USPACOR: UNIVERSAL SPARSITY-CONTROLLING OUTLIER REJECTION*

G. B. Giannakis, G. Mateos, S. Farahmand, V. Kekatos, and H. Zhu

Dept. of ECE, Univ. of Minnesota, 200 Union Street SE, Minneapolis, MN 55455, USA

ABSTRACT

The recent upsurge of research toward compressive sampling and parsimonious signal representations hinges on signals being sparse, either naturally, or, after projecting them on a proper basis. The present paper introduces a neat link between sparsity and a fundamental aspect of statistical inference, namely that of robustness against outliers, even when the signals involved are not sparse. It is argued that controlling sparsity of model residuals leads to statistical learning algorithms that are computationally affordable and universally robust to outlier models. Analysis, comparisons, and corroborating simulations focus on robustifying linear regression, but succinct overview of other areas is provided to highlight universality of the novel framework.

***Index Terms*—** Robustness, outlier rejection, sparsity, Lasso

1. INTRODUCTION

The information explosion propelled by the advent of computers, the Internet, and the global-scale communications has rendered *statistical learning* from data increasingly important for analysis and processing. Along with data that adhere to postulated models (inliers), present in large volumes of data are also those that do not (outliers). Resilience to outliers is of paramount importance in a plethora of tasks such as model selection, prediction, classification, estimation and tracking, to name a few. Due to its universal applicability, the method of least-squares (LS) is the workhorse of statistical learning. Unfortunately, LS is known to be very sensitive to outliers [9, 14].

Robust alternatives to LS include the M-estimators, which are maximum-likelihood (ML) optimal for a class of outlier models [9]. Other options are least-trimmed squares (LTS) estimators, which remove outliers from the LS fit [14]. LTS estimators have high breakdown point, but prohibitive complexity except for small sample sizes [13]. Random sample consensus (RANSAC) provides a computationally tractable, near-LTS alternative, especially popular in computer vision for coping with a large number of outliers [4, 7].

A universal sparsity-controlling outlier rejection (USPACOR) framework is introduced in this paper for robust learning. USPACOR is rooted at the crossroads of outlier-resilient estimation, the least-absolute shrinkage and selection operator (Lasso) for sparse regression, and convex optimization. It is shown that a sparsity-tuning parameter (λ_1) in Lasso controls the *degree of sparsity* in the estimator, and the *number of outliers* rejected by USPACOR.

Related approaches for robust linear regression can be found in [6, 10, 11]. The major difference is that λ_1 in these works is tied to a preselected outlier model, whereas here it is dictated by the data. This promotes universality and a systematic approach leveraging solvers for all *robustifaction paths* of Lasso; that is, for all values of λ_1 [2, 5, 17]. In this sense, USPACOR capitalizes on but *is not limited to* sparse settings (few outliers), since one can examine the gamut of sparsity levels along the robustification path. Due to space limitations, USPACOR is detailed only for linear regression. But its universality is highlighted through diverse generalizations pertaining to: i) the information used for selecting λ_1; ii) the inlier model; and iii) the criterion adopted to fit the chosen model. Simulated tests demonstrate that USPACOR outperforms RANSAC in a linear regression setup, especially when the percentage of outliers is high.

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

2. SPARSITY CONTROL FOR ROBUSTNESS

2.1. Robustifying linear regression

Consider the classical regression setup, where a real-valued scalar response y is to be predicted using p known variables (inputs) collected in the vector $\mathbf{x} := [x_1, \ldots, x_p]' \in \mathbb{R}^p$ ($'$ stands for transposition). A linear approximation of the mean-square error (MSE) optimal regression function $\mathbb{E}[y|\mathbf{x}]$ is $f(\mathbf{x}) = \mathbf{x}'\boldsymbol{\theta}$, where $\boldsymbol{\theta} := [\theta_1, \ldots, \theta_p]' \in \mathbb{R}^p$ comprises the regression coefficients.

Given a set $\mathcal{T} := \{y_i, \mathbf{x}_i\}_{i=1}^N$ of training data possibly contaminated with outliers, and supposing $\mathbf{X} := [\mathbf{x}_1, \ldots, \mathbf{x}_N]'$ has full column rank for simplicity, the goal is to develop a robust estimator of $\boldsymbol{\theta}$ that is universal with respect to the outlier model. The LTS estimator is universal in this sense, and is given by [14]

$$\hat{\boldsymbol{\theta}}_{LTS} := \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{s} r^2_{[i]}(\boldsymbol{\theta}) \tag{1}$$

where $r^2_{[i]}(\boldsymbol{\theta})$ is the i-th order statistic among the squared residuals $r_1^2(\boldsymbol{\theta}), \ldots, r_N^2(\boldsymbol{\theta})$, and $r_i(\boldsymbol{\theta}) := y_i - \mathbf{x}_i'\boldsymbol{\theta}$. The so-termed *coverage* s determines the breakdown point of LTS [14], since $N - s$ residuals are not present in (1). Even though (1) is nonconvex, existence of a minimizer $\hat{\boldsymbol{\theta}}_{LTS}$ can be established as follows: i) for each subset of $\mathcal{T}$ with cardinality s (there are $\binom{N}{s}$ such subsets), solve the corresponding LS problem to obtain a candidate estimator per subset; and ii) pick $\hat{\boldsymbol{\theta}}_{LTS}$ as the one among all $\binom{N}{s}$ candidates with the least cost. This solution procedure is combinatorially complex, and thus intractable except for small sample sizes N. Algorithms to obtain approximate LTS solutions are available; see e.g., [13].

Instead of discarding large residuals, the alternative approach here explicitly accounts for outliers in the regression model. To this end, consider the scalar variables $\{o_i\}_{i=1}^N$ one per training data point, which take the value $o_i = 0$ whenever datum i is an inlier, and $o_i \neq 0$ otherwise. This leads to the linear regression model

$$y_i = \mathbf{x}_i'\boldsymbol{\theta} + o_i + \varepsilon_i, \qquad i = 1, \ldots, N \tag{2}$$

where $\{\varepsilon_i\}_{i=1}^N$ are zero-mean i.i.d. random variables capturing inlier errors, while o_i can be deterministic or random with unspecified distribution. In the under-determined linear system of equations (2), both $\boldsymbol{\theta}$ as well as the $N \times 1$ vector $\mathbf{o} := [o_1, \ldots, o_N]'$ are unknown. The percentage of outliers dictates the degree of *sparsity* (number

of zero entries) in $\mathbf{o}$. Sparsity control will prove instrumental in efficiently estimating $\mathbf{o}$, rejecting outliers as a byproduct, and consequently arriving at a *robust* estimator of $\boldsymbol{\theta}$. A natural criterion for controlling outlier sparsity is to seek an estimator which solves

$$\min_{\boldsymbol{\theta},\mathbf{o}} \sum_{i=1}^{N} (y_i - \mathbf{x}_i'\boldsymbol{\theta} - o_i)^2 + \lambda_0 \|\mathbf{o}\|_0 \tag{3}$$

where $\|\mathbf{o}\|_0$ denotes the nonconvex ℓ_0-(pseudo)norm that is equal to the number of nonzero entries of $\mathbf{o}$. Sparsity in $\hat{\mathbf{o}}$ can be directly controlled by tuning the parameter $\lambda_0 \geq 0$.

As with compressive sampling and sparse modeling schemes that rely on the ℓ_0-norm [16], problem (3) is also NP-hard. In addition, the sparsity-controlling estimator (3) is intimately related to LTS, as asserted next (proofs are omitted due to space limitations).

Proposition 1: *If* $\{\hat{\boldsymbol{\theta}}, \hat{\mathbf{o}}\}$ *minimizes* (3) *with* λ_0 *chosen such that* $\|\hat{\mathbf{o}}\|_0 = N - s$*, then* $\hat{\boldsymbol{\theta}}$ *also solves* (1).

The importance of Proposition 1 is threefold. First, it formally justifies model (2) and its estimator (3) for robust linear regression, in light of the well documented merits of LTS [14]. Second, it further solidifies the connection between sparse linear regression and robust estimation. Third, problem (3) lends itself naturally to efficient (approximate) solvers based on convex relaxation. For instance, recall that the ℓ_1 norm $\|\mathbf{o}\|_1 := \sum_{i=1}^{p} |o_i|$ is the closest convex approximation of $\|\mathbf{o}\|_0$. This property also utilized by compressive sampling [16], provides the motivation to relax (3) to

$$\min_{\boldsymbol{\theta},\mathbf{o}} \sum_{i=1}^{N} (y_i - \mathbf{x}_i'\boldsymbol{\theta} - o_i)^2 + \lambda_1 \|\mathbf{o}\|_1 . \tag{4}$$

Being a (nondifferentiable) convex optimization problem, (4) can be efficiently solved by, e.g., resorting to an alternating minimization algorithm. The resulting iterations comprise a sequence of LS fits for $\boldsymbol{\theta}$, and coordinatewise soft-thresholded updates for $\mathbf{o}$. Alternatively, one can show that the solutions $\{\hat{\boldsymbol{\theta}}, \hat{\mathbf{o}}\}$ of (4) are respectively given by $\hat{\boldsymbol{\theta}} := \mathbf{X}^{\dagger}(\mathbf{y} - \hat{\mathbf{o}}_{\text{Lasso}})$ and $\hat{\mathbf{o}} := \hat{\mathbf{o}}_{\text{Lasso}}$, where $\mathbf{X}^{\dagger} := (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\hat{\mathbf{o}}_{\text{Lasso}}$ is given by

$$\hat{\mathbf{o}}_{\text{Lasso}} := \arg\min_{\mathbf{o}} \|(\mathbf{I}_N - \mathbf{X}\mathbf{X}^{\dagger})(\mathbf{y} - \mathbf{o})\|_2^2 + \lambda_1 \|\mathbf{o}\|_1 . \tag{5}$$

It is worth stressing at this point that selecting λ_1 is challenging because existing techniques such as cross-validation (CV) do not apply when outliers are present. USPACOR includes a general and systematic approach to selecting λ_1 by leveraging recent convex optimization solvers that yield the entire path of Lasso solutions, i.e., for all values of λ_1 in (5) [2, 5]. Based on these *robustification paths* and prior knowledge possibly available on the model (2), one can effectively select λ_1 – the subject dealt with in the next section.

Remark 1. The estimator obtained from (4) can be robust in the Huber sense [6]. However, this only holds for a specific choice of λ_1. The last point appears mundane, but is at the heart of the USPACOR novelty, since tuning λ_1 is tantamount to controlling the number of outliers rejected.

2.2. Selecting outlier sparsity

The ensuing methods for choosing λ_1 depend on prior information available about the inliers or the outliers (number or statistics).

Number of outliers is known. By direct inspection of the robustification paths one can determine the range of values for λ_1, so that the degree of sparsity in $\hat{\mathbf{o}}$ equals the number of outliers N_o. Specializing to the interval of interest, and after discarding the identified outliers, K-fold CV methods can be applied to determine the "best" λ_1^*. Note that N_o is also assumed known by RANSAC, in order to determine the number of random draws needed to attain a prescribed probability of success [4, 7].

Variance of the inlier noise is known. If the variance σ_ε^2 of the inlier noise ε_i in (2) is known, one can proceed as follows. Consider the estimates $\hat{\boldsymbol{\theta}}_g$ obtained using (4) and (5) after sampling the robustification path for each point $\{\lambda_g\}_{g=1}^G$ on a prescribed grid of size G. Based on $\{\hat{\boldsymbol{\theta}}_g\}_{g=1}^G$ and the data $\mathcal{T}$, find the sample variances $\{\hat{\sigma}_g^2\}_{g=1}^G$ after neglecting those training data $\{y_i, \mathbf{x}_i\}$ identified as outliers. The winner $\lambda_1^* := \lambda_{g^*}$ corresponds to the grid point

$$g^* := \arg\min_g |\hat{\sigma}_g^2 - \sigma_\varepsilon^2| . \tag{6}$$

This is an absolute variance deviation (AVD) criterion for selecting λ_1^*. Knowledge of σ_ε^2 is also required by RANSAC; see also Sec. 5.

Variance of the inlier noise is unknown. If σ_ε^2 is unknown, one can still compute a robust estimate of the variance $\hat{\sigma}_\varepsilon^2$, and repeat the previous procedure after replacing σ_ε^2 with $\hat{\sigma}_\varepsilon^2$ in (6). One simple option is based on the median absolute deviation (MAD) estimator, where $\hat{\sigma}_\varepsilon := 1.48 \times \text{median}_i \left(|\hat{r}_i - \text{median}_j \left(|\hat{r}_j|\right)|\right)$. The residuals $\hat{r}_i$ are formed based on a nonrobust estimate of $\boldsymbol{\theta}$, e.g., obtained via an LS fit using a small subset of the training data $\mathcal{T}$. The factor 1.48 provides an approximately unbiased estimate of σ_ε, when the inlier noise is Gaussian. In general, MAD requires knowledge of ε_i's symmetric pdf to determine the leading factor in $\hat{\sigma}_\varepsilon$ [14].

Contamination model. One may know a priori that the disturbances $\{o_i + \varepsilon_i\}$ in (2) adhere to Huber's contamination model [9]. Here ε_i can be thought of as nominal noise, and o_i as the contamination. If in this case λ_1 equals the threshold value in Huber's function, then $\hat{\boldsymbol{\theta}}$ enjoys asymptotic optimality in a well defined minimax sense [6].

Bayesian framework. Adopting a Bayesian perspective, one could model $\boldsymbol{\theta}$ as having i.i.d. entries obeying a non-informative (i.e., uniform) prior, independent of $\mathbf{o}$, which is assumed to have i.i.d. entries adhering to a common Laplacian distribution with parameter $2/\lambda_1^*$. Using $\lambda_1 = \lambda_1^*$ in (4), USPACOR yields estimates $\hat{\boldsymbol{\theta}}$ (and $\hat{\mathbf{o}}$) which are optimal in the maximum a posteriori sense; see also [10].

Building on (4), it is possible to envision a number of interesting generalizations beyond linear regression, which further justify the *universality* of the proposed USPACOR framework. These pertain to the: i) models adopted for the inliers; ii) loss functions chosen to penalize the fitting errors; and iii) regularization terms for $\boldsymbol{\theta}$ and $\mathbf{o}$.

3. UNIVERSALITY WITH RESPECT TO MODELS

This section shows how the USPACOR approach generalizes to models other than linear time-invariant regression in (2).

Errors-in-variables (EIV) and total least-squares (TLS). TLS extends ordinary LS to fully-perturbed linear models, such as the EIV one; see e.g., [12]. With $\bar{\mathbf{S}}$ denoting the sample covariance of the data vectors $\{[\mathbf{x}_i' \; y_i]'\}_{i=1}^N$, the TLS estimator corresponds to the eigenvector associated with the smallest eigenvalue of $\bar{\mathbf{S}}$. As such, TLS performs "orthogonal regression," which minimizes the sum of squared *orthogonal* distances from $[\mathbf{x}_i' \; y_i]'$ to the fitting hyperplane, as opposed to the *vertical* distance minimized by LS [12]. To robustify TLS against outliers, USPACOR can be applied to yield the desired robust estimator $\hat{\boldsymbol{\theta}}$ as solution of

$$\min_{\boldsymbol{\theta},\mathbf{o}} \sum_{i=1}^{N} \frac{(y_i - \mathbf{x}_i'\boldsymbol{\theta} - o_i)^2}{1 + \|\boldsymbol{\theta}\|_2^2} + \lambda_1 \|\mathbf{o}\|_1 . \tag{7}$$

Alternating minimization between variables $\boldsymbol{\theta}$ and $\mathbf{o}$ can converge to a stationary point of this nonconvex criterion. Each sub-problem per iteration reduces to either TLS or a scalar Lasso, and in both cases the solutions admit analytical forms.

Dynamical models for recursive (R)LS and Kalman smoothing. RLS schemes are of paramount importance for reducing complexity and memory requirements in estimating stationary signals as well as for tracking slowly varying processes, when no model is available for the variations and quadratic convergence is desired. Similar to LS, the quadratic cost minimized online by RLS is not robust against outliers. With data (2) becoming available sequentially, USPACOR can estimate outliers online and apply RLS to the outlier-compensated data $y_i - \hat{o}_i$. Specifically, at time $i = N$ it solves

$$\min_{\boldsymbol{\theta},\mathbf{o}} \sum_{i=1}^{N} \tau^{N-i} \left[(y_i - \mathbf{x}_i'\boldsymbol{\theta} - o_i)^2 + \lambda_1 |o_i| \right]$$

where $\tau \in (0, 1]$ denotes the forgetting factor. Since the cost here is convex, it can be solved using, e.g., coordinate descent (CD) [5].

The USPACOR approach can be tailored also for Kalman filtering and smoothing, when the time-varying parameters sought obey a model. The major novelty here is USPACOR's ability to cope with outliers present not only in the measurements but also in the state equation (the latter capture unmodeled dynamics of e.g., abrupt target maneuvering). To outline this doubly-robust approach over a smoothing horizon $i = 1, \ldots, N$, consider the state space model $\boldsymbol{\theta}_i = \mathbf{F}_i\boldsymbol{\theta}_{i-1} + \mathbf{o}_{\theta,i} + \mathbf{w}_i$, where $\mathbf{F}_i$ denotes the known state transition matrix, $\mathbf{w}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_i)$ the Gaussian process noise, $\boldsymbol{\theta}_0 \sim \mathcal{N}(\mathbf{m}_0, \boldsymbol{\Sigma}_0)$ the Gaussian initial state, and $\mathbf{o}_{\theta,i}$ ($\mathbf{o}_{y,i}$) the state (measurement) outliers. Extending (2) to the vector case yields the measurement equation $\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\theta}_i + \mathbf{o}_{y,i} + \boldsymbol{\varepsilon}_i$, where $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_i)$. The doubly-robust smooth estimate $\hat{\boldsymbol{\theta}} := [\hat{\boldsymbol{\theta}}_0', \ldots, \hat{\boldsymbol{\theta}}_N']'$ is given by

$$\min_{\boldsymbol{\theta},\mathbf{o}_\theta,\mathbf{o}_y} \sum_{i=1}^{N} \left[\|\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\theta}_i - \mathbf{o}_{y,i}\|^2_{\mathbf{R}_i^{-1}} + \|\boldsymbol{\theta}_i - \mathbf{F}_i\boldsymbol{\theta}_{i-1} - \mathbf{o}_{\theta,i}\|^2_{\mathbf{Q}_i^{-1}} \right] + \|\boldsymbol{\theta}_0 - \mathbf{m}_0\|^2_{\boldsymbol{\Sigma}_0^{-1}} + \sum_{i=1}^{N} \left[\lambda_{1,\theta}\|\mathbf{o}_{\theta,i}\|_1 + \lambda_{1,y}\|\mathbf{o}_{y,i}\|_1 \right]. \quad (8)$$

where $\|\mathbf{x}\|_{\mathbf{A}} := \mathbf{x}'\mathbf{A}\mathbf{x}$ for a positive definite matrix $\mathbf{A}$. Again, (8) can be solved via alternating minimization, and $\lambda_{1,\theta}$, $\lambda_{1,y}$ can be chosen along the lines outlined in Sec. 2.2.

Generalized linear models (GLM). The MSE-optimal regression function $\mathbb{E}[y|\mathbf{x}]$ is modeled here by the so-termed activation function $f(\mathbf{x}'\boldsymbol{\theta})$. A special case popular for (say binary) classification leads to logistic regression, where $f(u) := \left(1 + e^{-u}\right)^{-1}$, and y_i equals 1 when $\mathbf{x}_i$ belongs to the first class, and 0 otherwise [8, p. 119]. To robustify logistic regression USPACOR estimates $\boldsymbol{\theta}$ by

$$\min_{\boldsymbol{\theta},\mathbf{o}} -\sum_{i=1}^{N} y_i \log z_i + (1 - y_i)\log(1 - z_i) + \lambda\|\mathbf{o}\|_1 \quad (9)$$

where $z_i := f(\mathbf{x}_i'\boldsymbol{\theta} + o_i)$. Problem (9) is convex and can be efficiently solved by reweighted LS iterations [8, p. 120]. The result can be extended readily to: i) multiclass classification; and ii) probit regression, where $f(u)$ is replaced by the standard Gaussian cumulative distribution function.

Nonparametric (kernel) regression. Nonparametric regression is widely applicable to statistical learning problems, since it only assumes that the regression function f belongs to a (possibly infinite dimensional) space of e.g., "smooth" functions $\mathcal{H}$. As estimating $f \in \mathcal{H}$ from finite data is inherently ill-posed, the problem is typically solved by minimizing appropriately regularized criteria; see e.g. [8, p. 167]. USPACOR can be extended to this nonparametric context, to yield the desired robust estimate $\hat{f}$ as solution of

$$\min_{f\in\mathcal{H},\mathbf{o}} \sum_{i=1}^{N} (y_i - f(\mathbf{x}_i) - o_i)^2 + \mu\|f\|^2_{\mathcal{H}} + \lambda_1\|\mathbf{o}\|_1 \quad (10)$$

where $\mu \geq 0$ is chosen to tradeoff fidelity (to the outlier compensated) data for the degree of "smoothness" measured by $\|f\|^2_{\mathcal{H}}$. Interestingly, it can be shown that when $\mathcal{H}$ has the structure of a reproducing kernel Hilbert space, it suffices to solve a particular instance of Lasso as in (5), in order to obtain $\hat{f}$ in (10).

4. UNIVERSALITY WITH RESPECT TO CRITERIA

This section shows how flexible USPACOR is to encompass a number of criteria suitable for various statistical inference tasks.

4.1. Loss functions

Problem (4) relies on a square loss function $V(u) = u^2$ of the fitting errors $\{y_i - \mathbf{x}_i'\boldsymbol{\theta} - o_i\}_{i=1}^N$. If the inlier noise distribution is non-Gaussian and known, ML or MAP loss functions can replace the LS cost. Adopting $V(u) = |u|$ for instance, gives rise to ℓ_1 regression that is robust and enjoys ML optimality for Laplacian distributed (inlier) noise. In addition, USPACOR can be endowed with an inner layer of robustness by choosing V as Huber's function [9]. Alternatively, use of an ϵ-insensitive loss function $V(u) := \max(0, |u| - \epsilon)$ links USPACOR with robust support vector machine formulations. Upon departing from a square loss, Lasso can no longer be employed in the alternating minimization process.

Nonconvex loss functions could be of interest as well, such as the $\boldsymbol{\theta}$-dependent weighted loss arising with USPACOR-based TLS formulations [cf. (7)].

4.2. Regularization terms

Concave functions such as the SCAD penalty [3], or the sum-of-logs regularizer in [1, 11], can approximate better $\|\mathbf{o}\|_0$ in (3) but lead to nonconvex cost functions with multiple local minima. However, when initialized properly, e.g., with the USPACOR solution of (4), they typically provide considerable improvements after a few iterations. Noting that $\lambda_1\|\mathbf{o}\|_1$ biases $\hat{\mathbf{o}}$ towards zero, the performance gains due to nonconvex regularizers can be leveraged to bias reduction [3]. An appealing *convex* alternative is the weighted ℓ_1 norm of $\mathbf{o}$, which also corrects for bias errors in estimating $\mathbf{o}$ [8, p. 92].

USPACOR is also flexible to include group-Lasso counterparts of the ℓ_1-norm of $\mathbf{o}$. These are useful when one knows a priori that outliers are clustered, and collections of them can be (non)zero as a group; or, with high-dimensional data, e.g., images, where due to occlusion one may wish to discard the entire image instead of individual pixels. Group regularization terms does not sacrifice convexity and thus USPACOR's computational efficiency, since efficient group Lasso solvers are now available. In particular, the group LARS algorithm in [17] returns the entire robustification path. A different notion of grouping can be effected by superimposing ℓ_1-norms of different $\mathbf{o}$ terms appearing e.g., with USPACOR-based Kalman smoothing formulations [cf. (8)]. While in this case sparsity is not enforced at group level, each group has its own tuning parameter.

Regarding vector $\boldsymbol{\theta}$, ℓ_1-norm regularization is prudent if there is prior information that the unknown vector is sparse, thus robustifying the Lasso. Ridge penalties of the form $\lambda_2\|\boldsymbol{\theta}\|_2^2$ are also useful when the regression matrix $\mathbf{X}$ is ill-conditioned. A convex combination of ℓ_1 and ℓ_2 norms is known as the elastic net, which encourages sparsity while effectively dealing with strong correlation among variables [8, p. 662]. Note that being flexible to include these regularization terms, USPACOR can reject outliers even when the linear regression problem is under-determined. Group-Lasso counterparts can be incorporated as standalone regularizers, or jointly with the ℓ_1-norm of $\boldsymbol{\theta}$ to encourage hierarchical sparsity across and within groups [15]. If there is structure in the data such as smoothness or piecewise constancy, fused Lasso regularization can be adopted as well [8, p. 666]. The resulting convex cost may be challenging to optimize however, since coupling of variables renders CD solvers ineffective.

Remark 2. The limited space allows only for a closing comment on areas not covered here, which can also benefit from the USPACOR-based approach. Those that will be reported in the near future include robust nonlinear (e.g., Volterra) kernel regression, principal component analysis, and clustering.

5. NUMERICAL COMPARISON: USPACOR VS. RANSAC

A numerical experiment is carried out in this section, to compare the performance of USPACOR against RANSAC in a linear regression setting. For $N = 100$, inliers adhere to the linear Gaussian model $y_i = \mathbf{x}_i'\boldsymbol{\theta}_0 + \varepsilon_i$, where the "true" parameter vector $\boldsymbol{\theta}_0 \sim \mathcal{N}(10 \times \mathbf{1}_{10}, \mathbf{I}_{10})$, and $\mathbf{1}_{10}$ denotes the 10×1 vector of all ones. The i.i.d. data are $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}_{10}, \mathbf{I}_{10})$ and $\varepsilon_i \sim \mathcal{N}(0, 1)$. Outliers are Laplacian distributed with zero-mean and standard deviation $\sqrt{2} \times 10^3$, i.e., $y_i \sim \mathcal{L}(0, 10^3)$ and i.i.d.. Contamination levels ranging from 0% to 80% are examined. The inlier noise variance $\sigma_\varepsilon^2 = 1$ is assumed known.

For USPACOR, the optimum tuning parameter λ_1^* is obtained using an AVD criterion in (6). Ten samples of the robustification path are employed, equispaced on a logarithmic λ_1 scale. To further enhance the performance of USPACOR, a single iteration is carried out to minimize a concave sum-of-logs surrogate of (3). The refinement step is initialized with the solution to (4), for $\lambda_1 = \lambda_1^*$. The number of RANSAC iterations is fixed to either 1,000 or 10,000; and the threshold used to decide whether a data point is an outlier is set to $3 \times \sigma_\varepsilon$. RANSAC is enhanced with a follow-up Huber M-estimation step using the RANSAC-generated inlier set. The Huber function parameter is set to $1.345 \times \sigma_\varepsilon$ as suggested in [6].

Fig. 1 compares RANSAC with USPACOR in terms of root mean square error (RMSE), defined as RMSE $:= E[\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2]$, and approximated by sample averaging over 100 Monte Carlo runs. It is apparent that both methods generate very accurate results for small percentages of contamination. However, as the fraction of outliers increases, RANSAC breaks down resulting in large RMSEs with high variability. USPACOR provides accurate results up to 40% contamination, and degrades gracefully beyond this level. In terms of complexity, USPACOR falls in between RANSAC 1,000 and RANSAC 10,000. These results corroborate that USPACOR is a competitive alternative for robust linear regression, and outperforms state of the art RANSAC methods.

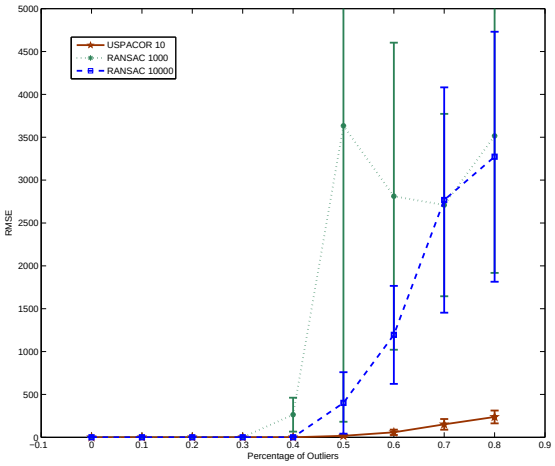

Fig. 1. USPACOR vs. RANSAC: RMSE comparison.

6. REFERENCES

[1] E. J. Candes, M. B. Wakin, and S. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimzation," *Journal of Fourier Analysis and Applications*, vol. 14, pp. 877–905, Dec. 2008.

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

[3] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *JASA*, pp. 1348–1360, 2001.

[4] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Comms. of the ACM*, pp. 381–395, 1981.

[5] J. Friedman, T. Hastie, and R. Tibshirani, "Regularized paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, 2010.

[6] J. J. Fuchs, "An inverse problem approach to robust regression," in *Proc. of ICASSP*, Phoeniz, AZ, Mar. 1999, pp. 180–188.

[7] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge Univ. Press, 2003.

[8] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. Springer, 2009.

[9] P. J. Huber, *Robust Statistics*. Wiley, 1981.

[10] Y. Jin and B. D. Rao, "Algorithms for robust linear regression by exploiting the connection to sparse signal recovery," in *Proc. of ICASSP*, Dallas, TX, Mar. 2010, pp. 3830–3833.

[11] V. Kekatos and G. B. Giannakis, "Robust layered sensing: From sparse signals to sparse residuals," in *Proc. of 44th Asilomar Conf. on Signals, Systems, and Comp.*, Pacific Grove, CA, Nov. 2010.

[12] I. Markovsky and S. V. Huffel, "Overview of total least-squares methods," *Signal Processing*, vol. 87, pp. 2283–2302, 2007.

[13] P. J. Rousseeuw and K. V. Driessen, "Computing LTS regression for large data sets," *Data Mining and Knowl. Disc.*, pp. 29–45, 2006.

[14] P. J. Rousseeuw and A. M. Leroy, *Robust regression and outlier detection*. New York, NY: Wiley, 1987.

[15] P. Sprechmann, I. Ramirez, G. Sapiro, and Y. C. Eldar, "Collaborative hierarchical sparse modeling," in *Proc. of CISS*, Princeton, 2010.

[16] J. Tropp, "Just relax: Convex programming methods for identifying sparse signals," *IEEE Trans. Inf. Theory*, pp. 1030–1051, Mar. 2006.

[17] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Royal. Statist. Soc B*, vol. 68, pp. 49–67, 2006.