Goodness of fit assessment of item response theory models

Alberto Maydeu-Olivares
University of Barcelona

Madrid
November 12, 2014
Outline

• Introduction
• Overall goodness of fit testing
• Two examples
• Assessing goodness of approximation
• Fine-tuning the model: identifying misfitting items
• Final remarks
Introduction

• Statistical modeling involves finding a model that may have generated the data

• What is (absolute) goodness of fit testing?
  o Testing whether the fitted model may have generated the observed data

• Why is this important?
  o Because inferences drawn on poorly fitting models are misleading

• How misleading?
  o It depends on a number of factors, among them the discrepancy between the true and fitted model
Why is fit important?
Outline

- Introduction
- Overall goodness of fit testing
How to assess goodness of fit in IRT models

- IRT are models for multivariate categorical data
  - The goodness of fit of IRT models is assessed using goodness of fit statistics for categorical data, such as Pearson’s $X^2$ statistic
  - There is nothing really special about testing IRT models
Pearson’s $X^2$

- **Notation:**
  - $N =$ number of respondents
  - $n =$ number of items
  - $K =$ number of response alternatives (number of categories)
  - $C = K^n =$ number of possible response patterns ($c = 1, \ldots, C$)
  - $p_c =$ observed proportion of a response pattern
  - $\pi_c =$ probability of a response pattern
  - $\theta =$ vector of item parameters (not latent traits!)
  - $q =$ number of item parameters

\[
X^2 = N \sum_c \left( \frac{p_c - \hat{\pi}_c}{\hat{\pi}_c} \right)^2 = N (p - \hat{\pi})' \hat{D}^{-1} (p - \hat{\pi}), \tag{1}
\]

where $\hat{\pi} = \pi(\hat{\theta})$, $p - \hat{\pi}$ are the cell residuals, and $\hat{D} = \text{diag} (\hat{\pi})$

- For the MLE, $X^2 \overset{d}{\rightarrow} \chi^2_{C-q-1}$
Two small examples

- A 1PL (a Rasch model where the latent trait is treated as a random effect) applied to a 15-item test measuring mathematical proficiency in Chilean adults
  - $n = 15; \ N = 3,000; \ K = 2; \ q = 16; \ C = 2^{15} = 32,768; \ df = 32,751$

- Samejima’s graded model applied to the 7 items of the short version of the PROMIS depression scale
  - $n = 7; \ N = 767; \ K = 4; \ q = 35; \ C = 4^7 = 16,384; \ df = 16,348$

- **Problem:**
  - (asymptotic) $p$-values for $X^2$ are only accurate when the expected frequencies of every pattern are large ($N\pi > 5$ is the usual rule of thumb)
  - In large models (large $C$) that’s impossible to accomplish
  - $p$-values for $X^2$ are only accurate in models with a small number of possible response patterns (300?)

- In IRT we are interested in much larger models than $C = 300$, how do we go about testing them?
Limited information goodness of fit testing

- There have been a number of proposals, most of them involve limited information test statistics

- Key references
  - Reiser (1996, Psychometrika)
  - Cai, Maydeu-Olivares, Coffman & Thissen (British J. of Math and Stat Psych, 2006)
What is limited information goodness of fit testing?

- Multivariate categorical data admits (at least) two representations

\[
\begin{array}{c|ccc}
Y_2 & 0 & 1 & 2 \\
\hline
Y_1 & \pi_{00} & \pi_{01} & \pi_{02} \\
0 & \pi_{10} & \pi_{11} & \pi_{12} \\
1 & \hat{\pi}^{(1)}_{12} & \hat{\pi}^{(1)(2)}_{12} & \hat{\pi}^{(1)}_1 \\
\end{array}
\]

- There is a one-to-one relationship between them
- All formulae in categorical data analysis can be written using cells or moments

\[
X^2 = N (p - \hat{\pi})' \hat{D}^{-1} (p - \hat{\pi}) = N \begin{pmatrix} \hat{p}_1 - \hat{\pi}_1 \\ \hat{p}_2 - \hat{\pi}_2 \end{pmatrix}' \begin{pmatrix} \Xi_{11} & \Xi_{21} \\ \Xi_{21}' & \Xi_{22} \end{pmatrix}^{-1} \begin{pmatrix} \hat{p}_1 - \hat{\pi}_1 \\ \hat{p}_2 - \hat{\pi}_2 \end{pmatrix}, \quad (2)
\]

- Pearson’s $X^2$ is a quadratic form in all moments up to order $n$
What is limited information goodness of fit testing?

\[ X^2 = N (\mathbf{p} - \hat{\pi})' \hat{\mathbf{D}}^{-1} (\mathbf{p} - \hat{\pi}) = N \begin{pmatrix} \hat{p}_1 - \hat{\pi}_1 \\ \hat{p}_2 - \hat{\pi}_2 \end{pmatrix}' \begin{pmatrix} \Xi_{11} & \Xi_{12} \\ \Xi_{21} & \Xi_{22} \end{pmatrix}^{-1} \begin{pmatrix} \hat{p}_1 - \hat{\pi}_1 \\ \hat{p}_2 - \hat{\pi}_2 \end{pmatrix}, \quad (3) \]

• A limited information test statistic is

\[ L_1 = N (\hat{p}_1 - \hat{\pi}_1)' \Xi_1^{-1} (\hat{p}_1 - \hat{\pi}_1) \quad (4) \]

• For chi-square distributed statistics degrees of freedom = number of statistics – number of item parameters

• Degrees of freedom for Samejima’s model are positive when univariate and bivariate moments are used \((r = 2)\)
  
  o Limited information testing requires using at least univariate and bivariate moments
Moments or margins?

- **Univariate moments** = univariate probabilities that do not involve category 0

- A researcher interested in univariate testing can use
  - All univariate moments: $\hat{\pi}_1^{(1)}, \hat{\pi}_2^{(1)}, \hat{\pi}_2^{(2)}$
  - All univariate margins: $\hat{\pi}_1^{(0)}, \hat{\pi}_1^{(1)}, \hat{\pi}_2^{(0)}, \hat{\pi}_2^{(1)}, \hat{\pi}_2^{(2)}$
  - The relationship is one-to-one

- It is convenient to use moments because they are mathematically independent
  - Fewer
  - Their covariance matrix is of full rank

- Moments up to order $r$ is equivalent to use all $r$-way margins

- Bivariate information = univariate and bivariate moments
Pros and cons of limited information testing

• Suppose bivariate information is used

**PROS:**

1) Accurate (asymptotic) $p$-values can be obtained with $N = 100$
   - Asymptotic approximation to the limited information statistic requires that 4-way frequencies are large
     - For $r$-way test, $2r$-way frequencies must be large

2) If misfit is located in the bivariate margins, higher power is obtained

**CON:**

• If misfit is not located in the bivariate margins, no power
Overall limited information test statistics

- Consider the quadratic form in univariate and bivariate moments

\[ Q = N \left( \mathbf{p}_2 - \hat{\mathbf{p}}_2 \right)' \hat{\mathbf{W}} \left( \mathbf{p}_2 - \hat{\mathbf{p}}_2 \right). \tag{5} \]

- \( Q \) is asymptotically distributed as a mixture of independent chi-square variates.

- When \( \hat{\mathbf{W}} \) is chosen so that

\[ \Sigma_2 \mathbf{W} \Sigma_2 = \Sigma_2 \mathbf{W} \Sigma_2 \tag{6} \]

where \( N \Sigma_2 \) is the asymptotic covariance matrix of \( \mathbf{p}_2 - \hat{\mathbf{p}}_2 \), \( Q \) is asymptotically chi-square

- Two ways to satisfy (6)
  - Choose \( \hat{\mathbf{W}} \) such that \( \Sigma_2 \) is a generalized inverse of \( \mathbf{W} \): \( M_2 \) of M-O and Joe (2005, 2006)

\[ M_2 = N \left( \mathbf{p}_2 - \hat{\mathbf{p}}_2 \right)' \hat{\mathbf{C}}_2 \left( \mathbf{p}_2 - \hat{\mathbf{p}}_2 \right), \quad \mathbf{C}_2 = \Xi_2^{-1} - \Xi_2^{-1} \Delta_2 \left( \Delta_2' \Xi_2^{-1} \Delta_2 \right)^{-1} \Delta_2' \Xi_2^{-1}, \tag{7} \]

  - Choose \( \hat{\mathbf{W}} \) such that \( \mathbf{W} \) is a generalized inverse of \( \Sigma_2 \): \( R_2 \) of Reiser (1996)

\[ R_2 = N \left( \mathbf{p}_2 - \hat{\mathbf{p}}_2 \right)' \hat{\Sigma}_2^+ \left( \mathbf{p}_2 - \hat{\mathbf{p}}_2 \right). \tag{8} \]
Third alternative:

Use an easily computable weight matrix \( \hat{W} = \hat{\mathbf{E}}_2^{-1}, \left( \text{diag}(\hat{\mathbf{E}}_2) \right)^{-1}, \mathbf{I} \) and compute \( p \)-values by adjusting \( Q \) by its asymptotic mean and variance

- With \( \hat{W} = \left( \text{diag}(\hat{\mathbf{E}}_2) \right)^{-1} \) and binary data, this is Cai et al. (2006) and Bartholomew and Leung (2002).
- Within the context of SEM, these are the Satorra-Bentler corrections.

Degrees of freedom:

- \( M_2 \): number of statistics – number of parameters
- \( R_2 \): estimated rank of \( \hat{\mathbf{E}}_2 \) (integer)
- MV: \( df = \frac{\text{tr}(\hat{W}\hat{\Sigma}_2)}{\text{tr}(\hat{W}\hat{\Sigma}_2)^2} \) (a real number)

MV corrected statistics can be transformed so that they have the same number of degrees of freedom as \( M_2 \).
The covariance matrix of the bivariate residuals (moments)

\[ \Sigma_2 = \Xi_2 - \Delta_2 \text{Acov}(\hat{\theta})\Delta_2' \]  \hspace{1cm} (9)

• Notation:
  o \( \Xi_2 \) = (asymptotic) covariance matrix of univariate and bivariate proportions (that exclude category 0)
  o \( \Delta_2 \) = derivatives of univariate and bivariate moments with respect to item parameters
  o \( \text{Acov}(\hat{\theta}) \) = (asymptotic) covariance matrix of item parameter estimates

• For multinomial ML (marginal ML), \( \text{Acov}(\hat{\theta}) = \hat{I}^{-1} \)
Estimating the covariance matrix of the item parameter estimates

- Expected information, $\mathcal{I}_E$
  - First order derivatives for all possible patterns
  - It can only be computed for small models

- Cross-products information, $\mathcal{I}_{XP}$
  - First order derivatives for observed patterns

- Observed information, $\mathcal{I}_O$
  - First and second order derivatives for observed patterns

- Sandwich (aka robust) information, $\mathcal{I}_S = \mathcal{I}_O^{-1} \mathcal{I}_{XP} \mathcal{I}_O^{-1}$

- Crossing the different estimators of the covariance matrix of the item parameter estimates with the different choices of bivariate statistics, there are many possible choices of test statistics!
Outline

- Introduction
- Overall goodness of fit testing
- Two examples
Two numerical examples

- A 1PL applied to Chilean mathematical proficiency data
  - $n = 15; \; N = 3,000; \; K = 2; \; q = 16; \; C = 2^{15} = 32,768; \; \text{df} = 32,751$

- Samejima’s graded model applied to the short version of the PROMIS anxiety scale
  - $n = 7; \; N = 767; \; K = 4; \; q = 35; \; C = 4^7 = 16,384; \; \text{df} = 16,348$

<table>
<thead>
<tr>
<th></th>
<th>PROMIS anxiety</th>
<th></th>
<th>Chilenan mathematical proficiency</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>stat</td>
<td>value</td>
<td>df</td>
<td>RMSEA</td>
<td>value</td>
</tr>
<tr>
<td>$M_2$</td>
<td>346.34</td>
<td>182</td>
<td>.034</td>
<td>121.32</td>
</tr>
<tr>
<td>$R_2$</td>
<td>386.45</td>
<td>196</td>
<td>.036</td>
<td>133.40</td>
</tr>
<tr>
<td>$\bar{Y}_2$</td>
<td>152.17</td>
<td>55.23</td>
<td>.048</td>
<td>22.88</td>
</tr>
<tr>
<td>$\bar{L}_2$</td>
<td>346.28</td>
<td>181.63</td>
<td>.034</td>
<td>121.47</td>
</tr>
<tr>
<td>$\bar{Y}_2$</td>
<td>357.98</td>
<td>182</td>
<td>.036</td>
<td>102.64</td>
</tr>
<tr>
<td>$\bar{L}_2$</td>
<td>346.81</td>
<td>182</td>
<td>.034</td>
<td>121.43</td>
</tr>
</tbody>
</table>
Type I errors, $N = 300$

<table>
<thead>
<tr>
<th>stat</th>
<th>$\mathcal{I}$</th>
<th>$K = 2, n = 10$</th>
<th>$K = 5, n = 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\alpha = .01$</td>
<td>$\alpha = .05$</td>
</tr>
<tr>
<td>$M_2$</td>
<td>–</td>
<td>.01</td>
<td>.04</td>
</tr>
<tr>
<td>$Y_2$</td>
<td>–</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>$\bar{Y}_2$</td>
<td>$O$</td>
<td>.01</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td>$XP$</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>$L_2$</td>
<td>–</td>
<td>.01</td>
<td>.04</td>
</tr>
<tr>
<td>$\bar{L}_2$</td>
<td>$O$</td>
<td>.01</td>
<td>.04</td>
</tr>
<tr>
<td></td>
<td>$XP$</td>
<td>.01</td>
<td>.05</td>
</tr>
<tr>
<td>$R_2$</td>
<td>$O$</td>
<td>.02</td>
<td>.04</td>
</tr>
<tr>
<td>$XP$</td>
<td></td>
<td>.00</td>
<td>.01</td>
</tr>
</tbody>
</table>

$$L_2 = N(p_2 - \hat{\pi}_2)' \hat{\Sigma}_2^{-1} (p_2 - \hat{\pi}_2), \quad Y_2 = N(p_2 - \hat{\pi}_2)' \left( \text{diag}(\hat{\Sigma}_2) \right)^{-1} (p_2 - \hat{\pi}_2), \quad (10)$$
Computational issues

- For large unidimensional models, computing the goodness of fit statistic can take longer than the estimation of the item parameters and their SEs.

- The number of univariate and bivariate moments is $s = n(K-1) + \frac{n(n-1)}{2} (K-1)^2$
  - For $n = 20$ and $K = 5$, $s = 3120$.
  - For $n = 30$ and $K = 7$, $s = 15840$ (probably unfeasible).

- When data is nominal, there is no solution.

- When data is ordinal one can use instead of a quadratic form in residual univariate and bivariate moments a quadratic form in residual (multinomial) means and cross-products.

- The number of means and cross-products is $s = \frac{n(n+1)}{2}$
  - One can test much larger models.
Means and cross products for multinomial variables

\[ \kappa_i = E[Y_i] = 0 \times \Pr(Y_i = 0) + \ldots + (K_i - 1) \times \Pr(Y_i = K_i - 1), \quad (11) \]

\[ \kappa_{ij} = E[Y_i Y_j] = 0 \times 0 \times \Pr(Y_i = 0, Y_j = 0) + \ldots + (K_i - 1) \times (K_j - 1) \times \Pr(Y_i = K_i - 1, Y_j = K_j - 1). \quad (12) \]

with sample counterparts \( k_i = \bar{y}_i \) (the sample mean), and \( k_{ij} = y_i'y_j' / N \) (the sample cross-product), respectively, where \( y_i' \) denotes the observed data for item \( i \).

- For our previous 2 \( \times \) 3 example, the elements of \( \kappa \) are

\[ \begin{align*}
\kappa_1 &= E[Y_1] = 1 \Pr(Y_1 = 1) = \pi_1^{(1)} \\
\kappa_2 &= E[Y_2] = 1 \Pr(Y_2 = 1) + 2 \Pr(Y_2 = 1) = \pi_2^{(1)} + 2\pi_2^{(2)} \\
\kappa_{12} &= E[Y_1 Y_2] = 1 \times 1 \Pr(Y_1 = 1, Y_2 = 1) + 1 \times 2 \Pr(Y_1 = 1, Y_2 = 2) = \pi_1^{(1)(1)} + 2\pi_1^{(1)(2)}. \end{align*} \quad (13) \]
Testing large models for ordinal data, $M_{ord}$

- Using these statistics one can construct a goodness of fit statistic for ordinal data

\[ M_{ord} = N (k - \hat{k})' \hat{C}_{ord} (k - \hat{k}), \quad C_{ord} = \Xi^{-1}_{ord} - \Xi^{-1}_{ord} \Delta_{ord} (\Delta'_r \Xi^{-1}_{ord} \Delta_{ord})^{-1} \Delta_{ord}' \Xi^{-1}_{ord}, \quad (14) \]

- It is just like $M_2$ except that $M_{ord}$ uses as summaries statistics that further concentrate the information
  - If the misfit is in the direction where the information has been concentrated, $M_{ord}$ will be more powerful than $M_2$

- It is good practice to check the rank of $\Delta_2$ or $\Delta_{ord}$ in applications. If the rank of these matrices is unstable, the estimated statistic is not reliable
  - The model is empirically underidentified from the information used for testing

- $M_{ord}$ cannot be computed when $K$ is large and $n$ is small due to lack of df, use $M_2$

- $M_{ord} = M_2$ for binary data
Outline

- Introduction
- Overall goodness of fit testing
- Two examples
- Assessing goodness of approximation
Assessing the goodness of approximation

• Finding a well-fitting model is more difficult as the number of variables increase
  o If I generate a 1-variable dataset and I ask you what statistical model I used to generate the data, it will be easier for you than if I give you a 10-variable dataset

• In our experience, finding a well-fitting model is more difficult as the number of categories increase
  o Are our models for polytomous data worse than our models for binary data?
  o Are personality/attitudinal/patient reported data more difficult to model than educational data?
Assessing the goodness of approximation

• I would argue that finding a well-fitting model in IRT is harder than in FA
  o In IRT we model all moments (univariate, bivariate, trivariate, ..., n-variate)
  o In FA we only model univariate and bivariate moments

• If we reject the fitted model, how do we judge how close we are to the true and unknown model?
  o I.e., how do we judge the goodness of approximation of our model?
Bivariate RMSEA and full information RMSEA

- The Root Mean Squared Error of Approximation is the standard procedure for assessing goodness of approximation in SEM

- A RMSEA can be constructed based on Pearson’s $X^2$ ($\text{RMSEA}_n$), $M_2$ (RMSEA$_2$), or any other statistic with known asymptotic distribution

- For members of the RMSEA$_r$ family

$$\varepsilon_r = \sqrt{\frac{\left(\pi_r^T - \pi_r^0\right)'C_r^0\left(\pi_r^T - \pi_r^0\right)_r}{df_r}}$$

$$\hat{\varepsilon}_r = \sqrt{\frac{M_r}{df_r}}$$

- RMSEA$_2$ will generally be larger than RMSEA$_n$ because $M_2$ is generally more powerful than $X^2$ and statistics with higher power have a higher $M_r / df_r$ ratio.
Choice of cutoff values for RMSEA$_2$ and RMSEA$_n$ in binary IRT models

- A cut-off of RMSEA$_2 \leq 0.05$ separates unidimensional from two-dimensional models
- A cut-off RMSEA$_n \leq 0.03$ separates unidimensional from two-dimensional models
RMSEA\textsubscript{n}, RMSEA\textsubscript{2}, and RMSEA\textsubscript{ord}

• RMSEA\textsubscript{n} is based on $X^2$
  - There is a limit in the size of models for which $X^2$ can be computed
    - Its sampling distribution will only be well approximated in small models (just as the sampling distribution of $X^2$)
    - It is not possible to offer an overall cut-off as population values of RMSEA\textsubscript{n} depend on the number of variables and the number of categories
RMSEA\(_n\), RMSEA\(_2\), and RMSEA\(_{ord}\)

- RMSEA\(_2\) is based on \(M_2\)
  - It can be computed for larger models
  - Its sampling distribution can be well approximated in large models
  - One can obtain confidence intervals and a test of close fit
  - It is possible to offer an overall cut-off

- RMSEA\(_{ord}\) is based on \(M_{ord}\)
  - It can be computed for even larger models
  - Its sampling distribution can be well approximated in large models
  - One can obtain confidence intervals and a test of close fit BUT
  - It is not possible to offer an overall cut-off as population values of RMSEA\(_{ord}\) depend on the number of variables and the number of categories
Effect of the number of variables and categories on RMSEA$_2$ and RMSEA$_{ord}$

- RMSEA$_2$ adjusted by the number of categories is stable when misfit is small.
- RMSEA$_{ord}$ decreases as $n$ and $K$ increase.
  - It is easier to obtain a low RMSEA$_{ord}$ for large $n$ and $K$. 
Goodness of approximation when the model is large

- If the model is so large that RMSEA\(_2\) cannot be computed and we should not use RMSEA\(_{ord}\) because we cannot offer an overall cutoff, what do we do?

- The RMSEAs are ill designed to assess *pure* goodness of approximation
  - They cannot be interpreted substantively
    - They are a weighted sum of residuals divided by df

- The RMSEAs by construction mix goodness of approximation and model selection

- How can we measure pure goodness of approximation?

- We want to measure the magnitude of the misfit (effect size)
  - In substantively interpretable units
  - Such that we can offer an overall cut-off valid for any \(n\) and \(K\)
The Standardized Root Mean Squared Residual (SRMSR)

- A more appropriate name is the Squared Root Mean Residual Correlation
  - It is only valid for binary or ordinal data (just as the RMSEA\textsubscript{ord})

\[
SRMSR = \sqrt{\sum_{i<j} \frac{(r_{ij} - \hat{\rho}_{ij})^2}{n(n-1)/2}}
\]  

(16)

- \(r_{ij}\) is just the product moment (Pearson) correlation and

\[
\hat{\rho}_{ij} = \frac{\hat{\kappa}_{ij} - \hat{\kappa}_i \hat{\kappa}_j}{\sqrt{\hat{\kappa}_{ii} - \hat{\kappa}_i^2} \sqrt{\hat{\kappa}_{jj} - \hat{\kappa}_j^2}}.
\]  

(17)

where the means (\(\kappa_i\) and \(\kappa_j\)) and the cross-product \(\kappa_{ij}\) were given in (11) and (12), and \(\kappa_{ii}\) is

\[
\kappa_{ii} = E\left[Y_i^2\right] = 0^2 \times \Pr(Y_i = 0) + \ldots + (K_i - 1)^2 \times \Pr(Y_i = K_i - 1).
\]  

(18)
Relationship between RMSEA\(_2\) and SRMR
Relationship between RMSEA$_2$ adjusted by the number of categories and SRMR
Proposal of cut-off values for assessing close fit in IRT

<table>
<thead>
<tr>
<th>criterion</th>
<th>RMSEA&lt;sub&gt;2&lt;/sub&gt;</th>
<th>SRMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>adequate fit</td>
<td>0.089</td>
<td>0.05</td>
</tr>
<tr>
<td>close fit</td>
<td>0.05</td>
<td>0.027</td>
</tr>
<tr>
<td>excellent fit</td>
<td>0.05 / (K – 1)</td>
<td>0.027 / (K – 1)</td>
</tr>
</tbody>
</table>

- Criteria for adequate fit and close fit are essentially identical to those proposed by Browne and Cudeck (1993) in the context of SEM.

- Criteria for excellent and close fit are equal for binary data.

- It is possible (but cumbersome) to compute CIs and tests for the SRMR.
  - It is best used as a goodness of fit index.

- When used as a goodness of fit index, the SRMR can be computed for models of any size.
  - It only requires computing means and cross-products under the IRT model.
Goodness of fit statistics are only summary measures

• It is possible that the model fits well overall but that it fits very poorly some items
  o It is good practice to inspect item and item pairs statistics and to report the largest observed statistics

• If the model misfits we wish to locate the source of the misfit
Outline

• Introduction
• Overall goodness of fit testing
• Two examples
• Assessing goodness of approximation
• Fine-tuning the model: identifying misfitting items
Identifying misfitting items in IRT

• One could compute Pearson’s statistic for every item and pair of items.

• For a pair of items we can write

\[ X^2_{ij} = N \left( p_{ij} - \hat{\pi}_{ij} \right)' \hat{D}^{-1}_{ij} \left( p_{ij} - \hat{\pi}_{ij} \right) \]  

where \( p_{ij}, \hat{\pi}_{ij} \) are vectors of dimension \( K^2 \).

○ Unfortunately \( X^2 \) does not follow an asymptotic chi-square distribution when applied to a subtable.

• As in the case of the overall goodness of fit statistics we can use

\[ M_{ij} = N \left( p_{ij} - \hat{\pi}_{ij} \right)' \hat{C}_{ij} \left( p_{ij} - \hat{\pi}_{ij} \right), \quad \hat{C}_{ij} = \hat{D}_{ij}^{-1} - \hat{D}_{ij}^{-1} \hat{\Lambda}_{ij} \left( \hat{\Lambda}_{ij}' \hat{D}_{ij}^{-1} \hat{\Lambda}_{ij} \right)^{-1} \hat{\Lambda}_{ij}' \hat{D}_{ij}^{-1} \]  

\[ R_{ij} = N \left( p_{ij} - \hat{\pi}_{ij} \right)' \hat{\Sigma}_{ij}^+ \left( p_{ij} - \hat{\pi}_{ij} \right) \]  

or a mean and variance corrected \( X^2_{ij} \)
Identifying misfitting items in IRT

- Alternatively, we could use z statistics for the residual cross-products

\[ z_{ord} = \frac{k_{ij} - \hat{\kappa}_{ij}}{SE(k_{ij} - \hat{\kappa}_{ij})} = \frac{k_{ij} - \hat{\kappa}_{ij}}{\sqrt{\hat{\sigma}_{ord}^2 / N}}. \]  

(22)

- For all statistics (but \( M_{ij} \)) the (asymptotic) covariance matrix of the residuals for the pair of items is needed. This is, for multinomial ML,

\[ \Sigma_{ij} = D_{ij} - \pi_{ij} \pi_{ij}' - \Delta_{ij} \mathcal{I}^{-1}_{(ij)} \Delta_{ij}'. \]  

(23)

and for the z statistics we need

\[ \hat{\sigma}_{ord}^2 = \mathbf{v}' \hat{\Sigma}_{ij} \mathbf{v}, \]  

(24)

where \( \mathbf{v}' \) is the \( 1 \times K^2 \) vector

\[ \mathbf{v}' = (0 \times 0, 0 \times 1, \ldots 0 \times (K - 1), \ldots, (K - 1) \times 0, (K - 1) \times 1, \ldots, (K - 1) \times (K - 1)). \]  

(25)
Identifying misfitting items in IRT: Type I errors for a 2PL

- **$X_{ij}^c$**
- **$\overline{X}_{ij}^2$**
- **$R_i$**
- **$z_i$**

**N = 300**
- Pair (1, 2)
- Pair (1, 3)
- Pair (2, 3)

**N = 1000**
- Pair (1, 2)
- Pair (1, 3)
- Pair (2, 3)

**Horizontal axis:** Nominal $\alpha$ level
**Vertical axis:** Empirical rejection rate

- **$-$** Observed information
- **$+$** Cross-product information
- **$-$** No information
- **$-$** Matched rejection rates
Identifying misfitting items: Power for Samejima’s model

\[ M_i \]

\[ \bar{X}_i \]

\[ R_i \]

\[ z_{rd} \]

2-dim same factor

2-dim different factors

Mixture

Guessing

Horizontal axis: Magnitude of slopes

Vertical axis: Rejection rate at \( \alpha = 0.05 \)

- Observed information
- Cross-product information
- No information
- Rejection rate = 0.05
Assessing the source of misfit in the PROMIS anxiety data

<table>
<thead>
<tr>
<th>Item</th>
<th>1</th>
<th>2</th>
<th>5</th>
<th>6</th>
<th>3</th>
<th>4</th>
<th>7</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td><strong>-3.48</strong></td>
<td><strong>-3.57</strong></td>
<td>-1.68</td>
<td>-1.54</td>
<td>-0.94</td>
<td>-1.98</td>
<td>1.88</td>
</tr>
<tr>
<td>2</td>
<td>-0.029</td>
<td></td>
<td>0.50</td>
<td>-2.10</td>
<td><strong>-3.21</strong></td>
<td>1.53</td>
<td>-0.87</td>
<td>1.67</td>
</tr>
<tr>
<td>5</td>
<td>-0.021</td>
<td>0.018</td>
<td></td>
<td>-2.41</td>
<td><strong>-3.27</strong></td>
<td>-1.37</td>
<td>-1.57</td>
<td>1.81</td>
</tr>
<tr>
<td>6</td>
<td>0.002</td>
<td>-0.016</td>
<td>-0.013</td>
<td></td>
<td>-0.79</td>
<td>-2.03</td>
<td>0.24</td>
<td>1.32</td>
</tr>
<tr>
<td>3</td>
<td>0.007</td>
<td>-0.031</td>
<td>-0.023</td>
<td>0.004</td>
<td></td>
<td>-2.20</td>
<td>-0.54</td>
<td>1.65</td>
</tr>
<tr>
<td>4</td>
<td>0.001</td>
<td>0.036</td>
<td>-0.008</td>
<td>-0.021</td>
<td>-0.020</td>
<td></td>
<td>-1.26</td>
<td>1.33</td>
</tr>
<tr>
<td>7</td>
<td>-0.012</td>
<td>-0.012</td>
<td>-0.018</td>
<td>0.003</td>
<td>-0.005</td>
<td>-0.016</td>
<td></td>
<td>0.92</td>
</tr>
</tbody>
</table>

- A 90% CI for RMSEA$_2$ is (0.029; 0.040); SRMSR = 0.016.
  - The model provides a close fit to the data
- $z_{ord}$ statistics are displayed above the diagonal, residual correlations below the diagonal.
- $z_{ord}$ statistics significant at the 5% level with a Bonferroni adjustment have been boldfaced.
Outline

• Introduction
• Overall goodness of fit testing
• Two examples
• Assessing goodness of approximation
• Fine-tuning the model: identifying misfitting items
• Final remarks
Recommendations

• For overall goodness of fit use $M_2$
  o Mean and variance diagonally weighted statistic may be more powerful to detect the presence of mixtures and guessing
    ▪ It requires the estimation of the covariance matrix of the item parameter estimates using the observed information matrix

• For assessing the magnitude of the misfit use RMSR (only with binary or ordinal data)

• For assessing goodness of approximation (taking into account model parsimony) use the RMSEA$_2$
  o Cut-off values for the RMSR and RMSEA$_2$ are available
Recommendations

• Report largest statistics for item pairs
  o Residual correlations and z statistics for binary and ordinal data
  o Mean and variance adjusted $X^2$ statistics for polytomous nominal data
    ▪ They require the estimation of the covariance matrix of the item parameter estimates
Concluding remarks

- Model-data assessment in IRT can now be performed as well (if not better) than in SEM

- Most often, our models will be misspecified to some extent
  - Research is needed to investigate the practical implications of different levels of misspecification

- The theory presented here is completely general
  - Applicable to any model for multivariate discrete data (e.g., cognitive diagnostic models)

- Better IRT models (or other measurement models) are needed
  - Particularly for polytomous data