Clever Randomization and Ensembling Strategies for Accommodating Multiple Data Pathologies in Conjoint Studies

Jeff Dotson (BYU) and Marc Dotson (BYU)

Southern Utah University

April 4, 2019

Outline

Conjoint Analysis in Marketing

- Products are defined by **attributes**, each with a number of **levels**
- Individuals choose from among a number of alternatives
- We infer **individual-level preferences** for each attribute level
- We can use these preferences to perform **counterfactual analyses**
- Instead of evaluating attributes separately, individuals are forced to make **tradeoffs** among combinations of various product attribute and price levels

Conjoint Analysis Example - Choice Task

Taking into consideration all of the features presented, please pick the **pregnancy monitoring device** that you would be **most likely to buy**.

	Option 1	Option 2	Option 3	Options 4
Tracks fetal heart rate	summer infant	MOTOROLA	bloomlife	Owlet
Tracks fetal heart rate	✓	✗	✓	✗
Records fetal heartbeat	✓	✓	✗	✗
Notifies if fetal health (heartbeat and kicks) is trending outside normal range	✗	✓	✗	✓
Counts kicks	✓	✓	✗	✓
Tracks contractions	✗	✓	✗	✓
Tracks maternal sleep position	✓	✗	✓	✓
Provides reports to share with friends and family	✗	✓	✓	✓
Payment Schedule	Rental	1 time payment	Rental	Monthly payments
Total Price	$174	$399	$374	$374
Refund if returned at end of pregnancy	$20	$60	$80	$10
	○	○	○	○

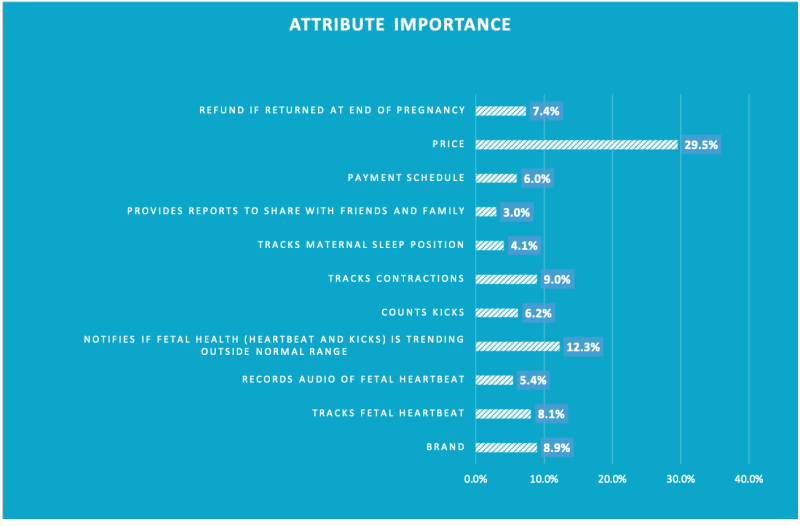
Hierarchical Models in Marketing

- Hierarchical Bayesian choice models enable both individual and aggregate-level preference parameter estimation
- This is possible even in the presence of few observations per individual by **sharing information** across individuals through an upper-level model

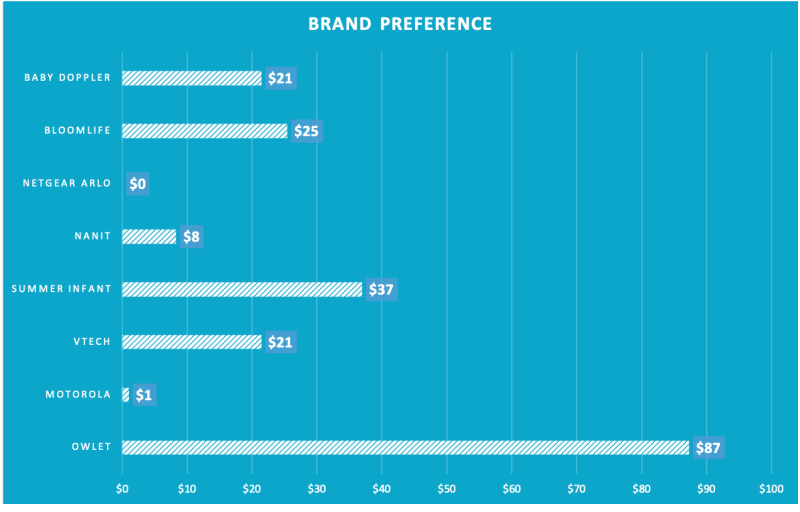
$$Pr(y_{h,j} = k|\beta_h) = \frac{\exp(x'_{k,j}\beta_h)}{\sum_{k=1}^{K}\exp(x'_{k,j}\beta_h)}$$

$$\beta_h = \Gamma z_h + \xi_h, \quad \xi_h \sim N(0, V_\beta)$$

Conjoint Analysis Example - Attribute Importance



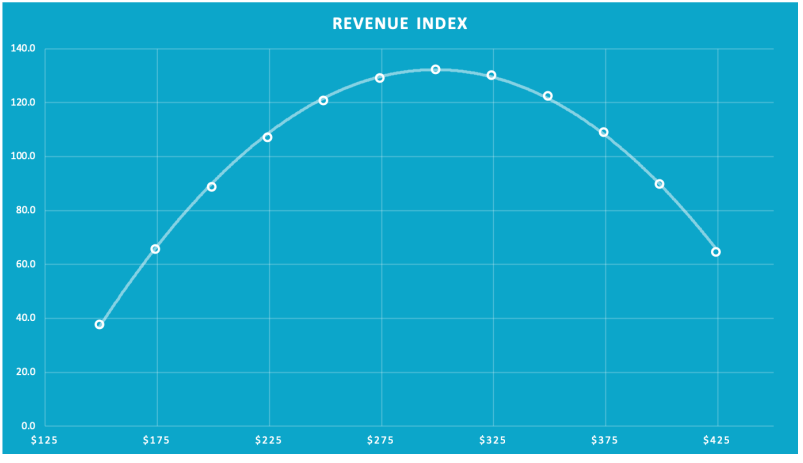
Conjoint Analysis Example - Brand Premium



Forecasting Demand with Market Simulators

- **Market simulators** make use of individual-level preference estimates to forecast the demand for **potential products** compared with the competition
- This can inform **new product development**, product line optimization, and go-no-go decision-making

Conjoint Analysis Example - Optimal Price



Motivation: Ensemble Approaches to Prediction

1. Ensemble-based approaches currently dominate the world of **competitive prediction** (e.g., Kaggle, Netflix Prize, etc.)

Wisdom from a Kaggle Superstar

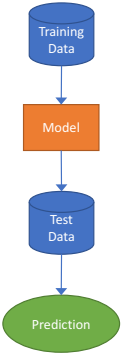
No matter how faithful and well-tuned your individual models are, you are likely to *improve the accuracy with ensembling*. Ensembling works best when the individual models are less correlated. Throwing a multitude of mediocre models into a blender can be counterproductive. Combining a few *well-constructed* models is likely to work better.

Motivation: Ensemble Approaches to Conjoint Prediction

1. Ensemble-based approaches currently dominate the world of competitive prediction (e.g., Kaggle, Netflix Prize, etc.)
2. With a few notable exceptions, most conjoint projects are still conducted using a **single model** for both inference and simulation

The Current Approach to Prediction

1. Training dataset
2. Estimate HBMNL
3. Simulate outcomes using new data
4. Generate predictions



Motivation: Ensemble Approaches to Conjoint Prediction

1. Ensemble-based approaches currently dominate the world of competitive prediction (e.g., Kaggle, Netflix Prize, etc.)
2. With a few notable exceptions, most conjoint projects are still conducted using a single model for both inference and simulation
3. A notable exception is Kevin Lattery's Sawtooth 2015 presentation and paper, "A Machine Learning Approach to Conjoint"

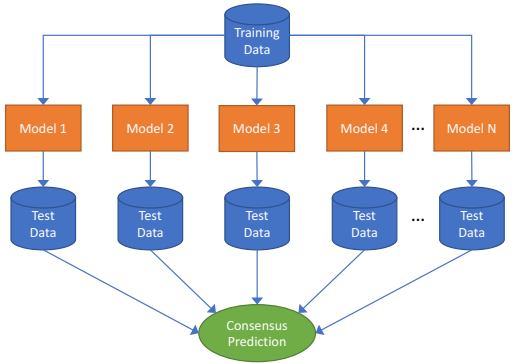
An Ensemble Approach to Conjoint Prediction

1. Training dataset

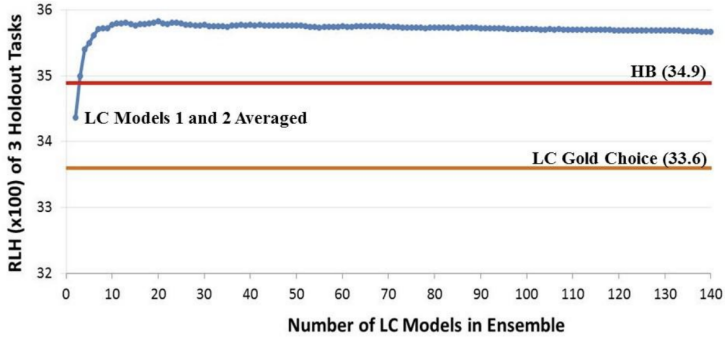
2. Estimate different types of models

3. Simulate outcomes on new data

4. Generate a consensus prediction



Results from Kevin Lattery's 2015 Paper



Why Do Ensembles Perform So Well?

- They strike an optimal trade-off on the bias/variance frontier
  - Wisdom of the Crowd
- They allow for efficient exploration of the model space
  - All models are wrong, but **some** models are useful
- They provide a hedge against **data conditions that are pathological** with respect to a single model (our hypothesis)

What is a Data Pathology?

Pathology: Structural and functional **deviations from the normal** that constitute disease or characterize a particular disease

Data Pathology: Any **respondent-level behavior** that deviates from the **normative model** of consumer choice, thus impeding accurate out-of-sample prediction

What is the Normative Model of Consumer Choice? (In Math)

For a given individual and choice task:

$$y \in \{1, 2, \cdots, J\}$$

$$U_j = V_j + \varepsilon_j$$

$$V_j = \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_k x_{kj}$$

$$\varepsilon_j \sim \text{iid Gumbel}$$

Across respondents:

$$\beta \sim N\left(\bar{\beta}, \Sigma\right)$$

What is the Normative Model of Consumer Choice? (In Words)

- Consumers are able to assess the "utility" of each alternative in a choice set
- They pick the alternative that provides the greatest level of utility
- Utility is made up of two components: A deterministic component and a random (to the researcher) component
- The deterministic component is expressed as a (linear, compensatory) function of the design of the alternative
- The random component is assumed to come from an independent and identically distributed Gumbel distribution
- While we estimate preferences at the individual level, we assume that the preferences of all individuals are drawn from a common Multivariate Normal Distribution

Models of Data Pathologies

- Most innovations in the academic conjoint literiture focus on addressing **specific data pathologies**
- These models fit better and provide marginal improvements in predictive fit
- Although prospectively useful, these models are **rarely used in practice**

Examples of Data Pathologies

- Attribute Non-Attendance
- Screening Rules and Other Forms of Non-Compensatory Choice
- Poor Respondent Quality
- Non-IIA Choice Behavior
- Respondent Fatigue
- Alternative Decision Rules (e.g., Regret Minimization)

Why Aren't These Models Used in Practice?

1. They are theoretically and computationally complex (i.e., difficult to understand and time-consuming to estimate and simulate)
2. We don't have high-quality commercial software that can be used to fit these models
3. Each model deals with a single data pathology
   - A priori how do I know which pathology will prove problematic (i.e., which model should I fit)?
   - What if multiple pathologies are present in a single dataset?

Our Proposed Approach

- Create an ensemble of models where **diversity** in the ensemble is created through various forms of **(clever) randomization**
- Induce diversity in the ensemble by thinking about how **specific forms of randomization** are related to **specific data pathologies**
- Generate a **consensus prediction** from the set of models in the ensemble

Note: Our focus in this study is to
improve **out-of-sample predictive validity**!

A Summary Thus Far...

- Conjoint analysis is a powerful tool
- Decisions are informed based on predictions using a market simulator
- Ensembles are cool and powerful for prediction
- We don't use them in the world of conjoint
- Data pathologies can lead to bad inference and prediction
- Cleverly designed ensembles of models might help hedge against data pathologies, thus improving inference
  - We don't have to fit complicated models, but we do need to fit a lot of models!
  - Trade-off between model complexity and computational intensity

Pathology 1: Attribute Non-Attendance (ANA)

Attribute Non-Attendance: Respondents ignore subsets of attributes when making decisions (i.e., $\beta = 0$ for all levels of the attribute)

Randomization Strategy for the Ensemble: Create an ensemble where we randomly set the coefficients for a full attribute to 0

Formal Model for Attribute Non-Attendance

Standard model for each individual h in a choice task:

$$U_{jh} = \beta_{1h} x_{1j} + \beta_{2h} x_{2j} + \beta_{kh} x_{kj} + \varepsilon_j \quad \text{and} \quad \varepsilon_j \sim \text{iid Gumbel}$$

Respondent h's attendance to attribute m is given by

$$\tau_{hm} \in \{1, c\}$$

Across respondents:

$$\beta_h \sim N\left(C_{\tau h}\, \bar{\beta}, C_{\tau h}\, \Sigma\, C_{\tau h}\right) \quad \text{with} \quad C_{\tau h} = \text{diag}(\tau_h)$$

$$Pr\left(\tau_{hm} = c\right) = \theta_m \quad \text{with} \quad \theta_m \sim \text{beta}(\alpha, \beta)$$

Ensemble Algorithm for Attribute Non-Attendance

1. Estimate an HB MNL on training data
2. Loop over respondent-level estimates ($\hat{\beta}_h$)
   - randomly select an attribute
   - with probability θ set all coefficients for that attribute to 0
3. Predict first choices (e.g., max utility) for each choice set in the out-of-sample holdout (test) data
4. Repeat 2 and 3 many times
5. Generate a consensus (e.g., most commonly selected) prediction

Pathology 2: Screening Behavior

Screening Behavior: Respondents use certain attribute levels to screen out alternatives from consideration (i.e., $\beta \approx -\infty$ for levels being screened on)

Randomization Strategy for the Ensemble: Create an ensemble where we randomly set the coefficients for attribute levels to $\approx -\infty$

Formal Model for Screening Behavior

For a given individual and choice task:

$$Pr(j) = Pr\left([\beta x_j + \epsilon_j] \cdot I_j > [\beta x_k + \epsilon_k] \cdot I_k \text{ for all } k \neq j\right)$$

where for each attribute level:

$$I_i = \prod_{l=1}^{L} \left[1 - \tau_l \cdot s_l\right]$$

Ensemble Algorithm for Screening Behavior

1. Estimate HB MNL on training data
2. Loop over respondent-level estimates ($\hat{\beta}_h$)
   - randomly select an attribute level
   - with probability ψ set the coefficient for that attribute level to $\approx -\infty$
3. Predict first choices (e.g., max utility) for each choice set in the validation data
4. Repeat 2 and 3 many times
5. Generate a consensus (e.g., most commonly selected) prediction

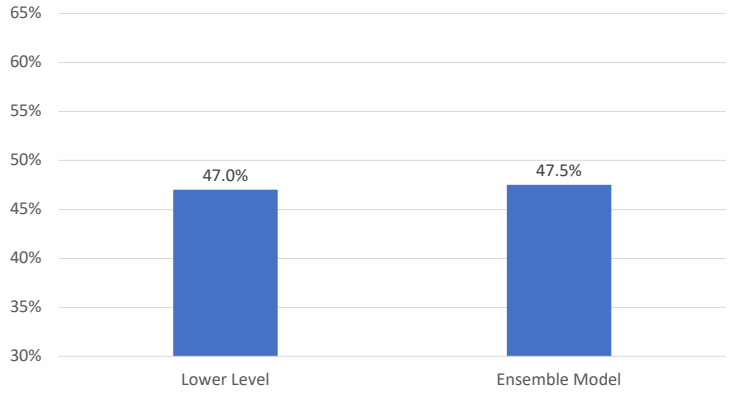
Simulation Study

- Simulate data with the presence or absence of pathologies: attribute non-attendance and screening
- 500 respondents; 12 choice tasks; 6 attributes
- Generate prediction from standard HBMNL results (lower level) and the joint ANA + Screening Ensemble

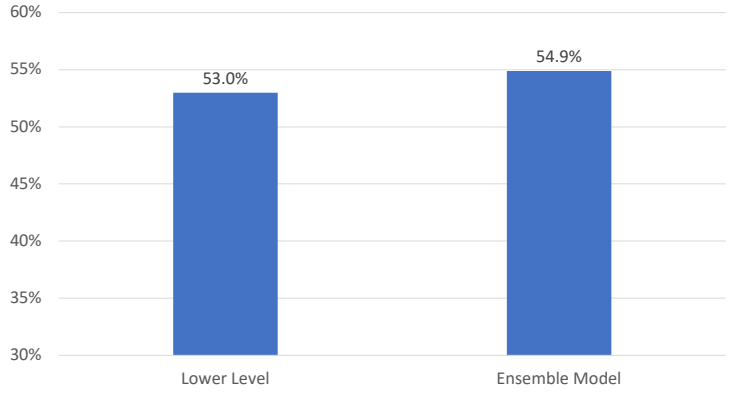
Joint Ensemble Algorithm for Attribute Non-Attendance and Screening

1. Estimate HB MNL on training data
2. Loop over respondent-level estimates ($\hat{\beta}_h$)
   - randomly select an attribute level
   - with probability ψ set the coefficient for that attribute level to $\approx -\infty$
   - randomly select an attribute
   - with probability θ set all coefficients for that attribute to 0
3. Predict first choices (e.g., max utility) for each choice set in the validation data
4. Repeat 2 and 3 many times
5. Generate a consensus (e.g., most commonly selected) prediction

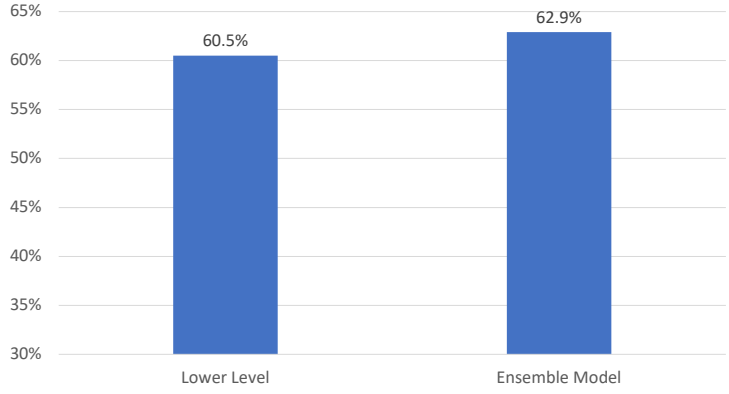
Simulated Data: No ANA + No Screening



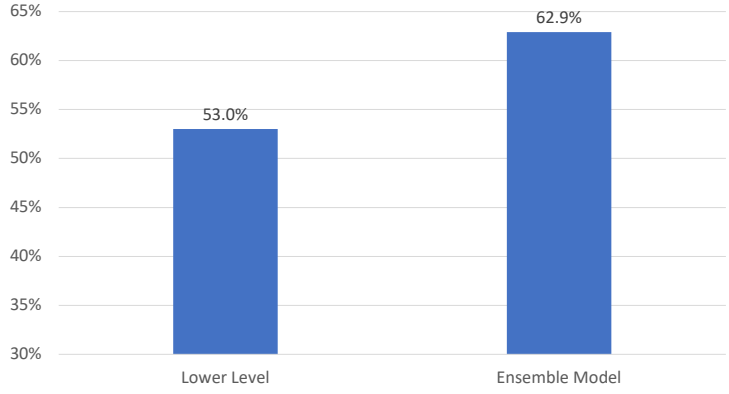
Simulated Data: ANA without Screening



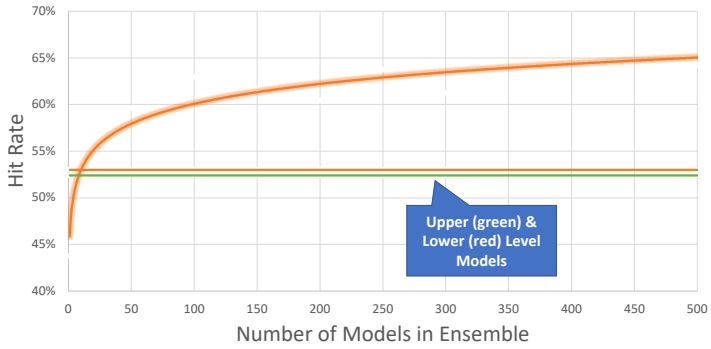
Simulated Data: Screening without ANA



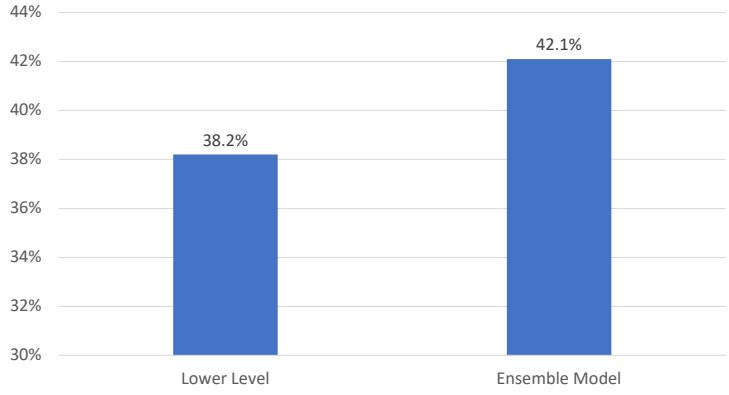
Simulated Data: ANA + Screening



Predictive Fit as a Function of Ensemble Size



Performance on Actual Data – Interior Paint



Conclusion

- Cleverly designed ensembles of models appear to help hedge against data pathologies and improve prediction
- The ensemble approach is especially powerful in the presence of multiple data pathologies

Next Steps

1. Faster computation (Stan)
2. More pathologies
3. Smarter ensembles: Can we employ principles of experimental design to (cleverly) cover the design space of multiple pathologies?
4. Better aggregation of prediction (e.g., model stacking via a meta-learner rather than consensus predictions)
5. Ensembles for inference: Can we learn about the severity of a particular pathology by examining the relative fit of ensemble members?

Thank You!