Journal of the Operational Research Society (2017)

www.palgrave.com/journals

Simultaneous preference estimation and heterogeneity control for choice-based conjoint via support vector machines

Julio López[1], Sebastián Maldonado[2*] and Ricardo Montoya[3]

[1]*Facultad de Ingeniería, Universidad Diego Portales, Ejército 441, Santiago, Chile;* [2]*Facultad de Ingeniería y Ciencias Aplicadas, Universidad de los Andes, Monseñor Álvaro del Portillo 12455, Las Condes, Santiago, Chile; and* [3]*Department of Industrial Engineering, Universidad de Chile, Av. República 701, Santiago, Chile*

Support vector machines (SVMs) have been successfully used to identify individuals' preferences in conjoint analysis. One of the challenges of using SVMs in this context is to properly control for preference heterogeneity among individuals to construct robust partworths. In this work, we present a new technique that obtains all individual utility functions simultaneously in a single optimization problem based on three objectives: complexity reduction, model fit, and heterogeneity control. While complexity reduction and model fit are dealt using SVMs, heterogeneity is controlled by shrinking the individual-level partworths toward a population mean. The proposed approach is further extended to kernel-based machines, conferring flexibility to the model by allowing nonlinear utility functions. Experiments on simulated and real-world datasets show that the proposed approach in its linear form outperforms existing methods for choice-based conjoint analysis.

Journal of the Operational Research Society (2017). doi:10.1057/s41274-016-0013-6

Keywords: conjoint analysis; heterogeneity control; support vector machines; OR in marketing; artificial intelligence

1. Introduction

Conjoint analysis is probably the most significant development in marketing research in the past few decades. It provides an useful technique to Marketing and Operations Research fields to identify customers' preferences. The firms introducing new products and services rely on conjoint modeling where the estimated preferences are used to evaluate different opportunities via market simulations (Tsafarakis *et al*, 2011). Originally developed by Green *et al* (2004), conjoint analysis has been widely used in various domains, such as Banking (Mankila, 2004), Higher Education (Irani *et al*, 2014), Transportation (Hensher *et al*, 1998), Tourism Management (Thyne *et al*, 2006), and Public Management (Venkatesh *et al*, 2012), among many others.

Conjoint estimation has two important Operations Research features: On the one hand several optimization methods have been applied to efficiently solve conjoint analysis problems (Camm *et al*, 2006), and on the other hand conjoint modeling has been used for multiattribute decision making (MADM) to assess customer preferences via multivariate analysis (Scholl *et al*, 2005). Regarding the first point, choice-based conjoint has been recently tackled using support vector machines (Chapelle and Harchaoui, 2005; Cui and Curry, 2005) and other advanced optimization techniques (Evgeniou *et al*, 2007). These techniques link conjoint analysis with business analytics tools which brings interesting research opportunities for both the Operations Research and Machine Learning communities.

One of the most interesting challenges in conjoint analysis is the modeling of heterogeneity in consumers' preferences. Given the limited information per customer usually collected by conjoint methods, the goal is to obtain robust individual-level estimates by leveraging population-level information across consumers. In this work, we present a novel SVM-based approach that obtains all individual partworths in a single optimization problem while simultaneously controlling for heterogeneity in the same step. We focus on choice-based conjoint, the most popular type of conjoint analysis, where respondents are asked to compare among different alternatives and choose one of them in repeated questions. Experiments on simulated and real-world data show that the proposed approach outperforms existing methods for choice-based conjoint analysis via SVM.

The rest of the paper is structured as follows: Recent developments on SVMs for choice-based conjoint analysis are reviewed in the next section. The proposed method for conjoint analysis based on SVMs is introduced in the section that follows. Section Results provides experimental findings

*Correspondence: *Sebastián Maldonado, Facultad de Ingeniería y Ciencias Aplicadas, Universidad de los Andes, Monseñor Álvaro del Portillo 12455, Las Condes, Santiago, Chile.*
E-mail: smaldonado@uandes.cl

using synthetic data and two empirical conjoint studies. The main conclusions can be found in the last section, together with future developments derived from this research.

2. Previous work on support vector machines for choice-based conjoint analysis

Originally developed for classification, support vector machines (Vapnik, 1998) provide important advantages for predictive modeling, such as a nonlinear decision function, absence of local minima, an superior generalization of new objects thanks to the *structural risk minimization principle* (Vapnik, 1998). This method has been successfully applied in several domains, such as credit scoring (Schebesch and Stecking, 2005) and churn prediction (Verbeke *et al*, 2012).

Several SVMs formulations have been presented in the last decade in order to achieve better predictive performance in conjoint estimation. Support vector machines were first adapted for choice-based conjoint analysis by Cui and Curry (2005), and subsequently improved to handle preference heterogeneity by Chapelle and Harchaoui (2005) and Evgeniou *et al* (2005, 2007). Additionally, Toubia *et al* (2007a) discussed the use of polyhedral optimization models to estimate customer preferences for choice-based conjoint.

Next, we describe support vector machines for individual utility estimation in a choice-based context, and subsequently present some remarks about modeling heterogeneity.

Customer i's preferences are modeled by a linear utility function $u_i(\mathbf{x}) = \mathbf{w}_i^\top \mathbf{x}$, $i = 1, \ldots, N$, where the weight vector $\mathbf{w_i}$ is called *partworth*. Each customer evaluates K different product profiles and chooses one in each of T choice occasions. Each product profile is characterized by J attributes, each one defined over n_j levels, $j = 1, \ldots, J$.

From the customer choices we obtain information of the form $\left([\mathbf{x_{it}^1}, \ldots, \mathbf{x_{it}^K}], \mathbf{y_{it}}\right)$, where $\mathbf{x_{it}^k} \in \Re^{\mathbf{J}}$ and $y_{it} \in \{1, \ldots, K\}$ for $1 \leq i \leq N$, $1 \leq t \leq T$, and $1 \leq k \leq K$. The choice $y_{it} = k$ means that consumer i prefers the k^{th} option among the K profiles described by $[\mathbf{x_{it}^1}, \ldots, \mathbf{x_{it}^K}]$, that is, $u_i(\mathbf{x}_{it}^{y_{it}}) \geq u_i(\mathbf{x}_{it}^b)$, $\forall b \in \{1, \ldots, K\} \setminus \{y_{it}\}$ (Chapelle and Harchaoui, 2005). Following previous research, we assume we can rearrange the data such that all customers choose the first profile at occasion t, i.e., $y_{it} = 1$, $1 \leq i \leq N$, and $1 \leq t \leq T$. Then, 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 $1 \leq t \leq T$. Following the structural risk minimization principle (Vapnik, 1998), the Euclidean norm of $\mathbf{w_i}$ controls the *shrinkage* of the partworth solution, which essentially limits the set of possible coefficients, reducing the complexity of the problem, and subsequently the risk of overfitting.

The Euclidean norm is a widely used method of regularization of ill-posed problems (Tikhonov and Arsenin, 1977), an issue that often arises in conjoint estimation (Evgeniou *et al*, 2007). Indeed, due to this property, it has been adopted as the default formulation for SVMs (Cui and Curry, 2005; Evgeniou *et al*, 2005). In its formulation for conjoint analysis, the use of the Euclidean norm has an interesting implication: having a set of constraints that reflect the agreement between the choice data and the estimated decision rule (the estimated utility of the chosen alternative should be higher than the estimated utilities of the remaining alternatives in the choice set), the l_2-regularizer chooses the parameters in the feasible polyhedron that are the furthest from all constraints. This allows satisfying the hardest comparison in a robust and efficient manner (Evgeniou *et al*, 2005). It can be shown that the *margin* with which the constraints are satisfied, which also corresponds to the radius of the largest inscribed sphere, is equal to $\frac{1}{\|\mathbf{w}\|^2}$ (Vapnik, 1998). Hence, by minimizing the Euclidean norm we maximize the *margin* with which the chosen alternative is preferred over the other alternatives, conferring robustness to the estimation process. It implies that the solution is robust to small variations of the estimated parameters.

A set of slack variables ξ_{kt} is introduced for noise penalization at fitting the estimated utilities compared to the actual choices. This leads to the following quadratic programming problem, which is estimated for each customer $i = 1, \ldots, N$ (Chapelle and Harchaoui, 2005; Evgeniou *et al*, 2005):

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

The parameter C determines the trade-off between model fit and shrinkage, which can be set via cross-validation (see eg, Evgeniou *et al*, 2005; Toubia *et al*, 2007a). The individual partworths $\mathbf{w_i}$ are obtained for each customer $i = 1, \ldots, N$ from this formulation, which has a single (global) optimum.

Formulation (2) can be transformed into a kernel-based problem by computing its dual and then applying the Kernel Trick to obtain nonlinear utility functions (Evgeniou *et al*, 2005). These nonlinear utility functions are linear in the parameters but nonlinear in the product characteristics. Thus, they reflect nonlinear changes in preferences due to changes in the product profile. The detailed derivation of this formulation can be found in Maldonado *et al* (2015).

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{t=1}^{T}\sum_{k=2}^{K}\alpha_{kt} - \frac{1}{2}\sum_{t,s=1}^{T}\sum_{k=2}^{K}\alpha_{kt}\alpha_{ks}\left(\mathcal{K}(\mathbf{x}_t^1, \mathbf{x}_s^1)\right. \\ & \left. + \mathcal{K}(\mathbf{x}_t^k, \mathbf{x}_s^k) - \mathcal{K}(\mathbf{x}_t^1, \mathbf{x}_s^k) - \mathcal{K}(\mathbf{x}_t^k, \mathbf{x}_s^1)\right) \\ \text{s.t.} \quad & 0 \leq \alpha_{kt} \leq C, \ t = 1, \ldots, T, \ k = 2, \ldots, K. \end{aligned} \tag{3}$$

The previous formulation has the following issue: the information needed to estimate individual partworths is usually not

sufficient due to the small size of the questionnaires. Consequently, several strategies have been proposed for the SVM formulation to pool information across consumers, similarly as Hierarchical Bayesian models do (see eg, Gelman and Pardoe, 2006). The idea is to capture general patterns at the population level and use them to adjust the individual partworths, reducing the risk of overfitting.

The heterogeneity in consumers' preferences can be modeled simply by constructing the individual partworths and then computing population partworths as their average, i.e., $\overline{\mathbf{w}} = 1/N \sum_i \mathbf{w_i}$ (Evgeniou *et al*, 2005). The final individual partworths are obtained based on a weighted sum between the population partworths and the original individual partworths. The trade-off between both terms is controlled by a parameter $\gamma_i \in [0, 1]$, i.e., $\gamma_i \mathbf{w_i} + (\mathbf{1} - \gamma_\mathbf{i})\overline{\mathbf{w}}$ (Evgeniou *et al*, 2005). Similar approach for heterogeneity control was followed by Maldonado *et al* (2015), where a feature selection strategy was proposed to identify consumer preferences and the most relevant attributes in the same process. Unlike our model, and following (Evgeniou *et al*, 2005), Maldonado *et al* (2015) use a two-step approach.

Alternatively, Chapelle and Harchaoui (2005) proposed a single optimization problem that simultaneously obtains all individual partworths based on population patterns. The model follows:

$$\begin{aligned} \min_{\mathbf{w}_i, \xi} \quad & \|\mathbf{w}_i\|^2 + \frac{C}{q_i} \sum_{t \in Q_i} \sum_{k=2}^{K} \xi_t^{k\,2} + \frac{\hat{C}}{\sum_{j \neq i} q_j} \sum_{t \notin Q_i} \sum_{k=2}^{K} \xi_t^{k\,2} \\ \text{s.t.} \quad & \mathbf{w}_i^\top \left(\mathbf{x}_{it}^1 - \mathbf{x}_{it}^k\right) \geq 1 - \xi_t^k, \ t = 1, \ldots, T, \ k = 2, \ldots, K, \\ & \xi_t^k \geq 0, \ t = 1, \ldots, T, \ k = 2, \ldots, K, \end{aligned} \tag{4}$$

where C and $\hat{C}$ are trade-off parameters that control heterogeneity: if $\frac{C}{\hat{C}} = 1$, then the partworths are modeled to be equal, i.e., the population is assumed to be homogeneous, while $\frac{C}{\hat{C}} \gg 1$ considers no heterogeneity (Chapelle and Harchaoui, 2005). The set Q_i contains the set of questions answered by customer i. The authors also propose using squared hinge loss instead of the traditional hinge loss function.

Evgeniou *et al* (2007) presented an approach called LOG-Het that jointly estimates the individual partworths based on information from all consumers. Unlike SVMs where the hinge loss is used to maximize fit and the Euclidean norm for shrinkage, they use the logistic error function and suggest shrinking the weights toward a vector $\mathbf{w_0}$, whose components are also decision variables. That is

$$\min_{\mathbf{w}_i, \mathbf{w}_0, D} \quad \sum_{i=1}^{N} (\mathbf{w}_i - \mathbf{w}_0)^\top D^{-1} (\mathbf{w}_i - \mathbf{w}_0) - \frac{1}{\gamma} \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{e^{\mathbf{x}_{it}^1 \mathbf{w}_i}}{\sum_{k=1}^{K} e^{\mathbf{x}_{it}^k \mathbf{w}_i}}, \tag{5}$$

where D is constrained to be a positive semidefinite matrix scaled to have trace equals to 1, and γ is a trade-off parameter which can be obtained via cross-validation.

The work by Evgeniou *et al* (2007) has appealing characteristics: instead of obtaining first the individual-level parameters and next updating such partworths, it incorporates the heterogeneity control in the model explicitly by shrinking the partworths toward a population mean. A single optimization problem takes into account all available information and allows the estimation of robust individual-level partworths. The model proposed by Chapelle and Harchaoui (2005) also uses a single optimization model, but it adds slack variables related to the other consumers in the objective function for each respondent. The heterogeneity control process is less explicit and intuitive for this model, where the relationship between C and $\hat{C}$ is not easy to determine.

Our methodology follows the ideas of LOG-Het to model preference heterogeneity, shrinking the weights toward a vector $\mathbf{w_0}$. However, our SVM approach minimizes the Euclidean norm of $\mathbf{w_i}$ to reduce complexity, and uses the hinge loss to maximize fit. Similarly to Evgeniou *et al* (2007), our approach shrinks the weights toward an aggregated partworth vector to model heterogeneity in consumers' preferences. Unlike LOG-Het, in our proposal we solve a single, strictly convex optimization problem to obtain all individual-level partworths. To solve the nonconvex LOG-Het formulation (5), Evgeniou *et al* (2007) solve several convex problems iteratively in order to find an adequate approximate solution. Our approach represents an important advantage in terms of efficiency, and it allows the inclusion of kernel functions to capture nonlinear preferences.

The proposed methodology solves a single optimization problem by combining the three objectives that are relevant in the estimation of preference decision models: model fit to the stated preferences, reduction of model complexity, and heterogeneity control. The gain of solving the problem simultaneously is threefold:

(i) *Efficiency* It is more efficient since it does not require the additional step of computing the population partworth a posteriori. This provides gains in computational time that are linear in the number of individuals.

(ii) *Shrinkage* The idea of the shrinkage is to borrow information across subjects. That is, each individual has a prior equal to the population parameter and deviates from it according to her stated preferences. As the population partworth is jointly optimized in our proposal, instead of using simply the average of all individual partworths (that should be obtained independently), the information of all subjects provided can be correctly incorporated in the estimation of the individual preferences.

(iii) *Multiobjective optimization* The single optimization approach represents a more coherent formulation of the multiobjective optimization problem, in which the trade-off between the three objectives should be tuned simultaneously according to the desired weights and/or via cross-validation using a single-grid search strategy.

The proposed procedure is further described in the following section.

3. Proposed method for support vector conjoint analysis

In this section, we present a novel choice-based conjoint approach based on support vector machines. The idea is to solve a single convex quadratic optimization problem based on three objectives: complexity reduction, model fit, and heterogeneity control. Our method relates the minimization of the Euclidean norm for the individual partworths and the shrinkage of them toward an aggregated partworth vector $\mathbf{w_0}$, while maximizing the fit by minimizing the hinge loss.

The main goal is to improve the SVM formulation for CBC presented in Evgeniou *et al* (2005), where the individual partworths are obtained independently. Our approach extends to some extent the ideas of LOG-Het (Evgeniou *et al*, 2007) for heterogeneity control to SVM, which has several advantages for dealing simultaneously with complexity control and model fit. Additionally, the kernel version of our approach represents, to the best of our knowledge, the first kernel-based approach reported in the literature for choice-based conjoint that controls for preference heterogeneity.

The primal formulation for linear partworth estimation is first presented, while the dual formulation of this problem is derived subsequently. The kernel version of the model for nonlinear preferences is described at the end of the section.

3.1. Support vector conjoint analysis: linear version

Let us consider the following quadratic programming problem:

$$\begin{aligned}\min_{\mathbf{w}_i,\mathbf{w}_0,\xi_{it}^k} \quad & \frac{1}{2}\sum_{i=1}^{N}(\|\mathbf{w}_i\|^2+\theta\|\mathbf{w}_i-\mathbf{w}_0\|^2)+C\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{k=2}^{K}\xi_{it}^k \\ \text{s.t.} \quad & \mathbf{w}_i^\top(\mathbf{x}_{it}^1-\mathbf{x}_{it}^k)\geq 1-\xi_{it}^k,\ i=1,\ldots,N,\ t=1,\ldots,T,\ k=2,\ldots,K, \\ & \xi_{it}^k\geq 0,\ i=1,\ldots,N,\ t=1,\ldots,T,\ k=2,\ldots,K,\end{aligned} \tag{6}$$

where θ and C are positive parameters that control the relationship between complexity, heterogeneity, and fit.

Proposition 1 (see Appendix A) shows that the proposed quadratic problem (6) is strictly convex. This result is key for our modeling approach since it implies that the optimization procedure behind our method guarantees a single optimal solution for partworth estimation.

3.2. Dual formulation for linear support vector conjoint analysis

Next, the dual formulation of problem (6) is presented (see Appendix B for the detailed derivation). This formulation is useful to obtain a kernel-based formulation to uncover nonlinear preferences. Such formulation enhances the flexibility of the model, potentially leading to higher predictive performance. This formulation is given by

$$\begin{aligned}\max_{\alpha_t^k\in\Re^N} \quad & \sum_{t=1}^{T}\sum_{k=2}^{K}\boldsymbol{\alpha}_t^{k\top}\mathbf{e}-\frac{1}{2N(\theta+1)}\left\|\widetilde{\mathcal{Q}}(\theta)^{1/2}\sum_{t=1}^{T}\sum_{k=2}^{K}X_t^k\boldsymbol{\alpha}_t^k\right\|^2 \\ \text{s.t.} \quad & 0\leq\boldsymbol{\alpha}_t^k\leq C\mathbf{e},\ t=1,\ldots,T,\ k=2,\ldots,K.\end{aligned} \tag{7}$$

3.3. Kernel-based support vector conjoint analysis

In order to obtain the kernel-based formulation associated to Problem (6), we first develop the quadratic term of the objective function of Problem (7). That is,

$$\begin{aligned}\left\|\tilde{\mathcal{Q}}(\theta)^{1/2}\sum_{t=1}^{T}\sum_{k=2}^{K}X_t^k\boldsymbol{\alpha}_t^k\right\|^2 &= \sum_{t=1}^{T}\sum_{k=2}^{K}\sum_{t'=1}^{T}\sum_{k'=2}^{K}(\boldsymbol{\alpha}_t^k)^\top(X_t^k)^\top\tilde{\mathcal{Q}}(\theta)X_{t'}^{k'}\boldsymbol{\alpha}_{t'}^{k'} \\ &= \sum_{t=1}^{T}\sum_{k=2}^{K}\sum_{t'=1}^{T}\sum_{k'=2}^{K}(\boldsymbol{\alpha}_t^k)^\top\Phi_\theta(\mathbf{x}_{1t}^k,\mathbf{x}_{1t'}^{k'})\boldsymbol{\alpha}_{t'}^{k'},\end{aligned}$$

where the elements of the matrix $\Phi_\theta(\mathbf{x}_{1t}^k,\mathbf{x}_{1t'}^{k'})\in\Re^{N\times N}$ are given by

$$\begin{aligned}&[\Phi_\theta(\mathbf{x}_{1t}^k,\mathbf{x}_{1t'}^{k'})]_{ij} \\ &=\begin{cases}(N+\theta)\left(\mathbf{x}_{it}^{1\top}\mathbf{x}_{it'}^1-\mathbf{x}_{it}^{1\top}\mathbf{x}_{it'}^{k'}-\mathbf{x}_{it}^{k\top}\mathbf{x}_{it'}^1+\mathbf{x}_{it}^{k\top}\mathbf{x}_{it'}^{k'}\right), & i=j,\\ \theta\left(\mathbf{x}_{it}^{1\top}\mathbf{x}_{jt'}^1-\mathbf{x}_{it}^{1\top}\mathbf{x}_{jt'}^{k'}-\mathbf{x}_{it}^{k\top}\mathbf{x}_{jt'}^1+\mathbf{x}_{it}^{k\top}\mathbf{x}_{jt'}^{k'}\right), & i\neq j.\end{cases}\end{aligned}$$

Since the training samples appear in the previous formulation only in the form of inner products of the form $\mathbf{x}_{it}^{1\top}\mathbf{x}_{it'}^{k'}$, we can apply the Kernel trick (Schölkopf and Smola, 2002) by replacing them with $\mathcal{K}(\mathbf{x}_{it}^1,\mathbf{x}_{it'}^{k'})$, where $\mathcal{K}:\Re^n\times\Re^n\to\Re$ is any function satisfying the Mercer's condition (Mercer, 1909). Typical choices for kernel functions are *the Gaussian kernel* defined by $\mathcal{K}(\mathbf{u},\mathbf{v})=\exp(-\|\mathbf{u}-\mathbf{v}\|^2/2\sigma^2)$ with $\sigma\in\Re$, and the *polynomial function* $\mathcal{K}(\mathbf{u},\mathbf{v})=(\mathbf{u}^\top\mathbf{v}+1)^d$ with $d\in\mathbb{N}$ (see eg, Maldonado *et al*, 2011; Schölkopf and Smola, 2002). After choosing this kernel function, the above relation can be rewritten as

$$[\Phi_\theta(\mathbf{x}_{1t}^k,\mathbf{x}_{1t'}^{k'})]_{ij}=\begin{cases}(N+\theta)\big(\mathcal{K}(\mathbf{x}_{it}^1,\mathbf{x}_{it'}^1)-\mathcal{K}(\mathbf{x}_{it}^1,\mathbf{x}_{it'}^{k'}) & \\ \quad -\mathcal{K}(\mathbf{x}_{it}^k,\mathbf{x}_{it'}^1)+\mathcal{K}(\mathbf{x}_{it}^k,\mathbf{x}_{it'}^{k'})\big), & i=j,\\ \theta\Big(\mathcal{K}(\mathbf{x}_{it}^1,\mathbf{x}_{jt'}^1)-\mathcal{K}(\mathbf{x}_{it}^1,\mathbf{x}_{jt'}^{k'}) & \\ \quad -\mathcal{K}(\mathbf{x}_{it}^k,\mathbf{x}_{jt'}^1)+\mathcal{K}(\mathbf{x}_{it}^k,\mathbf{x}_{jt'}^{k'})\Big), & i\neq j.\end{cases} \tag{8}$$

Hence, taking into account the relation (8), the kernel-based formulation for choice-based conjoint using SVMs is given by

$$\max_{\boldsymbol{\alpha}_t^k \in \Re^N} \sum_{t=1}^{T}\sum_{k=2}^{K} \boldsymbol{\alpha}_t^{k\top}\mathbf{e} - \frac{1}{2N(\theta+1)}\sum_{t=1}^{T}\sum_{k=2}^{K}\sum_{t'=1}^{T}\sum_{k'=2}^{K}(\boldsymbol{\alpha}_t^k)^\top \Phi_\theta(\mathbf{x}_{1t}^k, \mathbf{x}_{1t'}^{k'})\boldsymbol{\alpha}_{t'}^{k'}$$
$$\text{s.t.} \qquad 0 \le \boldsymbol{\alpha}_t^k \le C\mathbf{e},\ t = 1,\ldots,T,\ k = 2,\ldots,K. \tag{9}$$

Finally, from (20) it follows that the individual utility function has the following form:

$$u_i(\mathbf{x}) = \frac{1}{N(\theta+1)}\left((\theta+N)\sum_{t=1}^{T}\sum_{k=2}^{K}\alpha_{it}^k(\mathcal{K}(\mathbf{x}_{it}^1,\mathbf{x}) - \mathcal{K}(\mathbf{x}_{it}^k,\mathbf{x})) + \theta\sum_{j=1,j\neq i}^{N}\sum_{t=1}^{T}\sum_{k=2}^{K}\alpha_{jt}^k(\mathcal{K}(\mathbf{x}_{jt}^1,\mathbf{x}) - \mathcal{K}(\mathbf{x}_{jt}^k,\mathbf{x}))\right),$$
$$i = 1,\ldots,N. \tag{10}$$

4. Results

We applied the proposed SVM-based approach for choice-based conjoint to four simulated datasets and two empirical applications. We refer as **L-SVM** $_\theta$ and **NL-SVM** $_\theta$ to the proposed approach in its linear and kernel-based formulations, respectively.

We compare the proposed methods with SVMs for individual-level utility functions in its linear (**L-SVM**, Formulation (4)) and kernel-based form (**NL-SVM**, Formulation (3)), respectively. We also compared the proposed approach with the SVM formulation for heterogeneity control proposed by Evgeniou *et al* (2005) (**L-SVM** $_\gamma$), and the mixed logit model approach (linear compensatory by aspects or **LCA**) based on Hierarchical Bayesian Markov chain Monte Carlo (MCMC) estimation method (see eg, (Rossi *et al*, 2005)) and SVMs for choice-based conjoint. For the Mixed logit model, we use a Markov Chain Monte Carlo (MCMC) method to obtain random draws from the posterior density. The MCMC methods involve sampling parameter estimates from full conditional distributions of parameters. Because the full conditional distributions do not have closed-form expressions, we use the Metropolis–Hastings algorithm to draw the parameters. In the MCMC procedure, we iteratively draw from the full conditional distributions of each parameter. The Metropolis–Hastings algorithm requires the choice of the proposal distribution. Determining such proposal distribution is difficult, and it is usually implemented in an adhoc manner involving many trial and error steps (Rosenthal *et al*, 2011). To facilitate rapid mixing of the resulting Markov chain, we use an adaptive random walk Metropolis–Hastings algorithm (Atchade, 2006) to determine the tuning parameters. See (Rosenthal *et al*, 2011) for more details about optimal proposal distributions and adaptive MCMC methods.

4.1. Synthetic data and empirical datasets

In this section, we briefly present the datasets used in this work.

4.1.1. Simulated data We used the simulation procedure proposed by Arora and Huber (2001) and Toubia *et al* (2007b): we generated four different datasets varying the noise condition in consumer choices (low and high noise) and the sparseness in consumer preferences (low and high), as suggested in Maldonado *et al* (2015). In each dataset $N = 200$ respondents across $T = 12$ choice occasions among $K = 3$ product profiles were simulated. The profiles were constructed using an orthogonal design based on $J = 10$ attributes with $n_j = 4$ levels each $(j = 1,\ldots,10)$. The deterministic utility of each profile was simulated based on a multivariate 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.

The noise condition was varied by adjusting the parameter β: the values of $\beta = 0.5$ and $\beta = 2$ for "high" and "low" noise conditions were used, respectively (Arora and Huber, 2001). Additionally, the sparseness is operationalized as follows, two and six randomly selected attributes were generated to be irrelevant (low sparseness and high sparseness conditions, respectively) by their corresponding mean parameter of the Gaussian distribution to zero ($\boldsymbol{\mu} = \mathbf{0}$) for each individual (Maldonado *et al*, 2015). A high sparseness in consumer preferences can be interpreted as customers ignoring a high number of attributes (six out of ten) when evaluating the different product profiles.

For each respondent, the dataset was split for calibration purposes: the first 10 questions were used for training and model calibration, while the final two decisions were used for testing purposes.

Although in our proposal we explore linear and nonlinear approaches, we decided to simulate data generated from linear utility functions mainly for two reasons. First, we reported these experiments in order to make our proposal comparable with most of the CBC literature that consider this dataset, although it may favor linear methods over kernel-based ones. Secondly, nonlinear CBC estimation is a recent field of research, and there is no standard methodology to generate nonlinear utility functions.

4.1.2. Digital camera dataset This dataset comprises digital cameras described across $J = 5$ attributes with $n_j = 4$ levels each. The following attributes describe this dataset: Price (US$500, US$400, US$300, and US$200), Resolution (2, 3, 4, and 5 Megapixels), Battery Life (150, 300, 450, and 600 pictures), Optical Zoom (2×, 3×, 4×, and 5×), and Camera Size (SLR, Medium, Pocket, and Ultra Compact).

$N = 125$ subjects from a customer panel answered 20 questions in an online CBC experiment. In each question,

respondents evaluated four product profiles randomly assigned and chose one of them. The product attributes were introduced and described to the respondents before applying the questionnaire. The choices from the first 16 questions were used to calibrate the models, while the remaining four were used to test the estimated models. See Abernethy *et al* (2008) for further details about this experiment.

4.1.3. Study time dataset The second empirical dataset represents information collected for marketing research purposes for an important American company. $N = 602$ subjects evaluated products in an online CBC study. Each product profile was described by $J = 10$ unbalanced attributes with between three and 15 levels. Due to the proprietary nature of the dataset, the actual product and the specific attributes and attribute levels cannot be revealed. The products in this dataset are presented in choice sets with three alternatives, while each respondent answered 12 choice questions. Ten of these questions were used for training and calibration purposes, while the remaining two were considered for testing.

4.1.4. Parameters' calibration For the SVM-based approaches, we need to calibrate the following parameters: C, γ [only for the heterogeneity control procedure suggested by Evgeniou *et al* (2005)], and θ (only for our proposal). For this purpose, we use a leave-one-out cross-validation (LOOCV) strategy on the training/calibration data. For each individual, a subset of the training/calibration data comprising all questions but one is used to estimate the individual partworths (training step). Subsequently, the question left out is used to predict the response (calibration or validation step). This process is repeated so that each question in the training subset is left out once and used for calibration purposes. The mentioned parameters are set to the values that maximize the LOOCV hit ratio. After the calibration procedure, the partworths are estimated using the entire calibration set with the optimal parameters found in the validation step, and the final evaluation is performed in the test subset, which remains unseen during the calibration step. This strategy has been used previously in choice-based conjoint (see eg, Evgeniou *et al*, 2005; Evgeniou *et al*, 2007; Toubia *et al*, 2007a).

The hit ratio has been widely used in choice-based conjoint analysis in order to measure the accuracy of a solution (see eg, Evgeniou *et al*, 2005; Evgeniou *et al*, 2007; Toubia *et al*, 2007a). In particular, the out-of-sample hit rate demonstrates the capability of the corresponding model to predict individual preferences. It is arguably the most intuitive performance metric, since it computes the percentage of correct predictions across the sample. It also provides a simple way of capering the performance of different models, including nested and nonnested formulations. In addition, the use of the holdout hit rate based on an independent test set is particularly important to avoid overfitting.

The following values for C, γ, and θ were studied in the calibration step: $C,\ \gamma,\ \theta \in \{2^{-7}, 2^{-6}, \ldots, 2^{7}\}$. These exponentially growing sequences are recommended for grid search in the machine learning literature (see eg, Hsu *et al*, 2010; Maldonado and López, 2014). For kernel-based approaches we explored the following widths: $\sigma \in \{1, 2, 4, 8\}$.

4.2. Results summary

Tables 1 and 2 summarize the results for all approaches and all four simulated datasets. The best performance among all methods in terms of test hit ratio is highlighted in bold type. We use a Student t test for pairwise comparisons between the best average performance in terms of holdout hit rate and the remaining methods. The best approach and those that are not significantly worse than the best at a 1 % level are highlighted with an asterisk.

Table 1 Empirical Comparison of the Preference Models (in percentages), low noise condition

Models	*Low noise*			
	Sparsity			
	Low		*High*	
	Hit rate		*Hit rate*	
	In[a]	*Out*[b]	*In*[a]	*Out*[b]
LCA	87.5	56.3	82	50.3
L-SVM	98.3	54.0	98.1	51.8
L-SVM $_{\gamma}$	96.7	60.0*	96.0	58.5*
NL-SVM	98.4	54.0	100	54.3*
L-SVM $_{\theta}$	96.2	**65.5***	97.4	**59.0***
NL-SVM $_{\theta}$	80.3	62.8*	77.0	54.3*

*Best predictive hit rate or not significantly different than the best at the 1 % level

[a]In-sample hit rate

[b]Out-of-sample hit rate

Table 2 Empirical comparison of the preference models (in percentages), high noise condition

Models	*High Noise*			
	Sparsity			
	Low		*High*	
	Hit rate		*Hit rate*	
	In[a]	*Out*[b]	*In*[a]	*Out*[b]
LCA	76.4	43.5	73.3	41.2
L-SVM	98.0	47.3	99.7	44.0
L-SVM $_{\gamma}$	95.6	52.5*	95.6	**52.8***
NL-SVM	99.1	49.3*	100	46.5
L-SVM $_{\theta}$	88.3	**53.0***	93.9	51.8*
NL-SVM $_{\theta}$	71.9	51.5*	73.5	51.3*

*Best predictive hit rate or not significantly different than the best at the 1 % level

[a]In-sample hit rate

[b]Out-of-sample hit rate

Table 3 Empirical Comparison of the Preference Models (in percentages)

Models	*Camera*		*Study time*	
	Hit rate		*Hit rate*	
	In[a]	*Out*[b]	*In*[a]	*Out*[b]
LCA	84.5	58.0*	70.4	57.2*
L-SVM	92.3	56.4	95.0	58.6*
L-SVM $_{\gamma}$	91.6	58.4	91.4	59.6*
NL-SVM	95.1	58.6*	92.8	58.8*
L-SVM $_{\theta}$	89.3	**61.2***	97.6	**61.6***
NL-SVM $_{\theta}$	92.4	61.0*	73.5	54.3

*Best predictive hit rate or not significantly different than the best at the 1 % level

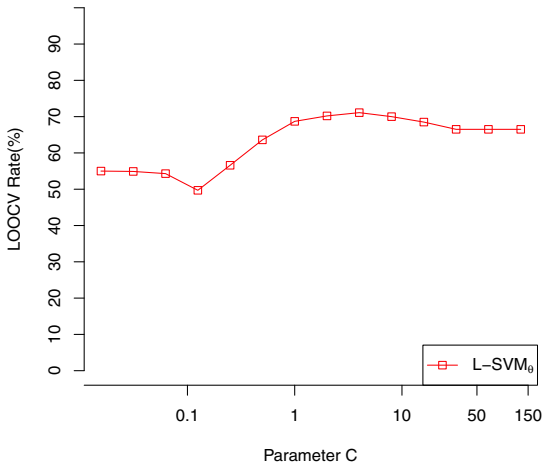
[a]In-sample hit rate

[b]Out-of-sample hit rate

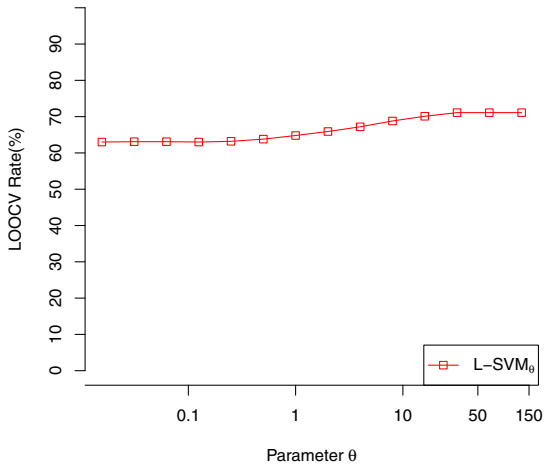
In Tables 1 and 2 it can be seen that the proposed approach in its linear form (**L-SVM** $_{\theta}$) achieves best average results in terms of out-of-sample hit rate on three out of the four simulated datasets, while SVMs with the heterogeneity correction proposed by Evgeniou *et al* (2005) (**L-SVM** $_{\gamma}$) has best predictive performance in the simulated data with high noise and high sparsity, but the differences are not significant compared to our proposal (second best). We can also observe that the kernel-based formulation for SVM (**NL-SVM**) has best in-sample performance, which is somehow expected since it provides more flexibility than linear models, but it shows signs of overfitting since it does not control for heterogeneity. The proposed approach and **L-SVM** $_{\gamma}$ reduce this risk by incorporating general patterns into the construction of the individual utility functions. The lower performance can be also explained by the fact that the datasets were simulated assuming linear decision rules.

Table 3 summarizes the results of all approaches applied to the two real-world applications. The best performance among all methods in terms of test hit ratio is highlighted in bold type. We also indicate with one asterisk the best predictive hit rate or not significantly different than the best at the 1 % level.

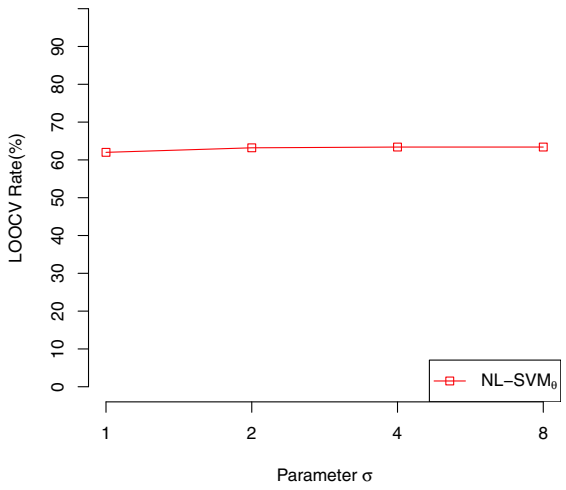
In Table 3, it can be seen that the proposed approach in its linear form has again the best predictive performance in the two empirical applications. The application in these two real datasets also confirms the analysis obtained earlier for simulated datasets: kernel methods and linear methods without heterogeneity control have better in-sample performance than the linear methods with heterogeneity control but worse predictive performance, demonstrating the advantage of pooling information across consumers to predict their preferences. For the first empirical application, both proposed methods achieve the best predictive performance, proving also that nonlinear models can be useful in real-world applications.



(a) Influence of parameter C



(b) Influence of parameter θ



(c) Influence of parameter σ

Fig. 1 LOOCV hit rates for L-SVM$_\theta$ for different values of C, and θ; LOOCV hit rates for NL-SVM$_\theta$ for different values of σ (Camera dataset).

We note that the small number of observations for each decision variable and the high number of parameter combinations for validation may have negatively affected the predictive performance of kernel methods due to overfitting. We observed that nonlinear methods achieved higher training and validation hit rates, but they also had a higher gap between validation and test hit rate. Additionally, the choice of the kernel function is still a matter of research, and an exhaustive grid search using a broader range of possible Kernel functions and their corresponding parameters may have led to different conclusions.

In sum, although there is not a unique method that completely outperform all others, we can conclude that overall our proposed approach in its linear form has the best average performance in terms of holdout hit ratio.

4.3. Sensitivity analysis to the tuning parameters and complexity

Next, we analyze the influence of the parameters C, θ, and σ on the performance of the proposed method. For illustration purposes, we vary these parameters and monitor the leave-one-out validation hit rates for the Camera dataset. To assess the influence of C and θ, we used the linear version of our approach, while the kernel-based approach with Gaussian kernel was used to explore the influence of parameter σ. Figure 1a–c present the LOOCV hit rates as a function of C, θ, and σ, respectively. We vary one parameter at a time, while the other parameters remained fixed on their optimal values.

Figure 1 presents relatively stable results for all parameters, although C presents a higher variance for values below the

unit. We observe an important influence of these parameters in the final outcome of the proposed approach, and therefore an adequate grid search is highly recommended.

As a reference, the training time is 22.6 seconds when $\theta = 0$ (standard L-SVM for CBC) and 23.1 seconds for our proposal ($\theta > 0$) for the Camera dataset, which has the size of typical CBC applications. For the Study time dataset, which could be considered large in CBC studies, the training time is 110.0 seconds when $\theta = 0$ (standard L-SVM for CBC) and 114.8 seconds for our proposal ($\theta > 0$). The experiments were performed on an HP Envy dv6 with 16 GB RAM, 750 GB SSD, an i7-2620M processor with 2.70 GHz, and using Microsoft Windows 8.1 Operating System (64-bits). We used the QPC solver for quadratic programming in Matlab 7.12.

We acknowledge that, although the running times are similar between L-SVM for CBC and our proposal, the introduction of an extra parameter leads to additional experiments to properly validate the model. We explored 15 different values for C and θ in the calibration step, which means a line search to estimate theta for a fixed C will require 15 more experiments, and the full grid search for our proposal involves $15 \times 15 = 225$ different runs. Given the training times reported before, these additional experiments can be performed in tractable running times.

5. Conclusions and future work

In this work, we present a novel choice-based conjoint analysis approach based on support vector machines. In contrast to the original SVM formulation for CBC, the proposed method solves a single optimization problem to construct all individual partworths simultaneously, and pools information across consumers by shrinking them toward a vector $\mathbf{w_0}$ that acts as an aggregated partworth. The proposed work can be seen as an extension to SVMs of LOG-Het, proposed by Evgeniou *et al* (2007), which follows a similar strategy for preference heterogeneity control, and improves the SVM formulations proposed by Evgeniou *et al* (2005) [and used in Maldonado *et al* (2015)] and Chapelle and Harchaoui (2005) by including a more appealing heterogeneity control strategy. We identified the following advantages of the proposed approach according to our experiments presented in the previous section:

- It has higher predictive performance than alternative choice-based conjoint approaches, thanks to its ability to handle simultaneously three objectives in one single optimization step: complexity reduction, model fit, and heterogeneity control.
- The method solves a strictly convex quadratic problem, which ensures a unique optimal solution for the problem. This is important since the proposed procedure avoids using time-consuming simulation strategies such as MCMC to estimate the partworths.
- The kernel-based version of the proposed model confers flexibility to the estimation process by allowing nonlinear preferences. Furthermore, to the best of our knowledge, this research represents the first work introducing kernel methods for conjoint analysis dealing with heterogeneity control. Strategies for heterogeneity control such as the one proposed by Evgeniou *et al* (2005) have not been extended to kernel methods.

We identify some research opportunities for future work. This approach can be extended to other conjoint applications such as menu-based conjoint or dynamic settings, such as adaptive methods for choice-based conjoint analysis. The superior predictive performance and computational efficiency are appealing features for these applications. Additionally, the method can be further extended to deal with clusters of consumers instead of an unimodal representation of preference heterogeneity. As suggested in Evgeniou *et al* (2007), the existence of multiple groups of respondents can be included in the modeling process by modifying the form of the shrinkage strategy and the loss function.

Acknowledgments—The authors thank Olivier Toubia and Bryan Orme for providing the data for the two empirical applications. The first author was funded by FONDECYT project 1160894 and by CONICYT Anillo ACT1106. The second author was supported by FONDECYT projects 1140831 and 1160738. The third author was supported by FONDECYT project 1151395 and FONDEF Project IT13I20031. This research was partially funded by the Complex Engineering Systems Institute, ISCI (ICM-FIC: P05-004-F, CONICYT: FB0816).

References

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

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

Atchade YF (2006). An adaptive version for the metropolis adjusted Langevin algorithm with a truncated drift. *Methodology and Computing in applied Probability* **8**(2):235–254.

Bertsekas D (1982). *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press: New York.

Camm JD, Cochran JJ, Curry DJ and 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 and Harchaoui Z (2005). A machine learning approach to conjoint analysis. In: Osherson DN, Kosslyn SM (eds) *Advances in Neural Information Processing Systems*, vol 17, pp. 257–264. MIT Press: Cambridge, MA.

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

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

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

Gelman A and Pardoe I (2006). Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics* **48**(2):241–251.
Green PE, Krieger AM and Wind Y (2004). Thirty years of conjoint analysis: Reflections and prospects. In Wind Y, Green PE *Marketing Research and Modeling: Progress and Prospects, International Series in Quantitative Marketing*, vol 14, pp. 117–139. Springer: Berlin.
Hensher D, Louviere J and Swait J (1998). Combining sources of preference data. *Journal of Econometrics* **89**(1):197–221.
Hsu C.-W., Chang C.-C. and Lin C.-J. (2010). A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University.
Irani S, Dwivedi Y and William M (2014). Analysing factors affecting the choice of emergent human resource capital. *Journal of the Operational Research Society* **65**(6):935–953.
Maldonado S and López J (2014). Alternative second-order cone programming formulations for support vector classification. *Information Sciences* **268**:328–341.
Maldonado S, Montoya R and Weber R (2015). Advanced conjoint analysis using feature selection via support vector machines. *European Journal of Operational Research* **241**(2):564–574.
Maldonado S, Weber R and Basak J (2011). Simultaneous feature selection and classification using kernel-penalized support vector machines. *Information Sciences* **181**(1):115–128.
Mankila M (2004). Retaining students in retail banking through price bundling: Evidence from the swedish market. *European Journal of Operational Research* **155**(2):299–316.
Mercer J (1909). Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London* **209**:415–446.
Rosenthal JS et al. (2011). Optimal proposal distributions and adaptive MCMC. In: Brook S, Gelman A, Jones GL, Meng X-L (eds) *Handbook of Markov Chain Monte Carlo*, pp. 93–112. Chapman and Hall: Boca Raton.
Rossi PE, Allenby GM and McCulloch R (2005). *Bayesian statistics and marketing*. Wiley: New York.
Schebesch K and Stecking R (2005). Support vector machines for classifying and describing credit applicants: detecting typical and critical regions. *Journal of the Operational Research Society* **56**(9):1082–1088.
Schölkopf B and Smola AJ (2002). *Learning with Kernels*. MIT Press: Cambridge.
Scholl A, Manthey L, Helm R and 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.
Thyne M, Lawson R and 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.
Tikhonov AN and Arsenin VY (1977). *Solution of Ill-posed Problems*. Winston & Sons: Washington.
Toubia O, Evgeniou T and Hauser J (2007a). Optimization-based and machine-learning methods for conjoint analysis: Estimation and question design. In: Gustafsson A, Herrmann A, Huber F (eds) *Conjoint Measurement: Methods and Applications*, pp. 231–258. Springer: New York.
Toubia O, Hauser J and Garcia R (2007b). Probabilistic polyhedral methods for adaptive choice-based conjoint anaysis. *Marketing Science* **26**(5):596–610.
Tsafarakis S, Grigoroudis E and Matsatsinis N (2011). Consumer choice behaviour and new product development: An integrated market simulation approach. *Journal of the Operational Research Society* **62**(7):1253–1267.
Vapnik V (1998). *Statistical Learning Theory*. Wiley: New York.
Venkatesh V, Chan FK and Thong JY (2012). Designing e-government services: Key service attributes and citizens' preference structures. *Journal of Operations Management* **30**(1):116–133.
Verbeke W, Dejaeger K, Martens D, Hur J and Baesens B (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research* **218**(1):211–229.
Yajima Y (2005). Linear programming approaches for multicategory support vector machines. *European Journal of Operational Research* **162**(2):514–531.

Appendix A

Strictly convexity of problem (6)

In order to prove that our Formulation (6) is strictly convex, we first rewrite it in a compact form. For this, we follow the derivation of Yajima (2005) for multicategory SVM. Let us denote by

$$\widetilde{\mathbf{w}} = [\mathbf{w}_0^\top, \mathbf{w}_1^\top, \ldots, \mathbf{w}_N^\top]^\top \in \Re^{J(N+1)},$$

and

$$\mathcal{Q}(\theta) = \begin{bmatrix} N\theta I_J & -\theta I_J & -\theta I_J & \cdots & -\theta I_J \\ -\theta I_J & (1+\theta) I_J & 0 & \cdots & 0 \\ -\theta I_J & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ -\theta I_J & 0 & \cdots & 0 & (1+\theta) I_J \end{bmatrix} \in \Re^{J(N+1)\times J(N+1)}, \tag{11}$$

where I_J denotes the identity matrix of size J. Then, the quadratic term in (6) can be expressed as

$$\sum_{i=1}^{N} (\|\mathbf{w}_i\|^2 + \theta \|\mathbf{w}_i - \mathbf{w}_0\|^2) = \tilde{\mathbf{w}}^\top \mathcal{Q}(\theta) \tilde{\mathbf{w}}. \tag{12}$$

Proposition 1 For any $\theta > 0$, the matrix $\mathcal{Q}(\theta)$ is symmetric definite positive. Moreover,

$$\mathcal{Q}(\theta)^{-1} = \frac{1}{N} \begin{bmatrix} \frac{\theta+1}{\theta} I_J & I_J & I_J & \cdots & I_J \\ I_J & \frac{\theta+N}{\theta+1} I_J & \frac{\theta}{\theta+1} I_J & \cdots & \frac{\theta}{\theta+1} I_J \\ I_J & \frac{\theta}{\theta+1} I_J & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \frac{\theta}{(\theta+1)} I_J \\ I_J & \frac{\theta}{\theta+1} I_J & \cdots & \frac{\theta}{(\theta+1)} I_J & \frac{\theta+N}{\theta+1} I_J \end{bmatrix}. \tag{13}$$

Proof It is clear that the matrix $\mathcal{Q}(\theta)$ is symmetric definite positive (cf. (12)). Now, we denote by $F_i \in \Re^{J\times J(N+1)}$ and $C_i \in \Re^{J(N+1)\times J}$ the i-th block (in row) and the i-th block (in column) of $\mathcal{Q}(\theta)$ and $\mathcal{Q}(\theta)^{-1}$, respectively. Then,

$$F_1C_1 = I_J, \quad F_1C_i = \frac{1}{N}\left(N\theta - \frac{N\theta+\theta^2N}{\theta+1}\right) = 0, \quad i = 2,\ldots,N+1,$$

$$F_iC_1 = 0, \quad F_iC_i = I_J, \quad F_iC_j = \frac{1}{N}(\theta-\theta) = 0, i \neq j.$$

Thus, the result follows. □

Appendix B

Dual formulation of problem (6)

Let us denote by $\boldsymbol{\xi}_t^k = (\xi_{1t}^k,\ldots,\xi_{Nt}^k) \in \Re^N$, and by $\mathbf{X}_t^k = \begin{bmatrix} 0 \\ X_t^k \end{bmatrix} \in \Re^{(N+1)J\times N}$ with

$$X_t^k = \begin{bmatrix} \mathbf{x}_{1t}^1 - \mathbf{x}_{1t}^k & 0 & & 0 \\ 0 & \ddots & \ddots & \vdots \\ & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \mathbf{x}_{Nt}^1 - \mathbf{x}_{Nt}^k \end{bmatrix} \in \Re^{NJ\times N}.$$

Then, the constraints of the problem (6) can be expressed as follows

$$\boldsymbol{\xi}_t^k \geq 0, \quad \mathbf{X}_t^{k\top}\tilde{\mathbf{w}} \geq \mathbf{e} - \boldsymbol{\xi}_t^k, \ t = 1,\ldots,T, \ k = 2,\ldots,K.$$

With this notation, the Lagrangian function associated to formulation (6) is given by

$$L(\tilde{\mathbf{w}}, \boldsymbol{\xi}_t^k, \boldsymbol{\alpha}_t^k, \mathbf{s}_t^k) = \frac{1}{2}\tilde{\mathbf{w}}^\top\mathcal{Q}(\theta)\tilde{\mathbf{w}} + \sum_{t=1}^{T}\sum_{k=2}^{K}\Big[C(\boldsymbol{\xi}_t^k)^\top\mathbf{e} - \boldsymbol{\alpha}_t^{k\top}(\mathbf{X}_t^{k\top}\tilde{\mathbf{w}} - \mathbf{e} + \boldsymbol{\xi}_t^k) - (\boldsymbol{\xi}_t^k)^\top\mathbf{s}_t^k\Big]. \tag{14}$$

Then, Problem (6) can be written equivalently as

$$\min_{\tilde{\mathbf{w}},\boldsymbol{\xi}_t^k}\max_{\boldsymbol{\alpha}_t^k,\mathbf{s}_t^k}\{L(\tilde{\mathbf{w}}, \boldsymbol{\xi}_t^k, \boldsymbol{\alpha}_t^k, \mathbf{s}_t^k) : \boldsymbol{\alpha}_t^k, \mathbf{s}_t^k \geq 0, \ t = 1,\ldots,T, \ k = 2,\ldots,K\}.$$

Hence, the dual formulation (see eg, Bertsekas, 1982) of (6) is given by

$$\max_{\boldsymbol{\alpha}_t^k,\mathbf{s}_t^k}\min_{\tilde{\mathbf{w}},\boldsymbol{\xi}_t^k}\{L(\tilde{\mathbf{w}}, \boldsymbol{\xi}_t^k, \boldsymbol{\alpha}_t^k, \mathbf{s}_t^k) : \boldsymbol{\alpha}_t^k, \mathbf{s}_t^k \geq 0, \ t = 1,\ldots,T, \ k = 2,\ldots,K\}.$$

The above expression enables us to compute the dual problem based only on the Lagrange multipliers $\boldsymbol{\alpha}$. The first-order conditions of the inner minimization problem yields to

$$\nabla_{\tilde{\mathbf{w}}}L(\tilde{\mathbf{w}}, \boldsymbol{\xi}_t^k, \boldsymbol{\alpha}_t^k, \mathbf{s}_t^k) = \mathcal{Q}(\theta)\tilde{\mathbf{w}} - \sum_{t=1}^{T}\sum_{k=2}^{K}\mathbf{X}_t^k\boldsymbol{\alpha}_t^k = 0, \tag{15}$$

$$\nabla_{\boldsymbol{\xi}_t^k}L(\tilde{\mathbf{w}}, \boldsymbol{\xi}_t^k, \boldsymbol{\alpha}_t^k, \mathbf{s}_t^k) = C\mathbf{e} - \boldsymbol{\alpha}_t^k - \mathbf{s}_t^k = 0. \tag{16}$$

Since $\mathbf{s}_t^k \geq 0$, from (16) it follows that $\boldsymbol{\alpha}_t^k \leq C\mathbf{e}$ for $t = 1,\ldots,T$, and $k = 2,\ldots,K$.

Remark 1 Note that using (1), (15), and the notation of $\mathbf{X}_t^k$, we have that

$$\mathbf{w}_0 = \frac{1}{N}\sum_{i=1}^{N}\mathbf{w}_i.$$

On the other hand, by using (15) and (16) in (14), we obtain that

$$L(\tilde{\mathbf{w}}, \boldsymbol{\xi}_t^k, \boldsymbol{\alpha}_t^k, \mathbf{s}_t^k) = \sum_{t=1}^{T}\sum_{k=2}^{K}\boldsymbol{\alpha}_t^{k\top}\mathbf{e} - \frac{1}{2}\left\|\mathcal{Q}(\theta)^{1/2}\tilde{\mathbf{w}}\right\|^2.$$

Since $\mathcal{Q}(\theta)$ is nonsingular, it follows from (15) that the above expression can be written as

$$L(\tilde{\mathbf{w}}, \boldsymbol{\xi}_t^k, \boldsymbol{\alpha}_t^k, \mathbf{s}_t^k) = \sum_{t=1}^{T}\sum_{k=2}^{K}\boldsymbol{\alpha}_t^{k\top}\mathbf{e} - \frac{1}{2}\left\|\mathcal{Q}(\theta)^{-1/2}\sum_{t=1}^{T}\sum_{k=2}^{K}\mathbf{X}_t^k\boldsymbol{\alpha}_t^k\right\|^2. \tag{17}$$

The following result allows us to rewrite the above equality.

Proposition 2 For $\theta > 0$, let $\widetilde{\mathcal{Q}}(\theta) = NI_{JN} + \theta\mathcal{J} \in \Re^{JN\times JN}$, where I_{JN} denotes the identity matrix of size JN and

$$\mathcal{J} = \begin{bmatrix} I_J & \cdots & I_J \\ \vdots & \ddots & \vdots \\ I_J & \cdots & I_J \end{bmatrix} \in \Re^{JN\times JN}.$$

Then, there exists a symmetric matrix $\widetilde{\mathcal{Q}}(\theta)^{1/2}$ satisfying $(\widetilde{\mathcal{Q}}(\theta)^{1/2})^2 = \widetilde{\mathcal{Q}}(\theta)$.

Proof Let

$$\widetilde{\mathcal{Q}}(\theta)^{1/2} = \sqrt{N}I_{JN} - \frac{1-\sqrt{1+\theta}}{\sqrt{N}}\mathcal{J}. \tag{18}$$

Note that $\mathcal{J}^2 = NJ$. Then,

$$(\widetilde{\mathcal{Q}}(\theta)^{1/2})^2 = NI_{JN} - 2(1-\sqrt{1+\theta})\mathcal{J} + \frac{(1-\sqrt{1+\theta})^2}{N}(N\mathcal{J}) = NI_{JN} + \theta\mathcal{J}.$$

□

By using the relation (13), Proposition 2, and the definition of $\mathbf{X}_t^k$, the expression (17) reduces to

$$L(\tilde{\mathbf{w}}, \xi_t^k, \boldsymbol{\alpha}_t^k, \mathbf{s}_t^k) = \sum_{t=1}^{T}\sum_{k=2}^{K} \boldsymbol{\alpha}_t^{k\top}\mathbf{e} - \frac{1}{2N(\theta+1)} \left\| \widetilde{\mathcal{Q}}(\theta)^{1/2} \sum_{t=1}^{T}\sum_{k=2}^{K} X_t^k \boldsymbol{\alpha}_t^k \right\|^2.$$

Hence, the dual formulation is given by

$$\begin{aligned} \max_{\boldsymbol{\alpha}_t^k \in \Re^N} \quad & \sum_{t=1}^{T}\sum_{k=2}^{K} \boldsymbol{\alpha}_t^{k\top}\mathbf{e} - \frac{1}{2N(\theta+1)} \left\| \widetilde{\mathcal{Q}}(\theta)^{1/2} \sum_{t=1}^{T}\sum_{k=2}^{K} X_t^k \boldsymbol{\alpha}_t^k \right\|^2 \\ \text{s.t.} \quad & 0 \le \boldsymbol{\alpha}_t^k \le C\mathbf{e},\ t = 1,\ldots,T,\ k = 2,\ldots,K. \end{aligned} \tag{19}$$

Remark 2 From (15) and (13), it follows that

$$\mathbf{w}_0 = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{k=2}^{K}(\mathbf{x}_{it}^1 - \mathbf{x}_{it}^k)\alpha_{it}^k,$$

and

$$\mathbf{w}_i = \frac{1}{N(\theta+1)}\left((\theta+N)\sum_{t=1}^{T}\sum_{k=2}^{K}(\mathbf{x}_{it}^1 - \mathbf{x}_{it}^k)\alpha_{it}^k + \theta\sum_{j=1, j\neq i}^{N}\sum_{t=1}^{T}\sum_{k=2}^{K}(\mathbf{x}_{jt}^1 - \mathbf{x}_{jt}^k)\alpha_{jt}^k \right), \tag{20}$$

for $i = 1,\ldots,N$.

Received 11 August 2015; accepted 1 June 2016