A Unified Robust Regression Model for Lasso-like Algorithms

Wenzhuo Yang a0096049@nus.edu.sg
Department of Mechanical Engineering, National University of Singapore, Singapore 117576

Huan Xu mpexuh@nus.edu.sg
Department of Mechanical Engineering, National University of Singapore, Singapore 117576

Abstract

We develop a unified robust linear regression model and show that it is equivalent to a general regularization framework to encourage sparse-like structure that contains group Lasso and fused Lasso as specific examples. This provides a robustness interpretation of these widely applied Lasso-like algorithms, and allows us to construct novel generalizations of Lasso-like algorithms by considering different uncertainty sets. Using this robustness interpretation, we present new sparsity results, and establish the statistical consistency of the proposed regularized linear regression. This work extends a classical result from Xu et al. (2010) that relates standard Lasso with robust linear regression to learning problems with more general sparse-like structures, and provides new robustness-based tools to to understand learning problems with sparse-like structures.

1. Introduction

In this paper we establish a unified relationship between robustness and regularization schemes for various sparse-like structures, in the context of linear regression. Linear regression aims to find a vector $\boldsymbol{\beta}$ such that $\mathbf{y} \approx \mathbf{X}\boldsymbol{\beta}$, for a given matrix $\mathbf{X} \in \mathcal{R}^{n\times m}$ and vector $\mathbf{y} \in \mathcal{R}^n$. From a learning perspective, each row of $\mathbf{X}$ represents a training sample, and the corresponding element of $\mathbf{y}$ is the target value or response of this observed sample. Each column of $\mathbf{X}$ corresponds to a feature, and the objective of linear regression is to obtain a set of weights so that the weighted sum of the feature values approximates the target value.

Regularized linear regression framework – where one finds the solution that minimizes a weighted combination of the residual norm and a certain regularization term (e.g., Tikhonov & Arsenin, 1977; Tibshirani, 1996) – is now a standard practice in machine learning and statistics for linear regression. Among different regularization schemes, the ℓ_1 regularized linear regression, also termed *Lasso* (Tibshirani, 1996; Chen et al., 1998; Efron et al., 2004), is increasingly popular due to its tendency to select sparse solutions. Indeed, Lasso has been extremely successful in the high-dimensional regime, as it allows recovering the true solution β^* where the samples are significantly outnumbered by the dimensionality by exploiting sparse structure of β^*. Extensive effort has been made to explain the success of Lasso (e.g., Tropp, 2006; Donoho, 2006; Wainwright, 2009; Bickel et al., 2009; Zhang, 2009, and many others), among which, one interesting result from Xu et al. (2010) showed that the success of Lasso is due to its robustness. In particular, they showed that Lasso is equivalent to a robust linear regression formulation, and such robustness interpretation implies the sparsity and the consistency of Lasso.

Inspired by the success of Lasso, numerous regularization schemes were proposed to select solutions with more general sparse-like structures. For example, domain knowledge may indicate that the solution is group sparse, i.e., features can be grouped, and the features belonging to one group is likely to be either all non-active (corresponding to the regressor having zero coefficients), or all active. One example of group sparsity appears is measuring gene expression, where experiments show that selecting a few genes that belong to the same functional groups can lead to increased interpretability of the predictive signature (Rapaport et al., 2007). A prominent algorithm proposed to enforce this sparse-like structure is the group Lasso formulation (Yuan & Lin, 2006), where the regularization term is the sum of the ℓ_2-norms of the different groups of features, also called the ℓ_1/ℓ_2-norm. This

Proceedings of the 30^th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. Copyright 2013 by the author(s).

formulation leads to a sparse selection of the *groups* of features. Other examples of Lasso-like algorithms include the fused Lasso (Tibshirani et al., 2005) that encourages sparsity of the coefficients and also sparsity of their differences, the sparse group Lasso (Friedman et al., 2010) that encourages solutions that are sparse at both the group and individual feature levels, and many others.

This paper attempts to explain the success of those Lasso-like algorithms in a unified way. Our approach is largely inspired by Xu et al. (2010) – we analyze these algorithms based on their robustness properties. In specific, our first result states that a wide range of regularized linear regression problems including the aforementioned ones, all have equivalent robust regression reformulations. This provides a robustness reinterpretation of a class of regularized linear regression formulations for sparse-like structured solutions, and generalizes similar results of standard Lasso showed in Xu et al. (2010). Moreover, our robustness interpretation leads to new formulation and new analysis. We derive new regularization variants of Lasso-like algorithms by considering different uncertainty sets of the robust linear regression formulation. We then present new sparsity results for the group Lasso, as well as proofs of consistency of Lasso-like algorithms, all based on the robustness interpretation. Since robustness is a geometric concept, our approach gives new analysis and new geometric intuition compared to previous methods.

Notations. We use lower-case boldface letters to denote column vectors and upper-case boldface letters to denote matrices. The operator vectorizing a matrix by stacking its columns is denoted by $\mathrm{vec}(\cdot)$. For simplicity, we use $\|\mathbf{X}\|_p$ to denote the ℓ_p-norm of $\mathrm{vec}(\mathbf{X})$, e.g. $\|\mathbf{X}\|_2$ is the Frobenius norm $\|\mathbf{X}\|_F$, and $\|\mathbf{X}\|_p^*$ to denote its dual norm. We denote the set $\{1, \cdots, m\}$ as $[m]$ and call a subset g of $[m]$ a *group*. The identity matrix is denoted by $\mathbf{I}$, the ith element of vector $\mathbf{x}$ is denoted by x_i, and the ith column of matrix $\mathbf{\Delta}$ is denoted by $\mathbf{\Delta}_i$. For vector $\mathbf{x}$ and group g, we denote $\mathbf{x}_g$ as the vector whose ith element is x_i if $i \in g$ or 0 otherwise. Similarly, for matrix $\mathbf{\Delta}$ and group g, we denote $\mathbf{\Delta}_g$ as the matrix whose ith column is $\mathbf{\Delta}_i$ if $i \in g$ or $\mathbf{0}$ otherwise.

2. Unified Robust Framework

This section presents the main result of this paper – there exists a strong relationship between robust linear regression and several widely applied variants of Lasso.

2.1. Preliminary

We start by briefly review the result from Xu et al. (2010) that connects standard Lasso with robust regression. Robust linear regression considers the case that the observed data is corrupted by some (potentially malicious) disturbance. To protect against such disturbance, the following min-max formulation is typically solved:

$$\min_{\boldsymbol{\beta}\in\mathcal{R}^m} \{\max_{\mathbf{\Delta}\in U} \|\mathbf{y} - (\mathbf{X}+\mathbf{\Delta})\boldsymbol{\beta}\|_p\}, \tag{1}$$

where U is the uncertainty set, or the set of admissible disturbances of the observed matrix $\mathbf{X}$. Xu et al. (2010) showed that the robust optimization above is equivalent to the ℓ_1-norm regularized linear regression (standard Lasso) when the uncertainty set is defined by *feature wise* norm constraints:

Theorem 1 (Xu et al. (2010))**.** *The robust regression problem (1) with the uncertainty set*

$$U = \{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_m) | \|\boldsymbol{\delta}_i\|_2 \leq c_i,\ i = 1, \cdots, m\},$$

for given $c_i \geq 0$, is equivalent to the following ℓ_1-norm regularized regression problem:

$$\min_{\boldsymbol{\beta}\in\mathcal{R}^m} \{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + \sum_{i=1}^{m} c_i |\boldsymbol{\beta}_i|\}.$$

It turns out Theorem 1 not only provides a new insight of Lasso from a robustness perspective, but is also a powerful tool to analyze the sparsity and consistency of Lasso, see Xu et al. (2010) for details.

2.2. Main Results

Given the success of the robust interpretation of Lasso, it is natural to ask whether different Lasso-like formulations such as the group Lasso or the fused Lasso can also be reformulated as robust linear regression problems by selecting appropriate uncertainty sets. We provide in this section an affirmative answer. To illustrate our general result, we first consider the overlapping group Lasso proposed in Yuan & Lin (2006). The following theorem shows that it is equivalent to a robust linear regression problem:

Theorem 2. *Let the uncertainty set be*

$$\begin{aligned} U = \{\mathbf{\Delta}^{(1)} + \cdots + \mathbf{\Delta}^{(t)} | \|\mathbf{\Delta}^{(i)}_{g_i}\|_2 \leq c_{g_i} \text{ and} \\ \|\mathbf{\Delta}^{(i)}_{g_i^c}\|_2 = 0,\ \forall i \in [t]\}, \end{aligned} \tag{2}$$

where matrix $\mathbf{\Delta}^{(i)} \in \mathcal{R}^{n\times m}$, $\bigcup_{i=1}^t g_i = [m]$ and $g_i^c = [m] \setminus g_i$, then the robust regression (1) with U is equivalent to

$$\min_{\boldsymbol{\beta}\in\mathcal{R}^m} \{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + \sum_{i=1}^{t} c_{g_i} \|\boldsymbol{\beta}_{g_i}\|_2\}. \tag{3}$$

Proof. See Appendix. Notice that here and after, the regression formulations we consider slightly differ from the more widely used ones, as we minimize the norm of the error, rather than the squared norm. It is known that these two coincide up to a change of the regularization coefficient since the empirical error terms and the regularization terms we discuss are all convex. □

Note that the groups defined in Theorem 2 are allowed to overlap. Theorem 2 shows that the group Lasso formulation is equivalent to the robust linear regression where the admissible disturbance is given by the norm constraints on each group g_i, as opposed to constraints on each feature in Theorem 1. Observe that by taking each feature as one group, Theorem 2 immediately implies Theorem 1.

We now present our main result that connects variants of Lasso-like algorithms with the robust linear regression framework. Consider the following uncertainty set:

$$\begin{aligned} U = \{&\boldsymbol{\Delta}^{(1)}\mathbf{W}_1 + \cdots + \boldsymbol{\Delta}^{(t)}\mathbf{W}_t | \\ &\forall i \in [t], \forall g \in G_i, \|\boldsymbol{\Delta}_g^{(i)}\|_p \le c_g\}, \end{aligned} \tag{4}$$

where matrix $\mathbf{W}_i \in \mathcal{R}^{m\times m}$ is fixed, G_i is the set of the groups, and c_g provides the norm bound of group g of the disturbance. Notice that G_i may contain more than one groups, and two different groups $g_1, g_2 \in G_i$ are allowed to overlap, i.e., $g_1 \cap g_2 \neq \emptyset$. It is easy to see that such set contains the uncertainty set considered in Theorem 2 as a special case, i.e. $G_i = \{g_i, g_i^c\}$ for $i \in [t]$. The next theorem shows that such uncertainty set provides a unified framework that "encodes" the ridge regression and many variants of Lasso-like algorithms.

Theorem 3. *The robust regression problem (1) with the uncertainty set (4) is equivalent to the convex regularized linear regression problem:*

$$\min_{\boldsymbol{\beta}\in\mathcal{R}^m} \{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \sum_{i=1}^t \max_{\forall g\in G_i, \|\boldsymbol{\alpha}_g^{(i)}\|_p \le c_g} {\boldsymbol{\alpha}^{(i)}}^\top \mathbf{W}_i\boldsymbol{\beta}\}. \tag{5}$$

Proof. For any fixed $\boldsymbol{\beta}$, we have

$$\begin{aligned} &\max_{\boldsymbol{\Delta}\in U} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_p \\ =&\max_{\boldsymbol{\Delta}\in U} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \sum_{i=1}^t \boldsymbol{\Delta}^{(i)}\mathbf{W}_i\boldsymbol{\beta}\|_p \\ \le&\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \max_{\boldsymbol{\Delta}\in U} \sum_{i=1}^t \|\sum_{j=1}^m (\mathbf{W}_i\boldsymbol{\beta})_j \boldsymbol{\Delta}_j^{(i)}\|_p \\ \le&\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \max_{\boldsymbol{\Delta}\in U} \sum_{i=1}^t \sum_{j=1}^m |(\mathbf{W}_i\boldsymbol{\beta})_j| \|\boldsymbol{\Delta}_j^{(i)}\|_p. \end{aligned}$$

For clarity, denote

$$\begin{aligned} &\boldsymbol{\alpha}^{(i)} \equiv \\ &[\mathrm{sign}((\mathbf{W}_i\boldsymbol{\beta})_1) \cdot \|\boldsymbol{\Delta}_1^{(i)}\|_p, \cdots, \mathrm{sign}((\mathbf{W}_i\boldsymbol{\beta})_m) \cdot \|\boldsymbol{\Delta}_m^{(i)}\|_p]^\top. \end{aligned}$$

From the definition of the uncertainty set U, we know that $\|\boldsymbol{\Delta}_g^{(i)}\|_p \le c_g$ for any $i \in [t]$ and $g \in G_i$. Thus, $\|\boldsymbol{\alpha}_g^{(i)}\|_p = \|\boldsymbol{\Delta}_g^{(i)}\|_p \le c_g$, and we have

$$\begin{aligned} &\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \max_{\boldsymbol{\Delta}\in U} \sum_{i=1}^t \sum_{j=1}^m |(\mathbf{W}_i\boldsymbol{\beta})_j| \|\boldsymbol{\Delta}_j^{(i)}\|_p \\ =&\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \max_{\boldsymbol{\Delta}\in U} \sum_{i=1}^t {\boldsymbol{\alpha}^{(i)}}^\top \mathbf{W}_i\boldsymbol{\beta} \\ \le&\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \sum_{i=1}^t \max_{\forall g\in G_i, \|\boldsymbol{\alpha}_g^{(i)}\|_p \le c_g} {\boldsymbol{\alpha}^{(i)}}^\top \mathbf{W}_i\boldsymbol{\beta}. \end{aligned}$$

On the other hand, let

$$\boldsymbol{\alpha}_0^{(i)} = \arg\max_{\forall g\in G_i, \|\boldsymbol{\alpha}_g^{(i)}\|_p \le c_g} {\boldsymbol{\alpha}^{(i)}}^\top \mathbf{W}_i\boldsymbol{\beta}$$

and

$$\mathbf{u} = \begin{cases} \frac{\mathbf{y} - \mathbf{X}\boldsymbol{\beta}}{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p} & \text{if } \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p \neq 0 \\ \text{any vector with unit } \ell_p \text{ norm} & \text{otherwise} \end{cases}$$

and then let

$$\boldsymbol{\Delta}^{(i)} = -\mathbf{u} \cdot {\boldsymbol{\alpha}_0^{(i)}}^\top$$

From the definition above, we know that $\|\boldsymbol{\Delta}_g^{(i)}\|_p = \|\boldsymbol{\alpha}_g^{(i)}\|_p \le c_g$. Thus, we have

$$\begin{aligned} &\max_{\boldsymbol{\Delta}\in U} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_p \\ \ge&\|\mathbf{y} - (\mathbf{X} + \sum_{i=1}^t \boldsymbol{\Delta}^{(i)}\mathbf{W}_i)\boldsymbol{\beta}\|_p \\ =&\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} + \mathbf{u}\sum_{i=1}^t {\boldsymbol{\alpha}_0^{(i)}}^\top \mathbf{W}_i\boldsymbol{\beta}\|_p \\ =&\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \sum_{i=1}^t \max_{\forall g\in G_i, \|\boldsymbol{\alpha}_g^{(i)}\|_p \le c_g} {\boldsymbol{\alpha}^{(i)}}^\top \mathbf{W}_i\boldsymbol{\beta}, \end{aligned}$$

which establishes the theorem. □

Indeed, the regularized linear regression (5) is a generalization for Lasso. By setting t, G_i, $\mathbf{W}_i$ and c_g to appropriate values, (5) can be reduced as *standard Lasso*, *group Lasso*, *fused Lasso*, *trend filtering*, among others.

Corollary 1 (Ridge Regression)**.** *Suppose that* $t = 1$, $p = 2$, $\mathbf{W}_1 = \mathbf{I}$, $G_1 = \{[m]\}$ *and* $c_g = c$, *then the robust regression problem (1) is equivalent to*

$$\min_{\boldsymbol{\beta}\in\mathcal{R}^m} \{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + c\|\boldsymbol{\beta}\|_2\}. \tag{6}$$

Ridge regression has been well studied. It shrinks the regression coefficients $\beta_1, \cdots, \beta_m$ by penalizing their sizes (in terms of ℓ_2-norm) to control the complexity of the regression model.

Corollary 2 (Standard Lasso)**.** *Suppose that $t = 1$, $\mathbf{W}_1 = \mathbf{I}$, $G_1 = \{\{1\}, \cdots, \{m\}\}$ and $c_i = c_{\{i\}}$, then the robust regression problem (1) is equivalent to*

$$\min_{\boldsymbol{\beta} \in \mathcal{R}^m} \{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \sum_{i=1}^{m} c_i |\boldsymbol{\beta}_i|\}. \tag{7}$$

The main difference between the ridge regression and the standard Lasso is that the Lasso penalizes the ℓ_1-norm of the coefficients. The Lasso's ability to recover sparse solutions has been extensively explored, and has found wide applications in statistics, signal processing, computer vision, bioinformatics, to name a few.

Corollary 3 (Non-overlapping Group Lasso)**.** *Suppose that $t = 1$, $\mathbf{W}_1 = \mathbf{I}$, $G_1 = \{g_1, \cdots, g_k\}$ and $g_i \cap g_j = \emptyset$ for any $i \neq j$, then the robust regression problem (1) is equivalent to*

$$\min_{\boldsymbol{\beta} \in \mathcal{R}^m} \{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \sum_{i=1}^{k} c_{g_i} \|\boldsymbol{\beta}_{g_i}\|_p^*\}. \tag{8}$$

The non-overlapping group Lasso is an extension of the standard Lasso, where non-overlapping group structure of features is known as the prior information. In particular, features are partitioned into known groups, and one seeks solutions that select few non-zero *groups*. Different from Lasso, group Lasso does not encourage sparsity inside each group.

Corollary 4 (Overlapping Group Lasso (Jacob et al., 2009))**.** *Suppose that $t = 1$, $\mathbf{W}_1 = \mathbf{I}$, $G_1 = \{g_1, \cdots, g_k\}$, and $\bigcup_{i=1}^k g_i = [m]$, then the robust regression problem (1) is equivalent to*

$$\min_{\boldsymbol{\beta} \in \mathcal{R}^m} \{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \min_{\sum \mathbf{v}_{g_i} = \boldsymbol{\beta},\ supp(\mathbf{v}_{g_i}) \subseteq g_i} \sum_{i=1}^{k} c_{g_i} \|\mathbf{v}_{g_i}\|_p^*\}. \tag{9}$$

Different from the overlapping group Lasso formulation (3) proposed in Yuan & Lin (2006) that encourages solutions whose supports are in the *complement of a union of groups* (i.e, many groups are all zero), Formulation (9) tends to select solutions whose support is contained in a union of potentially overlapping groups. This is motivated by applications in bioinformatics, e.g., predicting the class of a tumor from gene expression measurements with microarrays, and simultaneously select a few genes to establish a predictive signature. Figure 1 illustrates the difference between two group Lasso formulations.

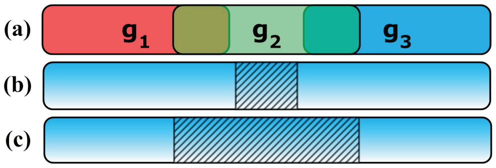

Figure 1. Preferred solutions of the two group Lassos. Hatched regions indicates non-zero coefficients and unhatched regions indicates zero coefficients. (a) Predefined groups of the coefficient $\boldsymbol{\beta}$; (b) One solution that Yuan & Lin (2006) tends to select; (c) One solution that Jacob et al. (2009) tends to select.

Corollary 5 (Fused Lasso (Tibshirani et al., 2005))**.** *Suppose that $t = 2$, $G_1 = G_2 = \{\{1\}, \cdots, \{m\}\}$, and*

$$\mathbf{W}_1 = \mathbf{I},\ \mathbf{W}_2 = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & -1 \\ 0 & \cdots & 0 & 0 & 0 \end{bmatrix},$$

then the robust regression problem (1) is equivalent to

$$\min_{\boldsymbol{\beta} \in \mathcal{R}^m} \{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \sum_{i=1}^{m} c_i |\boldsymbol{\beta}_i| + \sum_{i=1}^{m-1} c_i' |\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i+1}|\}, \tag{10}$$

where c_i and c_i' are the "$c_{\{i\}}$"s corresponding to the uncertainty sets of G_1 and G_2, respectively.

The fused Lasso is motivated by protein mass spectroscopy and gene expression profiling. After estimating an order of data and putting correlated data near one another, solving it not only encourages sparsity in the coefficients $\beta_1, \cdots, \beta_m$ but also encourages sparsity in their differences, which implies that it tends to select a sparse solution in which nearby coefficients are similar to each other.

Corollary 6 (Sparse Group Lasso (Friedman et al., 2010))**.** *Suppose that $t = k + 1$, $\bigcup_{i=1}^k g_i = [m]$ and $g_i^c = [m] \setminus g_i$. Let $\mathbf{W}_i = \mathbf{I}$, $G_i = \{g_i, g_i^c\}$, $c_{g_i^c} = 0$ for $i \in [k]$, and let $\mathbf{W}_{k+1} = \mathbf{I}$, $G_{k+1} = \{\{1\}, \cdots, \{m\}\}$, then the robust regression problem (1) is equivalent to*

$$\min_{\boldsymbol{\beta} \in \mathcal{R}^m} \{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \sum_{i=1}^{k} c_{g_i} \|\boldsymbol{\beta}_{g_i}\|_p^* + \sum_{i=1}^{m} c_i |\boldsymbol{\beta}_i|\} \tag{11}$$

where c_i is equal to $c_{\{i\}}$.

The sparse group Lasso blends the standard Lasso with the group Lasso, and encourages solutions that are sparse at both the group and the individual feature levels. Notice that Equation (11) is equivalent to the elastic net (Zou & Hastie, 2005) when $k = 1$ and $p = 2$.

Corollary 7 (Generalized Lasso (Tibshirani & Taylor, 2011))**.** *Suppose that* $t = 1$, $\mathbf{W}_1 = \mathbf{D}$, $G_1 = \{\{1\}, \cdots, \{m\}\}$, *and* $c_{\{i\}} = \lambda$, *then the robust regression problem (1) is equivalent to*

$$\min_{\boldsymbol{\beta}\in\mathcal{R}^m} \{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \lambda\|\mathbf{D}\boldsymbol{\beta}\|_1\}. \tag{12}$$

By making various choices of $\mathbf{D}$, the generalized Lasso can be reformulated as well-known problems in the literature: *trend filtering* (Kim et al., 2009), etc.

Remark. While the inner maximization of the robust linear regression problem (1) over the uncertainty set (4) is non-convex, Theorem 3 shows that it can be solved efficiently as it is equivalent to a convex optimization problem (5). In particular, by strong duality, the optimization problem (5) is equivalent to

$$\begin{aligned}
\min_{\beta, \mathbf{v}_g^{(i)}} \quad & \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \sum_{i=1}^{t}\sum_{g\in G_i} c_g \|\mathbf{v}_g^{(i)}\|_p^* \\
\text{s.t.} \quad & \sum_{g\in G_i} \|\mathbf{v}_g^{(i)}\|_p^* = \mathbf{W}_i\boldsymbol{\beta},\ \forall i \in [t] \\
& \mathbf{A}_g^{(i)}\mathbf{v}_g^{(i)} = 0,\ \forall i \in [t], \forall g \in G_i,
\end{aligned}$$

where $\mathbf{v}_g^{(i)} \in \mathcal{R}^m$ is a decision variable and $\mathbf{A}_g^{(i)} \in \mathcal{R}^{(m-|g|)\times m}$ is a constant matrix defined as $\mathbf{A}_g^{(i)} = (\mathbf{e}_{i_1}, \cdots, \mathbf{e}_{i_k})^\top$ where $k = m - |g|$, $\{i_1, \cdots, i_k\} = g^c$, and $\mathbf{e}_i$ is the i^{th} unit base vector. This is a linear constrained convex optimization problem which can be solved efficiently using off-the-shelf methods. In addition, for special case such as the non-overlapping group Lasso, more scalable codes are available (e.g., Meier et al., 2008; Roth & Fischer, 2008).

3. General Uncertainty Sets

As discussed above, we assume that the disturbance of each *group* is bounded individually, then the robust linear regression (1) can be reformulated as the regularized linear regression (5) which is a generalized formulation for Lasso-like algorithms. In this section, we provide a more generalized formulation of the uncertainty set.

Consider the following uncertainty set $\hat{U}$:

$$\begin{aligned}
\hat{U} = \{&\boldsymbol{\Delta}^{(1)}\mathbf{W}_1 + \cdots + \boldsymbol{\Delta}^{(t)}\mathbf{W}_t \mid \mathbf{c} \in Z; \\
&\forall i \in [t], \forall g \in G_i, \|\boldsymbol{\Delta}_g^{(i)}\|_p \le c_g\},
\end{aligned} \tag{13}$$

where G_i is the set of groups of disturbance $\boldsymbol{\Delta}^{(i)}$, $\mathbf{c}$ is the vector whose elements are the norm bounds c_g of all the groups contained in $G_1, \cdots, G_t$, e.g. $\mathbf{c} = (c_{g_1}, \cdots, c_{g_n})$, and Z is the feasible set of $\mathbf{c}$. If Z has only one element, then $\hat{U}$ is equivalent to the uncertainty set U which is defined as (4) where c_g is fixed. Hence, the set $\hat{U}$ is a very general formulation, and provides us with significant flexibility in designing uncertainty sets and equivalently new regression algorithms. In particular, we consider Z given by a set of convex constraints, i.e.,

$$Z = \{\mathbf{z} \in \mathcal{R}^k | f_i(\mathbf{z}) \le 0, \forall i \in [q];\ \mathbf{z} \ge 0\}, \tag{14}$$

where each $f_i(\mathbf{z})$ is a convex function and $k = \sum_{i=1}^t |G_i|$ ($|G_i|$ is the cardinality of G_i), and Z has non-empty relative interior.

Under these assumptions, we have the following theorem showing that the robust regression problem (1) with uncertainty set $\hat{U}$ can be converted to a tractable convex optimization problem.

Theorem 4. *The robust regression problem with the uncertainty set (13)*

$$\min_{\boldsymbol{\beta}\in\mathcal{R}^m} \{\max_{\boldsymbol{\Delta}\in\hat{U}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_p\}$$

is equivalent to

$$\min_{\boldsymbol{\lambda}\in\mathcal{R}_+^q, \boldsymbol{\kappa}\in\mathcal{R}_+^k, \boldsymbol{\beta}\in\mathcal{R}^m} \{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \upsilon(\boldsymbol{\lambda}, \boldsymbol{\kappa}, \boldsymbol{\beta})\} \tag{15}$$

where

$$\begin{aligned}
\upsilon(\boldsymbol{\lambda}, \boldsymbol{\kappa}, \boldsymbol{\beta}) = \max_{\mathbf{c}\in\mathcal{R}^k} \{&\sum_{i=1}^{t} \max_{\forall g\in G_i, \|\boldsymbol{\alpha}_g^{(i)}\|_p \le c_g} {\boldsymbol{\alpha}^{(i)}}^\top \mathbf{W}_i\boldsymbol{\beta} + \\
&\boldsymbol{\kappa}^\top\mathbf{c} - \sum_{i=1}^{q} \lambda_i f_i(\mathbf{c})\}.
\end{aligned}$$

Furthermore, the equivalent optimization problem (15) is convex and tractable.

Proof. We prove this theorem by using Theorem 3 and the duality. See Appendix for more details. □

One interesting implication of Theorem 4 is that by choosing "proper" uncertainty sets, we can simplify (15) and obtain new regularized linear regression formulations. We provide some examples to illustrate this in the rest of this section. The notations used follow those in Theorem 3.

Corollary 8. *Suppose that the uncertainty set* $\hat{U} = \{\boldsymbol{\Delta} | \exists\ \mathbf{c} \in \mathcal{R}^m$ *such that* $\mathbf{c} \ge 0$ *and* $\|\mathbf{c}_{g_i}\|_q^* \le s_i, \forall i \in [k]; \|\boldsymbol{\Delta}_j\| \le c_j, \forall j \in [m]\}$, *then the equivalent linear regularized regression problem is*

$$\min_{\boldsymbol{\beta}\in\mathcal{R}^m} \{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \sum_{i=1}^{k} s_i \|\boldsymbol{\beta}_{g_i}\|_q\},$$

where $\|\cdot\|_q^*$ is the dual norm of $\|\cdot\|_q$, $\bigcup_{i=1}^k g_i = [m]$, and $g_i \cap g_j = \emptyset$ for $i \neq j$.

Proof. From Theorem 3 and Theorem 4, we have

$$\begin{aligned}&\min_{\lambda \in \mathcal{R}_+, \boldsymbol{\kappa} \in \mathcal{R}_+^m} \upsilon(\lambda, \boldsymbol{\kappa}, \boldsymbol{\beta}) \\ = &\min_{\lambda \in \mathcal{R}_+, \boldsymbol{\kappa} \in \mathcal{R}_+^m} \max_{\mathbf{c} \in \mathcal{R}^m} \{\sum_{i=1}^m (\kappa_i + |\beta_i|)c_i - \\ &\sum_{i=1}^k \lambda_i (\|\mathbf{c}_{g_i}\|_q^* + s_i)\}.\end{aligned}$$

Define $\mathbf{r}_{g_i}$ as the vector whose j^{th} elements is $\kappa_j + |\beta_j|$ for all $j \in g_i$, then the equation above is equivalent to

$$\min_{\lambda \in \mathcal{R}_+, \boldsymbol{\kappa} \in \mathcal{R}_+^m | \|\mathbf{r}_{g_i}\|_q \leq \lambda_i, \forall i \in [k]} \boldsymbol{\lambda}^\top \mathbf{s} = \sum_{i=1}^k s_i \|\boldsymbol{\beta}_{g_i}\|_q,$$

which establishes the corollary. □

This corollary interprets arbitrary norm-based regularizers for the non-overlapping group Lasso from a robust regression perspective. By choosing different norms that bound $\mathbf{c}_{g_i}$ for $i \in [k]$, different regularization terms are obtained, which implies that the effect of the regularization term of Lasso is selecting a proper uncertainty set of the observed matrix.

Remark. For the overlapping group Lasso (Yuan & Lin, 2006), the same result holds by adding more disturbances to the overlapping columns of the observed matrix. See Appendix for more details.

We now consider a polytope uncertainty set in which there exists an additional constraint bounding the total disturbance besides the norm bound for disturbance on each group.

Corollary 9. *Suppose that $\hat{U} = \{\sum_{i=1}^t \boldsymbol{\Delta}^{(i)} \mid \exists\, 0 \leq \mathbf{c} \leq \mathbf{s} : \sum_{i=1}^t c_i/s_i \leq \theta;\ \|\boldsymbol{\Delta}_{g_i}^{(i)}\|_p \leq c_i, \|\boldsymbol{\Delta}_{g_i^c}^{(i)}\|_p = 0,\ \forall i \in [t]\}$, then the equivalent linear regularized regression problem is as follows*

$$\begin{aligned}\min_{\boldsymbol{\beta}, \lambda} \quad & \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \sum_{i=1}^t [s_i \|\boldsymbol{\beta}_{g_i}\|_p^* - \lambda]_+ + \lambda\theta \\ s.t. \quad & \lambda \geq 0\end{aligned} \tag{16}$$

where $[x]_+ = \max\{x, 0\}$.

Proof. From Theorem 2 and Theorem 4, we have

$$\begin{aligned}\upsilon(\boldsymbol{\lambda}, \boldsymbol{\kappa}, \boldsymbol{\beta}) = \max_{\mathbf{c}} \sum_{i=1}^t c_i (\|\boldsymbol{\beta}_{g_i}\|_p^* - \frac{\lambda}{s_i}) + \\ (\boldsymbol{\kappa} - \bar{\boldsymbol{\lambda}})^\top \mathbf{c} + \bar{\boldsymbol{\lambda}}^\top \mathbf{s} + \lambda\theta.\end{aligned}$$

Thus, $\upsilon(\boldsymbol{\lambda}, \boldsymbol{\kappa}, \boldsymbol{\beta}) = \bar{\boldsymbol{\lambda}}^\top \mathbf{s} + \lambda\theta$ and $\boldsymbol{\kappa}_i = \bar{\boldsymbol{\lambda}}_i + \lambda/s_i - \|\boldsymbol{\beta}_{g_i}\|_p^*$, $\forall i \in [t]$, which implies that the robust regression is equivalent to

$$\begin{aligned}\min_{\boldsymbol{\beta}, \lambda} \quad & \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \bar{\boldsymbol{\lambda}}^\top \mathbf{s} + \lambda\theta \\ s.t. \quad & \|\boldsymbol{\beta}_{g_i}\|_p^* - \lambda/s_i \leq \bar{\boldsymbol{\lambda}}_i,\ \forall i \in [t], \\ & \lambda \geq 0, \bar{\boldsymbol{\lambda}} \geq 0,\end{aligned}$$

which is also equivalent to (16). □

Notice that when $\theta = t$, the above formulation reduces to the overlapping group Lasso. On the other hand, when $\theta = 0$, it is equivalent to the linear least square problem. Hence, this formulation allows us to control the desired group sparsity level using only one parameter θ .

4. Sparsity

The standard Lasso's ability to recover spare solutions has been extensively studied (Chen et al., 1998; Feuer & Nemirovski, 2003; Candes et al., 2006; Tropp, 2004; 2006), and the sparsity properties of the group Lasso have also been explored (Huang et al., 2009a;b; Percival, 2011). These results typically take one of two approaches – treating the problem from either a statistical or optimization perspective. In this section, we investigate the sparsity properties of the robust regression and equivalently non-overlapping/overlapping group Lasso from a robust optimization perspective, and provides a geometric interpretation for sparsity. We consider first the overlapping group Lasso.

Theorem 5. *For the overlapping group Lasso*

$$\min_{\boldsymbol{\beta} \in \mathcal{R}^m} \{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + c \sum_{i=1}^t \|\boldsymbol{\beta}_{g_i}\|_2\}$$

where $\bigcup_{i=1}^t g_i = [m]$, if there exists $I \subset [t]$ such that for an orthonormal base $\mathbf{V}$ of span$(\{\mathbf{X}_j,\ j \in [m] \setminus \bigcup_{i \in I} g_i\} \cup \{\mathbf{y}\})$, we have $\|\mathbf{V}\mathbf{V}^\top \mathbf{X}_{g_i}\|_2 \leq c$ for $i \in I$, then any optimal solution $\boldsymbol{\beta}^*$ satisfies that $\boldsymbol{\beta}_{g_i}^* = 0$ for $i \in I$.

Proof. From Theorem 2, we know that the overlapping group Lasso is equivalent to

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in U} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2, \tag{17}$$

where the uncertainty set U is as follows

$$U = \{\sum_{i=1}^t \boldsymbol{\Delta}^{(i)} \mid \forall i, \|\boldsymbol{\Delta}_{g_i}^{(i)}\|_2 \leq c \text{ and } \|\boldsymbol{\Delta}_{g_i^c}^{(i)}\|_2 = 0\}.$$

Recall that it is allowed that $g_i \cap g_j \neq \emptyset$ for $i \neq j$. We define group $\hat{g}_i$ as

$$\hat{g}_i = \begin{cases} g_i & i \in I; \\ g_i - \bigcup_{j \in I} g_j & i \notin I, \end{cases}$$

and consider the following uncertainty set

$$\hat{U} = \{\sum_{i=1}^{t} \boldsymbol{\Delta}^{(i)} \mid \forall i, \|\boldsymbol{\Delta}^{(i)}_{\hat{g}_i}\|_2 \leq c \text{ and } \|\boldsymbol{\Delta}^{(i)}_{\hat{g}_i^c}\|_2 = 0\},$$

then we have $\hat{U} \subseteq U$ since $\hat{g}_i \subseteq g_i, \forall i \in [t]$. Thus,

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \hat{U}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2 \leq \min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in U} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2. \tag{18}$$

Let $\bar{\mathbf{X}}$ be a matrix whose ith column is

$$\bar{\mathbf{X}}_i = \begin{cases} \mathbf{X}_i & i \notin \bigcup_{j \in I} \hat{g}_j \\ \mathbf{X}_i - \mathbf{V}\mathbf{V}^\top \mathbf{X}_i & i \in \bigcup_{j \in I} \hat{g}_j, \end{cases} \tag{19}$$

then from the condition $\|\mathbf{V}\mathbf{V}^\top \mathbf{X}_{\hat{g}_i}\|_2 \leq c$ for $i \in I$, we have $\|(\mathbf{X} - \bar{\mathbf{X}})_{\hat{g}_i}\|_2 \leq c$ for $i \in I$. Now let

$$\bar{U} = \{\boldsymbol{\Delta}^{(1)} + \cdots + \boldsymbol{\Delta}^{(t)} \mid \|\boldsymbol{\Delta}^{(i)}_{\hat{g}_i}\|_2 \leq c \text{ and } \|\boldsymbol{\Delta}^{(i)}_{\hat{g}_i^c}\|_2 = 0 \text{ for } i \notin I; \|\boldsymbol{\Delta}^{(i)}\|_2 = 0 \text{ for } i \in I\},$$

and consider the following robust regression problem

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \bar{U}} \|\mathbf{y} - (\bar{\mathbf{X}} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2,$$

which is equivalent to

$$\min_{\boldsymbol{\beta}} \{\|\mathbf{y} - \bar{\mathbf{X}}\boldsymbol{\beta}\|_2 + c \sum_{i \notin I} \|\boldsymbol{\beta}_{\hat{g}_i}\|_2\}. \tag{20}$$

We denote the optimal solution of (20) as $\bar{\boldsymbol{\beta}}^*$. From the definition of $\bar{\mathbf{X}}$, we know that each column of $\bar{\mathbf{X}}_{\hat{g}_i}$ for $i \in I$ is orthogonal to the span of $\{\mathbf{X}_{\hat{g}_i},\ i \notin I\} \cup \{\mathbf{y}\}$. Hence by changing $\bar{\boldsymbol{\beta}}^*_{\hat{g}_i}$ to 0 for all $i \in I$, the minimizing objective does not increase. This implies that the optimal solution $\bar{\boldsymbol{\beta}}^*$ satisfies that $\bar{\boldsymbol{\beta}}^*_{g_i} = 0$ for $i \in I$.

We now prove that $\bar{\boldsymbol{\beta}}^*$ is also the optimal solution of the overlapping group Lasso. We first show that

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \bar{U}} \|\mathbf{y} - (\bar{\mathbf{X}} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2 \\ \leq &\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \hat{U}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2 \\ \leq &\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in U} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2. \end{aligned} \tag{21}$$

For any $\mathbf{X}$, $\bar{\mathbf{X}}$ (defined by (19)) and $\bar{\boldsymbol{\Delta}} \in \bar{U}$ such that $\|(\mathbf{X} - \bar{\mathbf{X}})_{\hat{g}_i}\|_p \leq c$ for $i \in I$, there exists $\boldsymbol{\Delta} \in \hat{U}$ such that $\bar{\mathbf{X}} + \bar{\boldsymbol{\Delta}} = \mathbf{X} + \boldsymbol{\Delta}$, which implies $\{\bar{\mathbf{X}} + \bar{\boldsymbol{\Delta}} | \bar{\boldsymbol{\Delta}} \in \bar{U}\} \subseteq \{\mathbf{X} + \boldsymbol{\Delta} | \boldsymbol{\Delta} \in \hat{U}\}$. Thus, Inequality (21) holds. On the other hand, since $\bar{\boldsymbol{\beta}}^*_{g_i} = 0$ for $i \in I$, we have

$$\begin{aligned} &\max_{\boldsymbol{\Delta} \in \bar{U}} \|\mathbf{y} - (\bar{\mathbf{X}} + \boldsymbol{\Delta})\bar{\boldsymbol{\beta}}^*\|_2 \\ = &\max_{\boldsymbol{\Delta} \in \hat{U}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\bar{\boldsymbol{\beta}}^*\|_2 \\ = &\max_{\boldsymbol{\Delta} \in U} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\bar{\boldsymbol{\beta}}^*\|_2, \end{aligned} \tag{22}$$

then for an arbitrary $\boldsymbol{\beta}$, the following inequality holds

$$\max_{\boldsymbol{\Delta} \in U} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\bar{\boldsymbol{\beta}}^*\|_2 \leq \max_{\boldsymbol{\Delta} \in U} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2,$$

which implies that $\bar{\boldsymbol{\beta}}^*$ is the optimal solution of the overlapping group Lasso. Hence we establish the theorem. □

Theorem 5 gives a geometric interpretation of the sparsity properties of the overlapping group Lasso based on its robustness. Indeed, it shows that a set of *groups of features* all receive zero weight if there exists an admissible perturbation of each group which makes their features orthogonal to the other ones. As a special case, if the groups are non-overlapping (i.e., $g_i \cap g_j = \emptyset$ for $i \neq j$), we have the following theorem that shows the sparsity properties of the non-overlapping group Lasso.

Corollary 10. *If there exists $I \subset [t]$ such that for an orthonormal base $\mathbf{V}$ of span($\{\mathbf{X}_{g_j},\ j \notin I\} \cup \{\mathbf{y}\}$), we have $\|\mathbf{V}\mathbf{V}^\top \mathbf{X}_{g_i}\|_2 \leq c$ for $i \in I$, then any optimal solution $\boldsymbol{\beta}^*$ of the non-overlapping group Lasso (8) satisfies that $\boldsymbol{\beta}^*_{g_i} = 0$ for $i \in I$.*

5. Consistency

In this section, we investigate statistical property of the regularized linear regression formulation (5), and show that it is asymptotically consistent by using the robust properties derived from its equivalence with the robust linear regression (1). The proofs of our results largely follow the same framework proposed in Xu et al. (2010). The main idea of the proofs is as follows: We show that the robust optimization formulation (1) can be seen to be the maximum expected error with respect to a class of probability measures. This class includes a kernel density estimator, and using this, we can prove that the regularized linear regression is consistent. However, because the uncertainty set we consider is more complicated than the one investigated in Xu et al. (2010) (which corresponds to the standard Lasso), the construction of the class of probability measures is more involved.

Using the same notation, we define $\bar{G}_i = \{g \in G_i | c_g \neq 0\}$ and assume that $\bigcup_{i=1}^{t} \bar{G}_i = [m]$, i.e., each feature is contained in at least one group to ensure that all features are regularized. We restrict our discussion to the case that $\mathbf{W}_i = \mathbf{I}$ for $i \in [t]$ and c_g for each group g equals either $\sqrt{n}c_n$ (n is the number of the samples) or 0, and establish the statistical consistency of the regularized linear regression (5) from a distributional robustness argument. Let P be a probability measure with bounded support that generates i.i.d samples $(b_i, \mathbf{r}_i^\top)$, and has a density $f(\cdot)$. Denote the set of the first n samples by S_n and define

$$\boldsymbol{\beta}(c_n, S_n) = \arg\min_{\boldsymbol{\beta}}\{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(b_i - \mathbf{r}_i^\top\boldsymbol{\beta})^2}+ \sum_{i=1}^{t} \max_{\forall g\in\bar{G}_i, \|\boldsymbol{\alpha}_g^{(i)}\|_2\le c_n} {\boldsymbol{\alpha}^{(i)}}^\top\boldsymbol{\beta}\},$$

$$\boldsymbol{\beta}(P) = \arg\min_{\boldsymbol{\beta}}\{\sqrt{\int_{b,\mathbf{r}}(b_i - \mathbf{r}_i^\top\boldsymbol{\beta})^2 dP(b,\mathbf{r})}\}.$$

Thus, $\boldsymbol{\beta}(c_n, S_n)$ is the solution to the regularized linear regression (5) with the tradeoff parameter set to $\sqrt{n}c_n$, and $\boldsymbol{\beta}(P)$ is the "true" optimal solution. We have the following consistency results.

Theorem 6. *Let $\{c_n\}$ be such that $c_n \downarrow 0$ and $\lim_{n\to\infty} n(c_n)^{m+1} = \infty$. Suppose there exists a constant H such that $\|\boldsymbol{\beta}(c_n, S_n)\|_2 \le H$ almost surely. Then*

$$\lim_{n\to\infty}\sqrt{\int_{b,\mathbf{r}}(b_i - \mathbf{r}_i^\top\boldsymbol{\beta}(c_n,S_n))^2 dP(b,\mathbf{r})} = \sqrt{\int_{b,\mathbf{r}}(b_i - \mathbf{r}_i^\top\boldsymbol{\beta}(P))^2 dP(b,\mathbf{r})},$$

almost surely.

Proof. We sketch the proof. We first show that the equivalent robust regression (1) over the training data is equal to the worst-case expected generalization error among a set of distributions. Then, we show that such set of distributions includes a kernel density estimator for the true (unknown) distribution of the samples. Finally, using the fact that the kernel density estimator converges to the true density function almost surely when $c_n \downarrow 0$ and $\lim_{n\to\infty} n(c_n)^{m+1} = \infty$, we can prove the consistency. See Appendix for more details. □

Remark. In the first step of the above proof, the set of distributions is the union of classes of distributions corresponding to disturbance in hyper-rectangle Borel sets $Z_1, \cdots, Z_n$ centered at $(b_i, \mathbf{r}_i^\top)$ with lengths depending on c_n and the constraints on the uncertainty set $\boldsymbol{\Delta}$. Since in Xu et al. (2010), only the constraint that the norm of each column of $\boldsymbol{\Delta}$ is bounded is considered, such Borel sets can be easily constructed for the standard Lasso. In contrast, in this paper, we consider the case where $\boldsymbol{\Delta} = \sum_{i=1}^{t}\boldsymbol{\Delta}^{(i)}$ and the constraints are imposed on feature groups $\boldsymbol{\Delta}_g^{(i)}$ for $g \in G_i$. Since two groups g_i and g_j may have overlapping elements, this case is much more general than Xu et al. (2010) and the construction of the Borel sets is more difficult. Yet, we can still show that such Borel sets can be constructed, and the kernel density estimator is included in the set of distributions formed by the constructed Borel sets. See Appendix for more details.

Indeed, the assumption that $\|\boldsymbol{\beta}(c_n, S_n)\|_2 \le H$ in Theorem 6 can be removed, and the consistency result still holds. See the Appendix for the theorem and the proof. Notice that Theorem 6 implies that *standard Lasso*, *group Lasso* and *sparse group Lasso* are all asymptotically consistent. Follow the same road map but with more involved analysis, one can show that that *fused Lasso* is also asymptotically consistent.

6. Conclusions

In this paper, we investigated a unified approach to explain the success of algorithms that encourage various sparse-like structures based on the concept of *robustness*. In particular, we considered robust linear regression where the perturbations are constrained with respect to *each group of features*, and show that this formulation is equivalent to a regularized linear regression framework that contains several widely used Lasso-like algorithms such as fused Lasso. This hence provides a robustness based interpretation of such algorithms. Moreover, we established sparsity property and statistical consistency of group Lasso from this robustness perspective. The main thrust of this work is to extend a classical result that relates standard Lasso with robust linear regression (Xu et al., 2010) to learning problems with more general sparse-like structures. Achieving this makes it possible to understand these problems by analyzing the respective uncertainty sets, and will eventually enable us to design new algorithms to specific learning tasks that has superior performance than existing approaches.

Acknowledgments

We thank the anonymous reviewers for the useful feedback that helped improve the paper. This work is partially supported by the Ministry of Education of Singapore through AcRF Tier Two grant R-265-000-443-112 and NUS startup grant R-265-000-384-133.

References

Bickel, P., Ritov, Y., and Tsybakov, A. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37:1705–1732, 2009.

Candes, E., Romberg, J., and Tao, T. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489C–509, 2006.

Chen, S., Donoho, D., and Saunders, M. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33C–61, 1998.

Donoho, D. L. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. *Annals of Statistics*, 32: 407–499, 2004.

Feuer, A. and Nemirovski, A. On sparse representation in pairs of bases. *IEEE Transactions on Information Theory*, 49(6):1579C–1581, 2003.

Friedman, J., Hastie, T., and Tibshirani, R. A note on the group Lasso and a sparse group Lasso. Technical report, Jan 2010.

Huang, J., Huang, X., and Metaxas, D. N. Learning with dynamic group sparsity. In *ICCV*, pp. 64–71, 2009a.

Huang, J., Zhang, T., and Metaxas, D. Learning with structured sparsity. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 417–424, 2009b.

Jacob, L., Obozinski, G., and Vert, J. Group Lasso with overlap and graph Lasso. In *ICML*, 2009.

Kim, S., Koh, K., Boyd, S., and Gorinevsky, D. ℓ_1 trend filtering. *SIAM Review*, 51(2):339–360, 2009.

Meier, L., Geer, S. Van De, and Bhlmann, P. The group Lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

Percival, D. Theoretical properties of the overlapping groups Lasso. *Electronic Journal of Statistics*, 2011.

Rapaport, F., Zinovyev, A., Dutreix, M., Barillot, E., and Vert, J. Classification of microarray data using gene networks. *BMC Bioinformatics*, 8, 2007.

Roth, V. and Fischer, B. The group Lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In *Proceedings of the 25th international conference on Machine learning*, pp. 848–855, 2008.

Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society Series B*, pp. 91–108, 2005.

Tibshirani, R. J. and Taylor, J. The solution path of the generalized Lasso. *The Annals of Statistics*, 39 (3), 2011.

Tikhonov, A. N. and Arsenin, V. Y. *Solutions of Ill-Posed Problems.* V. H. Winston & Sons, Washington, D.C.: John Wiley & Sons, New York,, 1977.

Tropp, J. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231C–2242, 2004.

Tropp, J. A. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 51(3):1030–1051, 2006.

Wainwright, M. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55:2183–2202, 2009.

Xu, H., Caramanis, C., and Mannor, S. Robust regression and Lasso. *IEEE Transactions on Information Theory*, 56(7):3561–3574, 2010.

Yuan, M. and Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006.

Zhang, T. Some sharp performance bounds for least squares regression with l1 regularization. *Annals of Statistics*, 37:2109–2144, 2009.

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.