CONNECTIONS BETWEEN SPARSE ESTIMATION AND ROBUST STATISTICAL LEARNING

Efthymios Tsakonas[*], *Joakim Jaldén* [*], *Nicholas D. Sidiropoulos* [‡], *Bjorn Ottersten* [*†]

[*] ACCESS Linnaeus Centre, Royal Institute of Technology (KTH), Stockholm, Sweden
[†] Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, Luxembourg
[‡] University of Minnesota, Minneapolis (UMN)

ABSTRACT

Recent literature on robust statistical inference suggests that promising outlier rejection schemes can be based on accounting explicitly for sparse gross errors in the modeling, and then relying on compressed sensing ideas to perform the outlier detection. In this paper, we consider two models for recovering a sparse signal from noisy measurements, possibly also contaminated with outliers. The models considered here are a linear regression model, and its natural one-bit counterpart where measurements are additionally quantized to a single bit. Our contributions can be summarized as follows: We start by providing conditions for identification and the Cramér-Rao Lower Bounds (CRLBs) for these two models. Then, focusing on the one-bit model, we derive conditions for consistency of the associated Maximum Likelihood estimator, and show the performance of relevant ℓ_1-based relaxation strategies by comparing against the theoretical CRLB.

Index Terms— Sparsity, robustness, outlier detection, Cramér-Rao lower bounds.

1. INTRODUCTION

Statistical learning of information from available data is of great importance in many signal processing and machine learning application areas, like speech or bioinformatics. Robustness against outliers (grossly corrupted data points that violate the postulated model) is an important requirement in this context, and often critical for successful application of learning methods in practice.

Robust estimators have been widely pursued in the context of linear least squares regression, which is well-known for being very sensitive to outliers. In the linear regression context, estimators based on *robust penalty functions* such as the so-called Huber function [1] and the now ubiquitous ℓ_1-norm are used, and numerous other alternatives such as the least trimmed squares (LTS) and RANSAC estimator have been proposed. The common underlying idea is to perform outlier rejection using penalty functions that discard or downweight large residuals, which are typically associated with outliers. While this is intuitive in the linear regression context, outlier rejection in more complicated nonlinear models is usually far less obvious.

In general, sparsity is linked to statistical inference in two fundamental ways. First, often the signal that one wishes to recover from measurements is sparse (even if high-dimensional) – either naturally, or after projecting it on a proper basis. Second, the number of outlier-contaminated measurements is usually small, hence outliers are sparse. Links between sparsity and robustness against outliers have recently been drawn in [2], and more recently in [3], where it was proposed to introduce a sparse auxiliary vector variable to account for outliers, as a *universal model robustification* strategy.

These ideas are further explored in this paper. In particular, the models considered here are the following. Consider a given sequence of vectors $\{\mathbf{d}_i\}_{i=1}^N$, where $\mathbf{d}_i \in \mathbb{R}^p$. We examine first a linear regression setup which incorporates explicitly additive outliers $\{o_i\}_{i=1}^N$ (one per measurement) and noise $\{e_i\}_{i=1}^N$, thus forming measurements $\{r_i\}_{i=1}^N$ as $r_i = \mathbf{d}_i^{\mathrm{T}}\mathbf{w} + e_i + o_i,\ \forall i$ **(M1)**. However, we shall be primarily interested in the natural one-bit counterpart of (M1), in which measurements are additionally quantized to a single-bit, thus forming binary outcomes $\{y_i\}_{i=1}^N$ as $y_i = \mathrm{sign}(\mathbf{d}_i^{\mathrm{T}}\mathbf{w} + e_i + o_i),\ \forall i$ **(M2)**. We treat $\mathbf{w} \in \mathbb{R}^p$ and the vector $\mathbf{o} \in \mathbb{R}^N$ of variables $\{o_i\}_{i=1}^N$ as sparse deterministic unknowns. Denoting as $||\cdot||_0$ the cardinality of a vector (i.e., the number of non-zeros), we assume that $||\mathbf{w}||_0 \leq \kappa_w$ and $||\mathbf{o}||_0 \leq \kappa_o$, where κ_w and κ_o are fixed known positive integers. Noise variables $\{e_i\}_{i=1}^N$ are assumed i.i.d. Gaussian $\mathcal{N}(0, \sigma^2)$ with known variance. For technical reasons which will become evident next, we also assume that all unknown parameters are bounded, i.e., that there exist positive constants R_w and R_o such that $\mathbf{w} \in \mathcal{B}_w \triangleq \{\mathbf{w} \in \mathbb{R}^p \mid ||\mathbf{w}||_\infty \leq R_w\}$ and similarly $\mathbf{o} \in \mathcal{B}_o \triangleq \{\mathbf{o} \in \mathbb{R}^N \mid ||\mathbf{o}||_\infty \leq R_o\}$.

Under either model **(M1)** or **(M2)**, we are interested in *estimating* $\mathbf{w}$ from available data, as well as *detecting* the measurements which are contaminated with outliers. From this estimation theoretic standpoint, our goal in this paper is to provide model identification conditions and the best achievable mean-square-error (MSE) performance for **(M1)** *and* **(M2)**, by providing the Cramér-Rao Lower Bound (CRLB) under sparsity constraints, building on earlier work on the CRLB computation in constrained parameter estimation [4], [5]. In particular, we describe how the identification and CRLB results from [6] and [5] respectively, apply to **(M1)**, and extend to **(M2)**. Finally, we focus on **(M2)** and show the potential of associated ℓ_1 relaxation strategies, by comparing against the CRLB.

Before presenting the results, it is worth noting some applications where the models are useful. Applications for **(M1)** abound, so we only focus here on briefly presenting a few recent applications for **(M2)**:

Robust Preference Measurement (PM) [7],[8]: In PM, often the objective is to estimate a consumer's response function based on his expressed *choice-based* data, that is, his expressed preferences among given products. A product α is usually represented by a *feature* vector $\mathbf{p}_\alpha \in \mathbb{R}^p$, and the underlying assumption is that the consumer's

The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement number 228044, SSF grant ICA08-0046, and NSF grant AST-1247885.

response is formed as a noisy linear combination of the product's features with weights $\{w_i\}_{i=1}^p$ given by the consumer's *partworth* vector $\mathbf{w}$ [9]. Given an option between two products α and β, a consumer chooses the one that has larger response value. Mathematically, his choice is produced by taking the sign of the difference of the two responses. The goal in choice-based PM is then to estimate the underlying partworth values from such choice-based data. Partworth values can be used to predict future preferences, but are also useful *per se* to the retailer, marketer, or product designer, e.g., for consumer sensitivity analysis. The proposed model **(M2)** is ideally suited for partworth estimation in modern (mostly web-based) PM systems, which may involve products with a very large number of features *and* grossly inconsistent response data due to the presence of outliers. The estimation theoretic approach provides an interesting alternative over other state-of-the-art approaches for choice-based preference analysis, which are primarily based on support-vector-machine (SVM) classifiers (see [7] for additional details).

Robust One-bit Compressed Sensing: The problem of recovering a sparse signal from one-bit quantized noisy linear measurements in AD conversion has gained recent attention by several researchers (see e.g., [10], [11] and references therein). In addition to small noise affecting all the entries of the received signal prior to one-bit quantization, there can also be sign-changes due to *impulsive noise*. This can be modeled as an outlier effect, rendering **(M2)** a suitable candidate model.

2. IDENTIFICATION & CRLB FOR MODEL M1

Consider first the linear regression setup which incorporates explicitly additive outliers. In matrix form, the model is the following

$$\mathbf{r} = \mathbf{D}^{\mathrm{T}}\mathbf{w} + \mathbf{e} + \mathbf{o} \text{ with } ||\mathbf{w}||_0 \leq \kappa_w,\ ||\mathbf{o}||_0 \leq \kappa_o, \quad (1)$$

where $\mathbf{r}$ is the vector that contains the measurements $\{r_i\}_{i=1}^N$ and $\mathbf{e}$ contains the i.i.d. Gaussian noise variables $\{e_i\}_{i=1}^N$. The matrix $\mathbf{D} \triangleq [\mathbf{d}_1, \cdots, \mathbf{d}_N] \in \mathbb{R}^{p\times N}$ is a (typically fat) matrix whose columns comprise the $\mathbf{d}_i$'s.

Due to the presence of the auxiliary variables $\{o_i\}_{i=1}^N$, the estimation problem of $(\mathbf{w}, \mathbf{o})$ is underdetermined; considering however that both $\mathbf{w}$ and $\mathbf{o}$ are sparse, it is important to determine when the estimation task is meaningful, i.e., when the parameters $(\mathbf{w}, \mathbf{o})$ are *identifiable*. Herein, the parameters are said to be identifiable if and only if for two different parameter vectors $(\mathbf{w}_0, \mathbf{o}_0) \neq (\mathbf{w}, \mathbf{o})$ the two corresponding random vectors $\mathbf{r}_0$ and $\mathbf{r}$ are not observationally equivalent, i.e., the distribution of the data conditioned on $(\mathbf{w}_0, \mathbf{o}_0)$ is different than the distribution conditioned on $(\mathbf{w}, \mathbf{o})$. Under this definition, a condition for identification follows from [6]. As in [6], for the matrix $\mathbf{Q} \triangleq \left[\mathbf{D}^{\mathrm{T}}\ \mathbf{I}_{N\times N}\right]$ define $\mathrm{Spark}\,(\mathbf{Q})$ as the minimum number of linearly dependent columns in $\mathbf{Q}$. Then, a *sufficient* condition for identification can be expressed in terms of $\mathrm{Spark}\,(\mathbf{Q})$ and the sum of the cardinality bounds κ_w and κ_o as (see [6])

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

In particular, if (2) is satisfied as $N \to \infty$, the limiting *log-likelihood function* associated with (1) will have a unique global maximum [15]. Note that (2) is only sufficient for identification in our context, and also observe the following: Generating the entries of $\mathbf{D}$ from a continuous distribution yields $\mathrm{Spark}(\mathbf{Q}) = N + 1$, almost surely. This essentially means that in the regime where $\kappa_w << \kappa_o$, one can have almost half of the measurements contaminated with outliers, while still retaining parameter identifiability.

Assuming that the errors $\{e_i\}_{i=1}^N$ in (1) are small and bounded, Candes *et al*. in [12] proposed convex optimization based estimators which, under suitable conditions on $\mathbf{D}^{\mathrm{T}}$, attain an estimation error upper-bounded by a constant times the error obtained had there been no outliers in the data $\{r_i\}_{i=1}^N$. On the other hand, under the Gaussianity assumption on $\mathbf{e}$, the corresponding CRLB on the MSE in the estimation of $\mathbf{w}$ can be derived using the approach in [5]. Specifically, the analysis in [5] shows that the CRLB in sparse linear regression equals the covariance of an *oracle* estimator (i.e., an estimator which assumes to know exactly the non-zero pattern in the unknown sparse vector), provided that the *degree of sparsity* (the number of non-zeros) is known a-priori. The model in (1) has a specifically structured dictionary matrix $\mathbf{Q}$, as well as structure on the non-zero pattern of $(\mathbf{w}, \mathbf{o})$. For this model it turns out that the CRLB at $\mathbf{w}$ depends *only* on $\mathbf{w}$ and the $\mathbf{d}_i$'s corresponding to the outlier-free measurements. We present this formally in Section 3, comparing the CRLB result for **(M1)** with the corresponding result for (**M2**).

3. EXTENSION TO MODEL M2

We now show how the identification result extends to **(M2)** and also present the CRLB. In matrix form, **(M2)** can be expressed as

$$\mathbf{y} = \mathrm{sign}(\mathbf{D}^{\mathrm{T}}\mathbf{w} + \mathbf{e} + \mathbf{o}) \text{ with } ||\mathbf{w}||_0 \leq \kappa_w,\ ||\mathbf{o}||_0 \leq \kappa_o, \quad (3)$$

where $\mathbf{y}$ is the vector containing the binary measurements $\{y_i\}_{i=1}^N$.

One complication arising from the sign operation in (3) as compared to the linear regression model in (1) is the following: When $\sigma^2 \to 0$ in (1), it is well-known that one can recover $\mathbf{w}$ exactly with a finite number of measurements – under suitable conditions on the matrix $\mathbf{D}^{\mathrm{T}}$ – provided that the fraction of corrupted entries of $\mathbf{D}^{\mathrm{T}}\mathbf{w}$ is not too large. In particular, [13] proved that if $\sigma^2 \to 0$ and the corruption $\mathbf{o}$ contains at most a fixed fraction of nonzero entries, then vector $\mathbf{w}$ is the unique solution of the minimum-ℓ_1 approximation problem

$$\underset{\mathbf{w}\in\mathcal{B}_w}{\text{minimize}}\ ||\mathbf{r} - \mathbf{D}^{\mathrm{T}}\mathbf{w}||_1. \quad (4)$$

While such recovery is possible in the linear regression case, in the case of the binary response model given in (3) the true parameter $\mathbf{w}$ cannot be recovered in the absence of noise, even with infinite number of measurements. [1] The scenario completely changes in the presence of noise, however.

Model identification is still possible for (3) in the same statistical sense as in Section 2. In fact, a sufficient condition for identification is (2), the same as in the linear regression case. To see this, let $\mathcal{I}_+$ be the set of indices $\{i \mid y_i = 1\}$, and similarly define $\mathcal{I}_- = \{i \mid y_i = -1\}$. 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

$$\begin{aligned} p_y &= \prod_{i\in\mathcal{I}_+} \Pr\left[\mathbf{d}_i^{\mathrm{T}}\mathbf{w} + o_i \geq e_i\right] \prod_{i\in\mathcal{I}_-} \Pr\left[\mathbf{d}_i^{\mathrm{T}}\mathbf{w} + o_i \leq e_i\right] \\ &= \prod_{i=1}^N \Phi\left(\frac{y_i\mathbf{d}_i^{\mathrm{T}}\mathbf{w} + y_i o_i}{\sigma}\right), \end{aligned}$$

where $\Phi\,(u)$ is the cumulative distribution function of the Gaussian density. The function $\Phi(u)$ is strictly monotonic, therefore the probability distribution of the data will be distinct for a specific parameter $(\mathbf{w}, \mathbf{o})$, as long as the condition in (2) is satisfied.

[1] If the magnitude $||\mathbf{w}||_2$ is provided, $\sigma^2 \to 0$, and $\mathbf{o} = \mathbf{0}$ in (3), lower/upper bounds on the possible reconstruction error of $\mathbf{w}$ were proven in [11], and these bounds go to zero as $N \to \infty$.

Now we present the CRLB result for (3). First of all, there is the Fischer Information Matrix (FIM) [14] for the unconstrained problem, i.e., the FIM for the problem of estimating $(\mathbf{w}, \mathbf{o})$ in (3) without making use of any deterministic prior cardinality constraints. This matrix is the expected value of the Hessian of the *log-likelihood* function

$$l(\mathbf{w}, \mathbf{o}) = \sum_{i=1}^{N} \log \Phi \left(\frac{y_i \mathbf{d}_i^{\mathrm{T}} \mathbf{w} + y_i o_i}{\sigma} \right), \tag{5}$$

where the expectation is taken with respect to the measurement vector $\mathbf{y}$ and the derivatives are taken with respect to $(\mathbf{w}, \mathbf{o})$. The unconstrained FIM is denoted here as $\mathbf{J} \triangleq \mathbb{E}_{\mathbf{y}} \left\{ \nabla^2 l(\mathbf{w}, \mathbf{o}) \right\}$.

Given the sequence $\{\mathbf{d}_i\}_{i=1}^N$, the unconstrained FIM for (3) (see [7] for a derivation) is given by $\mathbf{J} = \mathbf{Q}^{\mathrm{T}} \mathbf{\Lambda} \mathbf{Q}$, where $\mathbf{\Lambda} \in \mathbb{R}^{N \times N}$ is a positive diagonal matrix with elements

$$\begin{aligned} \mathbf{\Lambda}_{ii} = & \frac{1}{2\pi\sigma^2} \exp \left[-\frac{(\mathbf{d}_i^{\mathrm{T}} \mathbf{w} + o_i)^2}{\sigma^2} \right] \times \\ & \left[\Phi^{-1} \left(\frac{\mathbf{d}_i^{\mathrm{T}} \mathbf{w} + o_i}{\sigma} \right) + \Phi^{-1} \left(-\frac{\mathbf{d}_i^{\mathrm{T}} \mathbf{w} + o_i}{\sigma} \right) \right]. \end{aligned} \tag{6}$$

Note that $\mathbf{J}$ is singular because $\mathbf{Q}$ is fat by construction, so the unconstrained CRLB does not exist for the problem at hand. However, we are interested in the constrained CRLB, i.e., the CRLB for points known to obey the cardinality constraints in (3). The CRLB for such points typically exists. As pointed out in [5], albeit non-smooth, the set of vectors with restricted cardinality is *locally balanced*, meaning that it can be described locally (at any point) by a feasible subspace $\mathcal{F}$. In fact, one can assign at any given point with restricted cardinality a matrix of feasible directions $\mathbf{U}$, whose columns span $\mathcal{F}$. In the case of our model (3), we associate to each point $(\mathbf{w}, \mathbf{o})$ a feasible subspace spanned by

$$\mathbf{U} = \begin{bmatrix} \mathbf{U}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_o \end{bmatrix}, \tag{7}$$

where $\mathbf{U}_o$ is either the subset of columns of the identity matrix $\mathbf{I}_{N \times N}$ corresponding to the non-zero pattern of $\mathbf{o}$ if $||\mathbf{o}||_0 = \kappa_o$, or precisely the identity matrix $\mathbf{I}_{N \times N}$ if $||\mathbf{o}||_0 < \kappa_o$. Matrix $\mathbf{U}_s$ is determined in the same way. Once such description is found, the value of the constrained CRLB at point $(\mathbf{w}, \mathbf{o})$ depends *only* on the unconstrained FIM and the matrix of feasible directions $\mathbf{U}$ evaluated at the point $(\mathbf{w}, \mathbf{o})$, as the following lemma asserts [5]:

Lemma 1. *Let $\mathcal{R}(\mathbf{U})$ denote the range space of the matrix of feasible directions $\mathbf{U}$. If the condition $\mathcal{R}(\mathbf{U}\mathbf{U}^{\mathrm{T}}) \subseteq \mathcal{R}(\mathbf{U}\mathbf{U}^{\mathrm{T}}\mathbf{J}\mathbf{U}\mathbf{U}^{\mathrm{T}})$ holds, the covariance of any unbiased estimator for the point $(\mathbf{w}, \mathbf{o})$ satisfies $\mathrm{Cov}(\hat{\mathbf{w}}, \hat{\mathbf{o}}) \succeq \mathbf{U}(\mathbf{U}^{\mathrm{T}}\mathbf{J}\mathbf{U})^{\dagger}\mathbf{U}^{\mathrm{T}}$. Conversely, if the above condition does not hold, there is no* unbiased *finite-variance estimator for $(\mathbf{w}, \mathbf{o})$.*

Lemma 1 actually translates (after straightforward manipulations) to our model (3) in the following way (see [7] for the details – here we omit the proof due to lack of space):

Theorem 1. *Consider the estimation problem in (3) and assume that (2) holds. The CRLB on the MSE $\mathcal{E}(\mathbf{w}) \triangleq \mathbb{E}||\hat{\mathbf{w}} - \mathbf{w}||_2^2$ of any unbiased estimator $\hat{\mathbf{w}}$ for point $\mathbf{w}$ is given as follows*

$$\mathcal{E}(\mathbf{w}) \geq \mathrm{Tr}\left(\mathbf{U}_s^{\mathrm{T}} \mathbf{D}\mathbf{L}\mathbf{D}^{\mathrm{T}} \mathbf{U}_s \right)^{-1} \text{ when } ||\mathbf{w}||_0 = \kappa_w, ||\mathbf{o}||_0 = \kappa_o$$

$$\mathcal{E}(\mathbf{w}) \geq \mathrm{Tr}\left(\mathbf{D}\mathbf{L}\mathbf{D}^{\mathrm{T}} \right)^{\dagger} \text{ when } ||\mathbf{w}||_0 < \kappa_w, ||\mathbf{o}||_0 = \kappa_o,$$

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 $||\mathbf{o}||_0 < \kappa_o$.

Observe that the CRLB depends only on the outlier-free measurements and the columns of $\mathbf{D}^{\mathrm{T}}$ corresponding to the non-zero pattern of $\mathbf{w}$. Finally, we remark that the corresponding CRLB for the linear regression model in (1) is very similar to that above. In particular, using the above theorem and choosing $\mathbf{L}_{ii} = 1/\sigma^2$ if $\mathbf{o}_i = 0$ and $\mathbf{L}_{ii} = 0$ if $\mathbf{o}_i \neq 0$ recovers the CRLB result for (1). Theorem 1 allows one to gauge the MSE performance of the associated ML estimator for (3), showing also the performance capacity of relevant ℓ_1 relaxation strategies in this context.

4. ML ESTIMATOR PROPERTIES

Joint ML estimation of $(\mathbf{w}, \mathbf{o})$ in (3) amounts to solving the optimization problem

$$\underset{\mathbf{w} \in \mathcal{B}_w, \mathbf{o} \in \mathcal{B}_o}{\text{maximize}} \; l(\mathbf{w}, \mathbf{o}) \text{ subject to: } ||\mathbf{w}||_0 \leq \kappa_w, \; ||\mathbf{o}||_0 \leq \kappa_o, \tag{8}$$

a formulation reminiscent of the form of the ML estimator for the probit model [15]. Essentially, our work in this paper can be viewed as a natural robustification of such models against outliers (grossly corrupted data points) and/or datasets with a very large number of variables in $\mathbf{w}$ (necessitating variable selection to obtain meaningful estimates).

Each summand $\log \Phi(u)$ in $l(\mathbf{w}, \mathbf{o})$ in (5) is concave, increasing in u and tends to zero as $u \to \infty$, therefore the bounding boxes $\mathcal{B}_w$ and $\mathcal{B}_o$ ensure that (8) always has a maximizer. Now, an interesting question is whether the estimate $\hat{\mathbf{w}}$ provided by (8) is *consistent*, i.e., whether (8) yields the true vector of partworths as a solution in the limit. We treat this topic next, providing a sufficient condition for consistency:

Theorem 2. *Consider the estimation problem in (3) with unknown parameters $(\mathbf{w}_0, \mathbf{o}_0)$, $\sigma^2 > 0$, and assume that $\{\mathbf{d}_i\}_{i=1}^N$ are samples from an underlying probability distribution and satisfy the identifiability condition in (2). The ML estimate in (8) converges in probability to $\mathbf{w}_0$ as $N \to \infty$, assuming that the number of outlier-contaminated measurements increases sublinearly with N, i.e., that $\lim_{N \to \infty} \kappa_o / N = 0$.*

Proof. Given in [7]. □

On the practical side, it is clear that the ML estimation problem in (8) 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. Although this direct enumeration approach yields a consistent estimator, it is unfortunately often computationally intractable. Instead, one may formulate a tractable approximation to (8) by replacing the cardinality constraints in (8) with convex ℓ_1-norm constraints. This is motivated since the ℓ_1-norm is the tightest convex relaxation of the cardinality function [17]. Making this ℓ_1 replacement and penalizing the constraints at the objective yields the convex optimization problem

$$\underset{\mathbf{w}, \mathbf{o}}{\text{minimize}} \; \phi(\mathbf{w}, \mathbf{o}) = -l(\mathbf{w}, \mathbf{o}) + \lambda_w ||\mathbf{w}||_1 + \lambda_o ||\mathbf{o}||_1 \tag{9}$$

with fixed positive regularization parameters λ_w and λ_o. These control the trade-off between the value of $l(\mathbf{w}, \mathbf{o})$ and the number of non-zero elements of $\mathbf{w}$ and $\mathbf{o}$ respectively. These parameters are most often tuned in a heuristic fashion: One starts from a suitable

initial point $\left(\lambda_w^i,\ \lambda_o^i\right)$ and iterates until the desired sparsity/fit tradeoff is achieved. The convex alternative in (9) can be solved efficiently using a variety of polynomial-time algorithms, including interior point methods [17]. In particular, the work in [16] deals with distributed solution strategies for solving (9), which mainly become interesting when the number of observed samples N becomes exceedingly large.

5. NUMERICAL EXPERIMENTS

In our experiments, we benchmark the MSE performance of different MLE variants for (3), against the corresponding CRLBs. For the MSE comparison, the $\mathbf{d}_i$'s were generated as i.i.d. Gaussian vectors drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, each comprising $p = 20$ elements. The unknown partworth vector $\mathbf{w}_0$ was assumed sparse i.i.d. Gaussian with $\mathrm{NZ} = 3$ non-zero elements drawn from $\mathcal{N}(0, 1)$. The MSE of all tested estimators was evaluated using $\mathrm{MC} = 300$ Monte Carlo trials. For each trial, binary data were generated according to model **(M2)** in (3). The outlier percentage considered in the measurements was 1% (outliers correspond to uniform at random sign flips). The additive noise variables e_i's were assumed i.i.d. drawn from $\mathcal{N}(0, 1)$. In each trial, instead of solving (8) directly by enumerating all possible sparsity patterns for $(\mathbf{w}, \mathbf{o})$, all estimators obtain the estimate for $(\mathbf{w}, \mathbf{o})$ through relaxation. In particular, we first (i) solve the problem in (9) to obtain a plausible sparsity pattern for $(\mathbf{w}, \mathbf{o})$, and then (ii) we re-solve the problem having the sparsity pattern in $(\mathbf{w}, \mathbf{o})$ fixed.

The MLE variants tested here are the following: MLE-NPS performs outlier rejection but does not use the prior-information that $\mathbf{w}_0$ is sparse. This variant solves (9) using $\lambda_w = 0$ and $\lambda_o = 0.1||\nabla_o l(0, 0)||_\infty$ as initial regularization parameter values, and then iterates with respect to λ_o to achieve the desired sparsity level in the estimate $\hat{\mathbf{o}}$. MLE-NOR accounts for $\mathbf{w}$-sparsity but does not account for outliers, solving (9) using $\lambda_o = 0$, and $\lambda_w = 0.1||\nabla_w l(0, 0)||_\infty$ for initialization. MLE-PSOR performs simultaneously outlier rejection and also accounts for $\mathbf{w}$-sparsity using $\lambda_w = 0.1||\nabla_w l(0, 0)||_\infty$ and $\lambda_o = 0.1||\nabla_o l(0, 0)||_\infty$ as initial values for the regularization parameters. The (Root)-MSE results are depicted in Figure 1, where two additional CRLB curves are plotted as functions of the number of samples N. CRLB-PSOR is the CRLB of any unbiased estimator knowing the *number of outliers* and the fact 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 $\mathbf{w}$ can affect the expected estimation accuracy. One expects that the effect of the prior information regarding $\mathbf{w}$-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 slow. On the other hand, note that the MLE variant which does not perform outlier rejection fails miserably, while the other two estimators operate close to their respective CRLBs. This speaks for a key strength of the formulation in (3): The ability to detect and remove outliers efficiently.

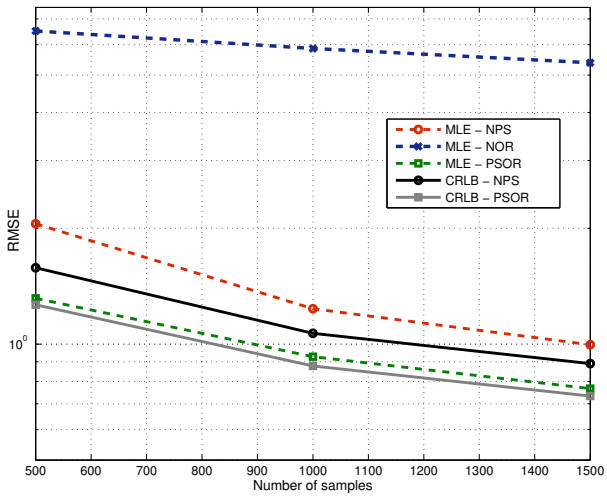

Fig. 1. RMSE comparison of the different MLE variants against their respective CRLBs for different sample sizes N. All estimators were implemented using the two-step procedure described in the text.

6. REFERENCES

[1] P. Huber, *Robust statistics*, John Wiley & Sons, 1981.

[2] Y. Jin and B. Rao, "Algorithms for robust linear regression by exploiting the connection to sparse signal recovery," in *Proc. of Intl. Conf. on Acoust., Speech, and Sig. Processing (ICASSP)*, Dallas, March 2010.

[3] G. 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 Sig. Processing (ICASSP)*, Prague, Czech Republic, May 2011.

[4] P. Stoica and B. C. Ng, "On the Cramér-Rao bound under parametric constraints," in *IEEE Sig. Process. Lett.*, vol. 5, no. 7, pp. 177-179, 1998.

[5] Z. Ben-Haim and Y. Eldar, "The Cramér-Rao Bound for estimating a sparse parameter vector," in *IEEE Trans. on Sig. Processing*, vol. 58, no. 6, June 2010.

[6] D. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," in *Proc. Natl. Acad. Sci.*, USA, 100:2197-2202, 2003.

[7] E. Tsakonas, J. Jaldén, N. Sidiropoulos, B. Ottersten, "Sparse conjoint analysis through maximum likelihood estimation," submitted to *IEEE Trans. on Sig. Processing*, Oct. 2012.

[8] G. Mateos and G. Giannakis, "Robust conjoint analysis by controlling outlier sparsity," in *Proc. of European Signal Processing Conference*, Barcelona, Spain, Aug. 29- Sep. 2, 2011.

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

[10] J. Haupt and R. Baraniuk, "Robust support recovery using sparse compressive sensing matrices" in *Proc. 45th Annual Conf. on Information Sciences and Systems*, Baltimore, MD, March 2011, pp. 16.

[11] L. Jacques, J. Laska, P. Boufounos and R. Baraniuk, "Robust 1-Bit Compressive Sensing via Binary Stable Embeddings of sparse vectors," available at *http://arxiv.org/abs/1104.3160v2*, February, 2012.

[12] E. Candes and P. Randall, "Highly robust error correction by convex programming," in *IEEE Trans. on Information Theory*, vol. 54, no. 7, pp. 28292840, 2008.

[13] E. Candes and T. Tao, "Decoding by linear programming," in *IEEE Trans. on Information Theory*, 51(12):4203- 4215, December 2005.

[14] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[15] W. Newey and D. McFadden, "Chapter 35: Large sample estimation and hypothesis testing," in *Handbook of Econometrics*, Vol. 4, Elsevier Science, pp. 2111–2245.

[16] E. Tsakonas, J. Jaldén, N. Sidiropoulos, B. Ottersten, "Maximum likelihood based sparse and distributed conjoint analysis," in *Statistical Signal Processing Workshop (SSP)*, Ann Arbor, USA, Aug. 5-8, 2012.

[17] S. Boyd, and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.