Avoiding Lemons in Search of Peaches:
Designing Information Provision∗

Pedro M. Gardete  Megan H. Antill

Working Paper
June 2018

Abstract
The increasing amount of data available to consumers has most likely aided in decision-making. However, it has also created an opportunity for sellers to design the information landscape that consumers navigate. This paper develops a novel search model for alternatives with multiple characteristics, and reports estimation results for an online used car seller. The model allows search over alternatives with multiple characteristics with arbitrary marginal distributions and correlation structures. For example, more expensive vehicles may feature fewer past owners, and vehicles with higher mileage may reveal more issues in their inspection reports. The model also allows for a rich set of consumer search behaviors, including (but not limited to) sequential search within vehicles and characteristic-by-characteristic search across. The estimated fundamentals are then used to consider different information design policies. We find that the choice of the characteristics to be made available to consumers upfront has significant economic implications. For example, featuring variance-reducing information upfront (in our application, vehicle histories) instead of other characteristics translates into an approximate conversion rate increase of 20%, in relative terms. In light of our results, we provide intuition on how different information design policies affect consumer and seller welfares. Additional counterfactual analyses confirm our intuition. Finally, we show that a simplified approach based on traditional choice models would produce low quality recommendations about information design policies.

∗Graduate School of Business, Stanford University, Stanford, California 94305. Correspondence: Gardete: gardete@stanford.edu, Antill: mkhunter@stanford.edu.
1 Introduction

In recent years, the increased availability of data surrounding the benefits and characteristics of sellers’ offerings has allowed consumers to make better-informed decisions. Consumers access diagnostic information across a variety of domains, including simple nutrition information of candy bars to the complex neighborhood characteristics of a potential new home. Simultaneously, sellers themselves have made a significant amount of data available to consumers: Real estate websites provide photo galleries and open house scheduling information and car dealership websites provide extensive information about the vehicles available on the lots. The result of these dynamics is a large data ecosystem that consumers seek to browse through efficiently. This, in turn, presents an information design problem for sellers. For example, a website featuring all the available information on its homepage (e.g., a ‘database dump’) is likely to turn most customers away. Instead, information provision must be designed to optimize customer learning: A used car dealership may present vehicle pictures, prices, and mileage on its homepage, and relegate additional information to individual vehicle detail pages. Consumers interested in specific models may then access the corresponding profile pages, and learn additional specific information, like the vehicle’s ownership history, the results of an inspection report, etc.

There is a clear interaction between the design of information provision and consumers’ information-seeking (search) behaviors. The first reason is that search actions are costly for consumers in terms of attention, time, and possibly money. Clearly, how accessible a piece of information is to a consumer may affect her search and purchase behaviors. The second reason is related to the fact that product characteristics are often correlated. For example, a used car, already sold multiple times, may feature higher mileage, may be older, and may be more likely to have been involved in an accident than a one-owner car; a newer vehicle may be priced higher by the seller, and it may have also fared better through inspection. These correlations can result in complex search behaviors, as consumers understand that learning good news about a given product attribute may mean good or bad news about the remaining ones.

This paper offers two main contributions. First, it advances the search literature by allowing for a rich characterization of consumer search behaviors over alternatives with multiple characteristics, each of which can be learned about. Moreover, the characteristics are
allowed to be correlated, such that learning about a given characteristic may affect beliefs and subsequent search decisions towards others. In addition to the observable characteristics consumers may search over, the model also allows for search over an unobserved utility component, which may be arbitrarily correlated with the levels of the observed characteristics. Our model leverages consumers’ search behaviors to recover this correlation, and deliver consistent estimates of the preference parameters. Identification stems from state-dependent search patterns, that vary according to the observable characteristics of the vehicles. Another unique characteristic of our model is that it allows for arbitrary marginal distributions (e.g., discrete and/or continuous; parametric and/or non-parametric) over product features. For example, consider a setting where prices are known to follow a log-normal distribution to the researcher, whereas other characteristics, like weight or color, may feature unique “empirical” distributions. The model relates the marginal distributions by implementing a statistical copula, which captures cross-characteristic relationships flexibly. The fourth distinguishing feature of the model vis-à-vis the current literature, is that it allows for consumer search to take place in a piecemeal fashion across alternatives: A consumer may search over the characteristics of offer 1 first, and then move on to search those of offer 2. Alternatively, she might instead search characteristic by characteristic, inspecting first a characteristic of products 1 and 2, and then move on to compare additional characteristics. The model is fully flexible, in that it allows for any search sequence over alternatives and their characteristics.

The second main contribution of this paper is to use the estimated search model to inform the design of information provision of the firm. Given the large data ecosystem available to sellers, a potentially important decision is which information to make immediately available to consumers, and which information to relegate to be potentially searched over. Consider, for example, the decision of moving information about a product characteristic from the main listing page down the page hierarchy. The reason the model described above is useful to analyze different information scenarios is that it takes into account that changes to the information disclosure policy affect consumers’ subsequent search decisions. After recovering the fundamental search and preference parameters, the model is used to consider counterfactual consumer search outcomes across different information disclosure scenarios. The model is also useful to shed light on how the information frictions affect market outcomes. For example, the presence of search costs may lead some vehicles to be searched more frequently than others. Specifically, outlier alternatives are likely to be searched inefficiently, because
their correlations over characteristics are less aligned with consumers’ beliefs. The model can be used to identify the specific dimensions of such outliers, and to suggest measures to alleviate search frictions.

In terms of the results, we first recover preference coefficients consistent with general expectations, in that consumers prefer lower prices, more recent cars with fewer miles, fewer past owners, etc. We identify two very different consumer segments. The first faces positive search costs in relation to all search activities. The second segment exhibits a negative search costs for browsing vehicle pictures, and very high search costs for the remaining activities. One explanation for this is that the second segment seems more attuned to extracting hedonic utility from browsing vehicle photos, and may be less interested in the informational content revealed by search activities. We also recover a positive cost related to switching focus between vehicle detail pages, suggesting that consumers are more likely to take within-vehicle search actions than proceeding by comparing alternatives one characteristic at a time.

We find non-zero estimated correlations between the observable characteristics and the unobserved utility component. The unobserved component is found to be strongly negatively associated with price, and strongly positively associated with vehicle age. One possible explanation is that the unobserved component summarizes consumers’ hopes of finding silver linings in relatively older, low-priced vehicles. We also find that the cross-correlations affect search patterns, but have less influence on actual conversion rates.

In terms of the counterfactual results, we find that eliminating search costs would increase sales from a predicted level of 4.6% to 5.4%. We also find positive predictions in terms of the effect of the design of the information disclosure policy on conversion rates. The predicted conversion rate increases in the range of 0.9% to 1.2% (absolute magnitudes) when information on the vehicle histories (specifically, the number of accidents and of owners) is provided, while relegating one of the characteristics currently disclosed on the main listing page (price, mileage, and year) to the vehicle detail page. Additional analyses reveal that the seller benefits from providing information on the most diagnostic characteristics upfront, and requiring the remaining ones to be searched over. Finally, we compare our results to a relatively simpler approach, based on traditional choice models and residual utility variances. We find that the simpler approach performs poorly, which highlights the benefit of accounting for search behaviors formally when considering decisions about the design of information presentation.
Albeit relatively recent to the marketing setting, the literature on consumer search is already extensive and diverse. The seminal methods and insights of consumer search have been established by both theoretical and empirical literature streams. Montgomery, Li, Srinivasan, and Liechty (2004) employ a dynamic multinomial probit model to characterize consumer online browsing for books, and find that first-order Markov approximations perform poorly in explaining the patterns in the dataset. Their finding highlights the need to consider at least second-order moments during estimation of search processes. The results by Koulayev (2013) alleviate this result, by proposing a Dirichlet-based model in which partial moment data can still be useful to estimate search models. Kim, Albuquerque, and Bronnenberg (2010) develop an empirical application of search based on Weitzman (1979), and use camcorder browsing behavior to estimate the search model. They also use the model to conduct welfare counterfactual analysis related to the availability of recommendation systems. Branco, Sun, and Villas-Boas (2012) characterize a search setting in which consumers can acquire information gradually, across multiple characteristics. Their model also takes into account the fact that the seller may choose price strategically to influence consumer search. More recently, Ke, Shen, and Villas-Boas (2016) characterize continuous-time search policies when consumers are faced with multiple products. Finally, Kim, Albuquerque, and Bronnenberg (2016) derive a probit-based search estimation approach, of particular benefit for settings with large search sets. Our model is tightly related to the work mentioned above. This paper can be thought of as advancing the empirical work on consumer search by incorporating gradual information acquisition. In our model, as idealized by the theoretical literature mentioned above, consumers can engage in piecemeal information acquisition about product characteristics.

In addition to the model-based insights presented by the work cited above, there has been a productive effort in describing consumer search patterns more flexibly. Johnson, Moe, Fader, Bellman, and Lohse (2004) document relatively short search sequences on websites selling books, c.d.'s and travel products, and also that consumer experience may lead to less informative search. See also Ke and Villas-Boas (2018) for a complete characterization of search across multiple alternatives, where search can be decreasingly informative.

1Although not always mentioned, the insights and advances available in the search literature are quite related to those of the learning literature (see, for example, Erdem and Keane (1996); Erdem (1998); Erdem, Imai, and Keane (2003)), in which consumers make purchase decisions while they are learning about product characteristics/benefits. In this case, search is not geared towards a unique purchase, but it is performed concurrently with purchase decisions. See also Dickstein (2018) for a recent example of learning whilst consuming.
search over time. Bronnenberg, Kim, and Mela (2016) analyze consumer search behavior for cameras, and find that consumers tend to focus on a small set of attributes, while also documenting state-dependence across searched attributes. Their findings also speak to those by Johnson, Moe, Fader, Bellman, and Lohse (2004), in that the chosen alternative is often first visited near the end of the search sequence. By allowing for arbitrary consumer search paths, our model can rationalize the state-dependence found in this literature, as well as the narrow consumer focus on a few attributes.³

There exist additional ways in which consumer search is relevant to firms. Seiler (2013) finds that 70% of consumers are unaware of the price of detergent, and that price promotions can increase consumers’ incentives to search. Also in a physical retail setting, Elberg, Gardete, Macera, and Noton (2017) use search behavior to characterize consumers’ sensitivities to future deals, induced by present ones. They go on to consider the competitive implications of such behavior to firms.⁴ The counterfactual policies we consider are different from the ones in the literature. In particular, our model is used to consider the information design problem of the seller, which can affect consumer search processes.

Characterizing the full dynamic problem in consumer search problems is currently unrealistic. This has led most of the literature to model search for a parsimonious utility component, usually assumed to be uncorrelated with the remaining product characteristics. Actual search for multiple characteristics induces a dynamic problem, which is most likely overly complex, in relation to how consumers actually search for information. In this paper we consider search through steepest-ascent, such that consumers behave much like standard numerical solvers: they consider the expected immediate benefits of all of the potential actions at their current information set. They are also able to exit search at any point, either by not purchasing or by converting to one of the alternatives. The steepest ascent method, also referred to as the knowledge gradient, has been shown to perform nearly- or exactly-

³Consumer search patterns can also be used to recover fundamental parameters. By adding formal structure to the search processes, or by considering additional sources of data, the work by Hong and Shum (2006), Honka (2014), Koulayev (2014), De Los Santos (2016), and Seiler and Pinna (2017) provides search cost estimates for consumers browsing for a variety of products, sold online and offline. Our model is also used to recover search costs. In our case, we recover the costs associated with the multiple search actions available to consumers, as well as the cost associated with switching the search focus across alternatives.

⁴Search behaviors have also been characterized in situations where users have access to ‘refinements’: sorting and filtering tools that allow searching for alternatives more efficiently. Chen and Yao (2016) find that refinement tools induce more search, have positive effects on consumer utility, and lead to less concentrated market structures. De Los Santos and Koulayev (2017) propose a method to design product rankings to be used by consumers who have access to search refinements.
optimally in theoretical analyses (e.g., Frazier, Powell, and Dayanik (2009) and Liang, Mu, and Syrgkanis (2017)), and has been documented to perform well across diverse empirical learning contexts (see Powell (2010)). Other search-related papers have also employed heuristic algorithms, sometimes suggesting that these may be more realistic than considering fully forward-looking consumers (e.g., Hodgson and Lewis (2017) and Dickstein (2018)). Research on the identification between specific search models includes De los Santos, Hortaçsu, and Wildenbeest (2012) and Honka and Chintagunta (2016), while the more general problem of identifying the actual underlying search processes remains open.

As already mentioned, a standard assumption held in most of the work in the search literature is that consumers face a set of uncorrelated characteristics/alternatives. This assumption alleviates the estimation burden, because information acquisition no longer requires completely recomputing the search problem consumers face, at each information set. Some exceptions include the work by Adam (2001), who considers the problem of an agent who faces a set of alternatives that may be correlated in an ex ante unknown way. In this case, searching an alternative yields information not only about that same alternative, but potentially also about some of the remaining ones. Santos, Hortaçsu, and Wildenbeest (2017) allow consumers to learn about their own utility distribution, and recover search cost bounds that can rationalize the observed online browsing and purchasing behavior. Hodgson and Lewis (2017) allow the inspection of a product to be informative about the utility that may be derived from inspecting others, and document that, because of learning, an initial bad product shown to a consumer can lead to ending search earlier. Our model also allows for learning, which follows from allowing for multiple observable characteristics, which may be related in arbitrary ways, as captured by their empirical joint distribution. Moreover, our model incorporates search over an unobservable characteristic, itself also potentially related to the observable ones. In our case, the information correlation structure induces learning, albeit in a different than in the papers cited above. In our case, learning about a characteristic of an alternative is informative about the remaining characteristics of that alternative (within-alternative correlation); whereas in the literature mentioned above, uncertainty about one’s own preferences induces cross-alternative correlations. Our model applies to contexts where consumers already understand their preferences well, and are mostly uncertain about different product attribute levels, whereas the assumptions in Santos, Hortaçsu, and Wildenbeest (2017) and Hodgson and Lewis (2017) apply to situations where consumers are first and
foremost uncertain about their own preferences, with incremental information acquisition being costless along characteristics of each alternative.\textsuperscript{5}

The next section describes our context and the dataset. Section 3 describes the search model as well as the steps of the empirical analysis, including consumer beliefs, identification and estimation details. Sections 4 and 5 present the empirical results and the counterfactual analyses, respectively, and Section 6 concludes.

2 Data

2.1 Browsing Behavior

Our dataset comprises all online browsing activity on the website of a North American used car dealership, between February and September 2016. The dealership operates in a number of geographic markets, listing more than 4,000 vehicles during the sample period. Upon arrival to the website, the user can click through a number of filters in order to focus on the vehicles of interest. The website lists a number of vehicle details. First, on the main listing page, each vehicle photo is accompanied by make/model information, year, price, and mileage data. Additional information is available in each vehicle’s detail page: Users can 1) browse through pictures, 2) access the report of the dealer’s vehicle inspection results, and 3) access the vehicle’s history, which includes additional information such as the number of previous owners and the number of accidents the vehicle has been involved in. Accessing each of these information opportunities (pictures, inspection reports, and vehicle histories) requires deliberate action by consumers. The dataset collects all such browsing information, including all user sessions, webpage visits, and click actions.\textsuperscript{6} Although vehicles are not sold directly online, the dealer uses a differentiated business model that provides a primary conversion variable. Upon settling on a vehicle, consumers can order a test drive by filling out a form with the relevant information. The company then follows up to confirm the

\textsuperscript{5}See also Kamenica (2008); Guo and Zhang (2012); Cao and Zhang (2017); Gardete and Guo (2018) for settings in which consumers learn about their own preferences, or about the fit between their preferences with the product characteristics.

\textsuperscript{6}At the time of our sample, all information-disclosing actions involved either clicks, when accessed via computer, or taps and swipes, when accessed through mobile and tablet devices. We restrict ourselves to browsing sessions via computer platforms since mobile and table swipes were found not to be consistently captured in the dataset. Moreover, we also eliminated consumers whose modal device was a phone or a tablet. We also conduct several data cleaning activities in order to eliminate ‘bouncing’ behavior and other noisy online activity, as we describe in the appendix.
time and location of the test drive, usually taking place near a location convenient to the customer. Successful test drive appointments are available in the dataset and are used to characterize consumer conversion.\(^7\) Given the broad product line offered by the dealer, we focus on the clickstream data related to browsing activity of sedan vehicles.\(^8\)

Table 1 presents descriptive statistics of the dataset, comprising information about the search behavior of 24,116 users on the website. The descriptive statistics point to two main patterns. First, search activity can take place over long periods of time. For example, although each user browsed through an average of 2.28 vehicle profile pages, it is not unusual to find users who browse tens or hundreds of vehicles. This is expected, given the complexity of the product involved, as well as its cost. Search activity is divided in browsing sessions, defined as sets of events taking place within sequential intervals of 30 minutes. On the high end, some consumers went through more than 700 sessions, spread across three quarters of a year.

The other insight taken from the descriptive numbers is that search behaviors can vary wildly, per inspection of the ‘Min’ and ‘Max’ columns of Table 1. While some users did not browse more than 1 vehicle, others browsed more than 100. The remaining statistics also indicate significant reasonable variation in search patterns.

Clearly, one of the key dependent variables in the dataset is whether consumers end up converting. It is natural to expect that consumers who order test drives behave differently during their search than the ones who do not. In order to assess this possibility, Figure 1 depicts the kernel densities of average session times across converters and non-converters. Across a relatively long tail, we find converters engage in longer browsing sessions, whereas non-converters concentrate their activities around sessions with durations below 15 minutes.

Figure 2 illustrates the correlational differences in behaviors between the two consumer segments by plotting histograms of four behavioral variables as a function of whether consumers ended up converting. Figures 2i)-iii) reveal relatively consistent patterns: converting

---

\(^7\)Also, the firm provides information on the vehicle prices on its website, which cannot be bargained over. Given the high conversion rates, we only model consumer behavior up to the first conversion. Related to this, we also eliminate redundant search decisions for objective information (i.e., search actions unrelated to vehicle photos). We believe these are related to consumer memory. Although these could be incorporated in the model, they are not the main focus of the analysis. The working assumption for the counterfactual is that consumers will always conduct the necessary searches to refresh their memories, if needed.

\(^8\)The sedan category is the largest one. Search behavior changes significantly depending on the focal category. For example, the sports cars category attracts a high number of users, but displays lower conversion/visit ratios.
customers are more likely to spend more time on the website, to trigger more behavioral events related to information collection, and to visit more vehicle detail pages. Figure 2iv), provides an exception to this ordering, in that converting customers are not always more likely to view more vehicle photos than non-converting customers. Although the previous pattern does hold at the right tail, the left region reveals that non-converters are much more likely to engage in an intermediate level of photo browsing activity. In other words, compared to non-converters, converting customers seem to opt for one of two strategies: either browse few pictures (e.g., less than 20) to decide whether to test drive a vehicle, or to perform extensive search.

These behavioral differences are likely to be driven by a number of factors. First, not all makes and models are stocked to the same extent, and consumers may be heterogeneous in terms of the vehicles that they like. In this case, those consumers who like well-stocked vehicles will have more attractive options to browse through. Another factor is that consumers may also be heterogeneous in terms of their opportunity costs of time and/or need for a vehicle. This would explain why, for the same consideration set, different consumers would exhibit search sequences of varying lengths.

The model captures the heterogeneity in search patterns in two ways. First, we restrict each consumer’s consideration set to the set of vehicles that was ever accessed by each individual. This means that any search action surrounding a vehicle, including browsing a single profile picture for example, implies that the vehicle belongs to the consumer’s consideration set. This assumption also allows us to abstract from the specific filters and searches consumers may have employed to initially discover the vehicles they are interested in. The implication for the counterfactual analysis is that consumers maintain their consideration sets constant, although they remain able to search freely across vehicles in those sets. We believe this is reasonable assumption, given that the dealer’s homepage is well designed in terms of providing customers with vehicle availability. However, understanding each vehicle’s specifications requires additional search efforts. The second way the model accounts for heterogeneity in search patterns is by incorporating a latent class distribution over search cost levels, such that some consumers may be willing to search certain aspects more than others.

Figure 3 characterizes the consumers’ search processes via the analysis of most frequent first-order transitions. Consumers exhibit high switching behavior across vehicles. The
patterns suggest two high-probability paths. First, consumers seem to initiate their vehicle browsing behavior by inspecting its photos. To see this, note that all search actions of vehicle \( j \) are highly likely to be followed by an inspection of a different vehicle’s photos. Second, we also observe consumers transitioning from one search action of vehicle \( j \) to the same search action in vehicle \( -j \). This suggests some users may engage in side-by-side comparisons, by maintaining multiple browser tabs open, for example.

From the descriptive statistics, we have learned that consumer search paths can be quite heterogeneous and often quite long. One implication is that it is important to model all information acquisition steps when characterizing search behavior. This is further emphasized by the fact that converters exhibit longer search activity paths than non-converters. Another implication, from the transition analysis, is that consumers often switch across vehicles, engaging in ‘non-linear’ search. They may start by inspecting a few aspects of vehicle \( j \), for example, then inspect other aspects of vehicle \( -j \), and then return to vehicle \( j \). Alternatively, they may also compare features sequentially across vehicles, followed by a purchase decision. These patterns highlight the importance of having a model that is able to explain information-gathering paths flexibly.

2.2 Vehicle Characteristics

We now turn our attention to the characteristic space of the 1,573 sedan vehicles in the dataset. Each vehicle is described by its make, model, color, price, mileage, number of inspection notes, age, number of accidents, and number of previous owners.\(^9\)

Figure 4 depicts the histograms and cross-scatter plots of vehicle characteristics, excluding the categorical variables make, model, and color. The main diagonal reveals that vehicles are relatively heterogeneous in terms of their characteristics. Moreover, characteristics follow very different marginal distributions. For example, price and mileage are continuous variables and exhibit relatively non-normal distributions. The remaining characteristics are countable, with seemingly different distributional properties as well: The number of previous owners starts at 1 and is strictly decreasing, whereas the number of inspection notes and the vehicle’s age (in years) are non-monotonic.

Table 2 depicts the correlation matrix across the (non-categorical) characteristics. With the exception of price, all characteristics are positively correlated, which is not surprising

---

\(^9\)Inspection reports feature one inspection note for each issue found by the dealer.
since they are all likely to be associated with the vehicle’s use, for example. The negative
correlation between price and these characteristics is also expected: It is an indication that
the seller prices the vehicles according to the appeal of their characteristics. While the actual
relationships across variables are non-linear, the results in Table 2 are in line with Figure 4. These analyses stress the need to allow for flexible marginal distributions of the vehicle
characteristics as well as to allow for flexible relationships across vehicle characteristics. As
we discuss later, we introduce a flexible method to fit the multivariate distribution of the
vehicles’ characteristics, while including the additional effects of make, model and color.

2.3 Preference Weights

In order to briefly characterize consumer preferences, we regress consumers’ conversions on
vehicle characteristics by use of a logit model. Table 3 summarizes the results. As expected,
consumers dislike increments to all vehicle characteristics. All results are highly significant,
except for the number of previous accidents a vehicle suffered in the past, whose parameter is
only significant at the 10% level. The second column presents the marginal effects measured
in odds ratios, which allows for a clear interpretation of the relative magnitude of the results.
Keeping the remaining characteristics constant, the model implies that a price increase of
$10,000 makes a vehicle’s purchase probability decrease to approximately a fourth of its initial
value, and an increase in 100,000 miles translates into the probability of a sale decreasing
approximately to a third. Affecting the remaining characteristics, by changing a vehicle’s
age by one year for example, decreases purchase rates to about 90% of the initial values.

While the results are economically significant, they translate into relatively low elas-
ticities. There may be two underlying reasons. First, an estimated low elasticity may be
partially induced by the fact that some consumers would not have been actually exposed
to all characteristics changes, due to search frictions. The logit model assumes perfect in-
formation, however, and explains the low response through low elasticity estimates. The
second reason affecting these estimates is the impact of unobserved utility factors, which
may be correlated with the observed vehicle characteristics. We later detail that considering
consumers’ search behaviors allows us to incorporate the correlations between unobserved
utility that consumers search over with the observable vehicle characteristics. For example,
a positive correlation between price and the unobserved component will likely bias price sen-
sitivity estimates toward zero. The search model can characterize the propensity of search
for the unobserved utility component conditional on vehicles’ price levels and remaining characteristics. In this way, search behavior is used to recover consistent estimates, through estimation of the correlations between the observable components and the unobserved one.

3 Model

3.1 Preliminaries

As in most discrete choice applications, we assume consumers’ derive linear utility from the characteristics of the alternatives. Under perfect information, consumer $i$’s indirect utility for purchasing vehicle $j$ is equal to

$$v_{ijt} = \sum_{k=1}^{K} \beta_k x_{jk} + \nu_{ij} + \epsilon_{ijt}$$

(1)

and earns utility

$$v_{i0t} = \epsilon_{i0t}$$

(2)

is she decides for the outside option. Above, $j$ indexes vehicles and $k$ indexes observable (by the econometrician) vehicle characteristics. Component $\nu_{ij}$ is an idiosyncratic preference for vehicle $j$, and $\epsilon_{ijt}$ is a preference shock, learned by the consumer at time $t$. Both terms $\nu_{ij}$ and $\epsilon_{ijt}$ are unknown to the econometrician, and the preference shock $\epsilon_{ijt}$ is uncorrelated across consumers, vehicles and time. Except for $\epsilon_{ijt}$, all vehicle characteristics may be correlated. Characteristics $\{x_{jk}\}$ are constant across consumers for the same vehicle (e.g., price, mileage, etc), and are likely to be correlated with each other. For example, as discussed in the descriptive analysis, we expect an older vehicle to have had more owners and be sold for less. These characteristics may also be correlated with the idiosyncratic characteristic $\nu_{ij}$ which, unlike $\{x_{jk}\}$, may be different across consumers. The introduction of term $\nu_{ij}$ captures the fact that, during search, consumers may learn more than the objective characteristics that the econometrician is able to observe directly in the dataset. Including term $\nu_{ij}$ takes into account that some actions reveal subjective information that is not only challenging to quantify, but also whose appreciation may vary significantly across users.

The website reveals some vehicle characteristics on its main listing page, including a picture, price, mileage, age, and make/model information of each vehicle. We include these
characteristics in the consumers’ initial information set. Because these specific characteristics are extremely cheap to observe, search for information is modeled over and above these elements, as we now explain.

Table 4 organizes the vehicle characteristics and presents the search action correspondence that allow users to discover specific characteristics. We assume consumers know their own preference parameters, as well as the vehicle characteristics included in the initial information set. Users can then learn additional vehicle characteristics by taking search actions with regards to each vehicle. Inspection of a vehicles’ histories reveals both the number of previous owners of the vehicle and the number of accidents the vehicle was involved in. Accessing inspection reports reveals information about the number of issues identified during inspection. Finally, browsing through vehicle photos informs users of the utility component $\nu_{ij}$.10

3.2 Distribution of Characteristics and Consumer Beliefs

We assume consumers hold beliefs consistent with the distribution of characteristics observed in the data. Let $X = [X_1..X_K]$ be the random vector of vehicle characteristics with realizations observable both to the consumer and to the researcher (i.e., all characteristics excluding $\nu_{ij}$ and $\epsilon_{ijt}$). We denote the multivariate cumulative distribution (c.d.f.) of these characteristics by

$$G_{\text{make}_\text{-model},\text{price},\text{mileage},\text{age},\text{color},\text{owners},\text{accidents},\text{notes}}$$

(3)

In order to estimate this distribution while preserving the different natures of the marginal distributions as well as the cross-correlation patterns, we employ a statistical copula. Sklar (1959) showed that all multivariate c.d.f.’s $F_{X_1..X_K}$ admit copula representations

$$C(F_{X_1}(x_1), ..., F_{X_K}(x_K))$$

(4)

where each $F_{X_k}(\cdot)$ is the marginal C.D.F. of variable $X_k$. Taking advantage of Sklar’s result is, however, complicated by two factors. First, variables such as color and make/model are categorical discrete r.v.’s. Given their lack of ordinality, cross-correlations are relatively

---

10We discretize the photo-browsing activity for the following rationale. Theoretically, the act of browsing each photo adds information to the consumer’s posterior utility about the focal vehicle. In practice, the resulting search sequences can become extremely long, increasing computation requirements to attain precise calculation of the likelihood function.
meaningless for these variables. To see this, suppose $X_1 \leq 0.2$ were to define a vehicle as yellow, and $X_1 \in (0.2,0.3)$ meant that the vehicle was white. It follows that proper estimation of a copula parameter between r.v.'s $X_1$ and $X_2$ ($X_2$ could represent mileage, for example) would require an unrealistically high degree of flexibility in order to capture the lack of ordinality in $X_1$. The second challenge is that the unobserved preference shock $\nu_{ij}$ has yet to be incorporated into this framework.

We start by addressing the first issue. Estimating the multivariate distribution with categorical variables could, in principle, be implemented by estimating a separate distribution for each level of variables \{make\_model, color\}. However, this method is cumbersome and imposes large data requirements. Because copula estimation can rely on parametric or non-parametric estimation of the marginal distributions, we employ a hybrid approach in that we use a parametric function to parse out the effects of color and make/model, and then use the empirical distribution of the residuals to estimate the empirical marginal distribution of the remaining attributes. We start by explaining c.d.f.

$$F_{\text{price},\text{mileage},\text{age},\text{owners},\text{accidents},\text{notes}|\text{make\_model},\text{color}}$$

which, as we now discuss, operates on different random variables than $G$.

Consider, for example, the conditional distribution of price. We assume the price of a vehicle is related to its color and make/model according to equation

$$p_j = \alpha_0 D_{\text{color}} j + \alpha_1 D_{\text{make\_model}} j + \varepsilon^p_j, \varepsilon^p_j \sim N(0,\sigma_{\varepsilon^p})$$

where $\alpha_0$ and $\alpha_1$ are parameters specific to the linkage between price with color and make/model, respectively, and $\varepsilon^p_j$ is a shock specific to vehicle price. Rather than introducing the empirical c.d.f. for price $p_j$ in copula $C$, we instead utilize the empirical distribution of residual $\tilde{\varepsilon}^p_j$, which is stripped of the effect of color and make/model. Moreover, note that correlations between shocks $\varepsilon^p_j$ and $\varepsilon^m_i$ are also uncontaminated by the influences of the conditioning variables. The estimates for these shocks are readily available by the following procedure

$$\tilde{\varepsilon}^p_j = p_j - E(p_j|D_{\text{color}} j, D_{\text{make\_model}} j)$$

$$= p_j - (\hat{\alpha}_0 D_{\text{color}} j + \hat{\alpha}_1 D_{\text{make\_model}} j)$$
and the empirical marginal distributions of these shocks are incorporated in the statistical copula.

While this methodology readily applies to continuous variables, it requires further adaptation for discrete ones. To illustrate the case of ordinal discrete random variables, let for example the number of accidents of vehicle $j$, conditional on its color and make/model, follow a Poisson distribution:

$$Pr \left( accidents_j | color_j, make_model_j \right) = \frac{\left( \lambda \left( x_j^t \beta \right) \right)^{accidents_j} \exp \left\{ -\lambda \left( x_j^t \beta \right) \right\}}{n_{accidents_j}}$$

where $\lambda \left( x_j^t \beta \right) = \exp \left\{ x_j^t \beta \right\} = \exp \left\{ \beta_0 D_{color_j} + \beta_1 D_{make_model_j} \right\}$. Define $\varepsilon_{acc}^j$ as the residual obtained by the difference between the realized value and the conditional expected value of the number of accidents:

$$\varepsilon_{acc}^j = accidents_j - E \left( n_{\_accidents_j} | color_j, make\_model_j \right)$$

$$= accidents_j - \lambda \left( x_j^t \beta \right)$$

Upon knowledge of a vehicle color, make/model information and a draw of $\varepsilon_{acc}^j$, the predicted number of accidents can be reconstructed according to

$$n_{\_accidents_j} | color_j, make\_model_j = \text{round} \left( \lambda \left( x_j^t \beta \right) + \varepsilon_{acc}^j \right)$$

where the rounding operation ensures the left-hand side is an integer. Error $\varepsilon_{acc}^j$ is continuous, given the continuity of $\lambda \left( x_j^t \beta \right)$, and addresses the challenge of fitting statistical copulas to discrete data. As before, error $\varepsilon_{acc}^j$ is likely to be correlated with the remaining variables, albeit not through the conditioned ones.

When the residuals above are evaluated at their empirical c.d.f.’s, they induce random

---

11 We found that the rounding operator performed well in maintaining the moments of $n_{\_accidents_j}$ similar to the ones of $n_{accidents_j}$, as well as for the remaining discrete variables. Simulation results are available from the authors.
variables \( u = [u_1...u_K]' \). These are introduced into the Gaussian copula according to

\[
c_\Sigma(u) = \frac{1}{\sqrt{\text{det} \Sigma}} \exp \left\{ -\frac{1}{2} \begin{pmatrix} \Phi^{-1}(u_1) \\ \vdots \\ \Phi^{-1}(u_K) \end{pmatrix}' . \Sigma^{-1} . \begin{pmatrix} \Phi^{-1}(u_1) \\ \vdots \\ \Phi^{-1}(u_K) \end{pmatrix} \right\}
\]

(11)

where

\[
u = F(\varepsilon) = \begin{pmatrix} F_1(\varepsilon_1) \\ \vdots \\ F_K(\varepsilon_K) \end{pmatrix}
\]

and where matrix \( \Sigma \) is a correlation matrix (its main diagonal is known to equal one), which moderates the relationships between the residuals, \( \Phi^{-1}(\cdot) \) is the inverse c.d.f. of the standardized normal, and \( F_k \)'s are the empirical marginal distributions of each characteristic.

Note that \( c_\Sigma(u) \) can be interpreted as a joint normal probability density function (p.d.f.) of random vector \( (\Phi^{-1}(u_1) ... \Phi^{-1}(u_K))' \). We simplify the notation by defining \( z_k = \Phi^{-1}(u_k) \), so that the copula can be readily interpreted as a p.d.f. of the random vector \( Z = \{z_1..z_K\} \):

\[
f_{Z_1..Z_K}(z) = \frac{1}{\sqrt{\text{det} \Sigma}} \exp \left\{ -\frac{1}{2} \begin{pmatrix} z_1 \\ \vdots \\ z_K \end{pmatrix}' . \Sigma^{-1} . \begin{pmatrix} z_1 \\ \vdots \\ z_K \end{pmatrix} \right\}
\]

(12)

Above, the elements of random vector \( Z \) follow non-independent standard normal distributions, with covariance matrix equal to \( \Sigma \). Estimation of the copula relies on the steps summarized in Figure A:

Figure A: Copula estimation and the marginal p.d.f. of \( z_k \)

First, the error estimates, orthogonal to make and color, are recovered through transformations \( \varepsilon_k = G_k^{-1}(x_k) \), as described in equations (7) and (9). Calculating the empirical
percentiles of $\varepsilon_k$ for each $k$ (or in different words, evaluating the empirical distributions $F_k$ at values $\varepsilon_k$), yields standard uniformly distributed random variables $u_k = F_k(\varepsilon_k)$. Then, applying the inverse standard normal c.d.f. yields normally-distributed random variables $z_k = \Phi^{-1}(u_k)$. Finally, $\Sigma$ is easily obtained by calculating the empirical correlation matrix of $(Z_1..Z_K)'$. The collection of steps described above allows us to characterize the relationships across variables, despite their different marginal distributions. As we explain later, the process depicted in Figure A is also used in reverse during estimation.

It remains to solve the second estimation challenge, which is to add the unobserved component $\nu_{ij}$ to this framework. We assume this component follows a normal distribution, such that it can be naturally added to the statistical copula by expanding the density appropriately:

$$f_{Z_1..Z_K,\nu}(z) = \frac{1}{\sqrt{\det \Sigma}} \exp \left\{-\frac{1}{2} \left( \begin{array}{c} z_1 \\ \vdots \\ z_K \\ \nu \end{array} \right)' \left( \begin{array}{cccc} \sum & \sigma_{z\nu} \\ \sigma_{z\nu}' & \sigma_{\nu}^2 \end{array} \right)^{-1} \left( \begin{array}{c} z_1 \\ \vdots \\ z_K \\ \nu \end{array} \right) \right\}$$

The covariance matrix $\Sigma$ is now bordered by the relationships between the unobserved component $\nu$ and each of the observable attributes. Vector $\sigma_{z\nu}$ contains the covariances between $\nu$ and each ‘normalized characteristic’ $z_j$, and $\sigma_{\nu}^2$ is a scalar corresponding to the variance of $\nu$. This specification allows us to identify correlations between observed and unobserved utility components. For example, it is possible that high-mileage vehicles are more likely to have lower values of the $\nu_{ij}$ component. The model leverages the search behavior on the unobserved component as a function of the values of the observed characteristics to estimate the covariance parameters in $\sigma_{z\nu}$, and produce consistent estimates.\footnote{One implicit assumption in the current formulation is that while the unobserved component may be correlated with the ordinal characteristics, it is orthogonal to the non-ordinal ones, namely make/model and color. Allowing for correlations between observable and unobservable vehicle characteristics, and recovering them from search data, has traditionally not been allowed for in the search literature, and so we believe our model still offers a significant contribution in this respect. Moreover, this potential concern is trivially solved through access to larger dataset (probably at the ‘big data’ level), by allowing for estimation of different joint distributions for each value of the non-ordinal variables.}

We discuss this further in the Identification section.
3.3 Search Behavior

During their searches, consumers decide whether to take subsequent information-acquisition actions about vehicles in their consideration sets, or to terminate search either by ordering a test drive or opting for the outside option. Given the discussion in the previous section, it is clear that search behaviors are cumulative processes, taking into account the information already available at each decision point. For example, a customer may stop or continue searching a vehicle depending on how many inspection issues she finds. Moreover, because features are correlated, a user may expect a vehicle to have had more owners after observing its high mileage. Clearly, expectations of unknown features should be conditioned on the known ones.

State Variables. We characterize the search problem of a consumer who acts ex-ante optimally according to her beliefs. Consumer $i$ considers alternatives $1..J_i$, each with $k$ characteristics, as described before. All consumers also have access to an outside option, with deterministic utility normalized to zero. Consumers have $S$ search actions available to them for each alternative, each of which maps into learning one or more characteristics. As before, there exist shocks $\epsilon_{ijt}$ that affect each vehicle’s utility in an idiosyncratic fashion. Moving towards a Bellman equation framework, we omit time subscripts, and assume the information-acquisition actions are also perturbed by preference shocks, denoted as $\epsilon_{ij}^s$, where $s \in S$ denotes a search action. These preference shocks are learned contemporaneously by consumers, and represent unobservable influences that may randomly affect consumers’ actions. We denote all of user $i$’s preference shocks as $\epsilon_i$.

We summarize the characteristics known by consumer $i$ about vehicle $j$ by set $\Omega_{ij}$, and denote the collection of all information known on all vehicles in consumer $i$’s consideration set by $\Omega_i$. For example, before initiating search for specific vehicles, user $i$ is endowed with information $\Omega_{ij} = \{\text{make}_j, \text{model}_j, \text{price}_j, \text{mileage}_j, \text{age}_j, \text{color}_j\}$ about vehicle $j$. The information set is then expanded according to the consumer’s search decisions. Finally, we denote the vehicle currently being viewed during the session as the scalar $J_i$. If the user has not yet visited a vehicle detail page during the session, we set $J_i$ equal to 0.
Decision-Making. At each decision point, consumer $i$ solves the problem:

$$V_F(\Omega_i, J_i, \epsilon_i) = \max \left\{ V_j(\Omega_i, \epsilon_i), V_{j,s}(\Omega_i, J_i, \epsilon_i) \right\}_{j=0..J_i, s=1..S}$$

(14)

where $V_F(\cdot)$ is a value function based on beliefs $F$. At each decision point, the consumer may decide to stop her search and make a final selection, or to continue learning by taking a search action. $V_j(\cdot)$ is the value of stopping search and selecting one of the $J_i+1$ alternatives, and $V_{j,s}(\cdot)$ is the utility of taking search action $s$ w.r.t. vehicle $j$. If the consumer stops her search, she earns expected utility

$$V_j(\Omega_i, \epsilon_i) = E \left( v_{ij} | \Omega_i, \epsilon_i \right)$$

(15)

$$= E \left( \sum_{k=1}^{K} \beta_k x_{jk} + \nu_{ij} \bigg| \Omega_i \right) + \epsilon_{ij}$$

(16)

conditional on the information available. If instead she decides to take a search action, say action $s$ w.r.t. vehicle $j$, she expects continuation utility

$$V_{j,s}(\Omega_i, J_i, \epsilon_i) = -c_s - 1 \left( l \neq J_i \right) c_{vd} + E_{\epsilon'_i, \omega_{js}} \left( V_F(\Omega_i \cup \omega_{js}, l, \epsilon'_i)|\Omega_i \right) + \epsilon'_{ij}$$

(17)

where $j$ and $s$ denote the vehicle and search actions to be maximized over. Vector $\omega_{js}$ contains the characteristics of vehicle $j$ to be learned with action $s$. The correspondence between search action and learned characteristics is presented in Table 5 (see also Table 4).

We allow consumers to incur different search costs depending on their actions, namely at rates $c_s$, $s \in \{1, 2, 3\}$. Moreover, we introduce parameter $c_{vd}$ to capture additional search costs (or savings) in case the consumer opts to move her search focus from one vehicle to another. This parameter is introduced to take into account that switching between vehicles may be more or less costly than browsing information sequentially within. For example, a consumer may move from one vehicle page to another by clicking the ‘back button’ on the browser first and then selecting the next vehicle. Alternatively, she may instead prefer to switch often among tabs of previously opened vehicles. This cost may be recovered positive, for example, if consumers tend to search characteristics within vehicles sequentially. It may also be negative, if consumers prefer to search each characteristic sequentially across vehicles, for example by comparing vehicle photos first, and then moving to vehicle histories, etc. We
later allow \( c_s \) to be heterogeneous across consumers.

Expression \( \Omega_i \cup \omega_j \) captures the updated information set the consumer will have access to, if she decides on search action \( s \) for vehicle \( j \). For example, if she starts out by searching vehicle \( j \)'s photos, her information set about vehicle \( j \) will transition from

\[
\Omega_{ij}^0 = \{ \text{make}_j, \text{price}_j, \text{mileage}_j, \text{age}_j, \text{color}_j \}
\]  

(18)

to

\[
\Omega_{ij}^1 = \{ \text{make}_j, \text{price}_j, \text{mileage}_j, \text{age}_j, \text{color}_j, \nu_{ij} \}
\]  

(19)

Consumer \( i \)'s decision to search incorporates the fact that she already has some information on vehicle \( j \). Hence, the expectation operator in expression (17) is taken with respect to the preference shocks \( \epsilon_i' \) as well as to the information \( \omega_{ls} \) the consumer will obtain through search, conditional on information set \( \Omega_i \).

3.4 Estimation

Search models are notably complex to estimate. In our case, estimation under optimal behavior, as expressed in (14), is feasible only in the presence of a limited number of alternatives and characteristics. The fact that some consumers engage in long search sequences implies using a significant number of simulations in order to calculate the likelihoods of search sequences. Moreover, in addition to the observable states, consumers exhibit an unobserved state for each vehicle in their search set. The result is a large dynamic problem with a large state space, part of which is unobservable. A related issue is that optimal search policies are challenging to derive, given the very large number of possible paths one has to integrate over. It is likely that actual consumers use relatively simple heuristics to make their search decisions. Based on these reasons, we employ a steepest-ascent decision procedure. Much like standard numerical solvers, we assume consumers base their decisions on one-period look-ahead policies.

**Steepest Ascent in Maximum Utility.** We focus on the case in which consumers take actions as a function of the steepest ascent in the expected maximum utility from searching. In this case, consumers compare the highest immediate expected utility from searching a
characteristic of an alternative with the status quo, and decide in favor of searching if the difference in utility is high enough. An equivalent idea is that consumers believe that their following action will be terminal, be it by selecting one of the vehicles or the outside option.

The utility of making a terminal decision toward alternative $j$ today equals

$$V_j(\Omega_i, \epsilon_i) = E(v_{ij} | \Omega_i, \epsilon_i)$$

whereas the utility of taking search action $s$ for some vehicle $j$ is equal to

$$V_{j,s}(\Omega_i, J_i, \epsilon_i) = -c_s - 1(j \neq J_i)c_{vdp} +
E_{\epsilon_i, \omega_{js}} \left[ \max \left\{ v_{i0}, E(v_1 | \Omega_i, \epsilon_i'), ..., E(v_j | \Omega_i \cup \omega_{js}, \epsilon_i'), ..., E(v_{J_i} | \Omega_i, \epsilon_i') \right\} | \Omega_i \right] + \epsilon_{ij}^s$$

In addition to integrating over the preference shocks, taking a search action involves calculating two sets of conditional expectations. The reason is that when consumer $i$ takes search action $s$, she first expects to learn the characteristics in $\omega_{js}$, which are most likely not known (hence the search). She forms an expectation of these characteristics by conditioning on the information already available, $\Omega_i$. The second expectation takes place because some characteristics may remain unknown after the current search action is taken, while being partially informed by $\omega_{js}$ in the next period. At that point, the expected utility of alternative $j$ will be of the form

$$E\left(v_j | \Omega_i \cup \omega_{js}, \epsilon_i'\right) = E\left(\sum_{k=1}^{K} \beta_k x_{jk} + \nu_{ij} | \Omega_{ij}, \omega_{js}\right) + \epsilon_{ij}'$$

where $\{\Omega_{il}, \omega_{ls}\}$ is the new information set, and the remaining variables in $\{x_{jk}, \nu_{ij}\} \setminus \{\Omega_{ij}, \omega_{js}\}$ remain uncertain.

Under the steepest ascent assumption, consumers make their present decision as if the next period were terminal. Consumer $i$ ’s optimal action solves problem

$$a^*(\Omega_i, J_i, \epsilon_i) = \arg \max \left\{ V_j(\Omega_i, \epsilon_i), V_{j,s}(\Omega_i, J_i, \epsilon_i) \right\}_{j \in J_i, s \in S}$$

Hence, at each decision point, the consumer takes the search action as long as it increases
the status quo utilities enough when compared to the implied search costs.

We assume that shocks $\epsilon$ follow a type 1 extreme value distribution (E.V.) with parameters $\{0, \sigma_\epsilon\}$. This idiosyncratic shock can be interpreted as a structural preference shock that may affect consumer decisions at any decision point. We define the standardized expected utility of a generic alternative $j$ conditional on some information set $\Omega_{ij}$, and net of preference shock $\epsilon_{ij}$, as

$$v_{ij}(\Omega_{ij}) = \frac{1}{\sigma_\epsilon} E[x_{jk}, \nu_{ij} \left( \sum_{k=1}^{K} \beta_k x_{jk} + \nu_{ij} \right) \mid \Omega_{ij}]$$  \hspace{1cm} (24)

We now take advantage of the E.V. distribution assumption to write the expected utility from taking search action $s$ w.r.t. alternative $j$:

$$V_{j,s}(\Omega_i, J_i, \epsilon_i) = -c(s) + \sigma_\epsilon E_{\omega_{js}} \left\{ \gamma + \log \left[ 1 + \sum_{i \neq j} \exp(v_{il}(\Omega_{il})) + \exp(v_{ij}(\Omega_{ij} \cup \omega_{js})) \right] \mid \Omega_{ij} \right\} + \epsilon^s_{ij}$$  \hspace{1cm} (25)

where

$$c(s) = -c_s - 1 (l \neq J_i) c_{edp}$$  \hspace{1cm} (26)

as before. The term ‘1’ in the logarithm in equation (25) follows from the normalization of the outside option. The sum of the transformed utilities of vehicles different from $j$ relates to the utility of the remaining alternatives individual $i$ may select at the next decision point. The final term in the logarithm relates to the expected utility of alternative $j$ with the additional knowledge of $\omega_{js}$, produced by search action $s$ of alternative $j$. Parameter $\gamma$ is Euler’s gamma constant and, as mentioned before, parameter $\sigma_\epsilon$ is the scale parameter of the preference shocks. Overall, the expression above takes into account that, by taking search action $\{j, s\}$, consumer $i$ expands her information set by $\omega_{js}$, after which expects to select the alternative with the highest utility. When considering whether to take search action $s$ w.r.t. vehicle $j$, consumer $i$ takes into account the information she is likely to learn, $\omega_{js}$, depending on the information she already possesses about vehicle $j$, $\Omega_{ij}$, as well as the potential consequences of learning $\omega_{j,s}$ on her beliefs about the remaining characteristics.

**Variance-Covariance Matrix.** Matrix $\Sigma$ moderates the relations between the vehicles’
observed characteristics, and its border includes the relations with the idiosyncratic utility component $\nu$. As we further discuss in the Identification section, we normalize parameter $\sigma^2_\nu$ to one. The remaining elements of the border of the variance-covariance matrix are estimated under the restriction that the resulting variance-covariance matrix is positive semi-definite (p.s.d.). First, we find an upper triangular matrix $U$ such that

$$U'U = \begin{pmatrix} \sum & \sigma_{z\nu} \\ \sigma'_{z\nu} & 1 \end{pmatrix}$$

(27)

Since $\Sigma$ is known, the process of finding $U$ follows from common Cholesky decomposition. In the appendix we explain how we parameterize matrix $U$ in order for it to be simultaneously consistent with the correlation matrix of the observable shocks in the data (see Table 6, discussed later) and remain p.s.d.

**Likelihood.** In our model, each observation is a sequence of search actions followed by a terminal action performed by a consumer. The likelihood function characterizes the probability of a sequence of actions conditional on a set of parameters and data. It is given by

$$L (\theta| X, A_i) = \prod_{i=1}^N Pr (A_i | \theta, X)$$

(28)

where $X$ contains vehicle characteristic data, $A_i = \{a^1_i, ..., a^{T_i}_i\}$ is the sequence of actions of individual $i$, $T_i$ is the number of actions performed by individual $i$, and $\theta$ is a vector of parameters. We now consider the likelihood of a given individual $i$ in more detail. Unlike consumers, the econometrician does not observe preference shocks nor the idiosyncratic component utilities. We first write the likelihood function conditional on these terms being known (by the econometrician), and later integrate them out. When $\nu_i$ and $\epsilon_i$ are known, the probability of sequence $A_i$ can be written as (henceforth omitting matrix $X$)

$$Pr (A_i | \theta, \nu_i, \epsilon_i) = 1 (a^1_i = a^{1*}_i (\theta, \Omega^1_i, \epsilon_i) \land ... \land a^{T_i}_i = a^{T_i*}_i (\theta, \Omega^{T_i}_i, \epsilon_{iT_i}) | \nu_i, \epsilon_i)$$

(29)
where \( \{a_1^{i*}, ..., a_{T_i}^{i*}\} \) are the policies implied by the model, to be compared with the actions observed in the data, \( \{a_1^i, ..., a_{T_i}^i\} \). When \( \nu_i \) and \( \epsilon_i \) are known, the likelihood of a given sequence is clearly equal to zero or one. Above, \( \Omega_i^1 \) is also conditioned on (but omitted for notation simplicity), and the transitions between information sets \( \Omega_i^t \) to \( \Omega_i^{t+1} \) are also perfectly known, given shocks \( \nu_i \) and \( \epsilon_i \).

The preference shocks \( \epsilon_i \) can be easily integrated out. First, define \( U_j(\Omega_i) \) to be equal to \( V_j(\Omega_i, \epsilon_i) \), net of the additive preference shock \( \epsilon_{ij} \), and the analogous definition for \( U_{j,s}(\cdot) \). Since each action is associated with one such shock, and moreover, since they are uncorrelated, the probability of action profile \( A_i \) when \( \epsilon_i \) is unknown can be written as

\[
Pr (A_i| \theta, \nu_i) = \int Pr (A_i| \theta, \nu_i, \epsilon_i) d\epsilon_i
\]

\[
= K \left( U_{a_1^i} (\theta, \Omega_i^1) \middle| \nu_i \right) \times K \left( U_{a_2^i} (\theta, \Omega_i^2) \middle| \nu_i \right) \times \cdots \times K \left( U_{a_{T_i}^i} (\theta, \Omega_i^{T_i}) \middle| \nu_i \right) \quad (30)
\]

where

\[
K \left( U_{a_i^i} \middle| \nu_i \right) = \left[ \exp \left( -\frac{1}{\sigma_e} U_{a_i^i} \right) + \sum_j \exp \left( \frac{1}{\sigma_e} (U_j - U_{a_i^i}) \right) \right]^{-1}
\]

\[
= \frac{1}{\sum_j \exp \left( \frac{1}{\sigma_e} (U_j - U_{a_i^i}) \right)}
\]

which is the familiar logit expression induced by the E.V. assumption. The denominator features the standardized utilities of all actions netted out by the standardized utility of the action observed in the data. Notation \( K (\cdot \middle| \nu_i) \) stresses the idea that the values of \( \nu_i \) are currently being ‘plugged in’, as if known deterministically by the econometrician. It remains to integrate them out. We turn to simulated maximum likelihood, by taking simulation draws of the idiosyncratic utility component \( \nu_i \). Because the shocks are related with the observable vehicle characteristics through a multivariate distribution, their distribution depends on the parameters \( \theta \). The likelihood function for a sequence of actions by consumer \( i \) is thus given by

\[
Pr (A_i| \theta) = \int Pr (A_i| \theta, \nu_i) dF_{\nu_i \mid X, \theta}
\]

\[
= \int \frac{Pr (A_i| \theta, \nu_i)}{F_{\nu_i \mid X, \theta}} dF_{\nu_i \mid X, \theta}
\]

where \( F_{\nu_i \mid X, \theta} \) is the c.d.f. of the idiosyncratic utility shocks conditional on data \( (X: \text{vehicle information}) \) and parameters \( (\theta: \text{the covariance parameters in } \Sigma) \). We approximate this
probability via simulation, according to

$$
\bar{P}_r (A_i | \theta) = \frac{1}{R} \sum_{r=1}^{R} P_r (A_i | \theta, \nu_r^i) \tag{33}
$$

and the simulated log-likelihood follows:

$$
l(\theta | X) = \log \left( \prod_{i=1}^{N} \bar{P}_r (A_i | \theta) \right) \\
= \sum_{i=1}^{N} \log \left( \bar{P}_r (A_i | \theta) \right) \tag{34}
$$

A final note related to the extreme value preference shocks is in order. These shocks have two main functions. First, they smooth out the accept/reject simulation, while maintaining the consistency properties of the maximum likelihood estimator (see McFadden (1989) and Train (2009)). Second, they also smooth out consumers’ expected values of searching, which involve a ‘max’ function, by greatly reducing the simulation burden during estimation. These assumptions operate at different levels. The inclusion of these shocks in the former case is a useful and harmless econometric aid, whereas in the latter case it is a ‘structural’ assumption, in the sense that consumers expect and incorporate the shocks into their utilities. In order to ensure that the preference shocks do not overwhelm search decisions, we set their variance to a low value (0.1). This provides smooth inclusive values without affecting the estimated parameters significantly, as confirmed by estimation runs. However, we increase the variance of the shocks to 1 when simulating accept/reject decisions. This provides a smooth simulated likelihood function on the parameter values. The interpretation of these assumptions is that the structural model is one in which consumers expect and earn utility shocks with a variance small enough to help smooth the ‘kinks’ of the ‘max’ function associated with computing future values, and to not affect the estimates of the remaining parameters significantly. The econometrician then includes additional noise with variance of 0.9, à la McFadden (1989), to smooth out the accept-reject estimator with no effect on parameter consistency.

Finally, we allow the search costs $c_s$ to be heterogeneous across consumers through a latent class model. We introduce parameter $\alpha$, such that $g(\alpha)$ proportion of consumers face search costs $c_s$, and $1 - g(\alpha)$ face search costs $c_s'$, where $g(\cdot)$ is the logistic ratio, used to perform unconstrained estimation of $\alpha$.\footnote{In practice, none of the estimation iterations induced a segment share near zero or one.}
Figure B: Obtaining draws of characteristics $x_l$, conditional on characteristics $x_k$.

![Diagram](image)

Notes: The process above applies to observable characteristics. The component $\nu$ is assumed to follow a normal distribution, from which draws can be readily taken, given knowledge of covariance matrix $\Sigma$.

**Simulation.** During the decision process, consumers need to integrate over the information they are likely to learn from different search actions, conditional on the information already available. These expectations are approximated via simulation. We now depict the procedure of taking draws of $\omega_{js}|\Omega_{ij}$, where $\omega_{js}$ are characteristics the individual may be searching over, and $\Omega_{ij}$ is the current information set.

We have already defined the joint distribution of the random variables $\{z_k, \nu\}$, underlying the vehicles' characteristics (see equation (13)). However, except for $\nu$, these variables do not enter utility functions directly. Additional steps are required to take draws of the characteristics they underlie, as depicted in Figure B. Taking draws of vehicle characteristics $\omega_{jk}$, conditional on characteristics $\Omega_j$, involves reversing the process outlined before. For example, suppose we intend to take draws of characteristic $x_2$, conditional on the information set $\Omega_j = \{x_1, \nu\}$. First, we convert $x_1$ to $z_1$ by following the steps depicted on the top arrows of Figure B. Second, we use the conditional normal p.d.f. to take draws $\{z_2^r\}_{r=1}^R$, conditional on $\{x_1, \nu\}$. We then carry out the reverse process, depicted on the bottom of Figure B, to produce draws of characteristic $x_2$.

This procedure is useful to calculate expected search utilities. Consider equation (25),
repeated below for convenience

\[ V_{j,s} (\Omega_i, J_i, \epsilon_i) = -c(s) + \sigma_x E_{\omega_{js}} \left\{ \gamma + \log \left[ 1 + \sum_{l \neq j} \exp (\nu_{il}(\Omega_{il})) + \exp (\nu_{ij}(\Omega_{ij} \cup \omega_{js})) \right] \right| \Omega_{ij} \} + \epsilon_{ij}^s \]  

(35)

Calculation of the expression above requires first taking draws of characteristics \( \omega_{js} \), conditional on the information set \( \Omega_{ij} \). Let these draws be denoted as \( \{ \omega^r_{js} \}_{r=1}^R \). For each element of this vector, we then take draws of the remaining characteristics, conditional on the simulated information sets \( \Omega^r_{ij} = \{ \Omega_{ij}, \omega^r_{js} \} \). For each draw of \( \Omega^r_{ij} \), we evaluate the integrand of the expectation operator in equation (35). We then average across integrands to approximate the outer conditional expectation.

One useful fact for estimation is that the first stages outlined on top of Figure B can be pre-calculated, and the results saved before estimation. Hence, the values of \( z_k \) are available for each vehicle before estimation. A second useful fact is that one can also save the draws of the remaining observable characteristics, in the cases where \( \nu \) is not yet known. For example, a consumer may start her search by browsing a vehicle’s inspection notes. The draws of these inspection notes, conditional on the observable characteristics, can be produced before estimation. The limitation is that once a consumer learns her idiosyncratic shock \( \nu_i \) for vehicle \( j \), her beliefs over the distribution of the remaining characteristics of that vehicle depends on an unknown shock to the econometrician. It is infeasible to save draws of characteristics conditional on all possible values of the idiosyncratic utility component, because it would entail having a different set of draws for each possible value of \( \nu_i \). Storing draws whenever possible (i.e., for the information sets not containing \( \nu_i \)) reduced estimation time significantly.

3.5 Identification

The identification of the search model is relatively complex, because most of the parameters affect each of the moments of the corresponding data generation process. The preference parameters are inferred through the relationships between their respective regressors and purchase rates, as in traditional discrete choice models. However, they also affect search behaviors. Notice, for example, that a consumer may be more willing to search a vehicle’s characteristic, say \( x_1 \), when her corresponding preference weight \( \beta_1 \) is also higher. The
reason is that higher values of $\beta_1$ induce higher variances of the expected search utility from evaluating $x_1$.

The search patterns in the data, together with the conversion decisions, allow us to characterize the preference parameters and the search cost parameters. The latter ones are intuitively identified by the average propensity of consumers to take each search action, conditional on the covariance parameters. This last statement is especially applicable to search actions involving vehicle characteristics that are known to the econometrician. In terms of the unobservable characteristic, there is an observational equivalence between a consumer who faces a low search cost and low variance with another who faces a high search cost and high variance surrounding the same characteristic. Simply put, more uncertainty surrounding a characteristic increases the option value of search, which in turn can be offset by higher search costs. In order to take this into account, we normalize the variance of the unobservable characteristic to one.

It remains to discuss the identification of the covariance parameters between the observed characteristics and the unobserved utility component $\nu$. Whereas search costs are identified by the average propensity of each search action being taken, the covariance parameters are identified by information-dependent search patterns concerning the unobserved characteristic. For example, suppose a consumer knows a given vehicle’s price. One can imagine that if the price is moderately high, she may opt for the outside option immediately, conditional on the preference and search cost parameters. If the price were lower, it is possible that the consumer would have continued her search of the vehicle’s attributes, including of component $\nu$. The price threshold at which the consumer is indifferent between continuing her search for $\nu$ identifies the covariance between price and the unobserved component: If the consumer selected the outside option without searching further, despite facing only a moderate price point, then it must have been that the residual variance of the unknown component was not high enough to justify further search. On the other hand, continuing search despite the current low prior on the vehicle’s expected utility is a signal of high variance of component $\nu$, conditional on the known characteristics. The parameters $\sigma_{z\nu}$ of the variance-covariance matrix thus identify the residual variance of component $\nu$, conditional on knowledge of the vehicles’ remaining characteristics.
4 Results

Before presenting the model estimates, we detail the correlation structure among the observable vehicle characteristics, net of the make/model and color effects. Table 6 presents the cross-correlations for the residuals discussed in equation (12). The results are in line with those in Table (2), which presents the correlations before the effects of make/model and color are removed. The main difference is related to price: The negative correlation between price and the remaining characteristics tends to increase as a result of removing the effects of the categorical variables.

Turning now to the model results, Table 7 presents the estimates of the preference coefficients. As expected, consumers dislike more expensive vehicles, with higher mileage, with inspection notes (found in the vehicle’s inspection report), older, with more past owners, and with previous accidents.\textsuperscript{14} Table 8 presents the estimates of the search cost parameters by segment. The size of the first segment is of approximately 34%. These consumers face positive search costs towards all activities. The highest search cost is related to inspecting vehicle inspection reports, and the lowest one is related to browsing through vehicles’ histories. This is consistent with the layout of the website at the time of the dataset, where the link to the vehicle’s inspection report was located at the bottom of the vehicle detail page, whereas the link to the vehicle history was located near the top. The second segment shows a similar pattern with regard to these activities, exhibiting higher values of search costs in general. In addition, unlike the first segment, it features a search benefit for browsing photo sets. Figure 6 presents the same data graphically. It follows that the second segment mostly engages in photo browsing, potentially deriving positive utility from this activity.\textsuperscript{15} These results point to two very different segments: The first engages in search behaviors across the board, and the second focuses primarily in browsing vehicle pictures. Given the relatively hedonic nature of the product in question, it is possible that the second segment derives utility from browsing through different vehicle profiles, potentially not being especially interested in an eventual purchase. Finally, the positive estimate of parameter

\textsuperscript{14}This manuscript reports the results based on a random sample of 1,000 consumers. Preliminary results on a larger sample based on parallel estimation are available. Current model fit statistics are provided in Table 11. The current estimates tend to underpredict the intensities of purchase and search behaviors observed in the data.

\textsuperscript{15}The reason that this analysis cannot unambiguously claim that segment 2 derives a net benefit from browsing photos is that a normalization to a higher level of the variance associated with the unobserved utility component would likely lead to the search cost estimate associated with browsing photos to increase.
reveals a tendency to browse vehicles one at a time, rather than by comparing vehicles characteristic-by-characteristic.

The estimates of the variance parameters are presented in Table 9. They are loaded into linear combinations, which then yield the estimated variance-covariance border presented in Table 10. We observe high correlations between the observed characteristics and the unobserved utility component. Of mention are the following high correlation results: The unobserved component is negatively correlated with price, but positively correlated with vehicle age. The identifying reason is that consumers are more likely to search this component (through picture browsing) when cars are cheaper and older. These results may indicate consumers’ hopes to find silver linings and/or deals by continuing to search such vehicles. The magnitude of these correlations relatively is high, and so we investigate their impacts in the next section.

5 Counterfactual Analyses

In this section we address two main questions. First, we analyze the role of the institutional features consumers face in our context, namely, of the search costs they face and of the correlations between observed vehicle characteristics and the unobserved utility component. Second, we consider the effects of different information sets, obtained by deciding which characteristics to disclose upfront, and which to have consumers search over.

5.1 Institutional Features: Search Costs and Unobserved Cross-Correlations

Table 12 contrasts consumer behavior indicators, as obtained by the model, with those obtained when consumer search costs are eliminated. The first result is that the purchase rate increases, from 4.6% to 5.4%. Given the relatively low conversion rate, the increase (approximately 17% in relative terms) is economically significant. The result is intuitive: The low base conversion rate is a result of most consumers not finding a desirable alternative while stopping search early. In some cases though, additional search would have revealed positive surprises and would have resulted in higher purchase rates. As expected, search behaviors

\[^{16}\text{We preserve the second segment’s negative search costs of browsing vehicle pictures.}\]
increase to their theoretical maxima, constrained by the size of consumers’ consideration sets as well as by the random preference shocks.

Search models are generally accepted to be complex, in that changes to the information structure can lead to relatively unpredictable search patterns. Figure 7 documents the impact of allowing for a correlated unobservable utility component (vs. the case where the unobserved component is independent from the remaining characteristics). We observe the actions preceding conversion being substituted from other vehicles the same vehicle, i.e., consumers appear to search more within a vehicle before deciding to purchase. The net effect on conversions (not show in the figure) is negligible, such that the cross-correlations appear to mainly affect how consumers look for information. We also observe searches over inspection reports resulting in more subsequent searches over vehicle photos, and fewer over vehicle histories. Overall, the resulting search effects are complex and not necessarily predictable from the correlations in Table 10.

5.2 Designing Information Provision

The problem we concern ourselves with is of a seller who decides which characteristics to include in consumers’ initial information sets, and which characteristics to require ‘effort’ to be learned. In our setting, the initial information set comprises all characteristics displayed in the main listing page, and the search set comprises the characteristics only available in the vehicle detail pages. Hypothetically, the seller could reduce all search costs to zero by displaying all attributes in the home page, thereby allowing consumers to identify their best matches ‘instantly’. In practice, however, moving all vehicle information to the main listing would result in a cluttered website, with effectively very high search costs. In this section we consider counterfactual analyses where the features in both levels (home page and vehicle detail page) may be exchanged, in order to increase consumer fit and conversion rates, ceteris paribus (for example, the level of website complexity and cluttering is held constant).

It is worth visiting the intuition of why the order of information disclosure matters. Consider a simple example, in which a consumer decides between purchasing a product of unknown utility $v_1$, or opt for the outside option, with known utility $v_0 = 1$. Moreover, let the inside good have two characteristics:

$$v_1 = x_1 + x_2$$

(36)
where $x_1 \in \{-2, 2\}$ and $x_2 \in \{-1.5, 1.5\}$ are independent, both with equiprobable outcomes. These parameters are chosen so that the mean of $x_1$ and $x_2$ are the same, and so that consumers would need to receive good news on both fronts in order to buy. This is consistent with the relatively low conversion rates we observe in our dataset.

Suppose the seller could move one of the characteristics to the consumer’s initial information set, but not both. Which one should she pick? Figure 8 depicts the possible utility transitions and outcomes from searching in each case. The left panel shows the case where the consumer is endowed with information about $x_1$, and so begins her search with expected utility of good 1 at utility level of 2 or -2. Notice that if she starts out at -2, even the best possible case only transitions her to posterior utility -0.5, which will never dominate the outside option. On the other hand if she begins her search having found that $x_1 = 2$, search may be valuable. If she decides to buy good 1 immediately, she earns expected utility of 2, but if she searches, she earns

$$E_{x_2}(\max\{E(v_1|x_1 = 2, x_2), 1\}) = 2.25$$

(37)

because if it turns out that $x_2 = -1.5$, she can still opt for the outside option, and earn utility equal to 1. The difference between points A and B in Figure 8 is the consumer’s option value of search, equal to 0.25. She will be willing to search, after learning that $x_1 = 2$, whenever the search cost falls below 0.25.

Consider now the right panel of Figure 8, in which the consumer learns about characteristic $x_2$ first. Notice that the final utility levels equal the ones in the left panel, because the order in which information is acquired does not affect utility realizations. As before, if she observes the lowest outcome, additional search can never yield utility higher than 1. If she observes $x_2 = 1.5$, however, additional search of good 1 produces more search utility than before, because the bad outcome (i.e., finding that $x_1 = -2$) is ‘insured’ by the outside good at a high rate. Because of this, her expected utility from searching good 2 increases from 1.5 to 2.25. When the consumer is initially aware of characteristic 2, she is willing to search characteristic 1 if the search cost falls below 0.75, a higher threshold than before, due to the higher variance of characteristic 1.

It is easy to confirm that when the search cost is equal to 0.5, disclosing $x_1$ rather than $x_2$ increases demand from 0.25 to 0.5. First, note that in this example, consumers are only
willing to buy if they learn that both attributes are high, or if they learn that one is high while the other is not searched over. In our example, if consumers observe a high level of \(x_1\) initially, they will refrain from searching because the residual variance in \(x_2\) offers little search value. Hence, 50% of consumers faced with information of \(x_1\) initially will buy. On the other hand, consumers who learn that \(x_2\) is high initially are still willing to search \(x_1\), and only those who received two positive signals will buy (mass of 25%). By designing the information structure, the seller was able to take advantage of the search costs and ensure higher profitability.

Notice also that this intuition does not imply that providing more information to consumers is always better. Had the seller provided information on both \(x_1\) and \(x_2\), the conversion rate would have remained 25%. In other words, there exists an optimal ‘interior point’, in terms of how much uncertainty to resolve first, and how much to leave potentially unresolved due to search costs.

This example suggests that it may be best for the seller to reveal the information with the highest variance first, and have consumers potentially search over the remaining characteristics. This result does not necessarily hold with interdependent characteristics, which highlights the need to characterize the characteristic cross-correlations and the resulting search process in detail.

Table 13 documents the effects of exchanging attributes between levels. In particular, we consider the counterfactual cases of switching information on price, mileage, and vehicle age with information obtained by accessing vehicles’ histories or inspection reports. The intuition for this one-to-one exchange is that it should be feasible for seller to replace information to the corresponding locations on the website, while keeping search costs relatively constant. For example, by exchanging mileage information by the number of inspection notes found, consumers would be likely to face the same search costs, which are related by the position and salience of the information to be searched on the vehicle detail pages.\(^{17}\) Table 13 reveals that all one-to-one exchanges benefit the seller. In particular, exchanging vehicle history elements (i.e., age and number of owners) with price translates into a 1.2% increase in sales. Given the relatively low conversion rate predicted by the model of 4.6%, this effect constitutes a 26% relative increase, of clear economic significance.

\(^{17}\) We do not consider the possibility of moving the vehicle photo galleries to the homepage, as that would be largely infeasible, requiring too much real estate and essentially changing the website look and feel.
In some cases, sellers may refrain from not listing pricing information on the listing pages. For example, not showing price upfront could induce unexpected consumer beliefs about the seller, currently not manifested in the data. Nonetheless, inspection of Table 13 further reveals that, independently of the feature moved to the vehicle detail page, moving vehicle history information to the homepage is most beneficial to the seller.\footnote{18 Additional analysis with the complete set of feature combinations is straightforward, but the economic intuition behind the results is less interesting and predictable.}

In order to provide insight into the underlying mechanism, Table 14 presents information on the impact of different information designs on search behavior. Specifically, it presents the percent change in the number of search actions as a result of changes to the information disclosure policy. We find that all counterfactual information scenarios, which according to Table 13 lead to higher profitability, also lead to less search. On average, the number of search actions taken by consumers decreases by 2.4% when price information is replaced by information on the vehicles’ histories. Moreover, in line with the motivating example discussed above, the increased profitability of the new information scenarios is negatively associated with the changes in search intensity. The correlation between the profitability increases and the decreases in search activities is equal to -98.5%, a strong indication that profitability can be increased by providing information of the most diagnostic characteristics first, thereby reducing subsequent search. The results in Table 15 further confirm this intuition: All information scenarios decrease the option value of the first search. Moreover, the ordering is similar to the ones presented in the previous tables. We make the relationship between the initial value of searching and conversion rates more explicit in Figure 9. The downward relationship shows that decreasing the value of search is associated with higher conversion rates. While search behavior is complex, in our setting, the option of value of search provided to consumers obtained by selecting which characteristics are initially known to consumers has an almost perfectly monotonic relationship with sales and profitability.

5.3 A Simplified Approach to the Design Problem

We now ask whether a relatively simpler analysis around (conditional) variances would have predicted the same profitability implications of the different information designs. If such a method performs well, it is an indication that sellers may be able to use residual correlation-based heuristics of product characteristics to design information disclosure policies.
reinforces the need to model search in a detailed fashion. In this alternative approach, we use the covariance matrix of the observable characteristics, and weight it by the estimated logit coefficient estimates. The preference coefficients recover the relative weights of the different characteristics as well as the effect of different scales/units. We first calculate the residual variance of utility, implied by the initial information set.

This benchmark variance is equal to

\[ V_0 = \text{Var}(Y|\beta, \text{price}, \text{mileage}, \text{age}) \]  

where \( Y = \sum_{k=1}^{6} \beta_k x_k \). We then calculate the difference between \( V_0 \) and the conditional variances induced by different information scenarios, much like in the analysis above. Our approach is to include a border to the original characteristic covariance matrix, with the covariances of random variable \( Y \). We then calculate the variance of \( Y \), conditional on different information sets.

Table 16 presents the results in terms of percent changes of the residual variances, and shows that this approach to the information design problem leads to very different recommendations. Based on the idea that the seller is better off opting for the initial information set that minimizes the residual variance, this analysis proposes that exchanging price with any of the remaining characteristics is beneficial. This is a different prediction than our model’s, in which the most beneficial design endows consumers with information about vehicle histories. Similarly, the current analysis suggests that moving mileage and/or age to ‘deeper’ information sets is undesirable, contrary to the results of our counterfactual analysis. These results suggest that even a moderately sophisticated approach to the design problem can yield poor quality recommendations.

6 Conclusion

We have developed and estimated a model in which consumers search across attributes of multiple alternatives before making their conversion decisions. The model is able to characterize complex search sequences, including sequential search across alternatives and along characteristics within each alternative. In addition, it allows for a rich joint distribution of the observable characteristics, as well as the characterization of the cross-correlations with
an unobserved utility component. The joint distribution informs consumer beliefs during search, such that good news about a specific attribute has implications for the remaining ones in consumers’ minds. For example, a consumer may dislike older vehicles because she believes they are likely to be in worse condition. As a result, she may be more likely to investigate their pictures and inspection reports more often than for newer vehicles. Allowing for arbitrary joint distributions of characteristics had thus far been ignored, due to the complexity they impose on the model and estimation as well as for lack of datasets detailing the mapping between search actions and the learned product features.

Another contribution of this paper is to consider the problem of designing information provision to consumers. By recomputing counterfactual search paths depending on different information sets, our model predicts that there is a reasonable potential growth opportunity from introducing changes to the information environment. Moreover, we find that a naive approach based on the decreases in residual variances of utilities calculated with traditional logit estimates produces poor quality recommendations.

The challenge of modeling search is far from being completely addressed. One challenge, partly addressed by ongoing research, is estimating consumers’ dynamic maps of which features translate to higher utilities. In this case, consumers learn about each product’s features as well as their own preferences. Insights on the identification and potential parsing of the learned objects will be certainly valuable to the field. Another open question is on identifying the actual updating process taking place. Despite some progress in recent years, it remains challenging to identify a parsimonious and estimable model that characterizes all the patterns observed in search datasets. Additional research on this topic, including investigating search with experimental methods, may provide valuable results.
### Figures

Figure 1: Densities of Avg. Session Times Across Users

Kernel Densities of Avg. Session Time

Red = Did not convert; Blue = Converted

Note: Density upper bounds selected to ensure columns are visually distinguishable across conditions.
Figure 2: Histograms of Search Behavior Statistics

Note: Histogram upper bounds selected to ensure columns are visually distinguishable across conditions.
Figure 3: First-Order Transitions of Search Behaviors

Note: Above, transitions among events of focal vehicle $j$ and also across to other vehicles ($-j$). The figure depicts all transitions probabilities above 10%, plus transitions toward converting.
Figure 4: Histograms and Cross Scatter Plots of Vehicle Characteristics

Note: Above, histograms in main diagonal and scatter plots in off-diagonals. Prices in (10^{-4} USD) and mileage in (10^{-5} miles).
Figure 5: Densities of Vehicle Purchase Probabilities with and without Search Costs

Kernel Densities of Purchase Probabilities
Red= Estimated Search Costs; Blue= Perfect Information
Figure 6: Search Cost Estimates

Search Cost Estimates by Segment

- c(photos)
- c(vehicle history)
- c(insp. report)
Note: Above, impact of setting unobservable covariances to zero. The figure depicts absolute changes above 2%.
Figure 8: Illustration of the Effect of Variance on Option Values

Note: Above, the left and right-hand sides denote the possible utility outcomes, starting with knowledge of $x_1$ or $x_2$ at moment 1, respectively. Note that the final utility levels are the same, but in the right-hand side, learning about $x_1$ (the characteristic with highest variance) induces a higher option value of searching (vertical distance CD, in relation to AB).
Figure 9: Effect of the Option Value of Search on Conversion Rates

Note: Above, each point combines the change in the value of the first search with the changes in conversion rates. The dashed line is the fitted non-linear exponential trend.
Tables

Table 1: Descriptive Statistics

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vehicle Detail Pages Viewed</td>
<td>2.28</td>
<td>3.47</td>
<td>1</td>
<td>129</td>
</tr>
<tr>
<td>Number of Sessions</td>
<td>6.03</td>
<td>21.14</td>
<td>1</td>
<td>762</td>
</tr>
<tr>
<td>Photo Sets Browsed</td>
<td>1.63</td>
<td>2.65</td>
<td>0</td>
<td>74</td>
</tr>
<tr>
<td>Vehicle Histories Browsed</td>
<td>0.69</td>
<td>2.09</td>
<td>0</td>
<td>101</td>
</tr>
<tr>
<td>Inspection Reports Browsed</td>
<td>0.30</td>
<td>1.27</td>
<td>0</td>
<td>53</td>
</tr>
<tr>
<td>Average Session Duration (min.)</td>
<td>14.2</td>
<td>13.39</td>
<td>0.1</td>
<td>282.04</td>
</tr>
<tr>
<td>Activity Range (days)</td>
<td>18.67</td>
<td>38.5</td>
<td>0</td>
<td>232</td>
</tr>
</tbody>
</table>

Number of users: 24,116

Note: Statistics are per user. The event “Photo Sets Browsed” equals one for each user-vehicle observation whenever a user browsed more 20 or more pictures of a vehicle.

Table 2: Correlation Table for Non-Categorical Characteristics

<table>
<thead>
<tr>
<th></th>
<th>Price</th>
<th>Mileage</th>
<th>Insp. notes</th>
<th>Vehicle age</th>
<th>No. accidents</th>
<th>No. owners</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mileage</td>
<td>-0.45</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insp. notes</td>
<td>-0.16</td>
<td>0.39</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vehicle age</td>
<td>-0.40</td>
<td>0.71</td>
<td>0.44</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. accidents</td>
<td>-0.07</td>
<td>0.10</td>
<td>0.09</td>
<td>0.12</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>No. owners</td>
<td>-0.08</td>
<td>0.21</td>
<td>0.09</td>
<td>0.17</td>
<td>0.09</td>
<td>1</td>
</tr>
</tbody>
</table>

Note: Number of vehicles: 1,573. All correlation estimates statistically significant, with p-values below 0.01.
Table 3: Logit Preference Coefficient Estimates

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Point Estimates</th>
<th>Odds Ratios</th>
</tr>
</thead>
<tbody>
<tr>
<td>price</td>
<td>-1.359**</td>
<td>0.257**</td>
</tr>
<tr>
<td></td>
<td>(0.035)</td>
<td>(0.009)</td>
</tr>
<tr>
<td>mileage</td>
<td>-1.189**</td>
<td>0.305**</td>
</tr>
<tr>
<td></td>
<td>(0.136)</td>
<td>(0.041)</td>
</tr>
<tr>
<td>age</td>
<td>-0.125**</td>
<td>0.883**</td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
<td>(0.012)</td>
</tr>
<tr>
<td>inspection notes</td>
<td>-0.06*</td>
<td>0.942*</td>
</tr>
<tr>
<td></td>
<td>(0.026)</td>
<td>(0.025)</td>
</tr>
<tr>
<td>n_accidents</td>
<td>-0.167†</td>
<td>0.846†</td>
</tr>
<tr>
<td></td>
<td>(0.093)</td>
<td>(0.079)</td>
</tr>
<tr>
<td>n_owners</td>
<td>-0.148**</td>
<td>0.863**</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
<td>(0.025)</td>
</tr>
</tbody>
</table>

Number of users: 24,166

Standard errors shown in parentheses. Significance levels: † p≤0.10, * p≤0.05, ** p≤0.01. Price in ($10^{-4}$ USD) and mileage in ($10^{-5}$ miles).

Table 4: Vehicle Information and State Variables

<table>
<thead>
<tr>
<th>Knowledge Level</th>
<th>Characteristic</th>
<th>Utility Component</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ex-ante Characteristics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observable ex-ante</td>
<td>Make/Model</td>
<td>make_model_j</td>
</tr>
<tr>
<td>&gt;&gt;</td>
<td>Price</td>
<td>price_j</td>
</tr>
<tr>
<td>&gt;&gt;</td>
<td>Mileage</td>
<td>mileage_j</td>
</tr>
<tr>
<td>&gt;&gt;</td>
<td>Age</td>
<td>age_j</td>
</tr>
<tr>
<td>&gt;&gt;</td>
<td>Color</td>
<td>color_j</td>
</tr>
<tr>
<td><strong>Vehicle History</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Learned upon visiting</td>
<td>Number of Previous</td>
<td>owners_j</td>
</tr>
<tr>
<td>a vehicle’s history</td>
<td>Owners</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Number of Accidents</td>
<td>accidents_j</td>
</tr>
<tr>
<td><strong>Vehicle Status</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Learned upon visiting</td>
<td>Number of Inspection</td>
<td>notes_j</td>
</tr>
<tr>
<td>inspection reports</td>
<td>Notes</td>
<td></td>
</tr>
<tr>
<td><strong>Perceived Quality/Fit</strong></td>
<td>Vehicle Photos</td>
<td></td>
</tr>
<tr>
<td>Learned after browsing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>one photo set</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: Vehicle indexes omitted. Ex-ante characteristics belong to users’ initial information set. The remaining characteristics are revealed through search actions.
Table 5: Correspondence of Search Actions and Learned Characteristics

<table>
<thead>
<tr>
<th>Search Action (s)</th>
<th>Meaning</th>
<th>Learned Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Vehicle History</td>
<td><em>owners</em>, <em>acc</em></td>
</tr>
<tr>
<td>2</td>
<td>Inspection Report</td>
<td><em>notes</em></td>
</tr>
<tr>
<td>3</td>
<td>Vehicle Photos</td>
<td><em>ν</em></td>
</tr>
</tbody>
</table>

Table 6: Estimated Correlation Table for the Errors Orthogonal to the Categorical Characteristics

<table>
<thead>
<tr>
<th></th>
<th>Price</th>
<th>Mileage</th>
<th>Insp. notes</th>
<th>Vehicle age</th>
<th>No. accidents</th>
<th>No. owners</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mileage</td>
<td>-0.66</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insp. notes</td>
<td>-0.31</td>
<td>0.37</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vehicle age</td>
<td>-0.73</td>
<td>0.71</td>
<td>0.41</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. accidents</td>
<td>-0.07</td>
<td>0.08</td>
<td>0.06</td>
<td>0.09</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>No. owners</td>
<td>-0.16</td>
<td>0.21</td>
<td>0.08</td>
<td>0.17</td>
<td>0.03</td>
<td>1</td>
</tr>
</tbody>
</table>

Note: Number of vehicles: 1,573. All correlation estimates significant with p-values below 0.01.

Table 7: Model Estimates: Preference Parameters

<table>
<thead>
<tr>
<th></th>
<th>Parameter Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Vehicle Characteristics</strong></td>
<td></td>
</tr>
<tr>
<td><em>price</em></td>
<td>-0.345 (0.008)</td>
</tr>
<tr>
<td><em>mileage</em></td>
<td>-0.746 (0.033)</td>
</tr>
<tr>
<td><em>notes</em></td>
<td>-0.451 (0.013)</td>
</tr>
<tr>
<td><em>age</em></td>
<td>-0.061 (0.003)</td>
</tr>
<tr>
<td><em>accidents</em></td>
<td>-0.775 (0.021)</td>
</tr>
<tr>
<td><em>owners</em></td>
<td>-0.391 (0.009)</td>
</tr>
<tr>
<td>Make/Model Dummies ✓</td>
<td></td>
</tr>
<tr>
<td>Color Dummies Dummies ✓</td>
<td></td>
</tr>
<tr>
<td>N= 1,000</td>
<td></td>
</tr>
</tbody>
</table>

Note: Standard errors in parentheses. All p-values below 1%. 

49
Table 8: Model Estimates: Search Cost Parameters

<table>
<thead>
<tr>
<th>Segment</th>
<th>Parameter Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Segment 1</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$c_{\text{photos}}$</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$c_{\text{vehicle,history}}$</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$c_{\text{insp.,report}}$</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Segment 2</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$c_{\text{photos}}$</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$c_{\text{vehicle,history}}$</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$c_{\text{insp.,report}}$</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Other Parameters</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\alpha$</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$c_{\text{vdp}}$</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Make/Model Dummies</td>
<td>✓</td>
</tr>
<tr>
<td>Color Dummies Dummies</td>
<td>✓</td>
</tr>
<tr>
<td>N= 1,000</td>
<td></td>
</tr>
</tbody>
</table>

Note: Standard errors in parentheses. All p-values below 1%. Size of segment 1, implied by $\hat{\alpha}$ is 33.8%.
Table 9: Model Estimates: Variance-Covariance Parameters

<table>
<thead>
<tr>
<th>Parameter Estimates</th>
<th>Variance-Covariance Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>-0.768 (0.002)</td>
</tr>
<tr>
<td>$s_2$</td>
<td>0.228 (0.007)</td>
</tr>
<tr>
<td>$s_3$</td>
<td>0.175 (0.004)</td>
</tr>
<tr>
<td>$s_4$</td>
<td>0.445 (0.005)</td>
</tr>
<tr>
<td>$s_5$</td>
<td>-0.306 (0.004)</td>
</tr>
<tr>
<td>$s_6$</td>
<td>0.057 (0.004)</td>
</tr>
</tbody>
</table>

Make/Model Dummies ✓
Color Dummies ✓
N = 1,000

Note: Standard errors in parentheses. All p-values below 1%.

Table 10: Estimated Variance-Covariance Matrix

\[
\hat{\Sigma}_0 = \begin{pmatrix}
1 & -0.66 & -0.31 & -0.73 & -0.07 & -0.16 & \mathbf{-0.77} \\
-0.66 & 1 & 0.37 & 0.71 & 0.08 & 0.21 & \mathbf{0.68} \\
-0.31 & 0.37 & 1 & 0.41 & 0.06 & 0.08 & \mathbf{0.45} \\
-0.73 & 0.71 & 0.41 & 1 & 0.09 & 0.17 & \mathbf{0.92} \\
-0.07 & 0.08 & 0.06 & 0.09 & 1 & 0.03 & \mathbf{-0.22} \\
-0.16 & 0.21 & 0.08 & 0.17 & 0.03 & 1 & \mathbf{0.22} \\
-\mathbf{0.77} & \mathbf{0.68} & \mathbf{0.45} & \mathbf{0.92} & -\mathbf{0.22} & \mathbf{0.22} & 1
\end{pmatrix}
\begin{pmatrix}
\text{Price} \\
\text{Mileage} \\
\text{Insp.notes} \\
\text{Vehicle age} \\
\text{No.accidents} \\
\text{No.owners} \\
\nu
\end{pmatrix}
\]

Note: In bold, covariance (cross-correlation) elements induced by estimated parameters $s_1$..$s_6$. Rightmost vector shows the corresponding characteristics.
Table 11: Search Moments from Dataset and Model

<table>
<thead>
<tr>
<th></th>
<th>Data</th>
<th>Model Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std. Dev.</td>
</tr>
<tr>
<td>Conversion Rate</td>
<td>.061</td>
<td>.239</td>
</tr>
<tr>
<td>Avg. N. Searches</td>
<td>2.202</td>
<td>3.22</td>
</tr>
<tr>
<td>Vehicle Histories</td>
<td>.656</td>
<td>1.53</td>
</tr>
<tr>
<td>Inspection Reports</td>
<td>.193</td>
<td>.900</td>
</tr>
<tr>
<td>Photo Sets</td>
<td>1.353</td>
<td>1.889</td>
</tr>
</tbody>
</table>

N: 1,000 consumers

Note: All model predictions are first averaged across simulations; then statistics are applied.

Table 12: Consumer Behavior Statistics With and Without Search Costs

<table>
<thead>
<tr>
<th></th>
<th>With Search Costs</th>
<th>No Search Costs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std. Dev.</td>
</tr>
<tr>
<td>Conversion Rate</td>
<td>0.046</td>
<td>0.197</td>
</tr>
<tr>
<td>Avg. N. Searches</td>
<td>1.532</td>
<td>2.016</td>
</tr>
<tr>
<td>Vehicle Histories</td>
<td>0.461</td>
<td>0.663</td>
</tr>
<tr>
<td>Inspection Reports</td>
<td>0.136</td>
<td>0.298</td>
</tr>
<tr>
<td>Photo Sets</td>
<td>0.934</td>
<td>1.242</td>
</tr>
</tbody>
</table>

N: 1,000 consumers

Note: All predictions are first averaged across simulations; then statistics are applied. Negative search cost parameters are held constant across scenarios.

Table 13: Conversion Effects of Exchanging Attributes Between Front and Detail Pages

<table>
<thead>
<tr>
<th>Front Page Attribute</th>
<th>Vehicle Detail Page Attribute</th>
<th>Vehicle History</th>
<th>Inspection Report</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price</td>
<td>+1.2%</td>
<td>+0.3%</td>
<td></td>
</tr>
<tr>
<td>Mileage</td>
<td>+0.9%</td>
<td>+0.4%</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>+0.8%</td>
<td>+0.3%</td>
<td></td>
</tr>
</tbody>
</table>

N: 1,000 consumers

Note: Above, impact of exchanging attributes in the front page with the ones in the vehicle detail page on mean conversions.
Table 14: Search Effects of Exchanging Attributes Between Front and Detail Pages

<table>
<thead>
<tr>
<th>Front Page Attribute</th>
<th>Vehicle History</th>
<th>Inspection Report</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price</td>
<td>-2.4%</td>
<td>-0.6%</td>
</tr>
<tr>
<td>Mileage</td>
<td>-2.1%</td>
<td>-1%</td>
</tr>
<tr>
<td>Age</td>
<td>-1.6%</td>
<td>-0.6%</td>
</tr>
</tbody>
</table>

N: 1,000 consumers

Note: Above, impact of exchanging attributes in the front page with the ones in the vehicle detail page on the number of searches.

Table 15: Percentage Changes to the First Search Option Values Resulting from Exchanging Attribute Orderings

<table>
<thead>
<tr>
<th>Front Page Attribute</th>
<th>Vehicle History</th>
<th>Inspection Report</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price</td>
<td>-6%</td>
<td>-3.4%</td>
</tr>
<tr>
<td>Mileage</td>
<td>-5.8%</td>
<td>-3.5%</td>
</tr>
<tr>
<td>Age</td>
<td>-4.9%</td>
<td>-2.6%</td>
</tr>
</tbody>
</table>

N: 1,000 consumers

Note: Above, impact of exchanging attributes in the front page with the ones in the vehicle detail page on search option values.

Table 16: Changes in Residual Variances from Exchanging Attribute Orderings

<table>
<thead>
<tr>
<th>Front Page Attribute</th>
<th>Vehicle History</th>
<th>Inspection Report</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price</td>
<td>-0.5%</td>
<td>-0.2%</td>
</tr>
<tr>
<td>Mileage</td>
<td>+13.4%</td>
<td>+15.2%</td>
</tr>
<tr>
<td>Age</td>
<td>+7.2%</td>
<td>+6%</td>
</tr>
</tbody>
</table>

Note: Above, impact of exchanging attributes in the front page with the ones in the vehicle detail page on mean search option values.
8 Appendix

8.1 Data Cleaning

A few unusual patterns emerged while analyzing the clickstream data. First, a few sessions were very short, resembling ‘bouncing’ behavior in which a user visits the website and leaves without triggering additional action flags, often adding up to a very short overall session time. This type of activity is typically generated by users who visited the website by mistake, or by ‘bots’ scraping the Internet for content. These users are likely to have very different objectives than purchasing a car on the platform. We control for this type of behavior by eliminating sessions that triggered only a page visit or that took less than 5 seconds in terms of the overall measured activity (e.g., a session that triggers two events, with only 3 seconds apart, is eliminated). We also removed users with very unusual browsing behaviors. For example, users who exhibited activity across multiple IP addresses belonging to different countries within a very short period of time were eliminated from the sample.

8.2 Variance-Covariance Matrix Decomposition

The variance-covariance matrix

\[
\Sigma_0 = \begin{pmatrix}
\sum_{z} & \sigma_{z\nu} \\
\sigma'_{z\nu} & \sigma^2_{\nu}
\end{pmatrix}
\]  \hspace{1cm} (39)

has a known component \( \Sigma \), given in Table 6, and an unknown border, to be estimated. We take the Cholesky decomposition of \( \Sigma_0 \), and parameterize its border with vector \( [s_1, s_2, ..., s_6, \sqrt{1 - \sum_{j=1}^{6} s_j^2}] \). This yields the decomposed matrix in Table 17.
Table 17: Cholesky Decomposition of Bordered Variance-Covariance Matrix

\[ U = \begin{pmatrix}
1 & -0.664415 & -0.305486 & -0.727454 & -0.0660357 & -0.159545 & s_1 \\
0 & 0.747364 & 0.217517 & 0.29733 & 0.0471439 & 0.143938 & s_2 \\
0 & 0 & 0.927019 & 0.128614 & 0.0370378 & 0.00187495 & s_3 \\
0 & 0 & 0 & 0.604867 & 0.0352993 & 0.016755 & s_4 \\
0 & 0 & 0 & 0 & 0.995389 & 0.00822777 & s_5 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.976461 & s_6 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \sqrt{1 - \sum_{j=1}^{6} s_j^2}
\end{pmatrix} \]

The matrix \( \Sigma \), obtained by

\[ \Sigma = U'U \] (40)

conforms to matrix \( \Sigma_0 \) in that the upper-left block is equal to \( \Sigma \), and the element \( \sqrt{1 - \sum_{j=1}^{6} s_j^2} \) in \( U \) implies that resulting lower-right parameter \( \sigma_\nu^2 \) is normalized to one. In addition, vector \( \sigma_\nu \) in \( \Sigma_0 \) is obtained by linear combinations of parameters \( \{s_1..s_6\} \). Finally, matrix \( \Sigma_0 \) is guaranteed to be p.s.d. by the variational theorem of linear algebra, as long as \( \sum_{j=1}^{6} s_j^2 \leq 1 \).
References


