Essays on Statistical Decision Theory and Econometrics

Bruno de Albuquerque Furtado

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2023

Abstract

Essays on Statistical Decision Theory and Econometrics

Bruno de Albuquerque Furtado

This dissertation studies statistical decision making in various guises. I start by providing a general decision theoretic model of statistical behavior, and then analyze two particular instances which fit in that framework.

Chapter 1 studies statistical decision theory (SDT), a class of models pioneered by Abraham Wald to analyze how agents use data when making decisions under uncertainty. Despite its prominence in information economics and econometrics, SDT has not been given formal choice-theoretic or behavioral foundations. This chapter axiomatizes preferences over decision rules and experiments for a broad class of SDT models. The axioms show how certain seemingly-natural decision rules are incompatible with this broad class of SDT models. Using those representation result, I then develop a methodology to translate axioms from classical decision-theory, a la Anscombe and Aumann (1963), to the SDT framework. The usefulness of this toolkit is then illustrated by translating various classical axioms, which serve to refine my baseline framework into more specific statistical decision theoretic models, some of which are novel to SDT. I also discuss foundations for SDT under other kinds of choice data.

Chapter 2 studies statistical identifiability of finite mixture models. If a model is not identifiable, multiple combinations of its parameters can lead to the same observed distribution of the data, which greatly complicates, if not invalidates, causal inference based on the model. High-dimensional latent parameter models, which include finite mixtures, are widely used in economics,

but are only guaranteed to be identifiable under specific conditions. Since these conditions are usually stated in terms of the hidden parameters of the model, they are seldom testable using noisy data. This chapter provides a condition which, when imposed on the *directly observable* mixture distribution, guarantees that a finite mixture model is non-parametrically identifiable. Since the condition relates to an observable quantity, it can be used to devise a statistical test of identification for the model. Thus I propose a Bayesian test of whether the model is close to being identified, which the econometrician may apply before estimating the parameters of the model. I also show that, when the model is identifiable, approximate non-negative matrix factorization provides a consistent, likelihood-free estimator of mixture weights.

Chapter 3 studies the robustness of pricing strategies when a firm is uncertain about the distribution of consumers' willingness-to-pay. When the firm has access to data to estimate this distribution, a simple strategy is to implement the mechanism that is optimal for the estimated distribution. We find that such an empirically optimal mechanism boasts strong profit and regret guarantees. Moreover, we provide a toolkit to evaluate the robustness properties of different mechanisms, showing how to consistently estimate and conduct valid inference on the profit generated by any one mechanism, which enables one to evaluate and compare their probabilistic revenue guarantees.

Table of Contents

List of Figures

List of Tables

Acknowledgements

First and foremost I would like to thank the members of my dissertation committee: Navin Kartik, Mark Dean and Jose Luis Montiel Olea. I could not have wished for a better advisor than Navin, who was always incredibly generous with his time and extremely precise in his insight. As I am writing this, I have been actively learning from Navin for six years already, but I feel that I have only scratched the surface of what he could teach.

I am also deeply indebted to Mark, who not only taught me much that I didn't know about economics, but also opened my mind to a much wider and richer interdisciplinary world. His questions and challenges always made me look at a problem through lenses that I would never have considered otherwise.

This work would not have been possible without Pepe, who first introduced me to both statistical decision theory and mixture models, and did so in a way that made those topics seem endlessly interesting. His incisive and insightful comments have greatly improved my work and made me a better researcher overall.

I would also like to thank Elliot Lipnowski, Laura Doval and Evan Sadler, for donating their precious time to improve my work with their brilliant comments. Insightful comments and productive discussions with Bernard Salanie, Yeon-koo Che, Serena Ng, Sokbae Lee, Duarte Goncalves, Mauricio Couri and Yu Fu Wong have greatly improved this dissertation.

Finally, I would like to thank Gil Riella, who taught me to be a researcher and without whom I would not be here.

Dedication

To my brilliant, supportive, inquisitive, funny and generally amazing wife. Without you, I would not be half the person (or researcher) I am today.

And to my parents, who set me out on the twin paths of economics and a good life.

Chapter 1: The behavioral implications of statistical decision theory

1.1 Introduction

Statistical decision theory (henceforth SDT) models decisions under uncertainty as a single agent decision problem. Nature selects a true parameter, or state of the world. Meanwhile, the decision maker (DM) chooses decision rules (strategies) and designs experiments (information structures) to maximize her objective function, without knowing which parameter Nature chose. The DM observes a signal drawn according to the distribution determined by the parameter and the experiment, and takes the action prescribed by the decision rule. The ultimate pay-off for the DM is based on the realised action and Nature's chosen parameter.

Models of this kind are ubiquitous in information economics. A clear example is the rational inattention literature, where a DM designs an experiment subject to constraints on its informativeness, simultaneously to choosing a decision rule. Moreover, many econometric problems can be formulated in statistical decision theoretic terms.

Given their importance, surprisingly little is known about the behavioral assumptions implicit in various statistical decision theoretic models. That is, there are no formal results describing which choices over decision rules and experiments are consistent with the predictions of SDT models. The main goal of this paper is to set SDT on a rigorous axiomatic — and behaviorally falsifiable — foundation. This is done by axiomatically characterizing preferences over the choice objects of SDT: pairs of decision rules and experiments. Formally, an experiment is a collection of probability distributions over signals, indexed by the parameters, and a decision rule is a function assigning an action to each possible signal arising from the experiment. The preferences over these objects, taken here as behavioral primitives, are assumed to have been elicited by observing the DM's choices from different feasible sets, following the revealed preference principle.

Besides being the natural choice environment for SDT, data on choices over decision rules and experiments is often readily available to empirical researchers. For example, a school, university or employer might be required to spell out its admissions policy as a function of standardized exam scores, transcripts, credentials, etc. This is common in the public university systems of many countries, for instance. Moreover, admissions policies may depend on the type of signal the candidate sends – e.g., whether or not they send their scores on an optional standardized test. Similarly, a regulator or licensing authority may be required to publicly commit to some approval criteria, which may depend on properties of the experiment (sample size, how to account for the placebo effect, etc.) that the applying firm runs. In either situation, the institution is required to publicly declare its choice of decision rule for a slate of different possible experiments.

Although much work has been done to behaviorally characterize various models of decisions under uncertainty, existing results overwhelmingly focus on preferences (or choices) over acts – functions mapping the unknown states of the world directly to final consequences. This is the framework of Savage (1954) and Anscombe and Aumann (1963). SDT differs from traditional decision theory under uncertainty in that observable information – in the form of signals from experiments – is treated separately from the parameters, which determine final outcomes. Therefore, unlike the traditional decision theoretic framework, SDT considers information acquisition an integral part of the model description. This makes it a more natural setting to model situations in which the DM expects to receive partial information about the parameter before making a decision.

To fix ideas, suppose our DM is a policymaker choosing whether to implement a costly social program (action a_1) or not (action a_0). The program should be implemented if and only if its mean value, θ, exceeds a known fixed cost c. Although the true value of the parameter θ is unknown, the DM can observe a sample of size n from a randomized controlled trial designed to assess the effectiveness of the program. Each sample is drawn independently from a normal distribution with mean θ and known variance. The full set of n samples is the observed signal, and the experiment can be characterized by a collection of random vectors, indexed by θ, having independent normally distributed components. An example of a decision rule could be to take action a_1 whenever the empirical mean of the signal is greater than c, and action a_0 otherwise. Suppose the DM can choose between two combinations of decision rules and experiments. The first pairs a sample of size 50 with the decision rule previously described, based on the empirical mean. The second pairs a sample of size 10 with a decision rule in which the policy is not implemented for any signal. A preference relation between these pairs could be, for instance, that the first combination is always chosen over the second one.

I impose axioms on the preference relation that fully characterizes a DM who chooses as if using a general SDT framework, which subsumes most models in applications. This establishes a tight link between the predictions of such a model and behaviorally falsifiable constraints (axioms) on preferences, which in turn allows us to tell what sets of choices are compatible with the SDT framework. In this essay, I focus on ex-ante preferences under the assumption that the DM is (or believes herself to be) dynamically consistent. That is, at the time of choosing the experiment, the DM believes that, upon observing the signal, she will follow the contingent plan described by the decision rule she chooses to pair with it.

Preferences that abide by such a framework are characterized by a state-dependent utility function u and a parameter aggregation functional I. The utility function $u : A \times \Theta \to \mathbb{R}$ — where Θ and A are the sets of parameters and actions respectively — determines the pay-offs of the actions at each parameter. For example, if A and Θ are subsets of $\mathbb{R}^n$, a common utility function is given by the negative of the squared error: $u(a, \theta) = -\|a - \theta\|_2$, where $\| \cdot \|_2$ denotes the Euclidian norm.

Using u, one can calculate the agent's risk functions, which define, for each parameter value, the expected utility implied by a decision rule–experiment pair. Formally, given a decision rule $\delta : X \to A$ – where X is the signal space – and an experiment $P = \{P_\theta : \theta \in \Theta\}$, I define the DM's risk function as the mapping $\theta \mapsto r_u(\delta, P)(\theta) \equiv \int_X u(\delta(x), \theta) \mathrm{d}P_\theta(x)$. Risk functions relate to lotteries, i.e., prospects involving objective probabilities. They represent taking the expected utility whenever the DM faces an objective distribution, while remaining completely agnostic about the subjective uncertainty captured by the parameters.

The second part of the representation, given by the aggregator I, summarizes the agent's attitude towards ambiguity, and can be viewed as an ordinal utility on the space of risk functions. It captures the way in which DMs deal with subjective uncertainty, i.e., the fact that they do not know the true parameter.

Formally, I characterize a DM who chooses decision rules and experiments to maximize the functional

$$V(\delta, P) = I(r_u(\delta, P)) = I\left(\left(\int_X u(\delta(x), \theta)\mathrm{d}P_\theta(x)\right)_{\theta\in\Theta}\right). \tag{1.1}$$

The aggregator I is only required to be continuous and monotone – basic properties that are satisfied by many models used in applications. Considering its particular form, I call this the monotone risk aggregation (MRA) model. Specific objective functions are obtained by specifying particular functional forms for I that satisfy these two properties.

Consider, for instance, two of the most well-known models in statistics and decision theory: the subjective expected utility (SEU) and the maximin expected utility (MEU) models. An SEU agent is assumed to have a prior belief π, which is a probability distribution on the set of parameters. Their ex-ante utility from decision rule δ and experiment P is $\int_\Theta \int_X u(\delta(x), \theta)\mathrm{d}P_\theta(x)\mathrm{d}\pi(\theta) = \int_\Theta r_u(\delta, P)\mathrm{d}\pi$. In the language of the MRA framework, I is the expectation operator with respect to the prior belief. On the other hand, an MEU agent maximizes the expected utility assuming the true parameter is the worst possible one for whichever decision rule they choose. Their ex-ante utility is given by $\min_{\theta\in\Theta} \int_X u(\delta(x), \theta)\mathrm{d}P_\theta(x) = \min_{\theta\in\Theta} r_u(\delta, P)(\theta)$, that is, $I(\cdot) = \min_{\theta\in\Theta}(\cdot)$. In both cases, I is monotone and continuous, thus SEU and MEU specialize the MRA model. While SEU and MEU models are widely applied in SDT, their exact behavioral implications were previously unknown. I provide characterizations of these and other models in Section 1.5.

The model in (1.1) is characterized by a set of six axioms on preferences. Two of those axioms are standard: they impose that preferences are rational (complete and transitive) and continuous. The remaining axioms are specific to my setting. The Consequentialism axiom states that the DM ultimately cares only about the parameter contingent probability distributions over actions induced by decision rules and experiments, not about the rules and experiments themselves. Independence of Irrelevant Parameters imposes that if one fixes the outcomes of two alternatives to be the same for every parameter but one, then preferences are completely determined by conditioning on the single parameter where they may differ. Taken together, these two axioms allow the ex-post utility u to be a function exclusively of actions and parameters, rather than depend on the particular decision rule and experiment being evaluated. Monotonicity says that if the DM prefers alternative 1 to alternative 2 conditional on any parameter being the truth, then she prefers alternative 1 unconditionally as well. This guarantees that the aggregator I is monotone. Finally, Conditional Mixture Independence requires that whenever the DM knows the true parameter, she chooses as if maximizing expected utility.

To see how the axioms can help us empirically test whether preferences satisfy the MRA model, recall the policymaker example above. For simplicity, suppose there are only two possible values of the parameter, $\theta_1 > c$ or $\theta_0 < c$. In principle, the DM can choose as her decision rule any signal-contingent distribution over the two possible actions. Suppose the decision rule that is strictly preferred by the DM is based on a likelihood ratio test: she specifies a significance level α and acts as if $\theta > c$ if, and only if, she rejects the null hypothesis that $\theta < c$. Consequently, she takes action a_1 (implements the program) whenever the null hypothesis is rejected, and takes action a_0 otherwise. This is an example of an inference-based decision rule – for further discussion of such rules, see Manski (2021). Furthermore, the DM has different preferred significance levels α_i for different sample sizes, perhaps feeling that more informative experiments (larger n) permit more stringent standards for implementing the policy. Each of these significance levels induce different decision rules δ_i^*, which are assumed to be strictly preferred to any other rule that could be paired with experiment i.

It is a priori unclear whether the apparently reasonable choice behavior just described is compatible with any objective function in the mold of eq. (1.1). As it turns out, a simple argument shows that such preferences violate the combination of Consequentialism and Independence of Irrelevant Parameters, making them incompatible with the MRA model. Indeed, suppose we observe that the DM strictly prefers significance level α_1 for her likelihood ratio test when the sample size is n_1, and α_2 if it is n_2. Tables 1.1a and 1.1b show the conditional action distributions P_i^*, for

(a) δ_1^*

$P_1^*(a\|\theta)$	a_0	a_1
θ_0	$1-\alpha_1$	α_1
θ_1	β_1	$1-\beta_1$

(b) δ_2^*

$P_2^*(a\|\theta)$	a_0	a_1
θ_0	$1-\alpha_2$	α_2
θ_1	β_2	$1-\beta_2$

(c) $\hat{\delta}_1$

$\hat{P}_1(a\|\theta)$	a_0	a_1
θ_0	$1-\alpha_2$	α_2
θ_1	β_1	$1-\beta_1$

(d) $\hat{\delta}_2$

$\hat{P}_2(a\|\theta)$	a_0	a_1
θ_0	$1-\alpha_1$	α_1
θ_1	β_2	$1-\beta_2$

Table 1.1: Conditional action distributions under different decision rules

$i = 1, 2$, resulting from decision rules δ_i^*, where β_i is the probability of type II error (i.e., the probability of wrongfully rejecting the policy). It is straightforward to construct alternative (mixed) decision rules $\hat{\delta}_1$ and $\hat{\delta}_2$, for sample sizes n_1 and n_2 respectively, such that $P_i^*(\cdot|\theta_1) = \hat{P}_i(\cdot|\theta_1)$ and $P_i^*(\cdot|\theta_0) = \hat{P}_{-i}(\cdot|\theta_0)$, as shown in tables 1.1c and 1.1d.

Since δ_i^* is the strictly preferred decision rule for experiment i, it is in particular strictly preferred to $\hat{\delta}_i$. Now note that the action distributions induced by each pair of decision rules, δ_i^* and $\hat{\delta}_i$, differ only on parameter θ_0. Moreover, the action distributions conditional on θ_0 induced by δ_1^* and δ_2^* are mirror images of each other. Therefore, either the DM cares about something other than parameter contingent action distributions, or preferences conditional on θ_0 depend on the action distribution at θ_1. In other words, if the DM satisfies Consequentialism, so that all that matters to her are the parameter contingent action distributions, then she must violate Independence of Irrelevant Parameters. The fact that we can not pinpoint exactly which of the axioms is violated in this example, is due to the fact that we only have data on two choice problems, rather than the full preference relation.

To obtain the MRA representation, I first recover the DM's utility function u from a particular incomplete binary relation that is implied by the DM's preferences. The aggregator I can then be viewed as a standard utility function on the space of the risk functions defined by u. In fact, using a direct analogy with consumer theory, risk functions can be interpreted as consumption bundles, with different parameters representing the different goods. The expected utility of a risk function at parameter $\theta \in \Theta$ is analogous to the quantity of good θ in the bundle. Therefore, showing the existence of an aggregator I is formally equivalent to proving the existence of a utility function in a generalized space of consumption bundles. This allows one to draw an analogy between a preference for hedging – where the DM intrinsically prefers risk functions that have a more balanced utility profile across parameters – to the concept of complementarity between different goods.

SEU agents act as if they can quantify the uncertainty about parameters with a single probability measure – in the terminology of Ellsberg (1961), they show no ambiguity aversion. Looking at it through the lens of consumer theory, they have additive utility across parameters, thus perceiving parameters as perfect substitutes. At the opposite end of the spectrum, MEU agents act as if they have no reason to believe one state is more likely than any other, and "play it safe" by planning for the worst case scenario. They have maximal ambiguity aversion. In the language of consumer theory, MEU agents have a parameter aggregator I that is similar to Leontieff preferences, implying perfect complements. This underscores the analogy between complementarity and ambiguity aversion that will be formalized in Section 1.4.

Characterizing the MRA model is an important step towards putting SDT on a sound axiomatic foundation. However, applications are usually couched on more structured models, which impose further constraints on DM's preferences. Fortunately, the representation of the MRA model is a key piece of a methodology that can be used to import representation results from the Anscombe-Aumann framework to SDT. Indeed, eq. (1.1) indicates that one can interpret risk functions as acts (mappings from states of the world to consequences), and that preferences over decision rule–experiment pairs induce a preference relation over such acts. Therefore, once I have characterized the MRA model, obtaining a representation of preferences in the SDT setting reduces to obtaining an Anscombe-Aumann representation of the corresponding preferences over risk functions. Using this technique, one can translate axioms on preferences over acts into axioms on preferences over decision rules and experiments. Hence if a model has a representation satisfying the analogue of eq. (1.1) in the Anscombe-Aumann setting (Cerreia-Vioglio et al., 2011b), I can easily obtain its corresponding SDT representation.

The second piece needed to establish this toolkit is Lemma 3, which connects properties of the risk functionals r_u to those of the decision rules and experiments that define them. The lemma contains three separate statements. First, the risk function arising from a convex mixture of decision rules is the convex mixture of the risk functions induced by each individual decision rule. Second, the lemma describes, in terms of the decision rule and experiment that induces it, the risk function obtained by substituting the value of one risk function, on a given set of parameters, by the corresponding values of another risk function. Most axioms in the Anscombe-Aumann framework are stated in terms of these two operations on acts. Finally, the lemma also characterizes the set of decision rule–experiment pairs yielding constant risk functions. Constant acts also tend to be important ingredients for Anscombe-Aumann axioms. In summary, Lemma 3 translates the main ingredients of Anscombe-Aumann representations into the SDT framework.

From a purely technical perspective, Consequentialism transforms the problem of representing a preference over decision rules and experiments into representing a preference over parameter contingent action distributions. Much like I did with risk functions, I can also identify such parameter contingent distributions with Anscombe-Aumann acts. But while risk functions map parameters directly to parameter-independent consequences (utility units), the utility of an action distribution itself depends on the true parameter. Therefore, viewed through the lens of traditional decision theory, my methodology assists in generating representations of state-dependent preferences (over action distributions) from state-independent ones (over risk functions). This is a useful technique, because state-independent representations have been studied much more thoroughly than their state-dependent counterparts. The rationale behind developing such a methodology when studying SDT is that parameter-dependence is often the whole point of statistical applications — e.g., when performing inference.

After establishing the formal connection between the two frameworks, I apply my methodology to obtain behavioral characterizations in SDT, of some models for which axiomatizations already exist in the Anscombe-Aumann setting. These applications illustrate how my main results provide tools than can be used to obtain the exact behavioral implications of a wide array of SDT models. Specifically, I provide characterizations of statistical decision theoretic versions of subjective expected utility (Anscombe and Aumann, 1963); multiple priors and maximin expected utility (Gilboa and Schmeidler, 1989a; Stoye, 2011); and variational and multiplier preferences (Maccheroni, Marinacci, and Rustichini, 2006; Hansen and Sargent, 2001; Strzalecki, 2011).

In some applications, data on preferences over decision rule–experiment pairs is not forthcoming. Hence, to expand the scope of my main results, I characterize the MRA model for two alternative types of behavioral data.

First, I consider data in the form of parameter dependent stochastic choices. This consists of action probabilities conditional on every parameter, and can be obtained by calculating action frequencies from repeated observations of choices from different decision problems. Such data is commonplace in the psychometric literature, and has recently received some attention in economics (Caplin and Martin, 2015; Caplin and Dean, 2015). I characterize when a parameter dependent stochastic choice function can be rationalized by a preference on decision rule–experiment pairs satisfying the MRA representation. This allows one to test whether choices are compatible with the MRA model without having to observe preferences over decision rules and experiments. I also discuss the identification problem using such data. That is, when can the particular decision rule and experiment chosen by the DM be recovered by observing only parameter dependent stochastic choice data. This turns out to be a problem of statistical identifiability of mixture models, for which answers are available in the econometrics literature.

Second, I examine the case where data comes in the form of two collections of preferences. One defines a preference over decision rules for each fixed experiment. This describes a DM who takes the experiment as given, and chooses the decision rule accordingly. The other collection consists of a preference over experiments for each menu M of decision rules. This models an agent, whom I call the Experimenter, who chooses an experiment, assuming that a decision rule will subsequently be chosen from M according to some choice procedure known to her in advance. I separately characterize the MRA model for each of these decision problems. These can be viewed as two separate DMs — one choosing the information structure, the other choosing the decision rule — or as two cross sections of the same DM's preferences over decision rule–experiment pairs.

It turns out that the axioms characterizing the MRA model for the DM who chooses over decision rules are closely related to those characterizing the same model for decision rule–experiment pairs. The same is true for the Experimenter, apart from two extra axioms. Consistency implies that the Experimenter is forward looking and correctly anticipates the decision rule that will be subsequently chosen from the feasible menu. Optimism says that the Experimenter assumes that if multiple decision rules might be chosen in the second stage of the decision process, the ultimate choice will be in her favor.

The remainder of the paper is structured as follows. Section 1.2 contextualizes the contributions of the present paper within the existing literature. Section 1.3 presents the decision theoretical setting and introduces some notation. Section 1.4 contains the main result: a representation of the MRA model. Section 1.5 develops a methodology that can be used to apply existing behavioral foundations of classic decision theory to SDT, and illustrates it with some applications. In Section 1.6 I characterize the MRA model for alternative data sets. All proofs are in the appendix.

1.2 Related Literature

Statistical decision theory was pioneered by Abraham Wald, who first applied it to the optimal choice of decision rules (Wald, 1939), and then to the design of experiments (Wald, 1947b). Wald himself framed his theory as a particular case of John von Neumann's theory of games, with Nature and the statistician as the players. As such, in SDT the statistician is endowed with an utility function that maps actions and unobserved parameters to ex-post pay-offs. Since the sampling distributions of experiments at every parameter are assumed to be known in advance, the statistician uses the expected utility criterion when assessing the ex-ante payoff at each parameter. Payoff aggregation across the parameter space — on which there is no objectively given probability distribution — can then be done in different ways. For example, Wald favored the maximin criterion:

maximizing utility conditional on the worst-case parameter.[1]

This canonical model of statistical decision making has since provided the conceptual framework for many important results in mathematical statistics, such as David Blackwell's equivalency result for the economic comparison of experiments (Blackwell, 1953; Marschak and Miyasawa, 1968; Crémer, 1982), and Wald's complete class theorem (Wald, 1947a; Kuzmics, 2017). A comprehensive review of theoretical results and applications of SDT to mathematical statistics can be found in Inoue (2009).

SDT has also become one of the preferred languages of information economics. Among the many influential economics models couched in the SDT framework are the Bayesian persuasion literature initiated by Gentzkow and Kamenica (2011), and the rational inattention and costly information acquisition models studied by Sims (2003), Caplin and Dean (2015) and Matêjka and McKay (2015), among others. As for other areas of economic inquiry, Manski (2021) argues for the application of SDT to econometrics, provides a good summary of the relatively recent efforts in this direction, and outlines the remaining obstacles to this approach.

As was mentioned in the Introduction, models in SDT make predictions in terms of choices over decision rules and experiments. For example, Sims (2003) models a Bayesian agent who jointly chooses a decision rule and an experiment, subject to a constraint on the mutual information between the chosen experiment and the prior distribution over states of the world. The present paper characterizes the exact behavioral implications of this and many other SDT models, by axiomatizing preferences over decision rule–experiment pairs.

A different strand of the literature, which I will simply call decision theory, has sought to derive the behavioral implications of models of choice under uncertainty, usually by axiomatizing a DM's preferences over acts, i.e., functions from states of the world to final consequences. The early results most relevant for this paper were presented by Savage (1954) and Anscombe and Aumann (1963). Both characterize SEU agents, but unlike Savage, Anscombe and Aumann assume that acts' consequences consist of lotteries with known probability distributions. This introduces

[1] When the objective is to minimize loss rather than maximize utility, as in Wald's original formulation, maximin becomes minimax, since loss is the negative of utility.

both objective and subjective probabilities into the model. In Section 1.5, I show that there is a natural formal connection between the SDT setting and the Anscombe and Aumann framework of preferences over acts. I then leverage this connection to develop a methodology that generates representation results in the SDT setting by importing analogous results from the Anscombe and Aumann framework.

A large literature has sought to characterize different models of choice under uncertainty by modifying the set of axioms put forth by Anscombe and Aumann (1963). I briefly cite a few that are specially relevant in the context of this paper. Gilboa and Schmeidler (1989a) characterize the multiple priors expected utility (MPEU) model, where the DM has a set of prior beliefs over the states of the world, and picks the act yielding the best outcome according to the worst prior in this set. Maccheroni, Marinacci, and Rustichini (2006) generalize MPEU by characterizing preferences that can be represented by an elementary variational problem. Strzalecki (2011) provides a representation theorem for the multiplier preferences model proposed by Hansen and Sargent (2001). This is a special case of variational preferences, and can interpreted as modelling a DM who cares about the robustness of choices to deviations from a prior distribution. Cerreia-Vioglio et al. (2011b) axiomatize a general class of preferences, which subsumes all models cited in this paragraph and can be viewed as a decision theoretic analogue of the MRA model.

Some decision theory papers have studied statistical models from the perspective of the Anscombe and Aumann framework. One approach is to take the standard view that both signals and parameters of the SDT framework are contained in the state space of decision theory, and then proceed by proving results in the Anscombe and Aumann setting. With this interpretation, any decision theoretic representation can be directly applied to statistical decision problems. However, this leaves out an important feature of SDT, which is the natural decomposition of the states into an informative but payoff irrelevant part (signals) and a payoff relevant but unobservable component (parameters). Cerreia-Vioglio et al. (2020) and Amarante (2009), for example, specifically use statistical applications to motivate representation theorems in the Anscombe and Aumann setting.

A second approach, exemplified by Epstein and Seo (2010), Cerreia-Vioglio et al. (2013) and

Al-Najjar and De Castro (2014), also takes acts as primitives, and provides conditions on preferences under which the DM can be thought as having a parametric representation of the state space. In SDT terms, under such conditions preferences over acts can be used to elicit the DM's subjective statistical model (the experiment).

A different approach is taken by Stoye (2011) and Stoye (2012), who characterizes versions of many widely used SDT models by essentially taking risk functions (also called utility acts) as decision theoretic primitives. As will become clear in Section 1.5, such an approach is complementary to the one I take in the present paper, and can be combined with my results to obtain novel representation theorems in the SDT framework.

Finally, the decision theory paper that comes closest to modelling choices between decision rules and experiments as I do here, is due to Jakobsen (2021). In it, the author axiomatizes the decision problems of two agents. The first chooses between experiments, knowing that for each realized signal an act will later be chosen by the second agent. The preferences of each agent depend on both agents' choices. This is similar to the approach taken in Section 1.6.2, except that Jakobsen (2021) works with acts rather than decision rules, focuses on Bayesian representations for both agents, and considers only the case of state-independent utility.

After separately characterizing each agent, Jakobsen (2021) gives conditions under which the utility functions and prior beliefs in the representations are identified. He also provides the conditions for when both representations coincide, which can be interpreted as the two decision problems describing the same agent. Such a case can thus be viewed as a DM who chooses an act-experiment pair, albeit under the constraint that each experiment must always be paired with the act that is optimal for it. By assuming richer observable data, I am able to characterize a single agent's preferences over decision rule–experiment pairs without any such constraints.

1.3 Setting and notation

Experiments. Consider a set Θ of parameters, assumed to be a compact topological space, with a σ-algebra Σ that includes the singletons. The set of all experiments the decision maker may be asked to choose from is denoted by $\mathscr{P}$. Each element $P \in \mathscr{P}$ is a function $\theta \mapsto P_\theta$ from parameters to probability distributions over a standard Borel signal space, i.e., a Polish space (separable complete metric space) endowed with its Borel σ-algebra. This can be viewed as the parameter contingent distribution of a random variable, denoted by X_P. Since all signal spaces are standard Borel, there is no loss of generality in assuming that all $P \in \mathscr{P}$ are defined on a common Polish *sample space* X, and I do so for the remainder of the paper. To guarantee that the sample space is rich enough to incorporate signals from a variety of distributions, I assume that X is uncountable.

For any Polish space Y with the usual topology, denote by $\Delta(Y)$ the set of probability distributions on its Borel σ-algebra $\mathscr{Y}$. This is a convex space, with mixture operation defined by $(\alpha p + (1-\alpha)q)(E) = \alpha p(E) + (1-\alpha)q(E)$ for all $E \in \mathscr{Y}$, $p, q \in \Delta(Y)$ and $\alpha \in [0,1]$. I endow $\Delta(Y)$ with the topology of weak convergence of measures. That is, a sequence $(p_n)_{n\geq 1} \in \Delta(Y)$ converges to p if $\int_Y f \mathrm{d}p_n \to \int_Y f \mathrm{d}p$ for every bounded continuous function $f : Y \to \mathbb{R}$. By Prokhorov's theorem, this topology induces a metric on $\Delta(Y)$.

The class of all functions $\Theta \ni \theta \mapsto P_\theta \in \Delta(Y)$ will be denoted by $\Delta(Y)^\Theta$. This is also a convex space, with mixture operation defined point-wise on parameters. That is, for all $P, Q \in \Delta(Y)^\Theta$ and $\alpha \in [0,1]$, $\alpha P + (1-\alpha)Q = (\alpha P_\theta + (1-\alpha)Q_\theta)_{\theta\in\Theta} \in \Delta(Y)^\Theta$. For any $E \in \mathscr{Y}$ and $P \in \Delta(Y)^\Theta$, denote $P(E) = (P_\theta(E))_{\theta\in\Theta}$. Convergence is also defined point-wise, equipping $\Delta(Y)^\Theta$ with the product topology. When Θ is finite, this topology is equivalent to the one induced by Euclidean distance. I assume that $\mathscr{P}$ is a convex subspace of $\Delta(X)^\Theta$, in the sense that if $P, Q \in \mathscr{P}$, then $\alpha P + (1-\alpha)Q \in \mathscr{P}$ for all $\alpha \in [0,1]$.

Not knowing realized signal $x \in X$, conditional on $\theta \in \Theta$, is called *risk*, while lack of knowledge about the parameter itself is referred to as *ambiguity*. The former is objectively quantifiable, since the probabilities for a given $\theta \in \Theta$ are deemed objectively given, while the latter is not. Assume throughout that there exists at least one $P^* \in \mathscr{P}$ with the *full information property*: $(\operatorname{supp} P_\theta) \cap (\operatorname{supp} P_{\theta'}) = \emptyset$ for all $\theta \neq \theta'$.[2] This experiment's signal realizations perfectly reveal the true parameter, so call it the *fully informative* experiment. On the other hand, I call any

[2] For example, if $\Theta \subseteq X$, then P^* such that $P^*_\theta(\{\theta\}) = 1$ for every $\theta \in \Theta$ is fully informative.

$P^0 \in \mathscr{P}$ *uninformative* if $P^0_\theta = P^0_{\theta'}$ for all $\theta, \theta' \in \Theta$. Since the likelihood of different signals of an uninformative experiment does not depend on the parameter, observing a signal from P^0 does not provide any information about the true parameter value.

Decision rules. The decision maker observes a signal $x \in X$ coming from some experiment $P \in \mathscr{P}$ and then chooses an *action* from a set A. I assume $(A, \mathscr{A})$ is a compact Polish space with its Borel σ-algebra $\mathscr{A}$. A decision rule $\rho = \{\rho_P\}_{P \in \mathscr{P}}$ is a family of Markov kernels from X to A, indexed by the experiments.[3] Decision rules can be viewed as mixed strategies, one for each experiment, assigning a distribution over actions to each signal realization.

Any decision rule defined by $\rho_P(x, \cdot) = \chi_{d_P(x)}(\cdot)$, where $d_P : X \longrightarrow A$ for every $P \in \mathscr{P}$ and $x \in X$, and χ_a denotes the degenerate distribution with full mass on $a \in A$, is called a *pure* decision rule. These are decision rules that assign each signal to a single action with certainty. Denote by $\mathscr{D}$ the set of all decision rules, and by $\mathscr{D}^X$ the set of all pure decision rules. I assume that $\mathscr{D}$ is the class of all families of Markov kernels from X to A, indexed by the set $\mathscr{P}$. In other words, I require the DM to have preferences over all experiment- and signal-contingent distributions over actions.

A decision rule $\rho \in \mathscr{D}$ is called *invariant* if $\rho_P = \rho_{P'}$ for all $P, P' \in \mathscr{P}$. Examples of invariant decision rules in statistics include the sample mean and the least squares estimators. On the other hand, estimators which make use of the specification of the statistical model, such as maximum likelihood, are not invariant. Let $\overline{\mathscr{D}}$ be the set of all *pure and invariant* decision rules. Thus $\overline{\mathscr{D}} \subset \mathscr{D}^X \subset \mathscr{D}$. When it is clear from context, I slightly abuse notation and omit P subscripts from ρ_P. When it does not lead to confusion, I may also call ρ_P, for a fixed $P \in \mathscr{P}$, a decision rule.

Parameter contingent action distribution. For any $\rho \in \mathscr{D}$, the *parameter contingent action distribution* induced by ρ and P, ρP, is the mapping $\theta \mapsto \rho P_\theta \equiv \int_X \rho_P(x, \cdot) \mathrm{d}P_\theta(x)$. This gives the distribution of actions induced, for each parameter, by the signal distribution and the decision rule.

[3] That is, for every $P \in \mathscr{P}$, the mapping $\rho_P : X \times \mathscr{A} \to [0, 1]$ satisfies: (i) $\rho_P(x, \cdot)$ is a probability for every $x \in X$; and (ii) $\rho_P(\cdot, E)$ is measurable for every $E \in \mathscr{A}$.

In particular, for $\delta \in \mathscr{D}^{X}$, we have that $\delta P(F) = (P_\theta \circ \delta_P^{-1}(F))_{\theta\in\Theta}$ are the (parameter contingent) pushforward measures of δ_P acting on P. Let $\mathscr{F} \equiv \Delta(A)^{\Theta}$ and note that $\rho P \in \mathscr{F}$ for all ρ and P. For any function $v : A \to \mathbb{R}$ and $\rho \in \mathscr{D}$, I denote by $v(\rho_P) : X \to \mathbb{R}$ its expectation under ρ_P, i.e., $v(\rho_P(x)) = \int_A v(a)\rho_P(x, \mathrm{d}a)$ for each $x \in X$.

Preferences. Let $\mathscr{S} = \mathscr{D} \times \mathscr{P}$ be the set of all decision rule–experiment pairs, with typical element $\sigma \in \mathscr{S}$ – which I simply call a *pairing*. I model the decision maker's preferences as a binary relation $\succsim$ on $\mathscr{S}$. If preferences were over decision rules alone, they could be interpreted either as ex-ante preferences over contingent plans, or as ex-post preferences over actions after observing each signal realization. By letting the DM simultaneously choose over experiments and decision rules, I am implicitly assuming that these are ex-ante preferences, since experiments must be chosen before observing the signal. Therefore, decision rules are appropriately interpreted as contingent plans rather than realized actions after observing each signal. Indeed, I focus on the ex-ante strategies of dynamically consistent DMs, who believe they will follow through on their contingent plans after observing a signal.

As usual, $\succ$ and $\sim$ denote the asymmetric and the symmetric parts of $\succsim$, respectively. The relation $\succsim$ is called *trivial* if $\sigma \succ \sigma'$ for no $\sigma, \sigma' \in \mathscr{S}$. By taking a single preference relation as a primitive, I am ruling out SDT models where preferences depend on the choice problem being considered. One prominent such model is minimax regret.

Example 1 (Minimax Regret): Let Θ be compact and $u : A \times \Theta \to \mathbb{R}$ be a continuous utility function. The minimax regret choice criterion C^R selects, for any choice problem $D \times \Gamma \subseteq \mathscr{S}$ such that $\{\rho P \in \mathscr{F} : \rho \in D,\ P \in \Gamma\}$ is compact,

$$C^R(D \times \Gamma) = \operatorname*{arg\,min}_{(\rho,P)\in D\times\Gamma} \max_{\theta\in\Theta} \left[\max_{a\in A_D} u(a,\theta) - \int_X u(\rho,\theta)\mathrm{d}P_\theta \right],$$

where $A_D = \{\rho(X) : \rho \in D\}$. In words, the DM chooses the alternative that minimizes, for the worst-case scenario, the difference between the optimal choice with perfect information and the expected utility of the pairing.

Clearly, the ranking of pairings under the minimax regret criterion may depend on the specific choice problem being considered, hence can not be represented by a single preference relation $\succsim$.‖

Let f and g be functions defined on some set Y. For any $E \subseteq Y$, let $f_{(E)}g$ denote the function given by $f_{(E)}g(y) = f(y)$ if $y \in E$ and $f_{(E)}g(y) = g(y)$ if $y \in Y \setminus E$. Recalling the definition of the fully informative experiment P^*, denote $S_\theta = \operatorname{supp} P^*_\theta$ and note that $S_\theta \cap S_{\theta'} = \emptyset$ for all $\theta \neq \theta'$. Moreover, define $S_T \equiv \bigcup_{\theta \in T} S_\theta$, for any $T \subseteq \Theta$. To economize on notation, let $\rho_{(S_T)}\tau \equiv \rho_T\tau$, for any $\rho, \tau \in \mathscr{D}$ and $T \subseteq \Theta$. Also denote by $\mathbf{1}_E$ the indicator function of $E \subseteq Y$, that is, $\mathbf{1}_E(y) = 1$ if $y \in E$ and $\mathbf{1}_E(y) = 0$ otherwise.

Risk functions. Given a utility function $u : A \times \Theta \to \mathbb{R}$ which is continuous in the first argument, define the *risk function* of the pairing $(\rho, P) \in \mathscr{S}$ under u as the function $r_u(\rho, P) : \Theta \to \mathbb{R}$ given by

$$r_u(\rho, P)(\theta) = \int_X u(\rho, \theta) \mathrm{d}P_\theta \text{ for every } \theta \in \Theta. \tag{1.2}$$

In words, the risk function describes the parameter contingent expected utility of a pairing. Define $\mathscr{R}_u \equiv \{r_u(\rho, P) \in \mathbb{R}^\Theta : (\rho, P) \in \mathscr{S}\}$, the set of all possible risk functions under u. Finally, a functional $I : \mathscr{R}_u \to \mathbb{R}$ is called *monotone* if $r(\theta) \geq r'(\theta)$ for all $\theta \in \Theta$ implies $I(r) \geq I(r')$; it is called *continuous* if it is continuous in the topology of point-wise convergence.

1.4 Main results

In this section, I present axioms that characterize a general model of statistical decisions. This will provide a foundation to obtain representations of many models of interest, including SDT versions of subjective expected utility (SEU), maximin expected utility (MEU) and multiplier preferences. We are interested in models of the following form.

Definition 1 (Monotone Risk Aggregation Representation). A binary relation $\succsim$ on $\mathscr{S}$ has *a monotone risk aggregation (MRA) representation* (u, I) if there exists a utility function $u : A \times \Theta \to \mathbb{R}$, continuous in the first argument, and a monotone and continuous functional $I : \mathscr{R}_u \to \mathbb{R}$, such that for all $\rho, \tau \in \mathscr{D}$ and $P, Q \in \mathscr{P}$,

$$(\rho, P) \gtrsim (\tau, Q) \iff I(r_u(\rho, P)) \geq I(r_u(\tau, Q)), \tag{1.3}$$

where $r_u : \mathscr{S} \to \mathbb{R}^\Theta$ is defined by eq. (1.2). ◇

A decision maker whose preferences are compatible with the MRA model evaluates prospects according to expected utility when probabilities are known, and aggregates unknown parameters in a way that favors point-wise improvements of the risk function on the parameters. Importantly for statistical applications, the ex-post utility u is allowed to be parameter-*dependent*. This is in contrast to most of the decision theory under uncertainty literature, where the Bernoulli utility is defined on ultimate consequences, and thus is assumed to be state-independent. In fact, the model of Definition 1 can be viewed as a state-dependent, statistical version of the monotone, Bernoullian and Archimedean (MBA) preferences studied by Cerreia-Vioglio et al. (2011b).

Unsurprisingly given its generality, most objective functions commonly used in statistical decision problems fit into this basic framework, differing only in the choice of utility function u and aggregator I.

Example 2 (Bayesian parameter estimation): A Bayesian statistician wants to estimate a parameter $\theta \in \Theta$ with the least possible mean squared error. Assume she has a prior belief $\pi \in \Delta(\Theta)$ over possible parameter values. In terms of the MRA framework, we have $A = \Theta$, and $u(a, \theta) = -(a - \theta)^2$ and $I(\cdot) = \mathbb{E}_\pi(\cdot)$. For a fixed experiment P, the statistician's preference is given by $(\rho, P) \gtrsim (\tau, P)$ if, and only if, $\mathbb{E}_\pi(\int_X (\rho - \theta)^2 \mathrm{d}P_\theta) \leq \mathbb{E}_\pi(\int_X (\tau - \theta)^2 \mathrm{d}P_\theta)$. ||

Example 3 (Ellsberg preferences): A decision maker is asked to place a bet on which color ball will be drawn from an urn containing 30 blue and 60 green or yellow balls. We can model this decision problem by setting $\Theta = \{(n_g, n_y) : n_g = 0, \ldots, 60,\ n_y = 60 - n_g\}$, $X = \{b, g, y\}$ and $A = \{0, 1\}$. The interpretation of an action $a \in A$ is of receiving a dollars. Thus, the decision rule $\delta_P = (\delta_P(b) = 1, \delta_P(g) = 0, \delta_P(y) = 1)$ corresponds to betting on either a blue or a yellow ball being drawn from experiment $P \in \mathscr{P}$.

Let $P \in \mathscr{P}$ be the experiment where the balls are drawn according to an uniform distribution. Then $P_\theta(b) = 1/3$ and $P_\theta(g) = \theta/90 = 2/3 - P_\theta(y)$. Typical Ellsberg choices for this decision problem imply $((1,0,0)_P, P) \succ ((0,1,0)_P, P)$ and $((0,1,1)_P, P) \succ ((1,0,1)_P, P)$, where $(b, g, y)_P$ denotes any decision rule $d \in \mathscr{D}^X$ such that $\delta_P = (b, g, y)$. Such choice data is not consistent with any version of SEU preferences, but can be rationalized by MEU preferences: $(\delta, P) \succsim (\delta', Q)$ if, and only if, $\min_{\theta\in\Theta} \int_X u(\delta) \mathrm{d}P_\theta \geq \min_{\theta\in\Theta} \int_X u(\delta') \mathrm{d}Q_\theta$ for an increasing, state-independent utility function u. Here, the aggregator $I(\cdot)$ corresponds to $\min_{\theta\in\Theta}(\cdot)$. $\|$

First, I impose a basic rationality postulate on the preference relation.

Axiom 1 (Weak Order): The preference relation $\succsim$ is complete and transitive.

The next axiom needed to characterize the MRA model captures a notion of the DM being a consequentialist. It states that the decision maker's preferences ultimately depend only on the parameter contingent action distribution induced by the pairings, not on the particular decision rule and experiment that generate this distribution.

Axiom 2 (Consequentialism): For all $(\rho, P), (\tau, Q) \in \mathscr{S}$: if $\rho P = \tau Q$, then $(\rho, P) \sim (\tau, Q)$.

Consequentialism is not a completely innocuous axiom. For instance, it may not hold in models where choosing different experiments entail different subjective costs.

Example 4: Consider a rationally inattentive DM of the kind postulated by Matêjka and McKay (2015). Suppose $A = \{-1, 1\}$, $\Theta = \{-1, 0, 1\}$ and $X = \{x_{-1}, x_0, x_1\}$. The DM is a Bayesian, with a uniform prior belief $\pi \in \Delta(\Theta)$, and utility function $u(a, \theta) = \theta a$. She also faces a cost to choosing experiment $P \in \mathscr{P}$ given by

$$k(P) = \sum_{\theta\in\Theta} \pi(\theta) \sum_{x\in X} P_\theta(x) \left[\log P_\theta(x) - \log \sum_{\theta\in\Theta} \pi(\theta) P_\theta(x) \right].$$

The DM chooses a decision rule δ and experiment P to maximize

$$V(\delta, P) = \sum_{\theta\in\Theta} \sum_{x\in X} u(\delta(x), \theta) P_\theta(x) \pi(\theta) - k(P).$$

If the DM can obtain conclusive information about whether $\theta = -1$, further telling $\theta = 0$ apart from $\theta = 1$ is costly, but does not affect the optimal decision.

For instance, take the fully informative experiment $P^*_j(x_j) = 1$ for $j = -1, 0, 1$, and compare it to P' such that $P'_{-1}(x_{-1}) = 1$ and $P'_0(x_0) = P'_1(x_0) = P'_0(x_1) = P'_1(x_1) = 1/2$. Then $k(P^*) > k(P')$. It can be easily verified that if δ' and δ^* denote optimal decision rules given P' and P^* respectively, then the DM strictly prefers (δ', P') to (δ^*, P^*), although they induce the same parameter contingent action distributions.

If the DM had instead been modelled as in Sims (2003), where she faces a hard constraint on the mutual information between the experiment and the prior and solves $\max_{\delta,P} V(\delta, P)$ subject to $k(P) \leq K$, $K > 0$, then Consequentialism would be satisfied. ||

The sets of pairings with the same induced distribution partition $\mathscr{S}$ into equivalence classes. For any $\sigma \in \mathscr{S}$, denote by $[\sigma]$ the equivalence class to which σ belongs, and let $\mathscr{S}/_*$ be the set of all such equivalence classes. The space of decision rules is rich enough that every equivalence class in $\mathscr{S}/_*$ has a representative involving the fully informative experiment. When given any pairing, the DM can mimic the parameter dependent action distribution arising from any garbling of the given experiment by appropriately randomizing the decision rule. This allows the DM to achieve any parameter contingent action distribution when given the fully informative experiment, since any any other experiment is a garbling of it (Blackwell, 1951). This is stated formally as follows.

Lemma 1. For all $[\sigma] \in \mathscr{S}/_*$, there exists $\rho^\sigma \in \mathscr{D}$ such that $(\rho^\sigma, P^*) \in [\sigma]$.

Axiom 2 guarantees that $\mathscr{S}/_*$ is a refinement of $\mathscr{S}/_\sim$, the quotient space of $\sim$. In other words, if two pairings induce the same distribution, they are deemed indifferent, but the converse is not necessarily true. Therefore, transitivity of $\succsim$ implies that, to obtain a characterization of the full preference relation, it suffices to characterize the restriction of $\succsim$ to a single representative of each member of $\mathscr{S}/_*$, since the DM is indifferent between all pairings within the same equivalence class. Now, Lemma 1 states that any parameter contingent distribution induced by an element of

$\mathscr{S}$ can be achieved by some decision rule acting on P^*. Therefore, in the presence of Axiom 2 and transitivity, $\succsim$ is completely determined by comparisons of pairings involving the fully informative experiment.

When the DM is provided the fully informative experiment P^*, then there is effectively no uncertainty about the parameter. Therefore, when comparing decision rules, she can focus on their induced actions at each parameter separately. This suggests that two decision rules that coincide on S_θ^c should be ranked in the same way, regardless of what specific actions each rule prescribes for this set of signals. This intuitive notion is captured by the next axiom.

Axiom 3 (Independence of Irrelevant Parameters): For all $\theta \in \Theta$ and $\rho, \tau, \gamma, \gamma' \in \mathscr{D}$:

$$(\rho_{\{\theta\}}\gamma, P^*) \succsim (\tau_{\{\theta\}}\gamma, P^*) \implies (\rho_{\{\theta\}}\gamma', P^*) \succsim (\tau_{\{\theta\}}\gamma', P^*).$$

In the presence of the fully informative experiment, Independence of Irrelevant Parameters (IIP) allows us to interpret $\{\rho_{\{\theta\}}\gamma : \rho \in \mathscr{D}\}$, for any $\gamma \in \mathscr{D}$, as the set of decision rules conditional on θ, since it implies that the ranking of such rules does not depend on what actions they prescribe on S_θ^c. Axiom 3 is the SDT formulation of a weaker version of the “sure thing principle”, presented by Savage (1954). The main conceptual difference is that Savage’s postulate holds when conditioning the decision rules on any subset $T \in \Sigma$, while IIP is only required to hold for singletons $\{\theta\} \in \Sigma$.

The following is a monotonicity axiom: if one decision rule is preferred to another after conditioning on every parameter, then it is preferred unconditionally.

Axiom 4 (Monotonicity): For all $\rho, \tau \in \mathscr{D}$: if $(\rho_{\{\theta\}}\gamma, P^*) \succsim (\tau_{\{\theta\}}\gamma, P^*)$ for every $\gamma \in \mathscr{D}$ and $\theta \in \Theta$, then $(\rho, P^*) \succsim (\tau, P^*)$.

If the DM is choosing between decision rules to pair with the fully informative experiment, the problem reduces to picking a terminal distribution over actions conditional on each parameter being the truth. Now suppose the actions induced by the available decision rules coincide on every parameter except one. Then only the objectively given probabilities, conditional on the relevant parameter, should matter. Since this choice only involves quantifiable risk, modifying the distributions conditional on the relevant parameter in the same way, for all decision rules under consideration, does not provide hedging value. Therefore, such a modification should not change preferences. This gives an intuitive interpretation to the following axiom.

Axiom 5 (Conditional Mixture Independence): For every $\rho, \tau \in \mathscr{D}$ and $\theta \in \Theta$: if $(\rho_{\{\theta\}}\gamma, P^*) \succsim (\tau_{\{\theta\}}\gamma, P^*)$, then $(\alpha\rho_{\{\theta\}}\gamma + (1-\alpha)\kappa_{\{\theta\}}\gamma, P^*) \succsim (\alpha\tau_{\{\theta\}}\gamma + (1-\alpha)\kappa_{\{\theta\}}\gamma, P^*)$ for all $\alpha \in (0, 1]$ and $\kappa \in \mathscr{D}$.

Conditional Mixture Independence (CMI) implies that the DM views randomization of actions conditional on a given parameter as objective risk. Therefore, she is able to evaluate distributions conditional on each parameter via expected utility.

Finally, I impose a form of continuity on the preference relation. This guarantees that the preference order is not reversed for pairings with arbitrarily similar induced distributions.

Axiom 6 (Continuity): The set $\{(\rho P, \tau Q) \in \mathscr{F}^2 : (\rho, P) \succsim (\tau, Q)\}$ is closed in $\mathscr{F}^2$.

Note that, while axioms 3 to 5 are essentially properties of preferences on decision rules alone, the other axioms act on both decision rules and experiments. Therefore, choice data that includes simultaneous variation of decision rules and experiments is needed to falsify axioms 1, 2 and 6. In Section 1.6, I show how to characterize a dominance representation from preferences defined exclusively on either decision rules or experiments.

For the remainder of this section, I show that if $\succsim$ satisfies the axioms above, then one can describe a sub-relation $\hat{\succsim} \subseteq \succsim$ which captures the decision maker's ex-post utility function u. This in turn reveals the set $\mathscr{R}_u$ of all risk functions available to the DM. The aggregator I in Definition 1 can then be viewed as an ordinal utility on the space of risk functions.

1.4.1 The dominance sub-relation

The first step in obtaining an MRA representation is to characterize a behavioral version of the concept of *dominance*, made precise in the following definition.

Definition 2 (Dominance Representation). A binary relation $\succsim$ on $\mathscr{S}$ has a *dominance representation* if there exists a utility function $u : A \times \Theta \to \mathbb{R}$, continuous in the first argument, such that for all $\rho, \tau \in \mathscr{D}$ and $P, Q \in \mathscr{P}$,

$$(\rho, P) \succsim (\tau, Q) \iff \int_X u(\rho, \theta) \mathrm{d}P_\theta \geq \int_X u(\tau, \theta) \mathrm{d}Q_\theta \text{ for all } \theta \in \Theta. \tag{1.4}$$

If $\succsim$ has a dominance representation, we call it a dominance relation. ◇

When an (incomplete) preference has a dominance representation, a pairing is preferred to another if and only if it yields higher expected utility for every parameter. If one views a statistical decision problem as a game where Nature chooses the true unknown parameter, as in the tradition of Wald (1949), then Definition 2 corresponds to (weak) dominance in mixed strategies. Furthermore, if one fixes an experiment and looks at the corresponding preferences over decision rules induced by a dominance representation, one obtains the statistical concept of *admissibility*.

Consider the following two axioms: one is the converse of Monotonicity, and the other is a weakening of completeness.

Axiom 7 (Admissibility): For all $\rho, \tau \in \mathscr{D}$: $(\rho, P^*) \succsim (\tau, P^*) \implies (\rho_{\{\theta\}}\gamma, P^*) \succsim (\tau_{\{\theta\}}\gamma, P^*)$ for every $\gamma \in \mathscr{D}$ and $\theta \in \Theta$.

Axiom 8 (Conditional Completeness): For all $\rho, \tau, \gamma \in \mathscr{D}$ and $\theta \in \Theta$: $(\rho_{\{\theta\}}\gamma, P^*) \succsim (\tau_{\{\theta\}}\gamma, P^*)$ or $(\tau_{\{\theta\}}\gamma, P^*) \succsim (\rho_{\{\theta\}}\gamma, P^*)$.

The following result characterizes the dominance representation in terms of preferences, and presents its uniqueness properties.

Theorem 1. *A binary relation $\succsim$ on $\mathscr{S}$ is a transitive and reflexive preference that satisfies Consequentialism, Independence of Irrelevant Parameters, Monotonicity, Conditional Mixture Independence, Continuity, Admissibility, and Conditional Completeness if, and only if, it has a dominance representation with utility function $u : A \times \Theta \to \mathbb{R}$.*

Furthermore, u is parameter-wise cardinally unique: if u' also represents $\succsim$, then there exist $\{(b_\theta, c_\theta) : \theta \in \Theta\}$ with $b_\theta > 0$ and $c_\theta \in \mathbb{R}$, such that $u'(\cdot, \theta) = b_\theta u(\cdot, \theta) + c_\theta$ for all $\theta \in \Theta$.

Dominance relations are of independent interest, as they behaviorally characterize the important concepts of strategic dominance and admissibility. However, my main purpose in proving Theorem 1 is to find a particular subrelation of the DM's full preferences that is a dominance relation, and thus can behaviorally elicit the DM's utility, and consequently their risk functions.

I say that σ is unanimously preferred to σ' if, conditional on each parameter, the outcome from σ is deemed preferred to that of σ'. Formally:

Definition 3. Let $\sigma, \sigma' \in \mathscr{S}$ and $\succsim$ be a preference. Then σ *is unanimously preferred to* σ', denoted by $\sigma \mathrel{\hat{\succsim}} \sigma'$, if there exist $(\rho, P^*) \in [\sigma]$ and $(\tau, P^*) \in [\sigma']$ such that $(\rho_{\{\theta\}}\gamma, P^*) \succsim (\tau_{\{\theta\}}\gamma, P^*)$ for all $\theta \in \Theta$ and $\gamma \in \mathscr{D}$. ◊

Let $\succsim$ satisfy axioms 1 to 6. A straightforward consequence of Lemma 1 is that $\hat{\succsim}$ is reflexive, therefore non-empty. It is also usually incomplete, but retains many of the properties of $\succsim$.

Proposition 1. *If $\succsim$ satisfies axioms 1 to 6, the following statements hold:*

1. *If $\sigma \mathrel{\hat{\succsim}} \sigma'$, then $\sigma \succsim \sigma'$. Moreover, $\hat{\succsim}$ is transitive.*

2. *$\hat{\succsim}$ is a dominance relation and, for every other dominance sub-relation $\succsim' \subseteq \succsim$, we have $\succsim' \subseteq \hat{\succsim}$.*

3. *$\hat{\succsim}$ satisfies the sure-thing principle: for all $\rho, \tau, \gamma, \gamma'$ and $T \in \Sigma$,*

$$(\rho_T\gamma, P^*) \mathrel{\hat{\succsim}} (\tau_T\gamma, P^*) \iff (\rho_T\gamma', P^*) \mathrel{\hat{\succsim}} (\tau_T\gamma', P^*).$$

4. *$\hat{\succsim}$ satisfies mixture independence: for all $\rho, \tau, \gamma \in \mathscr{D}$, $P \in \mathscr{P}$ and $\alpha \in (0, 1]$,*

$$(\rho, P) \mathrel{\hat{\succsim}} (\tau, P) \iff (\alpha\rho + (1-\alpha)\gamma, P) \mathrel{\hat{\succsim}} (\alpha\tau + (1-\alpha)\gamma, P).$$

5. *If $\hat{\succsim}$ is trivial, so is $\succsim$.*

The following result characterizes the unanimously preferred relation.

Lemma 2. Let $\succsim$ satisfy axioms 1 to 6. Then the unanimously preferred relation $\hat{\succsim}$ has a dominance representation with a utility $u : A \times \Theta \to \mathbb{R}$ that is continuous in the first argument. Moreover, u is parameter-wise cardinally unique.

In other words, the unanimously preferred relation $\hat{\succsim}$ allows us to elicit the DM's risk functions. In Section 1.4.2, I will use this fact to obtain a representation of the MRA model.

The representation in Lemma 2 also implies that, given the fully informative experiment, there exists unambiguously best and worst decision rules. Fix $\overline{a}_\theta \in \arg\max_{a \in A} u(a, \theta)$ and $\underline{a}_\theta \in \arg\min_{a \in A} u(a, \theta)$, and set $\overline{\rho}(x, \cdot) = \chi_{\overline{a}_\theta}$ and $\underline{\rho}(x, \cdot) = \chi_{\underline{a}_\theta}$ for all $x \in S_\theta$, $\theta \in \Theta$. Then $(\overline{\rho}, P^*) \mathrel{\hat{\succsim}} \sigma \mathrel{\hat{\succsim}} (\underline{\rho}, P^*)$, for all $\sigma \in \mathscr{S}$. I can then provide the following characterization, which will be useful going forward.

Definition 4. For any dominance relation $\hat{\succsim}$,

$$\mathscr{K}(\hat{\succsim}) = \left\{ \sigma \in \mathscr{S} : \sigma \mathrel{\hat{\sim}} (\alpha\overline{\rho} + (1-\alpha)\underline{\rho}, P^*),\ \alpha \in [0, 1] \right\} \tag{1.5}$$

defines its set of *constant-risk-equivalent* (CRE) pairings. ◇

The set $\mathscr{K}(\hat{\succsim})$ amounts to a behavioral characterization of constant-utility pairings, since members of $\mathscr{K}(\hat{\succsim})$ are behaviorally equivalent to pairings with a constant risk function. With state-independent utility, the role of $\mathscr{K}(\hat{\succsim})$ is played by pairings which induce parameter contingent distributions over actions that are constant across parameters. Since I am working within the more general framework of parameter-dependent utility, CRE pairings must be elicited from preferences. The interpretation of eq. (1.5) as a characterization of constant-utility acts will be formalized in Section 1.5, Lemma 3.

1.4.2 Attitudes toward ambiguity

I now construct the decision maker's risk functions from her preferences. Let $\succsim$ satisfy axioms 1 to 6 and fix a utility function u that represents its unanimously preferred relation $\hat{\succsim}$. Consider the mapping $r_u : \mathscr{S} \to \mathbb{R}^\Theta$, defined by eq. (1.2), and the corresponding set $\mathscr{R}_u$. Define a preference $\geq^u$ on $\mathscr{R}_u$ as follows:

$$\forall r, r' \in \mathscr{R}_u : \ r \geq^u r' \iff \exists \sigma \succsim \sigma' \text{ such that } r = r_u(\sigma) \text{ and } r' = r_u(\sigma'). \tag{1.6}$$

Let $\simeq^u$ and $>^u$ denote the symmetric and asymmetric parts of $\geq^u$, respectively. By construction, if $r \geq r'$, then $r \geq^u r'$, where $\geq$ denotes the usual order on $\mathbb{R}^\Theta$. Further, since $\succsim$ is complete, so is $\geq^u$. Also note that $\sigma' \in [\sigma]$ implies $r_u(\sigma) = r_u(\sigma')$.

The following is the main result of the paper.

Theorem 2. *A preference $\succsim$ satisfies Consequentialism, Weak Order, Independence of Irrelevant Parameters, Monotonicity, Conditional Mixture Independence and Continuity if, and only if, its unanimously preferred relation $\hat{\succsim}$ has a dominance representation with utility $u : A \times \Theta \to \mathbb{R}$ and there exists a monotone, continuous functional $I : \mathscr{R}_u \to \mathbb{R}$ such that, for all $\sigma, \sigma' \in \mathscr{S}$,*

$$r_u(\sigma) \geq^u r_u(\sigma') \iff I(r_u(\sigma)) \geq I(r_u(\sigma')). \tag{1.7}$$

Therefore, $\succsim$ has an MRA representation (u, I) if, and only if, it satisfies axioms 1 to 6. Moreover, u is parameter-wise cardinally unique and there exists a representation $(\tilde{u}, \tilde{I})$ such that $\tilde{u}(\cdot, \theta)$ is constant for no $\theta \in \Theta$.

Theorem 2 characterizes preferences that are compatible with the MRA model, and shows that the functional I in Definition 2 is essentially a utility on the space of risk functions. Indeed, the proof boils down to applying a general utility representation theorem, found in Herden (1989), to the preference $\geq^u$ on $\mathscr{R}_u$.

Attitudes toward uncertainty can then be viewed as patterns of complementarity and substitutability between risk under different parameters. Decision makers who want to hedge act as if risk is *complementary* across parameters: they favor pairings that yield more balanced pay-off profiles across parameters over those that have a particularly large expected utility on any specific event $T \in \Sigma$, but not on others. The following example formally illustrates this intuition.

Example 5 (Uncertainty attitudes as parameter complementarity): Take two preferences $\succsim_1$ and $\succsim_2$ such that $\hat{\succsim}_1 = \hat{\succsim}_2 = \hat{\succsim}$. Following Ghirardato and Marinacci (2002), I say that $\succsim_1$ is more averse to ambiguity than $\succsim_2$ if for all $\kappa \in \mathscr{K}(\hat{\succsim})$, $\sigma \succsim_1 \kappa$ implies $\sigma \succsim_2 \kappa$. In other words, whenever $\succsim_1$ prefers an uncertain pairing σ to a CRE pairing κ, so does $\succsim_2$. This is a widely accepted way of comparing preferences in terms of ambiguity aversion (Maccheroni, Marinacci, and Rustichini, 2006; Cerreia-Vioglio et al., 2011a).

Now consider a family of preferences $\{\succsim_s : s \in \mathbb{R} \cup \{-\infty\}\}$ with the same unanimously preferred relation — i.e., $\hat{\succsim}_s = \hat{\succsim}_{s'} = \hat{\succsim}$ for all $s, s' \in \mathbb{R} \cup \{-\infty\}$. Without loss of generality, normalize the representations $\{u_s\}$ of $\{\hat{\succsim}_s\}$ so that $\mathscr{R}_{u_s} = [0,1]^\Theta$ for all s. Under this normalization, $u_s = u$ for all s, and the risk functions associated with members of $\mathscr{K}(\hat{\succsim})$ are constant. Assume that $\{\succsim_s\}$ has a representation (I_s, u) by a Dixit and Stiglitz (1977) constant elasticity of substitution (CES) function, with (full support) shares $\mu \in \Delta(\Theta)$ and substitution parameter s:

$$I_s(r_u(\sigma)) = \begin{cases} \left(\int_\Theta r_u(\sigma)^s \mathrm{d}\mu\right)^{\frac{1}{s}}, & s \notin \{0, -\infty\} \\ \exp\left(\int_\Theta \ln r_u(\sigma) \mathrm{d}\mu\right), & s = 0 \\ \inf_{\theta \in \Theta} r_u(\sigma)(\theta), & s = -\infty \end{cases}$$

for all $\sigma \in \mathscr{S}$. The elasticity of substitution between any two goods (or parameters) of a CES utility function is given by $\varepsilon = 1/(1-s)$. Thus, $s = 1$ corresponds to unbounded elasticity of substitution, while $s = -\infty$ implies perfect complementarity. Moreover, it can be shown that $\lim_{s \to s^*} I_s = I_{s^*}$ for $s^* \in \{0, -\infty\}$. CES utility is a particular case of second order expected utility, which was axiomatized for the setting of Savage (1954) by Neilson (2010).

If $t > s$ and $t, s \neq 0$, we have, for all $\sigma \in \mathscr{S}$,

$$I_t(r_u(\sigma)) = \left(\int_\Theta [r_u(\sigma)^s]^{\frac{t}{s}} \mathrm{d}\mu\right)^{\frac{s}{ts}} \leq \left(\int_\Theta r_u(\sigma)^s \mathrm{d}\mu\right)^{\frac{s}{ts}\frac{t}{s}} = I_s(r_u(\sigma)),$$

where I have used Jensen's inequality. Since for all $\kappa \in \mathscr{K}(\hat{\succsim})$ and $t, s \in \mathbb{R} \cup \{-\infty\}$, we have $I_t(r_u(\kappa)) = I_s(r_u(\kappa))$, then $\sigma \succsim_s \kappa$ implies $\sigma \succsim_t \kappa \iff I_s(r_u(\sigma)) \geq I_t(r_u(\sigma)) \iff s \leq t$. Thus, ambiguity aversion in inversely related to the elasticity of substitution. In particular, an SEU agent ($s = 1$) shows no ambiguity aversion, treating parameters as perfect substitutes. On the other hand, an MEU agent ($s = -\infty$) has maximal ambiguity aversion, and thus perfect complementarity across parameters. ||

1.5 Applications

In the Anscombe-Aumann framework, an *act* is a measurable function from some set Ω of states of the world to a convex space of outcomes Y. We can thus identify each risk function of SDT with an Anscombe-Aumann act from $\Theta \equiv \Omega$ to $\mathscr{R}_u \equiv Y$. The representation in Theorem 2 implies that if $\succsim$ satisfies axioms 1 to 6, then $\succeq^u$ satisfies the following essential axioms of decision theory under uncertainty:

- **Weak Order**: the preference $\succeq^u$ is complete and transitive.
- **A-A Monotonicity**: for all $r, r' \in \mathscr{R}_u$, if $r \geq r'$, then $r \succeq^u r'$.
- **Risk Independence:** for all constant risk functions $r, r', q \in \mathscr{R}_u$ and $\alpha \in [0,1]$, $r \succeq^u r'$ implies $\alpha r + (1-\alpha)q \succeq^u \alpha r' + (1-\alpha)q$.
- **Mixture Continuity:** the sets $\{\alpha \in [0,1] : \alpha r + (1-\alpha)r' \succeq^u q\}$ and $\{\alpha \in [0,1] : q \succeq^u \alpha r + (1-\alpha)r'\}$ are closed for all $r, r', q \in \mathscr{R}_u$.

Most models of complete preferences in the Anscombe-Aumann framework satisfy these same properties, and impose extra axioms to obtain more structure on I. Thus there is hope that, by relating properties of $\succeq^u$ with the corresponding axioms on $\succsim$, I can import many of the existing representations from the Anscombe-Aumann framework to SDT. The following lemma is an important step towards establishing this connection.

Lemma 3. Suppose $\hat{\succsim}$ has a dominance representation with utility function u. The following statements hold:

1. $r_u(\alpha\rho + (1-\alpha)\tau, P) = \alpha r_u(\rho, P) + (1-\alpha) r_u(\tau, P)$ for all $\tau \in \mathscr{D}$ and $\alpha \in [0,1]$.

2. $r_u(\rho_T\gamma, P^*) = r_u(\rho, P^*)_{(T)} r_u(\gamma, P^*)$ for all $\rho, \gamma \in \mathscr{D}$ and $T \in \Sigma$.

3. There exists $\tilde{u}$ also representing $\hat{\succsim}$ such that $r_{\tilde{u}}(\sigma)$ is constant if, and only if, $\sigma \in \mathscr{K}(\hat{\succsim})$.

The usefulness of Lemma 3 stems from the fact that most axioms in the decision theory under uncertainty paradigm are stated in terms of (i) mixture of acts, as in the first statement; (ii) combinations of acts, i.e., substituting an act's consequences on a set of parameters for the consequences of a different act, as in statement 2; or (iii) properties of preferences over mixtures or combinations with constant-utility acts, as characterized by $\mathscr{K}(\hat{\succsim})$.

Next, I illustrate this methodology by applying Lemma 3 to a variety of existing axiomatizations of preferences in the Anscombe-Aumann framework, to obtain SDT versions of their representations. Throughout this section, I maintain the assumption that every experiment in $\mathscr{P}$ is measurable with respect to Σ, which makes every risk function also measurable.

1.5.1 Subjective expected utility

I begin by obtaining a SDT representation of the subjective expected utility (SEU) model. Recall that, in the Anscombe-Aumann setting, state-independent SEU is characterized by preferences over acts satisfying Weak Order, A-A Monotonicity, Mixture Continuity, and the following strengthening of Risk-Independence (which I state here in terms of $\geq^u$):

- **Independence:** for all $r, r', q \in \mathscr{R}_u$, if $r \geq^u r'$, then $\alpha r + (1-\alpha) q \geq^u \alpha r' + (1-\alpha) q$ for all $\alpha \in (0,1)$.

These four axioms imply the existence of a representation by SEU with *finitely additive* prior probability. To guarantee that the prior is *countably additive*, one needs an additional postulate, due to Arrow (1971).

- **Monotone Continuity:** if $r, r', q \in \mathscr{R}_u$, q is constant, $\{T_n\}_{n\geq 1} \in \Sigma$ with $T_1 \supseteq T_2 \supseteq \cdots$ and $\bigcap_{n\geq 1} T_n = \emptyset$, then $r \succ^u r'$ implies that there exists $n_0 \geq 1$ such that $q_{(T_{n_0})} r \succ^u r'$.

Given statement 1 in Lemma 3, it is straightforward to translate Independence of $\succeq^u$ into an axiom on $\gtrsim$:

Axiom 9 (Mixture Independence): For every $\rho, \tau, \gamma \in \mathscr{D}$ and $P \in \mathscr{P}$: if $(\rho, P) \gtrsim (\tau, P)$, then $(\alpha\rho + (1-\alpha)\gamma, P) \gtrsim (\alpha\tau + (1-\alpha)\gamma, P)$ for all $\alpha \in (0, 1]$.

With the aid of Lemma 1 and statements 2 and 3 of Lemma 3, it is also easy to obtain Monotone Continuity of $\succeq^u$ from preferences over pairings.

Axiom 10 (Monotone Continuity*): If $\rho, \tau, \kappa \in \mathscr{D}$, with $(\kappa, P^*) \in \mathscr{K}(\hat{\gtrsim})$, and $\{T_n\}_{n\geq 1} \in \Sigma$ with $T_1 \supseteq T_2 \supseteq \cdots$ and $\bigcap_{n\geq 1} T_n = \emptyset$, then $(\rho, P^*) \succ (\tau, P^*)$ implies that there exists $m \geq 1$ such that $(\kappa_{T_m}\rho, P^*) \succ (\tau, P^*)$.

We are ready to characterize SEU in the SDT framework. In what follows, we refer to axioms 1 to 6 as the MRA axioms.

Proposition 2 (SEU representation). *A preference $\gtrsim$ on $\mathscr{S}$ satisfies the MRA axioms, Mixture Independence and Monotone Continuity* if, and only if, there exists a bounded utility function $u : A \times \Theta \to \mathbb{R}$, continuous in the first argument, and a probability distribution $\pi \in \Delta(\Theta)$ such that, for all $\rho, \tau \in \mathscr{D}$ and $P, Q \in \mathscr{P}$,*

$$(\rho, P) \gtrsim (\tau, Q) \iff \int_\Theta \int_X u(\rho(x), \theta) \mathrm{d}P_\theta(x) \mathrm{d}\pi(\theta) \geq \int_\Theta \int_X u(\tau(x), \theta) \mathrm{d}Q_\theta(x) \mathrm{d}\pi(\theta).$$

Moreover, u is parameter-wise cardinally unique and, for a given u, π is unique.

Preferences that satisfy SEU are also called *Bayesian*, since for any $P \in \mathscr{P}$ and $\rho \in \mathscr{D}$, we have

$$\int_\Theta \int_X u(\rho(x), \theta) \mathrm{d}P_\theta(x) \mathrm{d}\pi(\theta) = \int_X \int_\Theta u(\rho(x), \theta) \mathrm{d}\pi_x(\theta) \mathrm{d}p(x),$$

where $p = \int_\Theta P_\theta \mathrm{d}\pi(\theta)$ and $\pi_x \in \Delta(\Theta)$ is the posterior distribution, conditional on $x \in X$: $\pi_x(T) = \mathbb{E}_\pi[\mathbf{1}_T \,|X_P = x]$ for all $T \in \Sigma$. When $p(x) > 0$, the posterior can be obtained via Bayes' rule.

1.5.2 Maximin expected utility

Maximin expected utility (MEU) — along with its close cousin, multiple priors expected utility (MPEU) — is perhaps the most thoroughly studied model of decision making under uncertainty, after SEU. Gilboa and Schmeidler (1989a) provided a characterization of MPEU for the Anscombe-Aumann setting that relies on Weak Order, A-A Monotonicity, Mixture Continuity, and two extra axioms:

- **Uncertainty Aversion:** for all $r, r' \in \mathscr{R}_u$ and $\alpha \in (0,1)$, $r \simeq^u r'$ implies $\alpha r + (1-\alpha) r' \succeq^u r$.
- **Certainty Independence:** for all $r, r', q \in \mathscr{R}_u$, with constant q: if $r \succeq^u r'$, then $\alpha r + (1 - \alpha) q \succeq^u \alpha r' + (1-\alpha) q$ for all $\alpha \in (0,1)$.

Uncertainty Aversion and Certainty Independence on the space of risk functions are induced by the following axioms on our primitive $\succsim$.

Axiom 11 (Hedging): For all $\rho, \tau \in \mathscr{D}$, $P \in \mathscr{P}$ and $\alpha \in (0,1)$: if $(\rho, P) \sim (\tau, P)$, then $(\alpha\rho + (1-\alpha)\tau, P) \succsim (\rho, P)$.

Axiom 12 (CRE-Independence): For all $\rho, \tau, \kappa \in \mathscr{D}$ and $P \in \mathscr{P}$ such that $(\kappa, P) \in \mathscr{K}(\hat{\succsim})$: if $(\rho, P) \succsim (\kappa, P)$, then $(\alpha\rho + (1-\alpha)\kappa, P) \succsim (\alpha\tau + (1-\alpha)\kappa, P)$ for all $\alpha \in (0,1]$.

Axiom 11 implies a preference for hedging: for any fixed experiment, the DM weakly prefers decision rules that provide a more balanced pay-off profile across parameters. Axiom 12 states that for any two pairings sharing the same experiment, mixing the decision rules with a third constant-risk-equivalent pairing does not change preferences.

I can now state the SDT version of the MPEU representation:

Proposition 3 (MPEU representation). *A preference $\succsim$ on $\mathscr{S}$ satisfies the MRA axioms, Hedging, and CRE-Independence if, and only if, there exists a bounded utility function $u : A \times \Theta \to \mathbb{R}$, continuous in the first argument, and a convex family of probability distributions $\Pi \subseteq \Delta(\Theta)$ such* *that, for all $\rho, \tau \in \mathscr{D}$ and $P, Q \in \mathscr{P}$,*

$$(\rho, P) \gtrsim (\tau, Q) \iff \inf_{\pi\in\Pi} \int_\Theta \int_X u(\rho(x), \theta) \mathrm{d}P_\theta(x) \mathrm{d}\pi(\theta) \geq \inf_{\pi\in\Pi} \int_\Theta \int_X u(\tau(x), \theta) \mathrm{d}Q_\theta(x) \mathrm{d}\pi(\theta).$$

Moreover, u is parameter-wise cardinally unique and, for a given u, Π is unique.

An MPEU agent can be interpreted as a pessimist, or as extremely cautious. He acts as if whichever pairing is chosen, the true parameter was drawn from the worst possible distribution among all the prior distributions he considers plausible.

The MEU representation can be seen as the special case of MPEU when $\Pi = \Delta(\Theta)$. Therefore, an MEU agent is maximally pessimistic, since he always consider the worst case for any distribution of the parameters. This model was axiomatized in the space of risk functions by Stoye (2012). The axiom required for the representation is translated to the SDT setting below.

Axiom 13 (Symmetry): Let $T, F \in \Sigma$ be such that $T \cap F = \emptyset$, and consider $\rho, \tau, \kappa, \gamma \in \mathscr{D}$ with $(\kappa, P^*), (\gamma, P^*) \in \mathscr{K}(\hat{\gtrsim})$. If $(\rho_T \kappa_F \gamma, P^*) \gtrsim (\tau_T \kappa_F \gamma, P^*)$, then $(\rho_T \gamma_F \kappa, P^*) \gtrsim (\tau_T \gamma_F \kappa, P^*)$.

Axiom 13 requires that the DM considers all events $T, F \in \Sigma$ as equally plausible: from an ex-ante perspective, reallocating constant pay-offs of different magnitudes across events does not change preferences. I can now state the following:

Proposition 4 (MEU representation). *A preference $\gtrsim$ on $\mathscr{S}$ has a MPEU representation and satisfies Symmetry if, and only if, there exists a bounded utility function $u : A \times \Theta \to \mathbb{R}$, continuous in the first argument, such that for all $\rho, \tau \in \mathscr{D}$ and $P, Q \in \mathscr{P}$,*

$$(\rho, P) \gtrsim (\tau, Q) \iff \inf_{\theta\in\Theta} \int_X u(\rho, \theta) \mathrm{d}P_\theta \geq \inf_{\theta\in\Theta} \int_X u(\tau, \theta) \mathrm{d}Q_\theta.$$

Moreover, u is parameter-wise cardinally unique.

1.5.3 Variational and multiplier preferences

Variational preferences models, put forth by Maccheroni, Marinacci, and Rustichini (2006), have also received considerable attention in the decision theoretic literature. In that setting, its characterization hinges on the same axioms as the MPEU model, except with a weakened version of C-Independence, which I translate into the SDT framework as follows:

Axiom 14 (Weak CRE-Independence): For all $\rho, \tau, \kappa, \kappa' \in \mathscr{D}$ and $P \in \mathscr{P}$ such that $(\kappa, P), (\kappa', P) \in \mathscr{K}(\hat{\succsim})$: if $(\alpha\rho + (1-\alpha)\kappa, P) \succsim (\alpha\tau + (1-\alpha)\kappa, P)$, then $(\alpha\rho + (1-\alpha)\kappa', P) \succsim (\alpha\tau + (1-\alpha)\kappa', P)$ for all $\alpha \in (0, 1]$.

Along with some previously defined axioms, Weak CRE-Independence suffices to characterize variational preferences in SDT. Before stating the result, we need the definition of a mathematical property: a function $c : \Delta(\Theta) \to [0, \infty)$ is said to be *grounded* if $\inf_{\pi \in \Delta(\Theta)} c(\pi) = 0$.

Proposition 5 (Variational representation). *A preference $\succsim$ on $\mathscr{S}$ satisfies the MRA axioms, Monotone Continuity*, Hedging and Weak CRE-Independence if, and only if, there exists a utility function $u : A \times \Theta \to \mathbb{R}$, continuous in the first argument, a convex set of probabilities $\Pi \subseteq \Delta(\Theta)$, and a grounded, convex and lower semicontinuous function $c : \Pi \to [0, \infty)$ such that, for all $\rho, \tau \in \mathscr{D}$ and $P, Q \in \mathscr{P}$,*

$$(\rho, P) \succsim (\tau, Q) \iff \inf_{\pi \in \Pi} \left\{ \mathbb{E}_\pi[r_u(\rho, P)] + c(\pi) \right\} \geq \inf_{\pi \in \Pi} \left\{ \mathbb{E}_\pi[r_u(\tau, Q)] + c(\pi) \right\}.$$

A DM whose preferences have a variational representation entertains a set of possible prior beliefs, but has different degrees of confidence in each of them, which is represented by the function c. Variational preferences generalize the MPEU model, which in turn generalizes SEU. This can be readily seen from the representation, by setting $c(\pi) = 0$ for all $\pi \in \Pi$ to get MPEU, and setting $\Pi = \{\pi\}$ to obtain SEU. It can also be seen from the axioms, since Weak CRE-Independence is weaker than CRE-Independence, which in turn is weaker than Mixture Independence.

A special case of this model, called *multiplier preferences* by Hansen and Sargent (2001), was characterized by Strzalecki (2011). This is done by imposing Savage's sure thing principle on a preference that has a variational representation.

Axiom 15 (Sure Thing Principle): For all $\rho, \tau, \gamma, \gamma' \in \mathscr{D}$ and $T \in \Sigma$: $(\rho_T\gamma, P^*) \succsim (\tau_T\gamma, P^*)$ implies $(\rho_T\gamma', P^*) \succsim (\tau_T\gamma', P^*)$.

Axiom 15 is clearly a stronger version of Independence of Irrelevant Parameters, and has a similar interpretation. It states that once the DM conditions the decision rule on a particular event $T \in \Sigma$, the action distributions on the remaining parameters $\theta \in T^c$ are inconsequential.

For any $\pi, \mu \in \Delta(\Theta)$, let $\pi \ll \mu$ denote that π is absolutely continuous of with respect to μ, i.e., $\mu(T) = 0$ implies $\pi(T) = 0$ for all $T \in \Sigma$. One can now obtain a representation by multiplier preferences:

Proposition 6 (Multiplier representation). *A preference $\succsim$ on $\mathscr{S}$ has a variational representation and satisfies the Sure Thing Principle if, and only if, there exists a utility function $u : A \times \Theta \to \mathbb{R}$, continuous in the first argument, a probability distribution $\mu \in \Delta(\Theta)$, and $k \in (0, \infty]$ such that, for all $\rho, \tau \in \mathscr{D}$ and $P, Q \in \mathscr{P}$,*

$$(\rho, P) \succsim (\tau, Q) \iff \min_{\pi \in \Delta(\Theta)} \{\mathbb{E}_\pi[r_u(\rho, P)] + kD(\pi\|\mu)\} \geq \min_{\pi \in \Delta(\Theta)} \{\mathbb{E}_\pi[r_u(\tau, Q)] + kD(\pi\|\mu)\},$$

where

$$D(\pi\|\mu) = \begin{cases} \int_\Theta \log \frac{\mathrm{d}\pi}{\mathrm{d}\mu} \mathrm{d}\pi, & \textit{if } \pi \ll \mu \\ \infty, & \textit{otherwise} \end{cases}$$

is the Kullback-Leibler divergence and $\frac{\mathrm{d}\pi}{\mathrm{d}\mu}$ is the Radon-Nikodym derivative of π with respect to μ.

Multiplier preferences describe a Bayesian DM who is worried about misspecification of the prior distribution. Though the DM's best guess for the parameter distribution is μ, she also entertains other specifications. Alternative candidates for the prior become less plausible as they diverge from the benchmark distribution μ. The degree of confidence in μ being correctly specified is regulated by the constant k, with larger values of k indicating more confidence in the benchmark distribution.

1.6 Alternative data sets

The previous sections focused on preferences over pairings. This is the most appropriate choice environment if the goal is to better understand the behavior of agents in different SDT models. It is also often observable as empirical data. For example, an econometrician who writes down a statistical model and chooses a particular estimator is ultimately revealing their choice of decision rule and experiment.

However, there are situations in which other types of choice data are forthcoming. In this section, I explore the behavioral implications of the MRA model to three alternative data sets. In Section 1.6.1, I consider state dependent stochastic choice data as assumed by Caplin and Martin (2015) and Caplin and Dean (2015). This is the standard data set in much of the psychometric literature, and keeps track of the probabilities of choosing each action conditional on the parameters.

In Section 1.6.2, I characterize the MRA model using choice data in the form of two different collections of preferences, each of which can be viewed as a cross-section of $\succsim$. The first describes, for each experiment $P \in \mathscr{P}$, a preference $\succcurlyeq_P$ on the set of decision rules. This corresponds to a DM who takes the experiment as given an chooses only among available decision rules. The second collection of preferences ascribes to each compact menu of decision rules $M \subseteq \mathscr{D}$, a preference $\succcurlyeq^M$ on the set of all experiments. This models a DM who chooses an experiment, assuming that in a second stage of the decision making process it will be paired with a particular decision rule.

1.6.1 Parameter dependent stochastic choice

Recall that the grand set of all parameter contingent action distributions is given by $\mathscr{F} = \Delta(A)^{\Theta}$. Each $f \in \mathscr{F}$ can be understood as a parameter dependent stochastic choice, i.e., the probability (or frequency) of choosing each action $a \in A$ for every parameter $\theta \in \Theta$. Consider a data set where one observes the DM choosing parameter dependent stochastic choices for a finite collection of *decision problems* $\Phi = \{F_1, \dots, F_n : F_i \subseteq \mathscr{F}\}$. Each $F \in \Phi$ represents the set of parameter contingent action distributions available to the DM in a particular choice situation. I call $(\mathscr{F}, \Phi)$ — where Φ consists of a finite collection of closed subsets of $\mathscr{F}$ — a choice space.

Formally, parameter dependent stochastic choice data consists of a choice function $c : \Phi \to \mathscr{F}$ such that $c(F) \in F$ for every $F \in \Phi$. Although the formalism is given in terms of a single observed choice for each feasible set, actual data is often better understood as repeatedly observing, for each decision problem $F \in \Phi$, the DM's choice of action for every parameter, and calculating action frequencies. To avoid questions of how to elicit multiple selections from the same decision problem in a setting where we only ever observe randomized choices, I assume that c is a function rather than a correspondence.

Given a parameter dependent stochastic choice function c, the goal is to determine whether it can be rationalized by a preference relation $\succsim$ on $\mathscr{S}$, in the following sense.

Definition 5. Given a choice space $(\mathscr{F}, \Phi)$, we say that a preference relation $\succsim$ on $\mathscr{S}$ rationalizes $c : \Phi \to \mathscr{F} \setminus \{\emptyset\}$, if the preference $\geq$ on $\mathscr{F}$ defined, for all $f, g \in \mathscr{F}$, by

$$f \geq g \iff (\rho, P) \succsim (\tau, Q) \; \forall (\rho, P), (\tau, Q) \in \mathscr{S} \text{ such that } f = \rho P, \; g = \tau Q, \qquad (1.8)$$

is such that for all $F \in \Phi$,

$$c(F) = \{f \in F : f \geq g \; \forall g \in F\}. \qquad \diamond$$

Definition 5 has the interpretation that the DM is actually choosing between pairings, but we (the modellers) only observe the induced parameter contingent action distributions. If c can be rationalized by $\succsim$, then $c(F)$ is the parameter contingent action distribution induced by the most preferred pairing (according to $\succsim$), among all pairings that can induce the distributions in F. Note that if the DM does not satisfy Consequentialism, then there is no hope of deducing properties of the preferences over pairings by observing only parameter dependent stochastic choices. Indeed, a failure of Consequentialism would imply that the preference $\geq$ on $\mathscr{F}$ given by (1.8) is not well defined.

An alternative is directly revealed preferred to another if it is chosen from a decision problem where the other is available. If there is a sequence of alternatives such that f_1 is directly revealed preferred to f_2 and so on, until f_{k-1} is directly revealed preferred to f_k, then I simply say that f_1 is *revealed preferred* to f_k. This is formalized in the definition below.

Definition 6. For any $f_1, f_k \in \mathscr{F}$, we say that f_1 is revealed preferred to f_k, and denote it by $f_1 \trianglerighteq f_k$, if there exist $\{F_1, \ldots, F_{k-1}\} \subseteq \Phi$ such that $\{f_i, f_{i+1}\} \subseteq F_i$ and $f_i = c(F_i)$ for all $i = 1, \ldots, k-1$. $\diamondsuit$

Consider the following axioms on the parameter dependent stochastic choice data set c.

Axiom C1 (GARP): For all $f, g \in \mathscr{F}$: if $f \trianglerighteq g$, then $g \neq c(F)$ for all $F \in \Phi$ such that $f \in F$.

The Generalized Axiom of Revealed Preference (GARP) is a basic rationality postulate, stating that the revealed preference relation is acyclic. That is, the DM's choices are required to be coherent, in the sense that if f is revealed to be strictly preferred to g, then g can not be chosen when f is available. This is the revealed preference analogue of Weak Order.

For each $\theta \in \Theta$, define a binary relation $\triangleright_\theta$ on $\Delta(A)$ by $f(\theta) \triangleright_\theta g(\theta)$ if, and only if, $f \trianglerighteq g$ and $f(\theta') = g(\theta')$ for all $\theta' \neq \theta$. This can be interpreted it as a directly revealed preference conditional on θ, since $f(\theta) \triangleright_\theta g(\theta)$ when f is chosen over g in a situation where the only relevant parameter is θ.

Let $\trianglerighteq_\theta$ be the convex hull of the transitive closure of $\triangleright_\theta$. That is, $p \trianglerighteq_\theta q$ if, and only if, there exist $\{(r_j, s_j)\}_{j=1}^k \subseteq \Delta(A)^2$ and $\{\lambda_j\}_{j=1}^k \subseteq \mathbb{R}_+$ with $\sum_{j=1}^k \lambda_j = 1$, such that $r_j \triangleright_\theta s_j$ for all $j = 1, \ldots, k$, $p = \sum_{j=1}^k \lambda_j r_j$ and $q = \sum_{j=1}^k \lambda_j s_j$. In more intuitive terms, $p \trianglerighteq_\theta q$ if p and q are convex combinations, with the same weights, of $\{r_j\}_{j=1}^k$ and $\{s_j\}_{j=1}^k$ respectively, and moreover each r_j is directly revealed preferred to s_j conditional on θ. Note that $\trianglerighteq_\theta$ can be elicited from choice data, by first taking the transitive closure and then the convex hull of $\triangleright_\theta$. This relation was first presented by Clark (1993), who also formulated the following axiom.

Axiom C2 (C-LARP): For all $f, g \in \mathscr{F}$, and $\theta \in \Theta$: $f(\theta) \trianglerighteq_\theta g(\theta)$ implies $g(\theta) \not\triangleright_\theta f(\theta)$.

The conditional linear axiom of revealed preference (C-LARP) says that in situations where there is a single relevant parameter, mixing action distributions on that parameter using the same weights does not reverse their conditional preference ranking. Thus, it is a parameter dependent stochastic choice analogue of Conditional Mixture Independence. Moreover, it implies that $\rhd_\theta$ is antisymmetric. That is, if $f(\theta) \neq g(\theta)$ and $f_{(\theta)}h \trianglerighteq g_{(\theta)}h$, then $g_{(\theta)}h' \trianglerighteq f_{(\theta)}h'$ does not hold for any $h' \in \mathscr{F}$. Therefore, C-LARP also encodes a version of Independence of Irrelevant Parameters.

Axiom C3 (RP-Monotonicity): If $f = c(F)$ and $g \in F \setminus \{f\}$, then there exists $\theta \in \Theta$ such that $g(\theta) \trianglerighteq_\theta f(\theta)$ does not hold.

RP-Monotonicity states that if f is directly revealed preferred to g, then there exists a parameter for which g is not conditionally revealed preferred to f. It can be considered a revealed preference version of Monotonicity.

Finally, we impose a continuity axiom on the conditional revealed preference relations.

Axiom C4 (Conditional Continuity): For all $\theta \in \Theta$, if $(p_n), (q_n) \in \Delta(A)$ are sequences such that $p_n \to p$, $q_n \to q$ and $p_n \trianglerighteq_\theta q_n$ for all $n \in \mathbb{N}$, then $q \not\rhd_\theta p$.

We can now state the main representation theorem of this section.

Theorem 3. *The parameter dependent stochastic choice function $c : \Phi \to \mathscr{F}$ satisfies GARP, C-LARP, RP-Monotonicity and Conditional Continuity if, and only if, there exists a preference $\succsim$ on $\mathscr{S}$ which rationalizes c and satisfies Consequentialism, Weak Order, Independence of Irrelevant Parameters, Monotonicity, Conditional Mixture Independence and Continuity.*

Moreover, if Θ is finite, so that $\mathscr{F}$ is metrizable, then $\succsim$ has an MRA representation (u, I) such that, for all $F \in \Phi$,

$$c(F) = \left\{ \rho P \in F : \operatorname*{arg\,max}_{\{(\tau,Q)\in\mathscr{S}:\tau Q\in F\}} I(r_u(\tau, Q)) \right\}.$$

Theorem 3 characterizes when parameter dependent stochastic choice data can be rationalized by postulating that the DM is actually facing a statistical decision problem, and making choices that maximize preferences compatible with the MRA framework. This kind of behavioral data differs from preferences over pairings in two main ways. First, it does not assume that we observe experiments and decision rules, only their final consequences — the induced parameter contingent action distributions. Second, it only assumes that choices from a finite number of decision problems — rather than the whole array of preferences — is observed. Therefore, Theorem 3 allows us to test whether the DM's choices are compatible with the MRA model using a data set that is sometimes easier to obtain.

It is natural to ask whether this type of data set can also be used to identify preferences over pairings, at least partially. That is, can the DM's choices over pairings be recovered by only looking at parameter dependent stochastic choice data? The following example shows that, at least under some conditions, this is indeed possible.

Example 6: Let A and Θ be finite, and $|\operatorname{supp} P_\theta| \leq \min\{|\Theta|, |A|\}$ for all $\theta \in \Theta$. The DM chooses pairings from an array of decision problems $\Psi = \{S_1, \ldots, S_n : S_j \subseteq \mathscr{S}\}$ so as to maximize a preference $\succsim$ on $\mathscr{S}$. Assume that the set of feasible decision rules remains fixed at some $D \subseteq \mathscr{D}$ for every $S \in \Psi$. Therefore, each decision problem only differs by the set of experiments available to the DM.

We do not directly observe the DM's choice of decision rule or experiment, only the realized action frequencies for each parameter $\theta \in \Theta$ and decision problem $S \in \Psi$. Let $\mathbb{P}(a|\theta, S)$ denote the probability of choosing action $a \in A$ from decision problem $S \in \Psi$, when the parameter is $\theta \in \Theta$. Letting $F(S) = \{\rho P : (\rho, P) \in S\}$ denote the set of parameter contingent action distributions that can be induced by members of S, we can express the parameter dependent stochastic choice data set as

$$c(F(S))(\theta) = \begin{pmatrix} \mathbb{P}(a_1|\theta, S) \\ \vdots \\ \mathbb{P}(a_{|A|}|\theta, S) \end{pmatrix} \quad \text{for all } \theta \in \Theta \text{ and } S \in \Psi.$$

Given the DM's postulated choice procedure, the realized action frequencies must satisfy

$$\mathbb{P}(a|\theta, S_j) = \sum_{x \in \operatorname{supp} P_\theta} \rho^j(x, a) P^j_\theta(x) \quad \text{for all } \theta \in \Theta \text{ and } j \in \{1, \ldots, n\}, \tag{1.9}$$

subject to $(\rho, P^j) \succsim (\tau, Q)$ for all $(\tau, Q) \in S_j$ — where I write $\rho^j \equiv \rho_{P^j}$ to simplify notation. With parameter dependent stochastic choice data, we only observe the left hand side of eq. (1.9): neither the chosen (ρ, P^j), nor the realized signals are deemed observable.

Note that for each $S_j \in \Psi$, varying the parameter changes the signal distribution, P^j_θ, but leaves the decision rule unchanged. Thus, eq. (1.9) describes a finite mixture model where varying the parameter shifts the mixture weights $P^j_\theta(x)$, but leaves the mixture distributions $\rho(x, \cdot)$ unchanged. If the model in (1.9) is point identified in the statistical sense, it is possible to recover the chosen decision rule and experiment from each $S \in \Psi$, given enough data.

Henry, Kitamura, and Salanié (2014) thoroughly studied conditions on P^j and ρ^j under which such models are partially identified. Adams (2016) provides necessary and sufficient conditions, again on P^j and ρ^j, for point identification. Such results are of limited applicability, since P^j and ρ^j are not directly observable. In Chapter 2, I present a sufficient condition on the observable parameter dependent stochastic choice data $\mathbb{P}$ that guarantees point identification of such mixture models. $\|$

1.6.2 Cross sectional preferences

In sections 1.4 and 1.5, I focused on preferences $\succsim$ over pairs of decision rules *jointly with* experiments. However, some applications are not directly concerned with preferences over pairings, but rather with one of two sets of preferences embedded in $\succsim$. The first and most ubiquitous kind of application involves finding an optimal decision rule for a given experiment. That is, given $P \in \mathscr{P}$, one wishes to characterize a preference $\succcurlyeq_P$ defined by $\rho \succcurlyeq_P \tau$ if, and only if, $(\rho, P) \succsim (\tau, P)$.

The second type of application involves ranking experiments, assuming that each will be paired with a decision rule that will be chosen in a second stage. Formally, given a menu of feasible decision rules $M \subseteq \mathscr{D}$ and a collection of choices $\{C_P(M) : P \in \mathscr{P}\}$ with $C_P(M) \subseteq M$, I wish to characterize the relation $P \succcurlyeq^M Q$ defined by $(\rho^P, P) \succsim (\rho^Q, Q)$ for every $P, Q \in \mathscr{P}$, assuming $\rho^j \in C^j(M)$, $j = P, Q$. This is the setting of the design of experiments literature.

In what follows, I will separately characterize the MRA model for these two collections of preferences.

Preferences over decision rules

Suppose choice data comes as a collection of preferences on decision rules $\succcurlyeq_{\cdot} = \{\succcurlyeq_P \subseteq \mathscr{D}^2 : P \in \mathscr{P}\}$, one for each (fixed) experiment. In the Bayesian persuasion literature, an agent described by such choice data is called a *Receiver*, since she draws inference from information received from a source beyond her control. I will still call this agent a DM, since she has preferences over *decision* rules.

Before presenting a representation theorem with this alternative data set, I define *conditional preferences*. Given P^*, the preference $\succcurlyeq_\theta$ conditional on θ is defined by $\rho \succcurlyeq_\theta \tau$ if, and only if, $\rho_{\{\theta\}}\gamma \succcurlyeq_{P^*} \tau_{\{\theta\}}\gamma$ for all $\gamma \in \mathscr{D}$. The following axiom imposes straightforward adaptations of axioms 1 to 6 for $\succcurlyeq_{\cdot}$, with similar interpretations:

Axiom D1: For every experiment $P \in \mathscr{P}$ and all decision rules $\rho, \tau \in \mathscr{D}$:

1 If $\rho \succcurlyeq_P \tau$, $\rho' P' = \rho P$ and $\tau' P' = \tau P$, then $\rho' \succcurlyeq_{P'} \tau'$.

2 $\succcurlyeq_P$ is complete and transitive.

3 $\rho_{\{\theta\}}\gamma \succcurlyeq_{P^*} \tau_{\{\theta\}}\gamma$ implies $\rho_{\{\theta\}}\gamma' \succcurlyeq_{P^*} \tau_{\{\theta\}}\gamma'$, for all $\gamma, \gamma' \in \mathscr{D}$ and $\theta \in \Theta$.

4 If $\rho \succcurlyeq_\theta \tau$ for all $\theta \in \Theta$, then $\rho \succcurlyeq_{P^*} \tau$.

5 If $\rho \succcurlyeq_\theta \tau$, then $\alpha\rho + (1-\alpha)\gamma \succcurlyeq_\theta \alpha\tau + (1-\alpha)\gamma$ for all $\alpha \in (0,1]$ and $\gamma \in \mathscr{D}$.

6 The set $\{(\rho P, \tau P) \in \mathscr{F}^2 : \rho \succcurlyeq_P \tau\}$ is closed.

I can now state the following representation theorem.

Theorem 4. *The system of preferences $\succcurlyeq_{\cdot} = \{\succcurlyeq_P \subseteq \mathscr{D}^2 : P \in \mathscr{P}\}$ satisfies Axiom D1 if, and only if, there exists a utility function $v : A \times \Theta \to \mathbb{R}$, continuous in the first argument, and a monotone, continuous functional $J : \mathscr{R}_v \to \mathbb{R}$ such that, for all $P \in \mathscr{P}$ and $\rho, \tau \in \mathscr{D}$,*

$$\rho \succcurlyeq_P \tau \iff J(r_v(\rho, P)) \geq J(r_v(\tau, P)). \tag{1.10}$$

We then say that $\succcurlyeq_{\cdot}$ has an MRA representation (v, J). Moreover, v is parameter-wise cardinally unique.

Theorem 4 characterizes the MRA model when behavioral data comes in the form of preferences over decision rules, for experiments that are exogenous from the DM's standpoint. This is useful because it provides a method to verify whether choices are compatible with the SDT framework, without having to vary decision rules and experiments simultaneously. For instance, in laboratory settings such data can be elicited by exogenously varying the experiment and available actions, while keeping track of action frequencies for each realized signal.

Preferences over experiments

Now turn to an agent who chooses an information structure for each menu M of available decision rules knowing that, in a second stage, a decision rule will be chosen for her (and, potentially, by her) from M. To differentiate this agent from the DM, who chooses decision rules, I call her the *Experimenter*.[4]

Formally, identify $\mathscr{D}$ with the space of measurable functions from X to $\Delta(A)$ and, as usual, endow it with the topology of point-wise convergence. A menu $M \in \mathscr{M}$ is some closed (thus compact) subset of $\mathscr{D}$. I now assume that the primitive is a system of preferences $\succcurlyeq^{\cdot} = \{\succcurlyeq^{M} \subseteq \mathscr{P}^2 : M \in \mathscr{M}\}$. I want to characterize which such preference systems are compatible with the MRA model.

Recall that a decision rule $\rho \in \mathscr{D}$ is called invariant if $\rho_P = \rho_{P'}$ for all $P, P' \in \mathscr{P}$. Denote by $\overline{\mathscr{M}}$ the class of all menus consisting only of invariant decision rules. Throughout this section, I assume that $\mathscr{P} = \Delta(X)^{\Theta}$. When this is the case, section A.1 shows that there exists an invariant decision rule $\delta^* \in \overline{\mathscr{D}}$ such that for all $(\rho, P) \in \mathscr{S}$, we have $\delta^* P' = \rho P$ for some $P' \in \mathscr{P}$.

First I state some postulates involving only preferences over experiments. These are versions of axioms 1 and 3 to 5 in such a space.

Axiom E1: For all menus $M \in \mathscr{M}$, $\overline{M} \in \overline{\mathscr{M}}$ and experiments $P, Q \in \mathscr{P}$:

[4] In Bayesian persuasion literature, such an agent is called a Sender.

1 **Weak Order:** $\succcurlyeq^{M}$ is complete and transitive.

2 **IIP:** $P_{(\theta)}R \succcurlyeq^{\overline{M}} Q_{(\theta)}R$ implies $P_{(\theta)}R' \succcurlyeq^{\overline{M}} Q_{(\theta)}R'$, for all $R, R' \in \mathscr{P}$ and $\theta \in \Theta$.

3 **Monotonicity:** if $P_{(\theta)}R \succcurlyeq^{\overline{M}} Q_{(\theta)}R$ for all $\theta \in \Theta$ and $R \in \mathscr{P}$, then $P \succcurlyeq^{\overline{M}} Q$.

4 **Experiment CMI:** if $P_{\theta'} = Q_{\theta'} = R_{\theta'}$ for every $\theta' \neq \theta \in \Theta$ and $P \succcurlyeq^{\overline{M}} Q$, then $\alpha P + (1 - \alpha)R \succcurlyeq^{\overline{M}} \alpha Q + (1 - \alpha)R$ for all $\alpha \in (0, 1]$.

5 **Continuity:** the set $\{(\delta P, \delta Q) \in \mathscr{F}^2 : P \succcurlyeq^{\{\delta\}} Q\}$ is closed for all $\delta \in \mathscr{D}$.

Now consider a family of *choice correspondences* $C = \{C_P : P \in \mathscr{P}\}$, where

$$\begin{aligned} C_P : \mathscr{M} &\to 2^{\mathscr{D}} \\ M &\mapsto C_P(M) \subseteq M. \end{aligned} \tag{1.11}$$

The inclusion $\rho \in C_P(M)$ means that, in a second stage, ρ is chosen from the menu M of feasible decision rules, when the experiment is P. The following axiom requires that information structures be chosen taking into account the decision rule selected by C, and that the Experimenter is a consequentialist, caring only about induced action distributions.

Axiom E2 (Consistency): If $P \succcurlyeq^{M} P'$, and for all $\rho \in C_P(M)$, $\rho' \in C_{P'}(M)$, $\tau \in C_Q(N)$ and $\tau' \in C_{Q'}(N)$ we have $\rho P = \tau Q$ and $\rho' P' = \tau' Q'$, then $Q \succcurlyeq^{N} Q'$.

Axioms E1 and E2 are sufficient to characterize the Experimenter's preferences given menus from which a single decision rule is chosen by C for every experiment (for instance, singleton menus). For all other menus, we need to elicit the Experimenter's tie-breaking rule from her choices. Since in this paper I am mostly concerned with a single sophisticated DM, it is natural to assume that indifference is decided in favor of the Experimenter. The next axiom guarantees just that, but first I introduce another bit of notation: for any $M \in \mathscr{M}$ and $\tau \in C_P(M)$, let $P^{\tau} \in \mathscr{P}$ satisfy $\delta^* P^{\tau} = \tau P$.

Axiom E3 (Optimism): If $\rho \in C_P(M)$ and $P^\rho \succcurlyeq^{\{\delta^*\}} P^\gamma$ for all $\gamma \in C_P(M)$, then $P \succcurlyeq^M Q$ ($Q \succcurlyeq^M P$) if, and only if, $P \succcurlyeq^{\tilde{M}} Q$ ($Q \succcurlyeq^{\tilde{M}} P$), where $\tilde{M} = \{\delta \in \mathscr{D} : \delta_P = \rho_P$ and $\forall R \neq P$, $\exists \gamma \in M$ s.t. $\delta_R = \gamma_R\}$.

Axiom E3 is conceptually straightforward. If decision rule τ induces an action distribution under P which is deemed worse by the Experimenter than that induced by ρ, then effectively removing from the menu the possibility of pairing τ with P does not change preferences. In other words, the Experimenter is an optimist, who believes the decision rule chosen in the second stage will be the best possible one, among the alternatives that are deemed "choosable" by C.

I can now characterize the Experimenter's preferences.

Theorem 5. *A system of preferences $\succcurlyeq^{\cdot} = \{\succcurlyeq^M \subseteq \mathscr{P}^2 : M \in \mathscr{M}\}$ satisfies axioms E1 to E3 if, and only if, there exists a utility function $w : A \times \Theta \to \mathbb{R}$, continuous in the first argument, and a monotone, continuous functional $H : \mathscr{R}_w \to \mathbb{R}$ such that, for all $P, Q \in \mathscr{P}$ and $M \in \mathscr{M}$,*

$$P \succcurlyeq^M Q \iff H(r_w(\rho, P)) \geq H(r_w(\tau, Q)) \quad \textit{subject to } \rho \in C_P(M),\ \tau \in C_Q(M). \tag{1.12}$$

We then say that $\succcurlyeq^{\cdot}$ has an MRA representation (w, H). Moreover, w is parameter-wise cardinally unique.

Note that Theorem 5 makes no assumptions about the properties of the choice correspondence. In particular, the decision rule that is ultimately chosen, for any given experiment, could differ from the one the Experimenter would prefer. The Experimenter can thus be seen as a sophisticated, but possibly dynamically inconsistent, agent. She correctly anticipates – but does not necessarily control – which decision rule will be ultimately chosen, and designs the information structure accordingly. Of course, if C takes the form

$$C_P(M) = \arg\max_{\delta \in M} H(r_w(\delta, P)) \quad \forall M \in \mathscr{M} \text{ and } P \in \mathscr{P}, \tag{1.13}$$

then the Experimenter's problem is equivalent to her sequentially choosing an experiment and then a decision rule for herself. To spell out sufficient conditions when (1.13) is true could be an interesting avenue for future research, but is beyond the scope of the current paper.

1.7 Discussion

In this paper, I have built an axiomatic foundation that can be used to obtain representation theorems for many statistical decision theoretic models. These theorems provide the full behavioral implications of said models, in terms of preferences over pairings of decision rules and experiments. Such representation results serve two main purposes in positive economics. First, they provide applied theorists working within the SDT framework with a deeper understanding of the behavior of agents in their models. Second, they allow empiricists to gauge the extent to which observed choice behavior deviates from the behavioral predictions of different SDT models. As SDT is the setting of much of information economics, results of this kind may find many uses in the wider economics literature.

Representation theorems such as the ones presented in sections 1.4 and 1.5 also serve a normative purpose. For instance, Manski (2021) advocates for the use of statistical decision theory in econometrics, but leaves open the issue of what objective function the econometrician should use. In this context, axioms can serve as a guide for the econometrician's choice of utility function and aggregator, when formulating an SDT model. That is, the econometrician should choose a statistical decision theoretic objective function that leads to desirable constraints on preferences, formalized as normatively appealing axioms.

For this paper, I have mainly focused on obtaining the behavioral implications of a canonical SDT model in which the DM takes the expected utility conditional on each parameter, and aggregates across parameters using a monotone functional. Therefore, although the methodology I presented in sections 1.4 and 1.5 is flexible enough to provide characterizations of a wide array of SDT models, it is by no means exhaustive. There are at least three particularly interesting classes of models which can be considered as belonging to the SDT family, but are ruled out by the approach taken here.

The first pertains to models in which not only choices, but DM's preferences themselves, depend on the menu of available options. The canonical example of this is regret-based preferences, as discussed in example 1. Second, by maintaining the Consequentialism axiom throughout the analysis, I am precluding models where the utility function depends not only on actions and parameters, but directly on decision rules and experiments. As discussed in example 4, one such model is the classic rational inattention framework with cognitive costs. A third class of SDT models not contemplated here is one in which, when the true parameter is known, the DM evaluates prospects using a criterion other than maximizing expected utility. For example, Manski and Tetenov (2014) evaluate the performance of a treatment rule based on maximizing a utility quantile, rather than its expectation, conditional on each parameter. Expanding the framework outlined here to any one of these classes of models presents a fruitful avenue for future decision theoretic research.

Another limitation of this essay is that it is concerned solely with ex-ante preferences. That is, our DM simultaneously chooses an experiment and a contingent plan of action, assuming that she will follow through with this plan once the information from a signal is revealed. The model is silent on how the DM updates their beliefs – or, more generally, their preferences – upon observing the signal. Hence, the present model could be misleading if one is interested in predicting which action the DM takes after observing a signal. Indeed, if the DM is not dynamically consistent, then after observing some signal realizations, they would prefer to take a different action than the one prescribed by their chosen decision rule. Extending the analysis to include how preferences and beliefs (should) change after observing a signal is an entire research agenda in itself. Indeed, how to update preferences and beliefs in view of new information has been an active area of research in traditional decision theory, starting with a seminal paper by Machina (1989). However, because it lacks a natural distinction between information-carrying signals and payoff-relevant parameters, this literature is constrained to considering only information of a particular kind. Hence, it could be very effective to re-frame the preference updating problem in SDT terms, using the methodology in this essay.

Since there are situations in which data on preferences over pairings is not easy to come by, in

Section 1.6 I expanded the scope of applicability of my main results by characterizing the MRA model assuming other data sets. In Section 1.6.1, I considered a data set consisting of action probabilities conditional on each parameter. This type of data is standard in psychometrics, and has recently received attention in economics. I showed when such data can be rationalized by a preference over pairings that has an MRA representation. Moreover, example 6 discusses sufficient conditions for preferences over decision rules and experiments to be identifiable from parameter dependent stochastic choice. This opens the possibility of extending the results of this paper to the case when the DM faces information costs, and thus building on the work of Caplin and Dean (2015) by (i) not assuming the existence of a utility function in their representations; and (ii) extending their results beyond Bayesian DMs.

Section 1.6.2 provides an MRA characterization for two systems of preferences. The first describes a DM's preferences over decision rules for each given experiment. The second describes a DM who correctly anticipates which decision rule will be chosen from a menu and ranks experiments accordingly. These preferences naturally arise in many applications, and can be viewed as two cross sections of the preference over pairings. Another interesting topic for future research will be to provide sufficient conditions for the MRA representations of these two systems of preferences to coincide, so that they can be interpreted as modelling the same agent.

Chapter 2: Testable identification of finite mixture models

2.1 Introduction

In many economic models, observed outcomes are assumed to be sampled from a mixture of a finite number of latent distributions. Models with heterogeneity characterized by a finite number of unobserved types is commonly modeled using finite mixtures, in fields ranging from industrial organization (Berry, Carnall, and Spiller, 1996) to labor economics (Keane and Wolpin, 1997). In models with misclassification error, observations are drawn from a finite mixture conditional on unobserved true covariates, while the observed values of the covariates suffer from measurement error (Chen, Hong, and Nekipelov, 2011). Mixture models are also used to describe the outcomes arising from games with multiple equilibria, where the equilibrium selection mechanism defines the mixture weights (e.g., Ciliberto and Tamer (2009)). The large class of models with latent discrete state variables, such as Markov regime switching, is also a prime example of finite mixtures (Kim and Nelson, 2018).

Due to the large number of unknown parameters, identification and consistent estimation of mixture models is known to be difficult. A large literature, summarized in book form by Frühwirth-Schnatter (2013), has dealt with this issue by assuming that the component distributions are in a parametric family. However, in parametric models the consistency of estimators for the parameters of interest hinges on the correct specification of the unobserved component distributions. Mispecified models may thus lead to inconsistent estimators and misleading inference. For this reason, a more recent literature has turned attention to non-parametric identification and estimation of finite mixture models. Many strategies have been proposed to achieve such identification. For example, Mahajan (2006) and Lewbel (2007) relied on instrumental variables to handle the misclassification problem, and Chen, Hong, and Tamer (2005) used auxiliary data. See Compiani and Kitamura

(2016) for a comprehensive review.

An exclusion restriction that is commonly imposed on the data generating process to achieve identification is that there exists an observable variable which shifts the mixture weights, while leaving the component distributions unchanged. Such is the case for Mahajan (2006) and Lewbel (2007), cited above, but also for Chen, Hu, and Lewbel (2008b), Chen, Hu, and Lewbel (2008a), and Hu (2008), just to cite a few examples. This exclusion restriction is widely applicable, since it can be derived from the Markov assumption in regime switching models (Cho and White, 2007), as well as from models with unobserved type heterogeneity where an observed instrument changes the type composition of the population, but does not change the outcomes conditional on types. As I show Section 2.2, even models of rational inattention can be framed as finite mixture models with such an exclusion restriction.

This paper contributes to the literature on non-parametric identification of finite mixture models with instrumental variables affecting only the mixtures weights. Henry, Kitamura, and Salanié (2014) were the first to fully characterize the identified set in such models. They prove that the model is only partially identified in general, and show how the identified set can be constructed from the mixture distribution. However, their characterization of the identified set relies on convex analysis methods and may not be well-suited for inference with finite data. On the other hand, Adams (2016) provided necessary and sufficient conditions for point-identification of finite mixture models, considering the same exclusion restriction. Unfortunately, such conditions are imposed on the latent parameters of the model, thus can not be checked without first estimating those parameters. But if the model is not identified in the first place, any estimator will be inconsistent, and inferences drawn from it will be misleading, even given a large sample.

The main result of this paper provides conditions, on the *observable* mixture distribution, which are sufficient to guarantee non-parametric point-identification of the *hidden* mixture weights and component distributions. Interestingly, the theorem uses a similar condition to the one proposed by Adams (2016), but applies it to the mixture distribution rather than the latent parameters. The result builds on the extensive mathematics and computer science literature regarding the (exact) non-negative factorization of matrices, in particular on the works of Huang, Sidiropoulos, and Swami (2014) and Gillis (2012). Indeed, the statistical identification result presented here follows almost directly from a purely mathematical result about uniqueness of a particular kind of non-negative matrix factorization: an intermediate finding that can be of independent interest.

An identification condition on finite mixture models which is formulated solely in terms of the mixture distribution has two main advantages. First, the condition is always interpretable in terms of observable quantities. This is particularly important when the mixture model is used mainly as a tool for dimensionality reduction, which implies that the components have no interpretation a priori. In such applications, it can be hard to qualitatively assess ex ante whether a given condition on the hidden component distributions and mixture weights is likely to hold. This is the case in some popular applications of mixture models, notably topic models (Blei, 2012).

Second, and most importantly, a condition that depends solely on observable quantities can be formally tested using statistical methods. Therefore, I leverage my sufficient condition to devise a Bayesian test of non-parametric identification of the mixture weights and component distributions. This allows the statistician to formally test whether a model is likely identified before estimating its hidden parameters. Moreover, I show that the proposed procedure is in fact a formal Bayes test, in the sense that it is the optimal decision rule for a properly defined Bayesian statistical decision problem.

The remainder of the essay is organized as follows. Section 2.2 presents the model, including how it can be made to fit the non-negative matrix factorization problem. Section 2.3 discusses identification of mixture models, and provides the main result. Section 2.4 is dedicated to developing a formal test of identification of the hidden parameters, and also briefly discusses the consistent estimation of identified parameters.

2.2 Model

Let Y, W and X be random variables defined on a common probability space $(\Omega, \mathscr{F}, P)$ and taking values on separable metric spaces $\mathbb{Y}$, $\mathbb{W}$ and $\mathbb{X}$, respectively. The variable Y is the outcome,

W is a covariate and X is an instrument. By working with an abstract space of outcomes, I am allowing Y, W and X to be random vectors with both continuous- and discrete-valued components.

In a finite mixture model, conditional on the covariate and instrument, each observed outcome is drawn from one of K component distributions corresponding to different types, where $1 < K \leq \min\{|\mathbb{Y}|, |\mathbb{X}|\}$. The probability of drawing a sample from component distribution k is given by $\lambda_k \geq 0$. Formally, for all measurable $A \subseteq \mathbb{Y}$,

$$P(Y \in A|W, X) = \sum_{k=1}^{K} \lambda_k(W, X) Q_k(Y \in A|W, X), \tag{2.1}$$

where $\{Q_k(\cdot|W, X) : k = 1, \ldots, K\}$ are conditional probability distributions on the outcomes, and $\sum_{k=1}^{K} \lambda_k(w, x) = 1$ for all $(w, x) \in \mathbb{W} \times \mathbb{X}$. I assume in most of what follows that K is known. We denote $\lambda = (\lambda_1, \ldots, \lambda_k)$ and $Q = (Q_1, \ldots, Q_k)$. The mixture distribution P is viewed as a reduced form parameter for the data generating process, while the component distributions and the mixture weights, $\{(\lambda_k, Q_k) : k = 1, \ldots, K\}$, are the parameters of interest. Note that no finite-dimensional parametric family was specified for the component distributions $\{Q_k : k = 1, \ldots, K\}$, so the model is non-parametric.

We say that a model is identifiable if there exists a bijective function between the parameter space and the data generating process, so that perfectly estimating the sampling distribution allows the statistician to pinpoint the underlying parameter. Clearly, the reduced form parameter P is non-parametrically identifiable. However, we are interested in identifying not only the mixture distribution, but the components and mixture wights.

Definition 7 (Identification of a finite mixture model). We say that the model (2.1) is identified if $(Q, \lambda) \neq (\tilde{Q}, \tilde{\lambda})$ implies $\sum_{k=1}^{K} \lambda_k(W, X) Q_k(\cdot|W, X) \neq \sum_{k=1}^{K} \tilde{\lambda}_k(W, X) \tilde{Q}_k(\cdot|W, X)$. $\diamond$

Absent further constraints it is clear that there are different combinations of λ and Q yielding the same mixture distribution. Indeed, the data generating process P could be obtained by setting $\lambda_k = 1$ and $Q_k = P$ for any $k = 1, \ldots, K$. Therefore, the parameters (Q, λ) are not identifiable in general. Our main identifying restriction is that the instrument affects only the mixture weights.

Assumption 1 (Exclusion Restriction). For every measurable event $A \subseteq \mathbb{Y}$ and $k \in \{1, \ldots, K\}$, we have $Q_k(Y \in A|W, X) = Q_k(Y \in A|W)$.

This assumption occurs naturally in many applications. Prominent examples include Markov switching models with finite hidden states, as in (Cho and White, 2007); models with misclassified discrete regressors (see Chen, Hong, and Nekipelov (2011)); and accounting for multiple equilibria in economic models. For an in-depth discussion of Assumption 1, see Henry, Kitamura, and Salanié (2014). The following examples illustrate the scope of applicability of models with such an exclusion restriction.

Example 7 (State-dependent stochastic choice with attention costs): In rational inattention models, a decision maker (DM) must take an action whose pay-off depends on an unknown state of the world. Both the set of available actions, A, and the set of all possible states, Σ, are assumed to be finite. The decision maker chooses an information structure, which is a stochastic mapping from the objective states of the world to a set of subjective signals S. Denote this information structure by $Q : \Sigma \longrightarrow \Delta(S)$ and assume that $|S| \leq \min\{|A|, |\Sigma|\}$.

Before taking an action, the DM observes the signal from their chosen information structure. Each choice of information structure Q incurs a subjective cost $\kappa(Q)$, which is assumed to increase with signal informativeness. Endowed with a prior μ over states of the world, the DM chooses a decision rule $\delta : S \longrightarrow \Delta(A)$ and an experiment $Q : \Sigma \longrightarrow \Delta(S)$ which jointly maximize their expected pay-off net of information costs.

Caplin and Dean (2015) consider the problem of testing the validity of such a model using only state-dependent stochastic choice data. This is a data set where one observes the probability with which the DM takes each action, conditional on every state of the world. That is, the data consists of $P(a|\sigma)$ for all $a \in A$ and $\sigma \in \Sigma$. In terms of the DM's chosen decision rule and information structure, we have

$$P(a|\sigma) = \sum_{s \in S} Q(s|\sigma)\delta(a|s).$$

Therefore, the problem of identifying the DM's choices of decision rule and experiment from state-dependent stochastic choice data can be formulated as a mixture model, where the actions are the outcomes, the observed states are the instruments and the signal realizations are the hidden types.

Dean and Neligh (2023) gathered experimental data to test the validity of the rational inattention model using a revealed preference approach, which does not hinge on identifying the parameters π and δ separately. Provided the model is identifiable, the same same data could in principle be used to recover the DM's choices of decision rules and experiments. ||

Example 8 (Collusion in auctions): Consider an auction with m buyers bidding for a single good. Each buyer i knows their own valuation $v_i \in [\underline{v}, \overline{v}]$ of the good, and believes that other buyers' valuations are independently drawn from distributions F_j, $j \in \{1, \ldots, m\} \setminus \{i\}$. This is the classic independent private values (IPV) framework with common beliefs.

Unbeknownst to the econometrician, each buyer may be a member of one of $K < m$ cartels. McAffe and McMillan (1992) and Marshall and Marx (2007), among others, show that bidding behavior will depend on the number of and membership in cartels. Let λ_k denote the probability that a bid comes from a member of cartel $k \in \{1, \ldots, K\}$. That is, λ_k is the proportion of buyers at the auction who belong to cartel k. Denote by β_k the distribution of bids by the members of cartel k, and by b the overall distribution of bids at the auction.

Let X denote the seller's revenue from a given auction. Observed seller revenue depends on the unobserved cartel membership of buyers taking part in the auction, but arguably does not directly affect each buyer's bid conditional on cartel membership. Therefore, we can denote the mixture bid distribution by

$$b(Y \in A|W, X) = \sum_{k \in K} \lambda_k(W, X) \beta_k(Y \in A|W),$$

where Y is a random variable representing the bids, and W are covariates such as the auction format (e.g., first price, second price, English auction, etc.) and the reserve price set by the seller. ||

Let $\mathscr{Y}$ and $\mathscr{X}$ be partitions of $\mathbb{Y}$ and $\mathbb{X}$ respectively, with $K \leq |\mathscr{Y}| = I < \infty$ and $K \leq |\mathscr{X}| = J < \infty$. Define $\mathscr{C} = \mathscr{Y} \times \mathscr{X}$, which is clearly a partition of $\mathbb{Y} \times \mathbb{X}$. Restricted to the elements of the partition $\mathscr{C}$, the model (2.1) can be represented in matrix form as

$$\mathbf{P}(W) = \mathbf{Q}(W)\mathbf{\Lambda}(W), \tag{2.2}$$

where for each $w \in \mathbb{W}$, $\mathbf{P}(w)$ is an $I \times J$ matrix, $\mathbf{\Lambda}(w)$ is a $K \times J$ matrix and $\mathbf{Q}(w)$ is an $I \times K$ matrix. In what follows, I omit the dependency of all distributions on the common conditioning covariate, with the understanding that all quantities are implicitly conditional on W.

If one hopes to recover (Q, λ) from the data, we need to be able to observe sufficient variability of mixture weights and outcomes.

Assumption 2 (Variability). There exists a partition $\mathscr{C}$ such that $\mathbf{P}$ has rank K.

The matrix $\mathbf{P}$ represents the distribution of Y conditional on X, for the events in the partition $\mathscr{C}$. In view of Assumption 2, I will assume throughout that $\mathbf{P}$ has rank K. If that is not the case, one need only consider a finer partition $\mathscr{C}$. The component distribution matrix $\mathbf{Q}$ tabulates the probabilities of the different outcome events $A \in \mathscr{Y}$, for each latent type, while $\mathbf{\Lambda}$ represents the component weights λ, conditional on different realizations of the instrumental variable X. In other words, partitioning the space of signals induces a partition of the mixture model into matrices defined by

$$\begin{aligned}
\mathbf{P}_{ij} &= P(Y \in A_i | X \in B_j);\\
\mathbf{Q}_{ik} &= Q_k(Y \in A_i);\\
\mathbf{\Lambda}_{kj} &= \lambda_k(X \in B_j),
\end{aligned}$$

for all $(A_i, B_j) \in \mathscr{C}$ with $i = 1, \ldots, I$ and $j = 1, \ldots, J$. By construction, $\mathbf{P}$, $\mathbf{Q}$ and $\mathbf{\Lambda}$ are *column-stochastic*: for all $i \in \{1, \ldots, I\}$, $j = \{1, \ldots, J\}$ and $k \in \{1, \ldots, K\}$, we have $\sum_{l=1}^{I} \mathbf{P}_{lj} = \sum_{l=1}^{I} \mathbf{Q}_{lk} = \sum_{l=1}^{K} \mathbf{\Lambda}_{jl} = 1$ and $\mathbf{P}_{ij}, \mathbf{Q}_{ik}, \mathbf{\Lambda}_{kj} \geq 0$. The pair $(\mathbf{Q}, \mathbf{\Lambda})$ is called a (column-)stochastic matrix factorization of $\mathbf{P}$.

Definition 8 (Stochastic matrix factorization). Given any matrix $\mathbf{M} \in \mathbb{R}^{I \times J}$, the pair of matrices

$(\mathbf{B}, \mathbf{C}) \in \mathbb{R}^{I \times K} \times \mathbb{R}^{K \times J}$ is called a stochastic matrix factorization of $\mathbf{M}$ if $\mathbf{B}$ and $\mathbf{C}$ are column-stochastic, and $\mathbf{M} = \mathbf{BC}$. $\diamond$

Equation (2.2) is not the unique factorization of $\mathbf{P}$ into two matrices of rank K, since for any non-singular $K \times K$ matrix $\mathbf{A}$,

$$\mathbf{P} = \mathbf{QAA}^{-1}\mathbf{\Lambda} = \tilde{\mathbf{Q}}\tilde{\mathbf{\Lambda}}. \tag{2.3}$$

However, if $\tilde{\mathbf{Q}}$ and $\tilde{\mathbf{\Lambda}}$ are to also represent discretized parameters of a mixture model, both $\tilde{\mathbf{Q}}$ and $\tilde{\mathbf{\Lambda}}$ must be column-stochastic. This constrains the possible values that $\mathbf{A}$ can take. In particular, it implies that $\sum_{k=1}^{K} \mathbf{A}_{lk} = 1$ (Adams, 2016). In the following section, I leverage this constraint to obtain conditions under which the stochastic factorization $\mathbf{P} = \mathbf{Q\Lambda}$ is unique, which in turn imply identification conditions for the latent parameters of the mixture model.

2.3 Identification

The finite mixture model restricted to the partition $\mathscr{C}$ is identifiable if, and only if, (2.2) is the unique factorization of $\mathbf{P}$ into column-stochastic matrices, up to simultaneous permutations (relabeling) of the columns of $\mathbf{Q}$ and rows of $\mathbf{\Lambda}$. By eq. (2.3), this condition can be expressed in terms of the matrix $\mathbf{A}$.

Definition 9 (Unique stochastic factorization). A stochastic matrix factorization $(\mathbf{Q}, \mathbf{\Lambda})$ of $\mathbf{P}$ is unique if for all other such factorizations $(\mathbf{Q}_\mathbf{A}, \mathbf{\Lambda}_\mathbf{A})$, where $\mathbf{Q}_\mathbf{A} = \mathbf{QA}$ and $\mathbf{\Lambda}_\mathbf{A} = \mathbf{A}^{-1}\mathbf{\Lambda}$, we have that $\mathbf{A}$ is a permutation matrix. $\diamond$

For any given $\mathbf{P}$, we define the identified set $\mathbb{I}(\mathbf{P})$ of its stochastic factorizations as all column-stochastic matrices $\mathbf{QA}$ and $\mathbf{A}^{-1}\mathbf{\Lambda}$ such that $\mathbf{P} = \mathbf{QAA}^{-1}\mathbf{\Lambda}$. Henry, Kitamura, and Salanié (2014) provides a full characterization of the identified set $\mathbb{I}(\mathbf{P})$ in terms of the convex hull of the rows of the matrix $\mathbf{H}$, where $\mathbf{H}_{ij} = \mathbf{P}_{ij} - \mathbf{P}_{i1}$. The identified set also has a characterization in terms of the matrix $\mathbf{A}$ alone, which was shown by Adams (2016):

Lemma 4. Let $\mathbf{Q}\mathbf{\Lambda}$ be any stochastic matrix factorization of $\mathbf{P}$ and define $\mathscr{A}(\mathbf{P}) = \{\mathbf{A} \in \mathbb{R}^{K\times K} : \forall i, j, k,\ \mathbf{P} = \mathbf{Q}\mathbf{A}\mathbf{A}^{-1}\mathbf{\Lambda},\ \sum_{l=1}^{K} \mathbf{A}_{kl} = 1 \text{ and } (\mathbf{Q}\mathbf{A})_{ik}, (\mathbf{A}^{-1}\mathbf{\Lambda})_{kj} \geq 0\}$. Then $\mathbb{I}(\mathbf{P}) = \{(\mathbf{Q}\mathbf{A}, \mathbf{A}^{-1}\mathbf{\Lambda}) \in \mathbb{R}^{I\times K} \times \mathbb{R}^{K\times J} : A \in \mathscr{A}(\mathbf{P})\}$.

From the constraint that each row of the matrix $\mathbf{A}$ sums to one, it is easily seen that the identified set has dimension at most $K(K-1)$. In particular, if there are two component distributions, any $\mathbf{A} \in \mathscr{A}$ can be expressed as

$$\mathbf{A} = \begin{bmatrix} 1+a_1 & -a_1 \\ -a_2 & 1+a_2 \end{bmatrix}$$

with inverse

$$\mathbf{A}^{-1} = \frac{1}{1+a_1+a_2} \begin{bmatrix} 1+a_2 & a_1 \\ a_2 & 1+a_1 \end{bmatrix}.$$

The non-negativity constraints $(\mathbf{Q}\mathbf{A})_{ik}, (\mathbf{A}^{-1}\mathbf{\Lambda})_{kj} \geq 0$ then imply that (a_1, a_2) must satisfy the following linear inequalities for all i, j:

$$\begin{aligned}
&a_2 \leq (1+a_1)\frac{\mathbf{Q}_{i1}}{\mathbf{Q}_{i2}}, \\
&a_2 \geq a_1\frac{\mathbf{Q}_{i1}}{\mathbf{Q}_{i2}} - 1, \\
&\operatorname{sgn}(1+a_1+a_2)a_2 \geq -\operatorname{sgn}(1+a_1+a_2)\left(1 + a\frac{\mathbf{\Lambda}_{2j}}{\mathbf{\Lambda}_{1j}}\right), \\
&\operatorname{sgn}(1+a_1+a_2)a_2 \geq -\operatorname{sgn}(1+a_1+a_2)(1+a)\frac{\mathbf{\Lambda}_{2j}}{\mathbf{\Lambda}_{1j}}.
\end{aligned}$$

The slopes of the linear constraints are likelihood ratios, e.g., $\frac{\mathbf{Q}_{i1}}{\mathbf{Q}_{i2}}$ measures how much less likely it is to observe the outcome event $A_i \in \mathscr{Y}$ coming from component distribution Q_1 than from Q_2. Intuitively, since the linear inequalities above must be satisfied for every pair of outcomes and instruments, the identified set shrinks as the component distributions and mixture weights diverge from each other for at least one pair of events $(A, B) \in \mathscr{C}$.

The discussion above indicates that the following property of matrices will be crucial to characterize which stochastic matrix factorizations are unique.

Definition 10 (*K*-sparsity). Let $\mathbf{M}$ be an $I \times J$ matrix and define the set

$$\mathscr{I}_j = \{i \in \{1, \ldots, I\} : \mathbf{M}_{ij} \neq 0\}. \tag{2.4}$$

We say that the matrix $\mathbf{M}$ is K-sparse if there exist columns $\mathscr{J}_K \subseteq \{1, \ldots, J\}$, with $|\mathscr{J}_K| = K$ and $K \leq \min\{|I|, |J|\}$, such that for all $j, j' \in \mathscr{J}_K$ with $j \neq j'$, we have $\mathscr{I}_j \nsubseteq \mathscr{I}_{j'}$. $\diamond$

Interpreted in terms of the matrix $\mathbf{P}$, K-sparsity says that there exists a subset of K different realizations of the instrument such that, for any pair of realized event B and B' in this subset, at least one outcome that happens with positive probability conditional on B, has zero probability conditional on B'.

By making use of a result by Huang, Sidiropoulos, and Swami (2014) and Lemma 4, Adams (2016) provided a condition on $\mathbf{Q}$ and $\mathbf{\Lambda}$ that is both necessary and sufficient for the stochastic factorization $\mathbf{P} = \mathbf{Q\Lambda}$ to be unique. This condition hinges on the K-sparsity property.

Lemma 5. The stochastic matrix factorization $(\mathbf{Q}, \mathbf{\Lambda})$ of $\mathbf{P}$ is unique if, and only if, both $\mathbf{Q}$ and $\mathbf{\Lambda}^T$ are K-sparse.

The condition in Lemma 5 is tight, so it completely characterizes the matrices $\mathbf{P}$ that have a unique stochastic matrix factorization. However, in the context of mixture models, having identification conditions in terms of the latent parameters of the model leads to at least two issues. First, such identification conditions are often difficult to interpret and to assess qualitatively, particularly when the types themselves do not have a natural interpretation. Second, to empirically test that the conditions hold, one must first estimate the model. But with finite data, the estimate of $\mathbf{P}$ will inevitably be noisy, thus the estimated latent parameters may not satisfy the identification conditions, even if the true parameters do, and vice-versa.

It is thus desirable to obtain a condition that guarantees identification of (2.1) and only imposes constraints on the observable mixture distribution P. It turns out that imposing K-sparsity on the discretized reduced form parameter matrix $\mathbf{P}$ is sufficient to guarantee uniqueness of the stochastic matrix factorization.

Lemma 6. If $\mathbf{P}$ is K-sparse and has a stochastic matrix factorization, then this factorization is unique.

Lemma 6 provides a sufficient condition for the stochastic matrix factorization of $\mathbf{P}$ to be unique. Its proof hinges on showing that K-sparsity imposed on $\mathbf{P}$ implies that the requirements of Lemma 5 on the matrices $\mathbf{Q}$ and $\mathbf{\Lambda}$ are satisfied. Importantly for the mixture model application, the K-sparsity condition pertains only to the observable matrix $\mathbf{P}$.

Typically, albeit sufficient for identification, K-sparsity is not necessary. Indeed, note that with $K \geq 3$, if the matrices $\mathbf{Q}$ and $\mathbf{\Lambda}$ are column-stochastic, and have main diagonals consisting of zeros and all other components strictly positive, then the stochastic factorization $\mathbf{P} = \mathbf{Q}\mathbf{\Lambda}$ satisfies the conditions in Lemma 5, and hence is unique. On the other hand, the resulting $\mathbf{P}$ consists of all strictly positive entries, and thus does not satisfy K-sparsity. However, there is one important case in which K-sparsity is both necessary and sufficient for identification: when there are only two latent components.

Corollary 1. *If $K = 2$, then any stochastic matrix factorization of $\mathbf{P}$ is unique if, and only if, it satisfies K-sparsity.*

The following is a direct result of Lemma 6, after accounting for some measure-theoretic considerations:

Theorem 6. *Let Y, W and X be random variables with distributions satisfying the mixture model* (2.1) *with latent parameters (Q, λ). For each $w \in \mathbb{W}$, if there exists a partition $\mathscr{C}$ of $\mathbb{Y} \times \mathbb{X}$ such that the model restricted to $\mathscr{C}$ can be represented by eq.* (2.2) *with $\mathbf{P}$ satisfying K-sparsity, then (Q, λ) are identifiable conditional on w. Conversely, if $K = 2$ and (Q, λ) is unique, then there exists a partition $\mathscr{C}$ such that $\mathbf{P}$ satisfies K-sparsity.*

Theorem 6 states that to guarantee that the mixture model (2.1) with K latent components is identifiable, it suffices to find a single discretization of the reduced form parameter P for which the matrix $\mathbf{P}$ is K-sparse. Because it only imposes constraints on the observable matrix $\mathbf{P}$, this sufficient condition has two major advantages.

First, its interpretation in terms of the joint distribution of Y and X is immediate, and does not depend on the nature of the hidden types. Even if types do not have an intuitive interpretation a priori, such as when the mixture model is mainly used as a tool for dimensionality reduction, Theorem 6 can still be used to gauge whether the identifiability assumption is likely to hold. This is the case, for instance, in the large literature on topic models.

Second, it opens up the possibility of devising a statistical test to assess whether the parameters (Q, λ) of the mixture model are identifiable, after observing a sample from the random vector (Y, W, X). The next section is dedicated to developing such a test and discussing other matters of inference on finite mixture models.

2.4 Inference

2.4.1 A test of identification

Suppose we observe a random sample $Z = (Y_l, X_l)_{l=1}^{n}$ drawn from the joint distribution P conditional on some value of X. Before attempting to estimate the latent parameters (Q, λ), it would be useful to know whether they can really be identified by observing only the reduced form parameter P. If that is not the case, then a consistent classical estimator of (Q, λ) does not exist, and any Bayesian estimator will be very sensitive to the prior, even in large samples (Ke, Luis Montiel Olea, and Nesbit, 2021).

Fix a partition $\mathscr{C} = \mathscr{Y} \times \mathscr{X}$ of $\mathbb{Y} \times \mathbb{X}$ with $|\mathscr{Y}| = I$ and $|\mathscr{X}| = J$, and consider the corresponding discretization $\mathbf{P}$ of the true reduced form parameter P. The simple conditional histogram, i.e., the frequency of samples falling on each element of the partition $\mathscr{Y}$ conditional on each element of $\mathscr{X}$, is a consistent estimator $\hat{\mathbf{P}}^n$ of $\mathbf{P}$. Let Θ be the set of all column-stochastic matrices of rank K and $\Theta_0 = \{\mathbf{M} \in \Theta : \mathbf{M} \text{ satisfies } K\text{-sparsity}\}$. Since we do not have a full characterization of identification conditions in terms of $\mathbf{P}$ alone, I will instead only consider testing the following, more restrictive null hypothesis, based only on the sufficient condition:

$$H_0 : \mathbf{P} \in \Theta_0 \quad vs. \quad H_1 : \mathbf{P} \in \Theta \setminus \Theta_0.$$

If Y and X are discrete, we may take $\mathscr{C}$ to be the finest possible partition of $\mathbb{Y} \times \mathbb{X}$. When either Y or X are continuous random variables, the discretization would ideally be chosen optimally, as a function of the non-parametric estimator of P, so as to maximize the power of the test. This is an interesting avenue for future research, but lies beyond the scope of the present paper, so $\mathscr{C}$ will remain fixed throughout this section.

Typically, one would test H_0 using either a χ^2 or likelihood-ratio test statistic based on the consistent estimator $\hat{\mathbf{P}}^n$. However, in this case the null hypothesis involves constraining some parameters to be on the boundary of the parameter space (the entries where $\mathbf{P}_{ij} = 0$ under H_0) and others to be *any point* in the interior of the space (where $\mathbf{P}_{ij} > 0$ under H_0). Testing such hypotheses is at best difficult (in the case of boundary constraints) or outright intractable (for strict inequality constraints) within the framework of classical asymptotic inference. Therefore, I take a Bayesian approach.

Let f_θ denote the sampling distribution of the estimator $\hat{\mathbf{P}}^n$ conditional on $\theta \in \Theta$, and π be an absolutely continuous prior probability over Θ. For each realization p of $\hat{\mathbf{P}}^n$, let $\pi(\theta|p) = f_\theta(p)\pi(\theta)$ be the posterior likelihood of θ given p. The proposed test rejects the null hypothesis if the posterior probability that the parameter is within an ε distance of Θ_0 is small, with the distance from a parameter θ to a set T being measured as

$$d(\theta, T) = \inf_{t \in T} \|\theta - t\|,$$

where $\|\cdot\|$ is any norm on $\mathbb{R}^{I \times J}$.

Definition 11 (Bounded distance test). The bounded distance test rejects H_0 whenever

$$Pr(d(\theta, \Theta_0) \leq \varepsilon | p) < \alpha,$$

where both $0 < \alpha < 1$ and $\varepsilon > 0$ are chosen by the statistician. ◊

If $\varepsilon > 0$ is small and $Pr(d(\theta, \Theta_0) \leq \varepsilon | p)$ is sufficiently close to 1, we can be confident that the true parameter $\mathbf{P}$ is not far from some identified parameter. This in turn guarantees that the identified set $\mathbb{I}(\mathbf{P})$ is not far from a single point, which implies that the lack of identifiability is not too severe. In this sense, $\mathbb{I} : \Theta \rightrightarrows [0,1]^{I\times K} \times [0,1]^{K\times J}$ is a continuous correspondence. This is formalized in the following result.

Proposition 7. *For all $\eta > 0$, there exists $\varepsilon > 0$ and an $I \times J$ matrix $\mathbf{B}$ having a unique stochastic factorization, such that*

$$d(\mathbf{P}, \Theta_0) \leq \varepsilon \implies \sup_{(\mathbf{Q},\boldsymbol{\Lambda})\in\mathbb{I}(\mathbf{P})} \|(\mathbf{Q}, \boldsymbol{\Lambda}) - \mathbb{I}(\mathbf{B})\|_F \leq \eta,$$

where $\|\cdot\|_F$ denotes some norm in the factorization space $\mathbb{R}^{I\times K} \times \mathbb{R}^{K\times J}$.

The main obstacle to implementing the proposed test is the characterization of the null set Θ_0. Note that, from the definition of K-sparsity, H_0 can be written as a finite set of inequality constraints involving all $I \times K$ submatrices of $\mathbf{P}$. Although finite, unfortunately the number of constraints characterizing H_0 can be very large, and listing all of them could be prohibitive. The following procedure implements the proposed test while bypassing this problem. It takes advantage of the fact that a generic matrix $\tilde{\theta}$ sampled from $\pi(\theta|p)$ will have all strictly positive elements, thus the matrix in Θ_0 which is closest to $\tilde{\theta}$ can be obtained by setting a carefully chosen set of K elements of the sampled matrix to zero.

1. Choose a tolerance parameter $\varepsilon > 0$ and a confidence level $1 - \alpha$.

2. Compute the posterior distribution from the data, given prior π. Note that if $\hat{\mathbf{P}}^n$ is a histogram, then nf_θ is a conditional multinomial distribution. In this case, we can use the well known Latent Dirichlet Allocation (LDA) estimator developed by Blei, Ng, and Jordan (2003), to obtain the posterior distribution $\pi(\theta|p)$.

3. Sample m column-stochastic matrices of rank K, $\{\theta^l\}_{l=1}^m$, from the posterior distribution.

   To calculate $d(\theta^l, \Theta_0)$ for each $l = 1, \ldots, m$, first note that since π is assumed to be absolutely continuous, θ^l will generically have all entries strictly positive. Therefore, one may find the distance by applying the following algorithm:

(i) If $k = 0$, let $\mathbf{M}^{(0)} = \theta^l$.

(ii) For all $k \in \{1, \ldots, K-1\}$, let

$$(i_k, \mathscr{J}_K) = \arg\min \left\{ \mathbf{M}^k_{ij} : i \in \{1, \ldots, I\} \setminus \{i_1, \ldots, i_{k-1}\},\ j = \{1, \ldots, J\} \setminus \{j_1, \ldots, j_{k-1}\} \right\}.$$

Define the matrix $\mathbf{M}^{(k+1)}$ by

$$\mathbf{M}^{(k+1)}_{ij} = \begin{cases} \mathbf{M}^{(k)}_{ij}, & \forall (i, j) \neq (i_k, \mathscr{J}_K). \\ 0, & \text{otherwise} \end{cases}$$

(iii) Let $\theta^* = \mathbf{M}^{(K)}$. We have $d(\theta^l, \Theta_0) = \|\theta^l - \theta^*\|$.

The algorithm proceeds by iteratively finding the smallest element in the sampled matrix, setting this element to zero, and repeating the process for the matrix obtained by ignoring the row and column corresponding to the null elements.

4. We reject the null hypothesis if, and only if,

$$\frac{|\{l \in \{1, \ldots, m\} : d(\theta^l, \Theta_0) \leq \varepsilon\}|}{m} < \alpha.$$

As with any statistical test, the procedure just described involves a somewhat arbitrary choice of constants by the statistician, in this case namely α and ε. In the Bayesian framework, it is thought that such choices should be guided by a decision-theoretic rationale. In other words, the test procedure should arise as the solution of a statistical decision problem, as presented in Chapter 1.

We will consider a DM who simultaneously chooses contingent plans for whether to accept or not the null hypothesis and for what tolerance region to consider. Let the set of actions available to the decision maker be $A = \{0, 1\} \times \mathscr{K}(\Theta)$, where $\mathscr{K}(\Theta)$ denotes the set of compact subsets of Θ.

Here, 0 denotes accepting the null hypothesis, 1 denotes rejecting it and $T \in \mathscr{K}(\Theta)$ is a tolerance region, which represents the values of the parameter that are deemed close enough to an identified set for classical inference on **P** not to be too misleading. Consider the following utility function:

$$u((a,T),\theta) = \begin{cases} -c_1, & \text{if } a = 1 \\ -c_2(1 - \mathbf{1}_{\{\theta \in T\}}) - c_3\, \mathbf{1}_{\{\theta \in T\}}\, d(\theta, \Theta_0), & \text{if } a = 0. \end{cases} \tag{2.5}$$

The constant $c_1 > 0$ denotes the cost for the DM of rejecting the null hypothesis and foregoing any claim of identification. This represents the opportunity cost of not performing inference on **P**. On the other hand, if the DM accepts the null hypothesis she incurs cost c_2 when the true parameter is not in the tolerance region, and a cost proportional to the distance between the true parameter and the identified set otherwise. If the statistician accepts the null hypothesis, c_2 is the disutility of the true parameter being intolerably far from identification, while c_3 is the marginal disutility of the distance between the true parameter **P** and the nearest identified parameter. When following the classical approach of assigning the null hypothesis a privileged status, we assume that $c_1 > c_2 > 0$ and $c_3 > 0$. The following result states that the test procedure in Definition 11 can arise as the optimal decision rule for the Bayesian statistical decision problem (u, A, π).

Theorem 7. *Let $\pi \in \Delta(\Theta)$ be a prior probability distribution. Then, the preferences that are represented by (u, π) satisfy the MRA axioms, Mixture Independence and Monotone Continuity* (see Section 1.5).*

Moreover, any optimal decision rule $\delta^* : \Theta \to A$ for (u, π), defined as

$$\delta^*(p) = (a(p), T(p)) \in \underset{\{\delta:\Theta\to A\}}{\arg\max}\, \mathbb{E}_\pi[u(\delta(p), \theta)],$$

satisfies, for all $p \in \Theta$,

$$T(p) = \{\theta \in \Theta : d(\theta, \Theta_0) \leq \varepsilon\}\,;$$

$$a(p) = \begin{cases} 1, & \textit{if } Pr\left(d(\theta, \Theta_0) \le \varepsilon | p\right) < \alpha(\varepsilon, p) \\ 0, & \textit{otherwise,} \end{cases}$$

where $\varepsilon = \frac{c_2}{c_3}$ *and* $\alpha(\varepsilon, p) = 1 - \frac{\frac{c_1}{c_3} - \int_T d(\theta,\Theta_0)\pi(\theta|p)\mathrm{d}\theta}{\varepsilon}$.

Note that Theorem 7 has interesting implications for the choice of ε and α. First, ε should not depend on the realized sample. This is intuitive, since the tolerance region is a measure of the degree of identifiability of the true model, not depending on any properties of the data generating process. On the other hand, the choice of α does depend on p, through the integral $\int_T d(\theta, \Theta_0)\pi(\theta|p)\mathrm{d}\theta$. This is somewhat unusual, but note that the integral in question can be numerically approximated, given $\varepsilon > 0$, by sampling from the posterior distribution and calculating the distance, similarly to the procedure described above. Moreover, $\int_T d(\theta, \Theta_0)\pi(\theta|p)\mathrm{d}\theta \le \varepsilon$ for all p, so a good rule of thumb is to take $\alpha(\varepsilon) = 2 - \frac{c_1}{c_2}$, especially when ε is small.

2.4.2 Estimation

When $(\mathbf{Q}, \mathbf{\Lambda})$ are identified, the approximate non-negative matrix factorization of the sample matrix is a consistent estimator for the discretized latent parameters.

Theorem 8. *Suppose P has a discretization $\mathbf{P}$ with a unique stochastic factorization $(\mathbf{Q}^*, \mathbf{\Lambda}^*)$ and let*

$$(\hat{\mathbf{Q}}^n, \hat{\mathbf{\Lambda}}^n) \in \arg\min \left\{ \|\hat{\mathbf{P}}^n - \mathbf{Q}\mathbf{\Lambda}\|_2 : \forall i, j, k, \ \sum_{l=1}^{I} \mathbf{Q}_{lk} = \sum_{l=1}^{K} \mathbf{\Lambda}_{jl} = 1 \ \textit{and} \ \mathbf{Q}_{ik}, \mathbf{\Lambda}_{kj} \ge 0 \right\}, \quad (2.6)$$

where $\|\cdot\|_2$ *denotes the Frobenius norm. Then,* $(\hat{\mathbf{Q}}^n, \hat{\mathbf{\Lambda}}^n) \xrightarrow{p} (\mathbf{Q}^*, \mathbf{\Lambda}^*)$ *as* $n \to \infty$.

The proof of Theorem 8 is a direct application of consistency results for extremum estimators, and will be ommitted. If we take increasingly finer partitions $\mathscr{C}^n$ of $\mathbb{Y} \times \mathbb{X}$ as n grows, where $|\mathscr{C}^n| = m_n$ and $\frac{m_n}{g(n)} \to 0$ for some concave function g which depends on the non-parametric estimator of

P, we obtain $(\hat{\mathbf{Q}}^n, \hat{\mathbf{\Lambda}}^n) \xrightarrow{p} (Q, \lambda)$. Therefore, by taking increasingly finer discretizations of P, one obtains a consistent estimator for (Q, λ).

Up until this point, I assumed that the number of types K was known. However, this parameter can also be estimated from the data. In fact, note that if P satisfies model (2.1) with K components, then for a fine enough partition $\mathscr{C}$, $\mathbf{P}$ has rank K. Thus to consistently estimate the number of types, we need only estimate the rank of $\mathbf{P}$. Fortunately, such estimators have been studied in the literature, including in a recent paper by Chen and Fang (2019).

Putting everything together, I propose the following estimation procedure for (Q, λ):

1. Non-parametrically estimate P from an i.i.d sample of size n and discretize it to obtain the estimate $\hat{\mathbf{P}}^n$. Let $\hat{\mathbf{P}}$ be an $I \times J$ matrix.

2. Estimate the rank of r of $\mathbf{P}$ by applying the estimator $\hat{r}_n$ in Chen and Fang (2019). Let $K = \hat{r}_n$.

3. Apply the test in Definition 11 assuming K latent types.

4. If H_0 can not be rejected, estimate (Q, λ) using the approximate non-negative matrix factorization (2.6).

2.5 Discussion

This paper presents a condition which guarantees identifiability of finite mixture models, assuming there exists an instrumental variable which shifts only the mixture weights. Unlike previous results, this sufficient condition is stated in terms of properties of the observable mixture distribution alone. This fact allowed me to devise a formal statistical test, which can be applied before estimating the component distributions and weights, to assess the identifiability of the hidden parameters. The test makes use of a prior on a discretized version of the mixture distribution, and is in fact a formal Bayes test, in the sense that it arises as the Bayes decision rule of a statistical decision problem.

The approach to identification and inference presented here relies on discretizing the mixture distribution. If the outcome random variable is already discrete — as is the case, for instance, in many data sets obtained from laboratory experiments — no additional complications are imposed by the procedures described in the paper. However, in applications where outcomes are drawn from a continuous random variable, the discretization must be done carefully.

If we take an excessively fine partition, there will be large variance in the estimated mixture distribution. On the other hand, if the partition is too coarse, the estimator will be biased. In both cases, the proposed test is liable to inflate the probability of type I error. One possible solution would be to start with a very fine discretization, and proceed to sum all the rows and columns of the resulting matrix that have zeros is the same positions, thus obtaining the coarsest discretization that preserves the original sparsity pattern. However, this heuristic alone does not solve the problem, and could itself create other issues. A fruitful avenue of future research is to develop bandwidth selection criteria tailored to the mixture model discretization problem.

Since the test of identification proposed in Section 2.4 only makes use of a sufficient, but not necessary condition, it will tend to understate the evidence in favor of the hypothesis that the model is identified. Therefore, it would be desirable to assess the false rejection probability of the test in situations where the model is known ex ante to be identified. This could be done by means of simulations: since the necessary and sufficient conditions for identifiability in terms of the hidden parameters are known, it is easy to simulate data from identified models and obtain the rejection probability of the test for such models. This could be used to fine tune the parameters of the test to obtain the desired false rejection probability.

Chapter 3: Statistical Mechanism Design: Robust Pricing and Reliable Projections

Joint work with Duarte Gonçalves

3.1 Introduction

When facing uncertainty about consumers' willingness-to-pay, pricing and projections are two central elements of a firm's business plan. As the profitability of any pricing strategy is uncertain, it may be desirable to pursue strategies that provide revenue guarantees for the firm. On the other hand, firms often depend on projections about their future revenue under different scenarios to inform budget planning, inventory management, capital investments, and more. Consequently, they need to be able not only to consistently estimate their expected profits but also to have reliable confidence bounds. While pricing and projections are conceptually intertwined, these can and are often considered separately.

In this paper, we propose a data-based approach to tackle both these issues. In our setup, the firm faces uncertainty about the true distribution of consumers' types in the otherwise canonical setup of Maskin and Riley (1984). As is standard, we allow the firm to design mechanisms – pairs of prices and quantities – though our results also hold in the more restrictive setting of uniform pricing. However, differently from what is typically assumed in the robust mechanism design literature, instead of holding precise information about features of the distribution of consumers' willingness-to-pay, we assume the firm observes a finite sample drawn from this distribution.

Our contribution is twofold. First, we study a specific pricing strategy, the *empirically optimal mechanism*, and its finite sample robustness properties. Second, we provide a toolkit to perform statistical inference on the profit obtained for any arbitrary mechanism, including the empirically optimal one, enabling a data-based approach to the evaluation and comparison of different pricing strategies.

Empirically optimal mechanisms are constructed in a simple and intuitive manner. Fixing any given distribution of consumer types, we start by obtaining a mechanism that is optimal for that distribution. This mapping from distributions to optimal mechanisms is then coupled with a consistent estimator for the true distribution. Hence, empirically optimal mechanisms maximize expected profit when the estimate is taken to be the true distribution. Importantly, this class of menus relies on a fully nonparametric, prior-free estimator of the type distribution.

We show that mechanisms constructed in this manner are asymptotically optimal, achieving the optimal profit with probability one as the sample size grows. Moreover, empirically optimal mechanisms are robust in the spirit of Bergemann and Schlag (2011), that is, small perturbations of the estimated distribution, induced by changes in the underlying data, do not affect the firm's expected profits. This follows from establishing Lipschitz continuity of the firm's value function in the distribution.

Empirically optimal mechanisms also entail strong probabilistic guarantees for both profit and regret – the difference between the optimal expected profit and the expected profit that a mechanism delivers under the true distribution – and we characterize how these relate to the sample size. For any given finite sample size, we obtain a probabilistic lower bound for expected profit and a probabilistic upper bound for regret. Crucially, these bounds are not asymptotic and depend only on known constants. These results are then used to establish how many samples the firm needs, from an *ex-ante* perspective, to obtain probabilistic bounds on expected profit and on maximal regret. Our findings are related to the growing literature on sample-based revenue guarantees (see e.g. Cole and Roughgarden 2014; Huang, Mansour, and Roughgarden 2018; Guo, Huang, and Zhang 2020), as we provide a non-asymptotic lower bound on sampling requirements for profit and regret guarantees.

We then provide tools for estimating expected profit and reliably conducting inference, not only for a mechanism that is empirically optimal given an estimate of the distribution, but for any fixed mechanism. We derive an estimator for expected profit that is consistent and unbiased, and, when appropriately rescaled, asymptotically normal. Further, we note the validity of a bootstrap implementation to conducting inference.

This approach enables estimation of the expected profit that any given mechanism attains for purposes such as budgeting, regardless of the criteria that guided the choice of the mechanism. Moreover, the ability to conduct data-based inference on the expected profit also expands the criteria that can be used to select a mechanism, namely by considering probabilistic revenue guarantees given by confidence intervals, by testing which in a set of alternative mechanisms achieves a higher expected profit, or by directly analyzing the distribution of the difference in profit between any two mechanisms. As such, this methodology of estimation and inference applies generally and opens up a new, data-based approach to deriving robust revenue guarantees.

In particular, we show how empirically optimal mechanisms can be used to estimate and conduct inference on the *optimal expected profit*, that is, the maximum expected profit that could be obtained were the firm to know the true distribution. Conducting inference on the optimal expected profit is useful for two reasons. First, as regret considerations constitute a standard criterion in the literature for selecting among competing mechanisms and as regret, by its definition, depends on knowledge of the optimal expected profit, a means to consistently estimate and conduct inference on the latter will enable the same for the former. Second, the maximum achievable profit can itself be an object of interest when, for instance, comparing the expected return of alternative investment possibilities.

While consistency of our suggested estimator follows from our results on asymptotic optimality of empirically optimal mechanisms, its asymptotic normality follows from a novel envelope theorem for the firm's value function. In other words, we show that the value function is Fréchet differentiable in the distribution of consumer types and that its Fréchet derivative equals that of the profit function at the empirically optimal mechanism. With Fréchet differentiability in hand, a Delta method for statistical functionals applies, and asymptotic normality of our estimator ensues.

We then again consider bootstrap implementations for conducting inference.

To study the finite sample properties of the proposed estimators, we conduct Monte Carlo simulations of their bootstrap implementation for the standard specification of sale of an indivisible good. We perform this exercise for the expected profit of a given mechanism as well as for the optimal profit. The evidence shows that the empirical coverage of our estimators approximates well the associated confidence intervals, even with relatively few samples.

Finally, we illustrate how our results on empirically optimal mechanisms partially extend to an auction setting. We consider the case where the firm auctions a single item to a finite number of risk-neutral bidders with independent private values drawn from the same distribution. In particular, analog versions of asymptotic optimality and profit and regret guarantees are shown to hold in this setting as well.

Empirically optimal mechanisms correspond to one of the simplest forms of statistically informed mechanism design: observing a sample, estimating a distribution, and implementing a mechanism that is optimal for the estimated distribution. As we demonstrate, this extremely simple mechanism not only has sound revenue guarantees, but also allows practitioners to estimate the maximum profit attainable. Our results on estimation of the expected profit enable designers to construct confidence intervals and perform hypothesis testing, constituting a useful tool for practitioners and empiricists. From a theoretical perspective, we hope our data-based perspective on robust mechanism design proves useful in the broader study of mechanism design under model uncertainty.

A brief outline of the paper is as follows. Section 3.2 places the paper within the existing literature. Section 3.3 introduces the main theoretical framework. In Section 3.4, we define our class of empirically optimal mechanisms and examine some of their main properties: asymptotic optimality and profit guarantees. After exploring this particular class of mechanisms, in Section 3.5 we turn to the question of providing a statistical toolkit to estimate and conduct inference on profit, including optimal expected profit. Section 3.6 illustrates an extension of our results to the auction setting with independent private values. Finally, we conclude with a discussion of specific suggestions for further work in Section 3.7. All omitted proofs are included in Chapter C.

3.2 Related Literature

The most directly related literature studies robust mechanism design with a monopolist who has perfect knowledge not about the whole distribution as in the more standard models (e.g. Maskin and Riley, 1984), but only about some features of this distribution. With such information, the firm can then narrow down the set of possible distributions to consider and adopt a pricing strategy that maximizes the worst-case profit or minimizes the worst-case regret. This approach tries to address the concern that the optimal mechanism is not robust to the firm having less than exact information on the distribution of consumers' willingness-to-pay, in line with the general research program of robust mechanism design.

The papers in this literature closest to ours are Bergemann and Schlag (2008; 2011) and Carrasco et al. (2018a). These papers model a firm that does not know the distribution of consumer types, but has access to imperfect information that allows it to refine the set of possible distributions. Focusing on a linear specification, they assume the firm acts as if it faces an adversarial nature that chooses a distribution to maximize regret.[1] Bergemann and Schlag (2008) assume that the firm knows only an upper bound for the support; Bergemann and Schlag (2011) study the case where the firm also knows that the true distribution of consumers' willingness-to-pay is in a given neighborhood of a given target distribution; Carrasco et al. (2018a) posits that the firm knows either the first moment and an upper bound for the support of the distribution,[2] or the first two or three moments of the distribution.

These papers then characterize the regret-minimizing mechanisms, whereby the firm hedges against uncertainty by randomizing over prices. In contrast, we assume that the firm does not know specific features of the distribution but instead has access to a sample drawn from the true distribution, from which these features could potentially be estimated and upon which an empirical

[1] Bergemann and Schlag (2011) also consider the case where nature minimizes profit and show the firm chooses a deterministic uniform pricing rule.

[2] This is also the case in a closely related paper, Carrasco et al. (2018b), where the assumptions on consumer's utility function of linearity in quantity and on linearity of the firm's cost are relaxed.

version of these mechanisms could be implemented. Our analysis shows that, in addition to its other desirable properties, the empirically optimal mechanism generates nearly minimal maximum regret in the sense of these papers.

In another closely related paper, Madarász and Prat (2017) allow for a firm that is uncertain over both the distribution of types and the functional form of consumers' utility functions, while at the same time endowing the firm with a possibly misspecified benchmark model of consumer demand. They provide a uniform bound on regret that depends on the distance between the firm's benchmark model and the truth, where the measure of distance between models is related to the largest absolute value of the difference of willingness-to-pay across all possible types and quantities. Instead of looking at the worst-case scenario solution, the authors show that adjusting the pricing strategy that is optimal for the misspecified model in a specific manner involving this distance leads to a uniform bound on regret.

The present study takes a similar approach to Segal (2003), in that both papers determine the optimal mechanism from an estimate of the distribution function. While we focus on issues of estimation and inference, Segal (2003) is concerned with designing a mechanism that makes optimal use of the information contained in consumers' reported valuations. This optimal mechanism implements a bidding system in which consumers receive the good if their bid is above a threshold price that depends on others' bids. This induces consumers to reveal their own type while simultaneously allowing the firm to use others' bids to infer the distribution. The paper shows that as the number of consumers increases, profits converge to the optimal profit with a known type distribution, and discusses rates of convergence. However, the dependency of the optimal price on others' valuations forces the imposition of strong assumptions on the estimated distribution function, rendering non-parametric methods problematic. Our framework, on the other hand, is singularly suited to rely on non-parametric estimation of and inference on the type distribution and does not require consumers to directly communicate their types.[3] Furthermore, we believe

[3] For instance, the sample may be obtained by implementing revealed-preference elicitation techniques. An example would be providing a menu that corresponds to a direct mechanism inducing perfect discrimination between all possible types – where the quantity is a strictly increasing function of the type – and infer consumers' types from their choices. This would then be incentive compatible if consumers are myopic, the good is durable or if it is possible to

that a key product of our paper is its novel toolkit for conducting inference and obtaining revenue guarantees for any mechanism.

Our work is also related to the literature on sample complexity. Huang, Mansour, and Roughgarden (2018) deals with the case where the firm observes independent samples from the unknown true type distribution and consumers have quasilinear-linear utility functions as above. The authors provide probabilistic asymptotic bounds on the number of samples from the true distribution of consumers' valuations that are necessary to achieve a share of $1 - \varepsilon$ of the optimal profit. Similarly, our sample complexity results provide non-asymptotic sample upper bounds for regret and profit for specific mechanisms, but for the more general class of utility functions. Fu, Haghpanah, and Hartline (2020), in an auction setup, characterize the number of samples from the true distribution of bidders' types that are necessary for the designer to achieve full surplus extraction. Opposite to this paper, their type space is finite and the unknown distribution of types belongs to a finite set of joint distributions over all the bidders' values such that the types are correlated in a specific manner.

Finally, our paper is complementary to the extensive econometrics literature on identification and estimation of mechanism design models. Such papers usually treat profits as observable and try to identify the underlying distribution of consumer types. For example, Athey and Haile (2002) studies the identification of buyers' type distributions and information structures in standard auction models; Agarwal and Somaini (2018) estimate the demand of students for different schools when the selection mechanism provides incentives to misreport their types; and Einav, Jenkins, and Levin (2012) estimate demand under adverse selection in credit markets with screening mechanisms. In contrast, the current paper takes as input a consistent estimator of the distribution of consumer types, and provides tools to perform statistical inference on both the optimal profit and the profit from any given mechanism. Therefore, the empirical mechanism design literature could provide a methodology to obtain an appropriate consistent estimator of consumer types, to be used as an input to our model.

exclude them from future purchases.

3.3 Setup

Let $\Theta := [\underline{\theta}, \bar{\theta}] \subset \mathbb{R}$, denote the set of feasible consumer types, which are distributed according to the cumulative distribution $F_0 \in \mathbb{F}$, where $\mathscr{F}$ corresponds to the set of all distributions on Θ endowed with the supremum-norm $\| \cdot \|_\infty$, i.e., $\|F\|_\infty := \sup_{t \in \mathbb{R}} |F(t)|$ for all $F \in \mathbb{F}$. Consumers' utility is given by $u(\theta, x, p) = v(\theta, x) - p$, where θ is the consumer's type, $x \in X := [0, \bar{x}]$ denotes quantity and $p \in \mathbb{R}_+$ price. We assume that v is twice continuously differentiable, concave in x, supermodular, $v(\underline{\theta}, x) = v(\theta, 0) = 0$ for all θ and x, increasing in both arguments and, wherever strictly positive, strictly so.

The firm can choose a menu or mechanism M from the set $\mathscr{M}$ of all compact menus $M' \subset X \times \mathbb{R}_+$ containing the element $(0, 0)$. These comprise pairs of quantity and prices that the consumers can choose, with the option of consuming nothing being always available. We impose the further restriction that if $(x, 0) \in M'$, then $x = 0$; that is, the firm does not give away strictly positive quantities of the good for free.[4] The firm incurs a twice differentiable, convex and strictly increasing cost for quantity sold, $c : X \to \mathbb{R}$, where $c(0) = 0$. When choosing menu $M \in \mathscr{M}$ subject to a distribution $F \in \mathbb{F}$ of consumer types, the firm's expected profit $\pi(M, F)$ is given by

$$\pi(M, F) := \int p(\theta) - c\left(x(\theta)\right) dF(\theta)$$

for some $(x(\theta), p(\theta)) \in \arg\max_{(x,p) \in M} u(\theta, x, p)$, with ties broken in favor of the firm.

Our setup encompasses many of the variations in the literature,[5] being mostly the same as that in the one buyer version of Myerson (1981) and Maskin and Riley (1984), except that there c is assumed to be linear and $\mathscr{F}$ corresponds to the distributions with a strictly positive density. Mussa and Rosen (1978), instead, assume that $v(\theta, x) = \theta \cdot x$ and specify c to be strictly convex. In Bergemann and Schlag (2011) and Carrasco et al. (2018a), $v(\theta, x) = \theta \cdot x$ and $\mathscr{F}$ is a subset of

[4] This restriction facilitates Section 3.5's inference exercise for an arbitrary fixed menu and is without loss of revenue from the firms' perspective.

[5] It does not include, for instance, the case where the firm's cost depends directly on the consumers' type as in Example 4.1 in Toikka (2011).

distributions that satisfy some pre-specified conditions.

We consider the case where neither the firm nor the consumers know the true distribution of types, $F_0 \in \mathbb{F}$, which motivates the choice of dominant strategies as our solution concept. Instead, the firm has access to a sample $S^n = (\theta_i,\, i = 1, ..., n) \in \Theta^n$, $n \in \mathbb{N}$, where each θ_i is independently drawn from F_0. This gives rise to the problem of selecting a menu depending on the realized sample. Let $\mathscr{S}$ denote the set of all samples, $\mathscr{S} := \bigcup_{n \in \mathbb{N}} \Theta^n$. A *sample-based mechanism* is then a mapping $M_S : \mathscr{S} \to \mathscr{M}$, which selects a specific menu depending on the realized sample.

The robust mechanisms mentioned earlier can easily be adjusted to incorporate the information in the sample in order to estimate the features of the distribution that they assume to be known. For example, given a particular sample S^n, the firm can then estimate the support of the true distribution and implement the mechanism given in Bergemann and Schlag (2008). Alternatively, it can estimate the first few moments of the true distribution and implement the mechanism in Carrasco et al. (2018a). A natural question is then whether, given sampling uncertainty, these sample-based mechanisms would exhibit probabilistic robustness properties akin to the deterministic ones that hold when these features are perfectly known. This question is relevant in practice, since *any* information about an unknown distribution is usually obtained from finite data. We therefore take a more direct and, arguably, simpler approach that makes full use of the sample itself to "learn" about the underlying true distribution and inform mechanism choice.

3.4 Empirically Optimal Mechanisms

In this section, we introduce the class of empirically optimal mechanisms. This class of mechanisms is defined by two elements: an estimator of the true distribution and a mapping that takes each estimated distribution to a menu that would be optimal were the estimate to coincide with the true distribution.[6]

Let $\widehat{\mathbb{F}}$ denote the set of estimators that are consistent for F_0, that is, the set of estimators $\hat{F}$ such

[6] Note that, if the seller did have a prior, $\mu \in \Delta(\Delta(\Theta))$, over the set of possible distributions, it would still be sufficient to consider only the expected distribution according to the seller's posterior, $\mathbb{E}_\mu[F|S^n] \in \mathscr{F}$, in order to determine which mechanism to choose. That is, due to linearity of the profit function on the distribution, $\max_{M \in \mathscr{M}} \mathbb{E}_\mu[\pi(M, F)|S^n] = \max_{M \in \mathscr{M}} \pi(M, \mathbb{E}_\mu[F|S^n])$.

that (i) $\hat{F} : \mathscr{S} \longrightarrow \mathscr{F}$; and (ii) $\|\hat{F}(S^n) - F_0\|_\infty \xrightarrow{p} 0$, for any F_0 in $\mathbb{F}$. Let $M^* : \mathbb{F} \longrightarrow \mathscr{M}$ be a fixed selection from the set of optimal menus for every distribution, that is, $\forall F \in \mathbb{F}$, $M^*(F) \in \mathscr{M}^*(F) := \arg\max_{M \in \mathscr{M}} \pi(M, F)$. An *empirically optimal mechanism* $\hat{M}^*$ is a sample-based mechanism that simply joins together a consistent estimator and a selection from the set of optimal menus, that is, $\hat{M}^* = M^* \circ \hat{F}$. We refer to the set of empirically optimal mechanisms as $\widehat{\mathscr{M}^*}$.

While extremely simple, nothing ensures us that such a sample-based mechanism is either well-defined in general environments or that it constitutes a reasonable approach to pricing under uncertainty. The purpose of this section is to address these issues and show that this "naive" approach to pricing delivers several desirable properties including strong probabilistic robustness guarantees.

3.4.1 Existence

In order to show that the set of empirically optimal mechanisms $\widehat{\mathscr{M}^*}$ is nonempty, we start by briefly noting that an optimal menu exists for any distribution $F \in \mathbb{F}$, that is, $\mathscr{M}^*(F) \neq \emptyset$ for all $F \in \mathbb{F}$. We include a formal proof of this statement in the appendix.

Since this implies that a selection $M^* : \mathbb{F} \longrightarrow \mathscr{M}$ such that $M^*(F) \in \mathscr{M}^*(F)$ exists and as there are consistent estimators for any $F_0 \in \mathbb{F}$, the set of empirically optimal mechanisms $\widehat{\mathscr{M}^*}$ is nonempty. An example of a consistent (and unbiased) estimator for F_0 is the empirical cumulative distribution, defined as $\hat{F}(S^n)(\theta) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{\theta_i \leq \theta\}}$, which, for any $F_0 \in \mathscr{F}$, has a uniform rate of convergence, that is, $\|\hat{F}(S^n) - F_0\|_\infty \xrightarrow{p} 0$ (in fact, by the Glivenko–Cantelli theorem, uniform convergence occurs almost surely).

Note that under some specific conditions, an exact characterization of the set of optimal menus is known. For instance, if $v(\theta, x) = \theta \cdot x$ and $c(x) = \overline{c} \cdot x$, it is well-known that any optimal mechanism is F-almost everywhere equal to an indicator function $\mathbf{1}_{\{p^* \geq \theta\}}$, where $p^* \in \arg\max_{p \in \text{supp}(F)} (p - \overline{c}) \cdot \int \mathbf{1}_{\{p \geq \theta\}} dF(\theta)$. Such an explicit solution simplifies the problem of characterizing empirically optimal mechanisms dramatically. When, instead, v is multiplicatively separable, and F_0 is absolutely continuous and has convex support, any optimal mechanism is almost everywhere equal to pointwise maximization of the ironed virtual value as shown in Toikka (2011). Hence, the problem of characterizing empirically optimal mechanisms can be made computationally tractable by ensuring not only consistency, but also absolutely continuity and convex support of estimates $\hat{F}(S^n)$. Although the empirical cumulative distribution is not absolutely continuous, one such estimator $\hat{F}$ can be easily obtained by adopting any smooth interpolation of the empirical cumulative distribution, for example a linear interpolation, a cubic spline or an interpolation relying on Bernstein polynomials.[7]

3.4.2 Asymptotic Optimality

Having defined our class of empirically optimal mechanisms, we establish in this section that they are asymptotically optimal: the realized expected profit given the mechanism converges in probability to the optimal expected profit as the sample size grows.

Such convergence is not guaranteed for arbitrary sample-based mechanisms, even for those that have desirable robustness properties, as it requires that the sample-based mechanism makes full use of the sample. For instance, if the mechanism relies only on statistics such as estimates for a finite number of moments or the support of the distribution, then it is immediate that it will not, in general, converge in probability to the optimal expected profit. Relying on an estimator for the true distribution itself (an infinite-dimensional parameter) is then key to obtaining asymptotic optimality.

We start by making an important observation:

Lemma 7. For any $M \in \mathscr{M}$, $\pi(M, F)$ is Lipschitz continuous in $F \in \mathbb{F}$, with a Lipschitz constant L that does not depend on M.

We defer the proof to the appendix, but highlight the main steps here. The proof of Lemma 7 first makes use of the revelation principle to focus on the elements of any arbitrary menu that are payoff relevant, as these are given by a bounded non-decreasing function, and hence of bounded

[7] In the appendix, we provide a simple proof for the fact that linear interpolations retain the uniform convergence (and therefore consistency) properties of the empirical distribution. See Babu, Canty, and Chaubey (2002) and Leblanc (2011) for details on interpolation of the empirical cumulative distribution using Bernstein polynomials.

variation. Then, we appeal to a result that provides an upper bound on Riemann–Stieltjes integrals of functions of bounded variation to obtain the result. In particular, we find a Lipschitz constant of at most

$$L = 2\left(v(\bar{\theta}, \bar{x}) + (\bar{\theta} - \underline{\theta}) \cdot \max_{\theta' \in \Theta} v_1(\theta', \bar{x}) + c(\bar{x}) \right).$$

For every $F \in \mathbb{F}$, define the firm's value function as

$$\Pi(F) := \sup_{M \in \mathscr{M}} \pi(M, F).$$

Lemma 7 leads to a further result, this time regarding (Lipschitz) continuity of the value function:

Lemma 8. Π is Lipschitz continuous, with Lipschitz constant L.

Proof: For any $F, G \in \mathbb{F}$,

$$\begin{aligned} |\Pi(F) - \Pi(G)| &= \left| \sup_{M \in \mathscr{M}} \pi(M, F) - \sup_{M \in \mathscr{M}} \pi(M, G) \right| \\ &\leq \sup_{M \in \mathscr{M}} |\pi(M, F) - \pi(M, G)| \leq L \cdot \|F - G\|_\infty \qquad \square \end{aligned}$$

Finally, the desired result of asymptotic optimality of our class of empirically optimal mechanisms follows immediately:

Proposition 8. *Let $\hat{M}^*$ be an empirically optimal mechanism given by $\hat{M}^* = M^* \circ \hat{F}$. Then, $|\pi(\hat{M}^*(S^n), F_0) - \Pi(F_0)| \xrightarrow{p} 0$.*

Proof: By Lemmas 7 and 8,

$$\begin{aligned} |\pi(\hat{M}^*(S^n), F_0) - \Pi(F_0)| &\leq |\pi(\hat{M}^*(S^n), F_0) - \pi(\hat{M}^*(S^n), \hat{F}(S^n))| + |\pi(\hat{M}^*(S^n), \hat{F}(S^n)) - \Pi(F_0)| \\ &\leq L \cdot \|\hat{F}(S^n) - F_0\|_\infty + |\Pi(\hat{F}(S^n)) - \Pi(F_0)| \leq 2L \cdot \|\hat{F}(S^n) - F_0\|_\infty \xrightarrow{p} 0. \qquad \square \end{aligned}$$

Proposition 8 provides a simple justification for using an empirically optimal mechanism to guide the firm's pricing strategy: As the sample size grows large, such sample-based mechanisms deliver an expected profit close to the optimal one. We stress the minimal informational assumptions made. In particular, this result does not depend on the firm knowing the support of the true distribution F_0 ex ante, as the empirically optimal mechanism is defined making use only of the estimated cumulative distribution and there are consistent estimators that require no assumptions on (and in fact, asymptotically learn) the support of F_0. Moreover, as $\Pi(F_0) - \pi(\hat{M}^*(S), F_0) \leq 2L\|\hat{F}(S) - F_0\|_\infty$, we conclude that empirically optimal mechanisms are robust in a sense akin to Bergemann and Schlag (2011), since for any $\varepsilon > 0$, samples inducing $\|\hat{F}(S) - F_0\|_\infty < \varepsilon$ imply that $\Pi(F_0) - \pi(\hat{M}^*(S^n), F_0) \leq 2L\varepsilon$.

3.4.3 Robustness Properties

While some sample-based implementations of existing robust mechanisms would not be asymptotically optimal, it is possible that others would. Thus, an obvious question is: Do empirically optimal mechanisms provide robustness guarantees with finite samples that render them especially appealing? In this section, we argue that this is indeed the case.

Robustness properties of mechanisms regard the worst-case scenarios. The existing literature has focused on two main properties. One corresponds to the worst-case profit that the firm can expect given that the true distribution lies in a specific set $A \subseteq \mathscr{F}$. This is in part motivated by appealing to the characterization of preferences exhibiting ambiguity aversion by Gilboa and Schmeidler (1989b), which entails a maxmin representation, whereby the decision-maker (here, the firm) evaluates each act (mechanism) by assuming the worst-case payoff. The robustness of a specific mechanism according to this criterion is then given by the lower bound on expected profit it can attain, $\min_{F \in A} \pi(M, F)$. The second robustness criterion that has been considered in the literature depends on the notion of regret: How much profit the firm may be forgoing by committing to mechanism M when the true distribution is F_0, that is, $R(M, F_0) := \Pi(F_0) - \pi(M, F_0)$.

A simple implication of Lemmas 7 and 8 is that we can immediately obtain probabilistic bounds on regret and on how far the realized expected profit may be from the profit the firm expects to obtain given its estimated distribution.

Proposition 9. *Let $\hat{M}^* = M^* \circ \hat{F}$ be an empirically optimal mechanism. Suppose that $\hat{F} \in \widehat{\mathbb{F}}$ is such that $\forall S^n \in \mathscr{S}$, $\mathbb{P}(\|\hat{F}(S^n) - F_0\|_\infty > \delta) \leq p(n, \delta)$ for some function $p : \mathbb{N} \times \mathbb{R}_+ \longrightarrow [0, 1]$. Then,*

(i) $\mathbb{P}\left(|\pi(\hat{M}^*(S^n), \hat{F}(S^n)) - \pi(\hat{M}^*(S^n), F_0)| > \delta\right) \leq p(n, \delta/L)$*; and*

(ii) $\mathbb{P}\left(R(\hat{M}^*(S^n), F_0) > 2\delta\right) \leq p(n, \delta/L)$*,*

where L denotes the Lipschitz constant from Lemma 7.

While Proposition 9 is a trivial observation, it enables the firm to obtain strong, non-asymptotic probabilistic bounds on both profit and regret whenever basing an empirically optimal mechanism on an estimator $\hat{F}$ with specific properties. Whenever the firm knows an upper bound $\bar{\theta}$ for the support of the true distribution, L can be obtained in a way that depends exclusively on known constants and these probabilistic profit and regret guarantees can be computed explicitly. The next two examples illustrate how this result can be applied by focusing on estimators $\hat{F}$ with well-known properties, such as the empirical cumulative distribution and smooth interpolations of it.

Example 9: Let $\hat{M}^* = M^* \circ \hat{F}$ be any empirically optimal mechanism such that $\hat{F}$ denotes the cumulative distribution estimator. By the Dvoretzky–Kiefer–Wolfowitz inequality (Dvoretzky, Kiefer, and Wolfowitz, 1956) with Massart's (1990) constant, we have that

$$\mathbb{P}(\|\hat{F}_n - F_0\|_\infty > \delta) \leq 2\exp\left(-2n\delta^2\right)$$

and hence $p(n, \delta) = 2\exp\left(-2n\delta^2\right)$. Then, Proposition 9 applies and we can obtain regret lower than 2δ with probability of at least $1 - p(n, \delta)$ and a confidence bound with range 2δ such that the true expected profit differs from $\pi(\hat{M}^*(S^n), \hat{F}(S^n))$ by less than δ also with probability greater than $1 - p(n, \delta)$.

Example 10: Suppose that $\mathscr{F}$ is restricted to the set of absolutely continuous distributions on Θ, of which F_0 is known to be an element of, and that v is multiplicatively separable in $\theta \in \Theta$ and $x \in X$. If we were to constrain $\hat{F}(S)$ to also be absolutely continuous, admitting a strictly positive density and having convex support, an analytic characterization of $M^*(\hat{F}(S))$ is known, given by pointwise maximization of the ironed virtual value given $\hat{F}(S)$.[8] This would then simplify the computational cost of finding the optimal mechanism. Take $\hat{F}$ to be any interpolation of the empirical cumulative distribution that results in a valid distribution function that is absolutely continuous and has convex support, such as the linear interpolation (see Lemma 27 in the appendix). Note that given that F_0 is atomless by assumption, $\mathbb{P}\left(\exists k, \ell \in \{0, 1, ..., n-1\} \, : \, s_k = s_\ell\right) = 0 \; \forall n \in \mathbb{N}$, and thus the linear interpolation is well-defined with probability 1. Furthermore, for any such interpolation $\hat{F}$, with probability 1, the estimate given by $\hat{F}(S^n)$ differs from the empirical cumulative distribution $\hat{F}^E(S^n)$ by at most $1/n$ at any given point. Hence,

$$\begin{aligned}
\mathbb{P}(\|\hat{F}(S^n) - F_0\|_\infty > \delta) &\leq \mathbb{P}(\|\hat{F}^E(S^n) - F_0\|_\infty + \|\hat{F}(S^n) - \hat{F}^E(S^n)\|_\infty > \delta) \\
&\leq \mathbb{P}(\|\hat{F}^E(S^n) - F_0\|_\infty > \delta - 1/n) \\
&\leq 2\exp\left(-2n(\delta - 1/n)^2\right),
\end{aligned}$$

where, again, the last inequality follows from the Dvoretzky–Kiefer–Wolfowitz inequality with Massart's constant. It follows that $p(n, \delta) = 2\exp\left(-2n(\delta - 1/n)^2\right)$. As $\hat{F} \in \widehat{\mathscr{F}}$, Proposition 9 applies and the regret and confidence bounds obtain.

As this next example shows, under some assumptions on the true distribution F_0 one can even obtain not only non-asymptotic, but also non-probabilistic regret and confidence bounds.

Example 11: Suppose that it is known that the true distribution F_0 admits a density f_0 with total variation bounded by $B < \infty$ and has support contained in $[0, 1]$.[9] Let V_B denote the set of all densities on $[0, 1]$ with total variation bounded by B. Take any kernel density estimator, given by

$$\hat{f}(S^n)(u) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{u - \theta_i}{h}\right),$$

[8] Toikka (2011) has shown that for any absolutely continuous distribution F on Θ with a strictly positive density, the set of maximizers $\mathscr{M}^*(F)$ pointwise maximize $\bar{J}(\theta)v(\theta, x) - c(x)$, where $\bar{J}(\theta)$ is the ironed version of $J(\theta) := \theta - \frac{1-F(\theta)}{f(\theta)}$.

[9] This can be generalized to some closed interval in $\mathbb{R}$.

where h is a smoothing parameter such that $h \to 0$ as $n \to \infty$, θ_i denotes the value of observation i, $K \geq 0$ and $\int K(u)du = 1$. From theorem 3 in Datta (1992), we know that, for any such kernel density estimator $\hat{f}_n$, one has

$$\sup_{f \in V_B} \|\hat{f}(S^n) - f\|_1 \leq (2B+1)k_1 h + \left(\frac{k_2}{nh}\right)^{1/2},$$

where $k_1 := \int |u|K(u)du$ and $k_2 := \int K^2(u)du$. Define $q(n,h) := (2B+1)k_1 h + \left(\frac{k_2}{nh}\right)^{1/2}$ and let $\hat{F}(S^n)(\theta) := \int \mathbf{1}_{u \leq \theta} \hat{f}(S^n)(u)du$ denote the estimated cumulative distribution.

Note that there are several kernel density estimators available such that

$$k_1 = \int |u|K(u)du < \infty \text{ and } k_2 = \int K^2(u)du < \infty,$$

e.g. if K is the uniform kernel, triangle, Epanechnikov, among others.

In order for $\hat{F}$ to belong to $\widehat{\mathscr{F}}$ we need for it to (1) be uniformly consistent and (2) to have compact support. There are several ways to achieve this, namely by relying on a kernel such that $\int_{|u|>\delta} K(u)du = 0$ for some finite $\delta > 0$ – which is satisfied by most of the standard kernels, namely the ones cited above. For any such kernel, $k_1, k_2 < \infty$ and $\forall S^n \in \mathscr{S}$, supp $\hat{F}(S^n) \subseteq \Theta$ is compact, by choosing lower and upper bounds $\underline{\theta}$, $\bar{\theta}$ appropriately.

Therefore,

$$\begin{aligned}\|\hat{F}(S^n) - F_0\|_\infty = \sup_{\theta \in \Theta} \left|\hat{F}(S^n)(\theta) - F_0(\theta)\right| &= \sup_{\theta \in \Theta} \left| \int \mathbf{1}_{u \leq \theta} \left(\hat{f}(S^n)(u) - f_0(u)ds\right)\right| \\ &\leq \sup_{\theta \in \Theta} \int \mathbf{1}_{u \leq \theta} |\hat{f}(S^n)(u) - f_0(u)|ds \leq \|\hat{f}(S^n) - f_0\|_1 \leq q(n,h),\end{aligned}$$

where $\|\cdot\|_1$ denotes the L^1 norm, i.e. $\|f\|_1 := \int |f(u)|du$. It follows immediately that if $h \to 0$ and $n \cdot h \to \infty$, $\|\hat{F}(S^n) - F_0\|_\infty \to 0$ and, consequently, $\hat{F} \in \widehat{\mathscr{F}}$. The argument above implies that

$$R(\hat{M}^*(S^n), F_0) \leq q(n,h)/2L;$$

$$|\pi(\hat{M}^*(S^n), \hat{F}(S^n)) - \pi(\hat{M}^*(S^n), F_0)| \leq q(n, h)/L.$$

As such, if f_0 has total variation bounded by B, we can obtain non-probabilistic bounds on regret and expected profit when implementing an empirically optimal mechanism based on kernel density estimation.

Other sample-based mechanisms could potentially yield even stronger robustness properties. However, for any sample-based mechanism M_S, one has

$$\begin{aligned} R(M_S(S^n), F_0) - R(\hat{M}^*(S^n), F_0) &= \pi(M_S(S^n), F_0) - \pi(\hat{M}^*(S^n), F_0) \\ &\leq \Pi(F_0) - \pi(\hat{M}^*(S^n), F_0) = R(\hat{M}^*(S^n), F_0). \end{aligned}$$

As the examples above show, choosing $\hat{F}$ appropriately ensures that $\mathbb{P}(R(\hat{M}^*(S^n), F_0) > \delta)$ declines exponentially with n. Hence, the gains in profit and regret guarantees from implementing alternative pricing policies are modest at best, in a formal sense.

To conclude this section, we note that Proposition 9 can instead be used to determine how many samples the firm requires in order to obtain specific robustness guarantees when relying on empirically optimal mechanisms. That is, another reading of Proposition 9 is that the firm only needs at most N samples – where N is the smallest integer such that $\alpha \geq p(N, \delta/L)$ – to secure at most 2δ of regret with probability $1-\alpha$. Alternatively, the same N samples provide a (conservative) confidence interval for profit with range 2δ with confidence level of $1 - \alpha$. This results in a non-asymptotic sample complexity bound for a specific class of sample-based mechanisms. In contrast to the sample-complexity bounds obtained in Huang, Mansour, and Roughgarden (2018), which pertain to the share of the optimal profit that the firm is able to secure, $R(M, F_0)/\Pi(F_0)$, we focus on bounding regret directly and our bounds are not asymptotic, that is, they hold for finite samples.

3.5 Inference and Robustness

While the revenue and regret guarantees derived in the previous section are useful, upon observing a particular sample and deciding on a menu, the firm may want to be able to obtain consistent projections for the expected profit. In this section, we show how to obtain consistent and unbiased estimates and conduct inference on the expected profit.

Our results enable unbiased and consistent estimation of and inference on the expected profit not only of empirically optimal mechanisms but of any given mechanism $M \in \mathscr{M}$. Moreover, we show how empirically optimal mechanisms serve a special purpose, in that they can be used to estimate and conduct inference on the optimal expected profit. With these tools, one can provide confidence intervals for the expected profit with specific asymptotic coverage for any one mechanism, calculate probabilistic bounds for regret, or test whether one mechanism yields higher expected profit than another.

3.5.1 Expected Profit

An immediate consequence of Lipschitz continuity of the firm's profit function with respect to the distribution is that, for any mechanism and any consistent estimator of the distribution of types, one can consistently estimate the expected profit that such a mechanism would generate.

Proposition 10. *For any true distribution $F_0 \in \mathscr{F}$, consistent estimator $\hat{F} \in \widehat{\mathscr{F}}$, and mechanism $M \in \mathscr{M}$, we have $\pi(M, \hat{F}(S^n)) \xrightarrow{p} \pi(M, F_0)$. Moreover, if $\hat{F}$ is an unbiased estimator such as the empirical distribution function, $\mathbb{E}\left[\pi(M, \hat{F}(S^n))\right] = \pi(M, F_0)$, that is, the plug-in estimator is also unbiased.*

Proof: By Lemma 7, we have that, $\forall \hat{F} \in \widehat{\mathscr{F}}$ and $\forall M \in \mathscr{M}$,

$$|\pi(M, \hat{F}(S^n)) - \pi(M, F_0)| \leq L\|\hat{F}(S^n) - F_0\|_\infty \xrightarrow{p} 0 \qquad \square$$

Moreover, if $\hat{F}$ is an unbiased estimator of F_0, by linearity of $\pi(M, F)$ in F, we have that

$$\mathbb{E}\left[\pi(M, \hat{F}(S^n))\right] = \pi(M, \mathbb{E}\left[\hat{F}(S^n)\right]) = \pi(M, F_0).$$

An important aspect in estimation is the ability to conduct inference. For instance, the firm could be interested in using statistical inference in order to compare different mechanisms, that is, to test whether a specific mechanism would deliver a higher expected profit than another. Another possible application would be to obtain valid confidence intervals for the expected profit under a particular mechanism, as it is an arguably crucial tool for the development of routine business activities such as drawing up budgets under different scenarios, with varying degrees of confidence.

While the firm could potentially derive confidence intervals for the expected profit by adjusting Proposition 9 and example 9 to the mechanism it is considering, these bounds would exhibit two drawbacks when used for this purpose. First, they require knowledge of $\bar{\theta}$, the upper bound on the type distribution. Second, and more critically, they generally do not provide the correct asymptotic coverage, that is, they would be exceedingly conservative.

In order to address these drawbacks, we suggest a simple estimation procedure that does yield asymptotically valid inference. First, we focus on the empirical distribution function as our estimator, since it admits a functional central limit theorem (by Donsker's theorem) when properly centered and rescaled. Then, because profit is linear and continuous in the type distribution, it is Fréchet differentiable, with derivative given by $\dot{\pi}_M(\cdot) = \pi(M, \cdot)$.[10] One can thus obtain the asymptotic distribution of our consistent estimator for the expected profit $\pi(M, \hat{F}(S^n))$ by appealing to a simple functional Delta method result.

Theorem 9. *Let $\hat{F}$ denote the empirical distribution estimator. Then, $\forall M \in \mathscr{M}$, $\forall F_0 \in \mathscr{F}$,*

$$\sqrt{n}\left(\pi(M, \hat{F}(S^n)) - \pi(M, F_0)\right) \xrightarrow{d} N(0, \sigma^2_{M,F_0}),$$

[10] The Fréchet derivative $\dot{\pi}_M$ is defined on the space of functions on Θ of bounded variation endowed with the supremum-norm. We refer to the appendix for details.

where $\sigma^2_{M,F_0} := \mathbb{E}\left[(\dot{\pi}_M(\delta_\theta - F_0))^2\right] = \mathbb{E}\left[(\pi(M, \delta_\theta - F_0))^2\right]$ *and* $\theta \sim F_0$.

The function δ_θ denotes the cumulative distribution associated with a Dirac measure at θ, that is, $\delta_\theta(x) = \mathbf{1}_{\{x \geq \theta\}}$.

Theorem 9 states that the distribution of the empirical process,

$$G_n := \sqrt{n}\left(\pi(M, \hat{F}(S^n)) - \pi(M, F_0)\right),$$

converges weakly to $N(0, \sigma^2_{M,F_0})$. A question then arises of how to estimate, in practice, the asymptotic distribution in a consistent manner, as it depends on the unknown distribution F_0. We provide two alternatives. One option is the use of a plug-in estimator for σ^2_{M,F_0}. This can be done directly – as the functional dependence of σ^2_{M,F_0} on F_0 is known and a consistent estimate for F_0 is readily available –, or by following other plug-in methods as those in Shao (1993). Another option is to rely on the classical bootstrap to approximate the distribution of G_n by the distribution of $\hat{G}_n := \sqrt{n}\left(\pi(M, \hat{F}(S^n_B)) - \pi(M, \hat{F}(S^n))\right)$, conditional on S^n, where S^n_B denotes the resampling of n observations from S^n with uniform weights. That this approach does in fact consistently estimate the limiting distribution was shown by Parr (1985b, Theorem 4).

We note that other bootstrap methods would also yield consistent estimates, such as subsampling (bootstrap without replacement) (Politis and Romano, 1994) or Jackknife procedures (Parr, 1985a). Moreover, given that the Fréchet derivative of profit is sufficiently well-behaved, under some smoothness assumptions on the true distribution F_0, smoothed versions of the bootstrap (Cuevas and Romo, 1997) can also be considered.

3.5.2 Optimal Profit and Regret

We now extend our statistical inference results, highlighting how empirically optimal mechanisms can be used to provide a consistent and asymptotically normal estimator for optimal profit.

Given that the firm's value function is Lipschitz continuous in the distribution and that empirically optimal mechanisms attain the optimal profit for a consistent estimate of the true distribution, it is easy to see that one can then use them as a tool to consistently estimate the optimal profit that the firm would obtain, were it to know F_0. We formalize this observation as follows:

Proposition 11. *For any true distribution $F_0 \in \mathscr{F}$ and empirically optimal mechanism $\hat{M}^*$ given by $\hat{M}^* = M^* \circ \hat{F}$, $\pi(\hat{M}^*(S^n),\ \hat{F}(S^n)) \xrightarrow{p} \Pi(F_0)$.*

Proof: Similarly to Proposition 10, we again have that, by Lemma 8, Π is Lipschitz continuous and therefore,

$$|\pi(\hat{M}^*(S^n), \hat{F}(S^n)) - \Pi(F_0)| = |\Pi(\hat{F}(S^n)) - \Pi(F_0)| \leq L\|\hat{F}(S^n) - F_0\|_\infty \xrightarrow{p} 0. \qquad \square$$

It is less straightforward that one could take an approach to conducting inference on the optimal profit similar to that derived for fixed mechanisms. Specifically, this would require proving that the firm's value function Π is also Fréchet differentiable.[11] We confirm that indeed such an approach is valid by proving an interesting technical result in this generalized Maskin-Riley setup: an envelope theorem for the firm's value function. In other words, our next result shows that the value function is Fréchet differentiable at any distribution $F \in \mathscr{F}$ and that its Fréchet derivative coincides with that of the expected profit at F with the optimal menu for F.

Theorem 10 (Envelope Theorem). *Π is Fréchet differentiable at all $F \in \mathscr{F}$. Moreover, its Fréchet derivative at F is given by $\dot{\Pi}_F = \dot{\pi}_{M_F}$, $\forall M_F \in \mathscr{M}^*(F)$.*

Then, defining the empirical process $\hat{G}_n := \sqrt{n}\left(\Pi(\hat{F}(S^n_B)) - \Pi(\hat{F}(S^n))\right)$, conditional on S^n, an adapted version of Theorem 9 ensues:

Theorem 11. *Let $\hat{F}$ denote the empirical distribution estimator. Then, $\forall F_0 \in \mathscr{F}$,*

$$\sqrt{n}\left(\Pi(\hat{F}(S^n)) - \Pi(F_0)\right) \xrightarrow{d} N(0, \sigma^2_{F_0}),$$

[11] In fact, the now standard functional Delta method requires only the weaker notion of Hadamard differentiability; see, e.g., Vaart and Wellner (1996, ch. 3.9). However, the stronger notion of Fréchet differentiability has the benefit of allowing us to bypass the measurability complications that arise when using weaker notions.

where $\sigma^2_{F_0} = \mathbb{E}\left[\left(\dot{\Pi}_{F_0}(\delta_\theta - F_0)\right)^2\right] = \mathbb{E}\left[(\pi(M_F, \delta_\theta - F_0))^2\right]$ *and* $\theta \sim F_0$*. Moreover,* $\hat{G}_n \xrightarrow{d} N(0, \sigma^2_{F_0})$.

In this case, opposite to the case of inference under a fixed mechanism, we do not have a valid (consistent) plug-in estimator for $\sigma^2_{F_0}$, which depends on F_0 and, more problematically, also on an optimal mechanism under the distribution F_0, $M_0 \in \mathscr{M}^*(F_0)$. An important argument in favor of a bootstrap approach to estimating the asymptotic distribution in this case is that it bypasses this issue.

Under some conditions, it may make sense to rely on different estimators. For instance, as discussed in example 10, when $v(\theta, x)$ is multiplicatively separable and F_0 is known to be absolutely continuous and with compact and convex support, the functional form of the solution is exactly known. This allows for a drastic simplification of the problem from a computational point of view, since it dispenses with the hurdle of finding the optimal mechanism for a given distribution. Especially in the context of implementing a bootstrap approach, the gains can be substantial. However, an estimate of the ironed virtual value depends on a suitable estimate of the density f. Therefore, we find it especially relevant that, when F_0 is known to be absolutely continuous, one can use as an estimator the simple linear interpolation of the empirical distribution discussed earlier to obtain a bootstrap estimator for the asymptotic distribution. Further, we note that this extremely simple approach is not only consistent for the true distribution F_0, but also for its density.

Proposition 12. *Let* $\hat{F}$ *denote the linear interpolation of the empirical distribution estimator. Then, for any absolutely continuous* $F_0 \in \mathscr{F}$*,*

(1) $\sqrt{n}\left(\Pi(\hat{F}(S^n)) - \Pi(F_0)\right) \xrightarrow{d} N(0, \sigma^2_{F_0})$*, where* $\sigma^2_{F_0} = \mathbb{E}\left[\left(\dot{\Pi}_{F_0}(\delta_\theta - F_0)\right)^2\right]$*;*

(2) $\hat{G}_n \xrightarrow{d} N(0, \sigma^2_{F_0})$*; and*

(3) $\|\hat{f}(S^n) - f_0\|_1 \xrightarrow{p} 0$*, where* $\hat{f}(S^n)$ *and* f_0 *denote the Radon-Nikodym derivatives of* $\hat{F}(S^n)$ *and* F_0*, respectively.*

While estimating the optimal expected profit may be relevant for investment decisions, as it provides an upper bound on the return of a given investment, these results can also be used to estimate regret. As regret, $R(M,F)$, is given by $R(M,F) = \Pi(F) - \pi(M,F)$, it is Fréchet differentiable at any distribution $F \in \mathscr{F}$, for any fixed $M \in \mathscr{M}$, as the sum of Fréchet differentiable functionals is itself Fréchet differentiable. Then, using similar arguments as those in Proposition 11 and Theorem 11, one can conduct inference on regret. Consequently, for any mechanism $M \in \mathscr{M}$, one can not only obtain asymptotically valid probabilistic bounds for expected profit, but also for regret.

3.5.3 Simulation Evidence

To conclude this section, we present empirical evidence on the finite sample properties of our estimators.

We conduct Monte Carlo simulations on the empirical coverage of the confidence intervals for expected profit under a fixed mechanism – uniform pricing, with the price set at $1/2$ – and for the optimal expected profit, using empirically optimal mechanisms for the empirical distribution. We use the approximation obtained by classic bootstrapping (N out of N), which we have showed to be asymptotically valid. We show simulation results for confidence levels $\alpha \in \{.1, .05, .01\}$, with varying sample size N. For each sample size, we draw $1,000$ samples and, for each sample, we estimate the confidence interval by drawing $1,000$ bootstrap samples from the original sample.

We focus on the case where consumers have quasilinear-linear utility and the unit cost is normalized to zero, as in Bergemann and Schlag (2011) and Carrasco et al. (2018a). We show results for three different parameterizations of F_0 relying on the Beta distribution: Beta(1/4,1/4), Uniform(0,1) and Beta(4,4).[12]

In table 3.1 we present evidence for the empirical coverage frequency at sample sizes of 500, 1,000 and 2,500. As is immediate upon inspection of the table, our estimators have extremely good finite sample properties, with the empirical coverage frequencies being very close to the theoretical asymptotic coverage probability, regardless of which of the three distributions is considered. We also investigated the behavior of our estimators under small samples. As fig. 3.1 shows, they fare

[12] The empirical coverage results were consistent across other parameterizations and using Beta mixtures or mixtures with degenerate distributions.

Table 3.1: Empirical Coverage Frequencies

(a) Profit with Fixed Mechanism

		Beta(1/4,1/4)			Unif(0,1)			Beta(4,4)		
	$1-\alpha$	.90	.95	.99	.90	.95	.99	.90	.95	.99
	500	.891	.944	.983	.892	.946	.987	.895	.943	.988
N	1,000	.901	.934	.985	.901	.946	.986	.896	.953	.985
	2,500	.909	.954	.989	.903	.953	.989	.889	.948	.985

(b) Optimal Profit

		Beta(1/4,1/4)			Unif(0,1)			Beta(4,4)		
	$1-\alpha$	.90	.95	.99	.90	.95	.99	.90	.95	.99
	500	.881	.933	.989	.894	.945	.983	.895	.950	.987
N	1,000	.886	.945	.985	.888	.943	.981	.894	.941	.984
	2,500	.910	.952	.988	.890	.941	.988	.884	.933	.984

Note: This table shows the frequency with which the estimated confidence interval with asymptotic coverage of $1-\alpha$ contained the true expected profit, $\pi(M, F_0)$ in the case of a fixed mechanism and $\Pi(F_0)$ in the case of the optimal expected profit. The fixed mechanism corresponds to uniform pricing at $1/2$. The estimated confidence interval followed a centered bootstrap procedure with 1,000 samples redrawn with replacement from the original sample with 1,000 iterations, for each sample size N.

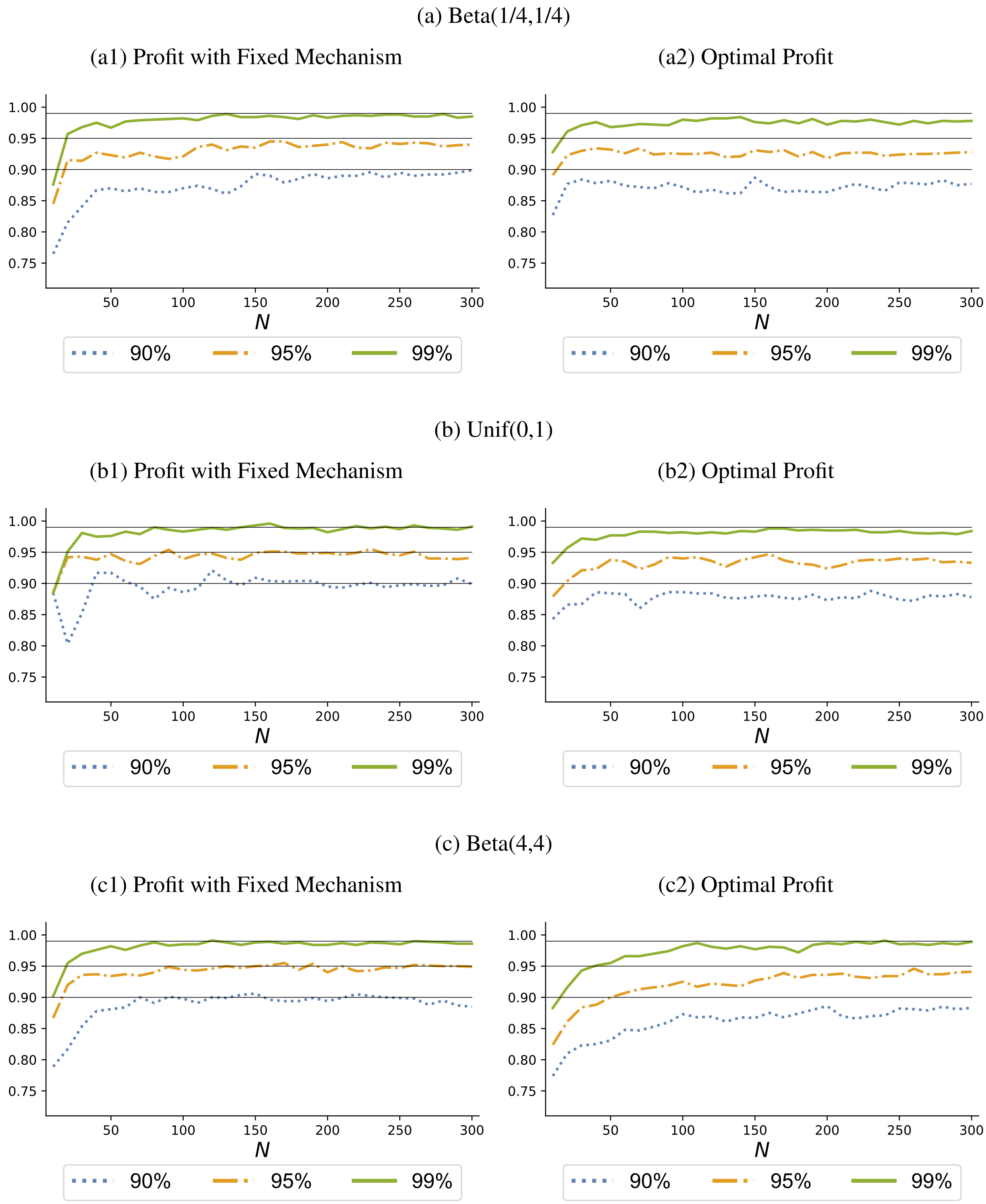
reasonably well for sample sizes between 50 and 300.

We also considered the regret incurred by adopting an empirically optimal mechanism that depends on the empirical distribution with finite samples. As illustrated in fig. 3.2, we empirically study the average regret as a share of the optimal expected profit, that is,

$$\left(\Pi(F_0) - \pi(\hat{M}^*(S^n), F_0)\right) / \Pi(F_0).$$

The average is taken across 1,000 samples of varying size in increments of 10 observations. Even with just 50 observations, the empirically optimal mechanism on average attains regret that is under 4% of the optimal expected profit. For the purpose of comparison, the robust mechanism in Carrasco et al. (2018a, Section 5.1), relying on an estimate of the mean and assuming knowledge of the upper bound of the distribution, exhibits average regret no lower than 20% of the optimal expected profit under any of the three distributions we consider.[13]

[13] We observe that for the minimax regret distribution derived in Carrasco et al. (2018a), by construction, the empirically optimal mechanism will attain the optimal profit with probability one and regardless of the number of samples, and, therefore, also attain the minimal regret.



(a) Beta(1/4,1/4)

(a1) Profit with Fixed Mechanism

(a2) Optimal Profit

(b) Unif(0,1)

(b1) Profit with Fixed Mechanism

(b2) Optimal Profit

(c) Beta(4,4)

(c1) Profit with Fixed Mechanism

(c2) Optimal Profit

Figure 3.1: Empirical Coverage Frequencies

Note: This figure shows the frequency with which the estimated confidence interval with asymptotic coverage of $1 - \alpha = .9,\ .95,\ .99$ contained the true expected profit, $\pi(M, F_0)$ in the case of a fixed mechanism and $\Pi(F_0)$ in the case of the optimal expected profit. The procedure is as described in the note to table 3.1. Sample size N varies between 10 and 300 with increments of 10 observations. The fixed mechanism corresponds to uniform pricing at 1/2.

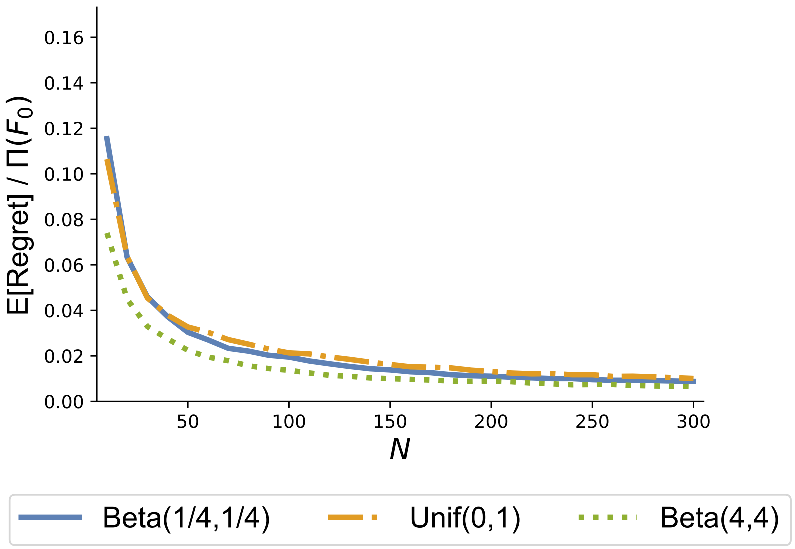


Figure 3.2: Regret of the Empirically Optimal Mechanism as a share of Optimal Profit

Note: This figure shows the average regret of the empirically optimal mechanism as a fraction of the optimal expected profit under the true distribution F_0. The average is taken over 1,000 samples for each sample size N between 10 and 300 with increments of 10 observations.

3.6 Extension to Single-Item Auctions

Before concluding, we discuss how to apply some of the insights developed in this paper to the related setting of single-unit auctions. In particular, we show how simple empirically optimal mechanisms in this context exhibit some of the desirable robustness features shown in Section 3.4.

Suppose that the firm has a single item to auction to $M \geq 2$ bidders. The firm values the item at $c > 0$, and each bidder $i = 1, ..., M$ is risk-neutral and values the item at θ_i, drawn independently from distribution F_0. To make matters simple, we assume that F_0 is absolutely continuous with convex and compact support. In such case, it is well-known that revenue equivalence holds and that a second-price auction with a reserve price is optimal for the firm, with bidders disclosing their types. Then, the optimal reserve price when the type distribution F satisfies the same assumptions solves

$$\max_{r \in \Theta} \pi(r, F),$$

where $\pi(r, F) = \int_r^{\bar{\theta}} dF_{(2;M)}$ and $F_{(2;M)}$ denotes the distribution of the second-highest willingness-to-pay, given a distribution of types F and M bidders. That is,

$$F_{(2;M)}(\theta) = M \cdot F(\theta)^{M-1}(1 - F(\theta)) + F(\theta)^M.$$

Consider the case where the firm has access to a sample of n observations drawn from F_0 and the reserve price is set before bids are submitted.[14] Similar to before, denote an empirically optimal reserve price $\hat{r}^*$ as the composition of a consistent estimator, $\hat{F}$, of the true distribution F_0, based on the realized sample S^n, and a selection from the set of reserve prices that are optimal for F, r^*.

The next proposition provides an analogue of Propositions 8 and 9 to this specific setting:

Proposition 13. *Let $\hat{r}^*$ be an empirically optimal reserve price given by $\hat{r}^* = r^* \circ \hat{F}$. Then, $|\pi(\hat{r}^*(S^n), F_0) - \Pi(F_0)| \xrightarrow{p} 0$. Moreover, if $\hat{F} \in \widehat{\mathbb{F}}$ is such that $\forall S^n \in \mathscr{S}$, $\mathbb{P}(\|\hat{F}(S^n) - F_0\|_\infty > \delta) \leq p(n, \delta)$ for some function $p : \mathbb{N} \times \mathbb{R}_+ \longrightarrow [0, 1]$, then,*

(i) $\mathbb{P}\left(|\pi(\hat{r}^*(S^n), \hat{F}(S^n)) - \pi(\hat{r}^*(S^n), F_0)| > \delta\right) \leq p(n, \delta/L)$; and

(ii) $\mathbb{P}\left(R(\hat{r}^*(S^n), F_0) > 2\delta\right) \leq p(n, \delta/L)$,

where $L = 2M(M - 1)$ and $\Pi(F) := \sup_{r \in \Theta} \pi(r, F)$.

The key insight is that the expected profit is linear in distribution of the second-order statistic and that this in turn is Lipschitz continuous in the distribution of types, $\|F_{(2;M)} - G_{(2;M)}\|_\infty \leq 2M(M - 1)\|F - G\|_\infty$. For $\hat{r}^*$ to be empirically optimal, though, we must have that $\hat{F}(S^n)$ is absolutely continuous and has convex and compact support. Similarly to example 10, when $\hat{F}$ is the linearly interpolated empirical distribution we have that $p(n, \delta) = 2\exp(-2n(\delta - 1/n)^2)$, delivering regret and confidence bounds.[15]

Our results on the properties on the robustness of empirically optimal mechanisms extend naturally to auction settings as this application illustrates. There is, however, a natural limitation in

[14] The same arguments apply when the reserve price is secret and takes into account the bids submitted, as these would just translate into a larger sample of $n + M$ observations.

[15] Cole and Roughgarden (2014) provide alternative sample-complexity bounds for this problem, characterizing the asymptotic number of samples needed to achieve $(1 - \epsilon)$ share of the optimal profit.

extending our results on inference: expected profit is linear in the distribution of the second-order statistic, not in the distribution of types themselves.

3.7 Discussion

This paper has studied two separate but related questions. The first is how a firm should price when uncertain about the distribution of consumers' willingness-to-pay. The second is how to conduct inference regarding the expected profit, both under any fixed pricing strategy and for the optimal profit. When the firm has access to a sample of consumers' valuations, we showed that adopting an extremely simple approach – estimating the distribution using the sample and then pricing optimally for the estimated distribution – yields attractive robustness properties, in particular obtaining probabilistic lower bounds both for the expected profit and for regret. On the other hand, we provided a toolkit to conduct inference for the expected profit. This enables practitioners to obtain confidence intervals not only for expected profit but also, for example, for the difference in profit that two different mechanisms induce. More generally, this allows for a data-based approach to robust mechanism design, where robustness properties are inferred from available data.

Two important concerns that we have not discussed are how to obtain such a sample and whether a mechanism that trades-off experimentation and exploitation performs better when sampling comes at a cost. Under the assumption of myopic consumer behavior, such a sample can be procured by means of a survey. If consumers are assumed to be forward-looking, they may gain from misrepresenting their type, but incentive compatibility can be restored if the firm can preclude surveyed customers from purchasing the item in the future (or supply it to them for free in the case of unit demand). While under some conditions, optimal experimentation asymptotically attains the optimal profit (see e.g. Aghion et al., 1991), our results show that empirically optimal mechanisms do so as well, insofar as the sample acquired grows. Therefore, the firm's overall loss of revenue will depend solely on the convergence rate of its estimator.

A different issue is how to optimality elicit types (i.e., generate a sample) through revealed choices. If consumers' types are not knowable but through their choices – for instance, consumers may not be aware of their types – the firm could conduct market research and elicit such a sample by means of a mechanism that induces each type to self-select to a different pair of quantity and price. This market research can then lead to the original sample or expand an existing sample – in which case a cost-benefit analysis on the net value of acquiring additional observations may come into play. Although existence and feasibility of such mechanisms is immediate, exploring how to optimally conduct such sample elicitation endeavors may be an interesting avenue for future research.

References

Adams, Christopher P. (2016). “Finite mixture models with one exclusion restriction”. In: *Econometrics Journal* 19.2, pp. 150–165.

Agarwal, Nikhil and Paulo Somaini (2018). “Demand Analysis Using Strategic Reports: An Application to a School Choice Mechanism”. In: *Econometrica* 86.2, pp. 391–444.

Aghion, Philippe et al. (1991). “Optimal Learning by Experimentation”. In: *Review of Economic Studies* 58.4, p. 621.

Al-Najjar, Nabil I. and Luciano De Castro (2014). “Parametric representation of preferences”. In: *Journal of Economic Theory* 150.1, pp. 642–667.

Amarante, Massimiliano (2009). “Foundations of neo-Bayesian statistics”. In: *Journal of Economic Theory* 144.5, pp. 2146–2173.

Anscombe, F. J. and R. J. Aumann (1963). “A Definition of Subjective Probability”. In: *The Annals of Mathematical Statistics* 34.1, pp. 199–205.

Arrow, Kenneth Joseph (1971). *Essays in the theory of risk-bearing*. Markham Pub. Co, p. 278. ISBN: 9780841020016.

Athey, Susan and Philip A. Haile (2002). “Identification of standard auction models”. In: *Econometrica* 70.6, pp. 2107–2140.

Babu, G. Jogesh, Angelo J. Canty, and Yogendra P. Chaubey (2002). “Application of Bernstein Polynomials for smooth estimation of a distribution and density function”. In: *Journal of Statistical Planning and Inference* 105.2, pp. 377–392.

Battauz, Anna, Marzia De Donno, and Fulvio Ortu (2015). “Envelope theorems in Banach lattices and asset pricing”. In: *Mathematics and Financial Economics* 9.4, pp. 303–323.

Beesack, Paul R. (1975). “Bounds for Riemann-Stieltjes Integrals”. In: *The Rocky Mountain Journal of Mathematics* 5.1, pp. 75–78.

Bergemann, Dirk and Karl H. Schlag (2008). “Pricing without Priors”. In: *Journal of the European Economic Association* 6.2-3, pp. 560–569.

— (2011). “Robust monopoly pricing”. In: *Journal of Economic Theory* 146.6, pp. 2527–2543.

Berry, Steven, Michael Carnall, and Pablo T. Spiller (1996). “Airline Hubs: Costs, Markups and the Implications of Customer Hetergoeneity”. Cambridge, MA.

Bickel, Peter J. and David A. Freedman (1981). “Some Asymptotic Theory for the Bootstrap”. In: *The Annals of Statistics* 9.6, pp. 1196–1217.

Blackwell, David (1951). “Comparison of Experiments”. In: *Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probability*, pp. 93–102.

— (June 1953). “Equivalent Comparisons of Experiments”. In: *Annals of Mathematical Statistics* 24.2, pp. 265–272.

Blei, David M. (2012). “Probabilistic topic models”. In: *Communications of the ACM* 55.4, pp. 77–84.

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). “Latent Dirichlet Allocation”. In: *Journal of Machine Learning Research* 3, pp. 993–1022.

Caplin, Andrew and Mark Dean (2015). “Revealed preference, rational inattention, and costly information acquisition”. In: *American Economic Review* 105.7, pp. 2183–2203.

Caplin, Andrew and Daniel Martin (2015). “A testable theory of imperfect perception”. In: *Economic Journal* 125.582, pp. 184–202.

Carrasco, Vinicius et al. (2018a). “Optimal selling mechanisms under moment conditions”. In: *Journal of Economic Theory* 177, pp. 245–279.

Carrasco, Vinicius et al. (2018b). “Robust mechanisms: the curvature case”. In: *Economic Theory* 68.1, pp. 203–222.

Cerreia-Vioglio, S. et al. (2014). “Niveloids and their extensions: Risk measures on small domains”. In: *Journal of Mathematical Analysis and Applications* 413.1, pp. 343–360.

Cerreia-Vioglio, S. et al. (2011a). “Uncertainty averse preferences”. In: *Journal of Economic Theory* 146.4, pp. 1275–1330.

Cerreia-Vioglio, Simone et al. (2013). “Ambiguity and robust statistics”. In: *Journal of Economic Theory* 148.3, pp. 974–1049.

Cerreia-Vioglio, Simone et al. (2011b). “Rational preferences under ambiguity”. In: *Economic Theory* 48.2, pp. 341–375.

Cerreia-Vioglio, Simone et al. (2020). “Making Decisions under Model Misspecification”. In: *SSRN Electronic Journal*.

Chateauneuf, Alain et al. (2005). “Monotone continuous multiple priors”. In: *Economic Theory* 26.4, pp. 973–982.

Chen, Qihui and Zheng Fang (2019). “Improved inference on the rank of a matrix”. In: *Quantitative Economics* 10.4, pp. 1787–1824.

Chen, Xiaohong, Han Hong, and Denis Nekipelov (2011). “Nonlinear models of measurement errors”. In: *Journal of Economic Literature* 49.4, pp. 901–937.

Chen, Xiaohong, Han Hong, and Elie Tamer (2005). “Measurement error models with auxiliary data”. In: *Review of Economic Studies* 72.2, pp. 343–366.

Chen, Xiaohong, Yingyao Hu, and Arthur Lewbel (2008a). “A note on the closed-form identification of regression models with a mismeasured binary regressor”. In: *Statistics and Probability Letters* 78.12, pp. 1473–1479.

— (2008b). “Nonparametric identification of regression models containing a misclassified dichotomous regressor without instruments”. In: *Economics Letters* 100.3, pp. 381–384.

Cho, Jin Seo and Halbert White (2007). “Testing for regime switching”. In: *Econometrica* 75.6, pp. 1671–1720.

Ciliberto, Federico and Elie Tamer (2009). “Market Structure and Multiple Equilibria in Airline Markets”. In: *Econometrica* 77.6, pp. 1791–1828.

Clark, Stephen A. (2000). “Revealed preference and expected utility”. In: *Theory and Decision* 49.2, pp. 159–174.

— (1993). “Revealed preference and linear utility”. In: *Theory and Decision* 34.1, pp. 21–45.

Cole, Richard and Tim Roughgarden (2014). “The Sample Complexity of Revenue Maximization”. In: *Proceedings of the 46th Annual ACM Symposium on Theory of Computing - STOC '14*. ACM Press.

Compiani, Giovanni and Yuichi Kitamura (2016). “Using mixtures in econometric models: a brief review and some new results”. In: *Econometrics Journal* 19.3, pp. C95–C127.

Crémer, Jacques (1982). “A simple proof of Blackwell’s "comparison of experiments" theorem”. In: *Journal of Economic Theory* 27.2, pp. 439–443.

Cuevas, Antonio and Juan Romo (1997). “Differentiable Functionals and Smoothed Bootstrap”. In: *Annals of the Institute of Statistical Mathematics* 49.2, pp. 355–370.

Darst, Richard and Harry Pollard (1970). “An inequality for the Riemann-Stieltjes integral”. In: *Proceedings of the American Mathematical Society* 25.4, pp. 912–912.

Datta, Somnath (1992). “Some Nonasymptotic Bounds for L_1 Density Estimation using Kernels”. In: *The Annals of Statistics* 20.3, pp. 1658–1667.

Dean, Mark and Nathaniel Neligh (2023). “Experimental Tests of Rational Inattention”. In: *Journal of Political Economy (Forthcoming)* 0.ja, null.

Dixit, Avinash K. and Joseph E Stiglitz (1977). “Monopolistic Competition and Optimum Product Diversity”. In: *American Economic Review* 67.3, pp. 297–308.

Dvoretzky, Aryeh, Jack Kiefer, and Jacob Wolfowitz (1956). “Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator”. In: *The Annals of Mathematical Statistics* 27.3, pp. 642–669.

Einav, Liran, Mark Jenkins, and Jonathan Levin (2012). “Contract Pricing in Consumer Credit Markets”. In: *Econometrica* 80.4, pp. 1387–1432.

Ellsberg, Daniel (1961). “Risk, Ambiguity, and the Savage Axioms”. In: *Quarterly Journal of Economics* 75.4, pp. 643–669.

Epstein, Larry G. and Kyoungwon Seo (2010). “Symmetry of evidence without evidence of symmetry”. In: *Theoretical Economics* 5.3, pp. 313–368.

Fishburn, Peter C. (1970). *Utility Theory for Decision Making*. 1st. New York: Research Analysis Corporation. ISBN: 471260606.

Frühwirth-Schnatter, Sylvia (2013). *Finite Mixtures and Markov Switching Models*. 1st. Vol. 9781461460. New York: Springer, pp. 1–171. ISBN: 9781461460220.

Fu, Hu, Nima Haghpanah, and Jason Hartline (2020). “Full Surplus Extraction from Samples”. In: *Working Paper*, p. 33.

Gentzkow, Matthew and Emir Kamenica (2011). “Bayesian Persuasion”. In: *American Economic Review* 101.6, pp. 2590–2615.

Ghirardato, Paolo and Massimo Marinacci (2002). “Ambiguity made precise: A comparative foundation”. In: *Journal of Economic Theory* 102.2, pp. 251–289.

Gilboa, Itzhak and David Schmeidler (1989a). “Maxmin expected utility with non-unique prior”. In: *Journal of Mathematical Economics* 18.2, pp. 141–153.

— (1989b). “Maxmin expected utility with non-unique prior”. In: *Journal of Mathematical Economics* 18.2, pp. 141–153.

Gillis, Nicolas (2012). “Sparse and unique nonnegative matrix factorization through data preprocessing”. In: *Journal of Machine Learning Research* 13, pp. 3349–3386.

Guo, Chenghao, Zhiyi Huang, and Xinzhi Zhang (2020). “Sample complexity of single-parameter revenue maximization”. In: *ACM SIGecom Exchanges* 17.2, pp. 62–70.

Hansen, Lars Peter and Thomas J Sargent (2001). “Robust Control and Model Uncertainty”. In: *American Economic Review* 91.2, pp. 60–66.

Henry, Marc, Yuichi Kitamura, and Bernard Salanié (2014). “Partial identification of finite mixtures in econometric models”. In: *Quantitative Economics* 5.1, pp. 123–144.

Herden, G. (1989). “On the existence of utility functions”. In: *Mathematical Social Sciences* 17.3, pp. 297–313.

Hewitt, Edwin (Aug. 1946). “A remark on density characters”. In: *Bulletin of the American Mathematical Society* 52.8, pp. 641–643.

Hu, Yingyao (2008). “Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution”. In: *Journal of Econometrics* 144.1, pp. 27–61.

Huang, Kejun, Nicholas D. Sidiropoulos, and Ananthram Swami (2014). “Non-Negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition”. In: *IEEE Transactions on Signal Processing* 62.1, pp. 211–224.

Huang, Zhiyi, Yishay Mansour, and Tim Roughgarden (2018). “Making the Most of Your Samples”. In: *SIAM Journal on Computing* 47.3, pp. 651–674.

Inoue, Lurdes Y T (2009). *Decision theory: principles and approaches*. Ed. by Vic Barnett, J. Stuart Hunter, and Jozef L. Teugels. 1st. Vol. 47. Chichester: John Wiley & Sons, Inc., p. 417. ISBN: 9780471496571.

Jakobsen, Alexander M. (2021). “An Axiomatic Model of Persuasion”. In: *Econometrica* 89.5, pp. 2081–2116.

Ke, Shikun, José Luis Montiel Olea, and James Nesbit (2021). “Robust Machine Learning Algorithms for Text Analysis”.

Keane, Michael P. and Kenneth I. Wolpin (1997). “The career decisions of young men”. In: *Journal of Political Economy* 105.3, pp. 473–522.

Kim, Chang-Jin and Charles R. Nelson (Nov. 2018). *State-Space Models with Regime Switching*. 1st ed. Cambridge: The MIT Press. ISBN: 9780262277112.

Kreps, David (2018). *Notes On The Theory Of Choice*. New York: Routledge. ISBN: 9780813375533.

Kuzmics, Christoph (2017). “Abraham Wald’s complete class theorem and Knightian uncertainty”. In: *Games and Economic Behavior* 104, pp. 666–673.

Leblanc, Alexandre (2011). “On estimating distribution functions using Bernstein polynomials”. In: *Annals of the Institute of Statistical Mathematics* 64.5, pp. 919–943.

Lewbel, Arthur (2007). “Estimation of average treatment effects with misclassification”. In: *Econometrica* 75.2, pp. 537–551.

Maccheroni, Fabio, Massimo Marinacci, and Aldo Rustichini (2006). “Ambiguity aversion, robustness, and the variational representation of preferences”. In: *Econometrica* 74.6, pp. 1447–1498.

Machina, Mark J (1989). “Dynamic consistency and non-expected utility models of choice under uncertainty”. In: *Journal of Economic Literature* 27.4, pp. 1622–1668.

Madarász, Kristóf and Andrea Prat (2017). “Sellers with Misspecified Models”. In: *Review of Economic Studies* 84.2, pp. 790–815.

Mahajan, Aprajit (2006). “Identification and estimation of regression models with misclassification”. In: *Econometrica* 74.3, pp. 631–665.

Manski, Charles F. (2021). “Econometrics for Decision Making: Building Foundations Sketched by Haavelmo and Wald”. In: *Econometrica* 89.6, pp. 2827–2853.

Manski, Charles F and Aleksey Tetenov (2014). “The Quantile Performance of Statistical Treatment Rules using Hypothesis Tests to Allocate a Population to Two Treatments”. In: *Cemmap Working Paper*. Cemmap Working Paper Series CWP44/14.January.

Marschak, Jacob and Koichi Miyasawa (1968). “Economic Comparability of Information Systems”. In: *International Economic Review* 9.2, p. 137.

Marshall, Robert C. and Leslie M. Marx (2007). “Bidder collusion”. In: *Journal of Economic Theory* 133.1, pp. 374–402.

Maskin, Eric and John Riley (1984). “Monopoly with Incomplete Information”. In: *The RAND Journal of Economics* 15.2, p. 171.

Massart, P. (1990). “The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality”. In: *The Annals of Probability* 18.3, pp. 1269–1283.

Matêjka, Filip and Alisdair McKay (2015). “Rational inattention to discrete choices: A new foundation for the multinomial logit model”. In: *American Economic Review* 105.1, pp. 272–298.

McAffe, R.Preston and John McMillan (1992). “Bidding Rings”. In: *The American Economic Review* 82.3, pp. 579–599.

Milgrom, Paul and Ilya Segal (2002). “Envelope Theorems for Arbitrary Choice Sets”. In: *Econometrica* 70.2, pp. 583–601.

Mirrlees, J. A. (1971). “An Exploration in the Theory of Optimum Income Taxation”. In: *The Review of Economic Studies* 38.2, pp. 175–208.

Mussa, Michael and Sherwin Rosen (1978). “Monopoly and Product Quality”. In: *Journal of Economic Theory* 18.2, pp. 301–317.

Myerson, Roger B. (1981). “Optimal Auction Design”. In: *Mathematics of Operations Research* 6.1, pp. 58–73.

Neilson, William S. (2010). “A simplified axiomatic approach to ambiguity aversion”. In: *Journal of Risk and Uncertainty* 41.2, pp. 113–124.

Niculescu, Constantin P. and Lars-Erik Persson (2018). *Convex Functions and Their Applications*. Springer International Publishing.

Nishimura, Hiroki, Efe A. Ok, and John K.H. Quah (2017). “A comprehensive approach to revealed preference theory”. In: *American Economic Review* 107.4, pp. 1239–1263.

Ok, Efe A. (2011). “Real analysis with economic applications”. In: *Real Analysis with Economic Applications*.

Ok, Efe A. and Gil Riella (2021). “Fully Preorderable Groups”. In: *Order* 38.1, pp. 127–142.

Parr, William C. (1985a). “Jackknifing Differentiable Statistical Functionals”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 47.1, pp. 56–66.

— (1985b). “The bootstrap: Some large sample theory and connections with robustness”. In: *Statistics & Probability Letters* 3.2, pp. 97–100.

Politis, Dimitris N. and Joseph P. Romano (1994). “Large Sample Confidence Regions Based on Subsamples under Minimal Assumptions”. In: *The Annals of Statistics* 22.4, pp. 2031–2050.

Savage, Leonard J. (1954). *Foundations of Statistics*. 1st. New York: John Wiley & Sons, Inc., p. 320. ISBN: 978-0486623498.

Segal, Ilya (2003). “Optimal Pricing Mechanisms with Unknown Demand”. In: *American Economic Review* 93.3, pp. 509–529.

Shao, Jun (1993). “Differentiability of Statistical Functionals and Consistency of the Jackknife”. In: *The Annals of Statistics* 21.1, pp. 61–75.

Sims, Christopher A. (2003). “Implications of rational inattention”. In: *Journal of Monetary Economics* 50.3, pp. 665–690.

Stoye, Jörg (2012). “New perspectives on statistical decisions under ambiguity”. In: *Annual Review of Economics* 4, pp. 257–282.

— (2011). “Statistical decisions under ambiguity”. In: *Theory and Decision* 70.2, pp. 129–148.

Strzalecki, Tomasz (2011). “Axiomatic Foundations of Multiplier Preferences”. In: *Econometrica* 79.1, pp. 47–73.

Toikka, Juuso (2011). “Ironing without control”. In: *Journal of Economic Theory* 146.6, pp. 2510–2526.

Vaart, Aad W. van der and Jon A. Wellner (1996). *Weak Convergence and Empirical Processes*. Springer New York.

Wald, Abraham (1947a). “An Essentially Complete Class of Admissible Decision Functions”. In: *The Annals of Mathematical Statistics* 18.4, pp. 549–555.

— (1939). “Contributions to the Theory of Statistical Estimation and Testing Hypotheses”. In: *The Annals of Mathematical Statistics* 10.4, pp. 299–326.

— (1947b). “Foundations of a General Theory of Sequential Decision Functions”. In: *Econometrica* 15.4, p. 279.

— (1949). “Statistical Decision Functions”. In: *The Annals of Mathematical Statistics* 20.2, pp. 165–205.

Appendix A: Proofs of results in Chapter 1

A.1 Proof of Lemma 1

We will prove the following, stronger result.

Lemma 9. For all $[\sigma] \in \mathscr{S}/_*$,

1. There exists $\rho^\sigma \in \mathscr{D}$ such that $(\rho^\sigma, P^*) \in [\sigma]$.
2. If $\mathscr{P} = \Delta(X)^\Theta$, then there exists an invariant $\delta^* \in \overline{\mathscr{D}}$ such that $(\delta^*, P^\sigma) \in [\sigma]$ for some $P^\sigma \in \mathscr{P}$.

Proof: Fix any $\sigma = (\rho, P) \in \mathscr{S}$. Recall that $\rho P_\theta(E) \equiv \int_X \rho_P(x, E) \mathrm{d}P_\theta(x)$ for all $\theta \in \Theta$ and $E \in \mathscr{A}$.

1. Construct ρ^σ by setting $\rho^\sigma_{P^*}(x, E) = \rho P_\theta(E)$ for all $E \in \mathscr{A}$, $\theta \in \Theta$ and $x \in S_\theta$. Then,

   $$\rho^\sigma P^*_\theta(E) = \int_X \rho^\sigma_{P^*}(x, E) \mathrm{d}P^*_\theta(x) = \int_{S_\theta} \rho P_\theta(E) \mathrm{d}P^*_\theta(x) = \rho P_\theta(E) \quad \forall E \in \mathscr{A}.$$

2. First assume that A is countable and take any surjective decision rule $\delta^* \in \overline{\mathscr{D}}$. For all $a \in A$, let $F_a = \{x \in X : \delta^*(x) = a\}$, and define $P^\sigma_\theta(F_a) = \rho P_\theta(\{a\})$ for all $\theta \in \Theta$ and $a \in A$. Then, $\delta^* P^\sigma(\{a\}) = P^\sigma(\delta^{*-1}(\{a\})) = P^\sigma_\theta(F_a) = \rho P_\theta(\{a\})$.

   Now suppose A is uncountable. Since A is standard Borel, there exists a bijective measurable $\delta^* : X \to A$ such that its inverse is also measurable. This implies that the σ-algebra generated by δ^* coincides with $\mathscr{B}(X)$, hence setting $P^\sigma_\theta(\delta^{*-1}(E)) = \rho P_\theta(E)$, for all $E \in \mathscr{A}$ and $\theta \in \Theta$, defines an experiment on $\mathscr{B}(X)$. By construction, $\delta^* P^\sigma = \rho P$. □

A.2 Proof of Theorem 1

From Axiom 2, we know that $\mathscr{S}/_*$ is finer than $\mathscr{S}/_\sim$. Therefore, in the presence of Axiom 2, providing a representation of $\succsim \subseteq \mathscr{S}^2$ is equivalent to characterizing the preference $\trianglerighteq$ on $\mathscr{F}$ defined in eq. (1.8). Denote the symmetric and asymmetric parts of $\trianglerighteq$ by $\simeq$ and $\triangleright$, respectively. It is easy to see that if $\succsim$ is transitive, so is $\trianglerighteq$. By Lemma 1, for all $f \in \mathscr{F}$, there exists $\rho \in \mathscr{D}$ such that $f = \rho P^*$. Therefore, completeness and reflexivity of $\succsim$ imply that the same properties hold for $\trianglerighteq$.

Consider the following axioms on $\trianglerighteq$:

Axiom A3: For all $f, g, h, h' \in \mathscr{F}$ and $\theta \in \Theta$: if $f_{(\theta)}h \trianglerighteq g_{(\theta)}h$, then $f_{(\theta)}h' \trianglerighteq g_{(\theta)}h'$.

Axiom A4: For all $f, g \in \mathscr{F}$, if $f_{(\theta)}h \trianglerighteq g_{(\theta)}h$ for every $\theta \in \Theta$ and $h \in \mathscr{F}$, then $f \trianglerighteq g$.

Axiom A5: For all $f, g, h \in \mathscr{F}$ and $\theta \in \Theta$, if $f(\theta') = g(\theta') = h(\theta')$ for all $\theta' \neq \theta$, then $f \trianglerighteq g$ implies $\alpha f + (1-\alpha)h \trianglerighteq \alpha g + (1-\alpha)h$ for all $\alpha \in (0, 1]$.

Axiom A6: The set $\{(f, g) \in \mathscr{F}^2 : f \trianglerighteq g\}$ is closed.

Axiom A7: For any $f, g \in \mathscr{F}$, $f \trianglerighteq g \implies f_{(\theta)}h \trianglerighteq g_{(\theta)}h$ for every $\theta \in \Theta$ and $h \in \mathscr{F}$.

Axiom A8: For any $f, g \in \mathscr{F}$ and $\theta \in \Theta$, if $f(\theta') = g(\theta')$ for all $\theta' \neq \theta$, then $f \trianglerighteq g$ or $g \trianglerighteq f$.

We have the following result linking the axioms above to axioms 3 to 7.

Lemma 10. Let $\succsim \subseteq \mathscr{S}^2$ satisfy Consequentialism. Then, the following statements hold:

1. If $\succsim$ satisfies Axiom 3, then $\trianglerighteq$ satisfies Axiom A3.
2. If $\succsim$ satisfies Axiom 4, then $\trianglerighteq$ satisfies Axiom A4.
3. If $\succsim$ satisfies Axiom 5, then $\trianglerighteq$ satisfies Axiom A5.
4. If $\succsim$ satisfies Axiom 6, then $\trianglerighteq$ satisfies Axiom A6.
5. If $\succsim$ satisfies Axiom 7, then $\trianglerighteq$ satisfies Axiom A7.

6. If $\succsim$ satisfies Axiom 8, then $\geq$ satisfies Axiom A8.

Proof:

1. Take any $f, g, h, h' \in \mathscr{F}$ such that $f_{(\theta)}h \geq g_{(\theta)}h$. By Lemma 1, there exists $\rho, \tau, \gamma, \gamma' \in \mathscr{D}$ such that $f = \rho P^*,\ g = \tau P^*,\ h = \gamma P^*$ and $h' = \gamma' P^*$. Fix any $\theta \in \Theta$. Then,

$$(\rho_{\{\theta\}}\gamma)P^*_{\theta'}(\cdot) = \int_{S_{\theta'}} \rho_{\{\theta\}}\gamma_{P^*}(x, \cdot)\mathrm{d}P^*_{\theta'}(x) = \begin{cases} \rho P^*_{\theta}(\cdot), & \text{if } \theta' = \theta \\ \gamma P^*_{\theta'}(\cdot), & \text{if } \theta' \neq \theta, \end{cases}$$

hence $f_{(\theta)}h = (\rho_{\{\theta\}}\gamma)P^*$, and similarly for $g_{(\theta)}h = (\tau_{\{\theta\}}\gamma)P^*,\ f_{(\theta)}h' = (\rho_{\{\theta\}}\gamma')P^*$, and $g_{(\theta)}h' = (\tau_{\{\theta\}}\gamma')P^*$. By Consequentialism and $f_{(\theta)}h \geq g_{(\theta)}h$, we have $(\rho_{\{\theta\}}\gamma, P^*) \succsim (\tau_{\{\theta\}}\gamma, P^*)$. By Axiom 3, this implies $(\rho_{\{\theta\}}\gamma', P^*) \succsim (\tau_{\{\theta\}}\gamma', P^*)$, and the definition of $\geq$ gives us $f_{(\theta)}h' \geq g_{(\theta)}h'$.

2. Let $f, g \in \mathscr{F}$ be such that $f_{(\theta)}h \geq g_{(\theta)}h$ for all $\theta \in \Theta$ and $h \in \mathscr{F}$. Lemma 1 implies that for all $h \in \mathscr{F}$, there exists $\gamma \in \mathscr{D}$ such that $h = \gamma P^*$. Let $\rho P^* = f$ and $\tau P^* = g$. Now, an argument analogous to the one in part 1 implies that, for all $\theta \in \Theta$ and $h \in \mathscr{F}$, there exists $\gamma \in \mathscr{D}$ with $f_{(\theta)}h = (\rho_{\{\theta\}}\gamma)P^*$ and $g_{(\theta)}h = (\tau_{\{\theta\}}\gamma)P^*$. Consequentialism and $f_{(\theta)}h \geq g_{(\theta)}h$ for all $h \in \mathscr{F}$ now imply that $(\rho_{\{\theta\}}\gamma, P^*) \succsim (\tau_{\{\theta\}}\gamma, P^*)$ for all $\gamma \in \mathscr{D}$. By Axiom 4, we have $(\rho, P^*) \succsim (\tau, P^*)$, hence $f \geq g$.

3. Fix $\theta \in \Theta$ and take any $f, g, h \in \mathscr{F}$ such that $f(\theta') = g(\theta') = h(\theta')$ for all $\theta' \neq \theta$. Assume $f \geq g$. By Lemma 1, there exists $\rho, \tau, \gamma, \kappa \in \mathscr{D}$ such that $f = (\rho_{\{\theta\}}\kappa)P^*,\ g = (\tau_{\{\theta\}}\kappa)P^*$ and $h = (\gamma_{\{\theta\}}\kappa)P^*$. By Consequentialism, $f \geq g$ implies $(\rho_{\{\theta\}}\kappa, P^*) \succsim (\tau_{\{\theta\}}\kappa, P^*)$. Applying Axiom 5 we get $(\alpha\rho_{\{\theta\}}\kappa + (1-\alpha)\gamma_{\{\theta\}}\kappa, P^*) \succsim (\alpha\tau_{\{\theta\}}\kappa + (1-\alpha)\gamma_{\{\theta\}}\kappa, P^*)$ for all $\alpha \in (0, 1]$. The result follows from how $\rho,\ \tau,\ \gamma$ and κ were defined.

4. This follows immediately from the definition of $\geq$.

5. Take $f, g \in \mathscr{F}$ and suppose $f \geq g$. Then there exists $\rho, \tau \in \mathscr{D}$ such that $f = \rho P^*$ and $g = \tau P^*$. By Axiom 2, $(\rho, P^*) \succsim (\tau, P^*)$, thus by Axiom 7, $(\rho_{\{\theta\}}\gamma, P^*) \succsim (\tau_{\{\theta\}}\gamma, P^*)$ for all

$\theta \in \Theta$ and $\gamma \in \mathscr{D}$. Lemma 1 and Axiom 2 then imply that $f_{(\theta)}h \geq g_{(\theta)}h$ for every $\theta \in \Theta$ and $h \in \mathscr{F}$.

6. Fix any $\theta \in \Theta$ and suppose $f, g \in \mathscr{F}$ are such that $f(\theta') = g(\theta')$ for all $\theta' \neq \theta$. By a now familiar argument, there exist $\rho, \tau, \gamma \in \mathscr{D}$ such that $f = (\rho_{\{\theta\}}\gamma)P^*$ and $g = (\tau_{\{\theta\}}\gamma)P^*$. Axiom 8 implies that $(\rho_{\{\theta\}}\gamma, P^*) \succsim (\tau_{\{\theta\}}\gamma, P^*)$ or $(\tau_{\{\theta\}}\gamma, P^*) \succsim (\rho_{\{\theta\}}\gamma, P^*)$. The result follows. □

We are now ready to prove Theorem 1. It is routine to check that the representation satisfies the axioms, so we only prove sufficiency. Let $\succsim$ be a reflexive and transitive preference satisfying axioms 2 to 8, and define $\geq$ as in (1.8). If $\succsim$ is trivial, it is clear that it has a dominance representation via a constant utility function u, and we are done. So assume that there exists $\sigma, \sigma' \in \mathscr{S}$ such that $\sigma \succ \sigma'$. By Lemma 10, $\geq$ is reflexive, transitive and satisfies axioms A3 to A8.

For each $\theta \in \Theta$, define the conditional preference $\geq_\theta$ on $\Delta(A)$ as follows:

$$\forall f(\theta), g(\theta) \in \Delta(A),\ f(\theta) \geq_\theta g(\theta) \iff f \geq g \text{ and } f(\theta') = g(\theta')\ \forall \theta' \neq \theta. \tag{A.1}$$

Axiom A3, and reflexivity and transitivity of $\geq$ guarantee that $\geq_\theta$ is well-defined, transitive and reflexive for every $\theta \in \Theta$, while Axiom A8 implies that it is also complete. By Axiom A5, $\geq_\theta$ satisfies *Independence*: if $p \geq_\theta q$, then $\alpha p + (1-\alpha)s \geq_\theta \alpha q + (1-\alpha)s$ for any $\alpha \in (0,1]$ and $s \in \Delta(A)$. Take sequences $(p_n), (q_n) \in \Delta(A)$ such that $p_n \to p$, $q_n \to q$ and $p_n \geq_\theta q_n$ for all $n \geq 1$. Axiom A6 implies that $p, q \in \Delta(A)$ and $p \geq_\theta q$. Since $\Delta(A)$ is a metric space, this implies that $\{(p,q) \in \Delta(A)^2 : p \geq_\theta q\}$ is closed. We can thus apply the classic expected utility representation theorem for compact prize spaces – see, e.g., Kreps (2018, Chapter 5) – to find, for every $\theta \in \Theta$, a continuous utility function $u_\theta : A \to \mathbb{R}$ such that, for all $f(\theta), g(\theta) \in \Delta(A)$,

$$f(\theta) \geq_\theta g(\theta) \iff \int_A u_\theta \mathrm{d}f(\theta) \geq \int_A u_\theta \mathrm{d}g(\theta). \tag{A.2}$$

Now take any $f, g \in \mathscr{F}$ such that $f \geq g$. By axioms A4 and A7, we must have $f(\theta) \geq_\theta g(\theta)$

for all $\theta \in \Theta$. Considering the representations in (A.2), this implies that for all $f, g \in \mathscr{F}$,

$$f \geq g \iff \int_A u_\theta \mathrm{d}f(\theta) \geq \int_A u_\theta \mathrm{d}g(\theta) \quad \forall \theta \in \Theta.$$

In view of (1.8) and Axiom 2, this implies that, for all $(\rho, P), (\tau, Q) \in \mathscr{S}$,

$$(\rho, P) \mathrel{\hat{\gtrsim}} (\tau, Q) \iff \int_A u_\theta \mathrm{d}\rho P_\theta \geq \int_A u_\theta \mathrm{d}\tau Q_\theta \ \forall \theta \in \Theta \iff \int_X u_\theta(\rho) \mathrm{d}P_\theta \geq \int_X u_\theta(\tau) \mathrm{d}Q_\theta \ \forall \theta \in \Theta.$$

The representation obtains by defining $u(\cdot, \theta) = u_\theta$ for each $\theta \in \Theta$. Its uniqueness properties derive from cardinal uniqueness of u_θ for every $\theta \in \Theta$. That is, if u'_θ also represents $\geq_\theta$, then there exist $b_\theta > 0$ and $c_\theta \in \mathbb{R}$ such that $u'_\theta = b_\theta u_\theta + c_\theta$. This is guaranteed by the expected utility theorem.

Consider the following definition.

Definition 12. We call the parameters in $\Theta^I \equiv \{\theta \in \Theta : (\rho_{\{\theta\}}\gamma, P^*) \mathrel{\hat{\sim}} (\tau_{\{\theta\}}\gamma, P^*) \ \forall \rho, \tau \in \mathscr{D}\}$ *irrelevant.* ◊

Irrelevant parameters get their name from the fact that changing the action distributions conditional on Θ^I does not affect the DM's valuation of a pairing. For future reference, note that $\geq_\theta$ is trivial for all $\theta \in \Theta^I$. This then implies that any utility representation u_θ of $\geq_\theta$, with $\theta \in \Theta^I$, is a constant function.

A.3 Proof of Lemma 2

Let $\gtrsim$ be a preference satisfying axioms 1 to 6 and define its unanimously preferred sub-relation $\hat{\gtrsim}$ as in Definition 3. We start by proving Proposition 1:

Proof of Proposition 1:

1. The inclusion $\hat{\gtrsim} \subseteq \gtrsim$ follows immediately. To prove transitivity, take $\sigma \mathrel{\hat{\gtrsim}} \sigma'$ and $\sigma' \mathrel{\hat{\gtrsim}} \tilde{\sigma}$. Then there exist $\rho, \tau, \gamma \in \mathscr{D}$, with $(\rho, P^*) \in [\sigma]$, $(\tau, P^*) \in [\sigma']$ and $(\gamma, P^*) \in [\tilde{\sigma}]$, such that $(\rho_{\{\theta\}}\kappa, P^*) \gtrsim (\tau_{\{\theta\}}\kappa, P^*)$ and $(\tau_{\{\theta\}}\kappa, P^*) \gtrsim (\gamma_{\{\theta\}}\kappa, P^*)$ for all $\kappa \in \mathscr{D}$ and $\theta \in \Theta$, . By

transitivity of $\succsim$, we have $(\rho_{\{\theta\}}\kappa, P^*) \succsim (\gamma_{\{\theta\}}\kappa, P^*)$, for every $\kappa \in \mathscr{D}$ and $\theta \in \Theta$. Thus, by definition, $\sigma \mathrel{\hat{\succsim}} \tilde{\sigma}$.

2. We want to show that $\hat{\succsim}$ satisfies axioms 2 to 8. First note that since $\succsim$ satisfies axioms 3 and 4, we have that $\hat{\succsim}$ is reflexive, thus non-empty. Axiom 2 follows directly from $\hat{\succsim} \subseteq \succsim$. Axioms 3 and 5 are implied by the Sure Thing Principle (STP) and Mixture Independence (MI) respectively, which we show in parts 3 and 4, below.

   Now, note that for any $T \subseteq \Theta$ and $\rho, \gamma, \gamma' \in \mathscr{D}$,

   $$\rho_T\gamma_{\{\theta\}}\gamma' = \begin{cases} \rho_{\{\theta\}}\gamma', & \text{if } \theta \in T \\ \gamma_{\{\theta\}}\gamma', & \text{if } \theta \in \Theta \setminus T. \end{cases} \tag{A.3}$$

   In particular, since $\succsim$ satisfies IIP, for any $\gamma \in \mathscr{D}$ and $\theta \in \Theta$, $(\rho_{\{\theta\}}\gamma, P^*) \succsim (\tau_{\{\theta\}}\gamma, P^*)$ implies $(\rho_{\{\theta\}}\gamma_{\{\theta'\}}\gamma', P^*) \succsim (\tau_{\{\theta\}}\gamma_{\{\theta'\}}\gamma', P^*)$ for all $\gamma' \in \mathscr{D}$ and $\theta' \in \Theta$, and thus $(\rho_{\{\theta\}}\gamma, P^*) \mathrel{\hat{\succsim}} (\tau_{\{\theta\}}\gamma, P^*)$. Since $\hat{\succsim} \subseteq \succsim$, we have that for all $\theta \in \Theta$ and $\rho, \tau, \gamma \in \mathscr{D}$, $(\rho_{\{\theta\}}\gamma, P^*) \mathrel{\hat{\succsim}} (\tau_{\{\theta\}}\gamma, P^*)$ if, and only if, $(\rho_{\{\theta\}}\gamma, P^*) \succsim (\tau_{\{\theta\}}\gamma, P^*)$. By Consequentialism, we also have that $(\rho, P^*) \mathrel{\hat{\succsim}} (\tau, P^*)$ if, and only if, $(\rho_{\{\theta\}}\gamma, P^*) \succsim (\tau_{\{\theta\}}\gamma, P^*)$ for all $\theta \in \Theta$ and $\gamma \in \mathscr{D}$. Taken together, these two facts facts imply axioms 4 and 7.

   Recalling the definition in (A.1) and since $\Delta(A)$ is a metric space, we have that $\{(f(\theta), g(\theta)) \in \Delta(A)^2 : f(\theta) \geq_\theta g(\theta)\}$ is closed if, and only if, it is sequentially closed. Also recall from Section A.2 that $(\rho, P) \mathrel{\hat{\succsim}} (\tau, Q) \iff f(\theta) \geq_\theta g(\theta)$ for all $f = \rho P$, $g = \tau Q$ and $\theta \in \Theta$. Therefore, $\{(\rho P, \tau Q) \in \mathscr{F}^2 : (\rho, P) \mathrel{\hat{\succsim}} (\tau, Q)\}$ is closed if, and only if, $\bigcap_{\theta\in\Theta}\{(f, g) \in \mathscr{F}^2 : f(\theta) \geq_\theta g(\theta)\}$ is closed. So consider sequences $\{f^n\}_{n\geq1}$ and $\{g^n\}_{n\geq1} \in \mathscr{F}$ such that $f^n \longrightarrow f$, $g^n \longrightarrow g$, and $f^n(\theta) \geq_\theta g^n(\theta)$ for all $\theta \in \Theta$, $n \geq 1$. Since $\geq_\theta$ is continuous by Section A.2, we have that $\{(f, g) \in \mathscr{F}^2 : f(\theta) \geq_\theta g(\theta)\}$ is closed for every $\theta \in \Theta$. Therefore, the intersection is also closed, which then implies continuity of $\hat{\succsim}$.

3. To prove that $\hat{\succsim}$ satisfies the Sure Thing Principle, first consider $\rho, \tau, \gamma \in \mathscr{D}$ such that

$(\rho_T\gamma, P^*) \mathrel{\hat{\succsim}} (\tau_T\gamma, P^*)$. Then, by eq. (A.3), $(\rho_{\{\theta\}}\kappa, P^*) \succsim (\tau_{\{\theta\}}\kappa, P^*)$ for all $\theta \in T$ and $\kappa \in \mathscr{D}$, which by Monotonicity of $\succsim$, implies $(\rho_T\gamma', P^*) \succsim (\tau_T\gamma', P^*)$ for any $\gamma' \in \mathscr{D}$. Again using eq. (A.3) and the fact that $(\rho_{\{\theta\}}\kappa, P^*) \succsim (\tau_{\{\theta\}}\kappa, P^*)$ for all $\theta \in T$ and $\kappa \in \mathscr{D}$, gives us $(\rho_T\gamma', P^*) \mathrel{\hat{\succsim}} (\tau_T\gamma', P^*)$.

4. Take any $P \in \mathscr{P}$ and $\rho, \tau \in \mathscr{D}$ with $(\rho, P) \mathrel{\hat{\succsim}} (\tau, P)$. Then, by definition, there exists $\rho', \tau' \in \mathscr{D}$ such that $\rho'P^* = \rho P$, $\tau'P^* = \tau P$ and $(\rho'_{\{\theta\}}\gamma, P^*) \succsim (\tau'_{\{\theta\}}\gamma, P^*)$ for all $\theta \in \Theta$ and $\gamma \in \mathscr{D}$. Now CMI implies that $(\alpha\rho'_{\{\theta\}}\gamma + (1-\alpha)\kappa'_{\{\theta\}}\gamma, P^*) \succsim (\alpha\tau'_{\{\theta\}}\gamma + (1-\alpha)\kappa'_{\{\theta\}}\gamma, P^*)$ for all $\theta \in \Theta$, $\alpha \in (0,1]$ and $\kappa', \gamma \in \mathscr{D}$. By Lemma 1, for any $\kappa P \in \mathscr{F}$, there exists $\kappa' \in \mathscr{D}$ such that $\kappa'P^* = \kappa P$. Therefore, for all $\kappa \in \mathscr{D}$ and $\alpha \in (0,1]$, there exists $\kappa' \in \mathscr{D}$ such that $(\alpha\rho + (1-\alpha)\kappa)P = (\alpha\rho' + (1-\alpha)\kappa')P^*$, $(\alpha\tau + (1-\alpha)\kappa)P = (\alpha\tau' + (1-\alpha)\kappa')P^*$, and $((\alpha\rho' + (1-\alpha)\kappa')_{\{\theta\}}\gamma, P^*) \succsim ((\alpha\tau' + (1-\alpha)\kappa')_{\{\theta\}}\gamma, P^*)$ for all $\theta \in \Theta$ and $\gamma \in \mathscr{D}$. We conclude that $\hat{\succsim}$ satisfies Mixture Independence.

5. Suppose $\succsim$ is non-trivial. Then, by completeness, there exists $(\rho, P) \succ (\tau, Q)$. Lemma 1 now guarantees that there exists $\rho', \tau' \in \mathscr{D}$ with $(\rho', P^*) \succ (\tau', P^*)$. Let $T = \{\theta \in \Theta : (\rho'_{\{\theta\}}\gamma, P^*) \succsim (\tau'_{\{\theta\}}\gamma, P^*)\ \forall\gamma \in \mathscr{D}\}$, and note that by IIP and Monotonicity, $T \neq \emptyset$. Monotonicity now implies that $(\rho'_T\tau', P^*) \succsim (\rho', P^*) \succ (\tau', P^*)$, and by transitivity, $(\rho'_T\tau', P^*) \succ (\tau', P^*)$. By construction of T and eq. (A.3), $(\rho'_T\tau', P^*) \mathrel{\hat{\succsim}} (\tau', P^*)$. Again by Monotonicity, there exists $\theta \in T$ and $\gamma \in \mathscr{D}$ such that $(\tau'_{\{\theta\}}\gamma, P^*) \not\succsim (\rho'_T\tau'_{\{\theta\}}\gamma, P^*)$. Therefore $(\tau', P^*) \mathrel{\hat{\not\succsim}} (\rho'_T\tau', P^*)$, which concludes the proof. □

Now, from Proposition 1, we know that $\hat{\succsim}$ is a dominance relation. By Theorem 1, it has a dominance representation by a parameter-wise cardinally unique utility function $u : A \times \Theta \to \mathbb{R}$.

A.4 Proof of Theorem 2

If $\succsim$ is trivial, it can clearly be represented by any utility function u paired with a constant aggregator functional I. So assume that $\succsim$ is not trivial. Let $\succsim$ satisfy axioms 1 to 6 and fix a utility function $u : A \times \Theta \to \mathbb{R}$ that represents its unanimously preferred sub-relation $\hat{\succsim}$. Consider the risk mapping $r_u : \mathscr{S} \to \mathbb{R}^\Theta$ under u defined by (1.2), and set $\mathscr{R}_u = \{r_u(\sigma) : \sigma \in \mathscr{S}\}$. Note that, as a product of (possibly degenerate) intervals, $\mathscr{R}_u$ is connected, and as a consequence of the Hewitt-Marczewski-Pondiczery theorem (Hewitt, 1946), it is also separable.

Consider the preference relation $\geq^u$ on $\mathscr{R}_u$ defined by (1.6). It is easy to see that $\geq^u$ is transitive, since $r_u(\sigma) \geq^u r_u(\sigma')$ and $r_u(\sigma') \geq^u r_u(\tilde{\sigma})$ imply that $\sigma \gtrsim \sigma'$ and $\sigma' \gtrsim \tilde{\sigma}$, hence $\sigma \gtrsim \tilde{\sigma}$ by transitivity of $\gtrsim$, which in turn implies $r_u(\sigma) \geq^u r_u(\tilde{\sigma})$. By definition, $\geq^u$ is also complete on $\mathscr{R}_u$.

Now we prove that $\geq^u$ is continuous, i.e., that $\{(r, r') \in \mathscr{R}_u^2 : r \geq^u r'\}$ is closed in the product topology. First note that for every $\theta \in \Theta$, the functional $\Delta(A) \ni f(\theta) \mapsto \int_A u(a, \theta) \mathrm{d} f(\theta)$ is continuous in the topology of weak convergence. Thus, $\tilde{r}_u : \mathscr{F} \to \mathbb{R}^\Theta$, defined as

$$\tilde{r}_u(f)(\theta) = \int_A u(a, \theta) \mathrm{d} f(\theta) \quad \forall f \in \mathscr{F},\ \theta \in \Theta, \tag{A.4}$$

is continuous in the product topology on $\mathbb{R}^\Theta$. Now note that $\{(r_u(\sigma), r_u(\sigma')) \in \mathscr{R}_u^2 : r_u(\sigma) \geq^u r_u(\sigma')\} = \{(\tilde{r}_u(\rho P), \tilde{r}_u(\tau Q)) \in \mathscr{R}_u^2 : (\rho, P) \gtrsim (\tau, Q)\}$. Let the map $\tilde{R}_u : \mathscr{F}^2 \to \mathscr{R}_u^2$ be defined by $\tilde{R}_u(f, g) = (\tilde{r}_u(f), \tilde{r}_u(g))$. Then $\tilde{R}_u$ is continuous, as a product of continuous functions, and

$$\tilde{R}_u(\{(\rho P, \tau Q) \in \mathscr{F}^2 : (\rho, P) \gtrsim (\tau, Q)\}) = \{(\tilde{r}_u(\rho P), \tilde{r}_u(\tau Q)) \in \mathscr{R}_u^2 : (\rho, P) \gtrsim (\tau, Q)\}.$$

Since A is compact, so is $\Delta(A)$. Thus, by Tychonoff's theorem, $\mathscr{F}$ is a compact Hausdorff space, and so is $\mathscr{F}^2$. Then, by Continuity of $\gtrsim$, $\{(\rho P, \tau Q) \in \mathscr{F}^2 : (\rho, P) \gtrsim (\tau, Q)\}$ is also compact, since it is a closed subset of a compact space. This makes $\tilde{R}_u(\{(\rho P, \tau Q) \in \mathscr{F}^2 : (\rho, P) \gtrsim (\tau, Q)\})$ compact as well – because $\tilde{R}_u$ is continuous – hence closed, since $\mathscr{R}_u^2$ is Hausdorff. We conclude that $\geq^u$ is continuous.

Therefore, $\geq^u$ is a continuous weak order on a connected and separable topological space. By Herden (1989, Corollary 3.2), there exists a continuous utility function $I : \mathscr{R}_u \to \mathbb{R}$ such that, for all $r_u(\sigma), r_u(\sigma') \in \mathscr{R}_u$, $r_u(\sigma) \geq^u r_u(\sigma') \iff I(r_u(\sigma)) \geq I(r_u(\sigma'))$. By definition of $\geq^u$, we obtain, for all $\sigma, \sigma' \in \mathscr{S}$,

$$\sigma \succsim \sigma' \iff r_u(\sigma) \geq^u r_u(\sigma') \iff I(r_u(\sigma)) \geq I(r_u(\sigma')).$$

Now we prove that $\succsim$ has a representation with an ex-post utility function $u : A \times \Theta \to \mathbb{R}$ such that $u(\cdot, \theta)$ is non-constant for all $\theta \in \Theta$. From Section A.2, for any dominance representation $\hat{u}$ of $\hat{\succsim}$, $\hat{u}(\cdot, \theta)$ is constant for all $\theta \in \Theta^I$. Let $(\hat{u}, \hat{I})$ be an MRA representation of $\succsim$ such that $\hat{u}(a, \theta) = k \in \mathbb{R}$ for all $a \in A$ and $\theta \in \Theta^I$. Such a representation exists by the uniqueness properties of Theorem 1. Set $u(\cdot, \theta) = \hat{u}(\cdot, \theta)$ for $\theta \in \Theta \setminus \Theta^I$, and let $u(\cdot, \theta)$ be any non-constant function with $k \in co(u(A, \theta))$ for all $\theta \in \Theta^I$, where $co(\cdot)$ denotes the convex hull of a set. Note that $\mathscr{R}_{\hat{u}} \subseteq \mathscr{R}_u$, and let $I : \mathscr{R}_u \to \mathbb{R}$ be given by $I(r) = \hat{I}(\hat{r})$ if $r(\theta) = \hat{r}(\theta)$ for all $\theta \in \Theta \setminus \Theta^I$, where $r \in \mathscr{R}_u$ and $\hat{r} \in \mathscr{R}_{\hat{u}}$. Clearly $I(r_u(\sigma)) \geq I(r_u(\sigma')) \iff \hat{I}(r_u(\sigma)_{(\Theta^I)}k) \geq \hat{I}(r_u(\sigma')_{(\Theta^I)}k)$, therefore (u, I) also represents $\succsim$.

A.5 Proofs of results in Section 1.5

Let $\succsim$ be a preference relation on $\mathscr{S}$ satisfying axioms 1 to 6, and take a utility function u that represents $\hat{\succsim}$. Consider the preference relation $\geq^u$ on $\mathscr{R}_u$ defined by (1.6). We first formally state and prove a result that was only alluded to in the main text.

Lemma 11. The preference $\geq^u$ on $\mathscr{R}_u$ satisfies the *basic A-A axioms*

- *Weak Order*: the preference $\geq^u$ is complete and transitive.
- *A-A Monotonicity*: for all $r, r' \in \mathscr{R}_u$, if $r \geq r'$, then $r \geq^u r'$.
- *Risk Independence:* for all constant risk functions $r, r', q \in \mathscr{R}_u$ and $\alpha \in [0, 1]$, $r \geq^u r'$ implies $\alpha r + (1 - \alpha)q \geq^u \alpha r' + (1 - \alpha)q$.
- *Mixture Continuity:* the sets $\{\alpha \in [0, 1] : \alpha r + (1 - \alpha)r' \geq^u q\}$ and $\{\alpha \in [0, 1] : q \geq^u \alpha r + (1 - \alpha)r'\}$ are closed for all $r, r', q \in \mathscr{R}_u$.

Proof: We showed in Section A.4 that $\geq^u$ satisfies Weak Order and that $\{(r, r') \in \mathscr{R}_u^2 : r \geq^u r'\}$ is closed. Take any $r, r', q \in \mathscr{R}_u$ and a sequence $(\alpha_n)_{n\geq 1} \in [0, 1]$, with $\lim_{n\to\infty} \alpha_n = \alpha$, such that $\alpha_n r + (1 - \alpha_n)r' \geq^u q$ for all $n \geq 1$. Since $\{(r, r') \in \mathscr{R}_u^2 : r \geq^u r'\}$ is closed, we have that $\lim_n(\alpha_n r + (1 - \alpha_n)r') = \alpha r + (1 - \alpha)r' \geq^u q$. Hence $\{\alpha \in [0, 1] : \alpha r + (1 - \alpha)r' \geq^u q\}$ is closed. An analogous argument proves that $\{\alpha \in [0, 1] : q \geq^u \alpha r + (1 - \alpha)r'\}$ is closed, thus $\geq^u$ satisfies Mixture Continuity.

A-A Monotonicity follows directly from Theorem 2, since I is monotone. To prove Risk Independence, note that if $r, r' \in \mathscr{R}_u$ are constant, then $r \geq^u r'$ if and only if $r \geq r'$, by A-A Monotonicity. The result follows immediately. □

Now we turn to Lemma 3.

Proof of Lemma 3: The proof of statement 1 consists of applying the definitions and doing simple algebra, thus will be omitted. To prove statement 2, recall that for all $\rho, \gamma \in \mathscr{D}$, $T \in \Sigma$ and $\theta \in \Theta$,

$$\begin{aligned} r_u(\rho_T\gamma, P^*)(\theta) = \int_X u(\rho_T\gamma, \theta)\mathrm{d}P_\theta^* &= \int_A \int_X u(a, \theta)\rho_T\gamma(x, \mathrm{d}a)P_\theta^*(\mathrm{d}x) \\ &= \begin{cases} \int_A u(a, \theta)\mathrm{d}\rho P_\theta^*(a), & \text{if } \theta \in T \\ \int_A u(a, \theta)\mathrm{d}\gamma P_\theta^*(a), & \text{if } \theta \notin T. \end{cases} \end{aligned}$$

Therefore, $r_u(\rho_T\gamma, P^*)(\theta) = r_u(\rho, P^*)(\theta)$ for $\theta \in T$, and $r_u(\rho_T\gamma, P^*)(\theta) = r_u(\gamma, P^*)(\theta)$ for $\theta \in \Theta \setminus T$, which proves the statement.

Now take another MRA representation $(\tilde{u}, \tilde{I})$ of $\succsim$, normalized in such a way that $\mathscr{R}_{\tilde{u}} = [0, 1]^\Theta$. It is possible to do this because, by Theorem 2, $\succsim$ has an MRA representation with parameter-wise non-constant utility. Then, parameter-wise cardinal uniqueness of u allows us normalize this utility function on each parameter. Note that $r_{\tilde{u}}(\overline{\rho}, P^*)(\theta) = 1$ and $r_{\tilde{u}}(\underline{\rho}, P^*)(\theta) = 0$ for all $\theta \in \Theta$.

If $\sigma \in \mathscr{K}(\hat{\succsim})$, Lemma 1 implies that there exists $\alpha \in [0, 1]$ such that $\sigma \in [(\alpha\overline{\rho} + (1 - \alpha)\underline{\rho}, P^*)]$, thus by statement 1, $r_{\tilde{u}}(\sigma)(\theta) = \alpha$, for all $\theta \in \Theta$. Conversely, take $r_{\tilde{u}}(\tau, P) \in \mathscr{R}_{\tilde{u}}$ such that $r_{\tilde{u}}(\tau, P)(\theta) = \alpha$ for all $\theta \in \Theta$, and take $\rho P^* = \tau P$. Since $r_{\tilde{u}}(\rho, P^*)_{(\{\theta\})} r_{\tilde{u}}(\gamma, P^*) = r_{\tilde{u}}(\alpha\overline{\rho} + (1 - \alpha)\underline{\rho}, P^*)_{(\{\theta\})} r_{\tilde{u}}(\gamma, P^*)$ for any $\gamma \in \mathscr{D}$ and $\theta \in \Theta$, we may apply statement 2 and the definition of

$\succeq^u$ to obtain $r_{\tilde{u}}(\rho_{\{\theta\}}\gamma, P^*) \simeq^u r_{\tilde{u}}((\alpha\overline{\rho} + (1-\alpha)\underline{\rho})_{\{\theta\}}\gamma, P^*)$ for all $\theta \in \Theta$. By definition of $\simeq^u$, $(\rho_{\{\theta\}}\gamma, P^*) \sim ((\alpha\overline{\rho} + (1-\alpha)\underline{\rho})_{\{\theta\}}\gamma, P^*)$ for all $\theta \in \Theta$, and by IIP, $(\rho, P^*) \mathrel{\hat{\sim}} (\alpha\overline{\rho} + (1-\alpha)\underline{\rho}, P^*)$. Since $\tau P_\theta = \rho P^*_\theta$ for all $\theta \in \Theta$, we have $(\tau, P) \in \mathscr{K}(\hat{\succsim})$, which finishes the proof. □

In view statement 3 of Lemma 3, for the remainder of this section we work with a representation (u, I) of $\succsim$ such that $\mathscr{R}_u = [0,1]^\Theta$, so that $r_u(\sigma)$ being constant implies $\sigma \in \mathscr{K}(\hat{\succsim})$. We start by proving the lemmas which translate SDT axioms on $\succsim$ to the corresponding Anscombe-Aumann axioms on $\succeq^u$.

Lemma 12. If $\succsim$ satisfies Mixture Independence, then $\succeq^u$ satisfies Independence.

Proof: Take $r, r', q \in \mathscr{R}_u$ and assume $\succsim$ satisfies Axiom 9. From Lemma 1 and the definition of $\succeq^u$, there exists $\rho, \tau, \gamma \in \mathscr{D}$ and $P \in \mathscr{P}$ such that, $r = r_u(\rho, P)$, $r' = r_u(\tau, P)$, $q = r_u(\gamma, P)$ and $(\rho, P) \succsim (\tau, P)$. By Mixture Independence, we have $(\alpha\rho + (1-\alpha)\gamma, P) \succsim (\alpha\tau + (1-\alpha)\gamma, P)$ for all $\alpha \in (0, 1]$. Therefore, from statement 1 of Lemma 3, we obtain $\alpha r + (1-\alpha)q \succeq^u \alpha r' + (1-\alpha)q$ for all $\alpha \in (0, 1]$. □

Lemma 13. If $\succsim$ satisfies Monotone Continuity*, then $\succeq^u$ satisfies Monotone Continuity.

Proof: Take any $r, r', q \in \mathscr{R}_u$ with q constant, and let $(T_n)_{n\geq 1} \in \Sigma$ be such that $T_1 \supseteq T_2 \supseteq \cdots$ and $\bigcap_{n\geq 1} T_n = \emptyset$. Suppose $r \succ^u r'$. By Lemma 1, there exist $\rho, \tau, \kappa \in \mathscr{D}$, with $(\kappa, P^*) \in \mathscr{K}(\hat{\succsim})$, such that $r = r_u(\rho, P^*)$, $r' = r_u(\tau, P^*)$ and $q = r_u(\gamma, P^*)$. From the definition of $\succeq^u$, $(\rho, P^*) \succ (\tau, P^*)$. Then Monotone Continuity* implies that there exists $m \geq 1$ such that $(\kappa_{T_m}\rho, P^*) \succ (\tau, P^*)$. The result then follows from statement 2 of Lemma 3. □

Lemma 14. If $\succsim$ satisfies Hedging, then $\succeq^u$ satisfies Uncertainty Aversion.

Proof: Let $r, r' \in \mathscr{R}_u$ be such that $r \simeq^u r'$. By Lemma 1, there exists $P \in \mathscr{P}$ and $\rho, \tau \in \mathscr{D}$ such that $r = r_u(\rho, P)$, $r' = r_u(\tau, P)$ and $(\rho, P) \sim (\tau, P)$. Then, by Hedging, $(\alpha\rho + (1-\alpha)\tau, P) \succsim (\rho, P)$ for all $\alpha \in (0, 1)$. This implies that $r_u(\alpha\rho + (1-\alpha)\tau, P) \succeq^u r_u(\rho, P)$, thus from statement 1 of Lemma 3, we obtain $\alpha r + (1-\alpha)r' \succeq^u r$. □

Lemma 15. If $\succsim$ satisfies CRE-Independence, then $\geq^u$ satisfies Certainty Independence.

Proof: Take any $r, r', q \in \mathscr{R}_u$ with constant q, and assume $r \geq^u r'$. A now familiar argument implies that there exists $P \in \mathscr{P}$ and $\rho, \tau, \kappa \in \mathscr{D}$ such that $r = r_u(\rho, P)$, $r' = r_u(\tau, P)$, $q = r_u(\kappa, P)$ and $(\rho, P) \succsim (\tau, P)$. Since q is constant, Lemma 3 guarantees that $(\tau, P) \in \mathscr{K}(\hat{\succsim})$. By CRE-Independence, $(\alpha\rho + (1-\alpha)\kappa, P) \succsim (\alpha\tau + (1-\alpha)\kappa, P)$ for all $\alpha \in (0, 1]$. Applying Lemma 3 once again, we obtain $\alpha r + (1-\alpha)q \geq^u \alpha r' + (1-\alpha)q$. □

Lemma 16. If $\succsim$ satisfies symmetry, then $\geq^u$ satisfies:

- *Risk Symmetry*: for all $r, r', q, k \in \mathscr{R}_u$, with q, k constant, and any $T, F \in \Sigma$ such that $T \cap F = \emptyset$, $r_{(T)}q_{(F)}k \geq^u r'_{(T)}q_{(F)}k$ implies $r_{(T)}k_{(F)}q \geq^u r'_{(T)}k_{(F)}q$.

Proof: Assume that $r_{(T)}q_{(F)}k \geq^u r'_{(T)}q_{(F)}k$, where $r, r', q, k \in \mathscr{R}_u$, with q, k constant, and $T \cap F = \emptyset$. There exist $\rho, \tau \in \mathscr{D}$ and $(\kappa, P^*), (\gamma, P^*) \in \mathscr{K}(\hat{\succsim})$ such that $r = r_u(\rho, P^*)$, $r' = r_u(\tau, P^*)$, $q = r_u(\kappa, P^*)$ and $k = r_u(\gamma, P^*)$. Then, by Lemma 3, $(\rho_T\kappa_F\gamma, P^*) \succsim (\tau_T\kappa_F\gamma, P^*)$. From Axiom 13, we get $(\rho_T\gamma_F\kappa, P^*) \succsim (\tau_T\gamma_F\kappa, P^*)$. Applying Lemma 3 once more, we obtain $r_{(T)}k_{(F)}q \geq^u r'_{(T)}k_{(F)}q$. □

Lemma 17. If $\succsim$ satisfies Weak CRE-Independence, then $\geq^u$ satisfies:

- *Weak Certainty Independence:* if $r, r', q \in \mathscr{R}_u$ and q is constant, then $\alpha r + (1-\alpha)q \geq^u \alpha r' + (1-\alpha)q$ implies $\alpha r + (1-\alpha)q' \geq^u \alpha r' + (1-\alpha)q'$ for any $\alpha \in (0, 1)$ and constant $q' \in \mathscr{R}_u$.

Proof: Suppose $\alpha r + (1-\alpha)q \geq^u \alpha r' + (1-\alpha)q$, where q is constant and $\alpha \in (0, 1)$. By a familiar argument, there exist $\rho, \tau, \kappa \in \mathscr{D}$ such that $r = r_u(\rho, P^*)$, $r' = r_u(\tau, P^*)$, $q = r_u(\kappa, P^*)$ and $\alpha(\rho, P^*) + (1-\alpha)(\kappa, P^*) \succsim \alpha(\tau, P^*) + (1-\alpha)(\kappa, P^*)$. Applying Weak CRE-Independence, we get $\alpha(\rho, P^*) + (1-\alpha)(\kappa', P^*) \succsim \alpha(\tau, P^*) + (1-\alpha)(\kappa', P^*)$ for all $(\kappa', P^*) \in \mathscr{K}(\hat{\succsim})$. Since any constant $q' \in \mathscr{R}_u$ correspond to $r_u(\kappa', P^*)$ for some κ' (by Lemma 1), we obtain $\alpha r + (1-\alpha)q' \geq^u \alpha r' + (1-\alpha)q'$ for any $\alpha \in (0, 1)$ and constant $q' \in \mathscr{R}_u$. □

Lemma 18. If $\succsim$ satisfies the Sure Thing Principle, then $\geq^u$ satisfies:

- *Savage's STP:* for all $r, r', q, q' \in \mathscr{R}_u$ and $T \in \Sigma$, if $r_{(T)}q \geq^u r'_{(T)}q$ then $r_{(T)}q' \geq^u r'_{(T)}q'$.

Proof: Take $r, r', q, q' \in \mathscr{R}_u$ and $T \in \Sigma$ with $r_{(T)}q \geq^u r'_{(T)}q$. As usual, there exists $\rho, \tau, \gamma, \gamma' \in \mathscr{D}$ such that $r = r_u(\rho, P^*)$, $r' = r_u(\tau, P^*)$, $q = r_u(\gamma, P^*)$, $q' = r_u(\gamma', P^*)$ and, by Lemma 3, $(\rho_T\gamma, P^*) \succsim (\tau_T\gamma, P^*)$. The Sure Thing Principle then implies that $(\rho_T\gamma', P^*) \succsim (\tau_T\gamma', P^*)$, and applying Lemma 3 once again we obtain the result. □

Now we can proceed to the proofs of the propositions in Section 1.5. Again, I will show only sufficiency of the axioms, since necessity is easily verifiable.

Proof of Proposition 2: Assume that $\succsim$ satisfies axioms 1 to 4 and 6, Mixture Independence and Monotone Continuity*. Then $\geq^u$ satisfies the basic A-A axioms, Independence and Monotone Continuity. Applying the classic SEU representation theorem (see, e.g., Fishburn (1970, Theorem 13.3)),[1] the axioms on $\geq^u$ are equivalent to the existence a cardinally unique linear function $U : [0,1] \to \mathbb{R}$ and a finitely additive probability distribution $\pi \in \Delta(\Theta)$ such that, for all $r, r' \in \mathscr{R}_u$,

$$r \geq^u r' \iff \int_\Theta U \circ r \mathrm{d}\pi \geq \int_\Theta U \circ r' \mathrm{d}\pi.$$

Monotone Continuity then guarantees that $\pi \in \Delta(\Theta)$ is countably additive (Arrow, 1971). From the uniqueness properties of the representation and linearity, we can take U to be the identity function, thus the result obtains. □

Denote by $B(\Sigma)$ the set of bounded real-valued Σ-measurable functions on Θ. Let $B(\Sigma)^*$ be its continuous dual space, i.e., the space of all linear functionals $\phi : B(\Sigma) \to \mathbb{R}$ that are continuous with respect to the topology induced by the uniform norm $\|r\|_\infty = \sup_{\theta\in\Theta} |r(\theta)|$. A well known fact is that $B(\Sigma)^*$ is isomorphic to the set $ba(\Sigma)$ of finitely additive real-valued set functions on Σ which are bounded in the total variation norm.

[1] Although SEU theorems in the Anscombe-Aumann framework are usually stated in terms of acts mapping Θ to some set Δ of probability distributions, the same proofs also work when acts take values in Y, where Y is any convex set.

Recall that we assumed $\mathscr{R}_u$ was endowed with the product topology, rather than the uniform topology. So denote by $B(\Sigma)^\times$ the continuous dual space of $B(\Sigma)$ in the product topology, and note that, since this topology is coarser than the uniform topology, we have that $B(\Sigma)^\times \subseteq B(\Sigma)^*$ (because any functional that is continuous in the former topology is also in the latter). For the same reason, any weak* compact subset of $B(\Sigma)^*$ remains compact as a subset of $B(\Sigma)^\times$.

Proof of Proposition 3: In view of Lemmas 14 and 15, if $\succsim$ satisfies axioms 1 to 4 and 6, Hedging and CRE-Independence, then $\geq^u$ satisfies the basic A-A axioms, Uncertainty Aversion and Certainty Independence. Following the proof of the MPEU representation theorem from Gilboa and Schmeidler (1989a, Theorem 1 and Proposition 4.1), $\geq^u$ satisfies these axioms if, and only if, there exists a convex set $\Phi \subseteq B(\Sigma)^\times$ of monotone continuous linear functionals on $B(\Sigma)$ such that, for all $r, r' \in \mathscr{R}_u$,

$$r \geq^u r' \iff \inf_{\phi\in\Phi} \phi(r) \geq \inf_{\phi\in\Phi} \phi(r').$$

Since $B(\Sigma)^*$ is isomorphic to $ba(\Sigma)$ and $B(\Sigma)^\times \subseteq B(\Sigma)^*$, there also exists an isomorphism between Φ and $\Pi \subseteq ba(\Sigma)$. Monotonicity of each $\phi \in \Phi$ implies that every $\pi \in \Pi$ is a (finitely additive) measure. Therefore, for all $r, r' \in \mathscr{R}_u$,

$$r \geq^u r' \iff \inf_{\pi\in\Pi} \int_\Theta r \mathrm{d}\pi \geq \inf_{\pi\in\Pi} \int_\Theta r' \mathrm{d}\pi.$$

Lemma 3.3 in Gilboa and Schmeidler (1989a) then guarantees that $\phi(\mathbf{1}_\Theta) = 1$ for all $\phi \in \Phi$, thus $\pi(\Theta) = 1$ for every $\pi \in \Pi$. Finally, since $\succsim$ satisfies Monotone Continuity*, then Monotone Continuity of $\geq^u$ guarantees that $\Pi \subseteq \Delta(\Theta)$ (Chateauneuf et al., 2005, Theorem 1). □

Many models in the Anscombe-Aumann setting are characterized only for simple acts, i.e., acts that have a finite range. To make use of such results, we first prove that the propositions in Section 1.5 hold when restricted to the set $\mathscr{R}_u^0 \equiv \{r \in \mathscr{R}_u : r(\Theta) \subseteq F,\ F \text{ finite}\}$ of all simple risk functions. Then we use standard results and continuity of I to approximate each $r \in \mathscr{R}_u$ by a sequence of functions in $(r_n^0)_{n\geq1} \in \mathscr{R}_u^0$. This leads to the lemma below.

Lemma 19. Suppose $\succsim$ has an MRA representation with utility function u, and $I_0 : \mathscr{R}_u^0 \to \mathbb{R}$ represents the restriction of $\geq^u$ to $\mathscr{R}_u^0$. Then I_0 has a unique continuous extension $I : \mathscr{R}_u \to \mathbb{R}$ and (u, I) represents $\succsim$.

Proof: Since $\succsim$ has an MRA representation, Theorem 2 applied to $\mathscr{R}_u^0$ implies that I_0 is monotone and continuous. Because I_0 represents the restriction of $\geq^u$ to $\mathscr{R}_u^0$, there exists (u, I) representing $\succsim$ such that I_0 is a restriction of I to $\mathscr{R}_u^0$. A well known result in functional analysis is that for any measurable function $r \in \mathscr{R}_u$, there exists a sequence $(r_n)_{n\geq 1} \in \mathscr{R}_u^0$ such that $r_n \leq r$ for all $n \geq 1$ and $r_n \to r$. Therefore, $\mathscr{R}_u^0$ is dense in $\mathscr{R}_u$.

Take any two continuous extensions $\hat{I}$ and $\tilde{I}$ of I_0. Fix an arbitrary $r \in \mathscr{R}_u \setminus \mathscr{R}_u^0$ and suppose that $\hat{I}(r) \neq \tilde{I}(r)$. Since $\mathbb{R}$ is Hausdorff, there exist open neighborhoods $\hat{N}$ of $\hat{I}(r)$ and $\tilde{N}$ of $\tilde{I}(r)$ such that $\hat{N} \cap \tilde{N} = \emptyset$. Since $\hat{I}$ and $\tilde{I}$ are continuous, $\hat{I}^{-1}(\hat{N})$ and $\tilde{I}^{-1}(\tilde{N})$ are open neighborhoods of r. However, $\hat{I}^{-1}(\hat{N}) \cap \tilde{I}^{-1}(\tilde{N}) \cap \mathscr{R}_u^0 = I_0^{-1}(\hat{N}) \cap I_0^{-1}(\tilde{N}) \cap \mathscr{R}_u^0 = I_0^{-1}(\hat{N} \cap \tilde{N}) \cap \mathscr{R}_u^0 = \emptyset$, which is absurd. We conclude that $\hat{I}(r) = \tilde{I}(r)$ for all $r \in \mathscr{R}_u$. Therefore, I is the unique continuous extension of I_0 and (u, I) represents $\succsim$. □

Proof of Proposition 4: Since $\succsim$ has a MPEU representation, $\geq^u$ satisfies the basic A-A axioms and Uncertainty Aversion. By Stoye (2011, Theorem 1(iii)), the restriction of $\geq^u$ to $\mathscr{R}_u^0$ satisfies Risk Symmetry if, and only if, it has a MEU representation. Using Lemma 19, we can then extend the representation to $\mathscr{R}_u$. □

Proof of Proposition 5: Suppose $\succsim$ satisfies axioms 1 to 4, 6 and 10, Hedging and Weak CRE-Independence. Then $\geq^u$ satisfies the basic A-A axioms, Uncertainty Aversion and Weak Certainty Independence, and so does the restriction of $\geq^u$ to $\mathscr{R}_u^0$. According to Maccheroni, Marinacci, and Rustichini (2006, Appendix B), this holds if, and only if, there exists a monotone functional $I_0 : \mathscr{R}_u^0 \to \mathbb{R}$ that represents $\geq^u$ on $\mathscr{R}_u^0$ and satisfies, for all $r, q \in \mathscr{R}_u^0$ and $\alpha, k \in [0, 1]$,

(a) *Translation Invariance:* $I_0(\alpha r + (1 - \alpha)k\,\mathbf{1}_\Theta) = I_0(\alpha r) + (1 - \alpha)k$;

(b) *Normalization:* $I_0(k\,\mathbf{1}_\Theta) = k$;

(c) *Concavity:* $I_0(\alpha r + (1-\alpha)q) \geq \alpha I_0(r) + (1-\alpha)I_0(q)$.

Moreover, by Theorem 2 and the fact that, for a given u, $I_0 : \mathscr{R}_u^0 \to \mathbb{R}$ is unique up to normalization, we have that I_0 is continuous in the topology of point-wise convergence.

Define $I : \mathscr{R}_u \to \mathbb{R}$ as the unique continuous extension of I_0. By Lemma 19, (u, I) represents $\gtrsim$. Clearly, I satisfies (b). To see that it satisfies (a), let $r \in \mathscr{R}_u$ and take any $(r_n) \in \mathscr{R}_u^0$ such that $r_n \to r$. Then $I(\alpha r + (1-\alpha)k\,\mathbf{1}_\Theta) = I(\lim_n \alpha r_n + (1-\alpha)k\,\mathbf{1}_\Theta) = \lim_n I_0(\alpha r_n + (1-\alpha)k\,\mathbf{1}_\Theta) = \lim_n I_0(\alpha r_n) + (1-\alpha)k = I(\alpha r) + (1-\alpha)k$, which implies that I is translation invariant. A similar argument implies that I is concave.

Applying the Fenchel-Moreau theorem as in Cerreia-Vioglio et al. (2014, Section 5), we can identify $I : \mathscr{R}_u \to \mathbb{R}$ with its convex biconjugate $I(r) = \inf_{\phi \in B(\Sigma)^\times} \{\phi(r) - I^*(\phi)\}$, where $\phi \mapsto I^*(\phi) = \inf_{r \in \mathscr{R}_u}\{\phi(r) - I(r)\}$ is concave, upper semicontinuous and $\sup_{\phi \in B(\Sigma)^\times} I^*(\phi) = 0$. Finally, recall that $B(\Sigma)^\times \subseteq B(\Sigma)^*$, thus $B(\Sigma)^\times$ is isomorphic to a subset $\Pi_0 \subseteq ba(\Sigma)$. Defining $c^* : \Pi_0 \to [0, \infty]$ by $c^*(\pi) = -\inf_{r \in \mathscr{R}_u}\{\int r\mathrm{d}\pi - I(r)\}$, we can thus write

$$I(r) = \inf\left\{\int r\mathrm{d}\pi + c^*(\pi) : \pi \in \Pi_0\right\} \quad \forall r \in \mathscr{R}_u.$$

From Monotone Continuity and Normalization, we have $c^*(\pi) = \infty$ for all $\pi \in \Pi_0 \setminus \Delta(\Theta)$ (see Maccheroni, Marinacci, and Rustichini (2006, Lemma 30)). Letting $\Pi = \{\pi \in \Pi_0 : c^*(\pi) < \infty\}$ and defining $c : \Pi \to [0, \infty)$ by the restriction of c^* to Π, we obtain the result. □

Proof of Proposition 6: Since $\gtrsim$ has a variational representation and satisfies Axiom 15, then $\geq^u$ satisfies the basic A-A axioms, Uncertainty Aversion, Weak Certainty Independence and Savage's STP. Strzalecki (2011, Theorem 1) implies that the restriction of $\geq^u$ to $\mathscr{R}_u^0$ satisfies the aforementioned axioms if, and only if, it has a multiplier representation (u, I_0), where $I_0(r) = \min_{\pi \in \Delta(\Theta)} \left\{\int_\Theta r\mathrm{d}\pi + kD(\pi \| \mu)\right\}$. From the SEOU representation of I_0 found in Strzalecki (2011), i.e., $I_0(r) = \int_\Theta -\exp(-k^{-1}r)\mathrm{d}\mu$ for $k < \infty$ and $\int_\Theta r\mathrm{d}\mu$ for $k = \infty$, it is easily checked that I_0 is continuous in the topology of point-wise convergence. We can thus extend I_0 to $I : \mathscr{R}_u \to \mathbb{R}$ using the same technique as in the proof of Proposition 5. □

A.6 Proofs of results in Section 1.6

A.6.1 Proof of Theorem 3

If $\succsim \subseteq \mathscr{S}^2$ satisfies axioms 1 to 6, then Lemma 10 implies that the relation $\geq$ defined by (1.8) is a weak order that satisfies axioms A3 to A6. Then it is a straightforward exercise to check that the state dependent stochastic choice function $c : \Phi \to \mathscr{F}$ given by $c(F) = \{f \in F : f \geq g \ \forall g \in F\}$ for all $F \in \Phi$, satisfies axioms C1 to C4.

Now assume $c : \Phi \to \mathscr{F}$ satisfies axioms C1 to C4. We want to construct a complete extension $\geq$ of $\trianglerighteq$ that satisfies axioms A3 to A6, which will allow us to define a preference $\succsim$ on $\mathscr{S}$ that rationalizes c and satisfies axioms 1 to 6.

Because $\trianglerighteq_\theta$ satisfies C-LARP and Conditional Continuity, we can apply Theorem 4 from Clark (1993) and Theorem 3 from Clark (2000) to obtain a continuous total order (a complete, transitive and antisymmetric preference) $\geq_\theta$ on $\Delta(A)$ that extends $\trianglerighteq_\theta$ and satisfies Independence. That is, if $p \geq_\theta q$, then $\lambda p + (1-\lambda)r \geq_\theta \lambda q + (1-\lambda)r$ for all $\lambda \in [0,1]$ and $r \in \Delta(A)$. By RP-Monotonicity, $f = c(F)$ implies that for all $g \in F$ with $g \neq f$, there exists $\theta \in \Theta$ such that $p \not\trianglerighteq_\theta f(\theta)$ for all $p \in \Delta(A)$ such that $g(\theta) \trianglerighteq_\theta p$. Therefore, a well-known argument (see, e.g., Ok and Riella (2021), Theorem 9.3) guarantees that $\{\geq_\theta : \theta \in \Theta\}$ can be constructed so that, if $f = c(F)$, $g \in F$ and $g \neq f$, there is a $\theta \in \Theta$ such that $f(\theta) \geq_\theta g(\theta)$.

Define, for all $f, g \in \mathscr{F}$, $f \mathrel{\hat{\geq}} g$ if, and only if, $f(\theta) \geq_\theta g(\theta)$ for every $\theta \in \Theta$. It is easily checked that, by construction, $\hat{\geq}$ is a partial order that satisfies axioms A3 to A5, A7 and A8. Suppose Θ is finite, so that $\mathscr{F}$ is metrizable. If $(f^m), (g^m) \in \mathscr{F}$ are such that $f^m \to f$, $g^m \to g$ and $f^m \mathrel{\hat{\geq}} g^m$ for all $m \in \mathbb{N}$, then $f^m(\theta) \to f(\theta)$ and $g^m(\theta) \to g(\theta)$ for all $\theta \in \Theta$, which by definition of $\hat{\geq}$ and the fact that $\geq_\theta$ is closed for all $\theta \in \Theta$, implies that $f \mathrel{\hat{\geq}} g$. Therefore, $\hat{\geq}$ satisfies axiom A6.

From GARP and the fact that we are working with choice functions, if $f^1 \in c(F_1) \cap F_2, f^2 \in c(F_2) \cap F_3, \ldots, f^k \in c(F_k) \cap F_1$, then $f^1 = \cdots = f^k = f$. Moreover, by construction, $g \mathrel{\hat{\not\geq}} f$ for any $g \in \bigcup_{i=1}^k F_k \setminus \{f\}$. In the language of Nishimura, Ok, and Quah (2017), $((\mathscr{F}, \hat{\geq}), \Phi)$ is a continuous choice environment and c satisfies cyclical $\hat{\succeq}$-consistency. Therefore, we can apply their Theorem 1 to obtain a weak order $\succeq$ that extends $\hat{\succeq}$ and, for all $F \in \Phi$,

$$c(F) = \{f \in F : f \succeq g \; \forall g \in F\}.$$

Furthermore, since $\succeq$ extends $\hat{\succeq}$, it is easy to see that it satisfies axioms A3 to A5.

Finally, define $\succsim$ on $\mathscr{S}$ by

$$\forall (\rho, P), (\tau, Q) \in \mathscr{S}, \; (\rho, P) \succsim (\tau, Q) \iff \rho P \succeq \tau Q.$$

Obviously, $\succsim$ satisfies Consequentialism. It is easy to check, using arguments similar to the proof of Lemma 10, that $\succsim$ also satisfies axioms 1 and 3 to 6. Therefore, by Theorem 2, it has an MRA representation (u, I), which completes the proof.

A.7 Proof of Theorem 4

Suppose $\succcurlyeq_{.}$ satisfies Axiom D1. Note that item 1 of Axiom D1 implies that each $\succcurlyeq_P$ satisfies a version of Consequentialism. Indeed, taking $P = P'$ in the statement, we have that $\rho P = \tau P$ implies $\rho \sim_P \tau$.

Consider the following relation on $\mathscr{D}$:

$$\forall \rho, \tau \in \mathscr{D} : \rho \mathrel{\hat{\succcurlyeq}} \tau \iff \rho \succcurlyeq_\theta \tau \text{ for all } \theta \in \Theta.$$

Since $\succcurlyeq_{P^*}$ is reflexive, $\hat{\succcurlyeq}$ is non-empty. Also define a preference relation $\hat{\succeq}$ on $\mathscr{F}$ by

$$f \mathrel{\hat{\succeq}} g \iff \exists \rho, \tau \in \mathscr{D} \text{ such that } \rho \mathrel{\hat{\succcurlyeq}} \tau, \; \rho P^* = f \text{ and } \tau P^* = g.$$

Substituting $\hat{\succeq}$ for $\succeq$ in the proof of Lemma 10, it can be seen that $\hat{\succeq}$ is reflexive, transitive, and satisfies axioms A3 to A8. Therefore, substituting $\hat{\succeq}$ for $\succeq$ in the Proof of Theorem 1, we obtain a dominance representation with utility index $v : A \times \Theta \to \mathbb{R}$, i.e., for all $\rho, \tau \in \mathscr{D}$,

$$\rho \mathrel{\hat{\succcurlyeq}} \tau \iff \int_X v(\rho, \theta) \mathrm{d}P^*_\theta \geq \int_X v(\tau, \theta) \mathrm{d}P^*_\theta \ \ \forall \theta \in \Theta.$$

Note that, due to Lemma 1, we have $\mathscr{R}_v = \{r_v(\delta, P^*) : \delta \in \mathscr{D}\}$. Define a preference $\geq_*$ on $\mathscr{R}_v$ by

$$\forall r, r' \in \mathscr{R}_v : \ r \geq_* r' \iff \exists \rho \succcurlyeq_{P^*} \tau \text{ such that } r = r_v(\rho, P^*) \text{ and } r' = r_v(\tau, P^*).$$

An argument entirely analogous to the Proof of Theorem 2 guarantees that $\geq_*$ is a continuous and monotone weak order. Therefore, there exists a monotone and continuous utility function $J : \mathscr{R}_v \to \mathbb{R}$ such that, for all $\rho, \tau \in \mathscr{D}$,

$$r_v(\rho, P^*) \geq_* r_v(\tau, P^*) \iff J(r_v(\rho, P^*)) \geq J(r_v(\tau, P^*)) \iff \rho \succcurlyeq_{P^*} \tau.$$

We now extend the representation to the collection $\succcurlyeq_\cdot = \{\succcurlyeq_P : P \in \mathscr{P}\}$. Take any $P \in \mathscr{P}$ and $\rho, \tau \in \mathscr{D}$. By Lemma 1, there exists $\rho', \tau' \in \mathscr{D}$ such that $\rho P = \rho' P^*$ and $\tau P = \tau' P^*$. From item 1 of Axiom D1, we obtain $\rho \succcurlyeq_P \tau \iff \rho' \succcurlyeq_{P^*} \tau'$. Therefore, $\rho \succcurlyeq_P \tau \iff J(r_v(\rho', P^*)) \geq J(r_v(\tau', P^*))$. Since $r_v(\rho', P^*) = r_v(\rho, P)$ and $r_v(\tau', P^*) = r_v(\tau, P)$, we have

$$\rho \succcurlyeq_P \tau \iff J(r_v(\rho, P)) \geq J(r_v(\tau, P)),$$

for all $\rho, \tau \in \mathscr{D}$ and $P \in \mathscr{P}$.

A.8 Proof of Theorem 5

Assume $\succcurlyeq^\cdot$ satisfies axioms E1 to E3. By Lemma 9, there exists $\delta^* \in \overline{\mathscr{D}}$ such that for any $f \in \mathscr{F}$, there is a $P^f \in \Delta(X)^\Theta$ such that $f = \delta^* P^f$.

For any given $f, h \in \mathscr{F}$, take $P^f, P^h \in \mathscr{P}$ such that $f = \delta^* P^f$ and $g = \delta^* P^h$. Since δ^* is invariant, we have:

1. For all $\alpha \in [0,1]$, $\delta^*(\alpha P^f + (1-\alpha)P^h) = \alpha\delta^* P^f + (1-\alpha)\delta^* P^h = \alpha f + (1-\alpha)h$;

2. For all $\theta \in \Theta$, $\delta^*(P^f{}_{(\theta)}P^h) = \delta^* P^f{}_{(\theta)}\delta^* P^h = f_{(\theta)}h$.

Define the preference $\succeq^*$ on $\mathscr{F}$ by

$$f \succeq^* h \iff \exists P^f, P^h \in \mathscr{P} \text{ such that } P^f \succcurlyeq^{\{\delta^*\}} P^h,\ \delta^* P^f = f \text{ and } \delta^* P^h = h.$$

Consistency guarantees that $\succeq^*$ is well-defined and reflexive, since by taking $M = N = \{\delta^*\}$, $Q = P'$ and $P = Q'$ in the statement of the axiom, we get that $\delta^* P = \delta^* P'$ implies $P \sim^{\{\delta^*\}} P'$. Item 1 of Axiom E1 and Lemma 9 immediately imply that $\succeq^*$ is complete and transitive, whereas item 5 readily implies that $\succeq^*$ satisfies Axiom A6. We now show that $\succeq^*$ satisfies axioms A3 to A6.

Indeed, take any $f, g, h, h' \in \mathscr{F}$ and $\theta \in \Theta$ such that $f_{(\theta)}h \succeq^* g_{(\theta)}h$. Then, $P^f{}_{(\theta)}P^h \succcurlyeq^{\{\delta^*\}} P^g{}_{(\theta)}P^h$ for some $\delta^* P^f = f$, $\delta^* P^h = h$ and $\delta^* P^g = g$. Item 2 of Axiom E1 then implies that $P^f{}_{(\theta)}R \succcurlyeq^{\{\delta^*\}} P^g{}_{(\theta)}R$ for all $R \in \mathscr{P}$, and thus $f_{(\theta)}h' \succeq^* g_{(\theta)}h'$, by Lemma 9. Thus $\succeq^*$ satisfies Axiom A3.

Now fix any $f, g, h \in \mathscr{F}$ such that $f_{(\theta)}h \succeq^* g_{(\theta)}h$ for all $\theta \in \Theta$. This implies that there exist $P^f, P^g, P^h \in \mathscr{P}$, where $\delta^* P^f = f$, $\delta^* P^g = g$ and $\delta^* P^h = h$, such that $P^f{}_{(\theta)}P^h \succcurlyeq^{\{\delta^*\}} P^g{}_{(\theta)}P^h$ for all $\theta \in \Theta$. By item 2 in Axiom E1, we have that $P^f{}_{(\theta)}R \succcurlyeq^{\{\delta^*\}} P^g{}_{(\theta)}R$ for all $\theta \in \Theta$ and $R \in \mathscr{P}$. From item 3, we obtain $P^f \succcurlyeq^{\{\delta^*\}} P^g$, which implies $f \succeq^* g$. Therefore, $\succeq^*$ satisfies Axiom A4.

Next fix any $\theta \in \Theta$ and take any $f, g, h \in \mathscr{F}$ such that $f(\theta') = g(\theta') = h(\theta')$ for all $\theta' \neq \theta$. Then there exist $P^f, P^g, P^h \in \mathscr{P}$ such that $P^f_{\theta'} = P^g_{\theta'} = P^h_{\theta'}$ for all $\theta' \neq \theta$, $f = \delta^* P^f$, $g = \delta^* P^g$ and $h = \delta^* P^h$. Suppose $f \succeq^* g$, implying $P^f \succcurlyeq^{\{\delta^*\}} P^g$. By item 4 of Axiom E1, we have $\alpha P^f + (1-\alpha)P^h \succcurlyeq^{\{\delta^*\}} \alpha P^g + (1-\alpha)P^h$ for all $\alpha \in (0,1]$, and thus $\alpha f + (1-\alpha)h \succeq^* \alpha g + (1-\alpha)h$. We conclude that $\succeq^*$ satisfies Axiom A5.

For each $\theta \in \Theta$, define the conditional preferences $\succeq_\theta$ as in (A.1). Since $\succeq^*$ is complete, so are $\succeq_\theta$. Section A.2 then shows that each $\succeq_\theta$ has a representation given by (A.2). Fix one such representation $w : A \times \Theta \to \mathbb{R}$ and define its associated risk functional $\tilde{r}_w : \mathscr{F} \to \mathbb{R}^\Theta$ as in eq. (A.4). Consider the set of all such risk functions $\mathscr{R}_w = \{\tilde{r}_w(f) : f \in \mathscr{F}\}$ and define a preference relation $\geq^w$ on $\mathscr{R}_w$ by $r \geq^w r' \iff \exists f, g \in \mathscr{F}$ such that $f \geq^* g$, $r = \tilde{r}_w(f)$ and $r' = \tilde{r}_w(g)$. Note that if $f = \delta^* P$, for any $f \in \mathscr{F}$ and $P \in \mathscr{P}$, then $\tilde{r}_w(f) = r_w(\delta^*, P)$. A straightforward adaptation of the argument in Section A.4 then implies that there exist $w : A \times \Theta \to \mathbb{R}$, continuous in the first argument, and a monotone and continuous $H : \mathscr{R}_w \to \mathbb{R}$, such that

$$P \succcurlyeq^{\{\delta^*\}} Q \iff H(r_w(\delta^*, P)) \geq H(r_w(\delta^*, Q)).$$

Moreover, we clearly have $C_P(\{\delta^*\}) = \{\delta^*\}$ for all $P \in \mathscr{P}$, hence $\succcurlyeq^{\{\delta^*\}}$ has the desired representation.

It remains to show that $\succcurlyeq^M$ has such a representation for each $M \in \mathscr{M}$, and that these representations coincide. To that end, first consider any $M \in \mathscr{M}$ such that $C_P(M)$ is a singleton for all $P \in \mathscr{P}$. Take any $P, Q \in \mathscr{P}$ and let $\{\rho\} = C_P(M)$ and $\{\tau\} = C_Q(M)$. By Lemma 9, there exist $P', Q' \in \mathscr{P}$ such that $\rho P = \delta^* P'$ and $\tau Q = \delta^* Q'$. Consistency then implies that $P \succcurlyeq^M Q \iff P' \succcurlyeq^{\{\delta^*\}} Q'$, and thus

$$P \succcurlyeq^M Q \iff H(r_w(\delta^*, P')) \geq H(r_w(\delta^*, Q')) \iff H(r_w(\rho, P)) \geq H(r_w(\tau, Q)).$$

We now extend the representation to menus M with $|C_P(M)| > 1$ for some $P \in \mathscr{P}$. Take any $M \in \mathscr{M}$ and $P \in \mathscr{P}$. By Lemma 9, for all $\rho \in C_P(M)$ there exists $P^\rho \in \mathscr{P}$ such that $\rho P = \delta^* P^\rho$. Since $\succcurlyeq^{\{\delta^*\}}$ is complete and transitive, there exists $\rho^*(P) \in C_P(M)$ such that $P^{\rho^*(P)} \succcurlyeq^{\{\delta^*\}} P^\rho$ for all $\rho \in C_P(M)$. Define a decision rule $\bar{\delta}$ by $\bar{\delta}_P = \rho^*(P)_P$ for every $P \in \mathscr{P}$ and note that $\bar{\delta} P = \rho^*(P) P$. Thus by Axiom E2, $P \succcurlyeq^{\{\bar{\delta}\}} Q \iff P \succcurlyeq^{\tilde{M}(P,Q)} Q$ for all $P, Q \in \mathscr{P}$, where $\tilde{M}(P, Q) = \{\delta \in \mathscr{D} : \delta_P = \rho^*(P)_P,\ \delta_Q = \rho^*(Q)_Q$ and $\forall R \neq P,\ \exists \gamma \in M$ s.t. $\delta_R = \gamma_R\}$. Applying Axiom E3 twice, we obtain the following equivalences:

$$P \succcurlyeq^{\{\bar{\delta}\}} Q \iff P \succcurlyeq^{\tilde{M}(P,Q)} Q \iff P \succcurlyeq^M Q.$$

Finally, note that since $r_w(\rho^*(P), P) = r_w(\bar{\delta}, P)$ for all $P \in \mathscr{P}$ and $\succcurlyeq^{\{\bar{\delta}\}}$ can be represented by

(w, H) we get that for all $P, Q \in \mathscr{P}$, $P \succcurlyeq^M Q \iff H(r_w(\rho^*(P), P)) \geq H(r_w(\rho^*(Q), Q))$ subject to $\rho^*(P) \in C_P(M)$, $\rho^*(Q) \in C_Q(M)$.

Appendix B: Proof of results in Chapter 2

B.1 Proof of Lemma 6

For any matrix $\mathbf{M}$, let $\Delta(\mathbf{M})$ denote the convex hull of the columns of $\mathbf{M}$. Also let Δ^I be the $(I-1)$-dimensional unit simplex. Note that if $\mathbf{P} = \mathbf{Q}\boldsymbol{\Lambda}$ is a stochastic matric factorization, then $\Delta(\mathbf{P}) \subseteq \Delta(\mathbf{Q})$. Moreover, since $\mathbf{Q}$ is column-stochastic, $\Delta(\mathbf{Q}) \subseteq \Delta^I$. This implies that if $\Delta^I \cap \Delta(\mathbf{P}) \subseteq \Delta(\mathbf{P})$, then $\Delta(\mathbf{Q}) = \Delta(\mathbf{P})$ and the stochastic matrix factorization is unique.

Since $\mathbf{Q}$ has K linearly independent columns, $\Delta(\mathbf{Q})$ is a polytope with K vertices inscribed in Δ^I. Now note that K-sparsity implies that K points of $\Delta(\mathbf{P})$ are on different facets of Δ^I. Moreover, since $\Delta(\mathbf{P}) \subseteq \Delta^I$, these must be extreme points of $\Delta(\mathbf{P})$. This immediately implies that $\Delta^I \cap \Delta(\mathbf{P}) \subseteq \Delta(\mathbf{P})$, since every column of $\mathbf{P}$ is a convex combination of at most K extreme points, due to the assumption that $\mathbf{P}$ has a stochastic matrix factorization of rank K.

B.2 Proof of Corollary 1

Suppose $\mathbf{P}$ has a unique stochastic factorization of rank $K = 2$. By Lemma 5, there exist $i_1, i_2 \in \{1, \ldots, I\}$ and $j_1, j_2 \in \{1, \ldots, J\}$, $i_1 \neq i_2$ and $j_1 \neq j_2$, such that $\mathbf{Q}_{i_l k_h}, \boldsymbol{\Lambda}_{k_l j_h} > 0$ if $l = h$ and $\mathbf{Q}_{i_l k_h} = \boldsymbol{\Lambda}_{i_l k_h} = 0$ if $l \neq h$. We then have

$$\begin{cases} \mathbf{P}_{i_1 j_1} = \mathbf{Q}_{i_1 k_1}\boldsymbol{\Lambda}_{k_1 j_1} + \mathbf{Q}_{i_1 k_2}\boldsymbol{\Lambda}_{k_2 j_1} > 0, \\ \mathbf{P}_{i_1 j_2} = \mathbf{Q}_{i_1 k_1}\boldsymbol{\Lambda}_{k_1 j_2} + \mathbf{Q}_{i_1 k_2}\boldsymbol{\Lambda}_{k_2 j_2} = 0, \\ \mathbf{P}_{i_2 j_1} = \mathbf{Q}_{i_2 k_1}\boldsymbol{\Lambda}_{k_1 j_1} + \mathbf{Q}_{i_2 k_2}\boldsymbol{\Lambda}_{k_2 j_1} = 0, \\ \mathbf{P}_{i_2 j_2} = \mathbf{Q}_{i_2 k_1}\boldsymbol{\Lambda}_{k_1 j_2} + \mathbf{Q}_{i_2 k_2}\boldsymbol{\Lambda}_{k_2 j_2} > 0. \end{cases}$$

Therefore, $\mathscr{I}_{i_1} \not\subseteq \mathscr{I}_{i_2}$ and $\mathscr{I}_{i_2} \not\subseteq \mathscr{I}_{i_1}$, implying that $\mathbf{P}$ is K-sparse.

B.3 Proof of Theorem 6

Recall the following definition:

Definition 13 (Ring of sets). For a given set $\mathbb{S}$, we call $\mathscr{S} \subseteq 2^{\mathbb{S}}$ a ring if it satisfies the following properties:

(i) $\emptyset \in \mathscr{S}$;

(ii) For all $A, B \in \mathscr{S}$, we have $A \cup B \in \mathscr{S}$;

(iii) For all $A, B \in \mathscr{S}$, $A \setminus B \in \mathscr{S}$. ◊

Carathéodory's extension theorem states that a pre-measure on a ring $\mathscr{S}$ can be uniquely extended to the σ-algebra $\sigma(\mathscr{S})$ generated by it.

Now take any finite partition $\mathscr{C}$ of $\mathbb{Y} \times \mathbb{X}$ such that the corresponding mixture probabilities restricted to $\mathscr{C}$ satisfy K-sparsity. By Lemma 6, there exists a unique stochastic factorization $\mathbf{P} = \mathbf{Q}\mathbf{\Lambda}$, which implies that both $\mathbf{Q}$ and $\mathbf{\Lambda}$ are K-sparse. Therefore, by the monotonicity property of probabilities, for any partition $\mathscr{C}'$ which is finer than $\mathscr{C}$, the probabilities in Q and λ restricted to $\mathscr{C}'$, $\mathbf{Q}'$ and $\mathbf{\Lambda}'$, will also be K-sparse.

Now denote by $\mathscr{S}(\mathscr{C})$ the ring generated by the family of all partitions which are finer than $\mathscr{C}$. Since $P, Q_1, \ldots, Q_K$ are pre-measures when restricted to every set in the generating class of $\mathscr{S}(\mathscr{C})$, we can apply Carathéodory's extension theorem to guarantee their uniqueness. Therefore, both $Q_1, \ldots, Q_K$ and P are unique, which implies that $\lambda_1, \ldots, \lambda_K$ are also unique.

B.4 Proof of Proposition 7

Since Θ is compact and $\mathbb{I} : \Theta \rightrightarrows \mathbb{R}^{I \times K} \times \mathbb{R}^{K \times J}$ has a closed graph by Lemma 4, we have that $\mathbb{I}$ is upper-hemicontinuous. Then a standard result, which can be found in Ok (2011, page 304), states that for any sequence $(\theta^m) \in \Theta$ such that $\theta^m \to \theta$, we have

$$\max\{d((\mathbf{Q}, \mathbf{\Lambda}), \mathbb{I}(\theta)) : (\mathbf{Q}, \mathbf{\Lambda}) \in \mathbb{I}(\theta^m)\} \to 0.$$

The result follows immediately.

B.5 Proof of Theorem 7

That preferences which are represented by (u, π) satisfy the MRA axioms, Mixture Independence and Monotone Continuity* is easily checked by applying Proposition 2.

For any choice of T, we have $\mathbb{E}_\pi[u((1,T),\theta)|p] = -c_1$ and

$$\begin{aligned}\mathbb{E}_\pi[u((0,T),\theta)|p] &= \int_\Theta \int_\Theta u(1,\theta)\pi(\theta|p)\mathrm{d}\theta \\ &= -c_2 + \int_T c_2 - c_3 d(\theta,\Theta_0)\pi(\theta|x)\mathrm{d}\theta \\ &= -c_2 + c_2 Pr(\theta \in T|p) - c_3 \int_T d(\theta,\Theta_0)\pi(\theta|x)\mathrm{d}\theta.\end{aligned}$$

Therefore,

$$\mathbb{E}_\pi[u(1,\theta)|p] > \mathbb{E}_\pi[u(0,\theta)|p] \iff Pr(\theta \in T|p) < 1 - \frac{c_1 - c_3 \int_T d(\theta,\Theta_0)\pi(\theta|p)\mathrm{d}\theta}{c_2}.$$

Now note that the choice of T is not relevant when $a = 1$. Conditional on accepting the null hypothesis:

$$\mathbb{E}_\pi[u((0,T),\theta)|p] = -c_2 + \int_T c_2 - c_3 d(\theta,\Theta_0)\pi(\theta|x)\mathrm{d}\theta.$$

This is maximized by choosing the region T such that $c_2 - c_3 d(\theta,\Theta_0) \geq 0$ for all $\theta \in T$. That is,

$$T = \left\{\theta : d(\theta,\Theta_0) \leq \frac{c_2}{c_3}\right\}.$$

Appendix C: Proof of results in Chapter 3

C.1 Existence of optimal mechanisms for arbitrary type distributions

Lemma 20. $\mathscr{M}^*(F) \neq \emptyset$ for all $F \in \mathbb{F}$.

Proof: Define $\bar{p} := \max_{(\theta,x)\in\Theta\times X} v(\theta,x)$ and endow the power set of $X\times[0,\bar{p}]$ with the Hausdorff metric. It is well-known that with the topology induced by the Hausdorff metric, the set of all compact subsets of $X \times [0,\bar{p}]$, denoted by $\mathscr{K}(X\times[0,\bar{p}])$, is itself compact. Define the set of mechanisms $\mathscr{M}_{IR}$ as

$$\mathscr{M}_{IR} := \left\{ M \in \mathscr{M} : \max_{(x,p)\in M} p \leq \bar{p} \right\} \subseteq \mathscr{K}(X\times[0,\bar{p}]).$$

Clearly, $\mathscr{M}_{IR}$ is closed, and therefore also compact. Since $v(\cdot,\cdot)$ is continuous and clearly so is the mapping $\Theta \times \mathscr{M}_{IR} \ni (\theta, M) \mapsto M \cup (0,0)$, we have by the Maximum Theorem that the correspondence $\mathscr{M}_{IC} : \Theta\times\mathscr{M}_{IR} \rightrightarrows X\times[0,\bar{p}]$ such that

$$(\theta, M) \mapsto \mathscr{M}_{IC}(\theta, M) = \operatorname*{arg\,max}_{(x,p)\in M\cup(0,0)} v(\theta,x) - p$$

is non-empty, compact valued and upper hemicontinuous. Applying the Maximum Theorem once again, we obtain that the value function $\gamma : \Theta\times\mathscr{M}_{IR} \to \mathbb{R}$ where

$$(\theta, M) \mapsto \gamma(\theta, M) = \max_{(x,p)\in\mathscr{M}_{IC}(\theta,M)} p - c(x)$$

is upper semicontinuous, i.e., for any $(\theta_0, M_0) \in \Theta\times\mathscr{M}_{IR}$, $\limsup_{(\theta,M)\to(\theta_0,M_0)} \gamma(\theta, M) \leq \gamma(\theta_0, M_0)$.

Now, since γ is bounded we can apply (reverse) Fatou's Lemma to obtain that for any $M_0 \in \mathscr{M}_{IR}$,

$$\limsup_{M_n \to M_0} \int \gamma(\theta, M_n) dF(\theta) \leq \int \limsup_{M_n \to M_0} \gamma(\theta, M_n) dF(\theta) \leq \int \gamma(\theta, M_0) dF(\theta).$$

This proves that the mapping $(F, M) \mapsto \int \gamma(\cdot, M) dF$ is upper semicontinuous in M for every $F \in \mathbb{F}$. Therefore, since $\mathscr{M}_{IR}$ is compact, the extreme value theorem guarantees the existence of $M^* \in \arg\max_{M \in \mathscr{M}_{IR}} \pi(M, F) = \arg\max_{M \in \mathscr{M}} \pi(M, F)$ for all $F \in \mathbb{F}$, where the latter equality comes from the fact that $(x, p) \in X \times (\bar{p}, \infty)$ will never be chosen over $(0, 0)$. □

C.2 Proof of Lemma 7

We first state two results that will prove useful in determining the Lipschitz continuity of $\pi(M, F)$ in $F \in \mathbb{F}$.

Let $\mathscr{P} := \left\{P = \theta_0, \ldots, \theta_{n_P} : \underline{\theta} = \theta_0 \leq \theta_1 \leq \cdots \leq \theta_{n_P} = \bar{\theta}\right\}$ and $V(f) := \sup_{P \in \mathscr{P}} \sum_{i=0}^{n_P - 1} |f(\theta_{i+1}) - f(\theta_i)|$, where $V(f)$ denotes the total variation of a function $f : \Theta \to \mathbb{R}$. We have:

Lemma 21 (Beesack-Darst-Pollard Inequality (Darst and Pollard, 1970; Beesack, 1975)). Let f, g, h be real-valued functions on a compact interval $I = [a, b] \subset \mathbb{R}$, where h is of bounded variation with total variation $V(h)$ on I and such that $\int_a^b f dg$ and $\int_a^b h dg$ both exist. Then,

$$m \int_a^b f dg + V(h) \sup_{a \leq a' \leq b' \leq b} \int_{a'}^{b'} f dg \geq \int_a^b h f dg \geq m \int_a^b f dg + V(h) \inf_{a \leq a' \leq b' \leq b} \int_{a'}^{b'} f dg,$$

where $m = \inf\{h(x) : x \in I\}$.

The following is a standard result in the mechanism design literature based on Mirrlees (1971) and Milgrom and Segal (2002).

Lemma 22. The choices from a menu $M \in \mathscr{M}$ satisfy

$$(x(\theta), p(\theta)) \in \arg \max_{(x,p) \in M \cup \{(0,0)\}} u(\theta, x, p)$$

if, and only if, $x(\cdot)$ is nondecreasing, $x(\underline{\theta}) = 0$ and $p(\theta) = v(\theta, x(\theta)) - \int_{\underline{\theta}}^{\theta} v_1(s, x(s))ds$.

For any $F \in \mathbb{F}$, we can thus rewrite the problem – with some abuse of notation – by choosing directly a function x from the set $\mathscr{X}$, where

$$\mathscr{X} := \left\{x : [\underline{\theta}, \bar{\theta}] \to X, \, x \text{ is nondecreasing and } x(\underline{\theta}) = 0\right\},$$

in order to maximize profit, given by

$$\pi(x, F) := \int_{\Theta} \left(v(\theta, x(\theta)) - \int_{\underline{\theta}}^{\theta} v_1(s, x(s))ds - c\,(x(\theta)) \right) dF(\theta).$$

Consider the normed vector space $(BV(\Theta), \|\cdot\|_{\infty})$, where $BV(\Theta) := \{g : \Theta \to \mathbb{R} \mid V(g) < \infty\}$. For any fixed $M \in \mathscr{M}$, consider its corresponding allocation function $x \in \mathscr{X}$ and extend the functional $\pi(M, \cdot)$ to $BV(\Theta)$ by defining

$$\bar{\pi}(M, H) = \int_{\Theta} \left(v(\theta, x(\theta)) - \int_{\underline{\theta}}^{\theta} v_1(s, x(s))ds - c\,(x(\theta)) \right) dH(\theta), \quad \forall H \in BV(\Theta).$$

Clearly, $\bar{\pi}(M, F) = \pi(M, F)$ for all $F \in \mathscr{F}$. Moreover, note that for all $F, G \in BV(\Theta)$,

$$\begin{aligned} |\bar{\pi}(M, F) - \bar{\pi}(M, G)| &= \left| \int_{\Theta} \left(v(\theta, x(\theta)) - \int_{\underline{\theta}}^{\theta} v_1(s, x(s))ds - c\,(x(\theta)) \right) d(F - G)(\theta) \right| \\ &\leq \left| \int_{\Theta} h_1 d(F - G)(\theta) \right| + \left| \int_{\Theta} h_2(\theta) d(F - G)(\theta) \right| \end{aligned}$$

where

$$\begin{aligned} h_1(\theta) &:= v(\theta, x(\theta)) \\ h_2(\theta) &:= \int_{\underline{\theta}}^{\theta} v_1(s, x(s))ds + c\,(x(\theta)) \end{aligned}$$

As both v and x are nondecreasing, we have that for any $x(\cdot)$,

$$\begin{aligned} V(h_1) &= v(\bar{\theta}, x(\bar{\theta})) - v(\underline{\theta}, x(\underline{\theta})) \\ &\leq v(\bar{\theta}, \bar{x}) =: L_1 < \infty \end{aligned}$$

As v is supermodular and nondecreasing in θ and c is increasing and convex,

$$\begin{aligned} V(h_2) &= \int_{\underline{\theta}}^{\bar{\theta}} v_1(s, x(s)) ds + c\left(x(\bar{\theta})\right) - c\left(x(\underline{\theta})\right) \\ &\leq (\bar{\theta} - \underline{\theta}) \cdot \max_{\theta' \in \Theta} v_1(\theta', \bar{x}) + c(\bar{x}) =: L_2 < \infty \end{aligned}$$

Moreover, we note that $\inf_{\theta \in \Theta} h_1(\theta) = v(\underline{\theta}, x(\underline{\theta})) = 0$ and $\inf_{\theta \in \Theta} h_2(\theta) = \int_{\underline{\theta}}^{\underline{\theta}} v_1(s, x(s)) ds + c\left(x(\underline{\theta})\right) = c(0) = 0$. Hence, for h_i, with $m_i := \inf_{\theta \in \Theta} h_i(\theta)$, $i = 1, 2$, and letting $H(\theta) = F(\theta) - G(\theta)$,

$$\begin{aligned} \left| \int_{\underline{\theta}}^{\bar{\theta}} h_i(\theta) dF(\theta) - \int_{\underline{\theta}}^{\bar{\theta}} h_i(\theta) dG(\theta) \right| &= \left| \int_{\underline{\theta}}^{\bar{\theta}} h_i(\theta) d[F(\theta) - G(\theta)] \right| \\ &\leq \max \left\{ m_i [H(\bar{\theta}) - H(\underline{\theta})] + V(h_i) \sup_{\underline{\theta} \leq \alpha < \beta \leq \bar{\theta}} \int_{\alpha}^{\beta} dH, \right. \\ &\qquad \left. -m_i [H(\bar{\theta}) - H(\underline{\theta})] - V(h_i) \inf_{\underline{\theta} \leq \alpha < \beta \leq \bar{\theta}} \int_{\alpha}^{\beta} dH \right\} \\ &= V(h_i) \cdot \max \left\{ \sup_{\underline{\theta} \leq \alpha < \beta \leq \bar{\theta}} \int_{\alpha}^{\beta} dH, \sup_{\underline{\theta} \leq \alpha < \beta \leq \bar{\theta}} - \int_{\alpha}^{\beta} dH \right\} \\ &= V(h_i) \cdot \max \left\{ \sup_{\underline{\theta} \leq \alpha < \beta \leq \bar{\theta}} H(\beta) - H(\alpha), \sup_{\underline{\theta} \leq \alpha < \beta \leq \bar{\theta}} H(\alpha) - H(\beta) \right\} \\ &\leq V(h_i) \cdot \sup_{\underline{\theta} \leq \alpha < \beta \leq \bar{\theta}} |H(\alpha)| + |H(\beta)| \\ &= 2 \cdot V(h_i) \cdot \|F - G\|_{\infty} \end{aligned}$$

Combining the preceding inequalities results in

$$|\bar{\pi}(M,F) - \bar{\pi}(M,G)| \leq 2(L_1 + L_2)\|F - G\|_\infty \quad \forall F, G \in BV(\Theta),\ M \in \mathscr{M}.$$

The result is obtained by restricting the domain to $\mathbb{F}$.

C.3 Proof of Theorem 10

Let $\mathscr{Y}$ be a normed vector space with an open subset $\mathscr{A} \subseteq \mathscr{Y}$. As is standard, we denote the dual space of $\mathscr{Y}$ by $\mathscr{Y}^*$. Before we prove the main result, let us recall two different notions of differentiability:

Definition 14 (Gâteaux Differentiability). A functional $T : \mathscr{Y} \to \mathbb{R}$ is Gâteaux differentiable at $F \in \mathscr{Y}$ if there exists a linear functional $\dot{T}_F \in \mathscr{Y}^*$ such that for all $H \in \mathscr{Y}$,

$$\dot{T}_F(H) = \lim_{t \to \infty} \frac{T(F + tH) - T(F)}{t}. \qquad \diamond$$

Definition 15 (Fréchet Differentiability). A functional $T : \mathscr{Y} \to \mathbb{R}$ is Fréchet differentiable at $F \in \mathscr{Y}$ if there is a linear continuous functional $\dot{T}_F$ defined on $\mathscr{Y}$ such that

$$\lim_{\|H\| \to 0} \frac{|T(F + H) - T(F) - \dot{T}_F(H)|}{\|H\|} = 0. \qquad \diamond$$

Note that if T is Gâteaux differentiable at $F \in \mathscr{Y}$, then its derivative is unique. Moreover, if it is Fréchet differentiable at $F \in \mathscr{Y}$, it is Gâteaux differentiable and its derivatives agree.

An important generalization of Gâteaux differential for the case of convex and continuous functionals is that of a subdifferential.

Definition 16. The subdifferential of $T : \mathscr{Y} \to \mathbb{R}$ at $F \in \mathscr{Y}$ is the set

$$\partial T(F) := \{D \in \mathscr{Y}^* : T(G) \geq T(F) + D(G - F) \text{ for each } G \in \mathscr{Y}\}. \qquad \diamond$$

Since the normed linear space $(BV(\Theta), \|\cdot\|_\infty)$ is a metric space, and thus $BV(\Theta)$ is open, the following lemma guarantees that ∂T is nonempty:

Lemma 23 (Theorem 3.3.1. Niculescu and Persson (2018)). If $T : \mathscr{A} \to \mathbb{R}$ is a continuous and convex functional, then $\partial T(a) \neq \emptyset$, for all $a \in \mathscr{A}$.

The next result shows why subdifferentials can be thought of as a generalization of Gâteaux differentials.

Lemma 24 (Proposition 3.6.9. Niculescu and Persson (2018)). Let $T : \mathscr{A} \to \mathbb{R}$ be a continuous and convex functional. T is Gâteaux differentiable at $a \in \mathscr{A}$ if and only if $\partial T(a)$ is a singleton.

In what follows, we borrow from the proof strategy of Theorem 2 in Battauz, Donno, and Ortu (2015), although their conditions do not directly apply to our problem.

First fix $M \in \mathscr{M}$ and notice that $\bar{\pi}(M, \cdot) : BV(\Theta) \to \mathbb{R}$ is a linear functional. Therefore, it has a Fréchet derivative. In fact, for all $F, H \in BV(\Theta)$,

$$\begin{aligned}\lim_{\|H\|_\infty \to 0} & \frac{|\bar{\pi}(M, F+H) - \bar{\pi}(M, F) - \bar{\pi}(M, H)|}{\|H\|_\infty} \\ &= \lim_{\|H\|_\infty \to 0} \frac{|\bar{\pi}(M, F) + \bar{\pi}(M, H) - \bar{\pi}(M, F) - \bar{\pi}(M, H)|}{\|H\|_\infty} = 0.\end{aligned}$$

This implies that the Fréchet derivative of $\bar{\pi}(M, F)$ is independent of F and given by $\dot{\pi}_M(\cdot) = \bar{\pi}(M, \cdot)$. Therefore, $\bar{\pi}(M, \cdot)$ is also Gâteaux differentiable, with Gâteaux derivative at F given by $\bar{\pi}(M, \cdot)$.

Define the functional $\bar{\Pi} : BV(\Theta) \to \mathbb{R}$ by

$$\bar{\Pi}(H) = \sup_{M \in \mathscr{M}} \bar{\pi}(M, H), \quad \forall H \in BV(\Theta).$$

Since the set $\{\bar{\pi}(M, H) : M \in \mathscr{M}\}$ is bounded for any $H \in BV(\Theta)$, it is clear that $\bar{\Pi}(H) < \infty$. Furthermore, it is immediate that $\bar{\Pi}(F) = \Pi(F)$ for all $F \in \mathscr{F}$.

As $\bar{\pi}(M, H)$ is linear in H for any M, one immediately has that $\bar{\Pi}$ is convex in $H \in BV(\Theta)$, as the supremum of a family of linear functionals. Moreover, from the proof of Lemma 8, it is easy to see that $\bar{\Pi}$ remains Lipschitz continuous. Therefore, by Lemma 23, $\partial\bar{\Pi}(H) \neq \emptyset$ for any $H \in BV(\Theta)$.

By Lemma 20, $\mathscr{M}^*(F) \neq \emptyset$ for any $F \in \mathscr{F}$. Fix any $F \in \mathscr{F}$. Take any $D \in \partial\bar{\Pi}(F)$ and any $M_F \in \mathscr{M}^*(F)$. Note, for any $G \in BV(\Theta)$, that $\bar{\pi}(M_F, F) - \bar{\pi}(M_F, G) \geq \bar{\Pi}(F) - \bar{\Pi}(G) \geq D(F - G)$, hence $\partial\bar{\Pi}(F) \subseteq \partial\bar{\pi}(M_F, F)$ and, as $\bar{\pi}(M_F, \cdot)$ is continuous and linear, by Lemma 24, $\partial\bar{\pi}(M_F, F) = \left\{\dot{\pi}_{M_F}\right\}$. Then, for any $G \in BV(\Theta)$,

$$\begin{aligned}
& \bar{\pi}(M_F, F) - \bar{\pi}(M_F, G) \geq \bar{\Pi}(F) - \bar{\Pi}(G) \geq \dot{\pi}_{M_F,F}(F - G) \\
\Longleftarrow \quad & \bar{\pi}(M_F, F) - \bar{\pi}(M_F, G) - \dot{\pi}_{M_F}(F - G) \geq \bar{\Pi}(F) - \bar{\Pi}(G) - \dot{\pi}_{M_F}(F - G) \geq 0 \\
\Longrightarrow \quad & |\bar{\pi}(M_F, F) - \bar{\pi}(M_F, G) - \dot{\pi}_{M_F}(F - G)| \geq |\bar{\Pi}(F) - \bar{\Pi}(G) - \dot{\pi}_{M_F}(F - G)| \geq 0.
\end{aligned}$$

By Fréchet differentiability of $\bar{\pi}(M_F, \cdot)$, we then have that $\forall \{G_n\}_n$ such that $\|G_n - F\|_\infty \to 0$,

$$0 \leq \frac{|\bar{\Pi}(G_n) - \bar{\Pi}(F) - \dot{\pi}_{M_F}(G_n - F)|}{\|G_n - F\|_\infty} \leq \frac{|\bar{\pi}(M_F, G_n) - \bar{\pi}(M_F, F) - \dot{\pi}_{M_F}(G_n - F)|}{\|G_n - F\|_\infty} \to 0$$

and, consequently, $\bar{\Pi}$ is Fréchet differentiable at $F \in \mathscr{F}$. As F was arbitrary, we have that $\bar{\Pi}$ is Fréchet differentiable at any $F \in \mathscr{F}$.

C.4 Proof of Theorems 9 and 11

We will prove Theorem 11. The proof for Theorem 9 is virtually the same.

By Theorem 10, $\Pi(F)$ is Fréchet differentiable at any $F \in \mathscr{F}$ and, thus, it can be written as $\Pi(F) = \Pi(F_0) + \dot{\Pi}_{F_0}(F - F_0) + o(\|F - F_0\|_\infty)$. Furthermore, we note that $\hat{F}$ is an unbiased estimator of F_0 and then, by linearity,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \pi(M^*(F_0), \delta_{\theta_i})\right] = \mathbb{E}\left[\pi(M^*(F_0), \hat{F}(S^n))\right] = \pi\left(M^*(F_0), \mathbb{E}\left[\hat{F}(S^n)\right]\right) = \pi(M^*(F_0), F_0),$$

where δ_θ denotes the cumulative distribution function associated with a Dirac delta measure at θ

and θ_i is the i-th observation in the sample S^n. Thus,

$$\begin{aligned}\sqrt{n}\left(\Pi(\hat{F}(S^n)) - \Pi(F_0)\right) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\pi(M^*(F_0), \delta_{\theta_i}) - \pi(M^*(F_0), F_0)\right) + \sqrt{n}\cdot o(\|\hat{F}(S^n) - F_0\|_\infty) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\pi(M^*(F_0), \delta_{\theta_i}) - \pi(M^*(F_0), F_0)\right) \\ &\qquad + \sqrt{n}\cdot\|\hat{F}(S^n) - F_0\|_\infty \cdot \frac{o(\|\hat{F}(S^n) - F_0\|_\infty)}{\|\hat{F}(S^n) - F_0\|_\infty} \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\pi(M^*(F_0), \delta_{\theta_i}) - \pi(M^*(F_0), F_0)\right) + O_p(1)\cdot o(1) \\ &\xrightarrow{d} N(0, \sigma^2_{F_0}),\end{aligned}$$

where $\sigma^2_{F_0} = \mathbb{E}\left[\left(\dot{\Pi}_{F_0}(\delta_\theta - F_0)\right)^2\right]$.

Finally, following Parr (1985b), we have that

$$\begin{aligned}\hat{G}_n =&\sqrt{n}\left(\Pi(\hat{F}(S^n_B)) - \Pi((\hat{F}(S^n))\right) \\ =&\sqrt{n}\left(\left\{\Pi(F_0) + \dot{\Pi}_{F_0}(\hat{F}(S^n_B) - F_0) + o(\|\hat{F}(S^n_B) - F_0\|_\infty)\right\}\right. \\ &\qquad \left. - \left\{\Pi(F_0) + \dot{\Pi}_{F_0}(\hat{F}(S^n) - F_0) + o(\|\hat{F}(S^n) - F_0\|_\infty)\right\}\right) \\ =&\sqrt{n}\left(\dot{\Pi}_{F_0}(\hat{F}(S^n_B) - \hat{F}(S^n)) + o(\|\hat{F}(S^n_B) - F_0\|_\infty) + o(\|\hat{F}(S^n) - F_0\|_\infty)\right) \\ =&\sqrt{n}\left(\dot{\Pi}_{F_0}(\hat{F}(S^n_B) - \hat{F}(S^n)) + o(\|\hat{F}(S^n_B) - \hat{F}(S^n)\|_\infty) + o(\|\hat{F}(S^n) - F_0\|_\infty)\right),\end{aligned}$$

where we used Fréchet differentiability to obtain a first-order von Mises expansion of Π in the second equality, linearity of $\dot{\Pi}_{F_0}$ in the third and the triangle inequality in the last. As, $\hat{F}(S^n_B)$ and $\hat{F}(S^n)$ denote empirical distributions of $\hat{F}(S^n)$ and F_0, by the Dvoretzky, Kiefer, and Wolfowitz (1956) inequality, we have that

$$\hat{G}_n = \sqrt{n}\dot{\Pi}_{F_0}(\hat{F}(S^n_B) - \hat{F}(S^n)) + o_p(1) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\pi(M^*(F_0), \delta_{\theta^B_i}) - \pi(M^*(F_0), \hat{F}(S^n))\right) + o_p(1)$$

As $\mathbb{E}\left[\pi(M^*(F_0), \delta_{\theta^B_i}) \mid S^n\right] = \pi(M^*(F_0), \hat{F}(S^n))$ and $\mathbb{E}\left[\left(\pi(M, \hat{F}(S^n_B))\right)^2\right] = \sigma^2_{F_0} < \infty$, by the central limit theorem in Bickel and Freedman (1981), we have that $\hat{G}_n \overset{d}{\to} G_0$. Finally, as the limiting distribution is continuous, convergence is uniform.

C.5 Proof of Proposition 12

Let $\hat{F}^E$ denote the empirical distribution estimator. By Lemmas 8 and 27, $\Pi(\hat{F}(S^n)) = \Pi(\hat{F}^E(S^n)) + O_p(n^{-1})$. As such, (1) follows from the observation that

$$\sqrt{n}\left\{\Pi(\hat{F}(S^n)) - \Pi(F_0)\right\} = \sqrt{n}\left\{\Pi(\hat{F}^E(S^n)) - \Pi(F_0)\right\} + O_p(n^{-1/2}),$$

which, together with Slutsky's theorem and Theorem 11 implies $\sqrt{n}\left\{\Pi(\hat{F}(S^n)) - \Pi(F_0)\right\} \overset{d}{\to} N(0, \sigma^2_{F_0})$.

(2) results from the analogous observation that:

$$\hat{G}_n = \sqrt{n}\left(\Pi(\hat{F}(S^n_B)) - \Pi((\hat{F}(S^n)))\right) = \sqrt{n}\left(\Pi(\hat{F}^E(S^n_B)) - \Pi((\hat{F}^E(S^n))\right) + O_p(n^{-1/2}) \overset{d}{\to} N(0, \sigma^2_{F_0}).$$

For (3), we first prove the following lemmata:

Lemma 25. Consider distribution functions $F,\ F_1,\ F_2, \ldots$ such that $\|F_n(t) - F(t)\|_\infty \xrightarrow{a.s.} 0$ for all $t \in \mathbb{R}$. For all $n \in \mathbb{N}$, let μ_n be the measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ induced by F_n and μ be the measure induced by F. Then $\mu_n(A) \to \mu(A)$ for all $A \in \mathscr{B}(\mathbb{R})$ almost surely.

Proof: Let $\left\{I_j := (a_j, b_j] \subseteq \mathbb{R} \mid j \in \mathbb{N}\right\}$ be a collection of disjoint intervals, and denote $I = \bigcup_{j=1}^\infty I_j$. Then, since for any $n \in \mathbb{N}$, μ_n and μ are finite measures,

$$\begin{aligned}
|\mu_n(I) - \mu(I)| &= \left|\sum_{j=1}^\infty \mu_n(I_j) - \sum_{j=1}^\infty \mu(I_j)\right| = \left|\sum_{j=1}^\infty F_n(b_j) - F_n(a_j) - \sum_{j=1}^\infty F(b_j) - F(a_j)\right| \\
&= \left|\sum_{j=1}^\infty \left(F_n(b_j) - F(b_j)\right) - \left(F_n(a_j) - F(a_j)\right)\right| \\
&\le \sum_{j=1}^\infty \left|F_n(b_j) - F(b_j)\right| + \left|F_n(a_j) - F(a_j)\right|
\end{aligned}$$

$$\leq \sum_{j=1}^{\infty} 2 \sup_{t\in\mathbb{R}} |F_n(t) - F(t)|$$

Let $c_n = \sup_{t\in\mathbb{R}} |F_n(t) - F| \to 0$. There exists a monotone convergent subsequence $c_{n_k} = 2\sup_{t\in\mathbb{R}} \left|F_{n_k}(t) - F(t)\right| \to 0$ as $k \to \infty$. Therefore, by the monotone convergence theorem,

$$\lim_{k\to\infty} \left|\mu_{n_k}(I) - \mu(I)\right| \leq \lim_{k\to\infty} \sum_{j=1} 2 \sup_{t\in\mathbb{R}} \left|F_{n_k}(t) - F(t)\right| = \sum_{j=1} \lim_{k\to\infty} 2 \sup_{t\in\mathbb{R}} \left|F_{n_k}(t) - F(t)\right| = 0.$$

This implies there exists a convergent subsequence $\mu_{n_k}(I) \to \mu(I)$.

Now take any convergent subsequence $\mu_{n_m}(I)$ of $\mu_n(I)$, which we will denote by $\mu_m(I)$. We have

$$|\mu_m(I) - \mu(I)| \leq \sum_{j=1}^{\infty} 2 \sup_{t\in\mathbb{R}} |F_m(t) - F(t)|.$$

Then $c_m = 2\sup_{t\in\mathbb{R}} |F_m(t) - F(t)| \to 0$ has a monotone convergent subsubsequence c_{m_r}. We can thus apply the monotone convergence theorem once again to conclude that

$$\left|\mu_{m_r}(I) - \mu(I)\right| \to 0 \quad \text{as } r \to \infty.$$

Since the subsequence $\mu_m(I)$ is convergent by assumption, it must converge to the limit of each of its subsequences, and we have

$$\lim_{m\to\infty} \mu_m(I) = \mu(I).$$

Therefore, every convergent subsequence of $\mu_n(I)$ converges to $\mu(I)$. Since $(\mu_n(I))_{n\in\mathbb{N}}$ is bounded, this implies that $\lim_{n\to\infty} \mu_n(I) = \mu(I)$, and thus $\lim_{n\to\infty} \mu_n$ is a pre-measure that agrees with μ in the ring formed by disjoint unions of intervals of the type $(a, b],\ b > a$. Therefore, since μ is σ-finite, Carathéodory's extension theorem implies that $\lim_{n\to\infty} \mu_n$ must agree with μ on $\mathscr{B}(\mathbb{R})$ almost surely. □

Lemma 26. Let $\{F_n : \Theta \to [0, 1] \mid n \in \mathbb{N}\}$ be a sequence of absolutely continuous distribution functions with Radon-Nikodym derivatives given by f_n. If there exists a distribution function F

such that $\|F_n - F\|_\infty \to 0$, then it is absolutely continuous with Radon-Nikodym derivative f and $\|f_n - f\|_1 \to 0$.

Proof: As $\forall n \in \mathbb{N}$, F_n is absolutely continuous and Θ has finite Lebesgue measure, then the Radon-Nikodym derivatives $\{f_n\}_{n\in\mathbb{N}}$ are uniformly integrable with respect to the Lebesgue measure. Let μ be the measure associated with F. By the Vitali-Hahn-Saks theorem and Lemma 25, we have that

$$\lim_{n\to\infty} \int_A f_n(\theta) d\theta = \mu(A)$$

for all $A \in \mathscr{B}(\Theta)$. Since $\{f_n\}_{n\in\mathbb{N}}$ is uniformly integrable, the Dunford-Pettis theorem implies that every subsequence of $\{f_n\}_{n\in\mathbb{N}}$ has a convergence subsubsequence converging to g in $L^1(\Theta)$. Denote such a subsubsequence by $\{f_{n_k}\}_{k\in\mathbb{N}}$. Then, for every $A \in \mathscr{B}(\mathbb{R})$,

$$\int_A g(\theta)d\theta = \lim_{k\to\infty} \int_A f_{n_k} d\theta = \lim_{k\to\infty} \mu_{n_k}(A) = \int_A f(\theta)d\theta.$$

This implies that $f = g$ almost surely and $\|f_n - f\|_1 \to 0$. □

Given that, $\{\hat{F}(S^n)\}_{n\in\mathbb{N}}$ is absolutely continuous with probability 1 and $\|\hat{F}(S^n) - F_0\|_\infty \overset{a.s.}{\to} 0$, the previous Lemmata imply that $\|f_n - f_0\|_1 \overset{p}{\to} 0$, which concludes the proof for (3).

C.6 Proof of Proposition 13

First, we note that $\pi(r, F)$ is Lipschitz continuous in F, with a Lipschitz constant that is independent of r. Note that

$$\begin{aligned} |\pi(r,F) - \pi(r,G)| &= \left| \int_\Theta \mathbf{1}_{\theta \geq r} d(F_{(2;M)} - G_{(2;M)}) \right| \\ &\leq \|F_{(2;M)} - G_{(2;M)}\|_\infty \\ &= \sup_{\theta\in\Theta} |M \cdot (F(\theta)^{M-1} - G(\theta)^{M-1}) + (M-1) \cdot (G(\theta)^M - F(\theta)^M)| \end{aligned}$$

$$\leq M \cdot \sup_{\theta\in\Theta} |F(\theta)^{M-1} - G(\theta)^{M-1}| + (M-1) \cdot \sup_{\theta\in\Theta} |F(\theta)^{M} - G(\theta)^{M})|$$

$$\leq 2M(M-1)\|F-G\|_\infty.$$

where the first inequality uses the Beesack-Darst-Pollard inequality – see Lemma 21. By the same arguments as in Lemma 8, we have that $\Pi(F) := \sup_{r\in\Theta} \pi(r, F)$ is also Lipschitz continuous in F and, by those made in Propositions 8 and 9, the result follows.

C.7 Other Proofs

Let the linearly interpolated empirical cumulative distribution be given by

$$\hat{F}(S^n)(\theta) = \sum_{k=0}^{n-1} \mathbf{1}_{\{\theta_{(k)}\leq\theta<\theta_{(k+1)}\}} \frac{1}{n} \frac{\theta - \theta_{(k)}}{\theta_{(k+1)} - \theta_{(k)}} + \mathbf{1}_{\{\theta_{(n)}\leq\theta\}},$$

where $\theta_{(k)}$ denotes the k-th smallest observation in the sample S^n and $\theta_{(0)} = \underline{\theta}$. The following holds:

Lemma 27. For any absolutely continuous $F_0 \in \mathscr{F}$, (1) $\|\hat{F}(S^n) - F_0\|_\infty \overset{a.s.}{\to} 0$ and (2) with probability 1, $\hat{F}(S^n) \in \mathscr{F}$, $\hat{F}(S^n)$ has convex support and is absolutely continuous.

Proof: Note that, with probability 1, any sampled value $\theta_i > \underline{\theta}$ as F_0 is absolutely continuous and so $\hat{F}(S^n)$ is well defined. By construction, $\hat{F}(S^n)$ has convex support and is absolutely continuous. As the probability that any two sampled observations have the same value is null, we have that, with probability 1, $\|\hat{F}(S^n) - \hat{F}^E(S^n)\|_\infty = 1/n$ where $\hat{F}^E$ denotes the empirical cumulative distribution and, with probability 1, $\|\hat{F}(S^n) - F_0\|_\infty = \|\hat{F}^E(S^n) - F_0\|_\infty + 1/n$. Consequently, as $\|\hat{F}^E(S^n) - F_0\|_\infty \overset{a.s.}{\to} 0$ and as F_0 is absolutely continuous, then the result follows. □