Separating Predicted Randomness from Residual Behavior

Jose Apesteguia and Miguel A. Ballester

Universitat Pompeu Fabra and University of Oxford

March 6, 2020

Introduction

- Choice data has often a probabilistic nature
  - Individual behavior has been shown to be stochastic
  - The result of aggregating heterogenous individual behavior

Introduction

- An analyst has a description of actual random behavior. Consider the ideal case where the dataset is statistically error-free.

Introduction

- An analyst has a description of actual random behavior. Consider the ideal case where the dataset is statistically error-free.
- The analyst aims at understanding the data from the perspective of her favorite stochastic choice model: predicted randomness

Introduction

- An analyst has a description of actual random behavior. Consider the ideal case where the dataset is statistically error-free.
- The analyst aims at understanding the data from the perspective of her favorite stochastic choice model: predicted randomness
- Empirically observed choice data is expected to deviate from the predictions of any stochastic choice model

Introduction

- An analyst has a description of actual random behavior. Consider the ideal case where the dataset is statistically error-free.

- The analyst aims at understanding the data from the perspective of her favorite stochastic choice model: predicted randomness

- Empirically observed choice data is expected to deviate from the predictions of any stochastic choice model

- If the model captures behavior sufficiently well, the analyst may be willing to sacrifice a perfect account of actual behavior in favor of the simpler explanation of choice provided by the imperfect stochastic model
    - This is so because the model may provide tractability, has sound behavioral foundations, facilitates a simple welfare analysis and allows prediction exercises in related environments
- Hence, even without statistical error, there is randomness in the data that is not predicted by the model, and therefore its

The Aim of this Paper is to:

- To present a methodology for separating the data consistent with the stochastic choice model (predicted randomness) from that which falls outside the model (residual behavior)

- We partition the data into two parts
  1. one representing the predictions of the model, in which a particular specification of the structured model is identified,
  2. and the remainder representing residual behavior, where a specification of unstructured residual behavior is singled-out

- We aim to maximize the portion that represents predicted randomness.

The Aim of this Paper is to:

- To present a methodology for separating the data consistent with the stochastic choice model (predicted randomness) from that which falls outside the model (residual behavior)

- We partition the data into two parts
  1. one representing the predictions of the model, in which a particular specification of the structured model is identified,
  2. and the remainder representing residual behavior, where a specification of unstructured residual behavior is singled-out

- We aim to maximize the portion that represents predicted randomness.

- This exercise provides us with three key elements:
  1. the maximal fraction of data explained by the model
  2. a particular specification of the model, and
  3. a description of the residual behavior.

Example

	x	y	z
$\{x, y, z\}$	.15	.6	.25
$\{x, y\}$	.25	.75	
$\{x, z\}$	.7		.3
$\{y, z\}$		.4	.6

- The Luce Model: $\delta_u(a, A) = \frac{u(a)}{\sum_{b \in A} u(b)}$
  - $\{x, y\} \Rightarrow u(y) > u(x)$; $\{x, z\} \Rightarrow u(x) > u(z)$; $\{y, z\} \Rightarrow u(z) > u(y)$. Impossible!
- Questions:
  - What is the fraction of the population that could be understood á la Luce?
  - Which Luce (which u)?
  - Which residual behavior?

Example: Preview

- What is the fraction of the population that could be understood á la Luce? .6
- Which Luce? $u = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$
- Which residual behavior?

	Data			*Luce*			Residual		
	x	y	z	x	y	z	x	y	z
$\{x, y, z\}$	.15	.6	.25	.25	.5	.25	0	.75	.25
$\{x, y\}$	.25	.75		.33	.66		.125	.875	
$\{x, z\}$	.7		.3	.5		.5	1		0
$\{y, z\}$		.4	.6		.66	.33		0	1

Outline

1. Framework and basic result
2. Discussion
3. Study of particular models of choice
4. Empirical application

1. Framework and basic result

- Non-empty finite set of alternatives X
- Arbitrary domain of menus $\mathcal{D}$
  - Observations: $\mathcal{O} = \{(a, A) : a \in A \in \mathcal{D}\}$
- $\texttt{SCF}$: maps from $\mathcal{O}$ to $[0, 1]$ such that $\sum_{a \in A} \sigma(a, A) = 1$
  - Data: a stochastic choice function ρ in the interior of $\texttt{SCF}$
  - Models: Any non-empty closed subset Δ of $\texttt{SCF}$

Separations

Primitives: data ρ and model Δ

Question: which are the specifications $\delta \in \Delta$ and residual behaviour $\epsilon \in \texttt{SCF}$, such that when combined at fraction $\lambda \in [0, 1]$ represent a separation of data ρ?

$$\rho = \lambda\delta + (1 - \lambda)\epsilon$$

In this case, we say that the triple $\langle \lambda, \delta, \epsilon \rangle \in [0, 1] \times \Delta \times \texttt{SCF}$ is a separation of data ρ

- $\mathcal{S}_\Delta$: the set of all separations
- Maximal Separation (largest λ)

Maximal Separations

- Existence: We have assumed closed, and given the nature of stochastic choice functions, compact models Δ. A basic argument using the continuity of λ guarantees existence.
- Characterization: can we identify maximal separations?

Main result

Proposition 1.

- $\lambda^* = \max_{\delta \in \Delta} \min_{(a,A) \in \mathcal{O}} \frac{\rho(a,A)}{\delta(a,A)}$

- $\delta^* = \arg\max_{\delta \in \Delta} \min_{(a,A) \in \mathcal{O}} \frac{\rho(a,A)}{\delta(a,A)}$

- $\epsilon^* = \frac{\rho - \lambda^* \delta^*}{1 - \lambda^*}$

Proof

Sketch of the proof:

- ρ must be in the segment connecting δ and ϵ
- $\langle 0, \delta, \rho \rangle$ is a separation
- When λ grows, ϵ must depart from ρ in the opposite direction of δ, reaching eventually the frontier of SCF. This happens at the frontier $\epsilon(a, A) = 0$.
- Since $\epsilon = 0$ is equivalent to $\lambda = \frac{\rho}{\delta}$, the frontier is reached for these observations that minimize the ratio $\frac{\rho(a,A)}{\delta(a,A)}$.
  - Data in these observations is too scarce with respect to the predictions of the model
- These are the **critical observations** $\mathcal{O}_\delta$

2. Discussion

- Maximum likelihood
- Model selection
- Convex models
- Constrained residual behavior

Maximum Likelihood (ML)

- ML identifies model instances, but does not quantify the fraction of the data explained, nor does it characterises the nature of the data that falls outside the model
  - Therefore, we must limit the comparison to the model instances identified by ML and MS
- Loss functions
  - $L_{MS}(\delta, \rho) = \max_{(a,A)\in\mathcal{O}} [1 - \frac{\rho(a,A)}{\delta(a,A)}]$
  - $L_{ML}(\delta, \rho) = \sum_{(a,A)\in\mathcal{O}} \rho(a, A) \log \frac{\rho(a,A)}{\delta(a,A)}$

Model Selection

Overfitting. Bigger models are better simply because they are bigger: $\Delta \subseteq \Delta'$

- Models have usually zero Lebesgue measure
- Worst-Best normalization usually ineffective $\frac{\lambda^* - \lambda^{\mathsf{min}}}{\lambda^{\mathsf{max}} - \lambda^{\mathsf{min}}} = \lambda^*$
- Dimension; Parameters

Convexity

If Δ is convex, the set of separations can be proved to be convex. Convex optimization problem

However, models are not usually convex (Deterministic, Luce). Mixture models (RUM)

Constrained residual behavior

Residual behavior is extreme in critical observations. Maybe too extreme?

$\Delta \cup \{\rho\} \subseteq \mathtt{RB} \subseteq \mathtt{SCF}$

Reformulate the concept of separation: $\langle \lambda, \delta, \epsilon \rangle \in [0,1] \times \Delta \times \mathtt{RB}$

3. Particular models Δ

Proposition 1 applies to any model Δ. We now study known models to reach more powerful results

- The standard, deterministic model: all randomness/heterogeneity is regarded as residual behaviour
- Three stochastic models, representing different sources of randomness:
  - Trembling: mistakes at the time of choosing
  - Luce: randomness in the utility evaluation of alternatives
  - Single-Crossing Random Utility Model: randomness in the ordinal preference

The model of Luce

- $u(x) > 0$ for all $x \in X$ and $\sum_{x \in X} u(x) = 1$

$$\delta_u(a, A) = \frac{u(a)}{\sum_{b \in A} u(b)}$$

- `Luce`: closure of this model

Proposition on Luce

Proposition. $\langle \min_{(a,A)} \frac{\rho(a,A)}{\delta_L^*(a,A)}, \delta_L^*, \frac{\rho-\lambda^*\delta_L^*}{1-\lambda^*} \rangle$ is a maximal separation for the Luce model if and only if $\mathcal{O}_{\delta_L^*}$ contains a sub-collection $\{(a_i, A_i)\}_{i=1}^I$ such that $\bigcup_{i=1}^I \{a_i\} = \bigcup_{i=1}^I A_i$.

Sketch of the proof

- Let $\mathcal{O}_{\delta_u} = \{(a_i, A_i)\}_i^I$ not be cyclical: $y \in \bigcup_{i=1}^I A_i \setminus \bigcup_{i=1}^I \{a_i\}$
- Move utilities in the direction of y: $v(\alpha) = \alpha \mathbf{1}_y + (1-\alpha)u$:
  - $\frac{\rho(a,A)}{\delta_{v(\alpha)}(a,A)}$ with $a \neq y \in A$: increase with α
  - $\frac{\rho(y,A)}{\delta_{v(\alpha)}(y,A)}$: decrease with α
  - $\frac{\rho(a,A)}{\delta_{v(\alpha)}(a,A)}$ with $y \notin A$: constant
- Minimal $\frac{\rho}{\delta_v}$ ratios increase for v, Proposition 1 implies u does not provide a maximal separation
- Only when $\bigcup_{i=1}^I A_i = \bigcup_{i=1}^I \{a_i\}$ there is no room for improvement

Luce algorithm

	x	y	z
$\{x, y, z\}$	.15	.6	.25
$\{x, y\}$	.25	.75	
$\{x, z\}$	.7		.3
$\{y, z\}$		.4	.6

Step 1: Start, e.g., with $u = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and compute the $\frac{\rho}{\delta}$ ratios. The only critical observation is $(x, \{x, y, z\})$, with ratio $\frac{.15}{1/3} = .45$

Luce algorithm

	x	y	z
$\{x, y, z\}$	.15	.6	.25
$\{x, y\}$	.25	.75	
$\{x, z\}$	.7		.3
$\{y, z\}$		.4	.6

Step 2: Critical observations do not satisfy condition Proposition 2. Select one free alternative, e.g., y. Move along the segment $v(\alpha) = \alpha(0, 1, 0) + (1 - \alpha)u$

Luce algorithm

	x	y	z
$\{x, y, z\}$	.15	.6	.25
$\{x, y\}$	.25	.75	
$\{x, z\}$	.7		.3
$\{y, z\}$		.4	.6

Step 3: How?

- $\min_{a \neq y \in A} \frac{\rho(a,A)}{\delta_{v(\alpha)}(a,A)} = \frac{\rho(x,\{x,y,z\})}{\delta_{v(\alpha)}(x,\{x,y,z\})} = \frac{.15}{\frac{1-\alpha}{3}}$ (increases with α)
- $\min_{A} \frac{\rho(y,A)}{\delta_{v(\alpha)}(y,A)} = \frac{\rho(y,\{y,z\})}{\delta_{v(\alpha)}(y,\{y,z\})} = \frac{.4}{\frac{1+2\alpha}{(2+\alpha)}}$ (decreases with α)

$\alpha^* = \frac{1}{4}$ equals the above two equations: $v = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$

Luce algorithm

	x	y	z
$\{x, y, z\}$	.15	.6	.25
$\{x, y\}$	.25	.75	
$\{x, z\}$	.7		.3
$\{y, z\}$		.4	.6

Step 4: The new critical observations, $(x, \{x, y, z\})$, $(z, \{x, z\})$ and $(y, \{y, z\})$, satisfy the condition, giving a $\frac{\rho}{\delta}$ ratio of .6

- Fraction of data explained: $\lambda_{\texttt{Luce}} = .6$
- Maximal separation:
  - δ_v with $v = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$
  - From $\rho = .6\delta_v + .4\epsilon$, we get ϵ

Luce algorithm

		ρ			δ_v			ϵ	
	x	y	z	x	y	z	x	y	z
$\{x, y, z\}$	**.15**	**.6**	**.25**	**.25**	**.5**	**.25**	**0**	**.75**	**.25**
$\{x, y\}$	**.25**	**.75**		**.33**	**.66**		**.125**	**.875**	
$\{x, z\}$	**.7**		**.3**	**.5**		**.5**	**1**		**0**
$\{y, z\}$		**.4**	**.6**		**.66**	**.33**		**0**	**1**

4. Empirical Application

- Experimental data from a different project, collected together with Syngjoo Choi at UCL in March 2013

Lotteries (equiprobable)

$l_1 = (17)$	$l_4 = (30, 10)$	$l_7 = (40, 12, 5)$
$l_2 = (50, 0)$	$l_5 = (20, 15)$	$l_8 = (30, 12, 10)$
$l_3 = (40, 5)$	$l_6 = (50, 12, 0)$	$l_9 = (20, 12, 15)$

- 87 individuals faced 108 different menus of lotteries: all 36 binary menus, and random samples of 36 menus of 3 and 5 alternatives
- Treatments NTL and TL
- Here: we focus on the binary menus, aggregate both treatments, which gives a total of 87 observations per menu overall

Maximal separations

Maximal Separations

Δ	λ	δ
Deterministic	.51	$P_{\texttt{DET}} = [l_1, l_5, l_4, l_8, l_7, l_9, l_3, l_6, l_2]$
Tremble	.68	$P_{\texttt{Tremble}} = [l_1, l_5, l_4, l_8, l_7, l_9, l_3, l_6, l_2]$; $\gamma = .51$
Luce	.74	$u = (0.22, 0.02, 0.09, 0.13, 0.25, 0.03, 0.07, 0.11, 0.08)$
SCRUM-CRRA	.78	$F(-4.15) = 0.21$, $F(-0.31) = 0.25$, $F(0.34) = 0.27$, $F(0.41) = 0.29$ $F(0.44) = 0.43$, $F(0.61) = 0.47$, $F(1) = 0.53$, $F(4) = 0.56$, $F(\infty) = 1$

SCRUM-CRRA

- CRRA expected utility preferences: $u_r(x) = \frac{x^{1-r}}{1-r}$
- Single-Crossing collection of ordinal preferences, represented by CRRA expected utility, ordered from most risk loving to most risk averse
  - $P_1 = [l_2, l_6, l_3, l_7, l_4, l_8, l_5, l_9, l_1]$ for $-\infty < r \leq -4.15$
  - $P_2 = [l_2, l_6, l_3, l_7, l_4, l_8, l_5, l_1, l_9]$ for $-4.15 < r \leq -0.52$
  - $P_3 = [l_2, l_3, l_6, l_7, l_4, l_8, l_5, l_1, l_9]$ for $-0.52 < r \leq -0.13$
  - ...
  - $P_{30} = [l_1, l_5, l_9, l_8, l_4, l_7, l_3, l_6, l_2]$ for $r \geq 4.7$
- μ: probability distribution over the set of preferences
- $\delta_\mu(a, A) = \sum_{P \in \mathcal{P} : a = m_P(A)} \mu(P)$

Empirical Application: MS vs ML

Maximal Separations and Maximal Likelihood

	Deterministic
MS	$P = [l_1, l_5, l_4, l_8, l_7, l_9, l_3, l_6, l_2]$
ML	$P = [l_1, l_5, l_4, l_8, l_7, l_9, l_3, l_6, l_2]$
	Tremble
MS	$P = [l_1, l_5, l_4, l_8, l_7, l_9, l_3, l_6, l_2]; \quad \gamma = .51$
ML	$P = [l_1, l_5, l_4, l_8, l_7, l_9, l_3, l_6, l_2]; \quad \gamma = .68$
	Luce
MS	$u = (0.22, 0.02, 0.09, 0.13, 0.25, 0.03, 0.07, 0.11, 0.08)$
ML	$u = (0.18, 0.04, 0.1, 0.14, 0.17, 0.04, 0.11, 0.13, 0.09)$
	SCRUM-CRRA
MS	$F(-4.15) = 0.21, F(-0.31) = 0.25, F(0.34) = 0.27, F(0.41) = 0.29$ $F(0.44) = 0.43, F(0.61) = 0.47, F(1) = 0.53, F(4) = 0.56, F(\infty) = 1$
ML	$F(-4.15) = 0.22, F(-0.31) = 0.29, F(0.44) = 0.44$ $F(1) = 0.50, F(-4) = 0.56, F(\infty) = 1$

MS vs ML: Prediction Exercise

- We take all the binary data except for one binary set, estimate the model instances by MS and ML using these data, and use the estimated instances to predict the behavior in the omitted binary set
- We do so for all the 36 binary sets
- For some of these binary sets, both MS and ML overestimate the probability of the same alternative in the binary menu
  - This makes comparing their ability to estimate the probabilities in this menu straightforward; one of the methods is unambiguously more accurate than the other
  - We therefore focus our comparison on these menus

Tremble				Luce				SCRUM-CRRA			
A	ρ	MS	ML	A	ρ	MS	ML	A	ρ	MS	ML
$\{l_9, l_5\}$	0.17	**0.26**	0.34	$\{l_9, l_5\}$	0.17	**0.24**	0.35	$\{l_6, l_3\}$	0.20	**0.21**	0.22
$\{l_6, l_3\}$	0.20	**0.26**	0.34	$\{l_6, l_3\}$	0.20	**0.26**	0.29	$\{l_6, l_8\}$	0.20	**0.24**	0.29
$\{l_6, l_8\}$	0.20	**0.26**	0.34	$\{l_6, l_8\}$	0.20	**0.23**	0.24	$\{l_6, l_9\}$	0.22	**0.26**	0.29
$\{l_6, l_9\}$	0.22	**0.26**	0.34	$\{l_6, l_9\}$	0.22	**0.29**	0.31	$\{l_2, l_8\}$	0.22	**0.26**	0.29
$\{l_2, l_8\}$	0.22	**0.26**	0.34	$\{l_2, l_7\}$	0.24	**0.26**	0.28	$\{l_6, l_4\}$	0.24	**0.24**	0.29
$\{l_6, l_4\}$	0.24	**0.26**	0.34	$\{l_9, l_1\}$	0.24	**0.26**	0.34	$\{l_2, l_7\}$	0.24	**0.26**	0.29
$\{l_2, l_7\}$	0.24	**0.26**	0.34	$\{l_6, l_7\}$	0.27	0.33	*0.27*	$\{l_2, l_5\}$	0.25	**0.26**	0.29
$\{l_9, l_1\}$	0.24	**0.26**	0.34	$\{l_3, l_8\}$	0.36	0.45	*0.43*	$\{l_2, l_1\}$	0.25	**0.26**	0.29
$\{l_2, l_5\}$	0.25	**0.26**	0.34	$\{l_9, l_8\}$	0.36	0.42	*0.41*	$\{l_2, l_9\}$	0.28	**0.28**	0.29
$\{l_2, l_1\}$	0.25	**0.26**	0.34	$\{l_3, l_9\}$	0.39	0.53	*0.52*	$\{l_3, l_8\}$	0.36	0.45	*0.44*
$\{l_6, l_5\}$	0.25	**0.26**	0.34	$\{l_5, l_1\}$	0.42	0.54	*0.48*	$\{l_9, l_8\}$	0.36	**0.47**	0.49
$\{l_8, l_7\}$	0.51	0.74	*0.66*	$\{l_9, l_7\}$	0.44	0.54	*0.45*	$\{l_3, l_9\}$	0.39	0.45	*0.44*
$\{l_5, l_4\}$	0.51	0.74	*0.66*	$\{l_8, l_4\}$	0.44	**0.45**	0.47	$\{l_3, l_1\}$	0.40	0.45	*0.44*
$\{l_7, l_3\}$	0.52	0.74	*0.66*	$\{l_8, l_7\}$	0.51	0.63	*0.54*	$\{l_5, l_1\}$	0.42	0.53	*0.51*
$\{l_1, l_4\}$	0.53	0.74	*0.66*	$\{l_5, l_4\}$	0.51	0.66	*0.54*	$\{l_9, l_7\}$	0.44	**0.55**	0.56
$\{l_5, l_3\}$	0.55	0.74	*0.66*	$\{l_1, l_4\}$	0.53	0.62	*0.56*	$\{l_9, l_4\}$	0.45	**0.47**	0.49
$\{l_4, l_9\}$	0.55	0.74	*0.66*	$\{l_5, l_3\}$	0.55	0.74	*0.63*	$\{l_4, l_1\}$	0.47	0.53	*0.51*
$\{l_6, l_2\}$	0.56	0.74	*0.66*	$\{l_4, l_9\}$	0.55	0.63	*0.61*	$\{l_3, l_7\}$	0.48	0.53	*0.51*
$\{l_4, l_8\}$	0.56	0.74	*0.66*	$\{l_4, l_3\}$	0.57	0.60	*0.59*	$\{l_4, l_5\}$	0.49	0.53	*0.51*
$\{l_7, l_9\}$	0.56	0.74	*0.66*	$\{l_1, l_3\}$	0.60	0.71	*0.64*	$\{l_8, l_7\}$	0.51	0.58	*0.56*
$\{l_4, l_3\}$	0.57	0.74	*0.66*	$\{l_3, l_2\}$	0.67	0.80	*0.70*	$\{l_5, l_3\}$	0.55	**0.55**	0.56
$\{l_1, l_5\}$	0.58	0.74	*0.66*	$\{l_4, l_2\}$	0.72	0.85	*0.77*	$\{l_6, l_2\}$	0.56	0.58	*0.56*
$\{l_1, l_3\}$	0.60	0.74	*0.66*	$\{l_1, l_6\}$	0.72	0.87	*0.81*	$\{l_4, l_8\}$	0.56	**0.56**	0.56
$\{l_9, l_3\}$	0.61	0.74	*0.66*	$\{l_1, l_2\}$	0.75	0.91	*0.81*	$\{l_5, l_8\}$	0.62	0.76	*0.71*
$\{l_5, l_8\}$	0.62	0.74	*0.66*	$\{l_5, l_6\}$	0.75	0.89	*0.80*	$\{l_4, l_7\}$	0.62	0.76	*0.71*
$\{l_4, l_7\}$	0.62	0.74	*0.66*	$\{l_5, l_2\}$	0.75	0.92	*0.79*	$\{l_1, l_7\}$	0.63	0.74	*0.71*
$\{l_1, l_7\}$	0.63	0.74	*0.66*	$\{l_4, l_6\}$	0.76	0.81	*0.78*	$\{l_5, l_7\}$	0.63	0.76	*0.71*
$\{l_5, l_7\}$	0.63	0.74	*0.66*					$\{l_1, l_8\}$	0.64	0.76	*0.71*
$\{l_8, l_9\}$	0.64	0.74	*0.66*					$\{l_3, l_2\}$	0.67	0.76	*0.71*
$\{l_1, l_8\}$	0.64	0.74	*0.66*					$\{l_1, l_9\}$	0.76	0.79	*0.78*
$\{l_8, l_3\}$	0.64	0.74	*0.66*					$\{l_5, l_9\}$	0.83	1.00	1.00

MS vs ML: Prediction Exercise

- The overestimation of small probabilities is less problematic for the maximal separation technique
- While maximum likelihood deals better with the overestimation of large probabilities

Conclusions

A new technique for:

- Separating predicted randomness from behaviour
- Determining the proportion of data explained by the model
- Identifying best parameters within a model
- Identifying which the limitations of a model are

Other settings

In a companion paper we analyse general settings, where X is a compact and convex subset of a vector space.

This allows us to cover cases like:

- Discrete choice: the Poisson model
- Ordered continuum: CDFs
- Cross-sectional regression: the linear case
- Time series: exponential moving average
- Game theory: the case of level-k