STA 431s15 Formulas[1]

Columns of $\mathbf{A}$ *linearly dependent* means there is a vector $\mathbf{v} \neq \mathbf{0}$ with $\mathbf{Av} = \mathbf{0}$.

$\mathbf{A}$ *positive definite* means $\mathbf{v}^\top \mathbf{Av} > 0$ for all vectors $\mathbf{v} \neq \mathbf{0}$.

$E(g(X)) = \int_{-\infty}^{\infty} g(x) \, f_{_X}(x) \, dx,$ $\qquad$ or $E(g(X)) = \sum_x g(x) \, p_{_X}(x)$

$Var(X) = E[(X - \mu_{_X})^2]$ $\qquad$ $Cov(X,Y) = E[(X - \mu_{_X})(Y - \mu_{_Y})]$

$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$ $\qquad$ $r = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2 \, \sum_{i=1}^n (Y_i - \overline{Y})^2}}$

$V(\mathbf{X}) = E\left\{(\mathbf{X} - \boldsymbol{\mu}_x)(\mathbf{X} - \boldsymbol{\mu}_x)^\top\right\}$ $\qquad$ $C(\mathbf{X},\mathbf{Y}) = E\left\{(\mathbf{X} - \boldsymbol{\mu}_x)(\mathbf{Y} - \boldsymbol{\mu}_y)^\top\right\}$

$\mathbf{L} = \mathbf{A}_1\mathbf{X}_1 + \cdots + \mathbf{A}_m\mathbf{X}_m + \mathbf{b}$ $\qquad$ $\overset{c}{\mathbf{L}} = \mathbf{A}_1 \overset{c}{\mathbf{X}}_1 + \cdots + \mathbf{A}_m \overset{c}{\mathbf{X}}_m$

$V(\mathbf{L}) = E(\overset{c}{\mathbf{L}}\overset{c}{\mathbf{L}}^\top)$ $\qquad$ $C(\mathbf{L}_1, \mathbf{L}_2) = E(\overset{c}{\mathbf{L}}_1 \overset{c}{\mathbf{L}}_2^\top)$

$f(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}$ $\qquad$ $f(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}} (2\pi)^{\frac{p}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$

If $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\mathbf{AX} + \mathbf{b} \sim N_p(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$.

$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-n/2}(2\pi)^{-np/2} \exp -\frac{n}{2}\left\{tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}) + (\overline{\mathbf{x}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}} - \boldsymbol{\mu})\right\}$

$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\top$ $\qquad$ $G^2 = -2 \ln \left(\frac{\max_{\theta \in \Theta_0} L(\theta)}{\max_{\theta \in \Theta} L(\theta)}\right) = -2 \ln \left(\frac{L(\widehat{\theta}_0)}{L(\widehat{\theta})}\right)$

If $W = X + e$, $\qquad$ Reliability is $Corr(W,X)^2 = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_e^2}$

The Double Measurement Model in centered form:

$\mathbf{Y}_i = \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\epsilon}_i$ $\qquad$ $V(\mathbf{X}_i) = \boldsymbol{\Phi}_x$, $V(\boldsymbol{\epsilon}_i) = \boldsymbol{\Psi}$

$\mathbf{F}_i = \left(\begin{array}{c} \mathbf{X}_i \\ \mathbf{Y}_i \end{array}\right)$ $\qquad$ $\mathbf{X}_i$ is $p \times 1$, $\mathbf{Y}_i$ is $q \times 1$, $\mathbf{F}_i$ is $(p+q) \times 1$

$V(\mathbf{F}_i) = \boldsymbol{\Phi}$

$\mathbf{D}_{i,1} = \mathbf{F}_i + \mathbf{e}_{i,1}$ $\qquad$ $V(\mathbf{e}_{i,1}) = \boldsymbol{\Omega}_1$, $V(\mathbf{e}_{i,2}) = \boldsymbol{\Omega}_2$

$\mathbf{D}_{i,2} = \mathbf{F}_i + \mathbf{e}_{i,2}$ $\qquad$ $\mathbf{X}_i$, $\boldsymbol{\epsilon}_i$, $\mathbf{e}_{i,1}$ and $\mathbf{e}_{i,2}$ are independent.

The General Structural Equation Model in centered form:

$\mathbf{Y}_i = \boldsymbol{\beta}\mathbf{Y}_i + \boldsymbol{\Gamma}\mathbf{X}_i + \boldsymbol{\epsilon}_i$ $\qquad$ $V(\mathbf{X}_i) = \boldsymbol{\Phi}_x$ and $V(\boldsymbol{\epsilon}_i) = \boldsymbol{\Psi}$

$\mathbf{F}_i = \left(\begin{array}{c} \mathbf{X}_i \\ \mathbf{Y}_i \end{array}\right)$ $\qquad$ $V(\mathbf{F}_i) = \boldsymbol{\Phi} = \left(\begin{array}{cc} \boldsymbol{\Phi}_{11} & \boldsymbol{\Phi}_{12} \\ \boldsymbol{\Phi}_{12}^\top & \boldsymbol{\Phi}_{22} \end{array}\right)$

$\mathbf{D}_i = \boldsymbol{\Lambda}\mathbf{F}_i + \mathbf{e}_i$ $\qquad$ $V(\mathbf{e}_i) = \boldsymbol{\Omega}$

$\mathbf{X}_i$, $\boldsymbol{\epsilon}_i$ and $\mathbf{e}_i$ are independent. $\qquad$ $\mathbf{X}_i$ is $p \times 1$, $\mathbf{Y}_i$ is $q \times 1$, $\mathbf{D}_i$ is $k \times 1$.

$\boldsymbol{\Phi}_x$, $\boldsymbol{\Psi}$ and $\boldsymbol{\Omega}$ are positive definite.

[1] This formula sheet was prepared by Jerry Brunner, Department of Statistics, University of Toronto. It is licensed under a Creative Commons Attribution - ShareAlike 3.0 Unported License. Use any part of it as you like and share the result freely. The LaTeX source code is available from the course website: http://www.utstat.toronto.edu/~brunner/oldclass/431s15

Rules for Parameter Identifiability[1]

Note: All the rules listed here assume that errors are independent of exogenous variables that are not errors, and that all variables have been centered to have expected value zero.

1. *Parameter Count Rule:* If a model has more parameters than covariance structure equations, the parameter vector can be identifiable on at most a set of volume zero in the parameter space. This applies to all models.

2. *Measurement model* (Factor analysis) In these rules, latent variables that are not error terms are described as "factors."

   (a) *Double Measurement Rule*: Parameters of the double measurement model are identifiable. All factor loadings equal one. Correlated measurement errors are allowed within sets of measurements, but not between sets.

   (b) *Three-Variable Rule for Standardized Factors*: The parameters of a measurement model will be identifiable if

   - Errors are independent of one another.
   - Each observed variable is caused by only one factor.
   - The variance of each factor equals one.
   - There are at least 3 variables with non-zero loadings per factor.
   - The sign of one non-zero loading is known for each factor.

   Factors may be correlated.

   (c) *Three-Variable Rule for Unstandardized Factors*: The parameters of a measurement model will be identifiable if

   - Errors are independent of one another.
   - Each observed variable is caused by only one factor.
   - For each factor, at least one factor loading equals one.
   - There are at least 2 additional variables with non-zero loadings per factor.

   Factors may be correlated.

[1] This handout was prepared by Jerry Brunner, Department of Statistical Sciences, University of Toronto. It is licensed under a Creative Commons Attribution - ShareAlike 3.0 Unported License. Use any part of it as you like and share the result freely. The LaTeX source code is available from the course website: http://www.utstat.toronto.edu/~brunner/oldclass/431s15

   (d) *Two-Variable Rule for Standardized Factors*: A factor with just two variables may be added to a measurement model whose parameters are identifiable, and the parameters of the combined model will be identifiable provided

   - The errors for the two additional variables are independent of one another and of those already in the model.
   - The two new variables are caused only by the new factor.
   - The variance of the new factor equals one.
   - Both new factor loadings are non-zero.
   - The sign of one new loading is known.
   - The new factor has a non-zero covariance with at least one factor already in the model.

   (e) *Two-Variable Rule for Unstandardized Factors*: A factor with just two variables may be added to a measurement model whose parameters are identifiable, and the parameters of the combined model will be identifiable provided

   - The errors for the two additional variables are independent of one another and of those already in the model.
   - The two new variables are caused only by the new factor.
   - At least one new factor loading equals one.
   - The other new factor loading is non-zero.
   - The new factor has a non-zero covariance with at least one factor already in the model.

   (f) *Four-variable Two-factor Rule*: The parameters of a measurement model with two factors and four observed variables will be identifiable provided

   - All errors are independent of one another.
   - Each observed variable is influenced by only one factor.
   - Two observed variables are influenced by one factor, and two are influenced by the other.
   - All factor loadings are non-zero.
   - For each factor, either the variance of the factor equals one and the sign of one new loading is known, or at least one factor loading equals one.
   - The covariance of the two factors does not equal zero.

   (g) *Combination Rule*: Suppose that the parameters of two measurement models are identifiable by any of the rules above. The two models may be combined into a single model provided that the error terms of the first model are independent of the error terms in the second model. The additional parameters of the combined model are the covariances between the two sets of factors, and these are all identifiable.

   (h) *Cross-over Rule*: Suppose that

   - The parameters of a measurement models are identifiable, and
   - For each factor there is at least one observable variable that is caused only by that factor (with a non-zero factor loading).

   Then any number of new observable variables may be added to the model and the result is a model whose parameters are all identifiable, provided that

   - The error terms associated with the new variables are independent of the error terms in the existing model.

   Each new variable may be caused by any or all of the factors, potentially resulting in a cross-over pattern in the path diagram. The error terms associated with the new set of variables may be correlated with one another. Note that no new factors are added, just new observable variables.

   (i) *Error-Free Rule*: A vector of observable variables may be added to the factors of a measurement model whose parameters are identifiable. Suppose that

   - The new observable variables are independent of the errors in the measurement model, and
   - For each factor in the measurement model there is at least one observable variable that is caused only by that factor (with a non-zero factor loading).

   Then the parameters of a new measurement model, where some of the variables are assumed to be measured without error, are identifiable. The practical consequence is that variables assumed to be measured without error may be included in the latent component of a structural equation model, provided that the measurement model for the other variables has identifiable parameters.

3. *Latent variable model*: $\mathbf{Y}_i = \boldsymbol{\beta}\mathbf{Y}_i + \boldsymbol{\Gamma}\mathbf{X}_i + \boldsymbol{\epsilon}_i$ Here, identifiability means that the parameters involved are functions of $V(\mathbf{F}) = \boldsymbol{\Phi}$.

   (a) *Regression Rule:* If no endogenous variables cause other endogenous variables, the model parameters are identifiable.

   (b) *Acyclic Rule*: Parameters of the Latent Variable Model are identifiable if the model is acyclic (no feedback loops through straight arrows) and the following conditions hold.

   - Organize the variables that are not error terms into sets. Set 0 consists of all the exogenous variables.
   - For $j = 1, \ldots, k$, each endogenous variable in set j is influenced by at least one variable in set $j-1$, and also possibly by variables in earlier sets.
   - Error terms for the variables in a set may have non-zero covariances. All other covariances between error terms are zero.

   These conditions are satisfied if $\boldsymbol{\Psi}$ is diagonal.

4. *Two-Step Rule*: This applies to models with both a measurement component and a latent variable component, including the full two-stage structural equation model.

   1: Consider the latent variable model as a model for observed variables. Check identifiability (usually using the Regression Rule and the Acyclic Rule).

   2: Consider the measurement model as a factor analysis model, ignoring the structure of $V(\mathbf{F})$. Check identifiability.

   If both identification checks are successful, the parameters of the combined model are identifiable.