ROBUST CONJOINT ANALYSIS BY CONTROLLING OUTLIER SPARSITY

Gonzalo Mateos and Georgios B. Giannakis

Dept. of ECE, University of Minnesota
200 Union St. SE, Minneapolis, MN 55455, USA
Emails: {mate0058,georgios}@umn.edu

ABSTRACT

Preference measurement (PM) has a long history in marketing, healthcare, and the biobehavioral sciences, where conjoint analysis is commonly used. The goal of PM is to learn the utility function of an individual or a group of individuals from expressed preference data (buying patterns, surveys, ratings), possibly contaminated with outliers. For metric conjoint data, a robust partworth estimator is developed on the basis of a neat connection between ℓ_0-(pseudo)norm-regularized regression, and the least-trimmed squared estimator. This connection suggests efficient solvers based on convex relaxation, which lead naturally to a family of robust estimators subsuming Huber's optimal M-class. Outliers are identified by tuning a regularization parameter, which amounts to controlling the sparsity of an outlier vector along the entire robustification path of least-absolute shrinkage and selection operator solutions. For choice-based conjoint analysis, a novel classifier is developed that is capable of attaining desirable tradeoffs between model fit and complexity, while at the same time controlling robustness and revealing the outliers present. Variants accounting for nonlinear utilities and consumer heterogeneity are also investigated.

1. INTRODUCTION

The growing volume of consumer-generated media provides ample testament to the urgent need for understanding the complex interactions between people and computers. Understanding the dynamics of the emergent socially-intelligent computational systems (SoICS), is a critical task to social and behavioral engineering towards desired collective objectives. SoICS involve human and computer 'actors' whose individual capabilities, values, and preferences determine modes of social engagement. Thus, a holistic approach to *preference measurement, analysis, and management* (PM for short) holds the keys to understanding and engineering SoICS.

PM has a long history in marketing, retailing, product design, healthcare, and also psychology and behavioral sciences, where *conjoint analysis* (CA - the PM 'workhorse') is commonly used [10, 11, 14]. In a nutshell, the goal of PM is to learn the utility function of an individual or group of individuals from expressed preference data (buying patterns, surveys, ratings, recommendations, etc). The pioneering idea behind CA is to decompose consumer preferences, into weights (partworths) of judiciously selected product attributes [10]. This not only allows one to understand the preferences of existing products, but also to *predict* utilities generated by new products obtained as combinations of the studied attributes. With few exceptions, PM has traditionally been an off-line task, assuming mostly 'rational' individuals, clean data, and linear utilities that depend on only a few product attributes. These are very restrictive for existing and forthcoming SoICS, which may involve thousands of underlying variables and include grossly inconsistent 'social liars' or even malicious actors.

To address some of these challenges, the present paper develops novel noise and outlier-robust partworth estimators for both metric and choice-based CA. For metric conjoint data, questionnaire responses (product ratings) are assumed generated from a linear regression model, which explicitly incorporates an unknown *sparse* vector of outliers. The proposed partworth estimator minimizes a tradeoff between fidelity to the training data, and sparsity of the outlier vector encouraged via a natural ℓ_0-(pseudo)norm regularization; or its convex ℓ_1-norm surrogate leading to the least-absolute shrinkage and selection operator (Lasso) [6, 15]. While regularization for model complexity control in conjoint estimation has well-documented merits in terms of generalization capability [7, 8, 5], the major innovative claim here is that *sparsity control* is tantamount to *robustness control*. This is indeed the case since a tunable parameter in Lasso, controls the degree of sparsity in the estimated vector of model outliers. Selection of tuning parameters could be at first thought as a mundane task. However, arguing on the importance of such task as well as devising principled methods to effectively carry out sparsity control, are at the heart of this paper's contribution to the field of CA.

For choice-based CA, a novel sparsity-controlling classifier is developed that is capable of attaining desirable tradeoffs between model fit and complexity, while at the same time controlling robustness and revealing the outliers present. Computer simulations show the effectiveness of the proposed methods.

2. PRELIMINARIES AND ROBUSTNESS

Consider I respondents (e.g., consumers) indexed by $i \in \{1, \ldots, I\}$, each rating J_i profiles represented by the $p \times 1$ vectors $\mathbf{x}_{ij}$, $j \in \{1, \ldots, J_i\}$. Each $\mathbf{x}_{ij}$ comprises p attributes of the profile (or question) j presented to respondent i. Parametric and linear utility functions $u(\mathbf{x})$ are typically adopted for modeling preference measurements [2, 18]. In these models responses $\{y_{ij}\}_{j=1}^{J_i}$ adhere to the linear regression $y_{ij} = \mathbf{x}_{ij}'\mathbf{w}_i + \varepsilon_{ij}$, where $(\cdot)'$ denotes transposition, $\mathbf{w}_i$ is the unknown $p \times 1$ vector of *partworths* for respondent i, and ε_{ij} captures random errors. Such a model describes the three most common types of conjoint data collection formats, namely: **(M1)** *full-profile* ratings, where one question per profile is presented to the respondent; **(M2)** *metric paired-comparison* ratings, where $\mathbf{x}_{ij}$ is replaced by the difference $\tilde{\mathbf{x}}_{ij} := \mathbf{x}_{ij}^{(1)} - \mathbf{x}_{ij}^{(2)}$ of a pair of profiles; and **(M3)** *choice-based* conjoint data, where in addition to taking pairwise differences of profiles, the measurement is the sign of y_{ij} [20]. In words, question j under (M3) asks respondent i to choose between profiles $\mathbf{x}_{ij}^{(1)}$ and $\mathbf{x}_{ij}^{(2)}$; whereas under (M2), the surplus utility of the preferred profile over the other one is also quantified. For simplicity of exposition, focus will be placed first on individual partworth estimates; that is, each $\mathbf{w}_i$ will be estimated separately without fusing information from individual respondents. Subscript i can clearly be dropped in this case. Once the homogeneous case is addressed, approaches to account for consumer heterogeneities are possible along the lines of [8, 20], as discussed in Section 4.3.

Given survey- or questionnaire-based training data $\mathcal{T} := \{y_j, \mathbf{x}_j\}_{j=1}^{J}$, modern statistical learning techniques have been developed to obtain $\mathbf{w}$. Under (M1) or (M2), the task amounts to parameter (or generally function) estimation, whereas under (M3)

† Work in this paper was supported by the NSF grants CCF-0830480, 1016605, and ECCS-0824007, 1002180.

it boils down to a binary classification problem [5, 7, 8]. Following either deterministic or Bayesian formulations, these state-of-the-art techniques rely on suitably regularized loss functions to "optimally" tradeoff complexity for error in the resultant model fit – an approach effecting the desirable generalization capability beyond $\mathcal{T}$ [20].

However, most existing partworth estimators have not accounted for outliers commonly present in large volumes of conjoint data. Outliers can be attributed to multiple factors, including: i) unintentional deviations from the adopted model of e.g., choice-based data; ii) behavioral effects of human respondents, e.g., response errors due to impatient or inattentive responders; and iii) intentional errors caused by malicious responders. Considering for simplicity (M1)[1], the starting point here is to develop a robust estimator of $\mathbf{w}$ that is universal with respect to the outlier model. One such approach is the least-trimmed squares (LTS) estimator given by [16]

$$\hat{\mathbf{w}}_{LTS} := \arg\min_{\mathbf{w}} \sum_{j=1}^{s} r^2_{[j]}(\mathbf{w}) \tag{1}$$

where $r^2_{[j]}(\mathbf{w})$ is the j-th order statistic among the squared residuals $r^2_1(\mathbf{w}), \ldots, r^2_J(\mathbf{w})$, and $r_j(\mathbf{w}) := y_j - \mathbf{x}_j'\mathbf{w}$. The so-termed *coverage* s determines the breakdown point of LTS [16], since $J-s$ profile ratings resulting in the largest residuals are not present in (1). Beyond this universal outlier-rejection property, the LTS estimator is an attractive option due to its high breakdown point and desirable theoretical properties, namely $\sqrt{J}$-consistency and asymptotic normality under mild assumptions [16].

Even though (1) is nonconvex, existence of a minimizer $\hat{\mathbf{w}}_{LTS}$ can be established as follows: i) for each subset of $\{y_j, \mathbf{x}_j\}_{j=1}^J$ with cardinality s (there are $\binom{J}{s}$ such subsets), solve the corresponding ordinary least-squares (LS) problem to obtain a candidate estimator per subset; and ii) pick $\hat{\mathbf{w}}_{LTS}$ as the one among all $\binom{J}{s}$ candidates with the least cost. This solution procedure is combinatorially complex, and thus intractable except for small number of profiles J. Algorithms to obtain (approximate) LTS estimates are available [17].

3. SPARSITY CONTROL FOR ROBUSTNESS

Instead of discarding large residuals, the proposed approach is to model outliers explicitly and estimate them jointly with $\mathbf{w}$. To this end, consider introducing scalar auxiliary variables $\{o_j\}_{j=1}^J$ one per question (rated profile), which take values $o_j \neq 0$ whenever rating j is outlier contaminated, and $o_j = 0$ otherwise. This leads to the preference model $y_j = \mathbf{x}_j'\mathbf{w} + o_j + \varepsilon_j$, where o_j can be deterministic or random with possibly unknown distribution. In this *under-determined* linear regression model, both $\mathbf{w}$ as well as the $J \times 1$ vector $\mathbf{o} := [o_1, \ldots, o_J]'$ are unknown. The percentage of outliers dictates the degree of *sparsity* (number of zero entries) in $\mathbf{o}$. Sparsity control will prove instrumental in efficiently estimating $\mathbf{o}$, rejecting outliers as a byproduct, and consequently arriving at a robust estimate $\hat{\mathbf{w}}$. A natural criterion for controlling outlier sparsity is to seek the estimator

$$(\hat{\mathbf{w}}, \hat{\mathbf{o}}) = \arg\min_{\mathbf{w},\mathbf{o}} \sum_{j=1}^{J} (y_j - \mathbf{x}_j'\mathbf{w} - o_j)^2 + \lambda_0 \|\mathbf{o}\|_0 \tag{2}$$

where $\|\mathbf{o}\|_0$ denotes the nonconvex ℓ_0-norm (equal to the number of nonzero entries of $\mathbf{o}$). Tuning $\lambda_0 \geq 0$ controls sparsity in $\hat{\mathbf{o}}$.

As with compressive sampling and sparse modeling schemes that rely on the ℓ_0-norm, e.g., [21], (2) is also NP-hard. In addition, the sparsity-controlling estimator (2) is intimately related to LTS, as asserted next [13] (proofs are omitted due to space limitations).

[1]Upon replacing $\mathbf{x}_{ij}$ with profile pair differences $\tilde{\mathbf{x}}_{ij} := \mathbf{x}^{(1)}_{ij} - \mathbf{x}^{(2)}_{ij}$, the estimators for model (M1) apply also to model (M2). A robust estimator for choice-based conjoint data (M3) is presented in Section 4.1.

Proposition 1: *If* $\{\hat{\mathbf{w}}, \hat{\mathbf{o}}\}$ *solves* (2) *with* λ_0 *chosen such that* $\|\hat{\mathbf{o}}\|_0 = J - s$, *then* $\hat{\mathbf{w}}_{LTS} = \hat{\mathbf{w}}$ *in* (1).

The importance of Proposition 1 is threefold: i) it formally justifies the additive contamination model and its estimator for robust CA; ii) it links sparse linear regression with robust estimation; and iii) it lends itself naturally to efficient (approximate) solvers based on convex relaxation. Recalling that the ℓ_1-norm $\|\mathbf{o}\|_1$ is the closest convex approximation of $\|\mathbf{o}\|_0$ [21], motivates relaxing (2) to

$$\min_{\mathbf{w},\mathbf{o}} \sum_{j=1}^{J} (y_j - \mathbf{x}_j'\mathbf{w} - o_j)^2 + \lambda_1 \|\mathbf{o}\|_1. \tag{3}$$

This estimator is universally robust, and subsumes Huber's M-estimator for a specific choice of λ_1; see e.g., [9]. Unlike Huber's formulation though, it is not confined to an assumed outlier contamination model. Albeit non-differentiable, (3) can be solved efficiently via e.g., alternating minimization (block-coordinate descent) iterations with guaranteed convergence to the global optimum. Iterations comprise a sequence of LS fits for $\mathbf{w}$, and coordinatewise soft-thresholded updates for $\mathbf{o}$ [13]. Alternatively, it is possible to show that the solutions $\{\hat{\mathbf{w}}, \hat{\mathbf{o}}\}$ of (3) are respectively given by $\hat{\mathbf{w}} := \mathbf{X}^\dagger(\mathbf{y} - \hat{\mathbf{o}}_{\text{Lasso}})$ and $\hat{\mathbf{o}} := \hat{\mathbf{o}}_{\text{Lasso}}$, where $\mathbf{y} := [y_1, \ldots, y_J]'$, $\mathbf{X}^\dagger := (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ with $\mathbf{X} := [\mathbf{x}_1, \ldots, \mathbf{x}_J]'$; and $\hat{\mathbf{o}}_{\text{Lasso}}$ is given by the Lasso esimator [13]

$$\hat{\mathbf{o}}_{\text{Lasso}} := \arg\min_{\mathbf{o}} \|(\mathbf{I}_J - \mathbf{X}\mathbf{X}^\dagger)(\mathbf{y} - \mathbf{o})\|_2^2 + \lambda_1 \|\mathbf{o}\|_1 \tag{4}$$

where $\mathbf{I}_J$ denotes the $J \times J$ identity matrix. Selecting λ_1 along the *robustification path* of Lasso solutions controls the number of outliers rejected. But this choice is challenging because existing techniques such as cross-validation are not effective when outliers are present [16]. To this end, systematic approaches were devised in [9], which leverage the (robustification) path of Lasso solutions available for all values of λ_1 [6, 15] to select the 'best' one dictated by the data. The methods [9] of require either a rough estimate of the percentage of outliers, or, robust estimates $\hat{\sigma}^2_\varepsilon$ that can be obtained using median absolute deviation schemes [16]. Before closing this section, a remark is in order.

Remark 1 In addition to $\mathbf{o}$ it is possible to also promote sparsity and/or smoothness of the partworth vector $\mathbf{w}$ by augmenting the cost in (3) with additional regularization terms entailing its ℓ_1-norm $\|\mathbf{w}\|_1$ and/or its ℓ_2-norm $\|\mathbf{w}\|_2^2$. The former promotes sparsifying the partworth vectors and retaining only the most critical attributes explaining the respondent's preferences. When the number of attributes p is large, parsimonious $u(\mathbf{x})$ can ease managerial decision-making. Ridge-type regularization allows to further control the (model) complexity of the solution, which is important when the responses J are few and p is considerably larger.

3.1 Estimator refinements

Nonconvex regularization. Instead of substituting $\|\mathbf{o}\|_0$ in (2) by its closest convex approximation, namely $\|\mathbf{o}\|_1$, letting the surrogate function to be nonconvex can yield tighter approximations. To this end, consider approximating (2) by the *nonconvex* formulation

$$\min_{\mathbf{w},\mathbf{o}} \sum_{j=1}^{J} (y_j - \mathbf{x}_j'\mathbf{w} - o_j)^2 + \lambda_0 \sum_{j=1}^{J} \log(|o_j| + \delta) \tag{5}$$

where $\delta \approx 0$ is introduced to avoid numerical instability.

Local methods based on iterative linearization of $\log(|o_j| + \delta)$ around the current iterate $o_j(k)$, can be adopted to minimize (5). Skipping details that can be found in [13], this procedure leads to the following iteration for $k = 0, 1, 2, \ldots$

$$[\mathbf{w}(k), \mathbf{o}(k)] = \arg\min_{\mathbf{w},\mathbf{o}} \sum_{j=1}^{J} \left[(y_j - \mathbf{x}_j'\mathbf{w} - o_j)^2 + \omega_j(k-1)|o_j| \right]$$

$$\omega_j(k) = \lambda_0 / \left(|o_j(k)| + \delta \right), \quad j = 1, \ldots, J$$

which altogether amounts to an iteratively reweighted version of (3). To avoid getting trapped in local minima, a good initialization for the iteration is the solution of (3). Numerical tests have shown that a couple iterations of this second-stage refinement suffices to yield improved partworth estimates $\hat{\mathbf{w}}$, in comparison to those obtained from (3). The improvements can be leveraged to bias reduction, also achieved by similar *weighted* norm regularizers.

Outlier rejection. From the equivalence between (3) and Huber's M-estimator, it follows that data $\{y_j, \mathbf{x}_j : j \text{ s.t. } \hat{o}_j \neq 0\}$ deemed as outliers are not completely discarded as with LTS. Instead, their effect is downweighted as per Huber's loss function [16]. Nevertheless, explicitly accounting for the outliers in $\hat{\mathbf{o}}$ provides the means of identifying and removing the contaminated data altogether, and thus possibly re-estimating partworths using the 'clean' data.

4. ROBUST CONJOINT ANALYSIS VARIANTS

4.1 Choice-based robust conjoint analysis

Over the last decade, choice-based CA has become a very popular alternative to metric analysis [11]. For the choice-based data model (M3) however, the approach to retrieve outliers and robustify the binary classifier for CA must be modified. Similar to [7] and for notational simplicity, assume without loss of generality that $\mathbf{x}_j^{(1)}$ is the preferred profile for all questions – otherwise profiles can be renamed accordingly. With this convention consumer responses become $y_j = 1, j = 1, \ldots, J$, and the proposed classifier is given by

$$\min_{\mathbf{w},\mathbf{o}} \sum_{j=1}^{J} \left[\left(1 - \tilde{\mathbf{x}}_j' \mathbf{w}\right)_+ - o_j \right]^2 + \lambda_o \|\mathbf{o}\|_1 + \lambda_w \|\mathbf{w}\|_2^2 \qquad (6)$$

where $(\cdot)_+ := \max(\cdot, 0)$. To gain further intuition as to why (6) is a suitable robust estimator for stated-preference data, introduce *slack* variables $\xi_j \geq 0$ collected in the vector $\boldsymbol{\xi} := [\xi_1, \ldots, \xi_J]'$, and note that (6) is equivalent to the linearly constrained formulation

$$\min_{\mathbf{w},\mathbf{o},\boldsymbol{\xi}} \sum_{j=1}^{J} \left(\xi_j - o_j\right)^2 + \lambda_o \|\mathbf{o}\|_1 + \lambda_w \|\mathbf{w}\|_2^2 \qquad (7)$$

$$\text{s.t. } \tilde{\mathbf{x}}_j' \mathbf{w} \geq 1 - \xi_j, \;\; \xi \geq 0, \quad j = 1, \ldots, J.$$

Because preference data can be contradictory (preferences change over time due to external factors, and unmodeled dynamics), it is often times impossible to find $\hat{\mathbf{w}}$ such that all inequalities $\tilde{\mathbf{x}}_j' \hat{\mathbf{w}} \geq 0$ are satisfied. It is thus prudent to allow for some 'slack', and try to minimize the inconsistencies ξ_j in the LS sense. This is exactly what (7) achieves in the absence of outliers. When outliers are present though, nonzero estimates $\hat{o}_j$ will ideally take values $\hat{o}_j \approx \hat{\xi}_j$, thus effectively removing the effect of the invalid responses in the estimation process. Note that 1 in the right-hand side of the first set of inequality constraints accounts for classifier margin; any other positive constant is equally good.

Problem (7) is a linearly-constrained quadratic program (QP), and is efficiently solved using general-purpose convex optimization software. In particular, it can be solved in the primal domain (advisable when p is small but J is large), or, in the dual domain (preferable when p is large and J is small). A result with ramifications to the robustness properties and computational advantages of (6), is asserted in the following proposition [13].

Proposition 2: *The robust CA classifier* (6) *is equivalent to*

$$\min_{\mathbf{w}} \sum_{j=1}^{J} h\left(\tilde{\mathbf{x}}_j' \mathbf{w}\right) + \lambda_w \|\mathbf{w}\|_2^2 \qquad (8)$$

where $h : \mathbb{R} \to \mathbb{R}$ *is the 'Huberized' square hinge loss function [15]*

$$h(z) := \begin{cases} \lambda_o(1-z) - \lambda_o^2/4, & z < 1 - \lambda_o/2, \\ (1-z)_+^2, & \textit{otherwise} \end{cases} . \qquad (9)$$

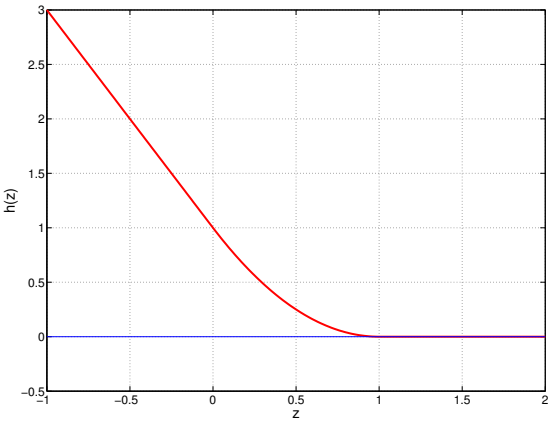

Figure 1: Huberized square hinge loss function for $\lambda_o = 2$.

Problem (8) is obtained after eliminating from (6), the optimized outlier variables $\hat{\mathbf{o}}(\mathbf{w})$. Examination of (9) (see also Fig. 1) reveals that (6) gives rise to three classification regions: r1) containing 'consistent' data for which $\tilde{\mathbf{x}}_j' \mathbf{w} \geq 1$; r2) comprising support vectors for which $1 - \lambda_o/2 \leq \tilde{\mathbf{x}}_j' \mathbf{w} \leq 1$; and r3) over which data satisfy $-\infty < \tilde{\mathbf{x}}_j' \mathbf{w} \leq 1 - \lambda_o/2$, and are deemed as contaminated with outliers. For $\lambda_0 = \infty$, $\hat{\mathbf{o}} = \mathbf{0}$ and h becomes the squared hinge loss function used in SVM variants.

When compared to the SVM used for CA [7, 8, 20], the key advantage of the classifier obtained via (6) is its ability to attain desirable tradeoffs between model fit and complexity, while at the same time controlling robustness and revealing the outliers present. Furthermore, convexity of the cost in (6) is not affected even when one chooses a different regularizer such as, e.g., $\lambda_w \|\mathbf{w}\|_1$ to encourage sparse partworth vectors and effect model complexity control. In fact, this could also be a wise choice from a computational standpoint, since the ℓ_1-norm regularized counterpart of (8) attains piecewise-linear solution paths as λ_w varies [15]. By capitalizing on this property, [15] shows that the entire path of solutions is efficiently obtained, using an algorithm that generalizes the LARS solver developed for Lasso [6].

4.2 Nonparametric utility function estimation

The linear utility function $u(\mathbf{x}) = \mathbf{x}'\mathbf{w}$ considered so far falls short in capturing *interdependencies* among the attributes of each profile (entries of vector $\mathbf{x}_j$) – customers preferring cell-phones with mp3 players, will also value highly those models with memory capacity above 4Gb, say. As these interdependencies are driven by complex mechanisms that are typically hard to model a priori, it is prudent to let the data dictate the form of the $u(\mathbf{x})$ sought. This motivates the *nonparametric regression* methods for PM modeling outlined in this section.

To ensure versatility, u is only assumed to belong to a (possibly infinite dimensional) space of e.g., "smooth" functions $\mathcal{H}$ [22]. As estimating $u \in \mathcal{H}$ from finite data is inherently ill-posed, one typically invokes properly regularized criteria [19]. Accordingly, u is robustly estimated from data adhering to (M1) by solving

$$(\hat{u}, \hat{\mathbf{o}}) := \arg \min_{u \in \mathcal{H}, \mathbf{o}} \sum_{j=1}^{J} (y_j - u(\mathbf{x}_j) - o_j)^2 + \mu R(u) + \lambda_o \|\mathbf{o}\|_1 \qquad (10)$$

where $R : \mathcal{H} \to \mathbb{R}$ is a convex smoothing regularization functional, and $\mu \geq 0$ is chosen to tradeoff fidelity to the (outlier compensated) data for the degree of smoothness measured by $R(u)$. Problem (10) is variational in nature, and in principle requires searching over the infinite-dimensional space $\mathcal{H}$.

There is a neat workaround however, if one lets $R(u) := \|u\|_{\mathcal{H}}^2$ in (10), and endows $\mathcal{H}$ with the structure of a reproducing ker-

nel Hilbert space [22]; with corresponding positive definite reproducing kernel function $K(\cdot,\cdot) : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$. The following proposition asserts that in this case, the unique solution of (10) is finitely parametrized, and it suffices to solve a single instance of Lasso to determine $\hat{u}$ along with the outliers $\hat{\mathbf{o}}$. Before stating the result, recall the conjoint data model (M1), the definition $\mathbf{y} := [y_1, \ldots, y_J]'$, and introduce the kernel matrix $\mathbf{K} \in \mathbb{R}^{J\times J}$ with ij-th entry $[\mathbf{K}]_{ij} := K(\mathbf{x}_i, \mathbf{x}_j)$. The proof relies on the Representer Theorem; see e.g., [22], and can be found in [13].

Proposition 3: *Consider* $\hat{\mathbf{o}}_{Lasso}$ *defined as*

$$\hat{\mathbf{o}}_{Lasso} := \arg\min_{\mathbf{o}} \|\mathbf{X}_\mu \mathbf{y} - \mathbf{X}_\mu \mathbf{o}\|_2^2 + \lambda_o \|\mathbf{o}\|_1 \tag{11}$$

where

$$\mathbf{X}_\mu := \begin{bmatrix} \mathbf{I}_J - \mathbf{K}(\mathbf{K} + \mu \mathbf{I}_J)^{-1} \\ (\mu \mathbf{K})^{1/2} (\mathbf{K} + \mu \mathbf{I}_J)^{-1} \end{bmatrix}. \tag{12}$$

Then the minimizers $\{\hat{u}, \hat{\mathbf{o}}\}$ *of* (10) *with* $R(u) := \|u\|_{\mathcal{H}}^2$ *are fully determined given* $\hat{\mathbf{o}}_{Lasso}$*, as* $\hat{\mathbf{o}} := \hat{\mathbf{o}}_{Lasso}$ *and* $\hat{u}(\mathbf{x}) = \sum_{j=1}^J \hat{\beta}_j K(\mathbf{x}, \mathbf{x}_j)$*, with* $\hat{\boldsymbol{\beta}} = (\mathbf{K} + \mu \mathbf{I}_J)^{-1}(\mathbf{y} - \hat{\mathbf{o}}_{Lasso})$.

Joint outlier sparsity and function complexity control mechanisms identify the best (μ^*, λ_o^*) in (11), trading-off optimally the number of outliers rejected and the predictive capability of $\hat{u}$. These methods extend naturally those outlined in [cf. the similarity between (11) and (4)], and require searching over a collection of robustifaction paths – one per μ value in a prescribed μ-grid. The end result yields estimates $\hat{u}$ with enhanced *ecological rationality*, yielding preference models better adapted to the shopping environment in which customers operate.

4.3 Distributed conjoint analysis

So far a single $\mathbf{w}$ was estimated, but multiple $\{\mathbf{w}_i\}$s are often needed to capture consumer heterogeneity, while improving estimation performance by fusing data from multiple respondents [20, 8, 11]. Traditional approaches have relied on hierarchical Bayes (HB) [1], and share with convex optimization based ones [8] the idea of shrinking the individual estimates $\{\hat{\mathbf{w}}_i\}_{i=1}^I$ towards the population mean $\bar{\mathbf{w}}$. Specifically for (M1), [8] suggests

$$\min_{\{\mathbf{w}_i, \mathbf{D}, \bar{\mathbf{w}}\}} \sum_{i=1}^I \sum_{j=1}^J (y_{ij} - \mathbf{x}_{ij}' \mathbf{w}_i)^2 + \gamma \sum_{i=1}^I \|\mathbf{w}_i - \bar{\mathbf{w}}\|_{\mathbf{D}}^2 \tag{13}$$

which is jointly convex in $\{\mathbf{w}_i, \mathbf{D}, \bar{\mathbf{w}}\}$, while the positive definite (PD) $\mathbf{D}$ is normalized to have $\mathrm{tr}(\mathbf{D}) = 1$; and $\|\mathbf{v}\|_{\mathbf{M}}^2 = \mathbf{v}'\mathbf{M}^{-1}\mathbf{v}$. Matrix $\mathbf{D}$ is related to the covariance matrix of the partworth estimators, so that pronounced shrinkage is effected to those $\mathbf{w}_i$'s far away from the mean $\bar{\mathbf{w}}$. MAP optimality is also apparent under a Gaussian nominal noise assumption, and identical Gaussian priors on the $\mathbf{w}_i$; see [8] for a detailed comparison between (13) and HB in [1]. Extension to choice-based data (M3) is possible by replacing the ℓ_2-error loss in (13) with e.g., the logistic error [8].

All existing works assume that the data $\{y_{ij}, \mathbf{x}_{ij}\}_{i,j=1}^{I,J}$ are available centrally to determine the estimates $\{\hat{\mathbf{w}}_i, \hat{\mathbf{D}}, \hat{\bar{\mathbf{w}}}\}$. However, collecting all data in a central location may be prohibitive in certain studies, simply because respondents are not collocated, or due to finite storage, limited complexity, or even privacy constraints. In CA-based healthcare studies carried out by pharmaceutical companies, physicians provide private patient information for the purpose of estimating partworth vectors. They may not be willing to share training data but only the learning results $\hat{\mathbf{w}}_i$. These reasons motivate well the *distributed* partworth estimator developed in this section, which is implementable through a cooperating network of processing units (agents) $\mathcal{I} := \{1, \ldots, I\}$, that exchange messages with directly connected neighbors. In the sequel, the network of agents will be modeled as a connected graph, and $\mathcal{N}_i \subseteq \mathcal{I}$ will denote the set of neighbors of agent i.

Algorithm 1 : DRCA

Agents $i \in \mathcal{I}$ initialize $\{\mathbf{w}_i(0), \bar{\mathbf{w}}_i(0), \mathbf{p}_i(0)), \mathbf{P}_i(0)\}$ to zero, $\{\mathbf{D}_i(0)\}$ to random unit-trace PD matrices , and locally run
for $k = 0, 1, \ldots$ **do**
  Exchange $\{\bar{\mathbf{w}}_i(k), \mathbf{D}_i(k)\}$ with neighbors in $\mathcal{N}_i$.
  Update $\{\mathbf{w}_i(k+1), \bar{\mathbf{w}}_i(k+1)\}$ using (15).
  Update $\mathbf{D}_i(k+1)$ using (16).
  $o_{ij}(k+1) = \mathcal{S}\left(y_{ij} - \mathbf{x}_{ij}' \mathbf{w}_i(k+1), \lambda_o/2\right), \quad j = 1, \ldots, J.$
  $\mathbf{p}_i(k+1) = \mathbf{p}_i(k) + c \sum_{i' \in \mathcal{N}_i} [\bar{\mathbf{w}}_i(k+1) - \bar{\mathbf{w}}_{i'}(k+1)].$
  $\mathbf{P}_i(k+1) = \mathbf{P}_i(k) + c \sum_{i' \in \mathcal{N}_i} [\mathbf{D}_i(k+1) - \mathbf{D}_{i'}(k+1)].$
end for

Towards distributing the centralized problem (13), introduce *local* auxiliary copies $\{\mathbf{D}_i, \bar{\mathbf{w}}_i\}_{i=1}^I$ of the *global* variables $\{\mathbf{D}, \bar{\mathbf{w}}\}$ per agent, along with constraints $\bar{\mathbf{w}}_i = \bar{\mathbf{w}}_{i'}$, $\mathbf{D}_i = \mathbf{D}_{i'}$, $i \in \mathcal{I}$, $i' \in \mathcal{N}_i$ to ensure consensus of these variables per neighborhood. Introducing the local quantities $\mathbf{y}_i := [y_{i1}, \ldots, y_{iJ}]'$, $\mathbf{X}_i := [\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iJ}]'$, and likewise for $\mathbf{o}_i$; the proposed approach to *distributed* and *robust* (DR) CA solves

$$\min_{\substack{\{\mathbf{w}_i, \bar{\mathbf{w}}_i, \\ \mathbf{D}_i, \mathbf{o}_i\}}} \sum_{i=1}^I \left[\|\mathbf{y}_i - \mathbf{X}_i \mathbf{w}_i - \mathbf{o}_i\|_2^2 + \lambda_o \|\mathbf{o}_i\|_1 + \gamma \|\mathbf{w}_i - \bar{\mathbf{w}}_i\|_{\mathbf{D}_i}^2 \right]$$
$$\text{s. t.} \qquad \bar{\mathbf{w}}_i = \bar{\mathbf{w}}_{i'}, \, \mathbf{D}_i = \mathbf{D}_{i'}, \, i \in \mathcal{I}, \, i' \in \mathcal{N}_i \tag{14}$$

with constraints $\mathrm{tr}(\mathbf{D}_i) = 1$, $i \in \mathcal{I}$, left implicit. Leaving aside robustness (cf. $\lambda_o = \infty$), problems (14) and (13) are equivalent since the network is connected. This property is instrumental because it ensures that the optimal local estimates coincide with the *global* minimizer of (13). Interestingly, the structure of (14) lends itself naturally to distributed implementation via the alternating-direction method of multipliers (AD-MoM), an iterative augmented Lagrangian method especially well-suited for parallel processing [3, 12]. AD-MoM iterations for $k = 0, 1, 2, \ldots$ entail: i) local optimization tasks to be run per agent; and ii) exchanges of local estimates $\{\bar{\mathbf{w}}_i(k), \mathbf{D}_i(k)\}$ only within $\mathcal{N}_i$, $i \in \mathcal{I}$. The latter are critical to percolate the spatially distributed data in $\mathcal{T}$ throughout the network, thus enabling agents to attain consensus on $\{\hat{\mathbf{w}}, \hat{\mathbf{D}}\}$ – the optimal solution of the centralized problem (13).

A detailed derivation of the DRCA algorithm (tabulated under Algorithm 1) can be found in [13]; see also [12]. At the beginning of iteration $k+1$, agent i collects its neighbors most up to date estimates $\{\bar{\mathbf{w}}_{i'}(k), \mathbf{D}_{i'}(k)\}_{i' \in \mathcal{N}_i}$, and updates its own ones by solving the following strictly convex optimization problems

$$\{\mathbf{w}_i(k+1), \bar{\mathbf{w}}_i(k+1)\} = \arg\min_{\{\mathbf{w}, \bar{\mathbf{w}}\}} \Big[\|\mathbf{y}_i - \mathbf{X}_i \mathbf{w} - \mathbf{o}_i(k)\|_2^2$$
$$+ \gamma \|\mathbf{w} - \bar{\mathbf{w}}\|_{\mathbf{D}_i(k)}^2 + \mathbf{p}_i'(k) \bar{\mathbf{w}} + c \sum_{i' \in \mathcal{N}_i} \left\| \bar{\mathbf{w}} - \frac{\bar{\mathbf{w}}_i(k) + \bar{\mathbf{w}}_{i'}(k)}{2} \right\|_2^2 \Big] \tag{15}$$

$$\mathbf{D}_i(k+1) = \arg\min_{\mathbf{D}} \Big[\gamma \|\mathbf{w}_i(k+1) - \bar{\mathbf{w}}_i(k+1)\|_{\mathbf{D}}^2$$
$$+ \mathrm{tr}(\mathbf{P}_i(k) \mathbf{D}) + c \sum_{i' \in \mathcal{N}_i} \left\| \mathbf{D} - \frac{\mathbf{D}_i(k) + \mathbf{D}_{i'}(k)}{2} \right\|_F^2 \Big]. \tag{16}$$

While (15) is an unconstrained QP with solution given in closed form, solving (16) requires an extra iterative procedure. Outliers are updated by parallel soft-thresholding of local residuals, where $\mathcal{S}(z, u) := \mathrm{sign}(z)(|z| - u)_+$ in Algorithm 1. Iteration $k+1$ is concluded after obtaining dual prices $\mathbf{p}(k+1)$ and $\mathbf{P}(k+1)$ through dual ascent updates (see Algorithm 1), where $c > 0$ is a stepsize which affects the convergence rate of the DRCA algorithm.

Table 1: Average partworth estimation errors

Response error	Questions	SVM [7]	Proposed (6)
Low	8	0.3791	0.3730
Low	16	0.2472	0.2445
High	8	0.4023	0.3901
High	16	0.2922	0.2831

To close this section, it is useful to mention that convergence of the DRCA algorithm to the minimizer of (13) is ensured – for any $c > 0$ – by virtue of AD-MoM's convergence theory [3, Prop. 4.2].

5. PRELIMINARY NUMERICAL TESTS

A simulated test is carried out here to corroborate the effectiveness of the proposed sparsity-controlling estimator for choice-based CA (cf. Section 4.1), and compare it with the SVM approach in [7]. Comprehensive numerical tests with both synthetic and real CA data can be found in [13].

The adopted simulation setup is standard for choice-based CA simulation studies under different (low-high) response-error levels, and (low-high) number of questions; see e.g. [7, 8]. Stated-preference questionnaires are simulated with $p = 10$ binary attributes per product profile, while the $\mathbf{x}_j^{(1)}$ were generated according to an orthogonal fractional-factorial design with $J = 16$. As per (M3), each of the questions comprises a pair of profiles to choose from, and given $\mathbf{x}_j^{(1)}$, the $\mathbf{x}_j^{(2)}$ were obtained through the shifting method of [4]. In the high number of questions setting, all $J = 16$ profiles pairs were presented to each respondent. For the reduced-size questionnaire condition, 8 profile pairs where randomly drawn from the complete set of 16. Each of the $I = 50$ respondents in a homogeneous population were given the same questionnaire, and 'true' partworths were drawn from a Gaussian distribution, i.e., $\mathbf{w}_i \sim \mathcal{N}(\mu \mathbf{1}_p, \sigma_{w_i}^2 \mathbf{I}_p)$, where $\mathbf{1}_p$ is the $p \times 1$ vector of all ones. The mean parameter μ takes the values 1.2 and 0.2, respectively in the low and high response error conditions. Since consumer heterogeneity is not considered here, values $\sigma_{w_i}^2 = \mu$ are adopted for $i = 1, \ldots, I$. Finally, logistic probabilities were used for the simulated nominal responses y_{ij}, i.e.,

$$\Pr(y_{ij} = 1) = \frac{\exp\left(\mathbf{w}_i' \mathbf{x}_j^{(1)}\right)}{\exp\left(\mathbf{w}_i' \mathbf{x}_j^{(1)}\right) + \exp\left(\mathbf{w}_i' \mathbf{x}_j^{(2)}\right)},$$

whereas outliers were generated by simulating y_{i3}, $i = 1, \ldots, I$, as the outcome of an unbiased coin toss.

The results are summarized under Table 1, the figure of merit being the average partworth estimation error across respondents $\sum_{i=1}^{I} \|\hat{\mathbf{w}}_i - \mathbf{w}_i\|_2 / I$, after normalizing partworths to have unit ℓ_1 norm. Results for the method of [7] are shown under the column labeled SVM. Interestingly, the proposed sparsity-controlling estimator (6) consistently outperforms the SVM alternative of [7]. Regardless of the number of questions, the performance edge is more significant under the high response error condition. This is a manifestation of the robustness properties of the novel estimator, not only against outliers but also against noisy (erroneous) responses. For all practical purposes, both schemes attain comparable estimation errors under the low response error regime.

6. CONCLUDING SUMMARY

Outlier-robust conjoint estimation methods were developed in this paper for both metric and choice-based conjoint data. Building on a neat link between the seemingly unrelated fields of robust statistics and sparse regression and classification, the novel estimators were found rooted at the crossroads of outlier-resilient estimation, statistical learning via the Lasso, and convex optimization.

REFERENCES

[1] G. M. Allenby and P. E. Rossi, "Marketing models of consumer heterogeneity," *J. Econometrics,* vol. 89, pp. 57-58, 1999.

[2] M. Ben-Akiva and S. R. Lerman, *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press, 1985.

[3] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Athena Scientific, 1999.

[4] D. S. Bunch, J. J. Louviere, and D. Anderson, "A comparison of experimental design strategies for multinomial logit models: The case of generic attributes," Working paper, Graduate School of Management University of California at Davis, 1994.

[5] D. Cui and D. Curry, "Prediction in marketing using the support vector machines," *Marketing Science,* vol. 24, no. 4 pp. 595-615, 2005.

[6] B. Efron, T. Hastie, I. M. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.,* vol. 32, pp. 407-499, 2004.

[7] T. Evgeniou, C. Boussios, and G. Zacharia, "Generalized robust conjoint analysis," *Marketing Science,* vol. 24, no. 3 pp. 415-429, 2005.

[8] T. Evgeniou, M. Pontil, and O. Toubia, "A convex optimization approach to modeling consumer heterogeneity in conjoint analysis," *Marketing Science,* vol. 26, no. 6, pp. 805-818, 2007.

[9] G. B. Giannakis, G. Mateos, S. Farahmand, V. Kekatos, and H. Zhu, "USPACOR: Universal sparsity-controlling outlier rejection," in *Proc. of Intl. Conf. on Acoust., Speech, and Signal Proc.*, Prague, Czech Republic, May 22-27, 2011.

[10] P. E. Green and V. R. Rao, "Conjoint measurement for quantifying judgmental data," *J. of Marketing Research,* vol. 8, pp. 355-363, 1971.

[11] J. R. Hauser and V. R. Rao, "Conjoint analysis, related modeling, and applications," *in Marketing Research and Modeling: Progress and Prospects,* Y. Wind and P. E. Green, Eds. New York, NY: Springer, 2005, pp. 141-168.

[12] G. Mateos, J. A. Bazerque, and G. B. Giannakis, "Distributed sparse linear regression," *IEEE Trans. Sig. Proc.*, vol. 58, no. 10, pp. 5262-5276, 2010.

[13] G. Mateos, V. Kekatos, and G. B. Giannakis, "Robust conjoint analysis by controlling outlier sparsity," 2011 (submitted).

[14] O. Netzer et al., "Beyond conjoint analysis: Advances in preference measurement," *Marketing Letters*, vol. 19, pp. 337-354, 2008.

[15] S. Rosset and J. Zhu, "Piecewise linear regularized solution paths," *Ann. Statist.,* vol. 35, pp. 1012-1030, 2007.

[16] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection.* New York, NY: Wiley, 1987.

[17] P. J. Rousseeuw and K. Van Driessen, "Computing LTS regression for large data sets," *Data Mining and Knowledge Discovery,* vol. 12, no. 1, pp. 29-45, 2006.

[18] V. Srinivasan and A. D. Shockern, "Linear programming techniques for multidimensional analysis of preferences," *Psychometrica,* vol. 38, no. 3, pp. 337-369, 1973.

[19] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-posed Problems.* Washington, DC: W. H. Winston, 1977.

[20] O. Toubia, T. Evgeniou, and J. Hauser "Optimization-based and machine-learning methods for conjoint analysis: Estimation and question design," *in Conjoint Measurement: Methods and Applications,* A. Gustafsson, A. Herrmann, and F. Huber, Eds. New York, NY: Springer, 2007, pp. 231-258.

[21] J. Tropp, "Just relax: Convex programming methods for identifying sparse signals," *IEEE Trans. Info. Theory*, vol. 52, no. 3, pp. 1030-1051, 2006.

[22] G. Wahba, *Spline Models for Observational Data.* Philadelphia, PA: SIAM, 1990.