Workshop on MPT Modeling

Richard Chechile: Tufts University

July 19, 2011

MPT Background, Development, Estimation and Use

Workshop Structure

Time	Instructor	General Topics
9:00 -10:20	Chechile	MPTs: what, who, why, & how
10:20-10.35	Coffee Break	
10:35-11:50	Batchelder	further theoretical development
11:50-12:00	Brief Break	
12:00-12:45	Hu	computer implementation
12:45-1:45	Lunch	
1:45-2:45	Chechile	estimation and software tools
2:45-2:55	Brief Break	
2:55-3:45	Batchelder	additional theoretical issues
3:45-4:00	Coffee Break	
4:00-5:00	Hu	more software implementation
5:00-5:30	All	general discussion

What is a multinomial processing tree (MPT) model?

- MPT models require one or more sets of categorical data with observed frequencies n_i, e.g.,

n_1	n_2

n_3	n_4	n_5	n_6

- MPT models also require a scientific model in the form of a probability tree that accounts for the observations in the various multinomial categories.

Model A: A four-cell model for old recognition

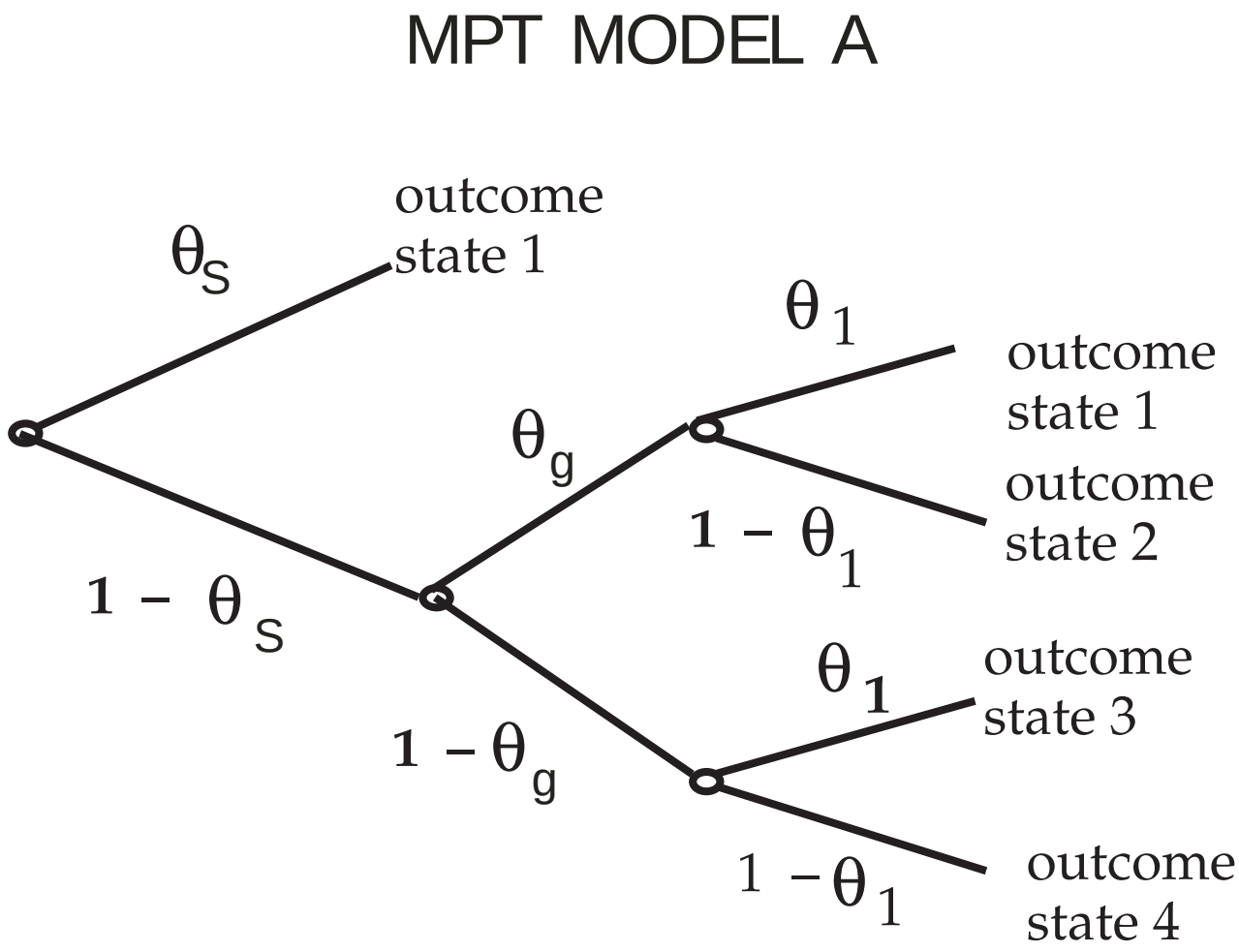
Consider the responses to an old recognition test:

n_1	n_2	n_3	n_4
ϕ_1	ϕ_2	ϕ_3	ϕ_4

- Cells 1 to 4 are respectively "yes high conf.", "yes low conf.", "no high conf.", and "no low conf."
- The ϕ_i value denote population proportion in the observational category

An illustration of Model A

Consider the following MPT model for a single four-cell multinomial:



Mapping Equations for Model A

$$
\begin{aligned}
\phi_1 &= \theta_S + (1-\theta_S)\theta_g\theta_1 \\
\phi_2 &= (1-\theta_S)\theta_g(1-\theta_1) \\
\phi_3 &= (1-\theta_S)(1-\theta_g)\theta_1 \\
\phi_4 &= (1-\theta_S)(1-\theta_g)(1-\theta_1)
\end{aligned}
$$

Likelihood Structure for Model A

$$L = C\,\phi_1(\theta_S, \theta_g, \theta_1)^{n_1}\phi_2(\theta_S, \theta_g, \theta_1)^{n_2}\phi_3(\theta_S, \theta_g, \theta_1)^{n_3}\phi_4(\theta_S, \theta_g, \theta_1)^{n_4}$$

where $C = \frac{n_o!}{n_1!\, n_2!\, n_3!\, n_4!}$ with $n_o = n_1 + n_2 + n_3 + n_4$.

The "Who" of MPT models: Very Brief History

- Outside of psychology similar approach used in genetics (see Elandt-Johnson, 1971)
- Correcting for guessing was a common feature in many studies in psychometrics

The "Who" of MPT models: Very Brief History

- Outside of psychology similar approach used in genetics (see Elandt-Johnson, 1971)
- Correcting for guessing was a common feature in many studies in psychometrics
- Storage-retrieval models Chechile (1973), Chechile & Meyer (1976); Batchelder & Riefer (1980); Chechile (2004).

The "Who" of MPT models: Very Brief History

- Outside of psychology similar approach used in genetics (see Elandt-Johnson, 1971)
- Correcting for guessing was a common feature in many studies in psychometrics
- Storage-retrieval models Chechile (1973), Chechile & Meyer (1976); Batchelder & Riefer (1980); Chechile (2004).
- The term MPT began with the Riefer & Batchelder (1988) *Psy. Rev.* paper

The "Who" of MPT models: Very Brief History

- Outside of psychology similar approach used in genetics (see Elandt-Johnson, 1971)
- Correcting for guessing was a common feature in many studies in psychometrics
- Storage-retrieval models Chechile (1973), Chechile & Meyer (1976); Batchelder & Riefer (1980); Chechile (2004).
- The term MPT began with the Riefer & Batchelder (1988) *Psy. Rev.* paper
- Software: GPT.EXE (Hu & Phillips 1999); AppleTree (Rothkegel, 1999); HMMTree (Stahl & Klauer, 2007); MultiTree (Moshagen, 2009).

The "Who" of MPT models: Very Brief History

- Outside of psychology similar approach used in genetics (see Elandt-Johnson, 1971)
- Correcting for guessing was a common feature in many studies in psychometrics
- Storage-retrieval models Chechile (1973), Chechile & Meyer (1976); Batchelder & Riefer (1980); Chechile (2004).
- The term MPT began with the Riefer & Batchelder (1988) *Psy. Rev.* paper
- Software: GPT.EXE (Hu & Phillips 1999); AppleTree (Rothkegel, 1999); HMMTree (Stahl & Klauer, 2007); MultiTree (Moshagen, 2009).
- Reviews of MPT models in the Batchelder & Riefer (1999) *Psychonomic Bull.* paper and in the Erdfelder et al. (2009) paper in *Zeitschrift für Psychologie*
- Hierarchical MPT models: Klauer (2006); Smith & Batchelder (2009)

Why Use a MPT Model? Reason 1 Measurement (Part A)

- Some of the basic problems in psychology deal with how to measure fundamental psychological processes
- For example, consider the state of memory literature in the 1970s.

Why Use a MPT Model? Reason 1 Measurement (Part A)

- Some of the basic problems in psychology deal with how to measure fundamental psychological processes
- For example, consider the state of memory literature in the 1970s.
- Theories and accounts from experimental reports could creditably account for experimental findings with either a storage or a retrieval explanation
- Without separate measures for storage and retrieval processes, then it was impossible to compile decisive evidence

Why Use a MPT Model? Reason 1 Measurement (Part A)

- Some of the basic problems in psychology deal with how to measure fundamental psychological processes
- For example, consider the state of memory literature in the 1970s.
- Theories and accounts from experimental reports could creditably account for experimental findings with either a storage or a retrieval explanation
- Without separate measures for storage and retrieval processes, then it was impossible to compile decisive evidence
- MPT models such as the models in Chechile (1973), Chechile & Meyer (1976) and Batchelder & Riefer (1980) papers provided a method to separable estimate storage and retrieval processes

Why MPT models?: Measurement (Part B)

- Basic DVs available are confounded measures in isolation.
- % correct, RT, ERP, BOLD are all measures that can be influenced by more than one psychological process (Chechile & Roder, 1998; Chechile, 2007)

Why MPT models?: Measurement (Part B)

- Basic DVs available are confounded measures in isolation.
- % correct, RT, ERP, BOLD are all measures that can be influenced by more than one psychological process (Chechile & Roder, 1998; Chechile, 2007)
- A model will be required to mold these DVs into measures for underlying psychological processes.
- Chechile & Roder (1998) used the term "model-based measurement" and Riefer et al. (2002) used the term "cognitive psychometrics" for this approach.

Why MPT models?: Measurement (Part B)

- Basic DVs available are confounded measures in isolation.
- % correct, RT, ERP, BOLD are all measures that can be influenced by more than one psychological process (Chechile & Roder, 1998; Chechile, 2007)
- A model will be required to mold these DVs into measures for underlying psychological processes.
- Chechile & Roder (1998) used the term "model-based measurement" and Riefer et al. (2002) used the term "cognitive psychometrics" for this approach.
- The measurement model need not be a MPT model (e.g., selective influence, SDT, etc.); nonetheless, the MPT approach provides a simple yet powerful way to obtain basic measures.

Why MPT models?: Simplicity (Part A)

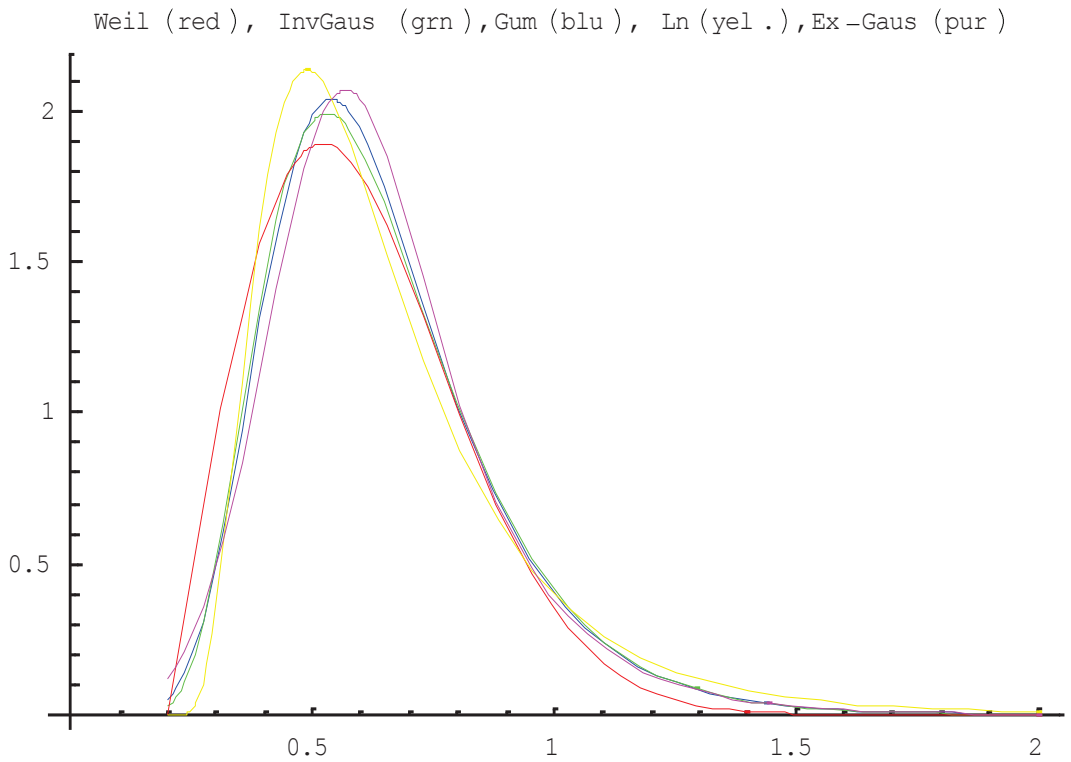
- The multinomial likelihood function is a consensus model of the statistical structure
- MPT modelers might have rival models for a latent psychological models for a task but there is no dispute about the statistical likelihood of the model

Why MPT models?: Simplicity (Part A)

- The multinomial likelihood function is a consensus model of the statistical structure
- MPT modelers might have rival models for a latent psychological models for a task but there is no dispute about the statistical likelihood of the model
- $L = \frac{n!}{n_1! \cdots n_{k+1}!} \prod_{i=1}^{k+1} \phi_i^{n_i}(\Theta)$

Likelihood Agreement More Challenging for Other Data

Consider the following distributions for RT data:



Why MPT models?: Simplicity (Part B)

- In contrast to strength models for multinomial data like SDT, MPT models do not require a detailed stipulation of a psychological scale.

Why MPT models?: Simplicity (Part B)

- In contrast to strength models for multinomial data like SDT, MPT models do not require a detailed stipulation of a psychological scale.
- The simplicity of MPT models also enable class level theorems and tools – more on this with Bill's talk.

Why MPT models?: Simplicity (Part B)

- In contrast to strength models for multinomial data like SDT, MPT models do not require a detailed stipulation of a psychological scale.
- The simplicity of MPT models also enable class level theorems and tools – more on this with Bill's talk.
- Despite the simplicity of the multinomial likelihood function, it enables a powerful and rigorous method for fitting the model and test the model.

Why MPT models?: Mixtures

- There are many reasons to expect that there are mixtures present in most psychological experiments.
- For example, trials where the participant is attentive and comes up with an effective memory encoding versus trials where the participant is inattentive, tired, or ineffective in developing a good encoding.

Why MPT models?: Mixtures

- There are many reasons to expect that there are mixtures present in most psychological experiments.
- For example, trials where the participant is attentive and comes up with an effective memory encoding versus trials where the participant is inattentive, tired, or ineffective in developing a good encoding.
- Mixtures are problematic for strength models like SDT and effectively need a tree-like structure to deal with the mixture.

Why MPT models?: Mixtures

- There are many reasons to expect that there are mixtures present in most psychological experiments.
- For example, trials where the participant is attentive and comes up with an effective memory encoding versus trials where the participant is inattentive, tired, or ineffective in developing a good encoding.
- Mixtures are problematic for strength models like SDT and effectively need a tree-like structure to deal with the mixture.
- Mixture are a natural feature for MPT models.

Is a MPT model a Psychological Theory? (Part A)

- With a measurement focus, MPT are more models of a specific task, and as such it is not a psychological theory.
- For example, the Chechile (2004) models are designed to obtain separate measures for storage and retrieval for a specific task, but these models are not a theory of memory in general.

Is a MPT model a Psychological Theory? (Part A)

- With a measurement focus, MPT are more models of a specific task, and as such it is not a psychological theory.
- For example, the Chechile (2004) models are designed to obtain separate measures for storage and retrieval for a specific task, but these models are not a theory of memory in general.
- Because MPT models are task models, there may be many parameters in the model because the experiment require many corrections, e.g., corrections for guessing, corrections for using a high confidence when information is not stored.
- Although all the parameters of the model are important, many of the parameters are nuisance parameters.

Is a MPT model a Psychological Theory? (Part B)

- The use of many parameters in scientific measurement is not uncommon – for example consider event models in high-energy physics.

Is a MPT model a Psychological Theory? (Part B)

- The use of many parameters in scientific measurement is not uncommon – for example consider event models in high-energy physics.
- Although a MPT model is not a general psychological theory, one can argue that psychological theories are numerous but short-lived.

Is a MPT model a Psychological Theory? (Part B)

- The use of many parameters in scientific measurement is not uncommon – for example consider event models in high-energy physics.
- Although a MPT model is not a general psychological theory, one can argue that psychological theories are numerous but short-lived.
- Yet solid measurement has the potential for being timeless!

Steps for Developing a Novel MPT Model

- Step (1) In the context of measurement goal, a task needs to be designed along with the MPT model for the task.
- Step (2) The proposed MPT model needs to be checked if it is identifiable. If not then redo Step 1.
- Step (3) Develop parameter estimates (MLE, standard Bayesian, Bayesian MCMC, PPM) or utilize a general software package.
- Step (4) Study test cases and Monte Carlo studies to learn about the statistical properties as a function of sample size.
- Step (5) Use the model for analyzing experimental studies. Check for goodness-of-fit and other tests of the model's performance. See if the experiments provide evidence for the validity of the model parameters.

Example 1: Storage/Retrieval Measurement Problem

- From an information processing framework, the concept of storage and retrieval are foundational psychological constructs.

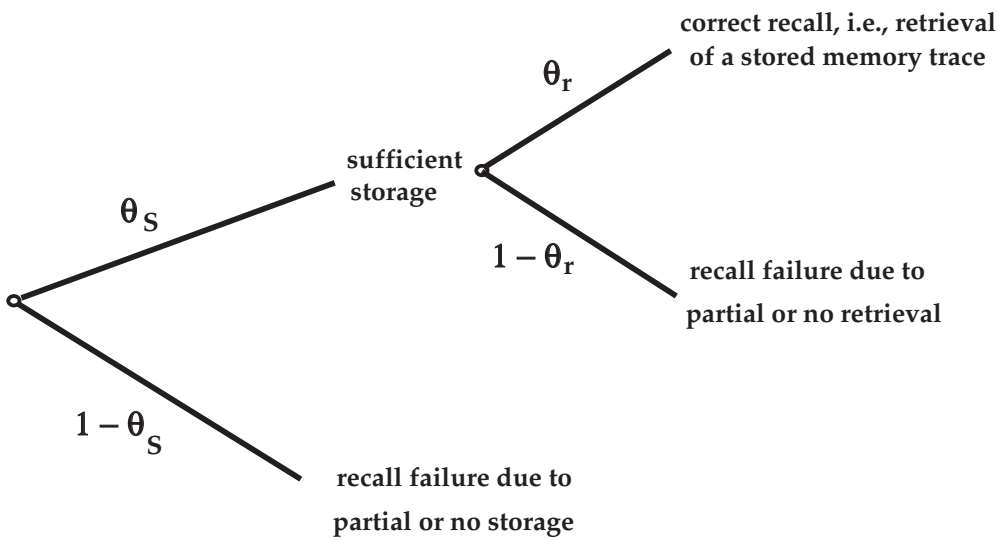
Example 1: Storage/Retrieval Measurement Problem

- From an information processing framework, the concept of storage and retrieval are foundational psychological constructs.
- When there is lost of memory retention, the lost could be caused by a retrieval defect (i.e., the items not remembered are still stored but not retrieved at the time of test) or the lost could be caused by a loss of some items from the memory system (i.e., the items are not sufficiently well stored to support the correct recall of the items).

Example 1: Storage/Retrieval Measurement Problem

- From an information processing framework, the concept of storage and retrieval are foundational psychological constructs.
- When there is lost of memory retention, the lost could be caused by a retrieval defect (i.e., the items not remembered are still stored but not retrieved at the time of test) or the lost could be caused by a loss of some items from the memory system (i.e., the items are not sufficiently well stored to support the correct recall of the items).
- Yet correct recall must be the result of both successful item storage and successful retrieval of that information. Hence, the proportion of correct recall must equal $\theta_S \, \theta_r$.

Recall Task Model is not Identifiable



Problem here is two model parameters and one degree of freedom for the recall data.

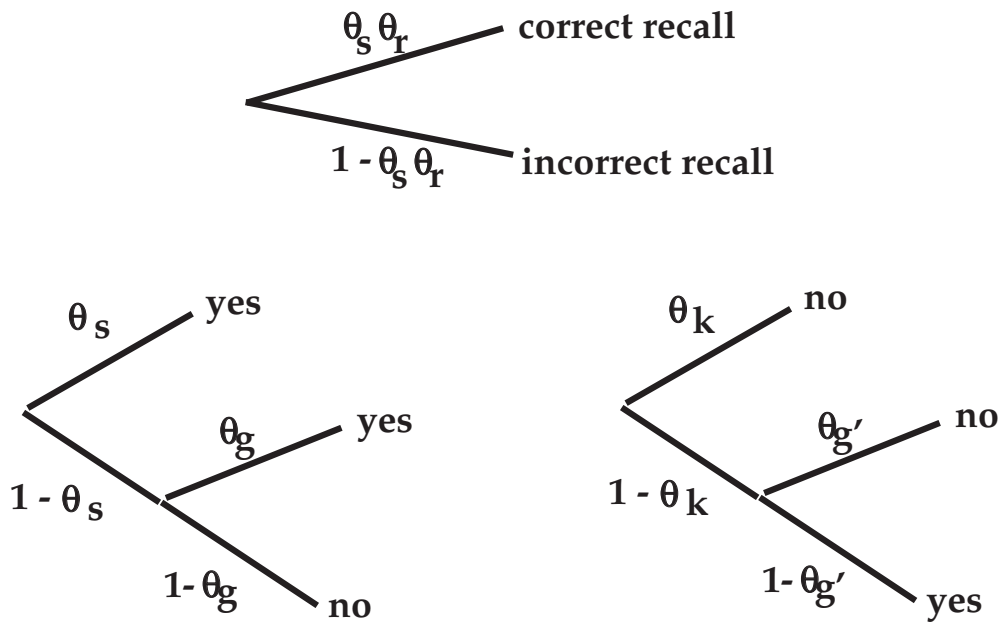
Will Recall Plus Recognition Work?

- Suppose we collect both recall data and old recognition data. In a recognition test we bypass the necessity to retrieve the memory representation by presenting the target again and simply ask for a yes versus no decision.
- If the item is stored in memory and the participant is tested with an old recognition probe, then the participant should respond "yes" because retrieval is not required. The concept of content addressable memory is a creditable idea from the memory literature.

Will Recall Plus Recognition Work?

- Suppose we collect both recall data and old recognition data. In a recognition test we bypass the necessity to retrieve the memory representation by presenting the target again and simply ask for a yes versus no decision.
- If the item is stored in memory and the participant is tested with an old recognition probe, then the participant should respond "yes" because retrieval is not required. The concept of content addressable memory is a creditable idea from the memory literature.
- But if we have old recognition it is necessary that we also have new recognition tests.
- Moreover, yes/no recognition testing can also be correct due to guessing. Guess might be different on old and new recognition test trial based on partial knowledge. So let see how these nuisance factors might affect the yes/no judgment.

Recall Plus Recognition MPT is Still Not Identifiable



Problem: There are five model parameters and three degrees of freedom.

The Chechile & Meyer (1976) Procedure

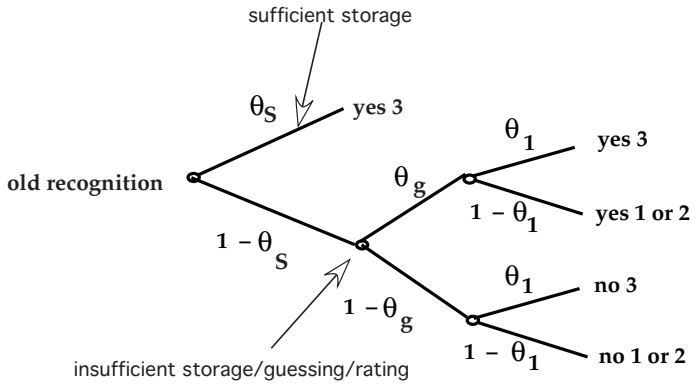
Random intermixing of recall, old, and new along with 3-point confidence rating on recognition test trials.

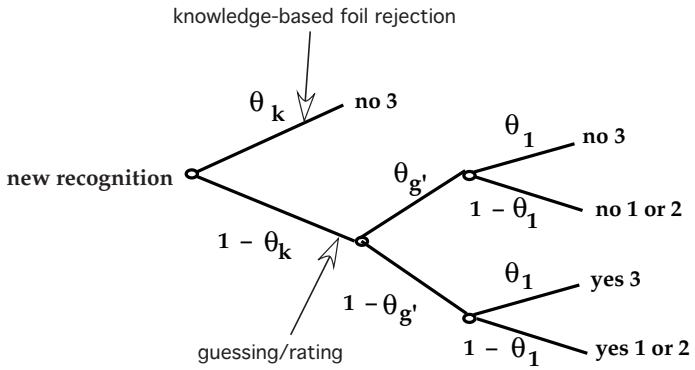
RECALL	
correct	incorrect
ϕ_1	ϕ_2

	RECOGNITION			
	no 3	no 1 or 2	yes 1 or 2	yes 3
old	ϕ_3	ϕ_4	ϕ_5	ϕ_6
new	ϕ_7	ϕ_8	ϕ_9	ϕ_{10}

Now there are seven degrees of freedom.

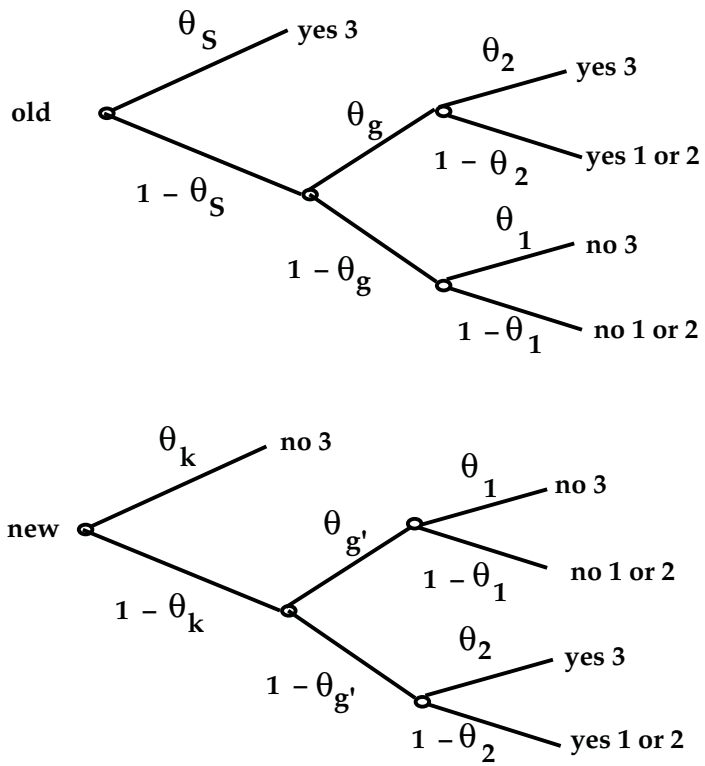
The Chechile (2004) 6P Model





The Chechile (2004) 7B Model

Chechile (2004)



MODEL 7 B

Task and Model Design is a Creative Process

- Step 1 (Task and Model Design) is a creative process that requires excellent domain content knowledge.
- Step 1 and an informal consideration of Step 2 (Identifiability) might require several iterations.

Task and Model Design is a Creative Process

- Step 1 (Task and Model Design) is a creative process that requires excellent domain content knowledge.
- Step 1 and an informal consideration of Step 2 (Identifiability) might require several iterations.
- Models can fail because of flawed conceptual framing, so do Step 1 very carefully

Task and Model Design is a Creative Process

- Step 1 (Task and Model Design) is a creative process that requires excellent domain content knowledge.
- Step 1 and an informal consideration of Step 2 (Identifiability) might require several iterations.
- Models can fail because of flawed conceptual framing, so do Step 1 very carefully
- Step 2 is more complex than just degrees of freedom counting.

Likelihood Identification and Global Identification

Definition: From Drèze (1975) and Chechile (1977), two model points $(\theta_i)_1 \neq (\theta_i)_2$, $i = 1, \cdots, m$, are called "observationally equivalent" if the likelihood functions for the two points are equal for any set of outcomes (n_j) for $j = 1, \cdots, k+1$, i.e., $P[(n_j) \,|\, (\theta_i)_1] = P[(n_j) \,|\, (\theta_i)_2]$.

Definition: A model is "likelihood identified" if for any point in the model space, there is no other point that is observationally equivalent to it, i.e., $\forall\, (\theta_i)_k \in \Theta$ there are no observationally equivalent points.

If the model is likelihood identified, then it is a "globally identified" model.

Identification Is Not Just a DF Issue

- A model that has more parameters than statistical degrees of freedom is not likelihood identifiable.

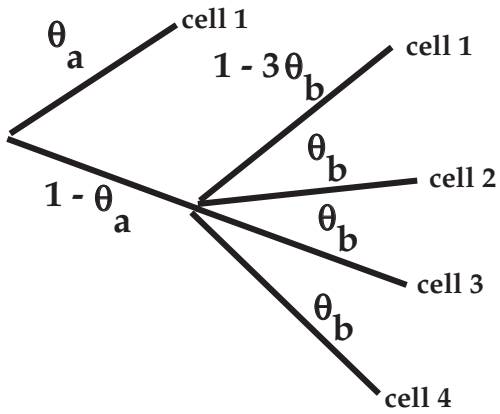
Identification Is Not Just a DF Issue

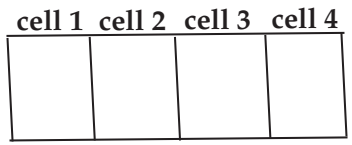
- A model that has more parameters than statistical degrees of freedom is not likelihood identifiable.
- Chechile (1977) did show that it is possible to still obtain separate Bayesian marginal estimates of parameters for some models that are not likelihood identified, but that approach has its limitations.

Identification Is Not Just a DF Issue

- A model that has more parameters than statistical degrees of freedom is not likelihood identifiable.
- Chechile (1977) did show that it is possible to still obtain separate Bayesian marginal estimates of parameters for some models that are not likelihood identified, but that approach has its limitations.
- Although we require the number of model parameters to be equal to or fewer than the statistical degrees of freedom for likelihood identification, **not all models that meet this constraint are identifiable**.

An Unidentifiable Model with $m < k$





For example, consider model points $(\theta_a = .6, \theta_b = .25)$ and $(\theta_a = .5, \theta_b = .2)$. These points are observationally equivalent and result in cell proportions of $(.7, .1, .1, .1)$.

How to proved that a MPT model is Identifiable

- Chechile (1998) showed that a MPT model is likelihood identified if the mapping of any vector of model values $(\theta_1, \cdots, \theta_m)$ to a vector of cell proportions $(\phi_1, \cdots, \phi_{k+1})$ is $1-$to-1, i.e., each $(\theta_1, \cdots, \theta_m)$ vector maps to a different $(\phi_1, \cdots, \phi_{k+1})$ vector.

How to proved that a MPT model is Identifiable

- Chechile (1998) showed that a MPT model is likelihood identified if the mapping of any vector of model values $(\theta_1, \cdots, \theta_m)$ to a vector of cell proportions $(\phi_1, \cdots, \phi_{k+1})$ is $1-$to-1, i.e., each $(\theta_1, \cdots, \theta_m)$ vector maps to a different $(\phi_1, \cdots, \phi_{k+1})$ vector.
- A model can be proved to be identifiable by assuming that the mapping of a (θ_i) to a (ϕ_j) vector is not a $1-$to-1 mapping and deriving a contradiction, i.e., proof by contradiction.

Model Identification via a Proof by Contradiction

- Suppose for example we consider the model where $\phi_1 = \theta_a$, $\phi_2 = (1 - \theta_a)\theta_b$, and $\phi_3 = (1 - \theta_a)(1 - \theta_b)$. We begin by assuming it is not identified, so there must be two observational equivalent points.

Model Identification via a Proof by Contradiction

- Suppose for example we consider the model where $\phi_1 = \theta_a$, $\phi_2 = (1 - \theta_a)\theta_b$, and $\phi_3 = (1 - \theta_a)(1 - \theta_b)$. We begin by assuming it is not identified, so there must be two observational equivalent points.
- Let the first be denoted as $\theta_a = a$ and $\theta_b = b$ and the second is $\theta_a = a + \epsilon_a$ and $\theta_b = b + \epsilon_b$ where $\epsilon_a \neq 0$ and $\epsilon_b \neq 0$.

Model Identification via a Proof by Contradiction

- Suppose for example we consider the model where $\phi_1 = \theta_a$, $\phi_2 = (1 - \theta_a)\theta_b$, and $\phi_3 = (1 - \theta_a)(1 - \theta_b)$. We begin by assuming it is not identified, so there must be two observational equivalent points.
- Let the first be denoted as $\theta_a = a$ and $\theta_b = b$ and the second is $\theta_a = a + \epsilon_a$ and $\theta_b = b + \epsilon_b$ where $\epsilon_a \neq 0$ and $\epsilon_b \neq 0$.
- From the first equation and the assumption of observational equivalence, it follows that $a = a + \epsilon_a$, but this implies that $\epsilon_a = 0$.

Model Identification via a Proof by Contradiction

- Suppose for example we consider the model where $\phi_1 = \theta_a$, $\phi_2 = (1 - \theta_a)\theta_b$, and $\phi_3 = (1 - \theta_a)(1 - \theta_b)$. We begin by assuming it is not identified, so there must be two observational equivalent points.
- Let the first be denoted as $\theta_a = a$ and $\theta_b = b$ and the second is $\theta_a = a + \epsilon_a$ and $\theta_b = b + \epsilon_b$ where $\epsilon_a \neq 0$ and $\epsilon_b \neq 0$.
- From the first equation and the assumption of observational equivalence, it follows that $a = a + \epsilon_a$, but this implies that $\epsilon_a = 0$.
- From second equation it also follows that $(1 - a)b = (1 - a - \epsilon_a)(b + \epsilon_b) = (1 - a)(b + \epsilon_b)$, which implies that $\epsilon_b = 0$.

Model Identification via a Proof by Contradiction

- Suppose for example we consider the model where $\phi_1 = \theta_a$, $\phi_2 = (1 - \theta_a)\theta_b$, and $\phi_3 = (1 - \theta_a)(1 - \theta_b)$. We begin by assuming it is not identified, so there must be two observational equivalent points.
- Let the first be denoted as $\theta_a = a$ and $\theta_b = b$ and the second is $\theta_a = a + \epsilon_a$ and $\theta_b = b + \epsilon_b$ where $\epsilon_a \neq 0$ and $\epsilon_b \neq 0$.
- From the first equation and the assumption of observational equivalence, it follows that $a = a + \epsilon_a$, but this implies that $\epsilon_a = 0$.
- From second equation it also follows that $(1 - a)b = (1 - a - \epsilon_a)(b + \epsilon_b) = (1 - a)(b + \epsilon_b)$, which implies that $\epsilon_b = 0$.
- Thus there cannot be observationally equivalent points for this model and, the mapping is $1-$to-1, and the model is likelihood identified.

Comments about Model Identification

- Likelihood identification or global identification is a strong claim that requires a proof and it holds for any data configuration.

Comments about Model Identification

- Likelihood identification or global identification is a strong claim that requires a proof and it holds for any data configuration.
- When investigator find a unique maximum likelihood estimate for parameter for a specific data set, then the model is often called locally identified. This is weaker claim because the maximum likelihood could be a local maxima.

Comments about Model Identification

- Likelihood identification or global identification is a strong claim that requires a proof and it holds for any data configuration.
- When investigator find a unique maximum likelihood estimate for parameter for a specific data set, then the model is often called locally identified. This is weaker claim because the maximum likelihood could be a local maxima.
- Some investigators use a non-identified model and later equate some parameters in the model in order to obtain to make the model identified.

Comments about Model Identification

- Likelihood identification or global identification is a strong claim that requires a proof and it holds for any data configuration.
- When investigator find a unique maximum likelihood estimate for parameter for a specific data set, then the model is often called locally identified. This is weaker claim because the maximum likelihood could be a local maxima.
- Some investigators use a non-identified model and later equate some parameters in the model in order to obtain to make the model identified.
- Alternatively other investigators change the task in order to obtain a more rich data structure to model and obtain an identified model.

Step 3: Parameter Estimation

- MLE
- Standard Bayesian Analysis
- MCMC Bayesian
- Population Parameter Mapping
- Software Tools

Ways to Obtain MLE Values

- (1) Solving the set of maximization equations, i.e., solve for $\hat{\theta}_1, \cdots, \hat{\theta}_m$ where $\frac{\partial L(\theta_1, \cdots, \theta_m)}{\partial \theta_i} = 0,\ i = 1, \cdots, m.$

Ways to Obtain MLE Values

- (1) Solving the set of maximization equations, i.e., solve for $\hat{\theta}_1, \cdots, \hat{\theta}_m$ where $\frac{\partial L(\theta_1, \cdots, \theta_m)}{\partial \theta_i} = 0,\ i = 1, \cdots, m.$
- (2) Constrained search of parameter space Θ to find the max. likelihood.

Ways to Obtain MLE Values

- (1) Solving the set of maximization equations, i.e., solve for $\hat{\theta}_1, \cdots, \hat{\theta}_m$ where $\frac{\partial L(\theta_1, \cdots, \theta_m)}{\partial \theta_i} = 0$, $i = 1, \cdots, m$.
- (2) Constrained search of parameter space Θ to find the max. likelihood.
- (3) A hybrid of (1) and (2).

Ways to Obtain MLE Values

- (1) Solving the set of maximization equations, i.e., solve for $\hat{\theta}_1, \cdots, \hat{\theta}_m$ where $\frac{\partial L(\theta_1, \cdots, \theta_m)}{\partial \theta_i} = 0$, $i = 1, \cdots, m$.
- (2) Constrained search of parameter space Θ to find the max. likelihood.
- (3) A hybrid of (1) and (2).
- (4) The EM algorithm
- (5) Software tools

Expectation-Maximization Algorithm

- The EM algorithm from Dempster, Laird, & Rubin (1977) maximize a function and is use widely as method to obtain MLE value.
- The algorithm is way to guess initial values for the model parameters and through a series steps have convergence on the maximum of the likelihood function.

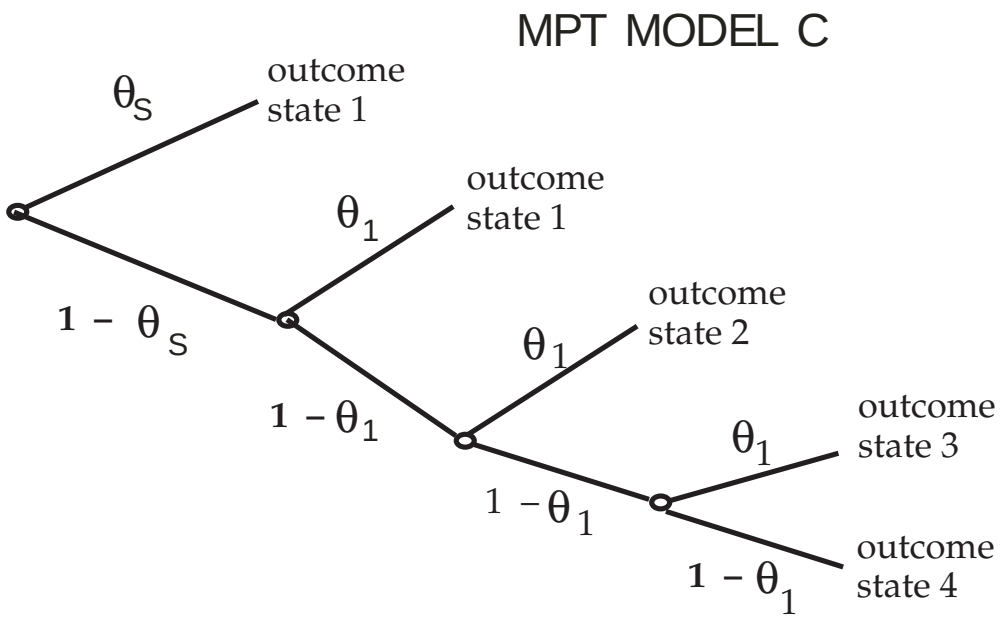
Expectation-Maximization Algorithm

- The EM algorithm from Dempster, Laird, & Rubin (1977) maximize a function and is use widely as method to obtain MLE value.
- The algorithm is way to guess initial values for the model parameters and through a series steps have convergence on the maximum of the likelihood function.
- In each iteration there is an expectation step (the E-step) where values for an augment data structure is estimated. This leads to a conditional maximization step (the M-step).

Expectation-Maximization Algorithm

- The EM algorithm from Dempster, Laird, & Rubin (1977) maximize a function and is use widely as method to obtain MLE value.
- The algorithm is way to guess initial values for the model parameters and through a series steps have convergence on the maximum of the likelihood function.
- In each iteration there is an expectation step (the E-step) where values for an augment data structure is estimated. This leads to a conditional maximization step (the M-step).
- The E-M algorithm is a key component in Xiangen's software (More details later)

Tree Model to Illustrate the EM algorithm



Equations for MPT Model

$$
\begin{aligned}
\phi_1 &= \theta_s + (1-\theta_s)\theta_1 \\
\phi_2 &= (1-\theta_s)(1-\theta_1)\theta_1 \\
\phi_3 &= (1-\theta_s)(1-\theta_1)^2\theta_1 \\
\phi_4 &= (1-\theta_s)(1-\theta_1)^3
\end{aligned}
$$

n_1	n_2	n_3	n_4

Augmented Data Structure

Replace the 4-cell multinomial with an augmented multinomial.

n_{11}	n_{12}	n_2	n_3	n_4

where $\phi_{11} = \theta_s$, $\phi_{12} = (1 - \theta_s)\theta_1$, $n_1 = n_{11} + n_{12}$ and

$L = K\theta_s^{n_{11}}(1 - \theta_s)^{n - n_{11}}\theta_1^{n_{12}+n_2+n_3}(1 - \theta_1)^{n_2+2n_3+3n_4}$

For fixed θ_1 the step max. for ith iteration is $\theta_s^{(i+1)} = \frac{n_{11}^{(i)}}{n}$.

Example of the EM Algorithm for θ_s

Suppose the frequencies are $(720, 84, 59, 137)$ $n = 1,000$ and we guess initially $\theta_s^{(0)} = .5$ and $\theta_1^{(0)} = .3$.

i	$\theta_s^{(i)}$	$n_{12}^{(i)} = n(1-\theta_s^{(i)})\theta_1^{(0)}$	$n_{11}^{(i)} = n_1 - n_{12}^{(i)}$	$\theta_s^{i+1} = \frac{n_{11}^{(i)}}{n}$
0	.5	150	570	.570
1	.570	129	591	.591
2	.591	122.7	597.3	.5973
3	.5973	120.81	599.19	.59919
4	.59919	120.243	599.757	.599757
5	.599757	120.0729	599.9271	.5999271

Standard Bayesian Estimation for Model C

Let the prior distribution for θ_s and θ_1 be uniform in the unit square, and let us denote the data as D, the joint posterior distribution as $f(\Theta|D)$, and the marginal on θ_s as $f(\theta_s|D)$. It follow that

$$
\begin{aligned}
f(\Theta|D) &= K(\theta_s + (1-\theta_s)\theta_1)^{n_1}(1-\theta_s)^{n-n_1}\theta_1^{n_2+n_3}(1-\theta_1)^{n_2+2n_3+3n_4} \\
f(\theta_s|D) &= K\sum_{i=0}^{n_1}\theta_s^{n_1-i}(1-\theta_s)^{n-n_1+i}\int_0^1 \theta_1^{i+n_2+n_3}(1-\theta_1)^{n_2+2n_3+3n_4}\,d\theta \\
&= K\sum_{i=0}^{n_1}\theta_s^{n_1-i}(1-\theta_s)^{n-n_1+i}\frac{(n_2+n_3+i)!\,(n_2+2n_3+3n_4)!}{(2n_2+3n_3+3n_4+i+1)!} \\
K &= \sum_{i=0}^{n_1}\frac{(n_1-i)!\,(n-n_1+i)!\,(n_2+n_3+i)!\,(n_2+2n_3+3n_4)!}{(n+1)!\,(2n_2+3n_3+3n_4+i+1)!}
\end{aligned}
$$

Remarks about Standard Bayesian Approach

- Point estimates for parameters are taken as central tendency measures from the posterior marginal distribution; usually the mean is the point estimate.

Remarks about Standard Bayesian Approach

- Point estimates for parameters are taken as central tendency measures from the posterior marginal distribution; usually the mean is the point estimate.
- Bayesian posterior distribution is a probability distribution, so interval estimates are just another property of the posterior distribution.

Remarks about Standard Bayesian Approach

- Point estimates for parameters are taken as central tendency measures from the posterior marginal distribution; usually the mean is the point estimate.
- Bayesian posterior distribution is a probability distribution, so interval estimates are just another property of the posterior distribution.
- Because there is a probability distribution for each parameter in each condition, we can, without further assumptions, test hypotheses. Consequently, we can directly compute a probability for a hypothesis like $P(\theta_s(2) > \theta_s(1)|D)$. More on this point in the PM.

Remarks about Standard Bayesian Approach

- Point estimates for parameters are taken as central tendency measures from the posterior marginal distribution; usually the mean is the point estimate.
- Bayesian posterior distribution is a probability distribution, so interval estimates are just another property of the posterior distribution.
- Because there is a probability distribution for each parameter in each condition, we can, without further assumptions, test hypotheses. Consequently, we can directly compute a probability for a hypothesis like $P(\theta_s(2) > \theta_s(1)|D)$. More on this point in the PM.
- Despite these desirable features, the standard Bayesian analysis is usually very computational challenging that requires numerical integration over a number of parameters.

Monte Carlo Methods

- Two methods use Monte Carlo methods to circumvent computational problems with the Bayesian approach, i.e., Markov Chain Monte Carlo (MCMC) and Population Parameter Mapping (PPM).
- With both methods, a random vectors of model parameter values are sampled from the posterior distribution.
- Monte Carlo sampling provide a way to obtain all the information that would be achieved with the standard Bayesian analysis but without the integration (analytic or numerical).

Monte Carlo Methods

- Two methods use Monte Carlo methods to circumvent computational problems with the Bayesian approach, i.e., Markov Chain Monte Carlo (MCMC) and Population Parameter Mapping (PPM).
- With both methods, a random vectors of model parameter values are sampled from the posterior distribution.
- Monte Carlo sampling provide a way to obtain all the information that would be achieved with the standard Bayesian analysis but without the integration (analytic or numerical).
- Monte Carlo methods provide a means to examine much more complex models. For example, hierarchical models where items and subjects are treated with separate parameters drawn from a hyper distribution. More about this later.

Monte Carlo Methods

- Two methods use Monte Carlo methods to circumvent computational problems with the Bayesian approach, i.e., Markov Chain Monte Carlo (MCMC) and Population Parameter Mapping (PPM).
- With both methods, a random vectors of model parameter values are sampled from the posterior distribution.
- Monte Carlo sampling provide a way to obtain all the information that would be achieved with the standard Bayesian analysis but without the integration (analytic or numerical).
- Monte Carlo methods provide a means to examine much more complex models. For example, hierarchical models where items and subjects are treated with separate parameters drawn from a hyper distribution. More about this later.
- PPM is an exact sampling procedure whereas MCMC is an approximate method. More of these features later.

Two Fundamental Monte Carlo Techniques

Monte Carlo sampling from a distribution can be based on either (1) the inverse transform method, or (2) the acceptance-rejection algorithm from Jonathan von Neumann. Both techniques rely on an existing method for selecting random values from the uniform $U(0,1)$.

Inverse transform method: Suppose there is a closed form function for the cumulative distribution $F(x)$, e.g., the standard logistic distribution $F(z) = \frac{1}{1+e^{-z}}$. To generate a random z_0, we select a random value from $x_0 \sim U(0,1)$ and $z_0 = \log(\frac{x_0}{1-x_0})$.

The acceptance-rejection method: Factor the density function such that $f(x) = c\, g(x)\, h(x)$ where $h(x)$ is a probability distribution that you know how to sample random values and $g(x)$ is a function on $(0,1)$. Sample x_0 from $h(x)$ and sample u_0 from $U(0,1)$. Accept x_0 only if $u_0 \leq g(x_0)$.

Example of the Acceptance-Rejection Method

Suppose we want to draw a random value from the standard normal, $f(z) = \frac{1}{(2\pi)^{\frac{1}{2}}} e^{-\frac{z^2}{2}}$. Factored into $f(z) = cg(z)h(z)$ form where

$$
\begin{aligned}
g(z) &= \frac{(1+e^{-z})^2}{4} \exp(-\frac{-z^2}{2} + z) \\
c &= \frac{4}{(2\pi)^{\frac{1}{2}}} \\
h(z) &= \frac{e^{-z}}{(1+e^{-z})^2}
\end{aligned}
$$

Draw random $z_0 \sim h(z)$, evaluate $g(z_0)$, draw random $u_0 \sim U(0,1)$, accept if $u_0 \leq g(z_0)$. Average acceptance rate is $\frac{1}{c} = .6267$.

MCMC Sampling from the Posterior Distribution

- Markov Chain Monte Carlo is a procedure (if properly implemented) that converges eventually on the correct target distribution. However, there are clearly preferable ways to do this for MPT modeling. Bad choices can result in either an interminably long time to generate a suitable sample from the posterior distribution or worse it convergence quickly to the wrong distribution. More on this in the PM.

MCMC Sampling from the Posterior Distribution

- Markov Chain Monte Carlo is a procedure (if properly implemented) that converges eventually on the correct target distribution. However, there are clearly preferable ways to do this for MPT modeling. Bad choices can result in either an interminably long time to generate a suitable sample from the posterior distribution or worse it convergence quickly to the wrong distribution. More on this in the PM.
- I find the Gibbs sampler to be impractical because it relies on sampling from conditional marginal distributions that are either not available or difficult to sample from.

MCMC Sampling from the Posterior Distribution

- Markov Chain Monte Carlo is a procedure (if properly implemented) that converges eventually on the correct target distribution. However, there are clearly preferable ways to do this for MPT modeling. Bad choices can result in either an interminably long time to generate a suitable sample from the posterior distribution or worse it convergence quickly to the wrong distribution. More on this in the PM.
- I find the Gibbs sampler to be impractical because it relies on sampling from conditional marginal distributions that are either not available or difficult to sample from.
- For my MPT modeling experience, the Metropolis-Hastings algorithm can converge quickly and be very easy to implement. It is a preferable MCMC algorithm for MPT models.

The Metropolis-Hastings Algorithm

Guess current state $(\theta_{1c}, \cdots, \theta_{mc})$ and for $i = 1, \cdots, m$, generate a random proposal θ_{ip} (e.g. $\theta_{ip} \sim U(0,1)$) and evaluate

$r_i = \frac{f(\theta_{1c}, \cdots, \theta_{ip}, \cdots, \theta_{mc})}{f(\theta_{1c}, \cdots, \theta_{ic}, \cdots, \theta_{mc})}$. If $r_i \geq 1$, then replace θ_{ic} with θ_{ip}. If $r_i < 1$, then generate $u_0 \sim U(0,1)$ and replace θ_{ic} with θ_{ip} if $u_0 \leq r_i$; otherwise θ_{ic} is unchanged.

Repeat above sampling for other i values, i.e., one parameter at a time. In some case with MPT models it is necessary to change more than one parameter at a time.

Repeat above for a "burn in" period, say 3 million cycles. Now the sampling should produce random values from the posterior distribution, but keep only a fraction of the samples in this asymptotic region in order to avoid autocorrelated samples. For example, skip J samples and take the $J + 1$ sample.

The PPM Method: Chechile (1998; 2004; 2010)

PPM begins by sampling random vectors $\langle \phi_1, \cdots, \phi_{k+1} \rangle$ from the posterior Dirichlet distribution where the Dirichlet is

$$f(\langle \phi_1, \cdots, \phi_{k+1} \rangle | D) = \frac{(n+k)!}{n_1! \cdots n_{k+1}!} \prod_{i=1}^{k+1} \phi_i^{n_i}$$

Map $\langle \phi_1, \cdots, \phi_{k+1} \rangle \longrightarrow \langle \theta_1, \cdots, \theta_m \rangle$ if possible. The successfully mapped samples to $\langle \theta_1, \cdots, \theta_m \rangle$ vectors determine all the inferential information for the MPT model. The proportion of successfully mapped vectors is the quantity $P(coh)$ which can be used to assess the model.

Comments on the PPM Method

- PPM is an exact Monte Carlo method, i.e., it does not rely on MCMC convergence and all the samples will not be autocorrelated.

Comments on the PPM Method

- PPM is an exact Monte Carlo method, i.e., it does not rely on MCMC convergence and all the samples will not be autocorrelated.
- The $P(coh)$ can be used to assess the model even in the case of a saturated model. It can also be used to select between rival MPT models.
- More on PPM in the PM.

Step 4: Sample Size Studies

Do repeated samples from the model space and in each case determine a set of frequencies from the corresponding multinomial distribution, i.e.,

$$\langle \theta_1, \cdots, \theta_m \rangle \sim \Theta \longrightarrow \langle \phi_1, \cdots, \phi_{k+1} \rangle$$

Compute ϕ_1, $\phi_1 + \phi_2$, $\phi_1 + \phi_2 + \phi_3 \cdots$ and base on n random values from $U(0, 1)$ determine the frequencies in the response categories n_1, n_2, $\cdots$, n_{k+1}. Given the "data" find $\hat{\theta}_1, \cdots, \hat{\theta}_m$. For each parameter compare $|\hat{\theta}_j - \theta_j|$ as a function of sample size n and method of point estimation.

Step 5: Model Use and GOF

- The parameter point estimates can be use to obtain expected frequencies for each multinomial cell.
- The contrast between the expected and observed frequencies is use to obtain a goodness-of-fit (GOF) measure which is compared with the critical chi-squared value where the degrees of freedom are equal to the statistical data structure DF minus the number of model parameters.

Step 5: Model Use and GOF

- The parameter point estimates can be use to obtain expected frequencies for each multinomial cell.
- The contrast between the expected and observed frequencies is use to obtain a goodness-of-fit (GOF) measure which is compared with the critical chi-squared value where the degrees of freedom are equal to the statistical data structure DF minus the number of model parameters.
- There are a number of tests that are possible, e.g., chi-squared, G-squared, power divergence I^{λ}.

Step 5: Model Use and GOF

- The parameter point estimates can be use to obtain expected frequencies for each multinomial cell.
- The contrast between the expected and observed frequencies is use to obtain a goodness-of-fit (GOF) measure which is compared with the critical chi-squared value where the degrees of freedom are equal to the statistical data structure DF minus the number of model parameters.
- There are a number of tests that are possible, e.g., chi-squared, G-squared, power divergence I^{λ}.
- Be careful with chi-squared and G-squared statistics when any of the model parameters is near either 0 or 1.

Step 5: Model Use and GOF

- The parameter point estimates can be use to obtain expected frequencies for each multinomial cell.
- The contrast between the expected and observed frequencies is use to obtain a goodness-of-fit (GOF) measure which is compared with the critical chi-squared value where the degrees of freedom are equal to the statistical data structure DF minus the number of model parameters.
- There are a number of tests that are possible, e.g., chi-squared, G-squared, power divergence I^{λ}.
- Be careful with chi-squared and G-squared statistics when any of the model parameters is near either 0 or 1.
- Read & Cressie (1988) showed that when $\lambda = \frac{2}{3}$ that their power divergence statistics is well represented by the chi-squared distribution.
- $I^{\frac{2}{3}} = \frac{9}{5} \sum_i n_i [(\frac{n_i}{n_{ei}})^{\frac{2}{3}} - 1]$.

Step 5: Model Use and Validation

- For actual experiments, does the MPT model have a reasonable assessment measure? Is the GOF measure and/or the $P(coh)$ measure reasonable for a good model? More on this point later.

Step 5: Model Use and Validation

- For actual experiments, does the MPT model have a reasonable assessment measure? Is the GOF measure and/or the $P(coh)$ measure reasonable for a good model? More on this point later.
- Construct Validation: Do the model parameters measure what they should?

Step 5: Model Use and Validation

- For actual experiments, does the MPT model have a reasonable assessment measure? Is the GOF measure and/or the $P(coh)$ measure reasonable for a good model? More on this point later.
- Construct Validation: Do the model parameters measure what they should?
- Are there experimental dissociations?

Step 5: Model Use and Validation

- For actual experiments, does the MPT model have a reasonable assessment measure? Is the GOF measure and/or the $P(coh)$ measure reasonable for a good model? More on this point later.
- Construct Validation: Do the model parameters measure what they should?
- Are there experimental dissociations?
- Model comparison issues: Bayes factor, $P(coh)$, and MDL. More on this in Bill talk.

End of Part I: Selective References

- Rationale for why use MPT models: see Chechile & Roder (1998), Chechile (2007), & Riefer et al. (2002)
- Reviews of the literature: Batchelder & Riefer (1999), & Erdfelder et al. (2009)
- PPM methodology: Chechile (1998; 2004; 2010)
- Model identification: Chechile (1977; 1998); Drèze (1975)
- GPT.exe software: Hu & Phillips (1999)
- EM algorithm: Dempter et al. (1977)
- Bayesian methods: Kruschke (2011), Box & Tiao (1973)
- Goodness-of-fit: Read & Cressie (1988)
- Monte Carlo Methods: Fishman (1996)