Separating Predicted Randomness from Noise

Jose Apesteguia
Miguel Angel Ballester

January 2018

Barcelona GSE Working Paper Series

Working Paper nº 1018

SEPARATING PREDICTED RANDOMNESS FROM NOISE*

JOSE APESTEGUIA† AND MIGUEL A. BALLESTER‡

Abstract. Given observed stochastic choice data and a model of stochastic choice, we offer a methodology that enables separation of the data representing the model's inherent randomness from residual noise, and thus quantify the maximal fraction of the data that are consistent with the model. We show how to apply our approach to any model of stochastic choice. We then study the case of four well-known models, each capturing a different notion of randomness. We conclude by illustrating our results with an experimental dataset.

Keywords: Stochastic Choice; Randomness; Noise.
JEL classification numbers: D00.

1. Introduction

Choice data often have a probabilistic nature. It has been systematically shown that individual behavior is stochastic, in that, when the same menu of options is presented repeatedly, the subjects choice varies.[1] There is also heterogeneity in individual preferences, and hence the aggregation of individual choices is often taken to be stochastic. It comes as no surprise, therefore, that some of the early research in decision theory adopted a probabilistic approach (see, e.g., Block and Marshak, 1960). Today, there is

Date: January, 2018.

*We thank Jay Lu, Paola Manzini, Marco Mariotti, Joel Sobel, and Ryan Webb for very helpful comments, and Angelo Gutierrez for outstanding research assistance. Financial support by Balliol College and "la Caixa" Foundation Research Grants on Socioeconomic Well-being (CG-2017-01) is gratefully acknowledged.

†ICREA, Universitat Pompeu Fabra and Barcelona GSE. E-mail: jose.apesteguia@upf.edu.

‡University of Oxford. E-mail: miguel.ballester@economics.ox.ac.uk.

[1] See, e.g., Mosteller and Nogee (1951) for an early experimental study and Agranov and Ortoleva (2016) for a recent one.

renewed interest in obtaining a better understanding of stochastic choice.[2] The literature offers a number of well-founded stochastic choice models incorporating randomness in various structured ways. We refer to this randomness as predicted randomness. In order to fix ideas, let us recall that in the model of Luce (1959), arguably the most influential stochastic model, predicted randomness is the result of the additive incorporation of a random variable, with an extreme type I distribution, to the utility values of the alternatives.

Since every meaningful stochastic choice model has empirical content, actual random behavior will often be inconsistent with the model. When this is the case, there is randomness in the data that is not predicted by the model, and therefore its origin is necessarily unknown to the analyst and left as unstructured. We refer to this randomness as noise or, alternatively, as residual behavior. The aim of this paper is to present, for the first time, to the best of our knowledge, a methodology for separating the data that is consistent with the stochastic choice model, i.e., the part that represents predicted randomness, from that which falls outside the model, i.e., the part that represents unknown noise.

In our approach, separating the data means partitioning it into two parts, one representing the predictions of the model, in which a particular specification of the model is identified, and the remainder representing unknown noise, where a specification of residual behavior is singled-out. Naturally, we aim to minimize the portion of the data that represents unknown noise, or, equivalently, to maximize the portion that represents predicted randomness. This exercise provides us with three key elements: namely, the maximal fraction of data explained by the model; a particular specification of the model; and a description of the residual behavior. We argue that these three elements constitute an important fund of information about the relationship between the data and the model. Firstly, the maximal fraction of data explained by the model indicates how accurately the model predicts the data. That is, the exercise provides a measure of the ability of the model to explain actual behavior. Secondly, the particular specification of the model identified in the maximal separation is another

[2] Recently published papers include Gul and Pesendorfer (2006), Dickhaut, Rustichini and Smith (2009), Caplin, Dean and Martin (2011), Ahn and Sarver (2013), Gul, Natenzon and Pesendorfer (2014), Manzini and Mariotti (2014), Fudenberg and Strzalecki (2015), Fudenberg, Iijima and Strzalecki (2015), Barseghyan, Molinari and O'Donoghue (2016), Brady and Rehbeck (2016), Caplin and Dean (2016), Apesteguia, Ballester and Lu (2017), Apesteguia and Ballester (forthcoming), Natenzon (forthcoming) and Webb (forthcoming).

potentially useful tool, when associated with large fractions of data explained, e.g., in counterfactual scenarios, such as those associated with prediction problems. Thirdly, the residual data identified in the maximal separation facilitates a better understanding of actual behavior. It endogenously enables identification of the menus and choices for which the model fails most dramatically, thereby delineating the source of inconsistent perturbations.

More formally, we consider the case of an analyst who has data in the form of a stochastic choice function ρ. In other words, the data show the probabilities of each alternative being chosen from the available menus. The aim of the analyst is to explain the data in the light of a given stochastic choice model Δ, which we define as a collection of stochastic choice functions. In the case of the Luce model, for instance, the set Δ contains the stochastic choice functions that arise from the different possible combinations of utility evaluations and extreme type I probability distributions. We define a separation as a pair $\langle \delta, \epsilon \rangle$, where δ is an instance of model Δ representing predicted randomness, and ϵ is an instance of the entire set of stochastic choice functions representing noise, such that the data ρ are the result of the convex combination of δ and ϵ. In other words, $\rho = \lambda_{\langle \delta, \epsilon \rangle}\delta + (1 - \lambda_{\langle \delta, \epsilon \rangle})\epsilon$, where the weight $\lambda_{\langle \delta, \epsilon \rangle}$ represents the fraction of the data that can be explained by the instance of model δ.

In section 2 we show how to implement our approach for the general case of any model of stochastic choice behavior. Let us first discuss how to obtain the maximal fraction of the data ρ that is explained by a particular instance δ of model Δ. We show that this exercise requires us to focus on the minimum ratio of ρ to δ within the data, that is, $\min_{(a,A)} \frac{\rho(a,A)}{\delta(a,A)}$, where $\rho(a, A)$ and $\delta(a, A)$ denote, respectively, the observed and predicted probabilities of choosing alternative a from menu A. The pairs (a, A) which minimize this ratio, and which we call critical observations, are those for which δ fails most severely; or, to put it more intuitively, those which are least consistent with the predictions of δ. Now, when considering model Δ as a whole, one merely needs to consider its best instance, i.e., the one that maximizes the ratio on the critical observations. That is, the instance of the model identified in the maximal separation is $\arg\max_{\delta \in \Delta} \min_{(a,A)} \frac{\rho(a,A)}{\delta(a,A)}$ and the maximal fraction of the data explained by the model is $\max_{\delta \in \Delta} \min_{(a,A)} \frac{\rho(a,A)}{\delta(a,A)}$. The residual stochastic choice function identified in the maximal separation follows immediately from these two elements. This is a simple method, applicable to any model, and potentially instrumental in the analysis of particular models, as will be shown in the following sections of the paper.

In section 3.1 we analyze the paradigmatic model of decision-making in economics: the deterministic choice model. In this model, the individual always selects the alternative that maximizes a preference relation, and hence there is no predicted randomness whatsoever. Thus, when a stochastic choice function is judged at the light of the deterministic model, any stochasticity in the data is regarded as noise. Given the overwhelming use of this model, it seems advisable to make it the first in our analysis of particular cases. Proposition 2 provides a simple recursive method over the sizes of the menus used to compute the maximal separations of the deterministic model.

We then turn to the study of three well-known stochastic choice models, each incorporating a different form of randomness. We start with the tremble model, where randomness represents the possibility of making mistakes at the time of choosing. In the tremble model, with probability $(1-\gamma)$ the decision-maker maximizes a preference relation, and with probability γ randomizes over all the available alternatives. Proposition 3 describes how to extend the results of the deterministic model to this case. We then analyze the model developed by Luce, which is also known as the logistic model. As mentioned, the Luce model incorporates randomness in the utility evaluation of the alternatives. Proposition 4 gives simplicity to the analysis of the Luce model by characterizing the structure of critical observations in the maximal separations. Finally, we study a class of random utility models incorporating randomness in the determination of the ordinal preference that governs choice. In particular, we study the class of single-crossing random utility models, that has the advantage of providing tractability, while also being applicable to a wide variety of important economic settings. Proposition 5 gives the corresponding maximal separations, following a recursive argument over the collections of preferences in support of the random utility model. In all three cases, our proofs provide algorithms to implement the maximal separation technique.

Section 4 reports on an empirical application of our approach. We use a previously-existing experimental dataset comprising 87 individuals making choices from binary comparisons of lotteries. We take the aggregate data of the entire population and illustrate the practicality of our results, obtaining the maximal separation results for all the models discussed in the paper. We first show that the fraction of the data explained by the deterministic model is .51, and that the preference relation identified in the maximal separation basically ranks the lotteries from least to most risky. The tremble model identifies exactly the same preference relation, together with a tremble probability of .51, which increases the fraction of data explained to .68. The Luce model also increases the fraction of data explained to .74, and identifies a utility function over lotteries that is ordinally close to the preference ranking of the deterministic and tremble models. Finally, we implement the single-crossing random utility model, assuming the utility functions given by CRRA expected utility. We obtain that the fraction of data explained increases further to .78, with the largest mass again being assigned to very high levels of risk aversion. In addition, we compare the instances identified by the maximal separation technique with other standard estimation techniques, such as maximum likelihood or least squares.[3] The empirical exercise neatly reveals interesting complementarities between maximal separation techniques and the standard ones for gaining a deeper understanding of the data.

Section 5 closes the main body of the paper by discussing several possible avenues of research that could be explored with the methodology we propose in this paper.

2. Maximal separations

Let X be a non-empty finite set of alternatives. Menus are non-empty subsets of alternatives and, in order to accommodate the diversity of existing settings, such as consumer-type domains or laboratory-type domains, we consider a non-empty arbitrary domain of menus $\mathcal{D}$. Pairs (a, A), with $a \in A$ and $A \in \mathcal{D}$ are called observations, and denoted by $\mathcal{O}$. A stochastic choice function is a mapping $\sigma : \mathcal{O} \to [0, 1]$ which, for every $A \in \mathcal{D}$, satisfies that $\sum_{a \in A} \sigma(a, A) = 1$. We interpret $\sigma(a, A)$ as the probability of choosing alternative a in menu A. We denote by $\texttt{SCF}$ the space of all stochastic choice functions. The data are represented by means of a stochastic choice function, that we denote by ρ and that we assume to be in the interior of $\texttt{SCF}$. Namely, $\rho(a, A) > 0$ for every $(a, A) \in \mathcal{O}$.[4] A model is a non-empty closed subset Δ of $\texttt{SCF}$ and an instance of the model is denoted by $\delta \in \Delta$.

We say that $\langle \delta, \epsilon \rangle \in \Delta \times$ $\texttt{SCF}$ is a separation of data ρ, whenever $\rho = \lambda_{\langle \delta, \epsilon \rangle} \delta + (1 - \lambda_{\langle \delta, \epsilon \rangle}) \epsilon$ for some $\lambda_{\langle \delta, \epsilon \rangle} \in [0, 1]$. We denote the set of all separations by $\mathcal{S}_\Delta$. In a separation, we write ρ as a convex combination of the stochastic choice function

[3] Appendix A formally compares these techniques. In addition, it discusses how the maximal separation approach compares with inconsistency indices. We stress therein that neither maximum likelihood, nor ordinary least squares techniques, nor inconsistency indices are suitable for addressing the issue of concern in this paper, namely the maximal separation of the part of the data representing predicted randomness, from that representing unknown noise.

[4] For expositional convenience, we assume ρ in the interior of $\texttt{SCF}$. The case of ρ in the boundary of $\texttt{SCF}$ can be trivially dealt with.

δ, which contains randomness consistent with model Δ, and the stochastic choice function ϵ, which contains the residual noise, with $\lambda_{\langle\delta,\epsilon\rangle}$ being the fraction of data explained in the separation. Notice that $\lambda_{\langle\delta,\epsilon\rangle}$ is uniquely determined by the separation $\langle\delta,\epsilon\rangle \in \mathcal{S}_\Delta$, except for the trivial case of $\delta = \epsilon = \rho$, for which any value in $[0,1]$ is possible. For convenience, in the latter case we set $\lambda_{\langle\rho,\rho\rangle} = 1$. We are particularly interested in the largest possible fraction of data explained by model Δ. Hence, we define maximal separations and the maximal fraction of data explained by model Δ as $\overline{\mathcal{S}}_\Delta = \arg\max\limits_{\langle\delta,\epsilon\rangle\in\mathcal{S}_\Delta} \lambda_{\langle\delta,\epsilon\rangle}$ and $\overline{\lambda}_\Delta = \max\limits_{\langle\delta,\epsilon\rangle\in\mathcal{S}_\Delta} \lambda_{\langle\delta,\epsilon\rangle}$, respectively.

The following proposition facilitates the computation of $\overline{\mathcal{S}}_\Delta$ and $\overline{\lambda}_\Delta$. In turn, it also shows the existence of maximal separations and, consequently, that the maximal fraction of data explained by the model is always well-defined. Notice that, when the value $\overline{\lambda}_\Delta$ is known, the description of maximal separations simply requires us to characterize one of the components (either the predicted randomness or the residual noise). To simplify the text, we focus on the first component of maximal separations, i.e., that of predicted randomness, henceforth denoted by $\overline{\mathcal{S}}^1_\Delta$.[5]

Proposition 1. $\overline{\mathcal{S}}^1_\Delta = \arg\max\limits_{\delta\in\Delta} \min\limits_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta(a,A)}$ *and* $\overline{\lambda}_\Delta = \max\limits_{\delta\in\Delta} \min\limits_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta(a,A)}$.

Proof of Proposition 1: Consider first the case where $\rho \in \Delta$. Then $\langle\rho,\rho\rangle \in \mathcal{S}_\Delta$, with $\lambda_{\langle\rho,\rho\rangle} = 1$. Given that $\min\limits_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta(a,A)} = 1$ if and only if $\rho = \delta$, the result follows. Let us now consider the case of $\rho \notin \Delta$. Fix $\delta \in \Delta$ and $\lambda \in [0,1)$. We claim that there exists $\langle\delta,\epsilon\rangle \in \mathcal{S}_\Delta$ for which $\lambda_{\langle\delta,\epsilon\rangle} = \lambda$ if and only if $\lambda \leq \min\limits_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta(a,A)}$. To prove the 'only if' part, let $\langle\delta,\epsilon\rangle \in \mathcal{S}_\Delta$ with $\lambda_{\langle\delta,\epsilon\rangle} = \lambda$. Then, it is the case that $\rho = \lambda\delta + (1-\lambda)\epsilon$, or equivalently, $\frac{\rho-\lambda\delta}{1-\lambda} = \epsilon \geq 0$. This implies that $\rho - \lambda\delta \geq 0$ and, ultimately, that $\lambda \leq \frac{\rho}{\delta}$. Hence, it must be that $\lambda \leq \min\limits_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta(a,A)}$, as desired.[6] To prove the 'if' part, suppose that $\lambda \leq \min\limits_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta(a,A)}$. We now prove that $\langle\delta, \epsilon = \frac{\rho-\lambda\delta}{1-\lambda}\rangle$ is a separation such that $\lambda_{\langle\delta,\epsilon\rangle} = \lambda$. Since by assumption $\delta \in \Delta$ and by construction $\rho = \lambda_{\langle\delta,\epsilon\rangle}\delta + (1-\lambda_{\langle\delta,\epsilon\rangle})\epsilon$, we are only required to prove that $\epsilon \in \texttt{SCF}$. We begin by checking that $\epsilon(a,A) \geq 0$ holds for every $(a,A) \in \mathcal{O}$. To see this, suppose by contradiction that this is not true. Then, there would exist $(b,B) \in \mathcal{O}$ such that $\frac{\rho(b,B)-\lambda\delta(b,B)}{1-\lambda} < 0$. This would imply that $\rho(b,B) - \lambda\delta(b,B) < 0$ and hence, that

[5] In order to avoid the discussion of indeterminacy in fractions throughout the text, we set the ratio $\frac{\rho(a,A)}{0}$ to be strictly larger than any real number. This is a harmless convention, since we could simply replace the expression $\min_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta(a,A)}$ with $\min_{(a,A)\in\mathcal{O},\delta(a,A)\neq 0} \frac{\rho(a,A)}{\delta(a,A)}$.

[6] Notice that, in dividing by δ, we are using the above-mentioned convention.

$\delta(b,B)>0$, with $\frac{\rho(b,B)}{\delta(b,B)}<\lambda\leq\min_{(a,A)\in\mathcal{O}}\frac{\rho(a,A)}{\delta(a,A)}$, which is a contradiction. Finally, it is also the case that $\sum_{a\in A}\epsilon(a,A)=\sum_{a\in A}\frac{\rho(a,A)-\lambda\delta(a,A)}{1-\lambda}=\frac{1-\lambda}{1-\lambda}=1$ for every $A\in\mathcal{D}$. Therefore $\epsilon\in\texttt{SCF}$ and the claim is proved. Now, given the definition of maximal fraction, we trivially have that $\overline{\lambda}_{\{\delta\}}=\min_{(a,A)\in\mathcal{O}}\frac{\rho(a,A)}{\delta(a,A)}$. This argument immediately implies the desired results on Δ, provided that maximal separations exist.

We now show the existence of maximal separations. Given the domain, any separation $\langle\delta,\epsilon\rangle$ of ρ is a vector in $\mathbb{R}^n$, with $n=2|\mathcal{O}|$. We prove that $\mathcal{S}_\Delta$ is a closed subset of $\mathbb{R}^n$. Consider a sequence $\langle\delta_t,\epsilon_t\rangle_{t=1}^\infty$ in $\mathcal{S}_\Delta$ and suppose that this sequence converges in $\mathbb{R}^n$. Given the finite-dimensionality and the fact that Δ and $\texttt{SCF}$ are closed, we have that $\lim_t\delta_t\in\Delta$ and $\lim_t\epsilon_t\in\texttt{SCF}$. Now consider the sequence of real values $\{\lambda_{\langle\delta_t,\epsilon_t\rangle}\}_{t=1}^\infty$. This sequence must converge and $\rho=\lim_t\rho=\lim_t[\lambda_{\langle\delta_t,\epsilon_t\rangle}\delta_t+(1-\lambda_{\langle\delta_t,\epsilon_t\rangle})\epsilon_t]=\lambda_{\langle\lim_t\delta_t,\lim_t\epsilon_t\rangle}\lim_t\delta_t+(1-\lambda_{\langle\lim_t\delta_t,\lim_t\epsilon_t\rangle})\lim_t\epsilon_t$, which shows that $\langle\lim_t\delta_t,\lim_t\epsilon_t\rangle$ is a separation of ρ. Thus we have proved that $\mathcal{S}_\Delta$ is closed. Also, $\mathcal{S}_\Delta$, as a subset of $[0,1]^n$, is bounded and hence, compact. It is immediate to see that, whenever $\rho\notin\Delta$, $\lambda_{\langle\delta,\epsilon\rangle}$ is a continuous function, thus rendering the existence of maximal separations and concluding the proof. ■

In order to grasp the logic implicit in Proposition 1, let us consider the non-trivial case where $\rho\notin\Delta$. Then, for a given instance of the model $\delta\in\Delta$, for $\langle\delta,\epsilon\rangle$ to be a separation of ρ, the three stochastic choice functions ρ, δ and ϵ must lie on the same line, with ρ in between δ and ϵ. We can always trivially consider the separation $\langle\delta,\rho\rangle$ with $\rho=0\delta+1\rho$. Increasing λ requires ϵ to depart from ρ in the opposite direction to that taken by δ, and hence λ will be maximal when reaching the frontier of $\texttt{SCF}$. Indeed, in the latter case, we need only consider frontier observations (a,A) for which $\epsilon(a,A)=0$, i.e., with $\rho(a,A)<\delta(a,A)$ or, equivalently, $\frac{\rho(a,A)}{\delta(a,A)}<1$. This is because, if $\epsilon(a,A)=1$ for some observation, we must also have that $\epsilon(b,A)=0$ for any other alternative $b\in A\setminus\{a\}$. Trivially, condition $\epsilon=0$ is equivalent to $\lambda=\frac{\rho}{\delta}$ and hence, the frontier is first reached by these observations which minimize the ratio $\frac{\rho(a,A)}{\delta(a,A)}$. Henceforth, we will call these observations critical and denote them by $\mathcal{O}_\delta$. Obviously, $\overline{\lambda}_{\{\delta\}}=\min_{(a,A)\in\mathcal{O}}\frac{\rho(a,A)}{\delta(a,A)}$, or, equivalently, $\overline{\lambda}_{\{\delta\}}=\frac{\rho(a,A)}{\delta(a,A)}$, with $(a,A)\in\mathcal{O}_\delta$. When considering the model Δ, the result follows.

3. Particular models of choice

The previous section characterizes maximal separations for every possible model Δ. We now work with specific choice models. In each case we use Proposition 1, together with the particular structure of the model being studied, to offer tighter results on maximal separations. The models we consider are the archetypical choice models in economics, i.e. the deterministic choice model, and three stochastic choice models incorporating different forms of randomness: the tremble model, the Luce model and the single-crossing random utility model.

3.1. **Deterministic rationality.**

The standard economic decision-making model contemplates no randomness whatsoever. Behavior is deterministic and described as the outcome of the maximization of a single preference relation. Thus, in the light of the deterministic model, all behavioral randomness must be regarded as residual noise. Given the fundamental role of this model, we begin our analysis with this special, limit case. Formally, denote by $\mathcal{P}$ the collection of all strict preference relations, that is, all transitive, complete and asymmetric binary relations on X. Maximization of $P \in \mathcal{P}$ generates the deterministic rational choice function δ_P, which assigns probability one to the maximal alternative in menu A according to preference P. We denote this alternative by $m_P(A)$, i.e., $m_P(A) \in A$ and $m_P(A) P y$ for every $y \in A \setminus \{m_P(A)\}$. Denote by `DET` the model composed of all the deterministic rational choice functions.

The following result shows that the maximal fraction and the maximal separation for `DET` can be easily computed using a simple recursive structure on subdomains of the data. For presenting the result, some notation will be useful. Given a subset $S \subseteq X$, denote by $\mathcal{D}|_S = \{A \in \mathcal{D} : A \subseteq S\}$ and $\mathcal{O}|_S = \{(a, A) \in \mathcal{O} : A \subseteq S\}$ the corresponding subdomains of menus and observations involving subsets of S. Then:

Proposition 2. *Let $\{\hat{\lambda}_S\}_{S:\mathcal{D}|_S \neq \emptyset}$ and $\hat{P} \in \mathcal{P}$ satisfy*

(1) $\hat{\lambda}_S = \max_{a \in S} \min \left\{ \{\rho(a, A)\}_{(a,A) \in \mathcal{O}|_S}, \hat{\lambda}_{S \setminus \{a\}} \right\}$,

(2) $m_{\hat{P}}(S) \in \arg\max_{a \in S} \min \left\{ \{\rho(a, A)\}_{(a,A) \in \mathcal{O}|_S}, \hat{\lambda}_{S \setminus \{a\}} \right\}$.[7]

Then, $\delta_{\hat{P}} \in \overline{\mathcal{S}}^1_{\mathtt{DET}}$ and $\hat{\lambda}_X = \overline{\lambda}_{\mathtt{DET}}$.

[7] Notice that equations (1) and (2) of Proposition 2 always compute a minimum over a non-empty collection of values. This is so because the computation only takes place when $\mathcal{D}|_S$ is non-empty and, hence, either $a \in A$ for some $A \subseteq S$, or $\mathcal{D}|_{S \setminus \{a\}} \neq \emptyset$.

Proof of Proposition 2: Let $\{\hat{\lambda}_S\}_{S:\mathcal{D}|_S\neq\emptyset}$ and $\hat{P}\in\mathcal{P}$ satisfy (1) and (2). Denote, for every S such that $\mathcal{D}|_S\neq\emptyset$, by $\mathtt{DET}_{\mathcal{D}|_S}$ the deterministic rational stochastic choice functions defined over the subdomain $\mathcal{D}|_S$. Similarly, denote by $\rho|_S$ the restriction of ρ to $\mathcal{D}|_S$. We start by proving, recursively, that the maximal fraction of data $\rho|_S$ explained by model $\mathtt{DET}_{\mathcal{D}|_S}$ is equal to $\hat{\lambda}_S$. Consider any subset S for which $\mathcal{D}|_S = S$. In this case, Proposition 1 guarantees that the maximal fraction of data $\rho|_S$ explained by model $\mathtt{DET}_{\mathcal{D}|_S}$ is $\max_{\delta\in\mathtt{DET}_{\mathcal{D}|_S}}\min_{(a,A)\in\mathcal{O}|_S}\frac{\rho|_S(a,A)}{\delta(a,A)}=\max_{P\in\mathcal{P}}\min_{(a,A)\in\mathcal{O}|_S}\frac{\rho(a,A)}{\delta_P(a,A)}=\max_{P\in\mathcal{P}}\min_{a\in S}\frac{\rho(a,S)}{\delta_P(a,S)}=\max_{P\in\mathcal{P}}\frac{\rho(m_P(S),S)}{\delta_P(m_P(S),S)}=\max_{P\in\mathcal{P}}\rho(m_P(S),S)=\max_{a\in S}\rho(a,S)=\max_{a\in S}\min_{(a,A)\in\mathcal{O}|_S}\rho(a,A)=\hat{\lambda}_S$. Now suppose that $\mathcal{D}|_S\neq S$ and that the result has been proved for any strict subset of S with non-empty subdomain. For any $a\in S$, denote by $\mathcal{P}_{aS}$ the set of preferences that rank a above any other alternative in S, i.e., $\mathcal{P}_{aS}=\{P\in\mathcal{P}:a=m_P(S)\}$, and by $\mathtt{aS}$ the subset of $\mathtt{DET}_{\mathcal{D}|_S}$ generated by preferences in $\mathcal{P}_{aS}$. Trivially, $\mathtt{DET}_{\mathcal{D}|_S}=\bigcup_{a\in S}\mathtt{aS}=\bigcup_{a\in S}\bigcup_{P\in\mathcal{P}_{aS}}\{\delta_P\}$. Since the only observations for which δ_P has a non-null value are those that are in the form $(m_P(A),A)$, Proposition 1 guarantees that the maximal fraction of data $\rho|_S$ explained by model $\mathtt{DET}_{\mathcal{D}|_S}$ is $\max_{a\in S}\max_{P\in\mathcal{P}_{aS}}\min_{A\in\mathcal{D}|_S}\rho(m_P(A),A)$. Since $P\in\mathcal{P}_{aS}$, we obtain that $m_P(A)=a$ whenever $a\in A$ and hence, the latter value is equal to $\max_{a\in S}\max_{P\in\mathcal{P}_{aS}}\min\left\{\{\rho(a,A)\}_{(a,A)\in\mathcal{O}|_S},\{\rho(m_P(B),B)\}_{B\in\mathcal{D}|_{S\setminus\{a\}}}\right\}$. This can be expressed as $\max_{a\in S}\min\left\{\{\rho(a,A)\}_{(a,A)\in\mathcal{O}|_S},\max_{P\in\mathcal{P}_{aS}}\min_{B\in\mathcal{D}|_{S\setminus\{a\}}}\rho(m_P(B),B)\right\}$ or, equivalently, as $\max_{a\in S}\min\left\{\{\rho(a,A)\}_{(a,A)\in\mathcal{O}|_S},\min_{B\in\mathcal{D}|_{S\setminus\{a\}}}\max_{P\in\mathcal{P}_{aS}}\min_{C\in\mathcal{D}|_B}\rho(m_P(C),C)\right\}$. Given that $a\notin B$, it is clearly the case that $\max_{P\in\mathcal{P}_{aS}}\min_{C\in\mathcal{D}|_B}\rho(m_P(C),C)=\max_{P\in\mathcal{P}}\min_{C\in\mathcal{D}|_B}\rho(m_P(C),C)$ and, by Proposition 1 and the structure of deterministic stochastic choice functions, the latter is the maximal fraction of data $\rho|_B$ explained by model $\mathtt{DET}_{\mathcal{D}|_B}$, which is equal to $\hat{\lambda}_B$ by hypothesis. Hence, the maximal fraction of data $\rho|_S$ explained by model $\mathtt{DET}_{\mathcal{D}|_S}$ must be also equal to $\hat{\lambda}_S$, as desired. As a corollary, we have that $\overline{\lambda}_{\mathtt{DET}}=\hat{\lambda}_X$. Also, it is evident from the recursive argument that $\overline{\lambda}_{\mathtt{DET}}=\overline{\lambda}_{\{\delta_{\hat{P}}\}}$, thus concluding the proof. ■

Proposition 2 enables a recursive computation of the maximal fraction of ρ explained by $\mathtt{DET}$, and a maximal separation for $\mathtt{DET}$. The algorithm starts with subdomains such that $\mathcal{D}|_S=\{S\}$, i.e., menus for which there are no available data in proper subsets. In these menus, only the highest choice frequency of an alternative must be considered. This value corresponds to the maximal fraction of the restriction of ρ to $\mathcal{D}|_S$ explained by the deterministic model. The separation can be constructed by considering the preference relation that places the alternative with the highest choice frequency above all other alternatives. For any other subdomain $\mathcal{D}|_S$, the algorithm analyzes the alternatives $a \in S$ one by one, again considering the consequences of placing a as the maximal alternative in S. It turns out to be the case that we just need to consider the following values: (i) the choice frequencies of a in subsets of S, and (ii) the maximal fractions over the subdomains where alternative a is not present.

Table 1. A stochastic choice function ρ

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

We now illustrate Proposition 2 with the example in Table 1, where the stochastic choice function ρ is defined on every non-singleton subset of $X = \{x, y, z\}$, i.e., $\mathcal{D} = \{\{x, y, z\}, \{x, y\}, \{x, z\}, \{y, z\}\}$. We can first calculate the maximal fraction for every set for which $\mathcal{D}|_S = \{S\}$, i.e., the binary sets:

$$\hat{\lambda}_{\{x,y\}} = \max\{\rho(x, \{x, y\}), \rho(y, \{x, y\})\} = .75,$$

$$\hat{\lambda}_{\{x,z\}} = \max\{\rho(x, \{x, z\}), \rho(z, \{x, z\})\} = .7, \text{ and}$$

$$\hat{\lambda}_{\{y,z\}} = \max\{\rho(y, \{y, z\}), \rho(z, \{y, z\})\} = .6.$$

We can then proceed to assign a value to menu X, for which we first analyze the alternatives in X one-by-one. For alternative x, we compute the minimum of $\left\{\{\rho(x, \{x, y\}), \rho(x, \{x, z\}), \rho(x, X)\}, \hat{\lambda}_{\{y,z\}}\right\} = \rho(x, \{x, y, z\}) = .15$. For alternative y, the minimum of $\left\{\{\rho(y, \{x, y\}), \rho(y, \{y, z\}), \rho(y, X)\}, \hat{\lambda}_{\{x,z\}}\right\} = \rho(y, \{y, z\}) = .4$ is the relevant value. Finally, for alternative z we are required to compute the minimum of $\left\{\{\rho(z, \{x, z\}), \rho(z, \{y, z\}), \rho(z, X)\}, \hat{\lambda}_{\{x,y\}}\right\} = \rho(z, \{x, y, z\}) = .25$. Thus, we get

$$\overline{\lambda}_{\texttt{DET}} = \hat{\lambda}_X = \max\{.15, .4, .25\} = .4.$$

Notice that the last value is obtained with alternative y. In subset $X \setminus \{y\}$, the key alternative is x. Hence, the second part of Proposition 2 guarantees that $\delta_{\hat{P}}$ with $y\hat{P}x\hat{P}z$ conforms to a maximal separation of ρ. From $\overline{\lambda}_{\texttt{DET}} = .4$, one can immediately obtain the residual noise, given by $\epsilon = \frac{\rho - .4\delta_{\hat{P}}}{.6}$, i.e., $\epsilon(x, X) = \frac{1}{4}, \epsilon(y, X) = \frac{1}{3}, \epsilon(x, \{x, y\}) = \frac{5}{12}, \epsilon(x, \{x, z\}) = \frac{1}{2}$, and $\epsilon(y, \{y, z\}) = 0$. To close the discussion of this example, notice that the frontier of $\texttt{SCF}$ is reached at $\epsilon(y, \{y, z\}) = 0$. This is precisely the critical observation, which in turn determines the maximal fraction of ρ explained by $\texttt{DET}$, i.e., $\frac{\rho(y,\{y,z\})}{\delta(y,\{y,z\})} = \frac{.4}{1} = .4$.

3.2. **Tremble model.** In tremble models, behavioral randomness is interpreted as a mistake at the moment of choice. In the simplest version, the individual contemplates a preference relation P. With probability $(1 - \gamma) \in [0, 1]$, the preference is maximized. With probability γ, the individual trembles and randomizes between all the alternatives in the menu.[8] This generates the tremble choice function $\delta_{[P,\gamma]}(a, A) = \frac{\gamma}{|A|}$ whenever $a \in A \setminus \{m_P(A)\}$ and $\delta_{[P,\gamma]}(m_P(A), A) = 1 - \gamma\frac{|A|-1}{|A|}$. Denote by $\texttt{Tremble}$ the model composed of all tremble choice functions.

The next result describes the maximal fraction of data explained by $\texttt{Tremble}$ and a maximal separation for $\texttt{Tremble}$. Given the immediate connection to the rational deterministic model, the result is a direct extension of Proposition 2.

Proposition 3. *Let* $\{\hat{\lambda}_S(\gamma)\}_{S:\mathcal{D}|_S \neq \emptyset}$ *and* $\hat{P}(\gamma) \in \mathcal{P}$ *satisfy, for every* $\gamma \in [0, 1]$*:*

(1) $\hat{\lambda}_S(\gamma) = \max\limits_{a \in S} \min \Big\{ \{\frac{|A|\rho(a,A)}{(1-\gamma)|A|+\gamma}\}_{(a,A)\in\mathcal{O}|_S}, \{\frac{|A|\rho(b,A)}{\gamma}\}_{\substack{(b,A)\in\mathcal{O}|_S \\ b\neq a\in A}}, \hat{\lambda}_{S\setminus\{a\}} \Big\}$,

(2) $m_{\hat{P}(\gamma)}(S) \in \arg\max\limits_{a \in S} \min \Big\{ \{\frac{|A|\rho(a,A)}{(1-\gamma)|A|+\gamma}\}_{(a,A)\in\mathcal{O}|_S}, \{\frac{|A|\rho(b,A)}{\gamma}\}_{\substack{(b,A)\in\mathcal{O}|_S \\ b\neq a\in A}}, \hat{\lambda}_{S\setminus\{a\}} \Big\}$.

Let γ^* *maximize* $\hat{\lambda}_X(\gamma)$*. Then* $\hat{\lambda}_X(\gamma^*) = \overline{\lambda}_{\texttt{Tremble}}$ *and* $\delta_{\hat{P}}(\gamma^*) \in \overline{\mathcal{S}}^1_{\texttt{Tremble}}$.

Proof of Proposition 3: Since the proof has the same structure as the proof of Proposition 2, we are able to skip some of the steps and use the same notation as before. We start by (recursively) proving that the maximal fraction of data $\rho|_S$ explained by the collection of stochastic choice functions in $\texttt{Tremble}_{\mathcal{D}|_S}$ with a fixed degree of tremble γ, which we denote by $\texttt{Tremble}_{\mathcal{D}|_S}(\gamma)$, is equal to $\hat{\lambda}_S(\gamma)$. We start with any subset S for which $\mathcal{D}|_S = S$. The maximal fraction of data $\rho|_S$ explained by $\texttt{Tremble}_{\mathcal{D}|_S}(\gamma)$ is $\max\limits_{\delta \in \texttt{Tremble}_{\mathcal{D}|_S}(\gamma)} \min\limits_{(a,A)\in\mathcal{O}|_S} \frac{\rho|_S(a,A)}{\delta(a,A)} = \max\limits_{P\in\mathcal{P}} \min \{ \frac{\rho(m_P(S),S)}{\delta_{[P,\gamma]}(m_P(S),S)}, \{\frac{\rho(b,S)}{\delta_{[P,\gamma]}(b,S)}\}_{b\in S\setminus\{m_P(S)\}} \} = \max\limits_{P\in\mathcal{P}} \min \{ \frac{|S|\rho(m_P(S),S)}{(1-\gamma)|S|+\gamma}, \{\frac{|S|\rho(b,S)}{\gamma}\}_{b\in S\setminus\{m_P(S)\}} \} = \max\limits_{a\in S} \min \{ \frac{|S|\rho(a,S)}{(1-\gamma)|S|+\gamma}, \{\frac{|S|\rho(b,S)}{\gamma}\}_{b\in S\setminus\{a\}} \} = \max\limits_{a\in S} \min \Big\{ \{\frac{|A|\rho(a,A)}{(1-\gamma)|A|+\gamma}\}_{(a,A)\in\mathcal{O}|_S}, \{\frac{|A|\rho(b,A)}{\gamma}\}_{\substack{(b,A)\in\mathcal{O}|_S \\ b\neq a}} \} = \hat{\lambda}_S(\gamma)$. Whenever $\mathcal{D}|_S \neq S$, we

[8] See Harless and Camerer (1994) for an early treatment of the trembling-hand concept in the stochastic choice literature.

can write the maximal fraction of data $\rho|_S$ explained by model $\texttt{Tremble}_{\mathcal{D}|_S}(\gamma)$ as $\max\limits_{a\in S}\max\limits_{P\in\mathcal{P}_{aS}}\min\big\{\{\frac{|A|\rho(m_P(A),A)}{(1-\gamma)|A|+\gamma}\}_{A\in\mathcal{D}|_S},\{\frac{|A|\rho(b,A)}{\gamma}\}_{(b,A)\in\mathcal{O}|_S,b\neq m_P(A)}\big\}$. Notice that we can decompose $\{\frac{|A|\rho(m_P(A),A)}{(1-\gamma)|A|+\gamma}\}_{A\in\mathcal{D}|_S}$ into $\{\frac{|A|\rho(a,A)}{(1-\gamma)|A|+\gamma}\}_{(a,A)\in\mathcal{O}|_S}$ and $\{\frac{|B|\rho(m_P(B),B)}{(1-\gamma)|B|+\gamma}\}_{B\in\mathcal{D}|_{S\setminus\{a\}}}$. Similarly, we can decompose $\{\frac{|A|\rho(b,A)}{\gamma}\}_{(b,A)\in\mathcal{O}|_S,b\neq m_P(A)}$ into components $\{\frac{|A|\rho(b,A)}{\gamma}\}_{\substack{(b,A)\in\mathcal{O}|_S\\ b\neq a\in A}}$ and $\{\frac{|B|\rho(b,B)}{\gamma}\}_{\substack{B\in\mathcal{D}|_{S\setminus\{a\}}\\ b\neq m_P(B)}}$. By the same reasoning as in the proof of Proposition 2, consideration of both $\{\frac{|B|\rho(m_P(B),B)}{(1-\gamma)|B|+\gamma}\}_{B\in\mathcal{D}|_{S\setminus\{a\}}}$ and $\{\frac{|B|\rho(b,B)}{\gamma}\}_{\substack{B\in\mathcal{D}|_{S\setminus\{a\}}\\ b\neq m_P(B)}}$ yields the value $\{\hat{\lambda}_B(\gamma)\}_{B\in\mathcal{D}|_{S\setminus\{a\}}}$. This proves the claim. As an immediate corollary, we have that $\overline{\lambda}_{\texttt{Tremble}(\gamma)}=\hat{\lambda}_X(\gamma)$ and the results follow directly from Proposition 1. ■

We now illustrate Proposition 3 using the example given in Table 1. Replicating the steps taken in the analysis of DET, we conclude that $y\hat{P}x\hat{P}z$ is the optimal preference relation for every given value of γ. In order to find the optimal value of γ, note that there are only two possible critical observations, depending on the value of γ. When γ is low, we know from the study of the deterministic case that the critical observation is $(y,\{y,z\})$, with a ratio ρ to δ equal to $\frac{.4}{1-\gamma+\frac{\gamma}{2}}$. When γ is high the critical observation is $(x,\{x,y,z\})$, with a ratio ρ to δ equal to $\frac{.15}{\frac{\gamma}{3}}$. By noticing that the first ratio is increasing and starts at a value below the second ratio, which is decreasing, it follows that the maximal fraction of data explained by the optimal tremble can be found by equating these two ratios, which yields $\gamma^*=.72$. Hence, $\overline{\lambda}_{\texttt{Tremble}}=\overline{\lambda}_{\{\delta_{[\hat{P},.72]}\}}=.625$. Now, one can immediately obtain the residual noise, given by $\epsilon=\frac{\rho-.625\delta_{\hat{P}}}{.375}$, i.e., $\epsilon(x,X)=0,\epsilon(y,X)=\frac{11}{15},\epsilon(x,\{x,y\})=\frac{1}{15},\epsilon(x,\{x,z\})=\frac{4}{5}$, and $\epsilon(y,\{y,z\})=0$.

3.3. **Luce model.** Denote by $\mathcal{U}$ the collection of strictly positive utility functions u such that, without loss of generality, $\sum_{x\in X}u(x)=1$. Given $u\in\mathcal{U}$, a strictly positive Luce stochastic choice function is defined by $\delta_u(a,A)=\frac{u(a)}{\sum_{b\in A}u(b)}$ with $a\in A\in\mathcal{D}$.[9] In order to accommodate the Luce model in our framework we consider the closure of the set of strictly positive Luce stochastic choice functions, which we denote by Luce. To denote a generic, not necessarily strictly positive, Luce stochastic choice function, we write δ_L. However, as we prove below, there are always instances of the model of Luce identified in the maximal separations that are strictly positive, and hence, the former assumption is inconsequential.

[9] It is well known that this definition of a Luce function is equivalent to the one we use in the Introduction.

We now describe the structure of maximal separations of $\texttt{Luce}$. From Proposition 1 we know that the study of a particular instance of model δ_L requires us to analyze its critical observations $\mathcal{O}_{\delta_L}$. It turns out to be the case that, under the Luce model, we only need to check for a simple condition on the set $\mathcal{O}_{\delta_L}$.

Proposition 4. *$\delta_L \in \overline{\mathcal{S}}^1_{\texttt{Luce}}$ if and only if $\mathcal{O}_{\delta_L}$ contains a sub-collection $\{(a_i, A_i)\}_{i=1}^I$ such that $\bigcup_{i=1}^I \{a_i\} = \bigcup_{i=1}^I A_i$. Moreover, $\overline{\lambda}_{Luce} = \frac{\rho(a,A)}{\delta_L(a,A)}$ where $(a, A) \in \mathcal{O}_{\delta_L}$ and $\delta_L \in \overline{\mathcal{S}}^1_{\texttt{Luce}}$.*

Proof of Proposition 4: We prove the characterization of maximal separations of $\texttt{Luce}$, and from this the second part of the claim follows immediately. To prove the 'if' part let $\delta_L \in \texttt{Luce}$ and suppose that there exists $\{(a_i, A_i)\}_{i=1}^I \subseteq \mathcal{O}_{\delta_L}$ such that $\bigcup_{i=1}^I \{a_i\} = \bigcup_{i=1}^I A_i$. Assume, by way of contradiction, that $\delta_L \notin \overline{\mathcal{S}}^1_{\texttt{Luce}}$. By Proposition 1 and the definition of $\mathcal{O}_{\delta_L}$, there exists $\delta'_L \in \overline{\mathcal{S}}^1_{\texttt{Luce}}$ such that, for every $i \in \{1, 2, \ldots, I\}$, it is the case that $\frac{\rho(a_i,A_i)}{\delta_L(a_i,A_i)} = \min\limits_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta_L(a,A)} < \min\limits_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta'_L(a,A)} \leq \frac{\rho(a_i,A_i)}{\delta'_L(a_i,A_i)}$. For every $i \in \{1, 2, \ldots, I\}$, we have that $\rho(a_i, A_i) > 0$ and hence, since the ρ/δ_L ratio is minimized at $\mathcal{O}_{\delta_L}$, it must be that $\delta_L(a_i, A_i) > 0$, making $\frac{\rho(a_i,A_i)}{\delta_L(a_i,A_i)} < \frac{\rho(a_i,A_i)}{\delta'_L(a_i,A_i)}$ equivalent to $\delta'_L(a_i, A_i) < \delta_L(a_i, A_i)$. Let $\{\delta'_{v_n}\}_{n=1}^\infty$ and $\{\delta_{u_n}\}_{n=1}^\infty$ be two sequences of strictly positive Luce stochastic choice functions that converge to δ'_L and δ_L, respectively. Select an m sufficiently large that $\delta'_L(a_i, A_i) < \delta_{u_m}(a_i, A_i)$ holds for every $i \in \{1, 2, \ldots, I\}$. Given m, now select an m' sufficiently large that, for every $i \in \{1, 2, \ldots, I\}$, $\delta'_{v_{m'}}(a_i, A_i) < \delta_{u_m}(a_i, A_i)$ holds. We then have that $\frac{1}{\sum_{x\in A_i} \frac{v_{m'}(x)}{v_{m'}(a_i)}} = \frac{v_{m'}(a_i)}{\sum_{x\in A_i} v_{m'}(x)} = \delta'_{v_{m'}}(a_i, A_i) < \delta_{u_m}(a_i, A_i) = \frac{u_m(a_i)}{\sum_{x\in A_i} u_m(x)} = \frac{1}{\sum_{x\in A_i} \frac{u_m(x)}{u_m(a_i)}}$, thus guaranteeing, for every $i \in \{1, 2, \ldots, I\}$, the existence of one alternative $x_i^* \in A_i \setminus \{a_i\}$ such that $\frac{v_{m'}(a_i)}{v_{m'}(x_i^*)} < \frac{u_m(a_i)}{u_m(x_i^*)}$. Given that $\bigcup_{i=1}^I \{a_i\} = \bigcup_{i=1}^I A_i$, there exists a subcollection $\{a_{i_h}\}_{h=1}^H$ of $\{a_i\}_{i=1}^I$ with the following properties: (i) $a_{i_{h+1}} \in A_{i_h}$, with $h = 1, \ldots, H-1$, and $a_{i_1} \in A_{i_H}$, and (ii) $\frac{v_{m'}(a_{i_h})}{v_{m'}(a_{i_{h+1}})} < \frac{u_m(a_{i_h})}{u_m(a_{i_{h+1}})}$ with $h = 1, \ldots, H-1$ and $\frac{v_{m'}(a_{i_H})}{v_{m'}(a_{i_1})} < \frac{u_m(a_{i_H})}{u_m(a_{i_1})}$. Obviously, $1 = \frac{v_{m'}(a_{i_H})}{v_{m'}(a_{i_1})}\Pi_{h=1}^{H-1}\frac{v_{m'}(a_{i_h})}{v_{m'}(a_{i_{h+1}})} < \frac{u_m(a_{i_H})}{u_m(a_{i_1})}\Pi_{h=1}^{H-1}\frac{u_m(a_{i_h})}{u_m(a_{i_{h+1}})} = 1$, which is a contradiction. This concludes the 'if' part of the proof.

To prove the 'only if' part, suppose that $\delta_L \in \overline{\mathcal{S}}^1_{\texttt{Luce}}$. Let $[x]$ be the set of all alternatives $x' \in X$ for which there exists a sequence of observations $\{(b_j, B_j)\}_{j=1}^J$, with: (i) $x = b_1$ and $x' \equiv b_{J+1} \in B_J$, and (ii) for every $j \in \{1, 2, \ldots, J\}$, $\delta_L(b_j, B_j) > 0$ and $\delta_L(b_{j+1}, B_j) > 0$. If there is no alternative for which such a sequence exists, let $[x] = \{x\}$. Clearly, $[\cdot]$ defines equivalence classes on X. Whenever there exists $A \in \mathcal{D}$

with $\{x, y\} \subseteq A$ and $\delta_L(x, A) > \delta_L(y, A) = 0$, we write $[x] \succ [y]$. We claim that $\succ$ is an acyclic relation on the set of equivalence classes. To see this, assume, by contradiction, that there is a cycle of pairs $\{a_q, b_q\}$, menus $A_q \supseteq \{a_q, b_q\}$, and equivalence classes $[x_q]$, $q \in \{1, 2, \dots, Q\}$, such that: (i) $\delta_L(a_q, A_q) > \delta_L(b_q, A_q) = 0$ for every $q \in \{1, 2, \dots, Q\}$, (ii) $a_q \in [x_q]$ for every $q \in \{1, 2, \dots, Q\}$, and (iii) $b_q \in [x_{q+1}]$ for every $q \in \{1, 2, \dots, Q-1\}$ and $b_Q \in [x_1]$. We can then consider a sequence of stochastic choice functions $\{\delta_{u_n}\}_{n=1}^{\infty}$ that converges to δ_L. Since b_q and a_{q+1} belong to the same equivalence class $[x_{q+1}]$, either $b_q = a_{q+1}$ or there exists a sequence of observations $\{(d_j, D_j)\}_{j=1}^{J}$ with: (i) $b_q = d_1$ and $a_{q+1} = d_{J+1} \in D_J$, and (ii) for every $j \in \{1, 2, \dots, J\}$, $\delta_L(d_j, D_j) > 0$ and $\delta_L(d_{j+1}, D_j) > 0$ (and the same holds for a_Q and b_1). Define the strictly positive constant $K_q = 1$ whenever $b_q = a_{q+1}$, and $K_q = \frac{1}{2}\Pi_{j=1}^{J}\frac{\delta_L(d_j, D_j)}{\delta_L(d_{j+1}, D_j)}$ otherwise (with a similar definition for K_Q relating a_Q and b_1). If $b_q = a_{q+1}$, then trivially $u_n(b_q) = u_n(a_{q+1})$ for every n. Otherwise, for an n sufficiently large in the sequence $\{u_n\}_{n=1}^{\infty}$, we have that $\frac{u_n(b_q)}{u_n(a_{q+1})} = \Pi_{j=1}^{J}\frac{u_n(d_j)}{u_n(d_{j+1})} = \Pi_{j=1}^{J}\frac{\delta_{u_m}(d_j, D_j)}{\delta_{u_m}(d_{j+1}, D_j)} \geq K_q$. Hence, in any case, $\frac{u_n(b_q)}{K_q} \geq u_n(a_{q+1})$ holds for any sufficiently large n (and the same holds for b_Q and a_1). Also, since $\delta_L(a_q, A_q) > \delta_L(b_q, A_q) = 0$ for every $q \in \{1, 2, \dots, Q\}$, we can find an n sufficiently large that $u_n(a_q) > \frac{u_n(b_q)}{K_q}$. Hence, we can find an m that is sufficiently large that $u_m(a_1) > \frac{u_m(b_1)}{K_1} \geq u_m(a_2) > \frac{u_m(b_2)}{K_2} \geq \cdots \geq u_m(a_Q) > \frac{u_m(b_Q)}{K_Q} \geq u_m(a_1)$. This is a contradiction which proves the acyclicity of $\succ$. We can then denote the equivalence classes as $\{[x_e]\}_{e=1}^{E}$, where $[x_e] \succ [x_{e'}]$ implies that $e < e'$. For an equivalence class $[x_e]$, define the vector $u_{[x_e]} \in \mathcal{U}$ such that $u_{[x_e]}(y) = 0$ if $y \notin [x_e]$ and, $\frac{u_{[x_e]}(y)}{u_{[x_e]}(y')} = \frac{\delta_L(y, A)}{\delta_L(y', A)}$ whenever $y, y' \in [x_e]$, $\delta_L(y, A) > 0$ and $\delta_L(y', A) > 0$. This is clearly well-defined due to the structure of Luce stochastic choice functions. Now consider the sequence of Luce stochastic choice functions $\{\delta_{v_n}\}_{n=1}^{\infty}$ given by $v_n = (1 - \sum_{e=2}^{E}(\frac{1}{2^e})^n)u_{[x_1]} + \sum_{e=2}^{E}(\frac{1}{2^e})^n u_{[x_e]}$, which clearly converges to δ_L. Consider the following three collections of observations $\mathcal{O}_1$, $\mathcal{O}_2$ and $\mathcal{O}_3$. $\mathcal{O}_1$ is composed of all observations $(a, A) \in \mathcal{O}$ such that $A \subseteq [a]$. $\mathcal{O}_2$ is composed of all observations $(a, A) \in \mathcal{O} \setminus \mathcal{O}_1$, such that $b \in A$, $a \in [a_i]$ and $b \in [a_j]$ imply $i \geq j$. $\mathcal{O}_3$ is composed of observations in $\mathcal{O} \setminus (\mathcal{O}_1 \cup \mathcal{O}_2)$. Notice that, for an n sufficiently large, for every $(a, A) \in \mathcal{O}_1$ we have that $\frac{\rho(a,A)}{\delta_{v_n}(a,A)} = \frac{\rho(a,A)}{\delta_L(a,A)}$ and for every $(a, A) \in \mathcal{O}_2$ we have that $\frac{\rho(a,A)}{\delta_{v_n}(a,A)} > \frac{\rho(a,A)}{\delta_L(a,A)}$. Also, for an n sufficiently large, $(\frac{1}{2})^n < \min\{\rho(a, A) : a \in A \in \mathcal{D}\}$, and hence $(a, A) \in \mathcal{O}_3$ implies that $\frac{\rho(a,A)}{\delta_{v_n}(a,A)} \geq \frac{\rho(a,A)}{(\frac{1}{2})^m} > 1$. In this case, we can fix an m sufficiently large that, from Proposition 1,

$\overline{\lambda}_{\{\delta_{v_m}\}} = \min\limits_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta_{v_m}(a,A)} = \min\limits_{(a,A)\in\mathcal{O}_1\cup\mathcal{O}_2} \frac{\rho(a,A)}{\delta_{v_m}(a,A)} \geq \min\limits_{(a,A)\in\mathcal{O}_1\cup\mathcal{O}_2} \frac{\rho(a,A)}{\delta_L(a,A)} \geq \overline{\lambda}_{\{\delta_L\}}$.[10] Indeed, since $\delta_L \in \overline{\mathcal{S}}^1_{\texttt{Luce}}$, it must be that $\overline{\lambda}_{\{\delta_{v_m}\}} = \overline{\lambda}_{\{\delta_L\}}$ and hence, $\mathcal{O}_1$ is non-empty, with $\mathcal{O}_{\delta_{v_m}} \subseteq \mathcal{O}_{\delta_L} \subseteq \mathcal{O}_1$.

Assume, by way of contradiction, that there is no subcollection $\{(a_i, A_i)\}_{i=1}^I \subseteq \mathcal{O}_{\delta_L}$ such that $\bigcup_{i=1}^I \{a_i\} = \bigcup_{i=1}^I A_i$. Then, for every subcollection $\{(a_i, A_i)\}_{i=1}^I \subseteq \mathcal{O}_{\delta_{v_m}}$ it must also be that $\bigcup_{i=1}^I \{a_i\} \neq \bigcup_{i=1}^I A_i$. Hence, there must exist at least one alternative x such that $x \neq a$ for every $(a, A) \in \mathcal{O}_{\delta_{v_m}}$ and $x \in A$ for some $(a, A) \in \mathcal{O}_{\delta_{v_m}}$. Consider the segment $\alpha\mathbf{1}_x + (1-\alpha)v_m$, with $\alpha \in [0, 1]$, where $\mathbf{1}_x$ is a function assigning a value 1 to x and a value 0 to any other alternative. Select the maximal separation in this segment, which can be identified as follows. Partition the set of observations into two classes $\mathcal{O}' = \{(a, A) \in \mathcal{O}, a \neq x \in A\}$ and $\mathcal{O}'' = \mathcal{O} \setminus \mathcal{O}'$ and then select the Luce utilities defined by the unique value $\alpha^* \in [0, 1]$ that solves $\min\limits_{(a,A)\in\mathcal{O}'} \frac{\rho(a,A)}{\delta_{\alpha\mathbf{1}_x+(1-\alpha)v_m}(a,A)} = \min\limits_{(a,A)\in\mathcal{O}''} \frac{\rho(a,A)}{\delta_{\alpha\mathbf{1}_x+(1-\alpha)v_m}(a,A)}$. Notice that, given the structure of the Luce model, the left-hand ratio increases with α, continuously and strictly, approaching infinity. Similarly, the right-hand ratio weakly decreases with α continuously. Notice also that, for $\alpha = 0$, the left-hand ratio is strictly below the right-hand ratio. This is because there exists at least one observation on the left-hand side that belongs to $\mathcal{O}_{\delta_{v_m}}$. Thus, α^* must exist and Proposition 1 guarantees that this provides the maximal separation in the segment. Then, consider the vector of Luce utilities $v = \alpha^*\mathbf{1}_x + (1-\alpha^*)v_m$. If alternative x is present in all the menus in $\mathcal{O}_{\delta_{v_m}}$, then $\overline{\lambda}_{\{\delta_v\}} > \overline{\lambda}_{\{\delta_{v_m}\}} = \overline{\lambda}_{\{\delta_L\}}$, thus contradicting the maximality of δ_L. If x is not present in some menu of $\mathcal{O}_{\delta_{v_m}}$, it must be the case that $\mathcal{O}_{\delta_v} \subsetneq \mathcal{O}_{\delta_{v_m}}$ and $\overline{\lambda}_{\{\delta_v\}} = \overline{\lambda}_{\{\delta_L\}}$. Given the finiteness of the data, we can repeat the same exercise for δ_v and, eventually, contradict the optimality of δ_L. This concludes the proof. ■

The proof of Proposition 4 describes a convenient method with which to identify maximal separations for the Luce model. To explain the intuition of the result, consider a strictly positive instance of `Luce` given by $u \in \mathcal{U}$. Then, we have δ_u with critical observations $\mathcal{O}_{\delta_u}$. Proposition 4 shows that if there is a subcollection $\{(a_i, A_i)\}_{i=1}^I \subseteq \mathcal{O}_{\delta_u}$ such that $\bigcup_{i=1}^I \{a_i\} = \bigcup_{i=1}^I A_i$, then δ_u constitutes a maximal separation, and the maximal fraction of the data explained by the Luce model is simply $\overline{\lambda}_{\texttt{Luce}} = \frac{\rho(a_i,A_i)}{\delta_u(a_i,A_i)}$ with $(a_i, A_i) \in \mathcal{O}_{\delta_u}$. Suppose that there exists a δ_v that explains a larger fraction of

[10] This shows, in addition, that there is always a strictly positive instance of Luce that is maximal.

the data, and hence assigns lower Luce probabilities to all the critical observations of δ_u. Given the structure of Luce, we can assume, without loss of generality, that $\sum_{i=1}^{I} u(a_i) = \sum_{i=1}^{I} v(a_i)$. Then, clearly, reducing the Luce probability in (a_1, A_1) requires that one alternative in A_1, say a_2, is such that $v(a_2) > u(a_2)$. However, since there exists a critical observation of the form (a_2, A_2), we need to find another alternative in A_2, say a_3, with $v(a_3) > u(a_3)$. Hence, the process must be cyclic, and consequently, we cannot improve the ρ/δ ratio of all the critical observations of δ_u, which shows its maximality. The situation is entirely different when there is $x \in \bigcup_{i=1}^{I} A_i \setminus \bigcup_{i=1}^{I}\{a_i\}$. In this case, we can find an improvement by moving the Luce values in the direction of alternative x, that is, by increasing the Luce utility of x and reducing all the rest by the same proportion. That is, the ρ/δ ratios of the critical observations of δ_u increase, some of them strictly, by moving in the segment $\alpha\mathbf{1}_x + (1-\alpha)u$. This process can be repeated in such a way that all the ρ/δ ratios of the critical observations of δ_u ultimately strictly increase, thereby leading to a separation of Luce that explains a strictly larger fraction of the data.

We now illustrate these ideas using the example in Table 1. To begin, consider the Luce utilities $u = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. The value $\overline{\lambda}_{\{\delta_u\}} = \min\limits_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta_u(a,A)} = .45$ is obtained only for observation $(x, \{x, y, z\})$. Since $\{x, y, z\} \setminus \{x\} = \{y, z\}$ is non-empty, we can select one of the alternatives in $\{y, z\}$, say y, and move within the segment $\alpha(0, 1, 0) + (1 - \alpha)u = (\frac{1-\alpha}{3}, \frac{1+2\alpha}{3}, \frac{1-\alpha}{3})$. In order to select the appropriate value of α, we consider the observations (a, A) with $a \neq y \in A$ and the observations (y, A). Among the former, the minimal ratio of the data to the Luce probabilities is obtained for $(x, \{x, y, z\})$, with value $\frac{.45}{1-\alpha}$. In the latter, the minimal ratio is reached at $(y, \{y, z\})$, with value $\frac{.4(2+\alpha)}{1+2\alpha}$. Equation $\frac{.45}{1-\alpha} = \frac{.4(2+\alpha)}{1+2\alpha}$ yields $\alpha^* = \frac{1}{4}$, which leads to $v = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$. The value $\overline{\lambda}_{\{\delta_v\}} = \min\limits_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta_v(a,A)} = .6$ is obtained for pairs $\{(x, \{x, y, z\}), (z, \{x, z\}), (y, \{y, z\})\}$. Notice that $\{x, y, z\} \cup \{x, z\} \cup \{y, z\} = X$ and $\{x\} \cup \{z\} \cup \{y\} = X$, therefore, the condition of Proposition 4 is met, and the process concludes. $\overline{\lambda}_{\texttt{Luce}} = \overline{\lambda}_{\{\delta_v\}} = .6$, with the Luce stochastic choice function generated by v and residual noise $\epsilon(x, X) = 0, \epsilon(y, X) = \frac{3}{4}, \epsilon(x, \{x, y\}) = \frac{1}{8}, \epsilon(x, \{x, z\}) = 1$, and $\epsilon(y, \{y, z\}) = 0$.

3.4. **Single-crossing random utility model.** In random utility models (RUMs), there exists a probability distribution μ over the set of all possible preferences $\mathcal{P}$. At the choice stage, a preference is realized according to μ, and maximized, thereby determining the choice probabilities $\delta_\mu(a, A) = \sum\limits_{P\in\mathcal{P}:a=m_P(A)} \mu(P)$, for every $(a, A) \in \mathcal{O}$.

In other words, the choice probability of a given alternative within a menu is given by the sum of the probability masses associated to the preferences where the alternative is maximal within the menu.

The literature has often considered these models as complex to work with, and offered models in restricted domains that facilitate their use in applications. Here, we focus on the single-crossing random utility models (SCRUMs), which are RUMs over a set of preferences satisfying the single-crossing condition.[11] Formally, SCRUMs consider probability distributions μ on a given ordered collection of preferences $\mathcal{P}' = \{P_1, P_2, \dots, P_T\}$, satisfying the single-crossing condition $P_j \cap P_1 \subseteq P_i \cap P_1$ if and only if $j \geq i$. That is, the preference over a pair of alternatives x and y reverses once at most in the ordered collection of preferences. We denote the set of SCRUM stochastic choice functions by $\mathtt{SC}$. Proposition 5 characterizes the maximal separations for SCRUMs.

Proposition 5. *Let* $\delta_{\hat{\mu}_1} = \delta_{P_1}$ *and* $\hat{\lambda}_1 = \min_{A \in \mathcal{D}} \rho(m_{P_1}(A), A)$ *and, for every* $i \in \{2, \dots, T\}$, *define recursively: (i)* $\delta_{\hat{\mu}_i} = (1 - \frac{\hat{\lambda}_{i-1}}{\hat{\lambda}_i})\delta_{P_i} + \frac{\hat{\lambda}_{i-1}}{\hat{\lambda}_i}\delta_{\hat{\mu}_{i-1}}$ *and (ii)* $\hat{\lambda}_i = \min_{A \in \mathcal{D}} \Big\{\rho(m_{P_i}(A), A) + \max_{j: j \leq i, m_{P_j}(A) \neq m_{P_i}(A)} \hat{\lambda}_j\Big\}$. *Then,* $\delta_{\hat{\mu}_T} \in \overline{\mathcal{S}}^1_{\mathtt{SC}}$ *and* $\hat{\lambda}_T = \overline{\lambda}_{\mathtt{SC}}$.

Proof of Proposition 5: We start by proving that $\hat{\lambda}_T \leq \overline{\lambda}_{\mathtt{SC}}$. The construction guarantees that $1 \geq \lambda_T \geq \lambda_{T-1} \geq \cdots \geq \lambda_1 \geq 0$. Whenever $\hat{\lambda}_T = 0$, the result follows immediately from the fact that $\overline{\lambda}_{\mathtt{SC}} \in [0, 1]$. Now assume that $\hat{\lambda}_T \in (0, 1)$. We prove that there exists $\langle \delta_{\hat{\mu}_T}, \epsilon \rangle \in \mathcal{S}_{\mathtt{SC}}$ such that $\lambda_{\langle \delta_{\hat{\mu}_T}, \epsilon \rangle} = \hat{\lambda}_T$. First, notice that the construction guarantees that $\delta_{\hat{\mu}_T} \in \mathtt{SC}$. Hence, by defining $\epsilon = \frac{\rho - \hat{\lambda}_T \delta_{\hat{\mu}_T}}{1 - \hat{\lambda}_T}$, we trivially have that $\rho = \hat{\lambda}_T \delta_{\hat{\mu}_T} + (1 - \hat{\lambda}_T)\epsilon$ and we are left to prove that $\epsilon \in \mathtt{SCF}$. To see this, consider $(a, A) \in \mathcal{O}$ and denote by i_* and i^* the integers of the first and last preferences in $\mathcal{P}'$, such that a is the maximal element in A. The construction guarantees that $\rho(a, A) \geq \hat{\lambda}_{i^*} - \hat{\lambda}_{i_*-1} = \hat{\lambda}_T \frac{\hat{\lambda}_{i^*} - \hat{\lambda}_{i_*-1}}{\hat{\lambda}_T}$. Now, the recursive equations can be written as $\hat{\mu}_T(P_i) = \frac{\hat{\lambda}_{P_i} - \hat{\lambda}_{P_{i-1}}}{\hat{\lambda}_T}$ for every $i \in \{1, 2, \dots, T\}$, with $\hat{\lambda}_0 = 0$ and hence, $\rho(a, A) \geq \hat{\lambda}_T \sum_{i=i_*}^{i^*} \hat{\mu}_T(P_i) = \hat{\lambda}_T \delta_{\hat{\mu}_T}(a, A)$. This implies that $\epsilon(a, A) \geq 0$. Notice also that $\sum_{a \in A} \epsilon(a, A) = \sum_{a \in A} \frac{\rho(a,A) - \hat{\lambda}_T \delta_{\hat{\mu}_T}(a,A)}{1 - \hat{\lambda}_T} = \frac{1 - \hat{\lambda}_T}{1 - \hat{\lambda}_T} = 1$, thus proving that $\epsilon \in \mathtt{SCF}$. This illustrates the claim and, hence, the desired inequality. Finally, suppose that $\hat{\lambda}_T = 1$. In this case, the construction guarantees that $\rho = \delta_{\hat{\mu}_T} \in \mathtt{SC}$, which implies that $\overline{\lambda}_{\mathtt{SC}} = 1$ and the desired inequality.

[11] See Apesteguia, Ballester and Lu (2017) for a study of this model. Other RUMs using restricted domains are Gul and Pesendorfer (2006) and Lu and Saito (2017).

We now show that $\hat{\lambda}_T \geq \overline{\lambda}_{\texttt{SC}}$, by showing that, for every separation for SCRUM $\langle \delta_\mu, \epsilon \rangle$, it must be the case that $\hat{\lambda}_T \geq \lambda_{\langle \delta_\mu, \epsilon \rangle}$. We proceed recursively to show that $\hat{\lambda}_i \geq \sum_{j=1}^{i} \lambda_{\langle \delta_\mu, \epsilon \rangle} \mu(P_j)$ holds, and hence, $\hat{\lambda}_T \geq \sum_{j=1}^{T} \lambda_{\langle \delta_\mu, \epsilon \rangle} \mu(P_j) = \lambda_{\langle \delta_\mu, \epsilon \rangle}$, as desired. Let $i = 1$ and A' be a menu solving $\min\limits_{A \in \mathcal{D}} \rho(m_{P_1}(A), A)$. Hence, $\hat{\lambda}_1 - \lambda_{\langle \delta_\mu, \epsilon \rangle} \mu(P_1) = \rho(m_{P_1}(A'), A') - \lambda_{\langle \delta_\mu, \epsilon \rangle} \mu(P_1) \geq \rho(m_{P_1}(A'), A') - \lambda_{\langle \delta_\mu, \epsilon \rangle} \sum_{j: m_{P_j}(A') = m_{P_1}(A')} \mu(P_j)$. By the definition of SCRUMs, the last expression can be written as $\rho(m_{P_1}(A'), A') - \lambda_{\langle \delta_\mu, \epsilon \rangle} \delta_\mu(m_{P_1}(A'), A')$, or equivalently as $(1 - \lambda_{\langle \delta_\mu, \epsilon \rangle}) \epsilon(m_{P_1}(A'), A')$. Since $\epsilon \in \texttt{SCF}$, the latter expression must be positive, thus proving the desired result. Suppose that the inequality is true for every P_j with $j < i$. We now prove this for P_i. Let A^* be a menu solving $\min_{A \in \mathcal{D}} [\rho(m_{P_i}(A), A) + \max_{j: j \leq i, m_{P_j}(A) \neq m_{P_i}(A)} \hat{\lambda}_j]$. Then, we have $\rho(m_{P_i}(A^*), A^*) = \lambda_{\langle \delta_\mu, \epsilon \rangle} \delta_\mu(m_{P_i}(A^*), A^*) + (1 - \lambda_{\langle \delta_\mu, \epsilon \rangle}) \epsilon(m_{P_i}(A^*), A^*) \geq \lambda_{\langle \delta_\mu, \epsilon \rangle} \delta_\mu(m_{P_i}(A^*), A^*) = \lambda_{\langle \delta_\mu, \epsilon \rangle} \sum_{P: m_P(A^*) = m_{P_i}(A^*)} \mu(P)$. If it is the case that $\{P : m_P(A^*) = m_{P_i}(A^*)\} \supseteq \{P_1, P_2, \ldots, P_i\}$, then clearly $\hat{\lambda}_i = \rho(m_{P_i}(A^*), A^*) \geq \lambda_{\langle \delta_\mu, \epsilon \rangle} \sum_{P: m_P(A^*) = m_{P_i}(A^*)} \mu(P) = \sum_{j=1}^{i} \lambda_{\langle \delta_\mu, \epsilon \rangle} \mu(P_j)$ and we have concluded the induction argument. Otherwise, the single-crossing condition guarantees that there exists $j^* \in \{1, \ldots, i-1\}$ such that $\{P : m_P(A^*) = m_{P_i}(A^*)\} \supseteq \{P_{j^*+1}, P_{j^*+2}, \ldots, P_i\}$ and $\rho(m_{P_i}(A^*), A^*) \geq \sum_{j=j^*+1}^{i} \lambda_{\langle \delta_\mu, \epsilon \rangle} \mu(P_j)$. In this case, the induction hypothesis also guarantees that $\hat{\lambda}_{j^*} \geq \sum_{j=1}^{j^*} \lambda_{\langle \delta_\mu, \epsilon \rangle} \mu(P_j)$. By combining these two inequalities, we are able to conclude that $\hat{\lambda}_i \geq \sum_{j=1}^{i} \lambda_{\langle \delta_\mu, \epsilon \rangle} \mu(P_j)$ and the induction step is complete. This implies, in particular, that $\lambda_{\langle \delta_\mu, \epsilon \rangle} \leq \hat{\lambda}_T$.

By combining the above two claims, we have shown that $\hat{\lambda}_T = \overline{\lambda}_{\texttt{SC}}$, and, by the construction, that $\delta_{\hat{\mu}_T} \in \overline{\mathcal{S}}^1_{\texttt{SC}}$, which concludes the proof. ■

Proposition 5 provides a smooth recursive method with which to obtain a maximal separation and the corresponding maximal fraction of data explained by SCRUM. It basically computes the fraction of data, $\hat{\lambda}_i$, that can be explained by SCRUMs using preferences up to P_i. Trivially, the maximal fraction of data explained by P_1 is $\min\limits_{A \in \mathcal{D}} \rho(m_{P_1}(A), A)$. Now consider any other preference $P_i \in \mathcal{P}'$ and assume that every preference P_j, $j < i$, has been analyzed. With the extra preference P_i, and for a given menu A, we can rationalize data $\rho(m_{P_i}(A), A)$ together with any other data $\rho(x, A)$, $x \neq m_{P_i}(A)$, that is rationalized by preferences preceding P_i. This can be achieved by considering the appropriate linear combination of the constructed SCRUM that uses preferences up to P_{i-1} with preference P_i.

We now illustrate how Proposition 5 works in the example of Table 1, where we assume the set of single-crossing preferences zP_1yP_1x, yP_2zP_2x, yP_3xP_3z and xP_4yP_4z.

We start with P_1. The fraction of data explained by P_1 is $\hat{\lambda}_1 = \min_{A \subseteq X} \rho(m_{P_1}(A), A) = \min\{\rho(z, X), \rho(y, \{x, y\}), \rho(z, \{x, z\}), \rho(z, \{y, z\})\} = \min\{.25, .75, .3, .6\} = .25$, where trivially $\mu_1(P_1) = 1$. We then consider preference P_2, where we have that $\hat{\lambda}_2 = \min\{\rho(y, X) + \hat{\lambda}_1, \rho(y, \{x, y\}), \rho(z, \{x, z\}), \rho(y, \{y, z\}) + \hat{\lambda}_1\} = \min\{.6 + .25, .75, .3, .4 + .25\} = .3$ with $\mu_2(P_1) = \frac{\hat{\lambda}_1}{\hat{\lambda}_2} = \frac{5}{6}$ and $\mu_2(P_2) = \frac{1}{6}$. For preference P_3, we have that $\hat{\lambda}_3 = \min\{\rho(y, X) + \hat{\lambda}_1, \rho(y, \{x, y\}), \rho(x, \{x, z\}) + \hat{\lambda}_2, \rho(y, \{y, z\}) + \hat{\lambda}_1\} = \min\{.6 + .25, .75, .7 + .3, .4 + .25\} = .65$, with $\mu_3(P_1) = \frac{\hat{\lambda}_2}{\hat{\lambda}_3}\mu_2(P_1) = \frac{5}{13}$, $\mu_3(P_2) = \frac{\hat{\lambda}_2}{\hat{\lambda}_3}\mu_2(P_2) = \frac{1}{13}$ and $\mu_2(P_3) = \frac{7}{13}$. Finally, we have that $\hat{\lambda}_4 = \min\{\rho(x, X) + \hat{\lambda}_3, \rho(x, \{x, y\}) + \hat{\lambda}_3, \rho(x, \{x, z\}) + \hat{\lambda}_2, \rho(y, \{y, z\}) + \hat{\lambda}_1\} = \min\{.15 + .65, .25 + .65, .7 + .3, .4 + .25\} = .65$ and hence $\mu_4 = \mu_3$. Thus, we conclude that $\overline{\lambda}_{\texttt{SC}} = .65$, with SCRUM δ_{μ_4} and residual noise $\epsilon(x, X) = \frac{3}{7}, \epsilon(y, X) = \frac{4}{7}, \epsilon(x, \{x, y\}) = \frac{5}{7}, \epsilon(x, \{x, z\}) = 1$, and $\epsilon(y, \{y, z\}) = 0$.

4. An empirical application

Here we use an experimental dataset to operationalize the maximal separation results obtained in the previous section.[12] There were nine equiprobable monetary lotteries, described in Table 1. Each of the 87 participants faced 108 different menus of lotteries, including all 36 binary menus and a random sample of larger menus.[13] There were two treatments. Treatment NTL was a standard implementation, with no time limit on the choice. In treatment TL, subjects had to select a lottery within a limited time. At the end of the experiment, one of the menus was chosen at random and the subject was paid according to his or her choice from that menu.[14]

Table 1. Lotteries

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

To ensure a sufficiently large number of observations, we focus on the choices made in the binary menus, which, when aggregating both treatments, gives a total of 87

[12] We collected the experimental data together with Syngjoo Choi at UCL in March 2013, within the context of another research project. This is the first completed paper to use this dataset. We are very grateful to Syngjoo for kindly allowing us to use this dataset.

[13] There were menus of 2, 3 and 5 alternatives, presented one at a time, in a randomized order. No participant was presented more than once with the same menu of alternatives. The location of the lotteries on the screen was randomized, as was the location of the monetary prizes within a lottery.

[14] Specifically, subjects had 5, 7 and 9 seconds for the menus of 2, 3, and 5 alternatives, respectively.

observations overall.[15] Table 2 reports the choice probabilities in each of the binary menus. It also reports the optimal and the residual stochastic choice functions identified in the maximal separation results, using the models described in the previous section. In SCRUM we use the CRRA expected utility representation, which is by far the most widely used utility representation for risk preferences.[16] There are several lessons to be learned from the table.

First note that the maximal fractions of the data explained by the respective models, $\overline{\lambda}_{\Delta}$, increase from the deterministic choice model, to the tremble model, to the Luce model and, finally, to the SCRUM-CRRA model. It is worth noting that the deterministic model already explains about half of the data, i.e., $\overline{\lambda}_{\texttt{DET}} = .51$. The identified optimal instance is the one associated with the preference $l_1 P l_5 P l_4 P l_8 P l_7 P l_9 P l_3 P l_6 P l_2$. The top alternative, lottery l_1, is the safest, since it gives £17 with probability one. The next is lottery l_5, which has the second lowest variance at the expense of a very low expected return. Lottery l_2, the one with the highest expected value and highest variance, is regarded as the worst alternative. Hence, the deterministic model pictures a population that is essentially highly risk-averse. The model reaches its explanatory limits with the critical observation $(l_8, \{l_7, l_8\})$ where, by Proposition 1, the ratio of observed to predicted probability is minimal. Specifically, the observed choice probability is .51 while the deterministic prediction is 1. The ratio of these two values gives the fraction of data explained by the model, $\overline{\lambda}_{\texttt{DET}} = .51$.

The tremble model identifies exactly the same preference as the deterministic model, while markedly increasing the maximal fraction of the data explained from $\overline{\lambda}_{\texttt{DET}} = .51$ to $\overline{\lambda}_{\texttt{Tremble}} = .68$. This is the result of using a relatively large tremble probability, $\gamma = .51$. The tremble model is characterized by critical observations $(l_8, \{l_7, l_8\})$ and $(l_9, \{l_5, l_9\})$. As in the deterministic case, choice data is scarce for l_8 versus l_7, but the problem is less severe thanks to the presence of a tremble, due to which, the individual is predicted to choose l_8 only with probability .74, thereby reducing the ratio of observed to predicted probabilities to .68. This ratio cannot be improved beyond this point. Although increasing the tremble probability would increase this ratio, it would also decrease the ratio of the other critical observation, $(l_9, \{l_5, l_9\})$, which has the same value of .68. To

[15] Due to the time limit in one of the treatments, the number is slightly lower for some menus. Specifically, there are 18 menus with 87 observations, 12 with 86, 3 with 85 and 3 with 84.

[16] The CRRA Bernoulli function is $\frac{x^{1-r}}{1-r}$, whenever $r \neq 1$, and $\log x$ otherwise, with x representing money. We have also studied the cases of CARA expected utility, and mean-variance utility, and obtained similar results, which are available upon request.

Table 2. Data and Δ-Maximal Separations

		DET		TREMBLE		LUCE		SCRUM-CRRA	
A	ρ	$\delta_{\texttt{DET}}$	$\epsilon_{\texttt{DET}}$	$\delta_{\texttt{Tremble}}$	$\epsilon_{\texttt{Tremble}}$	$\delta_{\texttt{Luce}}$	$\epsilon_{\texttt{Luce}}$	$\delta_{\texttt{SC-CRRA}}$	$\epsilon_{\texttt{SC-CRRA}}$
$\{l_1, l_2\}$	0.75	1.00	0.49	0.74	0.75	0.91	0.30	0.74	0.77
$\{l_1, l_3\}$	0.60	1.00	0.19	0.74	0.29	0.71	0.28	0.55	0.78
$\{l_2, l_3\}$	0.33	0.00	0.67	0.26	0.50	0.20	0.69	0.24	0.66
$\{l_1, l_4\}$	0.53	1.00	0.05	0.74	0.07	0.62	0.27	0.47	0.75
$\{l_2, l_4\}$	0.28	0.00	0.56	0.26	0.32	0.15	0.64	0.24	0.40
$\{l_3, l_4\}$	0.43	0.00	0.86	0.26	0.78	0.40	0.50	0.42	0.46
$\{l_1, l_5\}$	0.58	1.00	0.16	0.74	0.24	0.46	0.92	**0.47**	**1.00**
$\{l_2, l_5\}$	0.25	0.00	0.51	0.26	0.25	0.08	0.73	0.26	0.23
$\{l_3, l_5\}$	0.45	0.00	0.92	0.26	0.87	**0.26**	**1.00**	0.45	0.46
$\{l_4, l_5\}$	0.49	0.00	0.99	0.26	0.98	0.34	0.89	0.53	0.33
$\{l_1, l_6\}$	0.72	1.00	0.44	0.74	0.68	0.87	0.31	0.76	0.60
$\{l_2, l_6\}$	0.44	0.00	0.89	0.26	0.84	0.42	0.51	0.42	0.53
$\{l_3, l_6\}$	0.80	1.00	0.60	0.74	0.93	**0.74**	**1.00**	0.79	0.84
$\{l_4, l_6\}$	0.76	1.00	0.51	0.74	0.79	0.81	0.62	0.76	0.76
$\{l_5, l_6\}$	0.75	1.00	0.49	0.74	0.75	0.89	0.35	0.76	0.71
$\{l_1, l_7\}$	0.63	1.00	0.25	0.74	0.38	0.77	0.23	0.74	0.22
$\{l_2, l_7\}$	0.24	0.00	0.49	0.26	0.22	0.26	0.21	0.26	0.19
$\{l_3, l_7\}$	0.48	0.00	0.96	0.26	0.94	0.57	0.20	0.53	0.27
$\{l_4, l_7\}$	0.62	1.00	0.24	0.74	0.37	0.67	0.49	0.76	0.14
$\{l_5, l_7\}$	0.63	1.00	0.26	0.74	0.40	0.79	0.18	0.76	0.18
$\{l_6, l_7\}$	0.27	0.00	0.54	0.26	0.29	0.33	0.10	0.24	0.36
$\{l_1, l_8\}$	0.64	1.00	0.27	0.74	0.42	0.67	0.57	0.76	0.21
$\{l_2, l_8\}$	0.22	0.00	0.45	0.26	0.15	0.17	0.36	0.26	0.09
$\{l_3, l_8\}$	0.36	0.00	0.73	0.26	0.58	0.45	0.12	0.45	0.03
$\{l_4, l_8\}$	0.56	1.00	0.12	0.74	0.18	0.55	0.60	0.56	0.56
$\{l_5, l_8\}$	0.62	1.00	0.23	0.74	0.36	0.70	0.40	0.76	0.13
$\{l_6, l_8\}$	0.20	0.00	0.40	0.26	0.07	0.23	0.12	0.24	0.04
$\{l_7, l_8\}$	0.49	**0.00**	**1.00**	**0.26**	**1.00**	0.37	0.83	0.42	0.77
$\{l_1, l_9\}$	0.76	1.00	0.51	0.74	0.78	0.74	0.81	0.79	0.62
$\{l_2, l_9\}$	0.28	0.00	0.56	0.26	0.32	0.23	0.42	0.28	0.28
$\{l_3, l_9\}$	0.39	0.00	0.79	0.26	0.68	**0.53**	**0.00**	0.45	0.17
$\{l_4, l_9\}$	0.55	1.00	0.08	0.74	0.13	0.63	0.32	0.53	0.60
$\{l_5, l_9\}$	0.83	1.00	0.65	**0.74**	**1.00**	**0.76**	**1.00**	1.00	0.20
$\{l_6, l_9\}$	0.22	0.00	0.44	0.26	0.14	0.29	0.02	0.26	0.08
$\{l_7, l_9\}$	0.56	1.00	0.12	0.74	0.18	0.46	0.87	0.45	0.96
$\{l_8, l_9\}$	0.64	1.00	0.26	0.74	0.41	0.58	0.78	**0.53**	**1.00**
	λ_Δ	0.51		0.68		0.74		0.78	

Note: A denotes the binary menu of lotteries, ρ, δ_Δ and ϵ_Δ are the observed percentage, predicted choice of model Δ and error choice probability of Δ, respectively, associated with choosing lottery l_i from menu $\{l_i, l_j\}$, and λ_Δ reports the maximal fraction of the data explained by Δ with $\Delta \in \{\texttt{DET}, \texttt{Tremble}, \texttt{Luce}, \texttt{SC-CRRA}\}$. Data entries in bold refer to the menus containing the critical observations in the respective model.

see this, notice the choice prediction for alternative l_9, being worse than alternative l_5, corresponds entirely to the tremble probability, and hence, an increase in tremble would increase the predicted probability and thus decrease the ratio.

The Luce model is able to explain close to three quarters of the data. The optimal utility values for the lotteries are $u = (0.22, 0.02, 0.09, 0.13, 0.25, 0.03, 0.07, 0.11, 0.08)$, which again suggest a highly risk averse population. The alternative with the highest Luce utility value is lottery l_5, followed by lottery l_1, while lottery l_2, which is the riskiest, has the lowest Luce utility value. That is, although u does not represent $P_{\mathtt{DET}}$ exactly, it represents a preference very close to it. Interestingly, we see that the Luce model can accommodate a larger fraction of the data by allowing randomness to depend on the cardinal evaluation of alternatives. The model is hard pressed to explain observations $(l_5, \{l_3, l_5\})$, $(l_6, \{l_3, l_6\})$, $(l_3, \{l_3, l_9\})$ and $(l_9, \{l_5, l_9\})$, that represent the type of cyclical structure described in Proposition 4. In each of these observations, the ratio of observed to predicted probabilities is equal to .74. Increasing any of these ratios would require decreasing the utility of one alternative in $\{l_3, l_5, l_6, l_9\}$, but only, of course, at the expense of the ratio of another of these critical observations.

Finally, SC-CRRA explains as much as nearly 80% of the data. In so doing, it assigns positive masses to 9 of the 30 possible CRRA preferences, with the largest probability mass, .44, associated with the most risk averse CRRA preference, i.e., preference $l_1Pl_5Pl_9Pl_8Pl_4Pl_7Pl_3Pl_6Pl_2$, which is again very close to $P_{\mathtt{DET}}$. Since each preference compatible with CRRA corresponds to an interval of risk aversion levels, we can completely describe the optimal SC-CRRA instance by reporting the values of the cumulative distribution function at the upper bounds of these intervals. These are $F(-4.15) = 0.21$, $F(-0.31) = 0.25$, $F(0.34) = 0.27$, $F(0.41) = 0.29$, $F(0.44) = 0.43$, $F(0.61) = 0.47$, $F(1) = 0.53$, $F(4) = 0.56$ and $F(\infty) = 1$. Notice that, in addition to explaining a large fraction of the data, SC-CRRA is also rich enough to show that a quarter of the population is risk loving, $F(-0.31) = 0.25$. The limits of SC-CRRA in explaining the data are reached at observations $(l_5, \{l_1, l_5\})$ and $(l_9, \{l_8, l_9\})$. On the one hand, lottery l_5 is preferred over lottery l_1 by all CRRA levels with a risk aversion level below 2, which has an accumulated mass of .53. Given the observed choices, this leads to a critical ratio for observation $(l_5, \{l_1, l_5\})$ of .78. Improving this ratio would necessarily require us to assign a higher weight to levels of risk aversion above 2. However, this would immediately conflict with the ratio of l_9 to l_8, since l_9 is ranked above

l_8 at all levels of risk aversion above 1. As the ratio of observed to predicted data for $(l_9, \{l_8, l_9\})$ also has the critical value of .78, no improvement can take place.

To conclude the discussion of Table 2, we would like to emphasize that the four models are very consistent in the qualitative judgment of the population. All four models take the population of subjects to be highly risk averse. Then, we see that, by introducing different sources of randomness, it is possible to explain larger fractions of the data, and that the precise source of randomness affects the fraction of the data explained.

The identification of maximal separations of the data can be contrasted with other standard estimation techniques, such as maximum likelihood or least squares. Appendix A contains a formal comparison of these techniques. Here, we briefly comment on the differences that emerge when using these alternative techniques. First, the standard techniques identify instances of the models, but they do not quantify the fraction of the data explained by them, nor do they characterize the nature of the data that falls outside the models. Thus, in the exercise that follows, we must limit the comparison to the instances of the models identified by the different estimation techniques. Second, Appendix A shows how the structure of maximum likelihood and least squares loss functions are relatively similar to each other; both techniques are additive processes by which the deviations from all observations are aggregated. As a matter of fact, with our data, the estimates derived from both techniques are almost identical, and hence we concentrate, in what follows, on comparing our technique with one of them, the maximum likelihood technique.

Table 3 reports the instances of the models identified by the maximal separation and the maximum likelihood techniques. With respect to the deterministic model, no difference whatsoever is observed, as exactly the same preference relation is estimated. This ordinal similarity is preserved in the case of the tremble model, although our technique predicts a substantially smaller trembling coefficient, $.51 < .68$. The intuition for this difference is straightforward. Recall that, as we mentioned above, $(l_9, \{l_5, l_9\})$ is a critical observation in the maximal separation exercise for `Tremble`. The observed probability in this observation is small, .17, and the identified instance of the model for our technique predicts, due to the trembling parameter, a rather relative large frequency of .26. However, the maximum likelihood exercise is not severely affected by this local consideration and makes the estimation only by averaging over all the observations. Consequently, the estimation exercise in maximum likelihood is willing to sacrifice the

Table 3. Maximal Separation and Maximum Likelihood Instances of the Models

	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	$F(-4.15) = 0.22, F(-0.31) = 0.29, F(0.44) = 0.44$
	$F(1) = 0.50, F(-4) = 0.56, F(\infty) = 1$

Note: MS and ML denote maximal separation and maximum likelihood, respectively. P denotes the preference identified in the corresponding case, where the ranking declines from left to right, γ is the tremble probability in `Tremble`, u is the Luce utility vector associated with `Luce`, where the i-th entry in u corresponds to the utility value of lottery l_i, and finally $F(r)$ denotes the cumulative probability masses associated with the upper bounds of the intervals of the relative risk aversion coefficients r consistent with those CRRA preference relations that have a strictly positive mass in the corresponding estimation procedure.

prediction quality of this extreme observation in order to favor the prediction over other moderate ones. This is done by increasing substantially the trembling parameter and consequently the prediction in this particular observation $(l_9, \{l_5, l_9\})$, reaching a disproportionate value of .34, two times the observed value. A similar reasoning applies to the comparison of the cases of `Luce` and `SC-CRRA`.

The comparative analysis of both estimation techniques can be complemented with a prediction exercise that helps to explain the pattern of differences that emerges. We take all the binary data except for one binary set, estimate the instances of the models by maximal separation and maximum likelihood 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 maximal separation and maximum likelihood overestimate the probability of the same alternative in the binary

Table 4. Forecasting Results of Maximal Separation and Maximum Likelihood

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

Note: Every binary menu of lotteries $A = \{l_i, l_j\}$ reported in the table is ordered such that l_i is the lottery where the predictions of both maximal separation (MS) and maximum likelihood (ML) are above the observed choice data ρ. Those observations for which one of the predictions of MS or ML is above the observed choice data and the other below are not reported in the table. Then, for each one of the models, the binary menus of lotteries are ordered from lower to higher observed choice probabilities. Bold entries refer to the cases where MS is closer to the data and italicized entries refers to those cases where ML is closer to the data.

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, since the conclusions may otherwise depend on the particular distance function employed. Table 4 reports all the binary menus for which the predictions of both maximal separation and maximum likelihood can be unambiguously ranked, focusing on the observations for which both predictions are in excess of the observed data.[17] The theoretical analysis of the two techniques suggests a very intuitive conjecture; namely, that the maximal separation technique is very cautious and can therefore be expected to perform better in observations with low choice probabilities. This conjecture is largely confirmed in our analysis. In all three models, the overestimation of small probabilities is less problematic for the maximal separation technique, while maximum likelihood deals better with the overestimation of large probabilities. If one is interested in forecasting exercises, these results suggest that, to obtain a clear picture of the overall situation, it may be useful to apply both estimation techniques: maximal separation and maximum likelihood.

5. Final considerations

We have offered a novel methodology aimed at finding the maximal fraction of the data that can be explained by a stochastic choice model, by identifying the instance of the model that best explains the data, and discerning the anomalies involved in residual behavior. We have characterized the general results for any model, and investigated several prominent models used in the literature. Our approach may, in addition, prove instrumental in exploring several lines of research, some of which we now briefly comment.

Firstly, the study of maximal separations permits an intuitive use of mixture models. On the one hand, it allows us to define a model as the convex combination of other models (as in the case of SCRUMs, which are defined as the convex hull of some deterministic choice models), and directly provide maximal separations for the mixture model. However, if the analyst prefers to give priority to a certain model, she may sequentially repeat the maximal separation exercise, i.e., obtain a maximal separation for the preferred model, and take the residual data as the new observed stochastic choice function to be maximally separated using a different model. This enables the identification of a second behavioral type.

Secondly, one may be interested in restricting the space of possible residual choice functions to, e.g., a closed subset of stochastic choice functions containing the particular

[17] We do not report the results of the deterministic method, where the maximal separation and maximum likelihood predictions are exactly the same.

model of analysis and the actual data. This would enable the identification of residual stochastic choice functions with a certain structure of interest. It is clear that our techniques would, with minor modifications, apply to such a setting.

Finally, a related but alternative way of thinking would be to find the instance of model $\widetilde{\delta}$ and residual noise $\widetilde{\epsilon}$, such that $\widetilde{\delta} = \widetilde{\lambda}\rho + (1-\widetilde{\lambda})\widetilde{\epsilon}$, with minimal $1-\widetilde{\lambda}$. This would identify the minimal missing stochastic behavior, which, when combined with the data, rationalizes an instance of the model. It is easy to see that the logic of Proposition 1 applies here in basically the opposite direction. Now, the critical observations are those which maximize the ratio of the data to the instance of the model, that is, those cases where the data far exceed the predictions of the model. Consequently, the identification of the optimal instance of the model would require minimizing the ratios involved in the critical observations.

Appendix A. Deviation measures

Here we relate our approach to two important strands of literature involving loss functions and inconsistency indices. Essentially, both of these notions study deviations of actual behavior with respect to some benchmark. Loss functions measure the deviation between data and each instance of a stochastic model, with the ultimate purpose of identifying the closest instance, i.e., the one that minimizes the loss function, whereas inconsistency indices aim to measure the deviation of actual behavior when judged from the perspective of the standard deterministic choice model.

Loss functions. A loss function is a map $L : \Delta \times \texttt{SCF} \to \mathbb{R}$ that measures the deviation of data ρ from each of the instances δ of a stochastic model Δ. The loss function is then minimized in the space Δ, in order to identify the best instance of the model. Here, we formalize how the identification of the best instance in a maximal separation can be rewritten in terms of the minimization of a loss function. We then compare the resulting maximal-separation loss function with the well-known loss functions of other standard approaches in the literature, such as least squares and maximum likelihood. In order to facilitate the discussion, we focus on the case of δ being strictly positive.

Maximal separations are those that maximize the fraction of the data consistent with the model, that is $\overline{\lambda}_{\Delta}$, or alternatively, the ones that minimize the fraction representing residual noise, that is $1 - \overline{\lambda}_{\Delta}$. It is evident, therefore, that, within our approach, we can obtain optimal instances by simply minimizing the loss function $1 - \overline{\lambda}_{\{\delta\}}$ across all the instances δ of model Δ. Proposition 1 characterizes the structure of the maximal-separation loss function. Since $1-\min_{(a,A)\in\mathcal{O}}\frac{\rho(a,A)}{\delta(a,A)}=\max_{(a,A)\in\mathcal{O}}[1-\frac{\rho(a,A)}{\delta(a,A)}]$, the loss function can be written as

$$L_{MS}(\delta,\rho)=\max_{(a,A)\in\mathcal{O}}[1-\frac{\rho(a,A)}{\delta(a,A)}].$$

The least squares approach involves the minimization of the quadratic loss function

$$L_{LS}(\delta,\rho)=\sum_{(a,A)\in\mathcal{O}}(\delta(a,A)-\rho(a,A))^2.$$

First, notice that, for each observation $(a,A)\in\mathcal{O}$, L_{MS} and L_{LS} evaluate the existing gap between the data and the model in different ways. While L_{LS} evaluates the absolute difference between the data and the specified model, the function L_{MS} operates in a relative sense. Note, further, that, while L_{LS} aggregates the deviations across observations, L_{MS} focuses on the largest relative deviation.

Similarly, it is well known that the maximum likelihood approach is equivalent to the minimization of the loss function given by the Kullback-Leibler divergence from δ to ρ, which can be written as

$$L_{ML}(\delta,\rho)=\sum_{(a,A)\in\mathcal{O}}\rho(a,A)\log\frac{\rho(a,A)}{\delta(a,A)}.$$

The Kullback-Leibler divergence can be interpreted as the amount of information lost due to the use of δ instead of ρ. Comparing L_{MS} with L_{ML}, notice that, although both loss functions operate over the ratios of the data to the model, L_{MS} works over the largest deviation between the data and the specified model, while L_{ML} operates in the form of an expectation.

Table 2. L_{MS} versus L_{LS} and L_{ML}

	x	y	z	L_{MS}	L_{LS}	L_{ML}
ρ	.25	.45	.3			
δ_1	.15	.65	.2	.31	.06	.04
δ_2	.2	.35	.45	.33	.03	.02

Table 3. L_{LS} versus L_{MS} and L_{ML}

	x	y	z	L_{MS}	L_{LS}	L_{ML}
ρ	.1	.3	.6			
δ_1	.2	.1	.7	.5	.06	.07
δ_2	.15	.1	.75	.33	.065	.06

Table 4. L_{ML} versus L_{MS} and L_{LS}

	x	y	z	L_{MS}	L_{LS}	L_{ML}
ρ	.3	.6	.1			
δ_1	.05	.5	.45	.78	.19	.22
δ_2	.25	.25	.5	.8	.28	.18

Tables 2, 3 and 4 report simple examples of menus with three alternatives, showing that the differences in the structure of these loss functions may translate into the identification of different best instances. For each of the three loss functions, the tables give a dataset ρ and two model instances δ_1 and δ_2, such that the corresponding loss function ranks these two instances in complete reverse to the ranking given by the other two loss functions.

Inconsistency indices. Starting with Afriat (1973), there is a literature on measuring deviations of actual behavior with respect to the standard, deterministic, rational choice model. Formally, an inconsistency index can be defined as a mapping $I : \texttt{SCF} \to \mathbb{R}$ describing the inconsistency of a dataset $\rho \in \texttt{SCF}$ with the standard deterministic model, that is when the reference model is set as $\Delta = \texttt{DET}$. Despite some exceptions, most of the existing inconsistency indices are obtained throughout the minimization of a loss function. That is, the index starts by measuring the deviation of data with respect to every instance of the deterministic model, i.e., to every preference relation. In a second stage, the loss is minimized across all possible instances of the model, providing an inconsistency value and, at the same time, identifying the closest preference relation to the data.[18]

[18] See Apesteguia and Ballester (2015) for a characterization of this class and for a detailed review of the literature.

We can then immediately analyze the inconsistency index emerging from the maximal separation technique. Using the loss function discussed above, and the insights obtained in Section 3.1, we have

$$I_{MS} = 1 - \overline{\lambda}_{\texttt{DET}} = \min_{\delta_P} \max_{A} \sum_{\substack{a \in A: \\ \delta_P(a,A)=0}} \rho(a,A).^{19}$$

It is important to note that the nature of this index is unique in the literature. To illustrate this more clearly, we now compare the inconsistency index arising from the maximal separation approach with the well-known inconsistency index of Houtman and Maks (1985), which represents the closest index to I_{MS}. The Houtman and Maks index measures inconsistency by the minimal amount of data that needs to be removed in order to make the remainder of the data rationalizable by the standard choice model. The key difference is that the Houtman-Maks index enables different proportions of data to be removed from different menus of alternatives. Hence, using our notation, we can write the Houtman-Maks index as

$$I_{HM} = \min_{\delta_P} \sum_{A} \sum_{\substack{a \in A: \\ \delta_P(a,A)=0}} \rho(a,A).$$

These formulations provide a transparent comparison between the two approaches. Both methods remove data minimally until the surviving data is rationalizable. In the case of a maximal separation, since data must be removed at the same rate across all menus, the index focuses on the most problematic menu. In the case of Houtman and Maks, different proportions of data can be removed from different menus, therefore an aggregation across menus takes place.

Table 5. I_{MS} versus I_{HM}

	x	y	z
$\{x,y,z\}$	.25	.3	.4
$\{x,y\}$	.8	.2	
$\{x,z\}$	.4		.6
$\{y,z\}$		.7	.3

[19] Section 3.1 studies in detail the deterministic model and provides a convenient algorithm for computing $1 - \overline{\lambda}_{\texttt{DET}}$.

Table 5 reports an example of a choice function ρ with three alternatives and with data on all the relevant menus of alternatives. Taken from the perspective of I_{MS}, the data show the optimal preference to be $zPxPy$, while from the perspective of I_{HM} it is $xP'yP'z$.

To conclude this comparison of the maximal separation approach with both loss and inconsistency functions, it is crucial to emphasize that the loss function L_{MS} and the inconsistency index I_{MS} are the by-products of addressing a more general issue, which is to provide a methodology for maximally separating data representing predicted randomness, from that representing unknown noise. That is, the maximal separation approach is unique when it comes to quantifying the fraction of the data that is consistent with the model, and identifying the instance of the model and the residual-noise stochastic choice function that jointly explain the data. Thus, for anyone interested in the conceptual problem addressed in this paper, L_{MS} and I_{MS} are the way to go.

References

[1] Afriat, S.N. (1973). "On a System of Inequalities in Demand Analysis: An Extension of the Classical Method," *International Economic Reviw* 14:460–72.

[2] Agranov, M. and P. Ortoleva (2016). "Stochastic Choice and Preferences for Randomization," *Journal of Political Economy*, forthcoming.

[3] Ahn, D.S. and T. Sarver (2013). "Preference for Flexibility and Random Choice," *Econometrica*, 81(1):341–361.

[4] Apesteguia, J. and M.A. Ballester (2015). "A Measure of Rationality and Welfare," *Journal of Political Economy*, 123.6:1278–1310.

[5] Apesteguia, J. and M.A. Ballester (2017). "Monotone Stochastic Choice Models: The Case of Risk and Time Preferences," *Journal of Political Economy*, forthcoming.

[6] Apesteguia, J., M.A. Ballester and J. Lu (2017). "Single-Crossing Random Utility Models," *Econometrica*, 85.2:661–674.

[7] Barseghyan, L., F. Molinari, T. O'Donoghue and J.C. Teitelbaum (2016), "Estimating Risk Preferences in the Field," *Journal of Economic Literature*, forthcoming

[8] Brady, R.L. and J. Rehbeck (2016), "Menu-Dependent Stochastic Feasibility," *Econometrica*, 84.3:1203–1223.

[9] Block, H.D. and J. Marschak (1960). "Random Orderings and Stochastic Theories of Response," in I. Olkin et al, eds., Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, 97–132. Standford, Calif.: Standford University Press.

[10] Caplin, A., M. Dean and D. Martin (2011). "Search and Satisficing," *American Economic Review*, 101: 2899–2922.

[11] Caplin, A. and M. Dean (2016). "Revealed Preference, Rational Inattention, and Costly Information Acquisition," *American Economic Review*, forthcoming.

[12] Dickhaut, J, A. Rustichini and V. Smith (2009). "A Neuroeconomic Theory of the Decision Process," *Proceedings of the National Academy of Sciences*, 106.52:22146–22150.

[13] Fudenberg D. and T. Strzalecki (2015). "Dynamic Logit with Choice Aversion," *Econometrica*, 83.2:651–691.

[14] Fudenberg D., R. Iijima and T. Strzalecki (2015). "Stochastic Choice and Revealed Perturbed Utility," *Econometrica*, 83.6:2371–2409.

[15] Gul, F. and W. Pesendorfer (2006). "Random Expected Utility," *Econometrica*, 74.1:121–146.

[16] Gul, F., P. Natenzon, and W. Pesendorfer (2014). "Random Choice as Behavioral Optimization," *Econometrica*, 82.5:1873–1912.

[17] Harless, D.H. and C.F. Camerer (1994). "The Predictive Utility of Generalized Expected Utility Theories," *Econometrica*, 62.6:1251–1289.

[18] Lu, J. and K. Saito (2017). "Random Intertemporal Choice," mimeo.

[19] Luce, R.D. (1959). "Individual Choice Behavior: A Theoretical Analysis," New York, New York: Wiley.

[20] Manzini, P. and M. Mariotti (2014). "Stochastic Choice and Consideration Sets," *Econometrica*, 82(3):1153–1176.

[21] Mosteller, F. and P. Nogee (1951). "An Experimental Measurement of Utility," *Journal of Political Economy*, 59:371–404.

[22] Natenzon, P. (2017). "Random Choice and Learning," *Journal of Political Economy*, forthcoming.

[23] Webb, Ryan (2017). "The (Neural) Dynamics of Stochastic Choice," *Management Science*, forthcoming.