European Journal of Operational Research 241 (2015) 564–574

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Interfaces with Other Disciplines

Advanced conjoint analysis using feature selection via support vector machines

Sebastián Maldonado[a,*], Ricardo Montoya[b], Richard Weber[b]

[a] Universidad de los Andes, Mons. Álvaro del Portillo 12455, Las Condes, Santiago, Chile
[b] Department of Industrial Engineering, Universidad de Chile, Av. República 701, Santiago, Chile

ARTICLE INFO

Article history:
Received 22 January 2014
Accepted 25 September 2014
Available online 25 October 2014

Keywords:
Conjoint analysis
Feature selection
Support vector machines
Business analytics

ABSTRACT

One of the main tasks of conjoint analysis is to identify consumer preferences about potential products or services. Accordingly, different estimation methods have been proposed to determine the corresponding relevant attributes. Most of these approaches rely on the post-processing of the estimated preferences to establish the importance of such variables. This paper presents new techniques that simultaneously identify consumer preferences and the most relevant attributes. The proposed approaches have two appealing characteristics. Firstly, they are grounded on a support vector machine formulation that has proved important predictive ability in operations management and marketing contexts and secondly they obtain a more parsimonious representation of consumer preferences than traditional models. We report the results of an extensive simulation study that shows that unlike existing methods, our approach can accurately recover the model parameters as well as the relevant attributes. Additionally, we use two conjoint choice experiments whose results show that the proposed techniques have better fit and predictive accuracy than traditional methods and that they additionally provide an improved understanding of customer preferences.

1. Introduction

Conjoint analysis is one of the research techniques most widely used to identify customers' preferences (see e.g. Green, Krieger, & Wind, 2001). Firms' decisions regarding new product or service design (Kohli & Krishnamurti, 1989) as well as promotional and advertising campaigns increasingly rely on its outputs. Usually, the estimated preferences are the inputs for market simulation techniques that are then used to evaluate different market opportunities. Additionally, conjoint analysis allows estimating consumers' willingness to pay (WTP), defined as the price of indifference between buying and not buying (Gensler, Hinz, Skiera, & Theyson, 2012), and thus it helps to make important pricing decisions. Consequently, appropriate conjoint studies and their derived implications can determine the success or failure of new product introductions or marketing campaigns.

Originally developed in marketing (Green & Rao, 1971), conjoint analysis has had an increasing impact in many other disciplines such as health care (Bridge et al., 2011; Halme & Kallio, 2011), tourism management (Thyne, Lawson, & Todd, 2006), transportation (Hensher, Louviere, & Swait, 1998), and operations management (Dobson & Kalish, 1993), among others. Further applications where this technique has been successfully employed have been presented by Karniouchina, Moore, van der Rhee, and Verma (2009).

In addition, conjoint analysis is relevant for the Operations Research community for at least the following two reasons. First, conjoint analysis can be used in the context of multi-attribute decision making (MADM), since multiple attributes are considered in a preference measurement process. A comparison to an alternative MADM technique (Analytic Hierarchy Process) has been presented in Scholl, Manthey, Helm, and Steiner (2005). Second, optimization techniques are used in the context of conjoint analysis (see e.g. Camm, Cochran, Curry, & Kannan, 2006; Halme & Kallio, 2011, 2014). This fact constitutes an opportunity to develop different types of advanced optimization models to increase the applicability of conjoint analysis.

One of the main outputs of conjoint analysis is to identify the relevant attributes at the consumer level. That is, the (subset of) attributes that the consumer considers when evaluating the proposed alternatives. The usual approach to obtain such subset of attributes (or their ranking) is by post-processing the estimated parameters. For instance, the relative range of part-worths can be used to represent attribute importance when using additive models such as a mixed logit model. Such post-processing task implicitly assumes that consumers use all attributes when facing a conjoint decision. However, as shown later, traditional models can have problems eliminating irrelevant attributes across consumers, especially when there is limited individual-level data. Indeed, despite current developments

* Corresponding author. Tel.: +56 9 61704167.
E-mail address: smaldonado@uandes.cl (S. Maldonado).

http://dx.doi.org/10.1016/j.ejor.2014.09.051
0377-2217/© 2014 Elsevier B.V. All rights reserved.

Table 1
Summary of the relevant literature in Support Vector Machines (SVM), Feature Selection (FS), and Conjoint Analysis (CA)

References	SVM	FS	CA
Schoelkopf and Smola (2002); Vapnik and Chervonenkis (1991)	✓		
Blum and Langley (1997); Fan and Li (2001); Song, Smola, Gretton, Bedo, and Borgwardt (2012)		✓	
Arora and Huber (2001); Ben-Akiva and Lerman (1985); Dzyabura and Hauser (2011); Gelman and Pardoe (2006); Gilbride and Allenby (2006); Green et al. (2001); Hauser, Toubia, Evgeniou, Befurt, and Dzyabura (2010); Jedidi, Montoya, and Kohli (2013); Kohli and Krishnamurti (1989); Rossi, Allenby, and McCulloch (2005); Toubia, Hauser, and Garcia (2007b)			✓
Bradley and Mangasarian (1998); Guyon and Elisseeff (2003); Maldonado and Weber (2009); Maldonado, Weber, and Basak (2011)	✓	✓	
Chapelle and Harchaoui (2005); Cui and Curry (2005); Evgeniou et al. (2005); Evgeniou, Pontil, and Toubia (2007); Toubia, Evgeniou, and Hauser (2007a)	✓		✓
Argyriou, Evgeniou, and Pontil (2008)		✓	✓
This study	✓	✓	✓

in choice modeling that incorporate non-compensatory preferences, typical models based on conjoint analysis do not allow for “attribute non-attendance” (Hensher, Rose, & Greene, 2012). This occurs when customers completely neglect some attributes and focus their attention on a small subset of attributes. As conjoint analysis studies have been incorporating more complex products that are characterized by a larger number of attributes and at the same time more data are available, it is expected that many consumers be more selective regarding the attributes they really consider. Our proposed model contributes in filling this gap in the academic literature and also aims at providing a useful contribution to practitioners.

Several approaches from data mining and machine learning have been presented in the last decade in order to achieve better predictive performance in conjoint analysis (Evgeniou, Boussios, & Zacharia, 2005) and accurate representations of consumer preferences. These approaches have proved to provide important insights and consequently have gained reputation as valid methods to uncover customers’ preferences. However, they do not address the problem of effectively and efficiently selecting the relevant attributes used by consumers in their evaluation tasks. Attribute (or feature) selection has proved to be an important characteristic that predictive models need to include (see e.g. Blum & Langley, 1997; Guyon & Elisseeff, 2003). Not only because of a more parsimonious representation but also because it can better identify true underlying preferences that can lead to a higher predictive ability of consumer decisions. Table 1 presents an overview of the relevant literature studied in this work.

We present a novel technique based on Support Vector Machines (SVM) to determine the relevant attributes for estimating customer preferences. The identification of the relevant attributes that customers use to evaluate products, with the corresponding reduction in the dimensionality of customers’ utility functions, is achieved by a backward elimination of attributes procedure based on the individual part-worths. Therefore, such attribute selection is performed simultaneously to the estimation of customers’ preferences. An extensive simulation exercise shows that the proposed approach outperforms existing methods for attribute selection in the context of choice-based conjoint analysis.

The contribution of the paper is twofold: (i) it presents a framework that simultaneously identifies the most relevant attributes when estimating customer preferences, and (ii) it shows that the understanding of customers’ preferences and the predictive performance of the proposed approach can be enhanced considering the most relevant attributes.

The remainder of the paper is organized as follows. Section 2 discusses previous work. In particular, it describes SVM for CBC and provides a general overview of the different attribute selection approaches for SVM. The proposed method for attribute selection based on SVM for conjoint analysis is introduced in Section 3. In Section 4 we present the results of a simulation exercise that underline our method’s capabilities. Section 5 describes the application of the proposed approaches in two empirical conjoint studies highlighting the managerial implications that can be derived from the respective analyses. Section 6 summarizes the key results and discusses directions for future research.

2. Previous work

SVM were introduced to conjoint analysis by Evgeniou et al. (2005) and Cui and Curry (2005). Evgeniou et al. (2005) showed that SVM are accurate, robust to noise, and computationally efficient in a conjoint analysis context. Cui and Curry (2005) found that the predictive ability of SVM outperforms competing models such as multinomial logit models in consumer choice experiments. Later, Evgeniou et al. (2007) developed a convex approach for modeling consumer heterogeneity (Natter & Feurstein, 2002) in conjoint analysis, and compared it to Hierarchical Bayes (HB) methods. To the best of our knowledge, the present paper is the first work that adds feature selection to SVM for conjoint analysis.

Section 2.1 describes SVM in the context of choice-based conjoint analysis (CBC) (Chapelle & Harchaoui, 2005; Cui & Curry, 2005; Evgeniou et al., 2005). In Section 2.2 we present the state-of-the-art regarding feature selection using SVM.

2.1. Support vector machines for choice-based conjoint analysis

Consider a product profile with J attributes. Each attribute is defined over n_j levels, $j = 1, \ldots, J$. Suppose a consumer evaluates the profiles of K different products and chooses one profile in each of T choice occasions. Finally, consider a sample of N customers.

Customer i’s preferences are modeled by an additive utility function, which is assumed to be a linear combination of the partial utilities (part-worths): $u_i(\mathbf{x}) = \mathbf{w}_i{}^T \cdot \mathbf{x}$, $i = 1, \ldots, N$.

We consider CBC data with the following information $([\mathbf{x}_{it}^1, \ldots, \mathbf{x}_{it}^K], y_{it})$, where $\mathbf{x}_{it}^k \in \Re^J$ and $y_{it} \in \{1, \ldots, K\}$ for $1 \le i \le N$, $1 \le t \le T$, and $1 \le k \le K$. The choice $y_{it} = k$ indicates that at occasion t, consumer i prefers the kth alternative among the K product profiles described by $[\mathbf{x}_{it}^1, \ldots, \mathbf{x}_{it}^K]$. That is, $u_i(\mathbf{x}_{it}^{y_{it}}) \ge u_i(\mathbf{x}_{it}^b)$, $\forall b \in \{1, \ldots, K\} \setminus \{y_{it}\}$ (Chapelle & Harchaoui, 2005). Without loss of generality, and following previous research, we assume that for each choice occasion t all customers choose the first profile, i.e. $y_{it} = 1$, $1 \le i \le N$ and $1 \le t \le T$. Thus, the inequalities can be rewritten as

$$\mathbf{w}_i{}^T \cdot (\mathbf{x}_{it}^1 - \mathbf{x}_{it}^k) \ge 0, \tag{1}$$

where $1 \le i \le N$, $2 \le k \le K$, and $1 \le t \le T$.

To determine the weights $\mathbf{w}_i$ the *structural risk minimization principle* (Vapnik & Chervonenkis, 1991) has been considered. This approach minimizes the Euclidean norm of $\mathbf{w}_i$, with noise penalization via slack variables ξ_{kt} (l_2-soft margin formulation) that leads to the following quadratic programming problem for each customer $i = 1, \ldots, N$ (Chapelle & Harchaoui, 2005; Evgeniou et al., 2005):

$$\min_{\mathbf{w}_i, \xi} \quad \frac{1}{2}\|\mathbf{w}_i\|^2 + C\sum_{t=1}^{T}\sum_{k=2}^{K}\xi_{kt} \tag{2}$$

s.t.

$$\mathbf{w}_i{}^T \cdot (\mathbf{x}_{it}^1 - \mathbf{x}_{it}^k) \ge 1 - \xi_{kt} \quad t = 1, \ldots, T;\ k = 2, \ldots, K.$$

$$\xi_{kt} \ge 0 \quad t = 1, \ldots, T;\ k = 2, \ldots, K.$$

Model (2) minimizes ξ_{kt} that represent inconsistencies in the choice data. This formulation simultaneously controls for the complexity of the model by maximizing the margin ($\propto 1/\|\mathbf{w}_i\|^2$). The parameter C determines the trade-off between fitting the data and controlling for the model's complexity. It can be set exogenously by the researcher or endogenously, using, for example, a cross-validation procedure (see e.g. Toubia et al., 2007a). The components of the vector $\mathbf{w}_i$ (corresponding to the individual part-worths) satisfy the stated choice preferences (constraints) (Evgeniou et al., 2005). The solution to this optimization problem yields the part-worths $\mathbf{w}_i$ for each customer $i = 1, \ldots, N$.

CBC usually lacks sufficient information to estimate individual part-worths independently. Consequently, to allow for unobserved heterogeneity, the SVM formulation applied to CBC pools information across individuals in the same way as hierarchical Bayesian approaches do in discrete choice models (see e.g. Gelman & Pardoe, 2006). This pooling allows capturing general patterns at the population level and avoiding potential overfitting to each individual's choices. In the SVM literature several approaches have been proposed to deal with this issue. For instance, Evgeniou et al. (2005) suggested a regularization procedure that specifies a hierarchical functional form of the individual part-worths considering a population part-worth $\overline{\mathbf{w}} = 1/N \sum_i \mathbf{w}_i$. The trade-off between the individual and aggregated part-worth is controlled via cross-validation using a parameter $\gamma_i \in [0, 1]$ in the form of a weighted sum $\gamma_i \mathbf{w}_i + (1 - \gamma_i)\overline{\mathbf{w}}$. Alternatively, Chapelle and Harchaoui (2005) proposed an optimization formulation that simultaneously computes the individual part-worths for all respondents, considering general patterns in the population. Later, Evgeniou et al. (2007) introduced an alternative approach that jointly obtains the individual part-worths using the information from all customers. Unlike a ridge regression, where a quadratic loss function is used to maximize fit, they suggest shrinking the weights toward a vector $\mathbf{w}_0$, whose components are also decision variables.

2.2. Feature selection with support vector machines

Feature selection addresses the problem of finding the most compact and informative subset of the original attributes. This is based on the assumption that irrelevant and redundant attributes have a negative effect on supervised learning (Blum & Langley, 1997; Maldonado & Weber, 2009). Feature selection has three important general benefits that can be applied to a choice-based conjoint context (Guyon, Gunn, Nikravesh, & Zadeh, 2006). First, it improves the understanding of the decision process by obtaining a more parsimonious and meaningful representation of customer preferences. Second, it may improve the predictive performance of the model, especially in high-dimensional applications. This selection procedure can mitigate the *curse of dimensionality* that prescribes that as the number of attributes increases, an exponential increase in the number of observations is needed to maintain reliable model estimation (Stone, 1985). Additionally, the introduction of noise from irrelevant/redundant attributes results in less accurate predictors. And third, attribute selection limits storage requirements and increases the speed of the estimation algorithms. This is a critical issue in cases where accurate solutions are needed in a relatively short time.

2.2.1. Feature selection approaches

Three main approaches have been developed for feature selection: filter, wrapper, and embedded methods (Guyon et al., 2006).

Filter methods use statistical feature properties to filter out irrelevant attributes. This is usually performed before applying any supervised or unsupervised model. These methods have advantages, such as their simplicity, scalability, and reduced computational effort. In contrast, they ignore the interactions among attributes and their relationship with the classification algorithm (Guyon et al., 2006).

Wrapper methods explore the entire attribute space to score subsets of attributes according to their predictive power. Since the search for an optimal subset of attributes grows exponentially with the number of original variables, heuristic approaches have been suggested to address this combinatorial problem. The most commonly used wrapper strategies are the Sequential Forward Selection (SFS) and the Sequential Backward Elimination (SBE) (Guyon et al., 2006). In the first case, the strategy starts with few variables, and candidate variables are added sequentially to the set of selected features. At each iteration, the variable whose inclusion most improves the classifier's performance is added to the set of selected attributes. In contrast, SBE starts with the complete set of attributes, and eliminates attributes sequentially.

Embedded methods attempt to find an optimal subset of features while constructing the predictive model at the same time. In general, embedded methods present important advantages in terms of variable and model interaction, capturing the dependencies among variables, and being computationally less demanding than wrapper methods (Guyon et al., 2006). However, these techniques are more complex conceptually, and modifications to the classification algorithm may lead to poor performance. Guyon and Elisseeff (2003) presented a well-known embedded method for classification with SVM called *Recursive Feature Elimination* (SVM-RFE). The goal of this iterative approach, which inspired the method we propose next, is to find a subset of variables which maximizes the classifier's performance. The feature to be removed in each iteration is the one whose removal minimizes the variation of the objective function. One advantage of this method is the possibility to perform non-linear feature selection. In the following section we present an adaptation of this approach for CBC.

3. Proposed methods for relevant attribute identification in conjoint analysis

We propose methods for feature selection using SVM for conjoint analysis that build on Model (2) presented by Evgeniou et al. (2005). Our method is flexible enough to allow for sparseness in the datasets produced by a consumer partially ignoring the provided information. This flexibility could improve predictive performance by identifying relevant attributes and removing irrelevant ones when estimating customers' preferences. In Section 3.1 we describe the feature selection approach with linear SVM and present the sequential backward elimination procedure. Then, in Section 3.2 we present the proposed non-linear approach.

3.1. Attribute selection for conjoint analysis using linear SVM

For the linear case, we first formulate an SVM for each customer $i \in \{1, \ldots, N\}$ to obtain the individual part-worths (Model (2)). Each attribute j has associated n_j part-worths (one for each level), and the difference between the highest and the lowest part-worths can be considered as a measure of relevance as in traditional conjoint analysis (see e.g. Green & Rao, 1971). Formally, we define the attribute contribution AC_j for attribute j as:

$$\mathrm{AC}_j(\mathbf{w}_i^j) = \max \quad \mathbf{w}_i^j - \min \mathbf{w}_i^j, \tag{3}$$

where $\mathbf{w}_i^j = (w_{i1}^j, w_{i2}^j, \ldots, w_{in_j}^j)$ are the part-worths associated with each level of attribute j, while max(min) $\mathbf{w}_i^j :=$ max(min) $\{w_{i1}^j, \ldots, w_{in_j}^j\}$.

In the case of ordered levels in terms of consumer preferences (for example, from highest to lowest price), Eq. (3) becomes simply the difference between the part-worths of the last and the first levels of each attribute j, respectively: $\mathrm{AC}_j(\mathbf{w}_i^j) = w_{in_j}^j - w_{i1}^j$.

We propose the following algorithm for identifying the relevant attributes while estimating the individual customer's preferences. We follow the notation used by Song et al. (2012) where $\mathcal{S}$ denotes the full

Table 2
Illustrative example. Relevant attributes are √ marked.

Customers	Product attributes				
	Price	Brand	Screen size	Processor	Memory
A	√		√		
B	√			√	

Algorithm 1 Algorithm for relevant attribute selection with linear SVM for CBC

Input: The full set of features $\mathcal{S}$, threshold ϵ
Output: Individual part-worths for a subset of relevant features

```
1. For all respondents $i = 1, \ldots, N$ do:
2.   $\mathcal{S}_i \leftarrow \mathcal{S}$
3.   repeat
4.     $\mathbf{w}_i \leftarrow$ SVM Training, Formulation (2), using $\mathcal{S}_i$
5.     $\{j\} \leftarrow \text{argmin}_j \, AC_j(\mathbf{w}_i^j)$
6.     $\mathcal{S}_i \leftarrow \mathcal{S}_i \setminus \{j\}$
7.   until $AC_j(\mathbf{w}_i^j) > \epsilon$ or $|\mathcal{S}_i| = 1$
8. end.
```

set of features. For each respondent a backward algorithm eliminates those attributes that are least important in the construction of the utility functions at each stage, indicated by the attribute contribution criterion.

The parameter ϵ is a *relevance threshold* for the relative contribution of each attribute. This threshold needs to be sufficiently small to avoid the elimination of relevant attributes. The stopping criterion is reached when the contribution of all remaining attributes is above this threshold, or only one attribute remains.

We now introduce an illustrative example to describe the functioning of the proposed algorithm.

Illustrative example: Setting. *Suppose we have data from a conjoint study for tablet computers described by five attributes: price, brand, screen size, processor and memory. For simplicity assume we estimate one weight (part-worth) per attribute.*[1] *Let consider two customers A and B. Customer A values attributes price and screen size only, whereas customer B values attributes price and processor only (see Table 2). Both consumers use their underlying preferences to make choices.*

Illustrative example: Algorithm 1. *After collecting the choice-based conjoint data of the customers' decisions, Algorithm 1 proceeds as follows. In an iterative process, it builds a linear utility function for customer A using all available attributes (Step 4), then, it computes the contribution measure (Eq. (3)) for all available attributes, and identifies its minimum value (Step 5). If this minimum value is below a given threshold ϵ, it removes the corresponding attribute (Step 6). Then, it goes back to Step 4 and builds a new linear function with the remaining attributes and continues with Steps 5 and 6 until all remaining attributes surpass the minimum contribution threshold or only one attribute remains. As the model will most likely assign low weights for attributes brand, processor and memory, this will result in low contribution measures and therefore those attributes will be removed, resulting in a utility function for customer A that includes only price and screen size. The same procedure is performed next for customer B, removing attributes brand, screen size, and memory (see Table 3).*

Table 3
Illustrative example. Algorithm 1—Iterative removal of irrelevant attributes.

Iteration	Customer	Attributes remaining
1	A	Price (2), Brand (0.1), Screen size (2), Processor(0.2), Memory (0.3)
2	A	Price (2), Screen size (2), Processor (0.2), Memory (0.3)
3	A	Price (2), Screen size (2), Memory (0.3)
4	A	Price (2), Screen size (2)
5	B	Price (1), Brand (0.1), Screen size (0.2), Processor (1), Memory (0.3)
6	B	Price (1), Screen size (0.2), Processor (1), Memory (0.3)
7	B	Price (1), Processor (1), Memory (0.3)
8	B	Price (1), Processor (1)

In parentheses the hypothetical attribute contribution AC (ϵ=0.9).

Algorithm 2 is a variation of the previous algorithm that includes a part-worth regularization procedure that controls for heterogeneity. The algorithm follows:

Algorithm 2 Algorithm for relevant attribute selection with linear SVM for CBC with part-worth regularization

Input: The full set of features $\mathcal{S}$, threshold ϵ, regularization parameter γ
Output: Part-worths for a subset of relevant features

```
1.  For all respondents $i = 1, \ldots, N$ do:
2.    $\mathcal{S}_i \leftarrow \mathcal{S}$
3.    repeat
4.      $\mathbf{w}_i \leftarrow$ SVM Training, Formulation (2), using $\mathcal{S}_i$
5.      $\mathbf{w}_i^0 \leftarrow \mathbf{w}_i$ (First SVM Training, all available attributes)
6.      $\{j\} \leftarrow \text{argmin}_j \, AC_j(\mathbf{w}_i^j)$
7.      $\mathcal{S}_i \leftarrow \mathcal{S}_i \setminus \{j\}$
8.    until $AC_j(\mathbf{w}_i^j) > \epsilon$ or $|\mathcal{S}_i| = 1$
9.  end
10. For all respondents $i = 1, \ldots, N$ do:
11.   $\overline{\mathbf{w}} \leftarrow 1/N \sum_i \mathbf{w}_i^0$
12. For all attributes $j \in \mathcal{S}_i$ do:
13.   $\mathbf{w}_{ij}^* \leftarrow \gamma \mathbf{w}_{ij} + (1 - \gamma)\overline{\mathbf{w}}_j$
14. end.
```

Notice that Algorithm 2 differs from Algorithm 1 only in the last two instructions, where the part-worths are regularized after the feature selection procedure has been performed and the model has been estimated. The population weight vector $\overline{\mathbf{w}}$ is computed by averaging the individual weight vectors $\mathbf{w}_i^0$ that are obtained before the feature selection procedure is performed (first SVM training, Step 6). The regularization is conducted only for those attributes j that are relevant for a particular customer i ($j \in \mathcal{S}_i$, Step 13). Parameters C and γ are obtained via grid search; see Section 4.2. In order to avoid overfitting, the same values of C and γ are considered for all respondents.

Illustrative example: Algorithm 2. *Following with our example described for Algorithm 1, assume initial weights $\mathbf{w}_A^0$ =(2, 0.1, 2, 0.2, 0.3) and $\mathbf{w}_B^0$ =(1, 0.1, 0.2, 1, 0.3) for customers A and B, respectively (before feature selection). Assuming again that the weights of the relevant attributes do not change posterior to the backward elimination process, we obtain $\mathbf{w}_A$ =(2, 0, 2, 0, 0) and $\mathbf{w}_B$ =(1, 0, 0, 1, 0) for customers A and B, respectively. Following Step 11 in Algorithm 2, $\overline{\mathbf{w}}$ =(1.5, 0.1, 1.1, 0.6, 0.3) is the population weight vector. Next, the updated weights considering the pooling approach described in Step 13 and γ =0.5 yields $\mathbf{w}_A^*$ =(1.75, 0, 1.55, 0, 0) and $\mathbf{w}_B^*$ =(1.25, 0, 0, 0.8, 0). See Table 4.*

3.2. Kernel-based attribute selection for conjoint analysis

In this section we discuss the extension to non-linear utility functions. In the case of non-linear SVM, the data are mapped automatically into a higher-dimensional space H by a function $\phi : \mathbf{x} \rightarrow \phi(\mathbf{x}) \in H$ (Schoelkopf & Smola, 2002). This mapping allows efficiently capturing non-linear dependencies that linear utility functions are unable to uncover. Given that the only values one needs to compute are scalar products of the form $\phi(\mathbf{x}) \cdot \phi(\mathbf{y})$, the mapping is performed

[1] In this case $AC_j(\mathbf{w}_i^j) = w_i^j$. The extension to our general specification per attribute-level is straightforward.

Table 4
Illustrative example. Algorithm 2—Pooling of attribute weights after selecting relevant attributes ($\gamma = 0.5$).

Weights	(Price, Brand, Screen size, Processor, Memory)
$\mathbf{w}_A^0$	(2, 0.1, 2, 0.2, 0.3)
$\mathbf{w}_B^0$	(1, 0.1, 0.2, 1, 0.3)
$\mathbf{w}_A$	(2, 0, 2, 0, 0)
$\mathbf{w}_B$	(1, 0, 0, 1, 0)
$\overline{\mathbf{w}}$	(1.5, 0.1, 1.1, 0.6, 0.3)
$\mathbf{w}_A^*$	(1.75, 0, 1.55, 0, 0)
$\mathbf{w}_B^*$	(1.25, 0, 0, 0.8, 0)

by a kernel function $K(\mathbf{x},\mathbf{y}) = \phi(\mathbf{x}) \cdot \phi(\mathbf{y})$ that defines an inner product in H (Evgeniou et al., 2005; Schoelkopf & Smola, 2002).

The following formulation is solved in order to obtain non-linear utility functions. The detailed derivation of this formulation is given in the online appendix available at the *European Journal of Operational Research* website.

$$\max_{\boldsymbol{\alpha}} \sum_{t=1}^{T}\sum_{k=2}^{K}\alpha_{kt} - \frac{1}{2}\sum_{t,s=1}^{T}\sum_{k=2}^{K}\alpha_{kt}\alpha_{ks}\big(K(\mathbf{x}_t^1,\mathbf{x}_s^1) + K(\mathbf{x}_t^k,\mathbf{x}_s^k) - K(\mathbf{x}_t^1,\mathbf{x}_s^k) - K(\mathbf{x}_t^k,\mathbf{x}_s^1)\big) \quad (4)$$

s.t.

$$0 \le \alpha_{kt} \le C \quad t = 1, \ldots, T;\ k = 2, \ldots, K.$$

Subsequently, the estimated utility function has the following form:

$$u_i(\mathbf{x}) = \sum_{t=1}^{T}\sum_{k=2}^{K}\alpha_{kt}\big(K(\mathbf{x}_t^1,\mathbf{x}) - K(\mathbf{x}_t^k,\mathbf{x})\big) \quad (5)$$

Our approach for attribute selection is a variation of Algorithm 1, where attribute j's contribution (AC_j) has to be adapted for kernel functions as follows.

$$\mathrm{AC}_j(\boldsymbol{\alpha}) = |W^2(\boldsymbol{\alpha}) - W^2_{(j)}(\boldsymbol{\alpha})| \quad (6)$$

where:

$$W^2(\boldsymbol{\alpha}) = \sum_{t,s=1}^{T}\sum_{k=2}^{K}\alpha_{kt}\alpha_{ks}\big(K(\mathbf{x}_t^1,\mathbf{x}_s^1) + K(\mathbf{x}_t^k,\mathbf{x}_s^k) - K(\mathbf{x}_t^1,\mathbf{x}_s^k) - K(\mathbf{x}_t^k,\mathbf{x}_s^1)\big) \quad (7)$$

and

$$W^2_{(j)}(\boldsymbol{\alpha}) = \sum_{t,s=1}^{T}\sum_{k=2}^{K}\alpha_{kt}\alpha_{ks}\big(K(\mathbf{x}_t^{1(-j)},\mathbf{x}_s^{1(-j)}) + K(\mathbf{x}_t^{k(-j)},\mathbf{x}_s^{k(-j)}) - K(\mathbf{x}_t^{1(-j)},\mathbf{x}_s^{k(-j)}) - K(\mathbf{x}_t^{k(-j)},\mathbf{x}_s^{1(-j)})\big) \quad (8)$$

We note that only the second component of the objective function of Model (4) ($\frac{1}{2}W^2(\boldsymbol{\alpha})$) depends on the selected attributes. To identify the relevant attributes we eliminate the attributes whose removal does not significantly affect $W^2(\boldsymbol{\alpha})$.

The vectors $\mathbf{x}_t^{1(-j)}$ and $\mathbf{x}_t^{k(-j)}$ used in Eq. (8) result when removing attribute j from $\mathbf{x}_t^1$ and $\mathbf{x}_t^k$, respectively. As a consequence, $W^2_{(j)}(\boldsymbol{\alpha})$ is almost identical to $W^2(\boldsymbol{\alpha})$. The only difference is that it uses the reduced attribute vectors, i.e. the attribute vectors where the component j has been removed.

Following Model (4), one seeks to minimize the metric W^2 in the same manner as in the algorithm SVM-RFE (Guyon et al., 2006). In our attribute selection context, this implies that we want to eliminate those attributes whose removal keeps the value of metric W^2 relatively small, leading to a small attribute contribution value AC_j. Accordingly, the algorithm for CBC using non-linear SVMs for each customer i is provided in Algorithm 3.

Algorithm 3 Algorithm for relevant attribute selection with non-linear SVM for CBC

Input: The full set of features $\mathcal{S}$, threshold ϵ
Output: (Approximated) part-worths for a subset of relevant features

```
For all respondents $i = 1, \ldots, N$ do:
    $\mathcal{S}_i \leftarrow \mathcal{S}$
    repeat
        $\boldsymbol{\alpha}_i \leftarrow$ SVM Training, Formulation (4), using $\mathcal{S}_i$
        $\{j\} \leftarrow \mathrm{argmin}_j\ AC_j(\boldsymbol{\alpha}_i)$
        $\mathcal{S}_i \leftarrow \mathcal{S}_i \setminus \{j\}$
    until $AC_j(\boldsymbol{\alpha}_i) > \epsilon$ or $|\mathcal{S}_i| = 1$
end
```

Notice that in this approach the regularization procedure used to obtain heterogeneous part-worths is not considered since in non-linear SVM only an approximation of the part-worths can be obtained. As a consequence, the part-worths are not readily available, and therefore it is not feasible to apply the part-worth regularization directly. However, allowing for heterogeneity in non-linear methods could lead to an interesting venue for future research.

Several kernel functions are available for non-linear SVMs. The radial basis function (RBF, Gaussian kernel) is preferred in most applications (Maldonado et al., 2011) and has been used in our experiments: $K(\mathbf{x}_i,\mathbf{x}_z) = \exp(-\frac{||\mathbf{x}_i-\mathbf{x}_z||^2}{2\sigma^2})$, where $\sigma > 0$ is a parameter controlling the width of the kernel, which determines the shape of the implied non-linear function.

4. Simulation exercise

The objectives of the simulation exercise are assessing the effectiveness of the proposed estimation procedure in identifying relevant attributes and analyzing the performance of the model presented in Section 3 under different error conditions. Section 4.1 describes the simulation setup. Section 4.2 exhibits the preference models we applied and the performance measures used for their evaluation. Finally, Section 4.3 presents the results of our simulation exercise.

4.1. Simulation design

We generated different datasets varying the noise condition in consumer choices (low and high noise) and the number of irrelevant attributes (low and high). In each condition we simulated choice data for $N = 200$ subjects across $T = 12$ choice occasions. Each choice set had $K = 3$ product profiles. These product profiles were generated using an orthogonal design with $J = 10$ attributes, each attribute j having $n_j = 4$ levels, $(j = 1, \ldots, 10)$. The simulated data were then used to estimate all the different preference models by splitting the corresponding datasets into two subsamples: calibration and test. We used the first 10 simulated choice decisions for calibration, and the final two decisions for testing purposes.

To vary the amount of noise we use the following procedure used by Arora and Huber (2001) and Toubia et al. (2007b). We first draw a four-level symmetric design of linear part-worths for each attribute j $(j = 1, \ldots, 10)$, generated from a normal distribution with mean $\boldsymbol{\mu} = (-\beta, -\frac{\beta}{3}, \frac{\beta}{3}, \beta)$; and covariance matrix $\boldsymbol{\Sigma} = \beta I$, where I is the 4×4 identity matrix. Note that lower values of β imply higher noise in the choice data. Therefore, and following Arora and Huber (2001), we used the values of $\beta = 0.5$ and $\beta = 2$ for "high" and "low" noise conditions, respectively. As a consequence, *a priori*, all attributes are equally important. Next, we generated irrelevant attributes to study the effect of the implied sparseness on customers' preferences. We randomly selected two and six features for each individual and fixed their corresponding part-worths to zero ($\boldsymbol{\mu} = \mathbf{0}$ for those attributes) for the low sparseness and high sparseness conditions, respectively. Note that a high degree of sparseness corresponds to customers ignoring a high number of attributes when evaluating the different alternatives in each choice set.

4.2. *Preference models and performance measures*

We estimated the following preference models:

1. LCA: Linear compensatory by aspects, where each attribute level is an aspect which is represented by dummy coding.
2. L-SVMi: Linear SVM using individual part-worths (Formulation (2)).
3. L-SVMic: Linear SVM using individual part-worths, corrected with the aggregated part-worths.
4. NL-SVM: Non-linear SVM (Formulation (4)).
5. LBE-SVMi: Linear SVM with individual part-worths and linear backward elimination (Algorithm 1).
6. LBE-SVMic: Linear SVM with individual regularized part-worths and linear backward elimination (Algorithm 2).
7. NLBE-SVM: Non-linear SVM with kernel-based backward elimination (Algorithm 3).

For LCA we formulate a mixed logit model and estimate it using a hierarchical Bayesian Markov chain Monte Carlo method (MCMC; see Rossi et al., 2005). For the proposed approaches we need to calibrate four additional parameters: C, ϵ, γ (only for regularized methods that control heterogeneity), and σ (only for kernel-based methods) as will be described next.

We use a leave-one-out cross validation strategy to tune those parameters using only the training data. This procedure defines iteratively (for each individual) a subset of the training data comprising all questions but one. The individual part-worths are then estimated using this subset and subsequently used to predict the response to the question left out (validation subset). The predictive performance of the solution is assessed using a hit-rate metric. This procedure is repeated many times so that each question in the training sample is left out once and used for validation purposes. The parameters are set to the values that maximize the cross-validation hit rate. Finally, after the parameters have been tuned (and fixed), the utility functions are constructed using the entire calibration set, and the final evaluation is performed in a test set (holdout sample), which remains unused during the calibration process. This well-known machine learning procedure has been used previously in conjoint analysis (Evgeniou et al., 2005; Evgeniou et al., 2007; Toubia et al., 2007a).

For the tuning parameters, and based on previous research (Maldonado et al., 2011), we explore the following sets applying grid search:

$$C \in \{2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^{0}, 2^{1}, 2^{2}, 2^{3}, 2^{4}, 2^{5}\},$$

$$\gamma \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\},$$

$$\epsilon \in \{0.25, 0.5, 0.75, 1.0, 1.25, 1.5\}, \text{ and } \sigma \in \{2^{0}, 2^{1}, 2^{2}, 2^{3}, 2^{4}, 2^{5}\}.$$

We then use these parameters to estimate the preference parameters $\mathbf{w}_i$. We estimate the above mentioned preference models that are compared based on the following performance measures.

1. In-sample hit rate.
2. Out-of-sample hit rate.
3. Feature usage rate (FU-rate): The average number of attributes used by customers. We compute this measure as follows:

$$\text{FU-rate} = \frac{\sum_{i=1}^{N} |\mathcal{S}_i|}{N \cdot J}, \tag{9}$$

where $|\mathcal{S}_i|$ is the cardinality of $\mathcal{S}_i$, the subset of selected attributes for customer i ($i = 1, \ldots, N$), and J is the number of all available attributes.

Table 5
Results for preference models (in percentage).

	Low noise – Low sparsity			**Low noise – High sparsity**		
	Hit rate			**Hit rate**		
Models	**In**[b]	**Out**[c]	**FU-rate**[d]	**In**	**Out**	**FU-rate**
No feature selection						
LCA	87.5	56.3	100	82	50.3	100
L-SVMi	98.3	54.0	100	98.1	51.8	100
L-SVMic	**96.7**	**60.0**[a]	**100**	96.0	58.5[a]	100
NL-SVMi	98.4	54.0	100	100	54.3	100
Feature selection						
LBE-SVMi	95.3	54.5	44.7	91.4	56.5	33.0
LBE-SVMic	96.1	59.0[a]	67.3	**75.8**	**59.8**[a]	**11.6**
NLBE-SVM	96.8	55.8	35.8	100	52.8	63.5
	High noise – Low sparsity			**High noise – High sparsity**		
	Hit rate			**Hit rate**		
Models	**In**	**Out**	**FU-rate**	**In**	**Out**	**FU-rate**
No feature selection						
LCA	76.4	43.5	100	73.3	41.2	100
L-SVMi	98.0	47.3	100	99.7	44.0	100
L-SVMic	95.6	52.5[a]	100	95.6	52.8[a]	100
NL-SVMi	99.1	49.3	100	100	46.5	100
Feature selection						
LBE-SVMi	97.2	52.3	74.5	75.2	48.0	10.0
LBE-SVMic	**93.7**	**53.5**[a]	**93.0**	**96.9**	**54.0**[a]	**83.1**
NLBE-SVM	97.3	52.8	78.3	100	48.5	78.9

[a] Best predictive hit rate or not significantly different than the best at the 0.01 level.
[b] In-sample hit rate.
[c] Out-of-sample hit rate.
[d] Feature usage rate.

4.3. *Results of simulation exercise*

Section 4.3.1 displays the results for the different models under varying noise and sparseness conditions. The effectiveness for attribute selection is assessed in Section 4.3.2.

4.3.1. *Model performance*

Table 5 summarizes the results for the estimated preference models under different error conditions.

Following previous research, we use out-of-sample performance for model selection. Accordingly, the best model is the one that yields the highest out-of-sample hit rate (highlighted in bold). We compare the other models to the best model and test if their holdout performances are statistically different at 1 percent significance level. A t-test is used to make the corresponding pairwise comparisons between average hit rates across customers.

Several results can be derived from the simulation exercise. First, as expected, the performance of the different approaches increases as the level of noise (controlled by the magnitude of β) decreases. Second, the performance decreases as the level of sparseness (attributes ignored by customers) increases. Third, as seen in machine learning and forecasting applications, selecting the more relevant attributes improves the predictive performance (out-of-sample hit rate) without significantly decreasing the fit (in-sample hit rate).

Fourth, the L-SVMic model, which performs a regularization of individual part-worths, performs worse than L-SVMi in terms of in-sample performance but better when facing out-of-sample data, reducing the risk of overfitting by incorporating general patterns into the individual part-worths. A further step in that direction is given by feature selection, since prediction performance is subsequently improved in three out of four cases. The fact that LBE-SVMic outperforms LBE-SVMi demonstrates the usefulness of allowing for heterogeneity in both cases: with and without feature selection. Finally, linear SVM models tend to perform better than the non-linear (kernel-based) ones, since the data were simulated assuming linear decision rules.

4.3.2. *Identifying non-attended attributes—Attribute selection.*

We examine the effectiveness of the different feature selection models in identifying relevant and irrelevant attributes. Recall that

Table 6
KL divergence for simulated data (smaller is better).

Preference models	LL[a]	LH	HL	HH
LBE-SVMi	6.232[b,c]	8.271[c]	6.713[c]	8.722
LBE-SVMic	6.710[c]	7.776[b,c]	6.323[b,c]	9.200[c]
NLBE-SVM	6.060[b,c]	8.054[c]	6.511[c]	8.736[b,c]
Consider all attributes	7.219	9.710	7.219	9.710
Consider none attributes	7.219	9.710	7.219	9.710

[a] LL: Low Noise–Low Sparseness; LH: Low Noise–High Sparseness; HL: High Noise–Low Sparseness; HH: High Noise–High Sparseness.
[b] Best or not significantly different than the best at the 0.01 level.
[c] Significantly better than null models ($p < 0.01$).

the relevant attributes are the ones with $\boldsymbol{\mu} \neq \mathbf{0}$, and that two and six out of 10 attributes were simulated as irrelevant. We consider two types of errors: (i) selecting irrelevant attributes and (ii) eliminating relevant attributes. To evaluate the performance of these methods we consider both false-positive and false-negative predictions. We use the Kullback-Leibler divergence (KLD) measure proposed by Dzyabura and Hauser (2011) to deal with discrete predictions (see Appendix C in Dzyabura & Hauser, 2011). Like Dzyabura and Hauser (2011) we calculate divergence from perfect prediction, thus the smaller the KLD value, the better. Unlike Dzyabura and Hauser who use observed consideration data to build the model and predict considered profiles in a validation sample (see Dzyabura & Hauser, 2011), we use observed choices to build the model (without information about attribute relevance), and we predict ignored attributes.

Table 6 summarizes the corresponding results for the simulated conditions. We compare the proposed models to two null models that predict that all attributes are relevant and none of the attributes is relevant. We indicate with the superindex † when a model is significantly better (lower KLD) than the null models, and with the superindex $*$ to highlight the best model (or statistically similar to the best model). Again, *t*-tests were performed to make pairwise comparisons between the feature selection performance of the different approaches.

In Table 6 we observe a significantly lower divergence of SVM-based feature selection models compared to null models. Additionally, as expected, the accuracy of the methods decreases with the level of error and sparseness. Interestingly, the model LBE-SVMic performs better than the other models displayed in Table 6 when there is high error and a low number of irrelevant attributes (HL), and when there is low error and a high number of irrelevant attributes (LH). Otherwise, it performs worse than the other feature selection models. It is important to note that the overall performance achieved by the proposed approach in terms of the KL divergence metric is comparable to other studies in which the information of the elements included (typically profiles) is observable. As we mentioned before, our models do not use information of this kind (considered profiles) and predict relevant and irrelevant attributes from observed choices.

5. Illustrative empirical applications

In this section we illustrate the application and characteristics of the proposed approaches using two existing choice-based conjoint datasets. The first set, analyzed in Section 5.1, is comprised of products (digital cameras) described across five attributes with four levels each (20 aspects in total). The product profiles are presented in choice sets with four alternatives. The second dataset, studied in Section 5.2, is a larger set that represents information usually collected for marketing research purposes. It contains products described across 10 unbalanced attributes with between 3 and 15 levels (51 aspects in total). The products in this dataset are presented in choice sets with three alternatives. In Section 5.3 we provide academic and managerial insights of our work based on these two applications.

5.1. Empirical application to digital cameras

$N = 125$ subjects were asked to evaluate different digital cameras in an on-line CBC study. A digital camera in this study is described by $J = 5$ attributes with $n_j = 4$ levels ($j = 1, \ldots, 5$):

- Price ($500, $400, $300, and $200),
- Resolution (2, 3, 4, and 5 Megapixels),
- Battery life (150, 300, 450, and 600 pictures),
- Optical zoom (2x, 3x, 4x, and 5x), and
- Camera size (SLR, Medium, Pocket, and Ultra Compact).

Subjects responded to 20 choice questions, with each choice question comprised of four product profiles. See Abernethy, Evgeniou, Toubia, and Vert (2008) for further details about the conjoint experiment.

The proposed approach simultaneously uncovers consumer preferences and identifies the most important attributes even when customers completely neglect some attributes while evaluating the product profiles. This is equivalent to a non-compensatory decision process where bad performances in one attribute cannot be compensated with good performances in other attributes. Accordingly, and following the literature in decision process we estimate two types of preference models (i) compensatory approaches and (ii) non-compensatory approaches. For the compensatory models we estimate a state-of-the-art mixed logit model (LCA) and a q-compensatory model that is used to represent strict compensatory preferences (Hauser et al., 2010; Jedidi et al., 2013). The q-compensatory model is an additive model in which the importance of any aspect is no more than q times as large as the importance of any other aspect. Additionally, we include traditional SVM-based models that do not incorporate feature selection to highlight the differences in terms of predictive performance and identification of the relevant attributes. As non-compensatory benchmarks we choose two types of models: Elimination By Aspects (EBA) and Lexicographic. For EBA we use the Gilbride and Allenby (2006) approach, while for the Lexicographic models we use the Jedidi et al. (2013) approach. For these models we estimate three variants: Lexicographic by attributes (LBA), Lexicographic acceptance by aspects (LAL), and Lexicographic elimination by aspects (LEL). All benchmark models were specified as proposed by the authors and estimated using a hierarchical Bayesian MCMC approach (see e.g. Rossi et al., 2005). Our proposed SVM feature selection approach is included in this type of models.

We consider the first 16 questions to calibrate the models, tune the respective parameters in the case of SVM methods, and identify the relevant attributes. With the last four questions we test the estimated models.

5.1.1. Results

Table 7 summarizes the performance of the estimated methods for the digital cameras dataset. Similar to previous studies in the field we compare the performance of these methods in terms of predictive hit rates (see e.g. Cui & Curry, 2005; Dzyabura & Hauser, 2011; Evgeniou et al., 2007; Hauser et al., 2010). The best predictive performance among all methods is highlighted in bold type. We indicate with an asterisk the best predictive hit rate or not significantly different than the best at the 0.01 level.

As can be observed in Table 7, feature selection methodologies outperform the alternative approaches for conjoint analysis in terms of out-of-sample hit rate. Best results are obtained with the kernel-based approach (NLBE-SVM), using on average only 39 percent of the features across customers. The linear feature selection approaches with and without regularization achieve similar predictive performance, but using 57 percent of the features. This application confirms the analysis obtained earlier for simulated datasets: LBE-SVMi has better in-sample performance than LBE-SVMic but worse predictive performance, demonstrating the advantage of pooling information across

Table 7
Empirical comparison of the preference models (in percentage).

Models	Hit rate		
	In[b]	**Out**[c]	FU-rate[d]
Compensatory benchmarks			
LCA	84.5	58.0	100
q-compensatory	72.4	51.2	100
Non-compensatory benchmarks			
EBA	49.3	37.8	100
LBA	71.9	51.6	100
LAL	84.5	44.8	100
LEL	89.5	57.6	100
Traditional SVM			
L-SVMi	92.3	56.4	100
L-SVMic	91.6	58.4	100
NL-SVMi	95.1	58.6	100
SVM-feature selection			
LBE-SVMi	83.2	62.0[a]	57.0
LBE-SVMic	82.0	63.4[a]	57.0
NLBE-SVM	**78.5**	**63.6**[a]	**38.9**

[a] Best predictive hit rate or not significantly different than the best at the 0.01 level.
[b] In-sample hit rate.
[c] Out-of-sample hit rate.
[d] Feature usage rate.

individuals to predict preferences in this context. This is also true for the traditional SVM approaches that do not allow for feature selection.

Notice that the traditional SVM methods used previously in conjoint applications achieve high in-sample hit rate (95.1 percent in the case of NL-SVMi) but give predictive performances that are not statistically better than the additive model (58.6 percent vs 58.0 percent in the case of NL-SVMi), showing potential signs of overfitting if no feature selection procedures are performed. The kernel-based feature selection approach (NLBE-SVM) helps to effectively mitigate this problem eliminating irrelevant attributes that could have been ignored, improving the predictive performance while selecting on average approximately only two ($5 \times 0.389 = 1.945$) out of five attributes.

5.1.2. *Identifying relevant attributes and their relationships*

One of the main objectives of this work is to show that the proposed approaches can improve the interpretation of customer preferences by identifying the relevant attributes, analyzing individual part-worths, thus providing insights for more targeted marketing decisions. Accordingly, we compute the usage rate per attribute, which represents the percentage of customers using such an attribute to evaluate the alternatives.

Table 8 shows this metric for the methods LBE-SVMi and NLBE-SVM.

In Table 8 we observe a high degree of heterogeneity among customers in the use of relevant attributes for evaluating product profiles. For the nonlinear approach (NLBE-SVM, which achieves best predictive performance), 77.6 percent of the subjects consider the price as a relevant attribute for evaluating the product alternatives, whereas 41.6 percent consider the resolution as a relevant attribute.

Identifying relevant attributes can be important, e.g. for product design decisions. Knowing that the most important attribute (and the only relevant attribute for a third of the subjects) is “Price” may help to focus on improving the capabilities of digital cameras in a cost-efficient manner. In the case of assortment decisions, it can help to position the different digital cameras in the assortment in such a way that they can capture the heterogeneity across consumers. That is, some cameras can be positioned to target the mass of customers (by considering the lowest price) whereas other products can be targeting the long tail (with high resolution or long battery life cameras). Since, on the other hand, “Camera Size” does not appear to be relevant for most customers, it would make little sense to highlight this attribute, for example in advertisements.

We further studied the relationship in the use of these attributes by analyzing their relation at the individual level. Specifically, using the relevance estimates for each customer and applying association rules (Baesens, 2014; Shmueli, Patel, & Bruce, 2010), we investigated which attributes are used simultaneously and which attributes are used as substitutes. This analysis provides interesting results, some of which are presented below.

- The most frequent pattern of attribute usage corresponds to the case when “Price” appears as the only relevant attribute, which happens for 33.6 percent of the subjects. The next frequent patterns are “Price and Resolution” (16 percent), “Price and Zoom” (6.4 percent), and “Resolution” only (5.6 percent). The remaining combinations appear less than 5 percent of the time.
- The most frequent relationship among three attributes is “Price, Zoom, and Battery Life” (4.8 percent). This is a counterintuitive result given the importance of the attribute “Resolution” in the respondents’ preferences. We observe that customers who are looking not only for “Price”, are making the trade-off between “Resolution” and other attributes such as “Battery Life”.
- All five attributes are relevant for only one customer. Association among four relevant attributes appears for only 2.4 percent of the subjects (three out of 125 customers). This result confirms the importance of feature selection in this problem, and that low-dimensional solutions may lead to better results than using all available, albeit somewhat irrelevant, information.

5.2. *Empirical application with a large number of attributes*

The previous results show that even in applications with a moderate number of attributes, some customers seem to ignore part of the information presented. We can expect that in more complex settings, e.g. with more attributes, customers are even more selective in the information they use to evaluate alternatives. Accordingly, we now study a larger dataset in terms of the number of attributes that describe the products. Due to the proprietary nature of the data, the actual product and the specific attributes and attribute levels are disguised. We now provide some basic information that helps to interpret the results.

$N = 602$ subjects were asked to evaluate this product in an on-line CBC study. Each product is described by $J = 10$ unbalanced attributes, where three attributes have three levels, five have four levels, one has seven levels, and one has 15 levels. Subjects responded to 12 choice questions, each of which contained three product profiles. Ten questions were used for training and calibration purposes, while the remaining two were included for testing.

Table 8
Usage rate per attribute for the SVM-based feature selection approaches.

Models	**Attributes for digital cameras**				
	Price	**Resolution**	**Battery life**	**Optical zoom**	**Camera size**
LBE-SVMi	80.8	71.2	43.2	55.2	34.4
NLBE-SVM	77.6	41.6	17.6	30.4	15.2

Table 9
Empirical comparison of the preference models (in percentages).

Models	Hit rate In[b]	 Out[c]	FU-rate[d]
Compensatory benchmarks			
LCA	70.4	57.2	100
q-comp.	63.5	54.2	100
Non-compensatory benchmarks			
EBA	60.1	50.4	100
LAL	68.3	47.4	100
LEL	69.4	44.5	100
Traditional SVM			
L-SVMi	95.0	58.6	100
L-SVMic	91.4	59.6	100
NL-SVMi	92.8	58.8	100
SVM-feature selection			
LBE-SVMi	78.5	60.5[a]	10.0
LBE-SVMic	**75.1**	**61.3**[a]	**10.0**
NLBE-SVM	91.5	58.8	40.7

[a] Best predictive hit rate or not significantly different than the best at the 0.01 level.
[b] In-sample hit rate.
[c] Out-of-sample hit rate.
[d] Feature usage rate.

We analyze the same models as in the previous case except for LBA, since we treat all attributes as nominal and decide to analyze them at the aspect level. See the previous application in Section 5 for a description of each method and their references.

5.2.1. Results

Table 9 exhibits the fit and predictive statistics for the different models. This table also contains the percentage of the attributes identified as relevant. The best predictive performance among all methods is highlighted in bold type. We indicate with an asterisk the best predictive hit rate or not significantly different than the best at the 0.01 level.

The results presented in Table 9 are consistent with the previous ones in terms of the superior predictive performance of the proposed feature selection approaches. However, linear models outperform non-linear approaches in this application, and compensatory models perform better than non-compensatory models. Regarding the proposed approaches, the linear feature selection models outperform alternative approaches (including the proposed non-linear model). The best performance is achieved with LBE-SVMic, using on average only 10 percent of the features for all customers (one out of 10 across individuals).

5.2.2. Relationship among relevant attributes and marketing implications

The usage rate per attribute achieved by the methods LBE-SVMic and NLBE-SVM is shown in Table 10.

In Table 10 we again observe a high degree of heterogeneity in the use of the attributes. This heterogeneity is substantially lower in the case of the linear model (LBE-SVMic) where most of the respondents (61 percent) seem to use “Attribute Two” to evaluate the alternatives, whereas 15 percent of the customers use “Attribute Eight”. The rest of the attributes are used by less than 7 percent of the respondents. The kernel approach NLBE-SVM shows higher heterogeneity with most of the respondents using “Attribute Two” (78.9 percent), and 25.2 percent of them using “Attribute Ten” which is the least used attribute. Interestingly, both approaches coincide in the three most used attributes (Two, Eight, and Four). Finally, notice that if the attribute is described by too many levels, then consumers tend to ignore that attribute. This is the case with attributes Nine and Ten which have 15 and 7 levels, respectively. In these cases, the model LBE-SVMic predicts that only 0.7 percent and 0.5 percent of the customers use these attributes, respectively.

It is perhaps surprising that only 39 percent and 10 percent of the attributes in both applications respectively do as well as using all the information both in fitting the data and predicting choices. This is indeed an empirical issue, however we conjecture that this may be due to the complexity of the tasks that force individuals to concentrate on only a few attributes for evaluating the alternatives.

5.3. Academic and managerial insights

Feature selection is a standard tool from the machine learning literature that is now increasingly being used in other areas such as the ones presented in this paper. Indeed, both applications presented in this section underline the importance of an adequate selection of the relevant attributes. The results of both applications show that feature selection provides a more parsimonious representation of consumer preferences without sacrificing model performance and predictive ability. The proposed approach based on SVM for conjoint analysis shows a robust performance under different conditions (number of customers, number of product attributes) which enhances the fruitful conjoint analysis research area to simultaneously uncover preferences and identify the relevant attributes.

Managerial insights are mainly driven by marketing decisions that can benefit from our proposed models. Identifying customers’ preferences on an individual level is key when evaluating new product introductions and assortment decisions. Focusing on the most important attributes used by customers can leverage such decisions and produce more effective marketing communications. In that regard, our approach yields better predictive results than competing models with a substantially lower number of attributes used, which simplifies marketing communications and product positioning.

Finally, our applications also provide new empirical evidence of “attribute non-attendance”. That is, consumers seem to neglect a large number of attributes as the complexity of the product increases. A product in the first application (digital camera) is described by five attributes whereas a product in the second application is described by 10. As a consequence, whereas in the first application consumers use about 39 percent of the attributes, in the second application this number reduces to 10 percent on average. This is an important issue for practitioners who assume that consumers use all product attributes to evaluate products. This fact could misrepresent consumer preferences with the corresponding erroneous implications for managerial decisions that utilize such consumer preference information as input.

Table 10
Usage rate per attribute for the SVM-based feature selection approaches (in percentage).

	Product attributes									
	One 4[a]	Two 3	Three 4	Four 3	Five 4	Six 4	Seven 3	Eight 4	Nine 15	Ten 7
LBE-SVMic	4.8	61.0	1.5	6.8	3.7	2.2	3.8	15.1	0.7	0.5
NLBE-SVM	35.4	78.9	43.2	50.5	31.9	28.1	30.6	51.0	32.2	25.2

[a] Number of levels per attribute.

6. Conclusions and Future Work

In this work, we extend previous literature on conjoint analysis by allowing for feature selection and providing a new methodology to identify the relevant attributes in a choice-based conjoint setting. The proposed new methods are based on statistical learning techniques and perform feature selection simultaneously with estimating customers' preferences via a modified SVM approach. We adapt linear as well as non-linear SVMs to identify relevant attributes in CBC to improve both performance as well as interpretation of the implied results. A comparison between our approaches and other existing techniques shows several advantages for ours. First, the proposed models outperform the predictive ability of the traditional additive approach and the standard SVM for conjoint analysis. The main source of the improvement is their ability to identify relevant attributes at the individual level (and eliminate seemingly ignored attributes). This therefore reduces the number of attributes needed to represent customers' utility functions, avoiding the "curse of dimensionality" and consequently the risk of overfitting. Second, the proposed models can be extended to non-linear utility specifications by introducing kernel functions that improve predictive performance through the gain in flexibility. And third, analyzing individual partworths provides important insights into customers' preferences as revealed by simplifying heuristics that may lead customers to ignore some attributes. In particular, the results of our empirical applications show that consumers may use just one or two attributes to evaluate the alternatives. These relevant attributes, however, differ importantly across customers. Therefore, it is imperative that the feature selection step be performed individually instead of at the population level. Additionally, we confirm some well-known maxims in both marketing and machine learning fields, such as the importance of estimating individual utility functions to characterize customers' heterogeneity, and the need for low-dimensional models to avoid the curse of dimensionality. Furthermore, we show that attribute selection could replace—to a certain extent—a regularization procedure. Finally, we provide additional empirical evidence of the usefulness of using machine learning techniques such as SVM to analyze conjoint data.

Future work can be carried out in several directions. First, it would be interesting to apply this approach to other conjoint applications such as menu-based conjoint. Attribute selection can improve our understanding of the decision rules employed by customers in these more complex situations. These new contexts could even influence customers to ignore attributes in the different stages of the decision process. Second, the proposed approach could be applied to dynamic settings, where an attribute selection procedure could help to generate more parsimonious choice sets and potentially identify customers' preferences more efficiently. Third, our approach to identifying relevant attributes could be compared to adaptive methods for choice-based conjoint analysis (ACBC). The latter methods usually consider a Build-Your-Own (BYO) configuration, where a so-called screener generates specific (product or service) concepts. Since the analyzed concepts start from self-generated profiles (BYO), it would be interesting to see how the identification of relevant attributes would be affected. Fourth, we consider the study of different shrinkage specifications for our formulation as another interesting direction for future work. The main challenge is to perform feature selection at the individual level by solving a unique optimization problem. This requires introducing important modifications to state-of-the-art shrinkage specifications (Chapelle & Harchaoui, 2005; Evgeniou et al., 2007). Finally, the use of sparsity terms instead of a backward elimination procedure when selecting attributes could be explored. Although this procedure could be more parsimonious, it is expected to affect the efficiency of the proposed approach significantly.

Altogether this paper opens interesting research avenues based on the proposed techniques for simultaneous identification of consumer preferences and relevant attributes for the respective choice decisions.

Acknowledgments

The authors thank Olivier Toubia and Bryan Orme for providing the data for the two empirical applications. The first author was supported by FONDECYT project 11121196 and 1140831. The second author was supported by FONDECYT project 11110173. The third author was supported by FONDECYT project 1140831. The work reported in this paper has been partially funded by the Complex Engineering Systems Institute (ICM: P-05-004-F, CONICYT: FBO16).

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at doi:http://dx.doi.org/10.1016/j.ejor.2014.09.051

References

Abernethy, J., Evgeniou, T., Toubia, O., & Vert, J. (2008). Eliciting consumer preferences using robust adaptive choice questionnaires. *IEEE Transactions on Knowledge and Data Engineering*, *20*(2), 145–155.

Argyriou, A., Evgeniou, T., & Pontil, M. (2008). Convex multi-task feature learning. *Machine Learning*, *73*(3), 243–272.

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

Baesens, B. (2014). *Analytics in a big data world*. John Wiley and Sons.

Ben-Akiva, M., & Lerman, S. (1985). *Discrete choice analysis: Theory and application to travel demand*. Cambridge, MA: MIT Press.

Blum, A., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, *97*(1-2), 245–271.

Bradley, P., & Mangasarian, O. (1998). Feature selection via concave minimization and support vector machines. In J. Shavlik (Ed.), *Proceedings of the fifteenth international conference on machine learning (ICML'98)* (pp. 82–90). San Francisco, CA: Morgan Kaufmann.

Bridge, J., Hauber, A. B., Marshall, D., Lloyd, A., Prosser, L. A., Regier, D. A., et al. (2011). Conjoint analysis applications in health—A checklist: A report of the ispor good research practices for conjoint analysis task force. *Value in Health*, *14*(4), 403–413.

Camm, J. D., Cochran, J. J., Curry, D. J., & Kannan, S. (2006). Conjoint optimization: An exact branch-and-bound algorithm for the share-of-choice problem. *Management Science*, *52*(3), 435–447.

Chapelle, O., & Harchaoui, Z. (2005). *Advances in neural information processing systems (Vol. 17*, pp. 257–264). Cambridge, MA: MIT Press.

Cui, D., & Curry, D. (2005). Prediction in marketing using the support vector machine. *Marketing Science*, *24*(4), 595–615.

Dobson, G., & Kalish, S. (1993). Heuristics for pricing and positioning a product-line using conjoint and cost data. *Management Science*, *39*(2), 160–175.

Dzyabura, D., & Hauser, J. R. (2011). Active machine learning for consideration heuristics. *Marketing Science*, *30*(5), 801–819.

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

Evgeniou, T., Pontil, M., & Toubia, O. (2007). A convex optimization approach to modeling heterogeneity in conjoint estimation. *Marketing Science*, *26*(6), 805–818.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.

Gelman, A., & Pardoe, I. (2006). Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics*, *48*(2), 241–251.

Gensler, S., Hinz, O., Skiera, B., & Theyson, S. (2012). Willingness-to-pay estimation with choice-based conjoint analysis: Addressing extreme response behavior with individually adapted designs. *European Journal of Operational Research*, *219*(2), 368–378.

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

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

Green, P. E., & Rao, V. R. (1971). Conjoint measurement for quantifying judgmental data. *Journal of Marketing Research*, *8*, 355–363.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning research*, *3*, 1157–1182.

Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. A. (2006). *Feature extraction, foundations and applications*. Berlin: Springer.

Halme, M., & Kallio, M. (2011). Estimation methods for choice-based conjoint analysis of consumer preferences. *European Journal of Operational Research*, *214*(1), 160–167.

Halme, M., & Kallio, M. (2014). Likelihood estimation of consumer preferences in choice-based conjoint analysis. *European Journal of Operational Research*, In press.

Hauser, J. R., Toubia, O., Evgeniou, T., Befurt, R., & Dzyabura, D. (2010). Disjunctions of conjunctions, cognitive simplicity, and consideration sets. *Journal of Marketing Research*, *67*, 485–496.

Hensher, D., Louviere, J., & Swait, J. (1998). Combining sources of preference data. *Journal of Econometrics*, *89*(1), 197–221.

Hensher, D. A., Rose, J. M., & Greene, W. H. (2012). Inferring attribute non-attendance from stated choice data: Implications for willingness to pay estimates and a warning for stated choice experiment design. *Transportation*, *39*(2), 235–245.

Jedidi, K., Montoya, R., & Kohli, R. (2013). Probabilistic lexicographic models. *Working paper*, Columbia University.

Karniouchina, E. V., Moore, W. L., van der Rhee, B., & Verma, R. (2009). Issues in the use of ratings-based versus choice-based conjoint analysis in operations management research. *European Journal of Operational Research*, *197*(1), 340–348.

Kohli, R., & Krishnamurti, R. (1989). Optimal product design using conjoint analysis: Computational complexity and algorithms. *European Journal of Operational Research*, *40*(2), 186–195.

Maldonado, S., & Weber, R. (2009). A wrapper method for feature selection using support vector machines. *Information Sciences*, *179*(13), 2208–2217.

Maldonado, S., Weber, R., & Basak, J. (2011). Kernel-penalized svm for feature selection. *Information Sciences*, *181*(1), 115–128.

Natter, M., & Feurstein, M. (2002). Real world performance of choice-based conjoint models. *European Journal of Operational Research*, *137*(2), 448–458.

Rossi, P. E., Allenby, G. M., & McCulloch, R. (2005). *Bayesian statistics and marketing*. New York: Wiley.

Schoelkopf, B., & Smola, A. J. (2002). *Learning with kernels*. Cambridge, MA: MIT Press.

Scholl, A., Manthey, L., Helm, R., & Steiner, M. (2005). Solving multiattribute design problems with analytic hierarchy process and conjoint analysis: An empirical comparison. *European Journal of Operational Research*, *164*(1), 760–777.

Shmueli, G., Patel, N. R., & Bruce, P. C. (2010). *Data mining for business intelligence*. Hoboken, NJ: John Wiley and Sons.

Song, L., Smola, A., Gretton, A., Bedo, J., & Borgwardt, K. (2012). Feature selection via dependence maximization. *Journal of Machine Learning Research*, *13*, 1393–1434.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Annals of Statistics*, *13*, 689–705.

Thyne, M., Lawson, R., & Todd, S. (2006). The use of conjoint analysis to assess the impact of the cross-cultural exchange between hosts and guests. *Tourism Management*, *27*(2), 201–213.

Toubia, O., Evgeniou, T., & Hauser, J. (2007a). *Conjoint measurement: Methods and applications* (pp. 231–258). New York: Springer.

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

Vapnik, V., & Chervonenkis, A. (1991). The necessary and sufficient conditions for consistency in the empirical risk minimization method. *Pattern Recognition and Image Analysis*, *1*(3), 283–305.