Multidimensional Mastery Testing with CAT

Steven W. Nydick
April 6, 2014

Multidimensional Mastery Testing (MMT)

How do we conceptualize mastery testing?

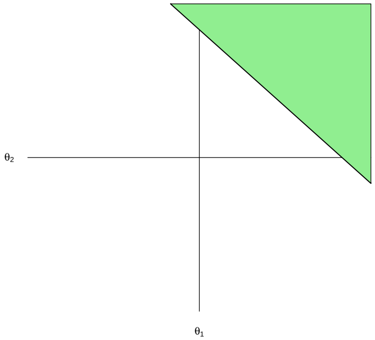
- Two regions of multidimensional space
  - $\mathrm{H}_0 : \boldsymbol{\theta}_i \in \boldsymbol{\Theta}_n$
  - $\mathrm{H}_1 : \boldsymbol{\theta}_i \in \boldsymbol{\Theta}_m$
- Classification bound *function* separating regions
  - Satisfies: $g(\boldsymbol{\theta}) = 0$.
  - $g(\boldsymbol{\theta}) = 0$ is a curve in two dimensions.
  - $g(\boldsymbol{\theta}) = 0$ is a surface in three dimensions.

Common classification bound functions:

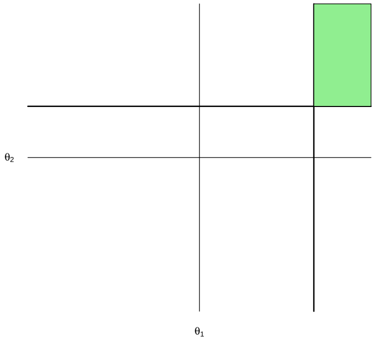
1. Compensatory/linear task
2. Noncompensatory/piecewise task

Multidimensional Mastery Testing (MMT)

Compensatory Task

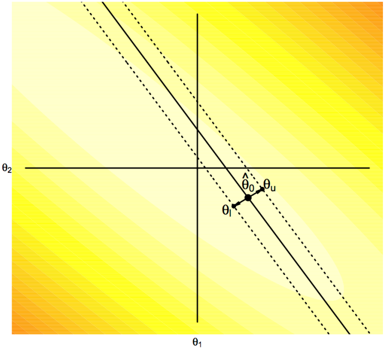


Noncompensatory Task

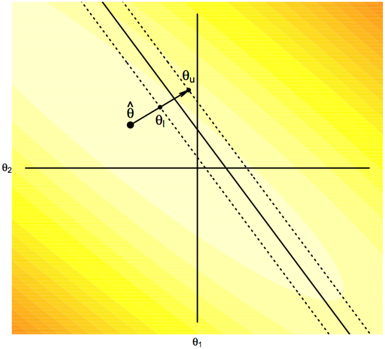


The Multidimensional SPRT: Graphically

Constrained SPRT

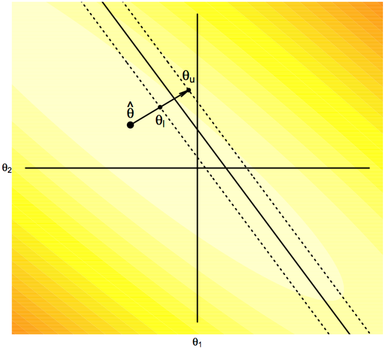


Projected SPRT

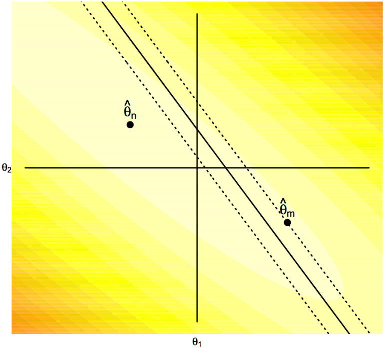


The Multidimensional GLR: Graphically

Projected SPRT



Multidimensional GLR



The Bayesian Credible Region Approach

An alternative option to point comparisons:

- Let α and β be Type I/II error rates.
- Find the posterior probability of being a master.

$$\Pr(m|\mathbf{y}_{i,j_{\text{tmp}}}) = \int_{\boldsymbol{\Theta}_m} \pi(\theta|\mathbf{y}_{i,j_{\text{tmp}}}) d\theta \tag{1}$$

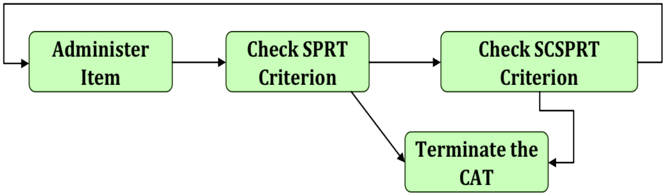
- If $\Pr(m|\mathbf{y}_{i,j_{\text{tmp}}}) > 1 - \beta$, classify as a master.
- If $\Pr(m|\mathbf{y}_{i,j_{\text{tmp}}}) < \alpha$, classify as a non-master.

Why does this work?

> *"In sequential scenarios, there is no need to 'spend α' for looks at the data" when using Bayesian tests because "posterior probabilities are not affected by the reason for stopping experimentation" (Berger, 2012).*

The Stocastically Curtailed SPRT

A primer on the SCSPRT:



Curtailed methods generalize directly to MIRT:

- The likelihood is a scalar function of $\boldsymbol{\theta}$.
- The expectation is a function of $L(\boldsymbol{\theta}|\mathbf{y}_{i,j})$ and $p_j(\boldsymbol{\theta})$.
- The variance is a function of $L(\boldsymbol{\theta}|\mathbf{y}_{i,j})$ and $p_j(\boldsymbol{\theta})$.

Curtailment depends on future items.

Fisher Information Selection Algorithms

Some commonly used FI item selection algorithms:

1. Minimize the volume of the confidence ellipsoid (D).
   - Determinant of inverse test information
2. Minimize the average variance across all dimensions (T).
   - Trace of inverse test information
3. Minimize the variance in a particular direction (L).
   - Quadratic form of $\boldsymbol{\lambda}^T \left(\sum \mathcal{I}\right)^{-1} \boldsymbol{\lambda}$

KL and the Expected Likelihood Ratio

How does KL divergence apply to mastery tests?

Typical method? Compare θ_u to θ_l:

$$\mathrm{KL}_j(\theta_u|\theta_l) = \mathbb{E}_{\theta_u}\left[\log\left[\mathrm{LR}(\theta_u, \theta_l|Y_{ij})\right]\right] \tag{2}$$

Equation (2) assumes a priori mastery.

New method? Take expectation w.r.t. current ability estimate:

$$\mathrm{ELR}_j(\hat{\theta}_i) = \mathbb{E}_{\hat{\theta}_i}\left[\log\left[\mathrm{LR}(\theta_u, \theta_l|Y_{ij})\right]\right] \tag{3}$$

Latter method shows promise in unidimensional simulations.

Application to Mastery Testing

A straightforward procedure to choose next MCMT item:

1. Let Θ_0 be the set of points on the classification bound.
2. Pick a function to maximize/minimize given any $\boldsymbol{\theta}_0 \in \Theta_0$.
   - L-Method FI with $\boldsymbol{\lambda}$ normal to the classification bound at $\boldsymbol{\theta}_0$
   - KL-divergence comparing $\boldsymbol{\theta}_u = \boldsymbol{\theta}_0 + \delta\boldsymbol{\theta}_\delta$ to $\boldsymbol{\theta}_l = \boldsymbol{\theta}_0 - \delta\boldsymbol{\theta}_\delta$
   - ELR comparing $\boldsymbol{\theta}_u$ to $\boldsymbol{\theta}_l$
3. Choose some appropriate weight function, w_{ij}.
   - $w_{ij} = \pi(\boldsymbol{\theta})$
   - $w_{ij} = \pi(\boldsymbol{\theta}|\mathbf{y}_{i,j})$
   - $w_{ij} = L(\boldsymbol{\theta}|\mathbf{y}_{i,j})$
4. Take the line/surface integral of the max/min function weighted by w_{ij} along the line/surface defined by Θ_0.

Item Bank and IRT Model

❶ Size of Item Bank

- $J = 900$

❷ Models

- C-MIRT

❸ Dimensions

- $K = 2$

❹ Item Banks

- Within-Item Multidimensionality
- Between-Item Multidimensionality

Latent Trait and Classification Bound

❶ Latent Trait Distribution

- $K = 2$
- $\boldsymbol{\theta} \sim N\left(\mathbf{0}, \left[\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right]\right)$ where $\rho \in \{.00, .33, .67\}$.

❷ Classification Bound Function

- Compensatory: $\theta_2 + \theta_1 = 0$
- Noncompensatory: $\theta_1 > 0$ and $\theta_2 > 0$

Item Selection Algorithms

❶ D-Method FI at $\hat{\boldsymbol{\theta}}_0$

❷ L-Method FI at $\hat{\boldsymbol{\theta}}_0$

❸ L-Method ELR at $\hat{\boldsymbol{\theta}}_0$

❹ L-Method KL at $\hat{\boldsymbol{\theta}}_0$

- λ normal to the classification bound function

❺ Surface KL

- λ normal to the classification bound function
- $w_{ij} = \pi(\boldsymbol{\theta}|\mathbf{y}_{i,j})$

Stopping Rules

1. P-SPRT, C-SPRT, and M-GLR
   - $\delta \in \{.15, .25\}$
   - $\alpha = \beta = .1$
2. M-SCSPRT
   - C-SPRT as the base SPRT method
   - $\delta \in \{.15, .25\}$
   - $\alpha = \beta = .1$
   - $\epsilon_1 = \epsilon_2 = .05$
3. BCR
   - $\alpha = \beta \in \{.05, .10\}$.

Procedure and Conditions Table

❶ Test Length

- $J = 4$ items randomly selected
- Minimum: $J = 10$
- Maximum: $J = 100$

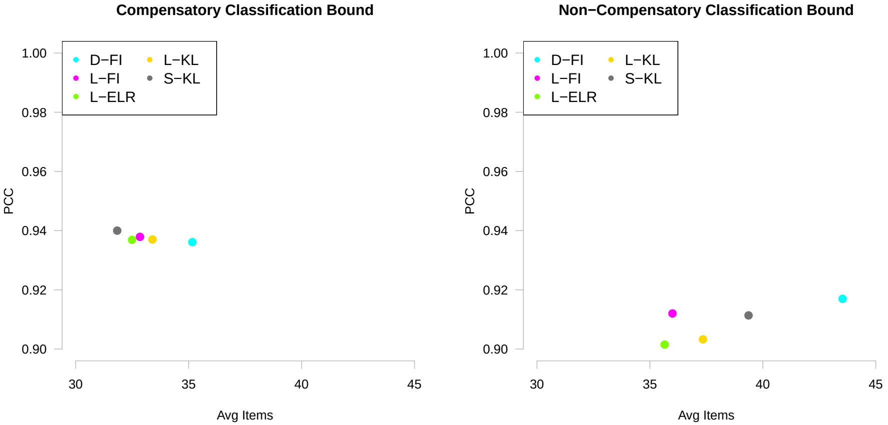
❷ Ability Estimation

- MLE bounded within $[-4, 4] \times [-4, 4]$

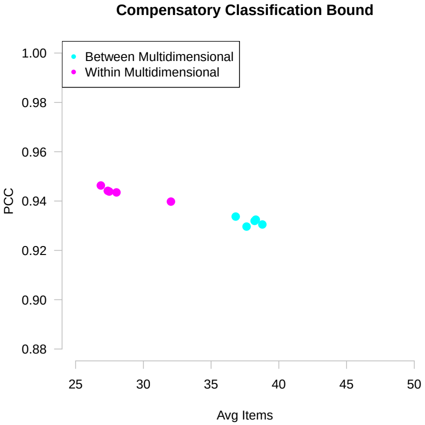
Conditions Table

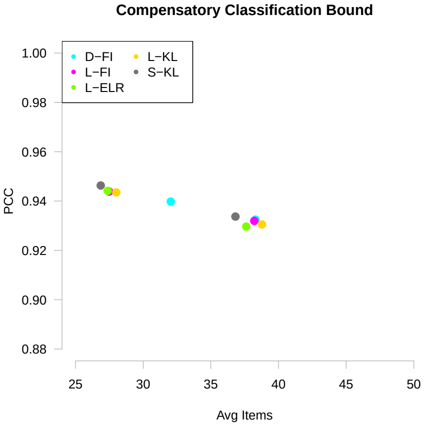
ρ	3 $(.00, .33, .67)$
Bound Functions	2 (Comp, Non-Comp)
Item Banks	2 (B/w, W/in Dimensions)
Item Selection	5 (D-FI, L-FI, L-ELR, L-KL, S-KL)
Stopping Rules	10 (5×2)
Overall	600

Length and Accuracy: Item Selection

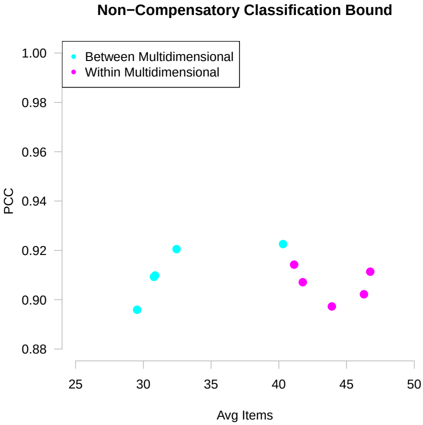


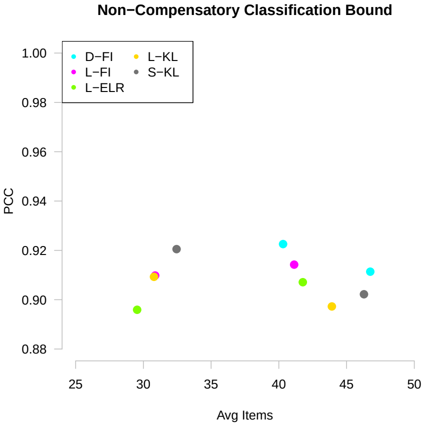
Length and Accuracy: Bank by Select



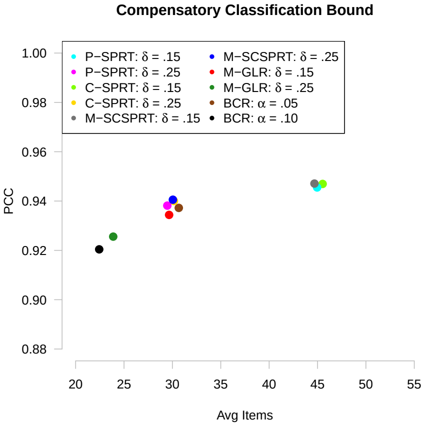


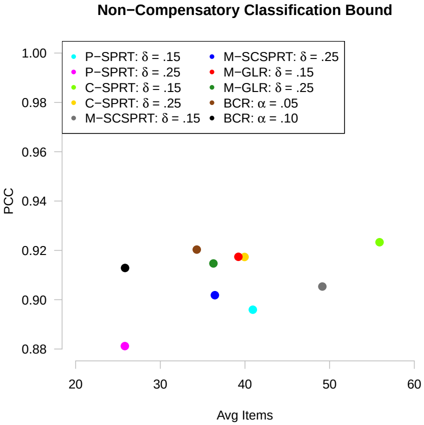
Length and Accuracy: Bank by Select



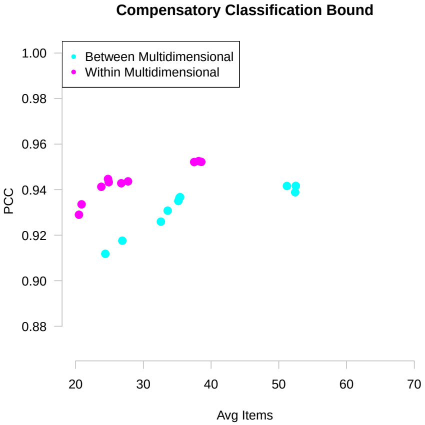


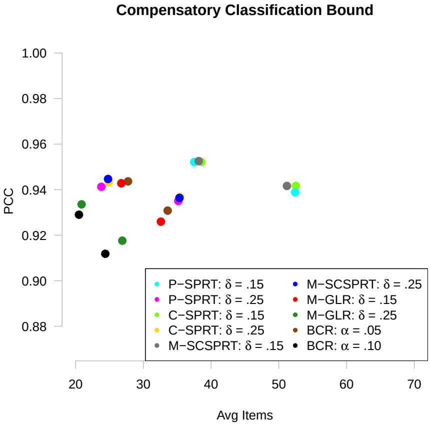
Length and Accuracy: Stopping Rule



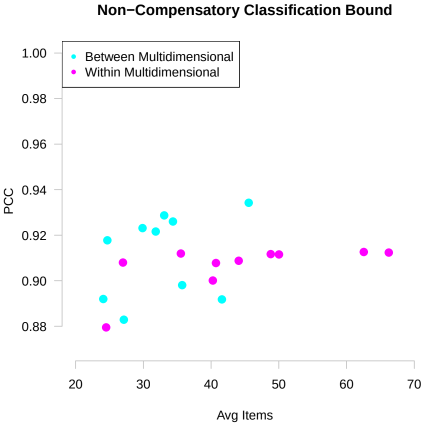


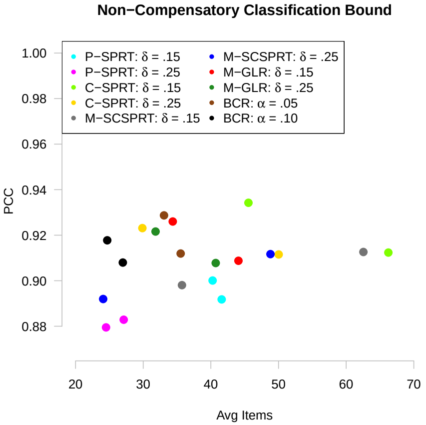
Length and Accuracy: Bank by Stop





Length and Accuracy: Bank by Stop





Summary of Results

What are the answers to the following questions?

1. **Have any of the stopping rules been adequately generalized to multidimensional classification problems?**
2. Are there differences between P- and C- SPRT algorithms in terms of test length and classification accuracy?
3. Do different item banks yield differential performance for different classification bound functions?
4. Do variables other than stopping rule or item bank affect test length or classification accuracy?

Yes. All of the stopping rules except for P-SPRT and M-SCSPRT appear to work in multidimensional mastery tests.

Summary of Results

What are the answers to the following questions?

1. All of the stopping rules except for P-SPRT and M-SCSPRT appear to work in multidimensional mastery tests.
2. **Are there differences between P- and C-SPRT algorithms in terms of test length and classification accuracy?**
3. Do different item banks yield differential performance for different classification bound functions?
4. Do variables other than stopping rule or item bank affect test length or classification accuracy?

Yes. P-SPRT and C-SPRT differ only when the classification bound does not align with the contours of the log-lik function.

Summary of Results

What are the answers to the following questions?

1. All of the stopping rules except for P-SPRT and M-SCSPRT appear to work in multidimensional mastery tests.
2. P-SPRT and C-SPRT differ only when the classification bound does not align with the contours of the log-lik function.
3. **Do different item banks yield differential performance for different classification bound functions?**
4. Do variables other than stopping rule or item bank affect test length or classification accuracy?

Yes. Between-item multidimensionality works best given a non-compensatory classification bound function.

Summary of Results

What are the answers to the following questions?

1. All of the stopping rules except for P-SPRT and M-SCSPRT appear to work in multidimensional mastery tests.
2. P-SPRT and C-SPRT differ only when the classification bound does not align with the contours of the log-likelihood function.
3. Between-item multidimensionality works best given a non-compensatory classification bound function.
4. **Do variables other than stopping rule or item bank affect test length or classification accuracy?**

Not really. Most of the item selection algorithms performed similarly given a reasonable item bank and stopping rule.

Conclusions

More research is needed in multidimensional CCT.

- Should one use Bayesian estimation procedures?
- Can one circumvent computational limitations?
- Do these methods generalize to more than two classification regions?
- How do practitioners consider selection constraints?

Thank You!