Random Choice and Learning

Paulo Natzenzon

First version: November 2010; Updated: April 2015

Abstract

The attraction effect and other decoy effects are often understood as anomalies and modeled as departures from rational inference. We show how these decoy effects may arise from simple Bayesian updating. Our new model, the *Bayesian probit*, has the same parameters as the standard multinomial probit model: each choice alternative is associated with a Gaussian random variable. Unlike the standard probit and any random utility model, however, the Bayesian probit can jointly accommodate similarity effects, the attraction effect, and the compromise effect. We provide a new definition of revealed similarity from observed choice behavior, and show that in the Bayesian probit revealed similarity is captured by signal correlation. We also show that signal averages capture revealed preference; and that signal precision captures the decision maker’s familiarity with the options. This link of parameters to observable choice behavior facilitates measurement and provides a useful tool for discrete choice applications.

---

*Department of Economics, Washington University in St. Louis. E-mail: pnatzenzon@wustl.edu. This paper is based on the first chapter of my doctoral dissertation at Princeton University. I wish to thank my advisor, Faruk Gul, for his continuous guidance and dedication. I am grateful to Wolfgang Pesendorfer for many discussions in the process of bringing this project to fruition. I also benefited from numerous conversations with Dilip Abreu, Roland Bénabou, Meir Dan-Cohen, Daniel Gottlieb, David K. Levine, Justinas Pelenis, and seminar participants at Alicante, Arizona State, Berkeley, the D-TEA Workshop, Duke, IESE, IMPA, Johns Hopkins, Kansas University, Haifa, the Harvard-MIT Theory Seminar, Hebrew University, NYU, Princeton, the RUD Conference at the Colegio Carlo Alberto, Toronto, UBC, WUSTL and the D-Day workshop at Yale. All remaining errors are my own.
1 Introduction

Suppose a consumer were equally likely to pick either option when presented with a choice of two comparable personal computers A and B. How would the introduction of a third option C that is similar to B affect the probabilities of choosing A and B? Tversky’s (1972) well-known similarity hypothesis offers one answer:

*The addition of an alternative to an offered set ‘hurts’ alternatives that are similar to the added alternative more than those that are dissimilar to it.*

According to this hypothesis, options that share more features compete more intensely with each other for the consumer’s attention. If computers B and C are closer substitutes, then introducing C should hurt B proportionally more than A.

Tversky’s similarity hypothesis holds in a large number of settings and underlies most of the tools currently used in discrete choice estimation (see for example McFadden (2001) and Train (2009)). But a sizable empirical literature, originating in marketing, has identified situations in which the hypothesis systematically fails. The most prominent example of this systematic failure is called the attraction effect.

The attraction effect refers to an increase in the probability of choosing computer B when option C is clearly dominated by B but not by A. For example, computer C may be almost identical to computer B while being clearly worse in minor ways, such as having a slightly smaller screen and costing a little more. While the dominated computer C is almost never chosen, consumers are more likely to choose B in the presence of C. Thus, the attraction effect is in exact opposition to the similarity hypothesis. Huber, Payne and Puto (1982) provide the first experimental evidence for the attraction effect. The effect has been shown to hold in many different settings, including choice over political candidates, risky alternatives, investment decisions, medical decisions, and job candidate evaluation (Ok, Ortoleva and Riella (forthcoming) provide references to many of these studies).

How can we reconcile Tversky’s similarity hypothesis with the observed attraction effect? In this paper, we propose a new model of individual decision making, the Bayesian probit. Our main contribution is a unified theory that explains under which conditions similarity plays the role predicted by Tversky’s hypothesis, and when it may generate the attraction effect and related phenomena (such as the compromise effect).

The Bayesian probit has the same parameters of the standard multinomial probit model—a collection of joint normally distributed random variables, one variable for each choice alternative. The random variables represent the decision maker’s incom-
plete information about the utility of the alternatives. Based on these noisy signals, the decision maker updates a symmetric prior using Bayes’ rule and chooses the alternative with highest posterior expected utility. In what is a key departure from random utility models, this decision maker sometimes picks an alternative that doesn’t have the highest signal realization.

The main reason for the ability of our model to account for the attraction effect, the compromise effect, and other context-dependent choice behaviors is that, under incomplete information, a Bayesian decision maker is more likely to choose among options that are easier to compare. We provide a simple example of this phenomenon in Section 2, and introduce our parametric model in Section 3.

In Section 4 we provide the behavioral content of each parameter in the model. First, we define a stochastic notion of revealed preference, and show that in the model it is represented by signal averages. Hence, our decision maker maximizes a single, rational preference over the alternatives, represented by a numerical utility function, albeit under incomplete information. Second, we note that signal precision captures how much the decision maker was able to learn about her true ranking of the options before making a choice. Our model nests the standard rational deterministic choice model in the limit as signal precision goes to infinity. Finally, we propose a new definition of revealed similarity, based only on the random choice function, and show that in the model it is represented by signal correlation.

Section 5 contains the main results. Using the Bayesian probit, we provide a unified explanation for several empirical regularities as arising from the interaction of preference, similarity, and information precision.

In Section 6 we show that the Bayesian probit has the same ‘portability’ as the standard probit to different discrete choice applications. The Bayesian probit model allows the analyst to formulate and test additional parametric assumptions that specify how utility and similarity depend explicitly on the location of the alternatives in the space of observable characteristics. We focus on a discrete choice setting of central interest in economics—individual choice over money lotteries. We use the experimental results of Soltani, De Martino and Camerer (2012) to illustrate several advantages of the Bayesian probit over existing models. We also connect the main results of Section 5 to the setting with observable attributes.

In Section 7 we present additional testable implications of the model. We also obtain novel closed-form expressions for the choice probabilities of the standard probit and the Bayesian probit in some special cases. Section 8 concludes.
1.1 Related Literature

In order to accommodate the attraction and the compromise effects, the bounded rationality literature has generated a large collection of formal models. For example, the attraction effect arises as a result of heuristics based on attribute salience in Bordalo et al. (2013), limited attention in Masatlioglu et al. (2012) and bargaining among multiple selves in de Clippel and Eliaz (2012). In contrast to these models of bounded rationality, the Bayesian probit provides an account of how these effects may arise from individual, rational inference under incomplete information.

The literature on random choice mostly consists of *random utility maximizers*. In this formulation, the agent is defined by a fixed probability distribution over utility functions. Each time the decision maker faces a choice problem, she selects a utility function from that distribution and chooses the element from the choice set that maximizes it. These include the classic models of Thurstone (1927), Luce (1959), Tversky (1972) and, more recently, the random consideration set rules of Manzini and Mariotti (2014) and the attribute rules of Gul, Natenzon and Pesendorfer (2014). They also include most econometric discrete choice models commonly used in applications, such as probit, logit and nested logit.

Every random utility maximizer satisfies a monotonicity property: the choice probability of existing alternatives can only decrease when a new alternative is introduced. Hence none of these models can account for empirical regularities of monotonicity violation such as the attraction effect. In order to accommodate these effects, our new model, the Bayesian probit, departs from random utility maximization.

Game-theoretic models with standard rational players have been shown to be compatible with decoy effects. The seminal paper is Kamenica (2008), which shows how the compromise effect may arise in equilibrium in a market with a monopolist firm and a continuum of standard rational consumers. Similarly, de Clippel and Eliaz (2012) show how the attraction and the compromise effects arise in equilibrium in collective decision making when agents use the Rawlsian arbitration rule as a bargaining protocol. In contrast to these models, our decision-theoretic model explains how decoy effects may arise from individual, rational inference, even in the absence of strategic interactions. This makes our new model particularly relevant for some discrete choice applications. We discuss in more detail the advantages of our model using an example of a particular experimental setting in Section 6.
2 Example

This example illustrates how a classic decoy effect, the attraction effect (Huber et al. (1982)), may arise from rational, Bayesian inference. Suppose a consumer is about to be presented with three options $A$, $B$ and $C$ for a new personal computer. This consumer has no information a priori that favors one alternative over another; her prior beliefs about the utility of the alternatives are given by any symmetric, absolutely continuous distribution. Hence, before receiving any information about the options, she assigns probability $1/6$ to each possible strict ranking involving $A$, $B$ and $C$, namely

\[
A \succ B \succ C \quad A \succ C \succ B \\
B \succ A \succ C \quad C \succ A \succ B \\
B \succ C \succ A \quad C \succ B \succ A
\]

Now suppose she obtains some information about the options (for example, by inspecting the three alternatives), which (i) conveys with a high degree of certainty that computer $B$ is better than computer $C$; and (ii) conveys very little or nothing about the relative position of computer $A$. To make it extreme, suppose she learns that the event ($B \succ C$) occurred and nothing else. Updating the prior with this information will result in probability $1/3$ given to each of

\[
A \succ B \succ C \\
B \succ A \succ C \\
B \succ C \succ A
\]

and hence computer $B$ now has $2/3$ probability of being the best option. If she had learned ($C \succ B$), then computer $C$ would have been the favored option. In summary: starting from a symmetric prior, any information that allows a precise ranking of computers $B$ and $C$ will make computer $A$ unlikely to be chosen.

The attraction effect arises in contexts of asymmetric dominance: when computer $C$ is clearly dominated by the computer $B$ but not by computer $A$ (Huber et al., 1982). Dominance makes the comparison of $B$ and $C$ much easier than comparing $A$ and $B$ or comparing $A$ and $C$. With limited information (for example, with limited time to evaluate the options before making a choice), the decision maker will have a higher degree of confidence in the ranking of computers $B$ and $C$ than about the relative position of computer $A$. Hence compared to a menu without dominance where the two only options are $A$ and $B$, introducing a third option $C$ that is clearly dominated by $B$ but not by $A$ will favor a choice of $B$ in detriment of $A$. 

5
The example above is an extreme case in which alternative $A$ is never chosen. Next we introduce our parametric model, the Bayesian probit, which takes a convenient functional form. Prior beliefs are Gaussian, the information obtained by the decision maker before making a choice takes the form of Gaussian signals, and therefore posterior beliefs are also Gaussian. The parametric model can accommodate the extreme example above (see Proposition 5), but also allow us to study less extreme cases.

3 Learning Process

Let $A$ be a non-empty set and $\mathcal{A}$ the collection of all non-empty finite subsets of $A$. The set $A$ is the universe of choice alternatives and each subset $B \in \mathcal{A}$ represents a choice situation (a menu of choices) that a decision maker may face. A function $\rho : A \times \mathcal{A} \to [0, 1]$ is a random choice rule if $\sum_{i \in B} \rho(i, B) = 1$ for all $B \in \mathcal{A}$ and $\rho(j, B) = 0$ for all $j \notin B$. The value $\rho(i, B)$ is the probability of choosing alternative $i \in A$, when the selection must be made from $B \in \mathcal{A}$.

In our model, the true preferences of the decision maker are given by a utility function $\mu : A \to \mathbb{R}$. We write $\mu_i := \mu(i)$ for the utility of alternative $i$. The decision maker chooses under incomplete information about $\mu$. A priori, every $\mu_i$ is seen as independently drawn from a standard normal distribution.\(^1\)

When facing a menu of choice alternatives $B = \{1, 2, \ldots, n\}$ the decision maker observes a noisy signal about utility $X_i = \mu_i + \varepsilon_i$ for each alternative $i \in B$. The random vector of signals $(X_i)_{i \in B}$ represents how much the decision maker was able to learn about her own ranking of the alternatives in $B$ before making a choice. The decision maker updates the prior according to Bayes’ rule and chooses the option $i \in B$ with the highest posterior mean belief $\mathbb{E}[\mu_i | (X_k)_{k \in B}]$.

The signals $(X_i)_{i \in B}$ are joint normally distributed for every menu $B \in \mathcal{A}$. The expectation of each signal $X_i$ is the true utility $\mu_i$. All signals have the same variance $1/t$. The correlation of signals $X_i$ and $X_j$ is given by $\sigma(i, j) \in [0, 1]$. We write $\sigma_{ij}$ instead of $\sigma(i, j)$. The function $\sigma : A \times A \to [0, 1]$ satisfies $\sigma_{ij} = \sigma_{ji}$ and $\sigma_{ii} = 1$ for every $i, j \in A$. In addition, the determinant of the matrix $(\sigma_{ij})_{i,j=1,\ldots,k}$ is strictly positive for every finite subset of options $\{1, \ldots, k\}$. We call the resulting random

\(^1\)While the normality assumption is at best an approximation, the symmetry assumption is natural in many settings. For example, in the experimental results we discuss in Section 6, the options are lotteries over money presented on a computer screen. The position of the lotteries on the screen is randomized on each choice trial. Hence there is no a priori reason to believe that the lottery on the left side of the screen came from a different distribution over utility than the lottery on the right side of the screen.
choice rule a *Bayesian probit* rule with utility parameters $\mu$, correlation parameters $\sigma$, and precision parameter $t$, and we denote it by $\rho_t^{\mu\sigma}$. In summary, for each $i \in B \subseteq A$,

$$\rho_t^{\mu\sigma}(i, B) := \mathbb{P}\{\mathbb{E}[\mu_i - \mu_j | (X_k)_{k \in B}] \geq 0 \ \forall j \in B\}.$$

A single parameter $t > 0$ captures the precision of the signals in the Bayesian probit rule. As precision $t$ increases, the choice rule $\rho_t^{\mu\sigma}$ more closely reflects the true utility of the alternatives given by $\mu$. In applications, the precision parameter $t$ facilitates the aggregation of choice data across subjects when there is heterogeneity in information precision or familiarity with the options. For example, the parameter $t$ can be used to describe choices conditional on a vector of observable subject characteristics, such as age, experience, ability, time available to contemplate the options before making a choice etc. This is particularly useful in experimental settings where the amount of information or the time available to contemplate the choice alternatives is manipulated (such as in Caplin, Dean and Martin (2011) or in Soltani, De Martino and Camerer (2012)). We refer to a family $(\rho_t^{\mu\sigma})_{t>0}$ of Bayesian probit rules indexed by precision $t$ as a *Bayesian probit process*.

### 3.1 Monotonicity and departure from random utility

A random choice rule $\rho$ is *monotone* if $\rho(i, B) \geq \rho(i, C)$ whenever $i \in B \subset C$. For monotone random choice rules, the probability of choosing an alternative $i$ can only decrease when new alternatives are added to a menu. The attraction effect is perhaps the most famous example of a violation of monotonicity: the introduction of a decoy alternative increases the probability that a similar existing option is chosen. This presents a challenge to every monotone model. Virtually all models used in the discrete choice estimation literature are monotone, as the following example shows.

**Example 1** (Random Utility Maximizers). For the sake of simple exposition, suppose the universe of choice alternatives $A$ is finite. Let $\mathcal{U}$ be a collection of strict utility functions on $A$. A random utility $\pi$ is a probability distribution on $\mathcal{U}$. A random choice rule maximizes a random utility $\pi$ if the probability of choosing alternative $i$ from menu $B$ equals the probability that $\pi$ assigns to utility functions that attain their maximum in $B$ at $i$. Most econometric models of discrete choice are random utility maximizers. Probit rules specify a random utility vector with a joint Gaussian (normal) distribution. Logit rules have identically and independently Gumbel distributed utilities. Generalizations of logit (such as nested logit) introduce correlations in utility. It is easy to verify that every random utility maximizer is monotone. ☐
The Bayesian probit allows violations of monotonicity and hence is not a random utility model. In the Bayesian probit the decision maker chooses the alternative $i$ in menu $B$ with the highest realization of $E[\mu_i|(X_k)_{k \in B}]$, the conditional expectation of $\mu_i$ given the realization of all the signals in the menu. When signals are correlated, this alternative may be different from the alternative that maximizes $X_j$.

### 3.2 Continuous learning process

The Bayesian probit can be equivalently described as arising from a gradual learning of utilities over time. We can model the signals obtained by the decision maker when facing the menu $B = \{1, \ldots, n\}$ as a continuous random process $X : \Omega \times [0, \infty) \to \mathbb{R}^n$, where $\Omega$ is the underlying probability space. The $n$-dimensional vector of signals starts from $X(0) = 0$ almost surely and follows a Brownian motion with drift given by

$$dX(t) = \mu dt + \Lambda dW(t)$$

where the constant drift vector $\mu = (\mu_1, \ldots, \mu_n)$ is given by the true utility of the alternatives, $\Lambda$ is a constant $n \times n$ matrix with full rank (\(\Lambda\) is the Cholesky decomposition of the correlation matrix with entries given by $\sigma_{ij}$) and $W(t) = (W_1(t), \ldots, W_n(t))$ is a Wiener process.

It is easy to verify that $X(t)/t$ has the same distribution and contains the same information about the utility of the alternatives as the signals in the static formulation. Hence the static formulation in which the decision maker observes a signal realization with precision $t$ and the continuous formulation in which the decision maker observes a signal process up to time $t$ are equivalent. Both lead to the same distribution for the mean posterior beliefs, and hence both generate the same random choice rule $\rho_{t,\sigma}^\mu$.

The continuous learning process formulation helps us interpret the assumption that the decision maker knows the value of the correlation parameters $\sigma_{ij}$. Since the continuous random process accumulates quadratic variation, even if the decision maker didn’t know the value of the correlation parameters $\sigma_{ij}$ at the start of the process, she could perfectly estimate them after an arbitrarily small time interval (see any standard textbook on Brownian motion, for example, Shreve (2004)).

---

2Drift diffusion models are often used in neuroscience to represent the noisy process by which the brain perceives the value of choice alternatives over time. See for example the drift diffusion model of Ratcliff (1978) and the decision field theory of Busemeyer and Townsend (1993). Fehr and Rangel (2011) provide an overview of this literature. See also Woodford (2014).
4 Revealed Preference and Revealed Similarity

Throughout this section $\rho^{\mu\sigma}_t$ is a Bayesian probit rule with utility $\mu$, correlation $\sigma$ and precision $t$. An advantage of the Bayesian probit model is that each parameter has a clear behavioral interpretation. Understanding the behavioral content of each parameter facilitates measurement and provides a useful tool for applications.

We begin with revealed preference. To simplify notation, we write $\rho(i, j) := \rho(i, \{i, j\})$ for choice probabilities in binary menus. Given a random choice rule $\rho$ we say that alternative $i$ is revealed preferred to alternative $j$ and write $i \succeq j$ when

$$\rho(i, j) \geq \rho(j, i).$$

As usual we write $\sim$ for the symmetric part and $\succ$ for the asymmetric part of the revealed preference relation. When $i \sim j$ we also say that $i$ and $j$ are on the same indifference curve.

**Proposition 1.** The revealed preference relation for $\rho^{\mu\sigma}_t$ is represented by $\mu$.

Proposition 1 shows that, for a Bayesian probit rule with parameters $\mu$, $\sigma$ and $t$, an alternative $i$ is revealed preferred to alternative $j$ if and only if $\mu_i \geq \mu_j$. Hence, the revealed preference relation $\succeq$ obtained from a Bayesian probit rule $\rho^{\mu\sigma}_t$ is always complete and transitive. It also follows that the revealed preference relation obtained from a Bayesian probit process $(\rho^{\mu\sigma}_t)_{t > 0}$ is well-defined and independent of $t$.

Next, we define revealed similarity as a binary relation $\succeq$ over the set of pairs of choice alternatives. Given a random choice rule $\rho$ we say the pair $\{i, j\}$ is revealed more similar than the pair $\{k, \ell\}$ and write $\{i, j\} \succeq \{k, \ell\}$, if

$$k \sim i \succ j \sim \ell$$

and

$$\rho(i, j)\rho(j, i) \leq \rho(k, \ell)\rho(\ell, k).$$

We write $\sim$ for the symmetric part and $\succ$ for the asymmetric part of the relation $\succeq$. When $i = k$ above we say that $i$ is revealed more similar to $j$ than to $\ell$. Likewise, when $j = \ell$ we say that $j$ is revealed more similar to $i$ than to $k$.

Condition (2) in the definition of similarity says that $i$ and $k$ are on the same indifference curve, and, likewise, $j$ and $\ell$ are on the same indifference curve. Insofar as preference is concerned, the pair $\{i, j\}$ is identical to the pair $\{k, \ell\}$. The inequality
in (3) means that $\rho(k, \ell)$ is less extreme, that is, closer to one-half than $\rho(i, j)$. In other words, the decision maker discriminates the pair $\{i, j\}$ at least as well as the pair $\{k, \ell\}$. A strict inequality in (3) indicates that the decision maker is less likely to make a mistake when choosing from $\{i, j\}$. Since the gap in desirability is the same in both pairs, the ability to better discriminate the options in $\{i, j\}$ reveals that $i$ and $j$ are more similar than $k$ and $\ell$.

For example, suppose that for a given decision maker we observe $\rho(i, j) = 1/2$, $\rho(i, k) = 6/7$ and $\rho(j, k) = 5/7$. According to our definition of revealed preference, objects $i$ and $j$ are on the same indifference curve. Both are strictly better than $k$. Hence when choosing from $\{i, k\}$ and when choosing from $\{j, k\}$, we interpret a choice of option $k$ as a mistake. Mistakes happen less often in $\{i, k\}$ than in $\{j, k\}$, thus the options in $\{i, k\}$ must be easier to compare. And since the difference in utility is the same in both pairs, we infer that options in $\{i, k\}$ are easier to compare because $k$ is more similar to $i$ than to $j$.

Does similarity always facilitate discrimination? Clearly the answer is no. It is easy to find examples where two options that are less similar are also easier to discriminate. But the key to our revealed similarity definition is that we are keeping utility constant. In that case, similarity does facilitate discrimination. This fact, which we call the discrimination hypothesis, is known in the psychology literature at least since Tversky and Russo (1969):

> The similarity between stimuli has long been considered a determinant of the degree of comparability between them. In fact, it has been hypothesized that for a fixed difference between the psychological scale values, the more similar the stimuli, the easier the comparison or the discrimination between them.\(^3\)

Adapted to a context of preferential choice, the discrimination hypothesis states that, when options 1 and 2 are equally desirable and option 3 more closely resembles 2 than 1, the decision maker is more likely to correctly discriminate the options in a pairwise choice between 2 and 3 than between 1 and 3. In other words, for a fixed gap in the desirability of the alternatives, increased similarity facilitates discrimination. In the following examples, we illustrate the wide applicability of this hypothesis and connect it to our definition of revealed similarity.

---

\(^3\)While Tversky and Russo call this hypothesis the ‘similarity hypothesis’, we should refer to it as the discrimination hypothesis to avoid confusion with the better known similarity hypothesis in Tversky (1972).
**Example 2** (Triangle Area). Suppose revealed preference analysis has determined that a decision maker’s tastes over triangles in the Euclidean plane are very simple: she always prefers triangles that have larger areas. In other words, her preferences over triangles can be represented by the utility function that assigns to each triangle the numerical value of its area. Among the options shown in Figure 1, which triangle should the decision maker choose?

![Figure 1: Which triangle has the largest area?](image)

Figure 1 illustrates that it is hard to discriminate among comparable options. Since the triangles in Figure 1 have comparable areas, they are close according to the decision maker’s preference ranking, and she may have a hard time picking the best (largest) one. If she were forced to make a choice with little time to examine the options, she would have a good chance of making a mistake.

Now suppose the same decision maker is faced with a choice from the pair shown in Figure 2. The triangle shown on the left in Figure 2 has exactly the same area as the triangle on the left in Figure 1, while the triangle on the right is the same in both Figures. Hence, from the point of view of desirability, the pair shown in Figure 1 is identical to the pair shown in Figure 2. However, she will be less likely to make a mistake when choosing among the options in Figure 2.

When presented with the choice in Figure 2, most subjects choose the triangle on the left; this choice task is in fact much easier than the task in Figure 1. This certainly has nothing to do with the gap in the desirability of the two options, since the pairs in both Figures exhibit exactly the same difference in area. In Figure 2 we increased the overlap in features between the triangles, while keeping them on the same indifference curves as in Figure 1. The increased overlap of features clearly helps improve the decision maker’s ability to discriminate among the options. This is captured by our definition of revealed similarity: based on this pattern of choices, the pair \( \{i', j\} \) of Figure 2 is revealed more similar than the pair \( \{i, j\} \) of Figure 1.

\[\diamond\]
In Euclidean geometry, two triangles are defined to be similar when they have the same internal angles (the pair in Figure 2 is such an example). This geometric definition of similarity is independent from size, and only depends on the shape of the objects. Likewise, our definition of revealed similarity departs from the everyday use of the word ‘similarity’ and is independent from utility. The next example involves choice objects commonly used in experimental economics.

**Example 3** (Lotteries over money). In the experimental study of Soltani, De Martino and Camerer (2012) each choice object is a simple lottery \( (p, m) \) described by a probability \( 0 < p < 1 \) of winning and a monetary prize \( m \geq 0 \). Suppose revealed preference analysis for one of the subjects determines that lotteries \( B, C, D, E \) are on the same indifference curve while \( A \) is superior, as depicted in Figure 3. The discrimination hypothesis in this case states

\[
\rho(A, E) > \rho(A, D) > \rho(A, C) > \rho(A, B) > 1/2.
\]

The difference in utility is the same in every pairwise choice, but mistakes are more likely when the options are less similar. In Figure 3, options \( A \) and \( B \) are the most difficult to compare, while options \( A \) and \( E \) are the easiest. Note that option \( A \) strictly dominates option \( E \), offering a larger prize and a larger probability of winning. The same difference in utility becomes more transparent when options are more similar. Hence strict dominance can be interpreted as an extreme form of similarity.

In many applications, a vector of observable characteristics of each choice object is available. For instance, the options in Example 3 can be objectively described as a point in \( \mathbb{R}^2 \). In such cases, the econometrician can formulate and test additional parametric assumptions about how utility and similarity translate to the location of objects in the space of observable characteristics. We use choice over lotteries to illustrate such an application of the Bayesian probit model in Section 6.
Figure 3: The two indifference curves represent the decision maker’s true preferences of over simple money lotteries. Lotteries $B, C, D, E$ lie on the same indifference curve, while lottery $A$ is superior. In pairwise choice tasks, mistakes are more likely when the options are less similar. The comparison of $A$ versus $B$ is the hardest, while the comparison of $A$ and $E$ is the easiest.

The next abstract example clearly illustrates how similarity can make a decision maker have a very precise belief about the relative ranking of two alternatives, even if she has a very imprecise belief about the true value of each individual alternative.\footnote{I am grateful to David K. Levine for suggesting this example.}

**Example 4** (Points on a star). Suppose the universe of choice alternatives is the set of all star-shaped figures. Revealed preference analysis has determined that the decision maker only cares about the number of points on a star. For example, when the options are a six-pointed star and a five-pointed star, the optimal choice is the six-pointed star. With just a glance at the options in Figure 4, which star should she pick?

Again, if pressed to make a choice in a short amount of time, she may be very likely to make a mistake. The probability of making a mistake would certainly decrease if she were given more time. She would also be more likely to make the correct choice if one of the alternatives was much better (had a much larger number of points) than the other. Given a limited amount of time, mistakes are likely, because the options are comparable and therefore difficult to discriminate.

Now suppose the same decision maker faces a choice between the options in Figure 5. Which of the two options should she choose? Here, similarity comes to the rescue. The
A quick comparison of Figures 4 and 5 illustrates the following distinction in the perception of utility and similarity: while utility is learned gradually as the level of familiarity with the options increases, the degree of similarity of the options is perceived instantly. The same point is illustrated by comparing Figures 1 and 2. This is captured in the model by the assumption that, while the decision maker gradually learns utility $\mu$ as $t$ increases, similarity $\sigma$ is either known or perfectly estimated by the decision maker.
from the beginning of the choice process (see the discussion at the end of Section 3).

As a final example, we can link our definition of revealed similarity to the computer choice example of the Introduction. Suppose that while computers $A$ and $B$ are very different in terms of features, they lie on the same indifference curve. In other words, computers $A$ and $B$ are equally desirable, but very ‘distant’ from each other in the space of relevant attributes. Computer $C$ shares many features with computer $B$, but is clearly worse in minor ways. In other words, computers $B$ and $C$ are very close to each other in the space of relevant attributes, but $B$ clearly dominates $C$.

Since $C$ is inferior, choosing computer $C$ is a mistake whenever $A$ or $B$ are available. In line with the discrimination hypothesis, we expect a decision maker to make a mistake when choosing between computers $A$ and $C$ more often than when facing a choice between $B$ and $C$. If this is the case in the choice data, then according to our definition the pair $\{B,C\}$ is revealed more similar than the pair $\{A,C\}$. Our definition declares that computer $C$ is more similar to computer $B$ than to computer $A$ based on observed choice data, and not based on any of their observable features.

**Proposition 2.** Let $\succsim$ be the revealed similarity relation for a Bayesian probit $\rho_t^{\mu\sigma}$. Then $\{i,j\} \succsim \{k,\ell\}$ if and only if $\mu_k = \mu_i > \mu_j = \mu_\ell$ and $\sigma_{ij} \geq \sigma_{k\ell}$.

Options are revealed more similar when the correlation of their noisy signals is higher. The intuition behind this result is that, for a fixed difference in utility $\mu_i - \mu_j$, a higher correlation parameter leads to more precise information about the true ranking of $i$ and $j$. We can write the difference in noisy signals as $X_i - X_j = (\mu_i - \mu_j) + (\varepsilon_i - \varepsilon_j)$. As the correlation in the error terms $\varepsilon_i$ and $\varepsilon_j$ increases, the variance of $(\varepsilon_i - \varepsilon_j)$ decreases, and signals become more informative about $(\mu_i - \mu_j)$. In the limit as signals become perfectly correlated, the noise terms cancel out.

Since our definition of revealed similarity requires fixing the indifference curves in condition (2), for some pairs of alternatives we may have neither $\{i,j\} \succsim \{k,\ell\}$ nor $\{k,\ell\} \succsim \{i,j\}$. To obtain a complete revealed similarity ranking from binary choices, it is necessary to introduce additional assumptions about the richness of the set of choice alternatives. We show in Section 7 that choices over menus of three alternatives are sufficient to reveal the complete similarity ranking in any triple. With either approach (using binary choice data under further richness assumptions; or using ternary choice data) the revealed similarity relation is represented by the correlation $\sigma$. Henceforward, we refer to the parameter $\sigma_{ij}$ as the similarity of pair $i,j$.  

15
5 Similarity Puzzle

This section contains the main results. Throughout the section \(\succeq\) denotes the revealed preference relation and \(\approx\) denotes the revealed similarity relation corresponding to the Bayesian Probit process \((\rho_t^\mu, \rho_t^\sigma)_{t>0}\).

**Proposition 3** (Small \(t\)). If \(\sigma_{23} > \sigma_{13}\) then for all \(t\) sufficiently close to zero,

\[
\frac{\rho_t^\mu(1, \{1, 2, 3\}) - \rho_t^\mu(2, \{1, 2, 3\})}{\rho_t^\mu(1, \{1, 2\}) - \rho_t^\mu(2, \{1, 2\})} < 1.
\]

Proposition 3 shows the effect of similarity for small values of information precision \(t\), that is, at the beginning of the choice process. When the new alternative 3 is more similar to 2 than to 1, the introduction of 3 hurts 1 proportionally more than 2. Hence with our notion of similarity, represented by the correlation parameter \(\sigma\), the exact opposite of Tversky’s similarity hypothesis holds when the decision maker is relatively uninformed. This effect holds regardless of the utility values.

It is easy to explain the intuition behind Proposition 3 (the proof can be found in the Appendix). Recall that the signal received by the decision maker about the utility of alternative \(i\) is given by \(X_i = \mu_i + \epsilon_i\) where \(\epsilon_i\) is normally distributed with mean zero and variance \(1/t\). Hence early in the learning process, when \(t\) is small, the variance \((1/t)\) is large and the information obtained about the numerical value of the utility of any given alternative is very imprecise.

But even when information precision \(t\) is small, she may be relatively well informed about the *difference* in value for a pair of alternatives. The difference in the signals for alternatives \(i\) and \(j\) is

\[X_i - X_j = \mu_i - \mu_j + (\epsilon_i - \epsilon_j)\]

where the constant \(\mu_i - \mu_j\) is the true difference in utility and the error term \(\epsilon_i - \epsilon_j\) has a normal distribution with variance \(2(1 - \sigma_{ij})/t\), which is decreasing in the correlation parameter \(\sigma_{ij}\) and vanishes when \(\sigma_{ij}\) approaches one.

The combination of low precision of information and varying degrees of similarity favors a choice among the more similar alternatives. To see this, note that, according to the prior, every strict ranking of the alternatives has the same probability:

\[
1 \succ 2 \succ 3 \quad 1 \succ 3 \succ 2 \\
2 \succ 1 \succ 3 \quad 3 \succ 1 \succ 2 \\
2 \succ 3 \succ 1 \quad 3 \succ 2 \succ 1
\]
when $\sigma_{23}$ is the highest among the correlation parameters, and $t$ is sufficiently low, the
decision maker will only be relatively certain about the difference in utility $\mu_2 - \mu_3$ and
hence about whether the true state of nature is in the left column (rankings in which
$2 \succ 3$) or in the right column (rankings in which $3 \succ 2$). In either case, alternative 1
is the best alternative in only one out of three possible rankings. The overall effect is
that alternative 1 is hurt more than alternative 2 when alternative 3 is introduced.

**Proposition 4 (Large $t$).** If $1 \sim 2 \sim 3$ and $\sigma_{23} > \sigma_{13}$, then for all $t$ sufficiently large,

$$\frac{\rho_t^{\mu_{13}}(1, \{1, 2, 3\})}{\rho_t^{\mu_{13}}(2, \{1, 2, 3\})} \geq \frac{\rho_t^{\mu_{23}}(2, \{1, 2\})}{\rho_t^{\mu_{23}}(2, \{1, 2\})}.$$  

Proposition 4 shows the effect of similarity for large values of information precision $t$. Large values of $t$ correspond to the behavior of a decision maker who has learned a great deal of information about the utility of the alternatives in the menu before making a choice. With high information precision, the decision maker will in general make very few mistakes. Irrespective of similarity values, the best option will be chosen with probability going to one as $t$ goes to infinity. For a given large value of $t$, similarity will have a negligible effect unless utility values are very close. Proposition 4 shows the effect of similarity when it is most relevant, that is, when it breaks the ties in the case of three equally desirable alternatives. In this case, introducing the new alternative 3 hurts the similar alternative 2 more than 1. Hence our notion of similarity (represented by the correlation parameter $\sigma$) breaks the ties in accordance to Tversky’s similarity hypothesis when the decision maker is sufficiently informed.

The intuition behind Proposition 4 can be easily explained. When information about the alternatives becomes arbitrarily precise, the weight of the prior vanishes and the ranking of the signals is very likely to reflect the true ranking of the alternatives. Hence as $t$ becomes arbitrarily large, the probability that the decision maker chooses the alternative with the highest signal realization goes to one. And in this case being in a pair with high correlation is bad for both alternatives: they tend to have high signals in the same states of nature, getting in each other’s way.

Since the mapping from parameters to choice probabilities is continuous, Proposition 4 is robust to a small perturbation of the parameters. For example, let $T > 0$ be sufficiently large, such that the conclusion of Proposition 4 holds for all $t > T$. Fixing any finite time $T' > T$ and truncating the random choice process up to time $T'$, there exists an $\varepsilon > 0$ such that the result holds for all combinations of parameters in an $\varepsilon$-neighborhood of $\mu$ and $\sigma$. In particular, the result holds for the choice process
restricted to $t \leq T'$ even if the alternatives are not equally desirable, as long as they are sufficiently close in utility.

We conclude from Propositions 3 and 4 that similarity interacts with information precision to determine the effect of introducing a new alternative to the choice menu. For low values of information precision, introducing a new alternative hurts dissimilar alternatives proportionally more than similar ones. With high values of information precision, similarity will have negligible effects, except when alternatives are close in utility. In this case, introducing a new alternative hurts similar alternatives more than dissimilar ones. To illustrate these two effects of similarity in their most extreme form, the next Proposition analyzes the case in which alternatives 1, 2 and 3 are equally desirable and the similarity parameter $\sigma_{23}$ for alternatives 2 and 3 is taken arbitrarily close to one.

**Proposition 5** (Duplicates). Let $1 \sim 2 \sim 3$. For every $\varepsilon > 0$ there exist $\delta, T, T' > 0$ such that if $\sigma_{23} > 1 - \delta$ then

(i) $\rho^\mu_t(2, \{1, 2, 3\}), \rho^\mu_t(3, \{1, 2, 3\}) \in (1/2 - \varepsilon, 1/2]$ for all $t < T$; and

(ii) $\rho^\mu_t(2, \{1, 2, 3\}), \rho^\mu_t(3, \{1, 2, 3\}) \in [1/4, 1/4 + \varepsilon)$ for all $t > T'$.

Item (i) shows how the Bayesian probit can accommodate the extreme example given in Section 2. It is a powerful illustration of how under low information precision the decision maker tends to choose among more similar alternatives. When $t$ is small and $\sigma_{23}$ is high, the decision maker learns whether $2 \succ 3$ or $3 \succ 2$ with a high level of certainty, but nothing else. In the absence of any other information, the winner among $\{2, 3\}$ is the most likely winner among $\{1, 2, 3\}$. Hence the probability of choosing the dissimilar alternative 1 becomes arbitrarily small. Note that item (i) is still true if we relax the assumption $1 \sim 2 \sim 3$; it holds independently of the true utility values.

Item (ii) says that, in the limit as $t$ goes to infinity, the alternatives in the very similar pair $\{2, 3\}$ are treated as a single option. In other words, they are treated as duplicates, in line with the behavior in the red-bus, blue-bus example of Debreu (1960). As the level of information precision increases, any small difference in the utility of the alternatives will eventually be learned by the decision maker, and the best alternative will be chosen with probability close to one. When the three alternatives have the exact same utility, it is similarity, rather than utility, that determines how ties are broken. Since the signals for alternatives 2 and 3 are highly correlated, it is very unlikely that the signal for alternative 1 lies between them. Approximately half of the time it will be above and half of the time it will be below. Hence alternative 1 is
chosen with probability close to one-half, and an alternative in \{2, 3\} will be chosen with probability close to one-half. Since the signals for alternatives 2 and 3 are highly but never perfectly correlated, there is enough orthogonal noise to break the ties and each alternative is chosen with approximately \(1/4\) probability.

The next result shows how similarity can generate the *compromise effect*: the tendency of a decision maker to avoid extreme options (Simonson (1989)). In the Bayesian probit there are three correlation parameters for every set of three alternatives \(i, j, k\), namely \(\sigma_{ij}, \sigma_{ik}\) and \(\sigma_{jk}\). When \(\sigma_{ij}\) is the smallest of the three correlation parameters, options \(i\) and \(j\) are each more similar to \(k\) than to each other. In this case we will say that alternative \(k\) is the *middle option*.

**Proposition 6** (Compromise effect). The middle option is chosen most often for every \(t\) sufficiently small.

Note that the result holds for any given utility values \(\mu_1, \mu_2\) and \(\mu_3\). In other words, the compromise effect arises for low levels of information precision, *independently of the true value of the alternatives*. The intuition behind the compromise effect is as follows. The middle alternative \(k\) is more similar to each of the two extreme alternatives, so the differences in signals \(X_i - X_k\) and \(X_j - X_k\) are more informative than the difference \(X_i - X_j\). Roughly, the decision maker will be two pieces of news about the middle alternative \(k\), but only one piece of news for each of \(i\) and \(j\). This makes the posterior beliefs about the middle alternative more volatile than the posterior beliefs of the other alternatives. Under low information precision, higher volatility of posterior beliefs makes the middle alternative \(k\) more likely to be chosen. To see this, it is be helpful to think of an extreme case, in which the variance of the posterior beliefs for the middle alternative could be made arbitrarily high: the middle alternative would have the highest posterior beliefs approximately half of the time, and the lowest posterior beliefs also half of the time.

The next result addresses the attraction effect. Recall that when alternatives 1 and 2 are equally desirable, each of them is chosen from the binary menu \{1, 2\} with probability one-half for any level of information precision \(t\). A violation of monotonicity obtains if one of these probabilities increases above one-half once alternative 3 is introduced. The Proposition below shows that the propensity of the Bayesian decision maker to choose among more similar alternatives can persist indefinitely when alternatives 2 and 3 are similar but not equally desirable.

---

5We relate the similarity parameter to the position of the objects in a space of observable attributes in Section 6.
Proposition 7 (Attraction effect). Let $1 \sim 2$, $\{2,3\} \succ \{1,3\}$ and $\varepsilon > 0$. If alternative 3 is sufficiently inferior, then alternative 2 is chosen from menu $\{1,2,3\}$ with probability strictly larger than one-half for all $t > \varepsilon$.

Hence adding a sufficiently inferior alternative 3 to the menu $\{1,2\}$ will boost the probability of the similar alternative 2 being chosen — the phenomenon known as the attraction effect. Proposition 7 shows conditions under which the attraction effect occurs arbitrarily early in the choice process (for arbitrarily low levels of information precision), and lasts indefinitely (even for arbitrarily high levels of information precision).

For a concrete illustration, consider the example in Figure 6. It plots choice probabilities as a function of information precision for the three alternatives once the similar but inferior alternative 3 is introduced. Prior beliefs are standard normally distributed. Utilities are given by $\mu_1 = \mu_2 = 3$, and $\mu_3 = -3$, so that 1 and 2 are equally desirable but 3 is inferior. Correlations are given by $\sigma_{12} = \sigma_{13} = 0$ and $\sigma_{23} = .5$, so that alternative 3 is more similar to alternative 2 than to alternative 1.

Figure 6 shows information precision in logarithmic scale, to better visualize the start of the choice process. The top curve (solid line) is the probability that alternative 1 is chosen; the middle curve (dotted line) corresponds to the choice probability for alternative 2; and the bottom curve (dashed line) corresponds to the choice probability of alternative 3.

At the beginning of the choice process (for $t$ close to zero), when choices are based on relatively little information about the alternatives, the higher similarity of pair $\{2,3\}$ makes these alternatives more likely to be chosen than alternative 1. As we saw in the discussion of Proposition 3, the correlation in the signals of alternatives 2 and 3 allows the decision maker to learn the relative ranking of alternatives 2 and 3 with much higher precision than the others, much earlier in the choice process. Since alternative 3 is inferior, the probability of the event $2 \succ 3$ according to her posterior beliefs will rapidly increase with $t$ (much faster than $1 \succ 3$), which boosts the choice probability of alternative 2 in detriment of alternative 3. This leads the probability of alternative 2 to raise above one-half, violating monotonicity, and the attraction effect obtains. Proposition 7 guarantees that this violation of monotonicity will happen arbitrarily early in the choice process, if we take alternative 3 sufficiently inferior.

Figure 6 also illustrates how the attraction effect persists as the decision maker becomes more familiar with the options. Since alternatives 1 and 2 are equally desirable, eventually each one is chosen with probability arbitrarily close to one-half. But
the probability of choosing alternative 1 never recovers from the initial disadvantage caused by the similarity of alternatives 2 and 3: while the choice probability of alternative 2 tends to one-half from above, the choice probability of alternative 1 tends to one-half from below. In other words, even as information precision $t$ goes to infinity, the attraction effect never disappears.

Note that since the mapping from parameters to choice probabilities in the Bayesian probit is continuous, Proposition 7 is robust to small perturbations of the parameters. Hence, for a small neighborhood around the initial parameter values, Figure 6 stays qualitatively the same. In particular, the result will hold even if alternatives 1 and 2 are not exactly equally desirable but just close in utility. Hence, we may obtain the attraction effect arbitrarily early in the choice process, and it may last for an arbitrarily long time, even if alternative 2 is slightly worse than alternative 1.

Finally, note that in statement of the main results we didn’t explicitly invoke the location of the choice objects in any given space of attributes. In other words, our model does not take a given vector of observable characteristics as a primitive. This makes the model independent of the particular set of observable attributes at hand, and gives the Bayesian probit the same portability of the standard multinomial probit and other discrete choice estimation models, as we show in the next section.
6 Observable Attributes

In experiments and applications, the analyst often has access to a vector of observable attributes \((x_{i1}, x_{i2}, \ldots, x_{im}) \in \mathbb{R}_+^m\) for each choice alternative \(i \in A\). For example, in Simonson (1989) cars are described by ‘ride quality’ and ‘miles per galon’. In Soltani, De Martino and Camerer (2012), experimental subjects choose among simple lotteries \((p, m)\) described by the value \(m\) of a cash prize and the probability \(p\) of winning the prize. In such settings, the Bayesian probit model allows the analyst to formulate and test additional parametric assumptions that specify how utility and similarity depend explicitly on the location of the alternatives in the attribute space. Hence the Bayesian probit has the same “portability” to different discrete choice settings as the standard probit, logit, etc.

Unlike random utility models, the Bayesian probit is able to capture context effects. And unlike many of the formal models proposed in the bounded rationality literature, the Bayesian probit can accommodate these effects without abandoning the assumption that decision makers are described by a single, rational preference relation. To illustrate these advantages, consider the experiment in Soltani, De Martino and Camerer (2012). The first part of the experiment determined, for each subject, the value of a cash prize \$M\) such that a low risk lottery \(\ell = (70\%, \$20)\) and a high risk lottery \(h = (30\%, \$M)\) are equally likely to be chosen from the menu \(\{h, \ell\}\). In other words, the high risk lottery prize was adjusted separately for each subject so that \(\ell\) and \(h\) lie on the same indifference curve.

The second part of the experiment measured the effect of a third decoy alternative on the likelihood of choosing \(h\) over \(\ell\). In each trial, subjects were given eight seconds to examine three lotteries on the computer screen; after the eight seconds one of the three lotteries was randomly removed; subjects then had two seconds to choose one of the two remaining lotteries by pressing a key.\(^6\) This allowed the experimenters to assess the effect of inspecting the decoy alternative over the likelihood of choosing \(\ell\) over \(h\). They found that decoys placed close to \(h\) and clearly dominated by \(h\) (region \(D1\) in Figure 7) significantly increased the likelihood of choosing \(h\). On the other hand, decoys placed close to \(h\) but dominating \(h\) (region \(D2\) in Figure 7) significantly decreased the likelihood of choosing \(h\) over \(\ell\). Analogous effects were found for decoys placed near lottery \(\ell\).

\(^6\)Prizes and probabilities were slightly perturbed and the order of the alternatives on the screen was randomized across trials. Subjects were only allowed to choose during the two-second period and were penalized if they failed to make a choice. See Soltani, De Martino and Camerer (2012) for a full description of the experimental design.
Figure 7: Experimental study from Soltani et al. (2012). In the absence of decoys, lotteries $h$ and $\ell$ are chosen with the same probability. Decoy lotteries in the $D2$ region decrease the likelihood of choosing $h$ over $\ell$, while decoys in $D1$ increase the likelihood of choosing $h$ over $\ell$.

Random utility models are, by design, incapable of capturing the effects of the decoy alternative. In any random utility model, the probability of choosing $h$ over $\ell$ is equal to the probability that the utility realization of alternative $h$ is higher than the utility realization of alternative $\ell$. This probability does not change in the presence of other alternatives. Hence in a random utility model the probability of choosing $h$ over $\ell$ is the same whether the decoy alternative is shown or not.

The Bayesian probit, on the other hand, can capture these decoy effects. The probability of choosing $h$ over $\ell$ equals the probability that the mean posterior belief about $h$ is higher than the mean posterior belief about $\ell$. The Bayesian decision maker uses the information obtained about every alternative to determine the beliefs about each alternative. Since, in general,

$$P \{ E[\mu_h|X_h, X_\ell, X_d] > E[\mu_\ell|X_h, X_\ell, X_d] \} \neq P \{ E[\mu_h|X_h, X_\ell] > E[\mu_\ell|X_h, X_\ell] \},$$

the presence of the decoy alternative $d$ adds relevant information to the Bayesian decision maker. The extra signal $X_d$ changes the distribution of posterior beliefs for alternatives $h$ and $\ell$. Hence the Bayesian probit is able to capture the effect of the decoy alternative over the likelihood of choosing $h$ over $\ell$.

To fit the Bayesian probit to the data, the econometrician formulates and tests additional parametric assumptions that relate the vector of observable characteristics of
subjects and choice objects to the utility, similarity, and information precision parameters. For example, when each alternative is a simple lottery \((p, m)\) the econometrician can formulate and test the assumption of expected utility preference over lotteries \(\mu(p, m) = p \times u(m, \beta)\) where \(u(\cdot, \beta)\) is a parametric Bernoulli utility function such as CARA or CRRA and \(\beta\) is a risk aversion parameter to be estimated.

The econometrician may specify, for example, that the similarity of two alternatives is given by cosine similarity: the correlation between the signals for alternatives \(i\) and \(j\) is equal to the cosine of the angle between \(i\) and \(j\). This additional parametric assumption is proposed for estimation of the standard probit model by Hausman and Wise (1978) and also has applications in computer science.\(^7\) Cosine similarity satisfies the discrimination hypothesis of Section 4 and can be varied independently from utility. In Figure 7 decoys placed in areas \(D1\) or \(D2\) will form a smaller angle (and therefore have larger cosine similarity) with \(h\) than with \(\ell\).

For any given level of information precision, a higher degree of similarity allows the decision maker to more easily compare the decoy and \(h\) than the other pairs. As in the example of Section 2, being able to more easily rank the decoy and \(h\) makes it more likely that posterior beliefs rank lottery \(\ell\) between the decoy and \(h\). When the decoy is inferior, this favors \(h\); on the other hand, when the decoy is superior, lottery \(\ell\) is more likely to be chosen over \(h\).

The evidence for the attraction and compromise effects is often presented in settings where objects have explicitly given attributes. The assumption of cosine similarity helps us connect the explanation of how the Bayesian probit addresses these effects to the multi-attribute setting. Consider, for example, the menu of options \(\{j, k\}\) represented in Figure 8. Adding a more extreme option \(i\) to the menu turns \(j\) into the middle option: since the angle formed by \(i\) and \(k\) is the largest among \(\{i, j, k\}\), the cosine of the angle formed by \(i\) and \(k\) is the smallest, and hence \(\sigma_{ik}\) will be smaller than \(\sigma_{ij}\) and \(\sigma_{jk}\). Proposition 6 then guarantees that \(j\) will be chosen most often from \(\{i, j, k\}\) when information precision is sufficiently low.

An important feature of the evidence for the compromise effect is that the effect happens in both directions. This is easily accommodated by the Bayesian probit model. For example, suppose that, instead of adding \(i\), we add option \(\ell\) to the menu \(\{j, k\}\) Figure 8. In this case it is \(k\) that becomes the middle option among \(\{j, k, \ell\}\), and Proposition 6 guarantees that option \(k\) is chosen most often when information precision

---

\(^7\)For example, using the frequency of words as the attributes of a text document, the cosine of the angle measures the similarity of two documents while normalizing document length during comparison.
is sufficiently low.

Now consider the attraction effect. Suppose we add an asymmetrically dominated alternative \( k' \) to the menu of alternatives \( \{j, k\} \) in Figure 8. The angle between \( k \) and \( k' \) is smaller than the angle between \( j \) and \( k' \), hence \( \sigma_{kk'} > \sigma_{jk'} \). Proposition 3 shows that for low levels of information precision, the addition of \( k' \) hurts the dissimilar alternative \( j \) more than the existing similar alternative \( k \). Moreover, Proposition 7 shows that if \( j \) and \( k \) are sufficiently close in utility and \( k' \) is sufficiently inferior, the attraction effect arises for arbitrarily low levels of information precision and persists even for arbitrarily high levels of information precision.

Again a key empirical finding about the attraction effect is that either initial option can increase in probability, depending on the location of the decoy alternative. The Bayesian probit can accommodate the attraction effect in both directions. For example, by adding alternative \( j' \) to the menu of alternatives \( \{j, k\} \) in Figure 8 we favor alternative \( j \) instead of \( k \).

Finally, a word of caution. In empirical applications of the Bayesian probit, several issues should be kept in mind by the analyst. For example, the parametric assumption of cosine similarity (as well as the assumption of expected utility over lotteries) may not provide the best fit to the data for a particular application. Likewise, even in settings where the symmetric prior assumption seems appropriate, the normal distribution assumption will be, at best, a good approximation.

It is beyond the scope of this paper to determine what are good parametric assumptions for any particular application. But we do provide a good starting point in the next section. In Proposition 8 we show that, when restricted to binary choice data, the Bayesian probit is equivalent to a particular version of the standard multinomial probit. In particular, when only two alternatives \( i \) and \( j \) are available, the Bayesian decision maker always picks the alternative with the highest signal realization. This means that, restricted to binary choice data, the two models not only have the same parameters, they are also fully equivalent. Hence, the Bayesian probit inherits the same identification properties for binary choice data. Any additional identifying assumptions and estimation procedures applied to the standard probit using binary choice data will also automatically apply to the Bayesian probit. Hence many lessons already obtained in the discrete choice estimation literature for the standard probit will be useful in applications of the Bayesian probit as well.
Figure 8: Cosine similarity (Hausman and Wise (1978)) assumes the correlation parameter $\sigma_{jk}$ equals the cosine of the angle formed by $j$ and $k$ in the space of observable attributes.

7 Naïve probit versus Bayesian probit

In this section we introduce the naïve probit process. The naïve probit is a standard multinomial probit with equal variance and positive correlations. It has the same parameters as the Bayesian probit, but it is a random utility maximizer and satisfies monotonicity. Since the behavioral implications of the multinomial probit have been explored in the literature, a careful comparison of the Bayesian and the naïve probit reveal further testable implications of our model. We show how choices over three alternatives reveal the complete similarity ranking. We also show that it is possible to obtain closed-form expressions for choice probabilities for the Bayesian and the standard probit in some special cases.

Let each alternative $i$ have utility $X_i$ normally distributed with mean $\mu_i$ and variance $1/t$. For every finite menu of alternatives $\{1, 2, \ldots, n\}$ the vector $(X_1, X_2, \ldots, X_n)$ has a joint normal distribution; the correlation of $X_i, X_j$ is given by $\sigma_{ij}$. There is no prior and no Bayesian updating: the decision maker simply picks the alternative with the highest realization of the random utility:

$$\rho^\mu_t(i, B) = \mathbb{P}\{X_i \geq X_j \text{ for all } j \in B\}$$

where we write $\rho^\mu_t$ for a naïve probit rule with parameters $\mu, \sigma$ and $t$. We refer to the ordered collection $(\rho^\mu_t)_{t>0}$ as a naïve probit process.
The distribution of the signals $X_1, \ldots, X_n$ is the same for the Bayesian and for the naïve probit. But instead of updating beliefs according to Bayes rule and maximizing posterior expected utility, the naïve probit’s decision maker simply chooses the object $i$ with the highest signal $X_i$. In other words, instead of solving a sophisticated noise filtering problem, the naïve decision maker simply chooses the object that “looks best”.

How well does this naïve decision rule do? The next proposition shows that it does just as well as the Bayesian decision rule when the menu of alternatives has only two options. In fact, it is impossible to distinguish a naïve probit process and a Bayesian probit process looking only at binary choice data.

**Proposition 8.** $\rho_t^{\mu \sigma}(i, j) = \rho_t^{\mu \sigma}(i, j)$ for every $i, j \in A$.

When the choice menu has two options, there are two signals with equal variance and only one correlation parameter at play; hence it is impossible for a bias towards more similar options to arise. The Bayesian decision rule has no advantage over the naïve; both choose the alternative with the highest signal realization.

Proposition 8 says that, when restricted to the domain of binary choice data, the Bayesian probit can be seen as particular choice of parameter restrictions on the standard multinomial probit, namely, that all Gaussian signals have equal variance and positive correlations. This equivalence of the Bayesian probit and the standard probit for binary choice data has two consequences worth mentioning.

First, the equivalence means that any axiomatic characterization of binary choice behavior for the standard probit will also automatically characterize the Bayesian probit. In particular, both models satisfy *moderate stochastic transtivity*:

$$\rho(i, j) \geq 1/2 \text{ and } \rho(j, k) \geq 1/2 \implies \rho(i, k) \geq \min\{\rho(i, j), \rho(j, k)\}$$

in other words, if $i \succeq j \succeq k$ then moderate stochastic transtivity says that the decision maker cannot do a worse job discriminating the options in $\{i, k\}$ than in both $\{i, j\}$ and $\{j, k\}$.

Second, both models have the same identifiability properties when restricted to binary choice data. Any estimation procedure that recovers the parameters of a standard probit is automatically also recovering the parameters for a Bayesian probit. Hence, as we illustrated in Section 6, we can build on the extensive literature on the estimation of the standard probit when estimating the Bayesian probit model. We can also reinterpret any estimation exercise of a binary standard probit in the literature as having estimated a version of the Bayesian probit.
Since the Bayesian probit and the naïve probit are equivalent when restricted to binary data, we now ask: when can the two models be distinguished? The next result shows that in menus of three alternatives, the Bayesian probit process and the naïve probit process exhibit radically different choice behavior from the very beginning. While choices for \( t = 0 \) are not defined, we abuse notation and write \( \rho_0^{\mu\sigma}(i, B) = \lim_{t \to 0^+} \rho_t^{\mu\sigma}(i, B) \) and likewise \( \ddot{\rho}_0^{\mu\sigma}(i, B) = \lim_{t \to 0^+} \ddot{\rho}_t^{\mu\sigma}(i, B) \).

**Proposition 9** (Closed-form choice probabilities at time zero). *In the limit as \( t \to 0^+ \),
\[
\ddot{\rho}_0(1, \{1, 2, 3\}) = \frac{1}{4} + \frac{1}{2\pi} \arctan\left(\frac{(1+\sigma_{23}-\sigma_{12}-\sigma_{13})}{\sqrt{4(1-\sigma_{12})(1-\sigma_{13})-(1+\sigma_{23}-\sigma_{12}-\sigma_{13})^2}}\right)
\]
\[
\rho_0(1, \{1, 2, 3\}) = \frac{1}{4} + \frac{1}{2\pi} \arctan\left(\frac{(1+\sigma_{12})(1+\sigma_{13})-\sigma_{23}(1+\sigma_{12}+\sigma_{13}+\sigma_{23})}{\sqrt{3+2\sigma_{12}+2\sigma_{13}+2\sigma_{23}}(1+2\sigma_{12}\sigma_{13}\sigma_{23}-\sigma_{12}^2-\sigma_{13}^2-\sigma_{23}^2)}\right)
\]
and analogous expressions hold for alternatives 2 and 3.*

Random choice models with normally distributed errors (such as the multinomial probit) generally require numerical approximation to calculate choice probabilities as a function of the parameters of the model (see for example Hausman and Wise (1978) and more recent simulation techniques in Train (2009)). Proposition 9 shows it is actually possible to obtain closed-form expressions for choice probabilities as a function of the parameters, for both the Bayesian and the naïve probit in the limit as \( t \to 0 \).

Note that choices do not depend on utility in the limit as precision \( t \) shrinks to zero. Since the variance of signals is given by \( 1/t \) the noise to signal ratio goes to infinity, and choices depend solely on the correlation parameters \( \sigma_{ij} \). Figure 9 shows choice probabilities for both the naïve and the Bayesian probit process for menu \( B = \{1, 2, 3\} \) in the limit as \( t \to 0 \). The Figure plots \( \rho_t^{\mu\sigma}(2, \{1, 2, 3\}) \) and \( \ddot{\rho}_t^{\mu\sigma}(2, \{1, 2, 3\}) \) as a function of the correlation \( \sigma_{23} \) between alternatives 2 and 3. The remaining correlation parameters \( \sigma_{12} \) and \( \sigma_{13} \) are fixed at zero. The solid line corresponds the Bayesian probit, and the dashed line corresponds to the naïve probit. Note that the curves meet when \( \sigma_{23} = 0 \), where noise is orthogonal and both the naïve and the Bayesian probit process converge to the uniform distribution, i.e., each alternative is chosen with probability \( 1/3 \).

A consequence of Proposition 9 is that choices from menus of three alternatives for arbitrarily low information precision (when \( t \to 0 \)) are sufficient to identify the entire ranking of the correlation parameters among three alternatives:

**Proposition 10.** \( \rho_0^{\mu\sigma}(i, \{i, j, k\}) \geq \rho_0^{\mu\sigma}(j, \{i, j, k\}) \) if and only if \( \sigma_{ik} \geq \sigma_{jk} \).
This suggests an alternative definition of the revealed similarity relation: the pair \{i, k\} is revealed more similar than \{j, k\} if \(\rho_{6}^{\mu\sigma}(i, \{i, j, k\}) \geq \rho_{6}^{\mu\sigma}(j, \{i, j, k\})\). Note that this definition extends the revealed similarity relation obtained from binary choice data and defined in Section 4. It is in line with our naming the correlation parameters \(\sigma_{ij}\) the similarity parameters.

When a menu has a unique utility maximizing alternative, it is perhaps not surprising that it is chosen with probability going to one for both the naïve and the Bayesian probit as information precision \(t\) grows arbitrarily large. The next Proposition shows more than that: even when a menu has two or more alternatives tied up in the first place according to utility, the naïve and the Bayesian probit process will break the ties in exactly the same way as \(t \to \infty\).

**Proposition 11.** \(\lim_{t \to \infty} [\rho_{t}^{\mu\sigma}(i, B) - \tilde{\rho}_{t}^{\mu\sigma}(i, B)] = 0\) for all \(i \in B \in A\).

Combining the results in Proposition 9 and Proposition 11, we obtain closed-form expressions for the Bayesian probit process in the special case of three equally desirable alternatives when \(t\) goes to infinity:

\[
\lim_{t \to \infty} \rho_{t}^{\mu\sigma}(i, \{1, 2, 3\}) = \tilde{\rho}_{t}^{\mu\sigma}(i, \{1, 2, 3\})
\]

for each \(i = 1, 2, 3\). In other words, in the knife-edge case in which the three alternatives are exactly equally desirable, the Bayesian probit will exhibit the same choice probabilities under an arbitrarily large amount of information, as the naïve probit when completely uninformed. Note that even though the three alternatives are equally desirable, they are not chosen with equal probability. Proposition 9 shows exactly how the similarity parameters break the tie between alternatives 1, 2 and 3.

**8 Conclusion**

We introduced a new model of random choice, the Bayesian probit. Random choice models finely capture the tendency of a decision maker to choose an alternative \(i\) over another alternative \(j\) and therefore allow us to go beyond classical revealed preference to model the ability of a decision maker to correctly discriminate the alternatives. In the Bayesian probit, this ability is determined by three factors: preference, similarity and information precision. In the model, preference is represented by a numerical utility function, similarity is represented by the correlation of the signals, and information precision is represented by the precision of the signals. The decision maker is less
likely to make a mistake when either the gap in utility, the correlation in the signals, or the precision of the signals increases.

Introducing a notion of similarity that is independent of preference allowed us to reconcile Tversky’s similarity hypothesis with the attraction effect. We showed that when two options $i$ and $j$ are close in utility, introducing a new alternative $k$ that is similar to $j$ will initially hurt the choice probability of the dissimilar alternative $i$ proportionally more than the choice probability of the similar alternative $j$, but eventually hurts $j$ more than $i$. So when information precision is low, similarity has the opposite effect as that predicted by Tversky’s similarity hypothesis. On the other hand, when information precision is sufficiently high, Tversky’s similarity hypothesis holds.

The Bayesian probit allowed us to explain how context-dependent behavior may
arise from rational, Bayesian learning. We showed how Bayesian updating and similarity can explain two classic decoy effects: the attraction effect and the compromise effect. While being able to accommodate violations of monotonicity and context-dependent behavior, the Bayesian probit retains the simplicity and the portability of the standard probit model. In particular, the Bayesian probit allows the econometrician to formulate and test additional parametric assumptions in applications where a vector of observable characteristics of choice objects and/or subjects is available. Hence the Bayesian probit has the same ability to be applied to different discrete choice settings as the standard probit, logit, etc.

To fully understand the effects of similarity and information precision on choice, we modeled a single decision maker in isolation and we treated the precision of the information obtained by the decision maker before making a choice as exogenous. This allowed us to explore the behavioral implications of each parameter of the model in isolation. A clear link between parameters and observable behavior facilitates measurement and provides a useful tool for applications. The model is now ripe for applications in which the decision maker also optimally decides when to stop acquiring information before making a choice, and to situations in which consumers and firms interact strategically. We leave these applications for future work.

References


9 Appendix: Proofs

Proof of Propositions 1 and 2

Consider an enumeration of a finite menu $B = \{1, 2, \ldots, n\}$ and let $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$ denote the vector of utilities for the alternatives in $b$. Prior beliefs about $\mu$ are jointly normally distributed with mean $m(0) = (m_0, m_0, \ldots, m_0)$ and covariance matrix $s(0) = s_0^2 I$, where $I$ is the $n \times n$ identity matrix. A standard application of Bayes’ rule (see for example DeGroot (1970)) gives that posterior beliefs about $\mu$ after observing the signal process $X$ up to time $t > 0$ (or, equivalently, after observing a single signal $X(t)$ with precision $t$) have a joint normal distribution with mean vector $m(t)$ and covariance matrix $s(t)$ given by:

$$s(t) = [s(0)^{-1} + t(\Lambda \Lambda')^{-1}]^{-1}$$

$$m(t) = s(t) [s(0)^{-1} m(0) + (\Lambda \Lambda')^{-1} X(t)]$$
where it is immediate that \( s(t) \) is a deterministic function of time. The posterior mean vector \( m(t) \) has a joint normal distribution with mean vector and covariance matrix given by

\[
\mathbb{E}[m(t)] = s(t) \left[ s(0)^{-1}m(0) + t(\Lambda \Lambda')^{-1}\mu \right]
\]

\[
\text{Var}[m(t)] = ts(t)(\Lambda \Lambda')^{-1}s(t)'
\]

Now consider a menu with only two alternatives enumerated as \( B = \{1, 2\} \). At any given \( t > 0 \), alternative 1 is chosen if and only if \( m_2(t) - m_1(t) < 0 \). Since \( m(t) \) has a normal distribution, \( m_2(t) - m_1(t) \) is also normally distributed with mean and variance given by

\[
\begin{pmatrix}
    ts_0^2(\mu_2 - \mu_1) & 2ts_0^4(1 - \sigma_{12}) \\
    ts_0^2 + 1 - \sigma_{12} & (ts_0^2 + 1 - \sigma_{12})^2
\end{pmatrix}
\]

hence

\[
\rho_t^{12} = \mathbb{P}\{m_2(t) - m_1(t) < 0\} = \Phi\left( \frac{\sqrt{t} - (\mu_1 - \mu_2)}{\sqrt{2(1 - \sigma_{12})}} \right)
\]

where \( \Phi \) denotes the cumulative distribution function of the standard normal distribution. This shows that \( \rho_t^{12} \geq 1/2 \) if and only if \((\mu_1 - \mu_2) \geq 0\), proving Proposition 1. It also shows that for fixed values of \( t \) and \((\mu_1 - \mu_2)\), \( \rho_t^{12} \) is strictly increasing in \( \sigma_{12} \) when \( \mu_1 > \mu_2 \) and strictly decreasing in \( \sigma_{12} \) when \( \mu_1 < \mu_2 \), proving Proposition 2.

**Proof of Proposition 8**

For all \( t > 0 \) we have \( \rho_t^{12}(i, j) = \mathbb{P}\{m_i(t) > m_j(t)\} = \Phi\left( \frac{\sqrt{t} - (\mu_i - \mu_j)}{\sqrt{2(1 - \sigma_{ij})}} \right) = \mathbb{P}\{X_i(t) > X_j(t)\} = \rho_t^{12}(i, j) \).

**Proof of Proposition 9**

Let \( B = \{1, 2, 3\} \) denote the menu of alternatives and write \( X(t) \) for the three-dimensional vector of utility signals corresponding to \( B \). For every time \( t > 0 \) we have \( X(t) \sim \mathcal{N}(t\mu, t\Lambda\Lambda') \) where \( \mu = (\mu_1, \mu_2, \mu_3) \) is the vector of utilities and \( \Lambda\Lambda' = \Sigma \) is the symmetric positive definite matrix formed by the correlation parameters \( \sigma_{ij} \) in each row \( i \) and column \( j \). \( \Lambda \) is the matrix square root of \( \Sigma \) obtained in its Cholesky factorization.

First consider the naïve probit model. Let \( L_1 \) be the \( 2 \times 3 \) matrix given by

\[
L_1 = \begin{bmatrix}
-1 & 1 & 0 \\
-1 & 0 & 1
\end{bmatrix}
\]
so that \( L_1 X(t) = (X_2(t) - X_1(t), X_3(t) - X_1(t)) \). Then
\[
\ddot{\rho}_t^{\mu\sigma}(1, B) = \mathbb{P}\{X_1(t) > X_2(t) \text{ and } X_1(t) > X_3(t)\} = \mathbb{P}\{L_1 X(t) \leq 0\}.
\]
The vector \( L_1 X(t) \) is jointly normally distributed with mean \( tL_1 \mu \) and covariance \( tL_1 \Lambda L_1' \). So \( L_1 X(t) \) has the same distribution as \( tL_1 \mu + \sqrt{t} M Z \), where the vector \( Z = (Z_1, Z_2) \) has a joint standard normal distribution and \( MM' = L_1 \Lambda L_1' \) has full rank. Take \( M \) as the Cholesky factorization
\[
M = \begin{bmatrix}
\sqrt{2(1 - \sigma_{12})} & 0 \\
\frac{1 + \sigma_{23} - \sigma_{12} - \sigma_{13}}{\sqrt{2(1 - \sigma_{12})}} & \sqrt{2(1 - \sigma_{12}) - \frac{(1 + \sigma_{23} - \sigma_{12} - \sigma_{13})^2}{2(1 - \sigma_{12})}}
\end{bmatrix}
\]
then we can write for each \( t > 0 \),
\[
\ddot{\rho}_t^{\mu\sigma}(1, B) = \mathbb{P}\{L_1 X(t) \leq 0\} = \mathbb{P}\{tL_1 \mu + \sqrt{t} M Z \leq 0\} = \mathbb{P}\{MZ \leq -\sqrt{t} L_1 \mu\}
\]
and taking \( t \to 0 \) we obtain
\[
\ddot{\rho}_0^{\mu\sigma}(1, B) = \mathbb{P}\{MZ \leq 0\}
\]
Now \( MZ \leq 0 \) if and only if
\[
\begin{cases}
0 \geq Z_1 \sqrt{2(1 - \sigma_{12})} \\
\text{and} \\
0 \geq \frac{1 + \sigma_{23} - \sigma_{12} - \sigma_{13}}{\sqrt{2(1 - \sigma_{12})}} Z_1 + \sqrt{2(1 - \sigma_{12}) - \frac{(1 + \sigma_{23} - \sigma_{12} - \sigma_{13})^2}{2(1 - \sigma_{12})}} Z_2
\end{cases}
\]
if and only if
\[
\begin{cases}
Z_1 \leq 0 \\
\text{and} \\
Z_2 \leq -Z_1 \frac{(1 + \sigma_{23} - \sigma_{12} - \sigma_{13})}{\sqrt{2(1 - \sigma_{12}) - (1 + \sigma_{23} - \sigma_{12} - \sigma_{13})^2}}
\end{cases}
\]
which describes a cone in \( \mathbb{R}^2 \) and, due to the circular symmetry of the standard normal distribution, we have
\[
\ddot{\rho}_0^{\mu\sigma}(1, B) = \frac{1}{4} + \frac{1}{2\pi} \arctan \left( \frac{(1 + \sigma_{23} - \sigma_{12} - \sigma_{13})}{\sqrt{2(1 - \sigma_{12}) - (1 + \sigma_{23} - \sigma_{12} - \sigma_{13})^2}} \right)
\]
with entirely analogous expressions for \( \ddot{\rho}_0^{\mu\sigma}(2, B) \) and \( \ddot{\rho}_0^{\mu\sigma}(3, B) \).
Now consider the Bayesian probit with the same utility signal $X(t)$ and prior beliefs $\mathcal{N}(m(0), s(0))$ where $m(0) = (m_0, m_0, m_0)$ and $s(0) = s_0 I$. Using the matrix $L_1$ defined in (4) we have for each $t > 0,$

$$\rho_t(1, B) = \mathbb{P}\{m_1(t) > m_2(t) \text{ and } m_1(t) > m_3(t)\} = \mathbb{P}\{L_1 m(t) \leq 0\}$$

The vector $L_1 m(t)$ has a joint normal distribution with mean given by

$$L_1 s(t) [s(0)^{-1} m(0) + t(\Lambda\Lambda')^{-1} \mu]$$

and covariance $t L_1 s(t) \Lambda\Lambda' s(t)' L_1'$. Hence $L_1 m(t)$ has the same distribution as

$$L_1 s(t) s(0)^{-1} m(0) + t L_1 s(t) (\Lambda\Lambda')^{-1} \mu + \sqrt{t} M(t) Z$$

where $Z = (Z_1, Z_2)$ is standard normally distributed and $M(t) m(t)' = L_1 s(t) \Lambda\Lambda' s(t)' L_1'$. Note the contrast with the naïve probit model, where the matrix $M$ does not depend on $t$. We can take $M(t)$ to be the Cholesky factorization

$$M(t) = \begin{bmatrix}
    M_{11}(t) & 0 \\
    M_{21}(t) & M_{22}(t) \\
\end{bmatrix}$$

given by

$$M_{11}(t) = \frac{C_1(t)}{\sqrt{(-1 - 3s^4 t^2 - s^6 t^3 + \sigma_{12}^2 + \sigma_{13}^2 + \sigma_{23}^2 - 2\sigma_{12} \sigma_{13} \sigma_{23} + s^2 t (\sigma_{12}^2 + \sigma_{13}^2 + \sigma_{23}^2 - 3))^2}}$$

where

$$C_1(t) = s^4 [-2s^8 t^4 (-1 + \sigma_{12}) - 2s^6 t^3 (-4 + 2\sigma_{12} (1 + \sigma_{12}) + (\sigma_{13} - \sigma_{23})^2)$$

$$- 4s^2 t (2 + \sigma_{12}) (-1 + \sigma_{12}^2 + \sigma_{13}^2 - 2\sigma_{12} \sigma_{13} \sigma_{23} + \sigma_{23}^2)$$

$$- s^4 t^2 (-12 + 7\sigma_{12}^2 + 2\sigma_{12} (\sigma_{12} (5 + \sigma_{12}) + \sigma_{13}^2) - 6 (1 + 2\sigma_{12}) \sigma_{13} \sigma_{23} + (7 + 2\sigma_{12}) \sigma_{23}^2)$$

$$- (-1 + \sigma_{12}^2 + \sigma_{13}^2 - 2\sigma_{12} \sigma_{13} \sigma_{23} + \sigma_{23}^2) (2 + 2\sigma_{12} - (\sigma_{13} + \sigma_{23})^2)]$$

and the expressions for $M_{21}(t)$ and $M_{22}(t)$ are similarly cumbersome and omitted.

Now we can write, for each $t > 0,$

$$\rho_t(1, B) = \mathbb{P}\{L_1 m(t) \leq 0\}$$

$$= \mathbb{P}\{L_1 s(t) s(0)^{-1} m(0) + t L_1 s(t) (\Lambda\Lambda')^{-1} \mu + \sqrt{t} M(t) Z \leq 0\}$$

$$= \mathbb{P}\{\sqrt{t} M(t) Z \leq -L_1 s(t) s(0)^{-1} m(0) - t L_1 s(t) (\Lambda\Lambda')^{-1} \mu\}$$

$$= \mathbb{P}\{M(t) Z \leq -\frac{1}{\sqrt{t}} L_1 s(t) s(0)^{-1} m(0) - \sqrt{t} L_1 s(t) (\Lambda\Lambda')^{-1} \mu\}$$
Lemma 12. In the limit as \( t \) goes to zero,
\[
\lim_{t \to 0^+} \frac{1}{\sqrt{t}} L_1 s(t) s(0)^{-1} m(0) = 0,
\]
\[
\lim_{t \to 0^+} M_{11}(t) = s^2 \sqrt{\frac{2 + 2 \sigma_{12} - \sigma_{13}^2 - 2 \sigma_{13} \sigma_{23} - \sigma_{23}^2}{1 + 2 \sigma_{12} \sigma_{13} \sigma_{23} - \sigma_{12}^2 - \sigma_{13}^2 - \sigma_{23}^2}} > 0,
\]
and
\[
\lim_{t \to 0^+} \frac{M_{21}(t)}{M_{22}(t)} = \frac{(1 + \sigma_{12})(1 + \sigma_{13}) - \sigma_{23}(1 + \sigma_{12} + \sigma_{13} + \sigma_{23})}{\sqrt{(3 + 2 \sigma_{12} + 2 \sigma_{13} + 2 \sigma_{23})(1 + 2 \sigma_{12} \sigma_{13} \sigma_{23} - \sigma_{12}^2 - \sigma_{13}^2 - \sigma_{23}^2)}}.
\]

Proof. Long and cumbersome, omitted.

Using Lemma 12 we have
\[
\rho_0^{\mu \sigma}(1, B) = \lim_{t \to 0^+} \rho_t^{\mu \sigma}(1, B)
\]
\[
= \lim_{t \to 0^+} \mathbb{P} \left\{ M(t) Z \leq -\frac{1}{\sqrt{t}} L_1 s(t) s(0)^{-1} m(0) - \sqrt{t} L_1 s(t)(\Lambda \Lambda')^{-1} \mu \right\}
\]
\[
= \mathbb{P} \left\{ Z_1 \lim_{t \to 0^+} M_{11}(t) \leq 0 \text{ and } Z_2 \leq -Z_1 \lim_{t \to 0^+} \frac{M_{21}(t)}{M_{22}(t)} \right\}
\]
and by the circular symmetry of the standard normal distribution we obtain
\[
\rho_0^{\mu \sigma}(1, B) = \frac{1}{4} + \frac{1}{2\pi} \arctan \left( \frac{(1 + \sigma_{12})(1 + \sigma_{13}) - \sigma_{23}(1 + \sigma_{12} + \sigma_{13} + \sigma_{23})}{\sqrt{(3 + 2 \sigma_{12} + 2 \sigma_{13} + 2 \sigma_{23})(1 + 2 \sigma_{12} \sigma_{13} \sigma_{23} - \sigma_{12}^2 - \sigma_{13}^2 - \sigma_{23}^2)}} \right)
\]
with analogous expressions for \( \rho_0^{\mu \sigma}(2, B) \) and \( \rho_0^{\mu \sigma}(3, B) \).

Proof of Proposition 10

By Proposition 9 we have \( \rho_0^{\mu \sigma}(i, \{i, j, k\}) \geq \rho_0^{\mu \sigma}(j, \{i, j, k\}) \) if and only if \( (1 + \sigma_{ij})(1 + \sigma_{ik}) - \sigma_{jk}(1 + \sigma_{ij} + \sigma_{ik} + \sigma_{jk}) > (1 + \sigma_{ij})(1 + \sigma_{jk}) - \sigma_{ik}(1 + \sigma_{ij} + \sigma_{ik} + \sigma_{jk}) \) if and only if \( \sigma_{ik} \geq \sigma_{jk} \).

Proof of Proposition 11

It is easy to show that if \( \mu_i < \mu_j \) for some \( j \in B \) then alternative \( i \) is chosen with probability going to zero for both the Bayesian and the na"ive probit as \( t \) goes to infinity. We will now show that the distribution of the posterior mean vector \( m(t) \) gets arbitrarily ‘close’ to the distribution of the utility signals \( X(t)/t \) and that in fact both
the naïve and the Bayesian probit will break ties in exactly the same way when there is more than one maximizer of \( \mu \) in menu \( B \).

First we use the matrix identity \((I + M^{-1})^{-1} = M(M + I)^{-1}\) to write the covariance of posterior beliefs as

\[
s(t) = [s(0)^{-1} + t\Sigma^{-1}]^{-1}
\]

\[
= s \left[ I + \left( \frac{1}{st} \Sigma \right)^{-1} \right]^{-1}
\]

\[
= s \left[ \frac{1}{st} \Sigma \left( \frac{1}{st} \Sigma + I \right)^{-1} \right]
\]

\[
= \frac{1}{t} \Sigma \left( \frac{1}{st} \Sigma + I \right)^{-1}
\]

and now the vector of posterior means can be written as

\[
m(t) = s(t)s(0)^{-1}m(0) + s(t)\Sigma^{-1}X(t)
\]

\[
= \frac{1}{st} \Sigma \left( \frac{1}{st} \Sigma + I \right)^{-1} m(0) + \frac{1}{t} \Sigma \left( \frac{1}{st} \Sigma + I \right)^{-1} \Sigma^{-1}X(t)
\]

Hence

\[
\mathbb{E} \left[ \sqrt{t} m(t) \right] - \sqrt{t} \mu = \frac{1}{s\sqrt{t}} \Sigma \left( \frac{1}{st} \Sigma + I \right)^{-1} m(0) + \\
+ \sqrt{t} \left[ \Sigma \left( \frac{1}{st} \Sigma + I \right)^{-1} \Sigma^{-1} - I \right] \mu
\]
and using the matrix identity \((I + M)^{-1} = I - (I + M)^{-1}M\) we have
\[
\mathbb{E} \left[ \sqrt{t} m(t) \right] - \sqrt{t} \mu = \frac{1}{s \sqrt{t}} \Sigma \left( \frac{1}{st} \Sigma + I \right)^{-1} m(0) + \\
+ \sqrt{t} \left[ \Sigma \left( I - \left( \frac{1}{st} \Sigma + I \right)^{-1} \frac{1}{ts} \Sigma \right) \Sigma^{-1} - \Sigma \Sigma^{-1} \right] \mu
\]
\[
= \frac{1}{s \sqrt{t}} \Sigma \left( \frac{1}{st} \Sigma + I \right)^{-1} m(0) + \\
+ \sqrt{t} \left[ \Sigma \left( I - \left( \frac{1}{st} \Sigma + I \right)^{-1} \frac{1}{ts} \Sigma - I \right) \Sigma^{-1} \right] \mu
\]
\[
= \frac{1}{s \sqrt{t}} \Sigma \left( \frac{1}{st} \Sigma + I \right)^{-1} m(0) + \\
- \frac{1}{\sqrt{t}} \left[ \Sigma \left( \frac{1}{st} \Sigma + I \right)^{-1} \frac{1}{s} \Sigma \Sigma^{-1} \right] \mu
\]
\[
= \frac{1}{\sqrt{t}} \left\{ \frac{1}{s} \Sigma \left( \frac{1}{st} \Sigma + I \right)^{-1} [m(0) - \mu] \right\}
\]
where since matrix inversion is continuous in a neighborhood of \(I\), the expression inside the curly brackets converges to
\[
\frac{1}{s} \Sigma [m(0) - \mu]
\]
as \(t\) goes to infinity and therefore is bounded. Since the expression in curly brackets is multiplied by \(1/\sqrt{t}\), we have
\[
\lim_{t \to \infty} \left| \mathbb{E} \left[ \sqrt{t} m(t) \right] - \sqrt{t} \mu \right| = 0.
\]
Moreover
\[
\lim_{t \to \infty} \text{Var} \left[ \sqrt{t} m(t) \right] = \lim_{t \to \infty} t \text{Var} \left[ m(t) \right]
\]
\[
= \lim_{t \to \infty} t^2 s(t) \Sigma^{-1} s(t)'
\]
\[
= \lim_{t \to \infty} t^2 \left[ \frac{1}{t} \Sigma \left( \frac{1}{st} \Sigma + I \right)^{-1} \right] \Sigma \left[ \frac{1}{t} \Sigma \left( \frac{1}{st} \Sigma + I \right)^{-1} \right]'
\]
\[
= \lim_{t \to \infty} \Sigma \left( \frac{1}{st} \Sigma + I \right)^{-1} \Sigma^{-1} \left( \frac{1}{st} \Sigma + I \right)^{-1} \Sigma
\]
\[
= \Sigma
\]
Since the vector $\sqrt{t} m(t)$ is normally distributed for all $t$, and since

$$
\rho_t^{\mu}(i, B) = \mathbb{P} \left( \bigcap_{j \in B \setminus \{i\}} \{ m_i(t) > m_j(t) \} \right) \\
= \mathbb{P} \left( \bigcap_{j \in B \setminus \{i\}} \{ \sqrt{t} m_i(t) > \sqrt{t} m_j(t) \} \right)
$$

the mean and covariance of the vector $\sqrt{t} m(t)$ fully determine choice probabilities.

By the limit calculations above, given any $\varepsilon > 0$ there exists $T > 0$ such that the mean and covariance matrix of $\sqrt{t} m(t)$ remain inside a $\varepsilon$-neighborhood of $\sqrt{t} \mu$ and $\Sigma$, respectively, for every $t \geq T$. Hence by continuity, $\rho_t^{\mu}(i, B) \to \rho^\mu(i, B)$ as desired. 

**Proof of Proposition 3**

Independently of the values $\mu_1$ and $\mu_2$, when the menu of available alternatives is $\{1, 2\}$, each alternative is chosen with probability $1/2$ at the start of the random choice process, i.e., in the limit as $t \to 0+$. By Proposition 9, when the menu is $\{1, 2, 3\}$ the probability that alternative 1 is chosen at the start of the random choice process is given by

$$
\frac{1}{4} + \frac{1}{2\pi} \arctan \left( \frac{(1+\sigma_{12})(1+\sigma_{13})-\sigma_{23}(1+\sigma_{12}+\sigma_{13}+\sigma_{23})}{\sqrt{(3+2\sigma_{12}+2\sigma_{13}+2\sigma_{23})(1+2\sigma_{12}\sigma_{13}\sigma_{23}-\sigma_{12}^2-\sigma_{13}^2-\sigma_{23}^2)}} \right)
$$

and the probability for alternative 2 is given by

$$
\frac{1}{4} + \frac{1}{2\pi} \arctan \left( \frac{(1+\sigma_{12})(1+\sigma_{23})-\sigma_{13}(1+\sigma_{12}+\sigma_{13}+\sigma_{23})}{\sqrt{(3+2\sigma_{12}+2\sigma_{13}+2\sigma_{23})(1+2\sigma_{12}\sigma_{13}\sigma_{23}-\sigma_{12}^2-\sigma_{13}^2-\sigma_{23}^2)}} \right).
$$

Since the function arctan is strictly increasing, the probability of alternative 2 is larger than the probability of alternative 1 if and only if

$$(1+\sigma_{12})(1+\sigma_{13})-\sigma_{23}(1+\sigma_{12}+\sigma_{13}+\sigma_{23}) < (1+\sigma_{12})(1+\sigma_{23})-\sigma_{13}(1+\sigma_{12}+\sigma_{13}+\sigma_{23})$$

which holds if and only if

$$(\sigma_{23} - \sigma_{13})(2 + 2\sigma_{12} + \sigma_{13} + \sigma_{23}) > 0$$

which holds since all $\sigma_{ij}$ are positive and since by assumption $\sigma_{23} > \sigma_{13}$. This shows that when $t > 0$ is sufficiently small, introducing alternative 3 hurts alternative 1 more than it hurts alternative 2. This holds for any utility values $\mu_1, \mu_2, \mu_3$.

**Proof of Proposition 4**

By Proposition 9 for the naïve probit we have $\tilde{\rho}_{0}^{\mu}(1, \{1, 2, 3\}) > \tilde{\rho}_{0}^{\mu}(2, \{1, 2, 3\})$ if and only if

$$
1 + \sigma_{23} - \sigma_{12} - \sigma_{13} > 1 + \sigma_{13} - \sigma_{12} - \sigma_{23}
$$

40
which holds since, by assumption, $\sigma_{23} > \sigma_{13}$. Under the assumption $1 \sim 2 \sim 3$ by Proposition 1 we have $\mu_1 = \mu_2 = \mu_3$ hence for every $t > 0$ we have

$$\tilde{\rho}^{\mu \sigma}_t(i, \{1, 2, 3\}) = \mathbb{P}\{X_i(t) \geq X_j(t) \text{ for all } j\}$$

$$= \mathbb{P}\{\sqrt{t}[X_i(t) - \mu_i] \geq \sqrt{t}[X_j(t) - \mu_j] \text{ for all } j\}$$

$$= \tilde{\rho}^{\mu \sigma}_0(i, \{1, 2, 3\})$$

since for every $t > 0$ the vector $\sqrt{t}[X(t) - \mu]$ is jointly normally distributed with mean vector zero and covariance matrix $\Sigma$. Hence for all $t > 0$,

$$\tilde{\rho}^{\mu \sigma}_t(1, \{1, 2, 3\}) - \tilde{\rho}^{\mu \sigma}_t(2, \{1, 2, 3\}) = \tilde{\rho}^{\mu \sigma}_0(1, \{1, 2, 3\}) - \tilde{\rho}^{\mu \sigma}_0(2, \{1, 2, 3\}) > 0$$

Finally, by Proposition 11, for all $t$ sufficiently large, we have $\rho^{\mu \sigma}_t(1, \{1, 2, 3\}) > \rho^{\mu \sigma}_t(2, \{1, 2, 3\})$ as desired. \qed

**Proof of Proposition 5**

Since $X(t)$ is jointly normally distributed, $\Sigma$ has a positive determinant

$$1 + 2\sigma_{12}\sigma_{23}\sigma_{13} - \sigma_{12}^2 - \sigma_{23}^2 - \sigma_{13}^2 > 0$$

hence we can’t fix any arbitrary values for $\sigma_{12}$ and $\sigma_{13}$ and take the limit $\sigma_{23} \rightarrow 1$. In particular, $\sigma_{23} \rightarrow 1$ implies $|\sigma_{12} - \sigma_{13}| \rightarrow 0$. We will assume that $\sigma_{12}$ and $\sigma_{13}$ stay bounded away from one. Hence as $\sigma_{23}$ approaches one, it eventually becomes the largest correlation parameter. This assumption rules out, for example, the case $\sigma_{12} = \sigma_{13} = \sigma_{23} \rightarrow 1$.

Item (i) in the Corollary follows from Propostion 9 once we show that, as $\sigma_{23} \rightarrow 1$ approaches one from below, we have

$$\frac{(1 + \sigma_{12})(1 + \sigma_{23}) - \sigma_{13}(1 + \sigma_{12} + \sigma_{23} + \sigma_{13})}{\sqrt{(3 + 2\sigma_{12} + 2\sigma_{23} + 2\sigma_{13})(1 + 2\sigma_{12}\sigma_{23}\sigma_{13} - \sigma_{12}^2 - \sigma_{23}^2 - \sigma_{13}^2)}} \rightarrow +\infty \quad (6)$$

since the numerator in (6) converges to a strictly positive number, while the denominator is positive and converges to zero.

The easiest way to show this is to fix $\sigma_{12} = \sigma_{23} = \sigma < 1$ and take the limit $\sigma_{23} \rightarrow 1$. Then the numerator in (6) goes to $2(1 - \sigma) > 0$; in the denominator, the term $(3 + 2\sigma_{12} + 2\sigma_{23} + 2\sigma_{13})$ goes to $(5 + 4\sigma) > 0$ while the other term is equal to the determinant of $\Sigma$ which converges to zero from above.
To prove item (ii) in the Corollary, recall from the proof of Proposition 4 that when 
\(1 \sim 2 \sim 3\) we have \(\rho_t^{\mu_\sigma}(i, \{1, 2, 3\})\) converges to \(\hat{\rho}_0^{\mu_\sigma}(i, \{1, 2, 3\})\) as \(t \to \infty\). The result then follows from Proposition 9 once we show that 
\[
\frac{(1 + \sigma_{13} - \sigma_{12} - \sigma_{23})}{\sqrt{4(1 - \sigma_{12})(1 - \sigma_{23}) - (1 + \sigma_{13} - \sigma_{12} - \sigma_{23})^2}} \to 0.
\]
The easiest way to show this is to fix \(\sigma_{12} = \sigma_{23} = \bar{\sigma} < 1\) in which case the expression simplifies to 
\[
\frac{\sqrt{1 - \sigma_{23}}}{\sqrt{4(1 - \bar{\sigma}) + (1 - \sigma_{23})}}.
\]
Taking the limit \(\sigma_{23} \to 1\), the numerator goes to zero while the denominator goes to \(\sqrt{4(1 - \bar{\sigma})} > 0\) and we are done.

**Proof of Proposition 6**

Follows immediately from Proposition 9 or from Proposition 10.

**Proof of Proposition 7**

Let \(\varepsilon > 0\) and recall \(m(t)\) is the (random) vector of posterior mean beliefs at time \(t\), which has a joint normal distribution. The probability that alternative 2 is chosen at time \(t\) is equal to the probability that \(m_2(t) > m_1(t)\) and \(m_2(t) > m_3(t)\). This happens if and only if the bi-dimensional vector \((m_1(t) - m_2(t), m_3(t) - m_2(t))\) has negative coordinates. This vector has a joint normal distribution with mean given by

\[
\mathbb{E}[m_1(t) - m_2(t)] = s^2 t \left[ \mu_3 (1 + s^2 t + \sigma_{12}) (\sigma_{23} - \sigma_{13}) + \mu_2 \left( (-1 - s^2 t)(1 + s^2 t + \sigma_{12}) + \sigma_{23} \sigma_{13} + \sigma_{13}^2 \right) + \mu_1 \left( 1 + \sigma_{12} + s^2 t (2 + s^2 t + \sigma_{12}) - \sigma_{23} (\sigma_{23} + \sigma_{13}) \right) \right] / \left[ (1 + s^2 t) (1 + s^2 t (2 + s^2 t) - \sigma_{12}^2 - \sigma_{23}^2 - \sigma_{13}^2) + 2 \sigma_{12} \sigma_{23} \sigma_{13} \right]
\]

and

\[
\mathbb{E}[m_3(t) - m_2(t)] = s^2 t \left[ \mu_1 (1 + s^2 t + \sigma_{23}) (\sigma_{12} - \sigma_{13}) + \mu_2 \left( (-1 - s^2 t)(1 + s^2 t + \sigma_{23}) + \sigma_{12} \sigma_{13} + \sigma_{13}^2 \right) + \mu_3 \left( 1 + \sigma_{23} + s^2 t (2 + s^2 t + \sigma_{23}) - \sigma_{12} (\sigma_{12} + \sigma_{13}) \right) \right] / \left[ (1 + s^2 t) (1 + s^2 t (2 + s^2 t) - \sigma_{12}^2 - \sigma_{23}^2 - \sigma_{13}^2) + 2 \sigma_{12} \sigma_{23} \sigma_{13} \right]
\]
The denominator is equal in both expressions and can be written as

\[
(1 + s^2t) (1 + s^2t (2 + s^2t) - \sigma_{12}^2 - \sigma_{23}^2 - \sigma_{13}^2) + 2\sigma_{12}\sigma_{23}\sigma_{13} = \\
 s^2t (s^2t (3 + s^2t) + 3 - \sigma_{12}^2 - \sigma_{23}^2 - \sigma_{13}^2) + \\
 1 + 2\sigma_{12}\sigma_{23}\sigma_{13} - \sigma_{12}^2 - \sigma_{23}^2 - \sigma_{13}^2
\]

which is clearly positive since \( s, t > 0, \sigma_{ij}^2 < 1 \) and \( 1 + 2\sigma_{12}\sigma_{23}\sigma_{13} - \sigma_{12}^2 - \sigma_{23}^2 - \sigma_{13}^2 \) is the positive determinant of the covariance matrix \( \Sigma \).

In both numerators, the expression multiplying the coefficient \( \mu_3 \) is positive. In the first case, note that \( 1 + \sigma_{12} + s^2t > 0 \) and that \( \sigma_{23} > \sigma_{13} \) since by assumption the pair \((2, 3)\) is more similar than the pair \((1, 3)\). In the second case, the expression multiplying \( \mu_3 \) can be written as

\[
[1 - \sigma_{12}^2] + [s^2 t (2 + \sigma_{23} + s^2 t)] + [\sigma_{23} - \sigma_{12}\sigma_{13}]
\]

where each expression in brackets is positive. Therefore for any fixed \( t \) both coordinates of the mean vector \( (\mathbb{E}[m_1(t) - m_2(t)], \mathbb{E}[m_3(t) - m_2(t)]) \) can be made arbitrarily negative by taking \( \mu_3 \) negative and sufficiently large in absolute value.

The covariance matrix \( \text{Var}[m(t)] = ts(t)(\Lambda\Lambda')^{-1}s(t)' \) does not depend on \( \mu \). Since \( 1 \sim 2 \) we have \( \mu_1 = \mu_2 \) we therefore both \( \rho_t^{(1, \{1, 2, 3\})} \) and \( \rho_t^{(2, \{1, 2, 3\})} \) converge to 1/2 when \( t \) goes to infinity. Note that, while increasing the absolute value of the negative parameter \( \mu_3 \) does not change \( \text{Var}[m(t)] \) for any \( t \), it decreases both \( \mathbb{E}[m_1(t) - m_2(t)] \) and \( \mathbb{E}[m_3(t) - m_2(t)] \) for every \( t > 0 \) and therefore increases \( \rho_t^{(2, \{1, 2, 3\})} \) for every \( t > 0 \). Moreover, for fixed \( t > 0 \), \( \rho_t^{(2, \{1, 2, 3\})} \) can be made arbitrarily close to 1 by taking \( \mu_3 \) sufficiently negative. This guarantees that we can have the attraction effect starting arbitrarily early in the random choice process. Moreover, since \( \mathbb{E}[m_1(t) - m_2(t)] \) above converges to zero from below as \( t \) goes to infinity, \( \rho_t^{(2, \{1, 2, 3\})} \) will converge to 1/2 from above, while \( \rho_t^{(1, \{1, 2, 3\})} \) will converge to 1/2 from below. □