View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Universidad Carlos III de Madrid e-Archivo

Universidad
Carlos III de Madrid
www.uc3m.es

TESIS DOCTORAL

Three essays on conjoint analysis: optimal design and estimation of endogenous consideration sets

Autor:

Agata Leszkiewicz

Director/es:

Mercedes Esteban-Bravo, PhD

Jose M. Vidal-Sanz, PhD

DEPARTAMENTO ECONOMÍA DE LA EMPRESA

Getafe, January 2014

Universidad
Carlos III de Madrid
www.uc3m.es

TESIS DOCTORAL

Three essays on conjoint analysis: optimal design and estimation of endogenous consideration sets

Autor: Agata Leszkiewicz

Director/es: Mercedes Esteban-Bravo, PhD

José M. Vidal-Sanz, PhD

Firma del Tribunal Calificador:

Firma

Presidente:

Vocal:

Secretario:

Calificación:

Getafe, de de

Acknowledgments

First and foremost, I owe a debt of gratitude to my PhD advisors, Mercedes Esteban-Bravo and Jose M. Vidal-Sanz. The influence of their guidance and motivation on the final form of this dissertation cannot be overstated. I feel very fortunate to have had the opportunity to work closely with Mercedes and Jose in many areas of the academic life. During these years I have gotten to know them not only as brilliant marketing modelers, but also as organized and trustworthy professionals. Collaboration with Mercedes and Jose shaped me into a mature researcher. I would also like to thank Mercedes for inviting me to participate in her research project, which provided financial support to this thesis.

I also wish to thank Don Lehmann for his hospitality and mentorship during my research visit at Columbia Business School. It was an invaluable opportunity to learn from him and I deeply appreciate his support, feedback and career advice. My thanks go to Leonard Lee, Oded Netzer and Nicholas Reinholtz for their warm reception in New York and comments on my work. I cannot emphasize enough how much I gained from (and enjoyed) my research stay at Columbia.

I am obliged to the entire marketing team at Carlos 3. I appreciate the support and feedback of Nora Lado and Alicia Barroso during the first years of my teaching experience. My thanks go to James Nelson, Lola Duque and Fabrizio Cesaroni for their words of encouragement. I am grateful to Goki and Vardan for clearing the paths for me on the PhD journey, for their sincere advice and for their friendship.

I wish to thank Manuel Bagües and Encarna Guillamón, with whom I worked closely in the Department, for many challenging discussions, advice and for being my referees. I would like to mention other faculty members who supported me at different moments at Carlos 3: Josep Tribó, Jaime Ortega, Pablo Ruiz-Verdú and Esther Ruíz. I warmly thank Agnieszka Szczepańska-Álvarez from the Poznań University of Life Sciences for the friendly review of my paper.

During the years I spent in Madrid I was fortunate to have met many extraordinary people. I enjoyed the friendship of Ana Laura, Dilan, Juliana and Su-Ping, who have been there for me through all the ups and downs. My special thanks go to Adolfo, Agnieszka, Ana Maria, Argyro, Borbala, Emanuele, Han-Chiang and Jonatan. I leave Madrid hoping that soon our paths will cross again.

I am grateful to Lukas for his patience. He has been a true partner on this journey, always supporting me in my objectives. I thank my whole family for their unlimited love, inspiration and constant encouragement. Finally, I am indebted to my mother for proofreading of parts of my work.

I thankfully acknowledge the financial support of the Department of Business Administration at the University Carlos III of Madrid, the University Carlos III of Madrid (Programa Propio de Investigación), the Ministry of Science and Innovation in Spain (research grant ECO2011-30198), and the Education Council of the Autonomous Community of Madrid (research grant S2009/ESP-1594).

I dedicate this dissertation to my family.
Agata

Abstract

Over many years conjoint analysis has become the favourite tool among marketing practitioners and scholars for learning consumer preferences towards new products or services. Its wide acceptance is substantiated by the high validity of conjoint results in numerous successful implementations among a variety of industries and applications. Additionally, this experimental method elicits respondents' preference information in a natural and effective way.

One of the main challenges in conjoint analysis is to efficiently estimate consumer preferences towards more and more complex products from a relatively small sample of observations because respondent's wear-out contaminates the data quality. Therefore the choice of sample products to be evaluated by the respondent (the design) is as much as relevant as the efficient estimation. This thesis contributes to both research areas, focusing on the optimal design of experiments (essay one and two) and the estimation of random consideration sets (essay three).

Each of the essays addresses relevant research gaps and can be of interest to both marketing managers as well as academicians. The main contributions of this thesis can be summarized as follows:

- The first essay proposes a general flexible approach to build optimal designs for linear conjoint models. We do not compute good designs, but the best ones according to the size (trace or determinant) of the information matrix of the associated estimators. Additionally, we propose the solution to the problem of repeated stimuli in optimal designs obtained by numerical methods. In most of comparative examples our approach is faster than the existing software for Conjoint Analysis, while achieving the same efficiency of designs. This is an important quality for the applications in an online context. This approach is also more flexible than traditional design methodology: it handles continuous, discrete and mixed attribute types. We demonstrate the suitability of this approach for conjoint analysis with rank data and ratings (a case of an individual respondent and a panel). Under certain assumptions this approach can also be applied in the context of discrete choice experiments.
- In the essay 2 we propose a novel method to construct robust efficient designs for conjoint

experiments, where design optimization is more problematic, because the covariance matrix depends on the unknown parameter. In fact this occurs in many nonlinear models commonly considered in conjoint analysis literature, including the preferred choice-based conjoint analysis. In such cases the researcher is forced to make strong assumptions about unknown parameters and to implement an experimental design not knowing its true efficiency. We propose a solution to this puzzle, which is robust even if we do not have a good prior guess about consumer preferences. We demonstrate that benchmark designs perform well only if the assumed parameter is close to true values, which is rarely the case, otherwise there is no need to implement the experiment. On the other hand, our worst-case designs perform well under a variety of scenarios and are more robust to misspecification of parameters.

- Essay 3 contributes with a method to estimate consideration sets which are endogenous to respondent preferences. Consideration sets arise when consumers use decision rules to simplify difficult choices, for example when evaluating a wide assortment of complex products. This happens because rationally bounded respondents often skip potentially interesting options, for example due to lack of information (brand unawareness), perceptual limitations (low attention or low salience), or *halo effect*. Research in consumer behaviour established that consumers choose in two stages: first they screen off products whose attributes do not satisfy certain criteria, and then select the best alternative according to their preference order (over the considered options). Traditional CA focuses on the second step, but more recently methods incorporating both steps were developed. However, they are always considered to be independent, while the halo effect clearly leads to endogeneity. If the cognitive process is influenced by the overall affective impression of the product, we cannot assume that the screening-off is independent from the evaluative step. To test this behavior we conduct an online experiment of lunch menu entrees using Amazon MTurk sample.

Resumen

A lo largo de los años, el “Análisis Conjunto” se ha convertido en una de las herramientas más extendidas entre los profesionales y académicos de marketing. Se trata de un método experimental para estudiar la función de utilidad que representa las preferencias de los consumidores sobre productos o servicios definidos mediante diversos atributos. Su enorme popularidad se basa en la validez y utilidad de los resultados obtenidos en multitud de estudios aplicados a todo tipo de industrias. Se utiliza regularmente para problemas tales como diseño de nuevos productos, análisis de segmentación, predicción de cuotas de mercado, o fijación de precios.

En el análisis conjunto, se mide la utilidad que uno o varios consumidores asocian a diversos productos, y se estima un modelo paramétrico de la función de utilidad a partir de dichos datos usando métodos de regresión en sus diversas variantes. Uno de los principales retos del análisis conjunto es estimar eficientemente los parámetros de la función de utilidad del consumidor hacia productos cada vez más complejos, y hacerlo a partir de una muestra relativamente pequeña de observaciones debido a que en experimentos prolongados la fatiga de los encuestados contamina la calidad de los datos. La eficiencia de los estimadores es esencial para ello, y dicha eficiencia depende de los productos evaluados. Por tanto, la elección de los productos de la muestra que serán evaluados por el encuestado (el diseño) es clave para el éxito del estudio. La primera parte de esta tesis contribuye al diseño óptimo de experimentos (ensayos uno y dos, que se centran respectivamente en modelos lineales en parámetros, y modelos no lineales). Pero la función de utilidad puede presentar discontinuidades. A menudo el consumidor simplifica la decisión aplicando reglas heurísticas, que de facto introducen una discontinuidad. Estas reglas se denominan conjuntos de consideración: los productos que cumplen la regla son evaluados con la función de utilidad usual, el resto son descartados o evaluados con una utilidad diferente (especialmente baja) que tiende a descartarlos. La literatura ha estudiado la estimación de este tipo de modelos suponiendo que la decisión de consideración está dada exógenamente. Pero sin embargo, las reglas heurísticas pueden ser endógenas. Hay sesgos de percepción que relacionan utilidad y la forma en se perciben los atributos. El tercer estudio de esta tesis considera modelos con conjuntos

de consideración endógenos.

Cada uno de los ensayos cubre problemas de investigación relevantes y puede resultar de interés tanto para managers de marketing como para académicos. Las principales aportaciones de esta tesis pueden resumirse en lo siguiente:

- El primer ensayo presenta una metodología general y flexible para generar diseños experimentales óptimos exactos para modelos lineales, con aplicación a multitud de variantes dentro del análisis conjunto. Se presentan algoritmos para calcular los diseños óptimos mediante métodos de Newton, minimizando el tamaño (traza o determinante) de la matriz de covarianzas de los estimadores asociados. En la mayoría de los ejemplos comparativos nuestro enfoque resulta más rápido que los softwares existentes para Análisis Conjunto, al tiempo que alcanza la misma eficiencia de los diseños. Nuestro enfoque es también más flexible que la metodología de diseño tradicional: maneja tipos de atributos continuos, discretos y mixtos. Demostramos la validez de este enfoque para el análisis conjunto con datos de rango de preferencias y valoraciones (un caso de un encuestado individual y un panel). Bajo ciertos supuestos, este enfoque puede también ser aplicado en el contexto de experimentos de elección discreta.

- En el segundo ensayo nos centramos en modelos de preferencia cuyos estimadores tienen matrices de covarianzas no pivotales (dependientes del parámetro a estimar). Esto sucede por ejemplo en modelos de preferencia no lineales en parámetros, así como modelos de elección como el popular Logit Multinomial. En tal caso la minimización de la matriz de covarianzas no es posible. La literatura ha considerado algunas soluciones como suponer una hipótesis acerca de este valor a fin de poder minimizar en el diseño la traza o determinante de la matriz de covarianzas. Pero estos diseños de referencia funcionan bien solo si el parámetro asumido es cercano a los valores reales (esto raramente sucede en la práctica, o de lo contrario no hay necesidad de implementar el experimento). En este ensayo proponemos un método para construir diseños robustos basados en algoritmos minimax, y los comparamos con los que normalmente se aplican en una gran variedad de escenarios. Nuestros diseños funcionan son más robustos a errores de los parámetros, reduciendo el riesgo de estimadores altamente ineficientes (que en cambio está presente en los otros métodos).

- El ensayo 3 aporta un método para estimar conjuntos de consideración que son endógenos a las preferencias de los encuestados. Conjuntos de consideración surgen cuando los consumidores usan reglas de decisión para simplificar la dificultad de las elecciones, lo cual requiere una significativa búsqueda de información y esfuerzos cognitivos (por ejemplo, evaluar una amplia variedad de productos complejos). Esto ocurre porque racionalmente limitados consumidores a menudo pasan por alto opciones potencialmente interesantes, por ejemplo, debido a una falta de información (desconocimiento de la marca), limitaciones de percepción (baja atención o prominencia), o efecto de halo. La investigación en el comportamiento de los consumidores establece que los consumidores eligen en dos fases: primero eliminan productos que no satisfacen ciertos criterios y luego seleccionan las mejores alternativas de acuerdo a su orden de preferencia (de acuerdo a las opciones consideradas). El Análisis Conjunto convencional, se centra en el segundo paso, pero recientemente, se han desarrollado métodos incorporando ambos pasos. Sin embargo, son siempre considerados independientes, mientras que el efecto de halo claramente lleva a la endogeneidad del proceso de consideración. Si el proceso cognitivo está influenciado por una impresión general afectiva del producto, no podemos asumir que la eliminación sea independiente del proceso evaluativo. Para probar este comportamiento llevamos a cabo un experimento online sobre entrantes en menús de comida usando una muestra desde Amazon MTurk.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Conjoint Analysis

Individual tastes and preferences are the starting point of customers' decision making and purchasing choices of products or services. It has become essential for consumer-oriented firms to understand how potential buyers value different product features and how they perceive the overall product offerings on the market. Decision makers need to listen in to the voice of the empowered consumers, while planning strategic marketing actions such as design of new products, repositioning, pricing or targeted advertising. However, in practice individuals are not capable of providing reliable information about their preferences when evaluating or assessing the importance of separate product characteristics.

A variety of methods embraced under the umbrella name "conjoint analysis" provide marketing practitioners with a reliable tool to elicit consumer preferences towards multi-attribute products and/or services. The work of Luce (1966) in psychometrics is traditionally viewed as the origin of conjoint analysis, while this method also has roots in multiattribute utility theory (Debreu 1960; Lancaster 1971). Its diffusion in marketing began with the seminal paper of Green and Rao (1971), followed by the work of Johnson (1974) and Louviere and Woodworth (1983) on discrete-choice experiments.

In a typical conjoint study, a product is perceived as a bundle of attributes and the researcher varies these features creating a number of unique product concepts (combinations of features). Respondents are then asked to evaluate each stimulus, which forces them to make difficult trade-offs between attributes, since all features are considered jointly. It is a decompositional approach: the contribution of each attribute in the overall product utility is inferred from evaluation of entire product concepts. This is an efficient way to learn about respondents' true preferences and the validity of this approach has been proven by many successful commercial applications (for a review see Wittink et al. 1982; Wittink and Cattin 1989; Wittink et al. 1994).

A variety of topics and problems addressed in real-life conjoint analysis studies is impressive, answering questions of strategic importance to marketing decision makers. Below I briefly list some a few examples of interesting CA applications in different areas of marketing mix:

- **Product**. Design of new products is the most straightforward CA application (Green et al. 1981; Hoeffler 2003; Drezè and Zufryden 1998; Kohli and Krishnamurti 1987; Wind et al. 1989), including product redesign and prediction of consumers' upgrading decisions (Kim 2000; Kim and Srinivasan 2006). The method can be applied to the problem of optimal composition of product lines (McBride and Zufryden 1988; Kohli and Sukumar 1990; Belloni et al. 2008; Chen and Hausman 2000), product bundles (Farquhar and Rao 1976; Chung and Rao 2003), as well as category assortment optimization (Bradlow and Rao 2000). Finally, there are implications for product positioning decisions (Green and Krieger 1993; Wind et al. 1989; Green and Krieger 1992) and benefit-based segmentation (Kamakura 1988; Green and Krieger 1991; Desarbo et al. 1995; Vriens et al. 1996).

- **Price**. The applications in the area of pricing are not limited to the study of price-demand relationship (Mahajan et al. 1982), or evaluation of willingness to pay for a product or service (Jain et al. 1999; Roe et al. 2001; Telser and Zweifel 2002). Other interesting topics include the estimation of reservation prices (Kohli and Mahajan 1991; Jedidi and Zhang 2002), construction of the pricing systems accounting for needs of different segments (Currim et al. 1981; Desarbo et al. 1995; Green and Krieger 1990), optimal construction of nonlinear pricing schemes (Iyengar et al. 2008), pricing of product bundles (Chung and

Rao 2003), game-theoretical models of price competition (Blattberg and Wisniewski 1989), or estimating price effects related to the budget constraint and the role of price as a signal of quality (Rao and Sattler 2007).

- **Distribution**. The research in this area studied the consumer choice of the shopping center (Oppewal et al. 1994), the tendency to combine shopping purposes and destinations (Dellaert et al. 1998), the choice of a vendor and supplier Wuyts et al. (2004), and purchase location influences respondents' preferences and willingness to pay (Martínez et al. 2006).

- **Advertising**. Some of the developments with implications to advertising include the optimal incentive scheme for sales force (Darmon 1979), formulating optimal push strategies (Levy et al. 1983), and relationship between advertising intensity and preferences (D'Souza and Rao 1995).

The critical milestone to the diffusion of the method was the development of a dedicated, easy-to-use software for conjoint analysis in 1980s. Bretton Clark's Conjoint Designer (Herman 1988) was the first tool for the implementation of the whole conjoint study from the design, data collection and estimation of preferences, and was also equipped in the simple market simulators. It was considered an "industry standard" for traditional conjoint analysis experiments and benchmark for validity of new methods Carroll and Green (1995). Another breakthrough was the Adaptive Conjoint Analysis (ACA) introduced by Johnson (1987) of Sawtooth Software - the first package for computer-assisted questionnaires (substituting the traditional pen-and-pencil methods). In ACA's adaptive questionnaire design the respondent is asked in detail only about attributes of the greatest importance to him. This way, ACA can study up to 30 attributes, each with up to 15 attribute levels (Sawtooth Software 2007).

Nowadays, conjoint analysis packages are available in many state-of-the art programs for data analysis such as SPSS or SAS. The field continues to flourish enriched by methodological contributions from many areas such as statistics, econometrics, and operations research (Toubia et al. 2003, 2004; Evgeniou et al. 2005; Yee et al. 2007).

1.2 Conjoint Analysis Process

The implementation of a typical conjoint analysis involves several steps and a process of subsequent, interdependent decisions on the researcher's part. Green and Srinivasan (1978) describe several experimental phases, emphasizing the interconnectedness of choices made throughout the conjoint experiment: the definition of product attributes and of the preference model, the choice of data collection method, stimuli assignment to respondents (experimental design), presentation of product concepts, selection of data collection method, choice of measurement scale, estimation and market simulation. Gustafsson et al. (2007) provide an updated flow diagram of conjoint analysis. The main focus of this thesis are the methodological issues in conjoint analysis: the experimental design and the estimation, which are discussed in depth in Chapters 2–4. However, the reader not acquainted with conjoint analysis may benefit at this point from a general overview of the process.

Model definition

The first modeling decision involves the definition of product attributes and the preference model. Traditionally, the researcher identifies relevant product features from the experimental pretests, focus groups or relies on managers' expertise (Green and Srinivasan 1978), but more recently there has been interest in text-mining techniques for extraction of attributes (and preference estimation) from user-generated content such as online forums or customer reviews (Decker and Trusov 2010; Lee and Bradlow 2011; Netzer et al. 2012). *Part-worth* preference model is the most flexible specification of utility function commonly used in conjoint analysis, which assumes that the total benefit towards the product is the sum of partial benefits related to product attributes and/or attribute levels, $u(x) = \sum_{j=1}^{K} \beta_j f_j(x_j)$ (Green and Rao 1971; Green and Srinivasan 1978). Ideal-vector and ideal-point specifications are special cases of the part-worth model, and are also used in conjoint modeling (comparison of different functional forms can be found in Krishnamurthi and Wittink 1991).

Stimuli presentation

Another decision in the process is how to present stimuli to the respondent. Ideally we would like to show full profiles, meaning products whose all attributes are described (Green and Rao 1971). However the respondents' task becomes difficult when evaluating complex products, therefore a variety of procedures have been proposed to handle a large number of attributes: 1) comparison of *pairs of products* (Thurstone 1927; Bemmaor and Wagner 2000); 2) comparison of *pairs of attributes*: trade-off procedure of Johnson (1974); 3) evaluation of products defined over a *subset* of attributes: partial profile method mentioned by Green (1974), and later improved by Alba and Cooke (2004); Bradlow et al. (2004); Rubin (2004); or 4) evaluation of single attributes: the *self-explicated* approach (Leigh et al. 1984; Srinivasan 1988; van der Lans and Heiser 1992). Although the latter is essentially a compositional approach, therefore does predict consumer trade-offs as well as full-profile conjoint, it performs well when the number of attributes is large (Srinivasan and Park 1997). Additionally, some of the hybrid procedures for preference elicitation are constructed as a combination of above approaches (Sawtooth Software 2007; Netzer and Srinivasan 2011). Various studies provide empirical evidence that ACA outperforms full-profile approach, specifically when the number of attributes is bigger than 5 and in the absence of a substantial warm-up task (Huber et al. 1991; van der Lans et al. 1992; Huber et al. 1993). On the other hand, Hauser and Toubia (2005) show that ACA's adaptive questionnaire lead to endogeneity and biased partworth estimates.

Experimental design

The experimental design determines which product profiles will be evaluated by every respondent and is critical to assure the quality of CA results. In conjoint analysis the design is constructed by varying features to create hypothetical products. The implementation of the complete design would require that each subject evaluates all possible product profiles, which is feasible only for simple products with very few attributes. Therefore, why is the experimental design relevant? A good design assures the reliable estimates of preferences from a small sample of stimuli and knowing the preference structure for different features the researcher can extrapolate the results to other product offerings on the market (not evaluated by the respondents). The number of alternative products to be evaluated by the individual is relevant because there are measurement errors due to respondent wear-out leading to poor data quality.

There are two big approaches for experimental design in the statistics literature: the Design of Experiments (DoE) theory and the optimal design approach. In marketing research, including conjoint analysis, researchers often use fractional factorial designs from DoE, because they possess desirable properties such as orthogonality. Fractional factorial are created by systematically reducing the complete design so that the attributes are kept independent (orthogonal). Such designs (and others) are available in ready-to-use experimental tables for specific, symmetric or asymmetric problems. The classic experimental design considers "good" designs for well-structured and well-defined problems. For a detailed review see e.g. Francis G. Giesbrecht (2004); Montgomery (2005), and the classic textbook of Cochran and Cox (1957).

The more flexible Optimal Design approach aims at finding the best possible design for a given research problem. An optimal design maximizes the information about the preferences (precision of estimates) given a certain sample size, or alternatively minimizes the covariance of parameters. This implies that suboptimal designs require a larger sample size to estimate the parameters with the same precision as the optimal design, increasing the market research cost and rating contamination caused by respondent's fatigue. An optimal design is obtained by minimizing the size of covariance matrix. The general theory was proposed by Kiefer (1959) and the most popular design criteria are: D-optimality - minimizing the determinant of covariance matrix; and A-optimality - minimizing the trace of covariance matrix. However, optimal design approach cannot be directly applied in CA, as it usually leads to experiments with repeated product profiles.

The design of experiments is a fundamental problem in marketing research. Green (1974) and Green and Srinivasan (1978) advocate the use of fractional factorial designs in conjoint analysis; Louviere (1988a) proposed a design construction method for conjoint analysis based on stated choices; Sándor and Wedel (2001, 2002, 2005) developed several utility-balanced choice designs for the logit and mixed logit model. A review and comparison of orthogonal and optimal designs can be found in Lazari and Anderson (1994); Kuhfeld et al. (1994).

A large part of this dissertation is devoted to the optimal design of conjoint experiments. In Section 1.3 I formally introduce the conjoint preference model and discuss the experimental design problem. Further contributions to this research area are made in Chapter 2, devoted to the optimal design for linear models, and Chapter 3, which presents the robust worst-case design for nonlinear models. Additional relevant developments and research gaps are discussed in both articles.

Profile presentation

As far as the presentation of incentives is concerned, they are frequently verbal or paragraph descriptions, often supplemented with graphical product representations (Green and Srinivasan 1978; Cattin and Wittink 1982; Wittink et al. 1994). Subject to the available experimental budget, respondents may test and evaluate actual experimentally designed product prototypes (Green et al. 2001). On the other hand, online administration of questionnaires facilitates the use of multimedia and less expensive virtual prototypes in form of images or video clips, see for example Dahan and Srinivasan (2000) and Intille et al. (2002). However, these methods seem to be specific to a given application. Finally, in a conjoint study about packaged apple snacks Jaeger et al. (2001) compared two forms of stimuli presentation (physical prototype vs. realistic pictorial representation) and did not find any significant differences in choice decisions.

Data collection

Conjoint analysis questionnaires can be administered using traditional survey channels such as personal interviews, mail surveys, and over the Internet. Initially, some CA tasks were also administered on the telephone (Cattin and Wittink 1982), however this method is suitable only for very simple studies. Until the 1980s conjoint analysis was almost exclusively done by paper and pencil in the laboratory or via mail (Wittink and Cattin 1989; Wittink et al. 1994), but the development of ACA (Johnson 1987) shifted the balance towards computerized questionnaires

(CAPI - computer assisted personal interviews). The end of the century marked the start of prevalence of Internet surveys (Witt 1997; Orme and King 1998): for example Foytik (1999) gives a list of "dos and don'ts" for online CA, and Melles et al. (2000) compares online CA with CAPI finding that both methods are equivalent in terms of reliability and predictive validity. Melles et al. (2000) also points out that data obtained from Internet CA needs more screening and cleaning, because the questionnaire is self-administered and respondent's cognitive abilities are low. With longer studies the subjects become quickly disinterested, which may heavily distort the results and lead to incorrect managerial decisions. There are methodologies specifically designed to obtain more data from the respondents in an online context: capturing behavior on the website (Drezè and Zufryden 1998), creating adaptive questionnaires in real time based on subjects' responses (Dahan et al. 2002; Netzer and Srinivasan 2011), or eliciting preferences from people's Internet behavior (De Bruyn et al. 2008). Finally, Ding et al. (2005) show that conjoint results can be improved if the study is conducted in the realistic setting, and the topic of the study is aligned with a prize for completing it.

Preference measurement scale

The next important decision for the conjoint process is to choose the preference measurement scale. Typically, the respondent is asked to evaluate the presented product profiles by: (1) ranking them top-to-bottom according to their preference; (2) rating them using a continuous or a Likert scale; (3) choosing one alternative from a set of available options.

The rating scale conveys more information than rankings, because apart from the preference order it also expresses the intensity. Therefore the rating data can be transformed into rankings but not the other way round. Additionally, the ratings are traditionally considered a metric scale (assuming approximately ordinal scale properties) therefore can be estimated by standard tools such as OLS, while the rankings require non-metric algorithms which will be discussed in the next section. There is mixed evidence which of the two scales performs better in terms of predictive validity: Carmone et al. (1978) provides the evidence that ratings have higher predictive validity, the results of Scott and Keiser's 1984 favor the rankings, while Green and Srinivasan

(1978) posits that both methods are equivalent.

However, transforming the ranking and rating data into choices is problematic. DeSarbo and Green (1984) point out that predictions of consumer's choice based on ranking or rating conjoint may not be accurate, because: (1) product profiles are never equal to real products, (2) the model usually estimates only main effects and maybe a few two-way interaction effects, and finally (3) conjoint analysis assumes equal effects of marketing variables across different suppliers. In conjoint analysis based on consumer choice (CBC) the respondents task is more similar to the way people behave in the marketplace, because the alternatives are presented in a competitive context.

Carroll and Green (1995) discuss several advantages and disadvantages of CBC over traditional conjoint. On the one hand, the choice tasks are more natural than ranking or rating and prediction of market shares does not require any deterministic rules. Moreover, the theory underlying the logit model is well-grounded (McFadden 1974) and choice probabilities can be directly and efficiently estimated. On the other hand, the estimation of choice models requires larger amount of data and only recently the usage of Bayesian methods permitted the estimation of individual-level parameters (see e.g. Cattin et al. 1983; Allenby et al. 2005; Toubia et al. 2007). Additionally, choice models provide little information about the non-chosen alternatives and IIA property of multinomial logit can be a serious limitation in marketing applications (see Kamakura and Srivastava 1984).

Finally, Elrod et al. (1992) provide an empirical comparison of different conjoint approaches (traditional and choice-based). Their results suggest that neither of approaches can be favored solely by their predictive ability, because on average they predict equally well. The choice of the method should depend rather on the purposes of conjoint study. If market share prediction is the central interest, choice-based approach may be more appropriate.

Estimation

The statistical analysis will depend on the preference model and the utility measurement scale for respondents. Initially, ordinal measurement was common (ranking of profiles), and to that end non-metric algorithms were developed. Estimation techniques for this kind of data include MONANOVA, a dedicated technique developed for CA by Kruskal (1965) that finds a monotone transformation of the data to achieve the highest possible percentage of variance accounted for by main effects; PREFMAP (Carroll 1972) – a mathematical programming model, which finds the respondent's ideal point from their preference rankings; LINMAP – a linear programming model to determine the attribute weights and the coordinates of consumer's ideal point (Srinivasan and Shocker 1973a, b); Johnson's non-metric trade-off procedure (Johnson 1974). In the classic, "metric" CA (rating of profiles) the coefficients are often estimated with OLS procedures, which with dummy variables is basically equivalent to the analysis of variance. On the other hand, choice-based CA models are usually estimated with Maximum Likelihood methods (we obtain the Multinomial Logit model assuming that y_t is a latent variable and ε_t has a type I extreme value distribution). Choice-based CA is nowadays widely applied, but from the econometric point of view the hypotheses about the distribution of ε_t are stronger than in the classic CA which is more robust to specification errors. For a literature review and description of the methods applications, see Gustafsson et al. (2007).

Market simulators

One of the important implications of CA is the forecasting of market shares for new products. There are three main deterministic rules to transform estimated utilities into consumer choice decisions: maximum utility (first-choice) rule, Bradley-Terry-Luce (BTL) model, and the logit model. In case of CBC it is not necessary to apply those rules because the choice probabilities are directly estimated from the model. All of above methods are incorporated in popular software packages for conjoint analysis, such as SPSS, SAS or Sawtooth Software.

The first choice rule assumes that each subject will buy the product of the highest utility to them (with certainty), and market shares are obtained by averaging the probabilities across subjects. Unlike the first-choice rule the BLT and logit method do not assign the whole probability mass to one product. The choice probabilities are rather a continuous function of predicted utilities. In case of BLT this is a linear function, and the probability is the ratio of a profile's utility to that for all simulation profiles, averaged across all respondents. The logit rule assumes a exponential function of predicted utilities and divides the exponentiated predicted utilities by the sum of exponentiated utilities (for every respondent).

The empirical evidence about predictive validity of those methods is mixed. DeSarbo and Green (1984) and Louviere (1988b) pointed out that application of maximum utility rule is problematic because a deterministic rule is applied to predict a probabilistic phenomenon. Further problems arise due to (1) intertemporal instability of tastes and beliefs of consumers, (2) an artificial assumption of perfect information about their attributes, and (3) assumption that there are no income, time or other constraints, which may influence individual's choice. On the other hand Green and Krieger (1988) and Finkbeiner (1988) demonstrated that first-choice rule is suitable for surveys about high-involvement products.

1.3 Benchmark Model in Conjoint Analysis

Let us turn to the formal specification of the preference model and to the methodological issues in conjoint analysis which arise from the choice of the design and measurement scale. The base model is the case of an individual respondent, however it is straightforward to extend the approaches developed in Chapter 2 and Chapter 3 to homogeneous consumer segments. Moreover, in Chapter 2.5 we discuss the case of consumer segments with heterogeneous intercept, while the method presented in Chapter 3 is robust to deviations of assumptions about consumer preferences.

Let's assume a multi-attribute product, x, defined by k continuous and L discrete attributes, each taking $J = [J_1, \ldots, J_L]'$ levels. A product profile shown to the respondent is represented by the $\left(k + \sum_{i=1}^{L} J_i\right) \times 1$ vector x_t of deterministic regressors in a compact set χ of an Euclidean space defining the attributes (discrete dummy and/or continuous variables). Individual's overall preferences for a product are described by a utility function parametric model $\left\{U\left(x, \beta^0\right) : \beta^0 \in \Theta \subset \mathbb{R}^p\right\}$. The vector β^0 is a $p \times 1$ vector of unknown parameters. Note that we allow $p \neq \left(k + \sum_{i=1}^{L} J_i\right)$, because consumer preference function may contain an intercept, depend on interactions between attributes (or other variable transformations such as squared regressors), and for estimation purposes we have to omit a level in each categorical variable to eliminate multicollinearity. The experimental sample, $\{x_t\}_{t=1}^{T}$, is composed of T profiles shown to the respondent and $T \geq p$. The responses, y_t, represent respondent's utility of each product profile, evaluated at the attitudinal scale (typically based on ratings, rankings or choice). Measures are affected by an error shock ε_t

$$y_t = U\left(x_t, \beta^0\right) + \varepsilon_t, \quad t = 1, ..., T,$$

where ε_t are regarded as mutually independent random shocks, satisfying $E[\varepsilon_t] = 0$ and $E[\varepsilon_t^2] = \sigma^2$. Stacking the data in matrices the model is $y = U\left(X, \beta^0\right) + \varepsilon$, where y, ε are $T \times 1$ vectors, $X = (f(x_1), ..., f(x_T))' \in \chi^T$ is a full rank design matrix, whose row t contains $f(x_t)'$. $f(\cdot)$ is a known continuous mapping from χ to $\mathbb{R}^p$ whose coordinates are linearly independent and may include an intercept, discrete interactions (products of dummies), or product of continuous regressors (to define multivariate polynomials similarly to surface response models). The function f could also have a known local maximum (self-explicated ideal point). The goal of CA is the estimation of the parameters β^0 from experimental data and as result to predict preferences towards different products versions.

This dissertation will focus generally on the methodological and statistical aspects of conjoint analysis: the design, the choice of measurement scale and the estimation. These linked decisions are essential for assuring the quality of conjoint results and we emphasize the rigorous approach towards the estimation issues. Therefore, what are the dependencies between these steps of the experiment and why is it relevant to consider them?

Different preference measurement scales and distributional assumptions can be considered, and based on this decision a variety of econometric methods can be used to estimate β^0, including ordinary or non linear least squares, several types of maximum likelihood estimators, least absolute deviations, etc. For example in the classic (metric) CA the coefficients are often estimated with OLS procedures, but choice-based models are usually estimated with Maximum Likelihood methods (we obtain the Multinomial Logit model assuming that y_t is a latent variable and ε_t

has a type I extreme value distribution). Other estimators may include generalized or non-linear least squares, other types of maximum likelihood estimators, least absolute deviations, etc.

Under regularity conditions, the appropriate estimators are consistent and when T grows the re-scaled sequence $V_T^{-1/2}\left(\hat{\beta}-\beta^0\right)$ converges in distribution to a standard normal distribution $N(0,I)$ where V_T is a positive definite matrix converging in probability to a limit asymptotic covariance matrix V. The distribution of the error $\left(\hat{\beta}-\beta^0\right)$ is generally unknown, and the main tool to justify inferences for a medium-to-large size T is the asymptotic distribution of the scaled error. Both covariance matrices, V_T and the limit V, depend on the design matrix X (or sequence, if we focus on V) with the product profiles $\{x_1,...,x_T\}$ shown in the experiment.

The efficiency of experimental estimators conveyed in the covariance matrices V_T, depends heavily on the product profiles evaluated by the respondents. Optimal experimental design maximizes the information elicited from the respondent, or equivalently minimizes the size of the covariance matrix. *Exact optimal designs* try to minimize $\phi(V_T)$ in the design matrix X, whilst *approximated optimal designs* try to minimize $\phi(V)$ in the limit frequencies w (which can be used to generate a $T \times p$ matrix X). The second approach was developed by Kiefer (1959) and his school.

Here $\phi(\cdot)$ denotes such a measure of the matrix "size" which is: (1) positively homogeneous: $\phi(\delta A) = \delta\phi(A)$ for $\delta > 0$ to ensure independence from scale factors; (2) non-increasing: $\phi(A) \leq \phi(B)$ when $(A-B)$ is non negative definite; and (3) convex to ensure that ϕ satisfies the condition that information cannot be increased through interpolation. The typical measures are the trace (A-optimality criterion), and the determinant (D-optimality criterion), therefore we will focus on these two methods. Other matrix size criteria have been considered, but they usually render equivalent solutions. This result was established by the Kiefer-Wolfowitz equivalence theorem for linear models and later extended to nonlinear models by White (1973).

Good designs use a matrix X that generates a small covariance matrix, V, meaning that the appropriate estimations will be reasonably accurate even if T is not very large, which reduces the burden on respondents. What are the consequences of using designs, which generate estimators with larger covariance matrices? Implementing suboptimal designs requires a larger T

to estimate the parameters with the same precision as an optimal design, which increases the market research cost and rating contamination caused by respondent's fatigue. Consequently, the design of conjoint experiments is a fundamental problem in marketing research.

1.4 Thesis Structure

Each of the chapters of this dissertation addresses relevant research questions for Conjoint Analysis practitioners and modelers. Chapters 2 and 3 are methodological in nature and are focused on the optimal design of CA experiments. Chapter 4 is devoted to the estimation of endogenous consideration sets and the endogeneity issue is tested with the data collected online using the Amazon's Mechanical Turk sample. Below I outline the scope of this dissertation by presenting the contents of every essay in more detail.

Chapter 2: Optimal experimental design with linear conjoint models.
In the first essay we develop a general approach for building exact optimal designs suitable for conjoint analysis using state-of-the-art optimization tools. We do not compute good designs, but the best ones according to the size of the information matrix of the associated estimators - trace and determinant. Such designs can be implemented by practitioners in various types of linear conjoint models: using product ranking data, rating-based, and under certain assumptions in discrete-choice experiments. Unlike previous methodologies, this approach flexibly handles continuous, discrete and mixed types of attributes. The essay also proposes a solution to the problem of repeated stimuli in optimal designs.

Classic CA considers that y_t is a utility ranking or a rating (measured either on a 0 to 100 attitude scale, a purchase probability scale, a strongly disagree to strongly agree scale, or some similar scale). The coefficients β^0 are estimated from an experimental setting, and the OLS estimator $\hat{\beta} = (X'X)^{-1}X'y$ is unbiased, with non-singular variance

$$Var\left(\hat{\beta}\right) = \sigma^2 \left(X'X\right)^{-1},$$

where X is the design matrix, whose rows define the product profiles shown to the respondent.

The experiments considered in CA are based on the classic statistical literature about optimal experimental designs. There are two big approaches: approximate optimal designs proposed by Kiefer (1959), and exact optimal designs. The former focuses on minimizing the size of the *asymptotic* covariance matrix, however it is not appropriate for CA, because it assumes that the optimal design consists of several stimuli replicated with optimal weights, while the same respondent should not be questioned several times about the same product.

Traditionally, the design of conjoint experiments is based on exact optimal designs. These designs minimize the size of the actual covariance matrix with the finite sample by solving the problem $\min_{X\in\chi}\phi\left(\left(X'X\right)^{-1}\right)$, where $\phi(\cdot)$ is a measure of matrix size: trace in case of A-optimality, and the determinant for D-optimality. However, also in the context of exact designs Box (1970) noticed that optimal designs may consist of a small number of duplicated profiles, which is not appropriate for CA.

Several procedures for computing exact designs have been proposed in the literature. These are mainly exchange algorithms, which sequentially add and delete one (Mitchell and Miller Jr 1970; Wynn 1972) or more (Mitchell 1974) profiles (rows in the matrix X) to improve the determinant of the information matrix. More advanced algorithms (Fedorov 1972; Cook and Nachtsheim 1980) at each iteration add an observation associated with the maximal improvement in the determinant. More recently, Meyer and Nachtsheim (1995) proposed the coordinate exchange algorithm, which instead of an entire product profile iteratively swaps attribute levels to ensure efficiency gains. A detailed comparison and evaluation of their computational performance can be found in Cook and Nachtsheim (1980). In general, none of these methods exploits satisfactorily the available numerical optimization tools.

The suitability of our approach for conjoint analysis is evaluated in a variety of of simulated scenarios. We compute optimal designs for experiments with continuous, discrete and mixed attributes, including the interactions between variables, the case of a single respondent and a panel of respondents exhibiting heterogenous intercepts. We additionally compare this method with the available conjoint software in the typical conjoint setting (a single respondent and discrete attributes only). In 3 out of 4 comparative examples our approach is faster, while achieving the same design efficiency as the available software for conjoint analysis. Moreover, this method is more flexible than traditional design procedures, which is implicit in the wide range of discussed applications and extensions.

Chapter 3: Robust designs for nonlinear conjoint analysis

In the second essay we generalize the problem presented in Essay 1 to optimal experimental design with nonlinear specifications, where the covariance matrix depends on unknown parameters. To this end, we use efficient computational methods profiting from the robustness property of worst-case optimization. The focus is on discrete choice experiments and compared with the benchmarks, the worst-case choice designs are more robust against misspecifications of unknown parameters and in majority of simulated scenarios are also more efficient. Therefore, such designs can be implemented when the risk-averse modeler does not have a good initial guess about consumer preferences.

Conjoint analysis literature has considered a variety of models which are nonlinear in parameters, for example choice-based CA, but also non-compensatory models, models with unknown ideal point, and others. In such cases the selection of optimal design is challenging because the covariance matrix $V_T = V\left(X,\beta^0\right)$ depends both on the deterministic regressors and the unknown parameters β^0 in a nonlinear way. To guarantee an efficient estimation of β^0 we need to compute an efficient experimental design X^* solving

$$\min_{X\in\chi}\phi\left(V\left(X,\beta^0\right)\right),$$

where the objective function is the size of covariance matrix for the usual estimators: maximum likelihood, nonlinear least squares, the generalized method of moments, and other related techniques.

In order to find an efficient design we need to know the value of β^0, which is unknown at the time the design is constructed - we want to estimate it from the experimental data! Therefore the design cannot be optimized without some assumptions about parameters and the data generating process. Since the size of covariance matrix is intrinsically linked to the unknown parameters, design efficiency is known only if the assumptions made on parameters are correct.

The experimental design and CA literature approached this puzzle in two distinct ways: 1) assuming a specific value (vector) for the unknown parameter β^0, predominantly under the "all-zero" parameters hypothesis, and 2) assuming a probability measure on the parametric space $\Theta \subset \mathbb{R}^K$ and weighting all possible values in $\beta \in \Theta$. We refer to the former as the local approach, and the latter as the Average-Optimum (AO) approach.

Perhaps the most common solution to the presented puzzle is the local approach suggested by Chernoff (1953), which is based on adopting a guess for the unknown parameters. This decision may be arbitrary, based on an inefficient pilot study, or using human prior beliefs about the preferences. With $\beta^0 = \overline{\beta}$, the local approach looks for a design X^+ defined as the solution to

$$\min_{X \in \chi} \phi\left(V\left(X, \overline{\beta}\right)\right),$$

where $\overline{\beta} \in \Theta$ is the assumed parameter vector. As the solution X^+ is specific to $\beta^0 = \overline{\beta}$, the resultant designs are locally optimal and are not optimal for values different from $\overline{\beta}$. Unfortunately, the efficiency of the locally optimal design, X^+, may be sensitive to even small perturbations in $\overline{\beta}$, and this initial guess is rarely close to the true β^0 (for if we had a good estimation, there would be no reason to run the experiment). In general we do not have any prior control over the efficiency of the design X^+ under the true β^0.

In CA context the local approach under null-hypothesis of $\overline{\beta} = 0$ has been used by Kuhfeld et al. (1994) for finding D-optimal choice designs for large conjoint applications through computerized search, and for discrete-choice experiments Kanninen (2002) suggested a procedure that leads to maximizing $|X'X|$ with continuous regressors. Huber and Zwerina (1996) have studied the effects of incorporating manager's prior beliefs into the optimal design, showing that under $\overline{\beta} \neq 0$ utility balance of choice sets remains an important property of efficient choice designs.

The Average-Optimum approach attempts to reduce the influence of $\overline{\beta}$, and considers an average of many values instead of the local design. This method involves a probability measure μ defined over the parametric space Θ, optimizing the weighted average of design efficiencies

$$\min_{X \in \chi} \int_{\Theta} \phi\left(V\left(X,\beta\right)\right)\mu\left(d\beta\right).$$

The solution X^{++} is not optimal under each scenario but hedged against the risk associated with all scenarios. The solution is quite sensitive to the choice of the weighting probability distribution μ (and its parameters). Unless μ is strongly concentrated near the true unknown β^0, little can we say about the true efficiency of the design, $\phi\left(V\left(X^{++},\beta^0\right)\right)$.

This approach has been used in CA to build exact optimal designs for choice models by Sándor and Wedel (2001) in the context of a single respondent, setting μ as a normal distribution representing managers' prior beliefs about product market shares. The Averaged Approach has also been applied in the Mixed Logit model (Arora and Huber 2001; Sándor and Wedel 2002). Sándor and Wedel (2005) extended the idea to panels of heterogeneous customers generating a different design for each customer.

Overall, the assumptions about unknown parameters β^0 are specific to a given application. Little is known about empirical validity or optimality claims of implemented designs when these assumptions are violated (Louviere et al. 2011). In Chapter 3 we propose a worst-case method to build efficient designs in CA experiments, where the covariance matrix depends on the unknown parameter. We solve this problem using efficient methods for robust optimization, and provide numerical examples for discrete-choice experiments, and other common nonlinear utility functions. This method is robust to misspecification of parameters, yields fewer designs with outlying (large) covariance, and is also more efficient in most of the scenarios considered.

Chapter 4: Estimation of endogenous consideration sets

The third essay is dedicated to the estimation of endogenous consideration sets. Consideration sets arise because rationally bounded consumers often skip potentially interesting options, for example due to perceptual limitations, lack of information, or halo effect. Therefore individuals choose in two stages: first they screen off products whose attributes do not satisfy certain criteria, and then *select the best alternative according to their preferences. Traditional CA methods focus on the second step, and more recent consideration set models assume that those steps are independent. However with halo effect present, we cannot assume that screening off stage is independent from evaluative step. We test this endogeneity with the data from an online experiment using Amazon MTurks.*

Actual consumers' choices are not always consistent with their preferences because rationally bounded individuals often skip potentially attractive products, for example due to the *lack of information*, or *perceptual limitations* or *halo effect*. Research in consumer behavior established that consumers choose in two stages: 1) they use heuristic rules to screen off products whose attributes do not satisfy certain criteria, often focusing on some key attributes (Bettman 1974; Montgomery and Svenson 1976; Payne 1976; Payne and Ragsdale 1978; Payne et al. 1993); 2) they select the best alternative from the considered options according to their preferences. If consideration rules are not taken into account the purchase decision might seem contradictory with preferences.

Whether or not consumers select a product depends on a screening-off consideration rule, and overall preferences are conditioned by this decision. The process can be described with a switching-preference model

$$y_t = \begin{cases} f(x_t)'\beta + \varepsilon_{1t} & x_t \in A(\gamma, u_t) \\ \alpha + \varepsilon_{2t} & x_t \notin A(\gamma, u_t) \end{cases}$$

where for each multiattribute product x_t, we observe individual preference ratings, y_t, and $(\varepsilon_{1t}, \varepsilon_{2t})$ are i.i.d. jointly distributed with $E(\varepsilon_i) = 0$ and $E(\varepsilon_i \varepsilon_i') = \sigma_i^2$. The consideration set $A(\gamma, u_t)$ depends on unknown parameter vector γ, and some random vector u_t. In marketing, the most common specifications of $A(\gamma, u_t)$ are: disjunctive, conjunctive, compensatory, and lexicographic heuristic (see e.g. Gilbride and Allenby 2004, 2006; Jedidi and Kohli 2005). Note that if $\Pr(u_t = 0) = 1$, the set is deterministic. However, the stochastic approach is more fruitful because situational factors, excitement, and attention can affect the consideration of a given product. Bettman and Zins (1977) finds out evidence that consumers build their consideration rules on-the-spot using memory fragments and situational elements.

Recently the marketing literature started to look at the heuristic consideration rules, building two-step models (Gensch 1987; Gilbride and Allenby 2004, 2006; Jedidi and Kohli 2005; Kohli and Jedidi 2007), where first the consideration set is specified, and then the utility function is analyzed conditionally over the considered options, assuming independence of those two steps. However, the halo effect is a clear reason for consideration sets to be endogenous with respect to the overall preferences (Beckwith and Lehmann 1975): if the cognitive process is influenced by the overall affective impression of the product, we cannot assume that the screening-off stage is independent from the evaluative step.

If we define a dummy consideration variable $C_t = I\left(x_t \in A\left(\gamma, u_t\right)\right)$, where $I(\cdot)$ denotes the indicator function (equal to 1 when $x_t \in A\left(\gamma, u_t\right)$ and zero otherwise), then the above model can be written as a regression equation

$$\begin{aligned} y_t &= C_t\, f(x_t)'\beta + (1-C_t)\ \alpha + \eta_t \\ \eta_t &= C_t\ \varepsilon_{1t} + (1-C_t)\ \varepsilon_{2t}. \end{aligned}$$

If u_t is independent of $(\varepsilon_{1t}, \varepsilon_{2t})$ then

$$E\left[\eta_t | x_t\right] = E\left[C_t | x_t\right] \times E\left[\varepsilon_{1t}\right] + E\left[(1-C_t)|x_t\right] \times E\left[\varepsilon_{2t}\right] = 0,$$

with $E\left[C_t|x_t\right] = \Pr(C_t = 1|x_t)$ and $E\left[(1-C_t)|x_t\right] = (1-\Pr(C_t = 1|x_t))$ and the model can be estimated using classical econometric tools for exogenous switching regression. The problem is much more difficult to handle if the consideration set $A\left(\gamma, u\right)$ is endogenously selected, and we cannot assume that u_t is statistically independent of $(\varepsilon_{1t}, \varepsilon_{2t})$. Now, the shock of the regression model satisfies

$$E\left[\eta_t|x_t\right] = \Pr(C_t = 1|x_t) \cdot E\left[\varepsilon_{1t}|x_t, C_t = 1\right] + (1-\Pr(C_t = 1|x_t)) \cdot E\left[\varepsilon_{2t}|x_t, C_t = 0\right],$$

which is in general different from zero. Ignoring this type of endogeneity will lead to inconsistent estimations, and a biased perspective on consumer preference formulations. Further difficulties arise when self-explicated information about consideration set is not observed (C_t is not available).

In the essay we illustrate the endogeneity of consideration sets with the conjoint experiment to evaluate customer preferences towards lunch entrées, which was conducted online on a sample of Amazon's Mechanical Turks. The empirical application involves a compensatory consideration set, the case when C_t is observed and the normal distribution of the shocks. A two-step procedure proposed by Heckman (1979) accounts for endogeneity in the consideration set and provides consistent, and asymptotically efficient estimates for all parameters.

Bibliography

Alba, J. W. and Cooke, A. D. J. (2004). When absence begets inference in conjoint analysis. *Journal of Marketing Research*, 41(4):382–387.

Allenby, G., Fennell, G., Huber, J., Eagle, T., Gilbride, T., Horsky, D., Kim, J., Lenk, P. J., Johnson, R. M., Ofek, E., Orme, B., Otter, T., and Walker, J. (2005). Adjusting choice models to better predict market behavior. *Marketing Letters*, 16(3/4):197–208.

Arora, N. and Huber, J. (2001). Improving parameter estimates and model prediction by aggregate customization in choice experiments. *The Journal of Consumer Research*, 28(2):273–283.

Beckwith, N. E. and Lehmann, D. R. (1975). The importance of halo effects in multi-attribute attitude models. *Journal of Marketing Research*, 12(3):265–275.

Belloni, A., Freund, R., Selove, M., and Simester, D. (2008). Optimizing product line designs: Efficient methods and comparisons. *Management Science*, 54(9):1544–1552.

Bemmaor, A. C. and Wagner, U. (2000). A multiple-item model of paired comparisons: Separating chance from latent preference. *Journal of Marketing Research*, 37(4):514–524.

Bettman, J. R. (1974). A threshold model of attribute satisfaction decisions. *Journal of Consumer Research*, 1(2):30–35.

Bettman, J. R. and Zins, M. A. (1977). Constructive processes in consumer choice. *Journal of Consumer Research*, 5(4):75–85.

Blattberg, R. C. and Wisniewski, K. J. (1989). Price-induced patterns of competition. *Marketing Science*, 8(4):291–309.

Box, M. J. (1970). Some experiences with a nonlinear experimental design criterion. *Technometrics*, 12(3):569–589.

Bradlow, E. T., Hu, Y., and Ho, T.-H. (2004). A learning-based model for imputing missing levels in partial conjoint profiles. *Journal of Marketing Research*, 41(4):369–381.

Bradlow, E. T. and Rao, V. R. (2000). A hierarchical bayes model for assortment choice. *Journal of Marketing Research*, 37(2):259–268.

Carmone, F. J., Green, P. E., and Jain, A. K. (1978). Roubustness of conjoint analysis: Some Monte Carlo results. *Journal of Marketing Research*, 15(2):300–303.

Carroll, J. D. (1972). Individual differences and multidimensional scaling. In Shepard, R., Romney, A., and Nerlove, S., editors, *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences*. Seminar Press.

Carroll, J. D. and Green, P. E. (1995). Psychometric methods in marketing research: Part I, Conjoint Analysis. *Journal of Marketing Research*, 32(4):385–391.

Cattin, P., Gelfand, A. E., and Danes, J. (1983). A Simple Bayesian Procedure for Estimation in a Conjoint Model. *Jounal of Marketing Research*, 20(1):29–35.

Cattin, P. and Wittink, D. R. (1982). Commercial use of conjoint analysis: A survey. *Journal of Marketing*, 46(3):44–53.

Chen, K. D. and Hausman, W. H. (2000). Mathematical properties of the optimal product line selection problem using choice-based conjoint analysis. *Management Science*, 46(2):327–332.

Chernoff, H. (1953). Locally optimal designs for estimating parameters. *The Annals of Mathematical Statistics*, 24(4):586–602.

Chung, J. and Rao, V. R. (2003). A general choice model for bundles with multiple-category products: Application to market segmentation and optimal pricing for bundles. *Journal of Marketing Research*, 40(2):115–130.

Cochran, W. G. and Cox, G. M. (1957). *Experimental Designs*. John Wiley & Sons, New York, 2nd edition.

Cook, R. D. and Nachtsheim, C. J. (1980). A Comparison of Algorithms for Constructing Exact D-Optimal Designs. *Technometrics*, 22(3):315–324.

Currim, I. S., Weinberg, C. B., and Wittink, D. R. (1981). Design of subscription programs for a performing arts series. *Journal of Consumer Research*, 8(1):67–75.

Dahan, E., Hauser, J. R., Simester, D., and Toubia, O. (2002). Application and test of web-based adaptive polyhedral conjoint analysis. Working papers, Massachusetts Institute of Technology (MIT), Sloan School of Management.

Dahan, E. and Srinivasan, V. (2000). The predictive power of Internet-based product concept testing using visual depiction and animation. *Journal of Product Innovation Management*, 17:99–109.

Darmon, R. Y. (1979). Setting sales quotas with conjoint analysis. *Journal of Marketing Research*, 16(1):133.

De Bruyn, A., Liechty, J. C., Huizingh, E. K. R. E., and Lilien, G. L. (2008). Offering Online Recommendations with Minimum Customer Input Through Conjoint-Based Decision Aids. *Marketing Science*, 27(3):443–460.

Debreu, G. (1960). Topological methods in cardinal utility theory. In Kenneth J. Arrow, S. K. and Suppes, P., editors, *Mathematical Methods in the Social Sciences, 1959: Proceedins of the First Stanford Symposium*, pages 16–26. Stanford University Press.

Decker, R. and Trusov, M. (2010). Estimating aggregate consumer preferences from online product reviews. *International Journal of Research in Marketing*, 27(4):293–307.

Dellaert, B. G. C., Arentze, T. A., Bierlaire, M., Borgers, A. W. J., and Timmermans, H. J. P. (1998). Investigating consumers' tendency to combine multiple shopping purposes and destinations. *Journal of Marketing Research*, 35(2):177–188.

DeSarbo, W. S. and Green, P. E. (1984). Choice-constrained conjoint analysis. *Decision Sciences*, 15(3):297–323.

Desarbo, W. S., Ramaswamy, V., and Cohen, S. H. (1995). Market segmentation with choice-based conjoint analysis. *Marketing Letters*, 6(2):137–147.

Ding, M., Grewal, R., and Liechty, J. (2005). Incentive-aligned conjoint analysis. *Journal of Marketing Research*, 42(1):67–82.

Drezè, X. and Zufryden, F. (1998). A web-based methodology for product design evaluation and optimisation. *The Journal of the Operational Research Society*, 49(10):1034–1043.

D'Souza, G. and Rao, R. C. (1995). Can repeating an advertisement more frequently than the competition affect brand preference in a mature market? *Journal of Marketing*, 59(2):32.

Elrod, T., Louviere, J. J., and Davey, K. S. (1992). An empirical comparison of ratings-based and choice-based conjoint models. *Journal of Marketing Research*, 29(3):368–377.

Evgeniou, T., Boussios, C., and Zacharia, G. (2005). Generalized robust conjoint estimation. *Marketing Science*, 24(3):415–429.

Farquhar, P. H. and Rao, V. R. (1976). A balance model for evaluating subsets of multiattributed items. *Management Science*, 22(5):528–539.

Fedorov, V. V. (1972). *Theory of optimal experiments*. Academic Press, New York.

Finkbeiner, C. (1988). Comparison of Conjoint Choice Simulators. In *Proceedings of the Sawtooth Software Conference*.

Foytik, M. (1999). Conjoint on the web – lessons learned. In *Proceedings of the Sawtooth Software Conference*.

Francis G. Giesbrecht, M. L. G. (2004). *Planning, construction, and statistical analysis of comparative experiments*. Wiley, New York.

Gensch, D. H. (1987). A two-stage disaggregate attribute choice model. *Marketing Science*, 6(3):223–239.

Gilbride, T. J. and Allenby, G. M. (2004). A choice model with conjunctive, disjunctive, and compensatory screening rules. *Marketing Science*, 23(3):391–406.

Gilbride, T. J. and Allenby, G. M. (2006). Estimating heterogeneous eba and economic screening rule choice models. *Marketing Science*, 25(5):494–509.

Green, P. E. (1974). On the design of choice experiments involving multifactor alternatives. *Journal of Consumer Research*, 1(2):61–68.

Green, P. E., Carroll, J. D., and Goldberg, S. M. (1981). A general approach to product design optimization via conjoint analysis. *Journal of Marketing*, 45(3):17–37.

Green, P. E. and Krieger, A. (1993). Conjoint analysis with product-positioning applications. In Eliashberg, J. and Lilien, G. L., editors, *Handbooks in OR/MS*. New York: Elsevier Science Publishers.

Green, P. E. and Krieger, A. M. (1988). Choice rules and sensitivity analysis in conjoint simulators. *Academy of Marketing Science Journal*, 16(1):114–128.

Green, P. E. and Krieger, A. M. (1990). A hybrid conjoint model for price-demand estimation. *European Journal of Operational Research*, 44(1):28 – 38.

Green, P. E. and Krieger, A. M. (1991). Segmenting markets with conjoint analysis. *Journal of Marketing*, 55(4):20–31.

Green, P. E. and Krieger, A. M. (1992). An application of a product positioning model to pharmaceutical products. *Marketing Science*, 11(2):117–132.

Green, P. E., Krieger, A. M., and Wind, Y. (2001). Thirty years of conjoint analysis: Reflections and prospects. *Interfaces*, 31(3):S56–S73.

Green, P. E. and Rao, V. R. (1971). Conjoint Measurement for Quantifying Judgmental Data. *Journal of Marketing Research*, 8(3):355–363.

Green, P. E. and Srinivasan, V. (1978). Conjoint analysis in consumer research: Issues and outlook. *The Journal of Consumer Research*, 5(2):103–123.

Gustafsson, A., Herrmann, A., and Huber, F. (2007). Conjoint analysis as an instrument of market research practice. In Gustafsson, A., Herrmann, A., and Huber, F., editors, *Conjoint Measurement: Methods and Applications*. Berlin: Springer Verlag.

Hauser, J. R. and Toubia, O. (2005). The impact of utility balance and endogeneity in conjoint analysis. *Marketing Science*, 24(3):498–507.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, 47(1):153–161.

Herman, S. (1988). Software for full-profile conjoit analysis. In *Proceedings on the Sawtooth Software Conference on Perceptual Mapping, Conjoint Analysis, and Computer Interviewing*.

Hoeffler, S. (2003). Measuring preferences for really new products. *Journal of Marketing Research*, 40(4):406–420.

Huber, J., Wittink, D. R., Fiedler, J. A., and Miller, R. (1991). An empirical comparison of aca and full profile judgments. In *Sawtooth Software Conference Proceedings*, pages 189–202.

Huber, J., Wittink, D. R., Fiedler, J. A., and Miller, R. (1993). The effectiveness of alternative preference elicitation procedures in predicting choice. *Journal of Marketing Research*, 30(1):105–114.

Huber, J. and Zwerina, K. (1996). The importance of utility balance in efficient choice design. *Journal of Marketing Research*, 33(3):307–317.

Intille, S., Kukla, C., and Ma, X. (2002). Eliciting user preferences using image-based experience sampling and reflection. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '02, pages 738–739, New York, NY, USA. ACM.

Iyengar, R., Jedidi, K., and Kohli, R. (2008). A conjoint approach to multipart pricing. *Journal of Marketing Research*, 45(2):195–210.

Jaeger, S. R., Hedderley, D., and MacFie, H. J. (2001). Methodological issues in conjoint analysis: a case study. *European Journal of Marketing*, 35(11/12):1217–1239.

Jain, D. C., Muller, E., and Vilcassim, N. J. (1999). Pricing patterns of cellular phones and phonecalls: A segment-level analysis. *Management Science*, 45(2):131–141.

Jedidi, K. and Kohli, R. (2005). Probabilistic Subset-Conjunctive Models for Heterogeneous Consumers. *Journal of Marketing Research*, 42(4):483–494.

Jedidi, K. and Zhang, Z. J. (2002). Augmenting conjoint analysis to estimate consumer reservation price. *Management Science*, 48(10):1350–1368.

Johnson, R. M. (1974). Trade-off analysis of consumer values. *Journal of Marketing Research*, 11(2):121–127.

Johnson, R. M. (1987). Adaptive Conjoint Analysis. In *Sawtooth Software Conference on Perceptual Mapping, Conjoint Analysis, and Computer Interviewing*, pages 253–265. r.

Kamakura, W. (1988). A least squares procedure for benefit segmentation with conjoint experiments. *Journal of Marketing Research*, 25(2):157–167.

Kamakura, W. A. and Srivastava, R. K. (1984). Predicting choice shares under conditions of brand interdependence. *Journal of Marketing Research*, 21(21):420–434.

Kanninen, B. J. (2002). Optimal design for multinomial choice experiments. *Journal of Marketing Research*, 39:214–227.

Kiefer, J. (1959). Optimum experimental designs. *Journal of the Royal Statistical Society. Series B (Methodological)*, 21(2):272–319.

Kim, S.-H. (2000). *Modeling buyers' upgrading behavior on successive versions of high-tech products: Theoretical and empirical analyses*. PhD thesis, Stanford University.

Kim, S.-H. and Srinivasan, V. (2006). A conjoint-hazard model of the timing of buyers' upgrading to improved versions of high technology products. Research Paper 1720, Standford.

Kohli, R. and Jedidi, K. (2007). Representation and inference of lexicographic preference models and their variants. *Marketing Science*, 26(3):380–399.

Kohli, R. and Krishnamurti, R. (1987). A heuristic approach to product design. *Management Science*, 33(12):1523–1533.

Kohli, R. and Mahajan, V. (1991). A reservation-price model for optimal pricing of multiattribute products in conjoint analysis. *Journal of Marketing Research*, 28(3):347–354.

Kohli, R. and Sukumar, R. (1990). Heuristics for product-line design using conjoint analysis. *Management Science*, 36(12):1464–1478.

Krishnamurthi, L. and Wittink, D. R. (1991). The value of Idiosyncratic Functional Forms in Conjoint Analysis. *International Journal of Research in Marketing*, 8:301–313.

Kruskal, J. B. (1965). Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 27(2):251–263.

Kuhfeld, W. F., Tobias, R. D., and Garratt, M. (1994). Efficient Experimental Design with Marketing Research Applications. *Journal of Marketing Research*, 31(4):545–557.

Lancaster, K. (1971). *Consumer demand: a new approach*. Columbia University Press, New York.

Lazari, A. G. and Anderson, D. A. (1994). Designs of discrete choice set experiments for estimating both attribute and availability cross effects. *Journal of Marketing Research*, 31(3):375.

Lee, T. Y. and Bradlow, E. T. (2011). Automated Marketing Research Using Online Customer Reviews. *Journal of Marketing Research*, 48(5):881–894.

Leigh, T. W., MacKay, D. B., and Summers, J. O. (1984). Reliability and validity of conjoint analysis and self-explicated weights: a comparison. *Jounal of Marketing Research*, 31(4):456–462.

Levy, M., Webster, J., and Kerin, R. A. (1983). Formulating push marketing strategies: A method and application. *Journal of Marketing*, 47(1):25.

Louviere, J. J. (1988a). *Analyzing decision making: metric conjoint analysis*. Newbury Park.

Louviere, J. J. (1988b). Conjoint analysis modelling of stated preferences. A review of theory, methods, recent developments and external validity. *Journal of Transport Economic and Policy*, 22(1):93–119.

Louviere, J. J., Carson, R., and Pihlens, D. (2011). Design of Discrete Choice Experiments: A Discussion of Issues That Matter in Future Applied Research. *Journal of Choice Modelling*, 4(1):1–8.

Louviere, J. J. and Woodworth, G. (1983). Design and analysis of simulated consumer choice or allocation experiments: An approach based on aggregate data. *Journal of Marketing Research*, 20(4):350–367.

Luce, R. D. (1966). Two extensions of conjoint measurement. *Journal of Mathematical Psychology*, 3:348–370.

Mahajan, V., Green, P. E., and Goldberg, S. M. (1982). A conjoint model for measuring self- and cross-price/demand relationships. *Journal of Marketing Research*, 19(3):334.

Martínez, L. M.-C., Mollá-Bauzá, M. B., Gomis, F. J. D. C., and Poveda, A. M. (2006). Influence of purchase place and consumption frequency over quality wine preferences. *Food Quality and Preference*, 17:315–327.

McBride, R. D. and Zufryden, F. S. (1988). An integer programming approach to the optimal product line selection problem. *Marketing Science*, 7(2):126–140.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In Zarembka, P., editor, *Frontiers in Econometrics*. Academic Press.

Melles, T., Laumann, R., and Holling, H. (2000). Validity and Reliability of Online Conjoint Analysis. In *Proceedings of the Sawtooth Software Conference*.

Meyer, R. K. and Nachtsheim, C. J. (1995). The coordinate-exchange algorithm for constructing exact optimal experimental designs. *Technometrics*, 37(1):60–69.

Mitchell, T. J. (1974). An algorithm for the construction of "D-optimal" experimental designs. *Technometrics*, 16(2):203–210.

Mitchell, T. J. and Miller Jr, F. (1970). Use of design repair to construct designs for special linear models. *Math. Div. Ann. Progr. Rept.(ORNL-4661)*, pages 130–131.

Montgomery, D. C. (2005). *Design and analysis of experiments*. John Wiley & Sons, New York, 6th edition.

Montgomery, H. and Svenson, O. (1976). On decision rules and information processing strategies for choices among multiattribute alternatives. *Scandinavian Journal of Psychology*, 17(1):283–291.

Netzer, O., Feldman, R., Goldenberg, J., and Fresko, M. (2012). Mine Your Own Business: Market-Structure Surveillance Through Text Mining. *Marketing Science*, 31(5):521–543.

Netzer, O. and Srinivasan, V. (2011). Adaptive self-explication of multi-attribute preferences. *Journal of Marketing Research*, 48(1):140–156.

Oppewal, H., Louviere, J. J., and Timmermans, H. J. P. (1994). Modeling hierarchical conjoint processes with integrated choice experiments. *Journal of Marketing Research*, 31(1):92–105.

Orme, B. K. and King, W. C. (1998). Conducting Full-Profile Conjoint Analysis over the Internet. Technical report, Sawtooth Software, Inc.

Payne, J. W. (1976). Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational behavior and human performance*, 16(2):366–387.

Payne, J. W., Bettman, J. R., and Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge University Press.

Payne, J. W. and Ragsdale, E. E. (1978). Verbal protocols and direct observation of supermarket shopping behavior: Some findings and a discussion of methods. *Advances in consumer research*, 5:571–577.

Rao, V. R. and Sattler, H. (2007). Measurement of Price Effects with Conjoint Analysis: Separating Informational and Allocative Effects of Price. In *Conjoint Measurement: Methods and Applications*. Springer.

Roe, B., Teisl, M. F., Levy, A., and Russell, M. (2001). US consumers' willingness to pay for green electricity. *Energy Policy*, 29(11):917 – 925.

Rubin, D. B. (2004). Design and modeling in conjoint analysis with partial profiles. *Journal of Marketing Research*, 41(4):390–391.

Sándor, Z. and Wedel, M. (2001). Designing conjoint choice experiments using managers' prior beliefs. *Journal of Marketing Research*, 38(4):430–444.

Sándor, Z. and Wedel, M. (2002). Profile construction in experimental choice designs for mixed logit models. *Marketing Science*, 21(4):455–475.

Sándor, Z. and Wedel, M. (2005). Heterogeneous conjoint choice designs. *Journal of Marketing Research*, 42(2):210–218.

Sawtooth Software (2007). *The ACA / Web v6.0 technical paper*. Sawtooth Software, Inc.

Scott, J. E. and Keiser, S. K. (1984). Forecasting acceptance of new industrial products with judgment modeling. *Journal of Marketing*, 48(2):54–67.

Srinivasan, V. (1988). A Conjunctive-Compensatory Approach to the Self-Explication of Multiattributed Preferences. *Decision Sciences*, 19:295–305.

Srinivasan, V. and Park, C. S. (1997). Surprising robustness of the self-explicated approach to customer preference structure measurement. *Journal of Marketing Research*, 34(2):286–291.

Srinivasan, V. and Shocker, A. D. (1973a). Estimating the weights for multiple attributes in a composite criterion using pairwise judgments. *Psychometrika*, 38(4):473–493.

Srinivasan, V. and Shocker, A. D. (1973b). Linear programming techniques for multidimensional analysis of preferences. *Psychometrika*, 38(3):337–369.

Telser, H. and Zweifel, P. (2002). Measuring willingness-to-pay for risk reduction: an application of conjoint analysis. *Health Economics*, 11(2):129–139.

Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review*, 34:278–286.

Toubia, O., Hauser, J., and Garcia, R. (2007). Probabilistic polyhedral methods for adaptive choice-based conjoint analysis: Theory and application. *Marketing Science*, 26(5):596–610.

Toubia, O., Hauser, J. R., and Simester, D. I. (2004). Polyhedral methods for adaptive choice-based conjoint analysis. *Journal of Marketing Research*, 41(1):116–131.

Toubia, O., Simester, D. I., Hauser, J. R., and Dahan, E. (2003). Fast polyhedral adaptive conjoint estimation. *Marketing Science*, 22(3):273–303.

van der Lans, I. A. and Heiser, W. J. (1992). Constrained part-worth estimation in conjoint analysis using the self-explicated utility model. *International Journal of Research in Marketing*, 9(4):325–344.

van der Lans, I. A., Wittink, D. R., Huber, J., and Vriens, M. (1992). Within- and across-attribute constraints in ACA and full profile conjoint analysis. Technical report, Sawtooth Software.

Vriens, M., Wedel, M., and Wilms, T. (1996). Metric conjoint segmentation methods: A monte carlo comparison. *Journal of Marketing Research*, 33(1):73–85.

White, L. V. (1973). An extension of the General Equivalence Theorem to nonlinear models. *Biometrika*, 60(2):345–348.

Wind, J., Green, P. E., Shifflet, D., and Scarbrough, M. (1989). “Courtyard by Marriott”: Designing a hotel facility with consumer-based marketing models. *Interfaces*, 19(1):25–47.

Witt, K. J. (1997). Best Practices in Interviewing via the Internet. In *Proceedings of the Sawtooth Software Conference*.

Wittink, D. R. and Cattin, P. (1989). Commercial use of conjoint analysis: An update. *Journal of Marketing*, 53(3):91–96.

Wittink, D. R., Krishnamurthi, L., and Nutter, J. B. (1982). Comparing derived importance weights across attributes. *Journal of Consumer Research*, 8(4):471–474.

Wittink, D. R., Vriens, M., and Burhenne, W. (1994). Commercial use of conjoint analysis in Europe: Results and critical reflections. *International Journal of Research in Marketing*, 11(1):41 – 52.

Wuyts, S., Stremersch, S., Bulte, C. V. D., and Franses, P. H. (2004). Vertical marketing systems for complex products: A triadic perspective. *Journal of Marketing Research*, 41(4):479–487.

Wynn, H. P. (1972). Results in the theory and construction of D-optimum experimental designs. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):133–147.

Yee, M., Dahan, E., Hauser, J. R., and Orlin, J. (2007). Greedoid-based noncompensatory inference. *Marketing Science*, 26(4):532–549.

Chapter 2

Reconsidering Optimal Experimental Design for Conjoint Analysis

2.1 Introduction

Since the seminar paper of Green and Rao (1971), Conjoint Analysis (CA) has become a widespread marketing research tool for marketing scholars and practitioners (see e.g. Cattin and Wittink 1982; Wittink and Cattin 1989). CA encompasses a variety of techniques designed to analyze consumer preferences over multi-attributed products, estimating preference trade-offs between attributes from experimental data. Respondents are presented with a series of stimuli (product/service descriptions, illustrations, samples, prototypes etc.), and are asked to rank or rate them (metric or "classic" CA), or to choose one from each subset of profiles shown to them (choice-based CA). The underlying preference model, regardless of the response measurement scale, is

$$y_t = f(x_t)'\beta + \varepsilon_t, \qquad t = 1, \ldots, T,$$

with a compensatory, linear-in-parameters utility function $U(x_t) = f(x_t)'\beta$, where y_t represents the consumer utility. Product profile x_t is a $k \times 1$ vector of deterministic regressors in a compact set χ in an Euclidean space representing attributes (discrete dummy and/or continuous variables), and sometimes other contextual block variables. f is a known continuous mapping from χ to $\mathbb{R}^p$ whose coordinates are linearly independent and may include an intercept, discrete interactions (products of dummies), or product of continuous regressors (to define multivariate polynomials similarly to surface response models). The function f could also have a known local maximum (self-explicated ideal point). We also allow $p < k$, if $f(x_t)$ is a projection of x_t on a linear space of a smaller dimension. The vector β is a $p \times 1$ vector of unknown parameters. The errors ε_t are regarded as mutually independent random shocks, satisfying $E[\varepsilon_t] = 0$, $E\left[\varepsilon_t^2\right] = \sigma^2$. The experiment considers $T \geq p$ independent observations. In a matrix notation the model can be written as $y = X\beta + \varepsilon$, where y and ε are $T \times 1$ vectors, $X = [f(x_1)', \ldots, f(x_T)']'$ is a $T \times p$ design matrix, with row t containing $f(x_t)'$, and $x = [x_1', \ldots, x_T']'$.

The statistical analysis will depend on the preference measure scale for respondents. For example in the classic CA the coefficients are often estimated with OLS procedures, but choice-based CA models are usually estimated with Maximum Likelihood methods (we obtain the Multinomial Logit model assuming that y_t is a latent variable and ε_t has a type I extreme value distribution). Choice-based CA is nowadays widely applied, but from the econometric point of view the hypotheses about the distribution of ε_t are stronger than in the classic CA which is more robust to specification errors. For a literature review and description of the methods and common CA applications, see Gustafsson et al. (2007). For a discussion of some problem areas in current CA methods, see Bradlow (2005) and Netzer et al. (2008).

In all cases, under regularity conditions the probability distribution of $\sqrt{T}\left(\hat{\beta} - \beta\right)$ can be approximated by a $N(0, V)$, where the variance V depends on the design matrix X. Good designs use a matrix X that generates a small covariance matrix V, meaning that the estimations will be reasonably accurate even if T is not very large, which reduces the burden on respondents and the study costs. The design of conjoint experiments is a fundamental problem in marketing research. The available methods are designed to provide (nearly) optimal efficiency (see e.g. Kuhfeld et al. 1994). In this essay we review existing algorithms for computing optimal experimental designs, and discuss their limitations and drawbacks. As we discuss later, these methods tend to choose designs with repeated product profiles, which is inconvenient in the CA context as we cannot show the same profile several times to a respondent. Also, available algorithms do not manage efficiently the models with continuous and discrete regressors. We propose a general approach to compute exact optimal designs for CA experiments with both continuous and discrete variables, furthermore we eliminate the problem of profile repetitions.

The structure of the essay is as follows. We begin with a discussion of the state of the art tools for the design of optimal experiments, and their limitations, particularly for the classic CA. Next we present a new approach to the design of experiments, and justify the use of appropriate constraints, which prohibit profile repetitions for the same respondent, ensuring its suitability for CA. We also present an integer and mixed version of the problem, followed by the case of a panel of consumers. For pedagogical reasons, we start with the discussion of CA based on ratings for individuals and panels. Then we discuss the estimation and optimal design for rank data under invariance to monotonous transformations. We conclude with the case of CA based on consumer choices. We also present some extensions such as the use of partial profiles for complex products with many attributes.

2.2 Literature Review on Optimal Experimental Design for Linear Models

In this section we review the tools available for the design of optimal experiments in linear regression models, and the drawbacks for their application to CA experiments. In classic experimental design y_t is an observable variable, and it is assumed that $E(\varepsilon) = 0$, $E(\varepsilon\varepsilon') = \sigma^2 I_T$ and $rank(X) = p$. In particular, classic CA considers y_t is a utility ranking or a rating (measured either on a 0 to 100 attitude scale, a purchase probability scale, a strongly disagree to strongly agree scale, or some similar scale). The coefficients β are estimated from an experimental setting, and the OLS estimator $\hat{\beta} = (X'X)^{-1}X'y$ is unbiased, with non-singular variance

$$Var\left(\hat{\beta}\right) = \sigma^2 \left(X'X\right)^{-1}.$$

Often, there are linear identities across the attributes which cause the $\left(X'X\right)$ matrices to be singular. For example, for continuous regressors this occurs when we consider compositional data (proportions of several ingredients) and an intercept (the sum of the proportions is identical to the intercept variable), and also when we have discrete dummies with an intercept. In these cases, the model is usually reformulated (e.g., omitting a regressor). We will assume that the necessary operations to eliminate collinearities have been already implemented in the considered formulation.

If the ratings were normally distributed we could perform inference analysis with a small T. But generally, this is not the case. If the deterministic matrix $Q_T = \left(X'X/T\right)$ converges to a positive definite matrix Q, then under regular conditions $\sqrt{T}\left(\hat{\beta} - \beta\right)$ converges in distribution to $N\left(0, \sigma^2 Q^{-1}\right)$. When ratings are not normally distributed, which is a common situation, the asymptotic approximation is the only way to justify inferences for medium-size to large T. The smaller the matrix Q_T^{-1}, (respectively Q^{-1}) the more (asymptotically) efficient is the OLS estimator. Classic experiments (Cochran and Cox 1957; Cox 1958) usually assume normality and discrete attributes, and for a small T statisticians try to make $X'X$ diagonal (i.e. X is an orthogonal matrix), albeit orthogonal designs are neither always possible (e.g., in models with squared regressors, which is common in polynomial specifications) nor optimal. Allowing for a small correlation between estimators we might obtain estimators with smaller variances.

An experiment X^* is (approximately) optimal if Q_T^{-1} (respectively Q^{-1}) is the smallest possible covariance matrix according to some appropriate criteria measuring the size of this matrix. Suboptimal designs require a larger T to estimate the parameters with the same precision as X^*, increasing the market research cost and rating contamination caused by respondent's fatigue. Notice that for models linear in parameters, optimal designs are not adaptive. In other words, even if data is collected and processed sequentially, we do not use what we learn to change the experimental setting. The reason is that neither the matrix Q_T^{-1} nor Q^{-1} are affected by collected information about the previous ratings, $\{y_1, \ldots, y_{t-1}\}$.

The experiments considered in CA are based on the classic statistical literature about optimal experimental designs. Broadly speaking there are two big approaches: approximate optimal designs proposed by Kiefer (1959) , and exact optimal designs. The Kiefer's approach, seeks designs where the asymptotic covariance $\sigma^2 Q^{-1}$ of the estimators is as small as possible, minimizing some function $\phi\left(\sigma^2 Q^{-1}\right)$ measuring the size of the matrix. By contrast, the second approach is focused on the actual covariance matrix with finite sample T, minimizing a measure $\phi\left(\sigma^2 Q_T^{-1}\right)$. In general, approximate optimal designs are not appropriate for CA, as often the method leads to repetition of the product profiles. Therefore, we will not discuss this approach in detail. Nevertheless, it is useful to understand approximated designs in order to obtain full perspective of the problem. In the Appendix A we provide an overview, and present some results that will be mentioned later in the essay.

The design of conjoint experiments has traditionally focused on exact optimal designs. These designs minimize some function of $\sigma^2 Q_T^{-1}$, solving the problem $\min_{X\in\chi} \phi\left(\left(X'X\right)^{-1}\right)$, where $\phi(\cdot)$ is a measure of matrix size: trace in case of A-optimality, and the determinant for D-optimality. In the first case the sum of variances of the estimators is minimized; in the second, researchers also pursue uncorrelated estimators in the vector $\widehat{\beta}$. Exact optimal designs have several advantages in CA. First, they minimize the actual covariance of the estimators instead of an approximation. Besides, for optimal exact designs we can consider not only such constraints as $x \in \chi^T$, but we can also include transversal constraints, linking characteristics of product profiles (levels of categorical variables, or simply values of continuous variables). For example we can consider the prices for every attribute, or level, and include them in a budget constraint over the whole experiment $\sum_{t=1}^{T} c'x_t \leq m$, where c is a $k \times 1$ vector of attribute prices, and m is the total budget. Without loss of generality we can impose that the stimulus belongs to the space $\chi' = \left\{x \in \chi^T : g(x) \leq 0\right\}$. Once an exact optimal design Q_T has been computed, any design used in practice should be compared to this benchmark.

Several procedures have been considered in the literature. Dykstra (1971) suggested the iterative inclusion of additional profiles, using the recursive expressions for partitioned matrix $|Q_{T+1}| = |Q_T|\left(1 + f(x_{T+1})' Q_T^{-1} f(x_{T+1})\right)$. The algorithm sequentially selects one observation to improve the determinant, therefore at each iteration the profile x_{T+1} is chosen to maximize $f(x_{T+1})' Q_T^{-1} f(x_{T+1})$. If χ is finite (with factorial designs), this is done by swapping alternative profiles and evaluating the change in the determinant. Johnson and Nachtsheim (1983) consider some other alternatives. The exchange algorithm can be also applied for trace minimization, using the Woodbury matrix inversion identity $tr\ \left(Q_{T+1}^{-1}\right) = tr\ \left(Q_T^{-1}\right) - tr\ \left(\frac{Q_T^{-1} x_{T+1} x'_{T+1} Q_T^{-1}}{1 + x'_{T+1} Q_T x_{T+1}}\right)$. Besides, some procedures initially developed for Kiefer's approximated optimal designs can be applied also in this context, such as the Fedorov method (Fedorov 1972).

Table 2.1: Exchange algorithms for computing exact designs

Algorithm	Description
Simple exchange algorithm (Mitchell and Miller Jr 1970; Wynn 1972)	Starts with an random n-point design. At each iteration one observation is added which maximizes the determinant, and then another observation deleted to maximize the efficiency gain.
DETMAX (Mitchell 1974)	Starts with an random n-point design. At each iteration the algorithm makes "excursions" from a n-point design: it is permitted to add/delete more than 1 observation until the determinant is improved.
Fedorov (1972)	Starts with an n-point nonsingular design. At each iteration the algorithm simultaneously adds one observation and deletes another so that the increase in determinant is maximal.
Modified Fedorov (Cook and Nachtsheim 1980)	Starts with an n-point nonsingular design. At each iteration the algorithm evaluates all pairs of design and candidate points, and selects the best candidate to switch with each design observation. Makes every swap that increases efficiency.
Coordinate exchange (Meyer and Nachtsheim 1995)	Does not use the candidate set. At each iteration, the initial design is improved by exchanging each point coordinate (attribute level) with every other possible coordinate. Exchanges which increase efficiency are maintained.

Table 2.1 presents a comparative summary of the commonly applied exchange algorithms. A detailed comparison and evaluation of their computational performance can be found in Cook and Nachtsheim (1980). In general, none of these methods exploits satisfactorily the available numerical optimization tools. But there is a more relevant drawback. After the optimal design is computed, we typically observe that a few rows (product profiles) are repeated several times, which

is a major problem for its application in CA. This is not a surprising result, since the arguments of Lemma 1 in the Appendix A also apply to the set $\mathbf{Q}_T = \{Q = X'X : X = [f(x_1)', \ldots, f(x_T)']'\}$. Therefore, with exact optimal designs we end up with repeated vertex questions with certain frequencies, not very differently from Kiefer's approximate designs. The problem of replications was initially recognized by Box (1970), who noticed that in many situations optimal designs consist of replications of a small number of distinct experimental profiles. This poses a problem in the context of individual based CA, where the same respondent should not be questioned several times for the same product.

2.3 A Direct Method for Optimal Exact Designs in Classic CA

In this section we propose an efficient approach for computing exact optimal designs without repeated stimuli, providing the basis for usability of this approach in CA. First, we analyze the properties of optimal exact design problems, and discuss the approach to solve this optimization problem efficiently with Newton-based methods. Next, we demonstrate how to create designs without duplicated treatments, which often appear in optimal designs. We also present some initial numerical results.

2.3.1 Using Newton-Based Algorithms

The general setup for computing exact optimal designs is the following optimization problem

$$\begin{aligned} \min_{x} \quad & \phi\left(\left(X'X\right)^{-1}\right) \\ \text{s.t.} \quad & X = [f(x_1)' \ldots f(x_T)']' \\ & x \in \chi^T, \end{aligned} \tag{2.1}$$

where ϕ is a measure of the size of a matrix, trace or determinant. It is a convex problem, since the objective function ϕ is convex, and we assume that the feasible set of experimental attributes is a nonempty, compact and convex set. The solution, x^*, is the exact optimal design matrix. Note that the optimal design x^* is not unique, as any permutation of the rows in x^* (reordering the questions or product profiles) renders the same matrix $Q_T = X'X$. All of these solutions are equivalent. Lower and upper bounds on x represent the set of feasible attributes, χ^T. Different types of constraints can be considered to handle flexibly a variety of marketing scenarios and managerial problems: linear and nonlinear equality, or inequality constraints.

Table 2.2: First and second order derivatives of the benchmark problems

	D-optimality	A-optimality
Objective	$\min \lvert (X'X)^{-1} \rvert$	$\min \operatorname{tr} \left(X'X\right)^{-1}$
Gradient	$-2\left\lvert (X'X)^{-1}\right\rvert \text{ vec } X\left(X'X\right)^{-1}$	$-2 \text{ vec } X\left(X'X\right)^{-2}$
Hessian[a]	$4\left\lvert (X'X)^{-1}\right\rvert \left((X'X)^{-1} \otimes X\left(X'X\right)^{-1}X'\right) +$ $2\left\lvert (X'X)^{-1}\right\rvert K\left(X(X'X)^{-1} \otimes (X'X)^{-1}X'\right) +$ $2\left\lvert (X'X)^{-1}\right\rvert K\left((X'X)^{-1}X' \otimes X(X'X)^{-1}\right) -$ $2\left\lvert (X'X)^{-1}\right\rvert \left((X'X)^{-1} \otimes I\right)$	$4\left((X'X)^{-1} \otimes X\left(X'X\right)^{-2}X'\right) +$ $4\left((X'X)^{-2} \otimes X\left(X'X\right)^{-1}X'\right) -$ $2\left((X'X)^{-2} \otimes I\right)$

[a] K is the commutation matrix, which transforms vec X into vec X'.

There are several Newton-based algorithms for constrained convex programming, which posses good theoretical properties. To solve Problem (2.1) with a Newton's method, we first calculated the first- and second-order derivatives. Unless stated otherwise, numerical examples considered here assume that $f(x_t) = x_t$, and the design matrix $x = X$. Objective functions, gradients and Hessians for minimization of A- and D-optimality criteria for this benchmark case are presented in Table 2.2. In case of discrete attributes we also include intercept and consider transformation of variables to eliminate dummy collinearities. The proof for a more general expression can be found in the Appendix B, and can be easily adapted for other specifications of f.

We have solved several numerical examples and observed that exact optimal designs in fact have repeated profiles, as expected from applying Lemma 1 to the set $\mathbf{Q}_T = \left\{Q = X'X : X = [x'_1 \dots x'_T]'\right\}$. Below we discuss how to overcome this problem.

2.3.2 Avoiding Repeated Questions

The issue of duplicated product profiles can be resolved by imposing simple quadratic constraints on Problem (2.1), which prohibit profile repetitions in the matrix x. Define a $T \times T$ similarity matrix $S = xx'$, whose elements are $S_{i,j} = x_i' x_j$, where x_i, x_j are product profiles i,j. Notice that the Euclidean distance between them, $d_{i,j} = \sqrt{\left(x_i - x_j\right)' \left(x_i - x_j\right)}$, satisfies $d_{i,j}^2 = S_{i,i} + S_{j,j} - 2S_{i,j}$, and the matrix $D = [d_{ij}^2]$ can be expressed as

$$D = \text{diag}\,(S)\ 1_T' + 1_T\ \text{diag}\,(S)' - 2S,$$

where diag (S) is a vector of main diagonal elements in S, and 1_T is a $T \times 1$ vector of ones. Both S and D are symmetric matrices, and the diagonal elements in D are zero. We consider a lower bound over the Euclidean distance between stimulus i and stimulus j, for all pairs of different questions shown to the same respondent: $L(D) \geq \underline{d}$. The linear operator $L(\cdot) : \mathbb{R}^{T \times T} \rightarrow \mathbb{R}^{T(T-1)/2}$ selects the lower triangle elements of a square matrix (excluding the diagonal elements equal to 0, and the symmetric upper triangle terms), and stacks them in a column vector; $\underline{d}$ is a $T(T-1)/2$ vector of positive distance tolerances, and the inequality is applied pointwise. Notice that $L(D) = H \cdot vec(D)$, where $vec(\cdot) : \mathbb{R}^{T \times T} \rightarrow \mathbb{R}^{T^2}$ is the operator that stacks the columns of a matrix, and H is a $T(T-1)/2 \times T^2$ sparse matrix

$$H = \left(\begin{array}{cccccccccc}
\mathbf{0}_{(T-1)\times 1} & \mathbf{I}_{T-1} & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \\
\cdots & \cdots & \mathbf{0}_{(T-2)\times 2} & \mathbf{I}_{T-2} & \cdots & \cdots & \cdots & \cdots & \cdots & \\
\cdots & \cdots & \cdots & \cdots & \mathbf{0}_{(T-3)\times 3} & \mathbf{I}_{T-3} & \cdots & \cdots & \cdots & \mathbf{0}_{T\times T} \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \ddots & \cdots & \cdots & \\
0_{1\times 1} & \mathbf{0}_{1\times(T-1)} & \mathbf{0}_{1\times 2} & \mathbf{0}_{1\times(T-2)} & \mathbf{0}_{1\times 3} & \mathbf{0}_{1\times(T-3)} & \cdots & \mathbf{0}_{1\times(T-1)} & \mathbf{I}_1 &
\end{array} \right)$$

where $\mathbf{I}_r$ is the $r \times r$ identity matrix, and blank spaces are adequately sized blocks of zeros (as shown in the last row). For example for $T = 3$,

$$L\begin{pmatrix} d_{11} & d_{12} & d_{13} \\ d_{21} & d_{22} & d_{23} \\ d_{31} & d_{32} & d_{33} \end{pmatrix} = \begin{pmatrix} \mathbf{0} & \mathbf{1} & \mathbf{0} & 0 & 0 & 0 & 0 & 0 & 0 \\ \mathbf{0} & \mathbf{0} & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{0} & \mathbf{0} & \mathbf{1} & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} d_{11} \\ d_{21} \\ d_{31} \\ d_{12} \\ d_{22} \\ d_{32} \\ d_{13} \\ d_{23} \\ d_{33} \end{pmatrix} = \begin{pmatrix} d_{21} \\ d_{31} \\ d_{32} \end{pmatrix}$$

The matrix H can be obtained from the identity matrix I_{T^2}, by eliminating rows that correspond to diagonal and upper-triangle elements (for example, with $T = 3$ these are the rows 1,4,5,6,7,8).

We begin with the benchmark case of continuous attributes. Formally, our approach to compute exact designs without stimuli repetitions is to solve the following optimization problem

$$\begin{aligned} \min_{x} \quad & \phi\left(\left(X'X\right)^{-1}\right) \\ \text{s.t.} \quad & X = x = [x_1' \ldots x_T']' \\ & L\big(\text{diag}\,(xx')\; 1_t' + 1_t \;\text{diag}\,(xx'' - 2xx'\big) \geq \underline{d} \\ & lb \leq x \leq ub. \end{aligned} \tag{2.2}$$

We have computed several examples for two versions of this problem: (1) the case of A-optimality, where we minimize the trace of the covariance matrix, $\phi(\cdot) = \text{tr}\,(\cdot)$; and (2) the case of D-optimality, where we minimize its determinant, $\phi(\cdot) = |\cdot|$. We add $T(T-1)/2$ distance constraints discussed above, and lower (lb) and upper bounds (ub) on values of continuous attributes, which represent the set of feasible attributes, χ^T. Below, we discuss in detail the computational performance of this approach, as well as the comparative behavior of the trace and determinant algorithm. The case of categorical variables is presented in the next Section.

2.3.3 Numerical Results for Some Benchmark Problems

We performed a series of simulations to test the performance of the algorithm and to compare the behavior of both criteria, the trace and the determinant. The algorithm was implemented using MATLAB 6.5 on Mobile Workstation, Intel Core$^{\text{TM}}$2 Duo 2.20 GHz, with machine precision 10e-16. Both problems have been solved using the MATLAB subroutine "fmincon" with the option "interior-point" algorithm, included in the Optimization toolbox.

Since the 1980s *interior point methods* have become popular for solving nonlinear constrained problems (also large-scale). They are very efficient, both in terms of theoretical worst-case complexity and practical performance. The interior-point approach to constrained minimization is to solve a sequence of minimization problems perturbed by some parameter. As this parameter decreases to zero, the minimum of perturbed minimization problem should approach the minimum of original minimization problem (for details see e.g. Byrd et al. 1999). To solve the perturbed problem, we consider a Newton framework using a line search. Solving the Karush-Kuhn-Tucker equations, we first compute the Newton search direction, $p_k = -H_k^{-1}g$, where H is the exact Hessian $\nabla^2\phi(x)$, and g is the gradient $\nabla\phi(x)$. To guarantee global convergence, we then compute a step size that determines the adjustment of the Newton direction, ensuring sufficient decrease and uniform progress towards a solution (Nocedal and Wright 2006).

We have solved Problem (2.2) for conjoint experiments of different sizes: small, medium, and large, with varying parameters for the number of stimuli (T) and product attributes (k). We have chosen sufficiently large T to ensure sufficient number of degrees of freedom for estimation of integer cases and interactions, which is the subject of the next Section. We have also checked that with small values of $\underline{d}$, we overcome the problem of stimuli repetitions. The parameter values used in the simulation are shown in Table 2.3.

Similarly to other local algorithms, the performance of this approach may be sensitive to initial points, and the algorithm may be trapped in a local minimum. To inspect this problem we have re-run the procedure 100 times for each of the scenarios, solving Problem (2.2) with random initial points. In general the simulation results for both trace and determinant optimization are

Table 2.3: Parameter values for simulation of benchmark problems

Problem size	Small	Medium	Large
# profiles (T)	10	16	25
# attributes (k)	3	5	8
# model parameters (p)	3	5	8
Lower bound (lb)	1	1	1
Upper bound (ub)	10	10	10
Distance ($\underline{d}$)	5	5	5

consistent and the local solutions lie close. Therefore, the sensitivity to initial points does not pose a big threat to our approach.

In each of the scenarios we have chosen the best of the 100 simulated results, that is the design which leads to the smallest covariance matrix (for both trace and determinant criteria). We evaluated them in terms of the quality of the attained solution and the computational cost. For the former we report the objective function value, the rank of the optimal design matrix, and the conditioning of the information matrix. To allow for comparability of A- and D-optimality measures we calculate $\phi_1(X_d) = \text{tr } (X_d^{*'} X_d^*)^{-1}$, where X_d^* is the solution to the determinant problem, and $\phi_2(X_a) = |(X_a^{*'} X_a^*)^{-1}|$, with X_a - solution to the trace problem. The evaluation of the algorithm's computational cost is based on the number of iterations, function evaluations and time needed for convergence. The comparative summary is outlined in Table 2.4.

For both trace and determinant criteria the convergence of algorithms takes a few seconds, and for the majority of scenarios the solution was found in less than a second. The determinant criterion converges faster than the trace in all cases, however its performance is suspicious. We can observe that for medium-to-large scenarios the solution obtained with the trace algorithm yields better determinant values than the solution to the determinant problem (compare the determinant values in the left and right panel). This suggests that the determinant algorithm gets easily stuck in a local minimum.

As the dimension of X grows, the function $|(X'X)^{-1}|$ rapidly approaches 0, so that the objective function value becomes smaller than the "machine epsilon" (the upper bound on the relative

Table 2.4: Simulation results for trace and determinant problems

Objective function	min tr $(X'X)^{-1}$			$\min \lvert (X'X)^{-1} \rvert$		
Problem size	Small	Medium	Large	Small	Medium	Large
Trace[a]	$\underline{0.0090}$	$\underline{0.0109}$	$\underline{0.0128}$	0.0153	0.0337	0.0478
Determinant[a]	1.60e-8	1.80e-14	9.41e-24	$\underline{4.64\text{e-}8}$	$\underline{1.24\text{e-}12}$	$\underline{4.05\text{e-}20}$
# iterations	25	22	27	11	0	0
# function evaluations	26	23	29	12	1	1
Time (s)	0.14	0.37	2.96	0.07	0.01	0.08
Rank	3	5	8	3	5	8
Condition	4.08	7.79	13.46	8.28	39.09	80.21

[a] Underlined values are objective function values.

error due to rounding in arithmetic operations). It means that for any sufficiently large matrix X, the determinant of the inverse of the information matrix will be essentially zero (rounding off at the 16th decimal place). The algorithm does not iterate because any initial point leads to function value equal to 0, and is therefore identified as the solution. These results imply that direct optimization of D-optimality criterion often does not work well in practice. To moderate these problems, one could consider a logarithmic transformation of the objective function. However, in this case numerical optimization could be troublesome as well since the gradient

$$\frac{\partial}{\partial x}\ln\left|(X'X)^{-1}\right| = \frac{-1}{|(X'X)|}\,\frac{\partial}{\partial x}\,|(X'X)|$$

is ill-conditioned as $|X'X|$ approaches 0. Therefore, the trace criteria usually works better.

We have also analyzed the optimal designs qualitatively. In case of trace algorithm, indeed the values of elements in X are close to lower and upper bounds. In the determinant case which was stuck in a local minimum, the solutions were included in the sampling interval further from the boundaries (see Table 2.5 for the designs computed in the "Medium" Scenario). Therefore, for the trace, the optimal solution lies relatively close to the bounds of the problem, confirming the intuition that evaluating extreme stimuli yield the most information. The optimal design matrix is of full rank in both cases, however the solution to trace problem has better conditioning than the solution to determinant problem.

Table 2.5: Design matrices computed in “Medium” scenario

	min tr $\left(X'X\right)^{-1}$					min $\lvert\left(X'X\right)^{-1}\rvert$				
Profile	A1	A2	A3	A4	A5	A1	A2	A3	A4	A5
1	1.00	10.00	10.00	1.00	10.00	5.25	2.30	8.78	5.92	3.79
2	1.00	10.00	10.00	10.00	1.00	2.73	7.15	3.51	8.03	8.49
3	1.00	1.00	10.00	10.00	10.00	7.57	5.38	3.77	6.74	2.52
4	1.00	1.00	10.00	10.00	1.00	2.86	9.73	1.06	9.27	1.41
5	10.00	10.00	10.00	1.00	1.00	6.79	8.09	5.58	6.28	9.92
6	10.00	10.00	1.00	1.00	10.00	5.69	8.33	5.48	8.96	2.18
7	1.00	1.00	10.00	1.00	10.00	8.99	7.70	9.11	1.32	8.05
8	1.00	1.00	1.00	10.00	10.00	8.83	2.03	1.09	4.41	9.07
9	1.00	10.00	1.00	10.00	10.00	1.06	7.41	7.54	8.46	5.33
10	10.00	1.00	10.00	1.00	1.00	8.62	8.51	1.43	1.03	5.51
11	10.00	10.00	1.00	10.00	1.00	8.78	1.92	5.39	9.70	2.72
12	10.00	1.00	10.00	1.00	10.00	3.99	2.25	8.04	8.60	3.61
13	10.00	1.00	10.00	10.00	1.00	5.99	1.47	8.83	5.01	6.86
14	1.00	10.00	10.00	1.00	1.00	3.45	9.44	4.45	3.54	7.88
15	10.00	1.00	1.00	1.00	10.00	3.62	9.46	7.75	3.34	4.94
16	10.00	1.00	1.00	10.00	1.00	9.65	2.97	2.13	4.71	6.66

Summarizing, the optimization of trace criterion is a more reliable approach, because its performance is not significantly affected by the increase in the problem dimension: the number of iterations and function evaluations remains relatively stable across scenarios, and the solutions obtained with different initial points are close. The convergence is fast and even for the largest problems it does not take longer than 3 seconds. Given the high chances of converging to a suboptimal local minimum, the stability of trace criterion becomes a useful advantage, outperforming the determinant in practice. Therefore, we will focus on the trace in the remainder of this chapter.

2.4 The Case of Discrete and Mixed Attributes

In CA we often find discrete attributes. For example, whether a certain material is used or not, or a component has been selected from a given catalogue. Continuous variables, like prices, are also often represented by a small number of meaningful levels and treated as discrete variables.

Typically, CA models have several categorical and perhaps also some continuous attributes. In an experimental context, the discrete attribute is known as a factor, and the alternative values that it can take are known as levels of the factor. The standard formulation in a model with $i = 1,....,L$ integer attributes, each of them having J_i levels is

$$y_t = \alpha + \sum_{i=1}^{L}\sum_{j=1}^{J_i} \gamma_{ij}\, d_{tij} + \beta' Z_t + \varepsilon_t, \tag{2.3}$$

where Z_t represents the set of p continuous regressors. The design matrix $x = [D_1,\ldots,D_L,Z]$ is partitioned in a way that every discrete attribute is represented by a matrix of indicator variables, $D_i = [d_{tij}]$, taking values 0 or 1 to indicate the absence or presence of a level in the profile. Linear regression models with discrete dummies, like the one defined in equation (2.3), are affected by collinearities as $\sum_{j=1}^{J_i} d_{tij} = 1,\ \forall\ i,t$. We consider the standard methods to eliminate multicollinearity from the model: (1) omission of one level in every factor, and (2) including dummy differences with respect to one factor. Depending on the selected method, the OLS estimators will be different as well as their covariance matrix, and we will obtain different optimal designs x^*.

(D1) The first approach involves substituting the regressor identity in the model. For example with $d_{tiJ_i} = 1 - \sum_{j=1}^{J_i-1} d_{tij}$ we can express

$$\begin{aligned} y_t &= \alpha + \sum_{i=1}^{L}\left(\sum_{j=1}^{J_i-1}\gamma_{ij}\, d_{tij} + \sum_{i=1}^{L}\gamma_{iJ_i}\left(1 - \sum_{j=1}^{J_i-1} d_{tij}\right)\right) + \beta' Z_t + \varepsilon_t \\ &= \left(\alpha + \sum_{i=1}^{L}\gamma_{iJ_i}\right) + \sum_{i=1}^{L}\sum_{j=1}^{J_i-1}\left(\gamma_{ij} - \gamma_{iJ_i}\right) d_{tij} + \beta' Z_t + \varepsilon_t, \end{aligned}$$

which is equivalent to **level omission**, and the interpretation of coefficients is relative to the parameter of a missing level. In a model with more factors, transformation by level omission can be conveniently written in a matrix form, $\tilde{f}(x) = xA$. The matrix A can be obtained from the identity matrix of the size $\left(\sum_{i=1}^{L} J_i + k\right)$ by eliminating columns associated to the omitted levels. This method is sometimes called *binary coding*.

(D2) In the second approach additional constraints are included, usually that $\sum_{j=1}^{J_i} \gamma_{ij} = 0$ (the dummy coefficients sum up zero). Substituting $\gamma_{iJ_i} = -\sum_{j=1}^{J_i-1} \gamma_{ij}$ in the model leads to

$$y_t = \alpha + \sum_{i=1}^{L} \sum_{j=1}^{J_i-1} \gamma_j \left(d_{tj} - d_{tiJ_i}\right) + \beta' Z_t + \varepsilon_t,$$

where new regressors are defined as **dummy differences**, $d'_{tij} = \left(d_{tij} - d_{tiJ_i}\right) \in \{-1, 0, 1\}$. Let $\tilde{f}(x) = x(I - B)A$ represent the transformation of dummy variables, which creates dummy differences with respect to the last level in each factor. In particular, I is an identity matrix of the size $\left(\sum_{i=1}^{L} J_i + k\right)$, A is defined above, and B is a square sparse matrix with value 1 at columns associated with the omitted levels and zero otherwise. This method is sometimes called *effects coding*.

In our examples, the design matrix will be partitioned as $X = f(x) = [1, \tilde{f}(x)]$, where the first column corresponds to the intercept, and $\tilde{f}$ represents the dummy coding method (D1 and D2).

Note that when the number of factors, L, is very small (one or two), and there are no continuous attributes, the number of different stimulus profiles that can be included in the experiment is small, and the experimental design problem is not relevant. All possible combinations of factor levels can be included in the experiment. Moreover, since replications are not allowed in CA, the inference analysis should be based on small sample analysis (typically under normality assumptions). But when the number of factors is large we may have larger size T, because the number of alternative stimuli increases multiplicatively, as $\prod_{i=1}^{L} J_i$, whereas the number of parameters increases additively. In this case, the experimental design does become important, as well as for the mixed CA (with both discrete and continuous attributes).

The selected procedure (D1) or (D2) affects the interpretation of the model parameters, but it does not essentially affect the efficiency (we can directly recover the exact OLS estimations from one method to other). We obtain the optimal design by minimizing the trace or the determinant of $\left(X'X\right)^{-1}$. The determinant is a more popular criterion, but it has limitations, which we discussed in the previous section. When there are no continuous attributes, the approach (D1) renders orthogonal designs, as the observation vectors for different dummies are naturally orthogonal. In the second approach, the columns in X often sum up to zero, rendering a balanced design (all attributes appear with the same frequency at each level). Nevertheless the trace/determinant values of the optimal matrix $\left(X'X\right)^{-1}$ might be quite different for both approaches. In any case, the relative efficiency of optimal solutions from each coding approach should not be directly compared as each procedure estimates different parameters.

To handle discrete attributes, we consider a branch-and-bound algorithm searching a tree, whose nodes correspond to continuous nonlinearly constrained optimization problems. The solvers have been compiled in both a sparse and a dense version, and they are commercially available with TOMLAB (http://tomopt.com/tomlab/) - a software package in MATLAB for practical solution of optimization problems. TOMLAB includes several solvers for the solution of all types of applied optimization problems. In particular we consider MINLP solver developed by Roger Fletcher and Sven Leyffer at the University of Dundee. MINLP implements a branch-and-bound algorithm and a sequential quadratic programming (SQP) trust region algorithm, using a recently developed filter technique to promote global convergence (Leyffer 2001).

Formally, the optimal design problem in the mixed-integer conjoint context is

$$\begin{aligned}
\min_{x} \quad & \text{tr } \left(X'X\right)^{-1} && (2.4)\\
\text{s.t} \quad & X = f(x) = [1, \tilde{f}(x)] \\
& x = [D_1, \dots, D_L, Z] \\
& \sum_{j=1}^{J_i} d_{tij} = 1, \ \forall \ i, t \\
& L\left(diag\left(xx'\right)1' + 1diag\left(xx'\right)' - 2xx'\right) \geq \ \underline{d} && (2.5)\\
& lb \leq Z \leq ub \\
& d_{tij} \in \{0,1\} \text{ are integer,}
\end{aligned}$$

where we choose optimally the value of continuous attributes, as well as the factor level to be shown in each stimulus. We include the intercept, as well as transformation to eliminate perfect collinearity in the dummy variables, $\tilde{f}(x)$. The third constraint requires that within each factor exactly one level is shown in a product profile, and is a simple linear equality constraint. We also impose the similarity constraint to avoid repetitions (equation (2.5)). Additionally lower and upper bounds on variables can be considered, which for dummy variables are naturally 0 and 1. To obtain subroutine inputs: the gradient and Hessian, we can apply directly the formula derived in the Appendix B. Note that the respective transformation matrices A,B are constant, therefore these expressions are further simplified.

We performed a series of simulations for one integer and two mixed examples for both trace and determinant approaches. However, we only report the results for the trace, because it is a more reliable and stable criterion. Table 2.6 summarizes parameter values for different scenarios.

Table 2.6: Parameter values for simulation of discrete scenarios

	Scenario 1	Scenario 2	Scenario 3
Type	integer	mixed	mixed
# profiles (T)	10	16	25
# continuous attributes (k)	0	2	5
# integer attributes (L)	3	3	3
# attribute levels (J_i)	[3,3,3]	[3,3,3]	[3,3,3]
# model parameters (p)	7	9	12
Lower bound[a] (lb)	-	1	1
Upper bound[a] (ub)	-	10	10
Distance ($\underline{d}$)	1	1	1

[a] Lower and upper bounds considered on the set of continuous attributes, Z.

As in the previous Section, for each of the scenarios and collinearity methods (D1 and D2) we first re-run the algorithm a few times to make sure it does not attain a local solution, and then we choose the design associated with the smallest size of the covariance matrix. Some characteristics of these designs are summarized in Table 2.7. Recall that the covariances of the two methods to eliminate collinearity in the dummies cannot be directly compared in terms of efficiency, as they estimate different parameters. As far as the quality of the solution is concerned, in all scenarios the transformed design matrix $X = f(x)$ is of full rank, and for Scenario 1 the collinearity problem is eliminated. The optimal design matrix obtained with "dummy differences" algorithm has in general better conditioning, than the one obtained when omitting one level. In terms of computational cost, the performance of both approaches is similar.

Some of the conclusions drawn from the continuous conjoint problem are confirmed here. As expected, including additional profiles, attributes and factor levels increases the optimization costs: more time and function evaluations are needed to converge to the optimum. The convergence for the pure integer scenario takes seconds. The mixed-integer problem is more complex and computationally challenging. It takes from above 1 minute to 6 minutes for the algorithm to converge, while the number of function evaluations remains quite stable in both scenarios.

Table 2.7: Simulation results for mixed and integer designs

Objective function	min tr $\left(X'X\right)^{-1}$					
	Scenario 1		Scenario 2		Scenario 3	
Approach to collinearity[a]	D1	D2	D1	D2	D1	D2
Trace[b]	4.0625	1.3333	2.6204	0.8998	1.6600	0.5938
Determinant	0.0023	3.18e-6	3.54 e-9	5.84e-12	2.58e-18	7.14e-21
# iterations	2	2	1	1	1	1
# function evaluations	103	70	68	92	131	141
Time (s)	6.88	9.77	81.49	75.71	362.01	342.920
Rank	7	7	9	9	12	12
Condition	16	4	439.94	123.87	1549.60	287.81

[a] D1 - level omission; D2 - including dummy differences.

[b] Underlined values are objective function values.

2.4.1 Model with Interactions: Fractional Factorial Designs

Consider a CA model with several factors (discrete attributes), where each factor may take different levels. A full factorial model considers all possible interactions for each dummy in the model (factors and levels).

$$y_t = \gamma + \sum_{j_1=1}^{J_1} .... \sum_{j_L=1}^{J_L} \beta_{j_1,j_2,...,j_K} \times \left(d_{tj_1} d_{tj_2} \cdots d_{tj_K}\right) + \varepsilon_t.$$

The number of parameters increases multiplicatively with $J_1 \times ...J_L$. The model can also include continuous attributes. Then we can have also interaction between the dummies and the continuous attribute.

In any case, the effort required to estimate a full factorial model is cost-prohibitive and tedious for the respondent. In practice researchers generally use fractional-factorial designs, containing just interactions of a few factors (e.g. products of pairs, or threesomes of dummies), and evaluating fewer product profiles. For an introduction see Addelman (1962), Green (1974), or Kuhfeld et al. (1994).

Our approach can also handle fractional factorial designs. To estimate the model we first have to eliminate multicollinearity, with either of the methods discussed in the previous section. Additionally, we need to assure sufficient number of degrees of freedom, meaning that the model with interactions requires more product profiles to be evaluated by the respondents (at least 1 observation per each interaction term). Then the optimal design for the model with two-way interactions is the solution to the following optimization problem

$$\begin{aligned} \min_x \quad & \text{tr } \left(X'X\right)^{-1} \\ \text{s.t} \quad & X = [1, \tilde{f}(x), W(x)], \\ & x = [D_1, \ldots, D_L, Z], \end{aligned} \tag{2.6}$$

where $W(x)$ is a matrix block representing all possible interaction elements between 2 variables, and the remaining constraints, dummy blocks, and functions are specified as in Problem (2.4). Note that the interaction terms are specified as a function of the design matrix, x. Mathematical details are given in the Appendix B.

We provide some examples to illustrate the behavior of the method, including Scenario 2 from the previous section, and omitting a factor level to eliminate multicollinearity (approach D1). The examples consider a model with two-way interactions between: 2 continuous attributes ("Continuous" case), a continuous and a categorical variable ("Mixed" case), and 2 categorical variables ("Integer" case). Table 2.8 summarizes the characteristics of the optimal design for the model with interactions.

The performance of the approach for the "Continuous" and "Mixed" case is very good. Including interactions does not result in the increase in computational cost, in comparison to the model without interactions. As far as the quality of the solution is concerned the optimal design matrix is of full rank. However, "the curse of dimensionality" affects the performance of the proposed approach as the standard Branch-and-Bound algorithm is considered. Other alternative algorithms can be considered to tackle this issue (Lawler and Wood 1966).

Table 2.8: Simulation results for a model with interactions

Objective function		$\min \operatorname{tr} \left(X'X\right)^{-1}$	
Type of interaction	Continuous	Mixed	Integer
Trace[a]	2.6754	2.6921	14.2166
Determinant	6.70e-12	6.77e-12	2.41e-7
# iterations	1	1	60
# function evaluations	50	31	1394
Time (s)	116.78	52.45	2457.00
Rank	10	11	13
Condition	3441.00	1758.80	5244.60

[a] Underlined values are objective function values.

2.4.2 A Comparison with Commonly Used Software

We have compared the performance of our approach with the software which is commonly used by practitioners in traditional conjoint experiments: Conjoint Value Analysis (CVA) by Sawtooth Software and %MktEx by SAS. Both programs allow only categorical attributes and rely on exchange algorithms (see Table 2.1) to optimize the determinant of the covariance matrix. Continuous attributes like prices are not explicitly permitted. Instead, they are usually discretized and represented by a few meaningful levels.

The setting is as follows. To ensure comparability of results with Sawtooth Software and SAS, we focus exclusively on the experiments with discrete attributes and begin with determinant minimization. The design matrix X has an intercept, and categorical variables are orthogonally coded, which is a common practice in CA (for details see Kuhfeld 2010). The values of orthogonal codes of dummy variables with 2 and 3 levels are presented in Table 2.9.

Table 2.9: Orthogonal coding of dummy variables

2-level dummy			3-level dummy				
Original		Orthogonal	Original			Orthogonal	
1	0	1.0000	1	0	0	1.3660	-0.3660
0	1	-1.0000	0	1	0	-0.3660	1.3660
			0	0	1	-1.0000	-1.0000

We have created 4 hypothetical conjoint experiments, with a varying number of product profiles (T), categorical product attributes (L), and attribute levels (J) (see the upper panel in Table 2.10). As far as the profile repetitions are concerned, the designs obtained with our approach will never have duplicated observations, for SAS we have activated the “no duplicates” options , and Sawtooth Software’s CVA does not take this issue into account. To achieve additional efficiency gains in the performance of our algorithms we used sparse versions of the constraints and their derivatives, taking into account patterns of non-zero elements in the corresponding matrices. For each of the scenarios, we have run 10 times our “determinant” algorithm, and chosen the design with the smallest objective function value.

If orthogonality is imposed in the design (meaning $X'X$ is diagonal), then the trace and determinant are closely related. Denote by v_i the sample variances of each regressor. For orthogonal regressors, the A-optimality criterion minimizes $\sum_{i=1}^{p}(1/v_i)$, and the D-optimality criterion minimizes $\prod_{i=1}^{p}(1/v_i)$. By the inequality of arithmetic and geometric means, we obtain that

$$tr\left(Q_T^{-1}\right) = \sum_{i=1}^{p} \frac{1}{v_i} \leq p \left(\prod_{i=1}^{p} \frac{1}{v_i} \right)^{1/p} \leq p \left| Q_T^{-1} \right|^{1/p}$$

holds with equality if and only if all variances are identical. In particular this occurs for pure factorial designs (discrete attributes only) with binary coding of dummies. Notice that for binary regressors the variance $v_i = p_i(1-p_i)/T$ where p_i is the frequency of level 1. Then $1/v_i$ is minimized when $p_i = 0.5$, i.e. when the same number of 0s and 1s is included for all regressors, and therefore both criteria are equal. This happens only at the minimum of both criteria. Nevertheless we have found that trace minimization renders better numerical results. When optimizing the determinant, we again encounter the problem of ill-conditioning of the information matrix. For larger conjoint experiments the determinant of the covariance matrix is smaller than machine epsilon, and therefore the round-off objective function value is 0. This prevents the algorithm from iterating towards a better solution. Minimizing the trace we search implicitly for the same optimum, but the problem has a much better numerical behavior.

The lower panel of Table 2.10 summarizes optimization results and design characteristics obtained with SAS, Sawtooth Software and our both approaches: minimizing the determinant and trace. In all cases we report two efficiency measures: determinant and trace of the optimal design. Recall, that the optimal design in SAS, Sawtooth Software, and "determinant" version of our approach is computed by minimizing the determinant. Finally, we also present the results from our approach based on trace minimization. When available, we provide a few measures of algorithm's computational cost: time to converge, number of iterations, function evaluations, and for SAS number of operations needed to find the design[1].

For small conjoint experiments (COMP1 and COMP2) our "determinant" approach achieves the same design efficiency as SAS and Sawtooth Software at a lower computational cost. For larger conjoint experiments numerical optimization of the determinant is problematic, and the algorithm gets stuck in a local minimum, which is a problem also for Sawtooth Software (COMP4). On the other hand, when minimizing the trace, which is a more stable criterion, in all scenarios we achieve the same design efficiency as SAS: the traces and determinants of covariance matrices calculated with SAS and our approach are equal. Moreover, our "trace" algorithm performs

[1] Number of operations is calculated as the sum of the following positions in SAS output: # algorithm searches, # design searches, # design refinements.

Table 2.10: Comparison with other software

Scenario		COMP1	COMP2	COMP3	COMP4
Parameters					
	# profiles (T)	8	10	16	18
	# attributes (L)	4	3	4	5
	# levels (J)	[2, 2, 2, 2]	[3, 3, 3]	[3, 3, 3, 3]	[3, 3, 3, 3, 3]
SAS					
	Determinant	3.05e-5	1.18e-7	1.91e-11	1.56e-14
	Trace^a	0.6250	0.7292	0.5972	0.6111
	Time (s)	2.60	3.84	4.00	3.25
	# Operations	1	82	61	1
Sawtooth Software					
	Determinant	3.05e-5	1.18e-7	1.91e-11	1.75e-14
	Trace^a	0.6250	0.7292	0.5972	0.6250
	Time (s)	1	1	4	15
Our approach minimizing determinant					
	Determinant	3.05e-5	1.18e-7	3.19e-11	1.19e-13
	Trace^a	0.6250	0.7292	0.6729	0.9395
	Time (s)	0.12	0.61	0.31	0.50
	# Iterations	1	3	1	1
	# Function evaluations	4	40	4	4
Our approach minimizing trace					
	Trace	0.6250	0.7292	0.5972	0.6111
	Determinant^b	3.05e-5	1.18e-7	1.91e-11	1.56e-14
	Time (s)	0.27	0.33	2.04	3.95
	# Iterations	1	1	1	1
	# Function evaluations	12	10	42	28

^{a} Trace of the covariance matrix of the D-optimal design.

^{b} Determinant of the covariance matrix of the A-optimal design.

faster than SAS in 3 out of 4 cases. Table 2.11 presents the optimal designs in COMP2 example obtained with SAS, Sawtooth Software, and our "determinant" and "trace" approach.

Table 2.11: Optimal designs computed in "Comparison 2"

	SAS			Sawtooth Software			"Det" approach			"Trace" approach		
Attribute	A1	A2	A3	A1	A2	A3	A1	A2	A3	A1	A2	A3
Profile												
1	1	3	1	1	2	1	2	2	1	1	2	1
2	3	3	2	1	3	2	1	2	2	3	1	1
3	2	2	1	3	3	3	1	1	2	2	1	1
4	3	2	3	1	1	3	3	3	1	1	3	3
5	2	1	2	2	2	3	3	2	3	1	1	2
6	1	2	3	2	1	2	1	3	3	2	3	2
7	3	1	1	3	2	2	2	3	2	3	1	3
8	1	1	3	2	3	1	2	1	1	2	2	3
9	2	3	3	3	1	1	3	2	1	3	2	2
10	1	2	2	1	3	3	2	1	3	3	3	1
						Level balance						
Level 1	4	3	3	4	3	3	3	3	4	3	4	4
Level 2	3	4	3	3	3	3	4	4	3	3	3	3
Level 3	3	3	4	3	4	4	3	3	3	4	3	3

In this section we presented only a part of functionality of our approach. To ensure comparability with available CA software, we have limited the scope of comparative examples to the experiments where the treatments are only categorical variables. Our "trace" approach achieves the same efficiency as SAS, and is faster in most of the examples considered. Despite the problems with numerical optimization of the determinant function, our "determinant" algorithm still matched %MktEx macro in terms of design efficiency in two of the scenarios, outperforming it in terms of the computational cost. Furthermore, our approach is far more flexible and provides functionalities which are not available either in SAS or Sawtooth Software. We can optimize the trace or determinant, and handle continuous and/or discrete variables. Additionally, we can solve problems by imposing quite general linear and nonlinear constraints, for example experimental budget restrictions. Our approach combines the flexibility and computational efficiency, which gives it an overall advantage compared to the existing software. Next we discuss extensions of the method to many other contexts, and we focus on the trace.

2.5 Optimal Designs: Extension to Customer Panels

There are many practical benefits of using consumer panels in conjoint studies. With a relatively homogeneous sample of respondents the experiment requires a few profile evaluations per respondent, reducing fatigue and learning effects. Homogeneous respondents may have identical preferences, but they might report their utility ratings with a random origin of coordinates. In other words, the response measure is an interval scale rather than a ratio scale in the taxonomy of Stevens (1951). CA researchers can handle this problem introducing heterogeneous intercepts. However, if we take a small number of measures for each individual, we cannot estimate the specific value of the intercept for each one. Alternatively, we can handle the problem using a random effects model. For $i = 1,...,N$ respondents, and T_i questions per individual, we consider the model

$$y_{it} = \eta_i + f(x_{it})'\beta + \varepsilon_{ti}, \tag{2.7}$$

where η_i are exogenous random variables with mean 0 and variance σ_η^2. If we include this effect in an overall shock $u_{it} = \eta_i + \varepsilon_{it}$, then the autocovariance matrix for each respondent is $E(u_i u_i') = \Omega = \left(\sigma_\eta^2 11' + \sigma_\varepsilon^2 I_T\right)$ has a special structure with $\sigma_\eta^2 + \sigma_\varepsilon^2$ as diagonal elements, and σ_η^2 otherwise. Finally, the panel is balanced if $T_i = T$ for each respondent (we assume this to simplify notation).

Consider the matrix notation $X = \left(f(x_{11})', \dots, f(x_{1T})', \dots, f(x_{N1})', \dots, f(x_{NT})'\right)' \in \mathbb{R}^{NT\times p}$, the vector of responses $Y = (y_{11}, \dots, y_{1T}, \dots, y_{N1}, \dots, y_{NT}) \in \mathbb{R}^{NT\times 1}$ and $u \in \mathbb{R}^{NT\times 1}$ analogously to Y. Then we can estimate consistently by OLS using $\widehat{\beta} = \left(X'X\right)^{-1} X'Y$. But this estimation is quite inefficient, as $Var[u] = (I_N \otimes \Omega)$. The method is not even consistent if η_i is correlated with some regressor (e.g. a socio-demographic block factor). In this Section we apply our approach to build exact optimal designs in the context of conjoint panels. We consider two of the most popular ways to estimate panels, which are consistent even when endogenous random effects are intrinsically eliminated. With a panel of consumers question repetitions are a concern only for an individual respondent. We can avoid them by introducing a lower bound on distances between product profiles for every individual. However, we do not forbid question repetitions across individuals.

2.5.1 Within-Groups (WG) Estimation

One commonly used way to eliminate the individual-specific effects, η_i, is to subtract time averages from the original panel model (2.7), leading to the within-groups (WG) model

$$\ddot{y}_{it} = \ddot{f}(x_{it})'\beta + \ddot{\varepsilon}_{it}$$

where $\ddot{y}_{it} = y_{it} - \bar{y}_i$, $\ddot{f}(x_{it}) = f(x_{itk}) - \overline{f(x_{ik})}$, and $\ddot{\varepsilon}_{it} = \varepsilon_{it} - \bar{\varepsilon}_i$. Stacking the observations for all individuals, such that $Y = \left(y'_{1t}, y'_{2t}, \ldots, y'_{Nt}\right)'$, $X = \left(f(x_{1t})', f(x_{2t})', \ldots, f(x_{Nt})'\right)'$, and $\varepsilon = \left(\varepsilon'_{1t}, \varepsilon'_{2t}, \ldots, \varepsilon'_{Nt}\right)'$ the equivalent compact form model is

$$MY = MX\beta + M\varepsilon$$

with $M = I_{NT} - \left(I_N \otimes \frac{1}{T} 1_T 1'_T\right) = I_{NT} - P$. Both M and P are idempotent matrices, and premultiplication by the matrix M creates deviation from the mean. We obtain mean centered data and the individual effect η_i disappears (because $\bar{\eta}_i = \eta_i$). Then OLS estimator is $\hat{\beta}_{OLS} = \left(X'MX\right)^{-1} X'MY$, with the variance

$$Var\left(\hat{\beta}_{OLS}\right) = \sigma^2_\varepsilon \left(X'MX\right)^{-1}.$$

Assuming vector preferences the design matrix is $X = f(x) = x$, and with a symmetric, constant matrix M we can directly apply the results of Proposition 3. Table 2.12 presents the analytical derivatives for the WG problem.

2.5.2 Estimation Based on Differences

Another way to eliminate the individual effect η_i is to take increments, so that

Table 2.12: Analytical derivatives for the WG problem

Objective	$\min_X \operatorname{tr} \left(X'MX\right)^{-1}$
Gradient	$-2 \operatorname{vec} MX\left(X'MX\right)^{-2}$
Hessian	$4\left[\left(X'MX\right)^{-1} \otimes MX\left(X'MX\right)^{-2} X'M\right]+$ $4\left[\left(X'MX\right)^{-2} \otimes MX\left(X'MX\right)^{-1} X'M\right]-$ $2\left[\left(X'MX\right)^{-2} \otimes M\right]$

$$\Delta y_{it} = \Delta X'_{it}\beta + \Delta\varepsilon_{it},$$

where $\Delta y_{it} = y_{it} - y_{i(t-1)}$, $\Delta X_{it} = \Delta f(x_{it}) = f(x_{it}) - f\left(x_{i(t-1)}\right)$, and $\Delta\varepsilon_{it} = \varepsilon_{it} - \varepsilon_{i(t-1)}$. Define the matrices

$$\Delta_T = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & \dots & 0 \\ & & \dots & & \\ 0 & \dots & 0 & -1 & 1 \end{bmatrix}_{(T-1)\times T} , \quad H = \begin{pmatrix} 2 & -1 & \dots & 0 & 0 \\ -1 & 2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \ddots & 2 & -1 \\ 0 & 0 & \dots & -1 & 2 \end{pmatrix}_{(T-1)\times(T-1)} .$$

Now, let's stack the observations for N individuals in a column to obtain a compact form of the transformed model

$$DY = DX\beta + D\varepsilon,$$

with a differentiation matrix $D = (I_N \otimes \Delta_T)$. To estimate the resulting model efficiently we have to apply GLS since $\{\Delta\varepsilon_{it}\}$ follows a non invertible MA(1), which implies that $E\left[\Delta\varepsilon_i\Delta\varepsilon'_i\right] = \sigma^2_\varepsilon H$. The GLS estimator with N customers and T questions for each one is

$$\hat{\beta} = \left(\sum_{i=1}^{N} \sum_{t=2}^{T} \Delta X_{it} H^{-1} \Delta X_{it}' \right)^{-1} \sum_{i=1}^{N} \sum_{t=2}^{T} \Delta X_{it} H^{-1} \Delta y_{it} = \left(X' D' \mathbf{H}^{-1} D X \right)^{-1} X' D' \mathbf{H}^{-1} D Y,$$

where $\mathbf{H} = (I_N \otimes H)$ is analogous to H but with dimension $N(T-1)$. Notice that $\hat{\beta}$ is an unbiased estimator with non singular variance

$$Var\left(\hat{\beta}\right) = \sigma_{\varepsilon}^2 \left(\sum_{i=1}^{N} \sum_{t=2}^{T} \Delta X_{it} H^{-1} \Delta X_{it}' \right)^{-1} = \sigma_{\varepsilon}^2 \left(X' D' \mathbf{H}^{-1} D X \right)^{-1},$$

An exact optimal design for this method should minimize $\phi \left[\left(X' D' \mathbf{H}^{-1} D X \right)^{-1} \right]$.

The analytical derivatives to minimize the trace in the GLS problem are explicitly given in Proposition 3, because $Z = D' H^{-1} D$ is a constant matrix. Table 2.13 presents the solution to the classic conjoint model with vector preferences.

Table 2.13: Analytical derivatives for the GLS estimator in a differenced model

Objective	$\min_X \ \text{tr} \left(X' D' H^{-1} D X \right)^{-1}$
Gradient	$-2 \ \text{vec} \, D' H^{-1} D X \left(X' D' H^{-1} D X \right)^{-2}$
Hessian	$4 \left[\left(X' D' H^{-1} D X \right)^{-1} \otimes D' H^{-1} D X \left(X' D' H^{-1} D X \right)^{-2} X' D' H^{-1} D \right] +$ $4 \left[\left(X' D' H^{-1} D X \right)^{-2} \otimes D' H^{-1} D X \left(X' D' H^{-1} D X \right)^{-1} X' D' H^{-1} D \right] -$ $2 \left[\left(X' H^{-1} X \right)^{-2} \otimes D' H^{-1} D \right]$

2.5.3 Numerical Results

We report some simulations for panels, analogous to the case where we considered a single consumer. Here we simulate a panel with $N = 10$ consumers, each of whom is shown 5 stimuli profiles (T). The remaining simulation parameters can be found in Table 2.3.

Table 2.14 presents the results of a simulation for the conjoint panel estimated with WG and a GLS-in-differences approach. We run the algorithm 100 times with random initial points, and select the most efficient design. Comparing the results with the case of the individual respondent (Table 2.4), we observe that using a panel of consumers leads to significant efficiency gains at a relatively small optimization cost. The algorithm converges in seconds in all cases, and the con-

Table 2.14: Simulation results for conjoint panels

	WG model			GLS in differences model		
Objective function	min tr $\left(X'MX\right)^{-1}$			min tr$\left(X'D'H^{-1}DX\right)^{-1}$		
Problem size	Small	Medium	Large	Small	Medium	Large
Trace[a]	$\underline{0.00309}$	$\underline{0.00516}$	$\underline{0.00829}$	$\underline{0.00028}$	$\underline{0.00076}$	$\underline{0.00187}$
Determinant	1.09e-9	1.16e-15	1.29e-24	6.28e-13	5.04e-20	3.73e-30
# iterations	15	16	20	44	63	184
# function evaluations	16	17	21	45	68	190
Time (s)	0.34	0.80	2.70	0.96	3.21	24.36
Rank	3	5	8	3	5	8
Condition	1.03	1.18	1.32	2.78	3.45	4.36

[a] Underlined values are objective function values.

ditioning of the full-rank design matrix is good. The performance of WG and GLS-in-differences approaches is similar. The within-groups optimal design is faster to compute, but as expected the objective function is worse. The problem for GLS in a differenced model is slower, but the solution is more stable - the minima lie very close.

2.6 Designing CA Studies with Invariance to Monotonous Transformations

Conceptually, a monotonous transformation of a utility function does not change the associated preorder of preferences. A drawback of regression models is that conditional means are not invariant to monotonous transformations of the response variable. Assume that $E[y|x] = f(x)'\beta$, then given a monotonous transformation h, in general

$$E[h(y)|x] \neq h(E[y|x]) = h\left(f(x)'\beta\right).$$

The lack of invariance to monotonous transformation is generally a nuisance, although for rating-based CA analysis it can accepted and OLS estimators are commonly considered in this context. But if the analysis is based on preference ordering, albeit OLS is a valid method to estimate $E[y|x]$, it is quite hard to accept the lack of invariance in the estimated utility function.

In this section we consider the question: is there any general method to build CA models, which is invariant to monotonous transformations of the response variable? The answer is positive: the conditional median or 0.5 conditional quantile. Note that the α-quantile of the conditional distribution of $y|x$ is

$$F_{y|x}^{-1}(\alpha) = inf\{t : \Pr(y \leq t|x) \geq \alpha\}.$$

Assuming that the conditional median is linear in parameters, $F_{y|x}^{-1}(0.5) = f(x_t)'\gamma$ (it happens under normality, but also for other distributions), we can consider the 0.5-quantile regression

$$y_t = f(x_t)'\gamma + \epsilon_t, \qquad t = 1, ..., T,$$

where $\{\epsilon_t\}$ are i.i.d. quantile innovations and $\epsilon|x$ has a conditional density function $g(\cdot|x)$ with zero median. Notice that quantiles are invariant to monotonous transformations, so that

$$F_{h(y)|x}^{-1}(0.5) = h\left(F_{y|x}^{-1}(0.5)\right) = h\left(f(x)'\gamma\right).$$

Another advantage of quantile regression is that the quantiles are identifiable under censure. For example, in CA using a positive ratio scale of preferences we would censure all products with disutility (negative ratings), as we only observe $y^c = \max\{0, y\}$. The quantile regression here is $F_{y^c|x}^{-1}(0.5) = \max\left\{0, F_{y|x}^{-1}(0.5)\right\} = \max\left\{0, h\left(f(x)'\gamma\right)\right\}$. By contrast, the conditional mean of censured variables is only identifiable with additional distributional assumptions (e.g., a Tobit model). How can we estimate the conditional median? The classic econometric solution is the Least Absolute Deviation (LAD) estimator $\hat{\gamma}$ minimizing

$$\sum_{t=1}^{T} \left|y_t - f(x_t)'\gamma\right|.$$

This procedure is older than OLS and under regularity conditions it is a consistent estimator of γ. If the conditional distribution of $y|x$ is symmetric, i.e. $g(\cdot|x)$ is symmetric in 0 for all x, then the 0.5 quantile regression model is equivalent of the standard linear regression, and OLS and LAD are two alternative estimators for the same parameters. OLS is more efficient, but less robust to outliers in y_t. However, if the conditional distribution of y is asymmetric (for example due to the choice of preference measurement scale in the questionnaire), the differences can be substantial. Then a conjoint modeler should use LAD instead of OLS. LAD estimators are traditionally computed solving the linear programing problem,

$$\begin{aligned} \min_{\{\gamma,u_1,\ldots,u_T\}} \quad & \sum_{t=1}^{T} u_t \\ s.t. \quad & u_t \leq y_t - f(x_t)'\gamma,\ t = 1,\ldots,T \\ & u_t \leq -\left(y_t - f(x_t)'\gamma\right),\ t = 1,\ldots,T \end{aligned}$$

which is easy to solve even with popular computational spreadsheets such as Microsoft Office Excel. The constraints force $u_t = |\widehat{\varepsilon}_t|$ in the optimum.

Our experimental design method can be adapted to conditional quantile estimators. The asymptotic normality of LAD estimators and quantile regressions has been studied by Koenker and Bassett (1978) and Pollard (1991). In particular, the asymptotic covariance matrix of LAD estimators can be consistently estimated by

$$V_T = \frac{1}{4}\left(\sum_{t=1}^{T} g(0|x_t) f(x_t) f(x_t)'\right)^{-1} \left(X'X\right) \left(\sum_{t=1}^{T} g(0|x_t) f(x_t) f(x_t)'\right)^{-1}.$$

We will consider the optimal design minimizing the trace of this matrix. In the CA setup the conditional density $g(0|x) = g(0)$ is independent of x and we obtain that

$$V_T = \frac{1}{4g(0)^2}\left(\sum_{t=1}^{T} f(x_t) f(x_t)'\right)^{-1} = \frac{1}{4g(0)^2} Q_T^{-1}.$$

Therefore, in order to minimize $\phi(V_T)$ we can use the same optimal designs minimizing $\phi\left(Q_T^{-1}\right)$, which we have proposed for Least-Squares estimators. In other words, the optimal designs and algorithms developed in this chapter can be directly implemented for LAD estimators.

2.7 Experimental Design for Choice-Based CA

The choice-based CA model is based on McFadden's (1974) work. Assume J alternatives in a choice set and each alternative is characterized by the attributes $\{x_j\}_{j=1}^{J}$. If the latent utility of the alternative j is $u_j = x_j'\beta + \varepsilon_j$, and ε_j has a type I extreme value distribution, then the probability that a consumer selects the alternative j from the set $\{x_1, ..., x_J\}$ is

$$\pi_j(x,\beta) = \frac{\exp\left(x_j'\beta\right)}{\exp\left(\sum_{l=1}^{J} x_l'\beta\right)}.$$

With $t = 1,2,..,T$ sets we codify choices $y_t = (y_{t1}, ..., y_{tJ})'$ into a vector of dummies so that y_{tj} is equal to 1 if alternative j is selected and zero otherwise. Then the model can be estimated by Maximum Likelihood, maximizing

$$L(X,\beta) = \sum_{t=1}^{T}\sum_{j=1}^{J} y_{tj} \ln \pi_j(x_t,\beta) = \sum_{t=1}^{T}\sum_{j=1}^{J} y_{tj}\left(x_{tj}'\beta - \ln\left(\sum_{l=1}^{J}\exp\left(x_{tl}'\beta\right)\right)\right).$$

The gradient and the information matrix are given by

$$\begin{aligned}
\frac{\partial L(X,\beta)}{\partial\beta} &= \sum_{t=1}^{T}\sum_{j=1}^{J}\left(y_{tj} - \pi_j(x_t,\beta)\right)\ x_{tj}, \\
I(X,\beta) &= E\left[\frac{\partial L(X,\beta)}{\partial\beta}\frac{\partial L(X,\beta)'}{\partial\beta}\right] = \sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{l=1}^{J} E\left[\left(y_{tj} - \pi_j(x_t,\beta)\right)\left(y_{tl} - \pi_l(x_t,\beta)\right)\right]\ x_{tj}x_{tl}' \\
&= \sum_{t=1}^{T}\sum_{j=1}^{J}\pi_j(x_t,\beta)\left(1-\pi_j(x_t,\beta)\right)\ x_{tj}x_{tj}' - \sum_{j\neq l}^{J}\pi_j(x_t,\beta)\,\pi_l(x_t,\beta)\ x_{tj}x_{tl}'
\end{aligned}$$

where we have used $cov\left(y_{tj}, y_{tl}\right) = \pi_j(x_t,\beta)\left(1-\pi_j(x_t,\beta)\right)$ for $l = j$, and $-\pi_j(x_t,\beta)\,\pi_l(x_t,\beta)$ for $l \neq j$.

The asymptotic covariance matrix of the maximum likelihood estimator can be estimated by the inverse of the Hessian, and it is upper-bounded since

$$I(X,\beta) \leq \sum_{t=1}^{T}\sum_{j=1}^{J}\pi_j(x_t,\beta)\ x_{tj}x_{tl}' \leq X'X,$$

using $\sum_{j=1}^{J}\sum_{l=1}^{J}\pi_j\left(x_t,\beta\right)\pi_l\left(x_t,\beta\right)x_{tj}x_{tl}'$ is non negative definite. Therefore,

$$\phi\left(Var\left(\hat{\beta}\right)\right)=\phi\left(I\left(X,\beta\right)^{-1}\right)\geq\phi\left(\left(X'X\right)^{-1}\right).$$

A commonly used procedure consists of minimizing the lower bound for the covariance matrix $\phi\left(\left(X'X\right)^{-1}\right)$. In particular Kanninen (2002) maximizes $\left|X'X\right|$, finding that the optimal design places the attributes at the extreme points of the domain χ. This is also the approach considered by Kuhfeld et al. (1994), based on the Fedorov algorithm. The presented algorithm can be used in this context. However, this is not a reliable solution, as $\phi\left(Var\left(\hat{\beta}\right)\right)$ could be much higher than $\phi\left(\left(X'X\right)^{-1}\right)$.

Another commonly used approach is to replace $\pi_j\left(x_t,\beta\right)$ by a known function, setting a value β_0 arbitrarily or based on prior information (Huber and Zwerina 1996). Then we can apply the presented algorithm to minimize $\phi\left(\left(\frac{\partial\ln L(X,\beta_0)}{\partial\beta\partial\beta'}\right)^{-1}\right)$ (updating the derivatives in the Newton algorithm). However, this procedure does not guarantee a small covariance matrix when the true parameter is distant from the considered value β_0. So far, neither the statistical nor the marketing literature has produced a robust solution to this problem.

2.8 Designing CA Experiments with Discretized Preference Scales

Standard CA models assume that the respondent's preferences over products are continuous, and given by the latent model $u_t=x_t'\beta^0+\varepsilon$, where ε_t are independent shocks with zero mean and cumulative distribution $F\left(\cdot/\sigma\right)$. However, in practice marketing researchers typically use discrete measurement scales, such as Likert scales, rankings, etc. Therefore, what we actually observe is not the continuous varying u_t, but a censored version of the true underlying preferences. Ordered regression models, introduced by McKelvey and Zavoina (1975) and popularized by McCullagh (1980), can be used to capture the influence of the nonlinear censuring transformation imposed by ordered discrete measurement scales.

If T alternatives are evaluated on a discrete scale with multiple ordered response categories $\{c_k\}_{k=1}^{m}$, we can study the relationship between these discrete measures and the continuous underlying model, using the ordered regression method. Assuming that the respondent allocates rating $y_t = c_k$ when the latent utility u_t falls in the scale interval $(c_{k-1}, c_k]$, where $c_0 = -\infty$ and $c_{m+1} = +\infty$. Then the log likelihood of the CA model is given by

$$L(X,\beta,\sigma) = \sum_{t=1}^{T} \sum_{k=0}^{m+1} y_{tk} \ln\left(F\left(\frac{c_k - x_t'\beta}{\sigma}\right) - F\left(\frac{c_{k-1} - x_t'\beta}{\sigma}\right)\right),$$

where we set $F\left((c_0 - x_t'\beta)/\sigma\right) = 0$ and $F\left((c_{m+1} - x_t'\beta)/\sigma\right) = 1$ for all t. Normality is often assumed, relying on the aggregation of innumerable small influences and the effect of the central limit theorem, but other distributions can be considered (such as the Logistic distribution) as well. Whenever $F(\cdot/\sigma)$ is continuously differentiable,

$$\frac{\partial L(X,\beta,\sigma)}{\partial \beta} = \sum_{t=1}^{T} \left(\sum_{k=0}^{m+1} y_{tk} \frac{\left(f\left(\frac{c_k - x_t'\beta}{\sigma}\right) - f\left(\frac{c_{k-1} - x_t'\beta}{\sigma}\right)\right)}{F\left(\frac{c_k - x_t'\beta}{\sigma}\right) - F\left(\frac{c_{k-1} - x_t'\beta}{\sigma}\right)} \right) x_t.$$

Notice that $E\left[y_{tk} y_{tj}\right] = E\left[y_{tk}^2\right] \times I(k=j) = E[y_{tk}] \times I(k=j)$, so that

$$I(X,\beta,\sigma) = E\left[\frac{\partial L(X,\beta,\sigma)}{\partial \beta} \frac{\partial L(X,\beta,\sigma)'}{\partial \beta}\right] = \sum_{t=1}^{T} \sum_{k=0}^{m+1} \frac{\left(f\left(\frac{c_k - x_t'\beta}{\sigma}\right) - f\left(\frac{c_{k-1} - x_t'\beta}{\sigma}\right)\right)^2}{F\left(\frac{c_k - x_t'\beta}{\sigma}\right) - F\left(\frac{c_{k-1} - x_t'\beta}{\sigma}\right)} x_{tj} x_{tl}'.$$

Setting $\beta = 0$ the covariance function can be expressed as $\left(X'BX\right)^{-1}$ for weights B that depend on the distribution $F(\cdot/\sigma)$ and the discrete measurement scale thresholds, therefore the algorithm, that we have presented, can be applied to compute optimal designs. But, similarly as in the case of choice-based CA the efficiency of the design will be good only if the true parameter is near the considered value.

2.9 Conclusions

Current methods to compute optimal experimental designs are typically inappropriate in CA, because they reduce treatments to a few, often repeated product profiles. Moreover, the most informative treatments (extreme vertices) are often dangerous to use or expensive. In these cases, the optimal design is not implemented, but it should be computed to be used as a reference to measure the efficiency of the implemented designs.

This essay proposes a general approach to compute the optimal matrix X^* with Newton type methods, avoiding repeated product profiles. Implementation results confirm the suitability of our approach to CA. We discuss cases with continuous and categorical product attributes, models for a single respondent and a panel of respondents, rating and choice preference measurement as well as studies with invariance to monotonous transformations.

The proposed procedure has the following advantages: (1) it is flexible to construct discrete-continuous designs; (2) it is easily implemented to the case of partial profiles in high dimensions; (3) the approach can handle easily other alternative linear regression estimators such as Stein's Shrinkage or ridge regression. Below we briefly review each of these issues.

2.9.1 The Discretization of Continuous Attributes

In many CA models, we often find continuous attributes reformulated as discrete ones. For example, prices are sometimes formulated as a continuous variable, but often just a few price levels are included in the model. What is the rationale for this type of specifications? In this section we introduce some remarks about this approach.

Discrete attributes are often introduced by the applied researchers as a way to approximate a nonlinear function effect of a continuous attribute. Assume that

$$y_t = \beta_0 + \beta_1 \ f_1(z_{1t}) + ... + \ \beta_k \ f_k(z_{Kt}) + \epsilon_t,$$

where z_t are continuous attributes and some of the functions in $f(\cdot)$ are unknown. If $f_i(z_i)$ is unknown we can absorb the coefficients β_i in this functions, and apply a semiparametric procedure for additively separable models.

One of the most elementary procedures is as follows. First, build a partition $\{A_{ij}\}_{j=1}^{k_i}$ of the range of variation of z_i, then we can approximate $f_i(z_i)$ by a simple function,

$$f_i(z_i) \simeq \sum_{j=1}^{k_i} \alpha_{ij}\, d_{ij},$$

where $d_{ij} = I\left(z_i \in A_{ij}\right)$ is a dummy variable, and I is the indicator function (equal to 1 if the event occurs, and 0 otherwise). The CA model can be written as

$$y_t = \beta_0 + \sum_{i=1}^{K} \sum_{j=1}^{k_i} \beta_{ij} d_{ijt} + \varepsilon_t$$

where $\beta_{ij} = \beta_i \alpha_{ij}$. If we have a single attribute, and we omit the intercept to avoid multicollinearity, the OLS estimator $\widehat{\beta}_{ij}$ is just the mean of all the data y_t for which $z_{it} \in A_j$. This way to specify and estimate a nonlinear regression model is known as a *regressogram*. It is the regression equivalent to the histogram for a density function - the most basic nonparametric regression estimator. To ensure consistency the partition must be thinner when the sample size increases at an appropriate rate (all k_i must growth with T, but slowly). The general case with several attributes is a standard semiparametric model for linear in components specifications. Notice that we might apply the same logic to an unknown general utility function $f(z_{1t}, ..., z_{Kt})$, then the nonparametric approximation would be given by the full factorial model, it can be see as a nonparametric estimator subject to the *curse of dimensionality* (the required sample size growths exponentially with K). This problem is not found in the semiparametric additively separable model.

Continuous attribute discretization are commonly used in CA but from a more primitive perspective: as a parametric model specification, which is sometimes problematic. Note that, if the number of levels k_i is too low we have an over-smoothing, and if it is too high - an over-fitting problem. The impact of the number of levels over CA estimations was pointed out by Currim et al. (1981) and Wittink et al. (1982). This is a well known problem in both, nonparametrics and semiparametrics, and it can be avoided using their fundamental principles.

From an econometric perspective, there are other approaches that can render better results than partitions. We recommend the use of semiparametric analysis (with the advantage that, for these models, there are much better tools for selecting the optimal level of smoothing than in the regressogram partitions), but even from a classic model perspective it is often better than discretization. For example, consider a simple case with regressors in $[0,1]$ and a Chebyshev Polynomials Basis $\{\phi_j(z) = \cos(j \arccos z)\}$, if we specify $f_i(z_i) = \sum_{j=1}^{k_i} \alpha_{ij}\ \phi_j(z_i)$ (in the semiparametric approach we take k_i as an increasing function of T). Then we can express

$$y_t = \beta_0 + \sum_{j=1}^{k_1} \beta_{1j}\ \phi_j(z_i) + \ldots + \sum_{j=1}^{k_K} \beta_{Kj}\ \phi_j(z_{Kt}) + \epsilon_t,$$

and the optimal experimental design for the OLS estimators of this model can be computed with the methodology presented in this essay.

2.9.2 Partial Profiles in High Dimensions

Standard CA models assume that all the stimuli attributes affecting utility ratings are included in the model. But the product complexity has increased over time, and consumer preference models often have to analyze categories described by a massive number of attributes and levels. It is unfeasible to study all of them. Some researchers use partial profiles, where each profile contains an experimentally designed subset of the attributes, as discussed by Bradlow (2005). Sometimes adaptive questionnaires are used to select a few important attributes.

However, the omission of other significant regressors generates biased estimations. Let us assume that the actual CA model is

$$y_t = f(x_t)'\beta + U_t + \varepsilon_t,$$

where $U_t = \gamma' f(z_t)$ is the utility associated to the omitted attributes z_t. If we omit the attributes $f(Z)$, estimating the model $y_t = f(x_t)'\beta + \varepsilon_t$ by OLS, some issues must be taken into account. First, notice that question repetitions in X can be accepted provided that omitted attributes are changing (the product profiles are actually different). The second and more crucial issue, is that OLS estimator $\widehat{\beta} = \left(X'X\right)^{-1} X'Y$ of the model with omitted variables is in general biased, with

$$E\left[\widehat{\beta}\right] = \beta + Q_T^{-1} X'U.$$

with $U = (U_1, ..., U_T)'$.

The presented algorithm we can used to generate optimal designs such that $\widehat{\beta}$ is unbiased. We just need to generate joint profiles $\{(x_t, z_t)\}$, including the additional constraint that $X'f(Z) = 0$. This constraints ensures that $X'U$ is zero, avoiding the bias problem, and the covariance matrix of $\widehat{\beta}$ will be determined by $\left(X'X\right)^{-1}$ under the standard assumptions.

2.9.3 Alternative Linear Regression Estimators

The Gauss-Markov Theorem ensures that OLS are the best linear unbiased estimators [BLUE], conditionally on the design matrix X. However, not all the design matrices render equally efficient estimators. We have focused on optimal experimental designs for OLS estimators, but the same method can be adapted to other increasingly popular estimators, such as Bayesian estimators for Gaussian Linear Regression. The classic model assumes that $Y \sim N\left(X'\beta, \sigma^2 I\right)$ with conjugate prior $\beta|\sigma^2 \sim N\left(\mu, \Sigma\right)$ and $1/\sigma^2$ distributed as a Gamma. In this case, β has a posterior distribution normal with $E\left(\beta|Y, \sigma^2\right) = \left(\Sigma^{-1} + X'X\right)^{-1}\left(\Sigma^{-1}\mu + X'Y\right)$ and covariance matrix $Var\left(\beta|Y, \sigma^2\right) = \sigma^2\left(\Sigma^{-1} + X'X\right)^{-1}$. The trace (or determinant) of $\left(\Sigma^{-1} + X'X\right)^{-1}$ can be minimized similarly to the trace (determinant) of $\left(X'X\right)^{-1}$ in OLS, subject to the required constraints preventing profile repetitions. In any case, the choice between OLS and the Classic Bayesian method is irrelevant with large T, as the distance between $E\left(\beta|Y, \sigma^2\right)$ and the OLS estimator converges faster than $\sqrt{T}$. Even if the researcher considers another prior distribution and non gaussian likelihood for $\varepsilon = Y - X'\beta$ (computing numerically the posterior), the Bernstein-von Mises Theorem ensures that the Bayes distribution a posteriori behaves asymptotically like the Maximum Likelihood estimator, under appropriate regularity conditions. With a normal likelihood function this estimator is precisely OLS. Therefore, the choice between Bayes or OLS matters essentially for relatively small T, which is precisely where the prior assumption has more impact.

The method can be also adapted to handle Stein's Shrinkage estimators that can have a smaller Mean Squared Error than OLS, reducing the variance in exchange for a small bias sacrifice. For example, a Ridge regression estimator $\widehat{\beta} = \left(X'X + \gamma I\right)^{-1} X'Y$ minimizes $\left\|Y - X'\beta\right\|_2^2 + \gamma\left\|\beta\right\|_2^2$, where γ is set as a minimizer of the MSE trace or determinant conditionally on the data.

Essentially the method penalizes complex models, and has a Bayesian interpretation. The algorithms considered in this paper can be readily adapted to these estimators, minimizing the trace or determinant of the appropriate covariance matrix.

Bibliography

Addelman, S. (1962). Symmetrical and Asymmetrical Fractional Factorial Plans. *Technometrics*, 4(1):47–58.

Box, M. J. (1970). Some experiences with a nonlinear experimental design criterion. *Technometrics*, 12(3):569–589.

Bradlow, E. T. (2005). Current Issues and a “Wish List” for Conjoint Analysis. *Applied Stochastic Models in Business and Industry*, 21(4-5):319–323.

Byrd, R. H., Hribar, M. E., and Nocedal, J. (1999). An Interior Point Algorithm for Large-Scale Nonlinear Programming. *SIAM Journal on Optimization*, 9(4):877–900.

Cattin, P. and Wittink, D. R. (1982). Commercial Use of Conjoint Analysis: A Survey. *Journal of Marketing*, 46(3):44–53.

Cochran, W. G. and Cox, G. M. (1957). *Experimental Designs*. John Wiley & Sons, New York, 2nd edition.

Cook, R. D. and Nachtsheim, C. J. (1980). A Comparison of Algorithms for Constructing Exact D-Optimal Designs. *Technometrics*, 22(3):315–324.

Cox, D. R. (1958). *Planning of Experiments*. New York: John Wiley & Sons.

Currim, I. S., Weinberg, C. B., and Wittink, D. R. (1981). Design of Subscription Programs for a Performing Arts Series. *Journal of Consumer Research*, 8(1):67–75.

Dykstra, Jr., O. (1971). The Augmentation of Experimental Data to Maximize $|X'X|$. *Technometrics*, 13(3):682–688.

Fedorov, V. V. (1972). *Theory of Optimal Experiments*. Academic Press, New York.

Green, P. E. (1974). On the Design of Choice Experiments Involving Multifactor Alternatives. *Journal of Consumer Research*, 1(2):61–68.

Green, P. E. and Rao, V. R. (1971). Conjoint Measurement for Quantifying Judgmental Data. *Journal of Marketing Research*, 8(3):355–363.

Gustafsson, A., Herrmann, A., and Huber, F. (2007). *Conjoint Measurement: Methods and Applications*. Berlin: Springer Verlag.

Huber, J. and Zwerina, K. (1996). The importance of utility balance in efficient choice design. *Journal of Marketing Research*, 33:307–317.

Johnson, M. E. and Nachtsheim, C. J. (1983). Some Guidelines for Constructing Exact D-Optimal Designs on Convex Design Spaces. *Technometrics*, 25(3):271–277.

Kanninen, B. J. (2002). Optimal design for multinomial choice experiments. *Journal of Marketing Research*, 39:214–227.

Kiefer, J. (1959). Optimum Experimental Designs. *Journal of the Royal Statistical Society. Series B (Methodological)*, 21(2):272–319.

Koenker, R. and Bassett, Gilbert, J. (1978). Regression quantiles. *Econometrica*, 46(1):pp. 33–50.

Kuhfeld, W. F. (2010). Experimental Design: Efficiency, Coding, and Choice Designs. Technical Report MR-2010C, SAS.

Kuhfeld, W. F., Tobias, R. D., and Garratt, M. (1994). Efficient Experimental Design with Marketing Research Applications. *Journal of Marketing Research*, 31(4):545–557.

Lawler, E. L. and Wood, D. E. (1966). Branch-and-bound methods: A survey. *Operations Research*, 14(4):699–719.

Leyffer, S. (2001). Integrating SQP and Branch-and-Bound for Mixed Integer Nonlinear Programming. *Computational Optimization and Applications*, 18:295–309.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2):pp. 109–142.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In Zarembka, P., editor, *Frontiers in Econometrics*. Academic Press.

McKelvey, R. D. and Zavoina, W. (1975). A Statistical Model for the Analysis of Ordered Level Dependent Variables. *The Journal of Mathematical Sociology*, 4(1):103–120.

Meyer, R. K. and Nachtsheim, C. J. (1995). The coordinate-exchange algorithm for constructing exact optimal experimental designs. *Technometrics*, 37(1):60–69.

Mitchell, T. J. (1974). An algorithm for the construction of "D-optimal" experimental designs. *Technometrics*, 16(2):203–210.

Mitchell, T. J. and Miller Jr, F. (1970). Use of design repair to construct designs for special linear models. *Math. Div. Ann. Progr. Rept.(ORNL-4661)*, pages 130–131.

Netzer, O., Toubia, O., Bradlow, E. T., Dahan, E., Evgeniou, T., Feinberg, F. M., Feit, E. M., Hui, S. K., Johnson, J., Liechty, J. C., Orlin, J. B., and Rao, V. R. (2008). Beyond Conjoint Analysis: Advances in Preference Measurement. *Marketing Letters*, 19(3):337–354.

Nocedal, J. and Wright, S. (2006). *Numerical Optimization*. Springer, New York.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7(2):186–199.

Stevens, S. S. (1951). *Handbook of Experimental Psychology*. John Wiley & Sons Inc.

Wittink, D. R. and Cattin, P. (1989). Commercial Use of Conjoint Analysis: An Update. *Journal of Marketing*, 53(3):91–96.

Wittink, D. R., Krishnamurthi, L., and Nutter, J. B. (1982). Comparing Derived Importance Weights Across Attributes. *Journal of Consumer Research*, 8(4):471–474.

Wynn, H. P. (1972). Results in the theory and construction of D-optimum experimental designs. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):133–147.

Appendix A: Approximate Optimal Designs

In this section we review the tools available for the design of approximate optimal experiments, and the drawbacks for their application to CA experiments.

What can we say about the matrix Q? We first consider the case with a finite number of explanatory variables (or treatments), $\chi = \{x_1,...,x_r\}$, meaning that the attributes are described by dummy variables. With $T > r$ some of the treatments are replicated, and let T_j denote the number of replications of treatment x_j. Then we can write $Q_T = \sum_{j=1}^{r} \left(T_j/T\right)\ f\left(x_j\right) f\left(x_j\right)'$, with $T = \sum_{j=1}^{r} T_j$, and the limit matrix must be of the form

$$Q = Q_\omega = \sum_{j=1}^{r} \omega_j\ f\left(x_j\right) f\left(x_j\right)',$$

where $\{\omega_j\}$ are limit relative frequencies that sum up one. Notice that the optimal ω_j are the continuous approach to treatments' relative frequencies T_j/T. An exact design for a given sample size T puts emphasis on setting T_j, and an approximate design on setting ω_j in the continuous limit (either generating T random profiles based on these probabilities so that and $Q = E_\omega\left[f(x) f(x)'\right]$, or setting an exact integer number T_j such that T_j/T is close to ω_j).

The theory of approximate designs was developed by Jack Carl Kiefer and his school (Kiefer 1959). They proposed to select optimally ω_j, minimizing some convex function measuring the size of Q_ω^{-1}. The most common procedures minimize:

1. generalized variance: $\left|Q^{-1}\right| = \prod_{r=1}^{k} \lambda_r\left(Q^{-1}\right) = \prod_{r=1}^{k} 1/\lambda_r(Q)$ where $\lambda_r(Q)$, $r = 1,...,k$ the eigenvalues of Q. Equivalently, the logarithm can be considered. D-optimality criterion minimizes the volume of the confidence ellipsoid of the model parameters. It is probably the most popular method;

2. average variance: $tr\left(Q^{-1}\right) = \sum_{r=1}^{k} \lambda_r\left(Q^{-1}\right) = \sum_{r=1}^{k} 1/\lambda_r(Q)$. A-optimality criterion (average-variance optimality) minimizes the mean of the variances of the estimates; and

3. worst possible prediction error: $d\left(Q^{-1}\right) = \max_{x\in\chi}\left\{xQ^{-1}x\right\}$. This is sometimes denoted G-optimality. The celebrated Kiefer-Wolfowitz equivalence theorem proved that G-optimal

and D-optimal designs are exactly the same.

4. the largest eigenvalue: $\max_r \left\{\lambda_r\left(Q^{-1}\right)=\right\} = \max_r\{1/\lambda_r(Q)\}$, called E-optimality or eigenvalue optimality.

More generally, we can minimize any non-negative function $\phi\left(Q^{-1}\right)$, provided that it is (1) positively homogeneous: $\phi(\delta A) = \delta\phi(A)$ for $\delta > 0$ to ensure that the factor σ^2/T is common to all designs; (2) non-increasing: $\phi(A) \le \phi(B)$ when $(A - B)$ is non negative definite; and (3) convex (to ensure that ϕ satisfies the condition that information cannot be increased through interpolation). This approach was developed by Kiefer (1959) inspired by the suggestion of Wald (1943) to compare designs using D-optimality, see also Kiefer and Wolfowitz (1960). Sometimes we are just interested in a subset or a combination of insightful coefficients, say $C\beta$ with a non singular matrix C. Then the optimal design minimizes the size of the corresponding covariance $\phi\left(CQC'\right)$ (Hausman 1982; Toubia and Hauser 2007). In particular, the L-optimality criteria minimizes $tr\left(CQ^{-1}\right)$ for an appropriate matrix C.

Following the Kiefer approach, we can randomly generate the designs with optimal probabilities ω^*, by minimizing a convex function ϕ

$$\min_{\omega}\phi\left(Q_{\omega}^{-1}\right) = \min_{\omega}\phi\left(\left[\sum_{j=1}^{r}\omega_j\, f\left(x_j\right) f\left(x_j\right)'\right]^{-1}\right) \tag{2.8}$$

subject to the constraint that ω is in the $\mathbb{R}_+^r$ simplex. We can add other convex constraints, e.g. a bound on the expected experiment cost $T \cdot c'\omega \le m$ where m is the available budget, and c is a $r \times 1$ vector, whose elements are costs associated with each treatment in χ, so that the expected cost of a single profile is $c'\omega$.

Instead of generating random designs with distribution ω^*, we can consider appropriate integer numbers T_j of repetitions, such that the optimal ω_j^* is approximated by T_j/T (Pukelsheim and Rieder 1992). Approximate designs are convenient from a theoretical and computational perspective, but in practice the results must be rounded off leading to the loss of design efficiency. Alternatively, we can try to optimize $\phi(Q_T)$ in $\left\{T_j\right\}$ directly (exact designs). Notice that in randomized experiments where individuals are allocated to a unique treatment (e.g. testing medical drugs), the individuals can be allocated based on the optimal probabilities ω^*.

For continuous treatments, we can generate treatments randomly from a probability distribution w, and consider a limit information matrix

$$Q(w) = \int f(x) f(x)' w(dx) \in \mathbb{R}^{p \times p}$$

where w is a probability distribution on χ. We need to select the optimal probability function. In practice this problem becomes similar to the case with finite number of treatments, focusing on a few extreme cases. This makes sense, as the extreme conditions in experiments usually render more information for inference decisions. The following result provides a theoretical basis for this and some more general statements. Let us denote by $vech$ the lower half-vectorization mapping (i.e., $vech(Q)$ is the column vector obtained by vectorizing only the lower triangular part of a symmetric matrix Q).

Lemma 1 *If χ is a convex compact set and the continuous f preserves convexity, then any feasible Q can be expressed as $\sum_{x_j \in \chi^e} \omega_j f(x_j) f(x_j)'$ where ω_j are discrete probabilities and $\chi^e \subset \chi$ are the profiles associated with the set of extreme points of $\{vech(f(x) f(x)') : x \in \chi\}$.*

Proof. The set $\mathbf{Q} = \{Q = Q(w) : w \geq 0, \int dw = 1\}$ is isomorphic to a non empty convex and compact set in $\mathbb{R}^{p(p+1)/2}$ using the lower half-vectorization mapping. The classic Krein–Milman Theorem ensures that if $\mathbf{Q}$ is a compact convex set of $\mathbb{R}^{p(p+1)/2}$, then any $Q \in \mathbf{Q}$ can be expressed as $\sum_{x_j \in \chi^e} \omega_j f(x_j) f(x_j)'$ where $\sum_j \omega_j = 1$ with $\omega_j \geq 0$, and χ^e are the profiles associated with the set of extreme points of $\{vech(f(x) f(x)') : x \in \chi\}$. ■

The literature has often considered an informal (and wrong) proof of this Lemma based on the Carathéodory's Theorem[2], suggesting also that the number r of elements in χ^e satisfies that $r \leq 1 + p(p+1)/2$. But notice that the subset of extreme points representing any specific element in the convex hull may change with the considered element, and as a result we cannot ensure that any point can be represented as a convex linear combination from elements of a finite set

[2] The Carathéodory's theorem states that if $y \in \mathbb{R}^d$ lies in the convex hull of a set P, there is a subset $P' \subset P$ consisting of no more than $d+1$ points such that y lies in the convex hull of P'.

with less than $1 + p(p+1)/2$ points.

Therefore, the search for optimal designs may be restricted to designs with a finite support. If the set $\{vech\left(f(x)f(x)'\right) : x \in \chi\}$ is a convex polytope in $\mathbb{R}^{p(p+1)/2}$ the first step is to compute the vertices, the second consists of solving a problem similar to (2.8), considering a frequency of repetitions for each vertex. Obviously, mixed models with continuous and discrete variables can be handled alike. These results can be directly adapted to experiments with heteroskedasticity, where $E(\varepsilon\varepsilon') = diag\left(\sigma^2(x_t)\right)$, considering information matrices $Q(w) = \int \sigma^2(x) f(x) f(x)' w(dx)$, and $Q = \sum_{j=1}^{r} \omega_j \sigma^2\left(x_j\right) f\left(x_j\right) f\left(x_j\right)'$. But in practice this cannot be applied unless we know $\sigma^2(x)$. To that end we can build a preliminary experiment to estimate this function but this is rarely considered.

In order to compute the approximate optimal designs solving (2.8), Kiefer's school has considered several algorithms. One of the most popular is the classic algorithm proposed by Fedorov-Wynn for D-optimality (Fedorov 1972; Wynn 1970), for the review see St. John and Draper (1975) and the references in Atwood (1973, 1976). These methods are variants from the steepest descent method algorithm, and they can be adapted for other criteria $\phi(\cdot)$ (Whittle 1973). However, the steepest descent methods converge very slowly. Atwood (1976) considers faster Newton directions. But the performance of these methods is not always good, and the search for optimal designs often restricts to low-dimensional models. As López Fidalgo (2009) states: "One may think the people working on optimal design must be good in optimization. They are not bad, but they are not experts in the topic. At the same time, people in optimization are sometimes far from statistics and even more from experimental designs. Therefore, there is a need of more co-operation between them." Several contemporary *constrained optimization* numerical algorithms can be implemented for a faster computation of ω^*, including classic *sequential quadratic optimization* algorithms, or the more recent *interior point algorithms* (see e.g. Vandenberghe et al. 1998; Boyd and Vandenberghe 2004, Ch.7). Solving the dual problem is a good strategy that often renders faster results.

Unfortunately, this approach is not adequate for CA. If the researchers use Kiefer's approximate design for a single customer, ω_j (respectively T_j) can be interpreted as the probability (absolute frequency) of times stimulus j is repeated. This is an entirely undesirable situation: if implemented, the repeated questions should be interspersed and presented separately over time to ensure that the respondent forgets the previous answers. Even then, the procedure could easily be cost-prohibitive and tedious for the respondent, leading to biased estimations[3]. Therefore, approximate optimal designs should not be implemented in CA in general.

Appendix B: Matrix Derivatives

This section presents the main results about matrix derivatives. First we introduce some concepts about functions of matrices and their derivatives. Let Z denote a $n \times q$ real matrix. We can consider a $m \times p$ real matrix valued function $\Phi(Z)$ (notice that scalar valued and vector valued functions are a particular case). We define the Jacobian of $\Phi(Z)$ as the $mp \times nq$ matrix

$$D\Phi(Z) = \frac{\partial vec(\Phi(Z))}{\partial(vec(Z))'}.$$

Using this definition, the properties of classic gradients and Jacobians are preserved. The differential of $\Phi(Z)$ will be given by $d\ \Phi(Z) = \Phi(Z)\ d\ vec\ Z = \Phi(Z)\ vec(dZ)$. Hessians, can be defined analogously as follows,

$$H\Phi(Z) = D(D\Phi(Z))' = \frac{\partial}{\partial(vec(Z))'} vec \left(\frac{\partial vec(\Phi(Z))}{\partial(vec(Z))'} \right)'.$$

The classic case where Φ is vector or scalar valued, is a particular case under this notation. For a detailed introduction to matrix derivatives see Magnus and Neudecker (1999).

Main derivatives

Consider a $T \times k$ design matrix, $X = f(x)$, where $f(\cdot)$ is a twice differentiable function. Let

[3] We can apply directly the Kiefer method for a homogeneous consumer sample, were we ask just one question to each different respondent. Then the optimal frequencies ω^* can be used for randomization, allocating different respondents to an specific question.

$$\begin{aligned} A &= \frac{\partial vec(f(x))}{\partial (vec(x))'} \\ B &= \frac{\partial}{\partial (vec(x))'} vec \left(\frac{\partial vec(f(x))}{\partial (vec(x))'} \right)'. \end{aligned}$$

the Jacobian and Hessian of f, and let Z be a constant positive definite weight matrix. We assume that Z is symmetric to simplify the notation, otherwise the derivatives become more involved. For example for the classic experimental regression model, with vector utility preferences $f(x) = x$, $A = I$, $B = 0$, and $Z = I$. Finally, let's define a commutation matrix K, such that vec $X' = K$ vec X.

Proposition 2 *Consider the objective function*

$$\min \left| \left(X'ZX \right)^{-1} \right|.$$

The gradient and Hessian are respectively, in a vec form,

$$\begin{aligned} D\phi(X) &= -2 \left| \left(X'ZX \right)^{-1} \right| \; vec\ AZX \left(X'ZX \right)^{-1}, \\ H\phi(X) &= 4 \left| \left(X'ZX \right)^{-1} \right| K \left(AZX(X'ZX)^{-1} \otimes (X'ZX)^{-1} X'ZA' \right) + \\ & \quad 4 \left| \left(X'ZX \right)^{-1} \right| \left(\left(X'ZX \right)^{-1} \otimes AZX \left(X'ZX \right)^{-1} X'ZA' \right) - \\ & \quad 2 \left| \left(X'ZX \right)^{-1} \right| \left(\left(X'ZX \right)^{-1} \otimes AZA' \right) - \\ & \quad \left| \left(X'ZX \right)^{-1} \right| \left(ZX \left(X'ZX \right)^{-1} \otimes B + \left(X'ZX \right)^{-1} X'Z \otimes B' \right). \end{aligned}$$

Proof. Note that the general first order derivative of $|X|$ is $d|X| = |X| \mathrm{tr}\, X^{-1} dX$, and the general first order derivative of the inverse is $X^{-1} = -X^{-1}(dX)X^{-1}$. Now recall main properties of trace: is invariant under cyclic permutations, the traces of a matrix and its transpose are equal, and additivity. The differential of $\left| \left(X'ZX \right)^{-1} \right|$ is

$$\begin{aligned}
d\left|\left(X'ZX\right)^{-1}\right| &= \left|\left(X'ZX\right)^{-1}\right| \text{ tr } \left(X'ZX\right) d\left(X'ZX\right)^{-1} = \\
&= -\left|\left(X'ZX\right)^{-1}\right| \text{ tr } \left(X'ZX\right)^{-1} d\left(X'ZX\right) = \\
&= -2\left|\left(X'ZX\right)^{-1}\right| \text{ tr } \left(X'ZX\right)^{-1} X'ZdX = \\
&= -2\left|\left(X'ZX\right)^{-1}\right| \text{ tr } \left(X'ZX\right)^{-1} X'ZA'dx.
\end{aligned}$$

Then the first order derivative is

$$D\phi(X) = -2\left|\left(X'ZX\right)^{-1}\right| AZX\left(X'ZX\right)^{-1}.$$

According to the first identification table (Magnus and Neudecker 1999, p. 176), the gradient in vec form is $-2\left|\left(X'ZX\right)^{-1}\right|$ vec $AZX\left(X'ZX\right)^{-1}$.

Recall, one of the trace properties: (tr U)(tr V) = tr $U \otimes V$, where U and V are square matrices. Then consider the Hessian

$$\begin{aligned}
d^2\left|\left(X'X\right)^{-1}\right| = d\left[-2\left|\left(X'ZX\right)^{-1}\right| \text{ tr } \left(X'ZX\right)^{-1} X'ZA'dx\right] = \\
= -2\, d\left|\left(X'ZX\right)^{-1}\right| \cdot \text{ tr } \left(X'ZX\right)^{-1} X'ZA'dx \\
-2\left|\left(X'ZX\right)^{-1}\right| \text{ tr } d\left(X'ZX\right)^{-1} X'ZA'dx \\
-2\left|\left(X'ZX\right)^{-1}\right| \text{ tr } \left(X'ZX\right)^{-1} (dX)'ZA'dx \\
-2\left|\left(X'ZX\right)^{-1}\right| \text{ tr } \left(X'ZX\right)^{-1} X'Z(dA)'dx = \\
= 4\left|\left(X'ZX\right)^{-1}\right| \left[\text{ tr } \left(X'ZX\right)^{-1} X'ZA'dx\right] \left[\text{ tr } \left(X'ZX\right)^{-1} X'ZA'dx\right] \\
+2\left|\left(X'ZX\right)^{-1}\right| \text{tr } \left(X'ZX\right)^{-1} d\left(X'ZX\right)\left(X'ZX\right)^{-1} X'ZA'dx \\
-2\left|\left(X'ZX\right)^{-1}\right| \text{tr } \left(X'ZX\right)^{-1} (dx)'AZA'dx \\
-2\left|\left(X'ZX\right)^{-1}\right| \text{ tr } \left(X'ZX\right)^{-1} X'Zdx'Bdx = \\
= 4\left|\left(X'ZX\right)^{-1}\right| \text{ tr } \left(X'ZX\right)^{-1} X'ZA'dx\; 1 \otimes \left(X'ZX\right)^{-1} X'ZA'dx
\end{aligned}$$

$$\begin{aligned}
&+4\left|\left(X'ZX\right)^{-1}\right|\text{tr }\left(X'ZX\right)^{-1}(dx)'AZX\left(X'ZX\right)^{-1}X'ZA'dx\\
&-2\left|\left(X'ZX\right)^{-1}\right|\text{tr }\left(X'ZX\right)^{-1}(dx)'AZA'dx\\
&-2\left|\left(X'ZX\right)^{-1}\right|\ \text{tr }\left(X'ZX\right)^{-1}X'Zdx'Bdx.
\end{aligned}$$

Using the Kronecker property $a \otimes A = aA$, the Hessian is:

$$\begin{aligned}
H\phi(X) = \quad & 4\left|\left(X'ZX\right)^{-1}\right|K\left(AZX(X'ZX)^{-1} \otimes (X'ZX)^{-1}X'ZA'\right) + \\
& 4\left|\left(X'ZX\right)^{-1}\right|\left(\left(X'ZX\right)^{-1} \otimes AZX\left(X'ZX\right)^{-1}X'ZA'\right) - \\
& 2\left|\left(X'ZX\right)^{-1}\right|\left(\left(X'ZX\right)^{-1} \otimes AZA'\right) - \\
& \left|\left(X'ZX\right)^{-1}\right|\left(ZX\left(X'ZX\right)^{-1} \otimes B + \left(X'ZX\right)^{-1}X'Z \otimes B'\right).
\end{aligned}$$

■

Proposition 3 *Consider the following objective function*

$$\min tr\ \left(X'ZX\right)^{-1}.$$

The gradient and Hessian are respectively in vec form

$$\begin{aligned}
D\phi(X) &= -2\ vec\ AZX\left(X'ZX\right)^{-2}\\
H\phi(X) &= 4\left(\left(X'ZX\right)^{-1} \otimes AZX\left(X'ZX\right)^{-2}X'ZA'\right) + \\
&\quad 4\left(\left(X'ZX\right)^{-2} \otimes AZX\left(X'ZX\right)^{-1}X'ZA'\right) - \\
&\quad 2\left(\left(X'ZX\right)^{-2} \otimes AZA'\right) - \\
&\quad \left(ZX\left(X'ZX\right)^{-2} \otimes B + \left(X'ZX\right)^{-2}X'Z \otimes B'\right).
\end{aligned}$$

Proof. Using the main properties of the trace, the differential of tr $\left(X'ZX\right)^{-1}$ is

$$d\ \mathrm{tr}\ \left(X'ZX\right)^{-1} = -\ \mathrm{tr}\ \left(X'ZX\right)^{-2} d\left(X'ZX\right) = -2\ \mathrm{tr}\ \left(X'ZX\right)^{-2} X'ZA'dx.$$

Following the identification table the first order derivative is

$$D\phi(X) = -2AZX\left(X'ZX\right)^{-2}.$$

and the gradient is simply the vec form of $D\phi(X)$.

For the Hessian, consider the second-order differential

$$\begin{aligned}
d^2\ \mathrm{tr}\ \left(X'ZX\right)^{-1} = d&\left(-2\ \mathrm{tr}\ \left(X'ZX\right)^{-2} X'ZA'dx\right) = \\
&= -2\ \mathrm{tr}\ d\left(X'ZX\right)^{-2} X'ZA'dx \\
&-2\ \mathrm{tr}\ \left(X'ZX\right)^{-2} (dX)'ZA'dx \\
&-2\ \mathrm{tr}\ \left(X'ZX\right)^{-2} X'Z(dA)'dx = \\
&= 2\ \mathrm{tr}\ \left(X'ZX\right)^{-1} d\left(X'ZX\right)\left(X'ZX\right)^{-2} X'ZA'dx \\
&+2\ \mathrm{tr}\ \left(X'ZX\right)^{-2} d\left(X'ZX\right)\left(X'ZX\right)^{-1} X'ZA'dx \\
&-2\ \mathrm{tr}\ \left(X'ZX\right)^{-2} (dx)'AZA'dx \\
&-2\ \mathrm{tr}\ \left(X'ZX\right)^{-2} X'Zdx'Bdx = \\
&= 4\ \mathrm{tr}\ \left(X'ZX\right)^{-1} (dx)'AZX\left(X'ZX\right)^{-2} X'ZA'dx \\
&+4\ \mathrm{tr}\ \left(X'ZX\right)^{-2} (dx)'AZX\left(X'ZX\right)^{-1} X'ZA'dx \\
&-2\ \mathrm{tr}\ \left(X'ZX\right)^{-2} (dx)'AZA'dx \\
&-2\ \mathrm{tr}\ \left(X'ZX\right)^{-2} X'Zdx'Bdx.
\end{aligned}$$

Then according to the second identification table the Hessian is

$$
\begin{aligned}
H\phi(x) = \; & 4\left(\left(X'ZX\right)^{-1} \otimes AZX\left(X'ZX\right)^{-2} X'ZA'\right) + \\
& 4\left(\left(X'ZX\right)^{-2} \otimes AZX\left(X'ZX\right)^{-1} X'ZA'\right) - \\
& 2\left(\left(X'ZX\right)^{-2} \otimes AZA'\right) - \\
& \left(ZX\left(X'ZX\right)^{-2} \otimes B + \left(X'ZX\right)^{-2} X'Z \otimes B'\right).
\end{aligned}
$$

■

Distance constraints

Proposition 4 *Consider the following distance constraints applied pointwise*

$$(1_{T\times T} - I_T)\varepsilon - diag\left(XX'\right)\mathbf{1}' - \mathbf{1}diag\left(XX'\right)' + 2XX' \leq 0$$

The gradient of the constraints in a matrix form is

$$
\begin{aligned}
dC &= -2\left(I_{T^2} + K_{T^2}\right)\left[(\mathbf{1}_{T\times 1} \otimes I_T)A - (X \otimes I_T)\right] \\
\textit{where } A_{i.} &= \left(vec\ e_i e_i' X\right)'.
\end{aligned}
$$

Proof. The constraint can be written as

$$(1_{T\times T} - I_T)\varepsilon - F - F' + 2S \leq 0$$

where $S = XX'$, and $F = diag(S)\mathbf{1}'$. In the constraint F,F' and S depend on X.

First, let's calculate the derivative of F. Note the the special structure of F (identical columns):

$$F = \begin{bmatrix} s_{11} & \dots & s_{11} \\ s_{22} & \dots & s_{22} \\ \dots & \dots & \dots \\ s_{tt} & \dots & s_{tt} \end{bmatrix}_{T\times T} = \begin{bmatrix} e_1' XX' e_1 & \dots & e_1' XX' e_1 \\ e_2' XX' e_2 & \dots & e_2' XX' e_2 \\ \dots & \dots & \dots \\ e_t' XX' e_t & \dots & e_T' XX' e_T \end{bmatrix}_{T\times T} = \sum_{i=1}^{T}\sum_{j=1}^{T} e_i' XX' e_i E_{ij}$$

where e_i is a unit vector containing 1 in the i-th element, and zeros otherwise. E_{ij} is an elementary matrix, containing 1 in the (i,j)-th element and zeros otherwise.

According to the first identification table (Magnus and Neudecker, p. 176), taking derivatives of a $T\times T$ matrix function $F(X)$ with respect to a $T\times k$ matrix X requires vectorizing both matrices

$$\text{d vec } F = A \text{ d vec } X \quad \Rightarrow DF(X) = A_{T^2\times Tk}.$$

Every row of a differential matrix A contains partial derivatives of each element of the vectorized F, taken with respect to vectorized X. Conveniently, in our case all columns are identical, therefore

$$\text{vec } F = (\mathbf{1}_{T\times 1} \otimes I_T) F_{.1}$$

is a column vector obtained by stacking T times first column of F. Each element of $F_{.1}$ is a scalar function of X, such that $F_{i1} = e_i' XX' e_i$, and its derivative is

$$d\ \phi(X) = d\left(e_i' XX' e_i\right) = e_i'(dX)X'e_i + e_i'X(dX)'e_i = 2 \text{ tr } X'e_ie_i'dX$$
$$\Leftrightarrow D\phi(X) = 2\left(\text{vec } e_ie_i'X\right)'.$$

Using the result from first identification table

$$\phi(X):\ \text{d}\phi = \text{tr } A'\text{d}X = (\text{vec } A)' \text{ d vec } X \Rightarrow D\phi(X) = (\text{vec } A)',$$

we obtain following derivative of F:

$$DF = (\mathbf{1}_{T\times 1} \otimes I_T) F_{.1} = 2(\mathbf{1}_{T\times 1} \otimes I_T) A$$

$$\text{where} \quad A = \begin{bmatrix} (\text{vec } e_1 e_1' X)' \\ (\text{vec } e_2 e_2' X)' \\ \cdots \\ (\text{vec } e_T e_T' X)' \end{bmatrix}.$$

It is straightforward to obtain the derivative of the second element, F', using the properties of vec operator

$$\text{vec } F' = K_{T^2} \text{ vec } F,$$

where K is a square commutation matrix. Then

$$d \text{ vec } F' = K_{T^2} d \text{ vec } F \Rightarrow DF' = K_{T^2} DF.$$

The last element in the constraint is $S = XX'$. If $S(X) = XX'$, then

$$dS(X) = (dX)X' + X(dX)'$$

and

$$\begin{aligned} \text{d vec } S(X) &= (X \otimes I_T) \text{ d vec } X + (I_T \otimes X) \text{ d vec } X' \\ &= (X \otimes I_T) \text{ d vec } X + (I_T \otimes X) K_{Tk} \text{ d vec } X \\ &= (X \otimes I_T) \text{ d vec } X + K_{T^2} (X \otimes I_T) \text{ d vec } X \\ &= \left(I_{T^2} + K_{T^2}\right)(X \otimes I_T) \text{ d vec } X. \end{aligned}$$

Therefore

$$DS(X) = \left(I_{T^2} + K_{T^2}\right)(X \otimes I_T).$$

Finally, combining all three results, the derivative of constraint on the distance matrix is

$$\begin{aligned}
DC &= -DF - K_{T^2}DF + 2DS = \\
&= -\left(I_{T^2} + K_{T^2}\right)DF + 2DS \\
&= -2\left(I_{T^2} + K_{T^2}\right)[(\mathbf{1}_{T\times 1} \otimes I_T)A - (X \otimes I_T)].
\end{aligned}$$

■

Interactions

In Problem (2.6) we have introduced a mixed-integer conjoint model with interactions. First, recall that in an experiment with L integer and k continuous variables, the design matrix $x = [D_1, \ldots, D_L, Z]$ is partitioned in such a way that each integer variable is represented by a dummy block D_l, and every continuous variable is represented by a column in a matrix Z. Next, we make the necessary operations to eliminate multicollinearity from the dummies, and that we add an intercept. We denote this transformation by $f_1(x) = xA + B$, where A is a sparse matrix, which eliminates the multicollinearity from the model (method D1 or D2 from a discussion of the case of integer attributes) and adds a column of zeros as a first column, and B is a sparse matrix with ones in the first column, and zeros otherwise. Let $\tilde{D}_l$ represent a transformed dummy variable from which we eliminated multicollinearity.

Consider a model with interactions between two variables, r and s, where (1) both of them can be integer, (2) both of them can be continuous, and (3) one can be integer and the other continuous. The interaction block containing all possible interaction elements between r and s can be expressed in terms of the design matrix x

$$W(x)=\sum_{t=1}^{T} e_t\,(R_t(x)\otimes S_t(x)),$$

where e_t is a unit vector with 1 in the position t and zeros otherwise, R and S are partitions of the design matrix x, representing variables r,s, and R_t,S_t is their tth row. Depending on the type of interaction: (1) $R=\tilde{D}_r$ and $S=\tilde{D}_s$ are corresponding dummy blocks; (2) $R=Z_r$ and $S=Z_s$ are corresponding columns from Z; (3) $R=\tilde{D}_r$ is a dummy block, and $S=Z_s$ is a column from Z. Hence, every row in $W(x)$ is calculated as the Kronecker product of either (1) two row vectors, (2) two scalars, or a (3) a row vector and a scalar, respectively.

More specifically, $R_t(x)=e'_t f_1(x)C_R$, and $S_t(x)=e'_t f_1(x)C_S$, where C_R,C_S are sparse matrices of dimensions equal to R and S respectively. Each column of C_R,C_S is a unit vector with a 1 in the position corresponding to a column in the design matrix x, so that $R=f_1(x)C_R$ and $S=f_1(x)C_S$. Then $X=[1,\tilde{f}(x),W(x)]$ is

$$X=f_2(f_1(x))=f_1(x)D_1+WD_2=xAD_1+BD_1+WD_2$$

D_1,D_2 being constant sparse matrices which add block of zeros to the back and front of the matrix, and W is defined above. Applying the result of Proposition 3, we get

$$d\ \mathrm{tr}\ \left(X'X\right)^{-1}=-2\ \mathrm{tr}\ \left(X'X\right)^{-2}X'dX=-2\ \mathrm{tr}\ \left(X'X\right)^{-2}X'd(xAD_1+BD_1+WD_2)=$$
$$=-2\ \mathrm{tr}\,AD_1\left(X'X\right)^{-2}X'dx-2\ \mathrm{tr}\,D_2\left(X'X\right)^{-2}X'dW.$$

The first element does not require any further calculations, so let's concentrate on the second element:

$$\mathrm{tr}\,D_2\left(X'X\right)^{-2}X'dW=\ \mathrm{tr}\,D_2\left(X'X\right)^{-2}X'd\left(\sum_{t=1}^{T}\ e_t\left(e'_t f_1(x)C_R\otimes e'_t f_1(x)C_S\right)\right)$$

$$= \sum_{t=1}^{T} \text{ tr } D_2 \left(X'X\right)^{-2} X' e_t \ d \ \left(e_t' f_1(x) C_R \otimes e_t' f_1(x) C_S\right).$$

Note that $D_2\left(X'X\right)^{-2} X' e_t$ is a column vector and the elements in Kroneker product are row vectors (for continuous variables they are scalars), therefore the Kronecker expression is also a row vector. This simplifies the algebra needed to compute the gradient:

$$\begin{aligned}
&\sum_{t=1}^{T} e_t' X \left(X'X\right)^{-2} D_2' \ d \ \left(e_t' f_1(x) C_R \otimes e_t' f_1(x) C_S\right)' && (2.9)\\
&\quad = \sum_{t=1}^{T} e_t' X \left(X'X\right)^{-2} D_2' \ d \ \left(C_R' A' x' e_t \otimes C_S' A' x' e_t\right) && (2.10)\\
&\quad = \sum_{t=1}^{T} e_t' X \left(X'X\right)^{-2} D_2' \ d \ \text{ vec } C_S' A' x' e_t e_t' x A C_R && (2.11)\\
&\quad = \sum_{t=1}^{T} e_t' X \left(X'X\right)^{-2} D_2' \text{ vec } \left(C_S' A' (dx)' e_t e_t' x A C_R + C_S' A' x' e_t e_t' (dx) A C_R\right) && (2.12)\\
&\quad = \sum_{t=1}^{T} e_t' X \left(X'X\right)^{-2} D_2' \left[\left(C_R' A' x' e_t e_t' \otimes C_S' A'\right) d \text{ vec } x' + \left(C_R' A' \otimes C_S' A' x' e_t e_t'\right) d \text{ vec } x\right] && (2.13)\\
&\quad = \sum_{t=1}^{T} e_t' X \left(X'X\right)^{-2} D_2' \left[\left(C_R' A' x' e_t e_t' \otimes C_S' A'\right) K_{T\times nucol} + \left(C_R' A' \otimes C_S' A' x' e_t e_t'\right)\right] d \text{ vec } x && (2.14)
\end{aligned}$$

In (2.9) we use the trace property for column vectors a, b that $\text{tr } ab' = a'b$. In (2.10) we apply $f(x) = xA + B$, and Kronecker property $(A \otimes B)' = (A' \otimes B')$. In (2.11), we use the Kronecker property for column vectors: $vec\ ab' = b \otimes a$. In (2.12) we take the derivative of a product $d(x'Ax) = (dx)'Ax + x'A(dx)$. In (2.13) apply vec $ABC = (C' \otimes A)$ vec B. Finally in (2.14) we use commutation matrix to get vec $x' = K$ vec x. Then the gradient for the problem which includes interactions is

$$\begin{aligned}
D\phi(x) = &-2(\text{vec } X \left(X'X\right)^{-2} D_1' A')' \\
&- 2 \sum_{t=1}^{T} e_t' X \left(X'X\right)^{-2} D_2' \left[\left(C_R' A' x' e_t e_t' \otimes C_S' A'\right) K_{T\times nucol} + \left(C_R' A' \otimes C_S' A' x' e_t e_t'\right)\right].
\end{aligned}$$

The Hessian of the problem has been computed numerically.

Bibliography

Atwood, C. L. (1973). Sequences Converging to D-Optimal Designs of Experiments. *The Annals of Statistics*, 1(2):342–352.

Atwood, C. L. (1976). Convergent Design Sequences, for Sufficiently Regular Optimality Criteria. *The Annals of Statistics*, 4(6):1124–1138.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Fedorov, V. V. (1972). *Theory of Optimal Experiments*. Academic Press, New York.

Hausman, J. A. (1982). The Effects of Time in Economic Experiments. In Hildenbrand, W., editor, *Advances in Econometrics*. Cambridge University Press.

Kiefer, J. (1959). Optimum Experimental Designs. *Journal of the Royal Statistical Society. Series B (Methodological)*, 21(2):272–319.

Kiefer, J. and Wolfowitz, J. (1960). The Equivalence of Two Extremum Problems. *Canadian Journal of Mathematics*, 12:363–366.

López Fidalgo, J. (2009). A Critical Overview on Optimal Experimental Designs. *Boletín de Estadística e Investigación Operativa*, 25(1):14–21.

Magnus, J. R. and Neudecker, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley.

Pukelsheim, F. and Rieder, S. (1992). Efficient Rounding of Approximate Designs. *Biometrika*, 79(4):763–770.

St. John, R. C. and Draper, N. R. (1975). D-Optimality for Regression Designs: A Review. *Technometrics*, 17(1):15–23.

Toubia, O. and Hauser, J. R. (2007). Research Note – On Managerially Efficient Experimental Designs. *Marketing Science*, 26(6):851–858.

Vandenberghe, L., Boyd, S., and Wu, S.-P. (1998). Determinant Maximization with Linear Matrix Inequality Constraints. *SIAM Journal on Matrix Analysis and Applications*, 19(2):499–533.

Wald, A. (1943). On the Efficient Design of Statistical Investigations. *The Annals of Mathematical Statistics*, 14(2):134–140.

Whittle, P. (1973). Some General Points in the Theory of Optimal Experimental Design. *Journal of the Royal Statistical Society. Series B (Methodological)*, 35(1):123–130.

Wynn, H. P. (1970). The Sequential Generation of D-Optimum Experimental Designs. *The Annals of Mathematical Statistics*, 41(5):1655–1664.

Chapter 3

Optimal Experimental Designs for Nonlinear Conjoint Analysis

3.1 Introduction

It has been 40 years since Green and Rao (1971) published their seminal paper on Conjoint Analysis (CA), but it still remains an active research area with enormous impact on practitioners (see Wittink and Cattin 1989; Wittink et al. 1994; Green et al. 2001; Gustafsson et al. 2007). Nowadays the expression CA encompasses a variety of techniques for modeling consumer preferences over multi-attributed stimuli, using experimental data to estimate the parameters of the specified utility function. Managerial contributions of CA are manifold and validated by thousands of commercial applications each year: from optimal design of products, through predicting market shares of brand offerings, to understanding how consumers make trade-offs between product features (and many more).

In CA experiments individual respondents are presented with a series of stimuli (product/service descriptions, illustrations, samples, prototypes, etc.), and are asked to rank or rate them, or to choose one from a set of alternatives. Let the vector x denote a multi-attribute product profile, and $\{U(x,\beta) : \beta \in \Theta \subset \mathbb{R}^p\}$ be a utility function parametric model. The profiles x_t are deterministic regressors in a compact set χ in an Euclidean space representing attributes (discrete dummy and/or continuous variables) and they form the sample $\{x_t\}_{t=1}^T$, with $T \geq p$. The responses, y_t, are evaluations of each product profile and are measured on the attitudinal scale (typically based on ratings, rankings or choice). Measures are affected by an error shock ε_t with zero mean and variance σ^2, satisfying

$$y_t = U\left(x_t, \beta^0\right) + \varepsilon_t, \quad t = 1, ..., T.$$

The first goal in CA is the estimation of the parameters β^0 in the utility model from experimental data. Stacking the data in matrices the model is $y = U\left(X, \beta^0\right) + \varepsilon$, where y, ε are $T \times 1$ vectors, $X = \left(x_1', ..., x_T'\right)' \in \chi^T$ is a full rank design matrix with all product profiles.

Different preference measurement scales and distributional assumptions can be considered, and based on this decision a variety of econometric methods can be used to estimate β^0, including ordinary or non linear least squares, several types of maximum likelihood estimators, least absolute deviations, etc. Under regularity conditions, the appropriate estimators are consistent and when T grows the re-scaled sequence $V_T^{-1/2}\left(\hat{\beta} - \beta^0\right)$ converges in distribution to a standard normal distribution $N(0, I)$ where V_T is a positive definite matrix converging in probability to a limit asymptotic covariance matrix V. The distribution of the error $\left(\hat{\beta} - \beta^0\right)$ is generally unknown, and the main tool to justify inferences for a medium-to-large size T is the asymptotic distribution of the scaled error. Both covariance matrices, V_T and the limit V, depend on the design matrix X (or sequence, if we focus on V) with the product profiles $\{x_1, x_2, ...\}$ shown in the experiment.

The efficiency of experimental estimators conveyed in the covariance matrices V_T, depends heavily on the product profiles evaluated by the respondents. Optimal experimental design maximizes the information elicited from the respondent, or equivalently minimizes the size of the covariance matrix. The goal of the researcher is usually to produce a good design matrix (or a sequence of profiles) so that $\phi(V_T)$ or the limit $\phi(V)$ respectively, are as small as possible. Here $\phi(\cdot)$ denotes such a measure of the matrix "size" which is: (1) positively homogeneous: $\phi(\delta A) = \delta\phi(A)$ for $\delta > 0$ to ensure independence from scale factors; (2) non-increasing: $\phi(A) \leq \phi(B)$ when $(A - B)$ is non negative definite; and (3) convex to ensure that ϕ satisfies the condition that information cannot be increased through interpolation. The typical measures are the trace (A-optimality criterion), and the determinant (D-optimality criterion), therefore we will focus on these two methods. Other matrix size criteria have been considered, but they usually render equivalent solutions. This result was established by the Kiefer-Wolfowitz equivalence theorem for linear models (Kiefer and Wolfowitz 1960) and later extended to nonlinear models by White (1973). What are the consequences of using designs, which generate estimators with larger covariance matrices? Implementing suboptimal designs requires a larger T to estimate the parameters with the same precision as an optimal design, which increases the market research cost and rating contamination caused by respondent's fatigue.

If the utility function is linear and classical regression model assumptions hold, typically $V_T = V(X)$ and optimal experimental designs minimizing $\phi(V(X))$ can be easily computed, for a literature overview with new results see Chapter 2. But the CA literature has considered a variety of models which are nonlinear in parameters, such as CA based on discrete choice, non-compensatory models, models with unknown ideal point, etc. In such cases the selection of optimal design is even more challenging because the covariance matrix $V_T = V\left(X, \beta^0\right)$ depends both on the deterministic regressors and the unknown parameters β^0 in a nonlinear way. To guarantee an efficient estimation of β^0 we need to compute an efficient experimental design X^* solving

$$\min_{X \in \chi} \phi\left(V\left(X, \beta^0\right)\right),$$

where the objective function is the size of covariance matrix for the usual estimators: maximum likelihood, nonlinear least squares, the generalized method of moments, and other related techniques.

In order to find an efficient design we need to know the value of β^0, which is unknown at the time the design is constructed - we want to estimate it from the experimental data! Therefore the design cannot be optimized without some assumptions about parameters and the data generating process. Since the size of covariance matrix is intrinsically linked to the unknown parameters, design efficiency is known only if the assumptions made on parameters are correct. The marketing research and the statistical experimental literature have made several attempts to solve this conundrum, for a literature review see the next section. So far a general solution has remained elusive.

In this essay we propose a novel general approach to construct designs for models in which the covariance matrix depends on unknown parameters. Although primarily we focus on discrete-choice CA, we also discuss how the method extends to other non-linear models considered in CA literature. This approach is generally more robust when the parameters deviate from the assumed values and is appropriate when there is no prior information about the unknown parameters.

3.2 Literature Review

Most of the literature about optimal experimental design is focused on linear models, but some of these ideas have been extended to deal with nonlinear specifications. In this section we consider the main approaches and their limitations. Consider a CA experiment, and a parametric estimator $\hat{\beta}$ such that $V_T^{-1/2}\left(\hat{\beta}-\beta^0\right)$ converges in distribution to a standard normal distribution $N(0,I)$, where V_T (and its limit V) depends on the design matrix (or sequence, if we focus on V) with the product profiles $\{x_1, x_2, ...x_T\}$ shown in the experiment. Typically, the covariance matrix can be expressed as a positive definite matrix

$$V_T = V_T\left(X,\beta^0\right) = \left(T^{-1}\sum_{t=1}^{T} v\left(x_t,\beta^0\right) v\left(x_t,\beta^0\right)'\right)^{-1}, \tag{3.1}$$

for some appropriate function $v\left(x,\beta^0\right)$, which holds for common estimators such as Maximum Likelihood or Nonlinear Least Squares. If $y_t|x_t$ are independent for different t with a regular conditional distribution $f\left(x_t,\beta^0\right)$, the MLE has an empirical covariance matrix

$$V\left(X,\beta^0\right) = T^{-1}\sum_{t=1}^{T} \frac{\partial \ln f\left(x_t,\beta^0\right)}{\partial\beta}\frac{\partial \ln f\left(x_t,\beta^0\right)'}{\partial\beta} = I_T\left(X,\beta^0\right)^{-1},$$

i.e. the inverse of the information matrix. If we estimate the CA model using NLS estimator with independent errors, then

$$V\left(X,\beta^0\right)=\sigma_\varepsilon^2\left(T^{-1}\sum_{t=1}^{T}\frac{\partial u\left(x_t,\beta^0\right)}{\partial\beta}\frac{\partial u\left(x_t,\beta^0\right)'}{\partial\beta}\right)^{-1}.$$

Under the standard regularity conditions, the limit variance of $V_T=V_T\left(X,\beta^0\right)$ in (3.1) is given by a positive definite matrix

$$V=\left(\int_\chi v\left(x,\beta^0\right)v\left(x,\beta^0\right)'\ w(dx)\right)^{-1} \tag{3.2}$$

where w is a probability measure obtained as the limit frequencies of using different profiles. If χ is finite (the typical case with discrete attributes), the integral can be written as an average

$$V=V\left(w,\beta^0\right)=\left(\sum_{j=1}^{r}w_j\ v\left(x_j,\beta^0\right)v\left(x_j,\beta^0\right)'\right)^{-1} \tag{3.3}$$

where the weights $\{w_j\}$ are non negative and sum up to one, and can be interpreted as limit frequencies with which r designs are replicated. For continuous regressors it is not so obvious, but if χ is a convex compact set we can obtain a similar representation applying the Krein–Milman Theorem as the inverse of a finite summation,where the r profiles $\{x_j\}_{j=1}^{r}$ are determined[1] by χ and $v\left(x,\beta^0\right)$, and the weights $\{w_j\}$ are limit frequencies of these profiles.

Exact optimal designs try to minimize $\phi\left(V_T\left(X,\beta^0\right)\right)$ in the design matrix X, whilst *approximated optimal designs* try to minimize $\phi\left(V\left(w,\beta^0\right)\right)$ in the limit frequencies w (which can be used to generate a $T\times K$ matrix X). The second approach was developed by Kiefer (1959) and his school. These two methodologies work only if V_T and V do not depend on unknown parameters β^0, which is the case for classical linear models but not for the models currently preferred by CA practitioners and researchers.

Some CA analysts use classical designs for linear models in the hope that Some CA analysts use classical designs for linear models in the hope that they will work well in a nonlinear con-

[1] The r points are the profiles associated to the set of extreme points of the set $\left\{vech\left(f\left(x,\beta^0\right)f\left(x,\beta^0\right)'\right):x\in\chi\right\}$, where $vech$ is the lower half-vectorization mapping.

text, but this is generally a wrong assumption. Optimal experimental designs are not robust to changes in the statistical model, as the structure of the covariance function changes dramatically. For example, Louviere and Woodworth (1983) considered a CA choice experiment suggesting that orthogonal and fractional-factorial *linear* designs are reasonably efficient for discrete-choice experiments. Street et al. (2005) consider a variety of ad hoc designs for choice CA.

The experimental design and CA literature approached this puzzle in two distinct ways: 1) assuming a specific value (vector) for the unknown parameter β^0, predominantly under the "all-zero" parameters hypothesis, and 2) assuming a probability measure on the parametric space $\Theta \subset \mathbb{R}^K$ and weighting all possible values in $\beta \in \Theta$. Hereafter we will refer to the former as the Local approach, and the latter as Average-Optimum (AO) approach. Below we discuss both methods in more detail.

3.2.1 Local Approach

Perhaps the most common solution to the presented puzzle is the local approach suggested by Chernoff (1953), which is based on adopting a guess for the unknown parameters. This decision may be arbitrary, based on an inefficient pilot study, or using human prior beliefs about the preferences. With $\beta^0 = \overline{\beta}$, the local approach looks for a design X^+ defined as the solution to

$$\min_{X \in \chi} \phi\left(V\left(X, \overline{\beta}\right)\right), \tag{3.4}$$

where $\overline{\beta} \in \Theta$ is the assumed parameter vector. As the solution X^+ is specific to $\beta^0 = \overline{\beta}$, the resultant designs are locally optimal and are not optimal for values different from $\overline{\beta}$. Unfortunately, the efficiency of the locally optimal design, X^+, may be sensitive to even small perturbations in $\overline{\beta}$, and this initial guess is rarely close to the true β^0 (for if we had a good estimation, there would be no reason to run the experiment). In general we do not have any prior control over the efficiency of the design X^+ under the true β^0.

For example, for Maximum Likelihood estimation the local optimal design is computed by minimizing $\min_{X \in \chi} \phi\left(I_T\left(X, \overline{\beta}\right)^{-1}\right)$ for a specific $\overline{\beta}$, where $I_T(\cdot)$ is the information matrix. The statistical literature on surface response experiments has considered the local approach in a nonlinear regression context by minimizing

$$\phi\left(\left(\sum_{t=1}^{T}\frac{\partial u\left(x_t,\overline{\beta}\right)}{\partial\beta}\frac{\partial u\left(x_t,\overline{\beta}\right)'}{\partial\beta}\right)^{-1}\right).$$

This is actually an old idea in the statistical literature. Chernoff (1953) linearized the nonlinear regression model using the first order Taylor expansion about a preliminary value, $\overline{\beta}$, applying the maximum trace criterion to obtain locally optimal designs for this linearized model. Box and Lucas (1959) also linearized the model applying the maximum determinant criterion. Abdelbasit and Plackett (1983) as well as Minkin (1987) used similar strategies in experiments with binary response. Similarly, Kiefer (1959) proposed a local version of approximate optimal designs by solving the problem

$$\min_{w}\phi\left(V\left(w,\overline{\beta}\right)\right)$$

in w for a given $\overline{\beta}\in\Theta$, where $V\left(w,\overline{\beta}\right)$ is matrix (3.3) considered for the nonlinear least squares estimator.

In CA context the local approach under null-hypothesis of $\overline{\beta}=0$ has been used by Kuhfeld et al. (1994) for finding D-optimal choice designs for large conjoint applications through computerized search, and for discrete-choice experiments Kanninen (2002) suggested a procedure that leads to maximizing $|X'X|$ with continuous regressors. Notice that in the context of choice models, the assumption $\overline{\beta}=0$ implies that all alternatives in the choice set have the same utility and probability to be selected from the set (all sets have perfect utility-balance). Huber and Zwerina (1996) have studied the effects of incorporating manager's prior beliefs into the optimal design, showing that under $\overline{\beta}\neq 0$ utility balance of choice sets remains an important property of efficient choice designs. Street et al. (2005) compared various strategies based on linearized designs to construct choice designs of pairs and triplets for experiments with a different number of attributes, levels and choice sets. The Local Approach has also been used by Marley and Louviere (2005) to elicit additional information about the ranking of alternatives within sets from the respondent's best-worst choices. This design was later extended in the experiment measuring attribute-level best-worst choices Marley et al. (2008). Bunch et al. (1996) developed a heuristic *cycling* procedure initiated with an orthogonal fractional factorial design treated as a list of one-element choice sets. However this approach is not robust to perturbations in parameters.

3.2.2 Average-Optimum Approach

In an attempt to reduce the influence of $\overline{\beta}$, some authors considered an average of many values instead of the local design. This method involves a probability measure μ defined over the parametric space Θ, optimizing the weighted average of design efficiencies

$$\min_{X\in\chi}\int_{\Theta}\phi\left(V\left(X,\beta\right)\right)\mu\left(d\beta\right). \tag{3.5}$$

The solution X^{++} is not optimal under each scenario but hedged against the risk associated with all scenarios. In general this method is unrelated to the Bayesian inference method, but some authors call it "the Bayesian approach", since both of them share the use of a prior belief μ. The solution is quite sensitive to the choice of the weighting probability distribution μ (and its parameters). Unless μ is strongly concentrated near the true unknown β^0, little can we say about the true efficiency of the design, $\phi\left(V\left(X^{++},\beta^0\right)\right)$.

The statistical literature developed this method several decades ago. It was introduced by Pronzato and Walter (1985), and a detailed analysis can be found in Pronzato and Pázman (2013, Ch.8). This approach has been used in CA to build exact optimal designs for choice models by Sándor and Wedel (2001) in the context of a single respondent, setting μ as a normal distribution representing managers' prior beliefs about product market shares. The Averaged Approach has also been applied in the Mixed Logit model (Arora and Huber 2001; Sándor and Wedel 2002). Sándor and Wedel (2005) extended the idea to panels of heterogeneous customers generating a different design for each customer. The Averaged Approach can also be applied in the Kiefer context, minimizing

$$\min_{w} \int_{\Theta} \phi\left(V\left(w,\beta\right)\right)\mu\left(d\beta\right)$$

in the relative frequencies $\{w_j\}$ defined over a given set of alternative profiles $\{x_j\}_{j=1}^{r}$.

In both, the local and the averaged approaches, the numerical optimization is generally not based on state-of-art algorithms developed in the operations research literature. The designs are often computed using either a Modified Fedorov algorithm, or the *cycling* procedure of Bunch et al. (1996). To improve the utility balance and design efficiency some authors additionally include a *swapping* and *relabeling* step: the former involves switching levels of alternatives within each set to improve their balance; the latter reassigns labels of levels in the design. Ferrini and Scarpa (2007) provide a review of the Logit CA experimental design literature and the algorithms implemented.

Overall, the assumptions about unknown parameters β^0 are specific to a given application. Little is known about empirical validity or optimality claims of implemented designs when these assumptions are violated (Louviere et al. 2011). There have been few studies, which tested the robustness of the approach with misspecified β^0. In this essay we propose a method to build the experimental designs that ensure robustness. We also provide simple comparative graphics to demonstrate the strength of this approach.

3.3 The Worst-Case (WC) Approach

In this section we present a general approach for optimal designs in problems where the covariance matrix depends on an unknown parameter. The underlying idea of this method is the robustness property of the worst case prevention: the optimal WC design is the one whose worst outcome is at least as good as the worst outcome of any other designs. In other words, the WC design is the best among any other designs under the worst-case scenario.

We define a WC design as the matrix of product profiles X^{wc} solving the problem

$$\min_{X\in\chi}\ \max_{\beta\in\Theta}\ \phi\left(V\left(X,\beta\right)\right), \tag{3.6}$$

with χ representing the set of feasible deterministic regressors, where we can introduce different types of constraints: bounds on continuous attributes, discrete regressors (for dummy attributes), or distance constraints to avoid repetition of profiles shown to the same respondent (see Chapter 2). Θ corresponds to the range of the unknown parameters. When χ and Θ are compact sets, and $V\left(X,\beta\right)$ is a continuous function in $\chi\times\Theta$, the existence of a solution is guaranteed by a standard application of the Weierstrass Theorem and the Maximum Theorem. The strategy behind minmax designs is to minimize the maximum size of variance-covariance matrix, where the maximum (maxima) is found over a specified range of the unknown parameters. Define $\mathbf{B}(X)$ as the correspondence allocating to each matrix $X\in\chi$ all maxima associated with it, ie. solutions to $\max_{\beta\in\Theta}\ \phi\left(V\left(X,\beta\right)\right)$, then WC designs solve $\min_{X\in\chi}\left\{\phi\left(V\left(X,\beta\right)\right):\beta\in\mathbf{B}(X)\right\}$.

The statistics literature has previously considered minimax design optimization for some specific nonlinear models. For example, Sitter (1992) considered minimax designs in the context of a binary choice model. However most literature considers Kiefer's approximate designs (Melas 1978; Fedorov 1980; Müller and Pázman 1998). The method is not fully developed and the strategies for their construction are somewhat ad hoc, (some examples are included in Wong 1992; Haines 1995; Imhof 2001). In this chapter, we do not apply the Kiefer's approach because design replications are not suitable in conjoint analysis, and we also consider general algorithms that allow the application of this idea to any suitable specification.

Why should risk averse CA modelers consider WC experimental designs? Although the true size of variance-covariance matrix is not identified when the parameters are unknown (or we do not have a preliminary estimator of β^0), WC designs guarantee that the size of the true covariance matrix $\phi\left(V\left(X^{wc},\beta^0\right)\right)$ is bounded by a known quantity,

$$\phi\left(V\left(X^{wc},\beta^0\right)\right)\leq\max_{\beta\in\Theta}\phi\left(V\left(X^{wc},\beta\right)\right)=\min_{X\in\chi}\ \max_{\beta\in\Theta}\ \phi\left(V\left(X,\beta\right)\right).$$

In other words, the minmax problem defined in (3.6) implicitly establishes an upper bound on the size of the true variance-covariance matrix $V\left(X^{wc},\beta^0\right)$. Therefore for any $\beta^0 \in \Theta$, the design efficiency associated with the WC design X^{wc} will be at least as good as the worst-case value, but it will never be worse. The local and average-optimum procedures do not guarantee any upper bound over the true variance of the design.

The optimal design in the local approach, X^+, obtained as a solution to the problem (3.4), does not guarantee small variances under the true β^0. On the contrary, $\phi\left(V\left(X^+,\beta^0\right)\right)$ can be arbitrarily large and the fitted model becomes unreliable. In contrast, WC designs guarantee an improvement of the true variance regardless of the value of β^0. This robustness is a key characteristic of WC methods. A practical advantage of this approach is that the researcher is only required to specify an appropriate range for the unknown parameters, instead of a specific value.

Due to similar reasons, designs obtained with the averaged approach are not robust when the assumptions about parameters are violated. The optimality of the design X^{++}, which minimizes the average size of variance-covariance matrix, is highly sensitive to the choice of μ. Unless this distribution is concentrated near the true unknown β^0, the design X^{++} can render estimators with quite large variances. Again, a cautious modeler should consider a conservative distribution:

$$\min_{X\in\chi}\max_{\mu\in\mathscr{M}}\int_{\Theta}\phi\left(V\left(X,\beta\right)\right)\mu\left(d\beta\right), \tag{3.7}$$

where $\mathscr{M}$ is the class of probability measures on Θ. Proposition 1 states that WC design is the solutions to the above problem.

Proposition 5 *Assume that $\phi\left(V\left(X,\beta\right)\right)$ is non negative, then*

$$\min_{X\in\chi}\max_{\mu\in\mathscr{M}}\int_{\Theta}\phi\left(V\left(X,\beta\right)\right)\mu\left(d\beta\right) = \min_{X\in\chi}\max_{\beta\in\Theta}\phi\left(V\left(X,\beta\right)\right). \tag{3.8}$$

Proof. For all $X \in \chi$ and $\mu \in \mathscr{M}$,

$$\int \phi\left(V\left(X,\beta\right)\right)\mu\left(d\beta\right) \leq \max_{\beta\in\Theta}\phi\left(V\left(X,\beta\right)\right)\int \mu(dy) = \max_{\beta\in\Theta}\phi\left(V\left(X,\beta\right)\right).$$

Furthermore, for any $X \in \chi$, it is satisfied that

$$\max_{\mu\in\mathcal{M}}\int_{\Theta}\phi\left(V\left(X,\beta\right)\right)\mu\left(d\beta\right) \geq \int_{\Theta}\phi\left(V\left(X,\beta\right)\right)\mu^{wc}\left(d\beta\right) = \max_{\beta\in\Theta}\phi\left(V\left(X,\beta\right)\right),$$

where μ^{wc} is any measure of probability with all its mass in

$$\beta(X) = \left\{\beta\in\Theta : \phi\left(V\left(X,\beta\right)\right) = \max_{\theta\in\Theta}\phi(V(X,\theta))\right\},$$

and the result follows. ■

The worst case approach is based on a general principle of the parametric robustness. The method is appropriate for standard parametric estimators, such as the nonlinear least squares or Maximum Likelihood (ML) estimators. We will not pay specific attention to the Bayes estimation methods, as the optimal designs computed for ML are also valid for Bayesian estimations. Notice that under regularity conditions, and regardless of the prior distribution, the Bernstein-von-Mises Theorem ensures that the posterior distribution of parameters has an asymptotically normal conditional mean with the same covariance matrix as the ML estimator. Therefore the optimal designs for Bayesian estimators are the same as the ones computed for maximum likelihood.

3.3.1 Computation of the Solution

To obtain WC designs, we face the issue of solving a minimax problem. The function $\phi\left(V\left(X,\beta\right)\right)$ is typically continuous in $\beta\in\Theta$, but not necessarily in χ. With continuous regressors it is usually differentiable in $X\in\chi$, but we often have discrete dummy regressors and then χ is a finite set.

Can we characterize the solution? Let us start discussing the case of continuous regressors. Even if $\phi\left(V\left(X,\beta\right)\right)$ is continuously differentiable, the function $\max_{\beta\in\Theta}\phi\left(V\left(X,\beta\right)\right)$ is not differentiable in X and standard optimization tools cannot be applied to compute the minimax solution

X^{wc}. However, if $\Theta = \{\beta \in \mathbb{R}^p : h(\theta) \leq 0\}$ where h is a continuously differentiable vector valued mapping and χ is a nonempty compact set, then there exists necessary conditions for the solution of a minimax problem (more involved than the classical Karush-Kuhn-Tucker), see Shimizu and Aiyoshi (1980, Theorem 1), or Demy̆anov and Malozemov (1974), and there are also sufficient conditions for a point satisfying the necessary conditions to be a minimax optimum (Bector and Bhatia 1985, see, e.g.). But, in practice, appropriate numerical methods are typically required to compute the worst-case solution. With discrete regressors X the analytical necessary conditions do not hold, and we need to use numerical methods in general.

The rest of this section deals with minimax computational tools to compute worst case designs. Essentially the steps of the algorithm are analogous for both continuous and discrete regressors but in the latter case the implementation is based on integer programming methods. Minimax optimization problems can be handled using algorithms for semi-infinite programming, because any continuous minimax problem such as (3.6) can be expressed as

$$\min_{X \in \chi,\, \rho \in \mathbb{R}} \left\{ \rho : \max_{\beta \in \Theta} \phi\left(V\left(X, \beta\right)\right) \leq \rho \right\},$$

which is equivalent to the semi-infinite problem:

$$\begin{aligned} \min_{X \in \chi,\, \rho \in \mathbb{R}} \quad & \rho \\ \text{s.t.} \quad & \phi\left(V\left(X, \beta\right)\right) \leq \rho, \text{ for all } \beta \in \Theta. \end{aligned} \tag{3.9}$$

To compute the WC design we implement a global optimization algorithm for minimax optimization proposed by Shimizu and Aiyoshi (1980) and later developed by Zakovic and Rustem (2003). It uses a global optimization approach with respect to $\beta \in \Theta$ and cutting planes to reduce the feasible region when constraint violation is encountered. Descriptively, each iteration consists of interchangeably solving the "min" and "max" problems: (i) solving the "min" problem with respect to X, subject to semi-infinite constraints associated with all global maxima with respect to

β, as in Problem (3.9); (ii) solving the global "max" problem with respect to β with X obtained in (i). Formally, the l-th iteration of this algorithm solves the problem:

$$\begin{aligned} \min_{\substack{X^{l+1}\in\chi \\ \rho^{l+1}\in\mathbb{R}}} \quad & \rho^{l+1} \\ \text{s.t.} \quad & \phi\left(V\left(X^{l+1},\beta_i\right)\right) \le \rho^{l+1}, \qquad i = 1,\dots,k_l, \end{aligned}$$

where $\{\beta_i\}_{i=1}^{k_l} \subset \mathbf{B}\left(X^l\right)$, and $\mathbf{B}\left(X^l\right)$ contains all global maxima computed with respect to $\beta \in \Theta$ at the previous iteration. Next we solve

$$\max_{\beta\in\Theta} \quad \phi\left(V\left(X^{l+1},\beta\right)\right)$$

computing all global maxima $\{\beta_i\}_{i=1}^{k_{l+1}} \subset \mathbf{B}\left(X^{l+1}\right)$, and check if the solution is feasible up to an arbitrary positive tolerance ε. If

$$\max_{\beta\in\Theta} \phi\left(V\left(X^{l+1},\beta\right)\right) > \rho^{l+1} + \varepsilon,$$

we iterate further, otherwise the algorithm terminates and $X^{wc} = X^{l+1}$ is the solution to the minimax problem. This algorithm converges in a finite number of iterations, on each one of them it is required to solve standard "min" and "max" problems. The global optimization approach is essential to guarantee the robustness of the solution of minimax problems because one of the crucial steps in solving the Problem (3.9) is to find a number of semi-infinite constraints by computing the global maximizers in the program $\max_{\beta\in\Theta} \ \phi\left(V\left(X^l,\beta\right)\right)$. In global optimization algorithms, all candidates for local maximizers must usually be bracketed by a comparison of function values $\max_{\beta\in\Theta} \ \phi\left(V\left(X^l,\beta\right)\right)$ on a sufficiently dense finite subset of Θ. To reduce the cost of computing global optima the domains χ and Θ should be restricted as much as possible given the information available. The monograph of Rustem and Howe (2002) is focused on the computation of minimax problems.

We have implemented the algorithm for WC-optimal designs in MATLAB 6.5 on a computer with Intel Core 2 Duo processor and machine precision 10^{-16}, programming directly the Zakovic and Rustem algorithm 2003 and applying standard optimization routines for solving the subproblems of the original algorithm (for both continuous and integer problems). To find the global maximizers subproblem we used the algorithm DIRECT for constrained mixed-integer global optimization (Jones 2001), which is available commercially with TOMLAB toolbox (http://tomopt.com/tomlab/). To handle discrete regressors in χ we consider MINLP solver developed by Roger Fletcher and Sven Leyffer (Leyffer 2001) also implemented in TOMLAB. Notice that purely continuous problems can be solved with a flexible "fmincon" routine from *Optimization toolbox* in MATLAB.

Additionally, MATLAB offers some specific routines to solve continuous minmax problems, such as the "fseminf" function from the *Optimization toolbox*. This is a dedicated subroutine for solving semi-infinitely constrained multivariate and nonlinear optimization problems, as the one defined in (3.9). The routine "fseminf" first estimates peak values in the semi-infinite constraints, which are later submitted as constraints in the minimization problem. The main disadvantage of this approach is the computational cost. Instead of searching for global maximizers over Θ, it requires that the intervals for β be discretized into a finite grid of values. Then the whole grid is submitted to the routine as the semi-infinite constraints and evaluated at each iteration. The number of the constraints depends on several factors: (1) number of product attributes, (2) size of Θ, and (3) how "fine" the grid is (the grid's step size). In practice, using this routine we could solve only very small conjoint scenarios in reasonable time. Another drawback is that "fseminf" solves only continuous problems.

3.4 A Paradigmatic Example of WC Designs: Multinomial Logit Model

Over the last few years CA experiments have been increasingly based on respondents' choices from a subset of alternatives. McFadden's 1974 Multinomial Logit model is the most widely accepted model of choice-based CA. Assume J alternatives in a choice set and each alternative is characterized by the attributes $\{x_j\}_{j=1}^{J}$. If the latent utility of the alternative j is $u_j = x_j'\beta + \varepsilon_j$, and ε_j has a type I extreme value distribution, then the probability that a consumer selects the alternative j from the set $\{x_1,...,x_J\}$ is

$$\pi_j(x,\beta) = \frac{\exp\left(x_j'\beta\right)}{\exp\left(\sum_{l=1}^{J} x_l'\beta\right)}.$$

With $t = 1,2,..,T$ sets we codify choices $y_t = (y_{t1},...,y_{tJ})'$ into a vector of dummies so that y_{tj} is equal to 1 if alternative j is selected and zero otherwise. Then the model can be estimated by Maximum Likelihood, maximizing

$$L(X,\beta) = \sum_{t=1}^{T}\sum_{j=1}^{J} y_{tj} \ln \pi_j(x_t,\beta) = \sum_{t=1}^{T}\sum_{j=1}^{J} y_{tj}\left(x_{tj}'\beta - \ln\left(\sum_{l=1}^{J}\exp\left(x_{tl}'\beta\right)\right)\right).$$

The gradient and the information matrix are given by

$$\begin{aligned}
\frac{\partial L(X,\beta)}{\partial\beta} &= \sum_{t=1}^{T}\sum_{j=1}^{J}\left(y_{tj} - \pi_j(x_t,\beta)\right)\ x_{tj}, \\
I(X,\beta) &= E\left[\frac{\partial L(X,\beta)}{\partial\beta}\frac{\partial L(X,\beta)'}{\partial\beta}\right] = \sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{l=1}^{J} E\left[\left(y_{tj} - \pi_j(x_t,\beta)\right)\left(y_{tl} - \pi_l(x_t,\beta)\right)\right]\ x_{tj}x_{tl}' \\
&= \sum_{t=1}^{T}\sum_{j=1}^{J}\pi_j(x_t,\beta)\left(1-\pi_j(x_t,\beta)\right)\ x_{tj}x_{tj}' - \sum_{j\neq l}^{J}\pi_j(x_t,\beta)\pi_l(x_t,\beta)\ x_{tj}x_{tl}'
\end{aligned}$$

where we have used that $cov\left(y_{tj},y_{tl}\right) = \pi_j(x_t,\beta)\left(1-\pi_j(x_t,\beta)\right)$ for $l = j$, and $-\pi_j(x_t,\beta)\pi_l(x_t,\beta)$ for $l \neq j$. The asymptotic covariance matrix of the maximum likelihood estimator can be estimated by the inverse of the Hessian, and it is upper-bounded since

$$I(X,\beta) \leq \sum_{t=1}^{T}\sum_{j=1}^{J}\pi_j(x_t,\beta)\ x_{tj}x_{tl}' \leq X'X,$$

using that $\sum_{j=1}^{J}\sum_{l=1}^{J}\pi_j(x_t,\beta)\pi_l(x_t,\beta)x_{tj}x_{tl}'$ is non negative definite. Therefore,

$$\phi\left(Var\left(\hat\beta\right)\right) = \phi\left(I(X,\beta)^{-1}\right) \geq \phi\left(\left(X'X\right)^{-1}\right).$$

A commonly used procedure consists of minimizing the lower bound for the covariance matrix $\phi\left(\left(X'X\right)^{-1}\right)$. In particular Kanninen (2002) maximizes $\left|X'X\right|$, finding that the optimal design places the attributes at the extreme points of the domain χ. This is also the approach considered by Kuhfeld et al. (1994), based on the Fedorov algorithm. The presented algorithm can be used in this context. However, this is not a reliable solution, as $\phi\left(I\left(X,\beta^0\right)^{-1}\right)$ could be much higher than $\phi\left(\left(X'X\right)^{-1}\right)$. By contrast, in this essay we consider a design X^{wc} solving $\min_{X\in\chi}\max_{\beta\in\Theta}\phi\left(I\left(X,\beta\right)^{-1}\right)$ which provides a robust solution with a bounded variance.

3.4.1 Comparison of Local, Average and WC Approach – a Simulated Example

This section provides an informal comparison of the approaches that have been proposed in CA literature to deal with uncertainty of designs in discrete-choice experiments. This uncertainty results from the fact, that the objective function (size of the covariance matrix) depends not only on the deterministic regressors, but also on unknown model parameters. Therefore, the researcher has to select an experimental design not knowing its true efficiency.

Conjoint literature has approached this problem by (1) assuming a specific value for β; or (2) postulating a probability distribution for unknown parameters. From the optimization point of view, the local approach (1) is a deterministic problem, because uncertainty is eliminated by assuming only a single scenario associated with the fixed parameter values. The design is optimal only for the assumed scenario, and we are not able to evaluate its efficiency if parameters are different from the assumed values. This is a naive approach, because if we knew real values we would not have to make an experiment in the first place. In the average-optimum approach (2) we optimize the expected design efficiency. Here unknown parameters are treated as random variables, and are described through a probability function. This approach suffers from the curse of dimensionality.

The proposed approach is based on robust optimization, and the objective is to minimize the worst case value of design efficiency. In this framework uncertainty related to unknown parameters is modeled in a deterministic way, based on bounded and convex sets. This approach has several advantages: we do not need to know the probability distribution for uncertainty, and it does not suffer from the curse of dimensionality. Its main limitation is that it is a conservative approach, because we look for the best efficiency under the worst-case scenario.

We have compared the performance of the WC approach with the local and average-optimum approach in a simple discrete-choice example. For clarity purposes we show the robustness of WC designs in a univariate setting, but the same idea applies to models with more parameters. The comparison setting is as follows. First, we solve Problem (3.4), Problem (3.5) and Problem (3.9) with the same initial point to obtain the local, AO and WC designs. Next, we calculate the efficiency of each of the designs, and plot it as a function of true parameter value. In this section we focus on the trace of covariance matrix as a measure of efficiency for these simulated designs (A-optimality criterion). In the next section we will demonstrate the robustness of WC approach using the determinant criterion.

Table 3.1: Parameters in simulated examples

	Local approach	AO approach	WC approach
# choice sets	4	4	4
# alternatives/set	4	4	4
# attributes	1	1	1
Assumptions about β			
Local approach	$\bar{\beta} = 0$	——	$\beta \in [-1,1]$
Average-optimum approach	——	$\beta_i \sim N(0,\sigma_i^2)$ $\sigma_i^2 = \{1,2,3\}$	$\beta \in [-2\sigma_i^2, 2\sigma_i^2]$

Table 3.1 summarizes the parameter values used for the simulation of comparative figures. In all examples we find designs for conjoint experiment with 1 continuous attribute, and 4 choice sets, each having 4 alternatives (1/4/4 design). For the local approach we compute the design assuming $\beta = 0$; for the AO approach we simulate β from normal probability distribution with mean 0 and different values of standard deviations: (1,2,3).

Figure 3-1 compares the efficiency of optimal designs obtained with the worst case approach and the local approach. For comparability, we solve both optimization problems with the same initial point. The horizontal axis represents different values of the true parameter, β^0; the true efficiency of the design (measured as the trace of the covariance matrix) is shown on the vertical axis. The efficiency of the optimal design obtained with the local approach is represented by the solid line, and the WC design - by a dashed line.

Figure 3-1: Efficiency comparison of the local and WC design. Simulated univariate case.

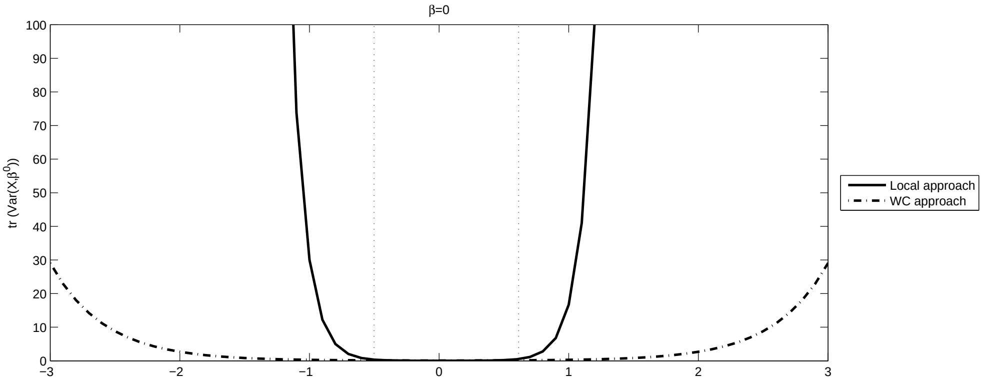
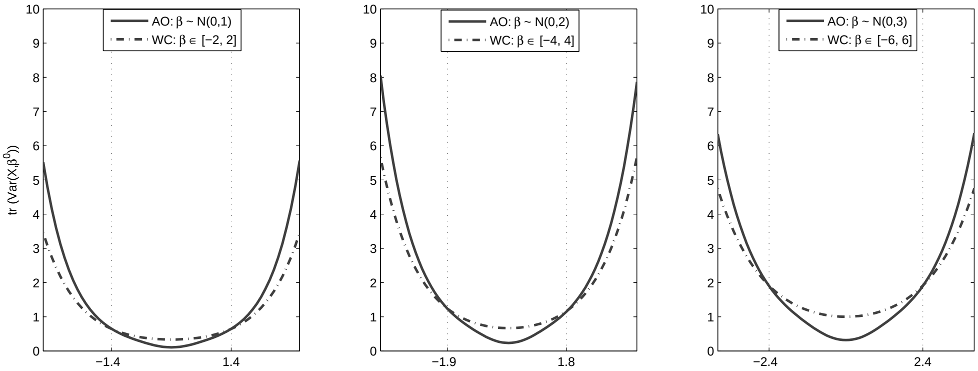


Figure 3-1 illustrates clearly the intuition behind WC design. The local approach performs better, when true parameters are close to the assumed values. The further the true parameters are from the assumed values, the bigger is the advantage of the robust WC approach. Vertical dashed lines indicate small intervals where the local approach dominates. The local design has been computed under the assumption that $\beta = 0$, and indeed the trace values of the covariance matrix are lower when the true parameter is close to 0 (values between -0.5 and 0.6). It is easy to see that the local approach is not robust, when the true parameter deviates even to the small degree from the assumed values. Worst-case design dominates even when the true values are outside of the assumed interval.

We have also compared the robust WC approach to the average-optimum approach. For the latter we optimize the expected design efficiency, assuming that parameters have a normal probability distribution with mean 0 and different values of standard deviation. The WC design has been computed assuming that the parameters lie in the interval within 2 standard deviations from the mean. As before, we start the optimization routines with the same initial point.

Similarly to the local approach, we observe that if the true parameter values are close to the assumed mean of the normal distribution then average-optimum approach performs better, but for the values far from 0, the WC approach is more robust. Dotted vertical lines indicate the interval where the AO approach performs better than the WC approach. Interestingly, the

Figure 3-2: Efficiency comparison of the AO and WC design. Simulated univariate case.



researcher can improve the robustness of average-optimum designs by postulating the model, which accounts for more uncertainty of parameters, ie. increasing the assumed standard deviation of the normal distribution. The design shown in the right panel of Figure 3-2 is more robust than the designs shown in the left and center panels. However, it is more computationally intensive and one still has to make sure that the assumed mean is close to the true parameter value.

3.4.2 Solved Designs and Comparison with Literature Benchmarks

In this section we present a couple of computed worst-case designs and we compare the robustness of such designs with their local and average-optimum counterparts proposed in conjoint literature. The scenarios include both continuous and discrete attributes and the optimization problem formulation for both types of regressors is essentially analogous. With continuous attributes the levels *per se* do not exist, and the algorithm looks for the solution throughout the whole attribute space without restriction. With discrete attributes the levels are recoded into 0-1 values, representing whether the level is respectively absent or present in the shown product profile. From the optimization point of view, we need to impose two restrictions on the solution to obtain a valid design matrix of discrete attributes: 1) the WC-optimal design matrix contains only zeros and ones; and 2) within each attribute exactly one level is shown. Apart from the

Table 3.2: Overview of the computed designs and references

Name	Type[a]	Size[b]	β_{WC}	Benchmark	Benchmark β
SCN1	C	2/2/6	$\beta_{wc} \in [0,1]$	Kanninen (2002)	$\overline{\beta} = 0$
SCN2	C	4/4/12	$\beta_{wc} \in [0,1]$	Kanninen (2002)	$\overline{\beta} = 0$
SCN3	C	8/8/24	$\beta_{wc} \in [0,1]$	Kanninen (2002)	$\overline{\beta} = 0$
SCN4	D	3^3/3/9	$\beta_{wc} \in [0,1]$	Huber and Zwerina (1996)	for each attribute: $\overline{\beta}_{HZ} = [-1,0,1]'$
SCN5	D	3^4/2/15	$\beta_{wc} \in [0,1]$	Huber and Zwerina (1996) Sándor and Wedel (2001)	as above $\beta \sim N(\beta_{HZ}, \Sigma_0)$

[a] C - continuous regressors, D - discrete regressors

[b] The notation represents: # attributes (or levels & attributes) / # alternatives per choice set/ # choice sets.

additional requirements about the solution, the formulation of the continuous and integer optimization problems is equivalent, although the implementation requires different optimization algorithms, which were described in Section 3.3.1.

For comparability purposes we compute designs for specific problems that have been solved using the local and AO approach in the standard McFadden's 1974 framework. Table 3.2 summarizes the scenarios addressed here, reporting the size parameters, corresponding literature benchmarks and the assumptions made about β in the local, AO and WC approach. The interval for parameter values relates to the literature benchmarks and is constant across different WC scenarios. Our comparison strategy consists of simulating the true unknown parameter values (vectors) and for each of them we compute the efficiency of the WC and the benchmark design, calculating the determinant of the covariance matrix $\left|V\left(X,\beta^0\right)\right|$. To inspect the robustness to misspecification of parameters and propensity to outliers, we plot the true design efficiency with respect to the distance between the true and the assumed parameter value. We also compare the pairs of designs individually, counting the number of cases when the WC design is more efficient than the corresponding benchmark, which is summarized in the Table 3.6 at the end of this section.

The first three scenarios involve WC design with purely continuous attributes. We use the designs reported by Kanninen (2002) in Table 1 (page 218) as the local approach benchmarks, as they are computed under the assumption that all parameters are zero. Hereinafter, we will refer to these designs as KAN1, KAN2 and KAN3. These benchmarks have the same number of choice sets as the number of continuous variables, $N^* = k$, and the implemented design is constructed by replication. Therefore, following Kanninen (2002) approach, KAN1-KAN3 have repeated choice sets, while our WC designs have unique choice sets. All continuous examples involve a binary choice setting, with 2, 4 and 8 continuous attributes (our scenarios SCN1, SCN2, and SCN3 respectively). To compute the WC designs, we set the lower and upper bounds of the attribute values to 1 and 5, and the assumed β lies between 0 and 1.

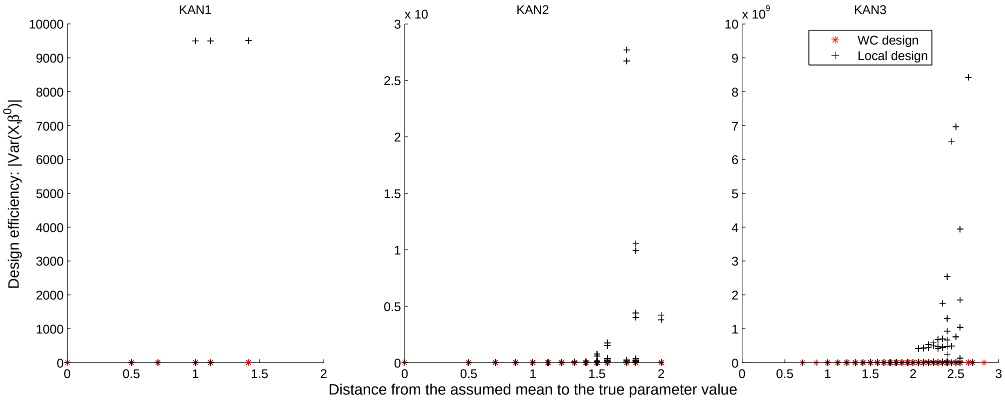
Table 3.3 presents the WC designs computed for the SCN1 and SCN2 examples, and their local counterparts (KAN1 and KAN2). The first apparent difference is the fact that our worst-case approach searches for the optimal design over the whole continuous attribute space without restrictions. It is expected that the solution to the nonlinear problem lies inside the feasible region, contrary to the linear problems whose solutions lie at the boundary of the feasible region. The approach proposed by Kanninen (2002) exhibits the properties of the linear design: all attributes except one lie on the boundary, while the first continuous attribute is used to scale response probabilities. Finally, an important issue is the duplication of choice sets: our WC approach can be implemented with the single respondent, while in Kanninen's approach each respondent replies only to a fraction N^*/N of choice questions.

We have also compared the behavior of WC approach and the local approach in terms of the robustness to misspecifications in the assumptions about the true parameters, β^0. Figure 3-3 shows the relation between the design efficiency and the distance to the true parameter value for continuous WC designs (SCN1-SCN3) and the local designs (KAN1-KAN3). To obtain the vectors of true parameters, β^0, we construct all possible vectors containing values $(-1,-0.5,0,0.5,1)$. Then, we compute the distance between the assumed parameter value and the simulated true parameters, $||\beta^0 - \overline{\beta}||$. Red stars correspond to WC design efficiency, $|V(X_{WC},\beta^0)|$, while black crosses represent the efficiency of KAN designs, $|V(X_{KAN},\beta^0)|$.

Table 3.3: Computed worst-case designs in SCN1 and SCN2 and their benchmarks

Set	KAN A		SCN1		Set	KAN B				SCN2			
1/1	1.54	5.00	5.00	1.00	1/1	1.04	5.00	5.00	5.00	1.77	3.81	2.25	2.50
1/2	0	1.00	1.00	5.00	1/2	0	1.00	1.00	1.00	4.27	1.62	3.15	4.11
2/1	1.54	1.00	5.00	2.30	2/1	1.04	5.00	1.00	1.00	3.12	4.76	3.19	1.00
2/2	0	5.00	1.00	4.68	2/2	0	1.00	5.00	5.00	4.36	1.69	2.39	5.00
3/1	1.54	5.00	5.00	1.00	3/1	1.04	1.00	5.00	1.00	4.76	4.98	1.00	4.99
3/2	0	1.00	1.00	5.00	3/2	0	5.00	1.00	5.00	1.16	1.22	5.00	1.05
4/1	1.54	1.00	3.62	3.54	4/1	1.04	1.00	1.00	5.00	1.00	4.77	3.15	4.37
4/2	0	5.00	2.29	2.46	4/2	0	5.00	5.00	1.00	3.47	1.43	1.48	3.41
5/1	1.54	5.00	2.36	2.44	5/1	1.04	5.00	5.00	5.00	3.95	1.00	5.00	4.56
5/2	0	1.00	3.64	3.56	5/2	0	1.00	1.00	1.00	1.94	4.96	1.00	1.37
6/1	1.54	1.00	2.18	5.00	6/1	1.04	5.00	1.00	1.00	3.12	4.54	3.51	2.89
6/2	0	5.00	4.57	1.00	6/2	0	1.00	5.00	5.00	4.94	1.37	3.21	2.78
					7/1	1.04	1.00	5.00	1.00	4.92	3.25	5.00	1.00
					7/2	0	5.00	1.00	5.00	1.08	3.78	1.00	5.00
					8/1	1.04	1.00	1.00	5.00	1.37	1.32	5.00	4.79
					8/2	0	5.00	5.00	1.00	2.71	4.02	1.00	1.00
					9/1	1.04	5.00	5.00	5.00	1.49	1.49	4.18	3.94
					9/2	0	1.00	1.00	1.00	3.41	4.52	1.27	2.95
					10/1	1.04	5.00	1.00	1.00	4.18	4.60	1.00	1.00
					10/2	0	1.00	5.00	5.00	2.30	2.20	4.76	5.00
					11/1	1.04	1.00	5.00	1.00	1.00	5.00	5.00	3.42
					11/2	0	5.00	1.00	5.00	5.00	1.00	1.00	3.41
					12/1	1.04	1.00	1.00	5.00	4.26	2.90	5.00	1.00
					12/2	0	5.00	5.00	1.00	1.49	2.11	1.00	4.87

Figure 3-3: Efficiency comparison of the local and worst-case designs: “KAN” scenarios .



Clearly, the worst-case approach is more robust to the misspecification of parameters: if the assumed parameter is far from the true value, Kanninen’s approach yields inefficient designs. The comparison of the three scenarios shows that for larger conjoint applications the risk of the local design being inefficient or even having infinite variance is severe (notice the number of outliers with huge variance in KAN3). The worst-case designs maintain robustness property in all scenarios considered regardless of the problem size.

The remainder of this section is devoted to conjoint designs, where the attributes are only categorical variables. In scenarios “SCN4” and “SCN5” we compute two discrete WC designs and compare them with the local benchmarks of Huber and Zwerina (1996), who propose a strategy to compute designs incorporating prior information about the parameters. Specifically, they consider a product whose all attributes take on 3 levels, and the parameters associated with those levels are assumed to be $\overline{\beta}_{HZ} = [-1\ 0\ 1]'$ (for all attributes). The right panel in Table 3.4 presents the WC design obtained in “SCN4” scenario and left panel - the literature benchmark (Table 1 on page 310 Huber and Zwerina 1996). Table 3.5 presents our worst-case solution to the “SCN5” example. For the benchmark we refer the reader to the Table 3 on page 313 in the original article (Huber and Zwerina 1996). Hereinafter, we will refer to the benchmarks as HZ1 and HZ2.

Table 3.4: Computed worst-case design in SCN4 and the benchmark

Original HZ design				WC design			
	Attributes				Attributes		
Set/Alt.	1	2	3	Set/Alt.	1	2	3
1/1	3	1	3	1/1	1	2	3
1/2	2	2	2	1/2	3	3	2
1/3	1	3	1	1/3	2	1	1
2/1	3	1	2	2/1	3	1	3
2/2	2	3	1	2/2	2	3	1
2/3	1	2	3	2/3	1	3	2
3/1	3	2	1	3/1	3	1	1
3/2	2	1	3	3/2	2	3	2
3/3	1	3	2	3/3	1	2	3
4/1	3	1	1	4/1	3	1	3
4/2	1	3	3	4/2	2	3	1
4/3	2	2	2	4/3	1	2	2
5/1	2	1	3	5/1	3	2	1
5/2	3	3	1	5/2	2	3	3
5/3	1	2	2	5/3	1	1	2
6/1	2	3	1	6/1	1	3	3
6/2	3	2	2	6/2	2	2	3
6/3	1	1	3	6/3	1	1	1
7/1	1	3	2	7/1	3	2	2
7/2	3	1	1	7/2	2	3	3
7/3	2	2	3	7/3	1	1	1
8/1	2	3	2	8/1	3	3	3
8/2	3	2	1	8/2	1	2	1
8/3	1	1	3	8/3	2	1	2
9/1	1	2	3	9/1	2	2	1
9/2	3	1	2	9/2	1	1	2
9/3	2	3	1	9/3	3	3	1

Table 3.5: Computed worst-case design in SCN5

	Attributes			
Set/Alt.	1	2	3	4
1/1	1	2	1	2
1/2	3	1	3	3
2/1	2	1	2	1
2/2	3	2	1	3
3/1	3	3	2	1
3/2	1	1	1	3
4/1	3	2	3	2
4/2	1	3	2	1
5/1	3	2	3	1
5/2	2	1	1	2
6/1	1	3	3	2
6/2	3	2	2	3
7/1	1	3	1	3
7/2	2	1	3	2
8/1	3	3	2	3
8/2	1	1	1	1
9/1	2	3	3	2
9/2	3	1	1	1
10/1	3	1	2	2
10/2	2	3	1	3
11/1	2	1	1	2
11/2	1	2	3	1
12/1	2	2	1	1
12/2	1	1	3	3
13/1	1	2	2	2
13/2	2	3	1	1
14/1	3	3	3	1
14/2	2	2	2	3
15/1	2	2	3	1
15/2	3	3	1	2

Figure 3-4: Efficiency comparison of the local and worst-case designs: “HZ” scenarios.

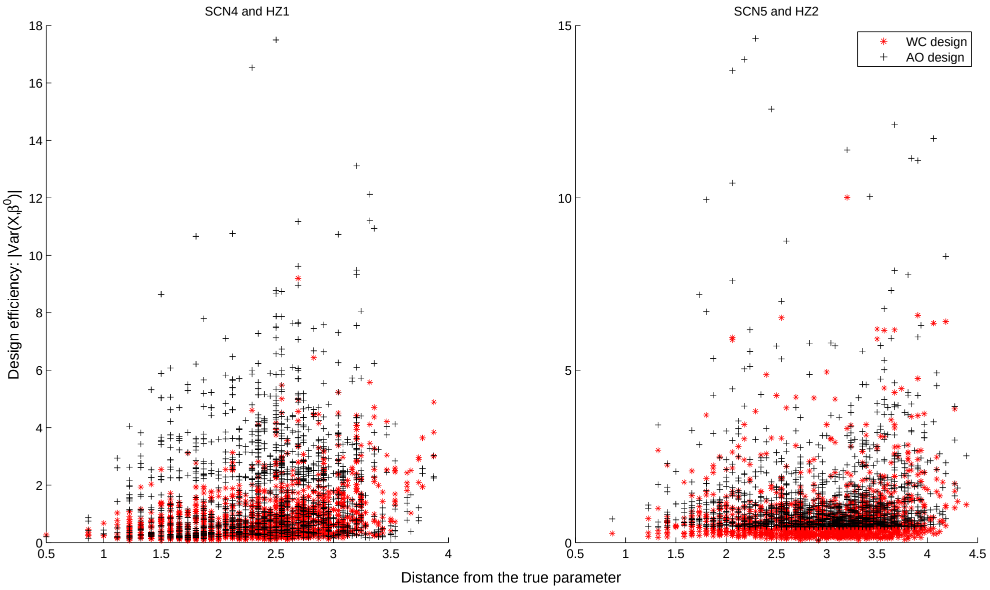
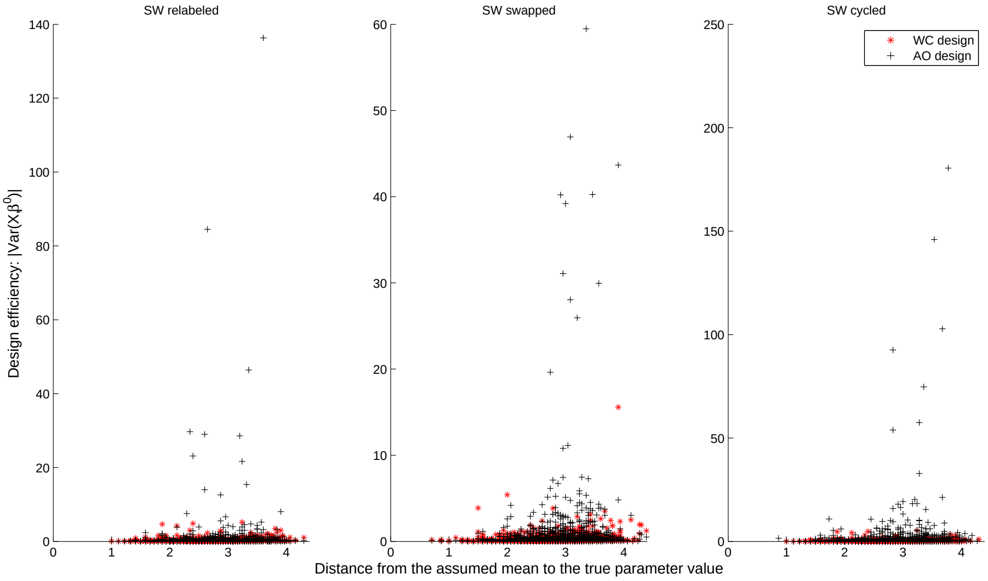


Figure 3-4 compares the efficiency of the WC designs, SCN4 and SCN5, with the local benchmarks, HZ1 and HZ2 (Tables 3.4 and 3.5). The left panel represents the efficiency of designs considered in the “SCN4” and the right panel - “SCN5”. We follow the same procedure as for continuous attributes case, first creating all possible vectors for β^0 with the values $(-1,-0.5,0,0.5,1)$, and then calculating $|V(X_{WC},\beta^0)|$ and $|V(X_{HZ},\beta^0)|$ represented by red stars and black crosses respectively. Then we plot the design efficiencies against the distances to the parameter values assumed by the benchmark, $||\beta^0-\beta_{HZ}||$. Again, the performance of WC designs is very good: the WC approach has much fewer outlying designs with large variance than the corresponding benchmarks, and the designs are efficient both when the true parameters are far and close to the assumed values. This result demonstrates how the assumed interval for β^0 influences the efficiency of WC design. A small interval means less uncertainty about the parameters, yielding as good designs as the benchmarks, which are optimal under the assumed parameter values. When setting a larger interval we account for more uncertainty in the model, therefore the design will be more robust to the misspecification of β^0 and will have good efficiency when the true values are far from the assumed values, but it can yield larger variances than local design when the

Figure 3-5: Efficiency comparison of the average-optimum and worst-case designs: "SW" scenarios.



true values are close to the ones assumed in the local approach.

We can compare the robustness the WC design presented in Table 3.5 (SCN5) and the AO approach benchmarks of Sándor and Wedel (2001, p.435), who consider the same $3^4/2/15$ choice experiment, reporting three Bayesian designs computed using relabeling, swapping and cycling strategies. The benchmarks, hereinafter called SW1-SW3, are constructed under the assumption that the true parameters have a normal probability distribution, with the mean fixed at β_{HZ}, and the square root of covariance matrix $\Sigma_0^{1/2}$ is an identity matrix. Figure 3-5 shows that the WC approach performs very well in comparison to all of three benchmark Bayesian designs. WC approach is very robust to the misspecification of parameters, yielding designs with small variance for different values of the true parameter. The Bayesian benchmarks are not robust and yield many outliers when the true parameter value is far from the assumed one.

We conclude this section with an additional comparison of the WC and the benchmark designs. For each of the compared pair we have simulated all possible vectors of β^0 with values $(-1,-0.5,0,0.5,1)$, mimicking the true parameter values. Then for each of these vectors we have

Table 3.6: Robustness of the worst-case and the benchmark designs

WC design	Benchmark	WC design advantage (%)[a]
SCN1	KAN1	72.00%
SCN2	KAN2	92.16%
SCN3	KAN3	98.86%
SCN4	HZ1	69.55%
SCN5	HZ2	66.68%
SCN5	SW1	64.80%
SCN5	SW2	80.91%
SCN5	SW3	92.60%

[a] (%) simulated parameter values for which the WC design is more efficient than the benchmark.

calculated the efficiency of the WC and the benchmark designs, counting the number of cases when the WC design was more efficient, ie. $|V(X_{WC}, \beta_i|) < |V(X_{BENCH}, \beta_i)|$. The results shown in Table 3.6 confirm the advantage of the WC approach. For every compared pair, the WC design was better than the benchmark at least in 64.80% of the cases. The performance of our designs is very good especially in comparison to all KAN designs, and the “swapped” and “cycled” SW designs. Together with the figures presented in this section, these results confirm our expectations about the robustness of WC approach. It is not always better than the benchmark design, but it performs better in case of misspecified parameters, yielding variance within reasonable bounds, while the benchmarks often produce designs associated with very large variance.

3.5 Concluding Remarks

This chapter presents a general approach to compute efficient exact designs for models in which the covariance matrix depends on unknown parameters. We have focused on conjoint experiments based on discrete-choice models, but the worst-case approach can be applied to many other contexts of interest for CA researchers. There is a variety of problems addressed in CA studies, that require specific estimators with a different covariance matrix. In many cases estimator covariance depends on the unknown parameters and the worst-case methods are capable to approach this type of problems in a robust way. In the concluding section we will mention a few interesting CA problems where WC strategy can be implemented to produce robust designs.

These topics actually open new lines of future work for CA users.

3.5.1 Example I: Classical Model with an Unknown Ideal Point

Additive models with an unknown ideal point are popular examples of nonlinear in parameter utility models, for example the model $u(x,\beta) = \alpha - \sum_{j=1}^{k} \delta_j (x - \gamma_j)^2$, where $\beta^0 = (\alpha, \delta', \gamma')'$ is a vector of unknown parameters. The model can be estimated by Nonlinear Least Squares. Unfortunately, in this case the asymptotic covariance

$$V(X,\beta^0) = \left[\frac{1}{T} \sum_{t=1}^{T} \frac{\partial u(x_t,\beta^0)\, u}{\partial \beta} \frac{\partial u(x_t,\beta^0)'}{\partial \beta} \right]^{-1},$$

depends on the unknown β^0. The worst-case approach can handle this problem.

3.5.2 Example II: Continuous Positive Scale

Consider a CA experiment, where the consumer has a latent utility function $y_t^* = f(x_t)'\beta + \varepsilon_t$, with ε_t Gaussian shocks. The respondents are asked to evaluate product profiles on a continuous positive scale $[0,\infty)$, therefore we observe $y_t = \max\{y_t^*, 0\}$. Due to the non-negativeness truncation, this is a Tobit model, with log-likelihood function

$$\ln L(\beta_1, \sigma_1) = \log(\sigma_1) \sum_{t=1}^{T} I(y_t > 0) + \sum_{t=1}^{T} \frac{(y_t - f(x_t)'\beta_1)^2}{2\sigma_1^2} + \sum_{t=1}^{T} \ln \Phi\left(\frac{f(x_t)'\beta_1}{\sigma_1} \right).$$

We compute the Hessian, and then for the estimation of the information matrix we replace $\sum_{t=1}^{T} I(y_t > 0)$ by $T \cdot E[I(y_t > 0)]$, where $E[I(y_t > 0)] = 1 - \Phi(f(x_t)'\beta_1 / \sigma_1)$. The Tobit ML estimator maximizes the likelihood function. Amemiya (1973) proved its consistency as well as the asymptotic normality, and the covariance matrix depends on the vector of unknown parameters $(\beta', \sigma^2)'$. Worst-case methods can be also applied to this context.

3.5.3 Example III: Interval Regression for Likert-Scale Ratings

Even in the classical additive utility context, non linearity arises when we use discretized ratings instead of measuring preferences on a continuous scale. Standard CA models assume that the respondent's preferences over products are continuous, and given by the latent model $u_t = x_t'\beta^0 + \varepsilon$, where ε_t are independent shocks with zero mean and cumulative distribution $F(\cdot/\sigma)$. However, in practice marketing researchers typically use discrete measurement scales, such as Likert scales, rankings, etc. Therefore, what we actually observe is not the continuous varying u_t, but a censored version of the true underlying preferences. Ordered regression models, introduced by McKelvey and Zavoina (1975) and popularized by McCullagh (1980), can be used to capture the influence of the nonlinear censuring transformation imposed by ordered discrete measurement scales.

If T alternatives are evaluated on a discrete scale with multiple ordered response categories $\{c_k\}_{k=1}^{m}$, we can study the relationship between these discrete measures and the continuous underlying model, using the ordered regression method. Assuming that the respondent allocates a rating $y_t = c_k$, the log likelihood of the CA model is given by

$$L(X,\beta,\sigma) = \sum_{t=1}^{T}\sum_{k=0}^{m+1} y_{tk} \ln\left(F\left(\frac{c_k - x_t'\beta}{\sigma}\right) - F\left(\frac{c_{k-1} - x_t'\beta}{\sigma}\right)\right),$$

where the latent utility u_t falls in the scale interval $(c_{k-1}, c_k]$, with $c_0 = -\infty$ and $c_{m+1} = +\infty$. Additionally, we set $F\left((c_0 - x_t'\beta)/\sigma\right) = 0$ and $F\left((c_{m+1} - x_t'\beta)/\sigma\right) = 1$ for all t. Normality is often assumed, relying on the aggregation of innumerable small influences and the effect of the central limit theorem, but other distributions can be considered (such as the Logistic distribution) as well. Whenever $F(\cdot/\sigma)$ is continuously differentiable,

$$\frac{\partial L(X,\beta,\sigma)}{\partial\beta} = \sum_{t=1}^{T}\left(\sum_{k=0}^{m+1} y_{tk} \frac{\left(f\left(\frac{c_k - x_t'\beta}{\sigma}\right) - f\left(\frac{c_{k-1} - x_t'\beta}{\sigma}\right)\right)}{F\left(\frac{c_k - x_t'\beta}{\sigma}\right) - F\left(\frac{c_{k-1} - x_t'\beta}{\sigma}\right)}\right) x_t.$$

Notice that $E\left[y_{tk}y_{tj}\right] = E\left[y_{tk}^2\right] \times I(k=j) = E\left[y_{tk}\right] \times I(k=j)$, so that

$$I(X,\beta) = E\left[\frac{\partial L(X,\beta,\sigma)}{\partial\beta}\frac{\partial L(X,\beta,\sigma)'}{\partial\beta}\right] = \sum_{t=1}^{T}\sum_{k=0}^{m+1} \frac{\left(f\left(\frac{c_k - x_t'\beta}{\sigma}\right) - f\left(\frac{c_{k-1} - x_t'\beta}{\sigma}\right)\right)^2}{F\left(\frac{c_k - x_t'\beta}{\sigma}\right) - F\left(\frac{c_{k-1} - x_t'\beta}{\sigma}\right)} x_{tj}x_{tl}'.$$

The asymptotic variance-covariance depends on β^0, and we can consider a robust design based on the worst-case method.

3.5.4 Example IV: Correlated Measurements

The presented approach is particularly useful in the context of nonlinear models, but there are some linear utility functions where the structure of the covariances impedes the computation of an optimal design. In such context WC designs may prove to be a sensible approach.

An important situation occurs when ratings are correlated, and the correlation between shocks is driven by the distance between the considered attributes. Consider the model $y_t = f(x_t)' \beta + \varepsilon_t$, where the systematic components ε_t are correlated for products with attributes perceived as similar. In particular, let us assume that

$$\varepsilon_{(T)} = \rho \; W \; \varepsilon_{(T)} + \eta_{(T)}$$

with $|\rho| < 1$, and $E\left[\eta_{(T)}\right] = 0$, $Var\left[\eta_{(T)}\right] = \sigma^2 I$ and where the subindex means that we consider a vector of dimension T. The matrix W depends on the distance between product profiles, meaning that products with similar attributes have correlated measurements. The coefficients $W_{t,s}$ in the matrix W are coded in the form of standardized weights matrices W, with a zero diagonal, and the off-diagonal non-zero elements often scaled to sum to unity in each row, with typical elements:

$$W_{t,s} = \frac{d(x_t, x_s)}{\sum_{l=1}^{T} d(x_t, x_l)}$$

for some discrepancy criteria d, symmetric and satisfying $d(x,x) = 0$. In particular, if d is the square of the Euclidean distance between attributes we will take into account that similar products tend to have correlated utilities measurements. But we can consider other phenomena. We set $W_{t,s} = w_{ts} / \sum_{l=1}^{T} w_{tl}$ where $w_{ts} = 1$ if t,s are consecutive and $w_{ts} = 0$ otherwise. The Generalized Least Squares (GLS) estimator is based on a two stage procedure: first estimate β

by OLS, then estimate ρ with the residuals. In the second step, $\hat{\beta}$ is re-estimated by

$$\hat{\beta} = \left(X' \left[(I - \widehat{\rho} W_x)' (I - \widehat{\rho} W_x)\right]^{-1} X\right)^{-1} X' \left[(I - \widehat{\rho} W_x)' (I - \widehat{\rho} W_x)\right]^{-1} y,$$

where $X = f(x)$. The estimation of the variance of $\hat{\beta}$ is given by $\sigma^2 \left(X' \left[(I - \widehat{\rho} W_x)' (I - \widehat{\rho} W_x)\right]^{-1} X\right)^{-1}$. If we do not have a good preliminary estimation of $|\rho| < 1$, in this case we also should minimize the worst-case design:

$$\min_{x} \max_{|\rho|<1} \phi \left\{ \left(X' \left[(I - \rho W_x)' (I - \rho W_x)\right]^{-1} X\right)^{-1} \right\}$$

where ϕ denotes the trace or determinant.

There are also other situations where this method can be useful, even in classical CA compensatory models. For example, consider a panel where individuals are heterogeneous in the error variance. The model is estimated by pooling the data and using GLS method and the covariance matrix of the estimator will depend on the unknown variance. Therefore worst-case methods can be used to design the experiment.

Bibliography

Abdelbasit, K. M. and Plackett, R. L. (1983). Experimental design for binary data. *Journal of the American Statistical Association*, 78(381):90–98.

Amemiya, T. (1973). Regression analysis when the dependent variable is truncated normal. *Econometrica*, 41(6):997–1016.

Arora, N. and Huber, J. (2001). Improving parameter estimates and model prediction by aggregate customization in choice experiments. *The Journal of Consumer Research*, 28(2):273–283.

Bector, C. and Bhatia, B. (1985). Sufficient optimality conditions and duality for a minmax problem. *Util. Math.*, 27:229–247.

Box, G. E. P. and Lucas, H. L. (1959). Design of experiments in non-linear situations. *Biometrika*, 46(1/2):77–90.

Bunch, D. S., Louviere, J. J., and Anderson, D. (1996). A Comparison of Experimental Design Strategies for Multinomial Logit Models: The Case of Generic Attributes. Technical Report 11, Graduate School of Management, University of California, Davis.

Chernoff, H. (1953). Locally optimal designs for estimating parameters. *The Annals of Mathematical Statistics*, 24(4):586–602.

Demẏanov, V. and Malozemov, V. (1974). Introduction to MINIMAX.

Fedorov, V. V. (1980). Convex design theory. *Mathematische Operationsforschung und Statistik. Series Statistics*, 11(3):403–413.

Ferrini, S. and Scarpa, R. (2007). Designs with a priori information for nonmarket valuation with choice experiments: A monte carlo study. *Journal of Environmental Economics and Management*, 53(3):342 –363.

Green, P. E., Krieger, A. M., and Wind, Y. (2001). Thirty years of conjoint analysis: Reflections and prospects. *Interfaces*, 31(3):S56–S73.

Green, P. E. and Rao, V. R. (1971). Conjoint Measurement for Quantifying Judgmental Data. *Journal of Marketing Research*, 8(3):355–363.

Gustafsson, A., Herrmann, A., and Huber, F. (2007). *Conjoint Measurement: Methods and Applications*. Berlin: Springer Verlag.

Haines, L. M. (1995). A geometric approach to optimal design for one-parameter non-linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(3):575–598.

Huber, J. and Zwerina, K. (1996). The importance of utility balance in efficient choice design. *Journal of Marketing Research*, 33(3):307–317.

Imhof, L. A. (2001). Maximin designs for exponential growth models and heteroscedastic polynomial models. *The Annals of Statistics*, 29(2):561–576.

Jones, D. R. (2001). Direct global optimization algorithm. In Floudas, C. A. and Pardalos, P. M., editors, *Encyclopedia of Optimization*, pages 431–440. Springer US.

Kanninen, B. J. (2002). Optimal design for multinomial choice experiments. *Journal of Marketing Research*, 39:214–227.

Kiefer, J. (1959). Optimum Experimental Designs. *Journal of the Royal Statistical Society. Series B (Methodological)*, 21(2):272–319.

Kiefer, J. and Wolfowitz, J. (1960). The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366.

Kuhfeld, W. F., Tobias, R. D., and Garratt, M. (1994). Efficient Experimental Design with Marketing Research Applications. *Journal of Marketing Research*, 31(4):545–557.

Leyffer, S. (2001). Integrating SQP and Branch-and-Bound for mixed integer nonlinear programming. *Computational Optimization and Applications*, 18:295–309.

Louviere, J. J., Carson, R., and Pihlens, D. (2011). Design of Discrete Choice Experiments: A Discussion of Issues That Matter in Future Applied Research. *Journal of Choice Modelling*, 4(1):1–8.

Louviere, J. J. and Woodworth, G. (1983). Design and analysis of simulated consumer choice or allocation experiments: An approach based on aggregate data. *Journal of Marketing Research*, 20(4):350–367.

Marley, A., Flynn, T. N., and Louviere, J. (2008). Probabilistic models of set-dependent and attribute-level best-worst choice. *Journal of Mathematical Psychology*, 52(5):281–296.

Marley, A. and Louviere, J. (2005). Some probabilistic models of best, worst, and best-worst choices. *Journal of Mathematical Psychology*, 49(6):464–480.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the royal statistical society. Series B (Methodological)*, 42(2):109–142.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In Zarembka, P., editor, *Frontiers in Econometrics*. Academic Press.

McKelvey, R. D. and Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *The Journal of Mathematical Sociology*, 4(1):103–120.

Melas, V. B. (1978). Optimal designs for exponential regression. *Mathematische Operationsforschung und Statistik. Series Statistics*, 9:45–59.

Minkin, S. (1987). Optimal designs for binary data. *Journal of the American Statistical Association*, 82(400):1098–1103.

Müller, C. H. and Pázman, A. (1998). Applications of necessary and sufficient conditions for maximin efficient designs. *Metrika*, 48(1):1–19.

Pronzato, L. and Pázman, A. (2013). *Design of Experiments in Nonlinear Models. Asymptotic Normality, Optimality Criteria and Small-Sample Properties*, volume 212 of *Lecture Notes in Statistics*. Springer.

Pronzato, L. and Walter, E. (1985). Robust experiment design via stochastic approximation. *Mathematical Biosciences*, 75(1):103–120.

Rustem, B. and Howe, M. (2002). *Algorithms for worst case design and applications to risk management*. Princeton University Press, Princeton NJ.

Sándor, Z. and Wedel, M. (2001). Designing conjoint choice experiments using managers' prior beliefs. *Journal of Marketing Research*, 38(4):430–444.

Sándor, Z. and Wedel, M. (2002). Profile construction in experimental choice designs for mixed logit models. *Marketing Science*, 21(4):455–475.

Sándor, Z. and Wedel, M. (2005). Heterogeneous conjoint choice designs. *Journal of Marketing Research*, 42(2):210–218.

Shimizu, K. and Aiyoshi, E. (1980). Necessary conditions for min-max problems and algorithms by a relaxation procedure. *Automatic Control, IEEE Transactions on*, 25(1):62–66.

Sitter, R. R. (1992). Robust designs for binary data. *Biometrics*, 48(4):1145–1155.

Street, D. J., Burgess, L., and Louviere, J. J. (2005). Quick and easy choice sets: constructing optimal and nearly optimal stated choice experiments. *International Journal of Research in Marketing*, 22(4):459–470.

White, L. V. (1973). An extension of the General Equivalence Theorem to nonlinear models. *Biometrika*, 60(2):345–348.

Wittink, D. R. and Cattin, P. (1989). Commercial Use of Conjoint Analysis: An Update. *Journal of Marketing*, 53(3):91–96.

Wittink, D. R., Vriens, M., and Burhenne, W. (1994). Commercial use of conjoint analysis in Europe: Results and critical reflections. *International Journal of Research in Marketing*, 11(1):41–52.

Wong, W.-K. (1992). A unified approach to the construction of minimax designs. *Biometrika*, 79(3):611–619.

Zakovic, S. and Rustem, B. (2003). Semi-infinite programming and applications to minimax problems. *Annals of Operations Research*, 124:81–110.

Chapter 4

Conjoint Analysis with Endogenous Consideration Sets

4.1 Introduction

Conjoint analysis is one of the most widespread tools to study consumers' preferences over multiattribute products and services, implemented through experiments where respondents are asked to rank, rate, or chose certain collections of alternative products to estimate a utility function – usually a compensatory (additive) model. Marketing managers rely on conjoint analysis for relevant processes such as new product development, packaging design, or pricing decisions. But actual consumers' choices are not always consistent with their preferences. They often take quick decisions evaluating a large number of alternative products in categories with large-dimensional multiattribute specifications, which requires significant information search and cognitive efforts.

Rationally bounded consumers often skip potentially interesting options, due to the *lack of information* (brand unawareness), or *perceptual limitations*. For example limited *attention* or low *salience*, which results from the fact that beliefs have different prominence in individual cognition, may lead to overlooking of unnoticed alternatives. Even more relevant is the *halo effect* which prompts the consumers to skip alternatives because their emotions distort the perception of attributes (e.g. consumers might reject attributes from products that they do not like). But it can play a positive role. The halo effect has been exploited by the automotive industry, where an iconic vehicle is used to promote sales of the whole range of vehicles under the same brand (for example the Volkswagen Beetle, the Hummer H1, the Ford GT – inspired by Ford's GT40 racing cars from the 1960s, or the SRT Viper – formerly Dodge Viper). Due to these biases, consumers do not perceive all available options as relevant to their decision problem, and as a result not all of them are evaluated in the same way as considered products. The final decision might seem contradictory with actual preferences if consideration rules are not taken into account. For example, a car attribute can be important in the utility function, but for a particular type of models it can have poor determinance on the decision due to lack of consideration.

Research in consumer behavior established that rationally bounded consumers often use heuristic rules to screen off products for future evaluative consideration, choosing in two stages. In the first step they use consideration set heuristic rules to screen off products whose attributes do not satisfy certain tolerance criteria, often focusing on some key attributes (Bettman 1974; Montgomery and Svenson 1976; Payne 1976; Payne and Ragsdale 1978; Payne et al. 1988, 1993). In the second one they select the best alternative according to their preference order over the considered options. Bettman and Park (1980) suggest that consideration sets are based on specific attributes even though the final selection is holistic. More recently the idea has been adopted in the economic literature, see e.g. Manzini and Mariotti (2013).

Most of the Conjoint Analysis literature focused on the second stage (the estimation of utility functions over considered options). But forecasting consideration sets is a managerially important issue. In the real marketplace consumers do not consider all available options (bear in mind categories with numerous close substitutes like personal care category), but rather focus on handful of alternative offerings (Hauser and Wernerfelt 1990; Urban and Hauser 2004), and for managers it is crucial to forecast if a given offering will pass the consideration threshold. More recently the literature started to look at the heuristic consideration rules, building two-step models (Gensch 1987; Gilbride and Allenby 2004; Jedidi and Kohli 2005; Kohli and Jedidi 2007). However, the literature always considered them as two independent steps: first consideration set is specified, and then the utility function is analyzed conditionally over the considered options. By contrast, in this essay we argue that it is to restrictive to assume such a clear directionality. The halo effect is a clear reason for consideration sets to be endogenous with respect to the overall preferences. If the cognitive process is influenced by the overall affective impression of the product, we cannot assume that the screening-off stage is independent from the evaluative step. Rationally bounded consumers might choose in two-stages, but they are mutually linked. The main contribution of this chapter is the joint estimation of consideration sets and preferences allowing for endogeneity between preferences and consideration sets.

4.2 Modeling Preferences with Consideration Thresholds

Consider a multiattribute product characterized by attributes vector x (either continuous, discrete, or both) varying in a compact subset $\chi \subset \mathbb{R}^k$, and let y_t be the evaluative response to x_t. In classical experiments y_t is an observable variable (ratings or rankings), whereas in choice experiments y is a latent variable and we simply observe the product with higher utility among a small set of alternatives. By pedagogical reasons we will discuss choice models later, and in the first part of the essay we will assume that y_t is observed.

Whether or not consumers select a product depends on a screening-off consideration rule, traditionally considered as a primitive of the decision procedure, and overall preferences are conditioned by this decision. Therefore, we postulate a switching-preference model where for each $x_t \in \chi$ we observe individual preference ratings y_t satisfying

$$y_t = \begin{cases} f(x_t)'\beta + \varepsilon_{1t} & x_t \in A(\gamma, u_t) \\ \alpha + \varepsilon_{2t} & x_t \notin A(\gamma, u_t) \end{cases} \tag{4.1}$$

where $\{(\varepsilon_{1t}, \varepsilon_{2t})\}$ are i.i.d. jointly distributed with zero mean and finite variances. Typically ε_{2t} has low variability and α is a small parameter. We can actually replace $f(x_t)'\beta$ by a more general nonlinear and twice continuously differentiable function $f(x_t, \beta)$. The consideration set $A(\gamma, u_t) \subset \chi$ depends on unknown parameter vector γ, and some random vector u_t. Notice that if $\Pr(u_t = 0) = 1$, we obtain a deterministic consideration set but this is a quite restrictive assumption. Consideration rules can be formulated as deterministic, but a stochastic approach is generally more fruitful. There are situational factors that can affect the final consideration decision of a given product. Bettman and Zins (1977) find out evidence that consumers build their consideration rules on-the-spot using memory fragments and situational elements. Excitement and attention might also affect consideration.

The marketing literature has considered a variety of consideration rules $A\left(\gamma, u_t\right)$, see e.g. Gilbride and Allenby (2004, 2006), Jedidi and Kohli (2005); Montgomery and Svenson (1976); Ordóñez et al. (1999); Payne et al. (1988); Olshavsky and Acito (1980); Bröder (2000); Yee et al. (2007) and Hauser et al. (2009). The most common specifications are:

- *Disjunctive rule* requiring that at least one attribute is above a threshold level. Typically it can be expressed as $A\left(\gamma,u\right)=\bigcup_{j=1}^{k}\left\{x\in\chi : x_j \geq \gamma_j + u_j\right\}$. But we can consider lower or upper bounds independently for every attribute.

- *Conjunctive rule*, assuming that all attributes of a considered profile exceed minimum threshold levels, so that $A\left(\gamma,u\right)=\bigcap_{j=1}^{k}\left\{x\in\chi : x_j \geq \gamma_j + u_j\right\}$. *Subset conjunctive rule* is a variation where a profile must have $r \leq j$ attributes above a threshold. In practice consumers can implement conjunctive rules sequentially (eliminating profiles using a given feature, and then moving to another attribute, this is known as the *Elimination-by-Aspects* process).

- *Compensatory or additive partworth rule*, where a combination of different attributes is above a threshold level, so that we can express $A\left(\gamma,u\right)=\left\{x\in\chi : x'\gamma_1 \geq \gamma_0 + u\right\}$. It can be applied when consumers screen-off products with holistic utility below some threshold. In this case the consideration function should be $A\left(\gamma,\varepsilon\right)=\left\{x\in\chi : f\left(x_t\right)'\beta+\varepsilon_1 \geq \gamma_0\right\}$ which is endogenous with respect to the preference shocks $(\varepsilon_1,\varepsilon_2)$, and the parameters in γ include β. We will argue later that there are additional reasons for endogeneity.

- *Lexicographic rule*. In a lexicographic consideration set, a profile is eligible if it satisfies a threshold level for the most important attribute $\left\{x_1 \geq \gamma_1 + u_1\right\}$, and if it doesn't, it should satisfy a similar threshold for the second attribute, and so on. This rule can be expressed as $A\left(\gamma,u\right)=\bigcup_{j=1}^{k}\left\{\{x_j \geq \gamma_j + u_j\}\bigcap_{l=1}^{j-1}\{x_l < \gamma_l + u_l\}\right\}$, where the order given to the attributes is essential to define the rule. This concept is more workable than lexicographic preferences[1], but still is relatively complex (Jedidi and Kohli 2005; Kohli and Jedidi 2007). A related method, combining unions and intersections, is a *disjunction of conjunctions* where one or more conjunctions are satisfied.

- *Single feature rule.* Sometimes the consideration set can be simpler, for example considering a single attribute $A\left(\gamma,u\right)=\left\{x_l \leq \gamma_l + u\right\}$. A relevant case is to consider the price, then such consideration set essentially imposes a reservation price.

The evaluative function $f\left(x_t\right)'\beta$ can depend on different set of attributes from those included in the consideration rule. In theory, γ can have elements in common with β, and also $\alpha = 0$. We will include all final parameters in a vector $\theta \in \Theta$, a compact set in the Euclidean space. Typically, to ensure identification of the parameters in the consideration set we need to assume that $Var(u) = 1$ and $E[u] = 0$ (or any other pre-fixed value, otherwise we can re-scale the consideration rule using an affine transformation).

More involved situations can also be explained within the proposed framework. For example, we can substitute $\alpha + \varepsilon_{2t}$ with a more complex model. A typical example is the conjoint model with a reference price r when the effect of losses is larger than that of gains, meaning that

$$y_t = \begin{cases} f\left(x_t\right)'\beta - \gamma_{gain}\cdot|p_t - r| + \varepsilon_{1t}, & p_t \leq r \\ f\left(x_t\right)'\beta - \gamma_{loss}\cdot|p_t - r| + \varepsilon_{2t}, & p_t > r \end{cases} \tag{4.2}$$

where $\gamma_{gain} < \gamma_{loss}$. In this case, $\{p_t \leq r\}$ is not exactly a consideration set, but it would have a similar interpretation if γ_{loss} is large. The model can be analyzed applying the methods discussed below, although we do not emphasize this specific case in our empirical analysis.

[1] If χ is uncountable, lexicographic preferences cannot be represented by a utility function.

4.2.1 Consideration Frequency

The presented model implies that overall preferences work in two steps, which can be characterized using conditional probabilities. Define a dummy consideration variable

$$C_t = I\left(x_t \in A\left(\gamma, u_t\right)\right),$$

where $I(\cdot)$ denotes the indicator function (equal to 1 when $x_t \in A\left(\gamma, u_t\right)$ and zero otherwise). In general we assume that the shocks u_t are i.i.d. with marginal density function $f_u(u)$. Then the probability of a product x_t being considered is precisely

$$\Pr(C_t = 1|x_t) = \Pr\left(x_t \in A\left(\gamma, u_t\right)|x_t\right) = \int_{A_{x_t}(\gamma)} f_u(u)\,du,$$

where $A_x\left(\gamma\right)$ is the set with all possible realizations of u for which x is considered

$$A_x\left(\gamma\right) = \left\{u : x_t \in A\left(\gamma, u\right)\right\}.$$

When u belongs to the complementary set $A_x\left(\gamma\right)^c = \left\{u : x_t \notin A\left(\gamma, u\right)\right\}$ the product x is discarded. For example, if $A\left(\gamma, u_t\right) = \left\{x \in \chi : x'\gamma \geq u\right\}$ then $A_x\left(\gamma\right) = \left(-\infty, x'\gamma\right]$. If $A\left(\gamma, u\right) = \left\{(x_1, x_2) : x_1'\gamma_1 \geq u_1,\right.$ $\left.x_2 \leq \gamma_2 + u_2\right\}$ then $A_x\left(\gamma\right) = \left(-\infty, x_1'\gamma_1\right] \times \left[x_2 - \gamma_2, \infty\right)$. Notice that in all the examples of consideration sets, the associated set $A_x\left(\gamma\right)$ is either a finite Cartesian product of semi-intervals (conjunctive rules), or a finite union of several of such sets (disjunctive rules). In the last case, we can decompose $A_x\left(\gamma\right)$ in non-overlapping rectangles, and integrate separately one each one of them. Notice that for the compensatory-consideration example, depending on the specification for $f(u)$ we can obtain different standard models such as Probits or Logits. For example, Ben-Akiva and Boccara (1995) specify a logit model for the consideration set probability with self-explicated consideration data $C_t = 1$. If the consideration set and the density function are more involved, we can use Monte Carlo simulations to compute these probabilities. Kau and Hill (1972) and Bettman (1974) already emphasized the importance of consideration probabilities $\Pr(C_t = 1|x_t)$.

Now we can examine the probabilistic structure of the problem. If we assume that ε_{it} are i.i.d with marginal distribution $f_i(\cdot)$, for $i=1,2$ then we can derive the joint density function for the evaluations and considerations as

$$\begin{aligned} f_{y,C}(y_t,C_t|x_t) &= \left(f_1\left(y_t-f(x_t)'\beta\right)\times\Pr(C_t=1|y_t,x_t)\right)^{C_t} \\ &\quad +(f_2(y_t-\alpha)\times(1-\Pr(C_t=1|y_t,x_t)))^{1-C_t}, \end{aligned} \tag{4.3}$$

If the evaluative shocks $(\varepsilon_{1t},\varepsilon_{2t})$ are independent from the consideration shocks u_t, then

$$\Pr(C_t=1|y_t,x_t)=\Pr(C_t=1|x_t). \tag{4.4}$$

It is important to mention that when u_t is statistically dependent with respect to $(\varepsilon_{1t},\varepsilon_{2t})$ the problem is much more difficult to handle. The consideration set literature traditionally considers exogenous consideration, $\alpha=0$ and that $f_{\varepsilon_2}(\cdot)$ is concentrated in this value, so that the density $f_{y,C}(y_t,C_t|x_t)$ collapses to

$$f_y(y_t|x_t)=f_1\left(y_t-f(x_t)'\beta\right)\times\Pr(C_t=1|x_t)$$

Then, if C_t is known, we can estimate the model by Maximum Likelihood, or a Bayesian method.

There is another way to write model (4.1), namely as a regression equation

$$\begin{aligned} y_t &= C_t\, f(x_t)'\beta+(1-C_t)\ \alpha+\eta_t \\ \eta_t &= C_t\ \varepsilon_{1t}+(1-C_t)\ \varepsilon_{2t}, \end{aligned} \tag{4.5}$$

where $C_t=I\left(x_t\in A\left(\gamma,u_t\right)\right)$ can be observed or not. If u_t is independent of $(\varepsilon_{1t},\varepsilon_{2t})$ then

$$E\left[\eta_t|x_t\right]=E\left[C_t|x_t\right]\times E\left[\varepsilon_{1t}\right]+E\left[(1-C_t)|x_t\right]\times E\left[\varepsilon_{2t}\right]=0,$$

with $E[C_t|x_t]=\Pr(C_t=1|x_t)$ and $E[(1-C_t)|x_t]=(1-\Pr(C_t=1|x_t))$ and the model can be estimated using classical econometric tools for exogenous switching regression based on the ideas in

(4.3) and (4.4). But if u_t is statistically dependent with respect to $(\varepsilon_{1t}, \varepsilon_{2t})$ we need a different approach.

Practitioners can face additional difficulties, as C_t is not always available. Jedidi et al. (1996) and Roberts and Lattin (1991) use respondent "self-explicated" survey measures about consideration of alternatives (we observe $C_t = 1$). But notice that the consideration decisions can be conscious or unconscious. In the last case, respondents would not be able to declare whether $C_t = 1$ or $C_t = 0$, and when asked directly they would be likely to make mistakes. For example, if a considered product receives low evaluation, respondents could mistakenly declare it as unconsidered. In such case consideration statements should be regarded as a noisy signal of consideration. Hauser et al. (2009, p.15) provide a thorough discussion about the limitations of Self-Explicated consideration variables. There are other contexts where one can find noisy signals of consideration. For example, researchers might estimate consideration using noisy signals such as previous brand purchase, or previous purchases of brands in promotion (see Fader and McAlister 1990).

If we consider C_t as an unobservable latent variable, and we consider some signals S_t related to consideration, but not entirely correct , the relationship between signals and consideration can be modeled using a matrix of parameters represented in Table 4.1, where $\pi_{ij} = \Pr(S_t = i | C_t = j)$, for $j, i = 0, 1$, and sum one by rows $(\pi_{11} + \pi_{10} = \pi_{01} + \pi_{00} = 1)$.

Table 4.1: Relationship between signals and consideration

	$S_t = 1$	$S_t = 0$
$C_t = 1$	π_{11}	$\pi_{10} = 1 - \pi_{11}$
$C_t = 0$	π_{01}	$\pi_{00} = 1 - \pi_{01}$

If we assume that the errors are independent from the specific attributes, then we can compute

$$\begin{aligned}
\Pr(S_t = 1 | x_t) &= \Pr(S_t = 1 | C_t = 1) \times \Pr(C_t = 1 | x_t) + \Pr(S_t = 1 | C_t = 0) \times (1 - \Pr(C_t = 1 | x_t)) \\
&= \pi_{11} \times \Pr(C_t = 1 | x_t) + \pi_{01} \times (1 - \Pr(C_t = 1 | x_t)) \\
\Pr(S_t = 0 | x_t) &= (1 - \pi_{11}) \times \Pr(C_t = 1 | x_t) + (1 - \pi_{01}) \times (1 - \Pr(C_t = 1 | x_t)) \qquad (4.6)
\end{aligned}$$

Lee and Porter (1984) considered related ideas in another context.

Appendix A provides the expressions for the Likelihood function for the different cases discussed above: observed C_t, observed signals S_t, and the case where nothing is observed. But any of these techniques require exogeneity of the consideration set.

4.3 Endogenous Consideration Sets

In practice it is difficult to accept that consideration sets are exogenous. One of the main reasons is the *halo effect* which may interfere with the perception of attributes. In particular we distinguish between the reverse-halo, with consumers either biasing or ignoring information about alternatives contradicting their affective attitude, and the direct-halo where consumers consider products with unacceptable attributes due to affective reasons. The phenomenon has a long history in marketing. Beckwith and Lehmann (1975, 1976) found the evidence of halo effect in multiattribute attitude models. Johansson et al. (1976) and Beckwith et al. (1978) discuss the phenomenon.

If the consideration set $A\left(\gamma,u\right)$ is endogenously selected, then we cannot assume that u_t is statistically independent of $(\varepsilon_{1t},\varepsilon_{2t})$, meaning that $C_t = I\left(x \in A\left(\gamma,u_t\right)\right)$ is endogenous with respect to these shocks. The classical methods render inconsistent estimations in models with endogenous consideration sets. Now, the shock of the regression model (4.5) satisfies

$$E\left[\eta_t|x_t\right] = \Pr(C_t = 1|x_t)\cdot E\left[\varepsilon_{1t}|x_t,C_t = 1\right] + (1-\Pr(C_t = 1|x_t))\cdot E\left[\varepsilon_{2t}|x_t,C_t = 0\right],$$

which is in general different from zero. Ignoring this type of endogeneity will lead to inconsistent estimations, and a biased perspective on consumer preference formulations. We can actually quantify the bias component $E\left[\varepsilon_{it}|x_t,C_t = i\right]$ for $i = 1,2$. If we denote by f_u the marginal density function of u,

$$E\left[\varepsilon_{1t}|x_t,C_t = 1\right] = E\left[E\left[\varepsilon_{t1}|x_t,u_t\right]\ |x_t \in A\left(\gamma,u_t\right)\right] = \int_{A_x(\gamma)} E\left(\varepsilon_{t1}|u\right)\frac{f_u(u)}{\Pr_F\left\{A_x\left(\gamma\right)\right\}}du$$

and similarly for $E[\varepsilon_{2t}|x_t, C_t = 0]$ replacing $A_x(\gamma)$ by its complement $A_x(\gamma)^c$.

To provide a clearer insight, let us consider a relatively standard example. If $(\varepsilon_{t1}, \varepsilon_{t2}, u_t)$ are normally distributed with zero mean and covariance matrix $\Sigma = \{\sigma_{ij}\}$, then for $i = 1,2$

$$\begin{aligned} \varepsilon_i|u &\sim N\left(0 + \frac{\sigma_{iu}}{\sigma_u^2}(u-0), \sigma_2^2 - \frac{\sigma_{2u}^2}{\sigma_u^2}\right) \\ E(\varepsilon_i|u) &= \frac{\sigma_{iu}}{\sigma_u^2} u \end{aligned}$$

where $\sigma_{iu}^2 = E[\varepsilon_{ti} u_t]$, meaning that $E(\varepsilon_i|u) = \frac{\sigma_{iu}}{\sigma_u^2} u$. To compute the expected value of ε_{it} conditionally on $C_t = 1$ we now need to consider only the type of screening-off rule.

4.3.1 Compensatory Consideration Setting

Consider $A(\gamma, u) = \{x : x'\gamma - u > 0\}$, with endogenous normal shocks, and we normalize $\sigma_u^2 = E[u_t^2] = 1$ to ensure that the regime parameters are identified. From the distribution of the shocks $(\varepsilon_{t1}, \varepsilon_{t2}, u_t)$ using the expressions that we computed for $E(\varepsilon_{t1}|u), E(\varepsilon_{t2}|u)$, we obtain that

$$\begin{aligned} E[\varepsilon_{t1}|x_t, u_t \leq x_t'\gamma] &= \int_{-\infty}^{x_t'\gamma} E(\varepsilon_{t1}|u) \frac{\phi(u)}{\Phi(x_t'\gamma)} du = \sigma_{1u} \int_{-\infty}^{x_t'\gamma} u \frac{\phi(u)}{\Phi(x_t'\gamma)} du \\ E[\varepsilon_{t2}|x_t, u_t > x_t'\gamma] &= \int_{x_t'\gamma}^{\infty} E(\varepsilon_{t2}|u) \frac{\phi(u)}{1-\Phi(x_t'\gamma)} du = \sigma_{2u} \int_{x_t'\gamma}^{\infty} u \frac{\phi(u)}{1-\Phi(x_t'\gamma)} du \end{aligned}$$

We will use also that for a standard normal density $\phi(\cdot)$ with cumulative distribution $\Phi(\cdot)$,

$$\begin{aligned} \frac{1}{\Phi(z)} \int_{-\infty}^{z} u\phi(u) du &= \frac{-\phi(z)}{\Phi(z)} = -\lambda(z) \\ \frac{1}{1-\Phi(z)} \int_{z}^{\infty} u\phi(u) du &= \frac{\phi(z)}{1-\Phi(z)} = \frac{\phi(z)}{\Phi(-z)} = \lambda(-z) \end{aligned}$$

where $\lambda(z) = \phi(z)/\Phi(z)$ is known as the inverse Mills ratio or Heckman's lambda. Therefore,

$$\begin{aligned} E[\varepsilon_{t1}|x_t, u_t \leq x_t'\gamma] &= -\sigma_{1u} \frac{\phi(x_t'\gamma)}{\Phi(x_t'\gamma)} = -\sigma_{1u}\ \lambda(x_t'\gamma) \\ E[\varepsilon_{t2}|x_t, u_t > x_t'\gamma] &= \sigma_{2u} \frac{\phi(x_t'\gamma)}{1-\Phi(x_t'\gamma)} = \sigma_{2u}\ \lambda(-x_t'\gamma) \end{aligned}$$

Clearly, the conditional expectation of the observed utility is given by

$$E\left[y_t|x_t,C_t=1\right] \;=\; f(x_t)'\beta-\sigma_{1u}\,\lambda\left(x_t'\gamma\right), \tag{4.7}$$

$$E\left[y_t|x_t,C_t=0\right] \;=\; \alpha+\sigma_{2u}\,\lambda\left(-x_t'\gamma\right), \tag{4.8}$$

Actually, if we know γ the preference parameters for each regime can be estimated consistently using these regression equations.

$$\Pr(C_t=1|x_t)=\int_{A_x(\gamma)} f_u(u)\,du=\Phi\left(-x_t'\gamma\right). \tag{4.9}$$

If C_t is known, for this specific switching-regression problem, Heckman (1979) proposed a two-step procedure, estimating γ from the probit (4.9), and in a second step estimate the remainder parameters applying OLS in the regression models (4.7) and (4.8) for each regime, using $\widehat{\lambda}_{t1}=\lambda\left(x_t'\widehat{\gamma}\right)$ and $\widehat{\lambda}_{t2}=\lambda\left(-x_t'\widehat{\gamma}\right)$ as regressors. Naturally, such estimation is less efficient than Maximum Likelihood.

For other types of consideration heuristic rules, endogeneity can generate even more complex biases. Notice that the expressions (4.7), (4.8) and (4.9), can be extended to consideration sets with several inequalities by considering a vector u_t, replacing the σ_{iu},σ_u^2 parameters by matrices, and taking multiple integrals instead of a simple integral. The situation can be far more complex when we do not know C_t, and we can also have intermediate situations where we observe some signal S_t (such as past brand purchase) positively dependent with respect to the unobserved C_t. The next section presents a simple empirical application based on the classical ideas by Heckman (1979), and the last part of this chapter is focused on the general model.

4.4 Empirical Application

We have conducted an experiment to test the endogeneity of consideration sets using an online survey with Amazon.com's Mechanical Turks (MTurks). This is a conjoint study of consumer preferences towards lunch dishes, providing a realistic setting where people screen off numerous available options and proceed with a choice from the reduced set of alternatives. At this stage we consider a model of compensatory additive consideration set that can be framed in the Heckman (1979) approach explained in Section 3.1. To this end we build a Heckman two-step estimator, and use it as a starting point for a Newton algorithm in a full maximum likelihood estimator.

4.4.1 Data Description

Conjoint data were collected online via Amazon's MTurk website, which allows cheap and fast recruitment of workers from a diverse subject pool. Berinsky et al. (2012) compare MTurks to other Internet and traditional samples, showing that in terms of representativeness and data quality they exceed the standards of published research. To recruit the subjects we have posted a job task (a so called HIT - Human Intelligence Task) with the survey link paying each respondent 95 cents for the completed 10-minute survey. We required that the MTurk workers have a history of at least 100 approved HITs and a 95% approval rate from all requesters. The survey was active for 4 weeks and a total of 3008 conjoint observations were collected from a sample of 188 respondents.

Table 4.2: Lunch entrée attributes and levels

Attributes	Levels
Type of meat	beef, chicken, fish, vegetarian
Price	$ 7.99, $ 11.99, $ 15.99
Preparation	grilled, fried, broiled
Sides	fries, rice, vegetables, salad
Gluten-free	yes, no
Organic	yes, no

The conjoint task included in the survey was designed to estimate respondents' preferences for lunch entrée features. The identified product were: type of meat, price, type of preparation, included sides, gluten-free and organic claims (see Table 4.2). Each subject was asked to evaluate the appeal of 16 different lunch options on a scale from 1 to 10, and the presented alternatives constituted an orthogonal fractional factorial design. We have also included self-explicated questions about respondent's choice and consideration set, asking directly which menu items they would order and were considering to order. Additionally, we collected the information about demographics, dietary issues, eating-out habits and fitness activity. Summarizing, the data set contains four types of variables: (1) lunch entrée characteristics (the experimental design described in Table 4.2), (2) demographic variables, (3) eating/fitness habits and food allergies, and (4) meat category ranking (ranking of the levels in the first attribute). Except for age, height, weight and BMI all variables are dummies.

Table 4.3: Overall evaluation of presented alternatives

ID and Description	Utilities (mean)	Choice (#)	Set (#)
1. Pan fried beef stripes with rice at $15.99. *O*[a]	5.58	6	33
2. Tenderloin beef roast with salad at $ 11.99. *GF,O*	6.67	10	54
3. Grilled steak with vegetables at $ 7.99. *GF*	7.20	31	74
4. Tenderloin beef roast with fries at $ 11.99.	6.47	19	38
5. Grilled chicken breast with fries at $ 11.99. *O*	6.97	24	74
6. Boneless buffalo chicken wings with rice at $ 7.99	6.00	11	41
7. Herb crusted fried chicken with salad at $ 11.99. *GF*	6.92	19	51
8. Boneless buffalo chicken wings with vegetables at $ 15.99. *GF,O*	5.60	6	21
9. Oven-baked grouper with salad at $ 7.99.	5.11	2	25
10. Grilled salmon with rice at $ 15.99. *GF*	5.44	17	46
11. Oven-baked grouper with fries at $ 15.99. *GF,O*	4.56	5	13
12. Fried shrimps platter with vegetables at $ 11.99. *O*	5.52	12	40
13. Fried vegetarian burger with fries at $ 7.99. *GF*	4.73	6	29
14. Vegetarian lasagna with vegetables at $ 15.99	5.45	8	34
15. Seasonal grilled veggies platter with salad at $ 15.99. *O*	5.34	5	26
16. Vegetarian lasagna with rice at $ 11.99. *GF,O*	5.18	7	27
Total:	5.80	188	626

[a] *O* – organic meal, *GF* – gluten-free meal

The presented lunch options are listed in Table 4.3, together with their corresponding mean utilities, the number of respondents who chose them, and the number of respondents who considered ordering the dish for lunch. The first look at the data suggests that in general the respondents are consistent in reporting their utilities, choices and consideration sets. Options 3, 5 and 7 are the top three lunch entrées in terms of all criteria: average appeal (utility score), number of choices, and number of considerations. Additionally, option 2 was also among the mostly considered dishes. On average, each respondent reported 3.33 dishes in their self-explicated consideration set, the average utility score of alternatives included in the consideration set is 8.5 (across alternatives and respondents), and it is 5.1 for unconsidered options. This informal empirical evidence suggests that indeed respondents evaluate differently the considered and unconsidered dishes and points to the possibility of a certain screening heuristic.

Additionally, we collected background information about respondents' dietary issues, eating out frequency and spending, as well as fitness activity habits. Interestingly, almost 30% of the respondents reported some diet restrictions, of which most frequent are low calorie diet (14 respondents), vegetarian diet (11), gluten allergy (7) and lactose allergy (6). These issues may have a significant impact on individual's preferences for the lunch entrée. Over 65% of the respondents eat out at least once per week, which indicates that the sample is adequate for the study. In terms of the average lunch spending the majority of subjects are concentrated in two categories: 56% of respondents pays for lunch an amount between $5.00 and $9.99, and 37% – between $10.00 to $14.99.

A number of questions allow us to assess the physical condition and fitness level of the respondents. Table 4.4 presents descriptive statistics for demographic variables in our data, including height age, gender, height, weight, and BMI[2]. The distribution of BMI values presented in Table 4.4 shows clearly that more than the half of the respondents are above the healthy weight threshold, including around 25% of respondents who are obese. In comparison, the prevalence of obesity in adult American population was 35.7% in 2012 (Ogden et al. 2012). Fitness activity level of respondents is rather low: 44% of them admitted to exercising less than once a week; around 33% of respondents exercises moderately (2-3 times per week), and only around 23% exercises at least 4 times weekly.

For the analysis, we use dummy coding of design variables, omitting one level within each attribute to eliminate multicollinearity: vegetarian for the meat-type attribute, $5.99 in the price attribute, fried for the preparation attribute, and rice for the side attribute. As a result, the estimated coefficients (part-worths) are interpreted relative to the omitted variable. The rest of categorical variables are recoded into the variables "excerciseY"/"eatoutY", indicating whether

[2] The Body Mass Index (BMI) is calculated as weight(kg)/height^2(m^2) $\approx$ 703 × weight(lb)/height^2(in^2), and is commonly used for weight assessment in adults, indicating: underweight (BMI< 18.5), healthy weight (18.5 < BMI < 24.9), overweight (25.0 < BMI < 29.9), obesity (BMI> 30).

Table 4.4: Demographic variables overview

Variable	Mean	SD	Min	Max	Q_1	Q_2	Q_3
Height (in)	66.88	4.78	59	86	64	66	69
Weight (lbs)	172.84	46.27	90	350	144	160	195
BMI	27.29	7.20	16.82	56.64	22.04	25.74	30.85
Age	37.52	11.62	20	71	29	35	44
Gender: female	60%						

the respondent exercises/eats out at least once per week, and “AvgPrice9.99” representing the average lunch spending up to $9.99. “Onlineorder” indicates whether or not the respondent ordered food online before, and “catrnk2” takes the value 1 if the lunch option belongs to respondents’ two favourite dish categories (and 0 otherwise).

4.4.2 Model Estimation

We have estimated the endogenous consideration set with compensatory selection rule described in Section 3.1 using Heckman two-step estimator implemented in Stata subroutine “heckman”, which provides consistent, asymptotically efficient estimates for all the parameters in the model. Note that the variables defining the consideration set and the utility need not be the same, otherwise it is implicitly assumed that there is no specific phenomenon determining whether or not an observation is considered.

In order to define the model, we considered a variety of alternative specifications with different types of variables and interactions for the consideration set and the utility equation. We performed diagnostic analysis of estimated models and tested the individual and overall significance of coefficients, obtaining the following specification. In the compensatory consideration set we included the design variables, “catrnk2”variable, and two meat-price interaction effects: “beef-$11.99”and “chicken-$11.99”. These interaction effects correspond to two most frequently considered lunch options on our menu. The utility evaluation equation is defined by the design and individual-specific variables: demographic variables, BMI, frequency of eating out and exercising, average lunch spending, and diet restrictions. In summary, the probability that a dish is included in the consideration set depends only on dish attributes and not on respondent-specific variables, but the way people evaluate lunch options depends on both respondent-specific effects and food characteristics.

Table 4.5 summarizes the results of the Heckman two-step procedure, where in the first stage we estimate the probability that a dish is considered by the respondent (probit), and in the second stage we estimate the utility of considered and unconsidered lunch options including the inverse Mills ratio variables to account for endogeneity (see equations 4.7 and 4.8). Each of the resulting regressions is globally statistically significant at the .000 level. Additionally, the standard residual errors in the "Utility" equation exhibit low variability: $\hat{\sigma} = 1.332$ in the first regime, and $\hat{\sigma} = 2.557$ in the second regime.

The results from the probit model are summarized in the left panel of Table 4.5, demonstrating that the majority of the estimates is significant at .01 level, and the insignificant coefficients are attributed to the smallest effects. The interaction elements between meat category and the second price level have the largest positive effect of a dish being considered, and are responsible for the increase in z-score[3] by 1.716 (chicken-$11.99 interaction term) and 1.505 (beef-$11.99 interaction term), relative to the omitted vegetarian category. Interestingly, they are larger than the main effects (see negative coefficients for prices, chicken, and beef), which suggests that respondents are more likely to consider both of the features simultaneously rather than separately. Furthermore a considerable positive estimate of a category ranking indicates that respondents more often consider dishes, which belong to one of the two of their favourite meat categories. Also, preparation of food on grill, and organic claims are factors which increase the probability of consideration. Finally, second and third price level decreases the probability that a dish will be included in the consideration set (relative to the lowest price), as well as the accompanying sides – salad and fries (relative to rice). The negative coefficients on beef and chicken dishes are relative to the omitted vegetarian category and represent an isolated main effect, separated from the influence of interaction terms and category rankings. The center and right panel of 4.5 present the results of the utility estimation of considered (regime 1) and unconsidered dishes (regime 2), receptively. Among design variables, beef dishes and two price levels contribute to

[3] Omitted in the output for space purposes.

Table 4.5: Heckman's two step estimation results

	Consideration		Utility		Utility	
	probability		Regime 1		Regime 2	
	β_j	SE	β_j	SE	β_j	SE
beef	-.774^c	.195	.447^b	.179		
chicken	-1.017^c	.199	.201	.181		
fish	-.258^c	.09	-.051	.181		
price \$11.99	-.933^c	.303	.3^a	.175		
price \$15.99	-.725^c	.15	.32^a	.17		
grilled	.254^c	.094	-.134	.164		
broiled	-.162	.1	-.051	.14		
salad	-.714^c	.184	.121	.166		
vegetables	.076	.082	.001	.157		
fries	-.825^c	.19	-.133	.178		
glutenY	.017	.059	.103	.116		
organic	.214^a	.12	-.386^c	.136		
catrnk2	.94^c	.067	-	-		
beef-\$11.99	1.505^c	.423	-	-		
chicken-\$11.99	1.716^c	.426	-	-		
female			.172	.113		
BMI			-.013^a	.007		
age			.03^c	.005		
onlineorder			-.028	.109		
eatoutY			.048	.117		
excerciseY			-.128	.117		
AvgPr9.99			-.239^b	.118		
AllGluten			-.973^c	.343		
AllLactose			-.637^a	.375		
AllFish			.827	1.279		
LowCalDiet			-.152	.349		
LowSodDiet			-.512	.334		
VegeDiet			.004	.388		
AllDiet0			-.31	.32		
AllDietOther			.255	.295		
Mills ratio			-.654^c	.196	3.551^c	.269
Intercept	-.400^a	.222	8.776^c	.559	3.929^c	.101
$\hat{\rho}$	-0.491					
σ				1.332	2.557	

^{a} $p < .1$, ^{b} $p < .05$, ^{c} $p < .01$

the increase in the utility score of a dish, while the organic claims reduce it. The majority of significant individual-specific estimates are negative, showing that respondents who are allergic to gluten and lactose evaluate the presented menu options less favourably than individuals who do not suffer from above allergies. The influence of age and BMI on appeal evaluation is small but significant. Finally, the disparity of intercept values for the utility equations in two regimes provides empirical evidence that individuals indeed evaluate the considered and overlooked lunch options in a different way.

A convenient advantage of the consideration set model is the ability to separate the effects on consideration probability and utility evaluation, because certain dish attributes may be important for the consideration step but not for the evaluative step (and vice versa). Observe an interesting behaviour of price variables: the two highest price levels have large, negative and significant estimates in the consideration probability equation, but a positive effect on utility evaluation of considered lunch options. Additionally, the negative estimate of AvgPrice9.99 suggests that a respondent, who on average spends above $9.99 on lunch, gives higher scores to the presented menu options than a price-conscious respondent. This suggests that individuals are more price sensitive in their decisions to include the dish in the consideration set, than in their utility evaluations. In other words, high price may exclude a dish from the consideration set, but concurrently it can have a positive impact on the utility evaluation (for example when price is a signal of quality).

There are several indicators that endogeneity exists between the consideration and evaluative step in respondents' decision making about menu entrées. Recall, that Heckman two-step estimator includes the Mills inverse ratio as a regressor in the utility model to correct for the endogeneity. The coefficient estimates of this variable represent effects of considerable size and are significant at the .001 level for both utility regimes. Additionally, the estimated correlation between the residuals from the selection step and evaluation step is -.491, confirming again that endogeneity is strong and persistent.

In this section we provided empirical evidence that people assess differently the considered and unconsidered lunch options and that this evaluation process is indeed endogenous to the formation of consideration sets. We focused on the compensatory consideration set, estimated with Heckman two-step estimator. Additionally, in Appendix C this specification is compared to the traditional conjoint analysis model, showing that the latter performs quite poorly because it does not account for existence of consideration sets. This simple supplementary analysis shows that Heckman two-step estimator correctly captures the compensatory formation of consideration set and corrects for the existing endogeneity between evaluation and consideration.

4.5 Conclusions and Future Research

Rationally bounded consumers often apply heuristic rules to screen-off products which are not considered. But this decision has a random component, and there is interference between the consideration and the evaluative step which can generate endogeneity. Given these arguments, researches should not force exogeneity, but to allow for statistical dependence between the shocks of utility functions and consideration sets.

The application described in Section 4.4.2 is based on two-step estimator proposed by Heckman (1979), which allows for consistent estimation of preference parameters in case of endogeneity between product evaluations and a *compensatory consideration* set, under the assumption of normality and perfect sample separation. Below we open the discussion of future research agenda, providing a theoretical framework towards a general solution for more versatile consideration sets, latent consideration variables C_t, and signals of considerations. We also discuss the case of modeling endogenous consideration sets in conjoint analysis based on choice data.

4.5.1 The General Consideration Model

Consider a random consideration set $A\left(\gamma,u\right)$, where the heuristic rules are dependent with the random shocks on the utility, i.e. u_t is statistically dependent of $(\varepsilon_{1t},\varepsilon_{2t})$. Let us denote by $f_{1,2,u}(\varepsilon_1,\varepsilon_2,u)$ the joint density function, with marginal densities

$$f_{iu}(\varepsilon_i,u) \quad = \quad \int f_{1,2,u}(\varepsilon_1,\varepsilon_2,u)d\varepsilon_j, \qquad j \neq i$$

$$\begin{aligned} f_i(\varepsilon_i) &= \int f_{1,2,u}(\varepsilon_1,\varepsilon_2,u)\,d\varepsilon_j du, \qquad j \neq i \\ f_u(u) &= \int f_{1,2,u}(\varepsilon_1,\varepsilon_2,u)\,d\varepsilon_1 d\varepsilon_2. \end{aligned}$$

Then, each regime of the evaluation model has a density $f_{1u}\left(y_t - f(x_t)'\beta, u\right)$ and $f_{2u}(y_t-\alpha,u)$ respectively. Typically we specify these models using a factorization

$$\begin{aligned} f_{1u}\left(y_t - f(x_t)'\beta, u\right) &= f_1\left(y_t - f(x_t)'\beta\right) \times f_{u|1}\left(u|y_t - f(x_t)'\beta\right), \\ f_{2u}(y_t-\alpha,u) &= f_2(y_t-\alpha) \times f_{u|2}(u|y_t-\alpha). \end{aligned}$$

Working with these expressions, we can build the Likelihood function. The idea is to marginalize integrating over the unobserved variables u in the appropriate event.

These densities are easily derived in the context of a joint normal distribution, for example. Now we can consider estimation of the model parameters $\theta = vec\left(\beta,\alpha,\gamma,\sigma_{1u}^2,\sigma_{2u}^2,\sigma_1^2,\sigma_2^2\right)$. We can consider several situations and methods.

Case 1) Self-explicated consideration. If we have a known sample separation (we observe C_t), then

$$f_{1,2,u|c}(\varepsilon_1,\varepsilon_2,u|c) = \left\{\int_{A_{x_t}(\gamma)} f_{1u}(\varepsilon_1,u)\,du\right\}^{C_t} \left\{\int_{A_{x_t}(\gamma)^c} f_{2u}(\varepsilon_2,u)\,du\right\}^{(1-C_t)}$$

and the full-information Likelihood function is given by

$$L_T(\theta) = \prod_{t=1}^{T} \left\{\int_{A_{x_t}(\gamma)} f_{1u}\left(y_t - f(x_t)'\beta, u\right) du\right\}^{C_t} \left\{\int_{A_{x_t}(\gamma)^c} f_{2u}(y_t-\alpha,u)\,du\right\}^{(1-C_t)},$$

and using the conditional densities, the logarithm can be written as

$$\begin{aligned} \ln L_T(\theta) &= \sum_{t=1}^{T} C_t \left\{\ln f_1\left(y_t - f(x_t)'\beta\right) + \ln \int_{A_{x_t}(\gamma)} f_{u|1}\left(u|y_t - f(x_t)'\beta\right) du\right\} \\ &\quad + (1-C_t)\left\{\ln f_2(y_t-\alpha) + \ln \int_{A_{x_t}(\gamma)^c} f_{u|2}(u|y_t-\alpha)\,du\right\}. \end{aligned}$$

Under differentiability assumptions this model can be estimated applying the Newton-Raphson method.

Case 2) Inferred consideration. If the consideration regime is not observed, then

$$f_{1,2}(\varepsilon_1,\varepsilon_2)=\int_{A_{x_t}(\gamma)} f_{1u}(\varepsilon_1,u)\,du+\int_{A_{x_t}(\gamma)^c} f_{2u}(\varepsilon_2,u)\,du$$

and the likelihood function Likelihood function is given by

$$\begin{aligned} L_T(\theta) &= \prod_{t=1}^{T}\left\{\int_{A_{x_t}(\gamma)} f_{1u}\left(y_t-f(x_t)'\beta,u\right)du+\int_{A_{x_t}(\gamma)^c} f_{2u}(y_t-\alpha,u)\,du\right\} \\ &= \prod_{t=1}^{T}\left\{f_1\left(y_t-f(x_t)'\beta\right)\int_{A_{x_t}(\gamma)} f_{u|1}\left(u|y_t-f(x_t)'\beta\right)du+f_2(y_t-\alpha)\int_{A_{x_t}(\gamma)^c} f_{u|2}(u|y_t-\alpha)\,du\right\}. \end{aligned}$$

Note that the log-likelihood function is relatively involved, as the logarithm is applied to the whole term in brackets.

Case 1 and case 2 are opposite extremes: either the sample separation (consideration) is completely known or unknown.

Case 3) Consideration Signals. If we have a signal S_t related to C_t as described in Table 1, then from (4.6), we conclude that

$$\begin{aligned} f_{1,2,S}(\varepsilon_1,\varepsilon_2,S) &= \left(\pi_{11}\times\int_{A_{x_t}(\gamma)} f_{1u}\left(y_t-f(x_t)'\beta,u\right)du+\pi_{01}\times\int_{A_{x_t}(\gamma)^c} f_{2u}(\varepsilon_2,u)\,du\right)^{S_t} \\ &\quad\times\left((1-\pi_{11})\times\int_{A_{x_t}(\gamma)} f_{1u}(\varepsilon_1,u)\,du+(1-\pi_{01})\times\int_{A_{x_t}(\gamma)^c} f_{2u}(\varepsilon_2,u)\,du\right)^{1-S_t} \end{aligned}$$

and we the conditional Likelihood function is given by

$$\begin{aligned} L_T(\theta) &= \prod_{t=1}^{T}\left\{\pi_{11}\times\int_{A_{x_t}(\gamma)} f_{1u}(\varepsilon_1,u)\,du+\pi_{01}\times\int_{A_{x_t}(\gamma)} f_{2u}\left(y_t-f(x_t)'\beta,u\right)du\right\}^{S_t} \\ &\quad\times\left\{(1-\pi_{11})\times\int_{A_{x_t}(\gamma)} f_{1u}(\varepsilon_1,u)\,du+(1-\pi_{01})\times\int_{A_{x_t}(\gamma)} f_{2u}\left(y_t-f(x_t)'\beta,u\right)du\right\}^{(1-S_t)} \end{aligned}$$

Based on the Likelihood function, researchers can either compute the Maximum Likelihood estimator (usually applying numerical methods), or the Bayes estimator (i.e. posterior mean from an arbitrary prior, generally applying numerical integration methods such as Monte Carlo approximations or quadrature formulas). Under regularity conditions, both types of estimators are consistent and have the same asymptotic distribution (by the Bernstein-von Mises Theorem), therefore we do not delve on this issue. In any case, the conditions required to ensure a good performance of Bayes estimators are usually stricter than for Maximum Likelihood, thus we will compute our estimators maximizing the log-likelihood.

The estimators based on self-explicated data (case 1) are typically more efficient, as they use more information than the estimators that do not use this data. However, such estimators will be inconsistent if the quality of self-explicated information is low. If researchers are doubtful, the validity can be tested using a Hausman-Wu test comparing both estimators.

If we do not have (or decide not to use) the consideration variables C_t, we must maximize the Likelihood function described in case 2, which is usually a harder task. There is an alternative trick that can be used to tackle this problem. We can consider an EM algorithm, taking arbitrary initial values for C_t (e.g. the half sample with larger values y_t), and following case 1) to estimate the parameters. Next, we update the consideration variables using the optimal forecast $\hat{C}_t = I(\Pr(C_t = 1|y_t, x_t) > 0.5)$ with

$$\Pr(C_t = 1|y_t, x_t) = \int_{A_{x_t}(\gamma)} f_{u|1}\left(u|y_t - f(x_t)'\beta\right) du,$$

iterating until the parameters converge or at least do not vary too much. The EM usually increases the true likelihood, but it is difficult to establish formal convergence to the maximum. On the other hand, the stopping values of the EM algorithm (even if we do not rigorously converge) should be closer to the maximum and can be used to initialize a Newton-Raphson algorithm for the true Likelihood (described in case 2).

In each of the considered scenarios, Maximum Likelihood estimates are consistent and asymptotically efficient, but the optimization can be cumbersome as often there are several local maxima. A variety of robust optimization algorithms can be used to solve this problem. If a locally convergent method is used (such as Newton-Raphson), it is convenient to initiate it at some preliminary consistent but inefficient estimator. For example, assuming normal distributions and observed C_t in the compensatory-consideration model, we can use the two-steps Heckman procedure to obtain an initial value (this actually the procedure applied in Section 4.4). For other consideration rules we can adapt the two step method of Heckman to the specific set $A_x(\gamma)$, and the resulting estimator for γ would be used to estimate the other parameters in the second step.

The computation of the gradients and Hessians is required for the Newton-Raphson method, and to estimate the asymptotic variance. In the Appendix B, we discuss the computation of these derivatives, these procedures should be applied to handle any of the discussed Maximum Likelihood methods.

4.5.2 Extension to Choice Models

We can also contemplate the existence of endogenous consideration sets in conjoint analysis based on choice. In the last decade, the experiments increasingly require respondents to choose for $t = 1, .., T$ times from a small group of J alternative products at each time (the product ratings $y_{1t}, ..., y_{Jt}$ are conceptualized as unobserved latent variables, and researchers record a group of dummy dependent variables $\{Y_{jt}\}_{j=1}^{T}$ related to the latent variable as $Y_{jt} = I\left(y_{jt} = \max\{y_{1t}, ..., y_{Jt}\}\right)$ where I denotes the indicator function). In particular, if we assumed normality, we can combine the Probit multinomial model with the endogenous likelihood. Notice that in choice models, we only observe choices and we do not even ask about consideration or not, nor use any signal.

Assume J_t alternatives are considered in a choice task t, and each alternative is defined by the attributes $\{x_{jt}\}_{j=1}^{J_t}$ and latent utility of the alternative (j,t) is y_{jt}. The final choice is taken over the considered alternatives. First, we will assume that for $x_{jt} \notin A\left(\gamma, u_{jt}\right)$ the evaluation is equal to 0 with probability one, so that we only need to compare considered options:

$$\begin{aligned}
\Pr\left(Y_{jt} = 1 | \{x_{tl}\}_{l=1}^{J_t}\right) &= \Pr\left(y_{jt} - y_{lt} \geq 0, l \neq j | x_t, \{C_{lt} = 1\}_{l=1}^{J_t}\right) \\
&= \Pr\left(\varepsilon_{jt} \leq \varepsilon_{lt} + \left(f\left(x_{jt}\right) - f\left(x_{lt}\right)\right)' \beta, l \neq j | x_t, \{C_{lt} = 1\}_{l=1}^{J_t}\right)
\end{aligned}$$

$$= \int\int_{A_{x_{lt}}(\gamma)} \left\{ \prod_{l\neq j} F_{\varepsilon_j,u_j} \left(\varepsilon_{lt} + \left(f\left(x_{jt}\right) - f\left(x_{lt}\right)\right)' \beta, u_j | C_{lt} = 1 \right) \right\} f\left(\varepsilon, u_j\right) d\varepsilon du_j$$

with conditional distributions $F_{\varepsilon_j,u_j|C_l=1}\left(\varepsilon_{lt} + \left(f\left(x_{jt}\right) - f\left(x_{lt}\right)\right)' \beta, u_j|C_{lt} = 1\right)$ determined by its density

$$f_{\varepsilon_j,u_j|C_l}\left(\varepsilon_j, u_j|C_l = 1\right) = \frac{\int_{A_{x_{lt}}(\gamma)} f_{\varepsilon_j,u_j,u_l}\left(\varepsilon_j, u_j, u_l\right)\right) du}{\int_{-\infty}^{+\infty} \int_{A_{x_{lt}}(\gamma)} f_{\varepsilon_j,u_j,u_l}\left(\varepsilon_j, u_j, u_l\right)\right) d\varepsilon du},$$

so that

$$F_{\varepsilon_j,u_j}\left(e, u_j|C_l = 1\right) = \int_{-\infty}^{e} \frac{\int_{A_{x_{lt}}(\gamma)} f_{\varepsilon_j,u_j,u_l}\left(\varepsilon_j, u_j, u_l\right)\right) du}{\int_{-\infty}^{+\infty} \int_{A_{x_{lt}}(\gamma)} f_{\varepsilon_j,u_j,u_l}\left(\varepsilon_j, u_j, u_l\right)\right) d\varepsilon du} d\varepsilon_j.$$

Usually this distribution is computable, e.g. under jointly normal shocks.

The problem becomes complex if we allow some chance to the unconsidered options. If we consider that a discarded option x_{lt} is (surprisingly) chosen in the evaluation step, then we should consider events

$$\Pr\left(\varepsilon_{jt} \leq \varepsilon_{lt} + f\left(x_{jt}\right)' \beta - \alpha\beta | x_t, C_{lt} = 0\right)$$

using $F_{\varepsilon_j,u_j|C_l=1}\left(\varepsilon_{lt} + f\left(x_{jt}\right)' \beta - \alpha, u_j|C_{lt} = 0\right)$. Howver, in most applications it would not have much sense.

Finally, the full-information Likelihood function is given by

$$L_T\left(\beta, \alpha, \gamma\right) = \prod_{i=1}^{T} \prod_{j=1}^{J_t} \Pr\left(Y_{jt} = 1|x_t\right)^{Y_{jt}}.$$

The computation of the Likelihood integral, and its derivatives might require numerical computation.

If a customers' Panel is used, we might include refinements, such as including a mixture integrating the parameters of the model to include unobserved heterogeneity. But the model is complex enough and we believe that it is better to include observable variables about customers in the evaluative and consideration equations. The introduction of unobservable parametric heterogeneity using mixtures often enhances overparametrization (i.e., overfitting).

Bibliography

Beckwith, N. E., Kassarjian, H. H., and Lehmann, D. R. (1978). Halo effects in marketing research: review and prognosis. *Advances in consumer research*, 5(1):465–467.

Beckwith, N. E. and Lehmann, D. R. (1975). The importance of halo effects in multi-attribute attitude models. *Journal of Marketing Research*, 12(3):265–275.

Beckwith, N. E. and Lehmann, D. R. (1976). Halo effects in multiattribute attitude models: An appraisal of some unresolved issues. *Journal of Marketing Research*, 13(4):418–421.

Ben-Akiva, M. and Boccara, B. (1995). Discrete choice models with latent choice sets. *International Journal of Research in Marketing*, 12(1):9–24.

Berinsky, A. J., Huber, G. A., and Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon. com's mechanical turk. *Political Analysis*, 20(3):351–368.

Bettman, J. R. (1974). A threshold model of attribute satisfaction decisions. *Journal of Consumer Research*, 1(2):30–35.

Bettman, J. R. and Park, C. W. (1980). Effects of prior knowledge and experience and phase of the choice process on consumer decision processes: A protocol analysis. *Journal of Consumer Research*, 7(3):234–248.

Bettman, J. R. and Zins, M. A. (1977). Constructive processes in consumer choice. *Journal of Consumer Research*, 5(4):75–85.

Bröder, A. (2000). Assessing the empirical validity of the" take-the-best" heuristic as a model of human probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(5):1332–1346.

Fader, P. S. and McAlister, L. (1990). An elimination by aspects model of consumer response to promotion calibrated on upc scanner data. *Journal of Marketing Research*, 27(3):322–332.

Gensch, D. H. (1987). A two-stage disaggregate attribute choice model. *Marketing Science*, 6(3):223–239.

Gilbride, T. J. and Allenby, G. M. (2004). A choice model with conjunctive, disjunctive, and compensatory screening rules. *Marketing Science*, 23(3):391–406.

Gilbride, T. J. and Allenby, G. M. (2006). Estimating heterogeneous eba and economic screening rule choice models. *Marketing Science*, 25(5):pp. 494–509.

Hauser, J. R., Ding, M., and Gaskin, S. P. (2009). Non-compensatory (and compensatory) models of consideration-set decisions. In *Proceedings of the Sawtooth Software Conference*.

Hauser, J. R. and Wernerfelt, B. (1990). An evaluation cost model of consideration sets. *Journal of consumer research*, 16(4):393–408.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, 47(1):153–161.

Jedidi, K. and Kohli, R. (2005). Probabilistic Subset-Conjunctive Models for Heterogeneous Consumers. *Journal of Marketing Research*, 42(4):483–494.

Jedidi, K., Kohli, R., and DeSarbo, W. S. (1996). Consideration sets in conjoint analysis. *Journal of Marketing Research*, 33(3):364–372.

Johansson, J. K., MacLachlan, D. L., and Yalch, R. F. (1976). Halo effects in multiattribute attitude models: some unresolved issues. *Journal of Marketing Research*, 13(4):414–417.

Kau, P. and Hill, L. (1972). A threshold model of purchasing decisions. *Journal of Marketing Research*, 9(3):264–270.

Kohli, R. and Jedidi, K. (2007). Representation and inference of lexicographic preference models and their variants. *Marketing Science*, 26(3):380–399.

Lee, L.-F. and Porter, R. H. (1984). Switching regression models with imperfect sample separation information–with an application on cartel stability. *Econometrica: Journal of the Econometric Society*, 52(2):391–418.

Manzini, P. and Mariotti, M. (2013). Stochastic choice and consideration sets. *Econometrica (forthcoming)*.

Montgomery, H. and Svenson, O. (1976). On decision rules and information processing strategies for choices among multiattribute alternatives. *Scandinavian Journal of Psychology*, 17(1):283–291.

Ogden, C. L., Carroll, M. D., Kit, B. K., and Flegal, K. M. (2012). Prevalence of Obesity in the United States, 2009–2010. Data Brief No.82.

Olshavsky, R. W. and Acito, F. (1980). An information processing prove into conjoint analysis. *Decision Sciences*, 11(3):451–470.

Ordóñez, L. D., Benson III, L., and Beach, L. R. (1999). Testing the compatibility test: How instructions, accountability, and anticipated regret affect prechoice screening of options. *Organizational Behavior and Human Decision Processes*, 78(1):63 – 80.

Payne, J. W. (1976). Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational behavior and human performance*, 16(2):366–387.

Payne, J. W., Bettman, J. R., and Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(3):534–552.

Payne, J. W., Bettman, J. R., and Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge University Press.

Payne, J. W. and Ragsdale, E. K. E. (1978). Verbal protocols and direct observation of supermarket shopping behavior: Some findings and a discussion of methods. *Advances in consumer research*, 5:571–577.

Roberts, J. H. and Lattin, J. M. (1991). Development and testing of a model of consideration set composition. *Journal of Marketing Research*, 28(4):429–440.

Urban, G. L. and Hauser, J. R. (2004). "listening in" to find and explore new combinations of customer needs. *Journal of Marketing*, 68(2):72–87.

Yee, M., Dahan, E., Hauser, J. R., and Orlin, J. (2007). Greedoid-based noncompensatory inference. *Marketing Science*, 26(4):532–549.

Appendix A: Likelihood Functions in the Case of Exogeneity

For the microeconomic context it is also useful to discuss the likelihood functions for models imposing exogeneity (for a variety of examples see Maddala 1984, 1986). For each x_t we consider that there is an unobserved exogenous random variable u_t, usually a normal variable with zero mean and variance σ_u^2 independent from $(\varepsilon_{1t}, \varepsilon_{2t})$. Below we derive the likelihood functions for the parameters $\theta = vec\left(\beta', \alpha, \gamma, \sigma_1, \sigma_2\right)$, depending on the type of the available information.

Case a) "Self-explicated"consideration data. When respondents are able to declare C_t, we can build the likelihood function

$$\begin{aligned} L_T(\theta) &= \prod_{t=1}^{T} \left\{ f_1\left(y_t - f(x_t)'\beta\right) \times \Pr(C_t = 1|x_t) \right\}^{C_t} \left\{ f_2(y_t - \alpha) \times (1 - \Pr(C_t = 1|x_t)) \right\}^{1-C_t} \\ &= \prod_{t=1}^{T} \left\{ \left[C_t f_1\left(y_t - f(x_t)'\beta\right) + (1 - C_t) f_2(y_t - \alpha) \right] \times \Pr(C_t = 1|x_t)^{C_t} (1 - \Pr(C_t = 1|x_t))^{1-C_t} \right\}. \end{aligned}$$

Note that in practice we maximize

$$\begin{aligned} \ln L_T\left(\beta, \alpha, \gamma\right) &= \sum_{t=1}^{T} \left[C_t \ln f_1\left(y_t - f(x_t)'\beta\right) + (1 - C_t) \ln f_2(y_t - \alpha) \right] \\ &\quad + \sum_{t=1}^{T} \left[C_t \ln \Pr(C_t = 1|x_t) + (1 - C_t) \ln(1 - \Pr(C_t = 1|x_t)) \right]. \end{aligned}$$

which requires the use of numerical optimization algorithms. Under differentiability assumptions we can apply the Newton-Raphson method. If there are not traversal constraints, the estimation of β and γ can be addressed independently.

Case b) Inferred consideration. If we ignore whether or not a product profile belongs to the consideration set, we need to consider the Likelihood function

$$L_T(\theta) = \prod_{t=1}^{T} \left\{ f_1\left(y_t - f(x_t)'\beta\right) \times \Pr(C_t = 1|x_t) + f_2(y_t - \alpha) \times (1 - \Pr(C_t = 1|x_t)) \right\}.$$

Once the parameters in the model have been estimated, we can compute the probability of an observation being considered. In practice it is not too different from the previous case. For example, with the compensatory consideration model we have $\Pr(C_t = 1|x_t) = \Phi\left(x_t'\gamma_1 - \gamma_0\right)$, therefore with a standard normal shock u, and for independent Gaussian random shocks $\varepsilon_1, \varepsilon_2$

$$\ln L_T(\theta) = \sum_{t=1}^{T} \ln\left\{\sigma_1 \phi\left(\frac{y_t - f(x_t)'\beta}{\sigma_1}\right)\Pr(C_t = 1|x_t) + \sigma_2\phi\left(\frac{y_t - \alpha}{\sigma_2}\right)(1 - \Pr(C_t = 1|x_t))\right\},$$

with $\theta = vec\left(\beta', \alpha, \sigma_1, \sigma_2\right)$. This can be solved using the Newton-Raphson method, but the problem is numerically difficult. Alternatively, some authors use the EM algorithm considered by Hartley (1977, 1978) and Kiefer (1980).

Case c) Signals of consideration. If we use signals S_t of consideration as described in Table 1, then

$$\begin{aligned} f_{1,2,S}(\varepsilon_1, \varepsilon_2, S) &= \left[f_1(\varepsilon_1) \times \pi_{11} \times \Pr(C_t = 1|x_t) + f_2(\varepsilon_2) \times \pi_{01} \times (1 - \Pr(C_t = 1|x_t))\right]^{S_t} \\ &\quad \times \left[f_1(\varepsilon_1) \times (1 - \pi_{11}) \times \Pr(C_t = 1|x_t) + f_2(\varepsilon_2) \times (1 - \pi_{01}) \times (1 - \Pr(C_t = 1|x_t))\right]^{1-S_t}, \end{aligned}$$

and the conditional Likelihood function is given by

$$\begin{aligned} L_T(\theta) &= \prod_{t=1}^{T} \left\{f_1\left(y_t - f(x_t)'\beta\right)\pi_{11}\Pr(C_t = 1|x_t) + f_2(y_t - \alpha)\pi_{01}(1 - \Pr(C_t = 1|x_t))\right\}^{S_t} \\ &\quad \times \left\{f_1\left(y_t - f(x_t)'\beta\right)(1 - \pi_{11})\Pr(C_t = 1|x_t) + f_2(y_t - \alpha)(1 - \pi_{01})(1 - \Pr(C_t = 1|x_t))\right\}^{(1-S_t)}. \end{aligned}$$

It can be also written as

$$\begin{aligned} L_T(\theta) &= \prod_{t=1}^{T} \Big\{S_t\left[f_1\left(y_t - f(x_t)'\beta\right)\pi_{11}\Pr(C_t = 1|x_t) + f_2(y_t - \alpha)\pi_{01}(1 - \Pr(C_t = 1|x_t))\right] \\ &\quad + (1 - S_t)\left[f_1\left(y_t - f(x_t)'\beta\right)(1 - \pi_{11})\Pr(C_t = 1|x_t) + f_2(y_t - \alpha)(1 - \pi_{01})(1 - \Pr(C_t = 1|x_t))\right]\Big\}. \end{aligned}$$

Finally, notice that exogeneity of consideration tests can be tested comparing the Maximum Likelihood estimators computed with and without endogeneity with a Haussman-Wu test.

Appendix B: Derivatives of the Likelihood Function

The computation of the probabilities in the Likelihood function depends on the particular distribution assumed for the different shocks of the model. A classical approach is to assume Normality.

Example 6 ***Compensatory consideration***. *If $A\left(\gamma,u\right)=\left\{x:x'\gamma-u>0\right\}$, we assume that $\sigma_u^2=1$ to ensure identification (as γ is estimable only up to a scalar factor), and shocks have a joint normal distribution. Then, $A_x\left(\gamma\right)=\left(-\infty,x_t'\gamma\right]$,*

$$u_t|\varepsilon_{it}\sim N\left(0+\frac{\sigma_{iu}}{\sigma_i^2}\left(\varepsilon_{it}-0\right),\sigma_u^2\left(1-\rho_{ui}^2\right)\right)$$

with $\rho_{ui}^2=\frac{\sigma_{iu}^2}{\sigma_i^2\sigma_u^2}$, and the required integrals can be expressed as

$$\begin{aligned}\int_{-\infty}^{x_t'\gamma}f_{u|1}\left(u|y_t-f\left(x_t\right)'\beta\right)du &= \frac{1}{\sqrt{\sigma_u^2\left(1-\rho_{ui}^2\right)}}\int_{-\infty}^{x'\gamma}\phi\left(\frac{\left(u-\frac{\sigma_{iu}}{\sigma_i^2}\left(y_t-f\left(x_t\right)'\beta\right)\right)}{\sqrt{\sigma_u^2\left(1-\rho_{ui}^2\right)}}\right)du\\ &= \frac{1}{\sqrt{\sigma_u^2\left(1-\rho_{ui}^2\right)}}\frac{\phi\left(\frac{\left(x_t'\gamma-\frac{\sigma_{iu}}{\sigma_i^2}\left(y_t-f\left(x_t\right)'\beta\right)\right)}{\sqrt{\sigma_u^2\left(1-\rho_{ui}^2\right)}}\right)}{\Phi\left(\frac{\left(x_t'\gamma-\frac{\sigma_{iu}}{\sigma_i^2}\left(y_t-f\left(x_t\right)'\beta\right)\right)}{\sqrt{\sigma_u^2\left(1-\rho_{ui}^2\right)}}\right)}.\end{aligned}$$

and similarly

$$\int_{x_t'\gamma}^{\infty}f_{u|2}\left(u|y_t-\alpha\right)du=\frac{1}{\sqrt{\sigma_u^2\left(1-\rho_{ui}^2\right)}}\frac{\phi\left(\frac{\left(x_t'\gamma-\frac{\sigma_{iu}}{\sigma_i^2}\left(y_t-\alpha\right)\right)}{\sqrt{\sigma_u^2\left(1-\rho_{ui}^2\right)}}\right)}{1-\Phi\left(\frac{\left(x_t'\gamma-\frac{\sigma_{iu}}{\sigma_i^2}\left(y_t-\alpha\right)\right)}{\sqrt{\sigma_u^2\left(1-\rho_{ui}^2\right)}}\right)}.$$

In practice, the main problem is not the computation of the Likelihood function, but the computation of the derivatives of the Likelihood function. In case of compensatory preferences, this is relatively simple because the integrals have a closed form. But the problem is more difficult for other types of consideration sets, and other distributional assumptions. This section is devoted to the analysis of the derivatives in a general context.

In order to compute the Likelihood function we need to take derivatives of the border of an integral. The likelihood functions contain expressions such as

$$\int_{A_x(\theta)} f(\theta,x,u)\,du$$

where θ is a general vector containing all the parameters in the model and f an integrable function in u. The goal of this section is the computation of gradients and Hessians for this type of integrals. Following condition is assumed.

Condition 7 *For each $x \in \mathscr{X}$, the support of the regressors, we assume that $A_x\left(\gamma\right)$ is a Cartesian product of rectangles (intervals or semi-intervals). In general we assume that θ belongs to a compact interval $\theta_0 \le \theta \le \theta_1$, and*

$$A_x(\theta) = \{u : a(\theta,x) \le u \le b(\theta,x)\},$$

where the inequalities $\theta_0 \le \theta \le \theta_1$, and $a(\theta,x) \le u \le b(\theta,x)$ must be considered in a pointwise coordinate sense. We also assume that given any fixed $x \in \mathscr{X}$, the density $f(\theta,x,u)$ satisfies that both $f(\theta,x,u)$ and and its partial derivative $\frac{\partial}{\partial\theta} f(\theta,x,u)$ are continuous in some region of the (θ,u)-plane including the set

$$\{(\theta,u) : a(\theta,x) \le u \le b(\theta,x), \theta_0 \le \theta \le \theta_1\}. \tag{4.10}$$

and the functions $a(\theta,x)$ and $b(\theta,x)$ are both continuous and both have continuous derivatives for $\theta_0 \le \theta \le \theta_1$.

Then, the following version of the Leibniz rule holds: for any $\theta_0 \le \theta \le \theta_1$,

$$\begin{aligned}\frac{d}{d\theta}\left(\int_{A_x(\theta)} f(\theta,x,u)\,du\right) &= f(x,\theta,b(x,\theta))\frac{\partial}{\partial\theta}b(\theta,x) - f(x,\theta,a(x,\theta))\frac{\partial}{\partial\theta}a(\theta,x)\\ &\quad + \int_{a(\theta,x)}^{b(\theta,x)} \frac{\partial}{\partial\theta} f(\theta,x,u)\,du\end{aligned}$$

This formula can be readily derived using the fundamental theorem of calculus. The idea can be extended to second derivatives.

Condition 8 *In addition, assume $\frac{\partial^2}{\partial\theta\partial\theta'}f(\theta,x,u)$ is continuous in region (4.10), and $a(\theta,x)$ and $b(\theta,x)$ have continuous second derivatives in this region.*

Then, applying the same argument we obtain that for any $\theta_0 \le \theta \le \theta_1$,

$$\begin{aligned}
\frac{d^2}{d\theta d\theta'}\left(\int_{A_x(\theta)} f(\theta,x,u)du\right) &= \frac{\partial}{\partial\theta}f(\theta,x,b(\theta,x))\frac{\partial}{\partial\theta}b(\theta)+f(\theta,x,b(\theta,x))\frac{\partial^2}{\partial\theta\partial\theta'}b(\theta,x) \\
&\quad -\left(\frac{\partial}{\partial\theta}f(\theta,x,a(\theta,x))\frac{\partial}{\partial\theta}a(\theta,x)+f(\theta,x,a(\theta,x))\frac{\partial^2}{\partial\theta\partial\theta'}a(\theta,x)\right) \\
&\quad +\frac{\partial}{\partial\theta}f(\theta,x,b(\theta,x))\frac{\partial}{\partial\theta}b(\theta,x)-\frac{\partial}{\partial\theta}f(\theta,x,a(\theta,x))\frac{\partial}{\partial\theta}a(\theta,x) \\
&\quad +\int_{a(\theta,x)}^{b(\theta,x)}\frac{d^2}{d\theta d\theta'}f(\theta,x,u)du.
\end{aligned}$$

The Leibniz integral rule can be extended to multidimensional integrals in more abstract sets $A_x(\theta)$, using tools from differential geometry. Without loss of generality we can take a limit for some of the integration boundaries in $a(\theta,x), b(\theta,x)$ where the boundary of an upper or lower bound moves to infinite. These are the required tools to compute gradient and Hessians in Newton formulas. In practice, the computation of derivatives involves the computation of

$$\begin{aligned}
&\int_{a(\theta,x)}^{b(\theta,x)}\frac{\partial}{\partial\theta}f(\theta,x,u)du \\
&\int_{a(\theta,x)}^{b(\theta,x)}\frac{d^2}{d\theta d\theta'}f(\theta,x,u)du.
\end{aligned}$$

If these integrals are analytically intractable, they can be numerically computed, e.g. using Quadrature formulas, or a Monte Carlo uniform sample on the set $A_x(\theta)$. Importance sampling can actually work better. In this case, it is convenient to keep the seed in the underlying pseudo-random sequence constant along all the iterations of the Newton-Raphson algorithm to avoid numerical instabilities.

Appendix C: Supplementary Empirical Analysis

As an additional empirical exercise, we report the estimation results for the traditional conjoint model, where it is assumed that no screening-off occurs, therefore all presented alternatives are included in the consideration set. With linear compensatory utility function and under standard assumptions this model can be estimated with OLS. Specifically, we include the same set of variables used for the Heckman two-step procedure in Section 4.4.

Table 4.6 summarizes the regression results estimating with OLS. The model explains around 21% of the variance in the data, which is not bad, and with the value of $F(30,2687) = 24.58$ is globally statistically significant at .000 level. The log-likelihood is equal to -6347.80 (around 3 times lower than the value attained by the Heckman two-step estimator). Most of the variables are significant, with the largest positive effects related to food allergies, indicating that people who suffer from fish, lactose and gluten allergy or are on the low-calorie diet tend to evaluate our menu more favourably than respondents who are not on these special diets. The estimate of age is also positive and significant but its influence is small. Among design variables (product features), "catrnk2" has the strongest influence on the utility score, meaning that respondents tend to evaluate the dishes from two of their preferred categories 1.9 point higher than dishes from other categories. Additionally, vegetarian and grilled dishes contribute to the increase in the overall dish utility, and the interaction effect representing chicken dishes at $11.99 has a strong and positive estimate, which is bigger (to the absolute value) than the associated main effects. On the other hand, vegetarians tend to give lower utility scores, as well as women, heavier people and respondents, who exercise at least once per week. Among design variables, chicken, fish, and second price level decrease the utility score of a dish. Note however, that a big part of the utility contribution of chicken and $11.99 is explained by significant and positive estimates of "catrnk2" and the interaction effect "chicken-$11.99".

Let's compare the average predicted utilities from the OLS estimation and a model with endogenous compensatory consideration set specified in Section 4.4. We see that globally OLS predicts the average utilities slightly better than Heckman two-step estimator: the fitted mean

Table 4.6: OLS estimation results

Utilities	β_j	SE
beef	.063	.323
chicken	-.583^a	.325
fish	-.398^c	.137
price \$11.99	.004	.510
price \$15.99	-.290	.249
grilled	.646^c	.158
broiled	.173	.158
salad	-.070	.305
veggies	.352^b	.137
fries	-.485	.305
glutenY	-.010	.096
organic	-.185	.208
female	-.363^c	.105
BMI	-.018^b	.007
age	.017^c	.005
onlineorder	.091	.102
eatoutY	-.107	.108
excerciseY	-.263^b	.108
AvgPr9.99	-.331^c	.109
AllGluten	.645^b	.326
AllLactose	.893^b	.402
AllFish	1.916^c	.735
LowCalDiet	.673^b	.337
LowSodDiet	.533	.330
VegeDiet	-1.625^c	.360
AllDiet0	.399	.323
AllDietOther	.297	.303
catrnk2	1.933^c	.106
beef-\$11.99	.659	.713
chicken-\$11.99	1.278^a	.713
Intercept	4.904^c	.558
R^2	0.2153	
Adj. R^2	0.2065	
LL	-6347.80	

^{a} $p < .1$, ^{b} $p < .05$, ^{c} $p < .01$

utilities are very close to the sample mean (see the Column "Total" in Table 4.7). However, the OLS model does not account well for existence of consideration sets: it overestimates the utility of unconsidered options (column $C_t = 0$), but more importantly the mean utilities for the considered options are significantly underestimated (column $C_t = 1$). On the other hand, the model with endogenous compensatory consideration is able to capture the consideration set formation much better that OLS, predicting the average utility equally well for considered and unconsidered lunch alternatives (the last row in Table 4.7).

Table 4.7: Mean utilities: data and fitted values

	$C_t = 0^a$	$C_t = 1^b$	Total
Sample	5.09	8.50	5.80
OLS	5.56	6.44	5.74
Heckman two-step	5.05	9.34	5.92

[a] Unconsidered options

[b] Considered options

Bibliography

Hartley, M. J. (1977). On the estimation of a general switching regression model via maximum likelihood methods. Discussion Paper 415, Department of Economics, State University of New York.

Hartley, M. J. (1978). Comment. *Journal of the American Statistical Association*, 73(364):738–741.

Kiefer, N. M. (1980). A note on switching regressions and logistic discrimination. *Econometrica: Journal of the Econometric Society*, 48(4):1065–1069.

Maddala, G. S. (1984). Estimation of the disequilibrium model with noisy indicators. Technical report, University of Florida.

Maddala, G. S. (1986). Disequilibrium, self-selection, and switching models. In Griliches, Z. and Intriligator, M., editors, *Handbook of Econometrics*, volume 3, pages 1633–1688. Elsevier.