Sparse Conjoint Analysis Through Maximum Likelihood Estimation

Efthymios Tsakonas, Joakim Jaldén, *Senior Member, IEEE*, Nicholas D. Sidiropoulos, *Fellow, IEEE*, and Bjorn Ottersten, *Fellow, IEEE*

***Abstract*—Conjoint analysis (CA) is a classical tool used in preference assessment, where the objective is to estimate the utility function of an individual, or a group of individuals, based on expressed preference data. An example is choice-based CA for consumer profiling, i.e., unveiling consumer utility functions based solely on choices between products. A statistical model for choice-based CA is investigated in this paper. Unlike recent classification-based approaches, a sparsity-aware Gaussian maximum likelihood (ML) formulation is proposed to estimate the model parameters. Drawing from related robust parsimonious modeling approaches, the model uses sparsity constraints to account for outliers and to detect the salient features that influence decisions. Contributions include conditions for statistical identifiability, derivation of the pertinent Cramér-Rao Lower Bound (CRLB), and ML consistency conditions for the proposed sparse *nonlinear* model. The proposed ML approach lends itself naturally to ℓ_1-type convex relaxations which are well-suited for distributed implementation, based on the alternating direction method of multipliers (ADMM). A particular decomposition is advocated which bypasses the apparent need for outlier communication, thus maintaining scalability. The performance of the proposed ML approach is demonstrated by comparing against the associated CRLB and prior state-of-the-art using both synthetic and real data sets.**

***Index Terms*—Conjoint analysis, maximum likelihood, estimation, sparse, CRLB, ADMM.**

I. Introduction

THE remarkable growth of the world-wide web has enabled large-scale (semi-)automated collection of *preference data*—that is, data containing information about people's preferences regarding products, services, other people, events, etc. Large-scale preference data collection is mainly driven by services such as online retailing, social networking, and personalized recommendation systems. The rapidly growing volume and diversity of preference data (purchases, choices, rankings, surveys, questionnaires) along with the need for accurate classification, personalization, and prediction, have spurred an increasing interest in preference modeling and analysis (PMA), a cross-disciplinary applied research area with a long history (e.g., early work in PMA includes [1], [2]). The goal of PMA, as suggested by the name, is to predict responses of individuals to products or services, based on already expressed preference data.

Conjoint analysis is a statistical technique commonly used in PMA, to determine how individuals value different features that make up a product or a service. CA is used in many social and applied sciences, including marketing, industrial design and economics. The modeling assumption behind CA is that responses are formed as noisy *linear combinations* of a product's features with weights given by the decision-maker's *partworths* [4]. By analyzing available preference data, CA techniques aim to estimate the underlying partworth values. These estimates can be used to predict future preferences and assess the profitability of new designs, but are also useful *per se* to the retailer/marketer, e.g., for consumer sensitivity analysis.

Traditional methods for partworth estimation for *choice-based* CA models (where preferences are only expressed in the form of binary choices), range from logistic regression [5] and hierarchical Bayes (HB) methods [6], [7], to methods based on support vector machine (SVM) classifiers [8]. Following either deterministic or Bayesian formulations, these state-of-the-art techniques rely on suitably regularized loss functions to "optimally" trade-off model fit for better generalization capability of the solution beyond the training data. See [9] for a compact description of these approaches; more detailed comparisons can also be found in [8].

Although the benefits of CA have been widely appreciated (see, e.g., [10]), the tacit assumption underlying most of the existing techniques is that the data is gathered under controlled conditions, i.e., there are no outliers, and responses regress upon a modest number of features. However, in modern preference modeling systems, especially in web-based collection, such controlled conditions are often not feasible. Therefore, solutions that are computationally efficient and offer robustness to gross errors are, if not necessary, at least highly desirable. In this direction, it has been noted in [8] that classification approaches to choice-based CA using SVMs are typically more robust against outliers, than HB methods, for example. A classification approach is sensible for a number of reasons, mainly because it avoids strong probabilistic assumptions. However,

Manuscript received October 31, 2012; revised April 23, 2013; accepted June 21, 2013. Date of publication August 15, 2013; date of current version October 16, 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jean Pierre Delmas. This work was supported by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement number 228044. The work of N. Sidiropoulos was supported by DTC/DTI, University of Minnesota. A portion of this paper was presented at the IEEE Statistical Signal Processing Workshop (SSP), Ann Arbor, MI, August 5–8, 2012 [3].

E. Tsakonas and J. Jaldén are with the ACCESS Linnaeus Centre, Royal Institute of Technology (KTH), Stockholm, Sweden.

N. D. Sidiropoulos is with the University of Minnesota, Minneapolis (UMN).

B. Ottersten is with the Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, Luxembourg. He is also with the ACCESS Linnaeus Centre, Royal Institute of Technology (KTH), Stockholm, Sweden.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2013.2278529

although SVMs perform very well in practice, the quality of the solution is difficult to quantify; for example, it is often difficult to benchmark classification performance. An outlier-aware SVM classifier for choice-based conjoint data is proposed in [8]. The SVM proposed in [8] solves an unconstrained optimization problem consisting of a convex, non-differentiable loss function combined with a suitable regularizing function whose addition aims to improve the generalization error of the classifier. Similar to [8], the authors in [11] follow an SVM approach to choice-based CA, the main difference being that sparse outliers are modeled explicitly using auxiliary variables.

Contributions: Unlike an SVM approach, we consider a statistical choice-based CA model which includes both standard errors and auxiliary variables that explicitly model sparse outliers. Our particular model was first proposed in [3]; here, we revisit the formulation in [3] and further investigate its properties, in an attempt to provide a more solid and well-rounded framework for partworth estimation. Links between sparsity and robustness against outliers exploiting connections with the ℓ_1-norm were drawn *in the linear regression context* in [12], [13], and more recently in [14] it was proposed to introduce a sparse auxiliary vector variable to account for outliers, as a *universal model robustification* strategy. The latter ideas are explored in our paper in the particular context of choice-based CA, where—unlike the focus of [11], [14]—the signal of interest (partworths) is also sparse, and an ML formulation is proposed for partworth estimation. A key contribution of our work is that we provide identifiability conditions and explore the best achievable mean-square-error (MSE) performance by deriving the CRLB under sparsity constraints, building on earlier work on the CRLB computation in constrained parameter estimation [16]–[19]. Our identifiability and CRLB results allow one to assess the performance of relevant relaxation strategies for our model. As a second step, we revisit the ML formulation we proposed in [3] and show that consistency holds for the partworths, under suitable conditions on the outliers. We show that the proposed ML formulation lends itself naturally to an ℓ_1-type relaxation (see, e.g., [15]) which is not only convex, but also naturally amenable to *distributed* implementation. Distributed solution strategies are interesting for two reasons: First, applications of interest usually involve large-scale datasets which may go beyond the reach of standard off-the-shelf solvers. Second, the proposed solutions are not only distributed, but also *decentralized*, meaning that the nodes in the distributed implementation need not share their private datasets to reach consensus to optimality. We derive a simple decentralized algorithm based on the alternating direction method of multipliers (ADMM), a method which has shown great potential in the area of distributed optimization [21]. An ADMM solution was pursued in [11] in the context of a *linear* CA model, whose convergence proof was deferred to another manuscript. Unlike [11], in this paper we focus on distributing *choice-based* CA and show how to directly embed our ML formulation into the consensus optimization framework of [21]. Finally, the efficacy of the proposed sparsity—aware ML estimator is assessed by comparing its MSE performance vis-a-vis the CRLB and the prior state-of-the-art using both simulated and real data from a conjoint choice experiment for coffee makers.

The rest of the paper is organized as follows. Section II describes the main problem setup, and the associated ML estimator that accounts for outliers and partworth sparsity. In Section III-A we give model identification conditions and derive the best achievable MSE performance for the estimation of partworths in the simple case where no outliers are present. The model identification and CRLB results are extended to the general outlier—contaminated case in Section III-B. The asymptotic properties of the proposed ML approach are discussed in Section IV. Section V describes a tractable convex relaxation of the ML estimator that is convenient for use in practice. Section VI gives a distributed implementation of the proposed relaxed ML estimator based on ADMM. Results of experiments are presented in Section VII, and conclusions are drawn in Section VIII, along with some discussion on future work.

II. Sparse CA Modeling & ML Estimation

We begin by describing the three basic CA models used in PMA. These are included in [3], but are also discussed here for completeness. The starting point is to represent the quantities over which preferences are expressed (and let us assume that these quantities are products, for simplicity) using associated profiles, i.e., p-dimensional vectors whose elements correspond to the different features. A profile captures all relevant information about the corresponding product's characteristics. Suppose there are J such profiles $\{\mathbf{p}_j\}_{j=1}^J$, to be evaluated by a single individual.[1] In CA it is customary to assume that responses $\{r_i\}_{i=1}^J$ obey a linear regression model (see, e.g., [4])

$$r_i = \mathbf{p}_i^{\mathrm{T}}\mathbf{w} + e_i, \tag{1}$$

where $(\,\cdot\,)^{\mathrm{T}}$ denotes transposition, $\mathbf{w}$ is the vector of partworths associated with the individual and e_i is a random variable modeling (usually small) random errors.

There are three different but related categories of models that link responses to preference measurements. In a *full-profile rating* model, the measurement is assumed to be directly the response r_i. Another category consists of the so-called *metric-paired rating* models, where the $\mathbf{p}_i$ in (1) is replaced by a difference $\mathbf{d}_i \triangleq \mathbf{p}_i^{(1)} - \mathbf{p}_i^{(2)}$ of a pair of profiles. Finally, we have also *choice-based* models, where in addition to using pairwise-differences of profiles in (1), the measurement is only the *sign* of r_i. In other words, in a choice-based CA model the individual is each time asked to indicate a preference between two profiles, but not the actual magnitude of this preference. Mathematically speaking, if we assume N given profile differences $\{\mathbf{d}_i\}_{i=1}^N$ the classical choice-based CA model is

$$y_i = \operatorname{sign}\left(\mathbf{d}_i^{\mathrm{T}}\mathbf{w} + e_i\right), \quad i = 1, \ldots, N. \tag{2}$$

Given J profiles, there are at most $J(J-1)/2$ unique profile differences, equating $\mathbf{d}_i$ and $-\mathbf{d}_i$, but N is typically selected smaller than this, reflecting that a subset of all possible questions are actually used in a survey.

[1]The term *individual* here can also be interpreted as a *homogeneous population* of individuals. Similar to [11] we focus on this homogeneous case for simplicity: Once this case is addressed, approaches to account for heterogeneous populations are possible along the lines of [22], [23].

There are several advantages of choice-based CA models as compared to models based on rating scales. One intuitive advantage is that choices are more realistic, resembling the real purchasing situation. Another advantage is that the problem of individual differences in interpreting rating scales is circumvented [24]. In this paper we deal exclusively with (2), and aim to robustify this model by utilizing structural information. Towards this end, we make two observations: The first is that responses can be grossly inconsistent due to a number of reasons, implying that it is more realistic to acknowledge that there can be gross errors in the measurement model in (2), in addition to the typically small errors $\{e_i\}_{i=1}^N$. We model the errors $\{e_i\}_{i=1}^N$ as i.i.d. normal variables $\mathcal{N}(0,\sigma^2)$ with known variance σ^2. On the other hand, the only assumption regarding the gross errors is *sparsity*, i.e., that there is a known upper bound on their number. Assuming that gross errors are sparse makes sense in this context, since intuitively, an individual will not be regularly inconsistent. The second observation which aims to robustify (2), is that the number of features p can be very large—modern products may have a very large number of relevant attributes and technical specifications—yet relatively few features will matter to any given individual, and even the 'typical' individual. Therefore, it makes sense to assume that the unknown partworth vector $\mathbf{w}$ will also be sparse, and this structural information can be exploited to facilitate the estimation.

In light of the above, the model in (2) can be re-stated by (a) explicitly modeling the gross errors using a sparse vector $\mathbf{o}\in\mathbb{R}^{N\times 1}$ of deterministic *outliers* $\{o_i\}_{i=1}^N$ and (b) utilizing the (deterministic) prior information that $\mathbf{w}$ itself is a sparse vector. Therefore, a conceptually appealing version of (2) is

$$y_i = \operatorname{sign}\left(\mathbf{d}_i^{\mathrm{T}}\mathbf{w} + e_i + o_i\right), \quad i=1,\ldots,N \tag{3}$$

coupled with the a-priori knowledge that $\operatorname{card}(\mathbf{w}) \le \kappa_w$ and $\operatorname{card}(\mathbf{o}) \le \kappa_o$. Here, the integers κ_w, κ_o are assumed fixed and given, and the function $\operatorname{card}(\cdot)$ stands for cardinality, i.e., it returns the number of non-zero elements of a vector. Unless explicitly stated otherwise, throughout the paper we focus on the case where one has at least as many measurements as partworths ($p \le N$). We are interested in that case because it provides intuition useful when describing the asymptotic properties of the model; however, the case where $p > N$ is explicitly discussed in some places as well. Finally, it is also assumed that all unknown parameters are *bounded*, i.e., that there exist positive constants R_w and R_o such that $\mathbf{w}\in\mathcal{B}_w \triangleq \{\mathbf{w}\in\mathbb{R}^p \,|\, \|\mathbf{w}\|_\infty \le R_w\}$ and $\mathbf{o}\in\mathcal{B}_o \triangleq \{\mathbf{o}\in\mathbb{R}^N \,|\, \|\mathbf{o}\|_\infty \le R_o\}$. This requirement is mostly technical, and its use will become evident later on, in our analysis.

Given N measurements from (3), we are interested in *estimating* the vector of partworths $\mathbf{w}$ as well as *detecting* the responses that have been contaminated with outliers. This joint estimation problem is challenging because the model in (3) is underdetermined; there are always more unknowns than measurements, so one expects that sparsity is the key to make the problem meaningful. Efficient outlier detection is critical for accurate partworth estimation in this context (and therefore accurate preference prediction), but can also have useful implications in the experimental design of the profile differences $\{\mathbf{d}_i\}_{i=1}^N$.

The ML estimator for the vector $(\mathbf{w},\mathbf{o})$ is derived as follows. Let $\mathcal{I}_+$ be the set of indices $\{i|y_i=1\}$, and similarly define $\mathcal{I}_- = \{i|y_i=-1\}$. Since noise samples e_i are independent, the probability of a random partition of the observations to $\mathcal{I}_+$ and $\mathcal{I}_-$ can be calculated explicitly to be

$$\begin{aligned} p_y(\mathbf{w},\mathbf{o}) &= \prod_{i\in\mathcal{I}_+} \Pr\left[\mathbf{d}_i^{\mathrm{T}}\mathbf{w} + o_i \ge -e_i\right] \prod_{i\in\mathcal{I}_-} \Pr\left[\mathbf{d}_i^{\mathrm{T}}\mathbf{w} + o_i \le -e_i\right] \\ &= \prod_{i\in\mathcal{I}_+} \Phi\left(\frac{\mathbf{d}_i^{\mathrm{T}}\mathbf{w}+o_i}{\sigma}\right) \prod_{i\in\mathcal{I}_-} \Phi\left(-\frac{\mathbf{d}_i^{\mathrm{T}}\mathbf{w}+o_i}{\sigma}\right), \end{aligned}$$

where $\Phi(u) := \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u} e^{-t^2/2}dt$ is the cumulative distribution function of the standardized Gaussian density. The log-likelihood function can be written compactly as

$$l_{wo}(\mathbf{w},\mathbf{o}) \triangleq \log p_y(\mathbf{w},\mathbf{o}) = \sum_{i=1}^{N} \log\Phi\left(\frac{y_i\mathbf{d}_i^{\mathrm{T}}\mathbf{w} + y_i o_i}{\sigma}\right). \tag{4}$$

Therefore, finding the ML estimate of the vector $(\mathbf{w},\mathbf{o})$ amounts to the optimization problem

$$\underset{\mathbf{w}\in\mathcal{B}_w,\mathbf{o}\in\mathcal{B}_o}{\text{maximize}} \quad l_{wo}(\mathbf{w},\mathbf{o}) \tag{5a}$$

$$\text{subject to: } \operatorname{card}(\mathbf{w}) \le \kappa_w, \quad \operatorname{card}(\mathbf{o}) \le \kappa_o. \tag{5b}$$

Observe that each summand $\log\Phi(u)$ in (4) is increasing in u and tends to zero as $u\to\infty$, therefore the bounding boxes $\mathcal{B}_w$ and $\mathcal{B}_o$ in (5) ensure that the maximizer $(\mathbf{w}^*,\mathbf{o}^*)$ will always exist. It is also well known that the objective $l_{wo}(\mathbf{w},\mathbf{o})$ in (5a) is concave (see e.g., [20, Ch. 3]), but the cardinality constraints on $\mathbf{w}$ and $\mathbf{o}$ are generally intractable [25]. Later in Section V, we propose a convex relaxation approach for dealing with the cardinality constraints in (5b).

It is worth noting that the above formulation is reminiscent of the form of the ML estimator for the probit model [26]. Therefore, our work in this paper can be seen as a natural robustification of such models against outliers (grossly corrupted data points) and/or datasets with a very large number of features per product (necessitating feature selection to obtain meaningful estimates). Further, our approach is based on explicit *structural* assumptions on the unknown parameters: We aim to quantify how sparsity affects the best achievable performance, as well as the performance of the proposed ML estimator.

The ML estimator is a plausible choice for many reasons, primarily because of its appealing asymptotic properties. Before analyzing these properties however, we discuss two issues related to the estimation problem posed in (3). First, we discuss conditions under which the model is *identifiable*, i.e., necessary and sufficient conditions for the estimation problem to be well-defined. Second, we explore the best achievable MSE performance for the estimation of the model parameters, by deriving the CRLB.

III. Identifiability and CRLB

A. Outlier-Free Model

To illustrate the main ideas, it is convenient to start with the simplest case where one assumes that outliers are not present. Such a simplification arises from the model in (3) by assuming

that all auxiliary variables $\{o_i\}_{i=1}^N$ are equal to zero. In such a case the choice-based CA model in matrix form becomes

$$\mathbf{y} = \text{sign}(\mathbf{D}^{\mathrm{T}}\mathbf{w} + \mathbf{e}) \quad \text{with } \mathbf{w} \in \mathcal{W} \triangleq \{\mathbf{u} \in \mathbb{R}^p | \text{card}(\mathbf{u}) \leq \kappa_w\} \tag{6}$$

where we have defined the vector $\mathbf{y}$ of measurements $\{y_i\}_{i=1}^N$ and the vector $\mathbf{e}$ of the i.i.d Gaussian noise variables $\{e_i\}_{i=1}^N$. The matrix $\mathbf{D} \triangleq [\mathbf{d}_1, \ldots, \mathbf{d}_N] \in \mathbb{R}^{p \times N}$ is a matrix whose columns comprise the profile differences. Note that there are no outliers in the model in (6), but only the cardinality constraint on the partworth vector.

The model is said to be statistically identifiable if and only if for $\mathbf{w} \neq \mathbf{w}_0$ the two corresponding random vectors $\mathbf{y}$ and $\mathbf{y}_0$ are not observationally equivalent, i.e., the distribution of the data at the true parameter is different than that at any other possible parameter value. The function $\log \Phi(u)$ is one-to-one, therefore it follows from the expression in (4) that it is necessary and sufficient to have $\mathbf{w} \neq \mathbf{w}_0 \Longrightarrow \mathbf{D}^{\mathrm{T}}\mathbf{w} \neq \mathbf{D}^{\mathrm{T}}\mathbf{w}_0$ to claim identifiability. We emphasize that, in contrast with the theory in linear compressed sensing [27], exact recovery of $\mathbf{w}$ is impossible in our case for finite N, because of the model non-linearity. The notion of statistical identifiability is instead employed, which requires that as $N \to \infty$ the log-likelihood function associated with (6) has a unique global maximum [26].

Therefore, if no sparsity constraints were assumed on $\mathbf{w}$, one would need to impose the condition that $\mathbf{D}$ should be full row rank (which necessitates $p \leq N$) for the estimation problem to be well-defined. Interestingly, when one utilizes the fact that $\mathbf{w}$ has restricted cardinality, one can replace the full row rank condition by a milder condition. To express such a necessary and sufficient condition in a convenient way, we follow ideas and definitions similar to the ones in [27]. Similar to [27], for the matrix $\mathbf{D}^{\mathrm{T}}$ we define the $\text{Spark}(\mathbf{D}^{\mathrm{T}})$ as the smallest integer s, such that there exist s linearly dependent columns in $\mathbf{D}^{\mathrm{T}}$. Then, following the derivation in [27], a necessary and sufficient condition can be expressed in terms of the $\text{Spark}(\mathbf{D}^{\mathrm{T}})$ and the cardinality bound κ_w as

$$\text{Spark}(\mathbf{D}^{\mathrm{T}}) > 2\kappa_w. \tag{7}$$

In other words, if (7) is true then any given vector $\mathbf{w}_0$ obeying the cardinality constraint in (6) will lead to a product $\mathbf{D}^{\mathrm{T}}\mathbf{w}_0$ which is unique. Interestingly, the identifiability condition is the same as if one was observing linear measurements directly, without taking the sign.

We now compute the CRLB for the model in (6). The CRLB is a lower bound on the variance of all unbiased estimators [28, Ch. 3], and therefore serves in practice as a useful exploratory tool. First of all, there is the Fischer Information Matrix (FIM) [28, Ch. 3] for the unconstrained problem, i.e., the FIM for the problem of estimating $\mathbf{w}$ in (6) without making use of the deterministic prior cardinality constraint on $\mathbf{w}$. This matrix is the expected value of the Hessian of the log-likelihood, where the expectation is taken with respect to the measurement vector $\mathbf{y}$. We denote the log-likelihood function for the model in (6) as $l_w(\mathbf{w})$. Naturally, $l_w(\mathbf{w})$ can be obtained from the expression in (4) by setting $o_i = 0 \forall i$. The FIM for the unconstrained problem is defined as $\mathbf{J} \triangleq \mathbb{E}_{\mathbf{y}}\{\nabla^2 l_w(\mathbf{w})\}$.

Given $\{\mathbf{d}_i, y_i\}_{i=1}^N$, the FIM for the unconstrained problem is

$$\mathbf{J} = \mathbf{D}\boldsymbol{\Delta}\mathbf{D}^{\mathrm{T}}, \tag{8}$$

where $\boldsymbol{\Delta} \in \mathbb{R}^{N \times N}$ is a positive diagonal matrix with elements

$$\boldsymbol{\Delta}_{ii} = \frac{1}{2\pi\sigma^2} \exp\left[-\frac{\left(\mathbf{d}_i^{\mathrm{T}}\mathbf{w}\right)^2}{\sigma^2}\right] \cdot \left[\Phi^{-1}\left(\frac{\mathbf{d}_i^{\mathrm{T}}\mathbf{w}}{\sigma}\right) + \Phi^{-1}\left(-\frac{\mathbf{d}_i^{\mathrm{T}}\mathbf{w}}{\sigma}\right)\right]. \tag{9}$$

The derivation is straightforward but is included in the Appendix for completeness. Inverting $\mathbf{J}$ yields the unconstrained CRLB for point $\mathbf{w}$, a bound on MSE which holds for all *unbiased* estimators [28].[2] Note that $\mathbf{J}$ will be singular if $\mathbf{D}$ is not full-row-rank, but we are interested in the constrained CRLB, i.e., the CRLB for points for which we know that they obey the cardinality constraint in (6). The claim is that for such points the bound will be typically lower.

The CRLB for constrained parameter sets is a well-studied topic, see, e.g., [16] and references therein. In essence, the constraint sets considered in [16] are sets of the form $\mathcal{U} \triangleq \{\mathbf{u} \in \mathbb{R}^n | f(\mathbf{u}) = \mathbf{0}, g(\mathbf{u}) \leq \mathbf{0}\}$ where f and g are smooth functions. It has been shown in [16] that smooth inequality constraints do not affect the CRLB; only active equality constraints yield a CRLB which is lower than the unconstrained one. The intuition behind this result is that active equality constraints restrict the unknown parameters into a lower dimensional manifold of the parameter space, leading to much looser requirements about the bias gradient of the estimators applicable. For example, when searching for unbiased estimators applicable to a specific point $\mathbf{u}_0 \in \mathcal{U}$, it suffices to consider estimators unbiased along the *feasible directions* only [16]. The feasible directions can be found at any point $\mathbf{u}_0 \in \mathcal{U}$ by approximating locally the manifold $\mathcal{U}$ by a tangent linear space, which in turn can be described by finding a basis for the nullspace of the gradient matrix associated with function f. Using these definitions, one can associate at each point $\mathbf{u}_0 \in \mathcal{U}$ a matrix $\mathbf{U}$ whose range space is the feasible subspace for $\mathbf{u}_0$. Once such description is found the value of the constrained CRLB depends *only* on the unconstrained FIM and the matrix of feasible directions $\mathbf{U}$ evaluated at the point $\mathbf{u}_0$.

The results of [16] were extended to the case of a singular unconstrained FIM in [17], [18], and later in [19], extensions were made towards the case of non-smooth constraint sets (non-differentiable functions f and g) encompassing also cardinality constraints. In particular, using the terminology of [19], the set $\mathcal{W}$ in (6) is *locally balanced*, meaning that it can be described locally at every point $\mathbf{w}_0 \in \mathcal{W}$ by the span of a set of feasible directions. In other words, one can again associate at every point a matrix of feasible directions $\mathbf{U}$, albeit this cannot be found by differentiation.

To introduce some notation, let $\mathbf{X} \succeq \mathbf{0}$ denote that matrix $\mathbf{X}$ is symmetric positive semidefinite, and let symbol $(\cdot)^{\dagger}$ denote

[2]It is of course possible to make the discussion more general by allowing estimators with a specified bias gradient (not necessarily equal to zero), but here we concentrate on unbiased estimators for simplicity.

the Moore-Penrose pseudoinverse. To state the CRLB for our model in (6), we use the following lemma from [18], [19]:

Lemma 1: Let $\mathcal{R}(\mathbf{U})$ denote the range space of the matrix of feasible directions $\mathbf{U}$. If the condition

$$\mathcal{R}(\mathbf{U}\mathbf{U}^{\mathrm{T}}) \subseteq \mathcal{R}(\mathbf{U}\mathbf{U}^{\mathrm{T}}\mathbf{J}\mathbf{U}\mathbf{U}^{\mathrm{T}}) \tag{10}$$

holds, the covariance of any unbiased estimator for the point $\mathbf{w}_0 \in \mathcal{W}$ satisfies $\mathrm{Cov}(\hat{\mathbf{w}}_0) \succeq \mathbf{U}(\mathbf{U}^{\mathrm{T}}\mathbf{J}\mathbf{U})^{\dagger}\mathbf{U}^{\mathrm{T}}$. Conversely, if the above condition does not hold, there is no *unbiased* finite-variance estimator for $\mathbf{w}_0$.

All that is required now is to be able to specify at any point $\mathbf{w}_0 \in \mathcal{W}$ the matrix $\mathbf{U}$ of feasible directions. Let $\mathrm{supp}(\mathbf{w}_0) = \{i_1, \ldots, i_k\}$ denote the support set of $\mathbf{w}_0$, i.e., the set of indices where the point $\mathbf{w}_0$ is non-zero. Following the same arguments as in [19] one may easily show the following:

- For points of maximal support, i.e., for points where $\mathrm{card}(\mathbf{w}_0) = \kappa_w$, a matrix $\mathbf{U}_s$ of feasible directions consists of the subset of columns of the identity matrix corresponding to the set $\mathrm{supp}(\mathbf{w}_0)$.
- For points of non-maximal support, i.e., for points where $\mathrm{card}(\mathbf{w}_0) < \kappa_w$, every direction of the identity matrix is a feasible direction, therefore $\mathbf{U} = \mathbf{I}_{p\times p}$.

The CRLB results concerning the model in (6) are summarized in the next theorem:

Theorem 1: *Consider the estimation problem in (6) with $\mathbf{w}_0 \in \mathcal{W}$ and assume that (7) holds. For a finite-variance unbiased estimator to exist, the FIM for the uncostrained problem must satisfy (10) whenever $\mathrm{card}(\mathbf{w}_0) < \kappa_w$. Furthermore, the covariance of any unbiased estimator for $\mathbf{w}_0 \in \mathcal{W}$ satisfies*

$$\begin{aligned}\mathrm{Cov}(\hat{\mathbf{w}}_o) &\succeq \mathbf{U}_s\left(\mathbf{U}_s^{\mathrm{T}}\mathbf{D}\Delta\mathbf{D}^{\mathrm{T}}\mathbf{U}_s\right)^{-1}\mathbf{U}_s^{\mathrm{T}} \\ &\qquad \textit{when } \mathrm{card}(\mathbf{w}_0) = \kappa_w, \textit{ and} \\ \mathrm{Cov}(\hat{\mathbf{w}}_o) &\succeq (\mathbf{D}\Delta\mathbf{D}^{\mathrm{T}})^{-1} \\ &\qquad \textit{when } \mathrm{card}(\mathbf{w}_0) < \kappa_w. \end{aligned} \tag{11}$$

Here, $\mathbf{U}_s$ comprises the columns of $\mathbf{I}_{p\times p}$ corresponding to $\mathrm{supp}(\mathbf{w}_0)$.

The condition in (10) ensures the existence of the inverses in (11), and note that it is automatically satisfied when $\mathrm{card}(\mathbf{w}_0) = \kappa_w$ and (7) holds. We observe here that the bounds in (11) are either identical to (i) the bounds that would have been obtained had there been no constraints in the problem (this is the case whenever $\mathbf{w}_0$ has non-maximal support), or (ii) the bounds for estimators with *oracle* performance, i.e., the best achievable MSE obtained by estimators that have perfect knowledge of the *support set* of the point to be estimated (whenever $\mathbf{w}_0$ has maximal support). This has also been observed to be the case for the simpler linear model considered in [19], but it is nice to see that it carries over for the nonlinear model in (6).

Remark 1: The above identifiability and CRLB results are also applicable when one has fewer measurements than partworths ($p > N$), which could be the case when p is very large, and/or choice-data collected from an individual are limited. In this case, however, the condition in (10) cannot be satisfied when $\mathrm{card}(\mathbf{w}_0) < \kappa_w$, as matrix $\mathbf{J}$ becomes rank-deficient, and therefore maximal support in $\mathbf{w}_0$ becomes critical to guarantee meaningful estimation.

B. Outlier-Contaminated Model

It is convenient to work with the model in matrix form, which in this general case becomes

$$\begin{aligned}\mathbf{y} &= \mathrm{sign}\left(\mathbf{D}^{\mathrm{T}}\mathbf{w} + \mathbf{e} + \mathbf{o}\right) \\ &\qquad \text{with } \mathrm{card}(\mathbf{w}) \le \kappa_w, \mathrm{card}(\mathbf{o}) \le \kappa_o. \end{aligned} \tag{12}$$

This case is interesting because there are always more model unknowns than measurements, so one expects sparsity to be the key which makes the problem meaningful. Defining the concatenated matrix $\mathbf{Q} \triangleq [\mathbf{D}^{\mathrm{T}}\mathbf{I}_{N\times N}]$ and following the reasoning in the previous section, one may easily determine a *sufficient* condition for identifiability of $\mathbf{w}$ and $\mathbf{o}$ expressed in terms of $\mathrm{Spark}(\mathbf{Q})$ and the cardinality bounds κ_w and κ_o as

$$\mathrm{Spark}(\mathbf{Q}) > 2(\kappa_w + \kappa_o). \tag{13}$$

The condition is not likely to be also necessary, in the sense that the bound in (13) might actually be tighter than necessary. To get a feel on how restrictive the condition in (13) is, note that generating $\mathbf{D}$ from a continuous distribution yields a $\mathrm{Spark}(\mathbf{Q}) = N + 1$, almost surely. Thus, roughly speaking, assuming that $\kappa_w \ll \kappa_o$ (note that in the regime that we are focusing on, $\kappa_w \ll p$ and $p < N$), one may have—when matrix $\mathbf{D}$ is designed analogously—almost half the measurements contaminated with outliers while still retaining identifiability.

Regarding the CRLB, the main difference for the model in (12) is that one always has more unknowns than measurement equations, therefore the *unconstrained* FIM is expected to be singular in this case. Indeed, it follows readily from Section III-A that the unconstrained FIM is given by

$$\mathbf{J} \triangleq \mathbf{Q}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{Q}, \tag{14}$$

where $\mathbf{\Lambda} \in \mathbb{R}^{N\times N}$ is a positive diagonal matrix with elements

$$\begin{aligned}\mathbf{\Lambda}_{ii} = \frac{1}{2\pi\sigma^2}\exp\left[-\frac{(\mathbf{d}_i^{\mathrm{T}}\mathbf{w} + o_i)^2}{\sigma^2}\right] \\ \cdot\left[\Phi^{-1}\left(\frac{\mathbf{d}_i^{\mathrm{T}}\mathbf{w} + o_i}{\sigma}\right) + \Phi^{-1}\left(-\frac{\mathbf{d}_i^{\mathrm{T}}\mathbf{w} + o_i}{\sigma}\right)\right], \end{aligned} \tag{15}$$

which is singular because $\mathbf{Q}$ is a fat matrix by construction.

The results for the unbiased *constrained* CRLB are of similar flavor to the previous ones given in Theorem 1. With a reasoning similar to that of Section III-A we associate to each point $(\mathbf{w}, \mathbf{o})$ a feasible subspace spanned by

$$\mathbf{U} = \begin{bmatrix}\mathbf{U}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_o\end{bmatrix} \tag{16}$$

where $\mathbf{U}_s$ and $\mathbf{U}_o$ are produced by sampling the columns of the identity matrices $\mathbf{I}_{p\times p}$ and $\mathbf{I}_{N\times N}$ corresponding to $\mathrm{supp}(\mathbf{w})$ and $\mathrm{supp}(\mathbf{o})$, respectively. We are primarily interested in the CRLB for the estimation of the partworth vector, which can be expressed conveniently as shown in the next theorem.

Theorem 2: *Consider the estimation problem in (12) and assume that (13) holds. For a finite variance unbiased estimator to exist, the FIM for the unconstrained problem in (12) must*

satisfy the condition in (10) with matrix $\mathbf{U}$ *defined as in (16). The CRLB on the MSE of any unbiased estimator* $\hat{\mathbf{w}}$ *for point* $\mathbf{w}$ *is given as follows*

$$\begin{aligned}\mathbb{E}\left\|\hat{\mathbf{w}}-\mathbf{w}\right\|_2^2 &\geq \mathrm{Tr}\left(\mathbf{U}_s^{\mathrm{T}}\mathbf{D}\mathbf{L}\mathbf{D}^{\mathrm{T}}\mathbf{U}_s\right)^{-1},\\ &\quad when\ \mathrm{card}(\mathbf{w})=\kappa_w, \mathrm{card}(\mathbf{o})=\kappa_o\\ \mathbb{E}\left\|\hat{\mathbf{w}}-\mathbf{w}\right\|_2^2 &\geq \mathrm{Tr}\left(\mathbf{D}\mathbf{L}\mathbf{D}^{\mathrm{T}}\right)^{-1},\\ &\quad when\ \mathrm{card}(\mathbf{w})<\kappa_w, \mathrm{card}(\mathbf{o})=\kappa_o \quad (17)\end{aligned}$$

where $\mathbf{L}\in\mathbb{R}^{N\times N}$ *is diagonal with* $\mathbf{L}_{ii}=\mathbf{\Lambda}_{ii}$ *if* $\mathbf{o}_i=0$ *and* $\mathbf{L}_{ii}=0$ *if* $\mathbf{o}_i\neq 0$. *No finite-variance unbiased estimator exists whenever* $\mathrm{card}(\mathbf{o})<\kappa_o$.

Proof: Suppose first that $\mathrm{card}(\mathbf{o})=\kappa_o$ and $\mathrm{card}(\mathbf{w})=\kappa_w$. With $\mathbf{J}$ and $\mathbf{U}$ defined in (14) and (16) respectively, observe that

$$\mathbf{U}^{\mathrm{T}}\mathbf{J}\mathbf{U}=\begin{bmatrix}\mathbf{U}_s^{\mathrm{T}}\mathbf{D}\mathbf{\Lambda}\mathbf{D}^{\mathrm{T}}\mathbf{U}_s & \mathbf{U}_s^{\mathrm{T}}\mathbf{D}\mathbf{\Lambda}\mathbf{U}_o\\ \mathbf{U}_o^{\mathrm{T}}\mathbf{\Lambda}\mathbf{D}^{\mathrm{T}}\mathbf{U}_s & \mathbf{U}_o^{\mathrm{T}}\mathbf{\Lambda}\mathbf{U}_o\end{bmatrix}, \quad (18)$$

and note that the product in (18) is non-singular because of the identifiability condition in (13) and because $\mathbf{\Lambda}\succ\mathbf{0}$. From Lemma 1, the CRLB for the point $(\mathbf{w},\mathbf{o})$ is given by the inverse of (18) multiplied by left and right with $\mathbf{U}$ and $\mathbf{U}^{\mathrm{T}}$ respectively. Note that we are interested in obtaining the upper left $p\times p$ block of $\mathbf{U}(\mathbf{U}^{\mathrm{T}}\mathbf{J}\mathbf{U})^{-1}\mathbf{U}^{\mathrm{T}}$, which can be expressed explicitly using block-wise inversion on (18), yielding (after straightforward manipulations) the bound

$$\begin{aligned}&\mathbb{E}\|\hat{\mathbf{w}}-\mathbf{w}\|_2^2 \geq \mathrm{Tr}\Big(\mathbf{U}_s^{\mathrm{T}}\mathbf{D}\sqrt{\mathbf{\Lambda}}\\ &\cdot\left[\mathbf{I}-\sqrt{\mathbf{\Lambda}}\mathbf{U}_o\left(\mathbf{U}_o^{\mathrm{T}}\mathbf{\Lambda}\mathbf{U}_o\right)^{-1}\mathbf{U}_o^{\mathrm{T}}\sqrt{\mathbf{\Lambda}}\right]\sqrt{\mathbf{\Lambda}}\mathbf{D}^{\mathrm{T}}\mathbf{U}_s\Big)^{-1}. \quad (19)\end{aligned}$$

Defining $\mathbf{L}\triangleq\sqrt{\mathbf{\Lambda}}[\mathbf{I}-\sqrt{\mathbf{\Lambda}}\mathbf{U}_o(\mathbf{U}_o^{\mathrm{T}}\mathbf{\Lambda}\mathbf{U}_o)^{-1}\mathbf{U}_o^{\mathrm{T}}\sqrt{\mathbf{\Lambda}}]\sqrt{\mathbf{\Lambda}}$ essentially completes the proof. The cases where $\mathrm{card}(\mathbf{w})<\kappa_w$ and/or $\mathrm{card}(\mathbf{o})<\kappa_o$ are proved by setting $\mathbf{U}_s=\mathbf{I}_{p\times p}$ and/or $\mathbf{U}_o=\mathbf{I}_{N\times N}$ respectively, and noting that in case where $\mathbf{U}_o=\mathbf{I}_{N\times N}$ the product in (18) becomes singular. ■

IV. ML Consistency

Consider the estimation problem in (12) with unknown parameters $(\mathbf{w}_0,\mathbf{o}_0)$. We now show that the ML estimator proposed in (5) will be consistent for the vector of partworths, assuming that the number of outlier-contaminated measurements *increases sublinearly* with N. For the consistency proof, we assume that $\{\mathbf{d}_i\}_{i=1}^N$ are samples from an underlying probability distribution and satisfy the identifiability condition in (13).

Define the set $\mathcal{T}_0\triangleq\mathrm{supp}(\mathbf{o}_0)$ and consider the normalized log-likelihood function

$$\begin{aligned}l_{wo}^N(\mathbf{w},\mathbf{o})=\frac{1}{N}\sum_{i\notin\mathcal{T}_0}\log\Phi\left(\frac{y_i\mathbf{d}_i^{\mathrm{T}}\mathbf{w}+y_io_i}{\sigma}\right)\\ +\frac{1}{N}\sum_{i\in\mathcal{T}_0}\log\Phi\left(\frac{y_i\mathbf{d}_i^{\mathrm{T}}\mathbf{w}+y_io_i}{\sigma}\right). \quad (20)\end{aligned}$$

In (5) we enforce that $\mathrm{card}(\mathbf{o})\leq\kappa_o$. Assuming that $\kappa_o=|\mathcal{T}_0|$, and that $|\mathcal{T}_0|\in\mathcal{O}(N^{1-\epsilon})$ for any positive ϵ, as $N\to\infty$ the law of large numbers implies

$$\lim_{N\to\infty} l_{wo}^N(\mathbf{w},\mathbf{o})\xrightarrow{\mathrm{p}}L_0(\mathbf{w})=\mathbb{E}[\log\Phi(y\mathbf{d}^{\mathrm{T}}\mathbf{w}/\sigma)], \quad (21)$$

where the expectation in (21) is taken with respect to y and $\mathbf{d}$ and the symbol $\xrightarrow{\mathrm{p}}$ denotes convergence in probabilty. By the well-known information inequality [28, pp. 211], $L_0(\mathbf{w})$ has a unique maximum at the true parameter $\mathbf{w}_0$, when this is identifiable. Now, to claim consistency, i.e., to claim that $\hat{\mathbf{w}}_{\mathrm{ML}}$ converges in probability to $\mathbf{w}_0$ as $N\to\infty$, one also needs additional technical conditions to hold. These are typically required to ensure that the limiting and maximization operations in (21) and (5) can be interchanged. Sufficient conditions for the maximum of the limit to be the limit of the maximum are that (*i*) the parameter space is compact and (*ii*) the normalized log-likelihood converges uniformly in probability to $L_0(\mathbf{w})$ as $N\to\infty$ [26].

The condition (*i*) follows immediately, since the parameter space is closed and bounded. To prove (*ii*), note that $l_{wo}^N(\mathbf{w},\mathbf{o})$ is continuous, therefore it suffices to prove the existence of a bounding function for $l_{wo}^N(\mathbf{w},\mathbf{o})$ [26], i.e., a function $\alpha(\mathbf{D},\mathbf{y})$ such that

$$\left|l_{wo}^N(\mathbf{w},\mathbf{o})\right|\leq\alpha(\mathbf{D},\mathbf{y})\ \text{for all points}(\mathbf{w},\mathbf{o}). \quad (22)$$

The existence of such a function $\alpha(\mathbf{D},\mathbf{y})$ along with the continuity of $l_{wo}^N(\mathbf{w},\mathbf{o})$ implies that the normalized log-likelihood converges uniformly in probability to $L_0(\mathbf{w})$ as $N\to\infty$, by the uniform law of large numbers [26]. To this end, note that the derivative $d\log\Phi(u)/du=\ell(u)=d\Phi(u)/\Phi(u)$ is convex and tends to $-u$ as $u\to-\infty$ and to zero as $u\to\infty$. Also, observe that the set $\mathcal{W}$ is a union of subspaces, therefore it contains all points in the line segment between a given point and the point $(\mathbf{w},o_i)=\mathbf{0}$. Hence, a mean value expansion of $\log\Phi(y_i\mathbf{d}_i^{\mathrm{T}}\mathbf{w}+y_io_i)$ around the point $(\mathbf{w},o_i)=\mathbf{0}$ yields

$$\begin{aligned}&|\log\Phi\left(y_i\mathbf{d}_i^{\mathrm{T}}\mathbf{w}+y_io_i\right)|=\\ &|\log\Phi(0)+\ell(y_i\mathbf{d}_i^{\mathrm{T}}\mathbf{w}'+y_io_i')\left(y_i\mathbf{d}_i^{\mathrm{T}}\mathbf{w}+y_io_i\right)|\leq\\ &|\log\Phi(0)|+\ell(y_i\mathbf{d}_i^{\mathrm{T}}\mathbf{w}'+y_io_i')|y_i\mathbf{d}_i^{\mathrm{T}}\mathbf{w}+y_io_i|\leq\\ &|\log\Phi(0)|+C(1+|\mathbf{d}_i^{\mathrm{T}}\mathbf{w}'+y_io_i'|)|y_i\mathbf{d}_i^{\mathrm{T}}\mathbf{w}+y_io_i|\leq\\ &|\log\Phi(0)|+C(1+\|\mathbf{d}_i\|_2\|\mathbf{w}'\|_2+|o_i'|)\left(\|\mathbf{d}_i\|_2\|\mathbf{w}\|_2+|o_i|\right),\end{aligned}$$

where $(\mathbf{w}',o_i')$ is some point in the line segment between $(\mathbf{w},o_i)$ and the zero-point and $C>0$ is a suitable constant. Since any point $(\mathbf{w},o_i)$ satisfies $\|\mathbf{w}\|_\infty\leq R_w$ and $o_i\leq R_o$ we have that

$$\begin{aligned}&\left|l_{wo}^N(\mathbf{w},\mathbf{o})\right|\leq|\log\Phi(0)|+\\ &\frac{C}{N}\sum_{i=1}^N\left[1+\frac{R_w\sqrt{\kappa_w}}{\sigma}\|\mathbf{d}_i\|_2+\frac{R_o}{\sigma}\right]\left[\frac{R_w\sqrt{\kappa_w}}{\sigma}\|\mathbf{d}_i\|_2+\frac{R_o}{\sigma}\right].\end{aligned} \quad (23)$$

Defining the right hand side of (23) as $\alpha(\mathbf{D},\mathbf{y})$ proves the desired bounding condition in (22).

Note that there are isolated cases where the ML estimator may still fail to be consistent due to, for example, insufficient randomness in the data. An interesting such case is when $\sigma\to0$ and $\mathbf{o}_0=\mathbf{0}$. The ML estimator cannot be consistent in this case and this is evident already from the CRLB: In fact, one can observe that there is no finite variance unbiased estimator for the vector of partworths when $\mathbf{o}_0=\mathbf{0}$ and $\sigma\to0$. This is since $\lim_{\sigma\to0}\mathbf{\Delta}_{ii}$ is zero for all $i\in\{1,\cdots,N\}$, and hence in this

noiseless case the FIM becomes singular. Indeed, one can make use of the following well known bounds on $\Phi(u)$

$$\Phi(u) \geq 1 - \frac{1}{2}\exp(-u^2/2)$$

and

$$\Phi(-u) \geq \frac{1}{\sqrt{2\pi}}\frac{u}{1+u^2}\exp(-u^2/2),\ u>0 \qquad (24)$$

to derive that (assuming that $\mathbf{d}_i^{\mathrm{T}}\mathbf{w} > 0$ without loss of generality)

$$\lim_{\sigma\to 0}\mathbf{\Delta}_{ii} \leq \lim_{\sigma\to 0}\frac{1}{\sqrt{2\pi}\sigma^2}\left(\frac{\sigma}{\mathbf{d}_i^{\mathrm{T}}\mathbf{w}}+\frac{\mathbf{d}_i^{\mathrm{T}}\mathbf{w}}{\sigma}\right) \cdot\exp\left[-\frac{(\mathbf{d}_i^{\mathrm{T}}\mathbf{w})^2}{\sigma^2}+\frac{(\mathbf{d}_i^{\mathrm{T}}\mathbf{w})^2}{2\sigma^2}\right]=0. \qquad (25)$$

The intuition behind this noiseless case is interrelated to identifiability: if there is a vector $\mathbf{w}$ consistent with all observations, any vector $c\mathbf{w}$ with $c>0$ will be consistent with the observations as well. Therefore, in the noiseless case there will be ambiguities regarding the magnitude of the true partworth vector, not resolvable by any algorithm not utilizing additional magnitude information. This is consistent with the results of [30], in which the authors provide additional theory and bounds regarding the reconstruction error of the vector $\mathbf{w}$ in this noiseless case.

V. Relaxed ML Estimator

In principle, the ML estimation problem in (5) can be solved exactly by enumerating all possible sparsity patterns for $(\mathbf{w},\mathbf{o})$, and for each sparsity pattern solving a convex optimization problem. Unfortunately however, this direct enumeration approach is often computationally intractable. Instead, one may formulate a tractable approximation to (5) by replacing the cardinality constraints in (5b) with convex ℓ_1-norm constraints. This is motivated since the ℓ_1-norm is the tightest convex relaxation of the cardinality function [20]. Such a replacement yields the convex optimization problem

$$\underset{\mathbf{w},\mathbf{o}}{\text{maximize}} \quad l_{wo}(\mathbf{w},\mathbf{o}) \qquad (26a)$$

$$\text{subject to:} \quad \|\mathbf{w}\|_1 \leq \kappa_w, \|\mathbf{o}\|_1 \leq \kappa_o \qquad (26b)$$

which can be solved efficiently, using, e.g., modern interior point methods [20]. The box-constraints can also be dropped when moving from (5) to (26), since the relaxed ML program (26) always has a maximizer. Further, a more compact way of expressing the relaxed ML estimator is

$$\underset{\mathbf{w},\mathbf{o}}{\text{minimize}}\ \phi(\mathbf{w},\mathbf{o}) = -l_{wo}(\mathbf{w},\mathbf{o}) + \lambda_w\|\mathbf{w}\|_1 + \lambda_o\|\mathbf{o}\|_1 \qquad (27)$$

since (26) and (27) can be shown to be equivalent for a suitable choice of the regularization parameters λ_w and λ_o. These control the trade-off between the value of $l_{wo}(\mathbf{w},\mathbf{o})$ and the number of non-zero elements of $\mathbf{w}$ and $\mathbf{o}$ respectively.

Remark 2: The minimizer of $\phi(\mathbf{w},\mathbf{o})$ may be viewed as a maximum a-posteriori probability (MAP) estimate of $\mathbf{w}$ and $\mathbf{o}$, under the assumption that both $\mathbf{w}$ and $\mathbf{o}$ are random with a Laplacian prior and $\mathbf{w},\mathbf{o}$ and $\mathbf{e}$ are jointly independent. MAP estimation is very commonly used in statistics [28].

The rest of the section is devoted to briefly discussing the choice of the regularization parameters λ_w and λ_o in practice. These parameters are most often tuned in a heuristic fashion: One starts from a suitable initial point $(\lambda_w^i,\lambda_o^i)$ and iterates until the desired sparsity/fit trade-off is achieved. Some assistance may be drawn from the following proposition.

Proposition 1: *The point* $(\mathbf{w}^\star,\mathbf{o}^\star)=\mathbf{0}$ *is optimal for problem (27) if and only if* $\lambda_w \geq \|\nabla_w l_{wo}(\mathbf{0})\|_\infty$ *and* $\lambda_o \geq \|\nabla_o l_{wo}(\mathbf{0})\|_\infty$, *where* $\nabla_w l_{wo}(\mathbf{0})$ *and* $\nabla_o l_{wo}(\mathbf{0})$ *denote the gradients of* $l_{wo}(\mathbf{w},\mathbf{o})$ *with respect to* $\mathbf{w}$ *and* $\mathbf{o}$ *respectively, evaluated at* $(\mathbf{w},\mathbf{o})=\mathbf{0}$.

The proof follows directly from subdifferential calculus and is omitted for brevity. Therefore, for $\lambda_w \geq \lambda_w^{\max} = \|\nabla_w l_{wo}(\mathbf{0})\|_\infty$ and $\lambda_o \geq \lambda_o^{\max} = \|\nabla_o l_{wo}(\mathbf{0})\|_\infty$, the minimization in (27) yields the sparsest possible pair $(\mathbf{w},\mathbf{o})$, the zero vector. A reasonable heuristic approach to tune the parameters is to initialize by choosing $\lambda_w = \lambda_w^{\max}/2$ and $\lambda_o = \lambda_o^{\max}/2$, and adjust to achieve the desired sparsity/fit trade-off. Devising systematic methods on how to choose the penalty parameters is an important topic on its own which deserves further investigation.

VI. Distributed Choice-Based CA

Although the relaxed ML formulation in (27) is a convex optimization problem in standard form (and therefore solvable by polynomial time algorithms), it is often of interest to solve it in a *distributed fashion*. This is because in applications of interest, data are often stored (or collected) in a distributed manner, simply because individuals are not collocated, or due to limited storage, complexity, or even confidentiality constraints. Even if data $\{y_i,\mathbf{d}_i\}_{i=1}^N$ are centrally available, often the number N of observed samples is extremely large, and standard interior point methods cannot handle efficiently the optimization in (27).

Interestingly, the structure of (27) lends itself naturally to distributed implementation via the alternating-direction method of multipliers (ADMM), an iterative Lagrangian method especially well-suited for parallel processing [21]. ADMM blends the benefits of dual decomposition and augmented Lagrangian methods. Essentially, the name derives from the fact that the algorithm alternates between optimizing different variables in the augmented Lagrangian function.

If we assume that the observed data are partitioned into M blocks $\{N_i\}_{i=1}^M$, then the goal is to split the objective function of (27) into M terms, and let each term to be handled by its individual processing unit (such as a thread or processor). To ensure the scalability properties of the algorithm, it is convenient to define the (convex) function $g_i:\mathbb{R}^p\to\mathbb{R}$ as

$$g_i(\mathbf{w}) \triangleq \sum_{j\in N_i}\inf_{o_j}\left[-\log\Phi\left(\frac{y_j\mathbf{d}_j^{\mathrm{T}}\mathbf{w}+y_jo_j}{\sigma}\right)+\lambda_o|o_j|\right]. \qquad (28)$$

Introducing M local auxiliary variable vectors $\mathbf{w}_i \in \mathbb{R}^p$ and the global variable $\mathbf{z} \in \mathbb{R}^p$, one can *equivalently* write problem (27) in its consensus form [21]

$$\underset{\{\mathbf{w}_i\}_{i=1}^M, \mathbf{z}}{\text{minimize}} \quad \sum_{i=1}^{M} g_i(\mathbf{w}_i) + \lambda_w \|\mathbf{z}\|_1 \tag{29a}$$

$$\text{subject to:} \quad \mathbf{w}_i - \mathbf{z} = \mathbf{0},\ i = 1, \cdots, M. \tag{29b}$$

Problem (29) is called the global consensus problem, owing to the consensus constraint [in (29b)] which enforces all the local variables to be equal. The optimization problem in (29) can be solved by applying the generic global variable consensus ADMM algorithm described in [21, Ch. 7]. The derivation of the distributed algorithm follows easily from the theory in [21, Ch. 7]; therefore, here we only present and explain the basic steps of the distributed algorithm. Upon defining the dual variables $\mathbf{u}_i \in \mathbb{R}^p$ and a fixed parameter $\rho > 0$ (often called the penalty parameter), each iteration of the algorithm comprises the following three main updates (k below denotes the iteration index):

$$\mathbf{w}_i{}^{k+1} := \underset{\mathbf{w}_i}{\arg\min} \quad g_i(\mathbf{w}_i) + (\rho/2) \left\| \mathbf{w}_i - \mathbf{z}^k + \mathbf{u}_i^k \right\|_2^2 \tag{30a}$$

$$\mathbf{z}^{k+1} := \underset{\mathbf{z}}{\arg\min}\ \lambda_w \|\mathbf{z}\|_1 + (M\rho/2) \|\mathbf{z} - \bar{\mathbf{w}}^{k+1} - \bar{\mathbf{u}}^k\|_2^2 \tag{30b}$$

$$\mathbf{u}_i^{k+1} := \mathbf{u}_i^k + \mathbf{w}_i^{k+1} - \mathbf{z}^{k+1}. \tag{30c}$$

The step in (30a) can be carried out in parallel for each data block. The second step requires gathering the vectors $\{\mathbf{w}_i{}^{k+1}\}_{i=1}^M$ and $\{\mathbf{u}_i{}^k\}_{i=1}^M$ to form their averages, which are denoted as $\bar{\mathbf{w}}^{k+1}$ and $\bar{\mathbf{u}}^k$, respectively. Note that the objective in (30b) is fully separable in the global variable $\mathbf{z}$, therefore the minimization can be carried out component-wise. In this case, a scalar z_i-update is

$$z_i^{k+1} := \underset{z_i}{\arg\min} \left(\lambda_w |z_i| + \frac{M\rho}{2} (z_i - \bar{w}_i^{k+1} - \bar{u}_i^k)^2 \right), \tag{31}$$

which admits a simple closed form solution. Explicitly, the solution of (31) is $z_i = \mathcal{S}_{\frac{\lambda_w}{M\rho}}(\bar{w}_i^{k+1} + \bar{u}_i^k)$, where $\mathcal{S}$ is the so-called *soft thresholding* operator defined as $\mathcal{S}_\kappa(\alpha) \triangleq \alpha \max\{1 - \kappa/|\alpha|, 0\}$ [21]. Thus, each iteration of the ADMM algorithm requires a *gather* and a *broadcast* operation: after the optimization in (30a), each node needs to communicate $\mathbf{w}_i^{k+1} \in \mathbb{R}^p$ along with $\mathbf{u}_i^k \in \mathbb{R}^p$ to the centralizer. The centralizer then gathers these variables, forms the necessary averages, updates the global variable $\mathbf{z}^{k+1} \in \mathbb{R}^p$, and broadcasts this updated global variable to the nodes. Note that the algorithm is scalable with respect to N, because outlier processing is strictly restricted to the individual nodes—outlier variables need not be shared for convergence. Overall, observe that the iterations produce an algorithm which is not only distributed, but also *decentralized*: A node does not need access to the individual data of another—only the consensus variable $\mathbf{z}^{k+1}$ is needed to be shared for convergence. Such decentralized solutions might be preferable from centralized ones for many reasons, even for modestly sized datasets (for example, due to the confidentiality requirements).

Following a random initialization, the iterations in (30) are guaranteed to converge to an optimal point for (29) as $k \to \infty$. In practice, although ADMM can be very slow to converge to high accuracy, it usually converges to modest accuracy within a few tens of iterations [21]. Thankfully, our simulation examples indicate that modest accuracy is sufficient in this context, motivating the practical use of this algorithm.

VII. Numerical Results

A. Estimation Performance Compared to the CRLB

1) Outlier-Free Measurements: In this part, we explore the MSE performance of two different ML estimator (MLE) variants, in the case where *outliers are not present* in the data. Profile differences $\mathbf{d}_i$ were generated as i.i.d Gaussian vectors drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, each comprising $p = 20$ elements. The unknown vector $\mathbf{w}_0$ was generated (sparse) i.i.d. Gaussian with $\mathrm{NZ} = 3$ non-zero elements drawn from $\mathcal{N}(0, 1)$. The MSE of both variants was evaluated using $\mathrm{MC} = 300$ Monte Carlo trials. For each trial, binary data were generated according to the model in (6). The additive noise e_i in the responses was assumed i.i.d. drawn from $\mathcal{N}(0, 1)$. The particular MLE variants chosen here are (α) a *sparsity-agnostic* ML estimator (MLE-SAG), which assumes that $\mathbf{o} = \mathbf{0}$ and ignores sparsity on $\mathbf{w}_0$, and (β) a sparsity-aware MLE (MLE-SAW) which assumes that $\mathbf{o} = \mathbf{0}$ and also knows that $\mathbf{w}_0$ is NZ-sparse. To implement the MLE-SAW, instead of solving (5) directly by setting $\mathbf{o} = \mathbf{0}$ and enumerating all possible sparsity patterns for $\mathbf{w}$, we obtain the estimate for the partworth vector through relaxation. In particular, we first (*i*) solve the problem in (27) with $\mathbf{o} = \mathbf{0}$ to obtain a plausible sparsity pattern for $\mathbf{w}$, and then (ii) we re-solve the problem having the sparsity pattern in $\mathbf{w}$ fixed. For carrying out (*i*) we choose $\lambda_w = 0.1\lambda_{max}$, where $\lambda_{max} \triangleq \|\nabla_w l_w(\mathbf{0})\|_\infty$, and retain the NZ largest elements as a plausible non-zero pattern.

The (Root)-MSE results are depicted in Fig. 1, where two additional CRLB curves are plotted as functions of the number of samples N. CRLB-PS is the CRLB of any unbiased estimator utilizing the knowledge that $\mathbf{w}_0$ is NZ-sparse, while CRLB-NPS is the CRLB of any unbiased estimator not utilizing the information that $\mathbf{w}_0$ is NZ-sparse. Observe the difference in the best achievable error performance, to get a feel on how sparsity in $\mathbf{w}_0$ can affect the expected estimation accuracy. One expects that the effect of the prior information regarding partworth sparsity on the best achievable MSE performance will diminish as N grows, and that the two CRLB curves will meet at some point, but we see that the rate of which this happens can actually be rather slow. Both estimators (and MLE-SAW in particular) operate close to their respective CRLBs, which is intuitively satisfying. The price paid by the estimator which does not account for sparsity in the pathworth vector is evident from the figure.

2) Outlier-Contaminated Measurements: Next, the case where *outliers are also present* in the responses is examined. In this experiment we consider an outlier percentage of 1% [outliers correspond to (uniform at random) sign changes]. Other than the outlier addition in the responses, we use the exact same setup as in the outlier-free case, to evaluate the MSE performance of our sparsity-aware ML formulation (MLE-OD) in (27) against the CRLB, and also the performance of the

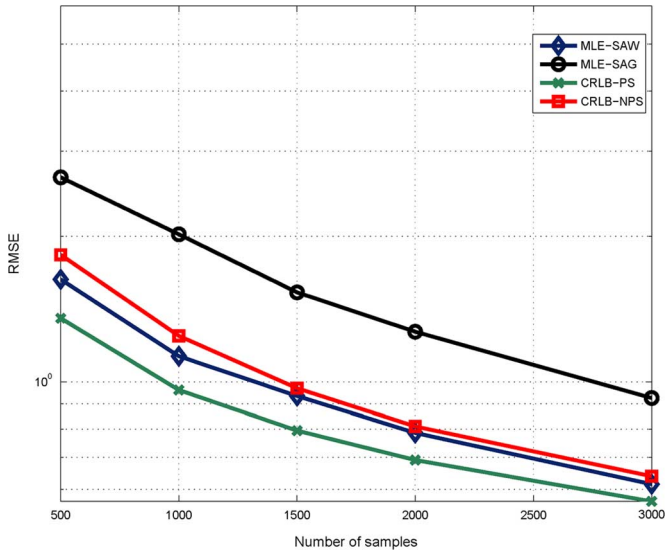


Fig. 1. RMSE comparison of the different MLE variants against their respective CRLBs for different sample sizes N. The MLE which accounts for partworth sparsity was implemented using the two-step procedure described in the text.

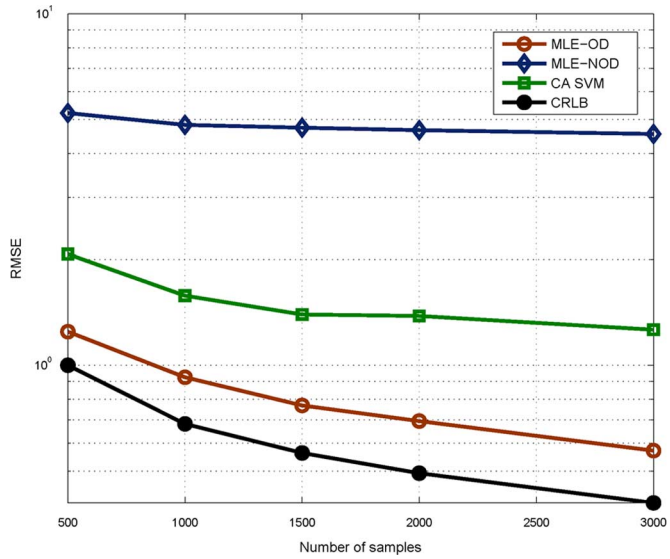


Fig. 2. RMSE comparison of the MLE versus the SVM estimator from [11] and the CRLB for different number of samples N, when outliers are present in the data [outlier percentage 1%]. Both the MLE and the SVM were implemented using the two-step procedure described in the text. Mean outlier efficiency of MLE-OD was found 99.2%, 95.7%, 96.3%, 93.7%, 94.2% for $N = 500, 1000, 1500, 2000, 3000$ samples, respectively.

SVM partworth estimator proposed in [11] (CA SVM), as this is another related method which deals explicitly with outliers. While both estimators know exactly the *degree* of outlier and partworth sparsity, for the SVM estimator of [11] we assume in addition exact knowledge of the *outlier support*, i.e., we provide the SVM with perfect knowledge of the outlying data points in every trial (thus eliminating the need to tune the regularization parameter for the SVM as far as outliers are concerned). To account for the sparsity in $\mathbf{w}_0$, the ℓ_1-norm regularized counterpart of the SVM of [11] was used (see in particular ([11], (6)) and ensuing discussion). The ℓ_1 regularization parameter for $\mathbf{w}$ was tuned in every trial using a five point equispaced grid ($\lambda_w \in [1, 5]$) so as to yield the closest to NZ-sparsity level in the estimate $\hat{\mathbf{w}}$. Upon obtaining a plausible sparsity pattern for $\mathbf{w}$, we re-solve the SVM problem having the sparsity pattern fixed. On the other hand, MLE-OD was implemented by first (i) solving (27) using $\lambda_w = 0.1\|\nabla_w l_{wo}(\mathbf{0}, \mathbf{0})\|_\infty$ and $\lambda_o = 0.1\|\nabla_o l_{wo}(\mathbf{0}, \mathbf{0})\|_\infty$ to obtain a plausible sparsity pattern for $(\mathbf{w}, \mathbf{o})$ (by retaining the NZ and $0.01N$ largest elements in $\mathbf{w}$ and $\mathbf{o}$ as non-zeros, respectively), and (ii) re-solving the problem having the sparsity pattern in $(\mathbf{w}, \mathbf{o})$ fixed.

The Root-MSE results are plotted in Fig. 2, as a function of the number of measurements. Observe that MLE-OD, which makes full use of the model where data are generated from, operates closer to the CRLB than CA SVM. The outlier-detection performance of the method is also reported in the figure text, measuring the average percentage of detected outliers (total number of outliers detected divided by the total number of outliers present) for different N. As it turns out, for this set of trials the method seems to exhibit consistently an outlier-detection performance of at least 93%. The method has outlier-misses, but these missed outliers seem to be relatively harmless to the estimation acuracy, as implied by the RMSE performance in Fig. 2. Note that CA SVM (provided with perfect outlier knowledge) identified the correct support of the partworth vector with an accuracy of 100%—still however, we see from Fig. 2 that its performance is limited by the model mismatch. The performance of an *outlier-agnostic* MLE variant (MLE-NOD)—an MLE variant which ignores the presence of outlying data points but still accounts for sparsity in $\mathbf{w}$—is also included. Observe how important the log-likelihood robustification can be in practice; the outlier-agnostic MLE essentially breaks down even from few badly corrupted data points.

B. Additional Comparisons With Other SVM Variants

In this section we compare our distributed implementation in (30) against a particular SVM variant inspired by [8]. The loss function associated with this variant is the so-called *hinge loss*, which yields the tightest convex relaxation of the classifier that attempts to minimize the number of misclassifications. The hinge loss is inherently robust against sparse outliers [8], and this is the reason why the comparison with this variant is also important. We use both synthetic and real data coming from a conjoint choice experiment for coffee makers. In the comparisons we always include the SVM variant proposed in [11], whose performance has been shown to be very competitive (and even superior) to that of [8].

1) Synthetic Data: The metric chosen for the comparison here is the Normalized Mean Squared Error (NMSE) between the estimated and "true" partworths, i.e., the MSE after partworths have been normalized in (ℓ_2-norm) magnitude. Normalized metrics are useful in some CA studies, especially when data are limited; similar metrics were also adopted in [8], [11]. The performance of the three methods was estimated using $\mathrm{MC} = 50$ Monte Carlo trials. For each trial, product profiles were generated as i.i.d Gaussian vectors, each comprising $p = 20$ attributes/features. For each trial we constructed $N = 500$ choice questions, by constructing vector differences randomly among

the generated profiles. We considered two different settings: (*i*) one where all $N = 500$ choice questions were used for the purpose of estimation, and (*ii*) a reduced-size (questionnaire) setting, where 50 choice questions were randomly drawn from this complete set of 500. Choice data were generated according to model (3). We experimented using two different outlier percentages in the responses, 4% and 10% (outliers correspond to sign change in y_i). The unknown partworth vector was assumed sparse (NZ $= 3$ non-zero entries) i.i.d. Gaussian.

To account for sparsity in the partworths, the ℓ_1-norm regularized counterparts of the SVMs proposed in [8] (abbreviated here as $L1$-SVM) and [11] (abreviated here as CA SVM) were used. For our distributed ML estimator we assumed $M = 5$ clusters of data of equal size, and a penalty parameter $\rho = 1$. For the ADMM, variables $\mathbf{z}$ and $\{\mathbf{u}_i\}_{i=1}^M$ were always initialized from zero. For the purpose of illustration, we demonstrate in Fig. 3 the objective suboptimality [the distance from the optimal value of (27)] of the distributed algorithm versus iterations for different values of penalty parameter ρ. This is for a typical problem instance with 50 choice questions, $M = 5$ and 4% outliers, where one can see that the algorithm converges in sufficient accuracy (on the order of 10^{-3}) in at most 60–70 iterations, depending on the value of the penalty parameter ρ. In this particular example one observes better convergence behavior for $\rho = 1$, but this in general depends on the particular data instance generated. We assume that the degree of sparsity in both $\mathbf{w}$ and $\mathbf{o}$ is known by all estimators, allowing to tune the parameters in every trial, using a grid and picking the values that yield sparsity levels closer to those known by the estimators. Proposition 1 was used to construct a grid for the MLE, with points equally spaced within the box $[0, 5\lambda_w^{\max}] \times [0, 5\lambda_o^{\max}]$ (10 values in each dimension). For the CA SVM 10 equispaced points for each one of the two parameters were used (from 1 to 10 for each parameter). For $L1$-SVM, 10 equispaced points (from 1 to 10) were used for its single partworth sparsity parameter. For every method, upon obtaining the best sparsity pattern for $(\mathbf{w}, \mathbf{o})$, the problem was re-solved using an interior point method, having the sparsity pattern fixed.

The results of the comparison are reported in Table I. Note that these are just *reference illustrations*: the performance of every method considered can perhaps be further improved by allowing more careful tuning, using denser grids and/or perhaps manual work. It is however evident from the trials that the methods which explicitly account for outliers in the responses perform better than those who do not, and that they exhibit highly competitive performance for all practical purposes, as far as NMSE is concerned. The ML estimator is slightly superior, stemming from better use of the statistical model used for data generation. The $L1$-SVM appears to be consistently inferior than the other two methods, although its performance in the reduced-sized questionnaire with a small percentage of outliers is competitive as well.

2) Real Data: In a similar comparison, we now use a real dataset where consumer responses might violate our modeling assumptions. We briefly describe the general setup; all details can be found in [24, Ch. 13.6].

Hypothetical coffee makers were defined using the following five attributes:

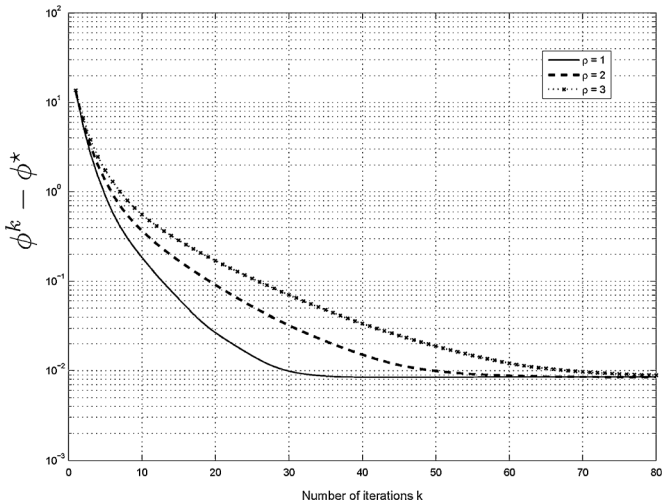


Fig. 3. Objective suboptimality of distributed MLE in (30) versus iteration for a typical sample run with $M = 5$, 50 choice questions and 4% outliers, for different values of penalty parameter ρ.

TABLE I
NMSE COMPARISON OF THE THREE METHODS: $L1$-SVM, CA SVM FROM [11], AND THE PROPOSED METHOD (30). THE METHOD THAT YIELDS LOWER NMSE IS MARKED WITH BOLD

Outliers	**Questions**	$L1$**-SVM**	**SVM[11]**	**Proposed** (30)
4%	50	0.0934	0.0450	**0.0115**
4%	500	0.0059	0.0007	**0.0006**
10%	50	0.1584	0.1447	**0.0620**
10%	500	0.0328	0.0021	**0.0015**

— [Brand] brand-name (Phillips, Braun, Moulinex)
— [Capacity] number of cups (6, 10, 15)
— [Price] price in Dutch Guilders f: (39, 69, 99)
— [Thermos] presence of a thermos flask (yes, no)
— [Filter] presence of a special filter (yes, no)

A total of sixteen profiles were constructed from combinations of the levels of these attributes using an incomplete design [24]. These sixteen profiles are represented mathematically as vectors in $\mathbb{R}^7$ (with three binary entries describing the brand of the product). In the choice experiment, respondents were asked to make choices out of sets of three profiles, each set containing the same base alternative [24]. Therefore, each choice expresses two strict preferences between different coffee makers. In total, 185 respondents were recruited in the experiment and each one provided data for 16 choices. Links for the actual dataset used in this part can be found in [24].

For all three estimators, our metric in this case was the predictive performance, or the "hit-rate" of each method, which we assessed by reserving the last out of the 16 choices for each individual and testing how often the estimated utility functions predict the correct winning product. A different partworth vector was assumed for each individual, which was estimated based on his/her choices alone. The parameters $(\sigma, \lambda_w, \lambda_o)$ for the MLE were tuned using a grid in each dimension (four equispaced points for $\sigma \in [0.1, 1]$, five points for $\lambda_w \in [1, 5]$, and five points for $\lambda_o \in [1, 5]$), and picking the values for which the predictive performance was maximized. This performance was assessed

from the (first fifteen) choices using the leave-one-out error approach of [8]. The choice of the parameters for the SVMs was carried out in a similar fashion, using the same grids as above for each associated parameter.

The observed classification performance was found very competitive for the MLE and CA SVM, $\approx$ 95% for the MLE (176/185) and $\approx$ 94% for the CA SVM (174/185). The $L1$-SVM resulted in lower classification accuracy $\approx$ 92% (171/185), suggesting that careful outlier modeling can have considerable implications in predictions as well.

VIII. Conclusion & Future Work

The paper proposes a new method for choice-based CA. The proposed framework allows for exploring choice-based CA through the scope of sparse estimation, giving insight into identifiability conditions and Cramér-Rao bounds which may serve as useful design tools. For the estimation of the model parameters, the proposed ML estimator leads to a formulation which can efficiently handle large scale datasets through a simple distributed implementation, which seems to perform very well in practice. There is a number of interesting topics arising as future work, including a theoretical analysis of the ML performance in finite sample sizes. Note that our CRLB results indicate that stable recovery of partworths (with a recovery error in the order of the standard errors) is indeed possible for finite N, despite the model non-linearity. Therefore, a theoretical performance analysis of the ML estimator for finite N could yield interesting results, under perhaps additional conditions. Exploring conditions under which the estimates obtained by (27) and (5) coincide is also an interesting research topic, which we do not touch upon in this paper.

Appendix

A. Derivation of the Unconstrained CRLB for (6)

The gradient and hessian of $l_w(\mathbf{w})$ are respectively (assuming that $\sigma = 1$ for simplicity)

$$\bullet\ \nabla l_w(\mathbf{w}) = \sum_{i=1}^{N} \frac{y_i/\sqrt{2\pi}}{\Phi\left(y_i \mathbf{d}_i^{\mathrm{T}}\mathbf{w}\right)} \exp\left[-\frac{(\mathbf{d}_i^{\mathrm{T}}\mathbf{w})^2}{2}\right] \mathbf{d}_i \text{ and}$$

$$\bullet\ \nabla^2 l_w(\mathbf{w}) = \sum_{i=1}^{N} \frac{y_i/\sqrt{2\pi}}{\Phi\left(y_i \mathbf{d}_i^{\mathrm{T}}\mathbf{w}\right)} \exp\left[-\frac{\left(\mathbf{d}_i^{\mathrm{T}}\mathbf{w}\right)^2}{2}\right] \left(\mathbf{d}_i^{\mathrm{T}}\mathbf{w}\right) \mathbf{d}_i\mathbf{d}_i^{\mathrm{T}} + \sum_{i=1}^{N} \frac{1/2\pi}{\Phi\left(y_i \mathbf{d}_i^{\mathrm{T}}\mathbf{w}\right)^2} \exp\left[-\frac{(\mathbf{d}_i^{\mathrm{T}}\mathbf{w})^2}{2}\right]^2 \mathbf{d}_i\mathbf{d}_i^{\mathrm{T}}.$$

Upon defining the matrix $\mathbf{D} = [\mathbf{d}_1 \cdots \mathbf{d}_N] \in \mathbb{R}^{p\times N}$, observe that the hessian of $l_w(\mathbf{w})$ can be written as $\nabla^2 l_w = \mathbf{D}\mathbf{M}\mathbf{D}^{\mathrm{T}}$, where $\mathbf{M} \in \mathbb{R}^{N\times N}$ is a diagonal matrix with elements

$$\mathbf{M}_{ii} = \frac{y_i/\sqrt{2\pi}}{\Phi\left(y_i \mathbf{d}_i^{\mathrm{T}}\mathbf{w}\right)} \exp\left[-\frac{(\mathbf{d}_i^{\mathrm{T}}\mathbf{w})^2}{2}\right] (\mathbf{d}_i^{\mathrm{T}}\mathbf{w}) + \frac{1/2\pi}{\Phi\left(y_i \mathbf{d}_i^{\mathrm{T}}\mathbf{w}\right)^2} \exp\left[-\left(\mathbf{d}_i^{\mathrm{T}}\mathbf{w}\right)^2\right]. \quad (32)$$

The probability density function for the y_i is

$$y_i \sim \begin{cases} -1, \text{with probability } \Phi\left(-\mathbf{d}_i^{\mathrm{T}}\mathbf{w}\right) \\ 1, \text{with probability } \Phi\left(\mathbf{d}_i^{\mathrm{T}}\mathbf{w}\right) \end{cases}. \quad (33)$$

Note that the so-called *regularity condition* [28] on the log-likelihood is satisfied because

$$\mathbb{E}\{\nabla l_w(\mathbf{w})\} = -\sum_{i=1}^{N} \mathbb{E}\left\{\frac{y_i/\sqrt{2\pi}}{\Phi\left(y_i \mathbf{d}_i^{\mathrm{T}}\mathbf{w}\right)}\right\} \exp\left[-\frac{(\mathbf{d}_i^{\mathrm{T}}\mathbf{w})^2}{2}\right] \mathbf{d}_i = \mathbf{0}, \quad (34)$$

therefore the CRLB for the unconstrained problem holds. The (unconstrained) FIM is the expected value of the $\nabla^2 l_w$ with respect to $\{y_i\}_{i=1}^N$, hence, it suffices to compute the expected value of each entry of the diagonal matrix $\mathbf{M}$. We get

$$\begin{aligned} \boldsymbol{\Delta}_{ii} = \mathbb{E}\{\mathbf{M}_{ii}\} &= \mathbb{E}\left\{\frac{1/2\pi}{\Phi\left(y_i \mathbf{d}_i^{\mathrm{T}}\mathbf{w}\right)^2} \exp\left[-(\mathbf{d}_i^{\mathrm{T}}\mathbf{w})^2\right]\right\} \\ &= \frac{1}{2\pi} \exp\left[-(\mathbf{d}_i^{\mathrm{T}}\mathbf{w})^2\right] \left[\Phi^{-1}\left(\mathbf{d}_i^{\mathrm{T}}\mathbf{w}\right) + \Phi^{-1}\left(-\mathbf{d}_i^{\mathrm{T}}\mathbf{w}\right)\right], \end{aligned} \quad (35)$$

and thereby proving the expression given in (8).

References

[1] R. Meyer and J. Pratt, "The consistent assessment and fairing of preference functions," *IEEE Trans. Syst. Sci. Cybern.*, vol. 4, no. 3, pp. 270–278, Sep. 1968.

[2] V. Srinivasan and A. Shocker, "Linear programming techniques for multidimensional analysis of preferences," *Psychometrica*, vol. 38, no. 3, pp. 337–369, 1973.

[3] E. Tsakonas, J. Jaldén, N. Sidiropoulos, and B. Ottersten, "Maximum likelihood based sparse and distributed conjoint analysis," in *Proc. Statist. Signal Process. Workshop (SSP)*, Ann Arbor, MI, Aug. 5–8, 2012.

[4] A. Gustafsson, A. Herrmann, and F. Huber, *Conjoint Measurement: Methods and Applications*. Berlin, Germany: Springer-Verlag, 2007.

[5] J. Louviere, D. Hensher, and J. Swait, *Stated Choice Methods: Analysis and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[6] W. DeSarbo and A. Ansari, "Representing heterogeneity in consumer response models," *Market. Lett.*, vol. 8, no. 3, pp. 335–348.

[7] P. Lenk, W. DeSarbo, P. Green, and M. Young, "Hierarchical Bayes conjoint analysis: Recovery of partworth heterogeneity from reduced experimental designs," *Market. Sci.*, vol. 15, no. 2, pp. 173–191, 1996.

[8] T. Evgeniou, C. Boussios, and G. Zacharia, "Generalized robust conjoint estimation," *Market. Sci.*, vol. 24, no. 3, pp. 415–429, 2005.

[9] O. Chapelle and Z. Harchaoui, "A machine learning approach to conjoint analysis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, vol. 17, pp. 257–264.

[10] J. Wind, P. Green, D. Shifflet, and M. Scarbrough, "Courtyard by Marriott: Designing a hotel facility with consumer-based marketing models," *Interfaces*, vol. 19, pp. 25–47, 1989.

[11] G. Mateos and G. Giannakis, "Robust conjoint analysis by controlling outlier sparsity," in *Proc. Eur. Signal Process. Conf.*, Barcelona, Spain, Aug. 29–Sep. 2 2011.

[12] J. Fuchs, "An inverse problem approach to robust regression," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Phoenix, AZ, USA, Mar. 15–19, 1999.

[13] J. Wright and Y. Ma, "Dense error correction via ℓ^1-minimization," *IEEE Trans. Inf. Theory*, vol. 56, no. 7, pp. 3540–3560, Jul. 2010.

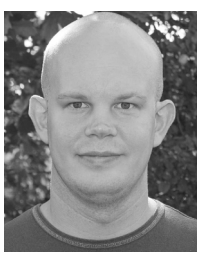
[14] G. Giannakis, G. Mateos, S. Farahmand, V. Kekatos, and H. Zhu, "USPACOR: Universal sparsity-controlling outlier rejection," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Prague, Czech Republic, May 2011.

[15] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Statist. Soc.*, pp. 267–288, 1996.

[16] J. D. Gorman and A. O. Hero, "Lower bounds for parametric estimation with constraints," *IEEE Trans. Inf. Theory*, vol. 26, no. 6, pp. 1285–1301, Jun. 1990.
[17] P. Stoica and B. C. Ng, "On the Cramér-Rao bound under parametric constraints," *IEEE Signal Process. Lett.*, vol. 5, no. 7, pp. 177–179, Jul. 1998.
[18] Z. Ben-Haim and Y. C. Eldar, "On the constrained Cramér-Rao bound with a singular Fisher information matrix," *IEEE Signal Process. Lett.*, vol. 16, no. 6, pp. 453–456, Jun. 2009.
[19] Z. Ben-Haim and Y. C. Eldar, "The Cramér-Rao bound for estimating a sparse parameter vector," *IEEE Trans. Signal Process.*, vol. 58, no. 6, Jun. 2010.
[20] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[21] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends in Mach. Learn.*, pp. 1–122, 2010.
[22] T. Evgeniou, M. Pontil, and O. Toubia, "A convex optimization approach to modeling consumer heterogeneity in conjoint analysis," *Market. Sci.*, vol. 26, no. 6, pp. 805–818, 2007.
[23] 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, USA: Springer, 2007, pp. 231–258.
[24] A. Skrondal and S. Rabe-Hesketh, *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton, FL, USA: Chapman & Hall/CRC.
[25] B. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, pp. 227–234, 1995.
[26] W. Newey and D. McFadden, "Chapter 35: Large sample estimation and hypothesis testing," in *Handbook of Econometrics*. New York, NY, USA: Elsevier Science, vol. 4, pp. 2111–2245.
[27] D. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," in *Proc. Natl. Acad. Sci. USA*, 2003, vol. 100, pp. 2197–2202.
[28] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.
[29] L. Wang, J. Zhu, and H. Zou, "The doubly regularized support vector machine," *Statist. Sinica*, vol. 16, pp. 589–615, 2006.
[30] L. Jacques, J. N. Laska, P. T. Boufounos, and R. G. Baraniuk, "Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors." Feb. 2012 [Online]. Available: http://arxiv.org/abs/1104.3160v2

Efthymios Tsakonas received the Diploma degree in electronics and computer engineering from the Technical University of Crete, Greece, 2008.

Since September 2009, he has been a member of the Signal Processing Laboratory, Royal Institute of Technology (KTH), where he is currently working toward the Ph.D. degree in signal processing. His current research interests are in particle filtering, robust optimization and convex approximation theory and algorithms, with particular emphasis on preference measurement and conjoint analysis applications.



Joakim Jaldén (S'03–M'08–SM'13) received the M.Sc. and Ph.D. degrees in electrical engineering from the Royal Institute of Technology (KTH), Stockholm, Sweden, in 2002 and 2007, respectively.

From July 2007 to June 2009, he held a postdoctoral research position at the Vienna University of Technology, Vienna, Austria. He also studied at Stanford University, CA, from September 2000 to May 2002, and worked at ETH, Zürich, Switzerland, as a visiting researcher, from August to September 2008. In July 2009, he joined the Signal Processing Lab, School of Electrical Engineering, KTH, as an Assistant Professor.

Dr. Jaldén was an Associate Editor for the IEEE COMMUNICATIONS LETTERS between 2009 and 2011, is as an Associate Editor for the IEEE TRANSACTIONS IN SIGNAL PROCESSING since 2012, and is a member of the IEEE Signal Processing for Communications and Networking Technical Committee (SPCOM-TC). For his work on MIMO communications, he was awarded the IEEE Signal Processing Society's 2006 Young Author Best Paper Award and the first prize in the Student Paper Contest at the 2007 International Conference on Acoustics, Speech and Signal Processing (ICASSP). He is also a recipient of the Ingvar Carlsson Award issued in 2009 by the Swedish Foundation for Strategic Research, and the coauthor of a Best Student paper at the 2012 International Symposium on Biomedical Imaging (ISBI).

Nicholas D. Sidiropoulos (F'09) received the Diploma in electrical engineering from the Aristotelian University of Thessaloniki, Greece, and the M.S. and Ph.D. degrees in electrical engineering from the University of Maryland—College Park, in 1988, 1990, and 1992, respectively.

He served as Assistant Professor at the University of Virginia (1997–1999); Associate Professor at the University of Minnesota—Minneapolis (2000–2002); Professor at the Technical University of Crete, Greece (2002–2011); and Professor at the University of Minnesota—Minneapolis (2011-). His current research focuses primarily on signal and tensor analytics, with applications in cognitive radio, big data, and preference measurement.

Dr. Sidiropoulos received the NSF/CAREER award (1998), the IEEE Signal Processing Society (SPS) Best Paper Award (2001, 2007, 2011), and the IEEE SPS Meritorious Service Award (2010). He served as IEEE SPS Distinguished Lecturer (2008–2009) and Chair of the IEEE Signal Processing for Communications and Networking Technical Committee (2007–2008).

Bjorn Ottersten (S'87–M'89–SM'99–F'04) was born in Stockholm, Sweden, 1961. He received the M.S. degree in electrical engineering and applied physics from Linkoping University, Linkoping, Sweden, in 1986. In 1989, he received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA.

In 1991, he was appointed Professor of Signal Processing at the KTH Royal Institute of Technology, Stockholm. He held research positions at the Department of Electrical Engineering, Linkoping University, the Information Systems Laboratory, Stanford University, the Katholieke Universiteit Leuven, Leuven, and the University of Luxembourg. During 1996–1997, he was Director of Research at ArrayComm Inc., a start-up in San Jose, CA, based on Ottersten's patented technology. From 1992 to 2004, he was head of the department for Signals, Sensors, and Systems at KTH and from 2004 to 2008, he was Dean of the School of Electrical Engineering, KTH. Currently, he is Director for the Interdisciplinary Centre for Security, Reliability and Trust at the University of Luxembourg. As Digital Champion of Luxembourg, he acts as an adviser to European Commissioner Neelie Kroes. His research interests include security and trust, reliable wireless communications, and statistical signal processing.

Dr. Ottersten coauthored journal papers that received the IEEE Signal Processing Society Best Paper Award in 1993, 2001, and 2006 and three IEEE conference papers receiving Best Paper Awards. He served as Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and on the editorial board of the IEEE SIGNAL PROCESSING MAGAZINE. He is currently editor-in-chief of EURASIP *Signal Processing Journal* and a member of the editorial board of EURASIP *Journal of Applied Signal Processing*. He currently serves on the IEEE Signal Processing Society Board of Governors and is a Fellow of EURASIP. In 2011 he received the IEEE Signal Processing Society Technical Achievement Award. He is a first recipient of the European Research Council advanced research grant.