

Cliff Notes for Bartholomew and Knott (BK) Fall 2003

Chapter 2

- 2.1 - 2.5 (very theoretical - deriving sufficient statistics in general latent variable model)
- 2.6 brief of latent trait and latent class
- 2.7 about normal model (still quite technical)
- 2.8 nonlinear models
- 2.9 distribution free fitting
- 2.10 very brief maximum likelihood (not very informative)
- ***** 2.11 generally about rotation, talks about orthogonal and oblique (OK for reading)
- ***** 2.12 naming latent variables (simple structure) (OK for reading)
- ***** 2.13 talks about using components X which are actually the factor score estimates. (OK for reading)
 - 2.14 brief discussion of probs of standard errors (difficult to read)
 - 2.15 tries to claim that standard normalis just fine for $h(y)$ (somewhat philisophical)
- ***** 2.16 Posterior analysis - obaining factor scores - gives a proof that factor score estimates have the same ordering as expected value of true latent variable (2 full pages of proof but still worth reading the rest)
- ***** 2.17 comparing populations in light of standardizing latent variables (OK for reading)
 - 2.18 reference to bayesian inference and the idea of sampling variables from a pop of all variables (very peripheral to our current study)

Chapter 3

• 3.1 The Model

The Normal factor analysis model is introduced. By Normal it means the distribution of the p observed variables \mathbf{X} are assumed to follow the Normal distribution. There are q underlying latent factors assumed and each observed variable is a linear combination of these factors \mathbf{f} plus error $\boldsymbol{\epsilon}$.

Here is my version of equation (3.3) in BK, note the change from \mathbf{y} to \mathbf{f} and the change from \mathbf{e} to $\boldsymbol{\epsilon}$.

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{f} + \boldsymbol{\epsilon} \quad (1)$$

It is assumed that this equation holds for each individual in the population and thus for the $i = 1 \dots n$ independently sampled individuals we have

$$\mathbf{x}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{f}_i + \boldsymbol{\epsilon}_i \quad (2)$$

The $q \times 1$ vectors of factors \mathbf{f}_i are considered i.i.d. multivariately Normally distributed with mean zero and variance Φ which in BK is set to $\mathbf{I}_{q \times q}$ and the $p \times 1$ vectors of errors ϵ_i are also assumed multivariately normally distributed with mean zero and variance Ψ where Ψ is assumed to be diagonal. Because the factors are normal and the errors are normal, that is why we can assume the observed \mathbf{x}_i are normal (i.e. linear combinations of normals are normal). It is further assumed that the factors are independent of the errors, i.e. since everything is normal it suffices to say $Cov(\mathbf{f}_i, \epsilon_i) = 0$.

The elements along the diagonal of $Var(\epsilon_i)$ are called "specific variances" or "specificities" or "uniquenesses". The reason for this name is that the variability due to each element ϵ is "specific" or "unique" to that particular observed variable, in contrast to variability coming from the factors which would be "shared" or "common" to several observed variables. Sometimes the factors \mathbf{f} are called "common factors".

The elements of the Λ matrix are called factor loadings (find out why the term "loading" is used).

NOTE: Because this entire chapter is actually focusing on the Exploratory factor analysis model (rather than the confirmatory factor analysis model), BK make the assumption that the factors are uncorrelated with variance 1 and assumes that the errors are uncorrelated with possibly different variances, when they write down the model. But in general this is not necessary. There are other ways of identifying the exploratory factor analysis model that do not require the factors to be orthogonal.

Equation (3.4) in BK will be used later when forming factor score estimates...

$$\mathbf{X} = \Lambda' \Psi^{-1} \mathbf{x} \quad (3)$$

This equation was derived back in chapter 2 when the authors were discussing "sufficient statistics". This equation will be used to create *factor score estimates*. Note that this equation takes the $p \times 1$ vector of observed variables \mathbf{x} and transforms them into a $q \times 1$ vector \mathbf{X} .

As an example of how to write equation (1), here are the equations when $p = 5$ and $q = 2$

$$\begin{matrix} x_{1i} & \mu_1 & & & & & & \epsilon_{1i} \\ x_{2i} & \mu_2 & & & & & & \epsilon_{2i} \\ x_{3i} & \mu_3 & + & \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \\ \lambda_{31} & \lambda_{32} \\ \lambda_{41} & \lambda_{42} \\ \lambda_{51} & \lambda_{52} \end{pmatrix} & \begin{matrix} f_{1i} \\ f_{2i} \end{matrix} & + & \begin{matrix} \epsilon_{3i} \\ \epsilon_{4i} \\ \epsilon_{5i} \end{matrix} \end{matrix} \quad (4)$$

$$Var(\mathbf{f}_i) = Var \begin{pmatrix} f_{1i} \\ f_{2i} \end{pmatrix} = \Phi = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad (5)$$

$$Var(\epsilon_i) = Var \begin{pmatrix} \epsilon_{1i} \\ \epsilon_{2i} \\ \epsilon_{3i} \\ \epsilon_{4i} \\ \epsilon_{5i} \end{pmatrix} = \Psi = \begin{pmatrix} \psi_1 & 0 & 0 & 0 & 0 \\ 0 & \psi_2 & 0 & 0 & 0 \\ 0 & 0 & \psi_3 & 0 & 0 \\ 0 & 0 & 0 & \psi_4 & 0 \\ 0 & 0 & 0 & 0 & \psi_5 \end{pmatrix} \quad (6)$$

• 3.2-3.3 Some Distributional Properties

This section has a lot of equations but only two of them are really important for you to understand

now.

The first is (3.5). Based on model (1) and the assumptions made about the distribution, variances and covariances of the \mathbf{f} and $\boldsymbol{\epsilon}$, we have

$$\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}) \quad (7)$$

Recall, BK is assuming $\boldsymbol{\Phi} = \mathbf{I}$, so they instead have $\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$. The importance of this equation is simply that you recognize that the variance of \mathbf{x} is separable into a part coming from the factors and a part coming from the errors. That leads to the second equation in this section that you should understand which is equation (3.10) of BK where they write (with slightly different subscript notation)

$$Var(x_j) = \sum_{k=1}^q \lambda_{jk}^2 + \psi_j \quad (8)$$

For the example given before with $p = 5$ and $q = 2$ we have for say the 4th observation x_4

$$Var(x_4) = \lambda_{41}^2 + \lambda_{42}^2 + \psi_4$$

The sum of the squared factor loadings is called the "communality", that is it is the part of variance of x_4 (in this example) that is coming from the factors that x_4 shares in "common" with the other observed variables. **Note if a particular observed variable has a very small "communality" then we may decide to drop it from the vector \mathbf{x} of indicators (or measurements) of the underlying factors.**

BK keeps foreshadowing to the eventual use of factor score estimation (or what they call Posterior Analysis in Section 3.23) and that is what all the other equations in this section are related to. In particular equation (3.6) is the conditional distribution of the underlying factors given the observed variables. We'll revisit this later.

• 3.4 Constraints on the Model

This very short section says several important things.

The first is about the use of an additional constraint on the factor analysis model (1) in order to remove the freedom of the $\boldsymbol{\Lambda}$ to arbitrarily rotate (See BK section 2.11 for more info about rotation). Even after restricting Φ_{ii} to be \mathbf{I} and Ψ_{ii} to be diagonal, the solution to the factor analysis model (1) is not unique. That is, there is more than one set of factor loadings $\boldsymbol{\Lambda}$ that will give the exact same answer. There are an infinite number of ways to fix this "rotation", each one corresponding to a different projection of the p dimensional space into a q dimensional space. To fix the solution so that one particular "rotation" can be examined, it is necessary to put an additional $q(q-1)/2$ restrictions on the elements of $\boldsymbol{\Lambda}$. These restrictions correspond to choosing which projection will be examined.

BK describe what may be considered the canonical restriction to fix rotation, that is they restrict $\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}$ to be diagonal. Hence, the off diagonal terms are being fixed to zero, i.e. $q(q-1)/2$ different restrictions. Of all the possible restrictions that could be placed (in order to fix the rotation), this

particular restriction yields some nice distributional properties for the conditional distribution of $\mathbf{f}|\mathbf{x}$ and also makes the eventual maximum likelihood estimation algorithm simpler.

Then BK foreshadow the use of confirmatory factor analysis and point out that when you pre-specify the pattern of the elements in $\mathbf{\Lambda}$ by fixing certain elements to zero, there will be no issue of rotation.

Finally BK talks about standardizing the \mathbf{x} variables. Although they do not point this out here, in general the \mathbf{x} variables should not be standardized before fitting the model. Although the maximum likelihood method produces compatible results for standardized data as compared to unstandardized data, not all discrepancy functions do. The most common advice given is to use unstandardized data for estimation and then standardize the estimates afterwards.

- **3.5 Maximum Likelihood Estimation**

Although BK does not go into it, we should be careful to think about what $\mathbf{\Sigma}$ represents. There is some population covariance matrix for the random variables \mathbf{x} and we want to model this covariance matrix. To do this we propose the factor analysis model (1) which yields a certain parametric covariance matrix $\mathbf{\Sigma}(\boldsymbol{\theta}) = \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}' + \mathbf{\Psi}$ where $\boldsymbol{\theta} = (\mathbf{\Lambda}, \mathbf{\Phi}, \mathbf{\Psi})$. The goal is to fit this parametric covariance matrix to the observed data taken as a random sample from the population. To do this, this section in BK talks about using Maximum likelihood.

Normal distribution and the likelihood function

Recall that maximum likelihood estimation asks us to find the $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ which maximize the likelihood L (or the log likelihood).

Given n i.i.d. random vectors $\mathbf{x}_i \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})$ the log likelihood is

$$\log L = \frac{pn}{2} \log 2\pi - \frac{n}{2} \log |\mathbf{\Sigma}| - \frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})' \mathbf{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})$$

After a bit of matrix algebra, this can be transformed into

$$\log L = \text{constant} + \frac{n}{2} \left[\log |\mathbf{\Sigma}^{-1}| - \text{trace}(\mathbf{S}\mathbf{\Sigma}^{-1}) \right] - \frac{n}{2} (\bar{\mathbf{x}} - \boldsymbol{\mu})' \mathbf{\Sigma}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu})$$

where \mathbf{S} is the sample covariance matrix. Since $\boldsymbol{\mu}$ only appears in the last term, we can maximize the likelihood with respect to $\boldsymbol{\mu}$ by minimizing the last term with respect to $\boldsymbol{\mu}$. This is clearly done when $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$

So what we are left with doing is maximizing the remaining part with respect to the parameters in $\mathbf{\Sigma}$, i.e. with respect to $\boldsymbol{\theta} = (\mathbf{\Lambda}, \mathbf{\Phi}, \mathbf{\Psi})$. So we want to maximize the following:

$$\log L = \frac{n}{2} \left[\log |\mathbf{\Sigma}^{-1}| - \text{trace}(\mathbf{S}\mathbf{\Sigma}^{-1}) \right] \tag{9}$$

This is equation (3.11) in BK. The next step is to actually perform the maximization.

Maximizing the Likelihood

Most of section 3.5 is describing how to maximize the likelihood (9). It is not crucial that you follow the rest of this section, but I have provided a brief sketch below:

- The maximization is done by iteration.
- Needs starting values
- The iteration starts by taking the eigenvalues of the following: $(\hat{\Psi}_{(0)}^s)^{-\frac{1}{2}}\mathbf{R}(\hat{\Psi}_{(0)}^s)^{-\frac{1}{2}} - \mathbf{I}_{(\mathbf{p} \times \mathbf{p})}$
This is implicit in the discussion of p. 47 of BK. It is the eigenvalues of this that appear on the first page of the PROC FACTOR output.
- I don't think I can provide any simple intuition about this iterative technique. Just trust that it converges to the Maximum Likelihood Estimators.

We will call the maximum likelihood estimators: $\hat{\theta} = (\hat{\Lambda}, \hat{\Psi}, \hat{\Phi})$, and $\hat{\Sigma} = \Sigma(\hat{\theta})$ will be referred to as the “fitted model covariance matrix” or “Implied covariance” (in AMOS) or “estimated covariance matrix” (in MPLUS).

• 3.6 Maximum likelihood by the EM algorithm

This section present another way to actually find the maximum likelihood estimators in (9). The EM algorithm stands for the Expectation Maximization algorithm (Dempster, Laird, and Rubin, 1977) and is a very useful method for models where there is missing information. The “missing” information in the factor analysis model can be thought of as the latent factors. If we actually knew the factor for each person then estimating $\theta = (\Lambda, \Psi, \Phi)$ would be trivial.

Why do we need two ways to maximize the likelihood? Well, actually for the normal factor analysis problem both of the methods work just fine, but for other kinds of latent variable model, in particular when the observed variables are not normal, the direct method described in section 3.5 will not work, and so the EM algorithm can be used. BK are simply introducing the idea of the EM algorithm here when the problem is still a nicely behaving normal problem.

The log likelihood function in (9) is a function of observed data and parameters. The way the EM algorithm works is to create a likelihood function that is a function of observed data, unobserved data (i.e. latent variables) and parameters. This is called the “complete data log likelihood”. The algorithm calculates the Expected value of the complete data log likelihood conditional on the observed data (using the current estimates of the parameters at that iteration), then this Expected value is maximized to get new estimates of the parameters and these two steps are iterated until convergence. The Dempster, et al. (1977) proves that at each step that the true likelihood (i.e. the observed data log likelihood) will never decrease at any iteration. This proof tells us that this algorithm is a valid way to maximize the log likelihood.

It is not crucial that you follow the formulas in this section.

• 3.7 Sampling Variation of Estimators

This section presents an example that they refer to as Johnson and Wichern's (1982) data and they talk about how we do not have a good way to test whether the factor loadings are significant or not in an exploratory factor analysis.

A common rule of thumb that is used to determine if a factor loading is “significant” or “big” is if it is greater than 0.3. BK refer to a paper by Cudeck and O'Dell (1994) “Applications of standard error estimates in unrestricted factor analysis: significance tests for factor loadings and correlations”

who show that this cut-off is not related to any kind of statistical significance, that is there can be smaller loadings which are significant and larger loadings which are not significant.

BK describe some of the work that has been done in order to come up with a way to test the significance in EFA. While the original work for standard errors in EFA were worked out for the rotation that fixes $\Lambda'\Psi\Lambda$ to be diagonal, there are now methods for calculating standard errors even for obliquely rotated loadings, in particular Ogasawara (1998) "Standard errors for rotation matrices with an application to the promax solution". There is a specialized program available from Michael Browne's website (at Ohio State University) that can calculate these standard errors. All of these results are based on asymptotic results, thus it is still not clear how useful they are for problems with small sample sizes.

The 7 variables (in Table 3.1) are measures taken on salesmen and it was desirable to see if a factor analysis could be used to reduce the dimension from 7 to something smaller so that the salesmen could be assessed. The first 3 variables are a measure of sales performance and the other four were from tests of aptitude. What BK are showing through this example is another way to calculate standard errors which is to use the "bootstrap" or "jackknife" technique. The factor loadings shown in Table 3.1 are from the $\Lambda'\Psi^{-1}\Lambda$ fixed to be diagonal rotation. Basically they show that factor loading of 0.646 on variable 4 is "not significant" using the bootstrap technique and thus emphasizing their point that loadings bigger than 0.3 are not necessarily "statistically significant".

BK point out that much more work is urgently needed to extend and consolidate our limited knowledge in this area of standard error estimation for EFA.

CROSS VALIDATION. If the sample size is large enough, a simpler way to gain an idea about the stability of the factor structure (rather than relying on rules of thumb) is to split the sample randomly into two equal parts and then fit the model to each part. If the factor structure is similar, this "tends to increase our confidence in the genuineness of the factors".

• **3.8-3.9 Goodness of Fit and Choice of q**

Besides giving estimates for Λ , Φ and Ψ , Maximum Likelihood Provides a GOODNESS OF FIT TEST.

- Use Likelihood Ratio Test, i.e.,

$$\begin{aligned} -2 \log \frac{L(\text{MLE of restricted model})}{L(\text{MLE of unrestricted model})} \\ &= 2\{\log(L(\mathbf{S})) - \log(L(\hat{\underline{\Sigma}}))\} \\ &= n\{\text{trace}\hat{\underline{\Sigma}}^{-1}\mathbf{S} - \log|\hat{\underline{\Sigma}}^{-1}\mathbf{S}| - p\} \end{aligned}$$

- IF the model fits the data well, this statistic should be small, thus the p-value will be big!
- Its distribution is asymptotically distributed χ^2 with degrees of freedom = $(\frac{p(p+1)}{2} - \text{number of unique parameters in model})$.
- (**SEE section 3.16**) For the EFA model the d.f. are specifically $\frac{p(p+1)}{2} - pq - p + \frac{q(q-1)}{2}$. The first term is the number of unique elements in the sample covariance matrix \mathbf{S} , then we are going to estimate pq factor loadings in Λ and p variances in Ψ and since we have to put restrictions in order to fix rotation we get those $\frac{q(q-1)}{2}$ d.f. back. These $\frac{q(q-1)}{2}$ d.f. correspond to fixing the $\Lambda'\Psi^{-1}\Lambda$ off diagonal elements to be zero.

- We can determine if the the statistic is “small” enough by comparing to the χ^2 distribution and obtaining a p-value.
- The Hypothesis being tested
 - H_0 : The model is correct (i.e. q factors sufficiently describe the p dimensional vector)
 - H_A : More factors are needed (i.e. a less restrictive model is needed)
- Thus we are looking for the model where we DO NOT REJECT the H_0 (i.e. find a big p-value)
- ”Starting with $q = 1$, we then take successive values in turn ntil the fit of themodel is udedged to be adequate. **Viewed as a testing procedure this is not stricly valid because it does not adjust the significance levels to allow for the sequential character of the test. It rather depends on regarding the p-value of the test as a measure of the adequacy of the model”**
- Because of the tendency for the approach described just above to keep adding more and more factors (since more factors will make the model fit better), BK suggest considering the AIC Akaike’s information criterion. This is simply a penalized version of the log likelihood where models with more parameters are penalized.
- FROM KLINE, page 209-210 Satorra Bentler correction to the Chi-square statistic when the data is non-normal.

- **3.10-3.11 Fitting without Normality Assumptions: Least Squares Methods**

A general class of discrepancy functions is given by

$$F = trace\{(\mathbf{S} - \mathbf{\Sigma})\mathbf{V}\}^2 \tag{10}$$

where \mathbf{V} represents different ways of weighting the difference between the observed covariance matrix and the model covariance matrix.

Mplus and AMOS will allows various choices for \mathbf{V} . Unweighted least squares, weighted least squares and asymptotically distribution free.

- **3.12 Approximate methods for estimating Ψ**

This section is not crucial

- **3.13-3.15 Goodness of fit and choice of q for least squares methods**

Section 3.13 talks about how the chi-square test has been proven to be asymptotically valid even when the data is not normally distributed (Amemiya and Anderson, 1985).

Section 3.14 describes a method of choosing q (i.e. the number of underlying factors) by taking it to be the number of eigenvalues ≥ 1 . They report a paper by Fachel 1986 (actually this is Fachel’s dissertation) that showed via simulation that this method tends to overestimate the number of factors.

Section 3.15 describes the use of the “scree test”. That is, plotting the ordered eigenvalues and then looking for the elbow in the plot. q is taken to be the numbered eigenvalue where the elbow appears.

- **3.16 Further Estimation Issues - Consistency**

This section gives the same formula that was given above for the d.f in the EFA model. It also turns this equation around to point out that since the d.f. must be greater than or equal to zero this places a restriction on the total number of factors that can be fit to p observed variables.

The formula given in (3.42) can be confusing because it is only a necessary condition, it is not sufficient. In order to check if you can fit say 3 factors to 5 variables, just plug into the d.f. given above to see if it is greater than or equal to zero, e.g.

$$p = 5, q = 3,$$

$$5 * 6/2 - 5 * 3 - 5 + 3 * 2/2 = -2$$

Thus we can't fit 3 factors, what about...

$$p = 5, q = 2,$$

$$5 * 6/2 - 5 * 2 - 5 + 2 * 1/2 = 1$$

Thus we **can** fit 2 factors

- **3.17 Further Estimation Issues - Scale invariance**

This section is describing the issue of whether to fit the observed covariance matrix or the observed correlation matrix. They point out that the same result is obtained when using maximum likelihood. The details of this section are not crucial.

- **3.17 Further Estimation Issues - Heywood Cases**

To be done

- **3.19-3.22 Rotation and related matters**

This section doesn't talk about the issue of rotation, it instead briefly talks about how the algorithms to find the different rotated factor loadings are done. It discusses the orthogonal rotation algorithm and oblique rotations (i.e. when the factors are allowed to correlate)

- **3.23-3.25 Posterior analysis "Factor Scores"**

To be done

- **3.26-3.28 Examples**

I will not create Cliff notes for these examples.

Parts of Chapters 4 and 5 will also be discussed for Factor analysis with categorical observed variables.