Spring 5-15-2019

Essays on Econometrics and Rational Choice

Junnan He
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS
Department of Economics

Dissertation Examination Committee:
Werner Ploberger (Chair)
Siddhartha Chib
George-Levi Gayle
Paulo Natenzon
Jonathan Weinstein

Essays on Econometrics and Rational Choice

by

Junnan He

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2019
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

Over the past six years, I have received support and encouragement from a great number of individuals. I am especially indebted to my advisor Werner Ploberger, who has been supportive of all my research interests, and at the same time actively providing valuable feedback, advice, and encouragement.

I am grateful to each of the committee members. In addition to my advisor, Siddhartha Chib, George-Levi Gayle, Paulo Natenzon, and Jonathan Weinstein, all have provided immense support in various stages during my graduate studies. In particular, my essays in rational choice have greatly benefited from the influence of John Nachbar, Jonathan Weinstein, and especially Paulo Natenzon, with whom I also share a rewarding collaboration experience. Moreover, I was fortunate to receive help from many other faculties in the university, including Lin Nan, Ian Fillmore, Yongseok Shin, and Carl Sanders among others.

I would also like to thank all my friends during the period, including Ting Yuen Terry Cheung, Inkee Jang, Jiahui Lyu, Yuping Chen, all the candidates in the 2018-19 job market, all the advisees of Werner's, and many other students from different cohorts or programs. Their company and encouragement have made my life in the past few years more colorful.

Junnan He

Washington University in St. Louis

May 2019

Dedicated to my parents, *Mǐn Líng Lín* and *Gān Huá Hé*,

and in memory of my grandpa *Cháo Hàn Lín*.

ABSTRACT OF THE DISSERTATION

Essays on Econometrics and Rational Choice

by

Junnan He

Doctor of Philosophy in Economics

Washington University in St. Louis, 2019

Professor Werner Ploberger, Chair

Decision and choice theory is a topic of interest in both econometrics and microeconomic theory. We contribute to the theory of decision under both contexts, that is, the theory of model selection in econometrics, and the theory of rational decision in microeconomics.

There is a long-lasting theoretical interest in model selection. More recently, research on sparse estimators, a class of estimation methods that select and estimate important parameters simultaneously, has been the central focus on model selection. The methods become especially relevant when the problem is of high-dimensional nature. Theoretically, sparse methods can perform well when the true data generating process (DGP) is assumed to have a low-dimensional structure. But empirically, a sparse estimator can be outperformed by some dense estimators when this assumption does not hold. In Chapter 1, we propose a test of sparsity for linear regression models. Our null hypothesis is that the number of non-zero parameters does not exceed a small preset fraction of the total number of parameters. It can be interpreted as a family of Bayesian prior distributions where each parameter equals zero with a large probability. For the alternative, we consider the case where all parameters are nonzero and of order $1/\sqrt{p}$ for all p number of parameters. Formally, the alternative is a normal prior distribution, the maximum entropy prior with the mean being zero, and the variance determined by the ANOVA identity. We derive a test statistic using the theory of

robust statistics. This statistic is minmax-optimal when the design matrix is orthogonal and can be used for general design matrices as a conservative test.

Sometimes, there is a natural ordering in which the importance among the parameters is arranged. Typical examples are the representation of a function by series or the estimation of a spectrum by a long autoregressive process. Chapter 2 and Chapter 3 are devoted to the analysis under this framework. In Chapter 2, we adapt concepts of the classical Hajek-Blackwell-Lecam theory to develop a theory of asymptotically optimal estimation of the parameters. In many of these cases, maximum likelihood estimators do not exist, and hence there is no canonical candidate for a good estimator. We define suitable loss functions for the estimation error, which allows us to uniquely characterize some estimators. In estimation procedures, it is quite common to assume higher order differentiability or smoothness conditions of the parameters. We construct some simple prior distributions that force the parameters to obey the smoothness conditions. We show that the class of shrunken sieve estimators is asymptotically efficient. I.e. the sieve estimator is multiplied with a matrix that shrinks the estimates towards zero, analogous to Ridge regressions or Bayesian estimators in a linear model.

In Chapter 3, we show that, in linear models with increasing dimension, the estimator resulting from the maximization of Akaike's Information Criterion is asymptotically equivalent to some Bayesian estimators. The family of prior distributions that generates our estimators is normal, defined on the space of all sequences, and is characterized by an exponential decay of the variance for the higher order components of the parameter.

The last two Chapters are devoted to decision theory in microeconomics. In contrast to the decision theory in econometrics where the loss (utility) function is predefined, the focus of microeconomics is to recover a well-defined preference (utility). A well-defined or a rational preference is one that satisfies certain consistency axioms. The most notable consistency axiom is arguably the transitive axiom. The most studied transitivity axiom in the stochastic

choice literature is the strong stochastic transitivity (SST). However, individual choice data often violate SST while conforming to moderate stochastic transitivity (MST). Chapter 4 focuses on the analysis of this axiom and its relevance to recovering the underlying preference. Our first theorem shows that a binary choice rule satisfies a slightly stronger version of the MST postulate, which we call MST+, if and only if it can be represented by a moderate utility model (MUM). Choices in the MUM are a function of utility difference divided by a distance metric, which determines the degree of comparability of the options. Our second theorem introduces the moderate expected utility model (MEM) and shows how our parameters can be identified from the choice data over lotteries.

Sometimes the choice data do not even satisfy the weakest form of transitivity and violate other classical axioms such as the independence of irrelevant alternatives. The main source of such observations comes from contextual choices. Chapter 5 is devoted to rationalizing such choice behaviors. We build a choice model with a fixed underlying utility function and explain contextual choices with a novel information friction: the agent's perception of the options is affected by an attribute-specific noise. Under this friction, the agent obtains useful information when additional options are introduced. Therefore, the agent chooses contextually, exhibiting intransitivity, joint-separate evaluation reversal, attraction effect, compromise effect, similarity effect, and phantom decoy effect. Nonetheless, because the noise is attribute-specific and common across alternatives, the agent chooses perfectly rationally whenever there is clear dominance between options.

Chapter 1

A Test for Sparsity

1.1 Introduction

The increase in availability of data has boosted a fast growing literature on variable selection. The main question in the literature is to find, in the vast number of different combinations, a small set of variables that can sufficiently explain the response variable. An estimator that contains many zeros in the estimated coefficients is called a *sparse* estimator. Such estimators include but not limited to AIC (Akaike, 1974), BIC (Schwarz, 1978), LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), Elastic Net (Zou and Hastie, 2005) etc. When the data generating process (DGP) is sparse, i.e. when the response is only significantly affected by a diminishing fraction of the variables, many sparse estimators can consistently find the important variables and estimate them efficiently (Meinshausen and Buhlmann, 2006; Zhao and Yu, 2006; Zhang and Huang, 2008).

However, it is known that a sparse estimator does not always dominate a non-sparse one. If the underlying DGP is not sparse, using a sparse method may result in inefficient estimates. Tibshirani (1996) observed in simulations that, when the DGP is dense, i.e. "a large number of small effects", the sparse estimator LASSO is significantly less efficient

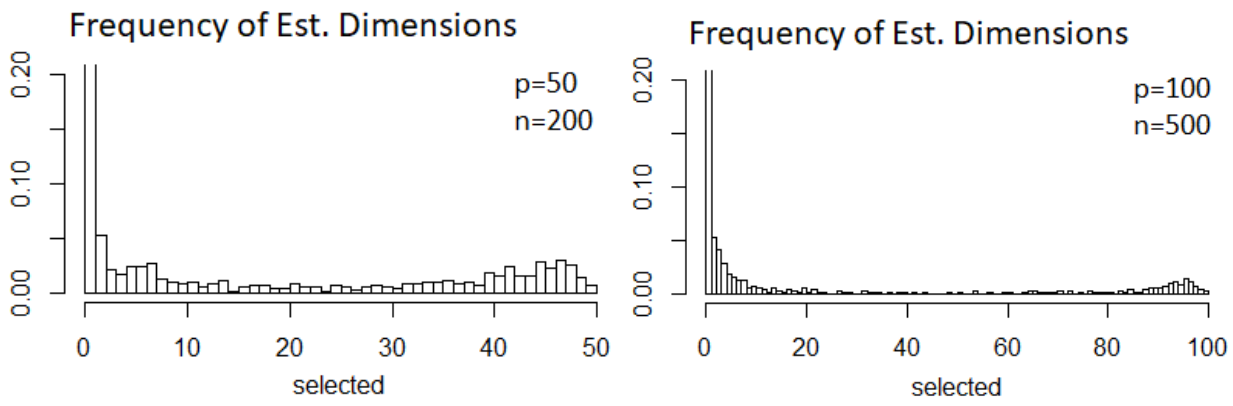


Figure 1.1: Dimension of Selected Models

In each graph, we simulate 1000 times the regression model $Y = X\beta + u$ and estimate the LASSO with the tuning parameter determined by optimizing the BIC (Zou, Hastie and Tibshirani, 2007). Then we plot the histogram of the number of non-zero coefficients estimated. The design matrix is always simulated from a multivariate standard normal and $u \sim \mathcal{N}(0,1)$. On the left penal, $\beta = (2,1,1,1,1,\ldots,2,1,1,1,1)/\sqrt{50}$ is a vector of length 50 and the number of observation is $n = 200$. On the right penal, $\beta = (2,1,1,1,1,\ldots,2,1,1,1,1)/\sqrt{100}$ is a vector of length 100 and the number of observation is $n = 500$.

than the ridge regression, a dense estimator. Apart from the loss in efficiency, consistency can also be compromised. When applied to a dense DGP, sparse estimators can be selection inconsistent by selecting significantly *too few* variables. Figure 1.1 shows that LASSO selects a very low dimensional model when the true DGP is dense. While all variables have non-zero coefficients in both simulations, 40% of the estimated models are of dimension less than four in the first panel, and 51% in the second panel. The typical problems for many sparse estimators is that they are too liberal by selecting too many irrelevant variables when the number of irrelevant regressors diverges in a sparse DGP (Chen and Chen, 2008). In contrast, the above simulation shows when the DGP is dense, sparse methods are likely too stringent.

In this paper, we provide a test to distinguish whether the DGP is sparse or dense for linear regression models.[1] The test can be used as a validation or diagnostics before or after

[1]We focus the analysis on the linear regression framework because it is the most popular statistical tool, and many sparse estimation techniques were first proposed for the regression context.For nonlinear problems, when the log-likelihood is sufficiently smooth, many estimators (e.g. maximum likelihood) can be locally approximated by a linear estimator. Generalizations to these problems are possible but outside the scope of

applying sparse estimators. One can interpret our test as a test between two families of Bayesian priors. The null hypothesis is a large set of prior distributions that each coefficient of interest is zero with high probability, and the alternative hypothesis is the prior that each coefficient is of the same magnitude. In comparison to pure Bayesian techniques such as Giannone, Lenza and Primiceri (2017), our null hypothesis consists of a large family of sparse data generating processes. Hence when the null is rejected, it is not subject to the specification of the prior distribution.

There is a need for determining the data sparsity when choosing between a dense estimator and a sparse estimator because economic variables may not have a sparse DGP. Giannone et al. (2017) used a Bayesian approach to estimate the model dimensions for a number of regressions with economic variables. Their posterior distributions were found to concentrate in high dimensional models for all the macroeconomic and financial examples in their paper. If the underlying DGP is dense, applying a dense estimator such as the ridge estimation is more efficient (see e.g., Hsu, Kakade and Zhang, 2014).

Informally, our procedure works in the following way. Let n and p respectively be the number of observations and the number of parameters to be estimated. Suppose $p < n$ while both n, p are allowed to diverge to infinity. The test statistic summarizes the number of coefficients estimated to be significantly far away from zero. Since the OLS is $\sqrt{n}$ consistent, all but a few estimated coefficients are close to zero of magnitude $1/\sqrt{n}$ under the null. When the alternative is true, the estimated coefficients are the sum of their values plus the estimator noise, hence they are of magnitude $\sqrt{1/p + 1/n}$. We borrow techniques from robust statistics to distinguish this difference between a smaller distribution with outliers and a wider distribution without outliers.

The organization of the paper is as follow. We formally introduce the hypotheses in the next section. The test statistic is derived in Section 3. The rejection region of the test is

this paper.

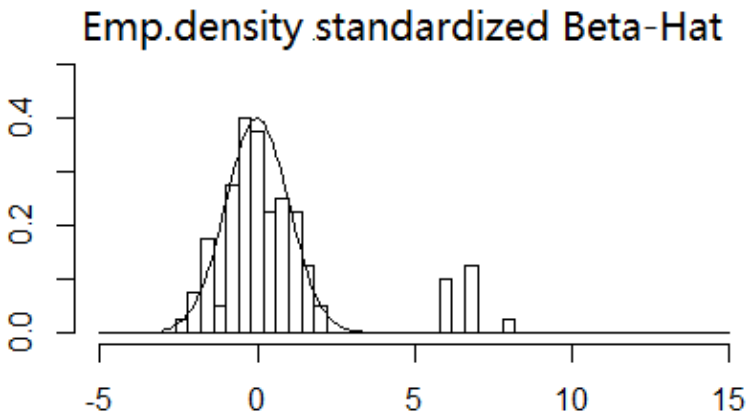


Figure 1.2: Histogram of Estimated Coefficients

The histogram of the standardized estimated values from a single regression. We standardize the OLS estimates of the model $Y = X\beta + u$ where $u \sim \mathcal{N}(0, 6)$ and X is simulated from multivariate normal with correlation $\rho(x_i, x_j) = 0.3^{|i-j|}$. The true parameter β is of dimension 100 and all but the first 10 entries are 0. The first 10 entries of β are 2's. The curve is a standard normal density super-imposed to the histogram.

simulated. The simulation method and a sufficiency result on asymptotic consistency is given in Section 4. We discuss issues related to implementation of the test in Section 5. Section 6 describes some simulation results and Section 7 provides two empirical applications of the test. Lengthy proofs are postponed to the appendix.

1.2 The Hypotheses

Consider the classical linear regression model

$$Y = X\beta + u$$

where X is independent from u_i where $u_i \sim_{iid} \mathcal{N}(0, \sigma^2)$ for $i = 1, \ldots, n$. The dimensions of Y and β are respectively n and p, both diverging to infinity. Without loss of generality we assume that Y and X are standardized to have mean zero and variance 1.

Our null hypothesis can be thought of a large family of prior distributions each describes β as a sparse vector. Let $\mathcal{F}$ be the set of all p dimensional distributions over the reals in $\mathbb{R}^p$.

Formally, the null hypothesis

$$
\begin{aligned}
H_0(\epsilon): \qquad & \forall i, \;\; \beta_i = z_i \gamma_i \\
& (\gamma_1, \ldots, \gamma_p) \sim F \text{ for some distribution } F \in \mathcal{F}; \\
& z_i \text{ is independent Bernoulli with success probability } \epsilon.
\end{aligned}
$$

In other words, each $\beta_i = 0$ whenever $z_i = 0$, which has probability $1 - \epsilon$. When $z_i \neq 0$, $\beta_i = \gamma_i$ which can be drawn from any distribution over the reals. When we have a fixed dispersed distributions F, the distribution for β is similar to a so-called "spike-and-slab" prior in Mitchell and Beauchamp (1988). Nonetheless, F can be an arbitrary distribution including a Dirac-delta measure, in which case each β_i is either 0 or a fixed constant, which may be better described as a "spike-and-spike" prior.

The above null hypothesis can be interpreted as a family of Bayesian priors for the vector β given the knowledge that "about at least $1 - \epsilon$ fraction of the entries in β are zero". Since $\beta_i = 0$ with probability $1 - \epsilon$, each such prior is rather informative about the location of each β_i. Naturally the alternative hypothesis should describe the contrary, a lack of information about the precise location. To this end, we take the alternative to be a maximum entropy distribution (see e.g. Jaynes, 1968). Imposing homogeneity and symmetry, each β_i is independent and identically distributed around zero. We take the second moment of the β_i's by the ANOVA identity $\mathbb{E}[Y'Y] = \mathbb{E}[\beta' X' X \beta] + \mathbb{E}[u'u]$, or equivalently

$$\mathbb{V}ar[Y] = \frac{1}{n} \mathbb{E}[\beta' X' X \beta] + \sigma^2.$$

These conditions pin down the alternative prior distribution for β. Formally,

$$H_a: \qquad \forall i, \;\; \beta_i \sim_{iid} \mathcal{N}\left(0, \frac{1 - \sigma^2}{p}\right).$$

Under this alternative, not only all β_i's are non-zero, but also nearly all parameters are of order $\frac{1}{\sqrt{p}}$. Hence this prior can naturally be interpreted as a hypothesis that there is a large number of small effects. Moreover, under the alternative hypothesis, the optimal estimator for square-loss is exactly a ridge regression estimator. This is in accordance with the observation that that the ridge regression is more efficient for a dense DGP.

1.3 The Test Statistic

Assuming the matrix $X'X$ is invertible, we derive a test statistic that bases on the OLS estimates.[2] The OLS estimator has variance $\sigma^2(X'X)^{-1}$. Let the ith diagonal element of the matrix $(X'X/n)^{-1}$ be s_i^2, and the respective positive square-root be s_i. Our test statistics come from the following intuitive observation. We standardize the estimated values to $\frac{\sqrt{n}}{\sigma s_i}\hat{\beta}_i$, so that the sequence of normalized estimated values have the same marginal variance conditional on β. When β is sparse, all but a few of the entries are zero. If we remove the indices i's where $\beta_i \neq 0$, the remaining estimated values all centers around 0 with variance 1.

As shown in Figure 1.1, other than a few $\frac{\sqrt{n}}{\sigma s_i}\hat{\beta}_i$'s for which $\beta_i \neq 0$, most other entries of normalized $\hat{\beta}$ lies under the standard normal density. Under some regularity conditions, Azriel and Schwartzman (2015) Theorem 1 shows that the empirical distribution of the estimated values, i.e. $\frac{\sqrt{n}}{\sigma s_i}\hat{\beta}_i$ for which $\beta_i = 0$, converges to the standard normal distribution. In our context, when the null is true, the (standardized) estimated vector can be thought of a normal vector with a few outliers. This allows us to interpret our null to be the following. Each $\frac{\sqrt{n}}{\sigma s_i}\hat{\beta}_i$ is drawn from a standard normal with $1-\epsilon$ probability, and with ϵ probability, it is drawn from an arbitrary unknown distribution. To put differently, the estimated values as random draws from a distribution in the epsilon-contamination neighborhood of standard

[2]For discussions about the case $p > n$, refer to Section 5.

normal distribution, as defined in Huber (2004). Hence a natural test statistic for the above hypotheses is the robust test for epsilon-contaminated neighborhood.

Before deriving the test statistics, we first examine the marginal distribution for $\hat{\beta}_i$ under both the null and the alternative. Under the $H_0(\epsilon)$, it is easy to derive that $\hat{\beta}_i$ equals in distribution to the following distribution

$$\hat{\beta}_i =_d (1 - z_i)u_i \frac{s_i}{\sqrt{n}} + z_i\gamma_i$$

where $u_i \sim \mathcal{N}(0, \sigma^2)$. Therefore, $\hat{\beta}_i$ follows $\mathcal{N}(0, \sigma^2 s_i^2/n)$ with $1 - \epsilon$ probability, and with ϵ probability following some arbitrary distributions. For this reason, we say that under $H_0(\epsilon)$, $\hat{\beta}_i$ is a random variable in the ϵ-contaminated neighborhood of $\mathcal{N}(0, \sigma^2 s_i^2/n)$. On the other hand, under the H_a we have

$$\forall i,\ \hat{\beta}_i =_d \mathcal{N}\left(0, \frac{1 - \sigma^2}{p} + \sigma^2 s_i^2/n\right).$$

To derive a likelihood-ratio type statistic for $\hat{\beta}_i$ under the null and the alternative, we start with the likelihood ratio without ϵ-contamination. This likelihood ratio between $\mathcal{N}(0, \frac{1-\sigma^2}{p} + \sigma^2 s_i^2/n)$ and $\mathcal{N}(0, \sigma^2 s_i^2/n)$ is proportional to

$$\exp\left(-\frac{x}{2\left(\frac{1-\sigma^2}{p} + \sigma^2 \frac{s_i^2}{n}\right)} + \frac{x}{2\sigma^2 \frac{s_i^2}{n}}\right).$$

Since the ratio is monotonically increasing in x, the normal variable squared, it is without loss of generality that we analyze only the squared variables according to Huber (2004). Denote by P_0 the cumulative distribution function (CDF) of the square of the $\mathcal{N}(0, \sigma^2 s_i^2/n)$ variable, and by P_a the square of the $\mathcal{N}(0, \frac{1-\sigma^2}{p} + \sigma^2 s_i^2/n)$ variable.[3] Observe that both P_0 and P_a are CDFs of some χ_1^2 variables with different scaling factors. Write their respective

[3]We suppress the index i here for simplicity.

densities as

$$p_0(x)\mathrm{d}x = \frac{e^{-\frac{x}{2\sigma^2 s_i^2/n}}}{\sqrt{2\pi\sigma^2(s_i^2/n)x}}\mathrm{d}x \qquad \text{and} \qquad p_a(x)\mathrm{d}x = \frac{e^{-\frac{x}{2v}}}{\sqrt{2\pi v x}}\mathrm{d}x$$

where $v = \frac{1-\sigma^2}{p} + \sigma^2 s_i^2/n$.

Every element in the ϵ-contamination neighborhood of P_0 can be written as $(1-\epsilon)P_0 + F'$ where F' is a distribution over $[0,\infty)$. Since under the null, $\hat{\beta}_i$ follows an ϵ-contaminated $\mathcal{N}(0, \sigma^2, s_i^2/n)$, we have

$$\hat{\beta}_i^2 \sim Q \text{ where } Q \text{ is a CDF on } [0,\infty) \text{ such that } Q(x) \geq (1-\epsilon)P_0(x)\ \forall x \geq 0.$$

For convenience, in the following of this section, we denote by $H_0(\epsilon)$ the set of distributions

$$\{Q \text{ is a CDF on } [0,\infty) | Q(x) \geq (1-\epsilon)P_0(x)\}.$$

As in Huber (2004), within H_0, we choose the following distribution represented by density q_0

$$q_0(x) = \begin{cases} (1-\epsilon)p_0(x) & \text{for } x \leq x^* \\ cp_a(x) & \text{for } x > x^* \end{cases}$$

for some constants x^* and c such that $\int q_0 = 1$ and $\frac{p_a(x^*)}{q_0(x^*)} = \frac{1}{c}$. The next lemma shows that for each s_i, there is a unique pair of x_i^* and c_i that satisfies these restrictions. The x_i^* would serve as a cut-off value to determine if $\hat{\beta}_i^2$ is "too large". For each $\hat{\beta}_i$, the log-likelihood ratio statistic between p_a and q_0 is $\left(\frac{1}{2\sigma^2 s_i^2/n} - \frac{1}{2v}\right)\min\{\hat{\beta}_i^2, x_i^*\}$ up to a deterministic constant. We take the following normalization of the average statistic over $i \in \{1,\ldots,p\}$ as our test

statistic.

$$T := \frac{1}{p}\sum_{i=1}^{p} \frac{(1-\sigma^2)n}{(1-\sigma^2)n + \sigma^2 s_i^2 p} \times \min\{\frac{\hat{\beta}_i^2}{\sigma^2 s_i^2/n}, \frac{x_i^*}{\sigma^2 s_i^2/n}\},$$

where x_i^* solves

$$\mathrm{erf}\left(\sqrt{\frac{x}{2\sigma^2 s_i^2/n}}\right) + \sqrt{\frac{v_i}{\sigma^2 s_i^2/n}} \exp\left(\left(\frac{1}{v_i} - \frac{1}{\sigma^2 s_i^2/n}\right)\frac{x}{2}\right) \mathrm{erfc}\left(\sqrt{\frac{x^*}{2v_i}}\right) = \frac{1}{1-\epsilon}.$$

The following Lemma describes the asymptotics of these cut-off values.

Lemma 1. *Let $v := \frac{1-\sigma^2}{p} + \sigma^2 s^2/n$. When $v > \frac{\sigma^2 s^2/n}{(1-\epsilon)^2}$, there is a unique pair of x^* and c that simultaneously solves the equations $\int q_0 = 1$ and $\frac{p_a(x^*)}{q_0(x^*)} = \frac{1}{c}$, and x^* satisfies*

$$\frac{x^*}{\sigma^2 s^2/n} \leq \frac{\sigma^2 s^2 p + (1-\sigma^2)n}{(1-\sigma^2)n} \ln\left(\frac{vn}{s^2\sigma^2}\frac{4}{\pi}\left(\frac{1-\epsilon}{\epsilon}\right)^2\right).$$

Let p, s^2, σ^2 and ϵ be functions in n. Suppose $\epsilon \to 0$, and for some constants $\kappa_1, \kappa_2 \in (0,1)$, $\kappa_1 < \sigma^2 < \kappa_2$, and for some constants $\kappa_3 > 0$, $\frac{ps^2}{n} \ln\frac{1}{\epsilon} < \kappa_3$. Then the solution x^* is bounded below by

$$\frac{\sigma^2 s^2 p + (1-\sigma^2)n}{(1-\sigma^2)n} \ln\left(\frac{vn}{s^2\sigma^2}\left(\frac{1-\epsilon}{\epsilon}\right)^2 C\right) \leq \frac{x^*}{\sigma^2 s^2/n},$$

whenever C is some constants independent of $p, s^2, \sigma^2, \epsilon$ and n.

From now on, we define x_i^* to be the solution of the equation

$$(1-\epsilon)\,\mathrm{erf}\left(\sqrt{\frac{x}{2\sigma^2 s_i^2/n}}\right) + (1-\epsilon)\sqrt{\frac{v_i}{\sigma^2 s_i^2/n}} e^{\left(\frac{1}{v_i} - \frac{1}{\sigma^2 s_i^2/n}\right)\frac{x}{2}} \mathrm{erfc}\left(\sqrt{\frac{x^*}{2v_i}}\right) = 1,$$

where $v_i := \frac{1-\sigma^2}{p} + \sigma^2 s_i^2/n$. When the solution does not exists, we set $x_i^* = 0$.

1.4 Rejection Regions and Asymptotic Consistency

Recall that in the test statistic

$$T := \frac{1}{p}\sum_{i=1}^{p} \frac{(1-\sigma^2)n}{(1-\sigma^2)n + \sigma^2 s_i^2 p} \times \min\{\frac{\hat{\beta}_i^2}{\sigma^2 s_i^2/n}, \frac{x_i^*}{\sigma^2 s_i^2/n}\},$$

each term in the sum is derived from the likelihood ratio between the densities p_a and q_0 for $\hat{\beta}_i$ under the alternative and the null respectively. Since the null contains a family of distributions each $\hat{\beta}_i$, the choice of the density q_0 is not arbitrary. Huber (2004) showed that this particular choice ensures that the likelihood ratio between p_a and q_0 is a max-min statistic for each $\hat{\beta}_i$. In particular, when the design matrix is orthogonal, $\hat{\beta}_i$ are independent conditional on β. In this case the test statistic T is max-min optimal.

The exact distribution of T under the null is difficult to express, however the following result allows us to simulate the rejection region.

Theorem 2. *Under $H_0(\epsilon)$, T is first order stochastically dominated by*

$$S := \frac{1}{p}\sum_{i=1}^{p} \frac{(1-\sigma^2)n}{(1-\sigma^2)n + \sigma^2 s_i^2 p} \times \left((1-z_i)\min\{e_i^2, \frac{x_i^*}{\sigma^2 s_i^2/n}\} + z_i \frac{x_i^*}{\sigma^2 s_i^2/n}\right),$$

where $z_i \sim_{iid} Bernoulli(\epsilon)$ and

$$e \sim \mathcal{N}\left(0, diag\left(\sqrt{\frac{n}{s_1^2}}, \ldots, \sqrt{\frac{n}{s_p^2}}\right)(X'X)^{-1} diag\left(\sqrt{\frac{n}{s_1^2}}, \ldots, \sqrt{\frac{n}{s_p^2}}\right)\right).$$

In particular, this first order stochastic upper bound is tight.

Therefore, a proper alpha level of the test can be defined as the region $T \geq t_\alpha$, where $\Pr(S \geq t_\alpha) \leq \alpha$. This region can be simulated. The order of the rejection region can be easily bounded using the Markov's inequality.

Proposition 3. *The random variable S is of order*

$$O_p\left(\mathbb{E}[S]\right) \leq O_p\left(1 + \frac{\epsilon}{p}\sum_{i=1}^{p}\frac{x_i^*}{\sigma^2 s_i^2/n}\right).$$

Proof. Observe that

$$\begin{aligned} S =& \frac{1}{p}\sum_{i=1}^{p}\frac{(1-\sigma^2)n}{(1-\sigma^2)n+\sigma^2 s_i^2 p} \times \left((1-z_i)\min\{e_i^2, \frac{x_i^*}{\sigma^2 s_i^2/n}\} + z_i\frac{x_i^*}{\sigma^2 s_i^2/n}\right) \\ \leq& \frac{1}{p}\sum_{i=1}^{p}(e_i^2 + z_i\frac{x_i^*}{\sigma^2 s_i^2/n}). \end{aligned}$$

Let Σ be the covariance matrix for e, we have

$$\mathbb{E}\sum_{i=1}^{p} e_i^2 = \mathbb{E}[e'e] = \mathbb{E}[e'\Sigma^{-1/2}\Sigma\Sigma^{-1/2}e] = p.$$

for the trace of Σ is p. The rest of the proposition follows directly from Markov's inequality.

□

Usually asymptotic consistency means a test rejects the null with probability approaching 1 as n diverges. However, since we allow both p and the design matrix (hence s_i^2's) to vary with n, the asymptotic consistency of this test means the rejection probability approaches 1 for a sequence of null and alternatives. In particular, the sequence of null is a sequence of models that are asymptotically sparse. Since p can increase as n increases, the number of non-zero coefficients in the sequence of models can potentially increase as a result. However we need to avoid the pathological case where the number of non-zero coefficients increases faster than n. Following Meinshausen and Buhlmann (2006), Zhao and Yu (2006) and Huang et al. (2008), we assume the fraction of non-zero coefficients goes to zero, and the number of non-zero coefficients grows at a rate less than one. Mathematically, we define *asymptotic sparsity* as follow.

Condition 4. *As n increases, there exist constants $\alpha_1 > 0, \alpha_2 \in [0, 1)$ such that $\epsilon = \alpha_1 n^{\alpha_2 - 1}$.*

This condition has implication for the test statistic under the null. The cut-off values are implicitly affected by the above assumption.

Proposition 5. *Let $n \to \infty$, if there exists a positive constant κ such that $\frac{n}{ps_i^2} \geq \kappa$ for all i, then Condition 4 implies that there exists c_1 and c_2 such that $0 < c_1 < c_2$ and for all i,*

$$c_1 \ln n \leq \frac{x_i^*}{\sigma^2 s_i^2 / n} \leq c_2 \ln n.$$

And hence $S = O_p(1 + \epsilon c_2 \ln n) = O_p(1)$ asymptotically.

Since under the null, S dominates the test statistic, therefore Condition 4 implies the test statistic under the null is of finite order. To obtain a consistency, we can show that the test statistic diverges to infinity under the alternative when some sufficiency condition holds. One sufficient condition is that

$$\frac{n\lambda}{p} \geq \kappa \ln n$$

where λ is the minimal eigenvalues of $X'X/n$. Since $X'X/n$ is a normalized, its minimal eigenvalue can be thought of as a measure of multiple-colinearty of the design matrix. The above condition requires the effective number of observations per coefficient diverges slowly.

Theorem 6. *Let the minimal eigenvalues of $X'X/n$ be λ. Suppose there exists some constant $\kappa > 0$ such that $\frac{n\lambda}{p} \geq \kappa \ln n$ always holds. Suppose Condition 4 holds. Then under H_a, T diverges to ∞ in probability as $n, p \to \infty$. Hence the test is consistent.*

1.5 Further Discussions

In this section, we discuss three questions related to the application of the test. They include the cases when σ^2 is unknown, when $p > n$ and the choice of ϵ.

1. Unknown σ^2.

When σ^2 is unknown, we can plug in the residual mean-squared error $\hat{\sigma}^2$ from OLS estimates. The plug-in test statistics is then

$$\hat{T} := \frac{1}{p}\sum_{i=1}^{p} \frac{(1-\hat{\sigma}^2)n}{(1-\hat{\sigma}^2)n + \hat{\sigma}^2 s_i^2 p} \times \min\{\frac{\hat{\beta}_i^2}{\hat{\sigma}^2 s_i^2/n}, \frac{\hat{x}_i^*}{\hat{\sigma}^2 s_i^2/n}\},$$

where $\hat{x}_i^*$ solves $(1-\epsilon)\,\mathrm{erf}\left(\sqrt{\frac{x}{2\hat{\sigma}^2 s_i^2/n}}\right) + (1-\epsilon)\sqrt{\frac{\hat{v}_i}{\hat{\sigma}^2 s_i^2/n}} e^{\left(\frac{1}{v_i} - \frac{1}{\hat{\sigma}^2 s_i^2/n}\right)\frac{x}{2}}\,\mathrm{erfc}\left(\sqrt{\frac{x^*}{2\hat{v}_i}}\right) = 1$, and $\hat{v}_i = \frac{1-\hat{\sigma}^2}{p} + \hat{\sigma}^2 s_i^2/n$. An application of Theorem 2 shows that under the null, the above test statistic is 1st order dominated by the following random variable.

$$S' := \frac{1}{p}\sum_{i=1}^{p} \frac{(1-\hat{\sigma}^2)n}{(1-\hat{\sigma}^2)n + \hat{\sigma}^2 s_i^2 p} \times \left((1-z_i)\min\{e_i^2\frac{\sigma^2}{\hat{\sigma}^2}, \frac{\hat{x}_i^*}{\hat{\sigma}^2 s_i^2/n}\} + z_i \frac{\hat{x}_i^*}{\hat{\sigma}^2 s_i^2/n}\right)$$

where $z_i \sim_{iid} Bernoulli(\epsilon)$, $e \sim \mathcal{N}(0, \Delta(X'X)^{-1}\Delta)$ for $\Delta := diag\left(\sqrt{\frac{n}{s_1^2}}, \ldots, \sqrt{\frac{n}{s_p^2}}\right)$. Since σ^2 is unknown, the rejection region is simulated from the random variable

$$\hat{S} := \frac{1}{p}\sum_{i=1}^{p} \frac{(1-\hat{\sigma}^2)n}{(1-\hat{\sigma}^2)n + \hat{\sigma}^2 s_i^2 p} \times \left((1-z_i)\min\{e_i^2, \frac{\hat{x}_i^*}{\hat{\sigma}^2 s_i^2/n}\} + z_i \frac{\hat{x}_i^*}{\hat{\sigma}^2 s_i^2/n}\right).$$

The following sufficiency result shows the difference between S' and $\hat{S}$ can be asymptotically negligible.

Proposition 7. *Let $\hat{S}$ and S' be defined as above. We have $\frac{\hat{S}-S'}{\sqrt{Var(\hat{S})}} \to 0$ as both ϵ and $\frac{p}{n\lambda} \to 0$.*

2. $p > n$.

For problems involving a data set where $p > n$, our test can be used as a post-selection test. Under the null hypothesis of a sparse DGP, one can split the data into two disjoint subsets perform any desired screening procedures to the first subset. For example, the Dantzig selector (Candes and Tao (2007)) and Sure Independence Screening (Fan and Lv (2008)) can be used to screen all the important variables while reduces the number of parameters to less than the number of observations. Methods to obtain a $\sqrt{n}$-consistent estimate for σ^2 is available in the literature as well.[4] Our test can be subsequently applied to the second part of the data.

3. Choice of ϵ.

If one has some preconception about which sparsity level to test for, one can fix such ϵ level and perform the test. When there is little preconception about the sparsity level, our test can be turned into a confidence set about the sparsity of the underlying model. See empirical application section for more detail.

1.6 Simulations

In this section we report the results of some simulation experiments. In each subsection we simulate datasets from the following model

$$Y = X\beta + \mathcal{N}(0, \sigma^2)$$

for various sizes and number of observations. In all of them, the covariates x_i $(i = 1, \ldots, n)$ is simulated from a p dimensional multivariate normal where p is the dimension of β.

[4]See Reid et al. (2016) for a survey and comparison of these estimators.

1.6.1 Simulation Under Alternative 1

In this simulation, we repeat the same setting as in Figure 1.1. β is fixed at the same levels respectively when the number of observations is 200 and 500. We simulate X from standard multivariate normal and the residual is standard normal $\mathcal{N}(0,1)$ resulting in a signal to noise ratio of roughly 1.6. For each level of ϵ, we simulate the data and perform the test 500 times and report the rejection rate below.

Table 1.1: Rejection probabilities for various ϵ values

ϵ	0.1	0.2	0.3	0.4	0.5	0.6
p=50, n=200	100%	100%	100%	97.5%	80.4%	12.5%
p=100, n=500	100%	100%	100%	100%	100%	82.6%

1.7 Empirical Application

1.7.1 Application I

We apply our test to the a cross country growth data set. In the data subset, there are observation on 135 countries each with observation of 67 characteristics plus the response variable, GDP growth rate from 60 to 96. Since there are many missing observations in the data, we apply our test to only a subset of the sample. We use the subset of the sample where there is no missing observations on the following 18 variables

East Asian dummy (EAST)	African dummy (SAFRICA)
Primary schooling 1960 (P60)	Latin American dummy (LAAM)
Investment price (IPRICE1)	Fraction GDP in mining (MINING)
GDP 1960 (log) (GDPCH60L)	Spanish colony (SPAIN)
Fraction of tropical area (TROPICAR)	Years open 1950-1994 (YRSOPEN)
Population density coastal 1960's (DENS65C)	Fraction Muslim (MUSLIM00)
Malaria prevalence in 1960's (MALFAL66)	Fraction Buddhist (BUDDHA)
Life expectancy in 1960 (LIFE060)	Ethnolinguistic fractionalization (AVELF)
Fraction Confucian (CONFUC)	Government consumption share 1960's (GVR61)

which are a number of economic and political factors, geographical and historical dummies, and several demographic characteristics that were described as potential important factors in explaining long-run GDP growth in Sala-I-Martin et al. (2004). There are 94 observations that have no missing observations in the above listed variables. These countries

or regions are

Algeria	Benin	Botswana	Burkina Faso
Burundi	Cameroon	Cent'l Afr. Rep.	Congo
Egypt	Ethiopia	Gabon	Gambia
Ghana	Kenya	Lesotho	Liberia
Madagascar	Malawi	Mali	Mauritania
Morocco	Niger	Nigeria	Rwanda
Senegal	Somalia	South Africa	Tanzania
Togo	Tunisia	Uganda	Zaire
Zambia	Zimbabwe	Canada	Costa Rica
Dominican Rep.	El Salvador	Guatemala	Haiti
Honduras	Jamaica	Mexico	Nicaragua
Panama	Trinidad & Tobago	United States	Argentina
Bolivia	Brazil	Chile	Colombia
Ecuador	Paraguay	Peru	Uruguay
Venezuela	Bangladesh	Hong Kong	India
Indonesia	Israel	Japan	Jordan
Korea	Malaysia	Nepal	Pakistan
Philippines	Singapore	Sri Lanka	Syria
Taiwan	Thailand	Austria	Belgium
Denmark	Finland	France	Germany, West
Greece	Ireland	Italy	Netherlands
Norway	Portugal	Spain	Sweden
Switzerland	Turkey	United Kingdom	Australia
New Zealand	Papua New Guinea		

Many economic models focus analyses on a couple of factors and their relation with long-run growth. For example Sala-I-Martin et al. (2004) focuses their arguments on primary schooling enrollment, investment price and initial GDP levels. Therefore, in this numerical exercise we will set the sparsity parameter to be $\epsilon = 3/18$, interpreted as whether the variation in longrun growh can be sufficiently explained by 3-variable (linear regression) model. We simulate 10k random draws from the upperbound distribution of the null (see Thm 7). The 5% rejection is defined as the upper 5% quantile of the simulated sample. The test statistic calculated from the data is above the 5% quantile and has a p-value of less than 1.7%. Hence we reject the null that the cross country long run GDP growth can be explained by a (three factors or fewer) sparse linear model, and accept the alternative that a non-sparse model of multiple (18) small effects is better supported by the data.

It might be interesting to know which variables pass our robust thressholds. They are Primary schooling enrollment in 1960, initial GDP level 1960, investment price, life expectancy in 1960 and fraction of GDP in mining. One can interpret it as an indication that these variables may be more important than others in determining long run GDP growth.

Although we set the benchmark case to be $\epsilon = 3/p$ from a modelling perspective, it would be interesting to see how the test would conclude if we apply a less strict sparsity parameter. To this end we report the p-values for several different sparsity below.

$\epsilon \times p$	1	2	3	4	5	6	7	8	9	10	11	12
p-value (%)	0.2	0.9	1.6	1.8	2.2	2.6	3.4	3.9	5.1	5.6	17.2	57.2

The above p-values shows strong evidence for at least 9 non-zero variables, indicating that the underlying DGP is not sparse. Nonetheless, it does not contradicts the proposal of Sala-I-Martin et al. (2004) that primary schooling, initial GDP level and investment prices are very important factors. It suggest that long-run GDP growth is a complex high dimensional object that is affected by many different country-level characteristics.

1.7.2 Application II

Ludvigson and Ng (2009, 2010) found that the excess return of U.S. government bonds is predictable using macroeconomic fluctuations. They found that macroeconomic fundamentals contain information about risk premia beyond those embedded in bond market data. In this section, we apply our test to their prediction problem. The macro-factor data is taken from the updated data file is provided on Ludvigson's website. The eight factors, $f_1, \ldots, f_8$ each have different interpretations based on their loading of the sample series. According to Ludvigson (2009), f_1 is the factor of economy activity in real-terms; f_2 loads on interest rate spreads; f_3 and f_4 are price factors; f_5 is mainly a combination of interest rates (but not so much of interest rate spreads); f_6 loads on housing; f_7 on money supply; and f_8 loads mainly on stock-related series. The response variable $r_{t+1}^{(n)}$, the continuously compounded (log) excess return on an n-year discount bond in year $t+1$, is also taken from Ludvigson's website. The response data span from 2-year excess return to 5-year excess return. Due to the availability of the response series and the lag-12 month regression, we have in total 468 observations.

Following their papers, we regress $r_{t+1}^{(n)}$ on CP_t, the forward rate factor used in Cochrane and Piazzesi (2005), and the eight macro factors plus their interaction terms up to the third order. In other words, as predictors we have $CP, f_1, \ldots f_8$ and all of the $f_i \times f_j$, and $f_i \times f_j \times f_k$ where $1 \leq i \leq j \leq k \leq 8$, totaling to 109 predictors. Ludvigson and Ng (2009) uses BIC and searched through low dimensional models, and conclude that the best model consists of CP, f_1, f_1^3, f_3, f_4 and f_5. Our test results for excess return for different periods are reported below.

Overall, the 5% level test rejects extremely sparse models. The 95% confidence interval varies, covering from as less as models with dimensions ≥ 7, to as much as models of dimensions ≥ 5. The size of the model found by Ludvigson and Ng (2009) barely lies in

Table 1.2: p-values (%) for various sparsity levels

$\epsilon \times p$	1	2	3	4	5	6	7	8	9	10	11
response: $r_{t+1}^{(2)}$	0.4	1.1	2.3	2.9	4.0	4.2	5.3	6.1	6.9	8.8	10
response: $r_{t+1}^{(3)}$	0.2	0.4	0.9	2.5	2.9	3.4	6.5	6.9	8.1	10.6	13.4
response: $r_{t+1}^{(4)}$	0.1	0.7	2.3	3.6	5.0	7.0	9.7	11	14	15	16.7
response: $r_{t+1}^{(5)}$	0.2	0.8	2.6	4.9	8.3	11	14	17	20	23	25

these confidence intervals. A closer examination of the confidence interval also indicate that the longer the maturity of the Treasury bonds, the sparser the regression model becomes. Potentially short term returns can be affected by more factors whereas in the longer terms, only the most important facts has lasting effects. Nonetheless, our test supports the use of a sparse model to predict such excess returns.

Bibliography

[1] Aguiar M., and Hurst E. (2013), "Deconstructing Life Cycle Expenditure", *Journal of Political Economy*, Vol 121, No. 3.

[2] Abramowitz, M. and Stegun, I.A. (1972). *Handbook of Mathematical Functions*,Dover Publications, pp 298.

[3] Akaike, H. (1974). "A new look at the statistical identification model". *IEEE. Trans. Auto. Control.* 19. (6) 716-723.

[4] Azriel, D., and A. Schwartzman (2015) "The Empirical Distribution of a Large Number of Correlated Normal Variables" *Journal of the American Statistical Association* 110:511, 1217-1228

[5] Blanciforti LA, Green RD, King GA (1986). "U.S. Consumer Behavior Over the Postwar Period: An Almost Ideal Demand System Analysis." Monograph Number 40 (August 1986), Giannini Foundation of Agricultural Economics, University of California.

[6] Candes, E., and T. Tao (2007) "The Dantzig selector: Statistical estimation when p is much larger than n" *Ann. Statist.* Volume 35, Number 6, 2313-2351.

[7] Chen, Jiahua, and Zehua Chen (2008) "Extended Bayesian Information Criteria for Model Selection with Large Model Spaces". *Biometrika*, Volume 95, Issue 3. Sep. 2008.

[8] Cochrane, J. H., and M. Piazzesi. (2005). "Bond Risk Premia". *American Economic Review* 95:138-60

[9] Deaton A, Muellbauer J (1980). "An Almost Ideal Demand System." *The American Economic Review*, 70(3), 312-326.

[10] Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association,* 96, 1348-1360.

[11] Fan, J. and Lv, J. (2008) "Sure independence screening for ultrahigh dimensional feature space" *J. R. Statist. Soc. B* **70**, pp. 849-911

[12] Giannone, D., M. Lenza and G.E. Primiceri (2017) "Economic Predictions with Big Data: the Illusion of Sparsity" *working paper*

[13] Hausman J., and Leonard G. (2002), "THE COMPETITIVE EFFECTS OF A NEW PRODUCTINTRODUCTION: A CASE STUDY", *Journal of Industrial Economics*, Vol L. No. 3

[14] Horn, A. (1954), "Doubly stochastic matrices and the diagonal of a rotation matrix", *American Journal of Mathematics* 76, 620-630.

[15] Huang, J., Ma, S. and Zhang, C., (2008) "Adaptive LASSO for Sparse High-dimensional Regression Models", *Statistica Sinca* 18, pp 1603-1618.

[16] Huber, P. (2004) *Robust Statistics*. New Jersy. John Wiley & Sons, Inc.

[17] Hsu, D., Kakade, S. M. and Zhang, T (2014). "Random Design Analysis of Ridge Regression" *Found. Comput. Math* 14. 569-600.

[18] Jaynes, E.T. (1968). "Prior Probabilities". *IEEE Trans. on Systems Science and Cybernetics*. 4 (3): 227.

[19] Ludvigson, S. and S. Ng, (2009) "Macro Factors in Bond Risk Premia". *The Review of Financial Studies*, 2009, 22(12): 5027-5067.

[20] Ludvigson, S. and S. Ng, (2010) "A Factor Analysis of Bond Risk Premia" . *Handbook of Empirical Economics and Finance*, 2010, e.d. by Aman Uhla and David E. A. Giles, pp. 313-372.

[21] Meinshausen, N. and Buhlmann, P. (2006). "High dimensional graphs and variable selection with the Lasso". *Ann. Statist.* 34 1436-1462.

[22] Mitchell, T. J., and J. J. Beauchamp (1988) "Bayesian Variable Selection in Linear Regression," *Journal of the American Statistical Association*, 83, 1023-1032.

[23] Reid, S., R. Tibshirani and J. Friedman (2016). "A Study of Error Variance Estimation in LASSO Regression". *Statistica Sinica* Vol. 26, No.1. pp 35-67.

[24] Schur, I. (1923) "Uber eine Klasse von Mittelbildungen mit Anwendungen auf die Determinantentheorie", *Sitzungsber. Berl. Math. Ges.* 22, 9-20.

[25] Schwarz, G. E. (1978), "Estimating the dimension of a model", *Annals of Statistics*, 6 (2): 461-464

[26] Tibshirani, R. (1996). "Regression Shrinkage and Selection via the lasso". *Journal of the Royal Statistical Society. Series B (methodological)*. Wiley. 58 (1): 267-88.

[27] Zhang, C. H. and Huang, J. (2008). "The sparsity and bias of the Lasso selection in high-dimensional linear regression." *Annals of Statistics*, 36, 1567-1594.

[28] Zhao, P. and Yu, B. (2006). "On model selection consistency of LASSO". *J. Machine Learning Research* 7 2541-2567.

[29] Zou, H. (2006). "The adaptive lasso and its oracle properties" *Journal of the American statistical association* 101 (476), 1418-1429

[30] Zou, Hui and Hastie, Trevor (2005). "Regularization and Variable Selection via the Elastic Net". *Journal of the Royal Statistical Society*, Series B: 301-320.

[31] Zou, H., Hastie, T. and Tibshirani R. (2007) "On the "degrees of freedom" of the lasso" *The Annals of Statistics* Volume 35, Number 5, 2173-2192.

Chapter 2

Optimal Estimation when the Parameter Space is of Infinite Dimension

Coauthored with *Werner Ploberger*

2.1 Introduction

There is a vast literature on nonparametric estimators, but relatively less is known about their optimality properties. In the seminal paper, Andrews (1991) investigated a lot of estimators for their asymptotic variances. Essentially the problem is to estimate the long-term variance of the score process. Right at the beginning of his paper, he states that "Currently the consistency of these estimators has been established, but their relative merits are unknown". Since then, much progress has been made. An overview of recent developments can be found in Gine and Nickl (2016), and Armstrong and Kolesar (2018). In this paper, we apply the methodology similar to that in Ploberger and Phillips (2003, 2012) to derive admissible estimators in general non-parametric settings.

In the case of finite dimensions, a well-established theory of "optimal" estimation is

developed by Le Cam, Blackwell and Hayek (cf. van der Vaardt, 2000; Strasser, 1996). This theory allows a characterization of the maximum-likelihood estimator as "best estimator" compared to a large class of competing estimators. In this paper, we investigate the case of a infinite dimensional parameter space. The theory of finite dimensional case does not immediately generalize, since the maximum-likelihood estimator is usually not well-defined in infinite-dimensional settings. When it is defined, there are examples where the maximum likelihood estimator is inconsistent.

We assume that our parameter space Θ is a subset of the $\boldsymbol{R}^{\boldsymbol{N}}$, where $\boldsymbol{N} = \{1, 2, \ldots\}$ is the set of all natural numbers. So our parameters θ are sequences,

$$\theta = (\theta_1, \theta_2, \ldots) .$$

We assume that we have given a squence of data - at time n, our information is contained in the σ-algebra $\mathfrak{F}_n$ and for each θ a measure P_θ on the σ-algebra $\mathfrak{F} \supseteq \sigma$-algebra $\mathfrak{F}_n$. Although this formulation seems different from many problems in nonparametric estimations, any classical nonparametric models can be formulated within this framework. By Interpreting the θ_i as Fourier coefficients, our methodology can be applied in any problems of estimating reasonable functions. As examples, consider three traditional nonparametric problems.

- General models for stationary Gaussian process.

  Assume that the data y_t are generated by an infinite autoregressive process

  $$y_t = \sum_{k \geq 1} \gamma_k y_{t-k} + u_t,$$

  where u_t is Gaussian white noise uncorrelated with y_{t-i}. Let the spectral density f be

the parameter of interest. Then

$$f(\exp(i\lambda)) = \left| \frac{\sigma_u^2}{1 - \sum_{k \geq 1} \gamma_k \exp(ik\lambda)} \right|^2.$$

So our parameter $\theta = (\theta_1, \theta_2, ..) = (\sigma_u^2, \gamma_1, \gamma_2, \ldots)$, and it is easy to set up the likelihood as a function of θ.

- Nonparametric regression with Gaussian errors.

  Assume for simplicity the case of a single scalar regressor x_t, which takes values in a fixed interval. Without limitation of generality, let this interval equals $[0, \pi]$. Let y_t be the dependent variable. Consider the model

$$y_t = f(x_t) + u_t$$

  where the u_t are i.i.d. $G(0, \sigma^2)$. The function f can be written as a Fourier series

$$f(x) = \sum_{n=0}^{\infty} \gamma_n \cos(nx).$$

  Again, our parameter $\theta = (\theta_1, \theta_2, ..) = (\sigma_u^2, \gamma_0, \gamma_1, \ldots)$, and the likelihood can be written down accordingly.

- Density estimation.

  Assume there is a sample of i.i.d random variables X_i, taking values in an interval $[a, b]$ and one tries to estimate the density f. Suppose $\ln f$ is a square integrable function. Then one can choose a complete orthonormal set of functions φ_n (e.g. trigoniometric functions), and write

$$\ln f(.) = C(\gamma_1, \ldots) + \sum \gamma_n \varphi_n(.).$$

where $C(.,\dots)$ is chosen in such a way that $\int \exp(\ln f) = 1$, and thus is a function of the γ_n. Then our $\theta = (\theta_1, \theta_2, ..) = (\gamma_1, \dots)$, and we can define a likelihood accordingly.

In all of the examples given above, the parameter vector θ, is related to a function. In nonparametric problems, it is often assumed that the function is smooth, i.e. differentiable up to a certain order or more. If the parameters are Fourier coefficients, then differentiability of the underlying function is determined by the decay of the coefficients. I.e. if $(\theta_1, \theta_2, \dots)$ represents the Fourier coefficients of the function

$$f(\omega) = \sum \theta_k exp(ik\omega),$$

then the Fourier coefficients of the mth derivative $f^{(m)}$ equal $(\dots, i^m k^m \theta_k, \dots)$. Hence for $f^{(m)}$ to be square-integrable,

$$\sum \theta_k^2 k^{2m} < \infty. \tag{2.1}$$

So imposing growth conditions on the coefficients is essentially equivalent to the requirements of varying degrees of smoothness of the underlying function. We will assume that the prior distributions on the set of parameters are essentially independent Gaussian distributions with expectations zero and variances c_k^2, where c_k^2 converges to zero. Consequently, parameters with larger index will be very small.

2.2 Main Theorems

Our primary goal is the estimation of θ, and define criteria to compare estimators, and especially finding the "best" estimator. We will use an adaption of a technique used quite often in the finite dimensional context. In order to find the asymptotically optimal estimator, one first establishes that the posterior distribution is asymptotically $G(\hat{\theta}, \hat{\Sigma})$, and thereby simplifying the problem. When the Gaussianity is established, it almost follows that $\hat{\theta}$ is the

"optimal" estimator. The importance of "conditional Gaussianity" was first recognized by Kim(1998). Ploberger and Phillips (2012) utilized this property to characterize estimator in cases of stochastic information matrices.

Our first task is to establish that the posterior distribution is asymptotically normal. To do so, we need only a few more than the standard assumptions.

First of all let us assume that the conditional log-likelihoods

$$\ell_t(\theta) = \ln p_\theta(x_t|x_{t-1}, ...)$$

are 3 times differentiable, all have uniformly bounded second moments, and the requirements for all the CLTs for scores and information-matrix-type theorems are fulfilled. Furthermore, we assume that the eigenvalues of the expected information matrix are bounded from above and below. We argue that this is quite a plausible requirement, since it guarantees that all unidimensional restrictions of the parameter (i.e. curves of parameters) allow for standard ML estimation. Furthermore we assume that the "long-run" variances for all the derivatives of the log-likelihood are uniformly bounded.

Heuristically, for large n the average of n such expression differs from its expectation by $O(1/\sqrt{n})$. We also assume that for some bounded "neighborhood" O of the parameter,

$$E_\theta \left(\sup_{\theta \in O} \left| \frac{\partial^3 \ell_t(\theta)}{\partial\theta_i \partial\theta_j \partial\theta_k} \right| \right) \leq M \tag{2.2}$$

We do not want to assume that the parameter space Θ is the whole $\mathbf{R^N}$ This may be inconvenient. Consider e.g. the case of z-transforms of autoregressive parameters. The parameters have to be such that there are no zeroes in or on the unit circle. It would be very inconvenient to describe this set of these parameters directly. We will later on describe some of the assumptions of the parameter set. We assume, however, that our parameterspace is

an open set in a topology where consistent estimation of the parameter θ is possible.

We construct our prior distribution by starting with a product of Gaussian distributions with zero mean and variances c_i^2. The support of theses measures is, however, the whole space $\mathbf{R^N}$. So we have to restrict our prior distribution to our parameter space Θ. Let $C^{-1} = diag(c_i^2)$. To exclude trivial cases, we assume that

$$G(0, C^{-1})(\Theta) > 0.$$

Let us suppose that

$$k^8 c_k = o(1). \tag{2.3}$$

This condition is a bit stringent. Essentially we assume that the function describing the parameter θ is differentiable eight times, a bit extreme even for nonparameterics. However it should be possible to reduce the smoothness requirements to a reasonable form. We will be rather wasteful when we compute bounds for matrices with increasing dimensions. Let us now define

$$(A_n)_{i,j} = \left(\sum_{t=1}^{n} \frac{\partial^2 \ell_t(\hat{\theta})}{\partial \theta_i \partial \theta_j} \right).$$

Then we have the following theorem.

Theorem 8. *Assume the above conditions are met. Let Π_n be the posterior distribution of the parameter θ, and let $\hat{\theta}_n$ be the ML-estimator for first $n^{1/7}$ entries, holding the rest zero. Then the total variation of the difference between Π_n and $G((A_n + C)_{-1} A_n \hat{\theta}_n, (A_n + C)^{-1})$ converges to zero in probability with respect to P_n.*

The Π_n depends on data, and is a random probability measure on $\mathbf{R^N}$. It can be seen that the "mean" of the posterior distribution is not the ML-estimator, but some kind of "shrinkage estimator". We can think of the estimator as a linear combination of ML-estimator and prior mean, which we have set to zero. Moreover, it can be seen that $(A_n + C)^{-1} A_n \hat{\theta}_n$ is

asymptotically equivalent to the penalized ML estimator that maximizes

$$\sum l_t(\theta) - \theta' C\theta/2. \tag{2.4}$$

The proof is very technical, and is postponed to the appendix. Any Bayesian estimator, especially the conditional mean of a parameter is admissible by construction. The conditional mean of a parameter is evidently the estimator with minimum variance. However, our estimator is not exactly the Bayesian estimator, but only so asymptotically. Hence it is not even known if the expectation of the estimator exists.

Let us take an arbitrary estimator $\tilde{\theta}$ and some matrices B_n (which select the components we are interested in) and consider the quadratic distance

$$Q(\tilde{\theta}) = (\theta - \tilde{\theta})' B_n B_n' (\theta - \tilde{\theta}).$$

Typical examples includes

- B_n projects on finitely many components of θ: finitely many parameters;
- $B_n' = (1,1,...)$: Sum of parameters, "long-term" parameters. It now would be natural to try to minimize "average" $Q(\tilde{\theta})$ over all estimators.

The main problem is the fact that for many estimators, $Q(\tilde{\theta})$ may have very nice asymptotic properties, but the expectation does not exist, or would be hard to compute. Typical examples are the usual ML estimators for finite-dimensional parameters. The estimation errors are asymptotically normal, but the exact moments are often unknown. Therefore we classify estimators according to

$$Ef(Q(\tilde{\theta})),$$

where E is the expectation wrt to prior(s) and f is from a class of "squashing function".

Each of such function is bounded, but the class should be large enough to approximate the identity function.

A "Classical Example" is Le Cam theory. A function $\phi(.)$ defined on a finite dimensional vector space is called "bowl-shaped" if it is bounded and its level-sets are symmetric and convex. In this context, Anderson's lemma (cf. Strasser, 1995) guarantees that the expectations of all bowl-shaped functions of estimation errors are minimized by the mean of the asymptotic normal distribution. In fact, Anderson's lemma allows us our next asymptotic result. Let $b = (b_1, b_2, \dots)$ be a sequence that

$$\sum b_k^2/c_k < \infty. \tag{2.5}$$

The difference of the posterior distribution of $b'\theta$ and $G(b'(A_n + C)^{-1}A_n\hat{\theta}_n, b'(A_n + C)^{-1}b)$ converges to zero. Let

$$\sigma_n = \sqrt{b'(A_n + C)^{-1}b}. \tag{2.6}$$

Then the posterior distribution of

$$b'\theta - b'(A_n + C)^{-1}A_n\hat{\theta}_n \tag{2.7}$$

converges to a standard normal. Now let $f(.)$ be a "bowl shaped" function defined over the real line. We have the following theorem.

Theorem 9. *Let $\hat{\mu}_n = b'(A_n + C)^{-1}A_n\hat{\theta}_n$, where b satisfies (5). Let σ_n be defined as (6). Then for any "bowl shaped" function f and any other estimator $\tilde{\mu}_n$, we have*

$$\limsup_{n\to\infty} Ef((\hat{\mu}_n - b'\theta)/\sigma_n) - Ef((\tilde{\mu}_n - b'\theta)/\sigma_n) \leq 0.$$

Hence the best estimator for $b'\theta$ is $b'(A_n+C)^{-1}A_n\hat{\theta}_n$. For the next result, we use a specific

class of loss functions but allow for more general "scaling matrices".

Let $f : [0, \infty) \to [0, \infty)$ be "completely monotone", i.e. differentiable infinitely often, and

$$(-1)^k f^{(k)}$$

is negative. Typical examples include $\exp(-sx)$, $\frac{a}{b+cx^a}$, and $f(x) = \frac{ax}{x+a}$ for arbitrary $a > 0$.

Let

$$A_n := \sum \left(\frac{\partial \ell}{\partial \theta}\right) \left(\frac{\partial \ell}{\partial \theta}\right)' \approx -\sum \frac{\partial^2 \theta}{\partial \theta^2}.$$

Then we call B_n "reasonably normed" if and only if

$$tr(B_n'(A_n + C)^{-1} B_n) = O(1).$$

Furthermore, observe that

$$B_n'(A_n + C)^{-1} B_n$$

$$= B_n' \sqrt{C}^{-1} (\sqrt{C}^{-1} A_n \sqrt{C}^{-1} + I)^{-1} \sqrt{C}^{-1} B_n.$$

We are now ready to state and prove the following theorem.

Theorem 10. *Assume that total variation of the difference between the posterior distribution for the parameter θ and $G(\hat{\theta}, (A_n + C)^{-1})$ converges to zero for some estimator $\hat{\theta}$. Let $\tilde{\theta}$ be an arbitrary estimator. Then the following propositions are equivalent.*

1. *For " reasonably normed" B_n, $(\tilde{\theta} - \hat{\theta})' B_n B_n' (\tilde{\theta} - \hat{\theta})$ does not converge to 0 stochastically wrt P_n*

2. *For one (nontrivial) of our loss functions f*

$$\lim \Big(Ef(Q_n(\tilde{\theta})) - Ef(Q_n(\hat{\theta})) \Big) > 0.$$

3. For all of our loss functions f,

$$\lim \Big(Ef(Q_n(\tilde{\theta})) - Ef(Q_n(\hat{\theta})) \Big) > 0.$$

Proof: The structure of completely monotonic functions is well known. Bernstein's Theorem (cf. Bernstein (1928)) guarentees that every completely monotonic function f can be written as

$$g(x) = \int_0^\infty \exp(-sx) d\mu(x).$$

Hence it suffices to analyse

$$E_n \exp(-sQ_n(\tilde{\theta})) - E_n \exp(-sQ_n(\hat{\theta}))$$

Observe that

$$\tilde{\theta} - \theta = (\tilde{\theta} - \hat{\theta}) + (\hat{\theta} - \theta)$$

and therefore

$$\begin{aligned} \exp(-sQ_n(\tilde{\theta})) = & \exp(-(\tilde{\theta} - \hat{\theta})' s B_n B_n' (\tilde{\theta} - \hat{\theta})) \\ & \exp(-(\tilde{\theta} - \hat{\theta})' 2 s B_n B_n' (\tilde{\theta} - \hat{\theta})) \\ & \exp(-(\tilde{\theta} - \theta)' s B_n B_n' (\tilde{\theta} - \theta)) \end{aligned} \tag{2.8}$$

Now compute

$$E_n(\exp(-sQ_n(\tilde{\theta}))/X_1, ...X_n).$$

The first factor of $\exp(-sQ_n(\tilde{\theta}))$ is $X_1, ...X_n$ measurable and therefore can be taken outside of the conditional expectation.

Moreover, the conditional expectation of a function of θ is exactly the expectation with

respect to the posterior distribution.

We did assume that the difference of the posterior and normal with expectation $\hat{\theta}$ and variance $(A_n + C)^{-1}$ converges to zero. Since all the function involved are bounded, we can asymptotically replace the posterior with the normal. For easier manipulation of the infinite-dimensional matrices, observe that $(A_n+C)^{-1} = \sqrt{C}^{-1}(\sqrt{C}^{-1}A_n\sqrt{C}^{-1}+I)^{-1}\sqrt{C}^{-1}$.

Then the integral of the exponentiated quadratic expression wrt a Gaussian can be calculated in closed form, as in the finite dimensional case.

The integral for the two factors

$$\exp((\tilde{\theta}-\hat{\hat{}})'(sB_nB_n')(A_n+2sB_nB_n'+C)^{-1}(sB_nB_n')(\tilde{\theta}-\hat{\hat{}})/2)$$

can be reduced to

$$\sqrt{\det(\sqrt{C}^{-1}A_n\sqrt{C}^{-1}+I)}/ \tag{2.9}$$

$$\sqrt{\det(\sqrt{C}^{-1}A_n\sqrt{C}^{-1}+I+\sqrt{C}^{-1}B_n\sqrt{C}^{-1})}. \tag{2.10}$$

Since we assumed that the diagonal elements of C^{-1} decrease rapidly, the corresponding determinants are well defined even if A_n and B_n are infinite matrices. See Lang (1993) for a detailed reference to the trace class operators.

Now observe that $E_n \exp(-sQ_n(\tilde{\theta}))$ is asymptotically equivalent to the product of

$$\exp(-(\tilde{\theta}-\hat{\theta})'sB_nB_n'(\tilde{\theta}-\hat{\theta}))+ \tag{2.11}$$

$$\exp((\tilde{\theta}-\hat{\theta})'(sB_nB_n')(A_n+2sB_nB_n'+C)^{-1}(sB_nB_n')(\tilde{\theta}-\hat{\theta})/2) \tag{2.12}$$

and

$$\frac{\sqrt{\det(\sqrt{C}^{-1}A_n\sqrt{C}^{-1}+I)}}{\sqrt{\det(\sqrt{C}^{-1}A_n\sqrt{C}^{-1}+I+\sqrt{C}^{-1}B_n\sqrt{C}^{-1})}}. \tag{2.13}$$

The second factor does not depend on $(\tilde{\theta}-\hat{\theta})$, and hence represents $E_n \exp(-sQ_n(\hat{\theta}))$. The first one is equal to

$$\exp(-(\tilde{\theta}-\hat{\theta})'H_n(\tilde{\theta}-\hat{\theta})),$$

where

$$H_n = sB_nB_n' - (sB_nB_n')(A_n + 2sB_nB_n' + C)^{-1}(sB_nB_n')/2,$$

which is positive definite. Hence the expectation is smaller than 1 if $\tilde{\theta}-\hat{\theta}$ are different. The first factor depends on B_n, our norming of B_n guarantees that H_n does not vanish asymptotically.

2.3 Conclusing Remarks

The theorems here establish some asymptotic optimality properties of a "shrunken" ML-estimator. The restrictions on the parameter are quite severe. We have to assume that, if the parameter is interpreted as a function, this function has to be differentiable 8 times. We believe that future research will make it possible to relax this rather stringent requirement.

Another promising line of research are empirical Bayesian methods. Assuming the c_k e.g. to be of the form $An^{-\gamma}$. Theoretically, with an astronomical amount of data, one should be able to estimate A and γ consistently. Obviously this is not realistic for many applications. Nevertheless, this raises the questions what kind of inference on the hyper-parameters is possible.

Another venue are priors where the c_k decay exponentially. Preliminary results indicate that the resulting estimators are sieve estimators with sieve length proportional to the loga-

rithm of time. More interestingly, this optimal length of the sieve would also be achieved by using AIC or BIC. Hence this theory could be used to derive optimality results for standard techniques of inference. Moreover, doing so also highlights the strong and possibly weak points of these techniques.

Bibliography

[1] Andrews, D.W.K. (1991): Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation, *Econometrica*, 59, 817-858.

[2] Bernstein, S.N.(1928): Sur les fonctions absolument monotones, *Acta Mathematica* 52, 1-66

[3] Armstrong, T.B. and M. Kolesar (2018). Optimal inference in a class of regression models. *Econometrica*

[4] Brown, L.D. and M. Low (1996). Asymptotic equivalence of non-parametric regression and white noise. *Annals of Statistics* 24, p.2384-98.

[5] Gine, E, and R. Nickl (2016). *Mathematical Foundations of Infinite Dimensional Statistical Models*, Cambridge University Press

[6] Golubev, G.K., M. Nussbaum and H.H. Zhou (2010). Asymptotic equivalence of spectral density estimation and Gaussian white noise. *Annals of statistics* 38, p. 181-214.

[7] Grama, I. and M. Neumann (2006). Asymptotic equivalence of nonparametric autoregression and non-parametric regression. *Annals of Statistics* 34, p. 1701-1732.

[8] Kim, J.Y. (1998): Large Sample Properties of Posterior Densities, Bayesian Information Criterion and the Likelihood Principle in Nonstationary Time Series Models, *Econometrica*, 66, 359-380

[9] Lang, S.(1993): Real and Functional Analysis, Springer

[10] Nussbaum, M (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Annals of Statistics* 24, p. 2399-430.

[11] Ploberger, W and P.C.B. Phillips.(2012) Optimal Estimation under Nonstandard Conditions. *Journal of Econometrics* 169, p. 258-265

[12] Reiss M. (2008). Asymptotic equivalence for nonparametric regression with multivariate and random design *Annals of Statistics*. 36, p. 1957-82.

[13] Strasser, H.(1985). *Mathematical Theory of Statistics*, de Gruyter

[14] Van der Vaardt, A.W.(2000). *Asymptotic Statistics*, Cambridge University Press.

Chapter 3

Optimal Model Dimension through the AIC

Coauthored with *Werner Ploberger*

3.1 Introduction

In the case of finite-dimensional parameters, the theory of optimal estimation is already well established and the theory is well presented in textbooks like van der Vaardt (2000) and Strasser (1996). However, there is no comprehensive theory of optimal estimators in the case of infinite number of estimators. One of the classical optimality results is that of Shibata (1973), which shows under some mild regularity conditions, the AIC criterion (Akaike, 1973) chooses the best estimator among sieve estimators.

In this paper, we show the equivalence between a class of Bayesian estimators and the AIC selection of sieve estimators. We make very strong assumptions on the class of prior distributions. Suppose the parameters are drawn from such a prior distribution, the best estimator is then simply the estimator that minimizes posterior risks given the data. There-

fore when such priors are appropriate, through the equivalence results, the AIC estimator is asymptotically optimal among all other estimators. Many papers in the literature have justified certain information criterion based on a Bayesian rationale. Such as the Bayesian Information Criterion (Schwarz, 1978) and the Posterior information criterion (Phillips and Ploberger, 1994). Our paper differs from these literature because our prior distribution is not dismissible asymptotically.

Consider the linear model of the following form

$$y_{(t)} = x_{(t)}\prime\beta + u_{(t)},^{1}$$

where the data is generated in the following way. There is a constant $\lambda \in (0,1)$ such that $\beta_{(i)} \sim \mathcal{N}(0, \lambda^i)$ independent over i, and $\beta_{(i)}$ does not vary with t. The random error term $u_{(t)}$ is iid $\mathcal{N}(0,1)$. We assume that there are a number of regressors $x_{(t)} = \{x_{(t,n)}\}_{n=1,...}$ and they are taken as given. Let the X be an $T \times n$ matrix with orthogonal columns, and each column has Euclidean norm $\sqrt{T}$. The number of regressors n is diverging in T that $n = O(\sqrt{T})$. We call the dth model the one that include all the first d regressors X_d. Our main result shows that asymptotically, the AIC optimal model choose the $-\ln T/\ln\lambda$-th model. If n grows slower than $-\ln T/\ln\lambda$, our result implies the largest model is the best model.

The assumptions on the design matrix can be naturally satisfied in several applications. For example, in estimating non-linear function using higher order polynomials, one can apply the Gram-Schmidt process to X and obtain a set of orthogonal regressors of polynomials in increasing degree. As an other example, in factor models, usually the factors are estimated as principal components, which are orthogonal regressors. The principal components are also ordered naturally according to their variances.

The family of prior distributions is the following data generating process.

[1]We use subscript parathesis $X_{(t)}$ to indicate the tth item in the vector. If we use a subscript X_t without parenthesis, it means the sub-vector consisting of the 1st to tth item.

3.2 AIC and the OLS regression problem

We use the simple OLS estimator, when we select the first d regressors, the estimator is

$$(X_d\prime X_d)^{-1}X_d\prime Y.$$

and the AIC for d regressors is defined to be $\frac{T}{2}\ln(\hat{\sigma}^2(d)) + d$ where $\hat{\sigma}^2(d)$ is simply the MLE estimator of the error when d regressors are included into the regression. Hence

$$\begin{aligned}\hat{\sigma}^2(d) =& \frac{1}{T}(Y - X_d(X_d\prime X_d)^{-1}X_d\prime Y)\prime(Y - X_d(X_d\prime X_d)^{-1}X_d\prime Y)\\ =& \frac{1}{T}\left((I - Proj_d)(X_n\beta_n + u)\right)\prime\left((I - Proj_d)(X_n\beta_n + u)\right)\\ =& \frac{1}{T}(\beta_n\prime X_n\prime Proj_d^{\perp}X_n\beta_n + u\prime Proj_d^{\perp}u - 2u\prime Proj_d^{\perp}X_n\beta_n)\end{aligned}$$

Where $Proj_d := X_d(X_d\prime X_d)^{-1}X_d$ and $Proj_d^{\perp} := I - Proj_d$. It follows from the orthogonality of X that

$$X_n\prime Proj_d^{\perp}X_n = T\begin{bmatrix}\mathbf{0}_d & * \\ * & \mathbf{1}_{n-d}\end{bmatrix}$$

where the $\mathbf{0}_d$ represents a $d \times d$ zero matrix and $\mathbf{1}_{n-d}$ is an $(n-d)\times(n-d)$ identity matrix. Hence

$$\hat{\sigma}^2(d) = \sum_{i=d+1}^{n}\beta_{(i)}^2 + \frac{1}{T}\sum_{i=d+1}^{T}u_{(i)}^2 - \frac{2}{T}u\prime Proj_d^{\perp}X_n\beta_n.^2$$

We observe that

[2] Since u is a standard normal vector, we can specify the coordinate in whichever way we want, hence the subscript (i) picks the dimensions that is exactly in the basis of $Proj_d^{\perp}$ and should not be confused with the subscript of the β. Hence for example, $au\prime Proj_d^{\perp}u - u\prime Proj_{d+C}^{\perp}u$ is the sum of $n-d$ independent $\chi^2(1)$ random variables scaled by $(a-1)$ plus another C independent $\chi^2(1)$ random variables.

- since $\beta_{(i)}$ is normal of variance λ^i, $\beta_{(i)}^2$ is a χ_1^2 distribution scaled by a factor λ^i, therefore it has mean λ^i and variance $2\lambda^{2i}$.

- Moreover, each $u_{(i)}^2$ is a χ_1^2 random variable hand has expectation 1 variance 2.

- Denote $\frac{2}{T}u\prime Proj_d^{\perp}X_n\beta_n$ by $\epsilon(d)$. It has expectation 0, whether one takes β_n fixed or not. Hence its variance is simply

$$\frac{4}{T^2}\boldsymbol{E}\left[\beta_n\prime X_n\prime Proj_d^{\perp}uu\prime Proj_d^{\perp}X_n\beta_n\right].$$

  When we take β_n as given and take expectation over u, we get the conditional variance given β_n. When we take expectation over β and u we get the unconditional variance. They are respectively

$$\frac{4}{T}\sum_{i=d+1}^{n}\beta_{(i)}^2 \text{ and } \frac{4}{T}\sum_{i=d+1}^{n}\lambda^i = \frac{4}{T}\frac{1-\lambda^{n-d}}{1-\lambda}\lambda^{d+1}.$$

3.2.1 Comparing asymptotic $AIC(d)$ and $AIC(d+C)$ when $d := -\ln T/\ln\lambda$

We want to give a bound on the probability that for a fixed large C, the probability that AIC is minized at $d+C$ as $T\to\infty$. Such probability is bounded above by $\lim_{T\to\infty}\Pr(AIC(d)\geq AIC(d+C))$. It is easy to see that $AIC(d)\geq AIC(d+C)$ for some constant C if and only if

$$\hat{\sigma}^2(d)\geq e^{2C/T}\hat{\sigma}^2(d+C),$$

In other words

$$\sum_{i=d+1}^{n}\beta_{(i)}^2+\frac{1}{T}\sum_{i=d+1}^{T}u_{(i)}^2-\epsilon(d)\geq e^{2C/T}\left(\sum_{i=d+C+1}^{n}\beta_{(i)}^2+\frac{1}{T}\sum_{i=d+C+1}^{T}u_{(i)}^2-\epsilon(d+C)\right).$$

We rewrite it with a normalization T/C on both hand sides and get

$$\begin{aligned}
&\frac{T}{C}\sum_{i=d+1}^{d+C}\beta_{(i)}^2 \\
\geq&\frac{T}{C}(e^{2C/T}-1)\left(\sum_{i=d+C+1}^{n}\beta_{(i)}^2+\frac{1}{T}\sum_{i=d+C+1}^{T}u_{(i)}^2\right)-\frac{1}{C}\sum_{i=d+1}^{d+C}u_{(i)}^2 \\
&-\frac{T}{C}e^{2C/T}\epsilon(d+C)+\frac{T}{C}\epsilon(d)
\end{aligned}$$

The expectation of LHS is

$$\frac{T}{C}\sum_{i=d+1}^{d+C}\lambda^i=\frac{T}{C}\lambda^{d+1}\frac{1-\lambda^C}{1-\lambda}=\frac{\lambda}{C}\frac{1-\lambda^C}{1-\lambda}$$

when we plug in $d:=-\ln T/\ln\lambda$. The variance of LHS is

$$\frac{T^2}{C^2}\sum_{i=d+1}^{d+C}2\lambda^{2i}=2\frac{T^2}{C^2}\lambda^{2d+2}\frac{1-\lambda^{2C}}{1-\lambda^2}=2\left(\frac{\lambda}{C}\right)^2\frac{1-\lambda^{2C}}{1-\lambda^2}$$

when we plug in $d:=-\ln T/\ln\lambda$. Hence the LHS is of order $O(\frac{1}{C})$.

On the other hand, the RHS has expectation

$$\begin{aligned}
&\frac{e^{2C/T}-1}{C}\left(T\sum_{i=d+C+1}^{n}\lambda^i+(T-d-C)\right)-1 \\
=&\frac{e^{2C/T}-1}{C}\left(T\lambda^{d+C+1}\frac{1-\lambda^{n-d-C}}{1-\lambda}+T-d-C\right)-1 \\
\to&\frac{2}{T}\left(\lambda^{C+1}\frac{1-\lambda^{n-d-C}}{1-\lambda}+T-d-C\right)-1\to 1
\end{aligned}$$

by first plug in $d:=-\ln T/\ln\lambda$ and take $T\to\infty$.

The variance of RHS is bounded by the following term multiplied by 2 (to take care of

the covariances)

$$\begin{aligned}
&\left(\frac{e^{2C/T}-1}{C}\right)^2\left(T^2 2\sum_{i=d+C+1}^{n}\lambda^{2i}+\sum_{i=d+C+1}^{T}2\right)+\sum_{i=d+1}^{d+C}\frac{2}{C}\\
&+\left(\frac{T}{C}\right)^2 e^{4C/T}\boldsymbol{V}ar[\epsilon(d+C)]+\left(\frac{T}{C}\right)^2\boldsymbol{V}ar[\epsilon(d)]\\
\to&\frac{4}{T^2}\left(2\lambda^{2C+2}\frac{1-\lambda^{2(n-d-C)}}{1-\lambda}+2(T-d-C)\right)+\frac{2}{C}\\
&+\left(\frac{T}{C}\right)^2 e^{4C/T}\frac{4}{T}\frac{1-\lambda^{n-d-C}}{1-\lambda}\lambda^{d+C+1}+\left(\frac{T}{C}\right)^2\frac{4}{T}\frac{1-\lambda^{n-d}}{1-\lambda}\lambda^{d+1}\\
\to&\frac{2}{C}+\frac{4}{C^2}\frac{1-\lambda^{n-d-C}}{1-\lambda}\lambda^{C+1}+\frac{4}{C^2}\frac{1-\lambda^{n-d}}{1-\lambda}\lambda
\end{aligned}$$

by first plug in $d := -\ln T/\ln\lambda$ and take $T\to\infty$.

It can be seen that LHS is of order $O(\frac{1}{C})$. The first two terms in the RHS equals $2-\frac{\chi^2(C)}{C}$ which is a Chi-square C degree of freedom variable multiplied by a factor of $-1/C$ and then translated two units to the right. The last two terms is of order $O(\frac{1}{C})$. Moreover, it is clear that the above limits are uniform for all $C\in[0,n]$ as long as $n/T\to 0$.

Therefore, for any large enough $C\le n$, the probability that LHS $\ge$ RHS is approximately

$$\begin{aligned}
\Pr(0\ge 2-1/C\chi^2(C)) =&\Pr(\chi^2(C)>2C)\\
=&\int_{2C}^{\infty}\frac{1}{2^{C/2}\Gamma(C/2)}x^{C/2-1}e^{-x/2}\boldsymbol{d}x\\
=&\frac{\Gamma(C/2,C)}{\Gamma(C/2)}\\
\le&\frac{\lceil C/2-1\rceil!}{\lfloor C/2-1\rfloor!}e^{-C}\sum_{k=0}^{\lceil C/2-1\rceil}\frac{C^k}{k!}\\
\le&2e^{-C}\sum_{k=0}^{\lceil C/2-1\rceil}\frac{C^k}{k!}\le 2e^{-C}\frac{C/2}{\sqrt{\pi C}}\left(\frac{e}{2}\right)^{C/2}=\sqrt{\frac{C}{\pi}}(2e)^{-C/2},
\end{aligned}$$

where we used properties of the incomplete Gamma function[3] and Stirling's approxi-

[3] Weisstein, Eric W., "Incomplete Gamma Function", *MathWorld*. (equation 2)

mation. Hence we conclude that as $T \to \infty$ the probability that $d + C$ minimizes AIC is bounded by $\sqrt{\frac{C}{\pi}}(2e)^{-C/2}$ for all large C.

3.2.2 Comparing asymptotic $AIC(d)$ and $AIC(d - C)$ when $d := -\ln T / \ln \lambda$

On the other hand, we want to give a bound on the probability that for a fixed large C, the probability that AIC is minized at $d - C$ as $T \to \infty$. Such probability is bounded above by $\lim_{T\to\infty} \Pr(AIC(d) \geq AIC(d - C))$.

$AIC(d) \geq AIC(d - C)$ if and only if $e^{2C/T}\hat{\sigma}^2(d) \geq \hat{\sigma}^2(d - C)$. Apply the scaling $\lambda^{C-1}T$ to both hand sides, we rewrite the inequality in the following way:

$$\begin{aligned}
&\lambda^{C-1}\left[T(e^{2C/T}-1)\left(\sum_{i=d+1}^{n}\beta_{(i)}^2+\frac{1}{T}\sum_{i=d+1}^{T}u_{(i)}^2\right)-\sum_{i=d-C+1}^{d}u_{(i)}^2\right]\\
&-\lambda^{C-1}Te^{2C/T}\epsilon(d)+\lambda^{C-1}T\epsilon(d-C)\\
\geq&\lambda^{C-1}T\sum_{i=d-C+1}^{d}\beta_{(i)}^2
\end{aligned}$$

Expectation of LHS is

$$\begin{aligned}
&\lambda^{C-1}\left[T(e^{2C/T}-1)\left(\sum_{i=d+1}^{n}\lambda^i+\frac{1}{T}(T-d)\right)-C\right]\\
=&\lambda^{C-1}\left[2C\left(\lambda^{d+1}\frac{1-\lambda^{n-d}}{1-\lambda}+\frac{T-d}{T}\right)-C\right]\\
\to&\lambda^{C-1}C
\end{aligned}$$

when we take $d = -\ln T / \ln \lambda$ and then take $T \to \infty$. The variance of LHS is bounded by

two times the following:

$$\begin{aligned}
&\lambda^{2C-2}\left[(T(e^{2C/T}-1))^2\left(\sum_{i=d+1}^{n}2\lambda^{2i}+\frac{1}{T^2}2(T-d)\right)+2C\right]\\
&+\lambda^{2C-2}\left[(T(e^{2C/T}))^2\frac{4}{T}\lambda^{d+1}\frac{1-\lambda^{n-d}}{1-\lambda}+T^2\frac{4}{T}\lambda^{d-C+1}\frac{1-\lambda^{n-d+C}}{1-\lambda}\right]\\
\to&\lambda^{2C-2}4C^2\left(2\frac{\lambda^2}{T^2}\frac{1-\lambda^{2n-2d}}{1-\lambda}+\frac{2(T-d)}{T^2}\right)+4\lambda^{2C-1}\frac{1-\lambda^{2n-2d}}{1-\lambda}+4\lambda^{C-1}\frac{1-\lambda^{n-d+C}}{1-\lambda}\\
\to&4\lambda^{2C-1}\frac{1}{1-\lambda}+4\lambda^{C-1}\frac{1}{1-\lambda}
\end{aligned}$$

when we take $d=-\ln T/\ln\lambda$ and then take $T\to\infty$. Hence the LHS is of order $O(\lambda^{C/2})$

Now consider RHS, it can be seen that

$$RHS=\lambda^{C-1}T\sum_{i=d-C+1}^{d}\beta_{(i)}^2>\lambda^{C-1}T\beta_{(d-C+1)}^2\sim\chi^2(1)$$

after taking $d=-\ln T/\ln\lambda$.

Hence it can be seen that for any fixed large C, the probability that LHS$\geq$RHS is bounded above by the probability that LHS$\geq\chi^2(1)$. This is approximately

$$\Pr(\lambda^{C/2}>\chi^2(1))=\int_0^{\lambda^{C/2}}\frac{1}{\sqrt{2}\Gamma(1/2)}x^{-1/2}e^{-x/2}\boldsymbol{d}x\leq\int_0^{\lambda^{C/2}}x^{-1/2}\boldsymbol{d}x=\lambda^{C/4}$$

for all large C. Hence we conclude that as $T\to\infty$ the probability that $d-C$ minimizes AIC is bounded by $\lambda^{C/4}$ for all large C.

3.3 The Bayesian problem

Our Bayesian problem can be formulated in a slightly more general case of the infinite dimensional space. However since there is no density function available in infinite dimensional situations, finding the posterior measure in infinite dimensional space requires Radon-Nikodym

derivative. A general treatment of the subject can be found in Stuart (2010).

The standard result gives that the posterior mean vector of β is

$$\hat{\beta} := \Sigma^{1/2}\left(\Sigma^{1/2}X\prime X\Sigma^{1/2}+I\right)^{-1}\Sigma^{1/2}X\prime(X\beta+u)$$

Let $Q := \Sigma^{1/2}\left(\Sigma^{1/2}X\prime X\Sigma^{1/2}+I\right)^{-1}\Sigma^{1/2}$, then $\hat{\beta} = QX\prime(X\beta+u)$.

and the posterior variance covariance

$$\boldsymbol{E}_{posterior}[(\beta-\hat{\beta})\prime(\beta-\hat{\beta})] = \Sigma^{1/2}(\Sigma^{1/2}X\prime X\Sigma^{1/2}+I)^{-1}\Sigma^{1/2}.$$

Since $X\prime X = TI$, $\hat{\beta}_{(i)} = \frac{T\lambda^i}{T\lambda^i+1}\beta_{(i)} + \frac{\lambda^i}{T\lambda^i+1}X_{(i)}\prime u$. For any given i, the second term goes to 0 as $T\to\infty$ since $X_{(i)}$ and u are not dependent. On the other hand $\frac{T\lambda^i}{T\lambda^i+1}$ is approximately 1 for large T and small i, and approximately 0 for large i. It can be easily check that the first term is approximately $\beta_{(i)}$ for the first $-\ln T/\ln\lambda - C$ coordinates, and they are approximately 0 for $i \geq -\ln T/\ln\lambda + C$ for some C depends only on λ. This shows that the AIC from the previous section would select approximately the same number of regressors asymptotically.

3.4 Asymptotic equivalence

3.4.1 l^2 equivalence

Let $\tilde{\beta}$ be the AIC estimate and $\hat{\beta}$ be the bayesian estimate. Then we have the following two theorems.

Theorem 11. *Under our assumptions, we have*

$$\boldsymbol{E}_u||\tilde{\beta}-\hat{\beta}||_2^2 = o\left(\boldsymbol{E}_u||\beta-\hat{\beta}||_2^2\right)$$

e.g. the difference between the estimators is a magnitude smaller than the estimation error.

For some d^* is optimally chosen by AIC between $1, 2, \ldots, n$, it is readily seen that

$$\tilde{\beta} = (X_{d^*}\prime X_{d^*})^{-1} X_{d^*}\prime (X\beta + u) \text{ and } \hat{\beta} = QX\prime(X\beta + u)$$

Notice that the two estimates is of different dimensions, $\tilde{\beta}$ has d^* non-trivial dimensions and we would fill the remaining dimensions with 0. Notice that by definition, Q is a diagonal matrix. Hence we can write Q_{d^*} be the top left $d^* \times d^*$ square submatrix and Q_{d^*+} be the $(n - d^*) \times (n - d^*)$ be the submatrix at the bottom right cornor. Hence

$$\hat{\beta}_{d^*} = Q_{d^*} X\prime_{d^*}(X\beta + u) \text{ and } \hat{\beta}_{d^*+} = Q_{d^*+} X\prime_{d^*+}(X\beta + u).$$

And therefore, $||\tilde{\beta} - \hat{\beta}||_2^2 = ||\tilde{\beta} - \hat{\beta}_{d^*}||_2^2 + ||\hat{\beta}_{d^*+}||_2^2$.

Proof. We can expand the expression and get

$$\begin{aligned}
||\tilde{\beta} - \hat{\beta}||_2^2 =& ||\tilde{\beta} - \hat{\beta}_{d^*}||_2^2 + ||\hat{\beta}_{d^*+}||_2^2 \\
=& ||(X_{d^*}\prime X_{d^*})^{-1} X_{d^*}\prime (X\beta + u) - Q_{d^*} X\prime_{d^*}(X\beta + u)||_2^2 + ||Q_{d^*+} X\prime_{d^*+}(X\beta + u)||_2^2 \\
\leq& ||((X_{d^*}\prime X_{d^*})^{-1} - Q_{d^*}) X_{d^*}\prime X\beta||_2^2 + ||Q_{d^*+} X\prime_{d^*+} X\beta||_2^2 \\
&+ ||((X_{d^*}\prime X_{d^*})^{-1} - Q_{d^*}) X_{d^*}\prime u||_2^2 + ||Q_{d^*+} X\prime_{d^*+} u||_2^2 \\
=& \sum_{i=1}^{d^*} \left(\frac{\beta_{(i)}}{1 + T\lambda^i} \right)^2 + \sum_{i=d^*+1}^{n} \left(\frac{T\lambda^i \beta_{(i)}}{1 + T\lambda^i} \right)^2 \\
&+ ||((X_{d^*}\prime X_{d^*})^{-1} - Q_{d^*}) X_{d^*}\prime u||_2^2 + ||Q_{d^*+} X\prime_{d^*+} u||_2^2
\end{aligned}$$

The third and the fourth term can be separated into the norm contributed from the first

d^* terms in u and the remaining terms. i.e.

$$\begin{aligned}&||((X_{d^*}\prime X_{d^*})^{-1} - Q_{d^*})X_{d^*}\prime u||_2^2 + ||Q_{d^*+}X\prime_{d^*+}u||_2^2\\ =&u\prime X_{d^*}\left((X_{d^*}\prime X_{d^*})^{-1} - Q_{d^*}\right)^2 X_{d^*}\prime u + u\prime X_{d^*+}Q_{d^*+}^2 X_{d^*+}\prime u\end{aligned}$$

Take any $C = \ln\ln d$ for $d := -\ln T/\ln\lambda$, by our previous analysis, we have that $d - C < d^* < d + C$ as $T \to \infty$. Hence for T large enough, we can bound the above expression by

$$\begin{aligned}&||((X_{d^*}\prime X_{d^*})^{-1} - Q_{d^*})X_{d^*}\prime u||_2^2 + ||Q_{d^*+}X\prime_{d^*+}u||_2^2\\ <&u\prime X_{d+C}\left((X_{d+C}\prime X_{d+C})^{-1} - Q_{d+C}\right)^2 X_{d+C}\prime u + u\prime X_{(d-C)+}Q_{(d-C)+}^2 X_{(d-C)+}\prime u\end{aligned}$$

Taking expectation over u the above expression can be expressed in terms of trace, i.e. from $\mathbb{E}[uu\prime] = I$ we have

$$\begin{aligned}&=tr\left(\left((X_{d+C}\prime X_{d+C})^{-1} - Q_{d+C}\right)^2 X_{d+C}\prime\mathbb{E}[uu\prime]X_{d+C}\right) + tr\left(Q_{(d-C)+}^2 X_{(d-C)+}\prime\mathbb{E}[uu\prime]X_{(d-C)+}\right)\\ &=\sum_{i=1}^{d+C}\left(\frac{1}{T} - \frac{\lambda^i}{1+\lambda^i T}\right)^2 T + \sum_{i=d-C+1}^{T}\frac{T\lambda^{2i}}{(1+\lambda^i T)^2}\\ &\le\sum_{i=1}^{d+C}\frac{1}{T(T\lambda^i+1)^2} + \sum_{i=d-C+1}^{T}\lambda^{2i}T\\ &\le\sum_{i=1}^{d+C}\frac{\lambda^{-i}}{T^2} + T\lambda^{2d-2C+2}\frac{1-\lambda^{T-d+C}}{1-\lambda}\\ &=\lambda^{-1}\frac{1}{T^2}\frac{\lambda^{-d}\lambda^{-C}-1}{\lambda^{-1}-1} + T\lambda^{2d}\lambda^{-2C}\lambda^2\frac{1-\lambda^{T-d+C}}{1-\lambda}\end{aligned}$$

Since $\lambda^d = 1/T$ and $\lambda^C = (\ln d)^{\ln\lambda}$, the above expression becomes

$$\frac{\lambda^{-1}}{\lambda^{-1}-1}\frac{(\ln d)^{-\ln\lambda}}{T} - \frac{\lambda^{-1}}{\lambda^{-1}-1}\frac{1}{T^2} + \frac{(\ln d)^{-2\ln\lambda}}{T}\frac{1-\lambda^{T-d+C}}{1-\lambda} = O\left(\frac{(\ln d)^{-2\ln\lambda}}{T}\right)$$

Therefore,

$$\begin{aligned}
\mathbb{E}_u||\tilde{\beta}-\hat{\beta}||_2^2 =& \sum_{i=1}^{d^*}\left(\frac{\beta_{(i)}}{1+T\lambda^i}\right)^2+\sum_{i=d^*+1}^{n}\left(\frac{T\lambda^i\beta_{(i)}}{1+T\lambda^i}\right)^2 \\
&+\mathbb{E}_u||((X_{d^*}\prime X_{d^*})^{-1}-Q_{d^*})X_{d^*}\prime u||_2^2+\mathbb{E}_u||Q_{d^*+}X\prime_{d^*+}u||_2^2 \\
\leq& \sum_{i=1}^{d+C}\left(\frac{\beta_{(i)}}{1+T\lambda^i}\right)^2+\sum_{i=d-C}^{n}\left(\frac{T\lambda^i\beta_{(i)}}{1+T\lambda^i}\right)^2+O\left(\frac{(\ln d)^{-2\ln\lambda}}{T}\right) \\
\leq& \sum_{i=1}^{d+C}\left(\frac{\beta_{(i)}}{T\lambda^i}\right)^2+\sum_{i=d-C+1}^{n}\left(T\lambda^i\beta_{(i)}\right)^2+O\left(\frac{(\ln d)^{-2\ln\lambda}}{T}\right)
\end{aligned}$$

for large enough T. The first term has mean $\frac{1}{T^2}\lambda^{-1}\frac{\lambda^{-d-C+1}-1}{\lambda^{-1}-1}=O\big(\frac{(\ln d)^{-\ln\lambda}}{T}\big)$ and variance $\frac{2}{T^4}\lambda^{-2}\frac{\lambda^{-2d-2C}-1}{\lambda^{-2}-1}=O\big(\frac{(\ln d)^{-2\ln\lambda}}{T^2}\big)$, hence the first term is of order $O\big(\frac{(\ln d)^{-\ln\lambda}}{T}\big)$. The second term has mean $T^2\lambda^{3d}\lambda^{-3C}\lambda^3\frac{1-\lambda^{3(n-d+C)}}{1-\lambda^3}=O\big(\frac{(\ln d)^{-3\ln\lambda}}{T}\big)$ and variance $2T^4\lambda^{6d}\lambda^{-6C}\lambda^6\frac{1-\lambda^{6(n-d+C)}}{1-\lambda^6}=O\big(\frac{(\ln d)^{-6\ln\lambda}}{T^2}\big)$, hence the second term is of order $O\big(\frac{(\ln d)^{-3\ln\lambda}}{T}\big)$. Therefore we conclude that

$$\mathbb{E}_u||\tilde{\beta}-\hat{\beta}||_2^2\leq O\left(\frac{(\ln d)^{-\ln\lambda}}{T}\right)+O\left(\frac{(\ln d)^{-3\ln\lambda}}{T}\right)+O\left(\frac{(\ln d)^{-2\ln\lambda}}{T}\right)=O\left(\frac{(\ln d)^{-3\ln\lambda}}{T}\right).$$

On the other hand, we can get a Chi-square lower bound by comparing the first d^* terms in the true parameter β and AIC estimate $\tilde{\beta}$.

$$\begin{aligned}
||\tilde{\beta}-\beta||_2^2 \geq& ||(X_{d^*}\prime X_{d^*})^{-1}X_{d^*}\prime(X\beta+u)-\beta_{d^*}||_2^2 \\
=& u\prime X_{d^*}(X_{d^*}\prime X_{d^*})^{-1}(X_{d^*}\prime X_{d^*})^{-1}X_{d^*}\prime u \\
\geq& u\prime X_{d-C}(X_{d-C}\prime X_{d-C})^{-1}(X_{d-C}\prime X_{d-C})^{-1}X_{d-C}\prime u,
\end{aligned}$$

hence the lower bound follows some scaled Chi-square distribution of $d-C$ degree of freedom.

Taking expectation over u we have

$$\begin{aligned}\mathbb{E}_u||\tilde{\beta}-\beta||_2^2 \geq & \mathbb{E}_u[u\prime X_{d-C}(X_{d-C}\prime X_{d-C})^{-1}(X_{d-C}\prime X_{d-C})^{-1}X_{d-C}\prime u] \\ = & tr((X_{d-C}\prime X_{d-C})^{-1}X_{d-C}\prime \mathbb{E}_u[uu\prime]X_{d-C}(X_{d-C}\prime X_{d-C})^{-1}) \\ = & \frac{d-C}{T}\end{aligned}$$

Therefore

$$\mathbb{E}_u||\tilde{\beta}-\hat{\beta}||_2^2 = o(\mathbb{E}_u||\tilde{\beta}-\beta||_2^2)$$

as $T \to \infty$. We have therefore shown that $\hat{\beta}$ and $\tilde{\beta}$ are asymptotically equivalent under l^2 norm. □

3.4.2 Equivalence under linear projections

Not only the global distance between the two estimators is smaller than the estimation error, this is also true for many of the linear projections of the estimator. For all vectors B, $B\prime(\beta-\hat{\beta}))$ is normal. We show that for all vectors B satisfying some restrictions $B\prime(\tilde{\beta}-\hat{\beta})$ is of smaller order than the standard deviation of $B\prime(\beta-\hat{\beta}))$. For this to hold, we need to require that the components of $B=(b_{(i)})_{i=1}^{\infty}$ are all of the same order of magnitude.

Definition 12. *We say the partial sum of a sequence $S_n := \sum_{i=1}^{n} b_i^2$ is of* slow growth *if for any constant C*

$$\lim_{n\to\infty}\frac{S_{n+C}}{S_n}=1.^4$$

Theorem 13. *If $B=(b_{(i)})_{i=1}^{\infty}$ whose squared partial sum is of slow growth, then $B\prime(\tilde{\beta}-\hat{\beta})$ is of smaller order than the standard deviation of $B\prime(\beta-\hat{\beta})$.*

[4] Kapoor and Nautiyal (1981) studied classes of functions of various speeds of growth, our definition of slow growth here satisfies the more general hypothesis (H, ii) in their paper, but not necessarily the more restrictive ones $(H, iii)-(H, v)$.

Recall that we have

$$\tilde{\beta}-\hat{\beta}=\begin{bmatrix} \cdots \\ \frac{1}{1+\lambda^i T}\beta_{(i)}+\frac{1}{1+\lambda^i T}\frac{1}{T}(X\prime u)_{(i)} \\ \cdots \\ -\frac{T\lambda^j}{1+\lambda^j T}\beta_{(j)}-\frac{\lambda^j}{1+\lambda^j T}(X\prime u)_{(j)} \\ \cdots \end{bmatrix} \text{ and } \beta-\hat{\beta}=\begin{bmatrix} \cdots \\ \frac{1}{1+\lambda^k T}\beta_{(k)}-\frac{\lambda^k}{1+\lambda^k T}(X\prime u)_{(i)} \\ \cdots \end{bmatrix}$$

where $1 \le i \le d^* < j \le n$ and $1 \le k \le n$.

When $B\prime = (b_1, b_2, \ldots)$ is just a row vector, consider $B\prime(\beta - \hat{\beta})$, it follows a mean zero normal distribution with variance

$$\begin{aligned}\sum_{i=1}^{n}\left(\left(\frac{b_i}{1+\lambda^i T}\right)^2\lambda^i+\left(\frac{\lambda^i b_i}{1+\lambda^i T}\right)^2 T\right) \ge & \sum_{i=1}^{d}\left(\frac{\lambda^i b_i}{1+\lambda^i T}\right)^2 T \\ \ge & O\left(\frac{1}{T}\right)\sum_{i=1}^{d} b_i^2\end{aligned}$$

hence $B\prime(\beta - \hat{\beta})$ is of order greater or equal to $O(\frac{1}{\sqrt{T}})\sqrt{\sum_{i=1}^{d} b_i^2}$ for $d := -\ln T/\ln\lambda$.

To prove the theorem, we will show that $B\prime(\tilde{\beta}-\hat{\beta}) = o\left(\sqrt{\frac{\sum_{i=1}^{d} b_i^2}{T}}\right)$ under the assumption that $\sum_{i=1}^{n} b_i^2$ is of slow growth. Before we present the proof, we first prepare the following lemma.

Lemma 14. *Suppose $S_d := \sum_i^d b_i^2$ is of slow growth, then for any $\lambda \in (0,1)$, and any constant C, and n such that $\lim_{d\to\infty} d/n \to 0$, the following limits hold as $d \to \infty$:*

$$\frac{\sum_{i=1}^{d+C} b_i^2\lambda^{d+C-i}}{\sum_{i=1}^{d} b_i^2} \to 0; \text{ and } \frac{\sum_{j=1}^{n-d+C} b_{j+d-C}^2\lambda^j}{\sum_{i=1}^{d} b_i^2} \to 0.$$

Proof. To establish the first limit, observe that for any $\lambda \in (0,1)$

$$\lambda \times \lambda^{k-i} = \lambda^{k+1} + (1-\lambda) \sum_{j=k-i+1}^{k} \lambda^j.$$

Hence

$$\begin{aligned}
\frac{\sum_{i=1}^{d+C} b_i^2 \lambda^{d+C-i}}{\sum_{i=1}^{d} b_i^2} =& \frac{\sum_{i=1}^{d+C} b_i^2 \lambda^{d+C+1}}{\lambda S_d} + \frac{(1-\lambda)\sum_{i=1}^{d+C} b_i^2 \sum_{j=d+C-i+1}^{d+C} \lambda^j}{\lambda S_d} \\
=& \frac{S_{d+C}\lambda^{d+C+1}}{\lambda S_d} + \frac{(1-\lambda)\sum_{j=1}^{d+C}\sum_{i=d+C-j+1}^{d+C} b_i^2 \lambda^j}{\lambda S_d}
\end{aligned}$$

The first term goes to 0 as $d \to \infty$. The second term can be decomposed into two parts

$$\begin{aligned}
& \frac{\sum_{j=1}^{d+C}\sum_{i=d+C-j+1}^{d+C} b_i^2 \lambda^j}{S_d} \\
=& \frac{\sum_{j=1}^{K}\sum_{i=d+C-j+1}^{d+C} b_i^2 \lambda^j}{S_d} + \frac{\sum_{j=K+1}^{d+C}\sum_{i=d+C-j+1}^{d+C} b_i^2 \lambda^j}{S_d} \\
\leq& \frac{\sum_{j=1}^{K} \lambda^j (S_{d+C} - S_{d+C-K})}{S_d} + \frac{\sum_{j=K+1}^{d+C} \lambda^j \sum_{i=d+C-j+1}^{d+C} b_i^2}{S_{d+C}} \frac{S_{d+C}}{S_d} \\
\leq& \frac{\sum_{j=1}^{K} \lambda^j (S_{d+C} - S_{d+C-K})}{S_d} + \left(\sum_{j=K+1}^{d+C} \lambda^j\right) \frac{S_{d+C}}{S_d}.
\end{aligned}$$

For any fixed K, the first term goes to 0, the second term can be arbitrarily small by choosing K large enough and that $\frac{S_{d+C}}{S_d} \to 1$. This establishes the first limit.

To obtain the second limit, observe the following identity

$$\lambda^j = \lambda^{k+1} + (1-\lambda)\sum_{i=j}^{k} \lambda^i.$$

Therefore

$$\begin{aligned}
&\frac{\sum_{j=1}^{n-d+C} b_{j+d-C}^2 \lambda^j}{\sum_{i=1}^{d} b_i^2} \\
=&\frac{\sum_{j=1}^{n-d+C} b_{j+d-C}^2 \lambda^{n-d+C+1}}{S_d} + \frac{(1-\lambda)\sum_{j=1}^{n-d+C} b_{j+d-C}^2 \sum_{i=j}^{n-d+C} \lambda^i}{S_d} \\
\leq&\lambda^{n-d+C+1}\frac{S_n}{S_d} + \frac{(1-\lambda)\sum_{i=1}^{n-d+C}\sum_{j=1}^{i} b_{j+d-C}^2 \lambda^i}{S_d} \\
\leq&\lambda^{n-d+C+1}\frac{S_n}{S_d} + \frac{(1-\lambda)\sum_{i=1}^{K}\sum_{j=1}^{i} b_{j+d-C}^2 \lambda^i}{S_d} + \frac{(1-\lambda)\sum_{i=K+1}^{n-d+C}\sum_{j=1}^{i} b_{j+d-C}^2 \lambda^i}{S_d} \\
\leq&\lambda^{n-d+C+1}\frac{S_n}{S_d} + \frac{(1-\lambda)\sum_{i=1}^{K}\lambda^i(S_{K+d-C} - S_{d-C})}{S_d} + \frac{(1-\lambda)\sum_{i=K+1}^{n-d+C}\lambda^i S_{i+d-C}}{S_d}
\end{aligned}$$

For any fixed K the second term goes to 0 as $d \to \infty$ due to the slow growth assumption. Now observe that

$$S_k = S_d \prod_{i=1}^{k-d} \frac{S_{d+i}}{S_{d+i-1}},$$

by the slow growth assumption, there exists $\underline{d}$ such that if $d > \underline{d}$, $\frac{S_d}{S_{d-1}} \leq \lambda^{-1/2}$. Let $k > d > \underline{d}$, then

$$S_k \leq S_d \lambda^{-(k-d)/2}.$$

Hence by choosing any $K > C$, the first and the third term is bounded by

$$\begin{aligned}
&\lambda^{n-d+C+1}\frac{S_n}{S_d} + \frac{\sum_{i=K+1}^{n-d+C}\lambda^i S_{i+d-C}}{S_d} \\
\leq&\lambda^{n-d+C+1}\lambda^{-(n-d)/2} + \sum_{i=K+1}^{n-d+C}\lambda^i\lambda^{-(i-C)/2}
\end{aligned}$$

where the first term goes to 0 as $n, d \to \infty$ since $d/n \to 0$. The second term can be arbitrarily small by choosing K large enough. This completes the proof. □

Now we proceed to the proof of Theorem 3.

Proof. Observe that $B\prime(\tilde{\beta}-\hat{\beta})$ can be separated into four terms. We will show that each of the four terms is of order $o\left(\sqrt{\frac{S_d}{T}}\right)$.

$$\begin{aligned}|B\prime(\tilde{\beta}-\hat{\beta})| \leq & \sum_{i=1}^{d^*}|\frac{b_i}{1+\lambda^i T}\beta_{(i)}| + \sum_{i=1}^{d^*}|\frac{b_i}{1+\lambda^i T}\frac{1}{T}(X\prime u)_{(i)}| \\ & + \sum_{j=d^*+1}^{n}|\frac{b_j T\lambda^j}{1+\lambda^j T}\beta_{(j)}| + \sum_{j=d^*+1}^{n}|\frac{b_j \lambda^j}{1+\lambda^j T}(X\prime u)_{(j)}|\end{aligned}$$

where the d^* is determined optimally by AIC.

The First Term

Consider the first term, to show that it is of order $o\left(\sqrt{\frac{S_d}{T}}\right)$, we need to show that for any $M>0$ and for any $\epsilon>0$ arbitrarily small, there exists $\underline{T}$ such that if $T>\underline{T}$,

$$\Pr\left(\sqrt{\frac{T}{S_d}}\sum_{i=1}^{d^*}|\frac{b_i}{1+\lambda^i T}\beta_{(i)}|>M\right)<\epsilon.$$

From Section 2.1, we know that for any given large C,

$$\Pr\left(\sum_{i=1}^{d^*}|\frac{b_i}{1+\lambda^i T}\beta_{(i)}|>\sum_{i=1}^{d+C}|\frac{b_i}{1+\lambda^i T}\beta_{(i)}|\right)\leq\delta^C$$

for some fixed $\delta\in(0,1)$, $d:=-\ln T\ln\lambda$ as $T\to\infty$. For convenience, denote $\sum_{i=1}^{d^*}|\frac{b_i}{1+\lambda^i T}\beta_{(i)}|$ by Σ^* and $\sum_{i=1}^{d+C}|\frac{b_i}{1+\lambda^i T}\beta_{(i)}|$ by Σ^C.

$$\begin{aligned}&\Pr\left(\sqrt{\frac{T}{S_d}}\Sigma^*>M\right)\\ =&\Pr\left(\{\sqrt{\frac{T}{S_d}}\Sigma^*>M\}\bigcap\{\Sigma^*>\Sigma^C\}\right)+\Pr\left(\{\sqrt{\frac{T}{S_d}}\Sigma^*>M\}\bigcap\{\Sigma^*\leq\Sigma^C\}\right)\\ \leq&\Pr\left(\Sigma^*>\Sigma^C\right)+\Pr\left(\sqrt{\frac{T}{S_d}}\Sigma^C>M\right)\leq\delta^C+\Pr\left(\sqrt{\frac{T}{S_d}}\Sigma^C>M\right)\end{aligned}$$

Hence it is sufficient to show that $\Pr\left(\sqrt{\frac{T}{S_d}}\Sigma^C > M\right)$ goes to 0 for any C, M. Since Σ^C is a positive random variable, by markov inequality,

$$\Pr\left(\sqrt{\frac{T}{S_d}}\Sigma^C > M\right) \leq \frac{\sqrt{\frac{T}{S_d}}\mathbb{E}[\Sigma^C]}{M}$$

Since Σ^C is a sum of half-normal random variable. We can calculate their expectations.

$$\begin{aligned}
\mathbb{E}[\Sigma^C] \leq & \sqrt{\frac{2}{\pi}} \sum_{i=1}^{d+C} \frac{|b_i|\lambda^{i/2}}{\lambda^i T} \\
= & \sqrt{\frac{2}{\pi}} \frac{\lambda^{-(d+C)/2}}{T} \sum_{i=1}^{d+C} |b_i| \lambda^{(d+C-i)/2} \\
\leq & \sqrt{\frac{2}{\pi}} \frac{\lambda^{-C/2}}{\sqrt{T}} \sqrt{\frac{\sum_{i=1}^{d+C} b_i^2 \lambda^{(d+C-i)/2}}{\sum_{i=j}^{d+C} \lambda^{(d+C-j)/2}}} \sum_{j=1}^{d+C} \lambda^{(d+C-j)/2} \\
= & \sqrt{\frac{2}{\pi}} \frac{\lambda^{-C/2}}{\sqrt{T}} \sqrt{\sum_{i=1}^{d+C} b_i^2 \lambda^{(d+C-i)/2}} \sqrt{\frac{1}{1-\lambda^{1/2}}}
\end{aligned}$$

where we applied quadratic mean inequality to get the second inequality. Therefore

$$\begin{aligned}
\Pr\left(\sqrt{\frac{T}{S_d}}\Sigma^C > M\right) \leq & \frac{\sqrt{\frac{T}{S_d}}\mathbb{E}[\Sigma^C]}{M} \\
\leq & \sqrt{\frac{2}{\pi}} \sqrt{\frac{1}{1-\lambda^{1/2}}} \frac{\lambda^{-C/2}}{M} \sqrt{\frac{\sum_{i=1}^{d+C} b_i^2 \lambda^{(d+C-i)/2}}{S_d}}
\end{aligned}$$

which goes to 0 by Lemma 1. Therefore, $\Pr\left(\sqrt{\frac{T}{S_d}}\Sigma^* > M\right)$ is arbitrarily small for any M as $T \to \infty$. This shows the first term is of order $o\left(\sqrt{\frac{S_d}{T}}\right)$.

Other terms

For other terms, we can use similar arguments as above, by observing all the following

probabilities are exponentially small in C. I.e. by Section 2, for all T large enough, there exists a fixed $\delta \in (0,1)$ such that

$$\Pr\left(\sum_{i=1}^{d^*} |\frac{b_i}{1+\lambda^i T}\frac{1}{T}(X\prime u)_{(i)}| > \sum_{i=1}^{d+C} |\frac{b_i}{1+\lambda^i T}\frac{1}{T}(X\prime u)_{(i)}|\right) \leq \delta^C;$$

$$\Pr\left(\sum_{j=d^*+1}^{n} |\frac{b_j T\lambda^j}{1+\lambda^j T}\beta_{(j)}| > \sum_{j=d-C+1}^{n} |\frac{b_j T\lambda^j}{1+\lambda^j T}\beta_{(j)}|\right) \leq \delta^C;$$

$$\Pr\left(\sum_{j=d^*+1}^{n} |\frac{b_j \lambda^j}{1+\lambda^j T}(X\prime u)_{(j)}| > \sum_{j=d-C+1}^{n} |\frac{b_j \lambda^j}{1+\lambda^j T}(X\prime u)_{(j)}|\right) \leq \delta^C.$$

In addition, observe that

$$\mathbb{E}\left[\sum_{i=1}^{d+C} |\frac{b_i}{1+\lambda^i T}\frac{1}{T}(X\prime u)_{(i)}|\right] \leq \sqrt{\frac{2}{\pi}}\frac{\lambda^{-C}}{\sqrt{(1-\lambda)T}}\sqrt{\sum_{i=1}^{d+C} b_i^2 \lambda^{d+C-i}};$$

$$\mathbb{E}\left[\sum_{j=d-C+1}^{n} |\frac{b_j T\lambda^j}{1+\lambda^j T}\beta_{(j)}|\right] \leq \sqrt{\frac{2}{\pi}}\frac{\lambda^{-C/2}}{\sqrt{(1-\lambda^{1/2})T}}\sqrt{\sum_{j=1}^{n-d+C} b_{j+d-C}^2 \lambda^{j/2}};$$

$$\begin{aligned}\mathbb{E}\left[\sum_{j=d-C+1}^{n} |\frac{b_j \lambda^j}{1+\lambda^j T}(X\prime u)_{(j)}|\right] &\leq \mathbb{E}\left[\sum_{j=d-C+1}^{n} |b_j \lambda^j (X\prime u)_{(j)}|\right] \\ &\leq \sqrt{\frac{2}{\pi}}\frac{\lambda^{-C}}{\sqrt{(1-\lambda)T}}\sqrt{\sum_{j=1}^{n-d+C} b_{d-C+j}^2 \lambda^j}.\end{aligned}$$

All these expectations go to 0 after multiplied with $\sqrt{\frac{S_d}{T}}$, hence by similar arguments for the first term, all four terms are of order $o\left(\sqrt{\frac{S_d}{T}}\right)$. □

We therefore conclude that

$$(\tilde{\beta}-\hat{\beta})\prime BB\prime(\tilde{\beta}-\hat{\beta}) = o\left((\beta-\hat{\beta})\prime BB\prime(\beta-\hat{\beta})\right)$$

for all B of finite number of columns and each column $B_{(i)}$ satisfies the slow growth condition, i.e. $\sum_{j=1}^{n} B_{(ij)}^2$ is of slow growth in n.

3.5 Conclusion

We have shown that the AIC model selection would select approximately the same number of parameters as the Bayesian method. The interpretation is that suppose the information we have about the data generating is as described in the introduction, then given our knowledge about the decreasing nature of the $\beta(i)$'s, our best estimator would be the bayesian estimator $\hat{\beta}$. However usually we cannot know the exact rate of decrease in the $\beta(i)$'s, and hence there is usually no way of constructing such bayesian estimator in practice. The above analysis shows that we do not need such a bayesian estimator because applying the AIC to sieve estimators results in a good approximation to the Bayesian estimator. Therefore it is optimal compared to every other estimator.

Moreover, although we have analyzed when $\sigma^2(\beta_{(i)}) = \lambda^i$ for $\lambda \in (0, 1)$, it can be seen that the above argument carries through as long as $\sigma^2(\beta_{(i)})$ is decreasing even faster than exponentially in i. Hence AIC is approximately the best estimator as long as the prior knowledge for $\beta(i)$ indicates an at least exponentially decreasing variances in i.

Bibliography

[1] Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Info. Theory*, 267-281, eds. B.N. Petrov and F. Csaki, Akademia Kiado, Budapest.

[2] Phillips, P. C. B. and Ploberger, W. (1994) Posterior Odds Testing for a Unit Root with Data-Based Model Selection. *Econometric Theory*, Vol. 10, 774-808.

[3] Kapoor, G.P. and Nautiyal, A. (1981) Polynomial Approximation of an Entire Function of Slow Growth. *Journal of Approximation Theory*, Vol. 32, 64-75.

[4] Kim, J. Y. (1998) Large Sample Properties of Posterior Densities, Bayesian Information Criterion and the Likelihood Principle in Nonstationary Time Series Models. *Econometrica*, 66, 359-380.

[5] Schwarz, Gideon E. (1978), "Estimating the dimension of a model", *Annals of Statistics*, 6 (2): 461?464

[6] Shibata, R. (1983) Asymptotic mean efficiency of a selection of regression variables. *Ann. Inst. Statisti. Math.* 35, 415-423.

[7] Strasser, H. (1985) *Mathematical theory of statistics: statistical experiments and asymptotic decision theory*, Walter de Gruyter.

[8] Stuart, A. M. (2010) Inverse problems: A Bayesian perspective. *Acta Numerica*, 19, 451-559.

[9] van der Vaart, A. W. (2000) *Asymptotic Statistics*, Cambridge University Press.

Chapter 4

Moderate Expected Utility

Coauthored with *Paulo Natenzon*

4.1 Introduction

Consider a decision maker who is most likely to choose option x in a binary comparison against y, and, in turn, most likely to choose option y in a binary comparison against z. Denoting by $\rho(x, y)$ the probability of choosing x over y and by $\rho(y, z)$ the probability of choosing y over z, we have

$$\rho(x, y) \geq 1/2 \text{ and } \rho(y, z) \geq 1/2. \tag{4.1}$$

A simple test of the transitivity of the decision maker's choices may require the decision maker to choose x most often in a binary comparison against z,

$$\text{If (4.1) holds, then } \rho(x, z) \geq 1/2. \tag{WST}$$

This basic postulate is known as *weak stochastic transitivity*. WST is the most permissive condition under which an analyst may obtain a coherent ranking over the choice options from binary choice data.

A more stringent transitivity criterion which is well-studied in the literature is *strong stochastic transitivity*:

$$\text{If (4.1) holds, then } \rho(x,z) \geq \max\{\rho(x,y), \rho(y,z)\}. \tag{SST}$$

Choice models that satisfy SST (such as the classic Logit model) are typically simple to analyze but fail to accommodate many empirically relevant phenomena.

In this paper, we consider a less studied, intermediate condition called *moderate stochastic transitivity*:

$$\text{If (4.1) holds, then } \rho(x,z) \geq \min\{\rho(x,y), \rho(y,z)\}. \tag{MST}$$

As we show in this paper, MST allows for many empirically relevant choice patterns ruled out by SST, and yet has significantly more empirical bite than WST.

Our main contribution is to characterize a family of parametric models of individual choice that generates the entire range of observable choice behavior that satisfies MST. This family can prove useful in applications where SST is violated, while at the same time allowing the analyst to make sharper predictions out of sample than WST models.

Our main results are two representation theorems for choice behavior that exhibits a moderate degree of transitivity. First, we write a slight strengthening of MST,

$$\text{If (4.1), then } \begin{cases} \rho(x,z) > \min\{\rho(x,y), \rho(y,z)\} \\ \text{or} \\ \rho(x,z) = \rho(x,y) = \rho(y,z) \end{cases} \tag{MST+}$$

which we call *moderate stochastic transitivity plus*, or MST+.

Theorem 18 shows that binary choice behavior over a finite set of alternatives satisfies MST+ if and only if it can be represented by a *moderate utility model* (MUM). A binary choice rule ρ is a MUM if there exists a utility function u and a distance metric d such that, for all $x \neq y$,

$$\rho(x, y) = F\left(\frac{u(x) - u(y)}{d(x, y)}\right) \tag{MUM}$$

where F is a strictly increasing transformation with $F(t) = 1 - F(-t)$ for all $t \in \mathbb{R}$. The MST and $MST+$ postulates were formulated by [5] and [16], while the MUM formula was proposed by [20]. Hence, the equivalence that we establish in Theorem 18 answers a question that has been open for several decades.

In a MUM, the decision maker's ability to discriminate between a pair of options x and y depends on the difference in value $u(x) - u(y)$ and on the dissimilarity of the options given by the distance $d(x, y)$. Note the role of the distance metric d: for a given difference in value $u(x) - u(y)$, larger values of the distance $d(x, y)$ drive choice probabilities closer to $1/2$. In other words, more dissimilar options are more difficult to compare. The abstract metric d does not have to be the standard metric of Euclidean space: in applications, d takes the form of a statistical distance between random variables, angular distance between vectors in multi-attribute settings, and so on. In Section 4.6, we show that specific functional forms of u, d and F yield several familiar models from the discrete choice estimation literature. These particular instances of MUM (such as the probit model with correlated shocks) were developed as practical solutions to address violations of SST observed in data. Hence, our Theorem 1 determines the empirical content of a class of models that have demonstrated relevance in applications.

We next turn to the important question of measurement. Under what conditions can the analyst separately measure utility and dissimilarity from observed choices? In Theorem 21, we provide an answer by enriching the domain of choice options to include lotteries over the

alternatives. By imposing continuity, linearity, and convexity assumptions, in addition to MST+, our *moderate expected utility model* (MEM) characterization identifies (i) a unique von Neumann-Morgenstern expected utility function over lotteries; (ii) a norm, induced by an inner product on the relevant linear space, that is unique up to two scaling factors; and (iii) a monotonic transformation F that is unique up to the same two scaling factors.

Section 4.6 relates our representation results to the existing literature. We show that the MUM nests several classic binary choice models as special cases. Some special cases of the MUM are also instances of the random utility model (RUM). We show that, despite having a non-empty intersection, neither the MUM nor the RUM nest each other. Likewise, we show that the MEM characterized in Theorem 21 neither nests nor is nested in the random expected utility model studied by [17]. We conclude with a discussion of the possible extensions of our model to incorporate choices over more than two options.

4.2 Moderate stochastic transitivity

Let Z be a finite set of choice options. A (binary, stochastic) *choice rule* on Z is a function $\rho : Z^2 \to [0, 1]$ such that $\rho(x, y) + \rho(y, x) = 1$ for every $x, y \in Z$. The number $\rho(x, y)$ denotes the probability that the decision maker selects option x in a binary comparison against y.

Let $\wedge$ and $\vee$ denote the min and max operators, respectively, so that $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. The two most commonly studied notions of transitivity for binary choice data are weak stochastic transitivity (WST) and strong stochastic transitivity (SST):

$$\text{(WST)}\quad \rho(x, y) \wedge \rho(y, z) \geq 1/2 \Longrightarrow \rho(x, z) \geq 1/2$$

$$\text{(SST)}\quad \rho(x, y) \wedge \rho(y, z) \geq 1/2 \Longrightarrow \rho(x, z) \geq \rho(x, y) \vee \rho(y, z)$$

In this paper we focus on a less studied, intermediate form of transitivity called moderate stochastic transitivity (MST):

(MST) $\rho(x,y) \wedge \rho(y,z) \geq 1/2 \Longrightarrow \rho(x,z) \geq \rho(x,y) \wedge \rho(y,z)$

The definitions clearly imply that SST $\Rightarrow$ MST $\Rightarrow$ WST. Our main results characterize the set of choice rules that satisfy a slightly stronger version of MST, namely

(MST+) $\rho(x,y) \wedge \rho(y,z) \geq 1/2 \Longrightarrow \rho(x,z) > \rho(x,y) \wedge \rho(y,z)$ or $\rho(x,z) = \rho(x,y) = \rho(y,z)$.

The strengthening is the key to obtain our representation results. Note, however, that the only difference between MST and MST+ is that the knife-edge case

$$\rho(x,y) \vee \rho(y,z) > \rho(x,z) = \rho(x,y) \wedge \rho(y,z)$$

is allowed by MST but ruled out by MST+. Hence, MST and MST+ are empirically indistinguishable: no finite amount of data allows an analyst to tell them apart.

Choice models that satisfy a moderate degree of transitivity in the form of MST or MST+ are convenient for two reasons. First, moderate transitivity holds in many applications in which the more restrictive SST condition is violated. Hence, a choice model that satisfies MST/MST+ but allows for violations of SST may provide the flexibility that is needed to accommodate empirically relevant choice phenomena. Second, MST/MST+ are significantly more restrictive than WST, and restricting the analysis to models that satisfy this stronger postulate results in greater out-of-sample predictive power.

The classic Example 15, below, provides the intuition for why violations of SST must be expected when some pairs of alternatives are easier to compare than others.

Example 15 (attributed to L. J. Savage, adapted from [35])**.** *An individual has a difficult time comparing a trip to Paris, denoted P and a trip to Rome, denoted R, so that she is equally likely to pick either option $\rho(P,R) = 1/2$. The individual still has trouble deciding if the trip to Paris is enhanced by a €5 bonus, denoted by P^+. In other words, $\rho(P^+,R)$ is still approximately $1/2$. But when pressed to decide between the two Paris trip options,*

the individual clearly prefers the bonus, so that $\rho(P^+, P)$ *is close to* 1. *SST requires that* $\rho(P^+, R) \geq \rho(P^+, P)$ *which is intuitively violated in this case, while MST+ only requires the more plausible inequality* $\rho(P^+, R) > \rho(P, R)$.

The lesson we glean from Savage's Example 15 is that utility values cannot be the only factor determining the difficulty of comparing two options. A small monetary bonus makes the choice comparison between the two Paris trips very easy. The same monetary bonus has negligible impact, however, on the difficulty of comparing a trip to Paris and a trip to Rome. An important consequence is that models in which choice probabilities depend solely on utility (such as the classic Logit model) fail to capture this intuitively plausible behavior.

Savage's Example 15 was anticipated in the context of the theory of consumer choice by [16]. [6] provides perhaps the earliest empirical demonstration of this intuition in an experimental setting, while [37] provide the first clear empirical demonstration of this idea in psychology. The evidence for systematic violations of SST in individual choices is very robust. Reviewing some of this evidence, [27] note that "weak and moderate stochastic transitivity are often satisfied, although a few exceptions have been noted", while "[s]trong stochastic transitivity is frequently violated." Figures 4.2 and 4.2, below, illustrate how ease of comparison drives violations of SST in individual choice experiments with human and non-human subjects alike.

Relaxing SST to MST/MST+ allows the analyst to address the range of empirical phenomena illustrated by the examples above. At the same time, MST/MST+ retain significantly more empirical bite than WST. To see this, suppose the choice rule ρ on Z satisfies WST. Enumerate the n options in $Z = \{x^1, x^2, \ldots, x^n\}$ in such a way that $\rho(x^i, x^j) \geq 1/2$ whenever $i \leq j$. For the sake of simplicity, let us assume that choice probabilities differ whenever possible, so that the set $\{\rho(x, y) \in [0, 1] : x \neq y\}$ has maximum cardinality with $n(n-1)$ elements.

When $Z = \{x^1, x^2, x^3\}$ has three alternatives, WST allows ρ to have six strict orderings.

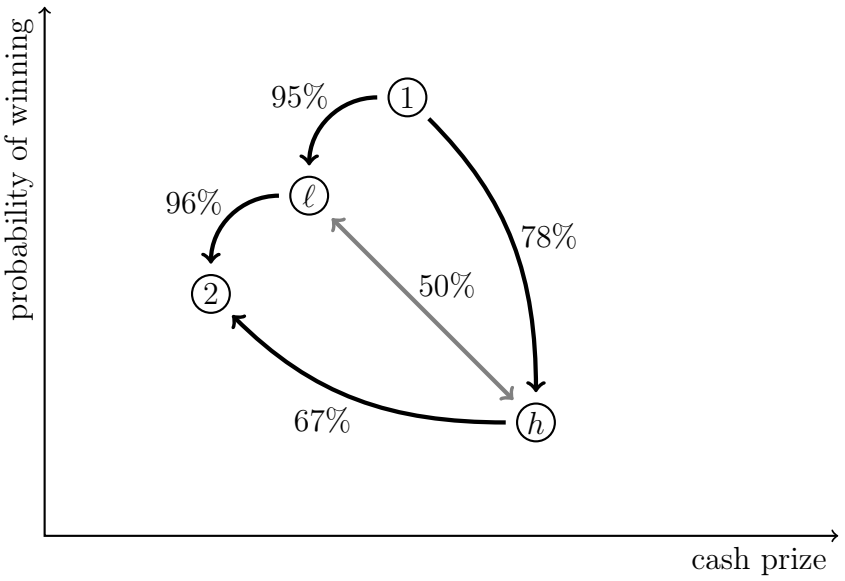


Figure 4.1: Binary choice frequencies violate SST but satisfy MST

[32] recorded thousands of choices by 21 male Caltech undergraduates using simple lotteries (p, m) that pay m dollars with probability p in the lab. A high risk lottery h and a low risk lottery ℓ were fine-tuned to each individual to be approximately indifferent, (i.e., equally likely to be chosen in a binary comparison). Slightly perturbed versions of h and ℓ were then offered for comparison against several types of 'decoy' lotteries. Figure 4.2 depicts the relative location of two decoy lotteries 1 and 2 with respect to h and ℓ. Decoy lottery 1 dominates ℓ and was chosen 95% of the time against ℓ but only 78% of the time against h. Thus, choice frequencies violate SST in the direction $1 \to \ell \to h$. Decoy lottery 2, on the other hand, is dominated by ℓ and was chosen 4% of the time against ℓ and 33% of the time against h. Hence, choice frequencies also violate SST in the direction $h \to \ell \to 2$. It is easy to verify that MST+ holds in both cases.

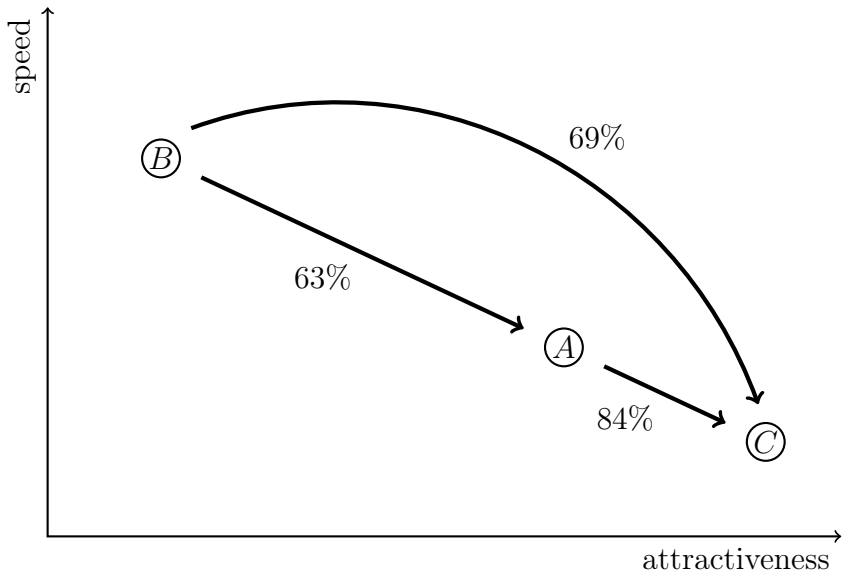


Figure 4.2: Binary choice frequencies violate SST but satisfy MST in [23].

[23] recorded hundreds of mating decisions by female túngara frogs. Female túngara frogs choose mates based on the sound of their call. Figure 4.2 shows how the calls of the three male options A, B and C were differentiated along two attributes. In the binary choice data, option B is chosen in 63% of the trials against A; option A is chosen in 84% of the trials against C; and option B is chosen in 69% of the trials against C. Choices therefore satisfy MST+ but violate SST.

MST+, which is equivalent to MST in this case, rules out the last two of the six strict orderings, where $\rho(x^1, x^3) < \rho(x^2, x^3) \wedge \rho(x^1, x^2)$. Let $\#WST(n) = [n(n-1)/2]!$ denote the number of strict orderings allowed by WST when Z has n options, and likewise, let $\#MST(n)$ denote the number of strict orderings allowed by MST+. The ratio $\#MST(n)/\#WST(n)$ can be interpreted as a measure of the restriction imposed on observable choice data by MST+ compared to WST. In the case $n = 3$ we just showed the ratio $\#MST(3)/\#WST(3)$ is equal to 2/3. This ratio decreases to less than 1/4 when $n = 4$ and less than 1/17 when $n = 5$. In fact, the ratio is arbitrarily small when n is large:

Proposition 16. $\lim_{n\to\infty} \#MST(n)/\#WST(n) = 0$.

We prove Proposition 16 in the Appendix. For completeness, we also show in the Appendix that the ratio between $\#SST(n)$ and $\#MST(n)$ goes to zero when n is large. To summarize, a model that covers the range of choice behavior allowed by MST+ is useful in two ways: first, it provides the flexibility that is needed to deal with empirical violations of SST. Second, it imposes significant restrictions out of sample —allowing the analyst to make sharper predictions— than the more lenient WST. We describe such a model in the next section.

4.3 Moderate utility model

A choice rule ρ on a finite set Z is a *moderate utility model (MUM)* if there is a utility function $u : Z \to \mathbb{R}$, a distance metric $d : Z^2 \to \mathbb{R}_+$ and a strictly increasing function F, such that for all $x \neq y$,

$$\rho(x, y) = F\left(\frac{u(x) - u(y)}{d(x, y)}\right) \tag{4.2}$$

where F satisfies $F(t) = 1 - F(-t)$ for all t. The utility u represents the value of each option. It is easy to see that $\rho(x, y) \geq 1/2$ if and only if $u(x) \geq u(y)$ for any $x, y \in Z$. The distance d

can be interpreted as a measure of the dissimilarity of the objects; for a given fixed difference in utility, more dissimilar objects are harder to compare. The ratio $[u(x) - u(y)]/d(x, y)$ can be interpreted as the strength of preference for option x over option y, while the function F maps strength of preference to choice probabilities.

The MUM formula is proposed in [20]. Taking the distance d in (4.2) to be the special case of the discrete metric $d(x, y) = 1$ if $x \neq y$ and $d(x, x) = 0$ for all x, we obtain the classic *Fechnerian utility model* as a special case

$$\rho(x, y) = F\left(u(w) - u(x)\right)$$

in which the ability to discriminate between x and y depends solely on the difference between the values of x and y [7, 9, 12, 14]. The role of a non-trivial distance metric d in a MUM is to make the choice probabilities of options that are more difficult to compare closer to $1/2$.

A non-trivial distance metric d gives the MUM the flexibility that is needed to deal with empirical violations of SST. For example, consider how the MUM accommodates the choices over trips in Example 15. Let the trip to Paris and the trip to Rome have utility $u(P) = U(R) = 0$, and let the trip to Paris with the €5 bonus have utility $u(P^+) = 1$. Let the distance metric be given by $d(P, P^+) = \varepsilon > 0$ and $d(P, R) = d(P^+, R) = 1/\varepsilon > 0$. Finally, let $F = \Phi$ be the standard Gaussian cdf. Applying (4.2) we have $\rho(P, R) = 1/2$, $\rho(P^+, P) = \Phi(1/\varepsilon)$ and $\rho(P^+, R) = \Phi(\varepsilon)$. Taking $\varepsilon > 0$ small, we obtain $\rho(P^+, P)$ close to one and $\rho(P^+, R)$ close to $1/2$ as desired. The dissimilarity of the two Paris trip options is small according to the metric d, which makes them easy to compare. The Rome option is very dissimilar from the other options according to d and therefore difficult to compare.

A MUM can also address examples in Figure 4.2 and 4.2 by explicitly mapping the abstract utility u and distance d to the attribute space depicted in Figures 4.2 and 4.2. It is important to note that d can differ from the standard Euclidean distance. For example, [21]

relate the difficulty of comparing two options to the angular distance between the vectors of their observable attributes. For instance, options 1 and ℓ in Figure 4.2 form a small angle with respect to the origin, so that $d(1, \ell)$ is small, while options 1 and h form a wider angle with respect to the origin, so that $d(1, h)$ is large. Hence, a decision maker may ascribe to h and ℓ the same utility values, and yet have an easier time comparing 1 to ℓ than to h. Similarly, options A and C are much closer in angular distance than options B and C in Figure 4.2. Options A and B may be close in value, but frogs find option C much easier to compare to A than to B. Hence, a MUM can address both situations in which the ease of comparison involves dominance (Figure 4.2) and non-dominance (Figure 4.2) in the attribute space. The next example is perhaps the most familiar special case of a MUM in the discrete choice estimation literature.

Example 17. *A concrete example of a MUM used in the discrete choice estimation literature is the* binary probit model*, first proposed by [34]. In a probit model there is a Gaussian vector $X = (X_1, \ldots, X_n)$, each coordinate X_i corresponding to an option $x^i \in Z$, such that $\rho(x^i, x^j) = \mathbb{P}\{X_i > X_j\}$ for all $x^i, x^j \in Z$. Note that*

$$\rho(x^i, x^j) = \mathbb{P}\left\{\frac{X_i - X_j - \mathbb{E}[X_i - X_j]}{\sqrt{Var(X_i - X_j)}} > \frac{\mathbb{E}[X_i - X_j]}{\sqrt{Var(X_i - X_j)}}\right\} = \Phi\left(\frac{\mathbb{E}[X_i - X_j]}{\sqrt{Var(X_i - X_j)}}\right)$$

which is a special case of (4.2) when $u(i) = \mathbb{E}[X_i]$ is the utility function, $d(i, j) = \sqrt{Var(X_i - X_j)}$ is the distance metric (once we rule out perfectly correlated variables), and $F = \Phi$ is the cdf of the standard Gaussian distribution.

[20] proposed the MUM definition (4.2) and showed that all MUMs satisfy MST. In our first characterization theorem, below, we show that MUMs also satisfy the stronger MST+ condition. In fact, we show that MST+ is both necessary and sufficient for a choice rule to be a MUM.

Theorem 18. *A choice rule ρ on a finite Z is a MUM if and only if it satisfies MST+.*

We prove Theorem 4.2 in the Appendix. Necessity is shown in two steps: first, we show every MUM satisfies MST (this is the step already proved by [20]). Then, we show that a MUM must also satisfy MST+. For sufficiency, we explicitly construct the utility u and distance d; we show that d satisfies the properties of a metric (the key property being the triangle inequality); and we show that an ordinal representation with u and d holds:

$$\rho(w,x) \geq \rho(y,z) \text{ if and only if } \frac{u(w)-u(x)}{d(w,x)} \geq \frac{u(y)-u(z)}{d(y,z)} \tag{4.3}$$

Then, it is straightforward to find a transformation F such that the cardinal representation of equation (4.2) holds.

It is easy to see why the stronger MST+ is needed in Theorem 1 instead of MST. Suppose $\rho(x,y) > \rho(y,z) = \rho(x,z) \geq 1/2$. A MUM representation would require

$$\frac{u(x)-u(y)}{d(x,y)} > \frac{u(y)-u(z)}{d(y,z)} = \frac{u(x)-u(z)}{d(x,z)},$$

which in turn would imply a violation of the triangle inequality:

$$d(x,z) = \frac{u(x)-u(y)+u(y)-u(z)}{u(y)-u(z)} d(y,z) > d(x,y) + d(y,z).$$

To obtain the identification of the MUM parameters, in the next section we enrich the choice domain to include all lotteries over the finite set Z. With a finite set of options, however, an analyst who observes ρ still obtains some ordinal information about u and d:

Proposition 19. *Let ρ be a MUM with parameters (u,d,F). Then*

1. *$\rho(x,y) \geq 1/2$ if and only if $u(x) \geq u(y)$;*

2. *$\rho(x,y) > \rho(x,z) > \rho(y,z) \geq 1/2$ implies $d(x,y) < d(x,z)$.*

Proof. From the MUM formula (4.2) it follows that $\rho(x, y) > 1/2$ if and only if $[u(x) - u(y)]/d(x, y) > 0$ if and only if $u(x) > u(y)$ proving (i). Suppose the assumption in item (ii) holds. Then, item (i) implies $u(y) \geq u(z)$ hence $u(x) - u(y) \leq u(x) - u(z)$. The MUM formula (4.2) implies $[u(x)-u(y)]/d(x, y) > [u(x)-u(z)]/d(x, z)$, hence $d(x, z) > d(x, y)$. □

Item (i) in Proposition 19 shows choices in a MUM reveal a complete and transitive ranking over the options represented by the utility parameter u. Item (ii) shows how every violation of SST is explained by the distance parameter d. To illustrate, consider the choice data from Figure 4.2. By (i) and (ii) in Proposition 19, the analyst concludes that any MUM that generates this data must satisfy $u(B) > u(A) > u(C)$ and $d(A, C) < d(B, C)$. Likewise, every MUM that generates the data in Figure 4.2 must satisfy $u(1) > u(\ell) = u(h) > u(2)$, $d(2, \ell) < d(2, h)$ and $d(1, \ell) < d(1, h)$.

It is worth noting that the inequalities $d(A, C) < d(B, C)$, $d(2, \ell) < d(2, h)$ and $d(1, \ell) < d(1, h)$ revealed by choice data agree with the inequalities an analyst would obtain by applying the standard Euclidean distance to measure the dissimilarity of the options in the space of observable attributes in both Figure 4.2 and Figure 4.2. In empirical applications, however, non-Euclidean distance functions —such as angle distance— may provide an even better fit than the standard Euclidean distance when the analyst maps the abstract distance parameter in the MUM model to the space of observable attributes.[1]

4.4 Moderate expected utility model

We continue to let Z be a finite set of objects and we extend the domain of choice alternatives to the set of all lotteries over Z, denoted by Δ. We identify Δ with the $n - 1$ dimensional simplex $\{x \in [0, 1]^n : x_1 + \cdots + x_n = 1\}$. The function $U : \Delta \to [0, 1]$ is an *expected*

[1]For a concrete example of the use of a non-Euclidean distance in applications see [21]. For the possible pitfalls that may arise when relating the abstract parameters to the attribute space, see the analysis of the issue of monotonicity in [1].

utility function if it is linear and onto. A choice rule $\rho : \Delta^2 \to [0,1]$ is a *moderate expected utility model* (MEM) if there exist an expected utility function U, a norm $\|\cdot\|$ induced by an inner product, and a strictly increasing and continuous transformation F, such that, for any lotteries $x \neq y$ in Δ,

$$\rho(x,y) = F\left(\frac{U(x) - U(y)}{\|x - y\|}\right). \tag{4.4}$$

Example 20. *For a concrete example of a MEM, extend the binary probit model of Example 17 to the set of lotteries over the finite set Z by letting*

$$\rho(x,y) = \mathbb{P}\{X'x > X'y\} = \Phi\left(\frac{u'x - u'y}{\sqrt{(x-y)'\Lambda'\Lambda(x-y)}}\right)$$

where $u = \mathbb{E}[X]$ is the mean and $\Lambda'\Lambda = Var(X)$ is the covariance matrix of the Gaussian vector X. This decision maker is a (random) expected utility maximizer, and her Bernoulli index is given by the random vector X. This model is a special case of (4.4), where $U(x) = u'x$ is the linear transformation given by the mean vector u, $F = \Phi$ is the cdf of the standard Gaussian distribution, and the norm $\|\cdot\|$ is induced by the inner product $\langle x, y\rangle = x\Lambda'\Lambda y$.

Every MEM satisfies MST+. This can be shown by repeating the argument for a MUM in the proof of Theorem 4.2. Compared to the MUM, however, the MEM is defined in the richer domain of lotteries contained in a linear vector space; it imposes linearity on the utility function U; and it requires the distance metric to be a norm induced by an inner product. These assumptions carry additional testable implications beyond MST+.

First, the requirement that U is onto $[0,1]$ implies that a MEM cannot be constant, that is $\rho(x,y) > 1/2$ for some lotteries x, y. Second, every MEM is *continuous* at every point in the domain except along the diagonal $\{(x,x) \in \Delta^2 : x \in \Delta\}$. Third, every MEM is *linear*, that is, for all $0 < \alpha < 1$ and any lotteries $x, y, z \in \Delta$ we have $\rho(x,y) = \rho(\alpha x + (1-\alpha)z, \alpha y + (1-\alpha)z)$. And finally, every MEM ρ is *convex*, that is, whenever $\rho(x,y) = 1/2$ and $\rho(x,z) = \rho(y,z) >$

$1/2$ for some $x \neq y$, we have $\rho\left(x/2 + y/2, z\right) > \rho(\alpha x + (1-\alpha)y, z)$ for all $\alpha \neq 1/2$.

Continuity and linearity are familiar postulates from the random choice literature [see, for example, 17], while convexity deserves some discussion. As in Example 15, suppose an individual is equally likely to choose a trip to Paris (P) and a trip to Rome (R) so that $\rho(P, R) = 1/2$. Suppose, moreover, that she is more likely to choose either trip over a trip to London (L), and that both Paris and Rome beat London with the exact same probability, that is, $\rho(P, L) = \rho(R, L) > 1/2$. The convexity postulate requires that, among all the lotteries $\alpha P + (1-\alpha)R$ that give a trip to Paris with probability α and a trip to Rome with probability $1-\alpha$, the even coin-flip ($\alpha = 1/2$) be the most likely lottery to be chosen in a binary comparison against London. In particular, the even coin-flip between Paris and Rome must be chosen more often against the trip to London than either Paris or Rome for sure.

In the context of a MEM, the convexity postulate can be interpreted as saying that the dissimilarity metric d must be strictly convex. To see this, note that, to address the example, above Paris and Rome must have the same value $u(P) = u(R)$. In fact, under linearity all lotteries $\alpha P + (1-\alpha)R$ have the same value. Note also that Paris and Rome must have the same degree of dissimilarity to London: $d(P, L) = d(R, L)$. As we change α, the value of the options does not change; any change in the choice probability $\rho(\alpha P + (1-\alpha)R, L)$ must come from the dissimilarity d. Hence, the convexity postulate says that, whenever $d(P, L) = d(R, L)$ holds, we must have $d(P/2 + R/2, L) < d(\alpha P + (1-\alpha)R, L)$.

Our next Theorem shows that these assumptions, in addition to MST+, are necessary and sufficient for a choice rule to be a MEM.

Theorem 21. *ρ is a MEM iff it is non-constant, continuous, linear, convex, and MST+.*

We prove Theorem 21 in the Appendix. Necessity is straightforward. To prove sufficiency, we first show that ρ has a unique linear extension to the $n-1$ dimensional hyperplane that

contains Δ. Transitivity, linearity and continuity allow us to invoke a standard result to obtain the expected utility function U. The indifference sets $I(y) := \{x \in \Delta : U(x) = U(y)\}$ are then parallel hyperplanes of dimension $n-2$. To construct the norm, we fix one indifference set I, and one lottery $\bar{y}$ with $U(x) > U(\bar{y})$ for all $x \in I$, as illustrated in Figure 4.4. The bulk of the work is showing that the contour sets $\{x \in I : \rho(x, \bar{y}) \geq \alpha\}$ must be compact, convex, dilations of one another, and centrally symmetric around the point $\hat{x}$ that maximizes $x \mapsto \rho(x, \bar{y})$ on I. Then, we take one such contour set to be the unit ball that defines the norm in the $n-2$ dimensional subspace parallel to I. We use the convexity postulate and a characterization of inner product spaces to show that this norm comes from an inner product. We then extend this inner product in one more dimension to obtain the MEM representation.

In the MEM representation, the expected utility U turns out to be unique, while the norm $\|\cdot\|$ and the transformation F are unique up to two scaling factors:

Proposition 22. *Suppose $(U_1, \|\cdot\|_1, F_1)$ is a MEM representation of ρ, and let the constant $T := F_1^{-1}(\max_{x,y} \rho(x,y))$. Then $(U_2, \|\cdot\|_2, F_2)$ represents the same MEM if, and only if, $U_2 = U_1 = U$, $[\ker(\mathbf{1}) \cap \ker(U)]^{\perp_2} = [\ker(\mathbf{1}) \cap \ker(U)]^{\perp_1}$, and there exist $A, B > 0$ such that:*

1. *$\|x\|_2 = A\|x\|_1$ for all $x \in \ker(\mathbf{1}) \cap \ker(U)$,*
2. *$\|x\|_2 = B\|x\|_1$ for all $x \in [\ker(\mathbf{1}) \cap \ker(U)]^{\perp}$, and*
3. *$F_2(t) = F_1\left(\frac{ATt}{\sqrt{T^2 + (A^2 - B^2)t^2}}\right)$ for all $-T/B \leq t \leq T/B$.*

We prove Proposition 22 in the Appendix. The uniqueness of the utility U up to an affine transformation comes from the standard vNM expected utility representation result. Full uniqueness comes from our normalization requiring that U is a function onto $[0,1]$.

Both norms $\|\cdot\|_1$ and $\|\cdot\|_2$ and their respective inner products $\langle\cdot,\cdot\rangle_1$ and $\langle\cdot,\cdot\rangle_2$ are defined on the $n-1$ dimensional subspace $\ker(\mathbf{1}) = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_1 + \cdots + x_n = 0\}$. The set

$\ker(\mathbf{1}) \cap \ker(U)$ is a subspace of $\ker(\mathbf{1})$ of dimension $n-2$, composed by vectors of the form $x - y$ with $U(x) = U(y)$. The orthogonal complement of this set, denoted $[\ker(\mathbf{1}) \cap \ker(U)]^{\perp}$ is therefore a single-dimensional subspace of $\ker(\mathbf{1})$. Proposition 22 says this orthogonal complement must be the same according to both inner products.

Items (i)–(iii) in Proposition 22 say the norm $\|\cdot\|$ and the transformation F are unique up to two scaling factors $A, B > 0$. The constant $A > 0$ rescales the norm along the $n-2$ dimensional subspace $\ker(\mathbf{1}) \cap \ker(U)$, while the constant $B > 0$ rescales the norm along its single dimensional orthogonal complement $[\ker(\mathbf{1}) \cap \ker(U)]^{\perp}$. It is easy to see that since

$$F_1\left(\frac{U(x) - U(y)}{\|x - y\|_1}\right) = \rho(x, y) = F_2\left(\frac{U(x) - U(y)}{\|x - y\|_2}\right)$$

we must have, for each $x \neq y$,

$$F_2^{-1}(\rho(x, y)) = \frac{\|x - y\|_1}{\|x - y\|_2} \times F_1^{-1}(\rho(x, y))$$

that is, F_2 can be obtained from F_1 by rescaling each point in the domain of F_1 by the ratio of the two norms. Item (iii) shows explicitly how one obtains F_2 from F_1 and the values of $A, B > 0$. In particular, when $A = B > 0$ we have $F_2(t) = F_1(t/A)$ for all t.

To provide some intuition for the existence of the two scaling factors $A, B > 0$, consider two lotteries x, x' on the same indifference plane I and a lottery $\bar{y}$ with lower utility, as shown in Figure 4.4. Suppose $\rho(x, \bar{y}) > \rho(x', \bar{y})$. In the representation, we have the inequality

$$\frac{U(x) - U(\bar{y})}{\|x - \bar{y}\|} > \frac{U(x') - U(\bar{y})}{\|x' - \bar{y}\|}$$

Let $\hat{x}$ be the projection of $\bar{y}$ onto the indifference plane I. By orthogonality of the projection,

we have $\|x-\bar{y}\|^2 = \|x-\hat{x}\|^2 + \|\hat{x}-\bar{y}\|^2$ and $\|x'-\bar{y}\|^2 = \|x'-\hat{x}\|^2 + \|\hat{x}-\bar{y}\|^2$, and hence

$$\begin{aligned}\frac{U(x)-U(\bar{y})}{\|x-\bar{y}\|} > \frac{U(x')-U(\bar{y})}{\|x'-\bar{y}\|} &\Leftrightarrow \|x-\bar{y}\| < \|x'-\bar{y}\| \\ &\Leftrightarrow \|x-\hat{x}\|^2 + \|\hat{x}-\bar{y}\|^2 < \|x'-\hat{x}\|^2 + \|\hat{x}-\bar{y}\|^2 \\ &\Leftrightarrow A^2\|x-\hat{x}\|^2 + B^2\|\hat{x}-\bar{y}\|^2 < A^2\|x'-\hat{x}\|^2 + B^2\|\hat{x}-\bar{y}\|^2\end{aligned}$$

where $A > 0$ rescales the norm of the components parallel to I and $B > 0$ rescales the norm of the components orthogonal to I. It is easy to see that the inequality is maintained for any $A, B > 0$. Hence, the rescaling preserves the ordinal representation for $\rho(\cdot, \bar{y})$ on the indifference plane I for any $A, B > 0$. By linearity, it preserves the ordinal representation in the entire domain of ρ. To preserve the cardinal representation, we must adjust F accordingly using the same factors $A, B > 0$ as stated in item (iii) of Proposition 22.

We can also describe how the two scaling factors work in terms of the inner product that generates the norm. The inner product can be written as $\langle x, y\rangle = xMy$ where M is a $n \times n$ matrix. Since the inner product is defined on the $n-1$ dimensional subspace $\ker(\mathbf{1})$, we can always find a matrix M with zeroes on the last row and the last column:

$$M = \begin{bmatrix} \tilde{M} & \mathbf{0} \\ \mathbf{0'} & 0 \end{bmatrix} \tag{4.5}$$

where $\tilde{M}$ is a symmetric, positive definite matrix of dimension $n-1$ by $n-1$. An implication of Proposition 22 is that, if the analyst fixes F in the MEM representation, then the utility U and the matrix $\tilde{M}$ in (4.5) are uniquely pinned down. In particular, when $F = \Phi$ is the standard Gaussian distribution, the parameters of the binary probit over lotteries in Example 20 are point-identified with the covariance matrix written as in (4.5).

4.5 The connection between MEM and MUM

Recall that a MUM is defined on a finite set of options $Z = \{x^1, \ldots, x^n\}$, while a MEM is defined on the richer set Δ of lotteries over Z. For convenience, we abuse notation and use the same symbol x^i to denote a degenerate lottery in Δ which gives prize $x^i \in Z$ with probability one. We say that a MEM representation $(U, \|\cdot\|, F')$ *is an extension* of a MUM representation (u, d, F) to Δ when $U(x^i) = u(x^i)$ for each $x^i \in Z$, $d(x^i, x^j) = \|x^i - x^j\|$ for each $x^i, x^j \in Z$ and $F'(t) = F(t)$ for each t in the domain of F. In this case, we also say that (u, d, F) is a *restriction* of the MEM representation $(U, \|\cdot\|, F')$ to Z.

Every MUM representation obtained as a restriction of a MEM satisfies three properties. First, u inherits the normalization from U. Since U is onto $[0, 1]$, we must have

$$\min_{z \in Z} u(z) = 0 \text{ and } \max_{z \in Z} u(z) = 1. \tag{4.6}$$

Second, d inherits the following property from $\|\cdot\|$:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} d(x^i, x^n)^2 \alpha_i \alpha_j < 0 \quad \text{for all } 0 \neq \alpha \in \mathbb{R}^n \text{ with } \alpha_1 + \cdots + \alpha_n = 0. \tag{4.7}$$

Letting D be the symmetric $n \times n$ matrix with entry (i, j) given by the square of the distance between x^i and x^j, equation (4.7) says the quadratic form $\alpha' D \alpha$ restricted to $\alpha \in \ker(\mathbf{1})$ must be negative definite. Too see why (4.7) holds, let $y^i = x^i - x^n$ for each $i = 1, \ldots, n-1$

and note that $y^1, \ldots, y^{n-1}$ belong to $\ker(\mathbf{1})$, the domain of $\|\cdot\|$. Therefore, we have

$$\begin{aligned}
-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} d(x^i, x^n)^2\alpha_i\alpha_j &= \frac{1}{2}\sum_{i=1}^{n-1}\sum_{j=1}^{n-1} \alpha_i\alpha_j \left[d(x^i, x^n)^2 + d(x^j, x^n)^2 - d(x^i, x^j)^2\right] \\
&= \frac{1}{2}\sum_{i}\sum_{j} \alpha_i\alpha_j \left(\left\langle y^i, y^i\right\rangle + \left\langle y^j, y^j\right\rangle - \left\langle y^i - y^j, y^i - y^j\right\rangle\right) \\
&= \sum_{i}\sum_{j} \alpha_i\alpha_j \left\langle y^i, y^j\right\rangle \\
&= \|\alpha_1 y^1 + \cdots + \alpha_{n-1} y^{n-1}\|^2 > 0.
\end{aligned}$$

[31] shows condition (4.7) has a geometric interpretation: it holds if and only if it is possible to map the n options in Z to the extreme points of a polytope of dimension $n-1$ in Euclidean space, in which the length of each vertex $[x^i, x^j]$ is equal to $d(x^i, x^j)$.

Finally, F clearly inherits continuity from F'. As it turns out, these three necessary properties are also sufficient for the existence of a MEM extension when the MUM represented by (u, d, F) satisfies positivity, that is, when $\rho(x, y) > 0$ for all $x, y \in Z$:

Proposition 23. *Let (u, d, F) be a MUM representation of $\rho > 0$ on Z. There exists a MEM extension of (u, d, F) to Δ if and only if u satisfies (4.6), d satisfies (4.7), and F is continuous.*

We established necessity above, and we prove sufficiency in the Appendix.

4.6 Related literature

[26] presented the MST postulate and attributed its formulation to [5, 6] and [16]. [5, 6] formulated restrictions on the distribution of tastes in a population of standard rational consumers. He showed the choice probabilities generated by consumers randomly drawn from the population satisfy MST if and only if those restrictions hold. [16] studied the

testable implications of his 'threshold model' [15] in a classical demand setting and showed that the model satisfies the slightly stronger MST+ postulate. As far as we are able to determine, Corollary 6 in Georgescu-Roegen [16, p. 161] is the first and only time that MST+ appeared in the literature before. Georgescu-Roegen [16, p. 160] also anticipated L. J. Savage's Example 15, explaining why it is natural to expect SST to be violated in a classical demand setting.

[20] proposed the definition (4.2) of a MUM, proved that every MUM satisfies MST, and left open the question of sufficiency. Our Theorem 18 answers the question posed by Halff by showing that, while MST is not sufficient for a choice rule to be a MUM, the slightly stronger MST+ condition is both necessary and sufficient.

The psychological foundations of the MUM formula can be traced back to [12] and [34]. The idea that the (dis)similarity of the alternatives should play the role captured by the distance metric in the MUM formula has been proposed in [37], [10], and [21]. A related literature highlights the non-metric nature of similarity judgements [see 36] and explores the role of similarity in intransitive choices [see 30].

The MUM characterized in Theorem 18 generalizes several nested models of stochastic binary choice in the literature, as shown in Figure 4.6. The most restrictive model in Figure 4.6 is the binary *Logit model* in which choice probabilities are given by the formula

$$\rho(x,y) = \frac{e^{u(x)}}{e^{u(x)} + e^{u(y)}} = \frac{1}{1 + e^{-[u(x)-u(y)]}} \tag{4.8}$$

for some utility function $u : Z \to \mathbb{R}$. [24] showed formula (4.8) is equivalent to the *product rule*

(PR) $\rho(x,y)\rho(y,z)\rho(z,x) = \rho(x,z)\rho(z,y)\rho(y,x)$

which can be interpreted as saying that the probability of observing a choice cycle in the direction $x \succ y \succ z \succ x$ is always equal to the probability of observing a choice cycle in

the opposite direction. [24] obtains this equivalence under the mild assumption of *positivity*, which requires that $\rho(x, y) > 0$ for all x, y.

A generalization of formula (4.8) is the *Fechnerian utility model* from psychophysics where

$$\rho(x, y) = F(u(x) - u(y)) \tag{4.9}$$

for some utility function $u : Z \to \mathbb{R}$ and some strictly increasing $F : \mathbb{R} \to (0, 1)$. The testable implications of formula (4.9) are well studied (see [9] and references therein). A result in [14] shows the Fechnerian formula is equivalent, under positivity, to the postulate of *acyclicity*. This postulate rules out cycles of the form $\rho(w^i, x^i) \geq \rho(y^i, z^i)$ for all $i = 1, \ldots, n$ with at least one strict inequality, whenever $\{w^i, x^i\} = \{y^{f(i)}, z^{f(i)}\}$ and $w^i = y^{g(i)}$ for some permutations $f, g : \{1, \ldots, n\} \to \{1, \ldots, n\}$.

Formula (4.9) can be further generalized to *simple scalability* [22] which requires

$$\rho(x, y) = F(u(x), u(y)) \tag{4.10}$$

for some utility function u and a real valued function F which is strictly increasing in the first argument and strictly decreasing in the second. [37] showed that simple scalability is equivalent to positivity and a slightly stronger version of SST:

(SST+) $\rho(x, y) \wedge \rho(y, z) \geq 1/2 \Longrightarrow \rho(x, z) \geq \rho(x, y) \vee \rho(y, z)$, and
$\rho(x, y) \wedge \rho(y, z) > 1/2 \Longrightarrow \rho(x, z) > \rho(x, y) \vee \rho(y, z)$

which, compared to the original SST postulate, imposes the additional requirement that a strict inequality in the hypothesis entails a strict inequality in the conclusion.

It can be seen immediately by inspecting the formulas that (4.8) $\Rightarrow$ (4.9) $\Rightarrow$ (4.10). To see that the simple scalability model (4.10) is nested in MUM, note that SST+ immediately implies MST+. The failure of the reverse implications is also easily seen by examples.

[9] considered an additional postulate, not shown in Figure 4.6, called the *quadruple condition*:

(QC) $\rho(w,x) \geq \rho(y,z)$ if and only if $\rho(w,y) \geq \rho(x,z)$

In a setting where Z is infinite, and under an additional stochastic continuity assumption, [9] showed that QC implies the Fechnerian utility model (4.9). It is also immediate from the definitions that a Fechnerian utility model (4.9) satisfies QC. When Z is finite, however, our next example shows that QC, while necessary, is not sufficient for ρ to be a Fechnerian utility model.

Example 24. *Let $Z = \{1,2,3,4,5\}$ and let ρ be a choice rule on Z with*

$$1 > \rho(5,1) > \rho(5,2) > \rho(4,1) > \rho(4,2) > \rho(3,1) >$$
$$\rho(5,3) > \rho(4,3) > \rho(3,2) > \rho(5,4) > \rho(2,1) > 1/2$$

Verifying that ρ satisfies QC is tedious but straightforward. This ρ does not admit a Fechnerian representation as in (4.9), since $\rho(5,4) > \rho(2,1)$ and $\rho(3,1) > \rho(5,3)$ would imply $u(5) - u(4) + u(3) - u(1) > u(2) - u(1) + u(5) - u(3)$ and the representation would require $\rho(3,2) > \rho(4,3)$, a contradiction.

QC is easily seen to imply SST+. Suppose $\rho(x,y) \wedge \rho(y,z) \geq 1/2$. Then QC and $\rho(y,z) \geq 1/2 = \rho(x,x)$ imply $\rho(y,x) \geq \rho(z,x)$ and hence $\rho(x,z) \geq \rho(x,y)$. Also, QC and $\rho(x,y) \geq 1/2 = \rho(z,z)$ imply $\rho(x,z) \geq \rho(y,z)$ and hence $\rho(x,z) \geq \rho(x,y) \vee \rho(y,z)$. The same argument with strict inequalities in the hypothesis implies a strict inequality in the conclusion and SST+ obtains. The converse implication fails, as our next example shows.

Example 25. *Let $Z = \{1,2,3,4\}$ and let ρ be a choice rule on Z with*

$$1 > \rho(4,1) > \rho(4,2) > \rho(3,1) > \rho(2,1) > \rho(4,3) > \rho(3,2) > 1/2$$

which is clearly satisfies SST+ but fails QC, since $\rho(4,2) > \rho(3,1)$ *but* $\rho(4,3) < \rho(2,1)$.

As Figure 4.6 shows, the MUM formula subsumes several models in the literature as special cases. In the other direction, note that by relaxing the triangle inequality property of the metric d in MUM, one obtains a more general model that is equivalent to WST. Hence, the empirical bite of the triangle inequality property of d in the MUM is exactly equal to the gap between WST and MST.

Several familiar discrete choice models used to address violations of SST in the literature are particular instances of MUM. Example 17 shows the classic multinomial probit is a MUM. The Bayesian probit model [28] restricted to binary choice is equivalent to a probit model, and therefore a MUM. Another example, below, is the elimination-by-aspects model proposed by [35].

Example 26 (Tversky's EBA)**.** *The choice rule* ρ *on a finite* Z *is an* elimination-by-aspects (EBA) *rule if there exist a mapping* A *that takes each option* $x \in Z$ *to a set of aspects* $A(x)$ *that* x *possesses, and a measure* m *over the set of all aspects such that*

$$\rho(x,y) = \frac{m\left[A(x)\right] - m\left[A(y)\right]}{m\left[A(x) \setminus A(y)\right] + m\left[A(y) \setminus A(x)\right]}.$$

Every EBA is a MUM with $u(x) = m[A(x)]$ *for all* $x \in Z$, $d(x,y) = m\left[A(x) \setminus A(y)\right] + m\left[A(y) \setminus A(x)\right]$ *and* F *given by the strictly increasing function* $F(t) = 1/2 + t/2$.

Probit and EBA are also instances of the *random utility model* (RUM). A choice rule ρ on a finite Z is a RUM if there exists a probability measure μ over the strict orderings on Z such that $\rho(x,y)$ equals the probability under μ of the event in which x beats y. [3] and [11] characterize the set of RUMs in an abstract setting of choice options when choice data for all finite menus is available. A review of the literature that tackles the characterization of binary choice RUMs is provided by [13]. Example 17 shows the MUM and RUM families have a non-empty intersection. Next, we show that neither MUM nor RUM nest each other.

Example 27. *We slightly modify an example given in [8] to obtain a choice rule that is a MUM but not a RUM. Let* $Z = \{1, 2, 3, 4, 5, 6\}$*, let* $0 < \varepsilon < 3/46$ *and let the choice rule* ρ *on* Z *be given by*

$$\begin{aligned}
\rho(4,5) = \rho(4,6) = \rho(2,5) = \rho(2,3) = \rho(1,6) = \rho(1,3) &= 1 - \varepsilon \\
\rho(2,6) = \rho(1,5) &= \frac{1}{2} + \varepsilon \\
\rho(2,4) = \rho(1,4) = \rho(3,5) = \rho(3,6) &= \frac{1}{2} + \frac{\varepsilon}{2} \\
\rho(3,4) = \rho(1,2) = \rho(5,6) &= \frac{1}{2} + \frac{\varepsilon}{3}
\end{aligned}$$

It is straightforward to verify that ρ *satisfies MST+. Now suppose* ρ *is a RUM generated by the probability* μ *on the set of strict orderings over* Z*. Since* $\rho(2,3) = \rho(4,6) = 1 - \varepsilon$*, the probability of the event* $\{2 \succ 3\} \cap \{4 \succ 6\}$ *is larger or equal to* $1 - 2\varepsilon$*. By transitivity, the event* $\{2 \succ 3\} \cap \{3 \succ 4\} \cap \{4 \succ 6\}$ *is contained in the event* $\{2 \succ 6\}$*. Hence the event* $\{3 \succ 4\} \cap \{6 \succ 2\}$ *has at most probability* 2ε*. By the same reasoning,* $\{3 \succ 4\} \cap \{5 \succ 1\}$ *has at most probability* 2ε*. And likewise* $\{6 \succ 2\} \cap \{5 \succ 1\}$ *has at most probability* 2ε*. Since* μ *is a probability, this implies* $\rho(3,4) + \rho(5,1) + \rho(6,2) \leq 1 + 6\varepsilon$*. But instead we have* $\rho(3,4) + \rho(5,1) + \rho(6,2) = 3/2 - 5\varepsilon/3 > 1 + 6\varepsilon$ *and therefore* ρ *cannot be a RUM.*

A converse example based on the well-known Condorcet paradox shows that RUM models can violate MST+. Let μ assign equal probability to three strict orderings $x \succ y \succ z$, $y \succ z \succ x$ and $z \succ x \succ y$ over the options x, y and z. Then the binary choice rule ρ generated by μ has $\rho(x, y) = \rho(y, z) = \rho(z, x) = 2/3$ which violates WST, and therefore also violates MST+. Similarly, some recent models proposed in the random choice literature including the random consideration set rule [25], the attribute rule [18], the single-crossing random utility rule [2], the deliberately stochastic choice rule [4] and the focus-then-compare procedure [29] can be easily verified to violate WST and therefore, their binary choice restrictions are not

nested by MUM.

In the richer set of lotteries, [17] introduce the random expected utility model: they characterize the choice behavior of an agent who is an expected utility maximizer with a stochastic Bernoulli index. In our Theorem 21, we impose the same linearity and continuity assumptions employed by [17], specialized to the binary choice domain. Let REM denote any binary choice rule that is obtained as the restriction of a random expected utility model to the domain of binary menus. The relationship between our MEM and REM mirrors the relationship between MUM and RUM: neither model nests the other.

First, it is straightforward to construct a REM that violates WST in the same spirit as the Condorcet paradox example above. By Theorem 21, this behavior cannot be accounted for by a MEM. On the other hand, consider the MUM in Example 27. The proof of Theorem 18 provides a constructive proof to obtain a MUM representation (u, d, F) for the ρ in Example 27. It is a straightforward exercise to verify that the metric d constructed in this manner satisfies condition (4.7). By Proposition 23, this ρ can be extended to a MEM. This MEM is obviously not a REM, for otherwise, its restriction to Z would be a RUM.

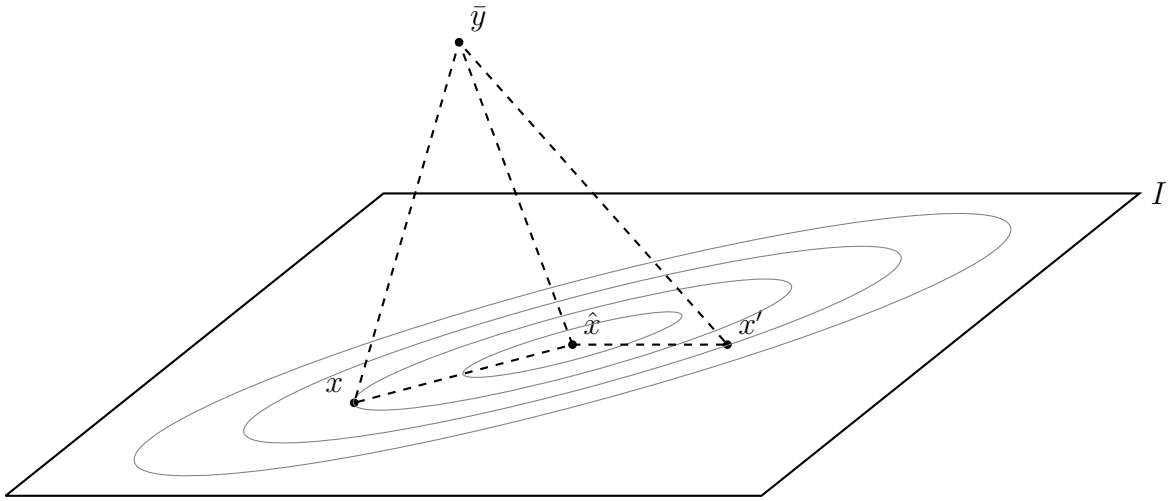


Figure 4.3: Illustration of the construction of the norm in the proof of Theorem 21.

All lotteries in the affine subspace I are stochastically indifferent to $\hat{x}$. Every lottery x in I is chosen with probability strictly larger than $1/2$ against $\bar{y}$. The maximum choice probability $\rho(\cdot, \bar{y})$ is obtained at lottery $\hat{x}$. The depicted contour sets in I given by $\{z \in I : \rho(z, \bar{y}) \geq \alpha\}$ are concentric ellipsoids centered at $\hat{x}$. We take one of these ellipsoids to be the unit ball that defines the norm on the $n-2$ dimensional subspace parallel to I.

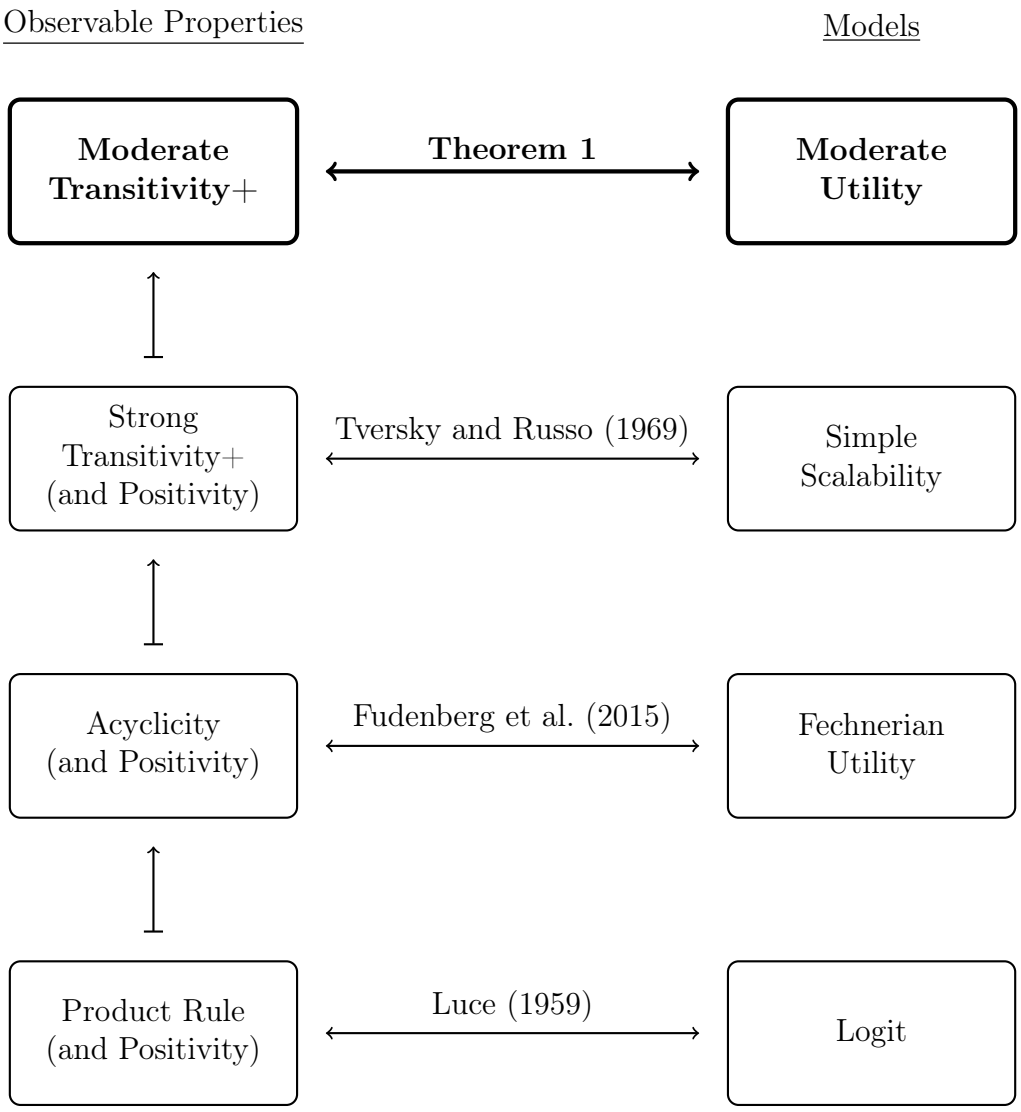


Figure 4.4: Relationship between models and postulates

Relationship between models and postulates on choice probabilities for binary stochastic choice over a finite set of options. A double arrow ($\leftrightarrow$) indicates equivalence while an arrow ($\mapsto$) indicates implication in the direction of the arrow and failure of implication in the opposite direction.

Bibliography

[1] **Apesteguia, Jose and Miguel A Ballester**, "Monotone stochastic choice models: The case of risk and time preferences," *Journal of Political Economy*, 2018, *126* (1), 74–106.

[2] **Apesteguia, Balleste, and Jay Lu**, "Single-Crossing Random Utility Models," *Econometrica*, 2017, *85* (2), 661–674.

[3] **Block, H.D. and Jacob Marschak**, "Random Orderings and Stochastic Theories of Response," Cowles Foundation Discussion Papers 66, Cowles Foundation for Research in Economics, Yale University 1959.

[4] **Cerreia-Vioglio, Simone, David Dillenberger, Pietro Ortoleva, and Gil Riella**, "Deliberately stochastic," *Working Paper*, 2017.

[5] **Chipman, John S.**, "Stochastic Choice and Subjective-Probability," *Econometrica*, 1958, *26* (4), 613–613.

[6] **Chipman, John S**, "Stochastic choice and subjective probability," in Dorothy Willner, ed., *Decisions, Values, and Groups*, Vol. I, Pergamon Press, 1960, pp. 70–95.

[7] **Davidson, Donald and Jacob Marschak**, "Experimental tests of a stochastic decision theory," *Measurement: Definitions and theories*, 1959, *17*, 274.

[8] **de Souza, Fernando Menezes Campello**, "Mixed models, random utilities, and the triangle inequality," *Journal of Mathematical Psychology*, 1983, *27* (2), 183 – 200.

[9] **Debreu, Gerard**, "Stochastic Choice and Cardinal Utility," *Econometrica*, 1958, *26* (3), 440–444.

[10] **Domencich, Thomas A. and Daniel McFadden**, *Urban travel demand: a behavioral analysis*, Vol. 93 of *Contributions to Economic Analysis*, North-Holland, 1975.

[11] **Falmagne, J. C.**, "A representation theorem for finite random scale systems," *Journal of Mathematical Psychology*, 1978, *18* (1), 52 – 72.

[12] **Fechner, Gustav Theodor**, *Elemente der Psychophysik*, Leipzig: Breitkopf & Hartel, 1859.

[13] **Fishburn, Peter C**, "Induced binary probabilities and the linear ordering polytope: A status report," *Mathematical Social Sciences*, 1992, *23* (1), 67–80.

[14] **Fudenberg, Drew, Ryota Iijima, and Tomasz Strzalecki**, "Stochastic Choice and Revealed Perturbed Utility," *Econometrica*, November 2015, *83* (6), 2371–2409.

[15] **Georgescu-Roegen, Nicholas**, "The pure theory of consumers behavior," *The Quarterly Journal of Economics*, 1936, *50* (4), 545–593.

[16] **Georgescu-Roegen**, "Threshold in Choice and the Theory of Demand," *Econometrica*, 1958, pp. 157–168.

[17] **Gul, Faruk and Wolfgang Pesendorfer**, "Random Expected Utility," *Econometrica*, January 2006, *74* (1), 121–146.

[18] **Gul, Faruk, Paulo Natenzon, and Wolfgang Pesendorfer**, "Random Choice as Behavioral Optimization," *Econometrica*, September 2014, *82* (5), 1873–1912.

[19] **Gurari, N.I. and Y.I. Sozonov**, "On normed spaces which have no bias of the unit sphere," *Math. Notes*, 1970, *7*, 187–189.

[20] **Halff, Henry M**, "Choice theories for differentially comparable alternatives," *Journal of Mathematical Psychology*, 1976, *14* (3), 244–246.

[21] **Hausman, Jerry A. and David A. Wise**, "A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences," *Econometrica*, 1978, *46* (2), pp. 403–426.

[22] **Krantz, David Harold**, "The scaling of small and large color differences." PhD dissertation, University of Pennsylvania 1964.

[23] **Lea, Amanda M and Michael J Ryan**, "Irrationality in mate choice revealed by túngara frogs," *Science*, 2015, *349* (6251), 964–966.

[24] **Luce, R. Duncan**, *Individual Choice Behavior: a Theoretical Analysis*, Wiley New York, 1959.

[25] **Manzini, Paola and Marco Mariotti**, "Stochastic Choice and Consideration Sets," *Econometrica*, May 2014, *82* (3), 1153–1176.

[26] **Marschak, Jacob**, "Binary-choice constraints and random utility indicators," in K. J. Arrow, S. Karlin, and P. Suppes, eds., *Mathematical methods in the social sciences*, Stanford University Press 1960, pp. 312–329.

[27] **Mellers, Barbara A, Shi jie Chang, Michael H Birnbaum, and Lisa D Ordonez**, "Preferences, prices, and ratings in risky decision making.," *Journal of Experimental Psychology: Human Perception and Performance*, 1992, *18* (2), 347.

[28] **Natenzon, Paulo**, "Random Choice and Learning," *Journal of Political Economy*, February 2019, *1* (127), 419–457.

[29] **Ravid, Doron and Kai Steverson**, "Focus, Then Compare," *SSRN Working Paper*, March 2019.

[30] **Rubinstein, Ariel**, "Similarity and decision-making under risk (Is there a utility theory resolution to the Allais paradox?)," *Journal of economic theory*, 1988, *46* (1), 145–153.

[31] **Schoenberg, Isaac J**, "Remarks to Maurice Frechet's Article"Sur La Definition Axiomatique D'Une Classe D'Espace Distances Vectoriellement Applicable Sur L'Espace De Hilbert," *Annals of Mathematics*, 1935, pp. 724–732.

[32] **Soltani, Alireza, Benedetto De Martino, and Colin Camerer**, "A Range-Normalization Model of Context-Dependent Choice: A New Model and Evidence," *PLoS computational biology*, 2012, *8* (7), e1002607.

[33] **Thompson, Anthony C**, *Minkowski geometry*, Cambridge University Press, 1996.

[34] **Thurstone, L. L.**, "A law of comparative judgment," *Psychological Review*, 1927, *34* (4), 273.

[35] **Tversky, A.**, "Choice by elimination," *Journal of Mathematical Psychology*, 1972, *9* (4), 341–367.

[36] **Tversky, Amos**, "Features of similarity.," *Psychological review*, 1977, *84* (4), 327.

[37] **Tversky A. and J. Edward Russo**, "Substitutability and similarity in binary choices," *Journal of Mathematical Psychology*, 1969, *6* (1), 1–12.

Chapter 5

Rational Contextual Choices under Imperfect Perception of Attributes

5.1 Introduction

Classically, rationality is defined through consistency axioms. Under consistency, rational preferences are transitive and independent of contexts, usually representable by utility functions. However, empirical research has long found violations of different aspects of consistency. For example, intransitivity was spotted as early as in [42] and more recently evidence is discussed in [33]. Other research suggesting contextual dependence includes [18], [31] and [16]. Here, by context dependence or contextual choices we refer the following type of intuitive observations. In different choice problems involving objects $\mathbf{x}$ or $\mathbf{y}$, the choice probabilities of $\mathbf{x}$ and of $\mathbf{y}$ differ in such a way that suggests the decision maker evaluates the objects differently. For instance, in [18] and [31], experimenters offer the subjects two choice problems. One involves only two options $\mathbf{x}$ and $\mathbf{y}$ and the other includes a third choice $\mathbf{z}$. They find that the inclusion of $\mathbf{z}$ can reverse the relative frequency between choosing $\mathbf{x}$ and $\mathbf{y}$, even though $\mathbf{z}$ itself is rarely chosen (attraction effect) or listed as unavailable (phantom

decoy effect). Another example is joint-separate valuation reversal (hence forth j-s reversal) in [16]. It is found that when the willingness to pay for each of $\mathbf{x}$ and $\mathbf{y}$ is elicited separately, $\mathbf{x}$ can be valued higher than $\mathbf{y}$, but when elicited together, $\mathbf{x}$ becomes inferior to $\mathbf{y}$. Intransitive choices can also be interpreted as a type of contextual dependence.

Within the rational choice literature, some contextual choice effects, such as the similarity effect ([44]), attraction effect ([18]) and compromise effect ([37]), can be explained (see e.g. [13], [20], [12], [26]). However, other contextual choice effects such as the phantom decoy effect ([31]), j-s reversal ([16]) and stochastic intransitivity ([42]) have not yet been explained in a classically rational framework.[1]

Our paper proposes a rational choice model that systematically predicts the aforementioned experimental findings. With a novel informational friction, our model generically exhibits both the stochastic intransitivity and the j-s reversal when there are trade-offs between attributes in the options, as is observed in data. Under a quite general noise structure and without assuming any parametric utility functions, our model predicts the *decoy choice pattern*, a comparative static that captures the attraction effect, the phantom decoy effect and the compromise effect.[2] Our model is relatively rigid because it does not explain the phenomena through calibrating free parameters. Instead, the phenomena are generic to the model. For example, no parametrization of the model accomodates the opposite of the compromise effect.[3] Apart from capturing these contextual choices, our general model also predicts that classical rational choice holds for a family of choice problems where there is no trade-off between attributes among the alternatives. Hence we also identify a subclass of

[1]The term "phantom decoy effect" is used differently here than in [26]. Here, it refers to a situation where an unavailable third option that is asymmetrically dominating, i.e. better than the target in all attributes but worse than the competitor in some attributes, *increases* the attractiveness of the target. Such results are found in [31] and later in [15], [28], [29] and [14] etc.

[2]In our model, these three different effects share the same underlying mechanism. This is similar in vein to the suggestion in [15], that the attraction effect and the phantom decoy effect may have the same cause.

[3]The compromise effect ([37]) finds that the option with moderate attribute levels becomes more popular when an alternative with extreme attributes is introduced. The "reversed phenomenon", not permitted by our model, would be when the moderate option becomes less popular instead.

choice problems where violations of the classical model does not occur.

Our novel information friction is that the decision maker faces a systematic noise in the perception of attributes. Each option $\mathbf{x}$ has precise attribute levels $\mathbf{x}^*$ over which the agent's utility function is defined. However the agent cannot observe these precise attributes, but a noisy signal $X|\mathbf{x}^*$. The noise is systematic in the sense that, conditional on the true attributes, the noisy signals across different alternatives are correlated. Therefore, although the distribution of $X|\mathbf{x}^*$ is unchanged, the agent makes different inference about $\mathbf{x}^*$ when she is presented with different alternatives. For example, in the choice problem $\{\mathbf{x}, \mathbf{y}\}$, the agent observes the signals X, Y but not the actual attribute levels $\mathbf{x}^*$ and $\mathbf{y}^*$. She forms a posterior belief, say about $\mathbf{x}^*$, conditional on the signals X, Y. When she faces the choice problem $\{\mathbf{x}, \mathbf{z}\}$, the posterior belief about $\mathbf{x}^*$ is conditional on X, Z. These two posterior beliefs about $\mathbf{x}^*$ are generally different, and so are the posterior expected utilities of $\mathbf{x}$ in $\{\mathbf{x}, \mathbf{y}\}$ and in $\{\mathbf{x}, \mathbf{z}\}$. Then intuitively, even if the agent is (stochastically) indifferent both in the choice problem $\{\mathbf{x}, \mathbf{y}\}$ and in the problem $\{\mathbf{x}, \mathbf{z}\}$, she would *not* be indifferent about $\{\mathbf{y}, \mathbf{z}\}$. In otherwords, the (stochastic) indifference curves can cross. Intransitivity is then a consequence of crossing indifference curves.

For the rest of the paper, we impose a specific type of correlated noise termed *imperfect perception of attributes*. Namely, the noisy signal is *common* across objects but may be idiosyncratic across attributes. Precisely, there is an attribute-specific error term (i.e. common across items) perturbing the perceived attribute levels of each item while keeping the relative differences unchanged. Under this noisy signal, if the agent over-perceives an attribute in an object, she over-perceives the same attribute in other objects. Although this error term directly induces a change in the utility levels, it is not equivalent to adding a common error term directly to the utilities. A common shock to the utilities will result in the classical rational choice model and no contextual choice occurs. In fact, our Proposition 30 shows that our model does not satisfies monotonicity, and hence cannot be interpreted

as any random utility model. A detailed discussion about the general framework and the empirical motivation of the information friction is provided in section 2.

The assumption of imperfect perception is intuitively sensible because correlated signal arises easily in perception tasks. Imagine in choosing apartments, an agent is looking for one with abundant natural light. She visits two apartments on the same day, and sees that apartment $\mathbf{x}$ is brighter than $\mathbf{y}$. Although the agent does not know how bright the apartments typically are (i.e. she does not observe $\mathbf{x}^*, \mathbf{y}^*$), she can use the visits as noisy signals for comparison. The noisy signals may not be accurate about the typical brightness in the apartments, but their difference can clearly indicates which room is typically brighter. Afterall, the agent is seeing both apartments at roughly the same time, under the same weather. There is naturally a common component in the noisy signals. The same intuition holds in perceiving other attributes such as noisiness of the neighborhood, length of commuting time etc. Such uncertainty in perception can also arise when the agent is learning about attributes *measured* in scientific units. The technical units can be difficult to interpret precisely, and over (under) interpreting a unit can lead to over (under) perceived attribute levels among the alternatives. In general, imperfect perception arises whenever the agent *believes* that there can be a common component in the uncertain perception of attributes. We will elaborate more on this assumption in section 2.

This imperfect perception causes a *contrast effect* in each attribute in the perception of a Bayesian agent. The contrast effect is a well-known psychological phenomenon that refers to the strengthening or weakening of the perception about any attribute when the object is contrasted with surrounding objects of different levels in the same attribute.[4] To illustrate with the apartment example, suppose that the agent on the same day also visited another apartment $\mathbf{z}$ that is much brighter than both $\mathbf{x}$ and $\mathbf{y}$. The Bayesian agent infers it is unlikely for any apartment to be so bright on every day, implying an upward bias in the

[4]See e.g. [36] and [30] page 38 - 41.

common component of all the signals. Hence after visiting the apartment $\mathbf{z}$, the agent revises downwards the perceived brightness of $\mathbf{x}$ and $\mathbf{y}$. The judgements about other attributes of the apartments can also be affected similarly. For example, an apartment can be perceived as quieter in the presence of a really noisy one.

When the agent's preference is determined by a single attribute monotonically, this contrast effect is inconsequential: she always chooses to maximize (or minimize) that attribute in the model.[5] However, if her preference involves at least two attributes, a different set of competing alternatives can simultaneously affect the perception of two attributes differently (increase one and decrease another). Hence the same two options can have different posterior utility when contrasted with different sets of alternatives.

The compromise effect is one such example. Suppose in choosing apartments, the agent faces a trade-off between natural lighting and quietness. She prefers better lighting as well as a quieter living place. As before, she observes correlated signals X and Y in both attributes of the two apartments $\{\mathbf{x}, \mathbf{y}\}$. Suppose $\mathbf{x}$ has good natural lighting but some sound of cars from the street can be heard, whereas $\mathbf{y}$ has a gloomy interior but it is very quiet. Suppose the agent is inclined to choose $\mathbf{y}$ between the two. Now introduce a third option $\mathbf{z}$ that is even brighter than $\mathbf{x}$ but is also much noisier. As explained previously, conditional on X, Y *and* Z, the (posterior) perceived brightness levels for both $\mathbf{x}$ and $\mathbf{y}$ are lower than those conditional on only X and Y (the contrast effect in perception of light). And similarly, with the additional signal Z the (posterior) perceived quietness for both $\mathbf{x}$ and $\mathbf{y}$ also increase. Now, reducing the perceived brightness of $\mathbf{x}$ and $\mathbf{y}$ affects both apartments negatively, but more so for $\mathbf{y}$ because of diminishing marginal utility in lighting. And increasing the perceived quietness of $\mathbf{x}$ and $\mathbf{y}$ affects both apartments positively, but more so for $\mathbf{x}$, due to the diminishing marginal utility in quietness. Consequently, $\mathbf{x}$ has a higher expected utility level relative to $\mathbf{y}$ after $\mathbf{z}$ is introduced.

[5] I.e, if the agent only cares about lighting, she always chooses the brightest apartment with certainty.

Besides the assumption of imperfect perception, Bayesian rationality is also an important component of our model. If there is no updating at all, presenting the alternative $\mathbf{z}$ will not affect the preference between $\mathbf{x}$ and $\mathbf{y}$. We use Bayesian updating because it is the rational benchmark in modeling information and learning. Despite the reliance of our model on Bayesian rationality, we do not claim that in reality people perform sophisticated Bayesian updating and calculates posterior expectations. Instead, we interpret the model as an as-if representation of the decision process. Nonetheless, the analysis of this as-if channel does parallel some intuitive explanations of contextual choices as illustrated above.

We organize the paper as below. The next section presents the general set up. In section 3, we apply a parametric special case of the model to explain intransitive choices, j-s reversal, and the compromise effect in detail. The analysis of the general model is presented in section 4, where the decoy choice pattern and the choice problems for which the agent exhibit classical rational choice are studied. Section 5 contains some further discussion. All proofs are contained in the appendix.

5.1.1 Related Literature

This paper contributes to the literature on rationalizing contextual choices by proposing a new and more disciplined informational channel that complements exisiting explanations. [20] studies a consumer-retailer game where the set of alternatives conveys information in equilibrium. In contrast, the our model focuses on a pure single agent decision environment when market interaction is not of major concern. Our paper is closer related to [12] and [26] in this sense, but the information structures differ. [12] assumes that the choice contexts do not provide different information. But because the incentive to acquire information depends on contexts, the agent eventually uses different (acquired) information in decision making. Different from [12], we do not study a model of information acquisition. Instead, we show how learning under a relatively general family of exogeneous information structure can predict

different contextual effects. However, because we do not model the information acquisition process, our channel is less relevant when the acquisition is of main consideration in the problems such as choice overload. [26] studies a transitive model where Bayesian updating is applied to the probit model ([13]). Different in nature, the uncertainty in our model lies in the more premitive attribute space. As a result, we explain different contextual choices, such as intransitivity, j-s reversal and the phantom decoy effects.

5.2 The Model, its Assumptions and Motivations

In the empirical research on contextual choices, choice problems consists of several options, each with a description in *two or more* different attributes. Therefore, we take the primitives of our model to be the attributes of each object. In particular, we use $\mathbb{R}^n$ for $n \geq 2$ to represent the attribute space. The attributes of each item $\mathbf{x}$ is represented as a vector $\mathbf{x}^* := (x_1^*, \ldots, x_n^*)$ in the space, with each coordinate given by the corresponding attribute level. The vector $\mathbf{x}^*$ is not directly observed by the agent. In many of the experiments, contextual choices are observed as long as there are two different attributes. Therefore we restrict our discussion to $\mathbb{R}^2$ in this paper for mathematical simplicity.[6]

In accordance with the classical theory, the agent is assumed to be rational in two senses. Firstly, she has a fixed preference over the attribute space that can be represented by a vNM utility function $u : \mathbb{R}^2 \to \mathbb{R}$. Following classical consumer theory, we assume that the utility function is monotonic over $\mathbb{R}^2$, and the marginal utilities are decreasing. In other words, the two attributes are both goods so that the utility function displays insatiability along each axis. Other standard assumptions from consumer theory include diminishing returns and weak complementarity between attributes. We call a preference *standard* if it displays these properties.

[6]The mechanisms for the main theorems can be extended to higher attribute dimensions.

Assumption 1 (Standard Preference)**.** *The decision maker's preference over distributions on* $\mathbb{R}^2$ *can be represented by a vNM utility function* $u : \mathbb{R}^2 \to \mathbb{R}$ *that is differentiable, increasing (i.e.* $u_1 > 0, u_2 > 0$*), and exhibits decreasing marginal sensitivity (i.e.* $u_{11} < 0, u_{22} < 0$*) and weak complementarity (i.e.* $u_{12} \geq 0$ *). Any utility function representing a standard preference is called a standard utility function.*

Secondly, she is Bayesian rational with a prior belief over $\mathbb{R}^2$. The prior distribution represents the agent's anticipation about the attribute levels before she observes any choice alternatives. We endow the agent with a normal prior distribution. Without loss of generality, we can translate and scale the attribute space and let the prior mean be the origin and the prior variance be $\Omega := \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$ for some correlation coefficient $r \in (-1, 1)$.[7] We adopt the conventional definition of Bayesian rationality, that is when the agent observes noisy signals, she chooses the object that maximizes posterior expected utility conditional on the signal.[8]

Assumption 2 (Normal-Bayesian Rationality)**.** *The decision maker is Bayesian with a normal prior* $\mathcal{N}(0, \Omega)$ *and maximizes posterior expected utility.*

Next, our main assumption proposes that there is noise in the perception of each attribute. The noise is only specific to the attribute's perception, hence is common across alternatives. We use capital letters (i.e. $X = (X_1, X_2)$) to denote the noisy signal of each object's attribute location. For example, in the choice set $\{\mathbf{x}, \mathbf{y}\}$, the attribute levels $\mathbf{x}^*, \mathbf{y}^*$ are perceived as $X = \mathbf{x}^* + \epsilon$ and $Y = \mathbf{y}^* + \epsilon$ with the same vector ϵ.[9] This implies that it is easier to perceive relative differences in attributes among the items, i.e. $X - Y = \mathbf{x}^* - \mathbf{y}^*$, but more difficult

[7] Such a correlation can arise when, for example, the two attributes are price and quality. One can interpret $r < 0$ as the agent having a prior belief that a good price is associated with low quality.

[8] See e.g. [35].

[9] Our model predictions only change marginally if we relax the assumption so that $X = \mathbf{x}^* + \epsilon + \epsilon_{\mathbf{x}}$ where $\epsilon_{\mathbf{x}}$ is a small i.i.d. noise for each object $\mathbf{x}$.

to perceive the absolute locations $\mathbf{x}^*$ and $\mathbf{y}^*$ in the attribute space. One way to interpret of this assumption is that the noise is a random anchoring effect for in each attribute. As a consequence, the agent over-perceive (or under perceive) the same attribute in every alternatives. Such a form of noisy perception is supported in experimental findings in [2], where they found participants underestimated the same attribute of several different objects (or overestimated all of them) if they were "anchored". As is summarized in their paper "we show that consumers' absolute valuation of experience goods is surprisingly arbitrary ... we also show that consumers' relative valuations of different amounts of the good appear orderly."

In some experiments, the attributes of each choice object are *measured in technical units and described numerically.* Nonetheless, there are experimental evidence suggesting that *even* when these descriptions are displayed in scientific units, the subjects are *not* able to perceive the numerical information precisely. For example, [2] finds that reading the volume of noise in units does not provide more information about the loudness than actually hearing the noise. It is intuitive to see that technical units of measurements can be difficult to interpret. And due to such difficulty, the precise measurements can only serve as noisy indicators of the attribute levels. For example, [20] argues that in choosing a personal computer, a decision maker usually cannot evaluate precisely in utils a given set of measurements in megahertz, gigabytes or other technical units. In general, depending on the decision maker's intuitive understanding of the technical units, the agent may under or over perceive the attribute levels even if they are described numerically. To further illustrate this point, in our apartments choice example, suppose the agent is also concerned with the safety of the respective neighborhoods. She can obtain a signal of this attribute by consulting the last-year crime statistics published by the same authority.[10] Even though for each neighborhood, this attribute is measured in simple units as "number of crimes per year per ten thousand

[10]E.g. local police department and city websites, or the *Uniform Crime Reports* by FBI in US.

people", it is still a noisy signal in the perspective of the agent, because for example, it is not clear how strict the definition of crime is in this context. The signal can be an exageration (understatement) for all neighborhoods if the local authority applies a broader (narrower) definition of crime than the agent understands. In general, our modeling of the noisy signal can arise *as long as the agent thinks* there can be uncertainty in her understanding of the unit of measurements. Formally, our *imperfect perception* assumption is as follows:

Assumption 3 (Imperfect Perception)**.** *For any n alternatives $\{\mathbf{x}^1, \ldots, \mathbf{x}^n\}$ each with attributes $\mathbf{x}^{1*}, \ldots, \mathbf{x}^{n*} \in \mathbb{R}^2$, the agent receives signals $X^1, \ldots, X^n$ where $X^i - \mathbf{x}^{i*} = \epsilon$ for all i. The noise term $\epsilon \sim \mathcal{N}(0, T^{-1})$ is normal with variance matrix*

$$T^{-1} = \begin{bmatrix} 1/t_1^2 & R/(t_1 t_2) \\ R/(t_1 t_2) & 1/t_2^2 \end{bmatrix} \text{ for some } \frac{1}{t_1^2} + \frac{1}{t_2^2} > 0, \text{ and some } R \in (-1, 1).$$

The usual assumption of normal noise is also conjugate to the normal prior. The specification that ϵ is common for all i can be expressed as a noise that is *perfectly* correlated across objects. If we relax the perfect correlation to high correlations, the change in model prediction is only marginal because the choice probability is continuous in the noise covariances. We allow the standard deviations in two attributes to differ as long as one of them is strictly positive (i.e. $\frac{1}{t_1^2} + \frac{1}{t_2^2} > 0$) even though one of them can be zero (e.g. $t_1 = \infty$). The assumption also allows the noise across attributes to have a non-zero correlation in R.[11]

We now summarize the notation used in the paper. Different letters denote different alternatives. Letters with an asterisk denote the true attribute levels of an object in $\mathbb{R}^2$. When there are more than 3 alternatives the superscripts (i.e. $\mathbf{x}^{1*}, \mathbf{x}^{2*}, \ldots$ etc.) are used. Capital letters denote the initial noisy perception by the agent. Calligraphic letters (i.e. $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$) are reserved for the agent's posterior belief about the true attributes. Subscripts distinguish

[11] Such a correlation can arise when attributes are closely related, such as the sugar content and calories in a soft drink, one might expect a correlation in the noise across these attributes.

the respective attribute-dimensions for a given vector. Choice behavior is a function that specifies the choice probability of an object when it is presented in a set of alternatives for which a subset is not available. We use the notation $C(\mathbf{x}, \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^i, (\mathbf{x}^{i+1}, \ldots, \mathbf{x}^{i+j})\})$ to denote the choice probability of x from the set $\{\mathbf{x}^1 \ldots \mathbf{x}^{i+j}\}$ in which $\{\mathbf{x}^{i+1}, \ldots, \mathbf{x}^{i+j}\}$ are unavailable. A $C(.,.)$ that assigns a probability for any $\mathbf{x}$ in every nonempty finite set of alternatives S, with any $S' \subsetneq S$ specifying the unavailable objects, is called the *choice behavior* of an agent. The choice behavior satisfies

$$\sum_{k=1}^{i} C(\mathbf{x}^k, \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^i, (\mathbf{x}^{i+1}, \ldots, \mathbf{x}^{i+j})\}) = 1.$$

5.3 A Parametric Special Case

In this section, we illustrate the stochastic intransitivity, the j-s reversal and the compromise effect with the following parametric settings. We take the simple exponential utility function $u : \mathbb{R}^2 \to \mathbb{R}$

$$u(x_1, x_2) = -e^{-3x_1} - e^{-3x_2}.$$

For the noise structure, we consider the simple case that the first attribute is perfectly perceived, and there is noise only in the perception of the second attribute. In other words, the noise has no variance in the first attribute,

$$\epsilon \sim \mathcal{N}\left(0, \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}\right).$$

For simplicity, we take the agent's prior distribution be the standard bivariate normal centered at the origin.

5.3.1 Violation of Weak Stochastic Transitivity

Weak stochastic transitivity refers to the proposition that if $C(\mathbf{x}, \{\mathbf{x}, \mathbf{y}\}) > 0.5$ and $C(\mathbf{y}, \{\mathbf{y}, \mathbf{z}\}) > 0.5$, then $C(\mathbf{x}, \{\mathbf{x}, \mathbf{z}\}) > 0.5$. Early evidence of intransitive choices can be found in [42], and more recently [33]. Some evidence discussed in the two papers suggests that weak transitivity can be violated when there is no clear domination among $\mathbf{x}, \mathbf{y}, \mathbf{z}$. In this subsection, a decision maker is said to *display intransitivity* if there are $\mathbf{x}, \mathbf{y}, \mathbf{z}$ such that the choice behavior C satisfies $C(\mathbf{x}, \{\mathbf{x}, \mathbf{y}\}) > 0.5$ and $C(\mathbf{y}, \{\mathbf{y}, \mathbf{z}\}) > 0.5$, and $C(\mathbf{z}, \{\mathbf{x}, \mathbf{z}\}) > 0.5$. In our model, intransitivity is a consequence of crossing stochastic indifference curves.

Due to the randomness ϵ in the information, the choice between any two objects $\mathbf{x}$ and $\mathbf{y}$ depends on their fixed attribute levels $\mathbf{x}^*, \mathbf{y}^*$ and the realization of ϵ. Hence given the attribute levels, we can determine the probability of choice, i.e. $C(\mathbf{x}, \{\mathbf{x}, \mathbf{y}\})$, from the distribution of ϵ. We say $\mathbf{x}$ is *stochastically indifferent* to $\mathbf{y}$ (writes $\mathbf{x} \sim \mathbf{y}$) if

$$C(\mathbf{x}, \{\mathbf{x}, \mathbf{y}\}) = 0.5.$$

Similarly, the *stochastic indifference curve* of $\mathbf{x}$ is the set of alternatives that that are stochastically indifferent to $\mathbf{x}$. On the space of attributes, this set of alternatives corresponds to the following set of attributes $\{\mathbf{y}^* \in \mathbb{R}^2 | \mathbf{x} \sim \mathbf{y}\}$.

Let us consider two alternatives $\mathbf{x}, \mathbf{y}$ such that $x_1^* > y_1^*$ and $y_2^* > x_2^*$. When is $\mathbf{x}$ chosen over $\mathbf{y}$? Since the agent is Bayesian rational, she chooses $\mathbf{x}$ whenever the posterior expected utility of $\mathbf{x}$ is greater than that of $\mathbf{y}$. Under the notation, the posterior beliefs about $\mathbf{x}^*$ and $\mathbf{y}^*$ are respectively the random variables $\mathcal{X}|X, Y$ and $\mathcal{Y}|X, Y$. So $\mathbf{x}$ is chosen over $\mathbf{y}$ if and only if

$$\mathbb{E}[u(\mathcal{X})|X, Y] > \mathbb{E}[u(\mathcal{Y})|X, Y].$$

We obtain the posterior belief from Bayesian updating, using the fact that $X - \mathbf{x}^* =$

$Y - \mathbf{y}^*$,

$$\mathcal{X}_1|X,Y = x_1^*, \text{ and } \mathcal{X}_2|X,Y \sim \mathcal{N}\left(\frac{1}{3}(2X_2 - Y_2), \frac{1}{3}\right).^{12}$$

The belief about the first attribute $\mathcal{X}_1|X,Y$ is equal to the true attribute level, because there is no noise in this dimension. The belief about the second attribute exhibits contrast effect. If $\mathbf{y}$ is very good in the second attribute (i.e. if Y_2 is very large), then in contrast, $\mathbf{x}$ is perceived to be very poor in the second attribute (i.e. then $\frac{1}{3}(2X_2 - Y_2)$ is very small). Substituting the belief into the expected utility formula gives $\mathbf{x}$ is chosen over $\mathbf{y}$ if and only if

$$\mathbb{E}[u(\mathcal{X})|X,Y] = -e^{-3x_1^*} - e^{-(2X_2 - Y_2)+3/2} > -e^{-3y_1^*} - e^{-(2Y_2 - X_2)+3/2} = \mathbb{E}[u(\mathcal{Y})|X,Y].^{13}$$

To obtain the choice probability, we substitute in the equality that $X - \mathbf{x}^* = Y - \mathbf{y}^* = \epsilon$ and get

$$-\frac{3}{2} + \ln\left(\frac{e^{-3y_1^*} - e^{-3x_1^*}}{e^{y_2^* - 2x_2^*} - e^{x_2^* - 2y_2^*}}\right) > -\epsilon_2.$$

Since the $\epsilon_2 \sim \mathcal{N}(0,1)$, the choice probability can be expressed using the normal c.d.f Φ,

$$C(\mathbf{x}, \{\mathbf{x}, \mathbf{y}\}) = \Phi\left(-\frac{3}{2} + \ln\left(\frac{e^{-3y_1^*} - e^{-3x_1^*}}{e^{y_2^* - 2x_2^*} - e^{x_2^* - 2y_2^*}}\right)\right).$$

For interpretation, first recall that $x_1^* > y_1^*$ and $y_2^* > x_2^*$. Therefore, both $e^{-3y_1^*} - e^{-3x_1^*}$ and $e^{y_2^* - 2x_2^*} - e^{x_2^* - 2y_2^*}$ are positive. Moreover, since both Φ and ln are increasing functions, the choice probability is increasing in x_1^* and x_2^*, and decreasing in y_1^* and y_2^*. Intuitively, the agent is more likely to choose $\mathbf{x}$ if the true attribute levels of $\mathbf{x}$ improves, less so if the

[12] Similarly $\mathcal{Y}_1|X,Y = \mathbf{y}_1^*$, and $\mathcal{Y}_2|X,Y \sim \mathcal{N}\left(\frac{1}{3}(2Y_2 - X_2), \frac{1}{3}\right)$.

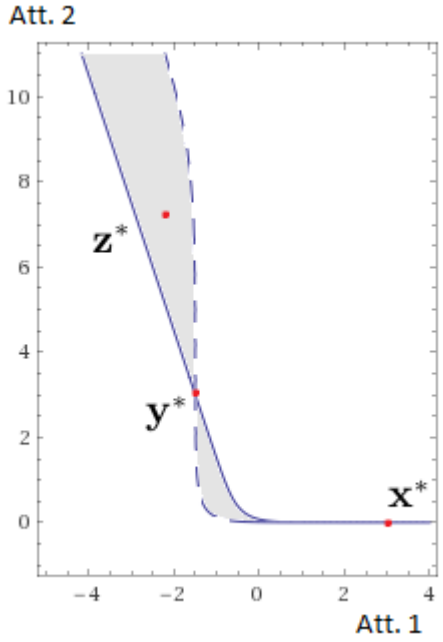
[13] This expression $\mathbb{E}[u(\mathcal{X})|X,Y] = -e^{-3x_1^*} - e^{-(2X_2 - Y_2)+3/2}$ can be interpreted as follow. The utility from the first attribute is clear due to perfect perception. We have mentioned the contrast effect influences perception, and hence the expected utility through $(2X_2 - Y_2)$. The better Y_2 is, the smaller the expected utility for $\mathbf{x}$. The constant in the exponent of the second term comes from the uncertainty. Because $\mathcal{X}_2|X,Y$ is normally distributed, $e^{-3\mathcal{X}_2}$ is log-normal, and its expectation involves a constant from the variance of $\mathcal{X}_2|X,Y$.

attributes of $\mathbf{y}$ becomes more desirable.[14]

The indifference curve can be traced out using the definition $C(\mathbf{x}, \{\mathbf{x}, \mathbf{y}\}) = 0.5$. Because $\Phi(0) = 0.5$, we have $\mathbf{x} \sim \mathbf{y}$ if and only if

$$0 = -\frac{3}{2} + \ln\left(\frac{e^{-3y_1^*} - e^{-3x_1^*}}{e^{y_2^* - 2x_2^*} - e^{x_2^* - 2y_2^*}}\right).$$

Any $\mathbf{x}$ and $\mathbf{y}$ with attributes satisfying the above equation are stochastically indifferent.



The horizontal asymptote comes from the noiseless perception in attribute one. For example, consider an alternative $\mathbf{w}$ *that is indifferent to* $\mathbf{x}$*. If* w_1^* *is large and positive,* $E[u(\mathcal{W})|X, W] = -\exp(-3w_1^*) - E[\exp(-3\mathcal{W}_2)|X, W] \approx -E[\exp(-3\mathcal{W}_2)|X, W]$*. So intuitively,* $\mathbf{w} \sim \mathbf{x}$ *requires* $-E[\exp(-3\mathcal{W}_2)|X, W]$ *to be close to the expected utility of* $\mathbf{x}$*. This restricts* w_2^* *close to a constant. On the other hand, if* w_1^* *is negative, then* $-\exp(-3w_1^*)$ *is a non-negligible negative number. In order to keep the indifference, a larger* w_2^* *is require to compensate for this negative utility.*

Another observation is that the indifference curve of $\mathbf{y}$ *is steeper than that of* $\mathbf{x}$ *when the first attribute is lacking. This is a result of noisy perception in the second attribute and the contrast effect that follows. To interpret the difference, notice that* $\mathbf{y}$ *is strong in attribute two while* $\mathbf{x}$ *is lacking. When an alternative* $\mathbf{w}$ *is evaluated in the context of* $\mathbf{x}$*, its second attribute is perceived as better compared to* $\mathbf{x}$*. However, in the context of* $\mathbf{y}$*, the same attribute level would appear less strong because* $\mathbf{y}$ *is strong in attribute two. Hence if* $w_1^* < y_1^*$*, a much stronger* w_2^* *is require for* $\mathbf{w}$ *to be comparable to* $\mathbf{y}$ *than to* $\mathbf{x}$*.*

Figure 5.1: Crossing Stochastic Indifference Curves

[14] We will show that if $x_1^* > y_1^*$ and $y_2^* > x_2^*$ does not hold, the dominating option will be chosen with probability 1 in next section.

Generically, if $\mathbf{x} \sim \mathbf{y}$, their indifference curves cross. For illustration, we let $\mathbf{x}^* = (3, 0)$ and $\mathbf{y}^* = (3 - \frac{1}{3}\ln(1 - e^{9/2} + e^{27/2}), 3)$ and check that $\mathbf{x} \sim \mathbf{y}$. As shown in Figure 5.1, the red dots are the corresponding true attribute levels, and the indifference curve of $\mathbf{x}$ is the solid curve, whereas the one of $\mathbf{y}$ is dashed. The two curves intersects at $\mathbf{x}^*$ and $\mathbf{y}^*$. The curves are indistinguishable for large values in first attribute. Because the curves are distinct, intransitivity can occur when we cansider any $\mathbf{z}$ with attributes in the shaded area. As in Figure 5.1, $\mathbf{z}^*$ is to the left of the $\mathbf{y}$-curve and the right of the $\mathbf{x}$-curve. So $C(\mathbf{y}\{\mathbf{y}, \mathbf{z}\}) > 0.5$ and $C(\mathbf{x}\{\mathbf{x}, \mathbf{z}\}) < 0.5$. But as readily seen, slight improving $\mathbf{x}^*$ in either attribute will cause $C(\mathbf{x}\{\mathbf{x}, \mathbf{y}\}) > 0.5$. Thereby strictly violating weak transitivity. The example is itself a proof of the following existence result.

Proposition 28. *Suppose there is imperfect perception in one of the attributes. There exists a normal-Bayesian rational agent with a standard preference who displays intransitivity.*

5.3.2 Joint-Separate Valuation Reversal

The phenomena refers to the following type of observations that the average willingness to pay (valuations) for two alternatives reverse in different contexts. As recorded in an experiment of Hsee (1996), the subjects (as company owners) were asked for their willingness to pay to hire different job candidates as programmers. Candidate $\mathbf{x}$ has a college GPA of 4.9 out of 5 and has written 10 programs in the computer language KY. Candidate $\mathbf{y}$ has a GPA of 3.0 from the same school, and has written 70 similar programs in the same language. When the subjects were asked to evaluate $\mathbf{x}$ alone, the average willingness to pay was about 32.7k dollars; when asked to evaluate $\mathbf{y}$ alone, the average willingness to pay was about 26.8k. However, when the two candidates were presented together, the inequality between the amounts reversed. The average willingness to pay for $\mathbf{x}$ in the presence of $\mathbf{y}$ became 31.2k, while that of $\mathbf{y}$ became 33.2k. With abuse of notation, we denote by $\$(\mathbf{x})$ and $\$(\mathbf{y})$

the average willingness to pay for $\mathbf{x}$ and $\mathbf{y}$ in dollars, and denote by $\$(\mathbf{x}|\mathbf{x},\mathbf{y})$ the average willingness to pay for $\mathbf{x}$ in the presence of $\mathbf{y}$, and $\$(\mathbf{y}|\mathbf{x},\mathbf{y})$ for $\mathbf{y}$ in the presence of $\mathbf{x}$. A decision maker is said to *display j-s reversal* if there are $\mathbf{x},\mathbf{y}$ such that both $\$(\mathbf{x}) > \$(\mathbf{y})$ and $\$(\mathbf{x}|\mathbf{x},\mathbf{y}) < \$(\mathbf{y}|\mathbf{x},\mathbf{y})$ holds.

In the experiment, the two attributes are the GPAs and programing experience. Since GPA is familiar to most people and relatively easy to interpret with a known scale, we take it to be the noiseless attribute. The programing experience, although explicitly measured in numbers of programs written, is more difficult to interpret. It is not clear how advanced the computer language KY is, and how difficult it is to write programs in. Hence we take this attribute to be confounded with imperfect perception. To demonstrate the reversal, we need a pair of $\mathbf{x}$ and $\mathbf{y}$ such that $x_1^* > y_1^*$ and $x_2^* < y_2^*$, and that $\$(\mathbf{x}) > \$(\mathbf{y})$ and $\$(\mathbf{x}|\mathbf{x},\mathbf{y}) < \$(\mathbf{y}|\mathbf{x},\mathbf{y})$ hold simultaneously. In this subsection, we use the *average posterior expected utility* as a proxy for average willingness to pay. That is, $\$(\mathbf{x})$ is understood as the average posterior expected utility of $\mathbf{x}$ in $\{\mathbf{x}\}$, $\$(\mathbf{y})$ that of $\mathbf{y}$ in $\{\mathbf{y}\}$, and $\$(\mathbf{x}|\mathbf{x},\mathbf{y})$ and $\$(\mathbf{y}|\mathbf{x},\mathbf{y})$ that of $\mathbf{x}$ and of $\mathbf{y}$ in $\{\mathbf{x},\mathbf{y}\}$.

When there is only one option, the posterior is based only on its own signal. From noiseless perception, $\mathcal{X}_1|X = x_1^*$. Standard bayesian update gives $\mathcal{X}_2|X \sim \mathcal{N}(\frac{1}{2}X_2, \frac{1}{2})$. Hence the average posterior expected utility is

$$\$(\mathbf{x}) := \mathbb{E}_X[\mathbb{E}_{\mathcal{X}_2}[-e^{-3x_1^*} - e^{-3\mathcal{X}_2}|X]] = -e^{-3x_1^*} - e^{-\frac{3}{2}x_2^* + \frac{27}{8}}.$$

A similar expression holds for $\mathbf{y}$. On the other hand, when there are two options, through similar analysis as previous subsection, we have

$$\$(\mathbf{x}|\mathbf{x},\mathbf{y}) := \mathbb{E}_{X,Y}[\mathbb{E}_{\mathcal{X}_2}[-e^{-3x_1^*} - e^{-3\mathcal{X}_2}|X,Y]] = -e^{-3x_1^*} - e^{-(2x_2^* - y_2^*)+2}.$$

Also, a similar expression holds for $\$(\mathbf{y}|\mathbf{x},\mathbf{y})$. Hence the two inequalities $\$(\mathbf{x}) > \$(\mathbf{y})$ and $\$(\mathbf{x}|\mathbf{x},\mathbf{y}) < \$(\mathbf{y}|\mathbf{x},\mathbf{y})$ become

$$\begin{cases} -e^{-3x_1^*} - e^{-\frac{3}{2}x_2^* + \frac{27}{8}} > -e^{-3y_1^*} - e^{-\frac{3}{2}y_2^* + \frac{27}{8}} \\ -e^{-3x_1^*} - e^{-(2x_2^* - y_2^*)+2} < -e^{-3y_1^*} - e^{-(2y_2^* - x_2^*)+2}. \end{cases}$$

There are many pairs of alternatives that satisfy both inequalities. For illustration, we let $\mathbf{x}^*$ be $(3, 0)$ as in the previous subsection, and Figure 5.2 plots the shaded region where both inequalities are satisfied. The dashed curve is the boundary defined by the first inequality, and the solid curve is the one by the second. Any $\mathbf{y}$ with attributes $\mathbf{y}^*$ in the shaded region is an example of the desired reversal.

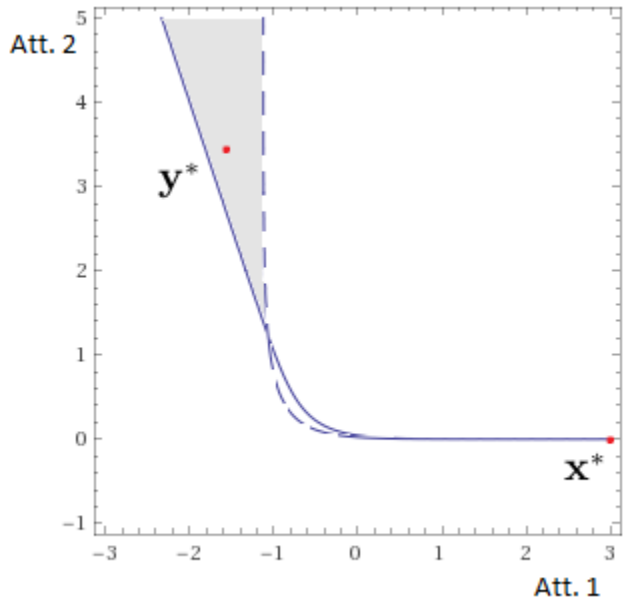


Figure 5.2: Joint-Separate Valuation Reversal

The solid curve is the set of all attributes that have the same average posterior utility as $\mathbf{x}$ *in a binary choice problem, and therefore naturally looks similar to the stochastic indifference curve of* $\mathbf{x}$ *in Figure 5.1. The dash line in 5.2 has a vertical asymptote, which is determined through* $-e^{-3\times 3} - e^{-\frac{3}{2}\times 0 + \frac{27}{8}} = -e^{-3y_1^*} - e^{-\frac{3}{2}y_2^* + \frac{27}{8}} \approx -e^{-3y_1^*}$ *for large values of* y_2^*.

This mechanism that causes the reversal is intuitive. An $\mathbf{y}$ that is bad in the first attribute receives low valuation in separate valuation. And because the utility function is concave and the perception is noisy, a strong second attribute does not effectively increase the overall valuation. However, in joint valuation, there is a clear contrast in the second attributes of $\mathbf{x}$ and $\mathbf{y}$. Under the contrast, $\mathbf{x}$ is perceived as much worse off, and $\mathbf{y}$ much better off, resulting in the reversal. The above example is itself a proof of the following existence result.

Proposition 29. *Suppose there is imperfect perception in one of the attributes. There exists a normal-Bayesian rational agent with standard preference who displays j-s reversal.*

5.3.3 Illustrating Ternary Choices Through the Compromise Effect

The compromise effect involves choice problems of two and three options. As in Figure 5.3, suppose there is a binary choice problem with options $\mathbf{x}, \mathbf{y}$ where $\mathbf{x}$ is better than $\mathbf{y}$ in the first attribute but $\mathbf{y}$ is better in the second. The *compromise effect* ([37]) refers to introducing a third $\mathbf{z}$ in or near the region C where $\mathbf{z}^*$ seems extremely favorable in the first attribute but extremely unfavorable in the second one. Empirically, at the introduction of $\mathbf{z}$, people are generally led to choose the "compromising option" $\mathbf{x}$, increasing its choice frequency. Mathematically, let the initial choice set be $\{\mathbf{x}, \mathbf{y}\}$ and the expanded choice set be $\{\mathbf{x}, \mathbf{y}, \mathbf{z}\}$ where $z_1^* > x_1^* > y_1^*$ and $y_2^* > x_2^* > z_2^*$. The compromise effect refers to $C(\mathbf{x}, \{\mathbf{x}, \mathbf{y}, \mathbf{z}\}) > C(\mathbf{x}, \{\mathbf{x}, \mathbf{y}\})$ for all $\mathbf{z}$ "inferior enough".

Let Pr denote the probability measure for ϵ. We have seen previously that

$$C(\mathbf{x}, \{\mathbf{x}, \mathbf{y}\}) = \Pr\left(\mathbb{E}[u(\mathcal{X})|X, Y] > \mathbb{E}[u(\mathcal{Y})|X, Y]\right) = \Pr\left(\epsilon_2 > \frac{3}{2} - \ln\left(\frac{e^{-3y_1^*} - e^{-3x_1^*}}{e^{y_2^* - 2x_2^*} - e^{x_2^* - 2y_2^*}}\right)\right), \tag{5.1}$$

Similarly, we can also express the ternary probability as

$$C(\mathbf{x}, \{\mathbf{x}, \mathbf{y}, \mathbf{z}\}) = \Pr\left(\left\{\mathbb{E}[u(\mathcal{X})|X, Y, Z] > \mathbb{E}[u(\mathcal{Y})|X, Y, Z]\right\} \cap \left\{\mathbb{E}[u(\mathcal{X})|X, Y, Z] > \mathbb{E}[u(\mathcal{Z})|X, Y, Z]\right\}\right),$$

where the first term in the intersection is the event that $\mathbf{x}$ is perceived better than $\mathbf{y}$,

$$\{\mathbb{E}[u(\mathcal{X})|X,Y,Z] > \mathbb{E}[u(\mathcal{Y})|X,Y,Z]\} = \{\epsilon_2 > \frac{3}{2} - \frac{4}{3}\ln\left(\frac{(e^{-3y_1^*} - e^{-3x_1^*})}{e^{-\frac{3}{4}(3x_2^*-y_2^*-z_2^*)} - e^{-\frac{3}{4}(3y_2^*-x_2^*-z_2^*)}}\right)\}, \tag{5.2}$$

and the second the event that $\mathbf{x}$ is perceived better than $\mathbf{z}$,

$$\{\mathbb{E}[u(\mathcal{X})|X,Y,Z] > \mathbb{E}[u(\mathcal{Z})|X,Y,Z]\} = \{\epsilon_2 < \frac{3}{2} - \frac{4}{3}\ln\left(\frac{e^{-3x_1^*} - e^{-3z_1^*}}{e^{-\frac{3}{4}(3z_2^*-x_2^*-y_2^*)} - e^{-\frac{3}{4}(3x_2^*-y_2^*-z_2^*)}}\right)\}. \tag{5.3}$$

In these two events, both fractions inside the logarithm are positive, because $z_1^* > x_1^* > y_1^*$ and $y_2^* > x_2^* > z_2^*$. It is clear that both sets are monotonic in the attributes of $\mathbf{x}$, the better the attributes for $\mathbf{x}$ are, the larger the event that $\mathbf{x}$ is the most preferred. Through a similar rationale, it is intuitive to see in Equation 5.3 that the event that $\mathbf{x}$ is preferred to $\mathbf{z}$ is monotonically decreasing in the attributes of $\mathbf{z}$.

More subtle is the influence of attributes of $\mathbf{z}$ on the preference between $\mathbf{x}$ and $\mathbf{y}$. From Equation 5.2, it is clear that the first attribute of $\mathbf{z}$ does not affect the preference between $\mathbf{x}$ and $\mathbf{y}$. This is because the first attribute is noiseless. The second attribute is not. The (main component of the) perceived second attribute of $\mathbf{x}$ is $3x_2^* - y_2^* - z_2^*$.[15] Hence the term $-e^{-\frac{3}{4}(3x_2^*-y_2^*-z_2^*)}$ is (the main component of) the posterior utility of $\mathbf{x}$ in the second attribute. A weak attribute level of z_2^* contrast with that of $\mathbf{x}$, increasing its perceived level as well as the posterior utility level. Therefore, $\mathbf{x}$ appears more appealing in the context of an undesirable $\mathbf{z}$. Similarly, such an undesirable $\mathbf{z}$ also increases the posterior utility of $\mathbf{y}$. However, $y_2^* > x_2^*$ and so $\mathbf{y}$ is more satiated than $\mathbf{x}$ in the second attribute. Hence the increase in perceived levels benefits $\mathbf{x}$ more. Mathematically, in attribute two, both the posterior utility $-e^{-\frac{3}{4}(3y_2^*-x_2^*-z_2^*)}$ of $\mathbf{y}$ and that $-e^{-\frac{3}{4}(3x_2^*-y_2^*-z_2^*)}$ of $\mathbf{x}$ increases as z_2^* decreases,

[15]The posterior belief is $\mathcal{X}_2 \sim \mathcal{N}(\frac{1}{4}(3X_2 - Y_2 - Z_2), \frac{1}{4})$.

but the gap

$$\begin{aligned} e^{-\frac{3}{4}(3x_2^*-y_2^*-z_2^*)} - e^{-\frac{3}{4}(3y_2^*-x_2^*-z_2^*)} &= -e^{-\frac{3}{4}(3y_2^*-x_2^*-z_2^*)} - \left(-e^{-\frac{3}{4}(3x_2^*-y_2^*-z_2^*)}\right) \\ &= \left(-e^{-\frac{3}{4}(3y_2^*-x_2^*)} - (-e^{-\frac{3}{4}(3x_2^*-y_2^*)})\right)\exp(\frac{3}{4}z_2^*) \end{aligned}$$

decreases. Therefore, from Equation 5.2, a low z_2^* reduces the gap, and makes it more likely that $\mathbf{x}$ is preferred to $\mathbf{y}$.

To show that the compromise effect occurs, we take the limit that $z_1^* \to x_1^*$ from the right and see from Equation 5.3 that $\mathbf{x}$ is perceived better thant $\mathbf{z}$ with probability approaching 1. I.e. $\Pr\left(\left\{\mathbb{E}[u(\mathcal{X})|X,Y,Z] > \mathbb{E}[u(\mathcal{Z})|X,Y,Z]\right\}\right) \to 1$ as $z_1^* \searrow x_1^*$. Moreover, for z_2^* small enough, the event in Equation 5.2 becomes a superset than the event in Equation 5.1. I.e. $\Pr\left(\left\{\mathbb{E}[u(\mathcal{X})|X,Y,Z] > \mathbb{E}[u(\mathcal{Y})|X,Y,Z]\right\}\right) > C(\mathbf{x},\{\mathbf{x},\mathbf{y}\})$ for z_2^* small enough. Therefore $C(\mathbf{x},\{\mathbf{x},\mathbf{y},\mathbf{z}\}) > C(\mathbf{x},\{\mathbf{x},\mathbf{y}\})$ for inferior enough $\mathbf{z}$. We have just proved the following result.

Proposition 30 (The Compromise Effect)**.** *Assume the parametrization in this section. For any* $\mathbf{x},\mathbf{y}$ *with* $x_1^* > y_1^*$ *and* $x_2^* < y_2^*$*,there exists* $\delta > 0$ *and* $D \in \mathbb{R}$ *such that for all* $\mathbf{z}$ *with* $z_1^* - x_1^* \in (0,\delta)$ *and* $z_2^* < D$*, the inequality* $C(\mathbf{x},\{\mathbf{x},\mathbf{y},\mathbf{z}\}) > C(\mathbf{x},\{\mathbf{x},\mathbf{y}\})$ *holds.*

The result above points out an important distinction between our model and a large class models that satify Monotonicity (also called Regularity). This include the class of all random utility models (see e.g. [4] and [8]). In the random utility framework, the utility of the options $\mathbf{x},\mathbf{y},\mathbf{z}$ are real-valued random variables $U_{\mathbf{x}}, U_{\mathbf{y}}, U_{\mathbf{z}}$, i.e. measurable functions from a probability space to $\mathbb{R}$. The decision maker chooses $\mathbf{x}$ if and only if the event $\{U_{\mathbf{x}} > U_{\mathbf{y}} \text{ and } U_{\mathbf{x}} > U_{\mathbf{z}}\}$ is realized. A very general random utility model allows $U_{\mathbf{x}}, U_{\mathbf{y}}$, and $U_{\mathbf{z}}$ to be correlated in arbitrary ways. Nonetheless it always holds that

$$\{U_{\mathbf{x}} > U_{\mathbf{y}}\} \subseteq \{U_{\mathbf{x}} > U_{\mathbf{y}} \text{ and } U_{\mathbf{x}} > U_{\mathbf{z}}\}, \text{ and hence } C(\mathbf{x},\{\mathbf{x},\mathbf{y}\}) < C(\mathbf{x},\{\mathbf{x},\mathbf{y},\mathbf{z}\}).$$

According to Proposition 30, our model directly violate this property, and hence it cannot be reinterpreted as any random utility model.

Through a similar mechanism, our model also capture two other effects in Figure 5.3. The *phantom decoy* effect ([31]) occurs in the situation when $\mathbf{z}$ is positioned near the area P. Usually, the phantom alternative is better than $\mathbf{x}$ in first attribute and no worse than $\mathbf{x}$ in the second. Also, it is worse than $\mathbf{y}$ in the second attribute. In experimental settings, the subjects are told that such $\mathbf{z}$ is unavailable to choose and hence the agent has to choose from $\{\mathbf{x}, \mathbf{y}\}$. Empirically, the phantom decoy increases the frequency of choosing $\mathbf{x}$.[16] The *attraction effect* ([18]) corresponds to introducing a third option $\mathbf{z}^*$ in or near the region A in Figure 2. In general, $\mathbf{z}$ needs to be inferior to $\mathbf{x}$ in the second attribute, and no better in the first. In addition, $\mathbf{z}$ needs to be better than $\mathbf{y}$ in the first attribute. Empirically, such a third option itself is hardly chosen.

Because our model predicts these two effects through a similar channel, it is suggestive that, at least there are some commonality among the effects, as observed by [15]. Here, we omit their formal proofs because they will be covered under the decoy choice pattern in the next section. Nonetheless, a proof of the attraction effect (phantom decoy effect) parallels the following intuition. Suppose there is imperfect perception in the second (first) attribute. Let $\mathbf{x}, \mathbf{y}$ be as before. $\mathbf{x}$ is inferior in the second attribute and $\mathbf{y}$ is inferior in the first attribute. Now introduce the third object $\mathbf{z}$ that is is extremely bad in the second attribute (good in the first attribute). In comparison, $\mathbf{z}$ in A (P) makes both $\mathbf{x}$ and $\mathbf{y}$ seem better in the second attribute (worse in the first attribute) than before. Such a change makes $\mathbf{x}$ relatively more favorable (less repulsive) since $\mathbf{y}$ was already good enough in the second attribute (barely acceptable in the first attribute) at the outset.

[16] See e.g. [31], [15], [28], [29] and [14] etc.

5.3.4 Remarks on the Parametric Model

In this section, we have illustrated intransitivity, the valuation reversal and the compromise effect through a parametric model. However we want to emphasize that the illustrated channel is not limited to the parameter chosen. In fact, similar calculation walks through with any utility function $u(x) := u(x_1, x_2) = -e^{\gamma x_1} - e^{\rho x_2}$ where $\gamma, \rho < 0$, and any noise distributed as $\epsilon \sim N\left(0, \begin{bmatrix} 1/t_1^2 & 0 \\ 0 & 1/t_2^2 \end{bmatrix}\right)$, $t_i \in (0, \infty]$ where one of the t_i's can be infinite.

Moreover, we also want to remark that given any family of parametrized utility functions, the parameters can be estimated easily from choice data. For example, in our parametrization, the choice probability for any binary problem is given analytically below.

Lemma 31. *For any* $\mathbf{x}, \mathbf{y}$ *where* $x_1^* > y_1^*$ *and* $y_2^* > x_2^*$*, the parametric model in the subsection gives* $C(\mathbf{x}, \{\mathbf{x}, \mathbf{y}\}) = \Phi\left(\theta(\gamma, \rho, \mathbf{x}^*, \mathbf{y}^*, t)\right)$*, where* Φ *is the standard normal c.d.f. function and* $\theta(\gamma, \rho, \mathbf{x}^*, \mathbf{y}^*, t)$ *is defined as*

$$\theta := \frac{1}{\sqrt{\left(\frac{\rho\sqrt{t_2}}{2+t_2^2}\right)^2 + \left(\frac{\gamma\sqrt{t_1}}{2+t_1^2}\right)^2}} \left[\frac{\gamma^2}{2(2+t_1^2)} - \frac{\rho^2}{2(2+t_2^2)} + \ln\left(\frac{\exp\left(\gamma \frac{(t_1^2+1)y_1^* - x_1^*}{2+t_1^2}\right) - \exp\left(\gamma \frac{(t_1^2+1)x_1^* - y_1^*}{2+t_1^2}\right)}{\exp\left(\rho \frac{(t_2^2+1)x_2^* - y_2^*}{2+t_2^2}\right) - \exp\left(\rho \frac{(t_2^2+1)y_2^* - x_2^*}{2+t_2^2}\right)} \right) \right].$$

When an attribute becomes noiseless (i.e. $t_1 \to \infty$), the above Lemma reduces to Equation 5.1. We have seen previously that an $\mathbf{x}$ with better attributes results in a higher θ and higher $C(\mathbf{x}, \{\mathbf{x}, \mathbf{y}\})$, and the reverse holds for $\mathbf{y}$. Because $x_1^* > y_1^*$, and γ is the preference parameter in the first attribute, a larger γ^2 implies the first attribute is more decisive, and hences more likely to choose $\mathbf{x}$.

As the Lemma specifies choice probabilities in terms of parameters, it can be used to estimate exponential utility functions when there are observations for different menus. When the parameters are estimated, the model can be used to predict choice probabilities in new menus. Here, we need to point out an implicit assumption similar to that in [22] is adopted.

To maintain empirical identifiability and avoid excessive degrees of freedom, the definitions and measurements of the attributes must be determined *before* fitting the model to data. They should not be free parameters but part of the data that the model seeks to explain.[17]

Although the complex expression can be useful for experimenters, the agent in the model does not evaluate this complicated algebra before making the choice. She simply chooses the choice item that maximizes her expected utility while being completely unaware of the choice probabilities her actions generate.

5.4 The General Results

We have now shown that our simple parametric model can explain and predict several contextual choice effects. These results are *not* the outcomes of model flexibility. On the contrary, the model is quite rigid in the sense that these types of contextual effects *have to occur* even without the parametric assumptions. One can view these as testable implications of the model. We first define the term "decoy choice pattern", as an abstraction of the attraction, compromise, and phantom decoy effect. We will then show that the general model predicts the decoy choice pattern under the general class of preferences and prior-signal distributions as described in section 2. We also show that although the model captures several contextual effects, the model also predicts classical rational choice behavior under some specific type of menus, and hence there are other rationality constraints on what the model can accomodate.

[17]While it is easier to satisfy this procedure in marketing experiments where the attributes of each object are specified by the experimenter, it is sometimes difficult to include other relevant attributes in real life decision-making processes. For example, when shopping (online or in person), individuals may base their decisions on attributes that are not listed on the product descriptions. For example, decisions may be made based on the retailer's customer service, which is usually not listed in the product labels. Hence it is difficult to account for these influences.

5.4.1 The Decoy Choice Pattern

The next definition is relevant to the phantom decoy effect, the compromise effect and the attraction effect. We start with a binary choice problem with $\mathbf{x}, \mathbf{y}$ where $\mathbf{x}$ is better than $\mathbf{y}$ in the first attribute but $\mathbf{y}$ is better in the second, as shown in Figure 5.3. As discussed previously, a third object $\mathbf{z}$ in the lower right corner of Figure 5.3 generally increases the choice probability of $\mathbf{x}$. Due to symmetry, it is also true empirically that a $\mathbf{z}$ in the upper left corner of the same Figure 5.3 will increase the choice probability of $\mathbf{y}$ (e.g. a compromise effect where $\mathbf{y}$ is the compromising option). These empirical effects share a commonality that $\mathbf{z}$ is either unavailable (as a phantom decoy) or rarely chosen (as in the compromise effect or attraction effect). We can reasonably conjecture that both the attraction effect and the compromise effect will remain qualitatively unchanged when the third option $\mathbf{z}$ is unavailable. We summarize these observation as follow. An *unavailable third option* $\mathbf{z}$ to the upper left area of the attribute space increases the choice probability of $\mathbf{y}$, and if the third option is to the lower right area of the space, it increases the choice probability of $\mathbf{x}$. We call this comparative statics the *decoy choice pattern*.

Definition 32. *The choice behavior is said to display the* decoy choice pattern *if there exists a vector $\Delta \in \mathbb{R}^2$ with $\Delta_1 > \Delta_2$, such that for any $\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{z}'$ with attributes in $\mathbb{R}^2$ satisfying $x_1^* > y_1^*$, $x_2^* < y_2^*$ and $\mathbf{z}'^* = \mathbf{z}^* + \Delta$, the inequality $C(\mathbf{x}, \{\mathbf{x}, \mathbf{y}, (\mathbf{z})\}) \leq C(\mathbf{x}, \{\mathbf{x}, \mathbf{y}, (\mathbf{z}')\})$ holds.*

Our model predicts the decoy choice pattern, which is an empirically testable implication in two ways. First, when there are at least two attributes under consideration, the agent does not satisfy the Luce's IIA over menus described in the decoy choice pattern. Second, the agent violates Luce's IIA in a specific way. E.g. making $\mathbf{x}$ the compromising option in the experiments *does not reduce* the choice probability of $\mathbf{x}$.

Theorem 33. *Any normal-Bayesian rational agent with standard preference and imperfection perception displays the decoy-choice pattern.*

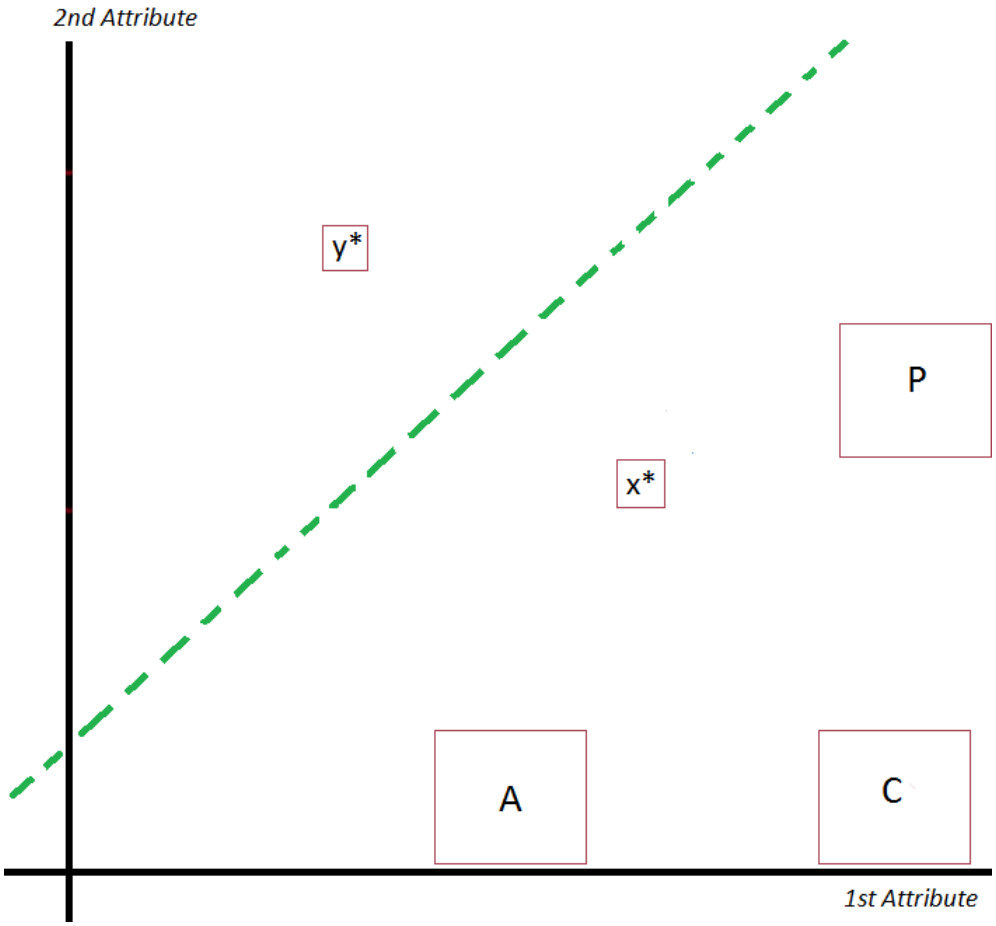


Figure 5.3: Areas for the phantom decoy effect (P), the compromise effect (C) and the attraction effect (A)

Observe that the theorem is a sufficiency result. Intuitively, it states that if $\mathbf{z}'^*$ is to the right or to the bottom of $\mathbf{z}^*$, such a $\mathbf{z}'$ affects the choice probability of $\mathbf{x}$ more positively than $\mathbf{z}$ does. Another interesting implication of the theorem is that the attraction effect and the compromise effects should still exist even when $\mathbf{z}$ is unavailable. Since $\mathbf{z}$ is infrequently chosen in experiments, such a prediction is reasonable to expect, but distinct for our model. Other choice models usually do not consider an unchoosable options.

5.4.2 Rational Content in the Model

We have seen previously that when there is a trade-off between the alternatives, i.e. some alternatives are better in the first attribute while others are better in the second, contextual choices arise in the model. A natural question is what would the model predict when such a trade-off is absent. Intuitively, if we are given two alternatives $\mathbf{x}$ and $\mathbf{z}$ where $\mathbf{z}^* > \mathbf{x}^*$, a rational agent should always choose $\mathbf{z}$ due to the monotonicity of the utility function.[18] The prediction of our model fits this intuition. Since the error ϵ in perception is the same for each of $\mathbf{x}$ and $\mathbf{z}$, the perturbed signal $X = \epsilon + \mathbf{x}^*$ and $Z = \epsilon + \mathbf{z}^*$ preserves the inequality: $Z > X$. A Bayesian rational agent can hence correctly infer the inequality and choose optimally.

Theorem 34. *For any* $\{\mathbf{x}, \mathbf{z}\}$ *with* $\mathbf{x}^*, \mathbf{z}^* \in \mathbb{R}^2$*, a normal-Bayesian rational agent with standard preference and imperfect perception chooses* $\mathbf{z}$ *with probability* 1 *if* $\mathbf{z}^* > \mathbf{x}^*$.

It is clear that the above theorem predicts the following intuitive choice effect described and observed in [43] and [44]. Consider an individual that is choosing between a trip to Paris ($\mathbf{x}$) and a trip to Rome ($\mathbf{y}$). If she is interested to see both places and doesn't have a strong preference for one over the other, let's say the choice probability for Paris ($\mathbf{x}$) would be 1/2. Now if we offer the individual a new choice problem with two alternatives, a trip to Paris ($\mathbf{x}$) or a trip to Paris plus a \$1 bonus ($\mathbf{z}$), he would probably not hesitate to choose the option with the extra dollar. In other words, choosing $\mathbf{z}$ over $\mathbf{x}$ is of probability 1. However, if we offer him a third choice problem that consists of a trip to Paris plus \$1 and a trip to Rome, it is intuitive that the choice probability should still be roughly 1/2.

An implication of the above theorem is that transitivity holds deterministically for a set of choice objects such that each one is dominated or dominates one another. Therefore, violation of weak stochastic transitivity can only happen when the alternatives do not dominates each other. We state this result formally and proof is immediate.

[18] The vector inequality $\mathbf{z}^* > \mathbf{x}^*$ means $z_1^* \geq x_1^*$ and $z_2^* \geq x_2^*$ with at least one inequality being strict.

Corollary 35. *Suppose* $\mathbf{x}, \mathbf{y}, \mathbf{z}$ *have attributes* $\mathbf{x}^* > \mathbf{y}^* > \mathbf{z}^*$*, then* $1 = C(\mathbf{x}, \{\mathbf{x}, \mathbf{y}\}) = C(\mathbf{y}, \{\mathbf{y}, \mathbf{z}\}) = C(\mathbf{x}, \{\mathbf{x}, \mathbf{z}\}) > 1/2$.

The Theorem 34 can also be generalized to the following statement. When $S = \{\mathbf{x}^1, \ldots, \mathbf{x}^n\}$ is the choice set involving multiple options, if $\mathbf{x}^i$ is dominated in the set S, then $C(\mathbf{x}^i, S) = 0$. In other words, objects are chosen with positive probability only when they are on the "attribute possibility frontier". This is a rationality condition that the agent has to satisfy, and it rules out many other types of irrational choice behaviors.

An interesting question would be if similar result holds for j-s reversal as well? The answer is not straight forward. The following Corollary follows immediately from the proof of Theorem 34.

Corollary 36. *For any* $\{\mathbf{x}, \mathbf{z}\}$ *with* $\mathbf{x}^* < \mathbf{z}^* \in \mathbb{R}^2$*, it holds that* $\$(\mathbf{x}|\mathbf{x}, \mathbf{z}) < \$(\mathbf{z}|\mathbf{x}, \mathbf{z})$.

However, it does not follow that $\$(\mathbf{x}) < \$(\mathbf{z})$ if $\mathbf{x}^* < \mathbf{z}^*$ when the correlations r and R are not restricted. To illustrate the intuition, consider the apartment choice problem and the two attributes are convenience and safty. Suppose the agent values both attributes, and she holds the prior that the two attributes are negatively correlated: a convenient location is usually less safe, and a safe location is farther away and hence less convenient on average. Let $\mathbf{x}$ and $\mathbf{z}$ be two apartments that are exactly of the same safty level, but $\mathbf{z}$ is more convenient, i.e., $\mathbf{x}^* < \mathbf{z}^*$. It clearly holds $\$(\mathbf{x}|\mathbf{x}, \mathbf{z}) < \$(\mathbf{z}|\mathbf{x}, \mathbf{z})$ in a joint valuation. However the agent cannot observe this comparison in the separate valuations. If she sees only $\mathbf{z}$, due to the prior, seeing how convenient $\mathbf{z}$ is causes her to believe that $\mathbf{z}$ is not very safe. If she values safty much more than convenience, her valuation for $\$(\mathbf{z})$ can be low. On the other hand, if she sees only $\mathbf{x}$, since $\mathbf{x}$ is not particularly convenient, her posterior easily trust the safty of the location. As a result, her valuation for $\$(\mathbf{x})$ can be decent. In this case, $\$(\mathbf{x}) > \$(\mathbf{z})$ is still allowed by our model. To see this numerically, let the prior be

$\mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}\right)$ and the noise be $\epsilon \sim \mathcal{N}(0, I_2)$. A signal $X = (0,0)$ would result in the posterior belief $\mathcal{X}|X \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{15}\begin{bmatrix} 7 & -2 \\ -2 & 7 \end{bmatrix}^{-1}\right)$. A dominating signal $Z = (1,0)$ would result in the posterior belief $\mathcal{Z}|Z \sim \mathcal{N}\left(\begin{bmatrix} 7/15 \\ -2/15 \end{bmatrix}, \frac{1}{15}\begin{bmatrix} 7 & -2 \\ -2 & 7 \end{bmatrix}^{-1}\right)$. If the utility function values the second attribute a lot more than the first attribute, it is possible that $\mathbb{E}[u(\mathcal{Z})|Z] < \mathbb{E}[u(\mathcal{X})|X]$.

5.4.3 Limiting Noise Structure

As shown is Proposition 30, our model does not satisfy Monotonicity, a fundamental property of all random utility models. Despite this difference, one interesting question might be whether such non-Monotonic predictions disappear in some limiting parameters of our model. For example, if the noise in the signal goes to zero, does our model converge to some well-known models? We discuss below that as the noise term becomes small, our model approximates the well-known conditional probit model of [13]. Because [13]'s model is a random utility model, it satisfies Monotonicity. We also remark that because the conditional probit model can explain the similarity effect, a corollary of this subsection is that our model can also explain the similarity effect.

Again, restrict our discussion to the exponential utility functions so that $u(x_1, x_2) = -e^{\gamma x_1} - e^{\rho x_2}$. Given a finite choice set $S = \{\mathbf{x}^1, \ldots \mathbf{x}^n\}$, the posterior belief of the ith alternative under imperfect perception is

$$\mathcal{X}^i|X^1, \ldots X^n \sim \mathcal{N}\left((T + n\Omega^{-1})^{-1}\left(TX^i + n\Omega^{-1}X^i - \sum_{j=1}^{n}\Omega^{-1}X^j\right), (T + n\Omega^{-1})^{-1}\right).$$

When the noise variance converges to zero, i.e. $T^{-1} \to \mathbf{0}$, the posterior belief $\mathcal{X}^i|X^1, \dots X^n$ is approximately $\mathcal{N}(X^i, T^{-1}) = \mathcal{N}(\mathbf{x}^{i*} + \epsilon, T^{-1})$. When the utility function is smooth enough near $\mathbf{x}^{i*}$, we approximate the expected utility using the utility of the expected attributes

$$\mathbb{E}[u(\mathcal{X})|X^1, \dots X^n] \approx u(\mathbf{x}^{i*} + \epsilon)$$

which is already a random utility model. To see this approximates the [13] under the exponential utility function, write

$$u(\mathbf{x}^{i*} + \epsilon) = -e^{\gamma(x_1^{i*} + \epsilon_1)} - e^{\rho(x_2^{i*} + \epsilon_2)} \approx u_1(x_1^{i*}) + u_2(x_2^{i*}) + \beta_1 u_1(x_1^{i*}) + \beta_2 u_2(x_2^{i*}),$$

where we have used the first order approximation at $\mathbf{x}^{i*}$ with the notation that $u_1(x_1) = -e^{\gamma x_1}, u_2(x_2) = -e^{\rho x_2}$ and $\beta_1 = \gamma\epsilon_1, \beta_2 = \rho\epsilon_2$. It is clear that the form of the approximation coincide with equation (3.6) in [13].

5.5 Discussion and Conclusion

We present a rational choice model with a novel information friction that leads to contextual choices. By rationality, we refer to the bench mark that there is a fixed underlying utility function which can be estimated given idealized data (e.g. Lemma 31). There are two key assumptions. Firstly, there is an attribute specific perception noise that is positively correlated across alternatives. Secondly, the decision maker chooses to maximize Bayesian posterior expectation over the fixed utility function. We show generically that the model predicts the compromise effect, the attraction effect and the phantom decoy effect, and the existence of choice cycles and j-s reversal.

To explain certain contextual choices through maximizing (posterior expected) utility, it is sometimes necessary for the (posterior expected) utility to depend on contexts. When

such dependence is abscent, such as in the random utility framework in [4] and [8], effects violating monotonicity are not allowed. This includes a vast amount of models such as [41], [23], [43] and [13] and more recently, [11].

Our model departs from the random utility framework because the (posterior expected) utility depends on the choice set. This dependence comes from Bayesian learning of each alternative's underlying attributes through a noisy signal. The distribution of the noisy signals is exogeneously given and *not* context dependent. This independency is in the same sense as in a random utility model where the random utilities themselves are not context dependent, even though only the ones in the menu is realized. In our model, the agent observe only the signals from alternatives that are in the menu. The posterior belief given the signals is, however, endogeneously dependent on the menu due to the specification of the noise distribution. This mechanism not only provide us the necessary dependence (of belief) on contexts to explain data, but also distinguishes ourselves from the class of reference dependent models. By reference dependent models, we refer to ones in which utilities are directly assumed to depend on menu, such as [37], [45], [47], [21], [3], [27] and [40] among many others. Because of the fixed underlying preference $u(.)$, our model can be used to perform welfare analysis, and identify choices that are potentially mistakes (i.e. failure to maximize the underlying $u(.)$) due to information frictions.

Our model provides one simple mechanism that is relevant for several contextual effects. In reality, it is likely that there are also other mechanisms at play and are not captured in our model. Imagine for example, *all* three options are presented simultaneously, but each time the availability is limited to a subset of the three. Our model does not allow intransitivity nor compromise effect to happen in these choice problems. In general, our model predicts that after seeing the set of all alternatives then restricting the choice set to any pairs, the resulting ordering will be the same as when the agent is asked to rank all alternatives at sight. This can be potentially inconsistent with certain empirical phenomena, such as the

the choice overload ([19]).[19] Intuitively, because our information structure is exogeneouly fixed and does not change in different contexts, it is assuming the agent can learn as much information as available. Therefore, our model is not relevant when the main mechanism is likely driven by the agent's limited information capacity. A more plausible mechanism then is likely to cover endogeneous attention. See [12] for one such approach in explaining choice overload.

Similarly, our model do not cover limited memory. After the agent has seen $\{\mathbf{x}, \mathbf{y}, \mathbf{z}\}$, the same choice behavior is predicted for the choice set $\{\mathbf{x}, \mathbf{y}, (\mathbf{z})\}$ where $\mathbf{z}$ is shown but not available, and the choice set $\{\mathbf{x}, \mathbf{y}\}$ where $\mathbf{z}$ is removed. This does not explain the findings in [38] that the choice probability of $\mathbf{x}$ (the target) is significantly reduced following the removal of $\mathbf{z}$, although it does not recover fully to the level for which $\mathbf{z}$ was never shown. One way to model such an observation is to assume that the agent partially forgets what she has learned when the stimuli are removed. Due to the limited scope we do not discuss in detail the modeling of forgetfulness in this paper.

There are many papers in the literature explaining different contextual choices. Such as the drift diffusion models in neuroscience (see e.g. [32], [5], [49], [48] and [10], and [9] for a survey), and extensions or variations of Luce's logit model (see e.g. [25], [1] [34] and [7]). We have also previously discussed [20] and [26] in earlier sections. Most of these models do not study attributes as primitives. In contrast, because our model fundamentals are the attributes, we have the advantage to naturally make strong predictions for clearly dominating alternatives as in Theorem 34.

[19]Thanks for pointing out by the associate editor and the referees.

Bibliography

[1] Aguiar, V. 2015. Stochastic choice and fuzzy attension. *working paper*

[2] Ariely, D., G. Loewenstein, D. Prelec. 2003. 'Coherent Arbitrariness': Stable Demand Curves Without Stable Preferences". *The Quarterly Journal of Economics* **118**, 73-106.

[3] Bordalo, P., N. Gennaioli, A. Shleifer. 2013. Salience and consumer Choice. *Journal of Political Economy.*

[4] Block, H.D. and Marschak, J. 1960. Random orderings and stochastic theories of responses. in *Contributions to probability and statistics.* I. Olkin, S. Ghurye, W. Hoeffding, W. Madow and H. Mann (Eds.). pp 97-132.

[5] Busemeyer, J.R. J.T. Townsend. 1993. Decision Field Theory: A Dynamic-Cognitive Approach to Decision Making in an Uncertain Environment, *Psychological Review* **100** 432-459.

[6] de Clippel, G. and E. Kfir. 2012. Reason-based choice: a bargaining rationale for the attraction and compromise effects *Theoretical Economics*, **7** 125-162.

[7] Echenique, Saito, Tserenjigmid. 2015. The Perception Adjusted Luce Model *working paper*

[8] Falmagne, J.C. 1978. A Representation Theorem for Finite Random Scale System. it Journal of Mathematical Psychology, **18**, 52-72

[9] Fehr, E., and A. Rangel. 2011. Neuroeconomic Foundations of Economic Choice-Recent Advances. *Journal of Economic Perspectives* **25** 3-30.

[10] Fudenberg,D., P. Strack, T. Strzalecki. 2015. Stochastic choice and optimal sequential sampling *working paper*

[11] Gul, F., P. Natenzon and W. Pesendorfer. 2014. Random Choice as Behavioral Optimization. *Econometrica* **82** 1873-1912.

[12] Guo, L. 2016. Contextual Deliberation and Preference Construction. *Management Science* **62** 2977-2993

[13] Hausman, J.A. D.A. Wise. 1978. A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences. *Econometrica*

[14] Hedgcock, W., A.R. Rao and H.A. Chen. 2009. Could Ralph Nader's Entrance and Exit Have Helped Al Gore? The Impact of Decoy Dynamics on Consumer Choice. *Journal of Marketing Research*

[15] Highhouse, S. 1996. Context-dependent selection: The effects of decoy and phantom job candidates. *Organizational Behavior and Human Decision Processes* **65** 68-76.

[16] Hsee, C. K. 1996. The evaluability hypothesis: An explanation of preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes* **67** 247-257.

[17] Hsee, C. K., G. F. Loewenstein, S. Blount and M. H. Bazerman. 1999. Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin* **125** 576-590.

[18] Huber, J., J. W. Payne, C. Puto. 1982. Adding asymmetrically dominated alternatives: Violation of regularity and similarity hypothesis. *Journal of Consumer Research*

[19] Iyengar S.S., Lepper M.R. (2000) When choice is demotivating: Can one desire too much of a good thing? *J. Personality Soc. Psych.* 79 (6): 995-1006.

[20] Kamenica, E. 2008. Contextual inference in markets: On the informational content of product lines. *American Economic Review* **98** 2127-2149

[21] Koszegi, B., Rabin, M. 2006. A model of reference-dependent preferences. *Quarterly Journal of Economics*

[22] Koszegi, B., Szeidl, A. 2013. A Model of Focusing in Economic Choice. *Quarterly Journal of Economics* 53-104

[23] Luce, R. D. 1959. *Individual Choice Behavior: a Theoretical Analysis*, Wiley New York.

[24] Manzini, P., Mariotti, M. 2007. Sequentially Rationalizable Choice. *American Economic Review* **97**. No. 5

[25] Masatlioglu, Y., D. Nakajima, E.Y. Ozbay. 2012 Revealed attention. *The American Economic Review* **102** 2183-2205.

[26] Natenzon, P. 2016. Random Choice and Learning. *Working paper*

[27] Ok E. A., P. Ortoleva, G. Riella. 2015. Revealed (P)Reference Theory. *AMERICAN ECONOMIC REVIEW* **105** 299-321.

[28] Pettibone, J. C., D. H. Wedell. 2000. Examining Models of Nondominated Decoy Effects across Judgment and Choice. *Organizational Behavior and Human Decision Processes* **81** No.2 300-328.

[29] Pettibone, J. C., D. H. Wedell. 2007. Testing Alternative Explanations of Phantom Decoy Effects. *Journal of Behavioral Decision Making* **20** 323-341

[30] Plous, Scott (1993). *The Psychology of Judgment and Decision Making*. McGraw-Hill. 38-41.

[31] Pratkanis, A. R., P. H. Farquhar. 1992. A brief history of research on phantom alternatives: Evidence for several empirical generalization about phantoms. *Basic and applied social psychology* **13** 103-122

[32] Ratcliff, R. 1978. A Theory of Memory Retrieval. *Psychological Review* **85** 59-108.

[33] Rieskamp, J., J. R. Busemeyer, B. A. Mellers. 2006. Extending the bounds of rationality: evidence and Theories of preferential choice *Journal of Economic Literature*

[34] Ravid, D. 2015. Focus, Then Compare *working paper*.

[35] Savage, L. J. (1954). *The Foundations of Statistics*. Wiley, New York.

[36] Schwarz, N., and Bless, H. (1992). *Scandals and the Public's Trust in Politicians: Assimilation and Contrast Effects*. Personality and Social Psychology Bulletin, **18**, 574-579.

[37] Simonson, I. 1989. Choice Based on Reasons: The Case of Attraction and Compromise Effects. *Journal of Consumer Research* **16** 158-174

[38] Sivakumar, K., J. Cherian. 1995. Role of product entry and exit on the attraction effect. *Marketing Letters* **6**, 45-51.

[39] Soltani, A., B. De Martino, C. Camerer. 2012. A RangeNormalization Model of Context-Dependent Choice: A New Model and Evidence. *PLoS computational biology* **8**

[40] Tserenjigmid, G. 2016. Choosing with the Worst in Mind: A Reference-Dependent Model. *working paper*

[41] Thurstone, L.L., 1927. Psychophysical analysis. *The American Journal of Psychology* **38** 368-389.

[42] Tversky, A. 1969. Intransitivity of Preferences. *Psychological Review* **76** (1), 31-48.

[43] Tversky, A. 1972. Elimination by Aspects: A Theory of Choice. *Psychological Review* **79** 281-299.

[44] Tversky, A., J.E. Russo. 1969. Substitutability and similarity in binary choices. *Journal of mathematical Psychology* 1-12.

[45] Tversky, A., I. Simonson. 1993. Context-Dependent Preferences *Management Science* **39** (10) 1179-1189.

[46] Wedell, D. H. 1991. Distinguishing among models of contextually induced preference reversals. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **17** (4) 767-778.

[47] Wernerfelt, B. 1995. A rational reconstruction of the compromise effect: Using market data to infer utilities. *Journal of Consumer Research* 627-633.

[48] Woodford, M. 2014. Stochastic choice: An optimizing neuroeconomic model *The American Economic Review* **104** 495-500.

[49] Usher, M., J. L. McClelland. 2004. Loss aversion and inhibition in dynamical models of multialternative choice. *Psychological review* **111**

Appendix A

Additional Proofs

A.1 A Test for Sparsity

A.1.1 Proof of Lemma 1

Proof. We devide the proof into three steps. The first step is to show there is a unique solution. The second step is to construct an upperbound and the third step is to construct a lower bound.

Step 1. We show the set of equations have a unique solution. Observe that $\frac{p_a(x)}{q_0(x)}$ is now increasing in x and reaches maximum $1/c$ on $[x^*, \infty)$. Let $v := \frac{1-\sigma^2}{p} + \sigma^2 s^2/n$. The relation between x^* and c can be solved from $\frac{p_a(x^*)}{q_0(x^*)} = \frac{1}{c}$ to be

$$\exp\left(\frac{x^*}{2\sigma^2 s^2/n} - \frac{x^*}{2v}\right) = \frac{1-\epsilon}{c}\sqrt{\frac{v}{\sigma^2 s^2/n}}.$$

So there is an inverse relationship between x^* and c. Substituting into $\int q_0 = 1$ we get

$$\int_0^{x^*} (1-\epsilon)p_0(t)\mathrm{d}t + (1-\epsilon)\sqrt{\frac{v}{\sigma^2 s^2/n}} e^{\left(\frac{1}{2v} - \frac{1}{2\sigma^2 s^2/n}\right)x^*} \int_{x^*}^{\infty} p_a(t)\mathrm{d}t = 1.$$

By differentiating the LHS with respect to x, we see that the LHS becomes

$$\begin{aligned}
&(1-\epsilon)\frac{e^{-\frac{x}{2\sigma^2 s^2/n}}}{\sqrt{2\pi\sigma^2(s^2/n)x}} - (1-\epsilon)\sqrt{\frac{v}{\sigma^2 s^2/n}}e^{\left(\frac{1}{2v}-\frac{1}{2\sigma^2 s^2/n}\right)x}\frac{e^{-\frac{x}{2v}}}{\sqrt{2\pi v x}} \\
&+(1-\epsilon)\sqrt{\frac{v}{\sigma^2 s^2/n}}e^{\left(\frac{1}{2v}-\frac{1}{2\sigma^2 s^2/n}\right)x}\left(\frac{1}{2v}-\frac{1}{2\sigma^2 s^2/n}\right)\int_x^\infty p_a(s)\mathrm{d}s \\
=&(1-\epsilon)\sqrt{\frac{v}{\sigma^2 s^2/n}}e^{\left(\frac{1}{2v}-\frac{1}{2\sigma^2 s^2/n}\right)x}\left(\frac{1}{2v}-\frac{1}{2\sigma^2 s^2/n}\right)\int_x^\infty p_a(s)\mathrm{d}s < 0,
\end{aligned}$$

because $v > \sigma^2 s^2/n$.

Since the LHS of the equation is decreasing in x, in order for a solution to exist, it is necessary that when $x = 0$ we have $LHS > 1$. In otherwords,

$$1-\epsilon > \sqrt{\frac{\sigma^2 s^2/n}{v}} \Rightarrow v > \frac{\sigma^2 s^2/n}{(1-\epsilon)^2}.$$

This is guarrenteed when $\epsilon \to 0$.

Before step 2, we observe from the x^* equation that

$$\begin{aligned}
&(1-\epsilon)\int_0^{x^*}\frac{e^{-t/(2\sigma^2 s^2/n)}}{\sqrt{2\pi\sigma^2(s^2/n)t}}\mathrm{d}t + (1-\epsilon)\sqrt{\frac{v}{\sigma^2 s^2/n}}e^{\left(\frac{1}{v}-\frac{1}{\sigma^2 s^2/n}\right)\frac{x^*}{2}}\int_{x^*/v}^\infty\frac{e^{-t/2}}{\sqrt{2\pi t}}\mathrm{d}t = 1 \\
\Rightarrow&(1-\epsilon)\int_0^{\sqrt{\frac{x^*}{2\sigma^2 s^2/n}}}\frac{2}{\sqrt{\pi}}e^{-t^2}\mathrm{d}t + (1-\epsilon)\sqrt{\frac{v}{\sigma^2 s^2/n}}e^{\left(\frac{1}{v}-\frac{1}{\sigma^2 s^2/n}\right)\frac{x^*}{2}}\int_{\sqrt{\frac{x^*}{2v}}}^\infty\frac{2}{\sqrt{\pi}}e^{-t^2}\mathrm{d}t = 1 \\
\Rightarrow&(1-\epsilon)\,\mathrm{erf}\left(\sqrt{\frac{x^*}{2\sigma^2 s^2/n}}\right) + (1-\epsilon)\sqrt{\frac{v}{\sigma^2 s^2/n}}e^{\left(\frac{1}{v}-\frac{1}{\sigma^2 s^2/n}\right)\frac{x^*}{2}}\,\mathrm{erfc}\left(\sqrt{\frac{x^*}{2v}}\right) = 1.
\end{aligned}$$

where $\mathrm{erfc}(x) := \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}\mathrm{d}t$ and $\mathrm{erf}(x) := 1 - \mathrm{erfc}(x)$.

Step 2. We now construct an upper bound. To construct the upperbound for x^*, we first substitute the following into the LHS of the above equation

$$x = \frac{v\sigma^2 s^2/n}{(1-\sigma^2)/p}\ln\left(\frac{vn}{s^2}\frac{a}{\sigma^2}\right) = \frac{\sigma^2 s^2}{n}\frac{\sigma^2 s^2 p + (1-\sigma^2)n}{(1-\sigma^2)n}\ln\left(\frac{vn}{s^2}\frac{a}{\sigma^2}\right),$$

where a depends on ϵ and is to be determined later. We have

$$LHS = (1-\epsilon)\,\mathrm{erf}\left(\sqrt{\frac{1}{2}(1+\xi)\ln(a(1+1/\xi))}\right) + (1-\epsilon)a^{-1/2}\,\mathrm{erfc}\left(\sqrt{\frac{1}{2}\xi\ln(a(1+1/\xi))}\right)$$

where $\xi = \frac{\sigma^2}{1-\sigma^2}\frac{ps^2}{n}$. Since LHS is monotonically decreasing in x, it suffices to show that for certain a, $LHS < 1$. Then we conclude that the solution for x^* would be less than this value. In the following, we use the following bounds for the function erfc derived from Formula 7.1.13 of Abramowitz and Stegun (1964). When $x \geq 0$,

$$\frac{2}{\sqrt{\pi}}\frac{e^{-x^2}}{2x+\sqrt{2}} \leq \frac{2}{\sqrt{\pi}}\frac{e^{-x^2}}{x+\sqrt{x^2+2}} < \mathrm{erfc}(x) \leq \frac{2}{\sqrt{\pi}}\frac{e^{-x^2}}{x+\sqrt{x^2+4/\pi}} < \frac{1}{\sqrt{\pi}}\frac{e^{-x^2}}{x},$$

and in particular, it is well-known that $\mathrm{erfc}(x) \leq \frac{2}{\sqrt{\pi}}e^{-x^2}$ for $x \geq 0$.

$$\begin{aligned}
\frac{LHS}{1-\epsilon} =& \,\mathrm{erf}\left(\sqrt{\frac{1}{2}(1+\xi)\ln(a(1+1/\xi))}\right) + a^{-1/2}\,\mathrm{erfc}\left(\sqrt{\frac{1}{2}\xi\ln(a(1+1/\xi))}\right) \\
\leq& 1 - \frac{2}{\sqrt{\pi}}\frac{\exp\left(-\frac{1}{2}(1+\xi)\ln(a(1+1/\xi))\right)}{\sqrt{2}+2\sqrt{\frac{1}{2}(1+\xi)\ln(a(1+1/\xi))}} + a^{-1/2}\frac{2}{\sqrt{\pi}}\exp\left(-\frac{1}{2}\xi\ln(a(1+1/\xi))\right) \\
=& 1 - \frac{a^{-1/2}2}{\sqrt{\pi}}\frac{(1+1/\xi)^{-1/2}\exp\left(-\frac{1}{2}\xi\ln(a(1+1/\xi))\right)}{\sqrt{2}+2\sqrt{\frac{1}{2}(1+\xi)\ln(a(1+1/\xi))}} + \frac{a^{-1/2}2}{\sqrt{\pi}}\exp\left(-\frac{1}{2}\xi\ln(a(1+1/\xi))\right)
\end{aligned}$$

Now let a satisfy $a^{-1/2} \times \frac{2}{\sqrt{\pi}} = \frac{\epsilon}{1-\epsilon}$, i.e. $a = \frac{4}{\pi}\left(\frac{1-\epsilon}{\epsilon}\right)^2$, and get

$$\begin{aligned}\frac{LHS}{1-\epsilon} =& 1 - \frac{\epsilon \exp\left(-\frac{1}{2}\xi \ln\left(a(1+1/\xi)\right)\right)}{(1-\epsilon)} \frac{(1+1/\xi)^{-\frac{1}{2}}}{\sqrt{2} + 2\sqrt{\frac{1}{2}(1+\xi)\ln\left(a(1+1/\xi)\right)}} \\ &+ \frac{\epsilon \exp\left(-\frac{1}{2}\xi \ln\left(a(1+1/\xi)\right)\right)}{(1-\epsilon)} \\ =& 1 + \frac{\epsilon \exp\left(-\frac{1}{2}\xi \ln\left(a(1+1/\xi)\right)\right)}{(1-\epsilon)} \left(1 - \frac{(1+1/\xi)^{-\frac{1}{2}}}{\sqrt{2} + 2\sqrt{\frac{1}{2}(1+\xi)\ln\left(a(1+1/\xi)\right)}}\right)\end{aligned}$$

Since $a > 1$ for all ϵ small enough and $\xi > 0$, we have

$$\exp\left(-\frac{1}{2}\xi \ln\left(a(1+1/\xi)\right)\right) \leq 1, \text{ and } 0 < \frac{(1+1/\xi)^{-\frac{1}{2}}}{\sqrt{2} + 2\sqrt{\frac{1}{2}(1+\xi)\ln\left(a(1+1/\xi)\right)}} < 1,$$

and therefore

$$LHS < (1-\epsilon)(1 + \frac{\epsilon}{1-\epsilon}) = 1.$$

Hence we conclude that

$$\frac{x^*}{\sigma^2 s^2/n} \leq \frac{\sigma^2 s^2 p + (1-\sigma^2)n}{(1-\sigma^2)n} \ln\left(\frac{vn}{s^2\sigma^2}\frac{4}{\pi}\left(\frac{1-\epsilon}{\epsilon}\right)^2\right).$$

Step 3. Now we establish an lowerbound for x^*. To do this, we show that by substituting in

$$x = \frac{v\sigma^2 s^2/n}{(1-\sigma^2)/p} \ln\left(\frac{vn}{s^2}\frac{a}{\sigma^2}\right) = \frac{\sigma^2 s^2}{n} \frac{\sigma^2 s^2 p + (1-\sigma^2)n}{(1-\sigma^2)n} \ln\left(\frac{vn}{s^2}\frac{a}{\sigma^2}\right),$$

for some other values of a (depending on ϵ) than before, we have $LHS > 1$ asymptotically.

As before, we start with

$$\begin{aligned}
\frac{LHS}{1-\epsilon} =& \operatorname{erf}\left(\sqrt{\frac{1}{2}\left(1+\xi\right)\ln\left(a(1+1/\xi)\right)}\right) + a^{-1/2}\operatorname{erfc}\left(\sqrt{\frac{1}{2}\xi\ln\left(a(1+1/\xi)\right)}\right) \\
\geq& 1 - \frac{1}{\sqrt{\pi}}\frac{\exp\left(-\frac{1}{2}\left(1+\xi\right)\ln\left(a(1+1/\xi)\right)\right)}{\sqrt{\frac{1}{2}\left(1+\xi\right)\ln\left(a(1+1/\xi)\right)}} + a^{-1/2}\operatorname{erfc}\left(\sqrt{\frac{1}{2}\xi\ln\left(a(1+1/\xi)\right)}\right) \\
=& 1 - \frac{a^{-\frac{1}{2}}}{\sqrt{\pi}}\left(1+1/\xi\right)^{-\frac{1}{2}}\frac{\exp\left(-\frac{1}{2}\xi\ln\left(a(1+1/\xi)\right)\right)}{\sqrt{\frac{1}{2}\left(1+\xi\right)\ln\left(a(1+1/\xi)\right)}} + a^{-\frac{1}{2}}\operatorname{erfc}\left(\sqrt{\frac{1}{2}\xi\ln\left(a(1+1/\xi)\right)}\right) \\
\geq& 1 - \frac{a^{-\frac{1}{2}}}{\sqrt{\pi}}\left(1+1/\xi\right)^{-\frac{1}{2}}\frac{\exp\left(-\frac{1}{2}\xi\ln\left(a(1+1/\xi)\right)\right)}{\sqrt{\frac{1}{2}\left(1+\xi\right)\ln\left(a(1+1/\xi)\right)}} + \frac{a^{-\frac{1}{2}}}{\sqrt{\pi}}\frac{\exp\left(-\frac{1}{2}\xi\ln\left(a(1+1/\xi)\right)\right)}{\sqrt{2}/2+\sqrt{\frac{1}{2}\xi\ln\left(a(1+1/\xi)\right)}} \\
=& 1 + \frac{a^{-\frac{1}{2}}}{\sqrt{\pi}}e^{-\frac{1}{2}\xi\ln(a(1+1/\xi))}\left(\frac{1}{\sqrt{2}/2+\sqrt{\frac{1}{2}\xi\ln\left(a(1+1/\xi)\right)}} - \frac{\left(1+1/\xi\right)^{-\frac{1}{2}}}{\sqrt{\frac{1}{2}\left(1+\xi\right)\ln\left(a(1+1/\xi)\right)}}\right) \\
=& 1 + a^{-\frac{1}{2}(1+\xi)}\frac{e^{-\frac{1}{2}\xi\ln(1+1/\xi)}}{\sqrt{\pi}}\left(\frac{1}{\sqrt{2}/2+\sqrt{\frac{1}{2}\xi\ln\left(a(1+1/\xi)\right)}} - \frac{\xi}{\left(1+\xi\right)\sqrt{\frac{1}{2}\xi\ln\left(a(1+1/\xi)\right)}}\right) \\
=& 1 + a^{-\frac{1}{2}(1+\xi)}\frac{e^{-\frac{1}{2}\xi\ln(1+1/\xi)}}{\sqrt{\pi}}\frac{\sqrt{\frac{1}{2}\xi\ln\left(a(1+1/\xi)\right)} - \xi\sqrt{2}/2}{\left(\sqrt{2}/2+\sqrt{\frac{1}{2}\xi\ln\left(a(1+1/\xi)\right)}\right)\left(1+\xi\right)\sqrt{\frac{1}{2}\xi\ln\left(a(1+1/\xi)\right)}} \\
=& 1 + a^{-\frac{1}{2}(1+\xi)}\frac{e^{-\frac{1}{2}\xi\ln(1+1/\xi)}}{\sqrt{\pi}}\frac{1}{\left(\sqrt{2}/2+\sqrt{\frac{1}{2}\xi\ln\left(a(1+1/\xi)\right)}\right)\left(1+\xi\right)}\left(1 - \frac{1}{\sqrt{\frac{1}{\xi}\ln\left(a(1+1/\xi)\right)}}\right).
\end{aligned}$$

Pick an arbitrary small $\delta \in (0,1)$, and let $a^{-\frac{1}{2}(1+\kappa)} := \left(\frac{\epsilon}{1-\epsilon}\right)^{1-\delta}$. In other words, $a :=$

$\left(\frac{1-\epsilon}{\epsilon}\right)^{\frac{2(1-\delta)}{1+\kappa}} > 1$ for any $\epsilon < 1/2$. Since $\xi \leq \kappa$, we have

$$
\begin{aligned}
& a^{-\frac{1}{2}(1+\xi)} \frac{e^{-\frac{1}{2}\xi \ln(1+1/\xi)}}{\sqrt{\pi}} \frac{1}{\left(\sqrt{2}/2 + \sqrt{\frac{1}{2}\xi \ln\left(a(1+1/\xi)\right)}\right)(1+\xi)} \left(1 - \frac{1}{\sqrt{\frac{1}{\xi}\ln\left(a(1+1/\xi)\right)}}\right) \\
\geq & a^{-\frac{1}{2}(1+\kappa)} \frac{e^{-\frac{1}{2}\kappa \ln(1+1/\kappa)}}{\sqrt{\pi}} \frac{1}{\left(\sqrt{2}/2 + \sqrt{\frac{1}{2}\kappa \ln\left(a(1+1/\kappa)\right)}\right)(1+\kappa)} \left(1 - \frac{1}{\sqrt{\frac{1}{\kappa}\ln\left(a(1+1/\kappa)\right)}}\right) \\
= & \frac{\epsilon}{1-\epsilon} \left(\frac{1-\epsilon}{\epsilon}\right)^{\delta} \frac{e^{-\frac{1}{2}\kappa \ln(1+1/\kappa)}}{\sqrt{\pi}} \frac{1}{\left(\sqrt{2}/2 + \sqrt{\frac{1}{2}\kappa \ln\left(a(1+1/\kappa)\right)}\right)(1+\kappa)} \left(1 - \frac{1}{\sqrt{\frac{1}{\kappa}\ln\left(a(1+1/\kappa)\right)}}\right) \\
> & \frac{\epsilon}{1-\epsilon} \text{ as } \epsilon \to 0
\end{aligned}
$$

since $\sqrt{\ln a} = \sqrt{\frac{2(1-\delta)}{1+\kappa} \ln\left(\frac{1-\epsilon}{\epsilon}\right)} = o\left(\left(\frac{1-\epsilon}{\epsilon}\right)^{\delta}\right)$ as $\epsilon \to 0$. Therefore $LHS > \left(1 + \frac{\epsilon}{1-\epsilon}\right)(1-\epsilon) = 1$ for all ϵ small enough.

This shows that for any $\delta \in (0,1)$

$$
\frac{x^*}{\sigma^2 s^2/n} \geq \frac{\sigma^2 s^2 p + (1-\sigma^2)n}{(1-\sigma^2)n} \ln\left(\frac{vn}{s^2\sigma^2} a\right),
$$

where $a := \left(\frac{1-\epsilon}{\epsilon}\right)^{\frac{2(1-\delta)}{1+\kappa}}$.

□

A.1.2 Proof of Theorem 2

Proof. By definition,

$$
diag\left(\sqrt{\frac{n}{\sigma^2 s_1^2}}, \ldots, \sqrt{\frac{n}{\sigma^2 s_p^2}}\right) \hat{\beta} =_d e + diag\left(\sqrt{\frac{n}{\sigma^2 s_1^2}}, \ldots, \sqrt{\frac{n}{\sigma^2 s_p^2}}\right) \beta
$$

where e is as defined above and $\beta_i = z_i\gamma_i$ where $\gamma_i \sim F$ for some $F \in \mathcal{F}$. Consider the i-th term in the test statistic

$$\min\{\frac{\hat{\beta}_i^2}{\sigma^2 s_i^2/n}, \frac{x_i^*}{\sigma^2 s_i^2/n}\}.$$

Under the null hypothesis, conditional on $z_i = 0$, $\beta_i = 0$ for $\beta_i = z_i\gamma_i$. And

$$\begin{aligned}\min\{\frac{\hat{\beta}_i^2}{\sigma^2 s_i^2/n}, \frac{x_i^*}{\sigma^2 s_i^2/n}\} =& \min\{e_i^2 + 2e_i\sqrt{\frac{n}{\sigma^2 s_i^2}}\beta_i + \frac{\beta_i^2}{\sigma^2 s_i^2/n}, \frac{x_i^*}{\sigma^2 s_i^2/n}\} \\ =&(1 - z_i)\min\{e_i^2, \frac{x_i^*}{\sigma^2 s_i^2/n}\} + z_i\frac{x_i^*}{\sigma^2 s_i^2/n}.\end{aligned}$$

On the other hand, conditional on $z_i = 1$, we have

$$\min\{\frac{\hat{\beta}_i^2}{\sigma^2 s_i^2/n}, \frac{x_i^*}{\sigma^2 s_i^2/n}\} \leq \frac{x_i^*}{\sigma^2 s_i^2/n} = (1 - z_i)\min\{e_i^2, \frac{x_i^*}{\sigma^2 s_i^2/n}\} + z_i\frac{x_i^*}{\sigma^2 s_i^2/n}.$$

Since this holds for each $i = 1, \ldots, p$, we conclude that under $H_0(\epsilon)$, $T \leq_{1st} S$.

To see the bound is tight, consider in $H_0(\epsilon)$ a sequence of $\{F_k\}_{k\in\mathbb{N}} \subseteq \mathcal{F}$ that diverges to infinity: for all $k \in \mathbb{N}$, $F_k(k) - F_k(-k) = 0$. For each $i = 1, \ldots, p$, conditional on$z_i = 0$,

$$\min\{\frac{\hat{\beta}_i^2}{\sigma^2 s_i^2/n}, \frac{x_i^*}{\sigma^2 s_i^2/n}\} = (1 - z_i)\min\{e_i^2, \frac{x_i^*}{\sigma^2 s_i^2/n}\} + z_i\frac{x_i^*}{\sigma^2 s_i^2/n}$$

as before. Conditional on $z_i = 1$, $\min\{\frac{\hat{\beta}_i^2}{\sigma^2 s_i^2/n}, \frac{x_i^*}{\sigma^2 s_i^2/n}\}$ converges in probability to $\frac{x_i^*}{\sigma^2 s_i^2/n} = (1 - z_i)\min\{e_i^2, \frac{x_i^*}{\sigma^2 s_i^2/n}\} + z_i\frac{x_i^*}{\sigma^2 s_i^2/n}$ for $\beta_i^2 = \gamma_i^2$ converges in probability to infinity along the sequence F_k. Since such a convergence holds for all $i = 1, \ldots, p$, the statistics T converges in distribution to S along F_k.

□

A.1.3 Proof of Proposition 5

We first introduce a lemma.

Lemma 37. *Let the minimal eigenvalues of $X'X/n$ be λ and let the ith diagonal entry of $(X'X/n)^{-1}$ be s_i^2. Then $1 \leq s_i^2 \leq 1/\lambda$ for all $i = 1, \ldots, p$.*

Proof. Let λ_i for $i = 1, \ldots, p$ be the eigenvalues of $X'X/n$. Write the eigenvalue decomposition as

$$X'X/n = Q\Lambda Q'$$

where $\Lambda = diag(\lambda_1, \ldots, \lambda_p)$. The ii-th entry of $X'X/n$ is $1 = \sum_{j=1}^p Q_{ij}^2 \lambda_j$. Similarly for $(X'X/n)^{-1} = Q\Lambda^{-1}Q'$, we have $s_i^2 = \sum_{j=1}^p Q_{ij}^2/\lambda_j$. By the inequality of weighted harmonic mean and arithmetic mean, we have

$$\frac{1}{s_i^2} = \frac{1}{\sum_{j=1}^p Q_{ij}^2/\lambda_j} \leq \sum_{j=1}^p Q_{ij}^2 \lambda_j = 1.$$

The other inequality follows from Schur-Horn Theorem (see Schur (1923) and Horn (1957)) that $1/\lambda \geq \max_i\{s_i^2\}$. □

Now we proceed to the proof of Proposition 5.

Proof. On one hand,by Lemma 1,

$$\frac{x_i^*}{\sigma^2 s_i^2/n} \geq \ln\left(\frac{v_i n}{s_i^2\sigma^2}\left(\frac{1-\epsilon}{\epsilon}\right)^{\frac{2(1-\delta)}{1+\kappa}}\right) \geq \frac{2(1-\delta)}{1+\kappa}\ln\frac{1}{\epsilon} \geq c_1 \ln n$$

for some $c_1 > 0$.

On the other hand, by Lemma 1, we have

$$\frac{x_i^*}{\sigma^2 s_i^2/n} = O\left(\ln\left(\frac{v_i n}{s_i^2\sigma^2}\frac{(1-\epsilon)^2}{\epsilon^2}\right)\right) \leq O\left(\ln\left(\left(\frac{1-\sigma^2}{\sigma^2}\frac{n}{p}+1\right)\frac{1}{\epsilon^2}\right)\right) = O(\ln n),$$

where in the inequality we used $s_i^2 \geq 1$ from the Lemma 37, and in the last equality we used the assumption of asymptotic sparsity. □

A.1.4 Proof of Theorem 6

We need to first prepare a lemma. This lemma may be of interest in its own. It states that the empirical distribution of $\hat{\beta}_i^2/v_i$ converges to the cdf of the χ_1^2 distribution.

Lemma 38. *Let the minimal eigenvalues of $X'X/n$ be λ and let $v_i = \frac{1-\sigma^2}{p} + \frac{\sigma^2 s_i^2}{n}$. Suppose there exists some constant $\kappa > 0$ such that $\frac{n\lambda}{p} \geq \kappa$. Then under the alternative, the empirical distribution of $\hat{\beta}_i^2/v_i$ converges to the cdf of χ_1^2 as $p(n) \to \infty$.*

Proof. Since under the alternative, the asymptotic distribution for $\hat{\beta}$ is

$$\hat{\beta} \sim \mathcal{N}\left(0, \frac{1-\sigma^2}{p} I_{p\times p} + \sigma^2 (X'X)^{-1}\right).$$

Under scaling by $D := diag\left(v_1^{-1/2}, \ldots, v_p^{-1/2}\right)$, the distribution becomes

$$D\hat{\beta} \sim \mathcal{N}\left(0, D\left(\frac{1-\sigma^2}{p} I_{p\times p} + \sigma^2 (X'X)^{-1}\right) D\right).$$

It is clear that in the above variance matrix, all entries on the main diagonal are 1. Let the minimal eigenvalues of $X'X/n$ be λ, it is clear that for any $\delta > 0$,

$$\begin{aligned} D\left(\frac{1-\sigma^2}{p} I_{p\times p} + \sigma^2 (X'X)^{-1}\right) D \leq & D\left(\frac{1-\sigma^2}{p} + \frac{\sigma^2}{n\lambda}\right) I_{p\times p} D \\ \leq & \frac{1-\sigma^2+\sigma^2/\kappa}{1-\sigma^2} I_{p\times p}. \end{aligned}$$

Let the eigenvalues of the variance matrix of $D\hat{\beta}$ be $r_1, \ldots r_p$. The normalized Frobenius norm of the above variance matrix is given by

$$||D\left(\frac{1-\sigma^2}{p} I_{p\times p} + \sigma^2 (X'X)^{-1}\right) D||_2 := \frac{1}{p}\left(\sum_{i=1}^{p} r_i^2\right)^{1/2} \leq \frac{1}{p}\left(\sum_{i=1}^{p}\left(\frac{1-\sigma^2+\sigma^2/\kappa}{1-\sigma^2}\right)^2\right)^{1/2} \to 0$$

as $p \to \infty$. Now we apply Theorem 1 of Azriel and Schwartzman (2015), and conclude that

the empirical distribution of the entries of $D\hat{\beta}$ converges to the standard normal, and hence the empirical distribution of $\hat{\beta}_i^2/v_i$ converges to χ_1^2 as $n, p \to \infty$. □

Now we proceed to the proof of Theorem 6.

Proof. We have seen that Lemma 37 implies $s_i^2 p/n \leq 1/\kappa$. Let $v_i := \frac{1-\sigma^2}{p} + \frac{\sigma^2 s_i^2}{n}$. There exists a constant $K_0 > 0$, such that

$$\begin{aligned} T =& \frac{1}{p}\sum_{i=1}^{p} \frac{(1-\sigma^2)n}{(1-\sigma^2)n + \sigma^2 s_i^2 p} \times \min\{\frac{\hat{\beta}_i^2}{\sigma^2 s_i^2/n}, \frac{x_i^*}{\sigma^2 s_i^2/n}\} \\ \geq& \frac{1}{p}\sum_{i=1}^{p} K_0 \min\{\frac{\hat{\beta}_i^2}{\sigma^2 s_i^2/n}, \frac{x_i^*}{\sigma^2 s_i^2/n}\} \\ \geq& \frac{1}{p}\sum_{\frac{\hat{\beta}_i^2}{v_i} \geq \frac{x_i^*}{v_i}} K_0 \frac{x_i^*}{\sigma^2 s_i^2/n} \geq \frac{1}{p}\sum_{\frac{\hat{\beta}_i^2}{v_i} \geq \frac{x_i^*}{v_i}} K_0 c_1 \ln n. \end{aligned}$$

Since Lemma 1 and Lemma 37 implies there exists constant K_1 that for all i,

$$\begin{aligned} \frac{x_i^*}{v_i} \leq& \frac{\sigma^2}{1-\sigma^2}\frac{p}{n\lambda} \ln\left(\left(1 + \frac{(1-\sigma^2)n\lambda}{\sigma^2 p}\right)\frac{4}{\pi}\left(\frac{1-\epsilon}{\epsilon}\right)^2\right) \\ \leq& K_1 \frac{p}{n\lambda} \ln\left(1 + \frac{(1-\sigma^2)n\lambda}{\sigma^2 p}\right) + K_1 \frac{p}{n\lambda}\ln n \\ \leq& K_1/\kappa \qquad \text{as } \frac{p}{n\lambda}\ln\left(1 + \frac{(1-\sigma^2)n\lambda}{\sigma^2 p}\right) \to 0 \end{aligned}$$

by the Condition 4 and the assumption that $\frac{n\lambda}{p} \geq \kappa \ln n$. Therefore

$$T \geq \frac{1}{p}\sum_{\frac{\hat{\beta}_i^2}{v_i} \geq K_1/\kappa} K_0 c_1 \ln n \to \Pr(\chi_1^2 \geq K_1/\kappa) K_0 c_1 \ln n \to \infty,$$

where $\frac{1}{p}\sum_{\frac{\hat{\beta}_i^2}{v_i} \geq K_1/\kappa} 1 \to \Pr(\chi_1^2 \geq K_1/\kappa)$ by Lemma 38. Since we have shown previously that the test statistics under the null is of order $O_p(1)$, the proof is complete.

□

A.1.5 Proof of Proposition 7

Before proving the proposition, we need to prepare the following lemma.

Lemma 39. *Let c_1, c_2 be two positive constants, and the random vector $(x, y)^t \sim \mathcal{N}(\mathbf{0}, M)$ where $M := \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$. For any $c_1, c_2 \geq 0$, we have $Cov\left(\min\{x^2, c_1\}, \min\{y^2, c_2\}\right) \geq 0$.*

Proof. The density of (x, y) can be written as

$$f(x)f(y|x) = \frac{\exp\left(-\frac{1}{2}x^2\right)}{\sqrt{2\pi}} \frac{\exp\left(-\frac{1}{2}\frac{(y-\rho x)^2}{1-\rho^2}\right)}{\sqrt{2\pi(1-\rho^2)}}$$

By definition,

$$\begin{aligned}
&Cov\left(\min\{x^2, c_1\}, \min\{y^2, c_2\}\right) \\
=&\mathbb{E}\left[\min\{x^2, c_1\} \times \min\{y^2, c_2\}\right] - \mathbb{E}[\min\{x^2, c_1\}]\mathbb{E}[\min\{y^2, c_2\}] \\
=&\int_{\mathbb{R}} \frac{\exp\left(-\frac{1}{2}x^2\right)}{\sqrt{2\pi}} \min\{x^2, c_1\} \int_{\mathbb{R}} \min\{y^2, c_2\} \frac{\exp\left(-\frac{1}{2}\frac{(y-\rho x)^2}{1-\rho^2}\right)}{\sqrt{2\pi(1-\rho^2)}} \mathrm{d}y\mathrm{d}x - \mathbb{E}[\min\{x^2, c_1\}]\mathbb{E}[\min\{y^2, c_2\}] \\
=&\int_{\mathbb{R}} \min\{x^2, c_1\} \int_{\mathbb{R}} \min\{(\sqrt{1-\rho^2}s + \rho x)^2, c_2\} \mathrm{d}\Phi(s)\ \mathrm{d}\Phi(x) - \mathbb{E}[\min\{x^2, c_1\}]\mathbb{E}[\min\{y^2, c_2\}] \\
=&\int_{\mathbb{R}} \min\{x^2, c_1\} h(x) \mathrm{d}\Phi(x) - \int_{\mathbb{R}} \min\{x^2, c_1\} \mathrm{d}\Phi(x) \int_{\mathbb{R}} \min\{y^2, c_2\} \mathrm{d}\Phi(y)
\end{aligned}$$

where Φ is the standard normal c.d.f., and $h(x) := \int_{\mathbb{R}} \min\{(\sqrt{1-\rho^2}s + \rho x)^2, c_2\} \mathrm{d}\Phi(s)$. It is clear that $h(x) = h(-x)$, we can write $h(x) = h(\sqrt{x^2})$. So with a change of variable $x^2 = t$, we have

$$\begin{aligned}
&Cov\left(\min\{x^2, c_1\}, \min\{y^2, c_2\}\right) \\
=&\int_{\mathbb{R}} \min\{t, c_1\} h(\sqrt{t}) \mathrm{d}\chi_1^2(t) - \int_{\mathbb{R}} \min\{t, c_1\} \mathrm{d}\chi_1^2(t) \int_{\mathbb{R}} \min\{t, c_2\} \mathrm{d}\chi_1^2(t)
\end{aligned}$$

where χ_1^2 is the Chi-square c.d.f of one degree freedom. Since we have $\int_{\mathbb{R}} h(\sqrt{t})\mathrm{d}\chi_1^2(t) = \int_{\mathbb{R}^2} \min\{y^2, c_2\} f(x) f(y|x)\mathrm{d}y\mathrm{d}x = \int_{\mathbb{R}} \min\{y^2, c_2\}\mathrm{d}\Phi(y)$, Chebyshev's sum inequality implies $Cov\left(\min\{x^2, c_1\}, \min\{y^2, c_2\}\right) > 0$ as long as if $h(\sqrt{t})$ is an increasing function in t for $t > 0$.

To see $h(\sqrt{t})$ is indeed increasing, observe that if s is a standard normal random variable, then $\left(s + \frac{\rho}{\sqrt{1-\rho^2}}\sqrt{t}\right)^2$ follows a noncentral Chi-squared distribution $\chi_1^2\left(\frac{\rho^2 t}{1-\rho^2}\right)$ with noncentrality parameter $\frac{\rho^2 t}{1-\rho^2}$. Since the noncentrality parameter has the monotone likelihood ratio property, it follows that for any $0 \le t_1 < t_2$, the distribution $\chi_1^2\left(\frac{\rho^2 t_2}{1-\rho^2}\right)$ first order stochastically dominates the distribution $\chi_1^2\left(\frac{\rho^2 t_1}{1-\rho^2}\right)$. Since

$$h(\sqrt{t}) := \int_{\mathbb{R}} \min\{(\sqrt{1-\rho^2}s + \rho\sqrt{t})^2, c_2\}\mathrm{d}\Phi(s) = \mathbb{E}[\min\{(1-\rho)s', c_2\}]$$

where $s' := \left(s + \frac{\rho}{\sqrt{1-\rho^2}}\sqrt{t}\right)^2$ is some $\chi_1^2\left(\frac{\rho^2 t}{1-\rho^2}\right)$ random variable. Hence $h(\sqrt{t})$ is increasing in t. □

Now we can proceed to the proof.

Proof. Since for each i, $\frac{(1-\hat{\sigma}^2)n}{(1-\hat{\sigma}^2)n + \hat{\sigma}^2 s_i^2 p}$ is bounded above by 1, $|\hat{S} - S'| \le \frac{1}{p}\sum_{i=1}^p e_i^2 |\frac{\sigma^2}{\hat{\sigma}^2} - 1|$. Since $\hat{\sigma}^2$ is $\sqrt{n}$-consistent, $|\hat{S} - S'|$ is of order $1/\sqrt{n}$ using Cauchy Schwarz.

For the rest of the proof, it suffices that to show that $Var(\hat{S})$ is of order at least $1/p$. For $i = 1, \ldots, p$, denote $S_i := \frac{(1-\hat{\sigma}^2)n}{(1-\hat{\sigma}^2)n + \hat{\sigma}^2 s_i^2 p} \times \left((1 - z_i)\min\{e_i^2, \frac{\hat{x}_i^*}{\hat{\sigma}^2 s_i^2/n}\} + z_i \frac{\hat{x}_i^*}{\hat{\sigma}^2 s_i^2/n}\right)$. By definition $Var(\hat{S}) = \frac{1}{p^2}\sum_{i,j=1}^p Cov(S_i, S_j)$. When $i \ne j$, we have

$$
\begin{aligned}
&\left(\frac{(1-\hat{\sigma}^2)n}{(1-\hat{\sigma}^2)n+\hat{\sigma}^2 s_i^2 p}\times\frac{(1-\hat{\sigma}^2)n}{(1-\hat{\sigma}^2)n+\hat{\sigma}^2 s_j^2 p}\right)^{-1} Cov(S_i,S_j)\\
=&Cov\left((1-z_i)\min\{e_i^2,\frac{\hat{x}_i^*}{\hat{\sigma}^2 s_i^2/n}\}+z_i\frac{\hat{x}_i^*}{\hat{\sigma}^2 s_i^2/n},\ (1-z_j)\min\{e_j^2,\frac{\hat{x}_j^*}{\hat{\sigma}^2 s_j^2/n}\}+z_j\frac{\hat{x}_j^*}{\hat{\sigma}^2 s_j^2/n}\right)\\
=&Cov\left((1-z_i)\min\{e_i^2,\frac{\hat{x}_i^*}{\hat{\sigma}^2 s_i^2/n}\},\ (1-z_j)\min\{e_j^2,\frac{\hat{x}_j^*}{\hat{\sigma}^2 s_j^2/n}\}\right)\\
=&Cov\left(\min\{e_i^2,\frac{\hat{x}_i^*}{\hat{\sigma}^2 s_i^2/n}\},\ \min\{e_j^2,\frac{\hat{x}_j^*}{\hat{\sigma}^2 s_j^2/n}\}\right)\geq 0
\end{aligned}
$$

where the last inequality follows from Lemma 39. Therefore

$$
Var(\hat{S})=\frac{1}{p^2}\sum_{i,j=1}^{p}Cov(S_i,S_j)\geq\frac{1}{p^2}\sum_{i}^{p}Var(S_i).
$$

Since for some $\delta>0$, $\frac{\hat{x}_i^*}{\hat{\sigma}^2 s_i^2/n}$ is uniformly bounded away from 0 for all $\epsilon\in(0,\delta)$ by Lemma 1. And as $\frac{\max_i\{s_i^2 p\}}{n}$ is bounded above, we have $\frac{(1-\hat{\sigma}^2)n}{(1-\hat{\sigma}^2)n+\hat{\sigma}^2 s_i^2 p}$ bounded away from zero uniformly over i. Therefore $Var(S_i)\geq C$ for some constant $C>0$. Hence $Var(\hat{S})\geq\frac{1}{p^2}\sum_{i=1}^{p}C=\frac{C}{p}$. This completes the proof. □

A.2 Optimal Estimation when the Parameter Space is of Infinite Dimension

A.2.1 Proof of Theorem 8

Proof. Let us define $k(n)=n^{1/7}$, and let us for $\theta\in\mathbf{R}^{\mathbf{N}}$ define by $\theta^{[k]}$ the vector

$$
\theta^{[k]}=(\theta_1,...\theta_k,0,...,0).
$$

Let us first observe that by integrating successively, we can conclude from (2.2) that, perhaps for a different M,

$$E_\theta \left(\sup_{\theta \in O} \left| \frac{\partial \ell_t(\theta)}{\partial \theta_i} \right| \right) \leq M \tag{A.1}$$

Then we have

$$\sup_{\theta \in O} \left| \sum_{t=1}^{n} \ell_t(\theta) - \sum_{t=1}^{n} \ell_t \left(\theta^{[k(n)]} \right) \right| \leq \sum_{m > k(n), 1 \leq t \leq n} \theta_m \sup_{\theta \in O} \left| \frac{\partial \ell_t(\theta)}{\partial \theta_i} \right| \tag{A.2}$$

and

$$\begin{aligned} E \sum_{m > k(n), 1 \leq t \leq n} |\theta_m| \sup_{\theta \in O} \left| \frac{\partial \ell_t(\theta)}{\partial \theta_i} \right| &\leq const \sum_{m > k(n), 1 \leq t \leq n} c_m M \\ &\leq const \sum_{m > n^{1/7}, 1 \leq t \leq n} o(1) m^{-8} M = o(1), \end{aligned} \tag{A.3}$$

Hence the probability measures of θ and $\theta^{[k(n)]}$ are asymptotically equivalent: The difference of the logarithms converges to zero in probability uniformly on O; hence the ratio converges to 1. So let us analyze $\ell_t \left(\theta^{[k(n)]} \right)$ as a function of θ.

Let us first analyze the ML-estimator. Since $\theta^{[k(n)]}$ only contains finitely many parameters, we can use the classical approach for linearization of the first order condition:

$$0 = \sum \ell_t^{(1)}(\theta^{[k(n)]}) + \sum \ell_t^{(2)}(\theta^{[k(n)]})(\hat{\theta} - \theta^{[k(n)]}) + R_n(\hat{\theta} - \theta^{[k(n)]}),$$

where R_n is the remainder term of the Taylor series expansion. With some tedious, but elementary calculations, we can show that (with $||.||$ denoting the usual matrix norm)

$$E||\frac{1}{n} \sum \ell_t^{(2)}(\theta^{[k(n)]}) - \frac{1}{n} E \sum \ell_t^{(2)}(\theta^{[k(n)]})||, \tag{A.4}$$

$$P[||R_n|| \leq const \cdot k^3(\hat{\theta} - \theta^{[k(n)]})] \to 1,$$

and for all $\epsilon > 0$, we can find a $C(\epsilon)$ so that

$$P[||\sum \ell_t^{(1)}(\theta^{[k(n)]})|| < \sqrt{k} \cdot \sqrt{n} \cdot C(\epsilon)] > 1 - \epsilon.$$

Since we did assume that the information matrix $\frac{1}{n} E \sum \ell_t^{(2)}(\theta^{[k(n)]})$ is well conditioned, we can conclude that

$$(\sqrt{n}/\sqrt{k})(\hat{\theta} - \theta^{[k(n)]} - \left(\sum \ell_t^{(2)}(\theta^{[k(n)]})\right)^{-1} \sum \ell_t^{(1)}(\theta^{[k(n)]}) \to 0 \tag{A.5}$$

(where the convergence is to be understood to be in probability) and when utilizing (A.4)

$$(\sqrt{n}/\sqrt{k})(\hat{\theta} - \theta^{[k(n)]})\text{remains } O_P(1) \tag{A.6}$$

A_n was defined as $\sum \ell_t^{(2)}(\hat{\theta})$. Using a third-order Taylor series expansion, and again using the fact that we assumed the information matrix to be well-conditioned, we may conclude that for all $\eta > 0$

$$P\left[(1-\eta)E\sum \ell_t^{(2)}(\theta^{[k(n)]}) < \sum \ell_t^{(2)}(\hat{\theta}) < (1+\eta)E\sum \ell_t^{(2)}(\theta^{[k(n)]})\right] \to 1 \tag{A.7}$$

We now have all the tools to compute posterior distribution. The posterior distribution is a random probability measure on Θ, and the density of this measure is proportional to the likelihood function. Let us denote this measure by Π_n. Then Π_n is measurable with respect to $\mathfrak{F}_n$ (the information available at time n). and is a measure defined on the σ-algebra $\mathfrak{I}$ of the measurable subsets of Θ. Let $D_n n$ be events from $\mathfrak{F}_n \times \mathfrak{I}$. Then define the random variables Δ_n by:

Let

$$\Delta_n(\omega) = \pi_n(\{\theta : (\omega, \theta) \in D_n\})$$

It is quite easy to show that Δ_n are random variables: It is trivial if D_n is a product set itself, and then apply a monotone-class argument. Then, trivially

$$P(D_n) = E(\Delta_n).$$

Si uf $P(D_n) \to 1$, $E(\Delta_n) \to 1$, too. As $0 \leq \Delta_n \leq 1, \Delta_n \to 1$ stochastically, too. Hence for all $\epsilon > 0$ we can find $F_n \in \mathfrak{F}_n$ with $P(F_n) > 1 - \epsilon$ so that for $\omega \in F_n$

$$\Delta_n(\omega) = \Pi_n(\{\theta : (\omega, \theta) \in D_n\}) \geq 1 - \epsilon.$$

So if we have given a sequence of events with probability converging to one. Then - automatically - the projections of this set will have - except for events from $\mathfrak{F}_n$ with arbitrary small probabilities - conditional probabilities arbitrarily near to one.

Let us now come back to our original problem, namely analyzing the poste- rior distribution. First of all let us observe that we postulated that our parame- ter can be estimated consistently. So we have estimators -$\mathfrak{F}_n$ valued functions -$\tilde{\theta}_n$ which converge to the true parameter. So $P[\tilde{\theta}_n \in O(\theta)] \to 1$, where O is the bounded neighbourhood we used in (2.2) and (A.1). Hence we can conclude that

$$\Pi_n[O(\theta)] \to 1. \tag{A.8}$$

in probability.

In principle, $\Pi_n(.)$ should be easy to construct. We know that the density is proportional to the likelihood, and we did derive some simplifying approx- imations to the likelihood. Our first problem is the normalizing factor. We would have to integrate the likelihood over the whole parameter space, which is inconvenient. When we use relations like (A.8), we can limit our averaging to e.g. $O(\theta)$. For all $\epsilon > 0$, for n large enough there exist sets $F_n \in \mathfrak{F}_n$

with $P(F_n) \geq 1-\epsilon$ and for $\omega \in F_n$, $\Pi_n[O(\theta)](\omega) \geq 1-\epsilon$. So with Π denoting the prior on Θ we have

$$\frac{d\Pi_n}{d\Pi} = \frac{\exp(\sum \ell_t(\theta))}{\int \exp(\sum \ell_t(\theta)) d\Pi(\theta)}.$$

For $\omega \in F_n$, however, $\Pi_n[O(\theta)](\omega) \geq 1-\epsilon$. Hence

$$\int_{O(\theta)} \frac{d\Pi_n}{d\Pi} d\Pi \geq 1-\epsilon,$$

and therefore

$$\frac{\int_{O(\theta)} \exp(\sum \ell_t(\theta)) d\Pi}{\int \exp(\sum \ell_t(\theta)) d\Pi} \geq 1-\epsilon.$$

Since the ration on the LHS is bounded by 1, we may conclude that

$$\lim \frac{\int_{O(\theta)} \exp(\sum \ell_t(\theta)) d\Pi}{\int \exp(\sum \ell_t(\theta)) d\Pi} = 1$$

which implies that

$$\lim \frac{\int_{O(\theta)^c} \exp(\sum \ell_t(\theta)) d\Pi}{\int \exp(\sum \ell_t(\theta)) d\Pi} = 0$$

Therefore we can conclude that total variation of the difference between Π_n and the random measure $\Pi_n^{(1)}$ defined by its with density

$$\frac{I_{O(\theta)} \exp(\sum \ell_t(\theta)) d\Pi}{\int I_{O(\theta)} \exp(\sum \ell_t(\theta)) d\Pi} n$$

converges to zero.

Let us now define the random measure $\Pi_n^{(2)}$ to have the density

$$\frac{I_{O(\theta)} \exp(\sum \ell_t(\theta^{[k]})) d\Pi}{\int I_{O(\theta)} \exp(\sum \ell_t(\theta^{[k]})) d\Pi} n$$

Then (A.2),(A.3) imply that the total variation of the difference converges to zero.

Now lets construc events of probability converging to one based on equation (A.5),(A.6) and (A.7). What we want to do is to approximate

$$\sum \ell_t(\theta^{[k]})$$

by its second order Taylor approximation around $\hat{\theta}_n$. First observe that in (A.6) implies that

$$P\left([\hat{\theta}_n \in O(\theta)]\right) \to 1.$$

We can apply our technique again to this sequence of events and construct measure $\Pi_n^{(3)}$ with the corresponding densities: this way we guarentee that $\hat{\theta}_n$ is in $O(\theta)$, so all our functions are differentiable. Next we analyze (10):

$$\left(\sqrt{n}/\sqrt{k}\right)(\hat{\theta} - \theta^{[k(n)]}) \text{ remains } O_p(1).$$

An equivalent formulation of this statement is: for any sequence $B_n \uparrow \infty$ define the events

$$S_n = \left[|| \left(\sqrt{n}/\sqrt{k}\right)(\hat{\theta} - \theta^{[k(n)]})|| \leq B_n\right]$$

Then $P(S_n) = 1$.

Then we have, again, we can construct measures $\Pi_n^{(4)}$ for which our density equals

$$\frac{I_S \exp(\sum \ell_t(\theta^{[k]}))d\Pi}{\int I_S \exp(\sum \ell_t(\theta^{[k]}))d\Pi} n.$$

For this density, however, we can use a Taylor series expansion with $\hat{\theta}$ as base value

$$\sum \ell_t(\theta^{[k]}) = \sum \ell_t(\hat{\theta} - (\theta^{[k]} - \hat{\theta}))' \left(\sum \ell_t^{(2)}(\hat{\theta})\right)(\theta^{[k]} - \hat{\theta}))/2 + r_n$$

where

$$r_n \leq \sum \sup_{\theta \in O} |\frac{\partial^3 \ell_t(\theta)}{\partial \theta_i \theta_j \theta_k}| \max |(\theta^{[k]} - \hat{\theta}))_i| \max |(\theta^{[k]} - \hat{\theta}))_j| \max |(\theta^{[k]} - \hat{\theta}))_k|.$$

Then for $\theta^{[k]} \in S_n$,

$$E|r_n| \leq \left(E \sup_{\theta \in O} |\frac{\partial^3 \ell_t(\theta)}{\partial \theta_i \theta_j \theta_k}| \right) nB^3 \frac{k^{3/2}}{n^{3/2}} = B_n^3 \frac{n^{3/(2\times 7)}}{n^{1/2}} = B_n^3 n^{-4/14}.$$

So choosing

$$B_n = o(n^{1/14})$$

guarentees that $E|r_n| \to 0$.

Hence we can again construct $\Pi_n^{(5)}$, being asymptotically equivalent to Π_n, with density

$$\frac{I_{S_n} \exp(\sum \ell_t(\hat{\theta}) + (\theta^{[k]} - \hat{\theta})' P_k' \left(\sum \ell_t^{(2)}(\hat{\theta}) \right) P_k (\theta^{[k]} - \hat{\theta})/2)}{\int I_{S_n} \exp(\sum \ell_t(\hat{\theta}) + (\theta^{[k]} - \hat{\theta})' P_k' \left(\sum \ell_t^{(2)}(\hat{\theta}) \right) P_k (\theta^{[k]} - \hat{\theta})/2) d\Pi}$$

$\sum \ell_t(\hat{\theta})$ does not depend on θ so the corresponding term $\exp(\sum \ell_t(\hat{\theta}))$ cancels out. Furthermore, observe that

$$\theta^{[k]} = P_k \theta.$$

where P_k is the matrix describing projection to the first k components of a vector. As $\hat{\theta}$ only contains k components, we have

$$\hat{\theta} = P_k \hat{\theta}$$

. Hence we can write our density as

$$\frac{I_{S_n} \exp((\theta - \hat{\theta})' P_k' \left(\sum \ell_t^{(2)}(\hat{\theta}) \right) P_k (\theta - \hat{\theta})/2)}{\int I_{S_n} \exp((\theta - \hat{\theta})' P_k' \left(\sum \ell_t^{(2)}(\hat{\theta}) \right) P_k (\theta - \hat{\theta})/2) d\Pi}$$

which looks very much like a Gaussian density. The only problem is the factor I_{S_n}. We know that this factor converges to 1 so it is sufficient to establish that the measure with density

$$\frac{\exp((\theta-\hat{\theta})' P_k' \left(\sum \ell_t^{(2)}(\hat{\theta})\right) P_k(\theta-\hat{\theta})/2)}{\int \exp((\theta-\hat{\theta})' P_k' \left(\sum \ell_t^{(2)}(\hat{\theta})\right) P_k(\theta-\hat{\theta})/2) d\Pi}$$

is

$$G\left(\left(C-\sum \ell_t^{(2)}(\hat{\theta})\right)^{-1}\left(\sum \ell_t^{(2)}(\hat{\theta})\right)\hat{\theta}, \left(C-\sum \ell_t^{(2)}(\hat{\theta})\right)^{-1}\right)$$

and

$$G\left(\left(C-\sum \ell_t^{(2)}(\hat{\theta})\right)^{-1}\left(\sum \ell_t^{(2)}(\hat{\theta})\right)\hat{\theta}, \left(C-\sum \ell_t^{(2)}(\hat{\theta})\right)^{-1}\right)\left(S_n^C\right) \to 0.$$

Where S_n^C is the complememt of S_n. Both are tedious but elementary calculations with normal random variables. so we will omit these proofs. □

A.3 Moderate Expected Utility

A.3.1 Proof of Proposition 16

Let Z be a finite set with n alternatives enumerated $x^1, x^2, \ldots, x^n$. Consider the set of choice rules ρ on Z which satisfy WST with $\rho(x^i, x^j) \geq 1/2$ whenever $i \leq j$ and for which the set $\{\rho(x,y) \in [0,1] : x \neq y\}$ has maximum cardinality with $n(n-1)$ elements. Each such ρ induces a strict ordering $\succ_\rho$ of the $n(n+1)/2$ pairs $P_n := \{(x^i, x^j) : n \geq i > j \geq 1\}$ given by $(x^i, x^j) \succ_\rho (x^k, x^\ell)$ if and only if $\rho(x^i, x^j) > \rho(x^k, x^\ell)$. This set of choice rules ρ induces $\#WST(n) = [n(n-1)/2]!$ different strict orderings $\succ_\rho$ on P_n.

MST and MST+ allow the same number of different strict orderings over P_n which we denote $\#MST(n)$. Now consider the addition of alternative x^{n+1} to the set Z.

Lemma 40. $\#MST(n+1) \leq [n(n-1)/2+1]^n \, \#MST(n)$

Proof. Take a single strict ordering over P_n compatible with MST. There are multiple ways to extend this strict ordering to incorporate the new pairs $(x^1, x^{n+1}), (x^2, x^{n+1}), \ldots, (x^n, x^{n+1})$ and obtain a strict ordering over P_{n+1} that is still compatible with MST. Since the original ordering has $n(n-1)/2$ pairs, there are $n(n-1)/2+1$ different positions to include (x^n, x^{n+1}). In this way we obtain $n(n-1)/2+1$ different strict orderings, all of which respect MST. The total number of strict orderings over $P_n \cup \{(x^n, x^{n+1})\}$ that satisfy MST is therefore $[n(n-1)/2+1] \, \#MST(n)$. Now we take one such strict ordering and extend it to incorporate a second pair (x^{n-1}, x^{n+1}). This pair can in principle be added into $n(n-1)/2+2$ different positions, but placing it in the very last position would violate MST, since MST requires $\rho(x^{n-1}, x^{n+1}) > \min\{\rho(x^{n-1}, x^n), \rho(x^n, x^{n+1})\}$. The total number of strict orderings over $P_n \cup \{(x^n, x^{n+1}), (x^{n-1}, x^{n+1})\}$ which satisfy MST must therefore be smaller or equal to $[n(n-1)/2+1]^2 \, \#MST(n)$. A simple inductive argument completes the proof. □

Lemma 41. $\lim_{n\to\infty} \left[\prod_{k=1}^{n} \frac{n(n-1)/2+k}{n(n-1)/2+1}\right] = e$

Proof. The result can be shown by verifying that, for each n,

$$\left(1+\frac{1}{n}\right)^{n-1} \leq \left[\prod_{k=1}^{n} \frac{n(n-1)/2+k}{n(n-1)/2+1}\right] \leq \left(1+\frac{1}{n}\right)^{n}$$

and taking the limit as $n \to \infty$. We leave the details to the reader. □

Lemma 40 implies that

$$\begin{aligned}\frac{\#MST(n+1)}{\#WST(n+1)} &\leq \frac{\#MST(n)}{\#WST(n)} \frac{[n(n-1)/2]!}{[n(n+1)/2]!} [n(n-1)/2+1]^n \\ &= \frac{\#MST(n)}{\#WST(n)} \left[\prod_{k=1}^{n} \frac{n(n-1)/2+1}{n(n-1)/2+i}\right]\end{aligned}$$

and by Lemma 41 the last expression in brackets goes to $1/e$ when n goes to infinity, where $e \approx 2.718$ is the base of the natural logarithm. Hence for all n sufficiently large the ratio $\#MST(n+1)/\#WST(n+1)$ is less than half of the ratio $\#MST(n)/\#WST(n)$, which completes the proof.

Finally, we prove the additional claim, stated after Proposition 16, that

$$\lim_{n\to\infty} \#SST(n)/\#MST(n) = 0.$$

The choice probability $\rho(x^1, x^n)$ must be the highest choice probability in every ρ that satisfies SST. For each strict ordering of choice probabilities satisfying SST, there exist at least $n-2$ strict orderings which violate SST but satisfy MST: for each $k = 2, 3, \ldots, n-1$ change the value of $\rho(x^1, x^n)$ to be equal to $\max\{\rho(x^1, x^k), \rho(x^k, x^n)\} - \varepsilon$ for $\varepsilon > 0$ sufficiently small. It is immediate to see that each resulting ranking violates SST. To see that MST still holds, note that every inequality required by SST holds except those involving $\rho(x^1, x^n)$. In addition, SST implies that for each $k, j = 2, \ldots, n-1$, $\max\{\rho(x^1, x^k), \rho(x^k, x^n\}) > \min\{\rho(x^1, x^j), \rho(x^j, x^n\})$ hence for ε small we have $\rho(x^1, x^n) > \min\{\rho(x^1, x^j), \rho(x^j, x^n)\}$. Thus, $\#SST(n)/\#MST(n) \leq 1/(n-1) \to 0$ when $n \to \infty$. □

A.3.2 Proof of Theorem 18

For necessity, assume there exist u, d and F satisfying (4.2), and assume $\rho(x, y) \geq 1/2$ and $\rho(y, z) \geq 1/2$. If it were the case that $\rho(x, z) < \min\{\rho(x, y), \rho(y, z)\}$, then by (4.2) and the

triangle inequality property of d it would follow that

$$\begin{aligned} u(x) - u(z) &< d(x,z) \min\left\{ \frac{u(x)-u(y)}{d(x,y)}, \frac{u(y)-u(z)}{d(y,z)} \right\} \\ &\leq [d(x,y) + d(y,z)] \min\left\{ \frac{u(x)-u(y)}{d(x,y)}, \frac{u(y)-u(z)}{d(y,z)} \right\} \\ &\leq d(x,y)\frac{u(x)-u(y)}{d(x,y)} + d(y,z)\frac{u(y)-u(z)}{d(y,z)} \\ &= u(x) - u(z) \end{aligned}$$

which is a contradiction. Hence, it must be the case that $\rho(x,z) \geq \min\{\rho(x,y), \rho(y,z)\}$. This first step of the necessity was also shown by [20].

Now suppose we have equality $\rho(x,z) = \min\{\rho(x,y), \rho(y,z)\}$. We consider the case $\min\{\rho(x,y), \rho(y,z)\} = \rho(x,y)$, while the remaining case is analogous and left to the reader. Representation (4.2) and the triangle inequality imply

$$\begin{aligned} u(x) - u(y) + u(y) - u(z) &= u(x) - u(z) \\ &= d(x,z)\left[\frac{u(x)-u(y)}{d(x,y)}\right] \\ &\leq [d(x,y) + d(y,z)]\left[\frac{u(x)-u(y)}{d(x,y)}\right] \\ &= u(x) - u(y) + d(y,z)\left[\frac{u(x)-u(y)}{d(x,y)}\right]. \end{aligned}$$

Subtracting $u(x) - u(y)$ from both sides we obtain

$$\frac{u(y)-u(z)}{d(y,z)} \leq \frac{u(x)-u(y)}{d(x,y)}$$

and therefore (4.2) yields $\rho(x,y) = \rho(y,z) = \rho(x,z)$ as desired.

For sufficiency, suppose ρ satisfies MST+. In particular, ρ satisfies WST, and hence, by letting $x \succcurlyeq y$ if and only if $\rho(x,y) \geq 1/2$, we obtain a complete and transitive relation $\succcurlyeq$ over

the finite set of options Z. The relation $\succcurlyeq$ induced by ρ divides the n alternatives in Z into $k \leq n$ indifference classes. Therefore, there exists a utility function $u : Z \to \{1, \ldots, k\}$ that is onto and represents $\succcurlyeq$, that is, $u(x) \geq u(y)$ if and only if $x \succcurlyeq y$ if and only if $\rho(x, y) \geq 1/2$.

Let $Y := \{\{x, y\} \subset Z : \rho(x, y) \neq 1/2\}$, and let m be the cardinality of the set $\{|\rho(x, y) - 1/2| : \{x, y\} \in Y\}$. Partition the set Y into m disjoint sets $Y_1 \cup Y_2 \cup \cdots \cup Y_m = Y$ such that for any two pairs $\{w, x\}$ and $\{y, z\}$ in Y we have $\{w, x\} \in Y_i$ and $\{y, z\} \in Y_j$ with $i \geq j$ if and only if $|\rho(w, x) - 1/2| \leq |\rho(y, z) - 1/2|$. Thus, the pairs in Y_1 have the highest value of $|\rho(x, y) - 1/2|$, while the pairs in Y_m have the lowest value of $|\rho(x, y) - 1/2|$ among the pairs in Y.

The result is trivial when Z has $n \leq 2$ alternatives so suppose $n \geq 3$. Define a constant $C = (n-1)^{[n(n-1)/2+1]} > 0$ and define the sequence $D_1, D_2, \ldots, D_m$ by:

$$D_1 = 0; D_j = (n-1)^{j-2} \text{ for } j = 2, \ldots, m.$$

Let $d : Z \times Z \to [0, \infty)$ be defined as follows:

$$d(x, y) = \begin{cases} 0, & \text{if } x = y \\ C, & \text{if } x \neq y \text{ and } \rho(x, y) = 1/2 \\ (C/2 + D_j) \, |u(x) - u(y)| \,, & \text{if } \{x, y\} \in Y_j \end{cases} \tag{A.9}$$

From the definition (A.9) it is immediate that d satisfies (i) $d(x, y) \geq 0$; (ii) $d(x, y) = 0$ if and only if $x = y$; and (iii) $d(x, y) = d(y, x)$ for all $x, y \in Z$. To show that d is a metric, it remains to verify the triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$. The inequality trivially holds when any two options among x, y, z are equal. Consider three distinct options $x, y, z \in Z$.

Case 1: $u(x) = u(y) = u(z)$. By the definition of u we have $\rho(x, y) = \rho(y, z) = \rho(x, z) = 1/2$. By the definition of d we have $d(x, z) = C < 2C = d(x, y) + d(y, z)$.

Case 2: $u(x) \neq u(y) = u(z)$. The definitions of u and d imply

$$\begin{aligned} d(x,y) + d(y,z) - d(x,z) &= (C/2 + D_i)\,|u(x) - u(y)| + C - (C/2 + D_j)\,|u(x) - u(z)| \\ &= (D_i - D_j)\,|u(x) - u(z)| + C \\ &\geq -(n-1)^{m-2}(n-1) + C \\ &= (n-1)^{[n(n-1)/2+1]} - (n-1)^{m-1} \\ &> 0 \end{aligned}$$

where the last inequality follows from the fact that we defined m to be the cardinality of $\{|\rho(x,y) - 1/2| : \{x,y\} \in Y\}$ which is smaller or equal to $n(n-1)/2$.

Case 3: $u(y) \neq u(x) = u(z)$. The definitions of u and d imply

$$\begin{aligned} d(x,y) + d(y,z) - d(x,z) &= (C/2 + D_i)\,|u(x) - u(y)| + (C/2 + D_j)\,|u(y) - u(z)| - C \\ &= (C + D_i + D_j)\,|u(y) - u(z)| - C \\ &\geq 0. \end{aligned}$$

Case 4: $u(z) \neq u(x) = u(y)$. The inequality follows from the same argument as in Case 2.

Case 5: $u(x) > u(y) > u(z)$. By the definition of u we have $\{x,y\} \in Y_i$, $\{y,z\} \in Y_j$, and $\{x,z\} \in Y_\ell$, for some i, j, ℓ. The definition of d implies

$$\begin{aligned} d(x,y) + d(y,z) - d(x,z) &= (C/2 + D_i)\,|u(x) - u(y)| + (C/2 + D_j)\,|u(y) - u(z)| \\ &\quad - (C/2 + D_\ell)\,|u(x) - u(y) + u(y) - u(z)| \\ &= (D_i - D_\ell)\,|u(x) - u(y)| + (D_j - D_\ell)\,|u(y) - u(z)| \end{aligned}$$

The definition of u implies $\rho(x,y) > 1/2$ and $\rho(y,z) > 1/2$. By MST+ we have either

$\rho(x,y) = \rho(y,z) = \rho(x,z)$ or $\rho(x,z) > \min\{\rho(x,y), \rho(y,z)\}$. The first case implies $D_i = D_j = D_\ell$ above and therefore $d(x,y) + d(y,z) - d(x,z) = 0$. The second case implies $D_\ell < \max\{D_i, D_j\}$. If $D_\ell \leq \min\{D_i, D_j\}$ then both $(D_i - D_\ell)$ and $(D_j - D_\ell)$ above are positive and the desired inequality holds. It remains to show the inequality holds when $\min\{D_i, D_j\} < D_\ell < \max\{D_i, D_j\}$, which implies

$$\begin{aligned} d(x,y) + d(y,z) - d(x,z) &\geq (\max\{D_i, D_j\} - D_\ell)\, 1 + (\min\{D_i, D_j\} - D_\ell)\,(n-2) \\ &\geq (n-1)^{\ell-1} - (n-1)^{\ell-2} + [0 - (n-1)^{\ell-2}](n-2) \\ &= 0. \end{aligned}$$

Case 6: $u(x) > u(z) > u(y)$. By the definition of u we have $\{x,y\} \in Y_i$, $\{y,z\} \in Y_j$, and $\{x,z\} \in Y_\ell$, for some i, j, ℓ. The definition of d implies

$$\begin{aligned} d(x,y) + d(y,z) - d(x,z) &= (C/2 + D_i)\,[u(x) - u(z) + u(z) - u(y)] \\ &\quad + (C/2 + D_j)\,[u(z) - u(y)] - (C/2 + D_\ell)\,[u(x) - u(z)] \\ &= (D_i - D_\ell)\,[u(x) - u(z)] + (C + D_i + D_j)\,[u(z) - u(y)] \\ &\geq \left(0 - (n-1)^{m-2}\right)(n-2) + (C + 0 + 0)\,1 \\ &= -(n-1)^{m-1} + (n-1)^{m-2} + (n-1)^{n(n-1)/2+1} \\ &> 0. \end{aligned}$$

Case 7: $u(y) > u(x) > u(z)$. By the definition of u we have $\{x,y\} \in Y_i$, $\{y,z\} \in Y_j$, and

$\{x, z\} \in Y_\ell$, for some i, j, ℓ. The definition of d implies

$$\begin{aligned} d(x,y) + d(y,z) - d(x,z) &= (C/2 + D_i)\,[u(y) - u(x)] \\ &\quad + (C/2 + D_j)\,[u(y) - u(x) + u(x) - u(z)] \\ &\quad - (C/2 + D_\ell)\,[u(x) - u(z)] \\ &= (C + D_i + D_j)\,[u(y) - u(x)] + (D_j - D_\ell)\,[u(x) - u(z)] \\ &> 0. \end{aligned}$$

Case 8: $u(y) > u(z) > u(x)$. Similarly to Case 7, we have

$$\begin{aligned} d(x,y) + d(y,z) - d(x,z) &= (C + D_i + D_j)\,[u(y) - u(z)] + (D_i - D_\ell)\,[u(z) - u(x)] \\ &> 0. \end{aligned}$$

Case 9: $u(z) > u(x) > u(y)$. Similarly to Cases 7 and 8, we have

$$\begin{aligned} d(x,y) + d(y,z) - d(x,z) &= (C + D_i + D_j)\,[u(x) - u(y)] + (D_j - D_\ell)\,[u(z) - u(x)] \\ &> 0. \end{aligned}$$

Case 10: $u(z) > u(y) > u(x)$. Since $d(x,y) + d(y,z) \leq d(x,z)$ if and only if $d(y,x) + d(z,y) \leq d(z,x)$, the inequality follows from the same argument as in Case 5.

By Cases 1 to 10 above, d satisfies the triangle inequality and is therefore a metric. Now, we verify that the utility u and the metric d constructed above provide an ordinal representation for ρ as in (4.3). First, $\rho(w,x) \geq \rho(y,z) > 1/2$ if and only if $\rho(w,x) > 1/2$, $\rho(y,z) > 1/2$, and $|\rho(w,x) - 1/2| \geq |\rho(y,z) - 1/2|$, if and only if $u(w) > u(x)$, $u(y) > u(z)$, $d(w,x) =$

$(C/2 + D_i)[u(w) - u(x)]$, $d(y,z) = (C/2 + D_j)[u(y) - u(z)]$, and $i \leq j$, if and only if

$$\frac{u(w) - u(x)}{d(w,x)} = \frac{1}{C/2 + D_i} \geq \frac{1}{C/2 + D_j} = \frac{u(y) - u(z)}{d(y,z)} > 0.$$

Second, $\rho(w,x) \geq 1/2 \geq \rho(y,z)$ if and only if $u(w) - u(x) \geq 0 \geq u(y) - u(z)$ if and only if

$$\frac{u(w) - u(x)}{d(w,x)} \geq 0 \geq \frac{u(y) - u(z)}{d(y,z)}.$$

And, finally, $1/2 > \rho(w,x) \geq \rho(y,z)$ if and only if $\rho(w,x) < 1/2$, $\rho(y,z) < 1/2$, and $|\rho(w,x) - 1/2| \leq |\rho(y,z) - 1/2|$, if and only if $u(w) < u(x)$, $u(y) < u(z)$, $d(w,x) = (C/2 + D_i)[u(x) - u(w)]$, $d(y,z) = (C/2 + D_j)[u(z) - u(y)]$, and $i \geq j$, if and only if

$$0 > \frac{u(w) - u(x)}{d(w,x)} = -\frac{1}{C/2 + D_i} \geq -\frac{1}{C/2 + D_j} = \frac{u(y) - u(z)}{d(y,z)}$$

hence the ordinal representation (4.3) holds. Finding a strictly increasing F such that the cardinal representation (4.2) holds is then straightforward and left to the reader. □

A.3.3 Proof of Theorem 21

To show necessity, let $U : \Delta \to [0,1]$ be linear and onto, let $\|\cdot\|$ be a norm defined on the subspace $\{x \in \mathbb{R}^n : x_1 + \cdots + x_n = 0\}$ and which is generated by an inner product $\|x\| = \sqrt{\langle x, x \rangle}$, and let F be a strictly increasing and continuous transformation such that the MEM representation (4.4) holds.

First, ρ must be non-constant since U is onto. Second, ρ must be continuous outside the diagonal since (i) U is linear; (ii) $\|\cdot\|$ is a norm hence $\|x - y\| > 0$ whenever $x \neq y$; and (iii)

F is continuous. Third, ρ must be linear since

$$\begin{aligned}\rho(\alpha x + (1-\alpha)z, \alpha y + (1-\alpha)z) &= F\left(\frac{U(\alpha x + (1-\alpha)z) - U(\alpha y + (1-\alpha)z)}{\|\alpha x + (1-\alpha)z - [\alpha y + (1-\alpha)z]\|}\right) \\ &= F\left(\frac{\alpha[U(x) - U(y)]}{\alpha\|x - y\|}\right) \\ &= \rho(x, y)\end{aligned}$$

whenever $0 < \alpha < 1$ and $x \neq y$, and the equality holds trivially when $x = y$. Finally, we show that ρ must be convex. Suppose $\rho(x, y) = 1/2$ and $\rho(x, z) = \rho(y, z) > 1/2$. By (4.4) we have $U(x) = U(y) > U(z)$ and $\|x - z\| = \|y - z\|$. The linearity of U and $U(x) = U(y) > U(z)$ imply that $x - z$ and $y - z$ are not collinear. Thus the Cauchy-Schwartz inequality implies

$$-1 < \frac{\langle x - z, y - z\rangle}{\|x - z\|\|y - z\|} < 1.$$

Since $\|x - z\| = \|y - z\|$, we have the equality

$$\frac{\|\alpha x + (1-\alpha)y\|^2}{\|x - z\|\|y - z\|} = 1 + 2(\alpha^2 - \alpha)\left(1 - \frac{\langle x - z, y - z\rangle}{\|x - z\|\|y - z\|}\right)$$

where the right hand side can be easily verified to have a strict minimum at $\alpha = 1/2$. Thus the mapping $\alpha \mapsto \|\alpha x + (1-\alpha)y\|$ also has a strict minimum at $\alpha = 1/2$ and by (4.4) the choice rule ρ must be convex.

To show sufficiency, let the non-constant choice rule ρ on Δ be linear, continuous (outside the diagonal), convex, and satisfy MST+. First, we show that ρ has a unique linear extension to the $n - 1$ dimensional hyperplane H that contains Δ.

Lemma 42. *ρ has a unique linear extension to $H = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_1 + \cdots + x_n = 1\}$.*

Proof. Let ρ' and ρ'' be two linear extensions of ρ and let $x, y \in \mathbb{R}^n$ with $x_1 + \cdots + x_n = y_1 + \cdots + y_n = 1$. Let $z = (1/n, \ldots, 1/n) \in \Delta$. Take $0 < \alpha < 1$ sufficiently small such that

$0 < \alpha x_i + (1-\alpha)/n < 1$ and $0 < \alpha y_i + (1-\alpha)/n < 1$ for each i. Then $\alpha x + (1-\alpha)z \in \Delta$, $\alpha y + (1-\alpha)z \in \Delta$ and, by linearity,

$$\begin{aligned}
\rho'(x,y) &= \rho'(\alpha x + (1-\alpha)z, \alpha y + (1-\alpha)z) \\
&= \rho(\alpha x + (1-\alpha)z, \alpha y + (1-\alpha)z) \\
&= \rho''(\alpha x + (1-\alpha)z, \alpha y + (1-\alpha)z) \\
&= \rho''(x,y)
\end{aligned}$$

hence ρ' and ρ'' must be equal. □

From this point on, we identify ρ with its unique linear extension. Define the relation $\succcurlyeq \subset \Delta \times \Delta$ by $x \succcurlyeq y$ if and only if $\rho(x,y) \geq 1/2$. Since ρ satisfies MST+, this $\succcurlyeq$ is complete and transitive. By linearity and continuity, $\succcurlyeq$ satisfies all the vNM axioms and admits an expected utility representation. Since ρ is non-constant, there is a unique linear function $U : \mathbb{R}^n \to \mathbb{R}$ which represents $\succcurlyeq$ with $U(\Delta) = [0,1]$.

For each lottery x, let $I(x) := \{y \in H : \rho(x,y) = 1/2\}$ denote the set of lotteries that are stochastically indifferent to x. Note that $I(x)$ is an affine subspace of dimension $n-2$. Since ρ is non-constant, there exist $\bar{x}, \bar{y} \in \Delta$ with $\rho(\bar{x}, \bar{y}) > 1/2$. By linearity, ρ is entirely determined by the values of the mapping $x \mapsto \rho(x, \bar{y})$ for $x \in I(\bar{x})$. For each $1/2 < p \leq 1$ let $B(p) := \{x \in I(\bar{x}) : \rho(x, \bar{y}) \geq p\}$ be the upper contour set of elements that are stochastically indifferent to $\bar{x}$ and that are chosen over $\bar{y}$ with probability greater or equal to p.

Lemma 43. *$B(p)$ is convex for all $1/2 < p \leq 1$.*

Proof. Let $x, x' \in B(p)$ and let $0 < \alpha < 1$. Since $I(\bar{x})$ is an affine subspace, $\alpha x + (1-\alpha)x' \in I(\bar{x})$. Linearity implies $\rho(\alpha x + (1-\alpha)x', \alpha\bar{y} + (1-\alpha)x') = \rho(x, \bar{y}) \geq p$. Linearity also implies $\rho(\alpha\bar{y} + (1-\alpha)x', \bar{y}) = \rho(x', \bar{y}) \geq p$. Then, MST+ implies $\rho(\alpha x + (1-\alpha)x', \bar{y}) \geq p$ and therefore $\alpha x + (1-\alpha)x' \in B(p)$. □

Lemma 44. *$B(p)$ is compact for all $1/2 < p \leq 1$.*

Proof. $B(p)$ is closed by continuity. Let $|\cdot|$ denote the standard Euclidean metric, not necessarily equal to the metric we are going to construct for the representation. If $B(p)$ were not bounded, there would exist a sequence $x(k)$ in $B(p)$ with $|x(k)-\bar{y}| \geq k$ for all $k \in \mathbb{N}$. For each k, by linearity $\rho(\bar{y}+(x(k)-\bar{y})/|x(k)-\bar{y}|, \bar{y}) = \rho(x(k), \bar{y}) \geq p$. By Bolzano-Weierstrass the sequence $\bar{y}+(x(k)-\bar{y})/|x(k)-\bar{y}|$ would have a subsequence converging to some $z \neq \bar{y}$. By the linearity of U we would have $U(z) = U(\bar{y})$ and $\rho(z, \bar{y}) = 1/2$, contradicting continuity. Hence $B(p)$ must be bounded and therefore compact. □

Lemma 45. *The mapping $x \mapsto \rho(x, \bar{y})$ has a unique maximizer $\hat{x}$ on $I(\bar{x})$.*

Proof. Since $\rho(\bar{x}, \bar{y}) > 1/2$ we have $B(p) \neq \emptyset$ for some $p > 1/2$. Since ρ is continuous outside the diagonal, the mapping $x \mapsto \rho(x, \bar{y})$ is continuous on $I(\bar{x})$. $B(p)$ is compact by Lemma 44, hence the maximum $\rho(\hat{x}, \bar{y}) = \bar{p}$ is attained at some $\hat{x} \in B(p)$. Hence $B(\bar{p})$ is not empty, and by the previous lemmas it is compact and convex. Since ρ is convex, $B(\bar{p})$ must be a singleton. □

For the rest of the proof, we denote by $\hat{x}$ the unique maximizer of $x \mapsto \rho(x, \bar{y})$ on $I(\bar{x})$.

Lemma 46. *$x \in I(\bar{x})$ and $\rho(x, \bar{y}) = p$ implies $\rho(2\hat{x} - x, \bar{y}) = p$.*

Proof. The statement trivially holds if $x = \hat{x}$, so suppose $x \neq \hat{x}$. First note $2\hat{x} - x = \hat{x} + (\hat{x} - x) \in I(\bar{x})$. If $\rho(2\hat{x} - x, \bar{y}) < p$, since $x \mapsto \rho(x, \bar{y})$ is continuous in the segment $[\hat{x}, \hat{x} + (\hat{x} - x)]$, by the intermediate value theorem we have $\rho(x', \bar{y}) = p$ for some x' in the open segment $(\hat{x}, 2\hat{x} - x)$. But then since $\hat{x}$ is the unique maximizer in $I(\bar{x})$ it is also the unique maximizer in the segment $[x, x']$. Since $\hat{x} \neq x/2 + x'/2$ this contradicts the fact that ρ is convex. Hence we must have $\rho(2\hat{x} - x, \bar{y}) \geq p$. The same argument shows that $\rho(2\hat{x} - x, \bar{y}) \leq p$. □

Recall that $\hat{x}$ is the unique maximizer $\rho(\hat{x}, \bar{y}) = \bar{p}$ on $I(\bar{x})$. Let $B = B(p) - \hat{x}$ for some fixed $p \in (1/2, \bar{p})$. We first define an auxiliary norm $\|\cdot\|_B$ on the $n-2$ dimensional subspace $I(\bar{x}) - \hat{x}$ using B as the unit ball.

Lemma 47. *$\|x\|_B := \inf\{\lambda \geq 0 : x \in \lambda B\}$ is a norm on $I(\bar{x}) - \hat{x}$.*

Proof. The *Minkowski functional* $\|\cdot\|_B$ defined above is a norm when B is a symmetric, convex set such that each line through zero meets B in a non-trivial, closed, bounded segment [33]. By definition $\|x\|_B \geq 0$ for all x. Moreover, if $\|x\|_B = 0$ then $x \in \lambda B$ for all $\lambda > 0$ and therefore $x = 0$. Now for each $\alpha \geq 0$ we have $x \in \lambda B$ if and only if $\alpha x \in \alpha\lambda B$ and therefore $\alpha\|x\|_B = \|\alpha x\|_B$. Lemma 46 implies $x \in \lambda B$ if and only if $-x \in \lambda B$ and therefore $\|x\|_B = \|-x\|_B$. To verify the triangle inequality, note that B is closed by Lemma A.5, and therefore $x/\|x\|_B \in B$ for all x. B is also convex by Lemma A.4, and therefore

$$\frac{x + x'}{\|x\|_B + \|x'\|_B} = \left(\frac{\|x\|_B}{\|x\|_B + \|x'\|_B}\right)\frac{x}{\|x\|_B} + \left(\frac{\|x'\|_B}{\|x\|_B + \|x'\|_B}\right)\frac{x'}{\|x'\|_B} \in B.$$

Thus,

$$\left\|\frac{x + x'}{\|x\|_B + \|x'\|_B}\right\|_B \leq 1$$

and the triangle inequality $\|x + x'\|_B \leq \|x\|_B + \|x'\|_B$ holds. □

Lemma 48. *If $\bar{p} \geq p \geq q > 1/2$ then $B(p) = \hat{x} + \lambda\,[B(q) - \hat{x}]$ for some $0 \leq \lambda \leq 1$.*

Proof. MST+ implies that, for any $x \neq \hat{x}$ in $B(p)$, the function $\alpha \mapsto \rho(\alpha\hat{x} + (1-\alpha)x, \bar{y})$ is strictly increasing for $0 \leq \alpha \leq 1$. It suffices to show that if $\rho(x^1, \bar{y}) = \rho(x^2, \bar{y})$ for $x^1, x^2 \in I(\bar{x})$ and $0 < \alpha < 1$, then $\rho(\alpha x^1 + (1-\alpha)\hat{x}, \bar{y}) = \rho(\alpha x^2 + (1-\alpha)\hat{x}, \bar{y})$. To see that equality must hold, suppose instead that we had $\rho(\alpha x^1 + (1-\alpha)\hat{x}, \bar{y}) < \rho(\alpha x^2 + (1-\alpha)\hat{x}, \bar{y})$. Continuity implies $\rho(\beta x^2 + (1-\beta)\hat{x}, \bar{y}) = \rho(\alpha x^1 + (1-\alpha)\hat{x}, \bar{y})$ for some $0 < \alpha < \beta < 1$.

Figure A.3.3 provides an illustration. Letting

$$z^1 = x^1 + \frac{\beta(1-\alpha)}{\beta-\alpha}(x^2 - x^1)$$
$$z^2 = x^1 + x^2 - z^1$$
$$z^3 = 2\hat{x} - z^1$$
$$z^4 = \alpha x^1 + \beta x^2 + (2 - \alpha - \beta)\hat{x} - z^1$$

we have that the line segment $[z^1, z^2]$ contains the line segment $[x^1, x^2]$; the line segment $[z^1, z^4]$ contains the line segment $[\alpha x^1 + (1-\alpha)\hat{x}, \beta x^2 + (1-\beta)\hat{x}]$ and

$$z^1/2 + z^2/2 = x^1/2 + x^2/2$$
$$z^1/2 + z^3/2 = \hat{x}$$
$$z^1/2 + z^4/2 = (\beta x^2 + (1-\beta)\hat{x})/2 + (\alpha x^1 + (1-\alpha)\hat{x})/2$$

so that, by convexity, we must have the equalities

$$\rho(z^1, \bar{y}) = \rho(z^2, \bar{y}) = \rho(z^3, \bar{y}) = \rho(z^4, \bar{y}) = r.$$

for some $1 \geq r > 1/2$. Now note that

$$0 < \frac{2\alpha\beta}{\alpha+\beta} < \frac{2\alpha\beta}{\alpha+\alpha} = \beta < 1$$

and let

$$y = \left(\frac{2\alpha\beta}{\alpha+\beta}\right) z^2 + \left(1 - \frac{2\alpha\beta}{\alpha+\beta}\right) z^3.$$

Since $B(r)$ is convex, by Lemma 43, we must have $\rho(y, \bar{x}) \geq r$. On the other hand, it is

straightforward to verify the equality

$$z^4 = \gamma y + (1-\gamma)z^1$$

where

$$\gamma = \frac{(1-\alpha+1-\beta)(\alpha+\beta)}{2\beta(1-\alpha)+2\alpha(1-\beta)} \in (0,1)$$

and by convexity we must have $\rho(y,\bar{x}) < r$, a contradiction. □

Lemma 49. *$\|\cdot\|_B$ is Euclidean, i.e., $\|x\|_B = \sqrt{\langle x,x\rangle_B}$ where $\langle\cdot,\cdot\rangle_B$ is an inner product.*

Proof. We use a characterization of inner product spaces by [19], who showed that a normed linear space is an inner product space if and only if

$$\left\|\frac{1}{2}x + \frac{1}{2}y\right\| \leq \|\alpha x + (1-\alpha)y\| \quad \text{whenever } \|x\| = \|y\| = 1 \text{ and } 0 \leq \alpha \leq 1. \tag{A.10}$$

If $\|x\|_B = \|y\|_B = 1$ then x, y are on the boundary of B, hence $\rho(x+\hat{x},\bar{y}) = \rho(y+\hat{x},\bar{y}) = p > 1/2$ and $\rho(x+\hat{x}, y+\hat{x}) = 1/2$. Since ρ is convex, for each $0 \leq \alpha \leq 1$ we must have

$$\rho(\alpha x + (1-\alpha)y + \hat{x}, \bar{y}) \leq \rho(x/2 + y/2 + \hat{x}, \bar{y})$$

thus $\alpha x + (1-\alpha)y$ is on the boundary of $B(q) - \hat{x}$ and $x/2 + y/2$ is on the boundary of $B(q') - \hat{x}$ for some $q \leq q'$. By Lemma 48, the norm $\|\cdot\|_B$ satisfies (A.10). □

Now we extend the inner product $\langle\cdot,\cdot\rangle_B$ on the $n-2$ dimensional subspace $I(\bar{x}) - \hat{x}$ obtained in the last Lemma to an inner product $\langle\cdot,\cdot\rangle$ on the $n-1$ dimensional subspace $H - \hat{x}$. Let $v_1, \ldots, v_{n-2}$ be an orthonormal base for the subspace $I(\bar{x}) - \hat{x}$ endowed with $\langle\cdot,\cdot\rangle_B$. Let $v_{n-1} := \hat{x} - \bar{y}$ and for every $1 \leq i, j \leq n-1$ let $\langle v_i, v_j\rangle = 0$ if $i \neq j$ and $\langle v_i, v_j\rangle = 1$ if $i = j$. We let the norm be induced by this inner product $\|x\| := \sqrt{\langle x,x\rangle}$ for all $x \in H - \hat{x}$.

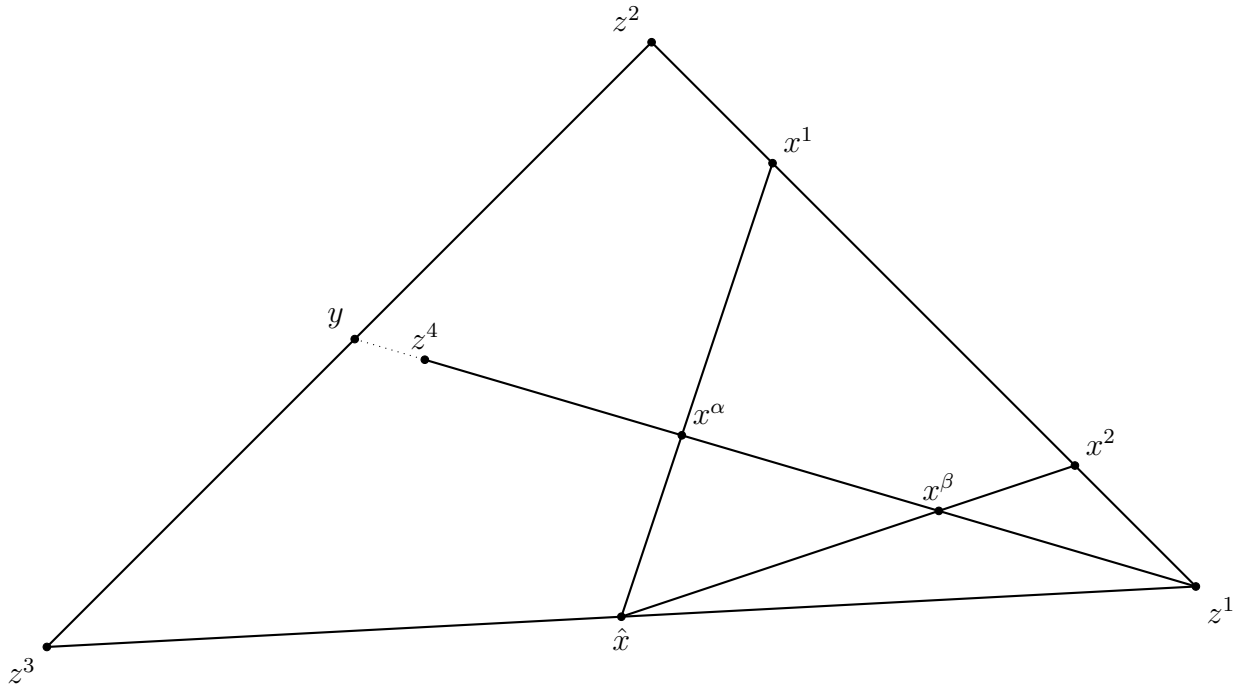


Figure A.1: Illustration of the proof of Lemma 48.

All lotteries shown are chosen fifty-fifty against each other in binary comparisons. Each lottery is also chosen with probability strictly larger than one-half against a lottery $\bar{x}$ (not shown). Lotteries x^1 and x^2 are each chosen with probability p against $\bar{x}$. Lotteries $x^\alpha = \alpha x^1 + (1-\alpha)\hat{x}$ and $x^\beta = \beta x^2 + (1-\beta)\hat{x}$ are each chosen with probability q against $\bar{x}$. When $\alpha < \beta$, the line through x^1 and x^2 must cross the line through x^α and x^β at a point z^1. The convexity postulate implies z^1, z^2, z^3, z^4 must each be chosen against $\bar{x}$ with the same probability r. But lottery z^4 is in the interior of the triangle formed by z^1, z^2, z^3, yielding a contradiction to the convexity postulate.

Lemma 50. *U and $\|\cdot\|$ provide an ordinal representation of ρ, that is, for any $w \neq x$ and $y \neq z$ we have*

$$\rho(w,x) \geq \rho(y,z) \iff \frac{U(w)-U(x)}{\|w-x\|} \geq \frac{U(y)-U(z)}{\|y-z\|}.$$

Proof. First, suppose $\rho(w,x) \geq \rho(y,z) > 1/2$. Then $w \succ x$, $y \succ z$ and since U represents $\succcurlyeq$ we have $U(w) > U(x)$ and $U(y) > U(z)$. Let

$$w' = \bar{y} + \frac{U(\hat{x}) - U(\bar{y})}{U(w) - U(x)}(w - x)$$
$$y' = \bar{y} + \frac{U(\hat{x}) - U(\bar{y})}{U(y) - U(z)}(y - z)$$

and note that $w', y' \in H$. Since U is linear, $U(w') = U(y') = U(\bar{x})$ and hence $w', y' \in I(\bar{x})$. By the linearity of ρ, $\rho(w', \bar{y}) = \rho(w,x) \geq \rho(y,z) = \rho(y', \bar{y})$. Hence $\|w' - \hat{x}\|_B \leq \|y' - \hat{x}\|_B$. By construction, $\hat{x} - \bar{y}$ is orthogonal to $I(\bar{x}) - \hat{x}$, and therefore

$$\begin{aligned}
\|w' - \bar{y}\|^2 &= \|w' - \hat{x} + \hat{x} - \bar{y}\|^2 \\
&= \|w' - \hat{x}\|^2 + \|\hat{x} - \bar{y}\|^2 \\
&\leq \|y' - \hat{x}\|^2 + \|\hat{x} - \bar{y}\|^2 \\
&= \|y' - \bar{y}\|^2
\end{aligned}$$

Thus

$$\left\| \frac{U(\hat{x}) - U(\bar{y})}{U(w) - U(x)}(w - x) \right\| = \|w' - \bar{y}\| \leq \|y' - \bar{y}\| = \left\| \frac{U(\hat{x}) - U(\bar{y})}{U(y) - U(z)}(y - z) \right\|$$

which implies

$$\frac{U(w) - U(x)}{\|w - x\|} \geq \frac{U(y) - U(z)}{\|y - z\|}.$$

Next, suppose $\rho(w,x) \geq 1/2 \geq \rho(y,z)$ with $w \neq x$ and $y \neq z$. Then $U(w) \geq U(x)$ and

$U(z) \geq U(y)$ which implies

$$\frac{U(w) - U(x)}{\|w - x\|} \geq 0 \geq \frac{U(y) - U(z)}{\|y - z\|}.$$

Finally, suppose $1/2 > \rho(w, x) \geq \rho(y, z)$. Then $\rho(z, y) \geq \rho(x, w) > 1/2$ and the desired inequality follows from the first step.

Reversing the argument to show that

$$\frac{U(w) - U(x)}{\|w - x\|} \geq \frac{U(y) - U(z)}{\|y - z\|} \implies \rho(w, x) \geq \rho(y, z)$$

is straightforward and left to the reader. □

Lemma 51. *The image of ρ is an interval $[1 - \bar{p}, \bar{p}]$.*

Proof. As we noted before, linearity implies ρ is entirely determined by the values of the mapping $x \mapsto \rho(x, \bar{y})$ for $x \in I(\bar{x})$. Hence, ρ achieves its maximum at $\bar{p} = \rho(\hat{x}, \bar{y})$. Linearity of ρ also implies ρ is entirely determined by the values of the mapping $x \mapsto \rho(x, \bar{y})$ for x in a unit sphere around $\bar{y}$. The continuity of ρ outside the diagonal implies $x \mapsto \rho(x, \bar{y})$ is continuous on the unit sphere around $\bar{y}$. The result then easily follows from the intermediate value theorem. □

To construct F, we first define an auxiliary function $f : [1 - \bar{p}, \bar{p}] \to \mathbb{R}$. Let $f(1/2) = 0$. For each $t \neq 1/2$, let $f(t) = [U(x) - U(y)]/\|x - y\|$ for any x, y such that $\rho(x, y) = t$. By Lemma 50 and Lemma 51, the function f is well defined. To see that the image of f must be a compact interval in $\mathbb{R}$, take any lottery $x \neq \hat{x}$ with $U(x) = U(\hat{x})$. Then we have $U(\hat{x} + t(x - \hat{x})) - U(\bar{y}) = U(\hat{x}) - U(\bar{y})$ for all $t > 0$ and $\|\hat{x} + t(x - \hat{x}) - \bar{y}\| \geq t\|x - \hat{x}\| - \|\hat{x} - \bar{y}\|$ which goes to infinity when t goes to infinity. Hence $[U(\hat{x} + t(x - \hat{x})) - U(\bar{y})]/\|\hat{x} + t(x - \hat{x}) - \bar{y}\|$ goes to zero when t goes to infinity. Thus and the image of f is the compact interval $[-T, T]$, where $T = [U(\hat{x}) - U(\bar{y})]/\|\hat{x} - \bar{y}\|$. By Lemma 50 f is strictly increasing and has an inverse.

Repeating the argument in the proof of Lemma 51 shows f is continuous. Leting $F = f^{-1}$ be the continuous inverse of f, it follows that $(U, \|\cdot\|, F)$ is a MEM representation of ρ. $\square$

A.3.4 Proof of Proposition 22

Let $\bar{x}, \bar{y}, \hat{x}$ be defined exactly as in the proof of Theorem 21. Let $(U, \|\cdot\|, F)$ be a MEM representation of ρ as in (4.4), and let $\langle \cdot, \cdot \rangle$ be the inner product that induces the norm.

Lemma 52. $\langle x - \hat{x}, \hat{x} - \bar{y} \rangle = 0$ *for all* x *with* $\rho(x, \hat{x}) = 1/2$.

Proof. This holds by construction for the particular representation obtained in the proof of Theorem 21, and now we show it holds for every representation. When $x = \hat{x}$ the statement is obviously true. Suppose $x \neq \hat{x}$. By Lemma 46 $\rho(x, \bar{y}) = \rho(2\hat{x} - x, \bar{y})$. By the representation (4.4) it must be $\|x - \bar{y}\| = \|2\hat{x} - x - \bar{y}\|$. Hence

$$\begin{aligned} \|x - \hat{x}\|^2 + 2\langle x - \hat{x}, \hat{x} - \bar{y} \rangle + \|\hat{x} - \bar{y}\|^2 &= \langle x - \bar{y}, x - \bar{y} \rangle \\ &= \langle 2\hat{x} - x - \bar{y}, 2\hat{x} - x - \bar{y} \rangle \\ &= \|x - \hat{x}\|^2 + 2\langle x - \hat{x}, \bar{y} - \hat{x} \rangle + \|\hat{x} - \bar{y}\|^2 \end{aligned}$$

which implies $4\langle x - \hat{x}, \hat{x} - \bar{y} \rangle = 0$ and we are done. $\square$

Lemma 53. $\rho(x, \hat{x}) = \rho(x', \hat{x}) = 1/2$ *and* $\rho(x, \bar{y}) = \rho(x', \bar{y})$ *implies* $\|x - \hat{x}\| = \|x' - \hat{x}\|$.

Proof. By the representation (4.4) we must have $\|x - \bar{y}\| = \|x' - \bar{y}\|$. By Lemma 52, $\langle x - \hat{x}, \hat{x} - \bar{y} \rangle = \langle x' - \hat{x}, \hat{x} - \bar{y} \rangle = 0$. Thus,

$$\|x - \hat{x}\|^2 + \|\hat{x} - \bar{y}\|^2 = \|x - \bar{y}\|^2 = \|x' - \bar{y}\|^2 = \|x' - \hat{x}\|^2 + \|\hat{x} - \bar{y}\|^2$$

and therefore $\|x - \hat{x}\| = \|x' - \hat{x}\|$ as desired. $\square$

To prove necessity, suppose $(U_1, \|\cdot\|_1, F_1)$ and $(U_2, \|\cdot\|_2, F_2)$ are two MEM representations of the same choice rule ρ. By the definition of MEM, both norms $\|x\|_1 = \sqrt{\langle x, x\rangle_1}$ and $\|x\|_2 = \sqrt{\langle x, x\rangle_2}$ are induced by inner products. The two norms and their respective inner products are defined on the linear subspace $\ker(\mathbf{1}) := \{x \in \mathbb{R}^n : x_1 + \cdots + x_n = 0\}$.

It is easy to see that the expected utility function $U_2 = U_1 = U$ is unique by the requirement that it is linear and that $U(\Delta) = [0, 1]$.

Let $\hat{x}, \bar{y}$ continue to denote the same elements fixed above. Consider any $x \in \ker(\mathbf{1})$ in the null space of U, that is $U(x) = 0$. Since $x \in \ker(\mathbf{1})$ we have $x + \hat{x} \in H$, where H is the hyperplane containing Δ. By linearity, we may assume without loss of generality that $x+\hat{x} \in \Delta$. Since $U(x+\hat{x}) = U(x)+U(\hat{x}) = U(\hat{x})$, the representation implies $\rho(x+\hat{x}, \hat{x}) = 1/2$. Lemma 52 implies

$$\langle x, \hat{x} - \bar{y}\rangle_1 = \langle (x + \hat{x}) - \hat{x}, \hat{x} - \bar{y}\rangle_1 = 0 = \langle (x + \hat{x}) - \hat{x}, \hat{x} - \bar{y}\rangle_2 = \langle x, \hat{x} - \bar{y}\rangle_2 .$$

Thus, every vector x in the null space of U is orthogonal to $\hat{x} - \bar{y}$ according to both inner products. In other words, $\ker(U)^{\perp_1} = \ker(U)^{\perp_2}$ is the single dimensional subspace of $\ker(\mathbf{1})$ given by $\{\alpha(\hat{x} - \bar{y}) \in \ker(\mathbf{1}) : \alpha \in \mathbb{R}\}$.

To show (i), fix a lottery $z \neq \hat{x}$ with $\rho(z, \hat{x}) = 1/2$ and let $A := \|z - \hat{x}\|_2/\|z - \hat{x}\|_1 > 0$. Now take any x with $U(x) = U(\hat{x})$. To show that $\|\cdot\|_2 = A\|\cdot\|_1$ on the null space of U, it suffices to show that $\|x - \hat{x}\|_2 = A\|x - \hat{x}\|_1$. This clearly holds if $x = \hat{x}$ so suppose $x \neq \hat{x}$. Let $q = \rho(x, \bar{y})$ and $p = \rho(z, \bar{y})$. Lemma 45 implies $p, q < \bar{p} = \rho(\hat{x}, \bar{y})$. Suppose wlog $p \geq q$. By Lemma 48 $B(p) = \hat{x} + \lambda[B(q) - \hat{x}]$ for some $0 \leq \lambda \leq 1$. Since $p < \bar{p}$ it must be $0 < \lambda \leq 1$. Then $\rho(\lambda x + (1 - \lambda)\hat{x}, \bar{y}) = p = \rho(z, \bar{y})$. By Lemma 53 we have

$$\|\hat{x} + \lambda(x - \hat{x}) - \hat{x}\|_2 = \|z - \hat{x}\|_2 = A\|z - \hat{x}\|_1 = A\|\hat{x} + \lambda(x - \hat{x}) - \hat{x}\|_1$$

hence $\lambda\|x-\hat{x}\|_2 = \lambda A\|x-\hat{x}\|_1$ and since $\lambda > 0$ we obtain $\|x-\hat{x}\|_2 = A\|x-\hat{x}\|_1$ as desired.

Since $\ker(U)^\perp$ is single-dimensional, (ii) must hold with $B := \|\hat{x}-\bar{y}\|_2/\|\hat{x}-\bar{y}\|_1$.

To see that (iii) holds, for each x with $U(x) = U(\hat{x})$, define $t(x) := [U(x)-U(\bar{y})]/\|x-\bar{y}\|_2$. Lemma 52 and items (i) and (ii) above imply

$$\|x-\bar{y}\|_2^2 = \|x-\hat{x}\|_2^2 + \|\hat{x}-\bar{y}\|_2^2 = A^2\|x-\hat{x}\|_1^2 + B^2\|\hat{x}-\bar{y}\|_2^2.$$

Let $T := F^{-1}(\max_{x,y}\rho(x,y)) = F^{-1}(\bar{p}) = [U(\hat{x})-U(\bar{y})]/\|\hat{x}-\bar{y}\|_1$. Substituting and rearranging we obtain for each x with $U(x) = U(\hat{x})$,

$$\frac{\|x-\hat{x}\|_1^2}{\|\hat{x}-\bar{y}\|_1^2} = \frac{T^2 - B^2t(x)^2}{A^2t(x)^2}.$$

And finally, by linearity of ρ for each $0 < t \leq T/B$ we have $t = t(x)$ for some x with $U(x) = U(\hat{x})$ thus

$$\begin{aligned}
F_2(t) = F_2(t(x)) &= \rho(x,\bar{y}) \\
&= F_1\left(\frac{U(x)-U(\bar{y})}{\|x-\bar{y}\|_1}\right) \\
&= F_1\left(\frac{U(\hat{x})-U(\bar{y})}{\sqrt{\|x-\hat{x}\|_1^2 + \|\hat{x}-\bar{y}\|_1^2}}\right) \\
&= F_1\left(\frac{T}{\sqrt{\frac{\|x-\hat{x}\|_1^2}{\|\hat{x}-\bar{y}\|_1^2}+1}}\right) \\
&= F_1\left(\frac{T}{\sqrt{\frac{T^2-B^2t(x)^2}{A^2t(x)^2}+1}}\right) \\
&= F_1\left(\frac{ATt}{\sqrt{T^2+(A^2-B^2)t^2}}\right),
\end{aligned}$$

and the results follows since $F_2(t) = 1 - F_2(-t)$.

Sufficiency is straightforward and left to the reader. □

A.3.5 Proof of Proposition 23

Necessity is shown in the main text. For sufficiency, suppose (u, d, F) is a MUM representation where u satisfies (4.6), d satisfies (4.7) and F is continuous. Let $x^1, \ldots, x^n$ denote the n degenerate lotteries in Δ. Since u satisfies (4.6), without loss of generality we can assume that $u(x^1) = 1$ and $u(x^n) = 0$. Let $U : \Delta \to [0, 1]$ be given by $U(x) = \sum_{i=1}^{n} u(x^i)x_i$ for each lottery $x = (x_1, \ldots, x_n)$ in Δ. Then U is automatically linear and onto.

To define the norm, note the $n - 1$ vectors $x^1 - x^n, x^2 - x^n, \ldots, x^{n-1} - x^n$ are linearly independent and span $\ker(\mathbf{1})$. We define an inner product on $\ker(\mathbf{1})$ by letting

$$\left\langle x^i - x^n, x^k - x^n \right\rangle := \frac{1}{2}\left[d(x^i, x^n)^2 + d(x^k, x^n)^2 - d(x^i, x^k)^2\right]$$

for each $i, k = 1, \ldots, n-1$ and extending it to $\ker(\mathbf{1})$ by linearity. To see that $\langle \cdot, \cdot \rangle$ is indeed an inner product, note that condition (4.7) implies the quadratic form

$$(\alpha_1, \ldots, \alpha_{n-1}) \mapsto \frac{1}{2}\sum_{i=1}^{n-1}\sum_{k=1}^{n-1}\left[d(x^i, x^n)^2 + d(x^k, x^n)^2 - d(x^i, x^k)^2\right]\alpha_i\alpha_k$$

is positive definite and hence $\langle x, x \rangle = 0$ implies $x = 0$. We let $\|x\| := \sqrt{\langle x, x \rangle}$ be the norm

on $\ker(\mathbf{1})$ induced by this inner product. Then for each $i, k = 1, \ldots, n-1$ we have

$$\begin{aligned}
\|x^i - x^k\|^2 &= \left\langle x^i - x^k, x^i - x^k \right\rangle \\
&= \left\langle x^i - x^n + x^n - x^k, x^i - x^n + x^n - x^k \right\rangle \\
&= \left\langle x^i - x^n, x^i - x^n \right\rangle + \left\langle x^n - x^k, x^n - x^k \right\rangle - 2\left\langle x^i - x^n, x^k - x^n \right\rangle \\
&= d(x^i, x^n)^2 + d(x^k, x^n)^2 - \left[d(x^i, x^n)^2 + d(x^k, x^n)^2 - d(x^i, x^k)^2\right] \\
&= d(x^i, x^k)^2
\end{aligned}$$

and $\|x^i - x^n\|^2 = (1/2)\left[d(x^i, x^n)^2 + d(x^i, x^n)^2 - d(x^i, x^i)^2\right] = d(x^i, x^n)^2$ as desired. □

A.4 Rational Contextual Choices under Imperfect Perception of Attributes

A.4.1 Proof of Lemma 31

Proof. Proof of Lemma 31 We calculate directly the expected utility

$$\begin{aligned}
\mathbb{E}[u(\mathcal{X})|X, Y] =& \mathbb{E}\left[-e^{\gamma \mathcal{X}_1} - e^{\rho \mathcal{X}_2} | X, Y\right] \\
=& -\exp\left(\gamma \frac{(t_1^2+1)X_1 - Y_1}{2+t_1^2} + \gamma^2 \frac{1}{2(2+t_1^2)}\right) - \exp\left(\rho \frac{(t_2^2+1)X_2 - Y_2}{2+t_2^2} + \rho^2 \frac{1}{2(2+t_2^2)}\right) \\
=& -\exp\left(\gamma \frac{(t_1^2+1)x_1^* - y_1^* + t_1^2\epsilon_1}{2+t_1^2} + \gamma^2 \frac{1}{2(2+t_1^2)}\right) - \exp\left(\rho \frac{(t_2^2+1)x_2^* - y_2^* + t_2^2\epsilon_2}{2+t_2^2} + \rho^2 \frac{1}{2(2+}\right.
\end{aligned}$$

where the second equality is due to the normally distributed exponents. The third equality is due to the identities $\mathbf{x}^* + \epsilon = X$, $\mathbf{y}^* + \epsilon = Y$. Similarly,

$$\mathbb{E}[u(\mathcal{Y})|X, Y] = -\exp\left(\gamma \frac{(t_1^2+1)y_1^* - x_1^* + t_1^2\epsilon_1}{2+t_1^2} + \gamma^2 \frac{1}{2(2+t_1^2)}\right) - \exp\left(\rho \frac{(t_2^2+1)y_2^* - x_2^* + t_2^2\epsilon_2}{2+t_2^2} + \rho^2 \frac{1}{2(2+t}\right.$$

Hence given $\mathbf{x}^*, \mathbf{y}^*$ and ϵ, the agent would choose $\mathbf{x}$ over $\mathbf{y}$ iff $\mathbb{E}[u(\mathcal{X})|X, Y] > \mathbb{E}[u(\mathcal{Y})|X, Y]$. Suppose $x_1^* > y_1^*$ and $y_2^* > x_2^*$, then we see that $\mathbf{x}$ is chosen over $\mathbf{y}$ iff

$$\exp\left(\frac{\gamma^2}{2(2+t_1^2)} - \frac{\rho^2}{2(2+t_2^2)}\right) \frac{\exp\left(\gamma\frac{(t_1^2+1)y_1^* - x_1^*}{2+t_1^2}\right) - \exp\left(\gamma\frac{(t_1^2+1)x_1^* - y_1^*}{2+t_1^2}\right)}{\exp\left(\rho\frac{(t_2^2+1)x_2^* - y_2^*}{2+t_2^2}\right) - \exp\left(\rho\frac{(t_2^2+1)y_2^* - x_2^*}{2+t_2^2}\right)} \geq \exp\left(\frac{\rho t_2^2\epsilon_2}{2+t_2^2} - \frac{\gamma t_1^2\epsilon_1}{2+t_1^2}\right). \tag{$\dagger$}$$

Since $x_1^* > y_1^*$ and $y_2^* > x_2^*$, we can take natural-log on both hand sides of ($\dagger$) to obtain the following equivalent condition

$$\frac{\gamma^2}{2(2+t_1^2)} - \frac{\rho^2}{2(2+t_2^2)} + \ln\left(\frac{\exp\left(\gamma\frac{(t_1^2+1)y_1^* - x_1^*}{2+t_1^2}\right) - \exp\left(\gamma\frac{(t_1^2+1)x_1^* - y_1^*}{2+t_1^2}\right)}{\exp\left(\rho\frac{(t_2^2+1)x_2^* - y_2^*}{2+t_2^2}\right) - \exp\left(\rho\frac{(t_2^2+1)y_2^* - x_2^*}{2+t_2^2}\right)}\right) \geq \frac{\rho t_2^2\epsilon_2}{2+t_2^2} - \frac{\gamma t_1^2\epsilon_1}{2+t_1^2}.$$

Notice that RHS follows a normal distribution $\mathcal{N}\left(0, \left(\frac{\rho}{2+t_2^2}\right)^2 t_2 + \left(\frac{\gamma}{2+t_1^2}\right)^2 t_1\right)$. We can standardize both hand side by multiplying $1/\sqrt{\left(\frac{\rho\sqrt{t_2}}{2+t_2^2}\right)^2 + \left(\frac{\gamma\sqrt{t_1}}{2+t_1^2}\right)^2}$. Hence $\mathbf{x}^*$ is chosen over $\mathbf{y}^*$ iff some standard normal random variable Z is below the threshold θ defined below:

$$\theta(\gamma, \rho, \mathbf{x}^*, \mathbf{y}^*, t) := \frac{1}{\sqrt{\left(\frac{\rho\sqrt{t_2}}{2+t_2^2}\right)^2 + \left(\frac{\gamma\sqrt{t_1}}{2+t_1^2}\right)^2}} \left[\frac{\gamma^2}{2(2+t_1^2)} - \frac{\rho^2}{2(2+t_2^2)} + \ln\left(\frac{\exp\left(\gamma\frac{(t_1^2+1)y_1^* - x_1^*}{2+t_1^2}\right) - \exp\left(\gamma\frac{(t_1^2+1)x_1^* -}{2+t_1^2}\right.\right.\right.$$
$$\left.\left.\left.\frac{}{}\right)}{\exp\left(\rho\frac{(t_2^2+1)x_2^* - y_2^*}{2+t_2^2}\right) - \exp\left(\rho\frac{(t_2^2+1)y_2^* -}{2+t_2^2}\right.\right.\right.$$

□

A.4.2 Proof of Theorem 33

Proof. Proof of Theorem 33 It suffices to show that under our assumptions, for every real-

ization of ϵ the following inequality holds

$$\mathbb{E}[u(\mathcal{X})|X,Y,Z+\Delta]-\mathbb{E}[u(\mathcal{Y})|X,Y,Z+\Delta]>\mathbb{E}[u(\mathcal{X})|X,Y,Z]-\mathbb{E}[u(\mathcal{Y})|X,Y,Z].$$

Conditional on X,Y,Z, the posterior for $\mathcal{X}$ is

$$\begin{aligned}
&\Pr(\mathcal{X}|X,Y,Z)\propto\exp\left(-\frac{\mathcal{X}'\Omega^{-1}\mathcal{X}}{2}\right)\exp\left(-\frac{\mathcal{Y}'\Omega^{-1}\mathcal{Y}}{2}\right)\exp\left(-\frac{\mathcal{Z}'\Omega^{-1}\mathcal{Z}}{2}\right)\exp\left(-\frac{(X-\mathcal{X})'T(X-\mathcal{X})}{2}\right)\times 1\\
&\propto\exp\left(-\frac{1}{2}\left[\mathcal{X}'\left(3\Omega^{-1}+T\right)\mathcal{X}-2\left(TX-\Omega^{-1}(Y+Z-2X)\right)'\mathcal{X}\right]\right)\\
&\propto\exp\left(-\frac{1}{2}\left(\mathcal{X}-\left(3\Omega^{-1}+T\right)^{-1}\left(TX-\Omega^{-1}(Y+Z-2X)\right)\right)'\left(3\Omega^{-1}+T\right)\left(\mathcal{X}-\left(3\Omega^{-1}+T\right)^{-1}\left(TX-\Omega^{-1}\right.\right.\right.
\end{aligned}$$

So we denote the above posterior distribution of $\mathcal{X}|X,Y,Z$ by $\mathcal{N}\left(\mu(\mathbf{x}^*;\mathbf{y}^*,\mathbf{z}^*,\epsilon),\hat{\Omega}\right)$, where

$$\begin{aligned}
\mu(\mathbf{x}^*;\mathbf{y}^*,\mathbf{z}^*,\epsilon)&:=\left(3\Omega^{-1}+T\right)^{-1}\left(T\mathbf{x}^*+T\epsilon-\Omega^{-1}(\mathbf{y}^*+\mathbf{z}^*-2\mathbf{x}^*)\right)\\
&=\left(3\Omega^{-1}+T\right)^{-1}\left(TX-\Omega^{-1}(Y+Z-2X)\right),\\
\text{and}\quad\hat{\Omega}&:=\left(3\Omega^{-1}+T\right)^{-1}.
\end{aligned}$$

Denote the density of $\mathcal{X}|X,Y,Z\sim\mathcal{N}(\mu,\Omega)$ by $\phi(\mathcal{X}-\mu,\Omega)$. The posterior expected utility is therefore

$$\begin{aligned}
\mathbb{E}[u(\mathcal{X})|X,Y,Z]&=\int_{\mathbb{R}^2}u(\mathcal{X})\times\phi\left(\mathcal{X}-\mu(\mathbf{x}^*;\mathbf{y}^*,\mathbf{z}^*,\epsilon),\hat{\Omega}\right)\mathrm{d}\mathcal{X}\\
&=\int_{\mathbb{R}^2}u(\mathbf{s}+\mu(\mathbf{x}^*;\mathbf{y}^*,\mathbf{z}^*,\epsilon))\times\phi\left(\mathbf{s},\hat{\Omega}\right)\mathrm{d}\mathbf{s}.
\end{aligned}$$

Similarly,

$$\mathcal{Y}|X,Y,Z\sim\mathcal{N}\left(\mu(\mathbf{y}^*;\mathbf{x}^*,\mathbf{z}^*,\epsilon),\hat{\Omega}\right).$$

Because

$$\begin{aligned}\mu(\mathbf{x}^*;\mathbf{y}^*,\mathbf{z}^*,\epsilon) :=& \hat{\Omega}\left(T\mathbf{x}^* + T\epsilon - \Omega^{-1}(\mathbf{y}^* + \mathbf{z}^* - 2\mathbf{x}^*)\right)\\ =& \mu(\mathbf{y}^*;\mathbf{x}^*,\mathbf{z}^*,\epsilon) - (\mathbf{y}^* - \mathbf{x}^*),\end{aligned}$$

we have

$$\mathbb{E}[u(\mathcal{Y})|X,Y,Z] = \int_{\mathbb{R}^2} u(\mathbf{s} + (\mathbf{y}^* - \mathbf{x}^*) + \mu(\mathbf{x}^*;\mathbf{y}^*,\mathbf{z}^*,\epsilon)) \times \phi\left(\mathbf{s},\hat{\Omega}\right) \mathrm{d}\mathbf{s}.$$

Recall that $\mu(\mathbf{x}^*;\mathbf{y}^*,\mathbf{z}^*,\epsilon) = \hat{\Omega}T\mathbf{x}^* + \hat{\Omega}T\epsilon - \hat{\Omega}\Omega^{-1}\mathbf{y}^* - \hat{\Omega}\Omega^{-1}\mathbf{z}^* + 2\hat{\Omega}\Omega^{-1}\mathbf{x}^*$. Substitute in $\mathbf{z}'^* := \mathbf{z}^* + \Delta$ for $\mathbf{z}^*$ we have

$$\begin{aligned}&\mathbb{E}[u(\mathcal{X})|X,Y,Z+\Delta] - \mathbb{E}[u(\mathcal{Y})|X,Y,Z+\Delta]\\ =& \int_{\mathbb{R}^2} u(\mathbf{s} + \mu(\mathbf{x}^*;\mathbf{y}^*,\mathbf{z}^*+\Delta,\epsilon)) \times \phi\left(\mathbf{s},\hat{\Omega}\right) \mathrm{d}\mathbf{s} - \int_{\mathbb{R}^2} u(\mathbf{s} + (\mathbf{y}^* - \mathbf{x}^*) + \mu(\mathbf{x}^*;\mathbf{y}^*,\mathbf{z}^*+\Delta,\epsilon)) \times \phi\left(\mathbf{s},\hat{\Omega}\right) \mathrm{d}\mathbf{s}\\ =& \int_{\mathbb{R}^2} \left[u\left(\mathbf{s} + \mu(\mathbf{x}^*;\mathbf{y}^*,\mathbf{z}^*,\epsilon) - \hat{\Omega}\Omega^{-1}\Delta\right) - u\left(\mathbf{s} + (\mathbf{y}^* - \mathbf{x}^*) + \mu(\mathbf{x}^*;\mathbf{y}^*,\mathbf{z}^*,\epsilon) - \hat{\Omega}\Omega^{-1}\Delta\right)\right] \times \phi\left(\mathbf{s},\hat{\Omega}\right) \mathrm{d}\mathbf{s}.\end{aligned}$$

Since u is standard, and $y_1^* < x_1^*$, and $y_2^* > x_2^*$, if $-\hat{\Omega}\Omega^{-1}\Delta \in (-\infty,0)\times(0,\infty)$, i.e. the second quadrent, then

$$\begin{aligned}&u\left(\mathbf{s} + \mu(\mathbf{x}^*;\mathbf{y}^*,\mathbf{z}^*,\epsilon) - \hat{\Omega}\Omega^{-1}\Delta\right) - u\left(\mathbf{s} + (\mathbf{y}^* - \mathbf{x}^*) + \mu(\mathbf{x}^*;\mathbf{y}^*,\mathbf{z}^*,\epsilon) - \hat{\Omega}\Omega^{-1}\Delta\right)\\ >& u\left(\mathbf{s} + \mu(\mathbf{x}^*;\mathbf{y}^*,\mathbf{z}^*,\epsilon)\right) - u\left(\mathbf{s} + (\mathbf{y}^* - \mathbf{x}^*) + \mu(\mathbf{x}^*;\mathbf{y}^*,\mathbf{z}^*,\epsilon)\right)\end{aligned}$$

for all $\mathbf{s}$ and ϵ. When we integrate out $\mathbf{s}$, we have $\mathbb{E}[u(\mathcal{X})|X,Y,Z+\Delta] - \mathbb{E}[u(\mathcal{Y})|X,Y,Z+\Delta] > \mathbb{E}[u(\mathcal{X})|X,Y,Z] - \mathbb{E}[u(\mathcal{X})|X,Y,Z]$ for every realization of ϵ.

Therefore, one sufficient condition is that $-\hat{\Omega}\Omega^{-1}\Delta \in (-\infty,0)\times(0,\infty)$. If this condition

holds, we have $-\hat{\Omega}\Omega^{-1}\Delta = \mathbf{w}$ for some $w_1 < 0$, and $w_2 > 0$. In order to show the decoy choice patter, we just need to show there exists Δ with $\Delta_1 > \Delta_2$ such that this condition holds.

Recall that we had normalized Ω so that for some $r \in (-1, 1)$,

$$\Omega = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$$

and the noise has variance

$$T^{-1} = \begin{bmatrix} 1/t_1^2 & R/(t_1 t_2) \\ R/(t_1 t_2) & 1/t_2^2 \end{bmatrix}.$$

We can calculate

$$\Omega^{-1} = \begin{bmatrix} 1/(1-r^2) & -r/(1-r^2) \\ -r/(1-r^2) & 1/(1-r^2) \end{bmatrix} \text{ and } T = \begin{bmatrix} t_1 & 0 \\ 0 & t_2 \end{bmatrix} \begin{bmatrix} 1/(1-R^2) & -R/(1-R^2) \\ -R/(1-R^2) & 1/(1-R^2) \end{bmatrix} \begin{bmatrix} t_1 & 0 \\ 0 & t_2 \end{bmatrix};$$

It follows that

$$\begin{aligned}
\Delta =& -\Omega\hat{\Omega}^{-1}\mathbf{w} = -\Omega(3\Omega^{-1} + T)\mathbf{w} = -(3I + \Omega T)\mathbf{w} \\
=& -\left(\begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix} + \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix} \begin{bmatrix} t_1 & 0 \\ 0 & t_2 \end{bmatrix} \begin{bmatrix} 1/(1-R^2) & -R/(1-R^2) \\ -R/(1-R^2) & 1/(1-R^2) \end{bmatrix} \begin{bmatrix} t_1 & 0 \\ 0 & t_2 \end{bmatrix} \right) \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \\
=& -\begin{bmatrix} 3 + \frac{t_1^2 - t_1 t_2 r R}{1-R^2} & \frac{t_2^2 r - t_1 t_2 R}{1-R^2} \\ \frac{t_1^2 r - t_1 t_2 R}{1-R^2} & 3 + \frac{t_2^2 - t_1 t_2 r R}{1-R^2} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}
\end{aligned}$$

Since $w_1 < 0$, and $w_2 > 0$, the sufficient condition to holds when Δ is some positive linear

combinations of the two vectors

$$\{\begin{bmatrix} 3(1-R^2)+t_1^2-t_1t_2rR \\ t_1^2r-t_1t_2R \end{bmatrix}, -\begin{bmatrix} t_2^2r-t_1t_2R \\ 3(1-R^2)+t_2^2-t_1t_2rR \end{bmatrix}\}.$$

And the decoy choice pattern holds when there exists such a Δ with $\Delta_1 > \Delta_2$. In other words, the decoy choice pattern holds if

$$\begin{cases} 3(1-R^2)+t_1^2-t_1t_2rR > t_1^2r-t_1t_2R \\ \text{or} \\ -(t_2^2r-t_1t_2R) > -(3(1-R^2)+t_2^2-t_1t_2rR), \end{cases} \Leftrightarrow \begin{cases} 3(1-R^2) > (r-1)(t_1^2+t_1t_2R) \\ \text{or} \\ 3(1-R^2) > (r-1)(t_2^2+t_1t_2R). \end{cases}$$

Because $r, R \in (-1,1)$ and $t_1, t_2 > 0$, it is impossible for both $t_1 + t_2R < 0$ and $t_2 + t_1R < 0$ to hold simultaneously. Therefore the decoy choice pattern holds.

A.4.3 Proof of Theorem 34

Proof. Proof of Theorem 34 As before, we start with the Bayesian posterior

$$\begin{aligned} \Pr(\mathcal{X}|X,Z) &\propto \exp\left(-\frac{\mathcal{X}'\Omega^{-1}\mathcal{X}}{2}\right)\exp\left(-\frac{\mathcal{Z}'\Omega^{-1}\mathcal{Z}}{2}\right)\exp\left(-\frac{(X-\mathcal{X})'T(X-\mathcal{X})}{2}\right)\times 1_{\{X-\mathcal{X}=Z-\mathcal{Z}\}} \\ &= \exp\left(-\frac{1}{2}\left[\mathcal{X}'\left(2\Omega^{-1}+T\right)\mathcal{X}-2\left(TX-\Omega^{-1}(Z-X)\right)'\mathcal{X}\ldots\right]\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\mathcal{X}-\left(2\Omega^{-1}+T\right)^{-1}\left(TX-\Omega^{-1}(Z-X)\right)\right)'\left(2\Omega^{-1}+T\right)\left(\mathcal{X}-\ldots\right)\right) \end{aligned}$$

Therefore, the posterior inference for $\mathbf{x}^*$ is

$$\begin{aligned}\mathcal{X}|X,Z \sim & \mathcal{N}\left(\left(2\Omega^{-1}+T\right)^{-1}\left(TX-\Omega^{-1}(Z-X)\right),\left(2\Omega^{-1}+T\right)^{-1}\right)\\ =& \mathcal{N}\left(\left(2\Omega^{-1}+T\right)^{-1}\left(T\mathbf{x}^*+T\epsilon-\Omega^{-1}(\mathbf{z}^*-\mathbf{x}^*)\right),\left(2\Omega^{-1}+T\right)^{-1}\right)\\ :=& \mathcal{N}\left(\mu(\mathbf{x}^*;\mathbf{z}^*,\epsilon),\hat{\Omega}\right)\end{aligned}$$

Similarly, $\mathcal{Z}|X,Z \sim \mathcal{N}\left(\mu(\mathbf{z}^*;\mathbf{x}^*,\epsilon),\hat{\Omega}\right)$. Observe that they have the same variance, and that

$$\begin{aligned}&\mu(\mathbf{z}^*;\mathbf{x}^*,\epsilon)-\mu(\mathbf{x}^*;\mathbf{z}^*,\epsilon)\\ =&\left(2\Omega^{-1}+T\right)^{-1}\left(T\mathbf{z}^*+T\epsilon-\Omega^{-1}(\mathbf{x}^*-\mathbf{z}^*)\right)-\left(2\Omega^{-1}+T\right)^{-1}\left(T\mathbf{x}^*+T\epsilon-\Omega^{-1}(\mathbf{z}^*-\mathbf{x}^*)\right)\\ =&\mathbf{z}^*-\mathbf{x}^*>0.\end{aligned}$$

Therefore the posterior inference distribution for $\mathbf{z}^*$ is that for $\mathbf{x}^*$ translated by the vector $\mathbf{z}^*-\mathbf{x}^*>0$. Since standard preference is increasing in both attributes, we have for every $\epsilon \in \mathbb{R}^2$

$$\mathbb{E}[u(\mathcal{X})|X,Z] < \mathbb{E}[u(\mathcal{Z})|X,Z].$$

Hence the rational agent chooses $\mathbf{z}$ over $\mathbf{x}$ with probability 1. □