ROBUST NONPARAMETRIC REGRESSION BY CONTROLLING SPARSITY*

Gonzalo Mateos and Georgios B. Giannakis

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

ABSTRACT

Nonparametric methods are widely applicable to statistical learning problems, since they rely on a few modeling assumptions. In this context, the fresh look advocated here permeates benefits from variable selection and compressive sampling, to robustify nonparametric regression against outliers. A variational counterpart to least-trimmed squares regression is shown closely related to an ℓ_0-(pseudo)norm-regularized estimator, that encourages *sparsity* in a vector explicitly modeling the outliers. This connection suggests efficient (approximate) solvers based on convex relaxation, which lead naturally to a variational M-type estimator equivalent to Lasso. Outliers are identified by judiciously tuning regularization parameters, which amounts to controlling the sparsity of the outlier vector along the whole *robustification* path of Lasso solutions. An improved estimator with reduced bias is obtained after replacing the ℓ_0-(pseudo)norm with a nonconvex surrogate, as corroborated via simulated tests on robust thin-plate smoothing splines.

Index Terms— Robustness, nonparametric regression, outlier rejection, sparsity, Lasso.

1. INTRODUCTION

Consider the classical problem of function estimation, in which an input vector $\mathbf{x} := [x_1, \ldots, x_p]' \in \mathbb{R}^p$ is given [$(\cdot)'$ denotes transpositon], and the goal is to predict the real-valued scalar response $y = f(\mathbf{x})$. The unknown function f is to be estimated from a training data set $\mathcal{T} := \{y_i, \mathbf{x}_i\}_{i=1}^N$. When f is assumed belonging to a family of finitely parameterized functions, standard (non)linear regression techniques can be adopted. If on the other hand, one is only willing to assume that f belongs to a (possibly infinite dimensional) space of "smooth" functions $\mathcal{H}$, then a nonparametric approach is in order, and this will be the focus of this work. Without further constraints beyond $f \in \mathcal{H}$, functional estimation from finite data is an ill-posed problem. To bypass this challenge, the problem is typically solved by minimizing appropriately regularized criteria, allowing one to control model complexity; see, e.g., the tutorial treatment in [3]. It is then further assumed that $\mathcal{H}$ has the structure of a reproducing kernel Hilbert space (RKHS), with corresponding positive definite reproducing kernel function $K(\cdot, \cdot) : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$, and norm denoted by $\|\cdot\|_{\mathcal{H}}$ [12].

The performance of traditional approaches that minimize the sum of squared model residuals regularized by a term of the form $\|f\|_{\mathcal{H}}^2$, is severely degraded in the presence of outliers. This is because the least-squares (LS) component of the cost is not robust [6]. Recent efforts have considered replacing the squared loss with a robust counterpart such as Huber's function, or its variations, but lack a systematic means of selecting the proper threshold that determines which datum is considered an outlier [14].

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

The starting point here is a variational least-trimmed squares (VLTS) estimator, suitable for robust function approximation in $\mathcal{H}$. VLTS is shown closely related to an (NP-hard) ℓ_0-(pseudo)norm-regularized estimator, adopted to fit a regression model that explicitly incorporates an unknown *sparse* vector of outliers [5]. As in compressive sampling (CS) [11], efficient (approximate) solvers are obtained by replacing the outlier vector's ℓ_0 (pseudo)norm with its closest convex approximant, the ℓ_1 norm. This leads naturally to a variational M-type estimator of f, also shown equivalent to a least-absolute shrinkage and selection operator (Lasso) [10] on the vector of outliers. A tunable parameter in Lasso controls the sparsity of the estimated vector, and the number of outliers as a byproduct. Hence, effective methods to select this parameter are of paramount importance.

The link between ℓ_1-norm regularization and robustness was also exploited for parameter (but not function) estimation in [5] and [7]. In [5] however, the selection of Lasso's tuning parameter is only justified for Gaussian training data; whereas a rigid value motivated by CS results is adopted in the Bayesian formulation of [7]. Here instead, a more general and systematic approach is pursued, building on contemporary algorithms that can efficiently compute all *robustification* paths of Lasso solutions, i.e., for all values of the tuning parameter [2, 4, 13]. In this sense, the method here capitalizes on but *is not limited to* sparse settings, since one can examine all possible sparsity levels along the robustification path. An estimator with reduced bias and improved generalization capability is obtained after replacing the ℓ_0-(pseudo)norm with a nonconvex surrogate, instead of the ℓ_1 norm that introduces bias [10, 15]. Simulated tests demonstrate the effectiveness of the novel robust estimation method.

2. ROBUST ESTIMATION PROBLEM

The training data in $\mathcal{T}$ comprises N *noisy* samples of f taken at the input points $\{\mathbf{x}_i\}_{i=1}^N$, and in the present context they can be possibly contaminated with multiple outliers. Building on LTS regression [9], the desired robust estimate $\hat{f}$ can be obtained as the minimizer of the following variational (V)LTS counterpart

$$\min_{f \in \mathcal{H}} \sum_{i=1}^{s} r_{[i]}^2(f) + \mu \|f\|_{\mathcal{H}}^2 \tag{1}$$

where $r_{[i]}^2(f)$ is the i-th order statistic among the squared residuals $r_1^2(f), \ldots, r_N^2(f)$, and $r_i(f) := y_i - f(\mathbf{x}_i)$. The so-termed trimming constant s determines the breakdown point of the VLTS estimator [9], since the largest $N - s$ residuals do not participate in (1). Ideally, one would like to make $N - s$ equal to the (typically unknown) number of outliers N_o in the sample. The tuning parameter $\mu \geq 0$ controls the tradeoff between fidelity to the (trimmed) data, and the degree of "smoothness" measured by $\|f\|_{\mathcal{H}}^2$.

Given that the first summand of the cost in (1) is a nonconvex functional, a nontrivial issue pertains to the existence of the pro-

posed VLTS estimator. Fortunately, a (conceptually) simple solution procedure suffices to show that a minimizer does indeed exist. Consider specifically a given subsample of s training data points, say $\{y_i, \mathbf{x}_i\}_{i=1}^s$, and solve $\min_{f\in\mathcal{H}} \left[\sum_{i=1}^s r_i^2(f) + \mu\|f\|_{\mathcal{H}}^2\right]$. A unique minimizer of the form $\hat{f}_j(\mathbf{x}) = \sum_{i=1}^s \beta_{i,j} K(\mathbf{x}, \mathbf{x}_i)$ is guaranteed to exist, where j is used here to denote the chosen subsample, and the coefficients $\{\beta_{i,j}\}_{i=1}^s$ can be obtained by solving a particular linear system of equations [12, p. 11]. This procedure can be repeated for each of the $J := \binom{N}{s}$ subsamples, to obtain a collection $\{\hat{f}_j(\mathbf{x})\}_{j=1}^J$ of candidate solutions of (1). The winner $\hat{f} := \hat{f}_{j^*}$ that yields the minimum objective value, is the desired VLTS estimator.

Even though conceptually simple, the solution procedure just described guarantees existence of (at least) one solution, but entails a combinatorial search over all J subsamples. This method is intractable for moderate to large sample sizes N, since the search space is combinatorially large.

2.1. VLTS as ℓ_0-(pseudo)norm regularized regression

A novel perspective to robust nonparametric regression is introduced in this section, which explicitly accounts for outliers in the regression model and allows to establish a neat connection between VLTS and ℓ_0-(pseudo)norm-regularized regression. To model the presence of outliers, consider the scalar variables $\{o_i\}_{i=1}^N$ one per training data point, which take the value $o_i = 0$ whenever data point i is an inlier, and $o_i \neq 0$ otherwise. The classic regression model can then be naturally extended to account for the outliers, using

$$y_i = f(\mathbf{x}_i) + o_i + \varepsilon_i, \qquad i = 1, \ldots, N \tag{2}$$

where $\{\varepsilon_i\}_{i=1}^N$ are zero-mean i.i.d. random variables modeling the observation errors. A similar model was advocated under different assumptions in [5] and [7] for robust parameter estimation in linear regression models. For an outlier-free data point i, (2) reduces to $y_i = f(\mathbf{x}_i) + \varepsilon_i$; thus, ε_i will be henceforth referred to as inlier noise. Note that in (2), both $f \in \mathcal{H}$ as well as the $N \times 1$ vector $\mathbf{o} := [o_1, \ldots, o_N]'$ are unknown, and they have to be jointly estimated. On the other hand, as outliers are expected to often comprise a small fraction of the training sample, vector $\mathbf{o}$ is typically *sparse*, i.e., most of its entries are zero. Sparsity provides valuable side-information when it comes to efficiently estimating $\mathbf{o}$, identifying outliers as a byproduct, and consequently performing *robust* estimation of f. To this end, the desired estimate $\hat{f}$ is obtained as the minimizer of

$$\min_{f\in\mathcal{H},\mathbf{o}} \sum_{i=1}^N (y_i - f(\mathbf{x}_i) - o_i)^2 + \mu\|f\|_{\mathcal{H}}^2 + \lambda_0\|\mathbf{o}\|_0. \tag{3}$$

where $\|\mathbf{o}\|_0$ denotes the ℓ_0-(pseudo)norm, which equals the number of nonzero entries of $\mathbf{o}$.

Sparsity is directly controlled through the selection of the tuning parameter $\lambda_0 \geq 0$. Unfortunately, analogously to related ℓ_0-(pseudo)norm constrained formulations in CS and sparse signal representations [11], problem (3) is NP-hard. Supposing that the number of outliers N_o is known, in principle a brute force approach could be adopted to tackle (3), by trying all $\binom{N}{N_o}$ support combinations for $\mathbf{o}$ such that $\|\mathbf{o}\|_0 = N_o$. Similar to VLTS, this procedure becomes intractable for moderate to large-size problems, yet it demonstrates the existence of a minimizer. Interestingly, the similarities between (1) and (3) transcend their complexity, since their solutions coincide for particular values of λ_0 in (3) [8].

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

Proposition 1 formally justifies model (2) and its estimator (3) for robust function approximation, in light of the well documented merits of LTS regression [9]. It further solidifies the connection between sparse linear regression and robust estimation. Most importantly, the ℓ_0-(pseudo)norm regularized formulation in (3) lends itself naturally to efficient (approximate) solvers based on convex relaxation, the subject dealt with next.

3. SPARSITY CONTROLLING OUTLIER REJECTION

To overcome the complexity hurdle in solving (3), one can resort to a suitable relaxation of the objective function. It is useful to recall that the ℓ_1 norm $\|\mathbf{x}\|_1 := \sum_{i=1}^p |x_i|$ of vector $\mathbf{x} \in \mathbb{R}^p$ is the closest convex approximation of $\|\mathbf{x}\|_0$. This property also utilized in the context of CS [11], provides the motivation to relax problem (3) to

$$\min_{f\in\mathcal{H},\mathbf{o}} \sum_{i=1}^N (y_i - f(\mathbf{x}_i) - o_i)^2 + \mu\|f\|_{\mathcal{H}}^2 + \lambda_1\|\mathbf{o}\|_1. \tag{4}$$

Being a convex optimization problem, (4) can be efficiently solved. The nondifferentiable ℓ_1-norm regularization term controls sparsity on the estimator of $\mathbf{o}$, a property that has been exploited in diverse problems in engineering, statistics and machine learning. A noteworthy representative is the Lasso [10], a popular tool for joint estimation and variable selection in linear regression problems.

It is pertinent to ponder on whether problem (4) still has the potential of providing robust estimates $\hat{f}$ in the presence of outliers. The answer is positive, since it is possible to show that (4) is equivalent to a variational M-type estimator [8]

$$\min_{f\in\mathcal{H}} \sum_{i=1}^N \rho(y_i - f(\mathbf{x}_i)) + \mu\|f\|_{\mathcal{H}}^2 \tag{5}$$

where ρ is a scaled version of Huber's convex loss function [6]

$$\rho(u) := \begin{cases} u^2, & |u| \leq \lambda_1/2 \\ \lambda_1|u| - \lambda_1^2/4, & |u| > \lambda_1/2 \end{cases}. \tag{6}$$

Existing works on linear regression have pointed out the equivalence between M-type estimators and ℓ_1-norm regularized regression [5]. However, they have not recognized the connection to LTS via convex relaxation of (3). Here, the treatment goes beyond linear regression by considering nonparametric functional approximation in RKHS. Linear regression is subsumed as a special case, when the linear kernel $K(\mathbf{x}, \mathbf{y}) := \mathbf{x}'\mathbf{y}$ is adopted.

3.1. Solving the relaxed convex problem

Because (4) is jointly convex in f and $\mathbf{o}$, a globally convergent alternating minimization (AM) algorithm can be adopted to solve (4). As shown in [8], AM iterations boil down to a sequence of linear systems and soft-thresholding operations. Such an algorithm is also conceptually interesting, since it explicitly reveals the intertwining between the outlier identification process, and the estimation of the regression function with the appropriate outlier-compensated data $\{y_i - o_i\}_{i=1}^N$. Next, it is established that an alternative to an AM algorithm, is to solve a single instance of Lasso [8].

Proposition 2: *Consider* $\hat{\mathbf{o}}_{Lasso}$ *defined as*

$$\hat{\mathbf{o}}_{Lasso} := \arg\min_{\mathbf{o}} \|\mathbf{X}_\mu\mathbf{y} - \mathbf{X}_\mu\mathbf{o}\|_2^2 + \lambda_1\|\mathbf{o}\|_1 \tag{7}$$

where $\mathbf{y} := [y_1, \ldots, y_N]'$ *and*

$$\mathbf{X}_\mu := \begin{bmatrix} \mathbf{I}_N - \mathbf{K}\left(\mathbf{K} + \mu \mathbf{I}_N\right)^{-1} \\ (\mu \mathbf{K})^{1/2}\left(\mathbf{K} + \mu \mathbf{I}_N\right)^{-1} \end{bmatrix}. \tag{8}$$

The kernel matrix $\mathbf{K} \in \mathbb{R}^{N \times N}$ *has entries* $[\mathbf{K}]_{ij} := K(\mathbf{x}_i, \mathbf{x}_j)$*, while* $\mathbf{I}_N$ *denotes the* $N \times N$ *identity matrix. Then the minimizers* $\{\hat{f}, \hat{\mathbf{o}}\}$ *of* (4) *can be specified given* $\hat{\mathbf{o}}_{Lasso}$*, as* $\hat{\mathbf{o}} := \hat{\mathbf{o}}_{Lasso}$ *and* $\hat{f}(\mathbf{x}) = \sum_{i=1}^{N} \hat{\beta}_i K(\mathbf{x}, \mathbf{x}_i)$*, with* $\hat{\boldsymbol{\beta}} = \left(\mathbf{K} + \mu \mathbf{I}_N\right)^{-1}\left(\mathbf{y} - \hat{\mathbf{o}}_{Lasso}\right)$.

The result in Proposition 2 opens the possibility for effective methods to select λ_1. These methods to be described in detail in the ensuing section, capitalize on recent algorithmic advances on Lasso solvers, which allow one to efficiently compute $\hat{\mathbf{o}}_{\text{Lasso}}$ for all values of the tuning parameter λ_1 [2, 4, 13]. This is crucial for obtaining satisfactory robust estimates $\hat{f}$, since *controlling the sparsity* in $\mathbf{o}$ by tuning λ_1 is tantamount to controlling the number of outliers.

3.2. Selection of the tuning parameters: robustification paths

The tuning parameters μ and λ_1 in (4) control the degree of smoothness in $\hat{f}$ and the number of outliers (nonzero entries in $\hat{\mathbf{o}}_{\text{Lasso}}$), respectively. In the contexts of regularization networks [3] and Lasso estimation for regression [10], corresponding tuning parameters are typically selected via model selection techniques such as cross-validation, or, by minimizing the prediction error over an independent test set, if available. However, these simple methods are severely challenged in the presence of multiple outliers.

The focus here is on an alternative method to overcome the aforementioned challenges, capitalizing on Proposition 2, and the possibility to efficiently compute $\hat{\mathbf{o}}_{\text{Lasso}}$ for all values of λ_1, given μ. To this end, consider a grid of K_μ values of μ in the interval $[\mu_{\min}, \mu_{\max}]$, evenly spaced in a logarithmic scale. Likewise, for each μ consider a similar type of grid consisting of K_λ values of λ_1, where $\lambda_{\max} := 2 \min_i |\mathbf{y}' \mathbf{X}'_\mu \mathbf{x}_{\mu,i}|$ is the minimum λ_1 value such that $\hat{\mathbf{o}}_{\text{Lasso}} \neq \mathbf{0}_N$ [4], and $\mathbf{X}_\mu = [\mathbf{x}_{\mu,1} \ldots \mathbf{x}_{\mu,N}]$ in (7). Note that each of the K_μ values of μ gives rise to a different λ grid, since $\lambda_{\max}$ depends on μ trough $\mathbf{X}_\mu$. Given the existing algorithmic alternatives to tackle the Lasso [2, 4, 13], it is safe to assume that (7) can be efficiently solved over the (nonuniform) $K_\mu \times K_\lambda$ grid of values of the tuning parameters. This way, for each value of μ one obtains K_λ samples of the Lasso path of solutions, which in the present context can be referred to as *robustification path*. As λ_1 decreases, more variables $\hat{o}_{\text{Lasso},i}$ enter the model signifying that more of the training points are considered as outliers.

Based on the robustification paths and the prior knowledge available on the outlier model (2), several alternatives are given next to select the best pair $\{\mu, \lambda_1\}$; additional ones can be found in [8].

Variance of the inlier noise is known: Under the assumption that the variance σ_ε^2 of the i.i.d. inlier noise random variables ε_i in (2) is known, one can proceed as follows. Using the solution $\hat{f}$ obtained for each pair $\{\mu_i, \lambda_j\}$ on the grid, form the $K_\mu \times K_\lambda$ sample variance matrix $\bar{\boldsymbol{\Sigma}}$ with ij-th entry $[\bar{\boldsymbol{\Sigma}}]_{ij}$ corresponding to a sample estimate of σ_ε^2, neglecting those training data points $\{y_i, \mathbf{x}_i\}$ that the method determined to be contaminated with outliers. The winner tuning parameters $\{\mu^*, \lambda_1^*\} := \{\mu_{i^*}, \lambda_{j^*}\}$ are such that

$$[i^*, j^*] := \arg\min_{i,j} |[\bar{\boldsymbol{\Sigma}}]_{ij} - \sigma_\varepsilon^2|. \tag{9}$$

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 (9). One 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 f, e.g., obtained after solving (4) with $\lambda_1 = 0$ and using a small subset of the training dataset $\mathcal{T}$. The factor 1.48 provides an approximately unbiased estimate of σ_ε, when the inlier noise is Gaussian.

3.3. Refinement via nonconvex regularization

Instead of substituting $\|\mathbf{o}\|_0$ in (3) by its closest convex approximation, namely $\|\mathbf{o}\|_1$, letting the surrogate function to be non-convex can yield tighter approximations. For example, $\|\mathbf{x}\|_0$ was surrogated in [1] by the logarithm of the geometric mean of its elements, or by $\sum_{i=1}^{p} \log|x_i|$. Adopting related ideas in the present nonparametric context, consider approximating (3) by the *nonconvex* formulation

$$\min_{f \in \mathcal{H}, \mathbf{o}} \sum_{i=1}^{N} (y_i - f(\mathbf{x}_i) - o_i)^2 + \mu \|f\|_{\mathcal{H}}^2 + \lambda_0 \sum_{i=1}^{N} \log(|o_i| + \delta) \tag{10}$$

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

Local methods based on iterative linearization of $\log(|o_i| + \delta)$, around the current iterate $o_i[k]$, can be adopted to minimize (10). Skipping details that can be found in [8], one such iteration is

$$\mathbf{o}[k] := \arg\min_{\mathbf{o}} \|\mathbf{X}_\mu \mathbf{y} - \mathbf{X}_\mu \mathbf{o}\|_2^2 + \lambda_0 \sum_{i=1}^{N} w_i[k] |o_i| \tag{11}$$

$$w_i[k] := \left(|o_i[k-1]| + \delta\right)^{-1}, \quad i = 1, \ldots, N. \tag{12}$$

which amounts to an iteratively reweighted version of (7). A good initialization for the iteration in (12) is $\mathbf{o}[-1] := \hat{\mathbf{o}}_{\text{Lasso}}$, which corresponds to the solution of (7) [and (4)] for $\lambda_0 = \lambda_1^*$ and $\mu = \mu^*$. The numerical tests in Section 4 will indicate that even a single iteration of (11) suffices to obtain improved estimates $\hat{f}$, in comparison to those obtained from (7). The improvements due to (11) can be leveraged to bias reduction, also achieved by similar *weighted* ℓ_1-norm regularizers proposed for linear regression [15].

4. NUMERICAL EXPERIMENTS

To validate the proposed approach to robust nonparametric regression, a simulated test is carried out here in the context of thin-plate smoothing spline approximation [12]. Specializing (4) to this setup, the robust thin-plate splines estimator can be formulated as

$$\min_{f \in \mathcal{S}, \mathbf{o}} \sum_{i=1}^{N} (y_i - f(\mathbf{x}_i) - o_i)^2 + \mu \int_{\mathbb{R}^2} \|\nabla^2 f\|_F^2 d\mathbf{x} + \lambda_1 \|\mathbf{o}\|_1 \tag{13}$$

where $||\nabla^2 f||_F$ denotes the Frobenius norm of the Hessian of $f : \mathbb{R}^2 \to \mathbb{R}$. The optimization is over $\mathcal{S}$, the space of Sobolev functions, for which the smoothing penalty in (13) is well defined [12]. RKHSs such as $\mathcal{S}$, with inner-products (and seminorms) involving derivatives are studied in detail in [12].

Noisy samples of the true function f_o comprise the training set $\mathcal{T}$. Function f_o is generated as a Gaussian mixture with two components, with randomly drawn mean vectors and covariance matrices; see also Fig. 1 (top left). The training data set comprises $N = 200$ examples, with inputs $\{\mathbf{x}_i\}_{i=1}^N$ drawn from a uniform distribution in the square $[0, 3] \times [0, 3]$. Without loss of generality, the corrupted data correspond to the first N_o training samples with $N_o = \{10, 20, 30, 40, 50\}$, for which the response values $\{y_i\}_{i=1}^{N_o}$ are independently drawn from a uniform distribution over $[-4, 4]$. Inliers are generated from the model $y_i = f_o(\mathbf{x}_i) + \varepsilon_i$, where the

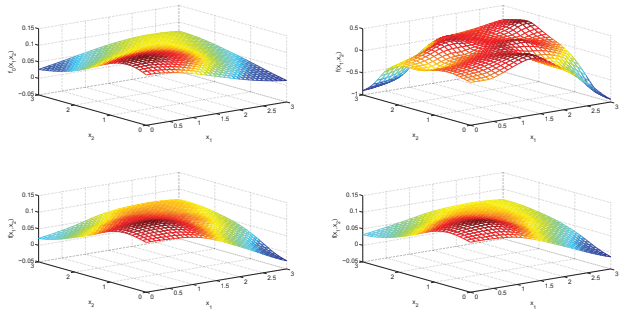

Fig. 1. (top-left) True function $f_o(\mathbf{x})$; (top-right) nonrobust predicted function; (bottom-left) predicted function after solving (13); (bottom-right) refined predicted function using a nonconvex penalty.

Table 1. Results for the thin-plate splines simulated test

N_o	λ_1^*	μ^*	$\mathbf{Err}_{\mathcal{T}}$ **for** (4)	$\mathbf{Err}_{\mathcal{T}}$ **for** (10)
10	3.9×10^{-2}	2.9×10^{-3}	2.4×10^{-5}	2.2×10^{-5}
20	3.8×10^{-2}	1.6×10^{-2}	4.3×10^{-5}	2.4×10^{-5}
30	2.3×10^{-2}	6.7×10^{-2}	2.9×10^{-5}	1.9×10^{-5}
40	2.8×10^{-2}	6.1×10^{-3}	1.6×10^{-5}	1.3×10^{-5}
50	2.5×10^{-2}	5.4×10^{-2}	1.2×10^{-5}	1.0×10^{-5}

independent additive noise terms $\varepsilon_i \sim \mathcal{N}(0, 10^{-3})$ are Gaussian distributed, for $i = N_o + 1, \ldots, 200$.

In the context of the present experiment, the inlier noise variance $\sigma_\varepsilon^2 = 10^{-3}$ is assumed known. A nonuniform grid of μ and λ_1 values is constructed, as described in Section 3.2. The relevant parameters are $K_\mu = K_\lambda = 200$, $\mu_{\min} = 10^{-9}$ and $\mu_{\max} = 1$. For each value of μ, the λ_1 grid spans the interval defined by $\lambda_{\max} := 2 \min_i |\mathbf{y}' \mathbf{X}_\mu' \mathbf{x}_{\mu,i}|$ and $\lambda_{\min} = \epsilon \lambda_{\max}$, where $\epsilon = 10^{-4}$. Each of the K_μ robustification paths corresponding to the solution of (7) is obtained using the SpaRSA toolbox in [13], exploiting warm starts for faster convergence. Fig. 2 depicts an example with $N_o = 20$ and $\mu^* = 1.55 \times 10^{-2}$. With the robustification paths at hand, it is possible to form the sample variance matrix $\bar{\mathbf{\Sigma}}$, and select the optimum tuning parameters $\{\mu^*, \lambda_1^*\}$ based on the criterion (9). Finally, the robust estimates are refined by running a single iteration of (11) as described in Section 3.3. The value $\delta = 10^{-5}$ was utilized, and several experiments indicated that the results are quite insensitive to the selection of this parameter.

The same experiment was conducted for a variable number of outliers N_o, and the results are listed in Table 1. In all cases, a 100% outlier identification success rate was obtained, for the chosen value of the tuning parameters. To assess quality of the estimated function $\hat{f}$, an approximation to the *generalization error* $\mathrm{Err}_{\mathcal{T}}$ was computed as $\mathrm{Err}_{\mathcal{T}} = E[(y - \hat{f}(\mathbf{x}))^2 | \mathcal{T}] \approx \sum_{i=1}^{\tilde{N}} (\tilde{y}_i - \hat{f}(\tilde{\mathbf{x}}_i))^2 / \tilde{N}$ were $\{\tilde{y}_i, \tilde{\mathbf{x}}_i\}_{i=1}^{\tilde{N}}$ is an independent test set generated from the model $\tilde{y}_i = f_o(\tilde{\mathbf{x}}_i) + \varepsilon_i$. For the results in Table 1, $\tilde{N} = 961$ was adopted corresponding to a uniform rectangular grid of 31×31 points $\tilde{\mathbf{x}}_i$ in $[0, 3] \times [0, 3]$. Inspection of Table 1 reveals that the nonconvex refinement (10) has an edge over (4) with regards to generalization capability, for all values of N_o. As expected, the bias reduction effected by the iteratively reweighting procedure of Section 3.3 improves considerably the generalization capability of the method.

A pictorial summary of the results is given in Fig 1, for $N_o = 20$ outliers. Fig 1 (top-left) depicts the true Gaussian mixture $f_o(\mathbf{x})$,

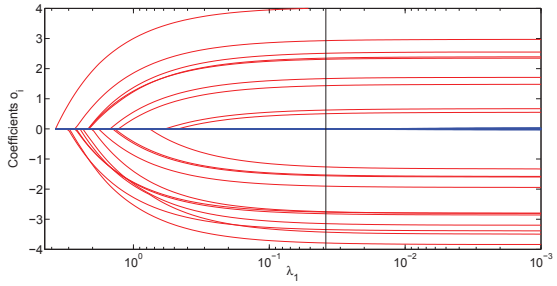

Fig. 2. Robustification paths. The coefficients $\hat{o}_i$ corresponding to the outliers are shown in red, while the rest are shown in blue. The vertical line indicates the selection of $\lambda_1^* = 3.83 \times 10^{-2}$.

whereas Fig. 1 (top-right) shows the nonrobust thin-plate splines estimate obtained after solving (13) with $\lambda_1 = 0$. Even though the thin-plate penalty enforces some degree of smoothness, the estimate is severely disrupted by the presence of outliers [cf. the difference on the z-axis ranges]. On the other hand, Fig. 1 (bottom-left and right), respectively, show the robust estimate $\hat{f}$ with $\lambda_1^* = 3.83 \times 10^{-2}$, and its bias reducing refinement. The improvement is apparent, corroborating the effectiveness of the proposed approach.

5. 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] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines," *Advances in Computational Mathematics*, vol. 13, pp. 1–50, 2000.

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

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

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

[7] 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.

[8] G. Mateos and G. B. Giannakis, "Sparsity-controlling robust nonparametric regression," *IEEE Trans. Signal Process.*, 2010 (submitted).

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

[10] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal. Statist. Soc B*, vol. 58, pp. 267–288, 1996.

[11] J. Tropp, "Just relax: Convex programming methods for identifying sparse signals," *IEEE Trans. Inf. Theory*, vol. 51, pp. 1030–1051, Mar. 2006.

[12] G. Wahba, *Spline Models for Observational Data*. SIAM, 1990.

[13] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Trans. Signal Process.*, vol. 57, pp. 2479–2493, 2009.

[14] J. Zhu, S. C. H. Hoi, and M. R. T. Lyu, "Robust regularized kernel regression," *IEEE Trans. Syst., Man, Cybern. B Cybern.*, vol. 38, pp. 1639–1644, Dec. 2008.

[15] H. Zou, "The adaptive lasso and its oracle properties," *J. Amer. Stat. Assoc.*, vol. 101, no. 476, pp. 1418–1429, 2006.