ROBUST RLS VIA THE NONCONVEX SPARSITY PROMPTING PENALTIES OF OUTLIER COMPONENTS

Longshuai Xiao, Ming Wu, Jun Yang and Jing Tian

The Key Laboratory of Noise and Vibration Research and The State Key Laboratory of Acoustics
Institute of Acoustics, Chinese Academy of Sciences, Beijing, 100190, China

ABSTRACT

Recursive least square (RLS) is a ubiquitous adaptive filtering algorithm used in general adaptive signal processing applications. However, it is well known that RLS is sensitive to outlier contaminated in measurements. Traditional robust RLS has difficulties to cope with correlated ambient noise. To provide robustness in such cases, a robust RLS via outlier pursuit (RRLSvOP) framework is proposed with outlier's sparsity control via possibly nonconvex penalties, such as the minimax concave penalty (MCP). Because of the nonconvexity of the proposed model, more advanced numerical procedures with convergence guarantees, such as multi-stage convex relaxation (MSCR), coordinate descent (CD), proximal gradient (PG) and PG with homotopy (PGH), are adapted in the online update stage, while the initialization stage is solved via MSCR strategy. Simulations demonstrate improved robustness of the model using nonconvex penalties via these procedures in comparison with that using ℓ_1 penalty.

Index Terms— Recursive Least Square, Sparsity Prompting Model, Robust Statistics, Nonconvex Penalty

1. INTRODUCTION

Adaptive filters have wide applications in the general signal processing applications, such as acoustic echo/feedback cancellation, active noise control, source localization and so on [1]. Among these adaptive filters, RLS enjoys fast convergence with acceptable computational and storage complexity. However, RLS is very sensitive to the interferences that are rarely appeared but with strong transient energy, i.e. outlier [2]. Based on robust statistics [3], robust RLS has been reported to recursively minimize some modified nonconvex cost function [4][5]. However,these models cannot handle the correlated ambient nominal noise with known or estimated covariance matrix.

On the other hand, instead to design proper cost functions to provide robustness in the paradigm of robust statistics, outlier' sparsity model has been explored to estimate outlier explicitly [6]. Successful applications are error correction [7], robust PCA[8] and outlier detection [9]. Particularly, this model has been introduced to robust RLS [10] and Kalman filter [11], where ℓ_1 norm penalty of outlier vector is used to solve a sparse optimization sub-problem.

However, it's well known that when the target sparse signal is strong, ℓ_1 norm model often introduces a sub-optimal solution, since it is only a loose relaxation of the ℓ_0 norm [12]. Naturally, typical concave penalties such as SCAD penalty[13] and MCP [14], hence, are motivated to use to solve the outlier sparsity prompting nonconvex optimization sub-problem. Moreover, recent computational advance in optimizing this kind of nonconvex penalization problems, such as multistage local linearized approximation (MLLA/MSCR) [15][16], CD [17], PG [18] and PGH[19], is another motivation to generalize original ℓ_1 norm penalized RRLSvOP to the possibly nonconvex penalized RRLSvOP.

1.1. Previous work

This work is to generalize ℓ_1 penalty used in [10] to possibly nonconvex ones. The usage of nonconvex penalties is highly motivated because of the bias problem induced by ℓ_1 penalty when target signals are strong. ADMM fails to deal with this nonconvex optimization problem. CD works well with convergence assured. MLLA, PG, PGH methods are also modified to solve the nonconvex model with different properties, such as high efficiency, with convergence guarantees and so on. Hence the generalized RRLSvOP in the work shares the same sparse outlier pursuit idea as [10], but differs in both the nonconvex penalty models and computational procedures.

2. THE PROPOSED RRLSVOP FRAMEWORK

2.1. Model

Assume that observations obey to the following linear system model

$$\mathbf{y}(n) = \mathbf{X}^T(n)\mathbf{w}_o(n) + \mathbf{o}(n) + \mathbf{v}(n), n = 1, 2, \ldots \quad (1)$$

where,$\mathbf{w}_o(n) \in \mathbb{R}^L$, $\mathbf{X}(n) \in \mathbb{R}^{L\times D}$, $\mathbf{y}(n), \mathbf{o}(n), \mathbf{v}(n) \in \mathbb{R}^D$, $\mathbf{v}(n) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. And $\mathbf{X}(k)$, $\mathbf{o}(k)$,$\mathbf{v}(k)$, $\forall k = 1, \ldots, n$,are independent from each other. Given measurements $\{\mathbf{y}(k), \mathbf{X}(k)\}_{k=1}^n$, our

This work is supported by National Natural Science Foundation of China(Grant No.11474306, No.11174317, No.11474307, No.11404367).
E-mail:xls.ioa@gmail.com, mingwu@mail.ioa.ac.cn, jyang@mail.ioa.ac.cn and tian@mail.ioa.ac.cn

task is, at time instant n, to estimate $\mathbf{w}_o(n)$ under the perturbation of ambient noise $\{\mathbf{v}(k)\}_{k=1}^n$ and outlier interference $\{\mathbf{o}(k)\}_{k=1}^n$.

The RLS solves,

$$\min_{\hat{\mathbf{w}}(n)\in\mathbb{R}^L} \frac{1}{2}\sum_{k=1}^{n}\beta^{n-k}\|\mathbf{y}(k)-\mathbf{X}^T(k)\mathbf{w}(n)\|_{\mathbf{\Sigma}^{-1}}^2. \tag{2}$$

.

The proposed robust RLS solves

$$(\hat{\mathbf{w}}(n),\hat{\mathbf{o}}(1:n)) = \arg\min_{\mathbf{w}(n),\mathbf{o}(1:n)} f(\mathbf{w}(n),\mathbf{o}(1:n)), \tag{3}$$

where,

$$f(\mathbf{w}(n),\mathbf{o}(1:n)) \triangleq \frac{1}{2}\sum_{k=1}^{n}\beta^{n-k}\left[\bar{\mathcal{L}}(\mathbf{w}(n),\mathbf{o}(k)) + \lambda\mathcal{P}(\mathbf{o}(k))\right] \tag{4}$$

with notations $\bar{\mathcal{L}}(\mathbf{w}(n),\mathbf{o}(k)) \triangleq \|\mathbf{y}(k)-\mathbf{X}^T(k)\mathbf{w}(n)-\mathbf{o}(k)\|_{\mathbf{\Sigma}^{-1}}^2$ and $\lambda\mathcal{P}(\cdot)$ as some sparsity prompting penalty defined in the next subsection.

2.2. Typical Penalties and Thresholding Functions

Assume $\lambda\mathcal{P}(\mathbf{o};\lambda,\theta) = \lambda\sum_{d=1}^{D}\mathcal{P}(o_d)$. Then we define several penalties, their concave-convex decompositions and proximal operators. Penalty definitions,

$$\lambda\mathcal{P}(o) = \begin{cases} \lambda|o| & \ell_1 \\ \lambda\int_0^{|o|}\left\{\mathbb{I}(z\le\lambda) + \frac{(\theta\lambda-z)_+}{(\theta-1)\lambda}\mathbb{I}(z>\lambda)\right\}dz & SCAD \\ \lambda\int_0^{|o|}\left(1-\frac{z}{\theta\lambda}\right)_+ dz & MCP, \end{cases} \tag{5}$$

where $(x)_+ = max(x,0)$, $\mathbb{I}(\cdot)$ is an indicator function, for SCAD $\theta > 2$, for MCP $\theta > 1$. SCAD and MCP can be formulated as ℓ_1 plus a concave part as $\lambda\mathcal{P}(o) = Q(o) + \lambda|o|$, where

$$Q(o) = \begin{cases} \frac{2\lambda|o|-o^2-\lambda^2}{2(\theta-1)}\mathbb{I}(\lambda<|o|\le\theta\lambda) + \frac{(\theta+1)\lambda^2-2\lambda|o|}{2}\mathbb{I}(|o|>\theta\lambda) \\ -\frac{o^2}{2\theta}\mathbb{I}(|o|\le\theta\lambda) + \left(\frac{\theta\lambda^2}{2}-\lambda|o|\right)\mathbb{I}(|o|>\theta\lambda). \end{cases} \tag{6}$$

The shrinkage thresholding function is the corresponding proximal operator, defined as $S_{\lambda\gamma^{-1}\mathcal{P}}(\bar{o}) = \arg\min_o \frac{1}{2}(\bar{o}-o)^2 + \lambda\gamma^{-1}\mathcal{P}(o)$. For ℓ_1, $S_{\lambda\gamma^{-1}\mathcal{P}}(\bar{o}) = sgn(\bar{o})(|\bar{o}|-\lambda\gamma^{-1})_+$. And MCP for example,

$$S_{\frac{\lambda}{\gamma}\mathcal{P}}(\bar{o}) = \begin{cases} \frac{\gamma\theta}{\gamma\theta-1}sgn(\bar{o})(|\bar{o}|-\frac{\lambda}{\gamma})_+ & |\bar{o}|\leqslant\theta\lambda \\ \bar{o} & |\bar{o}|>\theta\lambda. \end{cases} \tag{7}$$

2.3. Batch Initialization Stage

Initialization stage uses first n_0 samples to generate a good initial weight estimate and choose best λ_o for the corresponding penalty. Though the loss function $\bar{\mathcal{L}}$ is jointly convex over $(\mathbf{w}(n),\mathbf{o}(k))$, the penalty $\mathcal{P}$ may be concave. Multistage convex relaxation (MSCR or MLLA) [15] is used to relax the concave penalty to its local linear approximation. That is, we iteratively solve this cost function

$$(\hat{\mathbf{w}}^{(j)}(n_0),\hat{\mathbf{o}}^{(j)}) = \arg\min_{(\mathbf{w},\mathbf{o})} \bar{f}(\mathbf{w}(n_0),\mathbf{o};\hat{\mathbf{o}}^{(j-1)}) \tag{8}$$

where $\bar{f}(\mathbf{w}(n_0),\mathbf{o};\hat{\mathbf{o}}^{(j-1)}) \triangleq \frac{1}{2}\sum_{k=1}^{n_0}\beta^{n_0-k}\bar{\mathcal{L}}(\mathbf{w}(n),\mathbf{o}(k)) + \lambda\sum_{d=1}^{D}\mathcal{P}'(\hat{o}_d^{(j-1)}(k))|o_d(k)|$ and $\mathbf{o} \triangleq [\mathbf{o}(n_0)^T,\ldots,\mathbf{o}(1)^T]^T$. Thus, at j-th iteration, the new cost function is convex, and can be solved with CVX[20].This stage is executed several times to choose best λ according to the criterion used in [11].

Algorithm 1 Robust RLS via Outlier Pursuit (RRLSvOP)

Choose penalty $\mathcal{P} \in \{\ell_1, SCAD, MCP\}$ and sub-problem solver $\mathcal{M} \in \{MLLA-CVX, CD, PG, PGH\}$.
Batch Init.:
$(\hat{\mathbf{w}}(n_0),\hat{\mathbf{o}}(1:n_0),\lambda_o) \leftarrow$ Solve (8) with $n = n_0$ and $\lambda\mathcal{P}(\cdot)$, choose λ according to [11]
Online Update:
while $(n \geqslant n_0)$ **do**
  $n \leftarrow n+1$
  $\mathbf{e}(n) = \mathbf{y}(n) - \mathbf{X}^T(n)\hat{\mathbf{w}}(n-1)$
  Solve $\hat{\mathbf{o}}(\mathbf{e}(n),\mathcal{P},\mathcal{M})$
  Update RLS via (10)(11)(12).
end while

2.4. Online Update Stage

At the online update stage, it is impractical to re-estimate past outlier[10]. Therefore, we only estimate the most recent outlier. At time instant n, previous estimate $\hat{\mathbf{w}}(n-1)$ is used to estimate $\hat{\mathbf{o}}(n)$, that is,

$$\hat{\mathbf{o}}(n) = \arg\min_{\mathbf{o}} \frac{1}{2}\|\mathbf{e}(n)-\mathbf{o}\|_{\mathbf{\Sigma}^{-1}}^2 + \lambda\mathcal{P}(\mathbf{o}) \tag{9}$$

where $\mathbf{e}(n) = \mathbf{y}(n) - \mathbf{X}^T(n)\hat{\mathbf{w}}(n-1)$. This sub-problem will be deal with in the next section. With $\hat{\mathbf{o}}(n)$, we update weight as

$$\begin{aligned}\hat{\mathbf{w}}(n) &= \arg\min_{\mathbf{w}(n)} \frac{1}{2}\sum_{k=1}^{n}\beta^{n-k}\|\mathbf{y}(k)-\mathbf{X}^T(k)\mathbf{w}(n)-\hat{\mathbf{o}}(k)\|_{\mathbf{\Sigma}^{-1}}^2 \\ &= \hat{\mathbf{w}}(n-1) + \boldsymbol{\kappa}(n)\left[\mathbf{y}(n)-\mathbf{X}^T(n)\hat{\mathbf{w}}(n-1)-\hat{\mathbf{o}}(n)\right]\end{aligned} \tag{10}$$

$$\boldsymbol{\kappa}(n) \triangleq \beta^{-1}\mathbf{A}^{-1}(n-1)\mathbf{X}(n)\left[\mathbf{\Sigma} + \mathbf{X}^T(n)\beta^{-1}\mathbf{A}^{-1}(n-1)\mathbf{X}(n)\right]^{-1} \tag{11}$$

$$\mathbf{A}^{-1}(n) = \beta^{-1}\mathbf{A}^{-1}(n-1) - \boldsymbol{\kappa}(n)\mathbf{X}^T(n)\beta^{-1}\mathbf{A}^{-1}(n-1). \tag{12}$$

where $\mathbf{A}(n) \triangleq \sum_{k=1}^{n}\beta^{n-k}\mathbf{X}(k)\mathbf{\Sigma}^{-1}\mathbf{X}^T(k)$.

The proposed framework is listed in the **Algorithm** (1) where the sub-problem in the algorithm is solved in the next section.

3. THE OUTLIER PURSUIT SUB-PROBLEM SOLVERS

This section is to use several state-of-the-art solvers for nonconvex penalized least square regression task to solve the particular sub-problem in the proposed RRLSvOP framework. The sub-problem is to optimize

$$\hat{\mathbf{o}}(\bar{\mathbf{o}}, \mathcal{P}, \mathcal{M}) \triangleq \arg\min_{\mathbf{o}\in\mathbb{R}^D} \mathcal{L}(\mathbf{o}; \bar{\mathbf{o}}, \Sigma) + \lambda\mathcal{P}(\mathbf{o}), \tag{13}$$

where $\hat{\mathbf{o}}(\bar{\mathbf{o}}, \mathcal{P}, \mathcal{M})$ denote that $\hat{\mathbf{o}}$ is solved at $\bar{\mathbf{o}}$ with penalty $\mathcal{P}$ using method $\mathcal{M}$, the loss function $\mathcal{L}(\mathbf{o}; \bar{\mathbf{o}}, \Sigma) \triangleq \frac{1}{2}\|\bar{\mathbf{o}} - \mathbf{o}\|^2_{\Sigma^{-1}}$, Σ is noise covariance matrix and assumed to be positive definite. The loss function is strong convex and has Lipschitz continuous gradient with strictly positive Lipschitz constant $\lambda^{-1}_{max}(\Sigma)$. The good property of the loss function make it guaranteed to solve the sub-problem stably and/or efficiently with even nonconvex penalties. Four such kind of solvers with different properties are adapted to solve this problem.

3.1. MLLA-CVX

This method is the same one as that is used in the previous section for batch initialization, but with different cost function.

$$\hat{\mathbf{o}}^{(k)} = \arg\min_{\mathbf{o}} \mathcal{L}(\mathbf{o}; \bar{\mathbf{o}}, \Sigma) + \lambda \sum_{d=1}^{D} \mathcal{P}'(\hat{o}_d^{(k-1)})|o_d| \tag{14}$$

where $k = 1, \ldots, K$, K is the maximum iteration number, $\mathcal{P}'(\cdot)$ is the first order derivative of $\mathcal{P}(\cdot)$ defined in the previous section and when $k = 1$, set the weighted term of new penalty to 1. When the penalty function is the logarithmic function, this algorithm is corresponding to reweighted L1 (RWL1) [16], that is used in robust Kalman filter [11]. MLLA has been guaranteed to generate a better solution than ℓ_1 penalized form [15]. However, the computational cost of this method may be too high to be used, therefore, computational efficient methods must be considered.

3.2. CD

CD solves the sub-problem in the manner of one-at-a-time coordinate-wise update. Initialized with proper values, then iteratively cycle through each coordinate update [17].

$$\begin{aligned}\hat{o}_d^{(k)} &= \arg\min_{o_d} \frac{1}{2}\|\bar{\mathbf{o}} - \mathbf{M}(o_d)\|^2_{\Sigma^{-1}} + \lambda\mathcal{P}(o_d) \\ &= S_{\lambda\sigma_d^{-1}\mathcal{P}}(\tilde{o}_d)\end{aligned} \tag{15}$$

where $\mathbf{M}(o_d) \triangleq \left[\hat{o}_1^{(k)}, \ldots, \hat{o}_{d-1}^{(k)}, o_d, \hat{o}_{d+1}^{(k-1)}, \ldots, \hat{o}_D^{(k-1)}\right]^T$, $d = 1 : D$,$k = 1 : K$, $S_{\lambda\sigma_d^{-1}\mathcal{P}}(\cdot)$ has been defined in the previous section, $\sigma_d = (\Sigma^{-1})_{d,d}$ and

$$\tilde{o}_d = \sigma_d^{-1}((\Sigma^{-1}\bar{\mathbf{o}})_d - \sum_{i=1}^{d-1}(\Sigma^{-1})_{i,d}\hat{o}_i^{(k)} - \sum_{i=d+1}^{D}(\Sigma^{-1})_{i,d}\hat{o}_i^{(k-1)}). \tag{16}$$

Lemma 1 ([17]) *For possibly non-convex penalties, such as ℓ_1 norm, SCAD and MCP, the sequence $\hat{\mathbf{o}}^{(k)}$ generated by CD method* (15) *will converge to the stationary point of the outlier pursuit sub-problem* (13).

Though CD converges fast to the stationary point, there are no prediction error bounds available in the literature.

3.3. PG

Since the loss function in (13) has continuous Lipschitz gradient and penalties have simple structure, it is natural to consider Nesterov's PG method to optimize the problem [21]. Furthermore, PG has been used to solve nonconvex penalized optimization problem with convergence assured [22][18].

At k-th iteration, PG minimizes a quadratic approximation $\mathcal{L}_k(\mathbf{o}; \hat{\mathbf{o}}^{(k-1)}) \triangleq \mathcal{L}(\hat{\mathbf{o}}^{(k-1)}) + < \nabla\mathcal{L}(\hat{\mathbf{o}}^{(k-1)}), \mathbf{o} - \hat{\mathbf{o}}^{(k-1)} > + \frac{\gamma_k}{2}\|\mathbf{o} - \hat{\mathbf{o}}^{(k-1)}\|_2^2$ of the loss function near $\hat{\mathbf{o}}^{(k-1)}$,

$$\begin{aligned}\hat{\mathbf{o}}^{(k)} &= \arg\min_{\mathbf{o}} \mathcal{L}_k(\mathbf{o}; \hat{\mathbf{o}}^{(k-1)}) + \lambda\mathcal{P}(\mathbf{o}) \\ &= S_{\lambda\gamma_k^{-1}\mathcal{P}}(\mathbf{u}^{(k-1)})\end{aligned} \tag{17}$$

where $S_{\lambda\sigma_d^{-1}\mathcal{P}}(\cdot)$ has been defined in the previous section, $\mathbf{u}^{(k-1)} = \hat{\mathbf{o}}^{(k-1)} - \gamma_k^{-1} \nabla \mathcal{L}(\hat{\mathbf{o}}^{(k-1)})$, γ_k^{-1} is a step-size parameter that can be designed via monotone linear search criterion, $(\mathcal{L}(\hat{\mathbf{o}}^{(k)}(n)) \leqslant \mathcal{L}(\hat{\mathbf{o}}^{(k-1)}(n)) - \frac{\sigma}{2}\gamma_k\|\hat{\mathbf{o}}^{(k)}(n) - \hat{\mathbf{o}}^{(k-1)}(n)\|_2^2)$, where σ is a small constant. Convergence result is straightforward to be adapted from literature.

Lemma 2 ([22, 18]) *For possibly non-convex penalties, such as ℓ_1 norm, SCAD and MCP, the sequence $\hat{\mathbf{o}}^{(k)}$ generated by PG method* (17) *will converge to the stationary point of the outlier pursuit sub-problem* (13).

PG is simple to be implemented and scalable to large scale optimization, however, its convergence may be too slow.

3.4. PGH

Homotopy strategy can be used to accelerate PG's convergence rate, as fast local geometric convergence near sparse solutions has been observed [19]. We use a sequence of $\lambda_t = \eta^t\lambda_0, t = 1, \ldots, N$, to iteratively optimize the sub-problem (13). At stage t, given λ_t, a PG sub-procedure can be utilized with initial estimate and step-size as the output of the last stage. And at each stage, a gradually reduced termination bound $\epsilon^{(t)} = \delta_{prec}\lambda_t$ is used to check whether the current estimate gets close to a local optimum. To simplify computation, new loss function is formed as $\tilde{\mathcal{L}}(\mathbf{o}) = \mathcal{L}(\mathbf{o}) + \mathcal{Q}(\mathbf{o})$, where the later term is the concave part of the original penalty, the new penalty is ℓ_1 norm, hence the new subproblem is

$$\begin{aligned}\hat{\mathbf{o}}^{(k)} &= \arg\min_{\mathbf{o}} \tilde{\mathcal{L}}_k(\mathbf{o}; \hat{\mathbf{o}}^{(k-1)}) + \lambda_t\|\mathbf{o}\|_1 \\ &= S_{\lambda_t\gamma_k^{-1}\|\cdot\|_1}(\tilde{\mathbf{u}}^{(k-1)})\end{aligned} \tag{18}$$

where $\tilde{\mathcal{L}}_k$ is the local quadratic approximation of the new cost function, and $\tilde{\mathbf{u}}^{(k-1)} = \hat{\mathbf{o}}^{(k-1)} - \gamma_k^{-1} \triangledown \tilde{\mathcal{L}}(\hat{\mathbf{o}}^{(k-1)})$. It has been proven that PGH for this kind of possibly nonconvex optmization problems enjoys global geometric convergence to an oracle solution [19]. Numerical results in this work have confirmed the superior performance of this strategy. Besides the stop criterion, maximum iteration for PG is also used without degrading its performance.

4. NUMERICAL RESULTS

The setup is the same as the stationary case in [10]. $\theta = 2$ is used for MCP. We only consider stationary case, where $\mathbf{w}_o(n) = \mathbf{w}_o$. $L = 20, D = 10, \text{SNR} = 2L\sigma_x^2 = 2$, $\mathbf{w}_o$ is all-one vector, with i.i.d. standard gaussian noise added. Each element of the $\mathbf{X}(n)$ are i.i.d. Gaussian distribution with variance σ_x^2. At each time instant, colored Gaussian ambient noise vector $\mathbf{v}(n)$ is generated from a first order AR filter with pole at 0.95 to an independent standard white Gaussian distribution. Outlier are modeled by Bernoulli distribution with probability 0.2 to generate a non-zeros value, which is drawn from a uniform distribution with zero-mean and variance 20000. The root mean square error RMSE=$\|\hat{\mathbf{w}}(n) - \mathbf{w}_o\|_2$ is used as performance metric. 2000 samples are used in the simulation, where the first $n_0 = 20$ samples are used for the batch initialization, the rest is used for online update. $\theta = 2$ is used for MCP. All the results are averaged over 50 independent runs.

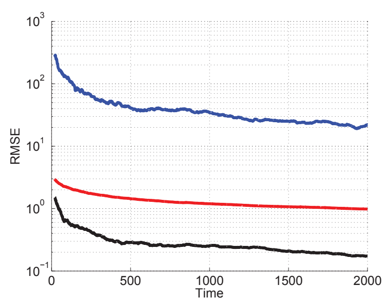

Fig. 1. RMSE convergence curves for RRLSvOP. The ℓ_1 penalty and MCP model are solved via MLLA-CVX solver at the initialization stage, then at the online update stage PGH are used. In the figure, each solid line corresponds to the result for, from top to down, RRLS, RRLSvOP-L1 and RRLSvOP-MCP respectively.

The RMSE convergence curves are plotted in Fig.1 and the run-time of each algorithm is summarized in Table 1. Since all four strategies, i.e., MLLA-CVX, CD, PG and PGH have similar curves, we only show the result for PGH in the figure. It can be seen from the figure that the proposed RRLSvOP framework with ℓ_1 penalty and MCP solved by PGH, can provide impressively robustness against outlier compared with the ordinary RLS result. Moreover, the results for ℓ_1 penalty are overwhelmingly outperformed by the advocated nonconvex penalties, such as MCP. From the Table 1 we can see that PGH has the best efficiency for ℓ_1 penalty and behaves similarly as CD and PG for MCP.

Table 1: Time consumption for different solvers(seconds)

	MLLA-CVX	CD(100)	PG(100)	PGH
ℓ_1	516.6	37.1	143.9	25.6
MCP	2401.4	39.7	76.5	118.5

The comparison between RRLSvOP with MCP and R-RLSvOP with ℓ_1 penalty is further performed with a large range of percentage of outlier. The result is shown in Fig. 2, where both the batch initialization results with MLLA-CVX solver and online update RMSE results with PGH solver are plotted. Up to 50 percentage of outlier, RRLSvOP can outperform RLS in both stages, where MCP results are remarkably superior to the ℓ_1 results [10].

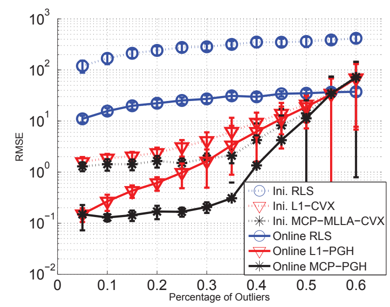

Fig. 2. RMSE vs percentage of outlier. Dotted lines are results for batch initialization stage, while solid ones are for online stage.

5. CONCLUTION

A robust RLS via outlier pursuit framework has been proposed to incorporate the original one [10] as the special case using ℓ_1 norm penalty. The proposed model generalize the original one in two aspects, firstly, concave penalties such as SCAD and MCP are also used to further improve the performance of RLS, especially when outlier are strong ,ℓ_1 known to be suboptimal. Secondly, to optimize nonconvex problems, advanced procedures, such as MLLA-CVX, CD, PG and PGH have been adapted with either strong theoretical guarantees or better computational efficiency. Furthermore, MLLA-CVX is used also in the batch initialization stage which further improves the performance of the online update with better initial weight estimate and optimal λ. The superior performance of the generalized model over the original one is verified in the numerical experiments, up to 50 percentage of outlier contaminations, the proposed model provides improved robustness over both non-robust one and the original one.

6. REFERENCES

[1] Ali H. Sayed, *Fundamentals of Adaptive Filtering*, Wiley-IEEE Press, 2003.

[2] Abdelhak M. Zoubir, Visa Koivunen, Yacine Chakhchoukh, and Michael Muma, "Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts," *IEEE Signal Processing Magazine*, vol. 29, no. 4, pp. 61–80, 2012.

[3] Peter J. Huber and Elvezio M. Ronchetti, *Robust Statistics*, Wiley, 2009.

[4] Shing-Chow Chan and Yue-Xian Zou, "A recursive least m-estimate algorithm for robust adaptive filtering in impulsive noise: fast algorithm and convergence performance analysis," *IEEE Transactions on Signal Processing*, vol. 52, no. 4, pp. 975–991, 2004.

[5] P. Petrus, "Robust huber adaptive filter," *IEEE Transactions on Signal Processing*, vol. 47, no. 4, pp. 1129–1133, 1999.

[6] G.B. Giannakis, G. Mateos, S. Farahmand, V. Kekatos, and H. Zhu, "Uspacor: Universal sparsity-controlling outlier rejection," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 1952–1955.

[7] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[8] Emmanuel J. Candes, Xiaodong Li, Yi Ma, and John Wright, "Robust principal component analysis?," *J. ACM*, vol. 58, pp. 11, 2011.

[9] Yiyuan She and Art B. Owen, "Outlier detection using nonconvex penalized regression," *Journal of the American Statistical Association*, vol. 106, no. 494, pp. 626–639, 2011.

[10] Shahrokh Farahmand and Georgios G. Giannakis, "Robust rls in the presence of correlated noise using outlier sparsity," *IEEE Transaction on Signal Processing*, vol. 60, no. 6, pp. 3308–3313, 2012.

[11] Shahrokh Farahmand, Georgios B. Giannakis, and Daniele Angelosante, "Doubly robust smoothing of dynamical processes via outlier sparsity constraints.," *IEEE Transactions on Signal Processing*, vol. 59, no. 10, pp. 4529–4543, 2011.

[12] Cun-Hui Zhang and Tong Zhang, "A general theory of concave regularization for high-dimensional sparse estimation problems," *Statistical Science*, vol. 27, no. 4, pp. 576–593, 2012.

[13] Jianqing Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, pp. 1348–1360, 2001.

[14] Cun-Hui Zhang, "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, vol. 38, no. 2, pp. 894–942, 2010.

[15] Tong Zhang, "Analysis of multi-stage convex relaxation for sparse regularization," *The Journal of Machine Learning Research*, vol. 11, pp. 1081–1107, 2010.

[16] E. J. Candes, M. B. Wakin, and S. Boyd, "Enhancing sparsity by reweighted l1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, 2008.

[17] Rahul Mazumder, Jerome H. Friedman, and Trevor Hastie, "Sparsenet: Coordinate descent with nonconvex penalties," *Journal of the American Association*, vol. 106, no. 495, pp. 1125–1138, 2011.

[18] Po-Ling Loh and Martin J. Wainwright, "Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima," *arXiv preprint arXiv:1305.2436*, pp. 1–45, 2013.

[19] Zhaoran Wang, Han Liu, and Tong Zhang, "Optimal computational and statistical rates of convergence for sparse nonconvex learning problems," *arXiv preprint arXiv:1306.4960v2*, pp. 1–49, 2013.

[20] Michael Grant and Stephen Boyd, "CVX: Matlab software for disciplined convex programming, version 2.0 beta," http://cvxr.com/cvx, sep 2013.

[21] Yu. Nesterov, "Gradient methods for minimizing composite objective function," CORE Discussion Papers 2007076, Universit catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2007.

[22] Pinghua Gong, Changshui Zhang, Zhaosong Lu, Jianhua Z. Huang, and Jieping Ye, "A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems," in *ICML*, 2013.