, F) = cov(QF + U, F) = cov(QF, F) = Q,$$

and since $var(X_j) = \sigma_{jj}$ and $var(F_j) = 1$, we deduce that

$$corr(X, F) = D^{-1/2}Q$$

where $D = diag(\sigma_{11} \ldots, \sigma_{pp})$.

☛ Like in PCA, by analysing these correlations, we can see which X_j's are strongly correlated with each factor, and interpret the factors.

☛ What if we change the scale of the X_j's? Suppose we use $Y = CX$ instead of X, where $C = diag(c_1, \ldots, c_p)$. Recalling that

$$X = \mu + QF + U,$$

we deduce that

$$Y = \mu_Y + Q_Y F + U_Y, \tag{3}$$

where we've defined

$$Q_Y = CQ, \quad U_Y = CU, \quad \mu_Y = C\mu.$$

☛ F hasn't changed, the new model (3) still orthogonally factorial:

- $E(F) = 0$, $var(F) = I_q$
- $E(U_Y) = 0$, $cov(U_{Y,i}, U_{Y,j}) = 0$ if $i \neq j$
- $var(U_Y) \equiv \Psi_Y = C\Psi C^T$ (still diagonal as C is diagonal)
- $cov(F, U_Y) = 0$.

☛ In many applications, the search of the loadings will be done through the scaled and centered data

$$Y = D^{-1/2}(X - \mu).$$

(Recall $D = diag(\sigma_{11} \ldots, \sigma_{pp})$)

☛ That is, we aim to estimate Q_Y and Ψ_Y in the model

$$Y = Q_Y F + U_Y.$$

under the same assumptions as before.

☛ Let $q_{Y,j\ell}$ denote the (j, ℓ)-th element of Q_Y. Then for $j = 1, \ldots, p$

$$\sum_{\ell=1}^{q} q_{Y,j\ell}^2 + \psi_{Y,j} = var(Y_j) = 1,$$

☛ If the communality $\sum_{\ell=1}^{q} q_{Y,j\ell}^2$ is close to 1, then the first q factors explain well the jth variable X_j.

☛ Recall in the non-scaled case, we found in (2) that

$$\sum_{\ell=1}^{q} q_{j\ell}^2/\sigma_{jj}$$

is the proportion of variance of X_j explained by the q factors.

☛ Since $q_{Y,j\ell} = q_{j\ell}/\sqrt{\sigma_{jj}}$,

$$\sum_{\ell=1}^{q} q_{Y,j\ell}^2 = \sum_{\ell=1}^{q} q_{j\ell}^2/\sigma_{jj}$$

is already the proportion of variance of X_j explained by the q factors.

☛ To interpret the factors, we can also compute the correlation matrix

$$corr(Y, F) = cov(Y, F) = cov(Q_Y F + U_Y, F) = Q_Y.$$

☛ The is the same as the correlation between F and the *original* X:

$$corr(X, F) = D^{-1/2}Q := Q_Y = corr(Y, F).$$

6.4 NON UNIQUENESS OF THE MATRIX Q

☛ Is the factor loading matrix Q unique? **No**.

☛ If $X = \mu + QF + U$, since $OO^T = I$ for any orthogonal $q \times q$ matrix O, we also have

$$X = \mu + Q_O F_O + U$$

holds with $Q_O = QO$ and $F_O = O^T F$.

☛ Still a valid factor model:

- $E(F_O) = 0$
- $var(F_O) = O^T O = I$
- $E(U) = 0$
- $cov(U_i, U_j) = 0$ if $i \neq j$
- $cov(F_O, U) = O^T cov(F, U) = 0$.

☛ Importantly, under the covariance model interpretation,

$$\Sigma = QQ^T + \Psi = Q_O Q_O^T + \Psi.$$

☛ From inference point of view: **Bad**, since one cannot uniquely recover the loading matrix (Q or Q_O?) from the covariance Σ.

☛ From a numerical point of view: **Bad too**, since it is difficult for an algorithm to find a solution.

☛ Solution: **Impose more restrictions on** Q.

☛ But how?

☛ How *unrestricted* are we in writing down a $q \times q$ orthogonal matrix? Must abide by these (algebraic) constraints:

All q columns must have unit lengths $\Rightarrow$ q constraints

Any *pair* of columns must give 0 dot-product. $\Rightarrow$ $\binom{q}{2}$ constraints

Hence, the degrees of freedom in writing a $q \times q$ orthogonal matrix (which has q^2 entries) is

$$q^2 - q - \binom{q}{2} = q(q-1)/2$$

☛ Without no restrictions on Q, the solution set for Q has degrees of freedom $q(q-1)/2$, because if Q is a solution to

$$\Sigma = QQ^T + \Psi$$

for a fixed Σ, then QO is too for any orthogonal matrix O.

☛ Hence, one generally needs to impose

$$q(q-1)/2 = \binom{q}{2} \text{ constraints}$$

to make Q identifiable.

☛ *Suggestion:* Lower triangular contraints on $Q = (q_{jl})_{\substack{1\le j\le p\\ 1\le l\le q}}$, that is, to

restrict $q_{jl} = 0$ for all $j < l$.

There are exactly $q(q-1)/2$ such zero constraints.

☛ It is motivated by the **Cholesky decomposition** theorem:

For any $p \times p$ *positive semidefinite* matrix A of rank $q \le p$, one can find a unique $p \times q$ matrix $L = (l_{jl})$ such that

$$A = LL^T,$$

and

L is lower triangular with positive diagonal entries,

i.e. $l_{jj} > 0$ for all $j = 1, \ldots q$, and $l_{jl} = 0$ for all $j < l$.

☛ In our context, QQ^T is the rank-q matrix A in question.

☛ Typically, one further impose diagonal positivity constraints, i.e.

$$\text{Restrict } q_{jj} > 0 \text{ for all } j = 1, \ldots, q.$$

☛ Without diagonal positivity, if $\Sigma = QQ^T + \Psi$ for a Q, then

$$\tilde{Q} = -Q$$

must also satisfie $\Sigma = \tilde{Q}\tilde{Q}^T + \Psi$, i.e. we can only identify Q up to sign.

☛ However, these diagonal +ve constraints **do not** further cut down the degrees of freedom.

☛Because the non-identifiability caused by sign change is of a different nature than that caused by multiplying an orthogonal matrix O.

☛(Intuitively, the whole real line and half the real line are both objects of geometric dimension 1.)

☛ The lower triangular contraints on Q are popular among some Bayesians.

☛ Alternatively, another common set of constraints found is to restrict the matrix

$$Q^T \Psi^{-1} Q$$

to be diagonal, with its diagonal elements distinct and arranged in decreasing order of magnitude. (suggested in Härdle and Simar)

☛ *Summary*: With a factor model for $X_1, \ldots, X_p$ with q factors, we generally need an additional set of non-redundant $q(q-1)/2$ constraints to make (Q, Ψ) identifiable from a given covariance matrix Σ.

☛ The <u>effective</u> parameter degrees of freedom of an **identifiable factor model** is

$$\underbrace{pq}_{\substack{\#\text{entries}\\ \text{in } Q}} + \underbrace{p}_{\substack{\#\text{diagonal}\\ \text{entries in } \Psi}} - \underbrace{q(q-1)/2}_{\substack{\#\text{ constraints}\\ \text{imposed on } Q\\ \text{for identifiability}}}$$

☛ The parameter degree of freedom of Σ under **no modelling** is simply

$$\underbrace{p(p+1)/2}_{\text{diagonal+the strict upper triangular part}}$$

☛ Factor modeling is for dimension reduction. To avoid over-parametrization, we require

$$p(p+1)/2 \geq pq + p - q(q-1)/2. \tag{4}$$

Otherwise, one may find ∞ many (Q, Ψ) that give the same Σ, *even after taking $q(q-1)/2$ identifiability constraints on Q into account.*

☛ **Implication on** (4): For p variables, there is a bound on the number of factors (q) to model the data, to render an identifiable model.

6.5 LIKELIHOOD METHODS UNDER NORMAL ASSUMPTION

☛ Estimating (Q, Ψ): **Maximum likelihood estimation** (MLE) method, assuming F and U are both normal (and hence so is X).

☛ Recall: the factor model models the covariance matrix as

$$\Sigma = \Sigma(Q, \Psi) = QQ^T + \Psi.$$

☛ For a dataset $\mathcal{X}$ with n samples, recall the normal log likelihood

$$l(\mathcal{X}; \mu, \Sigma) = -\frac{n}{2} \log |2\pi\Sigma| - \frac{1}{2} \sum_{i=1}^{n} (X_i - \mu)^T \Sigma^{-1} (X_i - \mu),$$

where $\bar{X}$ is the sample mean.

☛ The MLE for μ is always $\bar{X}$, so after simple algebra we have to maximize

$$
\begin{aligned}
l(\mathcal{X}; \bar{X}, \Sigma) &= -\frac{n}{2}\{\log|2\pi\Sigma| + tr(\Sigma^{-1}\hat{\Sigma})\} \\
&= -\frac{n}{2}\{\log|2\pi(QQ^T + \Psi)| + tr((QQ^T + \Psi)^{-1}\hat{\Sigma})\}
\end{aligned}
$$

with respect to (Q, Ψ), under the requisite constraints for identifiability of Q, where

$$
\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T.
$$

☛ Not an easy numerical problem; the `factanal` function in R implements it with the diagonal constraints on $Q^T\Psi^{-1}Q$, using a method described in Lawley & Maxwell (1971).

☛ We can also use likelihood ratio (LR) test to determine the number of factors q (in the allowable range that doesn't over-parametrize our data), by testing the null hypothesis

$$H_0 : q \text{ is the number of factors,}$$

for each allowed q.

☛ The LR statistic for a given q is then

$$-2\log\left(\frac{\text{maximized likelihood under } H_0}{\text{maximized likelihood under no modelling}}\right) = n\log\left(\frac{|\hat{Q}\hat{Q}^T + \hat{\Psi}|}{|\hat{\Sigma}|}\right), \quad (5)$$

where $\hat{Q}$ and $\hat{\Psi}$ are the MLEs, for this q.

☛ When H_0 is true, the LR statistic has an approximation

$$\chi^2_{\frac{1}{2}((p-q)^2-p-q)}$$

distribution, which can be used to calibrate the p-value for model fit.

☛ In practice, the n at the front in (5) is replaced by

$$n - 1 - (2p + 4q + 5)/6$$

to improve the χ^2 approximation. This is known as *Bartlett's correction*.

☛ Then go with the model with the largest p-value, if we must pick one.

6.6 Rotation

☛ One thing people often do is a varimax rotation of the factors, after they have already obtained an estimate $\hat{Q}$ for Q .

☛ Let $\hat{Q}^* = \hat{Q}G$ where G is an orthogonal matrix to be determined. They choose G that maximises the sum of the variances of the estimated squared loadings $(\hat{q}^*_{j\ell})^2$ within each column of $\hat{Q}^*$, i.e., they choose G to maximise:

$$\sum_{\ell=1}^{q} \sum_{j=1}^{p} \left[(\hat{q}^*_{j\ell})^2 - \frac{\sum_{j=1}^{p} (\hat{q}^*_{j\ell})^2}{p} \right]^2 .$$

☛ Often the resulting factors are easier to interpret: Groups of X_j's tend to be associated with less factors.

☛ The `factanal` function in R implements "`varimax`" by default.

Example (Kendall, 1975). In a job interview, 48 applicants were each judged on 15 variables. The variables were

(1) Form of letter of application
(2) Appearance
(3) Academic ability
(4) Likeability
(5) Self-confidence
(6) Lucidity
(7) Honesty
(8) Salesmanship
(9) Experience
(10) Drive
(11) Ambition
(12) Grasp
(13) Potential
(14) Keenness to join
(15) Suitability

Factor loadings before varimax rotation

Table 9.6.2 Maximum likelihood factor solution of applicant data with $k = 7$ factors, unrotated

Variable	Factor loadings						
	1	2	3	4	5	6	7
1	0.090	−0.134	−0.338	0.400	0.411	−0.001	0.277
2	−0.466	0.171	0.037	−0.002	0.517	−0.194	0.167
3	−0.131	0.466	0.153	0.143	−0.031	0.330	0.316
4	0.004	−0.023	−0.318	−0.362	0.657	0.070	0.307
5	−0.093	0.017	0.434	−0.092	0.784	0.019	−0.213
6	0.281	0.212	0.330	−0.037	0.875	0.001	0.000
7	−0.133	0.234	−0.181	−0.807	0.494	0.001	−0.000
8	−0.018	0.055	0.258	0.207	0.853	0.019	−0.180
9	−0.043	0.173	−0.345	0.522	0.296	0.085	0.185
10	−0.079	−0.012	0.058	0.241	0.817	0.417	−0.221
11	−0.265	−0.131	0.411	0.201	0.839	−0.000	−0.001
12	0.037	0.202	0.188	0.025	0.875	0.077	0.200
13	−0.112	0.188	0.109	0.061	0.844	0.324	0.277
14	0.098	−0.462	−0.336	−0.116	0.807	−0.001	0.000
15	−0.056	0.293	−0.441	0.577	0.619	0.001	−0.000

Factor loadings after varimax rotation

Table 9.6.3 Maximum likelihood factor solution of applicant data with $k = 7$ factors, varimax rotation (Kendall, 1975)

Variable	Factor loadings						
	1	2	3	4	5	6	7
1	0.129	0.074	0.665	−0.096	0.017	−0.042	0.267
2	0.329	0.242	0.182	0.095	0.611	−0.013	−0.006
3	0.048	−0.017	0.097	0.688	0.043	0.007	0.008
4	0.249	0.759	0.252	−0.058	0.090	−0.096	0.204
5	0.882	0.184	−0.082	−0.074	0.190	0.059	−0.045
6	0.907	0.266	0.136	0.046	−0.042	−0.290	−0.016
7	0.199	0.911	−0.224	−0.013	0.174	−0.094	−0.204
8	0.875	0.082	0.264	−0.076	0.140	0.043	−0.058
9	0.073	−0.027	0.718	0.158	0.069	0.036	0.009
10	0.780	0.197	0.386	0.026	−0.051	0.398	−0.023
11	0.874	0.036	0.157	−0.052	0.382	0.142	0.205
12	0.775	0.346	0.286	0.172	0.143	−0.159	0.111
13	0.703	0.409	0.354	0.329	0.140	0.070	0.193
14	0.432	0.540	0.381	−0.540	−0.013	0.099	0.275
15	0.313	0.079	0.909	0.049	0.142	0.027	−0.214

Interpretation (copied from that book):

- ☛ It is very difficult to interpret the unrotated loadings but easier to interpret the rotated loadings.

- ☛ The first factor is loaded heavily on variables 5, 6, 8, 10, 11, 12, and 13 and represents perhaps an outward and salesmanlike personality.

- ☛ Factor 2, weighting variables 4 and 7, represents likeability.

- ☛ Factor 3, weighting variables 1, 9, and 15 represents experience.

- ☛ Factors 4 and 5 each represent one variable, academic ability (3), and appearance (2), respectively.

- ☛ The last two factors have little importance and variable 14 (keenness) seemed to be associated with several of the factors.

6.7 Factor analysis versus PCA

☛ In PCA, our goal was to explicitly find linear combinations of the components of X. This is how we constructed the PCs $Y_1, \ldots, Y_p$ and this doesn't depend on any model .

☛ In factor analysis, the factors are not directly computable, they are latent. They appear after we model a structure on the covariance matrix. The whole factor analysis depends on the factor model we assumed. If the model is wrong, then the analysis will be spurious.

☛ In factor analysis, the factors are not linear combinations of the original variables, they are factors on their own which often represent characteristics that groups of variables may represent.

☛ In fact, in factor analysis, instead of taking the factors F to be functions of X, we express X as a function of F.

☛ In PCA the first few PCs explain the largest variability of the data (this is how we construct them).

☛ In factor analysis, the first factors are often those that are the most interpretable (after rotation).

☛ Sometimes, PCA and factor analysis give similar results, and we can understand why: as seen earlier, if only $q < p$ eigenvalues of Σ are nonzero, then we can almost write a factor model using the first q PCs, except without the specific factor U.