Embedded heterogeneous feature selection for conjoint analysis: A SVM approach using L1 penalty

Sebastián Maldonado[1] · Ricardo Montoya[2] · Julio López[3]

Published online: 10 November 2016

Abstract This paper presents a novel embedded feature selection approach for Support Vector Machines (SVM) in a choice-based conjoint context. We extend the L1-SVM formulation and adapt the RFE-SVM algorithm to conjoint analysis to encourage sparsity in consumer preferences. This sparsity can be attributed to consumers being selective about the attributes they consider when evaluating alternatives in choice tasks. Given limited individual data in choice-based conjoint, we control for heterogeneity by pooling information across consumers and shrinking the individual weights of the relevant attributes towards a population mean. We tested our approach through an extensive simulation study that shows that the proposed approach can capture the sparseness implied by irrelevant attributes. We also illustrate the characteristics and use of our approach on two real-world choice-based conjoint data sets. The results show that the proposed method has better predictive accuracy than competitive approaches, and that it provides additional information at an individual level. Implications for product design decisions are discussed.

Keywords Feature selection · Support vector machines · Conjoint analysis · L1 norm

1 Introduction

In today's highly competitive market, firms introducing new products require information on consumers' preferences to develop customized offers that meet consumers' needs [32]. Conjoint analysis [16] has proven to be a powerful methodology for identifying such customers' preferences on product features. After collecting conjoint data, diverse econometric methods are typically used to estimate the corresponding preference parameters. This task has been tackled more recently by using machine learning models such as Support Vector Machines (SVM) [8, 9].

One of the critical results of conjoint analysis is the identification of the most important attributes that consumers consider when evaluating product alternatives [27]. When identifying these attributes, researchers usually assume that consumers use all of the attributes when facing product evaluation tasks. However, respondents may ignore some attributes for several reasons, including: (i) lack of knowledge or uncertainty regarding some characteristics, (ii) use of a simple heuristic for choosing between profiles, and/or (iii) the fact that an attribute can truly be irrelevant for the choice [19]. It is expected that consumers might be selective regarding the attributes they take into account when facing complex products characterized by a large number of characteristics, a behavior that is known in the conjoint analysis literature as "attribute non-attendance" [19]. In addition, given the typical limited individual data in

✉ Sebastián Maldonado
smaldonado@uandes.cl

Ricardo Montoya
rmontoya@dii.uchile.cl

Julio López
julio.lopez@udp.cl

[1] Facultad de Ingeniería y Ciencias Aplicadas, Universidad de los Andes, Monseñor Álvaro del Portillo 12455, Las Condes, Santiago, Chile

[2] Department of Industrial Engineering, University of Chile, República 701, Santiago, Chile

[3] Facultad de Ingeniería, Universidad Diego Portales, Ejército 441, Santiago, Chile

conjoint analysis, researchers usually pool the information from the consumers. Therefore, in the context of preference sparsity, it is critical to control for heterogeneity and for simultaneously allowing sparsity at the individual level.

In this paper, we propose a novel technique based on SVM to determine the subset of attributes consumers use to evaluate alternatives in a choice-based conjoint. Instead of using the traditional l_2-SVM formulation for conjoint analysis [9], in which the shrinkage (complexity) is controlled by the Euclidean norm of the part-worths, we use the l_1-norm. This strategy has been used successfully in SVM for classification and provides a good compromise between reducing complexity and allowing for sparsity [35]. The identification of the relevant attributes that each customer uses to evaluate products, with the corresponding reduction in the dimensionality of the customers' preference representations, is achieved by a backward attribute elimination procedure based on the individual weights (part-worths). The final set of part-worths is obtained by using a linear programming approach that simultaneously handles model fit, complexity, and heterogeneity, pooling information across customers by shrinking the individual part-worths of the relevant attributes towards an estimated population mean. Thus, our main goal is to contribute to the consumer analytic literature by providing a new methodology that captures underlying sparsity in consumer preferences in the context of choice based conjoint.

The rest of this work is organized as follows: In Section 2, SVM approaches for conjoint analysis are presented and discussed. Section 3 presents a description of feature selection methods with SVM. Section 4 describes the proposed feature selection method for CBC. Experimental results in both simulated and real-world datasets are presented in Section 5. Finally, a conclusion is given in Section 6, with managerial implications in Analytics and suggested future developments.

2 Conjoint estimation via SVM

In this section, we discuss the relevant SVM formulations for conjoint analysis. First, the notation is introduced in Section 2.1. Section 2.2 presents the SVM formulation for preference estimation in choice-based conjoint (CBC-SVM). Finally, different approaches for heterogeneity control in CBC-SVM are discussed in Section 2.3.

2.1 Notation and preliminaries

Consider a set of N consumers that evaluate K different product profiles, randomly presented, and choose one profile at each choice occasion $t = 1, \dots, T$. Each profile is described by J attributes, and each attribute is defined on n_j levels, $j = 1, \dots, J$. SVM specifies an additive utility function for each customer $i = 1, \dots, N$ of the form $u_i(\mathbf{x}) = \mathbf{w}_i{}^\top \mathbf{x}$ that represents the utility that consumer i assigns to profile $\mathbf{x}$. In our formulation we specify any attribute using dummy coding representing an attribute level, as is usually done in SVM classification with nominal variables [20].

Consumer decisions can be modeled as tuples of the form $\left(\left[\mathbf{x}_{it}^1, \dots, \mathbf{x}_{it}^K\right], y_{it}\right)$, where $\mathbf{x}_{it}^k \in \Re^J$ represent the product profile and $y_{it} \in \{1, \dots, K\}$ represent the chosen profile. The condition $y_{it} = k$ indicates that at choice occasion t individual i prefers the k^{th} alternative among the K profiles described by $\left[\mathbf{x}_{it}^1, \dots, \mathbf{x}_{it}^K\right]$. This implies a series of inequalities $u_i(\mathbf{x}_{it}^{y_{it}}) \geq u_i(\mathbf{x}_{it}^b)$, $\forall b \in \{1, \dots, K\} \setminus \{y_{it}\}$ [8]. Following previous research, after all the responses are collected, the information can be rearranged such that the chosen profile at each occasion t is the first profile in our formulation, i.e. $y_{it} = 1$, $1 \leq i \leq N$ and $1 \leq t \leq T$. Thus, the inequalities can be rewritten as follows:

$$\mathbf{w}_i{}^\top \left(\mathbf{x}_{it}^1 - \mathbf{x}_{it}^k\right) \geq 0, \tag{1}$$

where $1 \leq i \leq N$, $2 \leq k \leq K$, and $1 \leq t \leq T$. This data processing does not affect the solution of the optimization problem.

2.2 CBC-SVM formulation

In order to determine the weights $\mathbf{w}_i$, also known as part-worths, SVM follows the *structural risk minimization principle* [33]. This approach minimizes the Euclidean norm of $\mathbf{w}_i$ with noise penalization performed by slack variables ξ_{kt} (l_2-soft margin formulation), leading to the following convex quadratic programming problem for each individual $i = 1, \dots, N$ [8, 11]:

$$\begin{aligned} \min_{\mathbf{w}_i, \xi} \quad & \frac{1}{2} \|\mathbf{w}_i\|^2 + C \sum_{t=1}^{T} \sum_{k=2}^{K} \xi_t^k \\ \text{s.t.} \quad & \mathbf{w}_i{}^\top \left(\mathbf{x}_{it}^1 - \mathbf{x}_{it}^k\right) \geq 1 - \xi_t^k, \; t = 1, \dots, T, \; k = 2, \dots, K, \\ & \xi_t^k \geq 0, \; t = 1, \dots, T, \; k = 2, \dots, K. \end{aligned} \tag{2}$$

Formulation (2) minimizes the complexity of the model by using the Euclidean norm as a regularizer while penalizing the inconsistencies in the choices. The trade-off between both objectives is controlled by the parameter C, which is usually set via cross-validation (see e.g. [30]). The solution to Formulation (2) provides the individual part-worths $\mathbf{w}_i$.

2.3 Heterogeneity control in CBC-SVM

Conjoint studies usually lack sufficient individual data to estimate individual preference models independently given the limited length of questionnaires. To overcome this issue, hierarchical approaches are used to capture general patterns at the population level by pooling information across customers and simultaneously estimating individual preferences (see e.g. [14]).

Some approaches have been proposed to perform this pooling in the context of SVM. The first strategy, proposed by [11], computes a population part-worth $\overline{\mathbf{w}} = 1/N \sum_i \mathbf{w}_i$ after performing the main optimization process. The final individual part-worths correspond to the weighted sum $\gamma_i \mathbf{w}_i + (1 - \gamma_i)\overline{\mathbf{w}}$ of the initial individual parth-worths and the a-posteriori population part-worth. Parameter $\gamma_i \in [0, 1]$ is usually obtained via a cross-validation procedure [11].

Another alternative for heterogeneity control was proposed by [8], in which a single optimization problem was proposed for individual part-worth estimation while pooling information across customers. Similarly, [12] suggest a single formulation to obtain all part-worths, in which the weights are shrunk towards a vector $\mathbf{w}_0$, whose components are also decision variables.

Our methodology follows the latter shrinkage approach. Specifically, after performing the attribute selection step, we estimate the part-worths by shrinking the weights towards an average weight vector $\mathbf{w}_0$ which is estimated simultaneously. To encourage sparseness, we consider the l_1-norm instead of the traditional Euclidean norm. This procedure is further described in Section 4.

The main difference between our approach and the one proposed in [12] is that the latter is designed for part-worth estimation exclusively, and does not include a regularization strategy to encourage feature selection. A second difference is that in [12] the authors include an additional matrix D, related to the covariance matrix of the partworths, which is estimated by using the calibration data. Since such a formulation is complex to solve due to non-linearity, the authors therefore use a simple iterative strategy to obtain $\mathbf{w}_i$ and $\mathbf{w}_0$ given D, and then D, given $\mathbf{w}_i$ and $\mathbf{w}_0$. Our proposal does not include this term since it does not provide additional information to our problem of individual sparsity, which allows us to solve a single LP without the need of performing a probably quite time-consuming iterative process.

3 Feature selection for SVM

Feature selection is a very important artificial intelligence task that has several applications in various domains, such as churn prediction [23], fault detection [7], or those in the environmental sciences [29]. The goal of feature selection is to find a subset of the original attributes in order to improve predictive performance, removing noisy, irrelevant, and redundant variables that could lead to overfitting [5]. Feature selection is of primary interest in Business Analytics since it improves the understanding of the decision process [23], leading to important managerial insights regarding customer preferences in the context of conjoint analysis [24].

There are several strategies for performing feature selection in SVM. The most common strategy is to remove irrelevant attributes before constructing a predictive model using statistical measures to assess relevancy or redundancy [18]. Another approach is the use of search strategies to evaluate various subsets of features using SVM models. The selection of those subsets is a combinatorial problem. Thus, different simplifying heuristics are used to select those subsets, and the decision is made according to the highest accuracy among the models. Although these strategies are simple to implement and powerful, because they consider the interactions among variables and their influence in the multivariate model, their main disadvantage is the high computational cost, especially in high-dimensional data sets [18].

One popular search strategy is the Sequential Backward Elimination approach that consists of a sequential elimination of features, starting with the complete set of variables. This strategy was adapted by [17] in their Recursive Feature Elimination (RFE) SVM method, in which the variable to be eliminated in each iteration is the one whose removal has less impact on the SVM solution. That is, the variable j to be eliminated is the one with the lowest

$$\left| \|\mathbf{w}\|^2 - \left\|\mathbf{w}_{(j)}\right\|^2 \right|, \tag{3}$$

where $\mathbf{w}_{(j)}$ is the computation of the weight vector with variable j removed. Feature selection can also be performed by encouraging variable elimination directly via a sparsity term in the objective function [6, 21]. For instance, the squared Euclidean norm from the classical SVM ($\|\mathbf{w}\|^2$ from Formulation (2)) can be replaced by the l_1-norm $\Omega(\mathbf{w}) = \sum_j |w_j|$, as presented in the l_1 Support Vector Machine (l_1-SVM) approach [2, 6]. The l_1-SVM method has been further extended to other SVM strategies, such as twin SVM [13], or sparse linear SVM [4]. Besides performing variable elimination while training the models, the l_1-norm can be used to speed up the training process by transforming the quadratic programming (QP) problem solved by standard SVM to a linear programming (LP) problem [6]. Alternatively, SVM can be solved efficiently via linear approximations in combination with advanced optimization techniques (e.g. stochastic gradient descent [10]), or by combining SVM with sampling strategies (see, e.g. [26]).

[24] proposed an embedded feature selection approach for CBC using SVM. SVM-RFE was extended to conjoint analysis, solving Formulation (2) for individual part-worth estimation and subsequently removing irrelevant attributes iteratively for each consumer. Heterogeneity control was

performed as suggested in [11], i.e. computing a population part-worth $\overline{\mathbf{w}}$ and controlling the trade-off via cross-validation using a parameter γ.

There are two main differences between the current approach and previous research: regularization and sparsity. In the present work, we introduce regularization by solving a single optimization problem to obtain all individual part-worths simultaneously while controlling for heterogeneity. This represents an improvement in this context as suggested in [12] in a related context. The second difference is the use of the l_1-norm instead of the Euclidean norm to induce sparsity and encourage attribute elimination. We investigated the relative contribution in performance of these two additions.

4 Proposed method l_1-SVM for conjoint estimation

In this section, we present an embedded feature selection algorithm for CBC using Support Vector Machines. The main contributions are two fold: the use of the l_1-norm instead of the Euclidean norm to induce sparsity by shrinking the individual part-worths towards zero, and the control for heterogeneity in both consumer preferences and attribute selection.

Allowing for individual sparsity and simultaneously controlling for heterogeneity presents an interesting challenge. The strategy of pooling individual part-worths towards a population mean could remove the capability of the method to select attributes at the individual level. Thus, feature selection and heterogeneity control might imply opposite goals. However, since our hypothesis is that most attributes are relevant at the population level, but that individuals may ignore some of them, both objectives cannot be accomplished simultaneously and successfully in one step. Indeed, a one-step procedure might kill individual sparsity unless the relative weight of the l_1-norm is high compared to fit. Still, this will most likely hurt the model fit badly. Consequently, we propose a sequential optimization procedure to accommodate these two goals.

In the first stage we perform feature selection at the individual level without heterogeneity control. At this stage, we estimate the part-worths only to determine the relative importance of each attribute. We need to remove all part-worths corresponding to the eliminated attribute to perform the feature selection. We use a backward elimination algorithm for each individual to remove the attributes whose contribution is small, based on the magnitude of their part-worths. In the second stage, we re-estimate the individual part-worths for the selected attributes using heterogeneity control. That is, the population estimates will affect only the estimates of the individuals for whom the attribute is relevant. Below, we present the sequential optimization procedure.

1. Initial part-worth estimation: For each customer $i \in \{1, \ldots, N\}$, the individual part-worths are obtained by solving the following optimization problem:

$$\begin{aligned}
\min_{\mathbf{w}_i, \xi_{it}^k} \quad & \|\mathbf{w}_i\|_1 + C \sum_{t=1}^{T} \sum_{k=2}^{K} \xi_{it}^k \\
\text{s.t.} \quad & \mathbf{w}_i^\top (\mathbf{x}_{it}^1 - \mathbf{x}_{it}^k) \geq 1 - \xi_{it}^k, \; i=1,\ldots,N, \; t=1,\ldots,T, \; k=2,\ldots,K, \\
& \xi_{it}^k \geq 0, \; i=1,\ldots,N, \; t=1,\ldots,T, \; k=2,\ldots,K.
\end{aligned} \tag{4}$$

Although the L1 norm is a non-smooth function, the previous formulation can be transformed into a linear optimization problem by adding a positive variable $\mathbf{z}_i$, which relates to the weight vector via the following constraints: $-\mathbf{z}_i \leq \mathbf{w}_i \leq \mathbf{z}_i$ for all $i = 1, \ldots, N$ (see [6] for details). The LP formulation for this step follows:

$$\begin{aligned}
\min_{\mathbf{w}_i, \mathbf{z}_i, \xi_{it}^k} \quad & \sum_{j=1}^{J} z_{ij} + C \sum_{t=1}^{T} \sum_{k=2}^{K} \xi_{it}^k \\
\text{s.t.} \quad & \mathbf{w}_i^\top (\mathbf{x}_{it}^1 - \mathbf{x}_{it}^k) \geq 1 - \xi_{it}^k, \; i=1,\ldots,N, \; t=1,\ldots,T, \; k=2,\ldots,K, \\
& \xi_{it}^k \geq 0, \; i=1,\ldots,N, \; t=1,\ldots,T, \; k=2,\ldots,K, \\
& -\mathbf{z}_i \leq \mathbf{w}_i \leq \mathbf{z}_i, \; i=1,\ldots,N.
\end{aligned} \tag{5}$$

2. Attribute elimination step: We assess the attribute's relevance by computing the difference between the highest and the lowest part-worths [15]. Formally, we define the attribute contribution AC_j for attribute j as:

$$AC_j(\mathbf{w}_i^j) = \max \; \mathbf{w}_i^j - \min \; \mathbf{w}_i^j. \tag{6}$$

For each customer $i \in \{1, \ldots, N\}$, we remove those attributes that are irrelevant (i.e. with a low AC) in the corresponding utility function using a backward elimination procedure. Algorithm 1 presents this procedure.

Algorithm 1 Algorithm for attribute elimination with linear l_1-SVM for CBC

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

(a) For each respondent $i = 1, \ldots, N$ do:
(b)   $\mathcal{S}_i \leftarrow \mathcal{S}$
(c)   repeat
(d)     $\mathbf{w}_i \leftarrow l_1$-SVM Training, Formulation (5), using $\mathcal{S}_i$
(e)     $\{j\} \leftarrow \text{argmin}_j \; AC_j(\mathbf{w}_i^j)$
(f)     if $AC_j(\mathbf{w}_i^j) < \epsilon$ then $\mathcal{S}_i \leftarrow \mathcal{S}_i \setminus \{j\}$
(g)   until $AC_j(\mathbf{w}_i^j) > \epsilon \; \forall j$ or $|\mathcal{S}_i| = 1$
(h) end.
```

The parameter $\epsilon \geq 0$ corresponds to a *relevance threshold* for the relative contribution of each attribute. This threshold needs to be sufficiently small to avoid eliminating relevant attributes. The stopping criterion is reached when the contribution of all remaining attributes is above this threshold, or when only one attribute remains.

3. Preference estimation: We applied the following model to obtain the final part-worths, considering only the relevant attributes for each customer:

$$\begin{aligned}
\min_{\mathbf{w}_i,\mathbf{w}_0,\xi_{it}^k} \quad & \sum_{i=1}^{N} (\|\mathbf{w}_i\|_1 + \theta\|\mathbf{w}_i - \mathbf{w}_0\|_1) + C\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{k=2}^{K}\xi_{it}^k \\
\text{s.t.} \quad & \mathbf{w}_i^\top(\mathbf{x}_{it}^1 - \mathbf{x}_{it}^k) \geq 1 - \xi_{it}^k, \ i=1,\ldots,N, \ t=1,\ldots,T, \ k=2,\ldots,K, \\
& \xi_{it}^k \geq 0, \ i=1,\ldots,N, \ t=1,\ldots,T, \ k=2,\ldots,K, \\
& \mathbf{w}_i^j = 0, \ i=1,\ldots,N, \ j \notin \mathcal{S}_i, \qquad (7)
\end{aligned}$$

with control parameters $\theta > 0$ and $C > 0$. Part-worths associated with attributes removed in the previous stage are not estimated but are fixed to zero and do not participate in the estimation of the population part-worths. The procedure yields individual utility functions that consider only the relevant attributes for each customer, and also yields the population pattern that reflects the preferences for individuals with non-zero preferences for each attribute. Note that solving Formulation (7) without performing feature selection leads necessarily to non-sparse solutions since the term $\|\mathbf{w}_i - \mathbf{w}_0\|_1$ will pool information across customers and shift part-worths linked to irrelevant attributes towards $\mathbf{w}_0$ instead of to zero.

Similarly to Formulation (5), the previous problem can be cast into an LP model by the inclusion of new variables. The linear version for Formulation (7) follows:

$$\begin{aligned}
\min_{\mathbf{w}_i,\mathbf{w}_0,\mathbf{z}_i,\mathbf{u}_i,\xi_{it}^k} \quad & \sum_{i=1}^{N}\sum_{j=1}^{J}\left(z_{ij} + \theta u_{ij}\right) + C\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{k=2}^{K}\xi_{it}^k \\
\text{s.t.} \quad & \mathbf{w}_i^\top(\mathbf{x}_{it}^1 - \mathbf{x}_{it}^k) \geq 1 - \xi_{it}^k, \ i=1,\ldots,N, t=1,\ldots,T, k=2,\ldots,K, \\
& \xi_{it}^k \geq 0, \ i=1,\ldots,N, t=1,\ldots,T, k=2,\ldots,K, \\
& \mathbf{w}_i^j = 0, \ i=1,\ldots,N, \ j \notin \mathcal{S}_i, \\
& -\mathbf{z}_i \leq \mathbf{w}_i \leq \mathbf{z}_i, \ i=1,\ldots,N, \\
& -\mathbf{u}_i \leq \mathbf{w}_i - \mathbf{w}_0 \leq \mathbf{u}_i, \ i=1,\ldots,N. \qquad (8)
\end{aligned}$$

The proposed approach has several advantages compared with other state-of-the-art techniques:

- Regularization and model fit: The approach reduces the complexity of the model by using the structural risk minimization principle [33] via the l_1-norm, providing a good compromise between model fit and robustness, which usually translates into better out-of-sample performance.
- Identification of the relevant attributes: The feature selection algorithm mitigates the effects of the *curse of dimensionality* by reducing the number of decision variables in the individual utility functions, and simultaneously provides a better understanding of the preference model.
- Heterogeneity control: The model incorporates general population patterns into each customer preference model, resulting in a general model that takes all available information into account to derive individual preferences.
- Model efficiency: The feature selection framework is based on several small linear programming models, which can be solved efficiently using standard optimization solvers.

5 Experimental results

We applied the proposed approach to four simulated datasets and two real-world conjoint applications. In Section 5.1, we describe the experimental setting and the conjoint data. We also describe the implementation of the proposed approach and the performance metrics. In Section 5.2, we present the results obtained for all the proposed and alternative approaches. Managerial insights and the relevance of this work in Analytics are discussed in Section 5.3.

5.1 Datasets and experimental setting

5.1.1 Simulated datasets

We designed a simulation exercise to study how our approach performs under different conditions. In particular we studied fit and predictive ability as a function of feature usage rate and response error. To study these effects, and for easy comparison with previous research findings [3, 31], we used the same standard simulation procedure that was used by these researchers, but modified it as needed to include irrelevant attributes. We generated various datasets varying the noise condition in consumer choices (low and high noise), and the number of irrelevant attributes (low and high).

For each of the four datasets, $N = 200$ consumers were simulated. Each individual chose its best alternative among $K = 3$ product profiles for each of the $T = 12$ choice occasions. From the 12 choice occasions, 10 questions were used for model training and calibration, while the remaining two were used for testing. For each profile, $J = 10$ variables were used to describe the simulated products, and each attribute j was further described by $n_j = 4$ levels.

We used the following standard procedure to vary the amount of noise and to create irrelevant variables: Each attribute was 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 denotes the identity matrix. Following [24], we created irrelevant attributes by setting $\boldsymbol{\mu} = \mathbf{0}$ for those attributes. Two and six irrelevant attributes were generated at an individual level for the low sparseness and high sparseness conditions, respectively. Following [3], values of $\beta = 0.5$ and $\beta = 2$ were used to create high and low noise conditions, respectively.

5.1.2 CBC dataset 1: Lower-dimensional dataset

For this dataset, $N = 125$ subjects responded to 20 choice questions about digital cameras, with each choice question comprised of four product profiles. 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). From the 20 choice occasions, 16 questions were used for model training and calibration, while the remaining four were used for testing. See [1] for further details about the conjoint experiment.

5.1.3 CBC dataset 2: Higher-dimensional dataset

This second CBC application consists of $N = 602$ individuals that face 12 choice occasions involving three product profiles. Each product is described by $J = 10$ attributes, with each attribute having a different number of levels: three attributes have 3 levels, five attributes have 4 levels, one attribute has 7 levels, and one attribute has 16 levels. Ten of the questions were used for training and the tuning of the parameters, while the remaining two were used for testing purposes. Due to the proprietary nature of the data, the actual product and the specific attributes and attribute levels are not mentioned.

5.1.4 Proposed and benchmark models

The following approaches were used in the experimental section:

1. l_1-SVM: Proposed 1-norm SVM model for feature selection and heterogeneity control, including the backward elimination process.

   For a more thorough comparison, in addition to our proposed model we estimated various preference models:
2. l_2-SVM: Traditional 2-norm Linear SVM using individual part-worths (Formulation (2)) without feature selection.
3. l_1-SVM$_{\epsilon=0}$: Proposed 1-norm SVM model for feature selection and heterogeneity control with $\epsilon = 0$, i.e. without the backward elimination process.
4. l_1-SVM$_{\theta=0}$: Proposed 1-norm SVM model for feature selection with $\theta = 0$, i.e. no heterogeneity control.
5. l_2-SVM-RFE: 2-norm Linear SVM allowing feature selection and including heterogeneity control.

   We also include the l_2 formulation of Problem (7) as an additional benchmark to assess the gain in terms of sparsity of our proposal compared to the proposed l_1-SVM model. We first solve the Formulation (9), and then Algorithm 1 is applied.

$$\begin{aligned} \min_{\mathbf{w}_i, \mathbf{w}_0, \xi_{it}^k} \quad & \frac{1}{2} \sum_{i=1}^{N} \left(\|\mathbf{w}_i\|_2^2 + \theta \|\mathbf{w}_i - \mathbf{w}_0\|_2^2 \right) + C \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{k=2}^{K} \xi_{it}^k \\ \text{s.t.} \quad & \mathbf{w}_i^\top (\mathbf{x}_{it}^1 - \mathbf{x}_{it}^k) \geq 1 - \xi_{it}^k, \; i=1,\ldots, N, \; t=1,\ldots, T, \; k=2,\ldots, K, \\ & \xi_{it}^k \geq 0, \; i = 1, \ldots, N, \; t = 1, \ldots, T, \; k = 2, \ldots, K. \end{aligned} \tag{9}$$

6. HB Mixed Logit: Mixed Logit model where each attribute level is represented by a binary variable. We use a hierarchical Bayesian Markov chain Monte Carlo (MCMC) procedure [28], which is the state-of-the-art approach for estimating the parameters of the Mixed Logit model (see the complete specification of the priors and full conditional distributions in Appendix A).

We used the SLINEARSOLVE and QUADSOLVE solvers for the linear and quadratic programming SVM implementations, respectively. Both solvers can be found in Matlab's Spider toolbox [34]. The HB Mixed Logit method was also implemented in Matlab as in [11] and [12].

5.1.5 Performance metrics

We estimate the preference models mentioned above and compare them based on the following performance measures:

1. In-sample hit rate: Average number of correctly predicted choices for the calibration data.
2. Out-of-sample hit rate: Average number of correctly predicted choices for the holdout data.
3. Feature usage rate ($FU - rate$): Average number of attributes used by customers. We compute this measure as follows:

$$FU - rate = \frac{\sum_{i=1}^{N} |\mathcal{S}_i|}{N \cdot J}, \tag{10}$$

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

5.1.6 Tuning of the parameters

Our method has the important advantage of being a linear programming problem with a single solution. In contrast to HB Mixed Logit, SVM does not make any assumptions regarding the part-worth distribution, automating the model identification process [9]. As a consequence, no random term is present, and the simulation step used in HB Mixed Logit, which can be expensive in terms of running times, is avoided. Nevertheless, SVM requires several parameters that need to be tuned via cross validation, which can be seen as a disadvantage compared to HB Mixed Logit or non-compensatory approaches.

We need to calibrate the parameter C for all SVM models. For the more general formulations we need to calibrate either the parameters ϵ, or θ, or both. We use a leave-one-out cross validation (LOOCV) strategy to tune these parameters using only the training data. This LOOCV procedure considers a subset of the training data comprising all questions but one, for each individual. The individual part-worths are then estimated using this subset, and used subsequently 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 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 a conjoint analysis context [11, 12, 30].

Based on previous research [22], we explored the following sets applying grid search, covering the combinations of a relatively wide range of possible values:

$$C, \theta \in \{2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^{0}, 2^{1}, 2^{2}, 2^{3}, 2^{4}, 2^{5}\},$$
$$\epsilon \in \{0.0, 0.1, 0.25, 0.5, 0.75, 1.0\}.$$

5.2 Results

Synthetic data The results of the simulation show that the performance of HB Mixed Logit decreases as the level of sparseness (attributes ignored by customers) increases, whereas, in contrast, the performance of the SVM approach with and without feature selection does not decrease with the level of sparseness. Additionally, consistent with machine learning and forecasting applications, selecting the most relevant attributes improves the predictive performance: the proposed method outperforms alternative approaches in terms of out-of-sample hit rate as the level of noise and sparseness increase.

Tables 1 and 2 presents the results for the proposed method and the alternative approaches for the four simulated datasets. We indicate with an asterisk the best predictive hit rate or that which is not significantly different from the best at the 5 % significance level.

The following results can be derived from these experiments:

- Interestingly, l_1-SVM performs similarly to l_1-SVM$_{\epsilon=0}$ in terms of predictive performance but provides a more parsimonious representation. That is, hold-out hit rate is not affected by the backward elimination procedure although the elimination of the less relevant attributes is substantial in l_1-SVM compared to l_1-SVM$_{\epsilon=0}$ (55-70 % vs 30-55 % of the attributes).
- In contrast, l_1-SVM achieves significantly higher performance compared to l_1-SVM$_{\theta=0}$ in two out of four conditions, demonstrating that pooling information across consumers is a crucial step of the algorithm.

Table 1 Results for Preference Models (in percentages)

	Low Noise – Low Sparsity			Low Noise – High Sparsity		
	Hit rate			Hit rate		
Models	In[a]	Out[b]	FU-rate[c]	In	Out	FU-rate
HB Mixed Logit	87.5	56.3*	100	82	50.3	100
l_2-SVM	98.1	51.8	100	98.3	54.0	100
l_1-SVM$_{\epsilon=0}$	98.8	54.0*	71.6	98.0	59.8*	65.4
l_1-SVM$_{\theta=0}$	98.8	52.0	72.6	97.7	59.8*	65.4
l_2-SVM-RFE	95.8	54.3*	54.4	91.2	59.0*	34.1
l_1-SVM	97.1	54.3*	48.0	93.0	60.5*	30.5

[a] In-sample hit rate. [b] Out-of-sample hit rate. [c] Feature usage rate.

Table 2 Results for Preference Models (in percentages)

	High Noise – Low Sparsity			High Noise – High Sparsity		
	Hit rate			Hit rate		
Models	In[a]	Out[b]	FU-rate[c]	In	Out	FU-rate
HB Mixed Logit	76.4	43.5	100	73.3	41.2	100
l_2-SVM	99.7	44.0	100	98.0	47.3*	100
l_1-SVM$_{\epsilon=0}$	95.0	45.3	61.9	92.3	50.0*	55.1
l_1-SVM$_{\theta=0}$	94.3	44.3	63.5	93.0	48.3*	49.6
l_2-SVM-RFE	83.1	47.0*	31.0	98.1	48.5*	70.5
l_1-SVM	86.6	49.8*	28.7	82.7	50.0*	27.6

[a] In-sample hit rate. [b] Out-of-sample hit rate. [c] Feature usage rate.

- l_1-SVM and l_2-SVM-RFE achieved relatively similar results (although the former has a slightly better performance), but the 2-norm formulation always selects more attributes than the 1-norm formulation when using the same value for parameter ϵ, demonstrating the advantage of the l_1-regularization in achieving sparser solutions.
- l_2-SVM presents some overfitting since the in-sample fit is the highest, but out-of-sample predictions are not better than those of the proposed method. The main source of overfitting seems to be the reduced number of questions per individual relative to the large number of parameters that needs to be calibrated.
- The performance of HB Mixed Logit decreases as the level of sparseness (attributes ignored by customers) increases, whereas, by contrast, the performance of the SVM approach with and without feature selection does not decrease with the level of sparseness.
- The performance of the different methods improves as the noise decreases, which is somehow to be expected.
- Finally, selecting the most relevant attributes improves the predictive performance: the proposed l_1-SVM method outperforms alternative approaches in terms of out-of-sample hit rate as the level of noise and sparseness increase.

In summary, the proposed method achieved the best predictive performance while having the lowest gap between in-sample and out-of-sample hit rate among the SVM approaches: the average gaps are 49.3, 43.8, 44.9, 39.9, and 36.2 for l_2-SVM, l_1-SVM$_{\epsilon=0}$, l_1-SVM$_{\theta=0}$, l_2-SVM-RFE, and the proposed l_1-SVM, respectively. These experiments demonstrate the advantages of using feature selection and heterogeneity control in reducing the magnitude of the overfitting by reducing the number of parameters to be estimated.

CBC datasets Recall that the CBC dataset 1 involves 125 subjects and product profiles described by 5 attributes whereas CBC dataset 2 is a higher-dimensional dataset that involves 602 subjects choosing among product profiles described by 10 attributes. Table 3 presents the results for both real-world CBC applications. We indicate the best predictive hit rate or those which are not significantly different from the best at the 5 % level with an asterisk.

Table 3 shows that the proposed feature selection methodology has the best average performance in terms of

Table 3 Comparison of the Preference Models in two real-world datasets

	CBC dataset 1			CBC dataset 2		
	Hit rate			Hit rate		
Models	In[a]	Out[b]	FU-rate[c]	In	Out	FU-rate
HB Mixed Logit	84.5	58.0*	100	70.4	57.2*	100
l_2-SVM	92.3	56.4	100	95.0	58.6*	100
l_1-SVM$_{\epsilon=0}$	91.6	58.4*	77.3	94.7	56.3	49.5
l_1-SVM$_{\theta=0}$	87.1	56.4	72.6	95.9	56.0	54.8
l_2-SVM-RFE	86.0	58.2*	65.6	92.9	54.5	33.3
l_1-SVM	85.9	59.8*	64.0	74.5	58.0*	14.3

[a] In-sample hit rate. [b] Out-of-sample hit rate. [c] Feature usage rate.

the out-of-sample hit rate. Best performance is achieved for the digital camera dataset, while the methods l_2-SVM, the proposed l_1-SVM, and the HB Mixed Logit models yield superior predictive performance for the higher-dimensional data.

Notably, the best results are obtained with l_1-SVM using only 64 % of the features on average across customers for the digital camera dataset. We note that the backward elimination procedure in the proposed method yields slightly higher predictive performance compared to l_1-SVM$_{\epsilon=0}$ although it considers a lower number of attributes (64.0 % vs 77.3 %). For the second dataset, it is noteworthy that the proposed model (l_1-SVM) considers on average only 14.3 % of the attributes across customers.

Another important issue is the influence of the different parameters. For the digital camera datasets, we explored the performance of parameters C, ϵ, and θ in the predictive performance (LOO validation hit rate). We observed a fairly stable performance for these parameters using a reasonably large set of values, concluding that our proposal is not strongly parameter-dependent. The detailed analysis is presented in Appendix B.

In both real-world datasets, l_1-SVM outperforms l_1-SVM$_{\epsilon=0}$ and l_1-SVM$_{\theta=0}$, demonstrating the importance of accounting for both feature selection and heterogeneity control simultaneously since the full method is the only one that achieves the same predictive performance as the best model using a significantly lower number of attributes. The most important gain compared to the full method, however, is given by the better interpretation of the results, as we will explain in the following section.

Regarding running times, the comparison between HB Mixed Logit and SVM-based approaches is not straightforward since the latter methods require an extra step in order to calibrate the parameters. Using the digital camera dataset (CBC dataset 1) as an illustrative example, the running time for HB was 1057.13 seconds, while the training time for our approach was 13.5 for a given configuration of C, ϵ, and θ. Although solving one instance of our approach is about 100 times faster than solving one instance of the HB method, we performed a grid search for these three parameters, therefore needing to run our method 11*11*6 times, leading to a total running time of approx. 9801 seconds. In both cases, running times are tractable and marginal compared to the effort of data collection and preprocessing.

5.3 Managerial insight and impact in consumer analytics

The proposed model achieved slightly better results in general compared to the alternative methods but used fewer attributes. In particular, it selected 3.2 out of the five, and 1.4 out of ten attributes for the first and second CBC datasets, respectively. Here, the gain for decision making is the identification of the relevant attributes at the individual level, allowing consumer segmentation according to the usage of the attributes.

A second analysis that is important for product design is the percentage of individuals that consider a given attribute to be useful according to the proposed model. For the digital cameras dataset, attribute **Price** was used in 84.0 % of the utility functions, **Resolution** was used in 73.6 % of the utility functions, **Battery Life** was used in 58.4 % of the utility functions,**Optical Zoom** was used in 63.2 % of the utility functions, and **Camera Size** was used in 40.8 % of the utility functions. This information provides a ranking of relevancy for the attributes, which can be useful in product design. These results confirm our previous finding that Camera size and Battery Life are less relevant attributes for consumers. Indeed, their elimination by most respondents does not affect the predictive ability of the proposed model.

It is equally important to note that 16 % of the subjects do not consider price at all. These price-insensitive customers constitute an important niche that premium brands should explore. It is important to note that this price insensitivity applies only within the price range presented to respondents. That is, it is possible that outside this price range, these customers might become more price sensitive.

We further studied the implied willingness to pay (WTP) estimates that the proposed model derives compared with

Table 4 Willingness to pay per changes in attribute levels for the proposed SVM-based feature selection approach and the HB Mixed Logit. The numbers are understood as follows: for the proposed model, a change from Resolution 2 to 5MP is equivalent to a change in $264.9

Models	Attributes for digital cameras			
	Resolution 2 vs 5MP	Battery Life 150 vs 600p	Optical Zoom 2 vs 5X	Camera Size SLR vs U. compact
l_1-SVM	$264.9	$158.9	$299.3	$128.5
HB Mixed Logit	$146.5	$53.4	$131.2	$39.3
HB Mixed Logit*	$182.9	$85.2	$209.3	$95.9

* Considering the same respondents as the l_1-SVM model.

those of the HB Mixed Logit. Following the standard method in the conjoint literature, we compared the equivalent change in utility points that correspond to changes in prices, and then used that relationship to compare changes in the other attributes. Table 4 reports the median estimates for both models. Note that for the proposed model the median is calculated only among consumers for whom the corresponding attribute is relevant. Consistent with previous research, we observed that the HB Mixed Logit that does not accommodate for irrelevant attributes yields substantially lower estimates for the WTP compared with our proposed model.

To study if this difference is due to selection of respondents, we also reported the results for the HB Mixed Logit, but included the same respondents that we used for the l_1-SVM model. Although the WTP estimates increased, they are still lower than the WTP indicated by the proposed method.

6 Discussion and conclusions

In this paper, we address the problem of identifying the relevant attributes consumers use when evaluating alternatives in a choice-based conjoint (CBC) task. We present a new methodology based on SVM to bring these relevant attributes to light. We adapt the l_1-SVM formulation to identify relevant attributes in CBC at the individual level, and to control for heterogeneity by pooling data across consumers simultaneously with the optimization procedure.

We compared the proposed approach with several benchmark models including nested versions of the proposed model. Our approach always yielded at least equivalent predictive performance results to that of the best benchmark model although it considers only a fraction of the attributes. The results of our illustrative examples show that consumers may use a small fraction of the available attributes to evaluate the alternatives. The relevant attributes, however, differ importantly across subjects. It is therefore imperative that the identification of the relevant attributes be performed individually instead of at the population level. At the same time, controlling for heterogeneity borrows information from the rest of the population to improve the recovery of individual preferences. Since these two objectives may contradict each other, we use a leave-one-out method to calibrate the weight of these components properly in the SVM formulation that yields high predictive performance on the holdout sample. We contribute to the consumer analytics literature by offering a method that can select the most relevant attributes at the individual level without sacrificing predictive performance. Further we provide additional evidence of the usefulness of using machine learning techniques, such as SVM, to analyze conjoint data in an attribute non-attendance context.

The present methodology is a first step toward using machine learning methods to address the problem of attribute non-attendance, a research problem that is attracting the attention of both researchers and practitioners. Current methods cannot easily accommodate this behavior and use survey data to augment the information with explicit information about non-neglected attributes. Other approaches try to infer relevant attributes using latent class methods, assigning customers to a set of predefined clusters based on a limited combination of attributes attended to. Compared to these approaches, our model offers several advantages. It does not require collecting additional information because it uses only choice data, and allows for any combination of relevant/irrelevant attributes at the individual level.

The feature selection method based on SVM that we developed provides a promising tool for addressing consumer behavior beyond non-attendance in CBC. For future work, first, it would be interesting to apply this approach to other conjoint applications such as menu-based conjoint analysis. These data collection methods could induce customers to use simplifying heuristics that ignore attributes at the different stages of the decision process. Second, it would be interesting to further explore if there is a correlation between the number of levels that characterize an attribute and the likelihood of ignoring such an attribute. This issue may have important managerial implications, especially for web retailers. Third, one could explore how data collected through eye-tracking methods can enhance the capabilities of the methodology we developed to identify relevant/irrelevant attributes. Finally, one could study the extension of our proposal to kernel-based SVM for CBC. This is not straightforward since the part-worths cannot be obtained explicitly in the kernel-based formulation, and therefore feature selection cannot be encouraged directly via L1 penalization.

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 projects 1140831 and 1160738. The second author was supported by FONDECYT project 1151395. The third author was supported by FONDECYT project 1160894 and CONICYT Anillo ACT1106. This research was partially funded by the Complex Engineering Systems Institute, ISCI (ICM-FIC: P05-004-F, CONICYT: FB0816).

Appendix A: HB mixed logit estimation

Prior and full conditional distributions

We denote by $\boldsymbol{\theta}_i$ the set of random-effect parameters.

Priors

Random-effect parameters $\boldsymbol{\theta}_i$

$$\begin{aligned}\boldsymbol{\theta}_i &\sim N(\boldsymbol{\mu}_\theta, \Sigma_\theta) \Rightarrow P(\boldsymbol{\theta}_i) \propto \exp\left(\frac{1}{2}(\boldsymbol{\theta}_i-\boldsymbol{\mu}_\theta)^\top \Sigma_\theta^{-1}(\boldsymbol{\theta}_i-\boldsymbol{\mu}_\theta)\right)\\ \boldsymbol{\mu}_\theta &\sim N(\boldsymbol{\mu}_0, \mathbf{V}_0) \Rightarrow P(\boldsymbol{\mu}_\theta) \propto \exp\left(\frac{1}{2}(\boldsymbol{\mu}_\theta-\boldsymbol{\mu}_0)^\top \mathbf{V}_0^{-1}(\boldsymbol{\mu}_\theta-\boldsymbol{\mu}_0)\right)\\ \Sigma_\theta^{-1} &\sim W(df_0, \mathbf{S}_0)\end{aligned}$$

Likelihood

$$L(\text{data}, \{\boldsymbol{\theta}_i\}, \boldsymbol{\mu}_\theta, \Sigma_\theta) = P(\text{data}|\{\boldsymbol{\theta}_i\})P(\{\boldsymbol{\theta}_i\}|\boldsymbol{\mu}_\theta, \Sigma_\theta)P(\boldsymbol{\mu}_\theta)P(\Sigma_\theta),$$

where $P(\text{data}|\{\boldsymbol{\theta}_i\})$ corresponds to the Multinomial Logit model.

Full conditionals

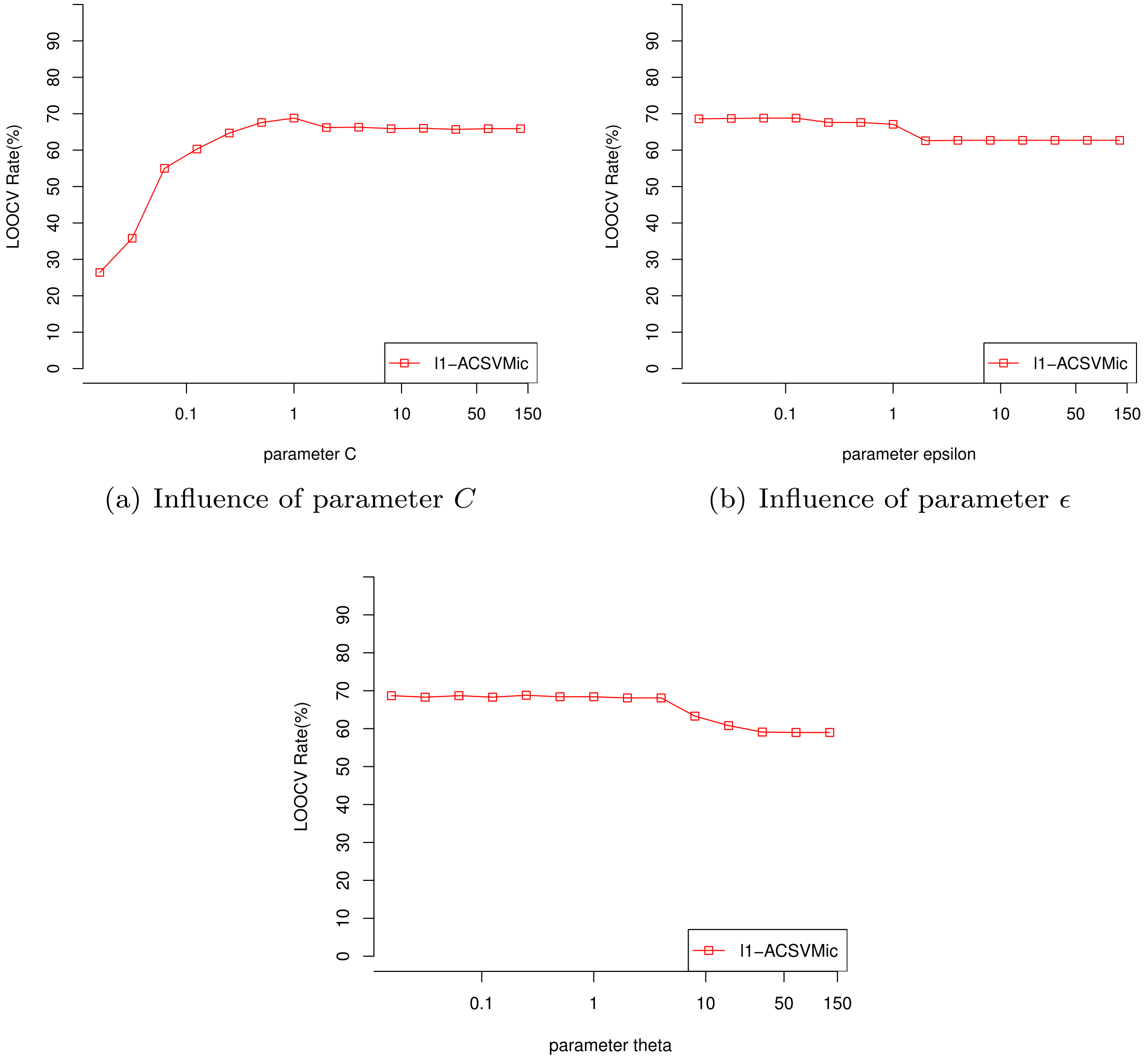
$$P(\boldsymbol{\theta}_i|\boldsymbol{\mu}_\theta, \Sigma_\theta, \text{data}_i) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\theta}_i-\boldsymbol{\mu}_\theta)^\top \Sigma_\theta^{-1}(\boldsymbol{\theta}_i-\boldsymbol{\mu}_\theta)\right) P(\text{data}_i|\boldsymbol{\theta}_i)$$

$$\boldsymbol{\mu}_\theta \sim N(\boldsymbol{\mu}_i, \mathbf{V}_i),\ \Sigma_\theta^{-1} \sim W(df_1, \mathbf{S}_1)$$

where

$$\begin{aligned}\mathbf{V}_i^{-1} &= [\mathbf{V}_0^{-1} + N\Sigma_\theta^{-1}]\\ \boldsymbol{\mu}_i &= \mathbf{V}_i[\boldsymbol{\mu}_0\mathbf{V}_0^{-1} + N\bar{\theta}\Sigma_\theta^{-1}]\\ df_1 &= df_0 + N\\ \mathbf{S}_1 &= \sum_{i=1}^{N}(\boldsymbol{\theta}_i - \boldsymbol{\mu}_\theta)(\boldsymbol{\theta}_i - \boldsymbol{\mu}_\theta)^\top + \mathbf{S}_0^{-1}.\end{aligned}$$

The MCMC procedure generates a sequence of draws from the posterior distribution of the model's parameters. Since the full conditionals for $\boldsymbol{\theta}_i$ do not have a closed form,



(a) Influence of parameter C

(b) Influence of parameter ϵ

(c) Influence of parameter θ

Fig. 1 Leave-one-out validation hit rates for l_1-SVM for different values of C, ϵ, and θ (Camera data set)

the Metropolis-Hastings (M-H) algorithm is used to draw the samples. In particular, we use a Gaussian random-walk M-H where the proposal vector of parameters $\boldsymbol{\varphi}^{(t)}$ for $\boldsymbol{\theta}_i$ at iteration t is drawn from $N(\boldsymbol{\varphi}^{(t-1)}, \sigma^2\Delta)$ and accepted using the M-H acceptance ratio. The tuning parameters σ and Δ are chosen adaptively to yield an acceptance rate of approximately 20 %.

We use the following uninformative prior hyperparameters: $\boldsymbol{\mu}_0 = 0, \mathbf{V}_0 = 10^3\mathbf{I}_{N\theta\times N\theta}, df_0 = N\theta + 5, \mathbf{S}_0 = df_0\mathbf{C}$, where N is the number of individuals, and $\mathbf{C}$ is an $N\theta \times N\theta$ matrix with 2 on the diagonal and 1 off the diagonal for the levels of each attribute. We assume that the parameters are a priori uncorrelated across attributes (see e.g. [25]).

Appendix B: Model calibration and influence of the parameters

In the proposed models, three parameters need to be calibrated: regularization parameter C, threshold ϵ, and shrinkage θ. We analyze how the performance of each model varies as a function of each parameter. For illustration purposes, we show the procedure used for the Camera data set. Similar analyses were conducted for the other data sets. Our goal was to assess whether the results are stable along different values of these parameters. A less rigorous validation strategy can be used in such a case. In contrast, a high variance in the performance requires an exhaustive model selection procedure such as LOOCV in order to find the best combination of parameters.

Figure 1 depicts the LOOCV hit rates as a function of C, ϵ, and θ for the proposed feature selection approach.

Figure 1 reveals the influence of parameters C, ϵ, and θ in the predictive performance (Leave-one-out validation hit rate). Results are relatively stable for small values of θ and ϵ, and values of C around the unit, although we observe an important influence of these parameters in the final outcome of the proposed method.

Performing an adequate grid search is highly recommended, varying the parameters C, ϵ, and θ along the suggested values in order to obtain the desired results. Additionally, the fact that the optimal values for these parameters are always above zero confirms the importance of feature selection and shrinkage to control for potential overfitting when a relatively small number of respondents is present.

References

1. Abernethy J, Evgeniou T, Toubia O, Vert J (2008) Eliciting consumer preferences using robust adaptive choice questionnaires. IEEE Trans Knowl Data Eng 20(2):145–155
2. Argyriou A, Evgeniou T, Pontil M (2008) Convex multi-task feature learning. Mach Learn 73(3):243–272
3. Arora N, Huber J (2001) Improving parameter estimates and model prediction by aggregate customization in choice experiments. J Consum Res 28:273–283
4. Bi J, Bennett K, Embrechts M, Breneman C, Song M (2003) Dimensionality reduction via sparse support vector machines. J Mach Learn Res 3:1229–1243
5. Blum A, Langley P (1997) Selection of relevant features and examples in machine learning. Artif Intell 97(1-2):245–271
6. Bradley P, Mangasarian O (1998) Feature selection via concave minimization and support vector machines. In: Shavlik J (ed) Proceedings of the Fifteenth International Conference on Machine Learning (ICML'98), Morgan Kaufmann, San Francisco, California, pp 82-90
7. Cerrada M, Sánchez RV, Pacheco F, Cabrera D, Zurita G, Li C (2016) Hierarchical feature selection based on relative dependency for gear fault diagnosis. Appl Intell 44(3):687–703
8. Chapelle O, Harchaoui Z (2005) A machine learning approach to conjoint analysis. Adv Neural Inf Proces Syst 17:257–264
9. Cui D, Curry D (2005) Prediction in marketing using the support vector machine. Mark Sci 24(4):595–615
10. Djuric N, Lan L, Vucetic S, Wang Z (2013) Budgetedsvm: A toolbox for scalable svm approximations. J Mach Learn Res 14:3813–3817
11. Evgeniou T, Boussios C, Zacharia G (2005) Generalized robust conjoint estimation. Mark Sci 24(3):415–429
12. Evgeniou T, Pontil M, Toubia O (2007) A convex optimization approach to modeling heterogeneity in conjoint estimation. Mark Sci 26(6):805–818
13. Gao S, Ye Q, Ye N (2011) 1-norm least squares twin support vector machines. Neurocomputing 74(17):35903597
14. Gelman A, Pardoe I (2006) Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. Technometrics 48(2):241–251
15. Green PE, Rao VR (1971) Conjoint measurement for quantifying judgmental data. J Mark Res 8:355–363
16. Green PE, Krieger AM, Wind Y (2001) Thirty years of conjoint analysis: Reflections and prospects. Interfaces 31(3):S56–S73
17. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
18. Guyon I, Gunn S, Nikravesh M, Zadeh LA (2006) Feature extraction foundations and applications. Springer, Berlin
19. Hensher DA, Rose JM, Greene WH (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
20. Hsu CW, Chang CC, Lin C (2010) A practical guide to support vector classification
21. Le Thi H, Pham Dinh T, Thiao M (2016) Efficient approaches for l2-l0 regularization and applications to feature selection in svm. Applied Intelligence In press 45(2):549–565
22. Maldonado S, Weber R, Basak J (2011) Kernel-penalized svm for feature selection. Inf Sci 181(1):115–128
23. Maldonado S, Flores A, Verbraken T, Baesens RBW (2015a) Profit-based feature selection using support vector machines - general framework and an application for customer churn prediction. Appl Soft Comput 35:740–748
24. Maldonado S, Montoya R, Weber R (2015b) Advanced conjoint analysis using feature selection via support vector machines. Eur J Oper Res 241(2):564–574
25. Orme B (2005) The cbc/hb system for hierarchical bayes estimation
26. Pan X, Xu Y (2016) Two effective sample selection methods for support vector machine. J Intell Fuzzy Syst 30:659–670
27. Rao VR (2014) Applied conjoint analysis. Springer

28. Rossi PE, Allenby GM, McCulloch R (2005) Bayesian statistics and marketing. Wiley, New York
29. Shen Q, Jensen R (2008) Approximation-based feature selection and application for algae population estimation. Appl Intell 28(2):167–181
30. Toubia O, Evgeniou T, Hauser J (2007a) Optimization-based and machine-learning methods for conjoint analysis: Estimation and question design. Conjoint Measurement p 231
31. Toubia O, Hauser J, Garcia R (2007b) Probabilistic polyhedral methods for adaptive choice-based conjoint anaysis. Mark Sci 26(5):596–610
32. Tsai HC, Hsiao SW (2004) Evaluation of alternatives for product customization using fuzzy logic. Inf Sci 158:233–262
33. Vapnik V, Chervonenkis A (1991) The necessary and sufficient conditions for consistency in the empirical risk minimization method. Pattern Recognit Image Anal 1(3):283–305
34. Weston J, Elisseeff A, BakIr G, Sinz F (2005) The spider machine learning toolbox. Software available at http://www.kyb.tuebingen.mpg.de/bs/people/spider/
35. Zhu J, Rosset S, Hastie T, Tibshirani R (2003) 1-norm support vector machines. In: Neural Information Processing Systems, MIT Press, pp 16–23

Sebastián Maldonado received his B.S. and M.S. degree from the University of Chile, in 2007, and his Ph.D. degree from the University of Chile, in 2011. He is currently Associate Professor at the School of Engineering and Applied Sciences, Universidad de los Andes, Santiago, Chile. His research interests include statistical learning, data mining and business analytics.

Ricardo Montoya is an Assistant Professor of Industrial Engineering at the University of Chile. He received his Ph.D. in Marketing from Columbia University and his Master in Operations Management and Industrial Engineering degrees from University of Chile. His research focuses on modeling and estimating consumer preferences, dynamic choice models, and optimal product design. His methodological interests lie in Bayesian econometrics, hidden Markov models, and stochastic dynamic programming.

Julio López received his B.S. degree in Mathematics in 2000 from the University of Trujillo, Perú. He also received the M.S. degree in Sciences in 2003 from the University of Trujillo, Perú and the Ph.D. degree in Engineering Sciences, minor Mathematical Modelling in 2009 from the University of Chile. Currently, he is an assistant Professor of Institute of Basic Sciences at the University Diego Portales, Santiago, Chile. His research interests include conic programming, convex analysis, algorithms and machine learning.