Multivariate Analysis (Slides 13)

- The final topic we consider is Factor Analysis.
- A Factor Analysis is a mathematical approach for attempting to explain the correlation between a large set of variables in terms of a small number of underlying *factors*.
- The main assumption of factor analysis is that it is not possible to observe the underlying factors directly, *i.e.*, they are 'latent'.

Sound familiar?

- Factor analysis is a dimension reduction technique and shares the same objective as principal components analysis.
- Both attempt to describe the relationships in a large set of variables through a smaller number of dimensions.
- However, the factor analysis model is much more elaborate.

Example

- The performance of a group children in Irish (x_1), English (x_2), and Maths (x_3) was recorded. The correlation matrix for their scores was:

$$\begin{pmatrix} 1 & 0.83 & 0.78 \\ & 1 & 0.67 \\ & & 1 \end{pmatrix}$$

- The dimensionality of this matrix can be reduced from $m = 3$ to $m = 1$ by expressing the three variables as:

$$\begin{aligned} x_1 &= \lambda_1 f + \epsilon_1 \\ x_2 &= \lambda_2 f + \epsilon_2 \\ x_3 &= \lambda_3 f + \epsilon_3 \end{aligned}$$

Example cont'd

- The f in these equations is an underlying *common factor*, the λ_i's are known as *factor loadings*, whilst the ϵ_i's are known as *errors* or *specific factors*.
- The common factor can often be given an interpretation, *e.g.*, 'general ability'.
- The specific factors ϵ_i will have small variance if x_i is closely related to general ability.

The factor model

- The observable random vector $\mathbf{X} = (X_1, X_2, \ldots, X_m)$ has mean μ and covariance matrix Σ.
- The factor model states that $\mathbf{X}$ is linearly dependent upon a few unobservable random variables $f_1, f_2, \ldots, f_p$ called **common factors** and m additional sources of variation $\epsilon_1, \epsilon_2, \ldots, \epsilon_m$ called **specific factors**.
- Hence:

$$\begin{aligned} X_1 - \mu_1 &= \lambda_{11} f_1 + \lambda_{12} f_2 + \cdots + \lambda_{1p} f_p + \epsilon_1 \\ &\vdots \qquad\quad \vdots \\ X_m - \mu_m &= \lambda_{m1} f_1 + \lambda_{m2} f_2 + \cdots + \lambda_{mp} f_p + \epsilon_m \end{aligned}$$

$$\Rightarrow \mathbf{X} - \mu = \Lambda \mathbf{f} + \epsilon.$$

The factor model

- The λ_{ij} value is called the **factor loading** of the i-th variable on the j-th factor.
- Λ is the matrix of factor loadings.
- Note that the i-th specific factor ϵ_i is associated only with the response X_i.
- Note that $f_1, f_2, \ldots, f_p, \epsilon_1, \epsilon_2, \ldots, \epsilon_m$ are all *unobservable random variables.*

Assumptions

Under the factor model it is assumed that:

1. $\mathbb{E}[\mathbf{f}] = \mathbf{0}$ $\qquad \mathbf{Cov}[\mathbf{f}] = \mathbf{I} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$

2. $\mathbb{E}[\epsilon] = \mathbf{0}$ $\qquad \mathbf{Cov}[\epsilon] = \mathbf{\Psi} = \begin{pmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{pmatrix}$

3. $\mathbf{f}$ and ϵ are independent and so $\mathbf{Cov}[\mathbf{f}, \epsilon] = \mathbf{0}$.

Orthogonal Factor Model

- Note the above is referred to as the 'orthogonal factor model'. However, if it is assumed that $\mathbf{Cov}[\mathbf{f}]$ is not diagonal, then the model is referred to as the 'oblique factor model'.
- The orthogonal factor model implies a specific covariance structure for $\mathbf{X}$:

$$\begin{aligned}
\mathbf{\Sigma} &= \mathbf{Cov}[\mathbf{X}] \\
&= \mathbb{E}[(\mathbf{X}-\mu)(\mathbf{X}-\mu)^T] \\
&= \mathbb{E}[(\Lambda\mathbf{f}+\epsilon)(\Lambda\mathbf{f}+\epsilon)^T] \\
&= \mathbb{E}[(\Lambda\mathbf{f})(\Lambda\mathbf{f})^T + \epsilon(\Lambda\mathbf{f})^T + (\Lambda\mathbf{f})\epsilon^T + \epsilon\epsilon^T] \\
&= \Lambda\mathbb{E}[\mathbf{f}\mathbf{f}^T]\Lambda^T + \mathbb{E}[\epsilon\mathbf{f}^T]\Lambda^T + \Lambda\mathbb{E}[\mathbf{f}\epsilon^T] + \mathbb{E}[\epsilon\epsilon^T] \\
&= \Lambda\Lambda^T + \mathbf{\Psi}
\end{aligned}$$

- It can also be shown that $\mathbf{Cov}[\mathbf{X},\mathbf{f}] = \mathbb{E}[(\mathbf{X}-\mu)(\mathbf{f}-\mathbf{0})] = \Lambda$, hence the covariance of the observed variable X_i and the unobserved factor f_j is the factor loading λ_{ij}.

Variance

- The variance of $\mathbf{X}$ can be split into two parts.
- The first portion of the variance for the i-th component arises from the m common factors, and is referred to as the i-th **communality**.
- The remainder of the variance for the i-th component is due to the specific factor, and is referred to as the **uniqueness**.
- Denoting the i-th communality by h_i^2, then:

$$\begin{aligned} \sigma_i^2 &= \lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{ip}^2 + \psi_i \\ &= h_i^2 + \psi_i \\ \mathbf{Var}[X_i] &= \text{communality} + \text{uniqueness} \end{aligned}$$

- The i-th communality is the sum of squares of the loadings of the i-th variable on the p common factors.

Variance cont'd

- The factor model assumes that the $m + m(m-1)/2$ variances and covariances of $\mathbf{X}$ can be reproduced from the mp factor loadings λ_{ij} and the m specific variances ψ_i.
- In the case where $p << m$ the factor model provides a simplified version of the covariation in $\mathbf{X}$ with fewer parameters than the $m(m+1)/2$ parameters in Σ.
- For example, if $\mathbf{X} = (X_1, X_2, \ldots, X_{12})$ and a factor model with $p = 2$ is appropriate, then the $12 \times 13/2 = 78$ elements of Σ are described in terms of the $pm + m = 2 \times 12 + 12 = 36$ parameters λ_{ij} and ψ_i of the factor model.
- Unfortunately, and as we will see, most covariance matrices cannot be uniquely factored as $\Lambda\Lambda' + \Psi$ where $p << m$.

Scale invariance

- Rescaling the variables of $\mathbf{X}$ is equivalent to letting $\mathbf{Y} = \mathbf{C}\mathbf{X}$, where $\mathbf{C} = \text{diag}(c_i)$.
- If the factor model holds with $\Lambda = \Lambda_x$ and $\Psi = \Psi_x$, then

$$\mathbf{Y} - \mathbf{C}\mu = \mathbf{C}\Lambda_x\mathbf{f} + \mathbf{C}\epsilon$$

  So,

$$\mathbf{Var}[\mathbf{Y}] = \mathbf{C}\mathbf{\Sigma}\mathbf{C} = \mathbf{C}\Lambda_x\Lambda_x^T\mathbf{C} + \mathbf{C}\Psi_x\mathbf{C}$$

- Hence the factor model holds for $\mathbf{Y}$ with factor loading matrix $\Lambda_y = \mathbf{C}\Lambda_x$ and specific variances $\Psi_y = \mathbf{C}\Psi_x\mathbf{C} = \text{diag}(c_i^2\psi_{ii})$.
- In other words, factor analysis (unlike PCA and every other technique covered) is not affected by a re-scaling of the variables.

Non-uniqueness: Proof

- To demonstrate that most covariance matrices Σ cannot be uniquely factored as $\Lambda\Lambda' + \Psi$, where $p << m$, let $\mathbf{G}$ be any $m \times m$ orthogonal matrix ($\mathbf{G}\mathbf{G}^T = \mathbf{I}$). Then,

$$\begin{aligned} \mathbf{X} - \mu &= \Lambda\mathbf{f} + \epsilon \\ &= \Lambda\mathbf{G}\mathbf{G}^T\mathbf{f} + \epsilon \\ &= \Lambda^*\mathbf{f}^* + \epsilon \end{aligned}$$

  Here $\Lambda^* = \Lambda\mathbf{G}$ and $\mathbf{f}^* = \mathbf{G}^T\mathbf{f}$.

- It follows that $\mathbb{E}[\mathbf{f}^*] = \mathbf{0}$ and $\mathbf{Cov}[\mathbf{f}^*] = \mathbf{I}$.

- Thus it is impossible, given the data $\mathbf{X}$, to distinguish between Λ and Λ^*, since they both generate the same covariance matrix Σ, *i.e.*,

$$\Sigma = \Lambda\Lambda^T + \Psi = \Lambda\mathbf{G}\mathbf{G}^T\Lambda' + \Psi = \Lambda^*\Lambda^{*T} + \Psi$$

Continued

- This issue leads to the idea of *factor rotations*, since orthogonal matrices correspond to rotations of the coordinate system of $\mathbf{X}$ (we'll return to this idea later).
- Usually we estimate one possibility for Λ and Ψ and rotate the resulting loading matrix (multiply by an orthogonal matrix $\mathbf{G}$) so as to ease interpretation.
- Once the loadings and specific variances are estimated, estimated values for the factors themselves (called **factor scores**) are constructed.

Maximum likelihood factor analysis

- When the data $\mathbf{x}$ are assumed to be normally distributed, then estimates for Λ and Ψ can be obtained by maximizing the likelihood.
- The normal location parameter is replaced by its MLE $\bar{x}$, whilst the log-likelihood of the data depends on Λ and Ψ through $\mathbf{\Sigma}$.
- To ensure Λ is well defined (invariance is caused by orthogonal transformations), the computationally convenient **uniqueness condition** is enforced:

$$\Lambda^T \Psi^{-1} \Lambda = \Delta$$

  Here Δ is a diagonal matrix.
- Numerical optimization of the log-likelihood can then be performed to obtain the MLEs $\hat{\Lambda}$ and $\hat{\Psi}$.

Maximum likelihood factor analysis

- However, problems may occur if $\psi_{ii} = 0$.
- This happens when the uniqueness is zero, *i.e.*, the variance in the variable X_i is completely accounted for by the factors f_i.
- Also, during estimation we may find that $\psi_{ii} < 0$. This solution is clearly improper and is known as a **Heywood case**.
- Such situations occur if there are too many common factors, too few common factors, not enough data, or the inappropriate application of the model for the dataset.
- In most computer programs this problem is resolved by re-setting ψ_{ii} to be a small positive number before progressing.

Factor rotation

- If the initial loadings are subject to an orthogonal transformation (*i.e.*, multiplied by an orthogonal matrix $\mathbf{G}$), the covariance matrix Σ can still be reproduced.
- An orthogonal transformation corresponds to a rigid rotation or reflection of the coordinate axes.
- Hence the orthogonal transformation of the factor loadings (and the implied transformation of the factors) is called a *factor rotation*.
- From a mathematical viewpoint, it is immaterial whether Λ or $\Lambda^* = \Lambda\mathbf{G}$ is reported since

$$\Sigma = \Lambda\Lambda^T + \Psi = \Lambda\mathbf{G}\mathbf{G}^T\Lambda^T + \Psi = \Lambda^*{\Lambda^*}^T + \Psi.$$

- However, when the objective is statistical interpretation, it may be that one rotation is more useful than alternatives.

Controversial?

- This feature of factor analysis is not without controversy.
- Some authors argue that by rotating the factors the analyst is in someway manipulating the results.
- Others argue that the technique is simply akin to sharpening the focus of a microscope so as to see the detail more clearly.
- In a sense the orientation of the factor solution reported is arbitrary anyhow.
- In any case, factor analysis is generally used as an intermediate step to reduce the dimensionality of a data set prior to other statistical analyses (similar to PCA).

A Simpler Structure

- The idea of a "simpler structure" needs further explanation.
- One ideal would be to rotate the factors so that each variable has a large loading on a single factor and small loadings on the others.
- The variables can then be split into disjoint sets, each of which is associated with one factor.
- A factor j can then be interpreted as an average quality over those variables i for which λ_{ij} is large.

Kaiser's Varimax Rotation

- First note that the squared loading λ_{ij}^2 is the proportion of the variance in variable i that is attributable to common factor j:

$$\begin{aligned} \mathbf{Var}[X_i] &= \lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{ip}^2 + \psi_i \\ &= h_i^2 + \psi_i. \end{aligned}$$

- We aim for a rotation that makes the squared loadings λ_{ij}^2 either large or small, *i.e.*, few medium-sized values.
- Let $\tilde{\lambda}_{ij}^* = \lambda_{ij}^*/h_i$ be the final rotated loadings scaled by the square root of the communalities.

Kaiser's Varimax Rotation

- The varimax procedure selects the orthogonal transformation $\mathbf{G}$ maximizing the sum of the column variances across all factors $j = 1, \ldots p$:

$$V = \frac{1}{m}\sum_{j=1}^{p}\sum_{i=1}^{m}\tilde{\lambda}_{ij}^{*4} - \frac{1}{m^2}\sum_{j=1}^{p}\left(\sum_{i=1}^{m}\tilde{\lambda}_{ij}^{*2}\right)^2$$

$$\propto \sum_{j=1}^{p}(\text{variance of the squares of scaled factor loadings for } j\text{th factor})$$

- Maximizing V corresponds to 'spreading out' the squares of the loadings on each factor as much as possible. Hence both groups of large and negligible coefficients are found in any column of Λ^*.

- Scaling the rotated loadings has the effect of giving variables with small communalities relatively more weight. After $\mathbf{G}$ has been determined the loadings $\tilde{\lambda}_{ij}^*$ are multiplied by h_i to ensure the original communalities are preserved.

Oblique Rotations

- The degree of correlation allowed between factors is generally small (two highly correlated factors $\equiv$ one factor).
- An oblique rotation corresponds to a *non-rigid* rotation of the coordinate system, *i.e.*, the resulting axes need no longer be perpendicular.
- Oblique rotations, therefore, relax the orthogonality constraint in order to gain simplicity in the interpretation.
- The **promax rotation** is an example of an oblique rotation. Its name derives from Procrustean rotation.
- Oblique rotations are much less popular than their orthogonal counterparts, but this trend may change as new techniques, such as independent component analysis, are developed.

Factor scores

- Interest usually lies in the parameters of a factor model (*i.e.*, λ_{ij} and ϵ_i) but the estimated values of the common factors (the **factor scores**, *i.e.*, $\hat{\mathbf{f}}$) may also be required.
- The location of each original observation in the reduced factor space is often necessary as input for a subsequent analysis.
- One method which can be used to estimate the factor scores is **the regression method**.
- From multivariate normal theory it can be shown that:

$$\hat{\mathbf{f}} = \mathbf{\Lambda}^T \Sigma^{-1} \mathbf{X}$$

Decathlon

- The men's decathlon event in the Seoul '98 Olympic games involved 34 competitors who each competed in the following disciplines:
  100 metres, Long jump, Shot Put, High jump, 400 metres , 110 metre hurdles, Discus, Pole vault, Javelin, 1500 metres.
- The result for each competitor in each event was recorded, along with their final overall 'score'.
- Factor analysis can be used to explore the structure of this data.
- In order to ease interpretation of results, the results of the 100m, 400m, hurdle event and the 1500m event were converted to the negative of the result (so a high score in any event implies a good performance).
- Most pairs of events are positively correlated to some degree.
- Factor analysis attempts to explain the correlation between a large set of variables in terms of a small number of underlying latent factors.

Decathlon

- A factor analysis model with four factors was fitted to the data with the following results:

```
Uniquenesses:
    m100m   LongJump    Shot   HighJump    m400m
    0.445      0.594   0.005      0.849    0.005

    mHurdles  Discus   PoleVault  Javelin  m1500m
       0.155   0.280       0.568    0.617   0.005

Loadings:
          Factor1 Factor2 Factor3 Factor4
m100m      0.652                   0.342
LongJump   0.483           0.261   0.316
Shot               0.970  -0.109   0.184
HighJump                           0.368
m400m      0.932  -0.148   0.323
mHurdles   0.654   0.111           0.630
Discus             0.806  -0.254
PoleVault  0.426   0.390           0.311
Javelin            0.609
m1500m     0.278  -0.222   0.925   0.112
```

```
                Factor1 Factor2 Factor3 Factor4
SS loadings       2.237   2.213   1.125   0.903
Proportion Var    0.224   0.221   0.113   0.090
Cumulative Var    0.224   0.445   0.557   0.648
```

Interpretation

- Note that the communalities are the sum of the squared loadings in each row of the loadings matrix.
- Also, the communality for each variable added to the uniqueness for each variable is the total variance for that variable (this is equal to 1 here as `R` initially standardizes the data).
- The SS loadings values are the sum of the squared loadings for that factor.
- The proportion of total standardized sample variance due to jth factor is:

$$= \frac{\hat{\lambda}_{1j}^2 + \hat{\lambda}_{2j}^2 + \cdots + \hat{\lambda}_{mj}^2}{\sigma_{11} + \sigma_{22} + \cdots + \sigma_{mm}} = \frac{\hat{\lambda}_{1j}^2 + \hat{\lambda}_{2j}^2 + \cdots + \hat{\lambda}_{mj}^2}{m}$$

  Hence the 'Proportion Var' values are the SS loadings values divided by $m = 10$.

- The four factors account for 64.8% of the total variance of the data.

Interpretation Continued

- The default setting in **R** is to perform a varimax rotation to the resulting factors (so there are few 'intermediate' loadings reported).
- Factor 1: 100 metres, 400 metres and the hurdles event load highly on this factor, so it could be interpreted as the *running speed* factor.
- Factor 2: The shot put, discus and javelin load highly on this factor, so it could be interpreted as *arm strength*.
- Factor 3: The 1500 metres event loads highly on this factor, so it could be interpreted as *endurance*.
- Factor 4: Hurdles, high jump, 100 metres, long jump and the pole vault all load higher on this factor than the other events, so it could be interpreted as *leg strength*.

PCA vs Factor Analysis

- Since both PCA and factor analysis have similar aims (*i.e.*, to reduce the dimensionality of a data set), it is worth highlighting the difference in the two approaches.
  - PCA looks for linear combinations of the data matrix $\mathbf{X}$ that are uncorrelated and of high variance, whilst factor analysis seeks unobserved linear combinations of the variables representing underlying fundamental quantities.
  - PCA makes no assumptions about the form of the covariance matrix, whilst factor analysis assumes that the data comes from a well-defined model in which specific assumptions hold (*e.g.*, $\mathbb{E}[\mathbf{f}] = 0$, *etc*).
  - PCA: data $\Rightarrow$ PCs. FA: factors $\Rightarrow$ data.
  - When specific variances are large they are absorbed into the PCs whereas factor analysis makes special provision for them. When the specific variances are small, PCA and FA give similar results.

PCA vs Factor Analysis

- The two analyses are often performed together. For example, you can conduct a principal components analysis to determine the number of factors to extract in a factor analytic study.