Likelihood and Fairness in Multidimensional Item Response Theory

or

What I Thought About On My Holidays

Giles Hooker and Matthew Finkelman

Cornell University, February 27, 2008

Educational Testing

Traditional model of a test:

- A test is comprised of N *items* (questions) to be responded to.
- A pool of n *subjects* (students) each provides answers to each of the questions.
- Answers are labeled *correct* (1) or *incorrect* (0); each student then has a *response vector* $\mathbf{y} = (y_1, \ldots, y_N)$.
- A *score* $S(\mathbf{y})$ is then assigned to each subject.

Classically:

$$S(\mathbf{y}) = \sum_{i=1}^{N} \alpha_i y_i$$

Item Response Theory

A more sophisticated approach:

- Each subject has an ability, given by $\theta \in \mathbb{R}$.
- The purpose of testing is to illicit information about θ.
- Model the probability of a correct response to item i by $P_i(\theta)$ (called the *response function*).
- Perform statistical inference to estimate θ.

Some reasons for doing this

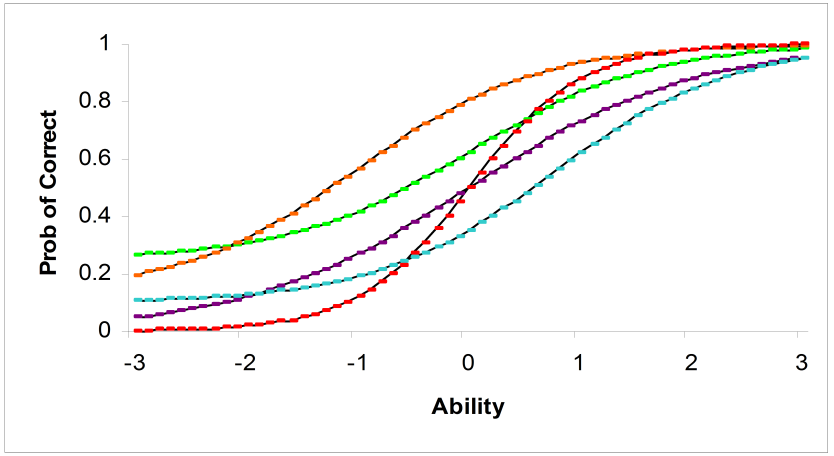
- Measures of confidence about ability estimates.
- Measures of test quality.
- Early stopping.
- Adaptive testing procedures.

Models and Methods

Item response models:

2-parameter logistic model: with $a_{i1} > 0$.

$$P_i(\theta) = \frac{1}{1 + \exp\left(- a_{i1}(\theta - a_{i0})\right)}$$



subset of ogive models:

$$P_i(\theta) = F\left(a_{i1}(\theta - a_{i0})\right)$$

Estimates for θ (ess. GLM with fixed intercept):

- Maximum Likelihood: $\hat{\theta}_{MLE} = \text{argmax}_\theta \, l^N(\theta; \mathbf{y})$
- Maximum *A-Posteriori*: $\hat{\theta}_{MAP} = \text{argmax}_\theta \, f^N(\theta|\mathbf{y})$
- Expected *A-Posteriori*: $\hat{\theta}_{EAP} = \int \theta f^N(\theta|\mathbf{y}) d\theta$

Multidimensional IRT

There is more than one type of ability (eg. language vs analysis vs computation).

Describe ability by $\boldsymbol{\theta} \in \mathbb{R}^d$. MIRT models become

- Compensatory:

$$P_i(\boldsymbol{\theta}) = F\left(a_{i0} + \boldsymbol{\theta}^T \mathbf{a}_{i1}\right)$$

- Non-compensatory:

$$P_i(\boldsymbol{\theta}) = \prod_{j=1}^{d} F(a_{ij0} + \theta_j a_{ij1})$$

Each $a_{ij1} \geq 0$ ensures that P_i is *increasing* for each θ_j.

Choosing F to be log-concave makes optimization easy.

Aside: Identifiability

In logistic regression, models unidentifiable if

- the design matrix is singular
- answer sequence makes classes linearly separable

In MIRT, intercepts are fixed $\Rightarrow$ separability is no problem.

However $a_{ij1} > 0$, $\Rightarrow$ unidentifiable for homogenous response sequences

Compensatory Models: if $\mathbf{y} = 1$ only confidence interval for θ_1 is $(-\infty, \infty]$, if $\mathbf{y} = 0$ it is $[-\infty, \infty)$.

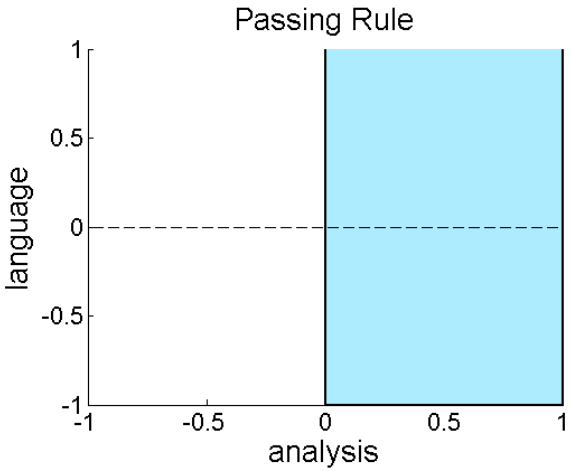
Non-compensatory Models: if $\mathbf{y} = 0$, only confidence interval for θ_1 is $[-\infty, \infty]$.

Only applies to likelihood-based estimates.
But even frequentists can use Bayesian methods in IRT!

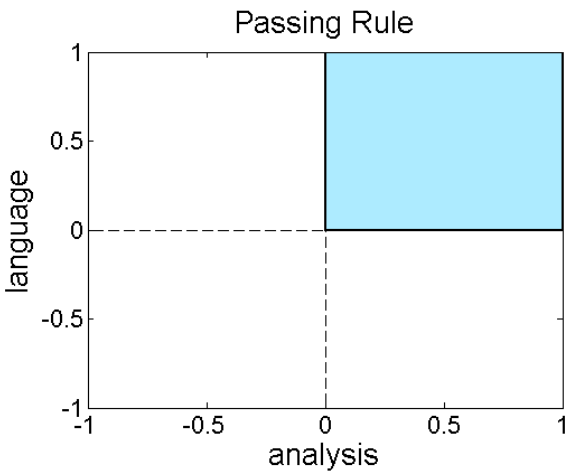
Possible Scoring Rules

Base scores on estimated $\hat{\boldsymbol{\theta}}$

Nuisance parameters:
$S(\mathbf{y}) = \hat{\theta}_1$.
Want to assess analysis skills, but control for language (Veldkamp and van der Linden 2002).



Multiple Hurdle:
$S(\mathbf{y}) = I(\hat{\theta}_1 > T_1, \hat{\theta}_2 > T_2)$.
Select only candidates with both high analysis and language skills (Segal, 2000).



A Story

A Story

- Jane and Jill are high school friends, taking a college entrance exam.

A Story

- Jane and Jill are high school friends, taking a college entrance exam.
- Comparing notes afterwards, they gave the same answers up until the final question.

A Story

- Jane and Jill are high school friends, taking a college entrance exam.
- Comparing notes afterwards, they gave the same answers up until the final question.
- Jill answered the last question correctly, Jane answered incorrectly.

A Story

- Jane and Jill are high school friends, taking a college entrance exam.
- Comparing notes afterwards, they gave the same answers up until the final question.
- Jill answered the last question correctly, Jane answered incorrectly.
- When the results come back, Jane passed but Jill failed!

A Story

- Jane and Jill are high school friends, taking a college entrance exam.
- Comparing notes afterwards, they gave the same answers up until the final question.
- Jill answered the last question correctly, Jane answered incorrectly.
- When the results come back, Jane passed but Jill failed!

The college's position:

- These are legitimate results.
- Exam questions required both language and analysis (MIRT Analysis).
- Candidates only admitted if they excel at both.
- Implies a multiple hurdle model, based on MLE estimates of $\boldsymbol{\theta}$.

An Explanation

An Explanation

- Jill and Jane got some questions right and some wrong.

An Explanation

- Jill and Jane got some questions right and some wrong.
- Final question heavily emphasized analysis.

An Explanation

- Jill and Jane got some questions right and some wrong.
- Final question heavily emphasized analysis.
- Jill answered correctly $\Rightarrow$ strong analysis skills.

An Explanation

- Jill and Jane got some questions right and some wrong.
- Final question heavily emphasized analysis.
- Jill answered correctly $\Rightarrow$ strong analysis skills.
- But she answered others incorrectly $\Rightarrow$ poor language skills.

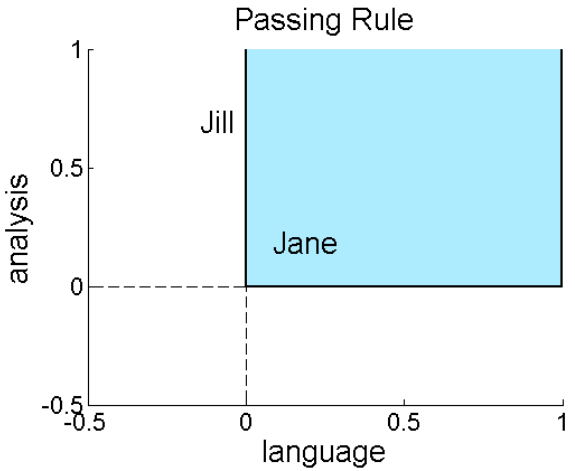
An Explanation

- Jill and Jane got some questions right and some wrong.
- Final question heavily emphasized analysis.
- Jill answered correctly $\Rightarrow$ strong analysis skills.
- But she answered others incorrectly $\Rightarrow$ poor language skills.
- Jane has fewer analysis skills; must have relied on language for correct answers.

An Explanation

- Jill and Jane got some questions right and some wrong.
- Final question heavily emphasized analysis.
- Jill answered correctly $\Rightarrow$ strong analysis skills.
- But she answered others incorrectly $\Rightarrow$ poor language skills.

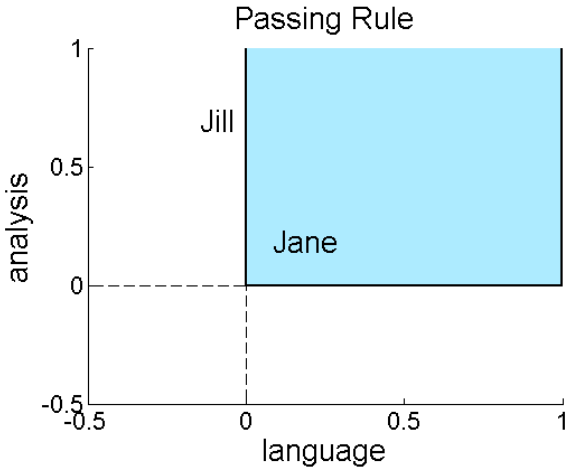
- Jane has fewer analysis skills; must have relied on language for correct answers.



An Explanation

- Jill and Jane got some questions right and some wrong.
- Final question heavily emphasized analysis.
- Jill answered correctly $\Rightarrow$ strong analysis skills.
- But she answered others incorrectly $\Rightarrow$ poor language skills.

- Jane has fewer analysis skills; must have relied on language for correct answers.



Jill's lawyer: *Is it really reasonable to put students in the position of second-guessing when their best answer might be harmful to them?*

Empirical Study

Describe $S(\mathbf{y})$ as *non-monotone* if there are answer sequences $\mathbf{y}_0 \prec \mathbf{y}_1$ and $S(\mathbf{y}_0) > S(\mathbf{y}_1)$.

In a real world data set:

- 67-question English test of 7500 grade 5 students.
- Responses modeled as measuring listening and reading/writing.
- Reading/writing skills most emphasized in test – treat this as "ability of interest".
- Test parameters estimated from 5000 students, 2500 remain to investigate non-monotonicity.
- $\hat{\theta}_1(\mathbf{y})$ taken to be EAP with standard bivariate normal prior.
- From 2500 students, 4 pairs match the "Jane and Jill" story.

Hypotheticals:

If only I had answered more questions wrong, I would have passed!

Greedy algorithm:
For each student

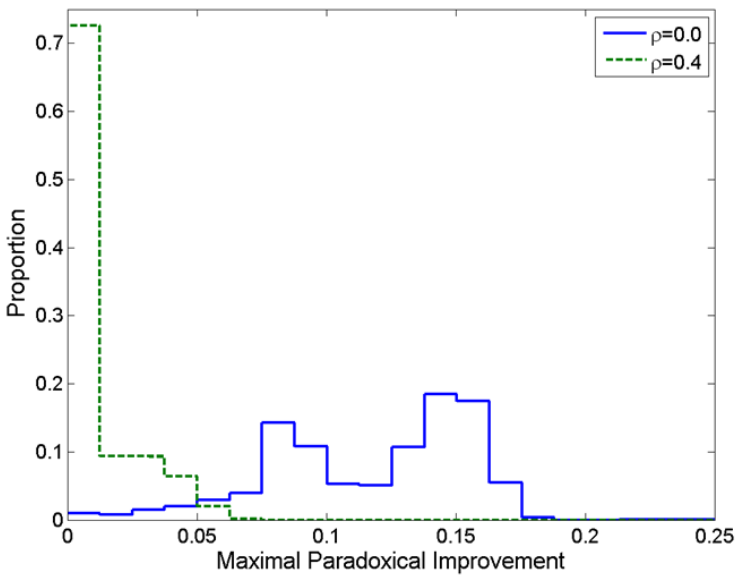
- Try changing each correct answer to incorrect.
- If at least one change leads to $\hat{\theta}_1$ increasing, choose the item that lead to largest increase.
- Repeat until $\hat{\theta}_1$ cannot be increased further
- Keep track of cumulative increase.

Paradoxical increase found in range $[0,\ 0.251]$ corresponding to $[0,\ 5.2]\%$ of range of estimated abilities.

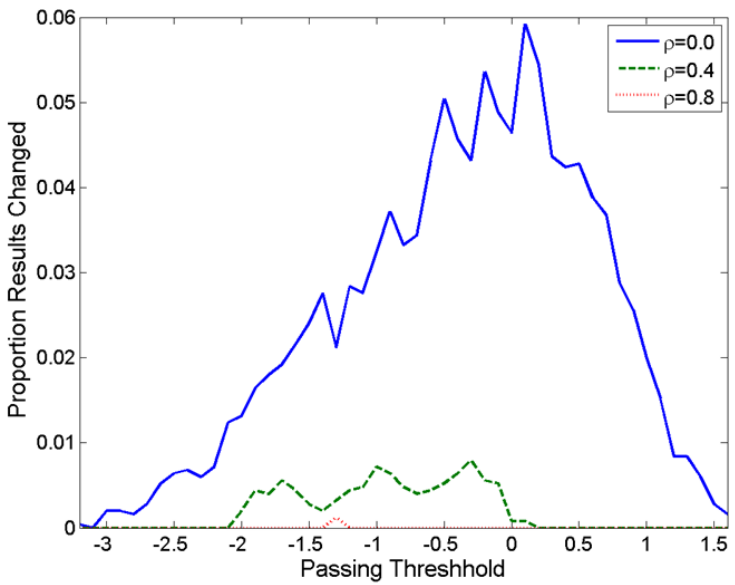
Graphically

What proportion of students would move from fail ($\hat{\theta}_1 < \theta_T$) to pass ($\hat{\theta}_1 \geq \theta_T$) by getting more questions wrong?

Paradoxical Increase



Paradoxical Classification



Mathematical Analysis

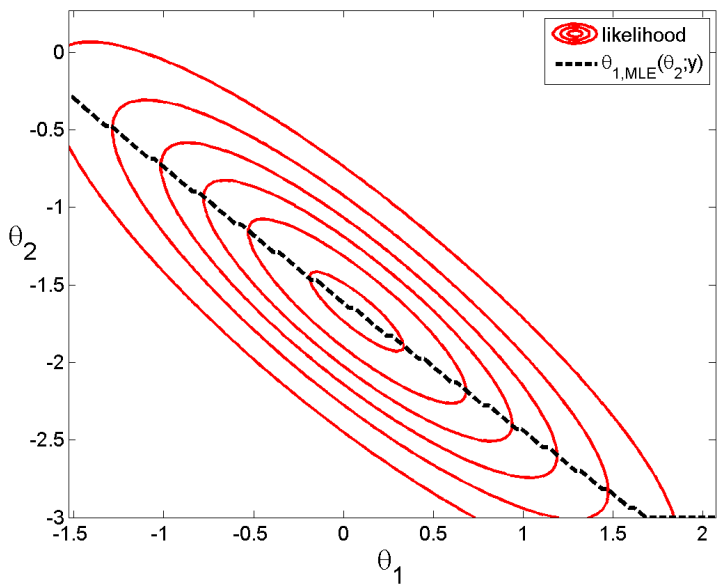
- Mathematically, how does non-monotonicity occur?
- How do we stop it?

  Possible variables:

  - MIRT models for response functions
  - statistical estimates
  - test design

An Intuition

- Contours of response functions always decreasing $\Rightarrow$ contours of likelihood look like negatively-oriented ellipses.
- *MLE* for θ_1 *conditional* on θ_2 is decreasing



If $\hat{\theta}_{2,MLE}$ increases without changing $\hat{\theta}_{1,MLE}(\theta_2)$, $\hat{\theta}_{1,MLE}$ must decrease.

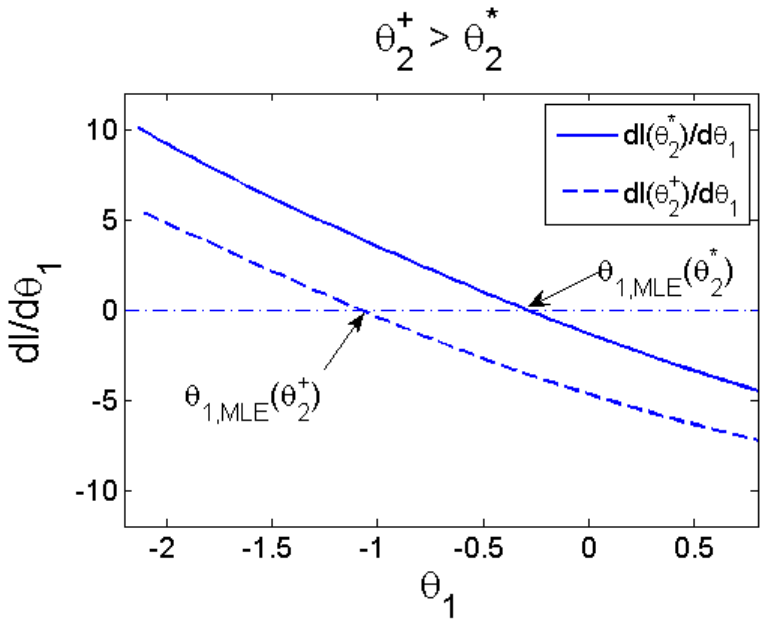
Achieve this with an item that is not affected by θ_1.

A Simple Observation

For log-concave ogives

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j} l(\boldsymbol{\theta};\mathbf{y}) = \frac{\partial^2}{\partial\theta_i\partial\theta_j} \sum_{i=1}^{N} y_k \log P_i(\boldsymbol{\theta}) + (1-y_i)\log(1-P_i(\boldsymbol{\theta})) < 0$$

for both compensatory and non-compensatory models.



In particular

$$\frac{\partial}{\partial\theta_1} l(\theta_1, \boldsymbol{\theta}^+_{-1};\mathbf{y}) < \frac{\partial}{\partial\theta_1} l(\theta_1, \boldsymbol{\theta}^*_{-1};\mathbf{y})$$

whenever $\boldsymbol{\theta}^+_{-1} > \boldsymbol{\theta}^*_{-1}$.

Formally

We restrict to two dimensions:

Theorem

For every test with bivariate compensatory or non-compensatory items based on log-concave ogives and for every response sequence $\mathbf{y}$ *such that* $\hat{\boldsymbol{\theta}}_{MLE}(\mathbf{y})$ *is well defined, a further item may be added to the test such that*

$$\hat{\theta}_{1,MLE}(\mathbf{y}, 1) < \hat{\theta}_{1,MLE}(\mathbf{y}, 0)$$

That further item can be defined by being any item that does not depend on θ_1.

Bounds on θ_1

Basing a threshhold on MLE values is silly, conduct a test!

Define a *profile likelihood lower bound* by:

$$\hat{\theta}_{1,PLB}(\mathbf{y}) = \min\left\{\theta_1 : \max_{\theta_2} l(\theta_1, \theta_2; \mathbf{y}) > l(\boldsymbol{\theta}_{MLE}; y) - \chi^2_1(\alpha)\right\}$$

Theorem

For every test with bivariate compensatory or non-compensatory items based on log-concave ogives and for every response sequence $\mathbf{y}$ *such that* $\hat{\boldsymbol{\theta}}_{MLE}(\mathbf{y})$ *is well defined, a further item may be added to the test such that*

$$\hat{\theta}_{1,PLB}(\mathbf{y}, 1) < \hat{\theta}_{1,PLB}(\mathbf{y}, 0)$$

Bayesian Methods

Consider an independence prior on $\boldsymbol{\theta}$:

$$\mu(\theta_1, \theta_2) = \mu_1(\theta_1)\mu_2(\theta_2)$$

with μ_1 and μ_2 log-concave.

The posterior $f(\boldsymbol{\theta}|\mathbf{y})$ has $\partial^2 \log f(\boldsymbol{\theta}|\mathbf{y})/\partial\theta_i\partial\theta_j < 0$

Theorem

For every test with bivariate compensatory or non-compensatory items based on log-concave ogives and a log-concave independence prior, for every response sequence $\mathbf{y}$ *a further item may be added to the test such that*

$$\hat{\theta}_{1,MAP}(\mathbf{y}, 1) < \hat{\theta}_{1,MAP}(\mathbf{y}, 0)$$

Marginal Inference

For univariate densities $f_1(x)$, $f_2(x)$ with cumulative distribution functions $F_1(x)$, $F_2(x)$,

$$\frac{d \log f_1}{dx} \prec \frac{d \log f_2}{dx} \Rightarrow F_1(x) \succ F_2(x)$$

As θ_2 increases, $\theta_1 | \theta_2, \mathbf{y}$ becomes stochastically smaller.

Theorem

For every test with bivariate compensatory or non-compensatory items based on log-concave ogives and a log-concave independence prior, for every response sequence $\mathbf{y}$ *a further item may be added to the test such that*

$$F(\theta_1 | (\mathbf{y}, 1)) > F(\theta_1 | (\mathbf{y}, 0))$$

Compensatory Models

Can we restrict the items in a test so that non-monotone results can't occur?

In compensatory models we have $P_i(\boldsymbol{\theta}) = F_i(\boldsymbol{\theta}^T \mathbf{a}_i)$.

Let $[A]_i = \mathbf{a}_i^T$ be the "design matrix" for this test.

Theorem

For any test employing compensatory response models, let $\mathbf{y}$ *be any response sequence such that* $\hat{\boldsymbol{\theta}}_{MLE}(\mathbf{y})$ *exists and is unique. Let* $F_N(\boldsymbol{\theta}^T \mathbf{b})$ *be a further item in the test. A sufficient condition for* $\hat{\theta}_{1,MLE}(\mathbf{y}, 1) < \hat{\theta}_{1,MLE}(\mathbf{y}, 0)$ *is*

$$e_1^T \left(A^T W A \right)^{-1} \mathbf{b} > 0$$

for all diagonal matrices $W \prec 0$.

In Bivariate Models

What does that condition mean?

$$e_1^T \left(A^T W A\right)^{-1} b = C\left(\sum w_i a_{i2}^2 b_1 - \sum w_i a_{i1} a_{i2} b_2\right)$$
$$= C \sum w_i a_{i2} \left(a_{i2} b_1 - a_{i1} b_2\right)$$

Corollary

Let a test comprised of bivariate compensatory items be ordered such that

$$\frac{a_{N2}}{a_{N1}} > \frac{a_{i2}}{a_{i1}}, \; i = 1, \ldots, N-1.$$

For any response pattern $\mathbf{y} = (y_1, \ldots, y_{N-1})$,
$\hat{\theta}_{1,MLE}(\mathbf{y}, 0) > \hat{\theta}_{1,MLE}(\mathbf{y}, 1)$ *so long as these are well defined.*

Some Consequences

- $\hat{\theta}_{1,MLE}(\mathbf{y})$ exhibits non-monotonicity for *every* bivariate test using compensatory models.
- For almost all answer sequences, $\hat{\theta}_{1,MLE}(\mathbf{y})$ can be made to increase by changing an answer from "correct" to "incorrect", or to decrease by changing an answer from "incorrect" to "correct".
- There is a simple rule for which answer needs to be changed.

Some Consequences

- $\hat{\theta}_{1,MLE}(\mathbf{y})$ exhibits non-monotonicity for *every* bivariate test using compensatory models.
- For almost all answer sequences, $\hat{\theta}_{1,MLE}(\mathbf{y})$ can be made to increase by changing an answer from "correct" to "incorrect", or to decrease by changing an answer from "incorrect" to "correct".
- There is a simple rule for which answer needs to be changed.

A student facing a test of their ability in mathematics which also includes language processing as a nuisance dimension may be well advised to find the question that appears to place least emphasis on mathematics and make certain to answer it incorrectly.

Some Consequences

- $\hat{\theta}_{1,MLE}(\mathbf{y})$ exhibits non-monotonicity for *every* bivariate test using compensatory models.
- For almost all answer sequences, $\hat{\theta}_{1,MLE}(\mathbf{y})$ can be made to increase by changing an answer from "correct" to "incorrect", or to decrease by changing an answer from "incorrect" to "correct".
- There is a simple rule for which answer needs to be changed.

A student facing a test of their ability in mathematics which also includes language processing as a nuisance dimension may be well advised to find the question that appears to place least emphasis on mathematics and make certain to answer it incorrectly.

But only if they can reliably identify this question!

The Story Continues

The Story Continues

During the trial, it was discovered that Jane's uncle worked for the agency that designed the exam.

The Story Continues

During the trial, it was discovered that Jane's uncle worked for the agency that designed the exam.

He denied any providing any improper assistance to his niece.

The Story Continues

During the trial, it was discovered that Jane's uncle worked for the agency that designed the exam.

He denied any providing any improper assistance to his niece.

I knew she was concerned about the language component, but I only told her not to get worried at the start of the test because we saved the toughest analysis question 'till last.

The Story Continues

During the trial, it was discovered that Jane's uncle worked for the agency that designed the exam.

He denied any providing any improper assistance to his niece.

I knew she was concerned about the language component, but I only told her not to get worried at the start of the test because we saved the toughest analysis question 'till last.

Nobody found the copy of Hooker and Finkelman (2008) in her computer's recycle bin.

The Analysis Continues

For Bayesian methods, suppose that

$$\frac{\partial \log \mu}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \succ K_-$$

Theorem

For any test employing compensatory response models, let $\mathbf{y}$ *be any response sequence. A sufficient condition for* $\hat{\theta}_{1,MAP}(\mathbf{y}, 1) < \hat{\theta}_{1,MAP}(\mathbf{y}, 0)$ *is*

$$e_1^T \left(A^T W A + K \right)^{-1} b > 0$$

for all diagonal matrices $W \prec 0$ *and* $K_- \prec K \prec 0$.

Bivariate Interpretation

Suppose μ bivariate normal with variances σ_1^2, σ_2^2 and correlation ρ.

Sufficient condition for $\hat{\theta}_{1,MAP}(\mathbf{y}, 1) < \hat{\theta}_{1,MAP}(\mathbf{y}, 0)$ is

$$\frac{b_2}{b_1} > \frac{\sum_{i=1}^{N} w_i a_{i2}^2 + \frac{1}{\sigma_2^2(1-\rho^2)}}{\sum_{1=1}^{N} w_i a_{i1} a_{i2} - \frac{\rho}{(1-\rho^2)\sigma_1\sigma_2}}, \ \forall w_i > 0$$

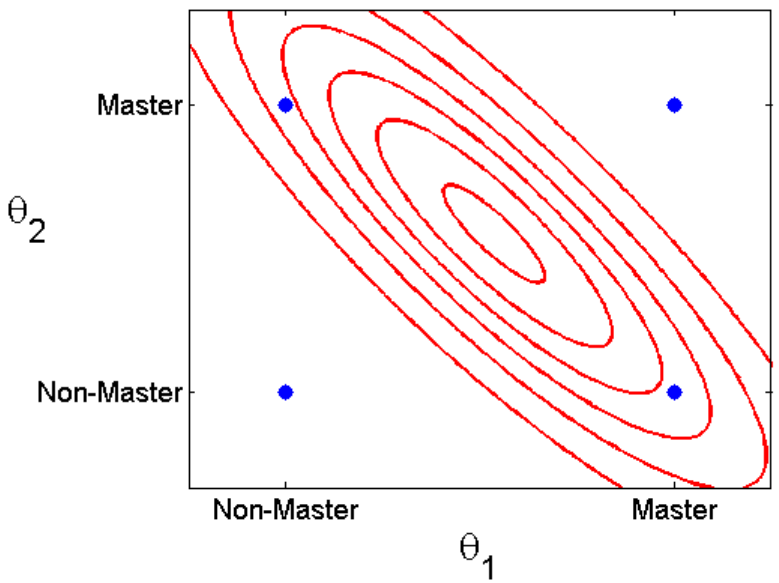
Same condition holds for EAP.

- Even with $\rho = 0$ the condition is more stringent.
- Making σ_2^2 small $\Leftrightarrow$ constraining the items to only vary with θ_1.
- Setting $\rho = 0.8$ did not entirely remove non-monotonicity in real data.

Extensions

Discrete Ability Spaces

Ability variables coded $master/non-master$ in each dimension.
Can construct numerical examples of non-monotonicity.
More difficult to analyze explicitly.



Extensions

Non Log-Concave Ogives

Common to use "guessing paramters" esp. for multiple-choice items

$$\tilde{P}_i(\boldsymbol{\theta}) = c_i + (1 - c_i)P_i(\boldsymbol{\theta})$$

- Second derivatives of the likelihood will not always be negative.
- Only need this condition near MLE or MAP.
- In real data, with independence prior, *all* subjects could have their EAP changed paradoxically.

Extensions

Three or more ability dimensions

For MLE's, if final question does not depend on θ_1,

$$\hat{\boldsymbol{\theta}}_{-1,MLE}(\mathbf{y}, 0) \prec \hat{\boldsymbol{\theta}}_{-1,MLE}(\mathbf{y}, 1) \Rightarrow \hat{\theta}_{1,MLE}(\mathbf{y}, 0) > \hat{\theta}_{1,MLE}(\mathbf{y}, 1).$$

But the left hand side is not guaranteed.

If LHS does not hold, $\hat{\boldsymbol{\theta}}_{MLE}$ behaves non-monotonically in some other dimension.

Problem for multiple-hurdle model; less clear for nuisance dimensions.

Constrained Estimates

One solution:

- *Jointly* estimate $\boldsymbol{\theta}_j$ for each subject j.
- Impose constraints that $\boldsymbol{\theta}_j \prec \boldsymbol{\theta}_i$ if $\mathbf{y}_j \prec \mathbf{y}_i$.
- Feasible (2min for real-world data using IPOPT routines and calculating $\boldsymbol{\theta}_{MAP}$).
- Moderate distortion: about the same as difference between MAP and EAP.
- Consistent: asymptotically $\mathbf{y}_j \prec \mathbf{y}_i$ doesn't happen.

But does not address "what if" questions, cannot add more subjects.

Estimate for *all possible* responses under constraints? Infeasible. (Inconsistent?)

Conclusions

Jane and Jill both took a test
To get into a college;
Jill beat Jane, but also failed
Yet Jane might pass, we allege.

Non-monotonicity

- has clear implications for test fairness.
- is sufficiently common in real-world data to be concerning.
- is theoretically inescapable in compensatory bivariate ability models using frequentist inference.
- is likely to occur in extensions of the models studied here.
- is unlikely to be eliminated by alternative statistical techniques without compromising properties such as consistency.

Perhaps we should re-think the goals of educational testing, at least when the student cares about the result.