Robust reconstruction with nonconvex subset constraints

Arthur Marmin, Marc Castella, Jean-Christophe Pesquet

2020 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 21–24, 2020, ESPOO, FINLAND

ROBUST RECONSTRUCTION WITH NONCONVEX SUBSET CONSTRAINTS: A POLYNOMIAL OPTIMIZATION APPROACH

Arthur Marmin[†] *Marc Castella*[⋆] *Jean-Christophe Pesquet*[†]

[†]Université Paris-Saclay, CentraleSupélec, Center for Visual Computing, Inria, Gif-sur-Yvette, France
[⋆]SAMOVAR, Télécom SudParis, Institut Polytechnique de Paris, France

ABSTRACT

In this paper, we are interested in the recovery of an unknown signal corrupted by a linear operator, a nonlinear function, and an additive Gaussian noise. In addition, some of the observations contain outliers. Many robust data fit functions which alleviate sensitivity to outliers can be expressed as piecewise rational functions. Based on this fact, we reformulate the robust inverse problem as a rational optimization problem. The considered framework allows us to incorporate nonconvex constraints such as unions of subsets. The rational problem is then solved using recent optimization techniques which offer guarantees for global optimality. Finally, experimental results illustrate the validity of the recovered global solutions and the good quality of the reconstructed signals despite the presence of outliers.

Index Terms— polynomial optimization, global optimization, robust estimation, union of subspaces/subsets, nonconvex constraints

1. INTRODUCTION

In order to reconstruct a signal from corrupted observations, approaches based on the maximum likelihood estimator are widely used. In many applications, the noise is modeled by a zero-mean normal distribution which is added to the noiseless signal. As a consequence, minimization of a mean square error appears as a ubiquitous technique in signal processing.

However, in practice, it is common that outliers are present in the observations, so altering the noise distribution. This results in poor performance of least squares based estimators. Indeed, outliers produce large errors making their corresponding weight prevalent in least squares fitting. Consequently, even very few of them can significantly decrease the performance of the estimator.

With the emergence of big data, manually discarding outliers is not a suitable solution. Moreover, it can be difficult to decide which data are outliers, especially in high dimensional problems. Hence, many robust fit functions have been proposed in order to reduce the impact of outliers on the estimate.

A standard approach in robust estimation is to cap the ℓ_2 function in order to keep the least squares behavior for small error values around zero and to apply a constant term in order to penalize equally errors and outliers above a given threshold [1]. However, the convexity of the fit function is lost and the resulting optimization problem becomes intricate. A convex surrogate is the ℓ_1 norm, or a smoothed version of it, which reduces the influence of the outliers as it gives a steadily increasing weight to the errors [2]. Nonetheless, it results in a shrinkage of low errors towards 0, which is not desirable as it induces biased estimates. In order to keep benefits both of least squares for low errors and of the least absolute values for high errors, the Huber function has been proposed [3]. The latter also has the advantage of being convex, which is an enjoyable property for optimization. Smoother version of Huber functions are also used such as the pseudo-Huber function [4]. Other M-estimators have been proposed such as Tukey's function for instance [5]. Furthermore, transpositions of robust estimators for sparse signals estimation [6] and for multivariate signals [5, 7] have also been proposed.

Exploiting the properties of the original signal is an important feature in inverse problems, that can also contribute to improve robustness. We here concentrate on an assumption corresponding to a union of subsets model. Such a model has been a topic of interest in signal processing, especially when the subsets are affine spaces. For instance, it has appeared in compressed sensing where one wants to reconstruct a signal of size T having only k nonzero components from linear observations [8]. Other examples that can be expressed as union of subspaces include reconstruction of a stream of Dirac impulses where both the location and the amplitude are unknown, determination of overlapping echoes with unknown delay and amplitude, or reconstruction of a signal whose Fourier transform is known to be located in an union of sub-bands [9]. Union of subspaces are also useful in matrix completion to express nonlinear connections between elements [10].

Working in a union of subsets often leads to a challenging problem as linearity and convexity are lost. By making additional assumptions on the shapes of the subspaces or on the model, some methods have been shown to be successful in

978-1-7281-6662-9/20/$31.00 ©2020 IEEE

solving specific problems [8,11–13]. Nevertheless, more general forms are still difficult to solve and the proposed methods do no extend easily to the union of general subsets.

A key observation is that many unions of subsets, can be expressed as polynomial constraints. Similarly, the capped ℓ_2 function used for robustness is piecewise polynomial. The approach in this paper is grounded on the versatility of polynomial modelling which emcompasses both contexts. We thus address a problem of robust signal reconstruction on an union of subsets. We add extra nonconvex constraints that force the amplitude of the signal to be greater than a threshold or identically zero. However, this constraint is difficult to handle. We propose to reformulate this nonconvex problem on an union of subsets as a rational optimization problem that is solved by using recent tools from polynomial optimization. The methodology was introduced in [14] but in a different context. We adapt it here to robust estimation. Its main benefit is to provide global optimization guarantees.

Our paper is organized as follows: Section 2 introduces the model for the observed signal. Section 3 details the optimization methodology and the formulation of the optimization approach we follow to reconstruct the initial signal. Simulation results are presented in Section 4, and Section 5 concludes our work.

2. OUR MODEL

2.1. Observations model

We consider a degraded version $\overline{\mathbf{y}}$ of an original signal $\overline{\mathbf{x}}$ of size T according to the following model

$$\overline{\mathbf{y}} = \phi(\mathbf{H}\overline{\mathbf{x}}) + \mathbf{w}\,, \tag{1}$$

where $\mathbf{H}$ is a $T \times T$ matrix corresponding to a linear operator, $\mathbf{w}$ is a zero-mean white Gaussian noise, and ϕ is a rational function, i.e. a ratio of two polynomials, that acts componentwise. The latter function gives flexibility to our model: for instance, ϕ can be used to model saturations of sensors. Note that assuming a rational ϕ is not restrictive. Indeed, most non-rational functions of practical interest can be closely approximated with rational or piecewise rational functions [15].

Moreover, some of the previous degraded samples are significantly perturbed. This models for instance the possibility of a sensor malfunction. It follows that the observation vector includes a certain number of outliers, i.e. values that differ significantly from the others. Finally, the observation model becomes

$$(\forall t \in [\![1,T]\!]) \quad y_t = \begin{cases} \overline{y}_t & \text{with probability } 1-\delta\,, \\ \overline{y}_t + n_t & \text{with probability } \delta\,. \end{cases}$$

In the above equation, $\overline{\mathbf{y}}$ comes from Model (1), δ is a small real between 0 and 1, and n_t is the realization of a noise with high amplitude compared to values of $\overline{\mathbf{y}}$.

2.2. Modelling signal with a union of subsets

Similarly to many methods for inverse problem, ours relies on both data fidelity and some assumptions about the original signal $\overline{\mathbf{x}}$. We consider the following prior knowledge on $\overline{\mathbf{x}}$: its components are either zero or have absolute value above a given positive threshold λ, i.e.

$$(\forall t \in [\![1,T]\!]) \quad x_t = 0 \quad \text{or} \quad |x_t| \geq \lambda\,. \tag{2}$$

Constraints (2) imply that each sample of the signal $\overline{\mathbf{x}}$ belongs to the union of three convex subsets, namely $\{0\} \cup]-\infty, -\lambda] \cup [\lambda, +\infty[$. The benefit of such a model is to reduce the space where we search for a solution. Let us emphasize that the union we consider here is composed of subsets that are not necessarily linear subspaces. Constraints (2) are nonconvex and thus result in a difficult optimization problem.

Our assumption on the signal model may be related to the standard union of subspaces approach from compressed sensing, where the dimension of the subspaces is low compared to the underlying space dimension. This accounts for sparsity. On the other hand, the union of subspaces model does not necessarily restrict the number of non-zero components nor impose sparsity on the components [16]. Hence, the method we propose can be applied to recover sparse signals as well as dense signals in the sense that only a few elements are null.

3. PROBLEM FORMULATION

3.1. Robust reconstruction criteria

To reconstruct the original signal $\overline{\mathbf{x}}$, we minimize a criterion matching the data to the model. A classic approach for dealing with Model (1) is to minimize the mean square error $\|\mathbf{y} - \phi(\mathbf{H}\mathbf{x})\|^2$ with respect to $\mathbf{x}$. However, the presence of outliers skews the solutions. To tackle this issue, several robust surrogates for the squared norm have been proposed. The latter are all written under the form $\sum_{t=1}^{T} \Psi_\theta(y_t - \phi((\mathbf{H}\mathbf{x})_t))$ where the function $\Psi_\theta : \mathbb{R} \to \mathbb{R}$ may depend on a positive real parameter θ. We obtain the following criteria, where $\mathbb{1}_{\mathcal{X}}$ denotes the characteristic function of a set $\mathcal{X}$:

- the previously mentioned ℓ_2 norm: $\Psi_\theta(x) = x^2$,
- the ℓ_1 norm: $\Psi_\theta(x) = |x|$,
- Huber function [3]:

$$\Psi_\theta(x) = \frac{1}{2}x^2 \mathbb{1}_{\{|x| \leq \theta\}}(x) + \theta\left(|x| - \frac{1}{2}\theta\right)\mathbb{1}_{\{|x|>\theta\}}(x)\,,$$

- the pseudo-Huber function [4]: $\Psi_\theta(x) = \sqrt{x^2 + \theta^2}$,
- and the capped ℓ_2 function:

$$\Psi_\theta(x) = \mathbb{1}_{\{|x| \geq 1/\sqrt{\theta}\}}(x) + \theta x^2 \mathbb{1}_{\{|x|<1/\sqrt{\theta}\}}(x)\,.$$

Figure 1 displays the considered robust fit functions Ψ_θ. The capped ℓ_2 and Huber functions aim at reducing the impact of the outliers by decreasing the penalization on the high errors while preserving mean squares on the low errors. The parameter θ represents here the threshold value on the error between these two penalties. It must be set carefully depending on the level of noise $\mathbf{w}$ and of the outliers.

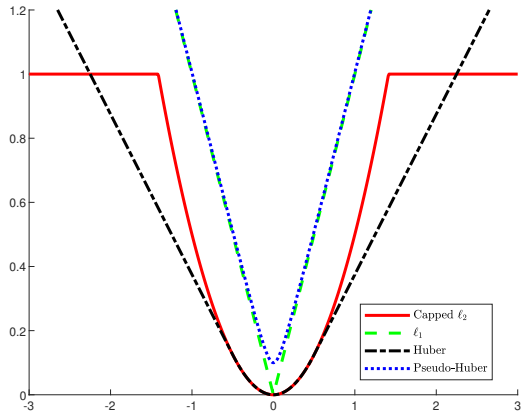

Fig. 1. Considered robust fit functions Ψ_θ ($\theta = 0.5$ for the capped ℓ_2 and Huber functions, $\theta = 0.1$ for the pseudo-Huber function)

Considering the fit criterion to be minimized and Constraints (2) simultaneously, we finally obtain the following optimization problem:

$$\begin{array}{ll} \underset{\mathbf{x}\in\mathbb{R}^T}{\text{minimize}} & \displaystyle\sum_{t=1}^{T} \Psi_\theta(y_t - \phi((\mathbf{H}\mathbf{x})_t)) \\ \text{s.t.} & (\forall t \in [\![1,T]\!]) \quad x_t = 0 \text{ or } |x_t| \geq \lambda\,, \end{array} \tag{3}$$

where $(\mathbf{H}\mathbf{x})_t$ denotes the t-th component of the vector $\mathbf{H}\mathbf{x}$. Problem (3) is difficult since the constraints are nonconvex. Moreover, the objective function is also nonconvex for the capped ℓ_2 fit. We propose here a general methodology to solve Problem (3) for any above choice of the function Ψ_θ. We first reformulate the optimization problem as a rational one, i.e. the minimization of a rational function under polynomial inequalities, before solving it. We start with the reformulation of the constraints.

3.2. Reformulation as a rational optimization problem

3.2.1. Constraints expressed as polynomial inequalities

The constraints in (3) can be expressed as polynomials ones by introducing extra binary variables ζ_t for all t in $[\![1,T]\!]$. We can indeed substitute the product $\zeta_t x_t$ for x_t in (3) and, letting $\zeta_t = 0$ account for vanishing x_t, the constraints in (3) reduce to the inequalities $|x_t| \geq \lambda$ for all t in $[\![1,T]\!]$. Finally, binary values can be imposed by the polynomial constraints $\zeta_t = \zeta_t^2$ for all t in $[\![1,T]\!]$. We thus obtain the following polynomial optimization problem

$$\begin{array}{ll} \underset{(\mathbf{x},\boldsymbol{\zeta})\in\mathbb{R}^T\times\mathbb{R}^T}{\text{minimize}} & \displaystyle\sum_{t=1}^{T} \Psi_\theta(y_t - \phi((\mathbf{H}(\mathbf{x}\odot\boldsymbol{\zeta}))_t)) \\ \text{s.t.} & (\forall t \in [\![1,T]\!]) \quad |x_t| \geq \lambda \\ & (\forall t \in [\![1,T]\!]) \quad \zeta_t = \zeta_t^2\,, \end{array} \tag{4}$$

where the operator $\odot$ denotes the element-wise Hadamard product. Note that the constraints $|x_t| \geq \lambda$ are still nonconvex and make Problem (4) difficult to solve with standard methods. However, under the transformation of the objective function from Section 3.2.2, it leads to the minimization of a polynomial function subject to polynomial constraints. Therefore it can be solved using polynomial optimization tools [17].

Nevertheless this formulation doubles the number of optimization variables and Problem (4) cannot be solved by state-of-the-art polynomial optimization solvers in a fair amount of time, even for signals of small size T. Instead, we suggest to relax the equality constraints in (2) into an inequality and we obtain the following constraints

$$(\forall t \in [\![1,T]\!]) \quad |x_t| \leq \epsilon \text{ or } |x_t| \geq \lambda\,, \tag{5}$$

where $\epsilon < \lambda$ is a small positive real. Notice that this constraint can also be used to expressed signals that are in two different bands as in [9] for instance. We now write the above constraints as polynomial inequalities and we obtain

$$\begin{array}{ll} (\forall t \in [\![1,T]\!]) & (\epsilon - r_t)(\lambda - r_t) \geq 0 \\ (\forall t \in [\![1,T]\!]) & r_t^2 = x_t^2\,,\ r_t \geq 0\,. \end{array}$$

The absolute value in the constraints of (5) is handled by adding the extra variable $\mathbf{r} = (r_1, \ldots, r_T)$ as detailed in [18].

3.2.2. Objective function expressed as a rational function

In the following, we study the reformulation for the capped ℓ_2 function since it is the most challenging case. The method can be easily adapted for other fit functions from Section 3.1.

As witnessed by the characteristic functions that appear in the definition of the capped ℓ_2 function, the objective in (3) is piecewise polynomial. Following [14], a characteristic function can be replaced by a binary variable z that takes identical values. For instance, if Ψ_θ is the capped ℓ_2 function as defined in Section 3.1, we introduce a binary variable z which takes value 0 when $|x|$ is smaller than $1/\sqrt{\theta}$, and 1 otherwise. This can be written as

$$(z - 1/2)\left(|x| - 1/\sqrt{\theta}\right) \geq 0\,. \tag{6}$$

Substituting similarly the characteristic functions for all t in $[\![1,T]\!]$ and the original constraints in Problem (3), we

finally obtain the following rational optimization problem

$$\begin{aligned}
\underset{(\mathbf{x},\mathbf{r},\mathbf{v},\mathbf{z})\in\mathbb{R}^{4T}}{\text{minimize}} \quad & \sum_{t=1}^{T} \; z_t + (1-z_t)\theta(y_t - \phi((\mathbf{H}\mathbf{x})_t))^2 \\
\text{s.t. } (\forall t \in [\![1,T]\!]) \quad & (\epsilon - r_t)(\lambda - r_t) \geq 0 \\
& r_t^2 = x_t^2 \;,\; r_t \geq 0 \\
& (z_t - 1/2)\left(v_t - 1/\sqrt{\theta}\right) \geq 0 \\
& v_t^2 = (y_t - \phi((\mathbf{H}\mathbf{x})_t))^2 \;,\; v_t \geq 0 \\
& z_t = z_t^2 \,,
\end{aligned} \tag{7}$$

where the extra variable $\mathbf{v}$ is used to handle the absolute value in Constraint (6). The objective function is now a rational function in $\mathbf{x}$ and $\mathbf{z}$ while the constraints are polynomial inequalities in $\mathbf{x}$, $\mathbf{r}$, $\mathbf{v}$, and $\mathbf{z}$.

Note that the radical in the pseudo-Huber function can be handled similarly to the absolute value. Indeed, we can replace $\sqrt{x^2+\theta^2}$ by an additional variable u and add the polynomial constraints

$$\begin{cases} u^2 = x^2 + \theta^2 \\ u \geq 0 \,. \end{cases}$$

3.3. Solving rational optimization problem

Problem (7) is a rational optimization problem and can be solved with a hierarchy of convex Semi-Definite Programming (SDP) problems known as Lasserre's hierarchy [17]. Solving the successive relaxations yields an increasing sequence of lower bounds converging to the optimal objective value of the polynomial optimization problem. Furthermore the convergence occurs at a finite order of relaxation generically [19], i.e. with probability one when the polynomial coefficients are drawn randomly from a continuous probability density law. Note that, as it is required by the method, we consider with no loss of generality that all optimization variables are bounded; in our case, their absolute value is smaller than one. A sufficient condition using the rank of the solution to the SDP [17] can be used to detect convergence in the hierarchy. In this situation, the minimizers of the polynomial objective function can be extracted from the solution of the SDP problem [20]. However, even when the sufficient condition for ensuring convergence in the hierarchy does not hold, one can compare the lower bound with the value of the criterion at the extracted minimizers. A small gap between both values is an indication of convergence. Therefore, this method guarantees that we find all global minimizers and the global optimal value of a polynomial optimization problem.

Note that the dimensions of the SDP problems also increase quickly with the order of the hierarchy. In practice, we start at a low order and increase it gradually until convergence is reached. For many applications, the convergence fortunately occurs during the first relaxation orders.

4. NUMERICAL RESULTS

For each test, we generate an initial signal $\overline{\mathbf{x}}$ of size $T = 50$ satisfying Constraints (2): a given percentage of elements of $\overline{\mathbf{x}}$, called the degree of sparsity, chosen randomly are set to 0 while the others are set to have their absolute value uniformly selected between $\lambda = 0.7$ and 1. We choose $\mathbf{H}$ as a convolution matrix associated to a finite impulse response filter $\mathbf{h}$ of length 3 whose elements are drawn uniformly in $[0,1]$, i.e. $\mathbf{H}$ is Toeplitz-band with $\mathbf{h}$ defining the elements on the band. The white noise $\mathbf{w}$ has a standard deviation of 0.15. The vector $\mathbf{y}$ contains 1% of outliers and their location have been drawn randomly with equal probability among the T samples. The impulse noise $\mathbf{n}$ has a fixed amplitude set to twice the maximum of $\overline{\mathbf{y}}$ and a random sign with equal probability. The parameter ϵ is set to 10^{-3} and the value of θ for the capped ℓ_2 and Huber functions is tuned to 0.4. The nonlinearity ϕ is chosen as a saturation expressed by the rational function

$$(\forall x \in \mathbb{R}) \quad \phi(x) = \frac{x}{0.3 + |x|} \,.$$

We use GloptiPoly [21] together with the SDP solver SDPT3 [22] to solve the rational problem. In all our simulations, we compute the order 3 SDP relaxation of Lasserre's hierarchy. The problem is formulated using the approach of [18] in order to use the structure of $\mathbf{H}$ and reduce the computational burden. All the simulations have been run on an Intel i7 CPU running at 1.90 GHz with 16 GB of RAM.

We compare our approach for the different fit functions listed in Section 3.1. We run 50 simulations and compare the different relative errors $\|\overline{\mathbf{x}} - \hat{\mathbf{x}}\| \,/\, \|\overline{\mathbf{x}}\|$ between the initial signal $\overline{\mathbf{x}}$ and the signal $\hat{\mathbf{x}}$ estimated with our method.

4.1. Convergence of the SDP hierarchy

We first look at the convergence of Lasserre's hierarchy. For the ℓ_2, the capped ℓ_2, and Huber functions, the convergence has occurred at relaxation order 3. This is certified by a sufficient rank condition implemented in GloptiPoly, which additionally certifies that the optimal point is unique. This is an important feature since, in contrast to many nonconvex optimization methods, we have the theoretical guarantee that the obtained solutions are exactly the global minimizers of Problem (7).

On the other hand, for the ℓ_1 and the pseudo-Huber functions, the convergence does not always occur at order 3. This is shown for the ℓ_1 function in Figure 2 where we draw for the 50 tests, the obtained lower bound in plain blue and the value of the criterion at the computed solution in dotted red. We observe a gap between the two curves that shows that a higher relaxation order would be required for convergence of the hierarchy. Moreover, a consequence for these two fit functions is that the imposed constraints are not always satisfied by the candidate approximate optimal point.

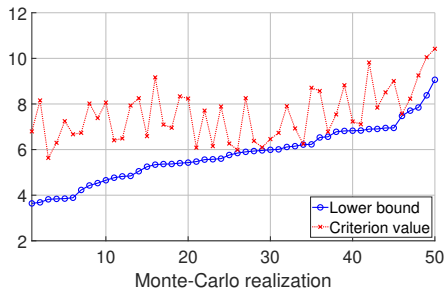

Fig. 2. Convergence study of the Lasserre's hierarchy for the ℓ_1 fit function on 50 tests: In plain blue the value of the lower bound, in dotted red the value of the criterion at the solution. For the sake of clarity, the results are ordered according to the values of the lower bound.

The convergence at a low relaxation order is important for the global minimum guarantee as well as for the applicability of the method by limiting the SDP to a fair size. Therefore in the following section, we focus our attention on the capped ℓ_2 and Huber functions against the ℓ_2 fit.

4.2. Comparison of the robust approach with least squares

We compare here the ℓ_2 with the capped ℓ_2 and Huber fit functions for three different degrees of sparsity of the original signal $\overline{\mathbf{x}}$. Table 1 shows the relative error for the different fit functions. We observe that both robust fit functions yield a smaller error than least squares but the capped ℓ_2 gives the smallest error. This confirms the interest of our methodology, which is able to deal with nonconvex penalty functions. Furthermore, similar results are observed for both sparse and dense signals. Table 2 shows both the true positive rate (TPR) and the false positive rate (FPR) for both ℓ_2 and the capped ℓ_2 functions. We use a threshold value of 0.1 for the peak detection. Note that since Constraints (2) are enforced when solving (7), the threshold value can be taken in $]\epsilon, \lambda[$ indifferently. We notice that for the capped ℓ_2 fit, the average TPR is close to 1, which means that almost all the peaks are well detected, and the FPR is close to zero, i.e. we do not detect peaks at samples originally equal to zero. This is in contrast with the results for the ℓ_2 fit.

Table 1. Statistics on the relative error between $\overline{\mathbf{x}}$ and $\hat{\mathbf{x}}$ with different degrees of sparsity for 50 tests.

	Degree of sparsity	80%	50%	30%
ℓ_2	Average	0.87	0.66	0.54
	Median	0.85	0.66	0.55
Huber	Average	0.65	0.48	0.35
	Median	0.69	0.47	0.33
Capped ℓ_2	Average	0.43	0.36	0.30
	Median	0.34	0.34	0.28

Table 2. Statistics on the TPR and FPR of peak detection between $\overline{\mathbf{x}}$ and $\hat{\mathbf{x}}$ with different degrees of sparsity for 50 tests.

Fit function		Capped ℓ_2			ℓ_2		
Degree of sparsity		80%	50%	30%	80%	50%	30%
TPR	Average	0.96	0.94	0.94	0.84	0.81	0.82
	Median	1.00	0.93	0.95	0.83	0.80	0.83
FPR	Average	0.01	0.10	0.09	0.15	0.22	0.22
	Median	0.04	0.07	0.10	0.12	0.20	0.20

The poor results above for the ℓ_2 fit are due to the outliers together with the convolution matrix $\mathbf{H}$ which makes the estimation inaccurate in the neighborhood of the outliers as illustrated in Figure 3. The latter shows a comparison between a robust and a least squares recovery on a single test. The red curve represents the initial signal $\overline{\mathbf{x}}$ and the blue and the green curves are the estimated signal using respectively the capped ℓ_2 and the ℓ_2 functions. For the capped ℓ_2 function, we observe that the locations of the non-zero samples are well recovered and the amplitudes are close to the ones of $\overline{\mathbf{x}}$. At locations far from any outliers, comparable results are observed for both the capped ℓ_2 and ℓ_2 fit functions. However, close to an outlier value, the estimated signal is prone to many errors using ℓ_2 in contrast with its robust counterpart. This illustrates the benefit of the capped ℓ_2 for robust reconstruction.

5. CONCLUSION

We tackle the issue of robust estimation for an inverse problem that involves both nonconvex constraints and objective function. We show that the resulting optimization problem can be reformulated into a polynomial optimization problem using additional variables. Moreover, our method extends without extra effort nor extra computational cost to a nonlinear model. This problem is then globally solved using recent polynomial optimization techniques that guarantee the global optimality of the solution. Finally, our simulations show the good quality of the signal reconstructed through our method.

6. REFERENCES

[1] G. Lan, C. Hou, and D. Yi, "Robust feature selection via simultaneous capped ℓ_2-norm and $\ell_{2,1}$-norm minimization," in *Proc. IEEE Int. Conf. on Big Data Analysis (ICBDA)*. Mar. 2016, IEEE.

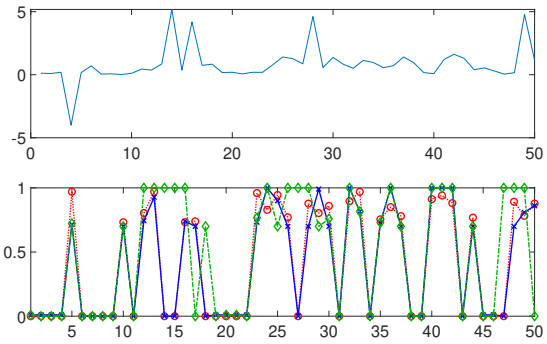

Fig. 3. Reconstruction of a signal with degree of sparsity of 50%. Top: the corrupted observation $\mathbf{y}$, Bottom: in red circle dotted curve, the initial signal $\overline{\mathbf{x}}$ and respectively in blue cross plain curve and green dotted curve, the estimated signal $\hat{\mathbf{x}}$ using the capped ℓ_2 and the ℓ_2 fit functions.

[2] M. Nikolova, "Minimizers of cost-functions involving nonsmooth data-fidelity terms. application to the processing of outliers," *SIAM J. Numer. Anal.*, vol. 40, no. 3, pp. 965–994, Jan. 2002.

[3] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, no. 1, pp. 73–101, Mar. 1964.

[4] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud, "Two deterministic half-quadratic regularization algorithms for computed imaging," in *Proc. Int. Conf. Image Process.* Nov. 1994, vol. 2, pp. 168–172 vol.2, IEEE Comput. Soc. Press.

[5] R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera, *Robust statistics: theory and methods (with R)*, Wiley, Hoboken, NJ, 2019.

[6] A. M. Zoubir, V. Koivunen, E. Ollila, and M. Muma, *Robust Statistics for Signal Processing*, Cambridge University Press, Oct. 2018.

[7] G. Drašković and F. Pascal, "New insights into the statistical properties of M-estimators," *IEEE Trans. Signal Process.*, vol. 66, no. 16, pp. 4253–4263, Aug. 2018.

[8] T. Blumensath, "Sampling and reconstructing signals from a union of linear subspaces," *IEEE Trans. Inform. Theory*, vol. 57, no. 7, pp. 4660–4671, July 2011.

[9] Y. Lu and M. Do, "A theory for sampling signals from a union of subspaces," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2334–2345, June 2008.

[10] G. Ongie, R. Willett, R. D. Nowak, and L. Balzano, "Algebraic variety models for high-rank matrix completion," in *Proc. Int. Conf. Mach. Learn.* Aug. 2017, vol. 70, pp. 2691–2700, PMLR.

[11] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Trans. Inform. Theory*, vol. 55, no. 11, pp. 5302–5316, Nov. 2009.

[12] C. Hedge, P. Indyk, and L. Schmidt, "Fast recovery from a union of subspaces," in *Advances in Neural Information Processing Systems 29*, pp. 4394–4402. Curran Associates, Inc., 2016.

[13] M. S. Asif and C. Hegde, "Phase retrieval for signals in union of subspaces," in *Proc. IEEE Global Conf. Signal Information Process.* Nov. 2018, pp. 356–359, IEEE.

[14] A. Marmin, M. Castella, and J.-C. Pesquet, "How to globally solve non-convex optimization problems involving an approximate ℓ_0 penalization," in *Proc. Int. Conf. Acoust. Speech Signal Process.* May 2019, pp. 5601–5605, IEEE.

[15] G. Dahlquist and Å. Björck, *Numerical Methods in Scientific Computing, Volume I*, Society for Industrial and Applied Mathematics, Jan. 2008.

[16] R. D. B. Brotto, K. Nose-Filho, and J. M. T. Romano, "Antisparse blind source separation," in *Proc. Signal Processing with Adaptive Sparse Structured Representations (SPARS) workshop*, July 2019.

[17] J. B. Lasserre, *Moments, Positive Polynomials and Their Applications*, Imperial College Press, 2009.

[18] M. Castella, J.-C. Pesquet, and A. Marmin, "Rational optimization for nonlinear reconstruction with approximate ℓ_0 penalization," *IEEE Trans. Signal Process.*, vol. 67, no. 6, pp. 1407–1417, Mar. 2019.

[19] J. Nie, "Optimality conditions and finite convergence of Lasserre's hierarchy," *Math. Programm.*, vol. 146, no. 1-2, pp. 97–121, May 2013.

[20] D. Henrion and J.-B. Lasserre, "Detecting global optimality and extracting solutions in GloptiPoly," in *Positive Polynomials in Control*, vol. 312, pp. 293–310. Springer Berlin Heidelberg, Sept. 2005.

[21] D. Henrion, J.-B. Lasserre, and J. Löfberg, "GloptiPoly 3: moments, optimization and semidefinite programming," *Optim. Methods Softw.*, vol. 24, no. 4-5, pp. 761–779, Oct. 2009.

[22] K. C. Toh, M. J. Todd, and R. H. Tütüncü, "SDPT3 — a Matlab software package for semidefinite programming, version 1.3," *Optim. Methods Softw.*, vol. 11, no. 1-4, pp. 545–581, Jan. 1999.