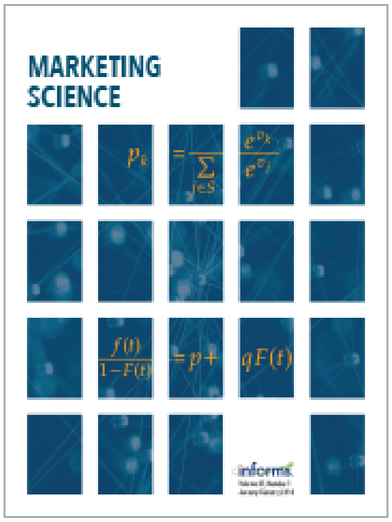


Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Generalized Robust Conjoint Estimation

Theodoros Evgeniou, Constantinos Boussios, Giorgos Zacharia,

To cite this article:
Theodoros Evgeniou, Constantinos Boussios, Giorgos Zacharia, (2005) Generalized Robust Conjoint Estimation. Marketing Science 24(3):415-429. https://doi.org/10.1287/mksc.1040.0100

MARKETING SCIENCE
Vol. 24, No. 3, Summer 2005, pp. 415–429
ISSN 0732-2399 | EISSN 1526-548X | 05 | 2403 | 0415

DOI 10.1287/mksc.1040.0100

Generalized Robust Conjoint Estimation

Theodoros Evgeniou
Technology Management, INSEAD, Boulevard de Constance, Fontainebleau 77300, France,
theodoros.evgeniou@insead.edu

Constantinos Boussios
Open Ratings, Inc., 200 West Street, Waltham, Massachusetts 02451, USA, and Laboratory for Information and Decision Systems, MIT, Cambridge, Massachusetts 02139, USA, boussios@openratings.com

Giorgos Zacharia
Open Ratings, Inc., 200 West Street, Waltham, Massachusetts 02451, USA, and Center for Biological and Computational Learning, MIT, Cambridge, Massachusetts 02139, USA, lysi@ai.mit.edu

We introduce methods from statistical learning theory to the field of conjoint analysis for preference modeling. We present a method for estimating preference models that can be highly nonlinear and robust to noise. Like recently developed polyhedral methods for conjoint analysis, our method is based on computationally efficient optimization techniques. We compare our method with standard logistic regression, hierarchical Bayes, and the polyhedral methods using standard, widely used simulation data. The experiments show that the proposed method handles noise significantly better than both logistic regression and the recent polyhedral methods and is never worse than the best method among the three mentioned above. It can also be used for estimating nonlinearities in preference models faster and better than all other methods. Finally, a simple extension for handling heterogeneity shows promising results relative to hierarchical Bayes. The proposed method can therefore be useful, for example, for analyzing large amounts of data that are noisy or for estimating interactions among product features.

Key words: choice models; data mining; econometric models; hierarchical Bayes analysis; marketing tools; regression and other statistical techniques
History: This paper was received May 28, 2003, and was with the authors 8 months for 2 revisions; processed by Duncan Simester.

1. Introduction

The amount of data capturing preferences of people for particular products, services, and information sources, has been dramatically increasing in recent years largely due, for example, to electronic commerce. Traditional preference modeling methods such as conjoint analysis (Carroll and Green 1995, Green and Srinivasan 1978, Green and Srinivasan 1990) have been used for many preference modeling applications (Wittink and Cattin 1989) typically with data gathered under controlled conditions such as through questionnaires. However, much of the available information today about choices of people, such as scanner or clickstream data, is not gathered in such a controlled way and therefore is more noisy (Cooley et al. 1997, Kohavi 2001). It is therefore important to develop new preference modeling methods that are (a) highly accurate, (b) robust to noise, and (c) computationally efficient in order to handle the large amounts of choice data available.

In this paper we present a family of preference models, from simple linear ones like existing ones (Ben-Akiva and Lerman 1985, Green and Srinivasan 1978, Srinivasan and Shocker 1973) to highly nonlinear ones that are robust to noise. They are developed based on the well-founded field of statistical learning theory and are shown to be almost equivalent to support vector machines (SVM) (Vapnik 1998), therefore bringing a number of new theories and tools to the preference modeling research community as also done by the recent work of (Cui and Curry 2005). Their estimation involves solving a quadratic programming optimization problem with simple box constraints, which is computationally efficient and leads to a unique optimal solution.

We compare our methods with logistic regression (Ben-Akiva and Lerman 1985, Louviere et al. 2000), hierarchical Bayes (HB) (DeSarbo and Ansari 1997, Allenby et al. 1998, Arora et al. 1998), and the polyhedral estimation methods of Toubia et al. (2004) using simulations as in Arora and Huber (2001), and Toubia et al. (2004). Our methods are shown experimentally to be more robust to noise than both logistic regression and the polyhedral methods, to either significantly outperform or never be worse than both logistic regression and the polyhedral methods, and to estimate nonlinear utility models faster and better than all methods including HB.

In this paper we do not address the issue of designing questionnaires as typically done in conjoint analysis. We plan to explore this issue in the future within the framework discussed here. We focus only on the utility estimation problem. We also deal only with full-profile preference data—full product comparisons—for the case of choice-based conjoint analysis (CBC) (Louviere et al. 2000) instead of other metric based ones. Extensions to the latter are possible like in the case of SVM regression (Vapnik 1998).

A number of practical issues arise when modeling preferences from unconstrained observations of choices. For example there are typically problems of taste heterogeneity (DeSarbo and Ansari 1997, Jedidi et al. 2003) among the subjects providing the data, or unobservable choice sets and incomplete information about choices (Manski 1977) which may also dynamically change over time (Pauwels 2005). These make the estimated models biased. In this paper we do not consider the latter issues, which is typical for conjoint analysis methods. In terms of handling heterogeneity across many individuals we only explore a simple ad hoc extension of the proposed methods to handle this issue, which shows promising results relative to HB. Extensions of the proposed methods to handle heterogeneity are part of future work.

Traditional conjoint estimation methods, such as logistic regression and HB, are developed assuming a particular probabilistic model of the data and the noise. Unlike those, our approach, like that of Toubia et al. (2003) and Cui and Curry (2005), is based on formulating the problem of preference modeling as an optimization problem where an appropriate cost function is minimized without assuming a particular probabilistic model for the data. The cost function is motivated by statistical arguments, namely by statistical learning theory, an empiricist's approach to developing models from data. We briefly discuss the statistical motivation of the optimization models. It is important to note that it is possible to justify the optimization models presented here using Bayesian arguments and reformulating the cost functions in terms of likelihood maximization (Evgeniou et al. 2000a). This is beyond the scope of this paper and we focus on providing methods and mathematical tools that can supplement existing ones that assume particular probabilistic models.

The work we present does not aim to replace existing methods for preference modeling, but instead to contribute to the field new tools and methods that can complement existing ones. One of the goals of this work is to bring to the field of conjoint analysis ideas from statistical learning theory and SVM which have been successfully used for other data analysis problems (Evgeniou et al. 2000a, b; Vapnik 1998).

The paper is organized as follows. In §2 we present our approach to modeling preferences. For simplicity we only show the basic linear model and briefly discuss some properties and the extension to nonlinear preference modeling. The latter are presented in more detail in the online appendix available at the *Marketing Science* website (http://mktsci.pubs.informs.org). In §3 experiments comparing the methods with other conjoint analysis methods, namely logistic regression, HB, and the estimation method of Toubia et al. (2004), are shown. Finally §4 is a summary and conclusions.

1.1. Previous Work

The market research community has traditionally approached utility estimation problems through function estimation. Conjoint analysis is one of the main methods for modeling preferences from data (Carroll and Green 1995, Green and Srinivasan 1978). A number of conjoint analysis methods have been proposed—see, for example, the Sawtooth Software website. Since the early 1970s conjoint analysis has been a very popular approach with hundreds of commercial applications per year (Wittink and Cattin 1989). In conjoint analysis designing questionnaires is a central issue (Arora and Huber 2001, Kuhfeld et al. 1994, Oppewal et al. 1994, Segal 1982), which, as mentioned above, we do not address here.

Within the discrete choice analysis area users' preferences are modeled as random variables of logit models (Ben-Akiva and Lerman 1985; Ben-Akiva et al. 1997; McFadden 1974, 1986). Both conjoint analysis and discrete choice methods have always faced the tradeoff between model (multinomial logit models) complexity and computational ease as well as predictive performance of the estimated model. This tradeoff is linked to the well-known "curse of dimensionality" (Stone 1985): as the number of dimensions increases an exponential increase in the number of data is needed to maintain reliable model estimation. The method we present in this paper can handle this issue, as already shown for other applications (Evgeniou et al. 2002, Vapnik 1998).

A different approach was implemented by Herbrich et al. (1999) who instead of trying to apply regression techniques for utility function estimation, they reformulated the problem as an ordinal regression estimation and used SVM to predict transitive ranking boundaries. More recently, Cui and Curry (2005) used SVM directly for predicting choices of consumers. Our methods are similar with those in Herbrich et al. (1999) and Cui and Curry (2005): in particular they are almost equivalent to SVM. Unlike Herbrich et al. (1999) and Cui and Curry (2005), we focus here on choice-based conjoint analysis and on the comparison with logistic regression, HB, and the estimation method of Toubia et al. (2004).

Finally, recent work by Toubia et al. (2003, 2004) addresses the problem of designing questionnaires and estimating preference models through solving polyhedral optimization problems which are similar to our methods as we discuss below. They develop methods for both metric (Toubia et al. 2003) and choice-based (Toubia et al. 2004) conjoint analysis. Our work, like that of Toubia et al. (2003, 2004), also aims at exploring the direction of developing new methods for preference modeling that are based on polyhedral optimization. A main difference from Toubia et al. (2004) is that they focus more on the design of individual-specific questionnaires while we focus on the estimation of a utility function from data. In the experiments below we only use the utility function estimation method of Toubia et al. (2004) and not the questionnaire design method they propose.

2. A Theoretical Framework for Modeling Preferences

2.1. Setup and Notation

We consider the standard (i.e., Louviere et al. 2000) problem of estimating a utility function from a set of examples of past choices all coming from a single individual—so from a single true underlying utility function. In the experiments we only deal with heterogeneity using a simple ad hoc approach described at the end of this section.

Formally we have data from n choices where, without loss of generality, the ith choice is among two products (or services, bids, etc.) $\{\mathbf{x}_i^1, \mathbf{x}_i^2\}$. To simplify notation we assume that for each i the first product $\mathbf{x}_i^1$ is the preferred one—we can rename the products otherwise. All products are fully characterized by m-dimensional vectors, where m is the number of attributes describing the products. We represent the jth product for choice i as $\mathbf{x}_i^j = \{x_i^j(1), x_i^j(2), \ldots, x_i^j(m)\}$ (notice that we use bold letters for vectors). So the ith choice is among a pair of m-dimensional vectors. We are now looking for a utility function that is in agreement with the data, namely a function that assigns higher utility value to the first product—the preferred one—for each pair of choices. This is the standard setup of choice-based conjoint analysis (Louviere et al. 2000). Variations of this setup (i.e., cases where we know pairwise relative preferences with intensities) can be modeled in a similar way.

2.2. A Robust Method for Linear Utility Function Estimation

We first make the standard assumption (Ben-Akiva and Lerman 1985, Srinivasan and Shocker 1973) that the utility function is a linear function of the values (or logarithms of the values, without loss of generality) of the product attributes: the utility of a product $\mathbf{x} = \{x(1), x(2), \ldots, x(m)\}$ is $U(\mathbf{x}) = w_1 \cdot x(1) + w_2 \cdot x(2) + \cdots + w_m \cdot x(m)$. We are looking for a utility function with parameters $w_1, w_2, \ldots, w_m$ that agrees with our data; that is we are looking for $w_1, \ldots, w_m$ such that $\forall\, i \in \{1, \ldots, n\}$,

$$w_1 \cdot x_i^1(1) + w_2 \cdot x_i^1(2) + \cdots + w_m \cdot x_i^1(m) \geq w_1 \cdot x_i^2(1) + w_2 \cdot x_i^2(2) + \cdots + w_m \cdot x_i^2(m). \quad (1)$$

Clearly there may be no w_f that satisfy all n constraints, since in practice the true utility function does not have to be linear and generally there are a lot of inconsistencies in data describing preferences of people. To allow for errors/inconsistencies we use slack variables, a standard approach for optimization methods (Bertsimas and Tsitsiklis 1997). For each of the n inequality constraints (1) we introduce a positive slack variable ξ_i which effectively measures how much inconsistency/error there is for choice i, as in Srinivasan and Shocker (1973). So we are now looking for a set of parameters $w_1, w_2, \ldots, w_m$ so that we minimize the error $\sum_{i=1,\ldots,n} \xi_i$ where $\xi_i \geq 0$ and satisfy $\forall\, i \in \{1, 2, \ldots, n\}$,

$$w_1 \cdot x_i^1(1) + w_2 \cdot x_i^1(2) + \cdots + w_m \cdot x_i^1(m) \geq w_1 \cdot x_i^2(1) + w_2 \cdot x_i^2(2) + \cdots + w_m \cdot x_i^2(m) - \xi_i. \quad (2)$$

Notice that one may require minimizing the L_0 norm of the slack variables ξ_i so that what is penalized is the number of errors/inconsistencies and not the "amount" of it. In that case the optimization problem becomes an integer programming problem which is hard to solve.

So in this simple model we are looking for a linear utility function that minimizes the amount of errors/inconsistencies on the *estimation* data. This, however, may lead to models that overfit the current data, are sensitive to noise, and can suffer from the curse of dimensionality and are, therefore, are less accurate and cannot handle well choice data that involve a large number of attributes m and are noisy (Vapnik 1998). It is therefore important to augment this model to avoid overfitting, hence improve accuracy performance and handle noise better.

Statistical learning theory suggests that this can be achieved by controlling the *complexity* of the model estimated: a very complex model (i.e., a polynomial of very high degree) may fit the estimation data perfectly but has the danger of overfitting and being sensitive to noise, while a very simple model may not be powerful enough to capture the relations in the data. Therefore one needs to control the complexity of the model in some way. Appropriate measures of complexity that are *not* necessarily related to the number of parameters estimated have been defined in the past (Vapnik and Chervonenkis 1971, Alon et al. 1993, Vapnik 1998). Discussing them is beyond the scope of this paper, and we refer the reader to Vapnik and Chervonenkis (1971) and Tapnik (1998) for more information. The main point to emphasize is that it is necessary to include a complexity control to avoid overfitting and handle noise and a large number of attributes better (Vapnik 1998).

We now present a way to do this as in the case of SVM, a method for classification and regression developed within the framework of statistical learning theory and widely used with a lot of success for other data analysis problems (Vapnik 1998). We briefly describe SVM in the Technical Appendix available at the *Marketing Science* website (http://mktsci.pubs.informs.org), and we refer the reader for more information on this rich area of research to (Vapnik 1998) and to www.kernel-machines.org.

As a final note, existing methods for preference modeling such as logistic regression fit the data without controlling for the complexity of the model; therefore, as we see in the experiments section below, they are less accurate and tend to be more sensitive to noise than the methods we discuss here.

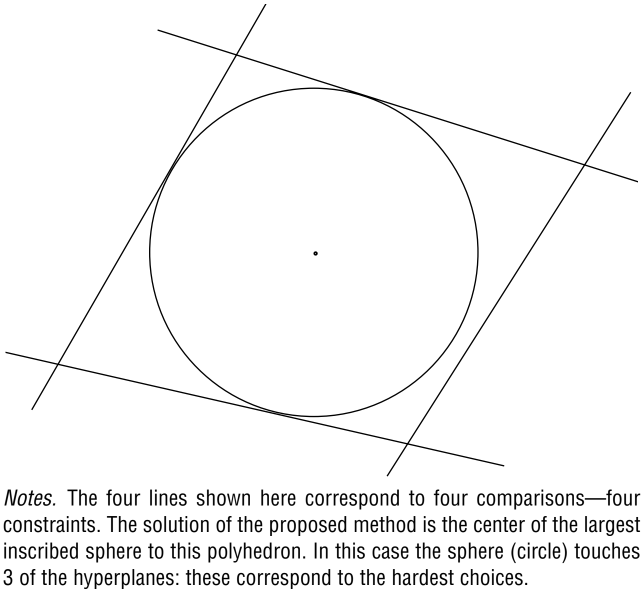
2.3. Robust Preference Modeling Methods

Choosing the appropriate complexity control is an important subject of ongoing research. In our case the main question is *what are the characteristics of a utility function that make it complex?* For example, in the case of regression, a standard measure of complexity of a model/function is how "smooth" the function is—formally the L_2 integral of its first derivative (Wahba 1990, Tikhonov and Arsenin 1977, Vapnik 1998, Girosi et al. 1995). This is in a sense a "natural" measure of complexity for regression. However, it is not clear if having a "smooth" utility function is necessarily "natural." We note that the dimensionality of the space where a function is estimated (i.e., the number of attributes m of the products) or the number of parameters used to represent the utility function can be, but *do not need to be*, measures of complexity (Vapnik 1998).

We use a model complexity control that is standard for other data analysis methods (Vapnik 1998, Wahba 1990, Girosi et al. 1995), such as for SVM (see the Technical Appendix at http://mktsci.pubs.informs.org). Intuitively, we require that constraints (1) hold (when they are feasible) with some "confidence margin." We would like to find a function that assigns to the preferred products utility that is larger than that assigned to the nonpreferred products *by as high an amount as possible*. The idea of requiring a margin is analogous to the method proposed by Srinivasan (1998). Geometrically the intuition is as follows. Consider first the case where the feasible area defined by the comparison constraints (1) is nonempty—so the ξ's in (2) are all 0. If we represent each constraint (1) as a hyperplane in the space of parameters w_f, as shown in Figure 1, then the feasible area is a polyhedron in that space. If we search among all functions that fit the data perfectly, then any solution $\mathbf{w}$ in this feasible area will do. Instead we choose the solution point in the feasible area that is the *furthest* from the constraints, therefore satisfying the *hardest* comparison constraints *the most*. This is the center of the largest inscribed sphere in the feasible area (Vapnik 1998) shown in Figure 1. It can be shown that the "margin" with which the constraints are satisfied, which is also the radius of the largest inscribed sphere, is equal to $1/\|\mathbf{w}\|^2$ (Vapnik 1998); hence by minimizing $\|\mathbf{w}\|^2$ we can maximize the "margin" with which the chosen products are preferred by the other ones. This is what we do below. In the case where empirical errors exist (there are nonzero slack variables ξ_i) the intuition is similar: we want to minimize the amount of error while satisfying the correct comparisons "as much as possible." We also refer the reader to Bennett and Bredensteiner (2000) for another intuitive geometric interpretation of both the case where the feasible area is nonempty and the case where it is empty.

The proposed method is to *simultaneously* minimize the error we make on the example data via minimizing the slack variables ξ_i *and* maximize the margin with which the solution satisfies the constraints. As in the case of SVM it can be shown (Vapnik 1998) that this is achieved through the following optimization problem. (For simplicity we omit the mathematical

Figure 1 Each of the Constraints (1) Defined by a Comparison is a Hyperplane in the Space of Parameters w_f



Notes. The four lines shown here correspond to four comparisons—four constraints. The solution of the proposed method is the center of the largest inscribed sphere to this polyhedron. In this case the sphere (circle) touches 3 of the hyperplanes: these correspond to the hardest choices.

derivation and we refer the reader to Vapnik 1998 for it.)

$$\min_{w_1,\ldots,w_m,\xi_i} \sum_{i=1,\ldots,n} \xi_i + \lambda \sum_{f=1,\ldots,m} w_f^2$$

subject to:

$$\begin{aligned} &w_1 \cdot x_i^1(1) + w_2 \cdot x_i^1(2) + \cdots + w_m \cdot x_i^1(m) \\ &\quad \geq w_1 \cdot x_i^2(1) + w_2 \cdot x_i^2(2) + \cdots + w_m \cdot x_i^2(m) + 1 - \xi_i \\ &\qquad \text{for} \quad \forall\, i \in \{1, \ldots, n\}, \text{ and } \quad \xi_i \geq 0. \qquad (3) \end{aligned}$$

Notice the following:

(1) The role of the constant 1 (clearly any other constant would also work, since the solution **w** is defined up to a scale factor, as long as we also appropriately change parameter λ) at the constraints: slack variables ξ_i are nonzero (hence we pay a cost) *both* for constraints corresponding to comparisons that are not satisfied by the estimated utility function, *and* for those satisfied but with "confidence margin" less than 1. For the scaling of the solution we found (Vapnik 1998).

(2) In the case where all ξ are zero, the radius of the inscribed sphere in Figure 1 is equal to $1/\|\mathbf{w}\|^2$, so, effectively, the smaller $\|\mathbf{w}\|^2$, the larger the radius, which is the reason we minimize $\|\mathbf{w}\|^2$ in (3) (Vapnik 1998). It turns out that the smaller $\|\mathbf{w}\|^2$ is, the smaller the complexity of the estimated model is (Vapnik 1998).

(3) Parameter λ controls the *trade off* between fitting the data ($\sum_i \xi_i$) and the complexity of the model ($\|\mathbf{w}\|^2$). There are a number of ways to choose parameter λ (Wahba 1990, Vapnik 1998). For example it can be chosen so that the prediction error in a small validation set is minimized or through cross validation (also called leave-one-out error) (Wahba 1990). Briefly, the latter is done as follows.

For a given parameter λ we measure its leave-one-out error as follows: for each of the n choice data, we estimate a utility function using (3) only with the remaining $n-1$ data and test if the estimated function correctly chooses the right product for the choice data point not used (left out). We then count the number of errors in the n choice points when they were left out. This is the cross validation (leave-one-out) error for the parameter λ. We then choose the parameter λ with the smallest cross validation error.

Cross validation is used when we can assume that the future data (choices) come from the same distribution as the data used for estimation (Vapnik 1998). However, in conjoint analysis this may not always be the case. Such is the situation, for example, when the estimation data come from an orthogonal design: the orthogonal design is not a sample from the probability distribution of future choices. So cross validation cannot be formally used with an orthogonal design, but in practice one can still use it, as we also do in the experiments below. Instead, a validation set approach can, for example, be used in such cases. We need to assume that the validation data come from the same probability distribution as the future data.

In the experiments below we tuned λ using cross validation. We chose, using line search, a lambda between 0.001 and 100 (samples every order of magnitude only). Because we have a few data for each individual we use the same λ for all individuals that we choose using the average cross validation error across the individuals.

The idea of finding a central point in the feasible area defined by the data-imposed constraints is not new (Srinivasan and Shocker 1973; Toubia et al. 2003, 2004). For example, Toubia et al. (2004) choose the analytic center of the polyhedron described above slightly modified to take into account other constraints. We believe that both choices lead to models that are robust to noise, as also shown by the experiments below. A key difference of (3) from the method of Toubia et al. (2004) is that in our case we optimize *both* the error on the data *and* the complexity of the solution $\|\mathbf{w}\|^2$ simultaneously using the trade-off parameter λ. Toubia et al. (2004) do not handle this trade off between error and complexity through a simultaneous optimization. Other than including a complexity control, the accuracy performance as well as the robustness of a method to noise depends on how this trade off between error on the data and complexity is handled (Vapnik 1998). We conjecture that the difference in performance between our method and the method of Toubia et al. (2004) shown in the experiments below is due to the difference between the way these two methods handle this trade off.

In our case we incorporate a trade off parameter λ that controls "how much" the constraints need to be satisfied, and we develop a family of methods that are almost equivalent to the well-known method of SVM classification (see the Technical Appendix) with similar useful characteristics, namely:

- the estimation is done through fast quadratic programming optimization with box constraints, namely constraints that give only upper and lower bounds to the parameters to be estimated;
- the estimated utility function turns out to depend only on certain data—the "hard choices" that are the hyperplanes touching the inscribed sphere in Figure 1—that are automatically detected;
- the generalization to highly nonlinear utility functions—that turn out to be linear in parameters (Vapnik 1998)—is straightforward and computationally efficient;
- probabilistic guarantees on the future performance of the estimated model can be given under

certain assumptions about the probability distribution of the data. In particular, it can be shown that the predictive performance of the estimated models—that is, how often the estimated utility assigns higher utility to the correct product for future choices—increases as $\|\mathbf{w}\|^2$ (which controls the confidence margin on the estimation data as discussed above) and the error $\sum_i \xi_i$ decrease and as the number of data n increases (Vapnik 1998).[1]

Next, we discuss these briefly and we refer the reader to the Technical Appendix for more details on each of these points.

Finally, we note that the solution $\mathbf{w}$ of (3) need not be positive. In practice we may want to impose the constraint that parameters w_f are positive or (equivalently) to incorporate prior knowledge about the base level for each product attribute when we use levels to describe the product attributes (Toubia et al. 2003). We show in Appendix A below how to augment model (3) using virtual examples (Scholkopf et al. 1996) to include such positivity constraints or to incorporate knowledge about the base level of an attribute. In the experiments below we added for all methods of positivity constraints capturing prior knowledge about the base level of the product attributes, since the polyhedral method of Toubia et al. (2004) requires the use of such constraints.

2.4. Characteristics of the Method and Nonlinear Extension

2.4.1. Dual Parameters and Hard Choices. It turns out that as in the case of SVM (see Technical Appendix at http://mktsci.pubs.informs.org) the utility function estimated through (3) can be written in the form:

$$U(\mathbf{x}) = \mathbf{w}\cdot\mathbf{x} = \sum_{i=1}^{n} \alpha_i^*(\mathbf{x}_i^1 - \mathbf{x}_i^2)\cdot\mathbf{x} \tag{5}$$

where α_i^* are the dual parameters (Bertsimas and Tsitsiklis 1997) corresponding to the dual optimization problem of (3). The dual problem is a quadratic programming problem with box constraints that has a unique optimal solution and is quickly solved in practice. See the Technical Appendix (http://mktsci.pubs.informs.org) and Cortes and Vapnik (1995). It can be shown (Vapnik 1998) that for the optimal solution (5), and for SVM in general, only a few of the coefficients α_i^* are nonzero. These are the coefficients α_i that correspond to the pairs of products $(\mathbf{x}_i^1, \mathbf{x}_i^2)$ hard to choose from. In other words *the utility function model developed from a set of choices is specified only by the "hard" choices, which are automatically found by the method*. This is in agreement with the intuition that preferences are shaped by the hard choices one has to make. Moreover, although we do not deal with this issue here, intuitively one could also use this characteristic of the proposed method to design questionnaires in the spirit of Toubia et al. (2003). For example there has been work in the area of active learning see for example (Tong and Koller 2000)—that can be used for this problem. It is interesting to note that the questionnaire design approach of Toubia et al. (2003) is similar in spirit with the active learning methods in the literature. We plan to explore this direction in future work.

2.4.2. Nonlinear Models. By estimating the utility function in its dual form (5), one can also estimate nonlinear functions efficiently even if the number of primal parameters w_f is very large even if it is infinite (Vapnik 1998, Wahba 1990). This is done by solving the dual optimization problem of (3), therefore always optimizing for the n free parameters α_i of (5) *independent of the dimensionality* of the "data" $\mathbf{x}$. Consider for example the case where we model the utility using a polynomial of degree 2, and assume that we only have 2-attribute products. The utility of a product $(x(1), x(2))$ is therefore $w_1x(1)^2 + w_2x(2)^2 + w_3x(1)x(2) + w_4x(1) + w_5x(2)$. Notice that we include the interaction of the attributes. Instead of estimating the 5 parameters w_f, we estimate the n dual parameters α_i, where n is the number of data (comparisons). The utility function is then of the form:

$$\begin{aligned} U(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i^*\big[&x_i^1(1)^2 - x_i^2(1)^2, x_i^1(2)^2 - x_i^2(2)^2,\\ &x_i^1(1)x_i^1(2) - x_i^2(1)x_i^2(2), x_i^1(1) - x_i^2(1),\\ &x_i^1(2) - x_i^2(2)\big]\\ &\cdot\big(x(1)^2, x(2)^2, x(1)x(2), x(1), x(2)\big). \end{aligned}$$

The dual formulation is always a quadratic programming optimization problem with box constraints and number of variables (α_i) equal to n, the number of constraints in (3) (Vapnik 1998). We discuss this further in the Technical Appendix available at the *Marketing Science* website (http://mktsci.pubs.informs.org). So the number of variables w_f in the primal

[1] The following theorem is well known for SVM (Vapnik 1998, Evgeniou et al. 2000a): With probability $1-\eta$, the probability ϵ that a future point is misclassified by a support vector machine classification solution that makes k misclassifications on n example data and has margin $\|w\|^2$ on this data is bounded by:

$$\epsilon < \frac{k}{n} + \Phi(\|\mathbf{w}\|^2, n, \eta), \tag{4}$$

where Φ is decreasing with n and η and increasing with $\|w\|^2$. One can also replace k with $\sum_i \xi_i$ and use a different Φ. For simplicity we do not give the form of Φ here and refer the reader for example to Vapnik (1998) and Evgeniou et al. (2000a). We note that this theorem holds only if the n data (product differences) are i.i.d., which is not necessarily the case for the preference modeling setup. It is an open question how to extend this theorem to the conjoint estimation case. This theorem currently provides only an informal motivation for the proposed approach.

formulation (3) is not important (Vapnik 1998): products with a very large number m of attributes, as well as highly nonlinear utility functions that, for example, include all high order interactions among attributes can be computationally efficiently estimated in a robust way.

2.4.3. Handling Heterogeneity. The method discussed so far assumes that the data come from a single true underlying utility function and we estimate one utility function. In practice the data may come from many individuals and therefore from different underlying utility functions. A state-of-the-art approach to handling such data is by assuming a priori that all utility functions come from a probability distribution, for example a (unknown) Gaussian, and then estimating all utility functions simultaneously through also estimating the parameters of this distribution, as it is done in the case of hierarchical Bayes (Lenk et al. 1996, DeSarbo et al. 1997, Allenby et al. 1998, Arora et al. 1998).

Developing methods along the lines of the ones presented here that can be used to simultaneously estimate many utility functions that are assumed to be related in some way (i.e., all come from the same, unknown, Gaussian distribution) is a direction for future research. Some possible directions can be, for example, along the lines of boosting (Freund and Schapire 1997, Friedman et al. 1998) or learning with heterogeneous kernels (Bennett et al. 2002). The issue is an open one also in the area of statistical learning theory.

In this paper we compare our approach with hierarchical Bayes even though we simply estimate one utility function for each individual independently: Hence HB has a relative advantage in the experiments since it combines information across all individuals. We also investigate along the direction of combining the models estimated for each individual following a simple ad hoc approach as in Toubia et al. (2004) briefly as follows.

First we estimate one model, for example one linear utility function $\mathbf{w}_k$, for each individual k independently. We then take the mean of the estimated models $\mathbf{w} = 1/N \sum_k \mathbf{w}_k$, where N is the number of individuals. Finally, for each individual we replace $\mathbf{w}_k$ with $\gamma_k \mathbf{w}_k + (1 - \gamma_k)\mathbf{w}$. Parameters γ_k are between 0 and 1 and we estimate them by minimizing the mean square error of $\gamma_k \mathbf{w}_k + (1 - \gamma_k)\mathbf{w}$ from the true utility function of each individual k. This gives an upper bound on the performance that can be achieved if we were to estimate γ_k using only the available data as should be done in practice (in practice a validation set can be used to set the parameters γ_k as for the case of parameter λ discussed above). Although this is a very simple and ad hoc approach to handling heterogeneity, the experiments show that the proposed direction is promising. We plan to explore this direction in the future.

3. Experiments

We run Monte Carlo simulations to study the performance of the methods under varying conditions. Simulations have often been used in the past to study preference modeling methods (i.e., Carmone and Jain 1978, Toubia et al. 2003, Andrews et al. 2002). They are useful, for example, in exploring various domains in order to identify strengths and weaknesses of methods. Below we explore domains that vary according to noise (magnitude) and respondent heterogeneity. We used simulations to compare our methods with logistic regression, the recently proposed polyhedral estimation method of Toubia et al. (2004), and with HB for heterogeneous data, considered a state-of-the-art approach. It is important to note that *in all cases we generated the data in a way that gives an advantage to logistic regression and HB*—that is, the data were generated according to the probability distributions *assumed* by these methods. Moreover, the comparison with HB is not well defined since our methods are for individual utility estimation while HB uses information across many individuals. The simple ad hoc extension to handle heterogeneity that we described above, labeled as "SVM Mix" below, is the only method that can be directly compared with HB.

3.1. Design of Simulations

For easy comparison with other work in the literature we followed the basic simulation design used by other researchers in the past. In particular we simply replicated the experimental setup of Toubia et al. (2004), which in turn was based on the simulation studies of Arora and Huber (2001). For completeness we briefly describe that setup.

We generated data describing products with 4 attributes, each attribute having 4 levels. Each question consisted of 4 products to choose from. The question design we used was either orthogonal or randomly generated. For the orthogonal design to be well defined we used 16 questions per individual as in Toubia et al. (2004). The random design is a closer simulation for data that are not from questionnaires, such as more unconstrained consumer choice data.

We simulated 100 individuals. The partworths for each individual were generated randomly from a Gaussian with mean $(-\beta, -\frac{1}{3}\beta, \frac{1}{3}\beta, \beta)$ for each attribute. Parameter β is the magnitude that controls the noise (response accuracy). As in Toubia et al. (2004) we used $\beta = 3$ for high magnitude (low noise) and $\beta = 0.5$ for low magnitude (high noise). We modeled heterogeneity among the 100 individuals by varying the variance σ^2 of the Gaussian from which the partworths were generated. The covariance matrix of the

Gaussian was a diagonal matrix with all diagonal elements being σ^2. We modeled high heterogeneity using $\sigma^2 = 3\beta$, and low heterogeneity using $\sigma^2 = 0.5\beta$, as in Toubia et al. (2004). As discussed in Arora and Huber (2001) and Toubia et al. (2003) these parameters are chosen so that the range of average partworths and heterogeneity found in practice is covered.

Notice that for each of the four attributes the mean partworths are the smallest for the first level and the largest for the fourth level—in increasing order. Because the method of Toubia et al. (2004) requires that constraints about the relative order of the actual partworths for each level relative to the lowest level are added (in the form of positivity constraints—Toubia et al. 2004), we incorporated this information to all other methods. For our method and for logistic regression this was done using the virtual examples approach discussed in Appendix A. For the case of HB this was done by simply constraining the sampling from the posterior during the HB estimation iterations to be such that we only use partworth samples for which the lowest levels are the same ones as the actual lowest levels. Adding constraints to HB can be done in other ways, too, as discussed in Sawtooth Software (see, for example, the Sawtooth Software website), but none of them is standard. Notice that the relative order may be changing as we sample the partworths for the four levels: We incorporated constraints about the actual lowest levels and not the lowest levels of the mean partworths.

Finally, all experiments were repeated five times, so a total of 500 individual utilities were estimated, and the average performance is reported.

3.2. Experimental Results

We compare the methods using the RMSE of the estimated partworths. Both estimated and true partworths were always normalized for comparability. In particular, as in Toubia et al. (2004), each attribute is made such that the sum of the levels is 0, and the utility vector is then normalized such as the sum of the absolute values is 1. We also measured the predictive performance (hit rate) of the estimated models by generating 100 new random questions for each individual and testing how often the estimated utility functions predict the correct winning product. In the table below we report the hit rates below the RMSE errors.

Table 1 shows the results. The format of the table is the same as that of Arora and Huber (2001) and Toubia et al. (2004). We label our method as "SVM" since it is very similar to SVM classification. The polyhedral method of Toubia et al. (2004) is labeled as "Analytic"—the method is called Analytic Center in Toubia et al. (2004). We also tested the simple method discussed in §2.4.3 to take into account information across the 100 individuals. The results are shown in the last column of Table 1 with label "SVM Mix."

We performed two significance tests: (a) One to compare only the analytic center method, logistic regression, and the method proposed here—the three methods that do not combine information across individuals. The best of the first three columns is reported in bold. (b) One to find the best among all columns (including HB and SVM Mix) which we report with a "*."

From Table 1 we observe the following:

- SVM significantly outperforms both the analytic center method and logistic regression, the latter for

Table 1 Comparison of Methods Using RMSE and Hit Rates in Parenthesis

Mag	Het	Design	Analytic	SVM	Logistic	HB	SVM Mix
L	H	Random	0.92	**0.69**	0.77	0.60*	0.64
			79.1%	81.8%	81.1%	84.5%	83.1%
L	H	Orthogonal	0.75	**0.66**	**0.67**	0.56*	0.61
			81.2%	82.7%	82.7%	85.5%	83.9%
L	L	Random	1.15	**0.86**	1.00	0.66*	0.69*
			74.5%	77.4%	75.7%	82.6%	81.7%
L	L	Orthogonal	0.89	**0.81**	**0.83**	0.62*	0.67
			76.9%	78.6%	78.3%	83.8%	82.3%
H	H	Random	0.67	**0.53**	**0.52**	0.46*	0.48*
			84.0%	85.9%	86.9%	88.2%	87.2%
H	H	Orthogonal	0.81	**0.61**	**0.59**	0.49*	0.51*
			80.4%	84.1%	84.9%	87.3%	86.3%
H	L	Random	0.65	**0.52**	**0.53**	0.35*	0.37*
			83.3%	86.0%	85.8%	90.3%	89.5%
H	L	Orthogonal	0.81	**0.68**	**0.65**	0.34*	0.53
			79.2%	81.3%	82.8%	90.6%	85.8%

Notes. The true utilities are linear and linear utility models are estimated. Bold indicates best or not significantly different than best at $p < 0.05$ among analytic center, SVM, and logistic regression—the first three columns only. With a * we indicate the best among all columns.

the random designs and when there is noise. It is never worse than logistic regression or the analytic center method.

- Both SVM and SVM Mix are relatively better for the random design. For example SVM is similar to logistic regression in all orthogonal design cases. We believe this is partly due to the problem with choosing parameter λ for the orthogonal design, as discussed above, and because in general the future data come from a different probability distribution than the estimation data. This limitation also indicates that it may be important to combine the proposed method with a similar method for designing questionnaires. As shown by Toubia et al. (2004) such an extension to questionnaire design can lead to significant improvements. We leave this as part of future work.
- The proposed method significantly outperforms both logistic regression and the analytic center method when there is noise for the random design. The performance drop from high magnitude to low magnitude, namely when noise increases, is significantly lower for SVM than for both logistic regression and the analytic center for the random design. It is significantly lower than logistic regression for the orthogonal design but larger than the analytic center method in that case. However the latter is always significantly worse than the proposed method. The proposed method is therefore overall *more robust to noise* than the other methods. We also note that for our method the performance drop from low to high noise is influenced by the relative λs used since different λs are used for the high and low magnitudes (chosen using cross-validation).
- Heterogeneity: the simple extension (SVM Mix) shows promising results. For the random design, HB is better only in the case of low magnitude and high heterogeneity, while in all other cases HB and SVM Mix perform similarly. This, coupled with the fact that the proposed method is computationally efficient while HB is not (Sawtooth Software), indicates the potential of the proposed approach.

3.3. Estimation of Nonlinear Models

The next set of experiments considers the case where the true underlying utility function of each individual (respondent) is nonlinear. In order to account for the nonlinear effect, we estimate nonlinear models as described in §2.4.2. and in the online appendix. A typical nonlinear effect in consumer preferences is simple interaction between two different product attributes (i.e., price and brand). In our case, adding interactions among all product attribute levels (all 16 dimensions) would lead to a large number of parameters to estimate ($15 * 16/2 = 120$) which would be computationally intractable for HB. Therefore we only added the interactions between the first two attributes. Since each attribute has 4 levels we added an extra $4 \times 4 = 16$ dimensions capturing all interactions among the 4 levels of the first attribute and the 4 levels of the second one. Thus, the utility function of each individual consisted of the original 16 parameters generated as before, plus 16 new parameters capturing the interactions among the levels of the first two attributes (clearly, without loss of generality, other choices could be made).

These new 16 parameters were generated from a Gaussian with mean 0 and standard deviation σ_{nl}. The size of σ_{nl} controls the size of the interaction parameters of the underlying utility functions. We assume that we do not know the sign of the interaction coefficients other than for the method of Toubia et al. (2004). The method of Toubia et al. (2004) requires prior knowledge of the least-desired level of each feature, so we incorporated this information (in the form of positivity constraints for the interaction coefficients) for the analytic center, effectively giving that method an advantage relative to the other ones. In practice we may not know the sign of the interaction coefficients, so we did not add this information to the other three methods.

Our objective is to experiment with two different levels of nonlinearity: low nonlinearity and high nonlinearity. Our definition of "level of nonlinearity" is given below in the form of a short sequence of computations. For a given σ_{nl}:

(1). Draw a random population of 1,000 utility functions (1,000 individuals);

(2). Generate the set of all $4 \times 4 \times 4 \times 4 = 256$ possible 4-attribute products;

(3). For each individual and each product, compute the absolute values of the nonlinear and linear parts of the utility of the product separately;

(4). For each individual, add all 256 absolute values of the nonlinear parts and the linear parts separately, and take the ratio between the sum-absolute-nonlinear and the sum-absolute-linear;

(5). Compute the average of this ratio over the 1,000 individuals.

We use the average ratio computed in the last step as a characterization of the relative size of the underlying nonlinear (interaction) effect in the simulated population. In the sequel, we present experiments for the cases where the average-ratio is 25% (low nonlinearity) and 75% (high nonlinearity). In other words, over all possible products the average nonlinear part of the utility is about 25% (low nonlinearity) or 75% (high nonlinearity) of the linear part. The values of σ_{nl} that result in the specified levels of nonlinearity are:

- for low magnitude and high heterogeneity: 0.61 and 1.84 (low and high nonlinearity, respectively);
- for high magnitude and low heterogeneity: 1.26 and 3.80 (low and high nonlinearity, respectively).

To estimate the nonlinear utilities using logistic regression and HB we represented the data using 32 dimensional vectors (16 linear plus 16 nonlinear). For the method of Toubia et al. (2004) we followed the suggestion in Toubia et al. (2003): we introduced an additional feature with 16 levels corresponding to the 16 nonlinear interaction parameters. Therefore the three methods (other than SVM which as we discussed always estimates the n dual parameters α_i) estimated 32 parameters for each individual. Notice that HB can hardly handle even this low dimensional nonlinear case (see, for example, the Sawtooth Software website), which in contrast is a computationally mild case for the polyhedral method, SVM, and logistic regression. For computational reasons (for HB) and to avoid cluttering, we only did experiments in two cases:

- high magnitude and low heterogeneity—the "easiest" case in practice;
- low magnitude and high heterogeneity—the "hardest" case in practice, and also the case where our method has the least advantage relative to HB as shown in Table 1.

For computational reasons we also simulated 100 individuals only once (instead of 5 times in the linear experiments case) for these experiments.

In Table 2 we compare only SVM Mix and HB, since the conclusions about the comparison of the polyhedral method, SVM, and logistic regression are similar as for the linear utility experiments. We show the performances of the logistic, SVM, and polyhedral methods in Appendix B. Although the actual utilities are nonlinear, we also estimated linear models to see if it is even worth estimating nonlinear models to begin with. To compare the linear and nonlinear models we use hit rates: the percentage of correct prediction of 100 out-of-sample choices. In Appendix B we report other RMSE errors. In Table 2 we also report the RMSE of the nonlinear parts of the utility functions, which captures the accuracy with which the 16 interaction coefficients are estimated. Therefore in Table 2 we show the hit rates of linear SVM Mix with the mixture parameter γ estimated as in the linear experiments; nonlinear SVM Mix where now we estimated two mixture parameters γ_l and γ_{nl} for the linear and nonlinear parts of the estimated utility using again the method outlined in the linear experiments; linear HB; and nonlinear HB. In parenthesis, for the nonlinear models, we report the RMSE of the interaction coefficients (the 16 coefficients for the nonlinear part of the utility function).

The results show the following:

- When the nonlinearity is low the linear models are generally better than the nonlinear ones.
- When nonlinearity is high, it is generally better to estimate nonlinear models both for HB and for SVM Mix.
- The best (among linear and nonlinear) HB outperforms the best (among linear and nonlinear) SVM Mix in the cases it outperformed it in the linear experiments (Table 1). However, the relative differences of the hit rates decrease as the amount of nonlinearity increases. For example, in the high nonlinearity case the nonlinear SVM Mix is similar to HB in three out of the four cases (Low-High or High-Low for random and orthogonal), while in Table 1, SVM Mix is similar to HB only in one out of the four cases. This indicates that the proposed approach has a relative advantage when there are nonlinearities.
- When we estimate nonlinear models, the RMSE of the nonlinear part of the estimated function is smaller for SVM Mix than for HB. In other words, the proposed method captures the nonlinear interactions better than HB. In Appendix B we show that a simple SVM (not "Mix") is also on average better than any other method in terms of capturing the nonlinear effects.

4. Conclusions

Preference modeling has been a central problem in the marketing community and is becoming increasingly important in other business areas such as in supply chain and procurement where the procurement

Table 2 **The True Utilities are Nonlinear**

Mag	Het	NL	Des	SVM Mix Lin	SVM Mix NL	HB Lin	HB NL
L	H	L	Rand	81.6%	81.1% (**1.43**)	**82.7%**	81.5% (1.56)
			Orth	81.7%	81.0% (**1.49**)	**82.7%**	80.7% (1.61)
L	H	H	Rand	75.3%	**78.1%** (**1.15**)	76.2%	**78.6%** (1.33)
			Orth	75.2%	**76.6%** (**1.30**)	75.4%	**76.1%** (1.50)
H	L	L	Rand	**87.9%**	87.6% (**1.48**)	**88.2%**	**88.4%** (1.57)
			Orth	85.5%	84.1% (**1.48**)	**89.0%**	88.3% (1.61)
H	L	H	Rand	78.6%	**82.6%** (**1.14**)	79.8%	**83.2%** (1.31)
			Orth	77.5%	78.6% (**1.27**)	79.8%	**81.1%** (1.41)

Notes. Hit rates and the RMSE of the estimated interaction coefficients in parenthesis are reported. Bold indicates best or not significantly different than best at $p < 0.05$ across all columns.

processes are automated and data describing past choices are captured. At the same time, the "democratization" of data, in the sense that data is captured everywhere and under any conditions, implies that companies often need to use preference modeling tools that do not assume the data is generated in a controlled environment, i.e., through questionnaires. As the conditions under which preference data are captured vary, and as more and more applications arise, there is an increasing need for new tools and approaches to the problem of preference modeling that are computationally efficient, have high accuracy, and can handle noise and high (multiattribute products) dimensional data. The work presented here aims at opening a direction of research in the area of preference modeling that can lead to such new approaches and tools. We did not discuss here issues such as how to use the proposed framework, for example, for designing questionnaires: We believe this is possible, as we briefly discussed and as is indicated by the work of Toubia et al. (2003), and we leave this for future research. Instead we focused on laying the foundations for methods and tools to solve a variety of preference modeling problems.

In this paper we presented a framework for developing computationally efficient preference models that have high accuracy and can handle noisy and large dimensional data. The framework is based on the well-founded field of statistical learning theory (Vapnik 1998). Highly nonlinear conjoint estimation models can also be computationally efficiently estimated. The models estimated depend only on a few data points, the ones that correspond to "hard choices." This can provide useful insights to managers by focusing their attention only on those choices. Moreover, this characteristic can be used to design individual specific questionnaires along the lines of Toubia et al. (2004).

The experiments showed that:

- The proposed approach significantly outperforms both the method of Toubia et al. (2004) and standard logistic regression, the latter when there is noise and for the random design. It is never worse than the best among these three methods.
- The proposed approach is less sensitive to noise—high response error—than both logistic regression and the method of Toubia et al. (2004). *It is therefore more robust to noise.*
- The proposed approach is relatively weaker when data from an orthogonal design are used. This limitation indicates that it may be important to combine the proposed method with a method similar in spirit for designing questionnaires. As shown by Toubia et al. (2004), such an extension to questionnaire design can lead to significant improvements. We leave this as part of future work.
- A simple method for handling heterogeneity lead to promising results with performance often similar to that of HB;
- When the true underlying utility function is nonlinear (for example there are interaction effects between the product attributes) it is better to estimate nonlinear models when the nonlinearity is high. Moreover the proposed method estimates the interaction coefficients significantly better than all other methods.

Estimation is also computationally efficient, so, for example, large datasets for products with large numbers of attributes can also be used, unlike the case of HB.

A number of extensions are possible within this framework for preference modeling. A clear direction for future work is to incorporate to the individual-specific models cross-respondent information in the case of heterogeneity. The experiments show that even a simple ad-hoc method for handling heterogeneity is already promising. Another important direction is to develop other preference modeling methods using the principles of Statistical Learning Theory: in particular, there is evidence (see for example Rifkin 2002) that the most important part of the proposed approach is the incorporation of the complexity control in the estimation process. It may be the case that logistic regression with complexity control, for example along the lines of Zhu and Hastie (2001), is a more appropriate approach than the one we tested here, since it may better capture the noise model of the data typically assumed in conjoint analysis.

The machinery developed for SVM as well as statistical learning theory can be used for solving problems in the field of conjoint analysis in new ways. For example, one can extend the use of virtual examples we used here (discussed in Appendix A) for adding positivity constraints on the utility function. Empirical evidence shows that if the original data used to estimate a model are extended to include virtual examples, then the performance of the estimated models improves (Scholkopf et al. 1996). Generally, virtual examples are data that are either added to the estimation data by the user because of prior knowledge about them, or are generated from the existing data using transformations that the user knows a priori do not alter their key characteristic (i.e., which product is the preferred one) (Scholkopf et al. 1996). Furthermore, models for metric-based conjoint analysis (Toubia et al. 2003) can also be developed within the framework in this paper, for example in the spirit of SVM regression instead of classification (Vapnik 1998). Finally, another direction of research is to develop active learning (Tong and Koller 2000) type methods for the problem of adaptively designing questionnaires, as for example in Toubia et al. (2003).

Once the problem of preference modeling is seen within the framework of statistical learning theory and SVM, a number of new methods can be developed for the conjoint analysis field. The work in this paper does not aim, by any means, to replace existing methods of preference modeling, but instead to contribute to the field new tools and frameworks that can be complementary to existing ones for solving preference modeling problems. Finally, the experiments presented here are by no means exhaustive: more experiments by other researchers will be needed to establish the relative strengths and weaknesses of the proposed approach, as is always the case with any newly developed method.

Acknowledgments
The authors thank John Hauser, Duncan Simester, and Olivier Toubia for all the support and useful conversations during this work and also thank Olivier Toubia for making available the software programs used for the experiments with the method proposed by Toubia et al. (2003) and with hierarchical Bayes.

Appendix A. Adding Positivity Constraints

The coefficients of a utility function, w_f, in problem (3), are sometimes assumed to be positive; if not, the values of the corresponding attributes can often be negative and have the corresponding coefficients be positive (Toubia et al. 2003). Therefore in practice it is often (but not always) important to add such constraints to the estimation of the utility function. To do so, the estimation method (for example in the simple linear case) should be modified by adding to (3) the extra constraints:

$$w_f \geq 0, \quad \forall f = 1, \ldots, m. \tag{6}$$

However, such a modification makes the generalization of the method to the nonlinear case using kernels impossible (Vapnik 1998). It is therefore not possible to add such constraints directly and still be able to efficiently estimate highly nonlinear models (Vapnik 1998).

To avoid this problem, we use virtual examples (Niyogi et al. 1998, Scholkopf et al. 1996). In particular, the positivity of the m parameters w_f is incorporated in the models by adding m (virtual) example difference vectors

$$\{(1, 0, \ldots, 0), (0, 1, \ldots, 0), \ldots, (0, 0, \ldots, 0, 1)\}.$$

These difference vectors correspond to pairs of products that have all attributes the same apart from one: the product with a higher value (by 1) for the one different attribute is preferred. Formally this modifies problem (3) as follows:

$$\min_{w_1, \ldots, w_m, \xi_i} \left(\sum_{i=1}^{n} \xi_i + \sum_{f=1}^{m} \xi_f \right) + \lambda \sum_{f=1 \ldots m} w_f^2$$

subject to:

$$\begin{aligned} & w_1 \cdot x_i^1(1) + w_2 \cdot x_i^1(2) + \cdots + w_m \cdot x_i^1(m) \\ & \quad \geq w_1 \cdot x_i^2(1) + w_2 \cdot x_i^2(2) + \cdots + w_m \cdot x_i^2(m) + 1 - \xi_i \\ & \qquad\qquad \text{for} \quad \forall i \in \{1, \ldots, n\}, \\ & w_f \geq 1 - \xi_f, \quad \forall f = 1, \ldots, m, \\ & \xi_f \geq 0, \\ & \xi_i \geq 0. \end{aligned} \tag{7}$$

Notice that m constraints of the form $w_f \geq 1 - \xi_f$, m new slack variables ξ_f, and m constraints $\xi_f \geq 0$ have been added. The slack variables ξ_f push the optimal w_f to be positive. Notice that we can further tune the proposed method by putting a different weight C on the ξ_f in the cost function so that we can have w_f being more or less pushed towards positivity. For example, if the cost function is

$$\min_{w_1, \ldots, w_m, \xi_i} \sum_{i=1}^{n} \xi_i + C \sum_{f=1}^{m} \xi_f + \lambda \sum_{f=1 \ldots m} w_f^2 \tag{8}$$

for a very large C, then all w_f will become positive (if there is a positive feasible solution for w_f). This way one can control the requirement that w_f be positive. *In the experiments we have used the simple method where $C = 1$ so equal weight is put on all slack variables.* Parameter C can in practice also be tuned using cross-validation or a validation set. In the nonlinear case—using kernels—the use of the virtual examples will not force only the linear effects of attribute f to be positive, but the overall effects of this attribute to be positive. For example, for a polynomial kernel of degree 2, the virtual examples will force $w_f + w_f^2 \geq 1 - \xi_f$. So in the general nonlinear case the virtual examples as used here will imply that for two products "all else being equal, more of a particular attribute by 1 is better," and this requirement can still be relaxed/controlled by the use of C for the slack variables ξ_f.

The experiments were designed similar to these in Toubia et al. (2004) where the positivity of the underlying utility function is used to capture the assumption that we know for each product attribute which level has the lowest partworth. One can remove that level and assume that all other partworths are positive. If, instead, the products are represented as binary vectors with each attribute corresponding to a number of dimensions equal to the number of levels for that attribute with a 1 at the location of the present level and a 0 elsewhere—often used in practice (Arora and Huber 2001, Toubia et al. 2004) and also in our experiments—then the virtual example corresponding to the prior knowledge that the partworth of a level is the smallest one would be, for example, of the form

$$(1, 0, 0, -1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$$

in the case of 4 attributes with 4 levels each for which we know that for the first attribute the fourth level has the smallest partworth—smaller than the first level in this case. This is the representation we used in the experiments for our method and for logistic regression. Finally we note that one can add other types of prior knowledge to constrain the estimation of the utility function through the use of virtual examples (Scholkopf et al. 1996).

Appendix B. Nonlinear Experiments: Detailed Results

We show all the results of the nonlinear experiments in Table 3. In each cell we report five performances: (a) the

Table 3 Comparison of Methods when Nonlinear Models are Estimated

Mag	Het	NL	Des	Performance	Analytic	SVM	Logistic	HB	SVM Mix
L	H	L	Rand	RMSE Lin	0.95	0.70	0.81	0.62	0.67
				Hit Lin	78.1%	80.8%	79.6%	82.7%	81.6%
				RMSE Lin/NL	0.86	0.69	0.77	0.66	0.65
				RMSE NL/NL	1.55	1.52	1.63	1.56	1.43
				Hit NL	77.6%	80.4%	77.8%	81.5%	81.1%
L	H	L	Orth	RMSE Lin	0.80	0.69	0.70	0.61	0.65
				Hit Lin	78.7%	80.6%	80.6%	82.7%	81.7%
				RMSE Lin/NL	0.83	0.71	0.72	0.65	0.66
				RMSE NL/NL	1.57	1.61	1.70	1.61	1.49
				Hit NL	78.7%	79.8%	80.2%	80.7%	81.0%
L	H	H	Rand	RMSE Lin	1.08	0.85	0.94	0.74	0.79
				Hit Lin	72.8%	75.1%	74.5%	76.2%	75.3%
				RMSE Lin/NL	0.96	0.79	0.81	0.76	0.74
				RMSE NL/NL	1.32	1.20	1.31	1.33	1.15
				Hit NL	77.0%	78.2%	77.1%	78.6%	78.1%
L	H	H	Orth	RMSE Lin	0.95	0.82	0.84	0.75	0.76
				Hit Lin	73.4%	74.4%	73.5%	75.4%	75.2%
				RMSE Lin/NL	0.95	0.79	0.78	0.78	0.74
				RMSE NL/NL	1.42	1.36	1.44	1.50	1.30
				Hit NL	75.6%	76.4%	75.9%	76.1%	76.6%
H	L	L	Rand	RMSE Lin	0.71	0.55	0.58	0.39	0.39
				Hit Lin	81.8%	84.7%	84.1%	88.2%	87.9%
				RMSE Lin/NL	0.78	0.54	0.59	0.41	0.40
				RMSE NL/NL	1.58	1.54	1.61	1.57	1.48
				Hit NL	80.4%	84.4%	83.7%	88.4%	87.6%
H	L	L	Orth	RMSE Lin	0.84	0.67	0.66	0.35	0.51
				Hit Lin	78.9%	81.5%	82.1%	89.0%	85.5%
				RMSE Lin/NL	0.92	0.71	0.70	0.39	0.56
				RMSE NL/NL	1.56	1.55	1.56	1.61	1.48
				Hit NL	76.6%	79.9%	80.7%	88.3%	84.1%
H	L	H	Rand	RMSE Lin	0.95	0.74	0.81	0.52	0.43
				Hit Lin	75.1%	77.1%	76.7%	79.8%	78.6%
				RMSE Lin/NL	0.88	0.67	0.67	0.54	0.44
				RMSE NL/NL	1.32	1.17	1.25	1.31	1.14
				Hit NL	78.6%	80.5%	80.8%	83.2%	82.6%
H	L	H	Orth	RMSE Lin	0.94	0.83	0.77	0.50	0.56
				Hit Lin	74.0%	75.5%	75.9%	79.8%	77.5%
				RMSE Lin/NL	1.03	0.75	0.75	0.49	0.62
				RMSE NL/NL	1.42	1.33	1.31	1.41	1.27
				Hit NL	74.2%	77.7%	77.5%	81.1%	78.6%

RMSE when we estimate a linear model; (b) the out-of-sample hit rate when we estimate a linear model; (c) the RMSE of the linear part of the utility when we estimate a nonlinear model; (d) the RMSE of the nonlinear part when we estimate a nonlinear model; (e) the out-of-sample hit rate of the nonlinear model estimated.

From the results we observe the following:

- **Nonlinear part estimation:** The key result is that SVM Mix estimates the nonlinear parts of the utility functions better than all other methods, including HB. SVM, without combining information across individuals, is also better than both the logistic regression and HB. It should be noted that all methods have large RMSE as compared to the linear estimations. We attribute this to the fact that each 32-dimensional vector describing a product includes just a single nonzero element out of the total 16 nonlinear elements (since only one of the four levels of the two attributes involved for the nonlinearity is nonzero for each product). In contrast, there are 4 nonzero elements out of the 16 linear ones. Effectively, there is little information about the nonlinear part of the utility functions.
- **Linear part estimation:** For the linear parts of the estimated utility function the comparison of SVM, logistic, and polyhedral is qualitatively similar as in the linear experiments (Table 1).
- **Linear part estimation comparison with HB:** The difference between HB and "SVM Mix" for the linear parts of the utility function is relatively smaller than in the linear utility experiments (Table 1). Our method is therefore less influenced by nonlinearities than HB.

References

Allenby, Greg M., Peter E. Rossi. 1999. Marketing models of consumer heterogeneity. *J. Econometrics* **89**(March/April) 57–78.

Allenby, Greg M., Neeraj Arora, James L. Ginter. 1998. On the heterogeneity of demand. *J. Marketing Res.* **35** 384–389.

Alon, Noga, Shai Ben-David, Nicolò Cesa-Bianchi, David Haussler. 1993. Scale-sensitive dimensions, uniform convergence, and learnability. *34th IEEE Sympos. Foundations Comput. Sci.*, Palo Alto, CA, October 1993.

Andrews, Rick, Asim Ansari, Imran Currim. 2002. Hierarchical Bayes versus finite mixture conjoint analysis models: a comparison of fit, prediction, and partworth recovery. *J. Marketing Res.* **39** 87–98.

Arora, Neeraj, Joel Huber. 2001. Improving parameter estimates and model prediction by aggregate customization in choice experiments. *J. Consumer Res.* **28** 273–283.

Arora, Neeraj, Greg Allenby, James Ginter. 1998. A hierarchical Bayes model of primary and secondary demand. *Marketing Sci.* **17**(1) 29–44.

Ben-Akiva, Moshe, Bruno Boccara. 1995. Discrete choice models with latent choice sets. *Internat. J. Res. Marketing* **12** 9–24.

Ben-Akiva, Moshe, Steven R. Lerman. 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press, Cambridge, MA.

Ben-Akiva, Moshe, Daniel McFadden, Makoto Abe, Ulf Bockenholt, Denis Bolduc, Dinesh Gopinath, Takayuki Morikawa, Venkatra Ramaswamy, Vithala Rao, David Revelt, Dan Steinberg. 1997. Modeling methods for discrete choice analysis. *Marketing Lett.* **8**(3) 273–286.

Bennett, Kristin, Erin Bredensteiner. 2000. Duality and geometry in SVM classifiers. Pat Langley, ed. *Proc. Seventeenth Internat. Conf. Machine Learning*. Morgan Kaufmann, San Francisco, CA, 57–64.

Bennett, Kristin, Michinari Momma, J. Embrechts. 2002. MARK: a boosting algorithm for heterogeneous kernel models. *Proc. SIGKDD Internat. Conf. Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada. ACM, New York.

Bertsimas, Dimitris, John Tsitsikilis. 1997. *Introduction to Linear Optimization*. Athena Scientific, Belmont, MA.

Bolduc, Denis, Moshe Ben-Akiva. 1991. A multinomial probit formulation for large choice sets. *Proc. Sixth IATBR Conf.*, Quebec, Canada.

Carmone, Frank, Arun Jain. 1978. Robustness of conjoint analysis: some Monte Carlo results. *J. Marketing Res.* **15** 300–303.

Carroll, Douglas, Paul Green. 1995. Psychometric methods in marketing research: Part I, Conjoint analysis. *J. Marketing Res.* **32** 385–391.

Cortes, Corinna, Vladimir Vapnik. 1995. Support vector networks. *Machine Learning* **20** 1–25.

Cooley R., J. Srivastava, B. Mobasher. 1997. Web mining: Information and pattern discovery on the World Wide Web. *Proc. 9th IEEE Internat. Conf. Tools with Artificial Intelligence (ICTAI'97)*.

Cui, Dapeng, David Curry. 2005. Prediction in marketing using the support vector machines. *Marketing Sci.* Forthcoming.

DeSarbo, Wayne, Asim Ansari. 1997. Representing heterogeneity in consumer response models. *Marketing Lett.* **8**(3) 335–348.

Devroye, Luc, Laszlo Györfi, Gabor Lugosi. 1996. *A Probabilistic Theory of Pattern Recognition, Applications of Mathematics*, No. 31. Springer, New York.

Evgeniou, T., M. Pontil, T. Poggio. 2000a. Regularization networks and support vector machines. *Adv. Comput. Math.* **13** 1–50.

Evgeniou, T., M. Pontil, T. Poggio. 2000b. Statistical learning theory: a primer. *Internat. J. Comput. Vision* **38**(1) 9–13.

Evgeniou, T., M. Pontil, T. Poggio, C. Papageorgiou. 2003. Image representations and feature selection for multimedia database search. *IEEE Trans. Knowledge Data Engrg.* **15** 911–920.

Freund, Yoav, Robert Schapire. 1997. A decision-theoretic generalization of online learning and an application to boosting. *J. Comput. System Sci.* **55**(1) 119–139.

Friedman, Jerome, Trevor Hastie, Robert Tibshirani. 2000. Additive logistic regression: a statistical view of boosting. *Ann. Statist.* **28**(2) 337–407.

Girosi, Federico, Michael Jones, Tomaso Poggio. 1995. Regularization theory and neural networks architectures. *Neural Comput.* **7** 219–269.

Green, Paul, V. Srinivasan. 1978. Conjoint analysis in consumer research: issues and outlook. *Consumer Res.* **5**(2) 103–123.

Green, Paul, V. Srinivasan. 1990. Conjoint analysis in marketing: new developments with implications for research and practice. *J. Marketing* **54**(4) 3–19.

Herbrich, Ralf, Thore Graepel, Klaus Obermayer. 1999. Large margin rank boundaries for ordinal regression. Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf, Dale Schuurmans, eds. *Advances in Large Margin Classifiers*. MIT Press, Cambridge, MA, 29–53.

Jedidi, Kamel, Sharan Jagpal, Puneet Manchanda. 2003. Measuring heterogeneous reservation prices for product bundles. *Marketing Sci.* **22**(1) 107–130.

Kearns, Michael, Robert Schapire. 1994. Efficient distribution-free learning of probabilistic concepts. *J. Comput. Systems Sci.* **48**(3) 464–497.

Kohavi, Ron. 2001. Mining E-commerce data: the good, the bad, and the ugly. *Proc. Seventh ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining*, San Francisco, California. ACM Press, New York, 8–13.

Kuhfeld, Warren F., Randall D. Tobias, Mark Garratt. 1994. Efficient experimental design with marketing research applications. *J. Marketing Res.* **31**(4) 545–557.

Lenk, Peter J., Wayne S. DeSarbo, Paul E. Green, Martin R. Young. 1996. Hierarchical Bayes conjoint analysis: recovery of partworth heterogeneity from reduced experimental designs. *Marketing Sci.* **15** 173–191.

Louviere, Jordan J., David A. Hensher, Joffre D. Swait. 2000. *Stated Choice Methods: Analysis and Applications*. Cambridge University Press, Cambridge, UK.

Manski, Charles F. 1977. The structure of random utility models. *Theory Decision* **8** 229–254.

McFadden, Daniel. 1974. Conditional logit analysis of qualitative choice behavior. Paul Zarembka, ed. *Frontiers in Econometrics*. Academic Press, New York, 105–142.

McFadden, Daniel. 1986. The choice theory approach to marketing research. *Marketing Sci.* **5**(4) 275–297.

Niyogi, Partha, Tomaso Poggio, Federico Girosi. 1998. Incorporating prior information in machine learning by creating virtual examples. *IEEE Proc. Intelligent Signal Processing* **86**(11) 2196–2209.

Oppewal, H., J. Louviere, H. Timmermans. 1994. Modeling hierarchical conjoint processes with integrated choice experiments. *J. Marketing Res.* **31** 92–105.

Pauwels, Koen. 2004. How dynamic consumer response, competitor response, company support, and company inertia shape long-term marketing effectiveness. *Marketing Sci.* **23**(4) 596–610.

Pontil, Massimiliano, Alessandro Verri. 1998. Properties of support vector machines. *Neural Comput.* **10** 955–974.

Rifkin, Ryan. 2002. Everything old is new again: a fresh look at historical approaches in machine learning. Ph.D. thesis, MIT Cambridge, MA.

Sawtooth Software, Inc. HB-Reg: Hierarchical Bayes regression. http://www.sawtoothsoftware.com/hbreg.shtml.

Scholkopf, Bernhard, Chris Burges, Vladimir Vapnik. 1996. Incorporating invariances in support vector learning machines. C. von der Malsburg, W. von Seelen, J. C. Vorbrüggen, B. Sendhoff, eds. *Artificial Neural Networks, ICANN'96, Lecture Notes in Comput. Sci.*, Vol. 1112. Springer, Berlin, Germany, 47–52.

Segal, Madhav N. 1982. Reliability of conjoint analysis: contrasting data collection procedures. *J. Marketing Res.* **19** 139–143.

Srinivasan, V. 1998. A strict paired comparison linear programming approach to nonmetric conjoint analysis. Jay E. Aronson, Stanley Zionts, eds. *Oper. Res. Methods, Models and Applications*. Quorum Books, Westport, CT, 97–111.

Srinivasan, V., Allan D. Shocker. 1973. Linear programming techniques for multidimensional analysis of preferences. *Psychometrica* **38**(3) 337–369.

Srinivasan, V., Arun Jain, Naresh Malhotra. 1983. Improving the predictive power of conjoint analysis by constrained parameter estimation. *J. Marketing Res.* **20** 433–438.

Stone, C. J. 1985. Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705.

Tikhonov, A. N., V. Y. Arsenin. 1977. *Solutions of Ill-Posed Problems*. W. H. Winston, Washington, D.C.

Tong, Simon, Daphne Koller. 2000. Support vector machine active learning with applications to text classification. *Proc. Seventeenth Internat. Conf. Machine Learning*,

Toubia, Olivier, John R. Hauser, Duncan I. Simester. 2004. Polyhedral methods for adaptive choice-based conjoint analysis. *J. Marketing Res.* **41** 116–131.

Toubia, Olivier, Duncan I. Simester, John R. Hauser, Ely Dahan. 2003. Fast polyhedral adaptive conjoint estimation. *Marketing Sci.* **22**(3) 273–303.

Tversky, Amos, Daniel Kahneman. 1974. Judgment under uncertainty: heuristics and biases. *Science* **185** 1124–1131.

Ulrich, Karl T., Steven D. Eppinger. 2000. *Product Design and Development*. McGraw-Hill, Inc., New York.

Vapnik, Vladimir. 1998. *Statistical Learning Theory*. Wiley, New York.

Vapnik, Vladimir, Alexey Chervonenkis. 1971. On the uniform convergence of relative frequences of events to their probabilities. *Theory Probab. Appl.* **17**(2) 264–280.

Wahba, Grace. 1990. Splines Models for Observational Data. *Series in Applied Mathematics*, Vol. 59. SIAM, Philadelphia, PA.

Wittink, Dick R., Philippe Cattin. 1989. Commercial use of conjoint analysis: an update. *J. Marketing* **53**(3) 91–96.

Zhu, Ji, Trevor Hastie. 2001. Kernel logistic regression and the import vector machine. *Proc. NIPS2001*, Vancouver, Canada.