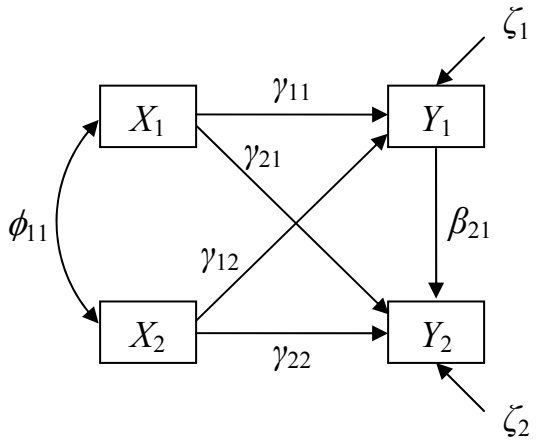
Covariance Structure and Factor Models

Mid-term exam answers

Andrej: All statements in brackets are supplementary, comments to you and/or given to help the class understand better, and so they are not necessary for a full credit.

1. The following diagram defines a structural equation model for data with two independent and two dependent manifest variables. For convenience, suppose all variables have a mean of zero. Write the structural equation for each of the dependent variables as a linear function exclusively of independent variables. For this and the following question, use an explicit notation for the individual variables and parameters, instead of using the vector/matrix notation that collects the same kinds of variables and parameters in vectors or matrices. For example, don't combine X_1 and X_2 into a two-element vector $\mathbf{x}' = [X_1, X_2]$.



There are four independent variables X_1, X_2, ζ_1 and ζ_2, and so the dependent variables Y_1 and Y_2 should be written as linear combinations (functions) of only these independent variables as follows:

$$Y_1 = \gamma_{11}X_1 + \gamma_{12}X_2 + \zeta_1$$

$$\begin{aligned} Y_2 &= \gamma_{21}X_1 + \gamma_{22}X_2 + \beta_{21}Y_1 + \zeta_2 \\ &= \gamma_{21}X_1 + \gamma_{22}X_2 + \beta_{21}(\gamma_{11}X_1 + \gamma_{12}X_2 + \zeta_1) + \zeta_2 \\ &= (\gamma_{21} + \beta_{21}\gamma_{11})X_1 + (\gamma_{22} + \beta_{21}\gamma_{12})X_2 + \beta_{21}\zeta_1 + \zeta_2 \end{aligned}$$

2. From the two equations you wrote for question (1), derive the equation to represent the covariance between the two dependent variables (call it COV(Y_1,Y_2)) exclusively by model parameters.

[Andrej: The validity of answer to this question depends on how students answered to question (1). And so the dependency should not be graded. That is, if a student correctly derived from a wrong answer to (a), a full credit should be given.] Covariance between Y_1 and Y_2 are derived from the equations given for question (1) as follows:

$$\begin{aligned}&\mathrm{COV}(Y_1, Y_2) = E(Y_1 Y_2)\\ &= E\Big[(\gamma_{11}X_1 + \gamma_{12}X_2 + \zeta_1)\big((\gamma_{21} + \beta_{21}\gamma_{11})X_1 + (\gamma_{22} + \beta_{21}\gamma_{12})X_2 + \beta_{21}\zeta_1 + \zeta_2\big)\Big]\end{aligned}$$

Since the error terms are not correlated with each other and with X_1 and X_2, it becomes

$$\begin{aligned}\mathrm{COV}(Y_1, Y_2) &= \gamma_{11}(\gamma_{21} + \beta_{21}\gamma_{11})E(X_1X_1) + \gamma_{11}(\gamma_{22} + \beta_{21}\gamma_{12})E(X_1X_2)\\ &\quad + \gamma_{12}(\gamma_{21} + \beta_{21}\gamma_{11})E(X_2X_1) + \gamma_{12}(\gamma_{22} + \beta_{21}\gamma_{12})E(X_2X_2)\\ &= \gamma_{11}(\gamma_{21} + \beta_{21}\gamma_{11})\phi_{11} + \big[\gamma_{11}(\gamma_{22} + \beta_{21}\gamma_{12}) + \gamma_{12}(\gamma_{21} + \beta_{21}\gamma_{11})\big]\phi_{12}\\ &\quad + \gamma_{12}(\gamma_{22} + \beta_{21}\gamma_{12})\phi_{22}\\ &= (\gamma_{11}\gamma_{21} + \gamma_{11}^2\beta_{21})\phi_{11} + (\gamma_{11}\gamma_{22} + \gamma_{12}\gamma_{21} + 2\gamma_{11}\gamma_{12}\beta_{21})\phi_{12}\\ &\quad + (\gamma_{12}\gamma_{22} + \gamma_{12}^2\beta_{21})\phi_{22}\end{aligned}$$

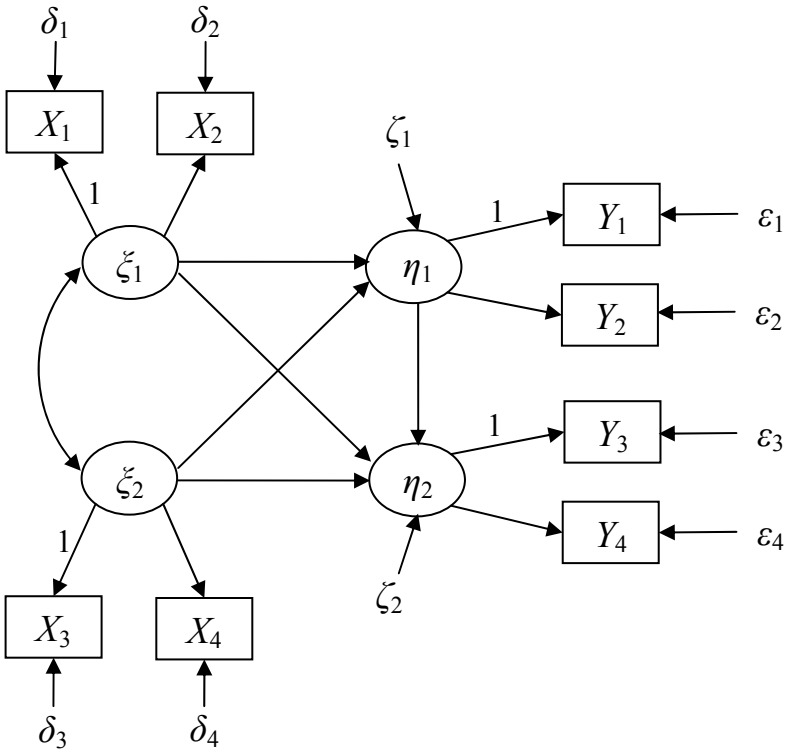
[As I announced during the exam, ϕ_{11} indicating the covariance between X_1 and X_2 in the diagram model was a typo. It should have been ϕ_{12}. If the mistaken ϕ_{11} was taken as it was, it can be considered as an equality constraint $\phi_{12} = \phi_{11}$. Then, the resulting derivation will be:

$$\begin{aligned}\mathrm{COV}(Y_1, Y_2) &= \big[\gamma_{11}(\gamma_{21} + \beta_{21}\gamma_{11}) + \gamma_{11}(\gamma_{22} + \beta_{21}\gamma_{12}) + \gamma_{12}(\gamma_{21} + \beta_{21}\gamma_{11})\big]\phi_{11}\\ &\quad + \gamma_{12}(\gamma_{22} + \beta_{21}\gamma_{12})\phi_{22}\\ &= (\gamma_{11}\gamma_{21} + \gamma_{11}\gamma_{22} + \gamma_{12}\gamma_{21} + \gamma_{11}^2\beta_{21} + 2\gamma_{11}\gamma_{12}\beta_{21})\phi_{11}\\ &\quad + (\gamma_{12}\gamma_{22} + \gamma_{12}^2\beta_{21})\phi_{22}\end{aligned}$$

In addition, it should not change the identifiability of the model: With the equality constraint, the model is over-identified and the recursive rule still holds, and so it's identifiable. Andrej: if a student does this, a full credit should be given as well.]

3. For the model shown above, determine whether the model is identifiable. Whichever you determine, explain why it is, or is not, identifiable. (3 points) (b) The following model is identical to the model shown above except that each manifest variable in question (1) is now measured by two indicators as follows. Treat all unnamed path coefficients, loadings and covariances as free parameters (except for the path from an

error term to its DV which is always set to 1). Determine the identifiability of this model by the two-step rule and explain your conclusion. (4 points)



(a) The path model given for question (1) is identifiable since it satisfies the recursive rule (i.e., there is no feedback loop of causal paths and the two error terms are not correlated with each other), which is a sufficient condition for identification. [In particular, it's just identifiable since the number of distinctive data points (= 4×5/2 = 10) is identical to the number of distinctive parameters: 5 path coefficients (γ_{11}, γ_{12}, γ_{21}, γ_{22}, β_{21}), 4 variances of IV's (ϕ_{11}, ϕ_{22}, ψ_{11}, ψ_{22}), and 1 covariance between the two IV's (ϕ_{12}).]

(b) In the first step of the two-step rule, we treat all 4 latent variables fully correlated and look at the identifiability of the resulting confirmatory factor model. This CFA model has a uni-factorial loading pattern with two indicators per factor (that is, every indicator is influenced by only one factor and every factor is measured by two indicators). In addition, all measurement error terms are mutually uncorrelated and all factors are allowed to correlate with each other. In consequence, this model satisfies the two indicator rule which is sufficient for identifiability and so passes the first step.

In the second step of the two-step rule, we only consider the identifiability of the path modeling part among the latent variables, ignoring all indicators, which reduces to the path model given for question (1) and is identifiable. Since the two-step rule is a sufficient condition for identification of a general model, this general model is identifiable. [Andrej: the validity of identifiability of the second step per se should not be graded since it was asked before. Instead, how it's used for the two-step rule should be graded. That is, if a student answers the

earlier question wrong, but correctly answers for the first step and accordingly concludes that this general model is not identifiable, then the student should get a full credit.]

4. For a comparison between a nesting and a nested model, one can use the likelihood ratio test, the Lagrangian multiplier test or the Wald test. While the LR test requires fitting both models, fitting only one model is sufficient for the other two tests. Specifically, the LM test only uses estimates for the nested model whereas the Wald test only uses estimates for the nesting model. Particularly when only one parameter is considered, the LM and the Wald statistic are readily available in most SEM fitting program, namely modification index and squared critical ratio $\left(\theta^2/\text{avar}(\theta)\right)$, respectively. For such a case of *df*-difference = 1 with a constraint $\theta = 0$ for a particular parameter in the nested model, state the null hypothesis for each of the LM and the Wald test. When statistical results suggest a rejection of both null hypotheses, what should you do with the currently fit model that is either the nesting or the nested model?

   For both statistics, the null hypothesis is the same: $\theta = 0$ [or equivalently, $\chi^2_{\text{nested}} - \chi^2_{\text{nesting}} = 0$, or MI = 0 and W = 0, all defined in the population]. However, the same null hypothesis is tested differently by the LM and the Wald statistic, though a rejection of them imply "necessarily" the same conclusion, $\theta \neq 0$. Since the LM statistic is based on the nested model with the wrong constraint $\theta = 0$ imposed, we should reject the fit model and respecify θ as a free parameter. The Wald statistic is computed with estimates for the nesting model where θ is correctly specified as a free parameter, and its specification is statistically supported. Thus, we should retain θ as a free parameter.

5. Comparative fit indices Δ_1 (a.k.a., Normed Fit Index) and ρ_2 (a.k.a., Tucker-Lewis Index) are defined as follows:

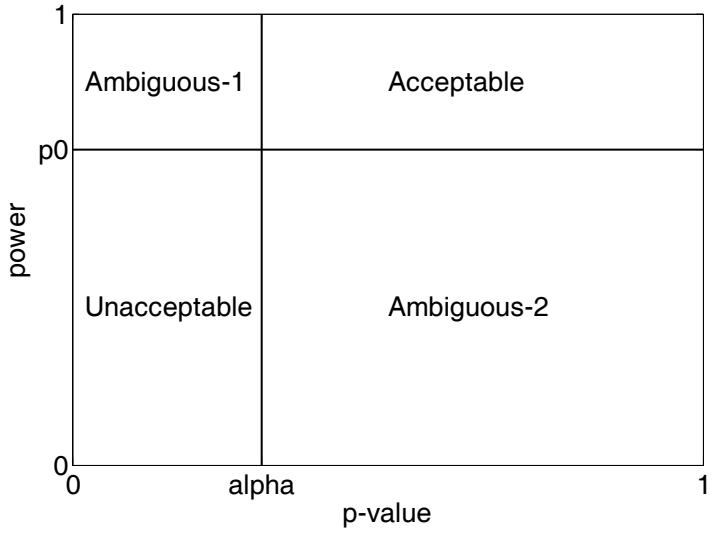
$$\Delta_1 = \frac{F_b - F_m}{F_b} = \frac{\chi_b^2 - \chi_m^2}{\chi_b^2}, \qquad \rho_2 = \frac{\frac{F_b}{df_b} - \frac{F_m}{df_m}}{\frac{F_b}{df_b} - \frac{1}{N-1}} = \frac{\frac{\chi_b^2}{df_b} - \frac{\chi_m^2}{df_m}}{\frac{\chi_b^2}{df_b} - 1}$$

   where F, χ^2, and df represent, respectively, fit function estimates, chi-square estimates and model degrees of freedom, and the subscripts b and m indicate, respectively, the baseline (or independence) and a hypothesized model. Both of these (goodness of fit) indices are relative to a baseline fit as a worst fit given the data. These indices differ in two aspects: df_m and N. That is, while Δ_1 does not take the model df and the sample size into account, ρ_2 incorporates both of them into the formula. Describe how different levels of df_m and N would affect ρ_2.

Once we have a dataset with a certain sample size, the basedline model's fit F_b (and χ_b^2) and degrees of freedom df_b are fixed (and so the denominator of the ρ_2 formula is fixed as well). In contrast, F_m and df_m will be determined by what kind of model is proposed. What determines the numerator of the ρ_2 formula is relative improvement of fit per degree of freedom from the fixed, worst model. Thus, for two alternative models with different df [not necessarily one nested in the other] the numerator will be smaller for the one with the smaller badness of fit per df (i.e., more effectively fitting model per df). [In this regard, Δ_1 is expected less than 1 (unless it's a just-identified model) even when the proposed model is correct, due to sampling error. However, ρ_2 is expected 1 if the proposed model is correct.]

To see the effect of different sample sizes, we consider a particular proposed model so as to fix df_m. Different sample sizes will mostly affect the denominator of the ρ_2 formula, smaller N resulting in larger ρ_2 if F_m doesn't change by N which is in practice unlikely. [More realistically, when the model is correct, as N grows F_m will become smaller and increase ρ_2, while the denominator becomes larger and decreases ρ_2. Thus, there will be a trade-off by increasing N: Increasing a small N is beneficial since the increasing numerator will be more than the increasing denominator. But such positive net outcome will become negative once N reaches some unknown optimal level. When the model is wrong, the effect of N is not obvious in that it involves the non-centrality in F_m or χ_m^2.]

6. The following diagram shows 4 possible cases for a chi-square test, where the horizontal and vertical axes represent computed p-value and power, respectively. The marked horizontal and vertical cut-off locations correspond to nominal levels of type I error and power (marked "alpha" and "p0", respectively). Thus, the region indicated by "Acceptable" is a situation where the p-value is greater than the alpha level while the test is estimated to have sufficient power, suggesting the proposed model to be accepted (or equivalently, whatever constraints that caused the tested chi-square statistic are correct). In contrast, the region indicated by "Unacceptable" is a situation where the p-value is small enough to be significant although the test is not sufficiently powerful, hence suggesting that the proposed model is not acceptable (or equivalently, the imposed constraints are not correct). There are 2 more cases left in the diagram, which are indicated to be ambiguous and are so for different reasons. Explain why these situations are ambiguous (use examples as relevant), and provide at least 2 remedies for each ambiguous situation.



In case of Ambiguous-1, computed p-value is less than alpha, which suggests the null hypothesis to be rejected (i.e., the constraints causing the test chi-square are not correct). But the test is sufficiently powerful, and so what's significant could be a tiny effect detected by an excessive power. A typical example of such a case is when sample size is "too" large. To alleviate such excessive power, we may (a) decrease the alpha level so that the rejection becomes harder. (b) [Though controversial,] we may reduce the same size to an optimal size if it's known. (c) [More problematic] remedy is to reduce the reliability of measurement so as to yield a weaker test.

The other ambiguous case arises when the computed p-value is too large to reject the null hypothesis, but the test is not powerful enough to detect a legitimate effect. The test could be weak because the sample is too small or because measures are not [internally] reliable enough. Thus, to increase the power of testing, we may (a) increase alpha so as to make it easier to reject the null hypothesis at the expense of tolerating a higher type I error rate (if the null hypothesis is indeed true). (b) If affordable, we may increase the sample size so as to reduce the sampling error. (c) Particularly when the measurement of latent variables is not reliable enough, we may add more indicators or replace the current less reliable indicators with more reliable ones.