Journal of Mathematical Psychology 57 (2013) 78–93

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Prior distributions for random choice structures

William J. McCausland [a,*,1], A.A.J. Marley [b,1]

[a] *Département de sciences économiques, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Québec H3C 3J7, Canada*
[b] *Department of Psychology, University of Victoria, Canada*

HIGHLIGHTS

- We develop a class of prior distributions on sets of discrete choice probabilities.
- We measure the strength of discrete choice axioms by their prior improbability.
- Prior analysis gives limits on how much support data can give a particular axiom.
- We set up a testing ground for extant and novel axioms of discrete choice.
- We argue for priors over all choice probabilities, even for binary choice data.

ARTICLE INFO

Article history:
Received 16 August 2012
Available online 14 June 2013

Keywords:
Bayesian inference
Choice axioms
Discrete choice
Probabilistic choice
Random utility

ABSTRACT

We study various axioms of discrete probabilistic choice, measuring how restrictive they are, both alone and in the presence of other axioms, given a specific class of prior distributions over a complete collection of finite choice probabilities. We do this by using Monte Carlo simulation to compute, for a range of prior distributions, probabilities that various simple and compound axioms hold. For example, the probability of the triangle inequality is usually many orders of magnitude higher than the probability of random utility. While neither the triangle inequality nor weak stochastic transitivity imply the other, the conditional probability that one holds given the other holds is greater than the marginal probability, for all priors in the class we consider. The reciprocal of the prior probability that an axiom holds is an upper bound on the Bayes factor in favor of a restricted model, in which the axiom holds, against an unrestricted model. The relatively high prior probability of the triangle inequality limits the degree of support that data from a single decision maker can provide in its favor. The much lower probability of random utility implies that the Bayes factor in favor of it can be much higher, for suitable data.

1. Introduction

1.1. Setting

Standard models for discrete choice specify choice probabilities $P_A(x)$, the probability of choosing a single option x from a set of available options A, for all $x \in A \subseteq T$, where T is a finite master set (or universe) of objects. We will assume for the sake of definiteness that these choice probabilities describe the choice behavior of a single agent; alternatively we could interpret them as choice probabilities of agents drawn at random from some population. An agent's choices are supposed to be statistically independent across choice situations—thus the probability of choosing x from A and then y from B is $P_A(x)P_B(y)$. We call the complete specification of the $P_A(x)$, for all non-empty $A \subseteq T$ and all $x \in A$, a *random choice structure* and denote it (T, P). This framework assumes that the agent is required to select exactly one of the available options from each presented choice set; however, one can also consider an experimental design where the agent has the additional option to select none of the available options, as in Corbin and Marley (1974); or, in the case of two options, to state indifference between the options, as in Davis-Stober (2012) and Regenwetter and Davis-Stober (2008).

Various axioms, conditions, properties and hypotheses about probabilistic choice behavior can be expressed as restrictions over the various choice probabilities. This is true of seven axioms we explicitly study in this paper: weak, moderate and strong stochastic transitivity, the triangle inequality, the multiplicative inequality, regularity and the random ranking hypothesis. The random ranking hypothesis is particularly important; it is equivalent to what is commonly known in Economics and Marketing as random utility, and is widely used there—see Appendix A for a discussion.

[*] Corresponding author.
E-mail address: william.j.mccausland@umontreal.ca (W.J. McCausland).

[1] Marley is Distinguished Professor (part time) and McCausland is External Affiliate of the Centre for the Study of Choice, University of Technology Sydney, Australia.

0022-2496/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmp.2013.05.001

1.2. Empirical testing

Over the years, there have been attempts to test whether or not specific behavioral properties hold in discrete choice data, the most famous being Tversky's (1969) study of (weak stochastic) transitivity. Frequentist approaches to the analysis of such data are challenging, due to the difficulty of deriving sampling distributions when the parameter space is truncated. Iverson and Falmagne (1985) is one early attempt to deal with the relevant issues. Very recently, much more satisfactory frequentist tools have been developed, which, in combination with relevant results on polytopes, give better statistical analyses for old and new data—see Regenwetter, Dana, and Davis-Stober (2011). These analyses strongly suggest that Tversky's original rejection of weak stochastic transitivity was premature; see Birnbaum (2011) for a discussion of the application of his *pattern model* to similar data and Regenwetter, Dana, Davis-Stober, and Guo's (2011) response to that approach.

Works describing Bayesian tests of various axioms for binary choice probabilities include Myung, Karabatsos, and Iverson (2005), some of the articles cited there, Cavagnaro and Davis-Stober (2013) and Zwilling, Cavagnaro, and Regenwetter (2011). These involve priors over binary choice probabilities. The prior we develop in this paper is a joint distribution of choice probabilities on all non-empty subsets of a finite master set, not just the doubletons. We will argue in the conclusions that this is desirable for tests of the random ranking hypothesis, even when the available data consist only of binary choice data.

1.3. Bayesian inference

We outline some standard concepts in Bayesian theory that will be useful in this paper. For further reading, see Berger (1985) and Bernardo and Smith (1994) and the references therein.

Bayesian methods are quite attractive in situations where we want to impose or test inequality restrictions on parameters or latent variables. We can easily impose these restrictions by truncating the prior distribution of parameters or the distribution of latent variables—direct application of Bayes' rule implies that the posterior for the restricted model is simply the truncation of the posterior for the unrestricted model to the same region.

Bayesian inference requires the full specification of the joint distribution of data (denoted y) and any unknown quantities of the model. The distinction between unknown parameters $\theta \in \Theta$ and latent variables $\gamma \in \Gamma$ is not important in Bayesian theory, but may be useful for expositional reasons. We usually express the complete model as $f(\theta, \gamma, y) = f(\theta)f(\gamma|\theta)f(y|\theta, \gamma)$, where $f(\theta)$ is the prior density for parameters, $f(\gamma|\theta)$ describes the latent part of a model and the data density $f(y|\theta, \gamma)$ describes observables.

We can also treat uncertainty about models in the same way as uncertainty about any other unknowns. Suppose we have two competing models M_1 and M_2. The first model gives $f(\theta_1, \gamma_1, y|M_1)$ and the second gives $f(\theta_2, \gamma_2, y|M_2)$. The posterior probabilities of the two models give their relative degrees of plausibility in the light of the data. The ratio of the two probabilities, the posterior odds ratio, is given by Bayes' rule as

$$\frac{\Pr[M_1|y]}{\Pr[M_2|y]} = \frac{\Pr[M_1]}{\Pr[M_2]} \cdot \frac{f(y|M_1)}{f(y|M_2)}.$$

The first ratio on the right hand side is the prior odds ratio. The second, denoted B_{12}, is the *Bayes factor* in favor of M_1 over M_2. Evaluated for the same given observed data y, $f(y|M_1)$ and $f(y|M_2)$ are the *marginal likelihoods* for the two models, and we can write them as

$$f(y|M_i) \equiv \int f(y|\theta_i, \gamma_i, M_i)f(\theta_i|M_i)f(\gamma_i|\theta_i, M_i)\, d\theta_i\, d\gamma_i, \quad i = 1, 2.$$

While for given data y, the likelihoods $f(y|M_1, \theta_1)$ and $f(y|M_2, \theta_2)$ are functions of the unknown parameters, the marginal likelihoods are known numbers. We can think of the maximum likelihood values $\max_{\theta_1 \in \Theta_1} f(y|M_1, \theta_1)$ and $\max_{\theta_2 \in \Theta_2} f(y|M_2, \theta_2)$ as measures of in-sample fit — they show how well the model predicts the data for parameters that we choose after observing the data. The marginal likelihoods, which make no reference to unknown quantities, are measures of out-of-sample fit — they average the likelihood over the prior, with the prior chosen before we observe data. The Bayes factor, then, compares the out-of-sample prediction records of the two models to measure the relative support that the data give M_1 compared with M_2.

We plan to investigate how restrictive axioms are — alone and in the presence of other axioms. This exercise relates to Bayes factors. Suppose we want to see how plausible an axiom is in the light of the data. We have an unrestricted model M_2 for choice probabilities, with parameters $\theta \in \Theta$ and latent variables $\gamma \in \Gamma$. Suppose further that a behavioral axiom can be expressed as a restriction on the space Γ of latent variables — choice probabilities satisfy the axiom for all $\gamma \in \Gamma_r \subseteq \Gamma$ and violate the axiom for all $\gamma \in \Gamma_r^c$, where Γ_r^c is the complement of Γ_r in Γ. Also suppose that the prior probability of the restriction holding in the unrestricted model is strictly positive. This will be the case for all the axioms we consider in this paper.

We can compare the unrestricted model with a restricted model M_1, which differs from M_2 in one respect: $f(\gamma|\theta, M_1)$ is the normalized truncation of $f(\gamma|\theta, M_2)$ to Γ_r. That is,

$$f(\gamma|\theta, M_1) = \begin{cases} \dfrac{f(\gamma|\theta, M_2)}{\int_{\Gamma_r} f(\gamma'|\theta, M_2)\, d\gamma'} & \gamma \in \Gamma_r \\ 0 & \text{otherwise.} \end{cases}$$

For this special case, a simple application of Bayes' rule gives the Bayes factor as

$$B_{12} = \frac{f(y|\gamma \in \Gamma_r, M_2]}{f(y|M_2)} = \frac{\Pr[\gamma \in \Gamma_r|y, M_2]}{\Pr[\gamma \in \Gamma_r|M_2]}. \quad (1)$$

The right hand side of the second equality is the ratio of the posterior to prior probability that the restriction holds in the unrestricted model. The Bayes factor 'rewards' an axiom for being consistent with data, but also for making the restricted model 'small', 'simple' or 'parsimonious', as measured by the axiom's low prior probability. The prior probability of a restriction gives an upper bound on the Bayes factor in favor of the restricted model—it can be no higher than $\Pr[\gamma \in \Gamma_r|M_2]^{-1}$, no matter how much data we collect, since the posterior probability $\Pr[\gamma \in \Gamma_r|y, M_2]$ can be no higher than 1.

As a practical matter, the computation of marginal likelihoods is often difficult. Bos (2002) describes and compares several approaches used in the literature, documenting some of these difficulties. Analytic integration is usually out of the question, and for generic simulation approaches it is difficult to obtain numerical standard errors that are small enough to be practical.

Approximating the Bayes factor in (1), however, is relatively easy, partly because the prior and posterior probabilities of axioms can be estimated in simulations as sample means of indicator functions, which are bounded. In the present paper, we compute prior probabilities of the form $\Pr[\gamma \in \Gamma_r|M_2]$ – the numerator in (1) – using independence Monte Carlo. We have no occasion in this paper to compute Bayes factors. In McCausland and Marley (2013), we address the practical problem of computing the denominator $\Pr[\gamma \in \Gamma_r|y, M_2]$. There, too, we use Monte Carlo, but since we know of no method for making independent draws from the posterior, we resort to Markov chain Monte Carlo (MCMC) methods. There is a very large literature on MCMC; popular texts include Gilks, Richardson, and Spiegelhalter (1996) and Robert and Casella (2010).

1.4. Outline of paper

In Section 2, we define a finite random choice structure, a flexible non-parametric framework for probabilistic discrete choice behavior over all non-empty subsets of a finite set T. We discuss several axioms in the literature governing random choice and illustrate some of the logical relations among them. In Section 3, we set up a probabilistic framework allowing us to measure how restrictive various simple and composite axioms are. This framework takes the form of a family of prior distributions over the set of random choice structures. We demonstrate several attractive features of the prior. In Section 4, we show the results of prior simulation exercises. Given a particular prior, we measure how restrictive an axiom is by the implied prior (im)probability of the axiom holding. In Section 5 we conclude and discuss extensions. As already noted, we focus on the random ranking hypothesis as an important example, but the techniques apply broadly, to various hypotheses and axioms, either extant or yet to be proposed.

2. Random choice: definitions, axioms and theorems

We first introduce some preliminary definitions. We will often refer to a finite *master set* $T \equiv \{x_1, \ldots, x_n\}$ of n choice objects. The order $x_1, \ldots, x_n$ is arbitrary, but will be useful to establish notation.

A *finite random choice structure* is a pair (T, P), where T is a finite set and the collection $P_A\colon T \to [0, 1]$, $\emptyset \neq A \subseteq T$, satisfies

(1) $\sum_{x\in A} P_A(x) = 1$ for any non-empty $A \subseteq T$.
(2) $P_A(x) = 0$ for any $A \subseteq T, x \notin A$.

We interpret $P_A(x)$ as the probability that an agent chooses object x when presented with choice set A, and suppose that choices are statistically independent across choice situations. For distinct $x, y \in T$, we use the standard shorthand notation $p(x, y)$ to mean $P_{\{x,y\}}(x)$.

For a given random choice structure (T, P) and non-empty $T' \subseteq T$, we define (T', P'), the *restriction* of (T, P) to T', as the random choice structure on T' such that for all $A \subseteq T'$, $P_A(\cdot) = P'_A(\cdot)$.

It will be helpful to represent random choice structures on the tripleton master set $T = \{x, y, z\}$ graphically. Fig. 1 gives, as an example, a representation of the random choice structure (T, P) consisting of the binary probabilities $p(x, y) = 0.4$, $p(y, x) = 0.6$, $p(y, z) = 0.8$, $p(z, y) = 0.2$, $p(x, z) = 0.7$ and $p(z, x) = 0.3$; and the ternary probabilities $P_T(x) = 0.4$, $P_T(y) = 0.2$ and $P_T(z) = 0.4$.

The triangle with vertices x, y and z is equilateral. The figure represents (T, P) by four points in the Barycentric coordinate system with respect to x, y and z. There is one point for each of the non-singleton choice sets. The vertices x, y and z have Barycentric coordinates $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$, respectively. We can also give Euclidean coordinates—taking the midpoint of the triangle's base as the origin and the height of the triangle as one unit, the Euclidean coordinates of the vertices x, y and z are $(0, 1)$, $(-\frac{1}{\sqrt{3}}, 0)$ and $(\frac{1}{\sqrt{3}}, 0)$.

The solid dot in the interior of the triangle represents the vector $(P_T(x), P_T(y), P_T(z)) = (0.4, 0.2, 0.4)$ of ternary probabilities. The point is a convex combination of x, y and z – in both Euclidean and Barycentric spaces – with weights 0.4, 0.2 and 0.4, respectively. The point is a Euclidean distance $P_T(x) = 0.4$ from the base of the triangle, $P_T(z) = 0.4$ from the left side and $P_T(y) = 0.2$ from the right. By this convention, the vertices x, y and z represent the degenerate distributions on T where objects x, y and z are chosen with probability one, respectively, from the choice set T.

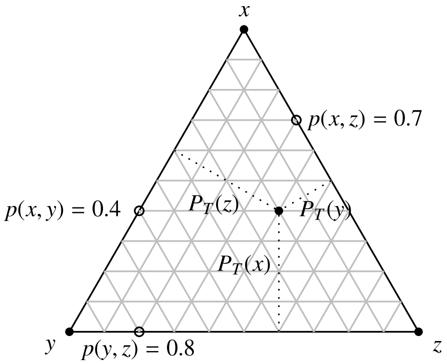

Fig. 1. Graphical illustration of a random choice structure on $T = \{x, y, z\}$.

The hollow dots on the left, right and bottom sides of the triangle represent the binary probabilities $p(x, y) = 0.4$, $p(x, z) = 0.7$ and $p(y, z) = 0.8$, respectively. For example, the dot on the left side of the triangle is the convex combination of the vertices labeled x and y, with weights $p(x, y) = 0.4$ and $p(y, x) = 0.6$ respectively. Throughout, we reserve hollow dots for binary probabilities and solid dots for ternary probabilities, to avoid any ambiguity for points on the boundary of the triangle.

We now define random rankings and consider the question of whether a given random choice structure (T, P) can be induced by some distribution over rankings.

For a given set $T = \{x_1, \ldots, x_n\}$, let $R(T)$ be the set of rankings (strict linear orders) on T. For notational convenience and without loss of generality, we write $R(T) = \{\succ_1, \ldots, \succ_{n!}\}$, with rankings in the lexicographic order where x_n changes position first and x_1 changes position last. For any ranking $\succ \in R$ and non-empty subset $A \subseteq T$, let $h_\succ(A)$ be the highest $\succ$-ranked object in A.

A *ranking distribution* is a pair (T, Π), where the master set T is finite and Π is a probability mass function on $R(T)$. Thus $\Pi(\succ)$ is the probability that ranking $\succ$ obtains. For a given ranking distribution (T, Π) and non-empty subset $T' \subseteq T$, we define (T', Π'), the *marginalization* of (T, Π) to T', as the ranking distribution that assigns probability to a ranking $\succ'$ on T' equal to the probability Π assigns to the set of rankings on T consistent with $\succ'$. That is, (T', Π') is the marginalization of (T, Π) to T' if for all $\succ' \in R(T')$,

$$\Pi'(\succ') = \sum_{\{\succ\in R(T):\forall x,y\in T', x\succ y\Rightarrow x\succ' y\}} \Pi(\succ).$$

For any ranking distribution (T, Π), we define (T, P^Π) as the random choice structure such that for all non-empty $A \subseteq T$, and all $x \in A$,

$$P^\Pi_A(x) = \sum_{\{\succ\in R(T):h_\succ(A)=x\}} \Pi(\succ).$$

We say that a ranking distribution (T, Π) *induces* a random choice structure (T, P) if $P = P^\Pi$.

2.1. Axioms for finite random choice structures

Numerous axioms (behavioral restrictions) have been proposed to constrain choice probabilities or to illustrate properties of other axioms through their logical relationships. We consider the following list of axioms.

The random choice structure (T, P) satisfies

TI: the *triangle inequality* if and only if for all distinct x, y, and z,

$p(x, y) + p(y, z) + p(z, x) \geq 1$,

Reg: *regularity* if and only if for all $A, B \subseteq T$ and for all $x \in A$,

$P_A(x) \geq P_{A\cup B}(x)$

RR: *the random ranking hypothesis* if there is some ranking distribution (T, Π) that induces it.

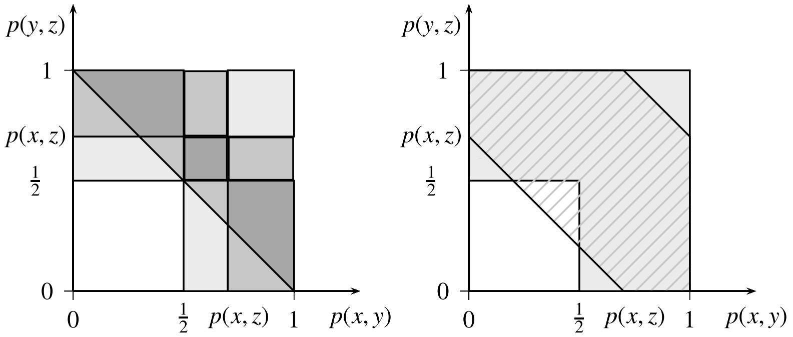

Fig. 2. Weak, moderate and strong stochastic transitivity and the triangle inequality. Both graphs indicate the sets of $(p(x,y), p(y,z))$ pairs that are consistent with various axioms, for a particular fixed value of $p(x,z)$ greater than 1/2. In the left graph, there are four regions: white indicates the region where none of WST, MST or SST are satisfied; light grey, the region where only WST is satisfied; medium grey, the region where only WST and MST are satisfied and dark grey where all versions of stochastic transitivity are satisfied. In the right graph, the region where WST is satisfied is shown in light grey; the region where TI is satisfied is hatched.

WST: *weak stochastic transitivity* if and only if for all distinct x, y, and z,

$$p(x,y) \geq \frac{1}{2} \quad \text{and} \quad p(y,z) \geq \frac{1}{2} \Rightarrow p(x,z) \geq \frac{1}{2},$$

MST: *moderate stochastic transitivity* if and only if for all distinct x, y, and z,

$$p(x,y) \geq \frac{1}{2} \quad \text{and} \quad p(y,z) \geq \frac{1}{2}$$
$$\Rightarrow p(x,z) \geq \min[p(x,y), p(y,z)],$$

SST: *strong stochastic transitivity* if and only if for all distinct x, y, and z,

$$p(x,y) \geq \frac{1}{2} \quad \text{and} \quad p(y,z) \geq \frac{1}{2}$$
$$\Rightarrow p(x,z) \geq \max[p(x,y), p(y,z)],$$

MI: the *multiplicative inequality* if and only if for all $A, B \subseteq T$ and all $x \in A \cap B$,

$$P_{A\cup B}(x) \geq P_A(x) \cdot P_B(x).$$

For MI, see Colonius (1983), Sattath and Tversky (1976) and Suck (2002). For the remaining conditions, see Luce and Suppes (1965). MI involves choice sets of different sizes; the multiplication *condition* in Luce and Suppes (1965) is a different axiom, involving only binary choice probabilities.

2.2. Known theorems for random choice

The following diagram illustrates the logical relationship between these axioms. The presence of a directed arrow from one axiom to another means that the first implies the second. The absence of a directed arrow means that there is no such implication.

$$\begin{tikzcd}
\text{SST} \arrow[r, Rightarrow] & \text{MST} \arrow[r, Rightarrow] \arrow[rd, Rightarrow] & \text{WST} \\
\text{RR} \arrow[r, Rightarrow] & \text{Reg} \arrow[r, Rightarrow] & \text{TI} \\
 & \text{MI} &
\end{tikzcd}$$

So, for example, regularity is necessary but not sufficient for the random ranking hypothesis. Neither the triangle inequality nor weak stochastic transitivity imply the other.

The diagram gives an exhaustive list of theorems for these axioms, in the sense that the various axioms can be satisfied or not in any combination not ruled out by these logical implications. We know this because in the simulations reported below, every such combination occurred. We can think of these occurrences as counterexamples to other candidate theorems for the axioms. So for example, we know the candidate theorem

Reg and not MI ⇒ MST

is not true because we generated counterexamples, random choice structures where Reg holds and MI and MST do not. Not all of these counterexamples were previously explicitly known.

The implications SST ⇒ MST ⇒ WST are obvious from the definitions of these axioms. The implication MST ⇒ TI becomes obvious when we write the MST and TI conditions as in Appendices C.3.1 and C.3.2. Fig. 2 graphically illustrates the constraints on binary probabilities implied by WST, MST, SST and TI. Points in various regions constitute counterexamples for all other candidate theorems involving WST, MST, SST and TI, such as WST ⇒ MST.

Fig. 3 illustrates the relationship between binary choice probabilities, TI and Reg. In the left panel, the three binary choice probabilities $p(x,y) = 0.6$, $p(z,x) = 0.7$ and $p(y,z) = 0.65$ are fixed. Upward sloped hatching indicates the region where the ternary probability vector $(P_T(x), P_T(y), P_T(z))$ is consistent with two necessary conditions for regularity implied by $p(x,y) = 0.6$: $P_T(x) \leq p(x,y) = 0.6$ and $P_T(y) \leq p(y,x) = 0.4$. Downward sloped hatching indicates the region where ternary probabilities are consistent with regularity and the fixed value $p(z,x) = 0.7$; horizontal hatching, the region consistent with regularity and $p(y,z) = 0.65$. Cross-hatching indicates that two pairs of necessary conditions for regularity hold. The black triangle is the intersection of the three regions, the region where all conditions for regularity are satisfied.

In the left panel, the binary probabilities exhibit a cycle, in the sense that WST is violated — we have $p(x,y) > 0.5$, $p(y,z) > 0.5$ and $p(z,x) > 0.5$. However, TI is satisfied, since $p(x,y) + p(y,z) + p(z,x) = 1.95 \geq 1$ and $p(z,y) + p(y,x) + p(x,z) = 1.05 \geq 1$. Ternary probabilities outside the black triangle give counterexamples to TI ⇒ Reg. But Reg is not ruled out — the ternary probabilities in the black triangle are consistent with Reg and the given binary probabilities.

When TI is satisfied in this example, we will call the minimum of $p(x,y) + p(y,z) + p(z,x) - 1$ and $p(z,y) + p(y,x) + p(x,z) - 1$ the amount of *slack*. Now imagine the gradual increase of one or more of the probabilities $p(x,y)$, $p(y,z)$ and $p(z,x)$. The amount of slack decreases and the black triangle shrinks. Eventually, the TI condition is violated and the black triangle vanishes. The right panel shows the case where $p(z,x)$ has changed from 0.7 to 0.8. The binary probabilities are now inconsistent with TI, since $p(z,y) + p(y,x) + p(x,z) = 0.95 < 1$. The intersection of the hatched regions is now the empty set—no ternary probabilities are consistent with Reg and the new binary probabilities.

For RR ⇒ Reg ⇒ TI, see Luce and Suppes (1965). Falmagne (1978) gives a set of conditions on choice probabilities that is necessary and sufficient for RR. Fiorini (2004) gives an alternate

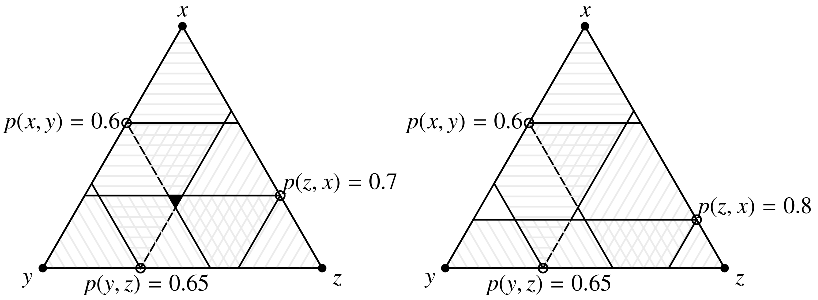

Fig. 3. Restrictions on regularity.

proof and identifies a strict subset of these conditions such that each condition is indispensable. These conditions are as follows: for all non-empty $A \subseteq T$ and all $x \in A$,

$$\sum_{B:A\subseteq B\subseteq T} (-1)^{|B\setminus A|} P_B(x) \geq 0. \tag{2}$$

For master sets of size $n = 3$, these conditions are identical to the set of regularity conditions and so Reg and RR are equivalent for this case. See McFadden and Richter (1990) for a counterexample to Reg ⇒ RR when the master set has size $n = 4$.

When the master set has no more than five elements, TI is necessary and sufficient for the set of binary choice probabilities to be consistent with RR. See Dridi (1980) for a proof, Koppen (1995) and literature cited there for additional necessary conditions for binary RR when the master set has more than five elements.

However, consider the discussion of Fig. 3 above. For fixed binary choice probabilities, TI is either satisfied or it is not. If it is, the volume of the region of non-binary choice probabilities consistent with Reg may be greater or smaller, according to the given binary choice probabilities. Intuitively, the amount of slack in the TI conditions plays a role: when there is little slack, the set of non-binary choice probabilities consistent with Reg is small. Since Reg is necessary for RR, the set consistent with RR will be just as small or smaller. In this sense, binary choice probabilities that just satisfy TI, together with the hypothesis of RR, imply strong predictions about non-binary choice probabilities. For reasonable prior distributions over the space of random choice structures, the posterior probability of RR will thus vary greatly over binary choice data sets whose choice frequencies are consistent with TI—the posterior probability will be lower if the binary choice frequencies are near the boundary of TI.

Fig. 4 shows the relationship between Reg and MI. As Sattath and Tversky (1976) note, Reg and MI are complementary, in the sense that they give upper and lower bounds, respectively, for choice probabilities. In the context $T = \{x, y, z\}$, we can write out the three inequalities in the definition of MI and the six inequalities in the definition of Reg, then combine them to obtain

$$p(x, y)p(x, z) \leq P_T(x) \leq \min[p(x, y), p(x, z)],$$
$$p(y, x)p(y, z) \leq P_T(y) \leq \min[p(y, x), p(y, z)],$$
$$p(z, x)p(z, y) \leq P_T(z) \leq \min[p(z, x), p(z, y)].$$

Fig. 4 illustrates the case where $p(x, y) = 0.6$, $p(y, z) = 0.6$ and $p(x, z) = 0.8$. The region of ternary probabilities consistent with Reg is the familiar equilateral triangle, here indicated with upward sloped hatching. The region of ternary probabilities consistent with MI is the intersection of the sets defined by the inequalities $P_T(x) \geq p(x, y)p(x, z) = 0.48$, $P_T(y) \geq p(y, x)p(y, z) = 0.24$ and $P_T(z) \geq p(z, x)p(z, y) = 0.08$. This region is the equilateral triangle with downward sloped hatching. The union of the regions where Reg and MI hold is an irregular six-pointed star; their intersection is an irregular hexagon. Points in various regions of the diagram give counterexamples to the candidate theorems MI ⇒ Reg, Reg ⇒ MI, MI ⇒ TI, and TI ⇒ MI. Since Reg and RR are equivalent for $n = 3$, some also give counterexamples for MI ⇒ RR and RR ⇒ MI.

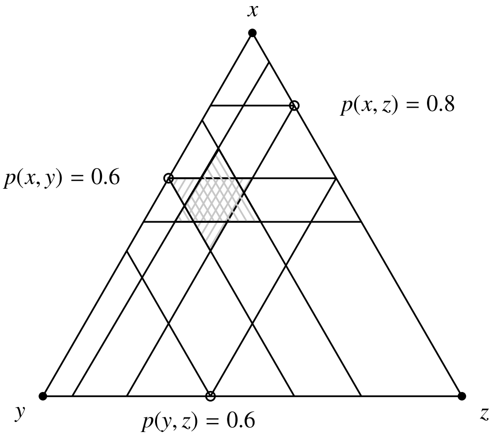

Fig. 4. Regularity and the multiplicative inequality.

MI is related to random utility and independent random utility, as they are defined in Appendix A. Sattath and Tversky (1976) show that MI is necessary for an independent random utility model and also show that Tversky's (1972) *elimination by aspects (EBA)* model satisfies MI. Tversky (1972, Theorem 7) shows that the EBA model is a random utility model and gives a counterexample showing that the EBA model is not necessarily an independent random utility model.

3. Prior distributions over random choice structures

Here we describe a parametric class of prior distributions over the set of random choice structures on a master set T. We formulated this class to have certain desirable properties, and we demonstrate that these properties indeed hold. Our class generalizes prior distributions in the literature, notably by allowing statistical dependence across choice probabilities $P_A(\cdot)$, $\emptyset \neq A \subseteq T$. However, it is not fully general, and we discuss below some possible extensions.

We begin by describing a class of prior distributions over the set of ranking distributions on T. We then extend this to a class of prior distributions on the entire set of random choice structures on T. Finally, we describe some properties of this class of priors. We make use of Gamma and Dirichlet distributions, which have some very useful properties. Appendix B lists some of these; for details, see Forbes, Evans, Hastings, and Peacock (2011).

3.1. Prior distributions over random choice structures induced by random rankings

A ranking distribution (T, Π) gives a probability distribution on the finite set $R(T)$ of rankings. This makes a prior distribution on the set of ranking distributions a distribution over distributions on $R(T)$. We will use a symmetric Dirichlet distribution for our prior, with a Dirichlet weight of $\alpha/n!$ for each ranking, where $\alpha > 0$ is a scalar parameter. This gives the prior distribution of the vector $(\Pi(\succ_1), \dots, \Pi(\succ_{n!}))$, where each $\Pi(\succ_i)$, $i = 1, \dots, n!$, is the probability of ranking $\succ_i$, as

$$(\Pi(\succ_1), \dots, \Pi(\succ_{n!})) \sim \mathrm{Di}\left(\frac{\alpha}{n!}, \frac{\alpha}{n!}, \dots, \frac{\alpha}{n!}\right),$$

where $\mathrm{Di}(\cdot)$ denotes the Dirichlet distribution. We will use the notation $H(\alpha, T)$ to denote this prior distribution and $(T, \Pi) \sim H(\alpha, T)$ to mean that the ranking distribution (T, Π) is drawn from this distribution.

We will now describe how to (randomly) construct a ranking distribution $(T, \Pi) \sim H(\alpha, T)$, for a given value of α. We first draw independent and identically distributed weights $\gamma(\succ)$, $\succ \in R(T)$:

$$\gamma(\succ) \sim \mathrm{Ga}\left(\frac{\alpha}{n!}, 1\right),$$

where $\mathrm{Ga}(\cdot, \cdot)$ denotes the Gamma distribution, and then construct probabilities $\Pi(\succ) = \gamma(\succ)/G$, where

$$G \equiv \sum_{\succ' \in R(T)} \gamma(\succ'). \tag{3}$$

We can think of the weights $\gamma(\succ)$ as latent variables, with the probabilities of rankings and (below) induced choice probabilities a deterministic function of them.

The parameter $\alpha \in (0, \infty)$ governs how close a ranking distribution $(T, \Pi) \sim H(\alpha, T)$ is likely to be to a degenerate distribution. For small α, the $\gamma(\succ)$ are likely to be very dissimilar. This means that for small α, a ranking distribution (T, Π) drawn from $H(\alpha, T)$ is likely to assign probability close to unity to one of the rankings in $R(T)$ and very low probabilities to all others. In the limit $\alpha = 0$, a ranking distribution (T, Π) drawn from $H(\alpha, T)$ is degenerate: it puts probability one on some ranking and probability zero on the rest. For large α, the $\gamma(\succ)$ are likely to be similar, which means that a ranking distribution (T, Π) drawn from $H(\alpha, T)$ is likely to assign probabilities close to $1/n!$ to all $n!$ rankings. In the limit $\alpha = \infty$, a ranking distribution drawn from $H(\alpha, T)$ assigns probability $1/n!$ to all $n!$ rankings in $R(T)$. See Kotz, Balakrishnan, and Johnson (2000, chapter 49) for limiting properties of the Dirichlet distribution.

Our prior has the following marginalization property. Take any non-empty $T' \subseteq T$ and let (T', Π') be the marginalization of (T, Π) to T'. Then $(T, \Pi) \sim H(\alpha, T)$ implies $(T', \Pi') \sim H(\alpha, T')$. This is a fairly direct application of the aggregation properties of independent Gamma random variables.

3.2. *Prior distributions over random choice structures*

Before defining a class of distributions over the set of random choice structures, we consider for a moment the implied distribution over random choice structures induced by the distribution $H(\alpha, T)$ over ranking distributions. We note four properties of this implied distribution. First, it assigns a probability of zero to the set of random choice structures that cannot be induced by a random ranking. This is true by construction. Second, the support of the implied distribution is the entire set of random choice structures that are induced by a random ranking. This follows from the fact that the Dirichlet distribution has full support on the simplex of ranking probabilities—see the expression for the Dirichlet density in Appendix B and the reference given there. Third, two choice distributions $P_A(\cdot)$ and $P_B(\cdot)$ will be statistically dependent whenever $A \cap B \neq \emptyset$. This is because they share common $\gamma(\succ)$ terms.

Finally, we can easily derive implied marginal distributions over choice probabilities. Take any choice set $A \subseteq T$ and order the elements of A as $x_1, \dots, x_{|A|}$. Then

$$\left(P_A(x_1), \dots, P_A(x_{|A|})\right) = \left(\frac{\gamma_A(x_1)}{G}, \dots, \frac{\gamma_A(x_{|A|})}{G}\right), \tag{4}$$

where the total weight G is given by (3) and where for every $x \in A$, $\gamma_A(x) \equiv \sum_{\{\succ \in R(T): h_\succ(A) = x\}} \gamma(\succ)$.

Each component $\gamma_A(x_i)$ in (4) is the sum of $n!/|A|$ independent Gamma random variables with shape parameter $\alpha/n!$. There are $|A|$ of them, they are mutually independent and they add to G. By a well known property – see Appendix B.3 – of the Gamma and Dirichlet distributions, we have

$$\left(P_A(x_1), \dots, P_A(x_{|A|})\right) \sim \mathrm{Di}\left(\frac{\alpha}{|A|}, \dots, \frac{\alpha}{|A|}\right),$$
$$\emptyset \neq A \subseteq T. \tag{5}$$

We now define a parametric class of distributions over the set of random choice structures on T. There are two parameters, $\alpha > 0$ and $\lambda \in [0, 1]$, and we use the notation $H(\alpha, \lambda, T)$ to denote the distribution for given parameter values. We will see that for all values $\alpha > 0$ and $\lambda \in [0, 1)$, the support of $H(\alpha, \lambda, T)$ is the entire set of random choice structures. For $\lambda = 1$, and any $\alpha > 0$, we have a prior distribution restricted to the set of random choice structures that are induced by a random ranking: $H(\alpha, 1, T)$ is the distribution of P^Π induced by the distribution $H(\alpha, T)$ over ranking distributions.

The parameter $\alpha > 0$ has a similar interpretation to the one it has in Section 3.1, and we will see that for all non-empty $A \subseteq T$, the marginal distribution of $P_A(\cdot)$ implied by $H(\alpha, \lambda, T)$ is the same as that implied by $H(\alpha, T)$, namely the distribution given by (5). It governs how close choice distributions are likely to be to degenerate choice distributions. In the limit $\alpha = 0$, all choice distributions are degenerate: for a given draw of $(T, P) \sim H(\alpha, \lambda, T)$, some element of each choice set is chosen with probability one. The element chosen with probability one will vary from one draw of $(T, P) \sim H(\alpha, \lambda, T)$ to another. In the limit $\alpha = \infty$, all choice distributions are uniform discrete distributions on their support.

Fig. 5 shows the marginal density of any binary choice probability, for various values of α. For $\alpha < 2$, the density is unbounded at zero and one, and there is a lot of probability mass near these extreme points. For $\alpha = 2$, the density is uniform; for $\alpha > 2$, the density is zero at the two extreme points. In the absence of much prior information to the contrary, we would recommend values of α less than two; standard approaches for the specification of so-called non-informative priors include those of Bernardo (1979) and Jeffreys (1946), both of which lead to a choice of $\alpha = 1$. The approach of Haldane (1948) gives $\alpha = 0$; the resulting density is improper.

Parameter $\lambda \in [0, 1]$ governs the degree of prior dependence between choice distributions on different choice sets. At the extreme $\lambda = 0$, choice distributions are *a priori* independent, with marginals given by (5).

We define the distribution $H(\alpha, \lambda, T)$ indirectly, by describing a random construction of P, the set of choice probability distributions. Suppose λ and α are now given. For each ranking $\succ \in R(T)$, we have a latent weight $\gamma(\succ)$, and for each non-empty $A \subseteq T$ and ranking $\succ \in R(A)$ we have a latent weight $\tilde{\gamma}_A(\succ)$. They are all mutually independent, with distributions

$$\gamma(\succ) \sim \mathrm{Ga}\left(\frac{\alpha\lambda}{n!}, 1\right), \quad \succ \in R(T), \tag{6}$$

and

$$\tilde{\gamma}_A(\succ) \sim \mathrm{Ga}\left(\frac{\alpha(1-\lambda)}{|A|!}, 1\right), \quad \succ \in R(A), \emptyset \neq A \subseteq T. \tag{7}$$

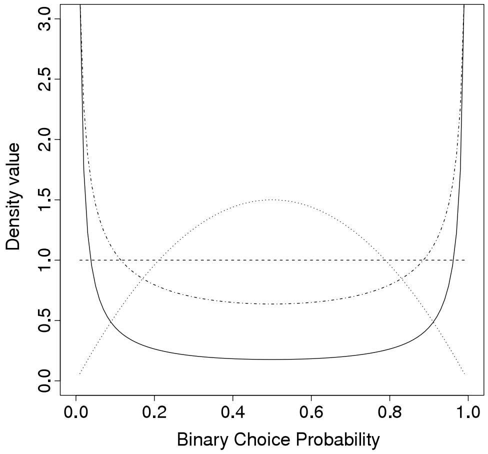

Fig. 5. Marginal density of any binary choice probability for $\alpha = 0.2$ (solid), $\alpha = 1.0$ (dot-dashed), $\alpha = 2.0$ (dashed), $\alpha = 4.0$ (dotted).

We construct choice probabilities from these weights as follows. For each non-empty $A \subseteq T$ and $x \in A$,

$$P_A(x) = \frac{\gamma_A(x) + \tilde{\gamma}_A(x)}{G + \tilde{G}_A}, \tag{8}$$

where $\gamma_A(x) \equiv \sum_{\{\succ \in R(T): x = h_\succ(A)\}} \gamma(\succ)$, $\tilde{\gamma}_A(x) \equiv \sum_{\{\succ \in R(A): x = h_\succ(A)\}} \tilde{\gamma}_A(\succ)$, $\tilde{G}_A \equiv \sum_{\{\succ \in R(A)\}} \tilde{\gamma}_A(\succ)$ and G is given by (3).

Eq. (8) resembles the representation of choice probabilities in Luce's (1959) choice model. Note, however, that the weights $\gamma_A(x)$ and $\tilde{\gamma}_A(x)$ are functions not only of x but also of A. Also, choice probabilities are jointly constrained because they have terms in common.

It is also important to understand that we introduce the $\gamma(\succ)$ and $\tilde{\gamma}_A(\succ)$ weights only as a device for specifying joint prior distributions over the various probabilities $P_A(x)$. Given the probabilities, the weights have no behavioral import. The fact that the weights are not identified – multiplying all the weights by a positive constant does not change choice probabilities – is no cause for concern since the probabilities themselves are identified.

We describe methods for posterior inference in McCausland and Marley (2013). Inducing a prior over probabilities by specifying a prior over weights does not present any serious problems. In that paper, we simulate the posterior distribution of weights and then use the simulation sample of weights to generate a posterior sample of probabilities. The fact that the prior is proper implies that the posterior is proper, despite the non-identification of the weights.

3.3. *Properties of prior*

This family of prior distributions has the following properties.

Marginal Distributions. For each choice set A, the marginal distribution of the distribution $P_A(\cdot)$ is given by (5), whatever the value of λ.

Proof. For fixed A and x, the two numerator terms in (8) are independent. The first is the sum of $n!/|A|$ independent Gamma random variables with shape $\alpha\lambda/n!$ and scale 1. The second is the sum of $|A|!/|A|$ independent Gammas with shape $\alpha(1-\lambda)/|A|!$ and scale 1. The numerator is therefore Gamma with shape $\alpha/|A|$ and scale 1. For fixed A, the various numerators, for $x \in A$, are independent and their sum is the denominator. The result is then a standard property of the Dirichlet distribution—see Appendix B.3. □

Marginalization. Suppose $P \sim H(\alpha, \lambda, T)$ and let T' be any non-empty subset of T. Then (P', T'), the restriction of (P, T) to T', satisfies $P' \sim H(\alpha, \lambda, T')$.

Proof. Let n' be the cardinality of T'. For all $\succ' \in R(T')$, let $\gamma'(\succ') = \sum_{\{\succ \in R(T): \forall x, y \in T',\, x \succ y \Rightarrow x \succ' y\}} \gamma(\succ)$.

Gathering terms, we can write, for all non-empty $A \subseteq T'$,

$$P_A(x) = \frac{\gamma'_A(x) + \tilde{\gamma}_A(x)}{G + \tilde{G}_A},$$

where $\gamma'_A(x) = \sum_{\{\succ' \in R(T'): x = h_{\succ'}(A)\}} \gamma'(\succ')$, and the other terms are defined as they are in (8). Since (P', T') is the restriction of (P, T) to T', P' agrees with P for non-empty $A \subseteq T'$.

This equation takes the same form as Eq. (8), so we just need to verify that the joint distribution of the choice probabilities agrees with $H(\alpha, \lambda, T')$. Each $\gamma'(\succ')$ is the sum of $n!/n'!$ independent terms, each with a Gamma distribution with shape $\alpha\lambda/n!$ and scale 1. Therefore, the $\gamma'(\succ')$ are Gamma with shape $\alpha\lambda/n'!$ and scale 1. Since they have no $\gamma(\succ)$ terms in common, they are independent. This matches the direct construction of $P \sim H(\alpha, \lambda, T')$. □

Invariance. The distribution of P is invariant to permutations of the elements of T. This is a symmetry property similar to exchangeability. The result is obvious.

4. Results

We now report prior simulation results, beginning with probabilities of simple axioms and then moving on to the probabilities of compound axioms and the relationships among axioms. We compute approximations to probabilities using independence Monte Carlo. For given λ, α and n, we draw M random choice structures from $H(\alpha, \lambda, T)$, where $T = \{x_1, \ldots, x_n\}$. The number of times an axiom holds, divided by M, is an approximation of the prior probability p that the axiom holds, and the simulation standard error of the approximation is $\sqrt{p(1-p)/M}$.

All probabilities described in this section are prior probabilities of axioms, as a function of α and λ.

4.1. *Prior probabilities of simple axioms*

Figs. 6–9 show probabilities of the four axioms on binary choice probabilities: weak, moderate and strong transitivity and the triangle inequality. Figs. 10–12 show probabilities of the three other axioms: regularity, random ranking and the multiplicative inequality.

Two or three panels in each figure show axiom probability contours for $n = 3, 4$ and sometimes $n = 5$. Each panel gives a contour plot of an axiom's probability as a function of λ and α. Plots are based on samples of $M = 5 \times 10^6$ draws for $n = 3, 4$ and $M = 2 \times 10^5$ draws for $n = 5$, one sample for each point on a grid of values.

For the binary choice axioms, which are relatively probable, probabilities are plotted over the range $0 \le \lambda \le 1$ and $0 \le \alpha \le 2$. When $\alpha = 2$, binary choice probabilities have a Dirichlet distribution with parameter values set to one. This is the same as the uniform distribution on the interval $[0, 1]$. The lower the value of α, the more probability mass is concentrated near zero and one.

For the other axioms, probabilities are plotted over the same range of α, $0 \le \alpha \le 2$, but for restricted ranges of λ. For $n = 3$, we use the full range $0 \le \lambda \le 1$; for $n = 4$, $0.97 \le \lambda \le 1$; and for $n = 5$, $0.994 \le \lambda \le 1$.

Probability values for various contour levels are labeled. The levels vary from panel to panel and figure to figure, but the grey scale does not. White represents a probability of zero, darker

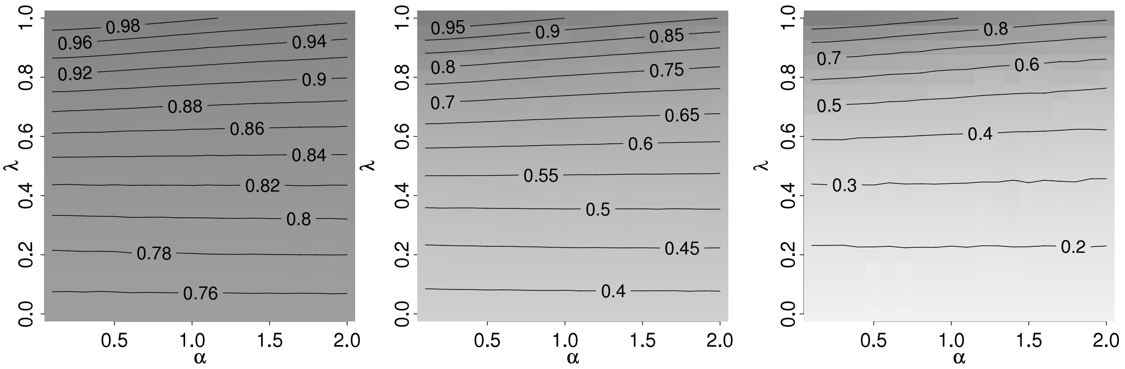

Fig. 6. Probability of weak stochastic transitivity, for $n = 3, 4, 5$.

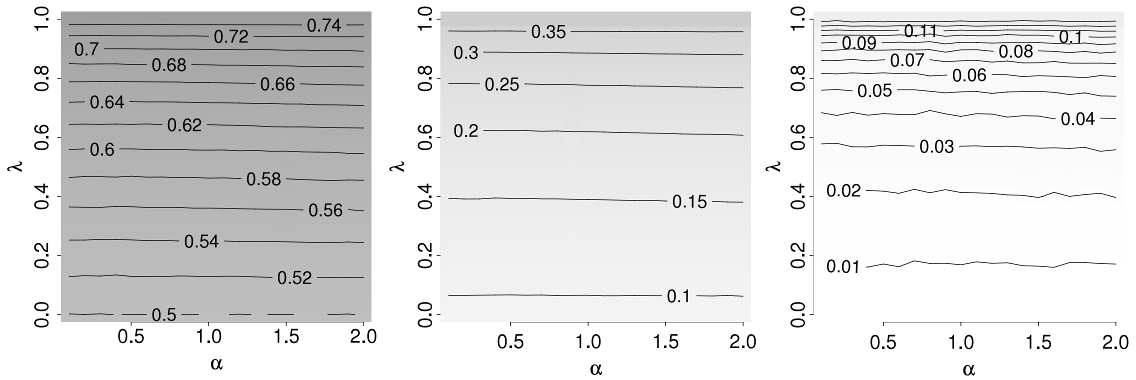

Fig. 7. Probability of moderate stochastic transitivity, for $n = 3, 4, 5$.

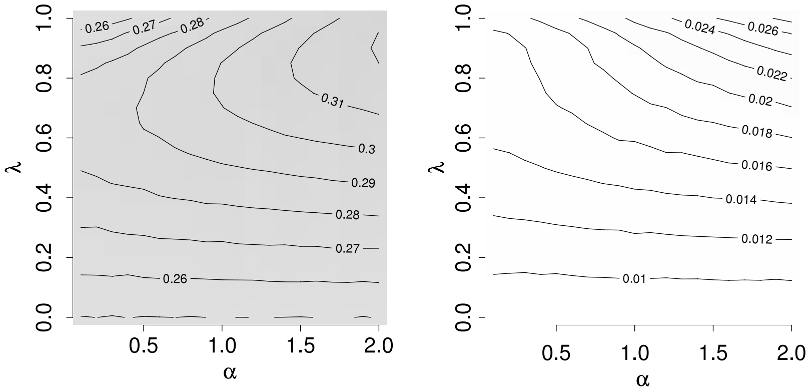

Fig. 8. Probability of strong stochastic transitivity, for $n = 3, 4$.

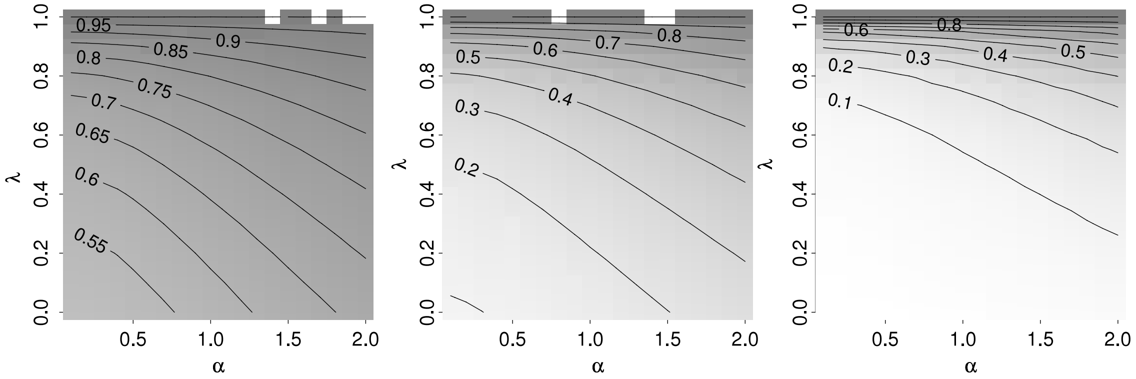

Fig. 9. Probability of triangle inequality, for $n = 3, 4, 5$.

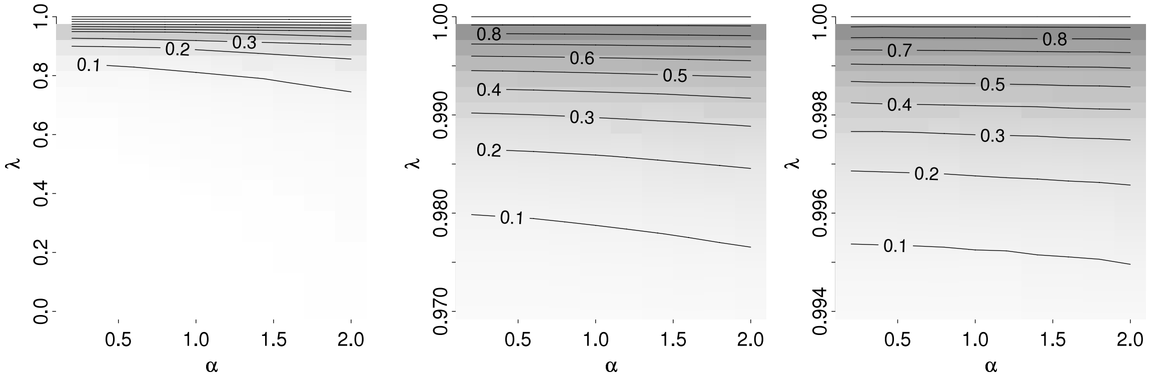

Fig. 10. Probability of regularity, for $n = 3, 4, 5$.

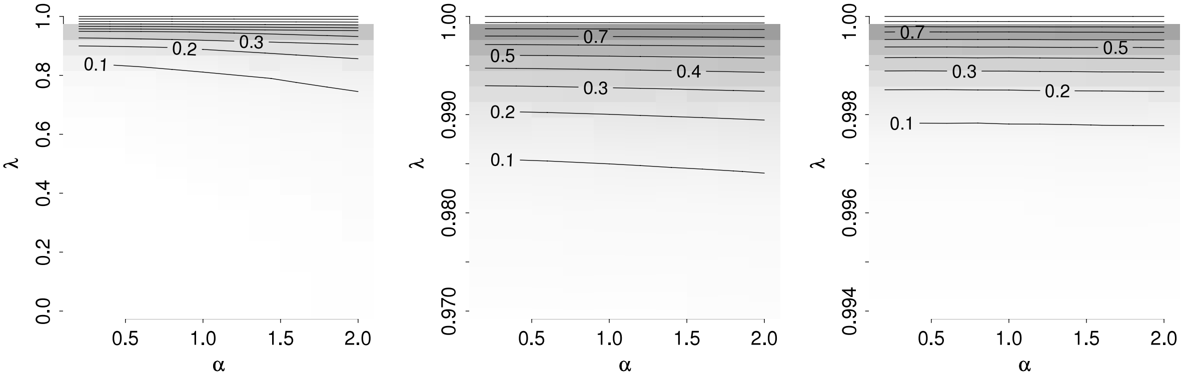

Fig. 11. Probability of random ranking, for $n = 3, 4, 5$.

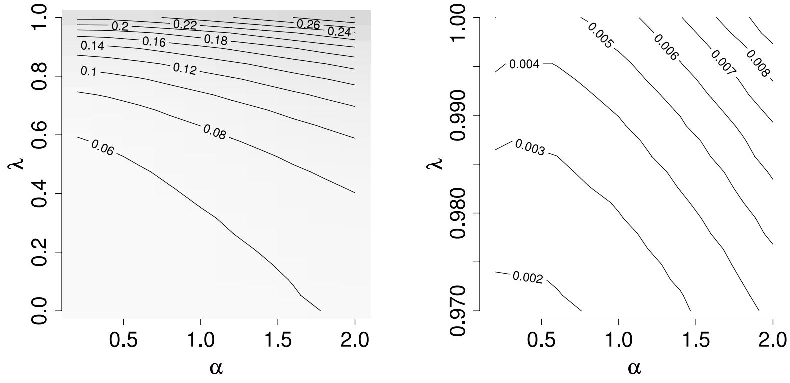

Fig. 12. Probability of multiplicative inequality, for $n = 3, 4$.

greys represent higher probabilities, and greys of equal darkness represent the same probabilities in each panel and figure.

We see that axioms for binary probabilities are much more probable than other axioms, except for λ very close to one, in which case regularity and random ranking have probability close to one. WST and TI are particularly probable, even for $n = 5$ objects. That Reg, RR and Mul are quite improbable (i.e. strong) is perhaps not surprising given that the number of constraints is larger and involve many choice probabilities other than the binary. On the other hand, the random ranking hypothesis is often regarded in Economics and Marketing as an innocuous condition. In fact, relative to unrestricted random choice, it is a very strong condition.

In Figs. 10 and 11, we see that regularity is an important part of the random ranking hypothesis, in the sense that in the region where Reg has reasonable probability, the probability of RR as a fraction of the probability of Reg – the conditional probability of RR given Reg – is not close to zero. At least for $n \leq 5$, violations of RR can often be attributed to violations of Reg. When we look at Fig. 9 as well, we see that we cannot say the same thing about the triangle inequality, which is also necessary for RR. Except for values of λ extremely close to 1, the probability of the triangle inequality is much higher than the probability of RR.

The probability of the multiplicative inequality, like those of Reg and RR, varies strongly with λ, becoming much higher close to $\lambda = 1$. However, unlike Reg and RR, Mul does not become certain when $\lambda = 1$. It is a very strong condition, even when we condition on RR. Independent random utility (as defined in Appendix A), frequently assumed in models applied in empirical work, is even stronger than Mul.

4.2. Prior probabilities of compound axioms

We illustrate the probabilities of compound axioms by reporting a normalized joint probability, which we will call the *probability*

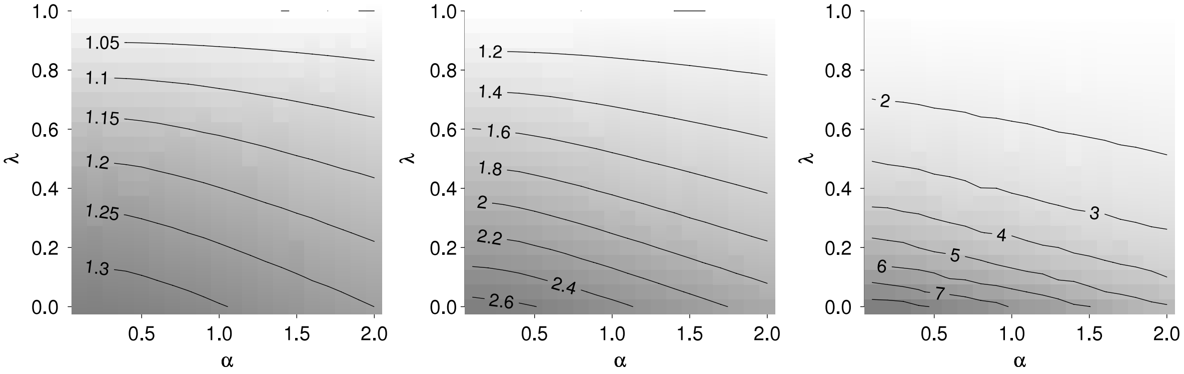

Fig. 13. Probability gain, weak stochastic transitivity and triangle inequality, for $n = 3, 4, 5$.

gain. For any two axioms, we define the probability gain β_{12} as

$$\beta_{12} \equiv \frac{\Pr[A_1 \cap A_2]}{\Pr[A_1]\Pr[A_2]},$$

where A_1 is the event that the first axiom holds and A_2 the event that the second holds. The probability gain β_{12} is equal to both $\Pr[A_1|A_2]/\Pr[A_1]$ and $\Pr[A_2|A_1]/\Pr[A_2]$, so it measures how the probability of one axiom changes when we condition on the other axiom. If events A_1 and A_2 are independent, then $\beta_{12} = 1$. If knowing that the first axiom holds increases the probability of the second holding, then $\beta_{12} > 1$; if it decreases it, then $\beta_{12} < 1$. The measure β_{12} is clearly symmetric, always equal to β_{21}. If the first axiom implies the second, then $A_1 \subseteq A_2$ and $\beta_{12} = 1/\Pr[A_2]$.

Fig. 13 shows how the probability gain varies with λ and α for the two axioms WST and TI. We see that the event that WST holds and the event that TI holds are nearly statistically independent when λ is close to one. The degree of dependence of these events – as measured by probability gain – increases as choice probabilities become more independent – as measured by λ. Dependence also increases with n.

5. Conclusions and possible extensions

We have formulated a class of prior distributions over the set of random choice structures. Each prior is a joint probability distribution over a collection $P_A(\cdot)$, $\emptyset \neq A \subseteq T$, of discrete choice probabilities. The class is somewhat flexible, with parameter α governing consistency of choices in repeated identical choice situations – whether choice probabilities are likely to be close to zero and one – and λ governing the degree of statistical dependence of choice probabilities across choice sets.

Our class of prior distributions is innovative in two dimensions, relative to priors described in the literature, such as those of Cavagnaro and Davis-Stober (2013), Myung et al. (2005) and Zwilling et al. (2011). First, each prior is a distribution over a larger collection of choice probabilities — it gives the joint distribution of all the choice probabilities $P_A(\cdot)$, for non-empty $A \subseteq T$, not just the ones where A is a doubleton set. Second, one can control the degree of statistical dependence across these choice probabilities. Setting $\lambda = 0$ gives mutual independence of $P_A(\cdot), \emptyset \neq A \subseteq T$; $\lambda = 1$ gives a prior with support equal to the set of random choice structures induced by random rankings. All intermediate values are possible.

At the same time, our prior has other desirable properties. The λ parameter influences only the dependence structure and not the marginal distributions of the $P_A(\cdot)$. These marginals are known distributions, with moments in closed form, which makes interpretation easier.

The marginalization property assures us that once we choose values for the prior parameters α and λ, our prior information about choice probabilities on subsets of T' does not depend on whether the master set is T' itself or some superset T. Whether $(T', P) \sim H(\alpha, \lambda, T')$ or $(T, P) \sim H(\alpha, \lambda, T)$, the joint distribution of the probabilities $P_A(\cdot)$, for non-empty $A \subseteq T'$, is the same. We emphasize that this consistency is a property of the prior distribution, not of the random choice structures themselves. If we, the authors, were to make predictions about how an agent would behave in choice situations involving only x, y and z, we might well want to allow for various kinds of inconsistency on the part of the agent, but our predictions would not depend on whether the master set is $\{x, y, z\}$ or $\{w, x, y, z\}$.

The invariance property is a form of symmetry or anonymity. A particular random choice structure drawn from one of the priors may assign very different choice probabilities to different objects. The symmetry lies in the fact that this draw is no more or less likely than the other $|T|!$ random choice structures obtained by relabelling the elements of the master set.

We believe this invariance is appropriate for the prior analysis in the present paper. For the purposes of data analysis, where one knows the identities and attributes of the choice objects, one might wish to dispense with invariance. One way to extend our class of priors to do this would be to allow the α parameter in the distributions of $\gamma_A(\succ)$ and $\tilde{\gamma}_A(\succ)$, given by (6) and (7), to depend on $\succ$. There are many ways in which one might do this; as long as the weights remain independent Gamma random variables with common rate parameter, distributions over choice probabilities remain Dirichlet, although no longer symmetric.

The prior could be generalized in other ways, using mixtures for example. But we would caution against adding flexibility for its own sake. In particular, one should keep in mind that the data generating process is already non-parametric in the sense that each choice distribution is fully flexible. The fact that there are a finite number of probabilities for a given master set comes from the finiteness of choice sets, not the arbitrary imposition of a parametric family of data densities. Generalizing the prior may turn out to be desirable in some cases, but it should proceed from an understanding of what regions of the space of random choice structures are over- and under-represented for particular choices of prior from the class we describe here.

Using our class of priors, we studied the strength of various axioms of discrete probabilistic choice, measuring how restrictive they are, both alone and in the presence of other axioms. We measure the strength of an axiom according to its prior (im)probability. We measure its relationship with other axioms according to what we call probability gain, a function of joint and marginal axiom probabilities. We obtained information about the relationship between WST and TI, neither of which implies the other. These exercises depend on the particular choice of a prior, but we reported results for a wide range of prior distributions.

RR is a very strong condition, in the following sense. For $n = 4$ and $n = 5$, and presumably for $n > 5$, its probability is very close to zero except for values of λ very close to one. On the other hand, the triangle inequality, a necessary condition for RR, is much more probable, for a wide range of parameter values. Regularity, also a necessary condition for RR, is much closer to RR in probability.

The multiplicative inequality is also a strong condition, even when RR is imposed. Some models used in Applied Economics, notably the multinomial logit model, are independent random utility models (as defined in Appendix A) and the multiplicative inequality is a necessary condition for independent random utility. Also, MI is a necessary condition for EBA. For these reasons, it would be interesting to see how behaviorally realistic it is.

Prior simulation gives us useful information on what we can hope to learn from data. Since the Bayes factor in favor of an axiom is the ratio of its posterior and prior probabilities, it is bounded above by the reciprocal of its prior probability. We obtain simulation consistent approximations of these prior probabilities, which we can use to compute these bounds.

Bounds on Bayes factors reveal the limits on how much support data can provide in favor of a particular axiom. For example, given the high prior probability of the triangle inequality across priors, the amount of evidence that data from a single decision maker can provide in its favor is quite limited. When one tests RR by testing its much more probable consequence, the triangle inequality, the evidence in favor of RR is likewise limited—even with an infinite amount of data on all doubleton sets, the Bayes factor cannot exceed the prior probability of the triangle inequality. A practical implication of this is that there are rapidly diminishing returns in collecting only binary data to test RR.

We offer two recommendations on testing RR using discrete choice data. First, even when only binary choice data are available and $n \leq 5$, it would be useful to compute Bayes factors in favor of RR based on a prior distribution over the complete random choice structure, not just the collection of binary choice probabilities. This is not the same as the Bayes factor in favor of the triangle inequality. We have seen that for given binary choice probabilities satisfying TI, RR implies constraints on non-binary probabilities that are more or less restrictive depending on what the values of those binary choice probabilities are. Therefore the truncation of a prior over all choice probabilities to the RR region induces important changes to the joint prior distribution of the various binary probabilities—the pre-truncation and post-truncation distributions over the binary probabilities are quite different, and the latter is not simply a truncation of the former to the region compatible with TI.

As we now show, an implication of this is that some configurations of binary probabilities consistent with TI give more or less evidence in favor of RR. Denote the vector of all binary probabilities by p and the events of TI and RR holding as TI and RR. Then Bayes' rule relates the probability of RR holding given p and TI to the ratio of post- to pre-truncated densities, evaluated at p.

$$\Pr[\mathrm{RR}|p, \mathrm{TI}] = \frac{f(p|\mathrm{RR}, \mathrm{TI})}{f(p|\mathrm{TI})} \Pr[\mathrm{RR}|\mathrm{TI}],$$

where $f(p|\mathrm{TI})$ is the marginal density of p implied by a TI-consistent prior over the random choice structure, and $f(p|\mathrm{RR})$ is the conditional density of p given RR, the result of truncation.

To take another approach, suppose $n \leq 5$ and we observe binary choice probabilities directly, as if we had an infinite amount of data. All configurations of binary choice probabilities satisfying the triangle inequality are consistent with RR. Here, consistency means that we can choose values for all non-binary choice probabilities in such a way that the complete random choice structure satisfies RR. However, different configurations of binary probabilities will be more or less easy to complete—the volume or prior probability of the consistent region (in the space of non-binary probabilities) will depend on what the observed binary choice probabilities are.

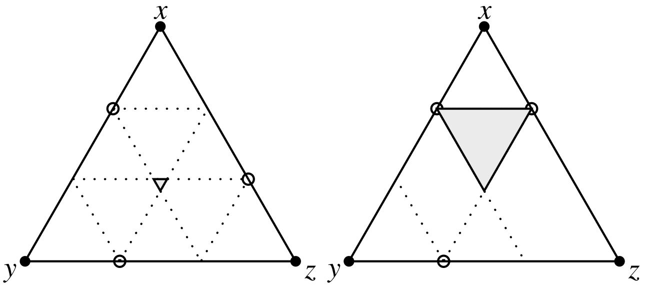

Fig. 14. Ternary probabilities compatible with RR and specified binary choice probabilities. Left panel: $p(x, y) = p(y, z) = p(z, x) = 0.65$. Right panel: $p(x, y) = p(y, z) = 0.65, p(z, x) = 0.35$.

We see these issues more clearly in an example, illustrated in Fig. 14, for the master set $T = \{x, y, z\}$. The two panels show two different, but similar, configurations of binary choice probabilities. In the left panel, $p(x, y) = 0.65, p(y, z) = 0.65$ and $p(z, x) = 0.65$. In the right panel, the first two probabilities are the same, but $p(z, x)$ is replaced by its complementary probability, 0.35. In both panels, the shaded triangle is the set of ternary probabilities compatible with regularity and the respective configurations of binary probabilities. Despite the similar binary choice probabilities, the compatible region on the right has an area 49 times as large as the region on the left. All this will be reflected in Bayes factors. For example, suppose that the prior assigns equal density to the two configurations of binary choice probabilities; independence and symmetry are sufficient conditions. Suppose further that the ternary probability $P_{\{x,y,z\}}(\cdot)$ is independent of the binary probabilities and uniform on the two-dimensional regular simplex. Then the limiting Bayes factor in favor of RR for binary data with choice frequencies $\tilde{p}(x, y) = 0.65, \tilde{p}(y, z) = 0.65$ and $\tilde{p}(z, x) = 0.35$ will be 49 times larger than the Bayes factor for choice frequencies $\tilde{p}(x, y) = 0.65, \tilde{p}(y, z) = 0.65$ and $\tilde{p}(z, x) = 0.65$. In contrast, the limiting Bayes factors in favor of TI are both equal to the reciprocal of the prior probability of TI.

The second recommendation is to test RR using data from choice sets of various sizes. The Falmagne conditions, Eq. (2), are linear inequalities in choice probabilities over various choice sets. Our results show that these are much more constraining than the triangle inequality for binary choice probabilities, so data on higher order choice probabilities should be very useful for testing RR. Such data are very rare, however, and new data would likely need to be collected. Collecting data is costly, of course, and choice becomes more difficult as the number and size of choice sets increases. But we would not be subject to the severely diminishing returns associated with collecting only binary data.

We have emphasized the random ranking hypothesis, partly because of its wide use in so-called random utility models—see Appendix A. This does not preclude us or others from investigating other models. Such an investigation would require framing a model or behavioral condition as a restriction on the choice probabilities of a random choice structure. In some cases, this might require providing further structure, to accommodate choice environments where the agent can choose none of the options or state indifference. We could extend our framework, currently based on assigning random weight functions to sets $R(T)$ and $R(A)$ of rank orders, to accommodate weight functions over other types of relations, such as weak or partial orders.

Our class of prior distributions on random choice structures serves several purposes beyond measuring the strength of axioms, and computing bounds on Bayes factors. We have described how prior simulation can be used to generate conjectures about the logical relationship between axioms. In future work we plan to use prior simulation to generate conjectures about best–worst choice probabilities — see Marley and Louviere (2005) for models of best–worst choice. This could help direct the search for theorems giving necessary and/or sufficient conditions for consistency of best–worst choice probabilities with random ranking. To give an example illustrating how this could proceed, we noticed in simulations we ran for the current paper (for best-only choice) that there was a subset of the Falmagne (1978) conditions for RR that was never pivotal — whenever one of these necessary conditions was false, there was always some necessary condition outside this set that was also false. The natural conjecture that these conditions are dispensable is in fact a theorem, although one which was previously known and proven in Fiorini (2004).

Of course the principal purpose of our class of priors is Bayesian data analysis. We plan to develop a testing ground for various choice axioms in an abstract choice setting, based on our class of priors. Axioms of interest include the axioms studied in this paper, other extant axioms and perhaps axioms that are yet to be proposed.

We hope to examine several data sets to explore the empirical support for various axioms in various choice contexts. We are interested in data sets on consumer choice, but also data sets specially constructed to elicit choice anomalies such as context effects or violations of stochastic transitivity. Marketing studies tend to use models satisfying the random ranking hypothesis to analyze discrete consumer choice. In this context, these tight models have reasonably good predictive performance. In the literature on choice anomalies, there are examples of context effects incompatible with the random ranking hypothesis as well as systematic violations of various forms of stochastic transitivity—see Busemeyer and Rieskamp (2013) and Rieskamp, Busemeyer, and Mellers (2006) for summaries of such effects and of stochastic dynamic models, incompatible with RR, that describe them. However, the choice situations that lead to such context effects are contrived – i.e. specifically constructed to lead to anomalies – and not very representative of naturally occurring choices. Furthermore, the models used to analyze these data are typically evaluated only by how well they capture these specialized data. Marketing practitioners are understandably reluctant to discard their models just because the data on context effects are incompatible with the random ranking hypothesis. In studying a wide range of data sets and a wide range of axioms, we hope to be able to shed light on what sets of axioms – or priors, more generally – on random choice structures can accommodate observed choice anomalies, while at the same time be disciplined enough to predict choices well in more naturally occurring choice situations. Recent research that studies rather limited extension of the random ranking framework suggest that such an approach may be successful: first, there are approaches that model differences in the scale factor (variance component) of a random ranking model across choice sets — see Louviere and Swait (2010) and Salisbury and Feinberg (2010a) for general discussions of the approach and Hutchinson, Zauberman, and Meyer (2010) and Salisbury and Feinberg (2010b) for value versus variance interpretations of temporal discounting data; second, there are approaches that can be roughly described as involving "context-dependent random ranking models" — see Trueblood's (2012) and Trueblood, Brown, and Heathcote's (2013) work on a multi-attribute linear ballistic accumulator (MLBA) model, which models both the choices made and the time to make them.

We plan to introduce methods for posterior analysis and report results of data analysis in a companion paper. We do not do this here, partly because the practical implementation of posterior simulation methods is not straightforward. A lengthy description of our posterior simulation methods would be required. While the marginal distribution of choice probabilities is Dirichlet, the conjugate distribution for multinomial observations, our joint prior is not conjugate for joint multinomial observations on multiple choice sets. This is because of prior dependence among the different choice probabilities. In fact, we do not even have a joint density in closed form for the various choice probabilities, only for the latent weights.

On the other hand, these difficulties are not insurmountable, as long as we restrict our attention to master sets with a small number of elements.[2] We have already noted that we can simulate the posterior distribution of the weights and use this posterior sample to generate a sample from the posterior distribution of probabilities. We have also noted that marginal likelihood computations, essential for computing Bayes factors, are fairly straightforward.

Acknowledgments

The authors gratefully acknowledge funding by the Social Sciences and Humanities Research Council of Canada in the form of Insight grant SSHRC 435-2012-0451 to the University of Victoria for Marley and McCausland. We thank Michel Regenwetter (Action Editor) and two referees for their very helpful comments.

Appendix A. On random utility

In Economics and Psychology, there is a relatively long history of theory and application of probabilistic discrete choice models. In both fields, most of these models are so-called *random utility models*, those in which agents select from each choice set as if they drew, independently and from the same continuous joint distribution, a collection $u_x, x \in T$, of utilities and then went on to choose the highest-utility element from that set. The assumption that utilities have a continuous distribution implies that the probability that any two utilities are equal is zero. If the definition is only asserted for the binary choice probabilities, then the model is called a *binary* random utility model. If the utilities $u_x, x \in T$, are mutually independent, then we say the model is an *independent* random utility model.

The sense above of the term "random utility" is equivalent to the random ranking hypothesis of Section 2, as shown by Block and Marschak (1960) and Luce and Suppes (1965, Theorem 49). While this meaning of "random utility" is extensively used, especially in Economics and Marketing, it is not universal, and is in fact quite restrictive relative to other definitions, such as that of Regenwetter and Davis-Stober (2008, Definition 11). This more general case has been applied to so-called *ternary choice*, where the agent can choose one of two options or state indifference between them, and is also applicable to various cases where the options might be considered to be partially ordered, as in Regenwetter and Davis-Stober (2008).

For the remainder of this section, we use the term "random utility" in the narrow sense above. For an example of a random utility model, suppose we have a master set $T = \{x, y, z\}$ and let the utility of each object $r \in T$ be given by $u_r = v_r + \epsilon_r$, where ϵ_x, ϵ_y and ϵ_z are independent extreme value errors, and assume that $v_x = v_y = \ln 2$ and $v_z = \ln 1 = 0$. This is a particularly simple example of the familiar multinomial logit (or MNL) model, for which the probability of choosing x when presented with x and y is $1/2$, and the probability of choosing x when presented with $\{x, y, z\}$ is $2/5$.

In Psychology, random utility models include the models of Luce and Thurstone, the most common discrete choice models in Psychology. See summaries in Luce (1977, 1994), Luce and Suppes (1965) and Marley's (1992a, 1992b, 2002) editorial introductions to special journal issues. Busemeyer and Rieskamp (2013) and Rieskamp et al. (2006) present excellent summaries of current

[2] Many of the operations we perform on random choice models involve iteration over all 2^n subsets of the master set and all $n!$ rankings on the master set, where n is the cardinality of the master set. This is prohibitive for all but small numbers of elements.

theory and data, and Salisbury and Feinberg (2010a), with commentary by Louviere and Swait (2010), present alternative theoretical interpretations of related data.

In Economics too, choice models are frequently random utility ones. Those of McFadden and his collaborators (see McFadden (1976, 2001) and Train (2009)) figure prominently in the Applied Economics literature, with much being made of the result of McFadden and Train (2000) showing a limiting equivalence of the set of mixed multinomial logit models and the set of random utility models. These models have been applied to choice options without any detailed structure, but also (and more often) to structured objects such as lotteries (gambles) or, say, cars of different make, color, etc. Clearly, the more structured models are special cases of the general class of random choice models.

Random utility is rarely questioned or defended in Economics, or even acknowledged to be restrictive. The situation is somewhat different in Psychology, where there are various very successful dynamic stochastic choice models that are not random utility models, as summarized in Busemeyer and Rieskamp (2013) and Rieskamp et al. (2006).

Falmagne (1978) and McFadden and Richter (1990, manuscript 1970) show that random utility is indeed restrictive. The following example makes it clear why: suppose we modify the choice process described above so that the "utility" of x is a function not only of x but also the choice set A that the decision maker faces. Let the universe of objects be as before, but now define

$$u_r = v_r + w_r 1_A(z) + \epsilon_r, \quad r \in T,$$

where A is the choice set presented and $1_A(z)$ is equal to one if $z \in A$ and zero otherwise. If we take $v_x = v_y = \ln 2$ and $v_z = \ln 1 = 0$, as before, and set $w_x = \ln 2$, $w_y = w_z = \ln 1 = 0$, we obtain that the probability of choosing x when presented with x and y, i.e., $P_{\{x,y\}}(x)$, is $1/2$, as before, and the probability of choosing x when presented with $T = \{x, y, z\}$, i.e., $P_T(x)$, is $4/7$.

This specification of "utility", by being a function of the choice set, is not a utility in the sense given in the definition of a random utility model. In fact, the induced choice probabilities cannot be generated by *any* random utility model. To see this, suppose that they can. Then $P_{\{x,y\}}(x)$ is the probability of the event $[u_x \geq u_y]$ and $P_T(x)$ is the probability of the event $[u_x \geq u_y] \cap [u_x \geq u_z]$. The latter, being a subset of the former, cannot have a higher probability, no matter what the dependence structure of the u_r is. But $P_T(x) > P_{\{x,y\}}(x)$ for the given example.

Appendix B. Some properties of the gamma and Dirichlet distributions

We mention here some well known properties of the Gamma and Dirichlet distributions. For details and other properties, see Forbes et al. (2011). There are two standard parameterizations of the Gamma distribution, we use the one for which the Gamma density $\mathrm{Ga}(\alpha, \beta)$ is

$$f(\gamma|\alpha,\beta) = \begin{cases} \dfrac{\beta^\alpha}{\Gamma(\alpha)} \gamma^{\alpha-1} e^{-\beta\gamma} & \gamma > 0, \\ 0 & \text{otherwise.} \end{cases}$$

The density of the J-dimensional Dirichlet distribution $\mathrm{Di}(\alpha_1, \ldots, \alpha_J)$ is

$$f(p_1, \ldots, p_J|\alpha_1, \ldots, \alpha_J) = \begin{cases} \dfrac{\Gamma(A)}{\prod_{j=1}^{J} \Gamma(\alpha_j)} \prod_{j=1}^{J} p_j^{\alpha_j - 1}, & p_1, \ldots, p_J \geq 0, \sum_{j=1}^{J} p_j = 1, \\ 0 & \text{otherwise,} \end{cases}$$

where $A = \sum_{j=1}^{J} \alpha_j$.

B.1. Aggregation property of the gamma distribution

If random variables γ_1 and γ_2 are independent, with $\gamma_1 \sim \mathrm{Ga}(\alpha_1, \beta)$ and $\gamma_2 \sim \mathrm{Ga}(\alpha_2, \beta)$, then $\gamma_1 + \gamma_2 \sim \mathrm{Ga}(\alpha_1 + \alpha_2, \beta)$.

B.2. Mean and variance of gamma distribution

If $\gamma \sim \mathrm{Ga}(\alpha, \beta)$, then $E[\gamma] = \alpha/\beta$ and $\mathrm{Var}[\gamma] = \alpha/\beta^2$. In this paper, only ratios of gammas are important, and we set the scale parameter $\beta = 1$. If $\gamma \sim \mathrm{Ga}(\alpha, 1)$, then $E[\gamma] = \mathrm{Var}[\gamma] = \alpha$. For values of α close to zero, the standard deviation is much higher than the mean, implying a highly skewed distribution.

B.3. Relationship between the gamma and Dirichlet distributions

If the random variables $\gamma_j, j = 1, \ldots, J$, are independent with $\gamma_j \sim \mathrm{Ga}(\alpha_j, \beta)$, where $\alpha_1, \ldots, \alpha_J$ and β are positive parameters, then

$$(\gamma_1/G, \ldots, \gamma_J/G) \sim \mathrm{Di}(\alpha_1, \ldots, \alpha_J),$$

where $G = \sum_{j=1}^{J} \gamma_j$.

B.4. Mean and covariance of Dirichlet distribution

If $(p_1, \ldots, p_J) \sim \mathrm{Di}(\alpha_1, \ldots, \alpha_J)$, then

$$E[p_j] = \frac{\alpha_j}{A}, \qquad \mathrm{Var}[p_j] = \frac{\alpha_j(A - \alpha_j)}{A^2(A+1)},$$

$$\mathrm{Cov}[p_i, p_j] = \frac{-\alpha_i\alpha_j}{A^2(A+1)}, \quad j \neq i,$$

where $A = \sum_{j=1}^{n} \alpha_j$.

B.5. Neutrality of Dirichlet distribution

If $(p_1, \ldots, p_J) \sim \mathrm{Di}(\alpha_1, \ldots, \alpha_J)$, then p_j and

$$\left(\frac{p_1}{1-p_j}, \ldots, \frac{p_{j-1}}{1-p_j}, \ldots, \frac{p_{j+1}}{1-p_j}, \ldots, \frac{p_J}{1-p_j}\right)$$

are independent.

Appendix C. Computational issues

The following appendix discusses some computational issues that are important for implementing the computational experiments described in the paper. For more on the C language, see Kernighan and Ritchie (1988). For more on the GNU Scientific Library, see Galassi et al. (2009).

C.1. Operations on sets

We program in C and represent sets using unsigned integers. We use C language bit operations to compute unions, intersections and complements; and to test conditions such as set membership and set inclusion.

In C, it is most convenient to index the elements of T as $i = 0, 1, \ldots, n-1$. We represent the singleton set containing an object i by the constant unsigned integer 2^i. The union between sets A and B is given by the bitwise 'or' operation `A|B`; their intersection, by the bitwise 'and' operation `A&B`. We represent the master set T by the integer $2^n - 1$ and the complement of set A in T by `T-A`. The C expression `(A&B)==B` is true if and only if B is a subset of A. Testing set membership is just a special case, using representations for singleton sets. To iterate through all subsets of the master set, we just iterate through the integers $0, 1, \ldots, 2^n - 1$.

C.2. Iteration over rankings

Iterating through all permutations of the integers $0, \ldots, n-1$ is a standard problem in combinatorics. We represent rankings as permutations and use routines provided in the GNU Scientific Library (GSL) for iterating through permutations in lexicographic order.

C.3. Numerically robust checks of axioms

We discovered in preliminary simulations that when probabilities are computed and compared in the direct and obvious way (see the example below), rounding errors sometimes led to false conclusions about whether a given random choice structure satisfied a given axiom. Here we describe the measures we took to make checking axioms more numerically robust.

To understand the issue, consider the comparison of $P_{\{x,y\}}(x) \equiv p(x,y)$ and $P_{\{x,y,z\}}(x)$, which arises when verifying the regularity axiom for the random choice structure P. The naive way of doing this is to construct both probabilities, then form the difference and test the sign of the result.

First consider the prior distribution over random choice structures that are random utility rational, the case $\lambda = 1$. We leave aside the obvious objection that in this case we know that regularity must hold. Then the difference $p(x,y) - P_{\{x,y,z\}}(x)$ is the probability that the random ranking $\succ$ satisfies $z \succ x \succ y$. This probability will often be much smaller than each of the probabilities $p(x,y)$ and $P_{\{x,y,z\}}(x)$. When it is, the rounding error of the computed difference will be about the same as the rounding errors of $p(x,y)$ and $P_{\{x,y,z\}}(x)$, but this will be a much greater fraction of its value. We obtain a much more numerically precise value of the difference of the probabilities by summing up the $\gamma(\succ)$ weights for the orders satisfying $z \succ x \succ y$ and dividing by the grand total G.

Now consider the more general prior distribution, the case $\lambda < 1$. In terms of weights, the difference $p(x,y) - P_{\{x,y,z\}}(x)$ is

$$\frac{\gamma_{\{x,y\}}(x) + \tilde{\gamma}_{\{x,y\}}(x)}{G + \tilde{G}_{\{x,y\}}} - \frac{\gamma_{\{x,y,z\}}(x) + \tilde{\gamma}_{\{x,y,z\}}(x)}{G + \tilde{G}_{\{x,y,z\}}}.$$

When $\gamma_{\{x,y\}}(x) - \gamma_{\{x,y,z\}}(x)$, $\tilde{\gamma}_{\{x,y\}}(x)$ and $\tilde{\gamma}_{\{x,y,z\}}(x)$ are very small compared to $\gamma_{\{x,y\}}(x)$, and $\tilde{G}_{\{x,y\}}$ and $\tilde{G}_{\{x,y,z\}}$ are very small compared to G, the sign of the computed difference may be incorrect, possibly leading to an incorrect conclusion about whether P satisfies regularity.

It might seem that such a case would arise with minuscule probability, and it is indeed rare, even for the small values of α for which we see it at all. (We noticed it by investigating cases where impossible combinations of axioms were 'detected'.)

To explain why this happens with non-negligible probability, we first point out some properties of the gamma distribution. When the shape parameter of a gamma random variable is close to zero, the standard deviation is much larger than the mean and a set of draws is likely to vary over many orders of magnitude. Recall that the weights $\gamma(\succ)$ have shape $\alpha\lambda/n!$ and the weights $\tilde{\gamma}_A(\succ)$ have shape $\alpha(1-\lambda)/|A|!$. The quantities $\gamma_{\{x,y,z\}}(x)$ and $\gamma_{\{x,y\}}(x) - \gamma_{\{x,y,z\}}(x)$ are independent Gamma random variables, the first with shape $\alpha\lambda/3$ (see result, Eq. (8)) and the second with shape $\alpha\lambda/6$ (exactly one sixth of the $\gamma(\succ)$ weights will be for orders $\succ$ satisfying $z \succ x \succ y$, whatever the value of n). For small enough α, the first, as a fraction of the second, can easily be smaller than the machine epsilon of a computer, the smallest increment to the value 1 that gives a sum that the computer can distinguish from 1.

While we have succeeded in making our axiom checks much more numerically robust, we suspect that many readers will have a lingering concern about priors putting so much probability mass so close to the boundary of the "regularity" region that classification errors have non-negligible frequency. Anticipating this concern, we offer the following comments.

First, this feature is only really extreme for priors near the boundary of the parameter space, in a region that we do not consider to be very plausible, as it implies high probability of binary choice probabilities being extremely close to zero or one. There is no bright line of plausibility, and we have chosen to report results down to very low values of α, doing what is necessary for reliable results.

Second, we expect this feature to arise for many reasonable priors, not just those in the parametric class we chose. First take the case of random choice structures induced by ranking distributions. Binary choice probabilities are aggregates of ranking probabilities, and for the prior distribution of the former to put significant probability mass close to zero *and* one, either the distributions of the latter need to be highly skewed, or one has to impose a strong dependence structure on the joint distribution of the ranking probabilities. Our choice not to impose such structure is completely appropriate given the non-parametric nature of our investigation. The consequence is highly skewed ranking probabilities, which implies lots of prior mass near the boundary of the regularity region. The general case of random choice structures is more complicated, but if we want a similarly unstructured prior that puts a fair amount of probability mass on structures that are random utility rational, then it is inevitable that a lot of prior mass will lie very near the boundary of the regularity region.

To make our checks of various axioms more efficient and robust to machine precision errors, we follow some useful guidelines. Instead of computing the difference of two probabilities that are both greater than 1/2, we compute minus the difference of their complements. If the two probabilities are very close to one, then the latter result will have much greater numerical precision, since the true difference will be small compared with the probabilities and less so compared with the complements. We avoid unnecessary divisions. For example, when comparing two ratios, we first cross multiply to clear the denominator so that only products remain. Finally, we compute linear combinations of terms $\gamma_A(x)$ directly in terms of sums of $\gamma(\succ)$ terms when this is more numerically robust.

C.3.1. Numerically robust checks of MST and SST

We first rewrite the MST condition as

$$p(x,y) \geq \frac{1}{2} \quad \text{and} \quad p(y,z) \geq \frac{1}{2} \\ \Rightarrow p(z,x) \leq \max[p(y,x), p(z,y)] \tag{9}$$

and the SST condition as

$$p(x,y) \geq \frac{1}{2} \quad \text{and} \quad p(y,z) \geq \frac{1}{2} \\ \Rightarrow p(z,x) \leq \min[p(y,x), p(z,y)].$$

We can then write a condition such as $p(z,x) \leq p(z,y)$ as

$$\frac{\gamma_{\{x,z\}}(z) + \tilde{\gamma}_{\{x,z\}}(z)}{G + \tilde{G}_{\{x,z\}}} - \frac{\gamma_{\{y,z\}}(z) + \tilde{\gamma}_{\{y,z\}}(z)}{G + \tilde{G}_{\{y,z\}}} \leq 0,$$

then as

$$\begin{aligned} &\left(\gamma_{\{x,z\}}(z) - \gamma_{\{y,z\}}(z)\right) G + \left(\tilde{\gamma}_{\{x,z\}}(z) - \tilde{\gamma}_{\{y,z\}}(z)\right) G \\ &\quad + \left(\gamma_{\{x,z\}}(z) + \gamma_{\{x,z\}}(z)\right) \tilde{G}_{\{y,z\}} \\ &\quad - \left(\gamma_{\{y,z\}}(z) + \gamma_{\{y,z\}}(z)\right) \tilde{G}_{\{x,z\}} \leq 0. \end{aligned}$$

The difference in the first term can be written

$$\gamma_{\{x,z\}}(z) - \gamma_{\{y,z\}}(z) = \sum_{\{\succ \in R(T): y \succ z \succ x\}} \gamma(\succ).$$

Computing it by adding up terms as indicated on the right hand side is more numerically robust than forming the difference of the sums $\gamma_{\{x,z\}}(z)$ and $\gamma_{\{y,z\}}(z)$. When this difference is many orders of magnitude less than the two terms, the advantage is important.

C.3.2. Numerically robust checks of TI

We first rewrite the TI condition as follows. For all tripleton subsets $\{x, y, z\} \subseteq T$,

$$p(x,y) + p(y,z) + p(z,x) \geq 1,$$
$$p(y,x) + p(z,y) + p(x,z) \geq 1.$$

At least three of the six binary probabilities in these two equations must be greater than or equal to 1/2. Therefore, in at least one equation, at least two of the binary probabilities are greater or equal to 1/2. Without loss of generality, suppose that these are $p(x, y)$ and $p(y, z)$. Then the sum in the first equation must hold and TI reduces to

$$p(z,x) \leq p(y,x) + p(z,y).$$

We can therefore say that the triangle inequality is equivalent to the following condition: for all $x, y, z \in T$,

$$p(x,y) \geq \frac{1}{2} \quad \text{and} \quad p(y,z) \geq \frac{1}{2} \Rightarrow p(z,x) \leq p(y,x) + p(z,y).$$

We note in passing that writing the TI condition in this form and writing the MST condition in the form of Eq. (9) makes it transparent that MST implies TI.

We can write the condition $p(z, x) \leq p(y, x) + p(z, y)$ as

$$\frac{\gamma_{\{x,z\}}(z) + \tilde{\gamma}_{\{x,z\}}(z)}{G + \tilde{G}_{\{x,z\}}} - \frac{\gamma_{\{x,y\}}(y) + \tilde{\gamma}_{\{x,y\}}(y)}{G + \tilde{G}_{\{x,y\}}} - \frac{\gamma_{\{y,z\}}(z) + \tilde{\gamma}_{\{y,z\}}(z)}{G + \tilde{G}_{\{y,z\}}} \leq 0,$$

then as

$$\begin{aligned}
&\{[\gamma_{\{x,z\}}(z) - \gamma_{\{x,y\}}(y) - \gamma_{\{y,z\}}(z)]G + [\gamma_{\{x,z\}}(z) - \gamma_{\{x,y\}}(y)]\tilde{G}_{\{y,z\}} \\
&\quad + [\gamma_{\{x,z\}}(z) - \gamma_{\{y,z\}}(z)]\tilde{G}_{\{x,y\}}\}G - [\gamma_{\{x,y\}}(y) + \gamma_{\{y,z\}}(z)]G\tilde{G}_{\{x,z\}} \\
&\quad + \gamma_{\{x,z\}}(z)\tilde{G}_{\{x,y\}}\tilde{G}_{\{y,z\}} - \gamma_{\{x,y\}}(y)\tilde{G}_{\{x,z\}}\tilde{G}_{\{y,z\}} - \gamma_{\{y,z\}}(z)\tilde{G}_{\{x,z\}}\tilde{G}_{\{x,y\}} \\
&\quad + \tilde{\gamma}_{\{x,z\}}(z)(G + \tilde{G}_{\{x,y\}})(G + \tilde{G}_{\{y,z\}}) \\
&\quad - \tilde{\gamma}_{\{x,y\}}(y)(G + \tilde{G}_{\{x,z\}})(G + \tilde{G}_{\{y,z\}}) \\
&\quad - \tilde{\gamma}_{\{y,z\}}(z)(G + \tilde{G}_{\{x,z\}})(G + \tilde{G}_{\{x,y\}}) \leq 0.
\end{aligned}$$

We can express the three quantities in brackets in the first line as

$$\begin{aligned}
\gamma_{\{x,z\}}(z) - \gamma_{\{x,y\}}(y) - \gamma_{\{y,z\}}(z) &= -\left(\sum_{\{\succ\, \in R(T): x \succ z \succ y\}} \gamma(\succ)\right) \\
&\quad - \left(\sum_{\{\succ\, \in R(T): y \succ x \succ z\}} \gamma(\succ)\right) - \left(\sum_{\{\succ\, \in R(T): z \succ y \succ x\}} \gamma(\succ)\right), \\
\gamma_{\{x,z\}}(z) - \gamma_{\{x,y\}}(y) &= \left(\sum_{\{\succ\, \in R(T): z \succ x \succ y\}} \gamma(\succ)\right) \\
&\quad - \left(\sum_{\{\succ\, \in R(T): y \succ x \succ z\}} \gamma(\succ)\right), \\
\gamma_{\{x,z\}}(z) - \gamma_{\{y,z\}}(z) &= \left(\sum_{\{\succ\, \in R(T): y \succ z \succ x\}} \gamma(\succ)\right) \\
&\quad - \left(\sum_{\{\succ\, \in R(T): x \succ z \succ y\}} \gamma(\succ)\right).
\end{aligned}$$

The computations implied by the right hand sides of these three equations are more numerically robust than those implied by the left hand sides. This is most obvious in the first case, where the weights on the right hand side all have the same sign. In the second two cases, no $\gamma(\succ)$ weights appear in more than one right hand side sum, so the differences on the right hand sides are extremely unlikely to be many orders of magnitude lower than the two terms in parentheses forming the differences.

C.3.3. Numerically robust checks of Reg

Testing regularity involves testing conditions of the form $P_A(x) - P_{A\cup\{y\}}(x) \geq 0$. All other conditions, such as $P_A(x) - P_B(x) \geq 0$, for $A \subset B$ and $|B| - |A| > 1$, are redundant, by simple induction.

We can write this condition as

$$\frac{\gamma_A(x) + \tilde{\gamma}_A(x)}{G + \tilde{G}_A} - \frac{\gamma_{A\cup\{y\}}(x) + \tilde{\gamma}_{A\cup\{y\}}(x)}{G + \tilde{G}_{A\cup\{y\}}} \geq 0,$$

or

$$\begin{aligned}
&G\left(\gamma_A(x) - \gamma_{A\cup\{y\}}\right) + G\left(\tilde{\gamma}_A(x) - \tilde{\gamma}_{A\cup\{y\}}\right) \\
&\quad + \tilde{G}_{A\cup\{y\}}\left(\gamma_A(x) + \tilde{\gamma}_A(x)\right) - \tilde{G}_A\left(\gamma_{A\cup\{y\}}(x) + \tilde{\gamma}_{A\cup\{y\}}(x)\right) \geq 0.
\end{aligned}$$

The first term can be written as

$$G\left(\gamma_A(x) - \gamma_{A\cup\{y\}}\right) = \sum_{\{\succ \in R(T): \forall z \in A \setminus \{x,y\}\ y \succ x \succ z\}} \gamma(\succ).$$

The right hand side is more numerically robust because there are weights appearing in both terms of the left hand side.

C.3.4. Numerically robust checks of MI

The multiplicative inequality consists of conditions of the form

$$P_{A\cup B}(x) \geq P_A(x)P_B(x). \tag{10}$$

Define the complementary probabilities $Q_A(x) = 1 - P_A(x)$, $Q_B(x) = 1 - P_B(x)$ and $Q_{A\cup B}(x) = 1 - P_{A\cup B}(x)$. We compute these directly in terms of the $\gamma(\succ)$ and $\tilde{\gamma}_A(\succ)$ weights to avoid loss of numerical precision when the complementary probabilities are close to zero.

We test one of four equivalent conditions, according to the values of $P_A(x)$ and $P_B(x)$. If $P_A(x) < 1/2$ and $P_B(x) < 1/2$, we check condition (10). If $P_A(x) \geq 1/2$ and $P_B(x) < 1/2$, we check

$$P_{A\cup B}(x) - P_B(x) + Q_A(x)P_B(x) \geq 0,$$

If $P_A(x) < 1/2$ and $P_B(x) \geq 1/2$, we check

$$P_{A\cup B}(x) - P_A(x) + Q_B(x)P_A(x) \geq 0.$$

Finally, if $P_A(x) \geq 1/2$ and $P_B(x) \geq 1/2$, we check

$$Q_A(x) + Q_B(x) - Q_A(x)Q_B(x) - Q_{A\cup B}(x) \geq 0.$$

C.3.5. Numerically robust checks of RR

Each term of (2) is computed as

$$(-1)^{|B\setminus A|}P_B(x) = \begin{cases} (-1)^{|B\setminus A|}\dfrac{\tilde{G}_B\gamma_B^c(x) - G\tilde{\gamma}_B^c(x)}{G(\tilde{G}_B + G)} & \tilde{\gamma}_B(x)/\tilde{G}_B > 1/2, \\ (-1)^{|B\setminus A|}\dfrac{G_B\tilde{\gamma}_B(x) - \tilde{G}\gamma_B(x)}{G(\tilde{G}_B + G)} & \text{otherwise,} \end{cases}$$

where $\gamma_B^c(x) = G - \gamma_B(x)$ and $\tilde{\gamma}_B^c(x) = \tilde{G} - \tilde{\gamma}_B(x)$. These complemented versions of the gamma weights are calculated directly as a sum of primitive gamma weights, not as a difference of sums.

References

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis* (2nd ed.). New York, NY: Springer-Verlag.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *JRSSB*, *41*(2), 113–147.

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory*. Chichester, England: John Wiley and Sons.

Birnbaum, M. H. (2011). Testing mixture models of transitive preference. Comment on Regenwetter, Dana, and Davis-Stober (2011). *Psychological Review*, *118*(4), 675–683.

Block, H. D., & Marschak, J. (1960). Random orderings and stochastic theories of responses. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), *Contributions to probability and statistics: essays in honor of Harold Hotelling* (pp. 97–132). Stanford, CA: Stanford University Press.

Bos, C. S. (2002). A comparison of marginal likelihood computation methods. Discussion paper. Tinbergen Institute.

Busemeyer, J. R., & Rieskamp, J. (2013). Psychological research and theories on preferential choice. In S. Hess, & A. Daly (Eds.), *Handbook of choice modelling*. Edward Elgar.

Cavagnaro, D.R., & Davis-Stober, C.P. (2013). Transitive in our preferences, but transitive in different ways: an analysis of choice variability. Manuscript. Mihaylo College of Business and Economics. California State University. Fullerton, CA.

Colonius, H. (1983). A characterization of stochastic independence by association, with an application to random utility theory. *Journal of Mathematical Psychology*, *27*(1), 103–105.

Corbin, R., & Marley, A. A. J. (1974). Random utility models with equality: an apparent, but not actual, generalization of random utility models. *Journal of Mathematical Psychology*, *11*, 274–293.

Davis-Stober, C. P. (2012). A lexicographic semiorder polytope and probabilistic representation of choice. *Journal of Mathematical Psychology*, *56*, 86–94.

Dridi, T. (1980). Sur les distributions binaires associées à des distributions ordinales. *Mathématiques et Sciences Humaines*, *69*, 15–31.

Falmagne, J. C. (1978). A representation theorem for finite random scale systems. *Journal of Mathematical Psychology*, *18*, 52–72.

Fiorini, S. (2004). A short proof of a theorem of Falmagne. *Journal of Mathematical Psychology*, *48*, 80–82.

Forbes, C., Evans, M., Hastings, N., & Peacock, B. (2011). *Statistical distributions*. John Wiley & Sons.

Galassi, M., Davis, J., Theiler, J., Gough, B., Jungman, G., & Alken, P. et al. (2009). GNU scientific library reference manual—third edition.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. Boca Raton: Chapman & Hall/CRC.

Haldane, J. B. S. (1948). The precision of observed values of small frequencies. *Biometrika*, *35*, 297–303.

Hutchinson, J. W., Zauberman, G., & Meyer, R. (2010). On the interpretation of temporal inflation parameters in stochastic models of judgment and choice. *Marketing Science*, *29*, 23–31.

Iverson, G. J., & Falmagne, J.-C. (1985). Statistical issues in measurement. *Mathematical Social Sciences*, *10*, 131–153.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London A*, *186*, 453–461.

Kernighan, B. W., & Ritchie, D. M. (1988). *The C programming language* (2nd ed.). Prentice Hall.

Koppen, M. (1995). Random utility representations of binary choice probabilities: critical graphs yielding critical necessary conditions. *Journal of Mathematical Psychology*, *39*, 21–39.

Kotz, S., Balakrishnan, N., & Johnson, N. L. (2000). Continuous multivariate distributions. In *Models and applications*: *vol. 1*. New York: Wiley.

Louviere, J., & Swait, J. (2010). Discussion of älleviating the constant stochastic variance assumption in decision research: theory, measurement and experimental test". *Marketing Science*, *29*(1), 18–22.

Luce, R. D. (1959). *Individual choice behavior: a theoretical analysis*. New York, NY: John Wiley & Sons.

Luce, R. D. (1977). The choice axiom after twenty years. *Journal of Mathematical Psychology*, *15*, 215–233.

Luce, R. D. (1994). Thurstone and sensory scaling: then and now. *Psychological Review*, *107*, 271–277.

Luce, R. D., & Suppes, P. (1965). Preference, utility, and subjective probability. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology Vol. 3* (pp. 249–410). New York, NY: John Wiley & Sons (chapter 19).

Marley, A. A. J. (1992a). A selective review of recent characterizations of stochastic choice models using distribution and functional equation techniques. *Mathematical Social Sciences*, *23*, 5–29.

Marley, A. A. J. (1992b). Stochastic models of choice and reaction time: New developments. *Mathematical Social Sciences*, *23*, 1–3, 147–149, 251–253.

Marley, A. A. J. (2002). Random utility models and their applications: recent developments. *Mathematical Social Sciences*, *43*, 289–302.

Marley, A. A. J., & Louviere, J. J. (2005). Some probabilistic models of best, worst, and best–worst choices. *Journal of Mathematical Psychology*, *49*, 464–480.

McCausland, W. J., & Marley, A. A. J. (2013). Bayesian inference and model comparison for random choice structures. Université de Montréal working paper.

McFadden, D. (1976). Quantal choice analysis: a survey. *Annals of Economic and Social Measurement*, *5*, 363–389.

McFadden, D. (2001). Economic choices, Nobel lecture, December 2000. *American Economic Review*, *91*, 351–378.

McFadden, D., & Richter, M. K. (1990). Stochastic rationality and revealed stochastic preference. In J. S. Chipman, D. McFadden, & M. K. Richter (Eds.), *Preferences, uncertainty and optimality* (pp. 161–186). Boulder CO: Westview Press (chapter 6).

McFadden, D., & Train, K. (2000). Mixed MNL models for discrete response. *Journal of Applied Econometrics*, *15*, 447–470.

Myung, J. I., Karabatsos, G., & Iverson, G. J. (2005). A Bayesian approach to testing decision making axioms. *Journal of Mathematical Psychology*.

Regenwetter, M., Dana, J., & Davis-Stober, C. P. (2011). Transitivity of preferences. *Psychological Review*, *118*, 42–56.

Regenwetter, M., Dana, J., Davis-Stober, C. P., & Guo, Y. (2011). Parsimonious testing of transitive and intransitive preferences. *Psychological Review*, *118*(4), 684–688.

Regenwetter, M., & Davis-Stober, C. P. (2008). There are many models of transitive preference: a tutorial review and current perspective. In T. Kugler, J. C. Smith, T. Connolly, & Y. J. Son (Eds.), *Decision modeling and behavior in uncertain and complex environments* (pp. 99–124). New York: Springer.

Rieskamp, J., Busemeyer, J. R., & Mellers, B. A. (2006). Extending the bounds of rationality: evidence and theories of preferential choice. *Journal of Economic Literature*, *44*, 631–661.

Robert, C. P., & Casella, G. (2010). *Monte Carlo statistical methods*. Springer.

Salisbury, L. C., & Feinberg, F. M. (2010a). Alleviating the constant variance assumption in decision research: theory, measurement, and experimental test. *Marketing Science*, *29*, 1–17.

Salisbury, L. C., & Feinberg, F. M. (2010b). Temporal stochastic inflation in choice-based research. *Marketing Science*, *29*, 32–39.

Sattath, S., & Tversky, A. (1976). Unite and conquer: a multiplicative inequality for choice probabilities. *Econometrica*, *44*, 79–89.

Suck, R. (2002). Independent random utility representations. *Mathematical Social Sciences*, *43*, 371–389.

Train, K. (2009). *Discrete choice methods with simulation* (2nd ed.). Cambridge University Press.

Trueblood, J. S. (2012). An investigation of contexts effects in multi-alternative choice behavior through experimentation and cognitive modeling. Ph.D. Thesis. Cognitive Science Program. Indiana University.

Trueblood, J. S., Brown, S. D., & Heathcote, A. (2013). The multiattribute linear ballistic accumulator model. Manuscript. Department of Cognitive Sciences. University of California Irvine.

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, *76*(1), 31–48.

Tversky, A. (1972). Elimination by aspects: a theory of choice. *Psychological Review*, *79*(4), 281–299.

Zwilling, C., Cavagnaro, D., & Regenwetter, M. (2011). Quantitative testing of decision theories: a Bayesian counterpart. *Presentation at the Annual Meeting of the Society for Mathematical Psychology*. Boston. July 15.