# Active Regression with Compressive-Sensing based Outlier Mitigation for Both Small and Large Outliers

Jian Zheng Xiaohua Li

Electrical and Computer Engineering
Binghamton University, State University of New York

IEEE Global Conference on Signal & Information Processing 2016

# Outline

# Major contributions

- Proposed a new active learning scheme for linear regression problems with the objective of resolving the unreliable training data labeling problem
    - Proposed two small outlier models
    - Developed a way to convert non-sparse small outliers to sparse large outliers
    - Successfully removed sparse large outliers and non-sparse small outliers

# Motivations

- Active regression: minimize the amount of training data used in regression problems by looking for the most informative ones
  - It outperforms conventional passive regression
  - But may introduce heavier labeling errors
- Human labeling errors
  - Sparse large outliers
  - Non-sparse small outliers

# General linear regression

- We consider the general linear regression

$$y_i = \mathbf{x}_i'\boldsymbol{\theta} + \epsilon_i + h_i v_i + o_i, \tag{1}$$

$y_i$: data label
$\mathbf{x}_i$: $N \times 1$ data vector
$\boldsymbol{\theta}$: $N \times 1$ regression vector
$h_i v_i$: small outliers with a scalar factor $h_i$
$o_i$: large outliers
$\epsilon_i$: noise with zero-mean and variance $\sigma_\epsilon^2$

- Select $T$ of the most informative training samples out of $I$ data vectors using a pool-based active learning method proposed by Sugiyama, etc.

# Large outlier mitigation

- Conventionally, use the joint optimization to estimate $\mathbf{o}$

$$\{\hat{\boldsymbol{\theta}}, \hat{\mathbf{o}}\} = \arg \min_{\{\boldsymbol{\theta}, \mathbf{o}\}} \|\mathbf{y}_{tr} - \mathbf{o} - \mathbf{X}_{tr}\boldsymbol{\theta}\| + \lambda_1 \|\mathbf{o}\|_1. \tag{2}$$

where

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{X}'_{tr}\mathbf{X}_{tr}\right)^{-1} \mathbf{X}'_{tr}(\mathbf{y}_{tr} - \hat{\mathbf{o}}). \tag{3}$$

- When substitute $\hat{\boldsymbol{\theta}}$ to (2), the problem becomes

$$\hat{\mathbf{o}} = \arg \min_{\mathbf{o}} \left\| \left(\mathbf{I} - \mathbf{X}_{tr}(\mathbf{X}'_{tr}\mathbf{X}_{tr})^{-1}\mathbf{X}'_{tr}\right)(\mathbf{y}_{tr} - \mathbf{o}) \right\| + \lambda_1 \|\mathbf{o}\|_1. \tag{4}$$

# Small outliers

- Assume the $T$ training data are labeled by $L$ labelers and each labeler labels $T_L = \frac{T}{L}$ data. The $\ell$th labeler labels the training data set $(\mathbf{X}_\ell, \mathbf{y}_\ell)$.
- Each of the $\ell$th labeler has a common outlier value $v_\ell$, which is added to the labeling values via the weighting vector

$$\mathbf{h}_\ell = [h_{(\ell-1)T_L+1}, \cdots, h_{\ell T_L}]', \quad \ell = 1, \cdots, L. \qquad (5)$$

- We assume that $\|\mathbf{h}_\ell\| = 1$ and $|v_\ell| \gg \sigma_\epsilon$ if $v_\ell \neq 0$.

# Small outlier models

- Small outlier model 1: Assume that the weighting vector $\mathbf{h}_\ell$ of each labeler $\ell$ is unknown, but all the labelers have the same weighting vector, i.e.,

$$\mathbf{h}_\ell = \mathbf{h} = [h_1, \cdots, h_{T_L}]'. \tag{6}$$

- Small outlier model 2: Assume that the weighting vectors $\mathbf{h}_\ell$ for the $L$ labelers are different from each other, where $\ell = 1, 2, \cdots, L$. But the weighting vectors are assumed known a priori.

# Weighting vector estimation for model 1

- For Model 1, by collecting all the $L$ users' labeled data $\mathbf{y}_\ell$, where $\ell = 1, \cdots, L$, we can estimate the common weighting vector $\mathbf{h}$ by solving the following maximization

$$\hat{\mathbf{h}} = \arg\max_{\mathbf{h}} E\left[\|\mathbf{h}'(\mathbf{y}_\ell - \hat{\mathbf{o}}_\ell)\|^2\right] \quad (7)$$

$$= \arg\max_{\mathbf{h}} \ \mathbf{h}'\mathbf{R}_y\mathbf{h}, \quad \text{s.t.,} \quad \|\mathbf{h}\| = 1 \quad (8)$$

where $\mathbf{R}_y$ is the correlation matrix

$$\mathbf{R}_y = E\left\{(\mathbf{y}_\ell - \hat{\mathbf{o}}_\ell)(\mathbf{y}_\ell - \hat{\mathbf{o}}_\ell)'\right\} \quad (9)$$

- The solution to the optimization (8) is the eigenvector of $\mathbf{R}_y$ corresponding to its maximum eigenvalue.

# Non-sparse small outliers to sparse large outliers

- We can see that

$$
\begin{aligned}
&E[\|\hat{\mathbf{h}}'(\mathbf{y}_\ell - \hat{\mathbf{o}}_\ell)\|^2] \\
&\quad = \hat{\mathbf{h}}' E\left[(\mathbf{X}_\ell\boldsymbol{\theta} + \boldsymbol{\epsilon}_\ell + \mathbf{h}v_\ell)'(\mathbf{X}_\ell\boldsymbol{\theta} + \boldsymbol{\epsilon}_\ell + \mathbf{h}v_\ell)\right]\hat{\mathbf{h}} \\
&\quad = \hat{\mathbf{h}}'\left(E[\boldsymbol{\theta}'\mathbf{X}_\ell'\mathbf{X}_\ell\boldsymbol{\theta}]\right)\hat{\mathbf{h}} + \sigma_\epsilon^2\hat{\mathbf{h}}'\hat{\mathbf{h}} + \hat{\mathbf{h}}'\mathbf{h}'v_\ell^2\mathbf{h}\hat{\mathbf{h}} \\
&\quad = \hat{\mathbf{h}}'\left(E[\boldsymbol{\theta}'\mathbf{X}_\ell'\mathbf{X}_\ell\boldsymbol{\theta}]\right)\hat{\mathbf{h}} + \sigma_\epsilon^2 + v_\ell^2. \qquad (10)
\end{aligned}
$$

where the noise power $\sigma_\epsilon^2$ stays unchanged, while the outlier power is enhanced from $|h_{(\ell-1)T_L+k}v_\ell|^2$ to $|v_\ell|^2$.

# Non-sparse small outliers to sparse large outliers

- For Model 1, with the estimated weighting vector $\hat{\mathbf{h}}$, we can construct $L$ new labeled training data with up to $L$ large outliers

$$z_\ell = \hat{\mathbf{h}}'(\mathbf{y}_\ell - \hat{\mathbf{o}}_\ell) = \hat{\mathbf{h}}'\mathbf{X}_\ell\boldsymbol{\theta} + \hat{\mathbf{h}}'\boldsymbol{\epsilon}_\ell + \hat{\mathbf{h}}'\mathbf{h}v_\ell. \tag{11}$$

- For model 2, since the weighting vectors $\mathbf{h}_\ell$ are assumed known, the new labeled data is calculated directly as

$$z_\ell = \mathbf{h}_\ell'(\mathbf{y}_\ell - \hat{\mathbf{o}}_\ell), \quad \ell = 1, \cdots, L. \tag{12}$$

# Non-sparse small outliers to sparse large outliers

- With (11) and (12), construct $L$ new labeled training data
- Append these $L$ new training data $(\mathbf{h}'_\ell \mathbf{X}_\ell, z_\ell)$ to the original training data set ($T + L$ in total), with up to $L$ new large outliers contained in the data $z_\ell$ with the magnitude of $v_\ell$, which guarantees the sparsity of the large outliers.

# Small outliers mitigation

- Define the new $T + L$ training data set as $(\tilde{\mathbf{X}}, \tilde{\mathbf{y}})$, where $\tilde{\mathbf{X}} = \left[\mathbf{X}'_{tr}, \mathbf{h}'_1\mathbf{X}_1, \cdots, \mathbf{h}'_1\mathbf{X}_L\right]'$, $\tilde{\mathbf{y}} = [\mathbf{y}'_{tr}, z_1, \cdots, z_L]'$. We have

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\theta + \tilde{\epsilon} + \tilde{\mathbf{v}}, \tag{13}$$

where

$$\begin{aligned} \tilde{\epsilon} &= [\epsilon', \mathbf{h}'_1\epsilon_1, \cdots, \mathbf{h}'_L\epsilon_L]', \\ \tilde{\mathbf{v}} &= [(\mathbf{Hv})', v_1, \cdots, v_L]'. \end{aligned} \tag{14}$$

- The new outliers in (13) are sparse and large enough in magnitude.

# Small outliers mitigation

- Therefore, we can use the compressive sensing method again to estimate $\boldsymbol{\theta}$ and $\tilde{\mathbf{v}}$ jointly as

$$\{\hat{\boldsymbol{\theta}}, \hat{\mathbf{v}}\} = \arg \min_{\{\boldsymbol{\theta}, \tilde{\mathbf{v}}\}} \|\tilde{\mathbf{y}} - \tilde{\mathbf{v}} - \tilde{\mathbf{X}}\boldsymbol{\theta}\| + \lambda_1 \|\tilde{\mathbf{v}}\|_1. \tag{15}$$

where,

$$\hat{\boldsymbol{\theta}} = \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1} \tilde{\mathbf{X}}'(\tilde{\mathbf{y}} - \tilde{\mathbf{v}}), \tag{16}$$

- The solution to the joint optimization of (15) is

$$\hat{\mathbf{v}} = \arg \min_{\tilde{\mathbf{v}}} \left\| \left( \mathbf{I} - \tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}' \right) (\tilde{\mathbf{y}} - \tilde{\mathbf{v}}) \right\| + \lambda_1 \|\tilde{\mathbf{v}}\|_1. \tag{17}$$

# The pseudo code for the proposed scheme

| **New Robust Regression Algorithm** |
|---|
| i) Input: Data pool $\{\mathbf{x}_i, y_i, i = 1, 2, \cdots, I\}$, $\lambda_1$, $T$, $T_L$ |
| ii) Pool-based active learning: Select $T$ training data out of the data pool; |
| iii) Large outlier mitigation: Estimate and remove $\hat{\mathbf{o}}$ with (3) and (4); |
| iv) Small outlier mitigation: |
| 1) Construct new training data with (11) and (12), and form the $T + L$ new training data $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{y}}$;<br>2) Estimate $\hat{\mathbf{v}}$ and $\hat{\theta}$ with (17) and (16); |
| v) Output: $\hat{\theta}$ for test data prediction. |

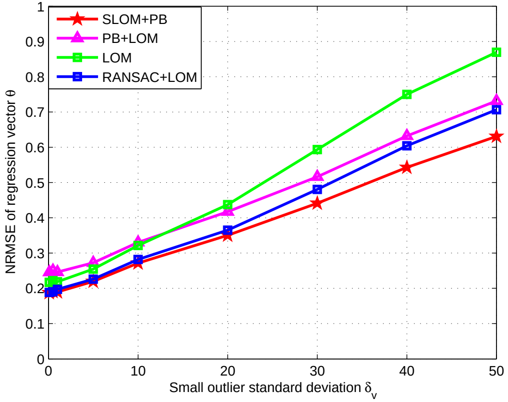


Figure: 1 Regressor estimation performance with the small outlier model 1.

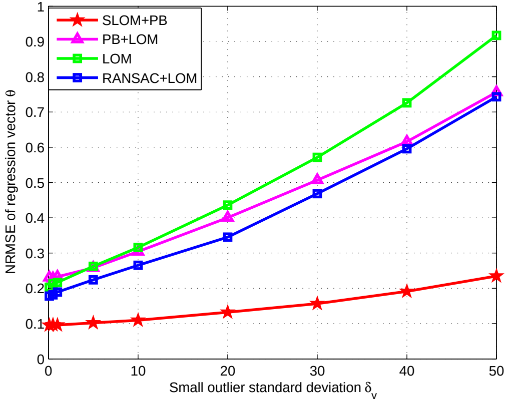


Figure: 2 Regressor estimation performance with the small outlier model 2.

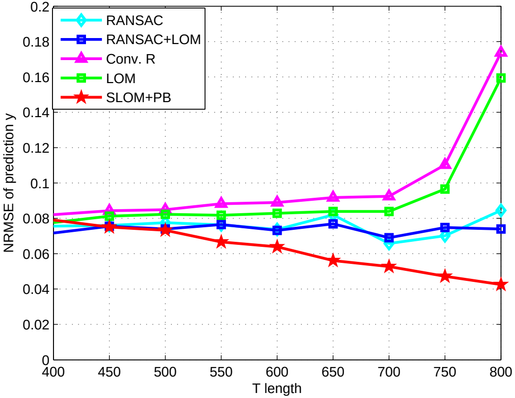


Figure: 3 Prediction performance with the small outlier model 1 in the Air Quality data set.

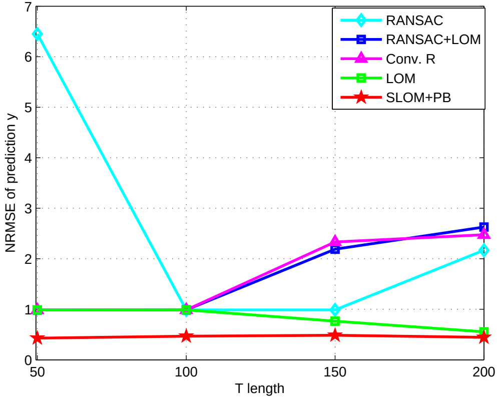


Figure: 4 Prediction performance with the small outlier model 1 in the survey data.

# Conclusions:

- Developed a new robust regression scheme by integrating active learning with compressive sensing to make the data labeling in linear regression problems more robust to both sparse large outliers and non-sparse small outliers;
- Proposed two small outlier models for converting non-sparse small outliers to sparse large outliers;
- Verified the robustness of the new algorithm by extensive simulations with artificial data, UCI benchmark data, as well as real survey data.

*Thank You*