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. We show how, unlike any random utility model, the Bayesian probit can jointly accommodate similarity effects, the attraction effect and the compromise effect. We also provide a new definition of revealed similarity based only on the choice rule and show that in the Bayesian probit (i) signal averages capture revealed preference; (ii) signal precision captures the decision maker’s familiarity with the options; and (iii) signal correlation captures our new definition of revealed similarity. 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: pnatenzon@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, D-TEA Workshop, IESE, IMPA, Johns Hopkins, Kansas University, Haifa, Harvard/MIT Theory Seminar, Hebrew University, NYU, Princeton, the RUD Conference at the Colegio Carlo Alberto, SBE Meetings, Toronto, UBC, Washington University in St. Louis and the D-Day workshop at Yale. All remaining errors are my own.
1 Introduction

Suppose a consumer would be equally likely to pick either option when presented with a choice of two comparable personal computers A and B. If we introduce a third option C that is similar to B, how should the probabilities of choosing A and B change? 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 many features should compete more intensely 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 modern tools used in discrete choice estimation (see for example McFadden (2001) and Train (2009)). But a large empirical literature, originating in marketing, has identified situations in which the hypothesis systematically fails, the most prominent example being the attraction effect.

The attraction effect refers to an *increase* in the probability of choosing computer B when we introduce an option C that 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. Huber et al. (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 et al. (2012) provide many references).

In the attraction effect, introducing an asymmetrically dominated option C increases the probability of choosing a similar alternative B, in detriment of A. This is the exact opposite of what Tversky’s similarity hypothesis states. What is missing in the similarity hypothesis? Why does the attraction effect happen? How can the two be reconciled? In this paper, we propose a new model of individual decision making, *the Bayesian probit*, which allows us to give precise answers to these questions. Our main contribution is to provide a unified theory which explains under which conditions similarity plays the role predicted by Tversky’s hypothesis, and also when it may generate the attraction effect and related phenomena (such as the compromise effect).
Previous answers to these questions have generated a large collection of formal
models based on biases, heuristics, multiple selves and bounded rationality. For ex-
ample, 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 Section 2, we show with
a non-parametric example that the attraction effect may arise from simple Bayesian
updating. Our new model, the Bayesian probit, can accommodate this example and
many other kinds of context-dependent choice behavior as a result of individual, ratio-
nal inference.

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 max-
imizes 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 (forthcoming). They
also include most econometric models of discrete choice such as probit, logit, nested
logit etc.

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

The Bayesian probit retains the simplicity of the standard probit model: each choice
alternative is associated with a normally distributed random variable. This variable is a
noisy signal about the utility of each alternative. It captures the information available
to the decision maker when a choice is made. Hence, the precision of these signals
captures how much the decision maker was able to learn about the alternatives in a
menu before making the choice. The decision maker uses all the information provided
by the signals in a menu of alternatives to determine the best choice. In Section 4
we provide a new definition of similarity based only on the random choice function
and show that in the model it is captured by the correlation of the signals. Context
dependent behavior arises when varying degrees of similarity among the alternatives
allow the decision maker to more quickly and easily rank some pairs of alternatives
than others.
We introduce the model in Section 3 and our new definition of similarity in Section 4. Section 5 contains the main results: we show how similarity and information precision interact to explain and reconcile several empirical regularities. In Section 6 we write the Bayesian probit as a function of a vector of observables to illustrate how the model is amenable to the same discrete choice estimation exercises in which the standard probit is used. Section 7 further explores the testable implications of the model and Section 8 concludes.

Game-theoretic models with standard rational players have been shown to be compatible with decoy effects. For example, Kamenica (2008) 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 collective decision making when agents use the Rawlsian arbitration rule as a bargaining protocol. In contrast, our decision-theoretic model explains how decoy effects may arise from individual, rational inference, even without the presence of strategic interactions. We therefore provide a model of individual choice behavior with the potential to account for the empirical findings in job candidate evaluation, choice of political candidates, investment decisions, medical decisions and other settings in which individual choice behavior has been found to be context-dependent.

2 Example

This example illustrates how a classic decoy effect, the attraction effect, may arise from rational, Bayesian inference. Suppose a consumer is about to be presented with three options for a new personal computer. For simplicity let us call the three computers $A$, $B$ and $C$. 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 computers $A$, $B$ 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) says absolutely nothing about the relative position of computer $A$. In other words, upon receiving this information 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: any information that allows a precise ranking of computers $B$ and $C$ will make computer $A$ unlikely to be chosen by the Bayesian decision maker.

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). With limited information (for example, under limited time to inspect the options before making a choice), asymmetric dominance will allow a decision maker to have a higher degree of confidence about 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 $A$.

This example illustrates how the context may influence choice even for a rational, Bayesian decision maker. While the beliefs of the decision maker about the utility of the options are a priori symmetric, a combination of very precise information about the true relative ranking of computers $B$ and $C$ and the absence of information about the ranking of other pairs clearly favored a choice of computer $B$.

In the Bayesian probit model, this extreme example is captured by noisy utility signals for computers $B$ and $C$ with a correlation parameter close to one, while the overall precision of the signals for computers $A$, $B$ and $C$ being close to zero. Imprecise signals are uninformative about the utility of computer $B$ or computer $C$, but high correlation makes them very informative about the difference in utility between $B$ and $C$. Intuitively, when we take the difference of two perfectly correlated noisy signals, the noise cancels. Thus upon observing the realization of those signals the decision maker learns with a high degree of certainty whether the event ($B \succ C$) or ($C \succ B$) occurred, and little else. The Bayesian probit model, introduced in the next Section, can also accommodate less extreme examples.
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 our universe of choice alternatives and $\mathcal{A}$ is the collection of choice situations (menus) that a decision maker may face. A function $\rho : A \times \mathcal{A} \rightarrow [0, 1]$ 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 $b \in \mathcal{A}$.

A random choice rule $\rho$ is monotone if $\rho(i, a) \geq \rho(i, b)$ whenever $a \subset b$. Under 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.

Our model is a departure from random utility maximization and allows violations of monotonicity. A random choice rule is a Bayesian Probit rule when choice probabilities are generated as follows. When facing a menu of choice alternatives $b = \{1, 2, \ldots, n\}$ the decision maker observes a signal $X_i$ about the utility of each alternative $i \in b$, updates his beliefs about the utility of the alternatives using Bayes’ rule and chooses the alternative with the highest posterior mean.

The true value of the alternatives is given by a utility function $\mu : A \rightarrow \mathbb{R}$. We write $\mu_i := \mu(i)$ for the utility of alternative $i$. Prior beliefs about the utility of the alternatives are identically and independently normally distributed. The signals $X_1, X_2, \ldots, X_n$ have a joint normal distribution. The expectation of each signal $X_i$ is
the true utility $\mu_i$. All signals have the same variance $1/t$, hence a single parameter $t > 0$ captures the precision of signals. 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 $\sigma : A \times A \rightarrow [0,1]$ satisfies $\sigma_{ij} = \sigma_{ji}$ and $\sigma_{ii} = 1$ for every $i,j \in A$. In addition, the determinant of the matrix $(\sigma_{ij})_{i,j=1,...,k}$ is strictly positive for every finite subset of options $\{1,\ldots,k\}$. We denote by $\rho^{\mu_\sigma}_t$ the Bayesian Probit rule with utility parameters $\mu$, correlation parameters $\sigma$, and precision parameter $t$.

A single parameter $t > 0$ captures the precision of the signals in the Bayesian Probit rule. As $t$ increases, the choices in $\rho^{\mu_\sigma}_t$ will more closely reflect the true utility of the alternatives given by $\mu$. We refer to a family $(\rho^{\mu_\sigma}_t)_{t>0}$ of Bayesian Probit rules indexed by precision $t$ as a Bayesian Probit Process.

Equivalently, we can model the signals obtained by the decision maker when facing the menu $b = \{1,\ldots,n\}$ as a continuous random process $X : \Omega \times [0,\infty) \rightarrow \mathbb{R}^n$. 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 (which 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.\footnote{Drift 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.} It is immediate to see that $X(t)/t$ has the same distribution for the signals and conveys the same information about the utility of the alternatives. Since the 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 textbook on Brownian motion).

### 4 Revealed Similarity

Random choice models finely capture the tendency of a decision maker to pick an alternative $j$ over an alternative $k$. This allows us to go beyond classical revealed preference theory to explain discriminability: how difficult is it for a decision maker to choose the best alternative out of the pair $j,k$?
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 $i \succ j$ 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.** Let $\succ$ be the revealed preference relation for $\rho_t^{u_\sigma}$. Then $i \succ j$ if and only if $\mu_i \geq \mu_j$.

Proposition 1 shows that the revealed preference relation $\succ$ obtained from a Bayesian probit rule $\rho_t^{u_\sigma}$ is always complete and transitive. Note that the revealed preference relation is solely represented by the utility parameters $\mu_i$. Therefore the revealed preference relation corresponding to a Bayesian probit process $(\rho_t^{u_\sigma})_{t>0}$ is well-defined and independent of $t$.

We define the revealed similarity relation $\succsim$ 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 $\{i, j\} \succsim \{k, \ell\}$, when

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

and

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

As usual we will write $\sim$ for the symmetric part and $\succsim$ for the asymmetric part of the relation $\succsim$. 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 closer to $1/2$ 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 $\{i, j\}$ is a more similar pair.
Our definition of similarity is best motivated with the following geometric example. Suppose the universe of choice alternatives is the set of all triangles in the Euclidean plane. Suppose further that the decision maker’s tastes over triangles are very simple: she always prefers triangles that have larger areas. So 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?

![Triangles](image)

**Figure 1: Which triangle has the largest area?**

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.

If the difference in utility were more pronounced, i.e., if one triangle were much better (larger) than the other, she probably wouldn’t have any trouble discriminating among the options. Likewise, a consumer presented with two products, of which one is significantly cheaper and of much better quality, should seldom make a mistake. Hence, any probabilistic model of choice should include the gap in desirability (measured as the difference in utility) as an important explanatory variable for the ability of subjects to discriminate among the options.

But the gap in utility alone cannot tell the whole story. Even if the utility of a pair of objects is kept exactly the same, it is possible to increase the decision maker’s ability to discriminate among a pair of options by making their features overlap more (i.e., increase their resemblance while keeping them on their original indifference curves). For example, suppose she still prefers larger triangles, and she is offered 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.

![Figure 2: Which triangle has the largest area?](image)

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—as we pointed out, the pairs in both Figures exhibit exactly the same difference in area. So what happened?

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 helped improve the decision maker’s ability to discriminate among the options. Building on this intuition, according to our formal definition the pair \{i’, j\} of Figure 2 is *more similar* than the pair \{i, j\} of Figure 1.

In Euclidean geometry, two triangles are defined to be similar when they have the same internal angles (the pair in Figure 2 is an example). This geometric definition of similarity abstracts from size, and only refers to the shape of the objects being compared. In the same spirit, our formal definition departs from the everyday use of the word ‘similarity’ and abstracts from utility.

The next abstract example clearly shows that (i) the same principle operates more generally than for triangles and areas; and (ii) while utility takes time to be perceived, similarity is perceived instantly.\(^2\) Suppose the universe of choice alternatives is the set of all star-shaped figures. The decision maker’s tastes over star-shaped figures are again very simple. This time, she 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, she always

\(^2\)I am grateful to David K. Levine for suggesting this example.
strictly prefers the six-pointed star. With just a glance at the options in Figure 3, which star should she pick?

![Figure 3: Which star has more points?](image)

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 consider the options in Figure 4. If the decision maker still prefers stars that have more points, which of the two options should she choose? Here, similarity comes to the rescue. The star on the left is the same in both Figures and the star on the right has the same number of points in both Figures. Hence the gap in desirability is the same in both Figures. But even though the difference in utility is the same, there is a much greater overlap of features in the pair shown in Figure 4. In other words, while stars on the right of each Figure lie on the same indifference curve, the stars are much ‘closer’ to each other in Figure 4. This clearly makes it easier to discriminate among the options in Figure 4. Accordingly, our formal definition says that the pair in Figure 4 is more similar than the pair in Figure 3.

The visual examples above offer a very clear-cut decomposition of utility and similarity in terms of observable attributes. In economic applications, this decomposition will be more involved: every feature of the choice alternatives may contribute to utility, and the modeler may be unable to observe some or even all of the attributes that are relevant for choice. But once revealed preference analysis determines the partition of the alternatives into indifference curves, we are in business: similarity can be identified by the increase in the ability of the decision maker to discriminate a pair of options, even when options are kept within the same indifference curves.
We infer the varying degrees of similarity across pairs of alternatives as a residual: similarity captures the ability of the decision maker to discriminate a pair of alternatives above and beyond what can be explained by differences in utility. In other words, if a set of choice probabilities can be perfectly explained by the utility parameters alone, then all pairs of alternatives are revealed to be equally similar and all similarity parameters can be taken equal to zero. In this sense, our behavioral notion of similarity is orthogonal to utility.

To link our behavioral definition of 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 they very distant in the space of relevant attributes. Now suppose that computer $C$ shares many features with computer $B$, but is clearly inferior. 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 $C$ lies on an inferior indifference curve, choosing $C$ when any of the other options are available would be a mistake. Even with a limited amount of information about computers, the decision maker should be less likely to make a mistake when choosing from the pair $\{B, C\}$ than from the pair $\{A, C\}$; after all, $B$ clearly dominates $C$. As in the previous visual examples of Figures 1 to 4, this pattern of choices reveals that the pair $\{B, C\}$ is more similar than the pair $\{A, C\}$. In other words, 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 $p_t^{\mu_\sigma}$. Then $(i, j) \succsim (k, \ell)$ if and only if $\mu_k = \mu_i > \mu_j = \mu_\ell$ and $\sigma_{ij} \geq \sigma_{k\ell}$.

Since the definition of revealed similarity requires condition (2), for some pairs of alternatives we may have neither $(i, j) \succsim (k, \ell)$ nor $(k, \ell) \succsim (i, j)$. Binary choices will
only reveal the full similarity ranking if we introduce further assumptions about the richness of the set of choice alternatives. Alternatively, the complete and transitive similarity ranking can be identified from choices over menus of three alternatives. We will follow this last approach in Section 7. With either approach (using binary choice data under further richness assumptions; or using ternary choice data) the full similarity relation is represented by the correlation parameters $\sigma_{ij}$. Henceforward, we refer to the parameter $\sigma_{ij}$ as the similarity of pair $i,j$.

A quick comparison of Figures 3 and 4 illustrates the following feature of similarity: while utility is learned gradually as the level of familiarity with the options increases, similarity is perceived practically 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 known from the beginning of the choice process.\(^3\)

### 5 Similarity Puzzle

This section contains the main results. Throughout the section $\succ$ denotes the revealed preference relation and $\succcurlyeq$ denotes the revealed similarity relation corresponding to the Bayesian Probit process $(\rho^\mu_t, \rho^\sigma_t)_{t>0}$.

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

\[
\frac{\rho^\mu_t(1, \{1, 2, 3\})}{\rho^\mu_t(2, \{1, 2, 3\})} < \frac{\rho^\mu_t(1, \{1, 2\})}{\rho^\mu_t(2, \{1, 2\})}.
\]

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 $X_i = \mu_i + \varepsilon_i$ where $\varepsilon_i$ is normally distributed with mean zero and variance $1/t$. Hence early in the learning process, when $t$ is small, the variance

---

\(^3\)As we pointed out at the end of Section 3 there is also a second, mathematical justification for the assumption of known $\sigma$: since Brownian motion accumulates quadratic variation, $\sigma$ can be perfectly estimated after observing the continuous signal process $X(t)$ for an arbitrarily small time interval.
(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

$$X_i - X_j = \mu_i - \mu_j + (\varepsilon_i - \varepsilon_j)$$

where the constant $\mu_i - \mu_j$ is the true difference in utility and the error term $\varepsilon_i - \varepsilon_j$ 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 makes the alternatives with higher similarity stand out. According to the prior, every strict ranking of the alternatives has the same probability:

- $1 \succ 2 \succ 3$
- $2 \succ 1 \succ 3$
- $2 \succ 3 \succ 1$
- $3 \succ 1 \succ 2$
- $3 \succ 2 \succ 1$

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 what 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,

$$\frac{\rho_t^{\mu \sigma}(1, \{1, 2, 3\})}{\rho_t^{\mu \sigma}(2, \{1, 2, 3\})} > \frac{\rho_t^{\mu \sigma}(1, \{1, 2\})}{\rho_t^{\mu \sigma}(2, \{1, 2\})}.$$

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 the 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 results will hold 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 the overall amount of information the decision maker has 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, the effect of similarity will only be detectable if alternatives are close in utility. In this case, introducing a new alternative hurts similar alternatives more than dissimilar ones. To illustrate the 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) $\rho_t^{\mu_2} (2, \{1, 2, 3\}), \rho_t^{\mu_3} (3, \{1, 2, 3\}) \in (1/2 - \varepsilon, 1/2]$ for all $t < T$; and

(ii) $\rho_t^{\mu_2} (2, \{1, 2, 3\}), \rho_t^{\mu_3} (3, \{1, 2, 3\}) \in [1/4, 1/4 + \varepsilon)$ for all $t > T'$.

Item (i) is a powerful illustration of how low information and varying degrees of similarity combine to make similar alternatives stand out. 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.

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 (i.e., they are duplicates). In general, 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 causes 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 at the beginning of the choice process, independently of the true value of the alternatives.

Our final 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 initial stand-out effect of similarity can persist indefinitely when alternatives 2 and 3 are similar but not equally desirable.
Proposition 7 (Attraction effect). Let \( 1 \sim 2, (2,3) \sim (1,3) \) 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 that the attraction effect occurs arbitrarily early in the choice process, and lasts indefinitely.

For a concrete illustration, consider the example in Figure 5. It shows the choice process 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 5 shows time 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 faster than any other. 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 5 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 converging to one-half. But the probability of choosing alternative 1 never recovers from the initial disadvantage caused by similarity: 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, the attraction effect never disappears.

Finally 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 5 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.

### 6 Observable Attributes

In experiments and applications, each choice alternative \( i \in A \) is often described by a vector of observable attributes \( x_i = (x_{i1}, x_{i2}, \ldots, x_{im}) \in \mathbb{R}_+^m \). For example, in Simonson (1989) cars are described by ‘ride quality’ and ‘miles per gallon’; in Soltani et al. (2012) subjects choose among simple lotteries \( (m, p) \) described by the value \( m \) of a cash prize and the probability \( p \) of winning the prize. This extra structure allows us to build a model in which utility and similarity depend explicitly on the location of the alternatives in the attribute space. It will also allow us to make explicit the
intuition about how the overlap of attributes facilitates discrimination, and the sense
in which our definition of similarity departs from the everyday use of the word and is
independent of utility.

In order to facilitate estimation procedures in the standard multinomial probit,
Hausman and Wise (1978) propose the following parameterization of the covariance
matrix. Let the random utility for alternative $x_i$ be given by

$$U(x_i) = \beta_1 x_{i1} + \beta_2 x_{i2}$$

where $\beta_1, \beta_2$ are independent and normally distributed with $E[\beta_1] = \bar{\beta}_1$, $E[\beta_2] = \bar{\beta}_2$
and $\text{Var}[\beta_1] = \text{Var}[\beta_2] = \sigma^2 > 0$. This can be interpreted as a consumer with a
linear utility and random weights given to each attribute. Since the same realization
of the weights $\beta_1$ and $\beta_2$ will enter the utility of every choice alternative, utilities are
positively correlated. The correlation between the utilities of a pair of alternatives $i$ and $j$ is given by

$$\text{Corr}(U(x_i), U(x_j)) = \frac{x_{i1}x_{j1}\sigma^2 + x_{i2}x_{j2}\sigma^2}{\sqrt{\sigma^2(x_{i1} + x_{i2})\sqrt{\sigma^2(x_{j1} + x_{j2})}}}
= \frac{x_{i1}x_{j1} + x_{i2}x_{j2}}{||x_i||||x_j||}
= \cos(x_i, x_j)$$

which corresponds to the cosine of the angle formed by the two alternatives in the
two-dimensional attribute space. Figure 6 provides an illustration.

![Figure 6](image)

Figure 6: In Hausman and Wise (1978) the correlation parameter $\sigma_{ij}$ equals the cosine of the angle
formed by $i$ and $j$. 

19
In the standard multinomial probit $\sigma_{ij}$ denotes the correlation in the true value of the alternatives $i$ and $j$. In the Bayesian probit, on the other hand, the true value of the alternatives is uncorrelated (the prior is iid). The parameter $\sigma_{ij}$ instead captures the correlation in the noise of the signals of alternatives $i$ and $j$.

In Section 7 we show that the standard multinomial probit and the Bayesian probit are equivalent when choices are restricted to binary menus. In particular, when only $i$ and $j$ are available the Bayesian decision maker always picks the alternative with the highest signal realization. Hence, in both models the probability that $i$ is chosen out of $\{i, j\}$ is given by
\[
\mathbb{P}(X_i > X_j) = \Phi\left(\frac{\mu_i - \mu_j}{\sqrt{1 - \sigma_{ij}^2}}\sqrt{t/2}\right)
\]
where $\Phi$ denotes the standard normal cdf.

This close connection between the models allows us to use the parametric example of Hausman and Wise (1978) for the Bayesian probit, with minor modifications. The power of the decision maker to discriminate the options in $\{i, j\}$ depends on two independent components: the gap in utility measured by $\mu_i - \mu_k$ and the similarity measured by $\sigma_{ik}$.

For example, suppose that in Figure 6 we move alternative $k$ continuously towards the location of object $i$. How should the ability of the decision maker to discriminate the options in the pair $\{i, k\}$ be affected? The overall effect of moving $k$ closer to $i$ in the attribute space will depend on the relative magnitude of two effects. First, as the distance between $i$ and $k$ in the attribute space decreases, the gap in utility $\mu_i - \mu_k$ will narrow, and the alternatives $i$ and $k$ will be closer to indifference. This makes discrimination harder; the decision maker is more likely to make a mistake when the gap in utility $\mu_i - \mu_k$ is smaller. Second, as the objects get closer in the attribute space, their similarity $\sigma_{ik}$ increases. This makes discrimination easier; for a given gap in utility, the decision maker is less likely to make a mistake when $\sigma_{ik}$ is higher.

Since the similarity of $i$ and $j$ is captured by correlation, two alternatives are maximally similar when they lie on the same ray departing from the origin. In this case they form an angle of zero radians, and the correlation $\sigma_{ij}$ is equal to the cosine of zero, which is one. This captures the intuition that when one alternative clearly dominates another in all attributes, they are easier to discriminate. On the other hand,

---

4While the normally distributed prior in the Bayesian probit necessarily includes negative utility values, all the alternatives in this example have positive utility values. This can be easily fixed by taking logs in the utility specification; we ignore this issue to keep the exposition simple.
if alternatives form an angle close to ninety degrees, the cosine of their angle is close to zero, and noise is uncorrelated. Intuitively, since the alternatives do not share any attributes, it is harder to discriminate them.

The link between the standard multinomial probit and the Bayesian probit breaks down once we introduce menus of three alternatives. In particular, the Bayesian probit allows violations of monotonicity and can account for the attraction and compromise effects. This can again be illustrated for the case of observable attributes using Figure 6.

Suppose for a moment that \(i\) and \(j\) are located on the same indifference curve so that each one is chosen with probability \(1/2\) at any date from the menu \(\{i, j\}\). Alternative \(k\) forms a tighter angle with alternative \(j\) than with alternative \(i\) and therefore \(\sigma_{jk} > \sigma_{ik}\). Moreover, alternative \(k\) is in a lower indifference curve. Proposition 7 says that introducing a sufficiently inferior \(k\) will boost the probability of choosing \(j\) (the attraction effect).

5 Again, taking logs would guarantee that alternative \(k\) can be taken sufficiently inferior.

On the other hand, consider what happens when we introduce \(k'\) to the menu \(\{i, j\}\). Since the angle between \(i\) and \(k'\) is larger than both the angle between \(i\) and \(j\) and the angle between \(j\) and \(k'\), Corollary 6 says that the middle option \(j\) will be chosen more often than \(i\) or \(k\) at the start of the random choice process (the compromise effect). Which alternative is in the middle will depend on the choice context. Hence, while alternative \(j\) is chosen more often from \(\{i, j, k\}\), instead alternative \(i\) will be the chosen more often from \(\{i, j, \ell\}\).

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 variances 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 cases.

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

\[
\bar{\rho}_t^{\mu\sigma}(i, b) = \mathbb{P}\{X_i(t) \geq X_j(t) \text{ for all } j \in b\}
\]

where we write \( \bar{\rho}_t^{\mu\sigma} \) for a naïve probit rule with parameters \( \mu, \sigma \) and \( t \). We refer to the ordered collection of choice rules \( (\bar{\rho}_t^{\mu\sigma})_{t>0} \) as the naïve probit process.

The distribution for the signal \( X(t) \) is the same for the Bayesian and for the naïve probit. But instead of updating beliefs according to Bayes rule and maximizing expected utility according to the posterior, 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.** \( \rho_t^{\mu\sigma}(i, j) = \bar{\rho}_t^{\mu\sigma}(i, j) \) for every \( i, j \in A \) and \( t > 0 \).

When only two options are present, there is only one correlation parameter at play, and 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.

If the models are equivalent when restricted to binary data, 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 \( \rho_0^{\mu\sigma}(i, a) = \lim_{t \to 0^+} \rho_t^{\mu\sigma}(i, a) \) and likewise \( \bar{\rho}_0^{\mu\sigma}(i, a) = \lim_{t \to 0^+} \bar{\rho}_t^{\mu\sigma}(i, a) \).

**Proposition 9** (Closed-form choice probabilities at time zero). Let \( b = \{1, 2, 3\} \). In the limit as \( t \to 0^+ \), the naïve probit rule and the Bayesian probit rule choose alternative 1 with probability

\[
\bar{\rho}_0(1, b) = \frac{1}{4} + \frac{1}{2\pi} \arctan \left( \frac{(1+\sigma_{23}-\sigma_{12}-\sigma_{13})}{\sqrt{4(1-\sigma_{12})(1-\sigma_{13})-(1+\sigma_{23}-\sigma_{12}-\sigma_{13})^2}} \right)
\]

\[
\rho_0(1, b) = \frac{1}{4} + \frac{1}{2\pi} \arctan \left( \frac{(1+\sigma_{12})(1+\sigma_{13})-\sigma_{23}(1+\sigma_{12}+\sigma_{13}+\sigma_{23})}{\sqrt{3+2\sigma_{12}+2\sigma_{13}+2\sigma_{23}+\sigma_{12}^2+\sigma_{13}^2+\sigma_{23}^2}} \right)
\]

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 \( t \) 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 7 illustrates how choices for both the naïve and the Bayesian probit process start at time zero from menu \( b = \{1, 2, 3\} \). The Figure plots \( \rho_t^{\mu}(2, \{1, 2, 3\}) \) and \( \tilde{\rho}_t^{\mu}(2, \{1, 2, 3\}) \) 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 start as a 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 at the very beginning of the choice process (when \( t \to 0 \)) are sufficient to identify the entire ranking of the correlation parameters:

**Proposition 10.** \( \rho_0^{\mu}(i, \{i, j, k\}) \geq \rho_0^{\mu}(j, \{i, j, k\}) \) 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 \( \rho_0^{\mu}(i, \{i, j, k\}) \geq \rho_0^{\mu}(j, \{i, j, k\}) \). This extends the revealed similarity relation defined in Section 4 to a complete and transitive binary relation represented by the correlation coefficients: \((i, j) \succ (k, \ell)\) if and only if \( \sigma_{ij} \geq \sigma_{kl} \). While this definition has the advantage of generating a complete and transitive binary relation, the definition based on binary choices given in Section 4 is more intuitive. In any case, Proposition 10 justifies calling 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 the amount of information \( t \) grows arbitrarily large. The next Proposition shows more: 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 \( t \to \infty \). Note that
Proposition 11. \( \lim_{t \to \infty} [\rho_t^{i,b} - \tilde{\rho}_t^{i,b}] = 0 \) for all \( i \in b \in \mathcal{A} \).

Putting together 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:

\[
\lim_{t \to \infty} \rho_t^{i,\{1,2,3\}} = \tilde{\rho}_0^{i,\{1,2,3\}}
\]

for all \( i = 1,2,3 \). 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 alternatives are equally desirable, ties are not broken uniformly: Proposition 9 shows the choice probabilities for 1, 2 and 3 depend on the similarity parameters.

8 Conclusion

We proposed 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 explain discrimination. In the Bayesian probit, how well the decision maker is able to discriminate among two alternatives depends on 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 is sufficiently high, Tversky’s similarity hypothesis holds.

The Bayesian probit allowed us to explain how the attraction effect may arise from rational, Bayesian learning. We showed that when the new alternative \( k \) is similar to
Figure 7: Probability of alternative 2 being chosen from menu $b = \{1, 2, 3\}$ as a function of the correlation $\sigma_{23}$ between the utility signals of alternatives 2 and 3, in the limit as $t$ is taken arbitrarily close to zero. The solid line corresponds to the Bayesian probit. The dashed line corresponds to the naïve probit. The correlations $\sigma_{12}$ and $\sigma_{13}$ are fixed equal to zero. By symmetry, an identical graph would describe the choice probabilities for alternative 3 for both models. When $\sigma_{23}$ is zero the utility signals are orthogonal and similarity effects disappear: choices correspond to the uniform distribution for both models, and therefore the two lines meet at exactly one third. As $\sigma_{23}$ increases to one, alternatives 2 and 3 become maximally similar and the naïve probit chooses each of them with probability $1/4$. In contrast, the Bayesian probit chooses alternatives 2 and 3 with probability $1/2$ and alternative 1 with probability zero.

If $j$ and sufficiently inferior, the attraction effect occurs arbitrarily early in the choice process, and it never disappears. We also used the Bayesian probit to explain the tendency of a decision maker to avoid extreme options (the compromise effect).

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 their measurement and provides a useful tool for applications. The model is now ripe for applications in which the decision maker decides on the optimal amount of information to be obtained before a choice is made, 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. An application of Bayes’ rule 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:

\[
\begin{align*}
    s(t) &= \left[ s(0)^{-1} + t(\Lambda \Lambda')^{-1} \right]^{-1} \\
    m(t) &= s(t) \left[ s(0)^{-1} m(0) + (\Lambda \Lambda')^{-1} X(t) \right]
\end{align*}
\]

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

\[
\begin{align*}
    \mathbb{E}[m(t)] &= s(t) \left[ s(0)^{-1} m(0) + (\Lambda \Lambda')^{-1} \mu \right] \\
    \text{Var}[m(t)] &= ts(t)(\Lambda \Lambda')^{-1} s(t)'
\end{align*}
\]

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

\[
\begin{align*}
    \left( \frac{ts_0^2(\mu_2 - \mu_1)}{ts_0^2 + 1 - \sigma_{12}}, \frac{2ts_0^4(1 - \sigma_{12})}{(ts_0^2 + 1 - \sigma_{12})^2} \right)
\end{align*}
\]

hence

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

where \( \Phi \) denotes the cumulative distribution function of the standard normal distribution. This shows that \( \rho_t^\mu(1, 2) \geq 1/2 \) if and only if \( (\mu_1 - \mu_2) \geq 0 \), proving Proposition 1. It also shows that for fixed values of \( t \) and \( (\mu_1 - \mu_2) \), \( \rho_t^\mu(1, 2) \) 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 \( \rho_t^\mu(i, j) = \mathbb{P}\{m_i(t) > m_j(t)\} = \Phi \left( \frac{\sqrt{t} \left( \frac{\mu_i - \mu_j}{\sqrt{2(1 - \sigma_{12})}} \right)} {\sqrt{2(1 - \sigma_{12})}} \right) = \mathbb{P}\{X_i(t) > X_j(t)\} = \tilde{\rho}_t^\mu(i, j) \).

**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

\[
L_1 = \begin{bmatrix}
-1 & 1 & 0 \\
-1 & 0 & 1
\end{bmatrix}
\]

(4)

so that \( L_1X(t) = (X_2(t) - X_1(t), X_3(t) - X_1(t)) \). Then

\[
\hat{\rho}_t^{\mu\sigma}(1, b) = \mathbb{P}\{X_1(t) > X_2(t) \text{ and } X_1(t) > X_3(t)\} = \mathbb{P}\{L_1X(t) \leq 0\}.
\]

The vector \( L_1X(t) \) is jointly normally distributed with mean \( tL_1\mu \) and covariance \( tL_1\Lambda\Lambda'L_1' \). So \( L_1X(t) \) has the same distribution as \( tL_1\mu + \sqrt{t}MB \), where the vector \( B = (B_1, B_2) \) has a joint standard normal distribution and \( MM' = L_1\Lambda\Lambda'L_1' \) has full rank. Take \( M \) as the Cholesky factorization

\[
M = \begin{bmatrix}
\sqrt{2(1 - \sigma_{12})} & 0 \\
\frac{1+\sigma_{23}-\sigma_{12}-\sigma_{13}}{\sqrt{2(1-\sigma_{12})}} & \sqrt{2(1 - \sigma_{12}) - \frac{(1+\sigma_{23}-\sigma_{12}-\sigma_{13})^2}{2(1-\sigma_{12})}}
\end{bmatrix}
\]

(5)

then we can write for each \( t > 0, \)

\[
\hat{\rho}_t^{\mu\sigma}(1, b) = \mathbb{P}\{L_1X(t) \leq 0\}
= \mathbb{P}\{tL_1\mu + \sqrt{t}MB \leq 0\}
= \mathbb{P}\{MB \leq -\sqrt{t}L_1\mu\}
\]

and taking \( t \to 0 \) we obtain

\[
\hat{\rho}_0^{\mu\sigma}(1, b) = \mathbb{P}\{MB \leq 0\}
\]

Now \( MB \leq 0 \) if and only if

\[
\begin{cases}
0 \geq B_1\sqrt{2(1 - \sigma_{12})} \\
\text{and} \\
0 \geq \frac{1+\sigma_{23}-\sigma_{12}-\sigma_{13}}{\sqrt{2(1-\sigma_{12})}}B_1 + \sqrt{2(1 - \sigma_{12}) - \frac{(1+\sigma_{23}-\sigma_{12}-\sigma_{13})^2}{2(1-\sigma_{12})}}B_2
\end{cases}
\]

if and only if

\[
\begin{cases}
B_1 \leq 0 \\
\text{and} \\
B_2 \leq -B_1 \frac{(1+\sigma_{23}-\sigma_{12}-\sigma_{13})}{\sqrt{2(1-\sigma_{12})2(1-\sigma_{13})-(1+\sigma_{23}-\sigma_{12}-\sigma_{13})^2}}
\end{cases}
\]

29
which describes a cone in $\mathbb{R}^2$ and, due to the circular symmetry of the standard normal distribution, we have

$$\tilde{\rho}_0^{\mu\sigma}(1, b) = \frac{1}{4} + \frac{1}{2\pi} \arctan \left( \frac{(1 + \sigma_{12} - \sigma_{13})}{\sqrt{2(1 - \sigma_{12})^2(1 - \sigma_{13}) - (1 + \sigma_{23} - \sigma_{12} - \sigma_{13})^2}} \right).$$

with entirely analogous expressions for $\tilde{\rho}_0^{\mu\sigma}(2, b)$ and $\tilde{\rho}_0^{\mu\sigma}(3, b)$.

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 $t > 0$,

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

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

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

where $B = (B_1, B_2)$ is standard normally distributed and $M(t)M(t)' = L_1 s(t) \Lambda\Lambda' s(t)' L_1'$. Note the contrast with the naive 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

$$C_1(t) = s^4 \left[ -2s^8 t^4 (-1 + \sigma_{12}) - 2s^6 t^3 (-4 + 2\sigma_{12}(1 + \sigma_{12}) + \sigma_{13} - \sigma_{23}) \\ -4s^2 t (2 + \sigma_{12}) (-1 + \sigma_{12}^2 + \sigma_{13}^2 - 2\sigma_{12}\sigma_{13}\sigma_{23} + \sigma_{23}^2) \\ -s^4 t^2 \left( -12 + 7\sigma_{12} - 2\sigma_{12}(5 + \sigma_{12}) + \sigma_{13}^2 \right) - 6(1 + 2\sigma_{12})\sigma_{13}\sigma_{23} + (7 + 2\sigma_{12})\sigma_{23}^2 \right]$$

and the expressions for $M_{21}(t)$ and $M_{22}(t)$ are similarly cumbersome and omitted.
Now we can write, for each $t > 0$,
\[
\rho_t(1, b) = \mathbb{P}\{L_1 m(t) \leq 0\} \\
= \mathbb{P}\{L_1 s(t) s(0)^{-1} m(0) + t L_1 s(t)(\Lambda\Lambda')^{-1} \mu + \sqrt{t} M(t) B \leq 0\} \\
= \mathbb{P}\{\sqrt{t} M(t) B \leq -L_1 s(t) s(0)^{-1} m(0) - t L_1 s(t)(\Lambda\Lambda')^{-1} \mu\} \\
= \mathbb{P}\{M(t) B \leq -\frac{1}{\sqrt{t}} L_1 s(t) s(0)^{-1} m(0) - \sqrt{t} L_1 s(t)(\Lambda\Lambda')^{-1} \mu\}
\]

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
\[
\lim_{t \to 0^+} \frac{M_{21}(t)}{M_{22}(t)} = \frac{(1 + \sigma_{12})(1 + \sigma_{13}) - \sigma_{23}(1 + \sigma_{12} + \sigma_{13} + \sigma_{23})}{(3 + 2\sigma_{12} + 2\sigma_{13} + 2\sigma_{23})(1 + 2\sigma_{12}\sigma_{13}\sigma_{23} - \sigma_{12}^2 - \sigma_{13}^2 - \sigma_{23}^2)}.
\]

Proof. Long and cumbersome, omitted. \hfill \square

Using Lemma 12 we have
\[
\rho_0^{\mu\sigma}(1, b) = \lim_{t \to 0^+} \rho_t^{\mu\sigma}(1, b) \\
= \lim_{t \to 0^+} \mathbb{P}\left\{M(t) B \leq -\frac{1}{\sqrt{t}} L_1 s(t) s(0)^{-1} m(0) - \sqrt{t} L_1 s(t)(\Lambda\Lambda')^{-1} \mu\right\} \\
= \mathbb{P}\left\{B_1 \lim_{t \to 0^+} M_{11}(t) \leq 0 \text{ and } B_2 \leq -B_1 \lim_{t \to 0^+} \frac{M_{21}(t)}{M_{22}(t)}\right\} \\
= \mathbb{P}\left\{B_1 \leq 0 \text{ and } B_2 \leq -B_1 \frac{(1 + \sigma_{12})(1 + \sigma_{13}) - \sigma_{23}(1 + \sigma_{12} + \sigma_{13} + \sigma_{23})}{(3 + 2\sigma_{12} + 2\sigma_{13} + 2\sigma_{23})(1 + 2\sigma_{12}\sigma_{13}\sigma_{23} - \sigma_{12}^2 - \sigma_{13}^2 - \sigma_{23}^2)}\right\}
\]
and by the circular symmetry of the standard normal distribution we obtain
\[
\rho_0^{\mu\sigma}(1, b) = \frac{1}{4} + \frac{1}{2\pi} \arctan\left(\frac{(1 + \sigma_{12})(1 + \sigma_{13}) - \sigma_{23}(1 + \sigma_{12} + \sigma_{13} + \sigma_{23})}{(3 + 2\sigma_{12} + 2\sigma_{13} + 2\sigma_{23})(1 + 2\sigma_{12}\sigma_{13}\sigma_{23} - \sigma_{12}^2 - \sigma_{13}^2 - \sigma_{23}^2)}\right)
\]
with analogous expressions for $\rho_0^{\mu\sigma}(2, b)$ and $\rho_0^{\mu\sigma}(3, b)$. \hfill \square

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 naïve 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 $(I + A^{-1})^{-1} = A(A + I)^{-1}$ to write the covariance of posterior beliefs as

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

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

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

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

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 \((I + A)^{-1} = I - (I + A)^{-1}A\) we have

\[
\mathbb{E} \left[ \sqrt{t} m(t) \right] - \sqrt{t} \mu = \frac{1}{s \sqrt{t}} \Sigma \left( \frac{1}{st} \Sigma + I \right)^{-1} m(0) + \\
+ \sqrt{t} \left[ \Sigma \left( I - \left( \frac{1}{st} \Sigma + I \right)^{-1} \frac{1}{ts} \Sigma \right) \Sigma^{-1} - \Sigma I \Sigma^{-1} \right] \mu
\]

where since matrix inversion is continuous in a neighborhood of \(I\), the expression inside the curly brackets converges to

\[
\frac{1}{s} \Sigma \left[ m(0) - \mu \right]
\]
as \(t\) goes to infinity and therefore is bounded. Since the expression in curly brackets is multiplied by \(1/\sqrt{t}\), we have

\[
\lim_{t \to \infty} \mathbb{E} \left[ \sqrt{t} m(t) \right] - \sqrt{t} \mu = 0.
\]

Moreover

\[
\lim_{t \to \infty} \operatorname{Var} \left[ \sqrt{t} m(t) \right] = \lim_{t \to \infty} t \operatorname{Var} \left[ m(t) \right]
\]

\[
= \lim_{t \to \infty} t^2 s(t) \Sigma^{-1} s(t)'
\]

\[
= \lim_{t \to \infty} t^2 \left[ \frac{1}{t} \Sigma \left( \frac{1}{st} \Sigma + I \right)^{-1} \right] \Sigma \left[ \frac{1}{t} \Sigma \left( \frac{1}{st} \Sigma + I \right)^{-1} \right]'
\]

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

\[
= \Sigma
\]
Since the vector $\sqrt{t} m(t)$ is normally distributed for all $t$, and since
\[
\rho_t^{\mu a}(i, a) = \mathbb{P}\left( \bigcap_{j \in a \setminus \{i\}} \{ m_i(t) > m_j(t) \} \right)
= \mathbb{P}\left( \bigcap_{j \in a \setminus \{i\}} \{ \sqrt{t} m_i(t) > \sqrt{t} m_j(t) \} \right)
\]
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, $\rho_t^{\mu a}(i, a) \to \tilde{\rho}_t^{\mu a}(i, a)$ as desired. \[\Box\]

**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 $\mu_1, \mu_2, \mu_3$. \[\Box\]

**Proof of Proposition 4**

By Proposition 9 for the naïve probit we have $\tilde{\rho}_0^{\mu a}(1, \{1, 2, 3\}) > \tilde{\rho}_0^{\mu a}(2, \{1, 2, 3\})$ if and
only if
\[
1 + \sigma_{23} - \sigma_{12} - \sigma_{13} > 1 + \sigma_{13} - \sigma_{12} - \sigma_{23}
\]
\[34\]
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

$$
\bar{\rho}_t^{\mu \sigma}(i, \{1, 2, 3\}) = \mathbb{P}\{X_i(t) \geq X_j(t) \text{ for all } j\}
= \mathbb{P}\{\sqrt{t}[X_i(t) - \mu_i] \geq \sqrt{t}[X_j(t) - \mu_j] \text{ for all } j\}
= \bar{\rho}_0^{\mu \sigma}(i, \{1, 2, 3\})
$$

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 $t > 0$,

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

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

**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 Proposition 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 $\sigma_{12} = \sigma_{23} = \bar{\sigma} < 1$ and take the limit $\sigma_{23} \rightarrow 1$. Then the numerator in (6) goes to $2(1 - \bar{\sigma}) > 0$; in the denominator, the term $(3 + 2\sigma_{12} + 2\sigma_{23} + 2\sigma_{13})$ goes to $(5 + 4\bar{\sigma}) > 0$ 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 \(\rho_1 \sim_2 \rho_3\) and \(\rho_0 \sim_{\{1, 2, 3\}} \rho_0\). The result then follows from Proposition 9 once we show that
\[
\frac{(1 + \sigma_{13} - \sigma_{12} - \sigma_{23})}{\sqrt{4(1 - \sigma_{12})(1 - \sigma_{23}) - (1 + \sigma_{13} - \sigma_{12} - \sigma_{23})}} \to 0.
\]
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. \(\Box\)

**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

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

and

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

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

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 \( 1 + \sigma_{12} + s^2 t > 0 \) 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 \( \rho_t^{\mu_3}(1, \{1, 2, 3\}) \) and \( \rho_t^{\mu_3}(2, \{1, 2, 3\}) \) 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 \( \rho_t^{\mu_3}(2, \{1, 2, 3\}) \) for every \( t > 0 \). Moreover, for fixed \( t > 0 \), \( \rho_t^{\mu_3}(2, \{1, 2, 3\}) \) 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, \( \rho_t^{\mu_3}(2, \{1, 2, 3\}) \) will converge to 1/2 from above, while \( \rho_t^{\mu_3}(1, \{1, 2, 3\}) \) will converge to 1/2 from below. \( \square \)