Journal of Machine Learning Research 18 (2017) 1-24 Submitted 10/14; Revised 2/17; Published 7/17

Two New Approaches to Compressed Sensing Exhibiting Both Robust Sparse Recovery and the Grouping Effect

Mehmet Eren Ahsen MEHMET.ERENAHSEN@MSSM.EDU
Icahn School of Medicine at Mount Sinai
1468 Madison Avenue
New York, NY 10029, USA

Niharika Challapalli NIHARIKA15C@GMAIL.COM
Department of Electrical Engineering
University of Texas at Dallas
Richardson, TX 75080, USA

Mathukumalli Vidyasagar M.VIDYASAGAR@UTDALLAS.EDU, M.VIDYASAGAR@IITH.AC.IN
Department of Systems Engineering
University of Texas at Dallas
Richardson, TX 75080, USA, and
Department of Electrical Engineering
Indian Institute of Technology Hyderabad
Kandi, Telangana, India 502285

Editor: Saharon Rosset

Abstract

In this paper we introduce a new optimization formulation for sparse regression and compressed sensing, called CLOT (Combined L-One and Two), wherein the regularizer is a convex combination of the ℓ_1- and ℓ_2-norms. This formulation differs from the Elastic Net (EN) formulation, in which the regularizer is a convex combination of the ℓ_1- and ℓ_2-norm *squared.* It is shown that, in the context of compressed sensing, the EN formulation *does not achieve* robust recovery of sparse vectors, whereas the new CLOT formulation achieves robust recovery. Also, like EN but unlike LASSO, the CLOT formulation achieves the grouping effect, wherein coefficients of highly correlated columns of the measurement (or design) matrix are assigned roughly comparable values. It is already known LASSO does not have the grouping effect. Therefore the CLOT formulation combines the best features of both LASSO (robust sparse recovery) and EN (grouping effect).

The CLOT formulation is a special case of another one called SGL (Sparse Group LASSO) which was introduced into the literature previously, but without any analysis of either the grouping effect or robust sparse recovery. It is shown here that SGL achieves robust sparse recovery, and also achieves a version of the grouping effect in that coefficients of highly correlated columns belonging to the same group of the measurement (or design) matrix are assigned roughly comparable values.

Keywords: Sparse regression, compressed sensing, LASSO, Sparse Group LASSO, Elastic Net

1. Introduction

The LASSO and the Elastic Net (EN) formulations are among the most popular approaches for sparse regression and compressed sensing. In this section, we briefly review these two problems and their current status, so as to provide the background for the remainder of the paper.

1.1 Sparse Regression

In sparse regression, one is given a measurement matrix (also called a design matrix in statistics) $A \in \mathbb{R}^{m \times n}$ where $m \ll n$, together with a measurement or measured vector $y \in \mathbb{R}^m$. The objective is to choose a vector $x \in \mathbb{R}^n$ such that x is rather sparse, and Ax is either exactly or approximately equal to y. The problem of finding the most sparse x that satisfies $Ax = y$ is known to be NP-hard (Natarajan, 1995); therefore it is necessary to find alternate approaches.

For the sparse regression problem, the general approach is to determine the estimate $\hat{x}$ by solving the minimization problem

$$\hat{x} = \operatorname*{argmin}_z \|y - Az\|_2^2 \text{ s.t. } \mathcal{R}(z) \leq \gamma, \tag{1}$$

or in Lagrangian form,

$$\hat{x} = \operatorname*{argmin}_z \|y - Az\|_2^2 + \lambda \mathcal{R}(z), \tag{2}$$

where $\mathcal{R} : \mathbb{R}^n \rightarrow \mathbb{R}$ is known as a "regularizer," and γ, λ are adjustable parameters. Different choices of the regularizer lead to different approaches. With the choice $\mathcal{R}_{\text{ridge}}(z) = \|z\|_2^2$, the approach is known as ridge regression (Hoerl and Kennard, 1970), which builds on earlier work (Tikhonov, 1943). The LASSO approach (Tibshirani, 1996) results from choosing $\mathcal{R}_{\text{LASSO}}(z) = \|z\|_1$, while the Elastic Net (EN) approach (Zou and Hastie, 2005) results from choosing

$$\mathcal{R}_{\text{EN}}(z) = \alpha_1 \|z\|_1 + \alpha_2 \|z\|_2^2, \tag{3}$$

where α_1, α_2 are adjustable parameters. For later use, we redefine the EN regularizer as

$$\mathcal{R}_{\text{EN}}(z) = (1 - \mu)\|z\|_1 + \mu \|z\|_2^2, \tag{4}$$

where

$$\mu = \frac{\alpha_2}{\alpha_1 + \alpha_2} \in [0, 1]$$

is an adjustable parameter, and the constant $\alpha_1 + \alpha_2$ can be absorbed into the Lagrange multiplier λ in (2). Note that the EN regularizer function interpolates ridge regression and LASSO, in the sense that EN reduces to LASSO if $\mu = 0$ and to ridge regression if $\mu = 1$. A very general approach to regression using a convex regularizer is given in (Negabhan et al., 2012).

The LASSO approach can be shown to return a solution $\hat{x}$ with no more than m nonzero components, under mild regularity conditions; see (Osborne et al., 2000). There is no such bound on the number of components of $\hat{x}$ when EN is used. However, when the columns of the matrix A are highly correlated, then LASSO chooses just one of these columns and ignores the rest. Measurement matrices with highly correlated columns occur in many practical situations, for example, in microarray measurements of messenger RNA, otherwise known as gene expression data. The EN approach was proposed at least in part to overcome this undesirable behavior of the LASSO formulation. It is shown in (Zou and Hastie, 2005, Theorem 1) that if two columns (say i and j) of the matrix A are highly correlated, then the corresponding components $\hat{x}_i$ and $\hat{x}_j$ of the EN solution are nearly equal. This is known as the "grouping effect", and the point is that EN demonstrates the grouping effect whereas LASSO does not.

1.2 Compressed Sensing

In compressed sensing, the objective is to *choose* the measurement matrix A (which is part of the data in sparse regression), such that whenever the vector x is nearly sparse, it is possible to nearly recover x from noise-corrupted measurements of the form $y = Ax + \eta$. Let us make the problem formulation precise. For this purpose we begin by introducing some notation.

Throughout, the symbol $[n]$ denotes the index set $\{1, \ldots, n\}$. The **support** of a vector $x \in \mathbb{R}^n$ is denoted by $\mathrm{supp}(x)$ and is defined as

$$\mathrm{supp}(x) := \{i \in [n] : x_i \neq 0\}.$$

A vector $x \in \mathbb{R}^n$ is said to be ***k*-sparse** if $|\mathrm{supp}(x)| \leq k$. The set of all k-sparse vectors is denoted by Σ_k. The ***k*-sparsity index** of a vector x with respect to a given norm $\| \cdot \|$ is defined as

$$\sigma_k(x, \| \cdot \|) := \min_{z \in \Sigma_k} \|x - z\|. \tag{5}$$

It is obvious that $x \in \Sigma_K$ if and only if $\sigma_k(x, \| \cdot \|) = 0$ for every norm.

The general formulation of the compressed sensing problem given below is essentially taken from (Cohen et al., 2009). Suppose that $A \in \mathbb{R}^{m \times n}$ is the "measurement matrix", and $\Delta : \mathbb{R}^m \to \mathbb{R}^n$ is the "decoder map", where $m \ll n$. Suppose $x \in \mathbb{R}^n$ is an unknown vector that is to be recovered. The input to the decoder consists of $y = Ax + \eta$ where η denotes the measurement noise, and a prior upper bound in the form $\|\eta\|_2 \leq \epsilon$ is available; in other words, ϵ is a known number. In this set-up, the vector $\hat{x} = \Delta(y)$ is the approximation to the original vector x. With these conventions, we can now state the following.

Definition 1 *Suppose $p \in [1, 2]$. The pair (A, Δ) is said to achieve **robust sparse recovery** of order k with respect to $\| \cdot \|_p$ if there exist constants C and D that might depend on A and Δ but not on x or η, such that*

$$\|\hat{x} - x\|_p \leq \frac{1}{k^{1-1/p}} [C \sigma_k(x, \|\cdot\|_1) + D\epsilon]. \tag{6}$$

The restriction that $p \in [1, 2]$ is tied up with the fact that the bound on the noise is for the Euclidean norm $\|\eta\|_2$. The usual choices for p in (6) are $p = 1$ and $p = 2$.

Among the most popular approaches to compressed sensing is ℓ_1-norm minimization, which was popularized in a series of papers, of which we cite only (Candès and Tao, 2005; Candès et al., 2006; Candès, 2008; Donoho, 2006). The survey paper (Davenport et al.,

2012) has an extensive bibliography on the topic, as does the recent book (Foucart and Rauhut, 2013). In this approach, the estimate $\hat{x}$ is defined as

$$\hat{x} := \operatorname*{argmin}_{z} \|z\|_1 \text{ s.t. } \|Az - y\|_2 \leq \epsilon. \tag{7}$$

Note that the above definition does indeed define a decoder map $\Delta : \mathbb{R}^m \to \mathbb{R}^n$. In order for the above pair (A, Δ) to achieve robust sparse recovery, the matrix A is chosen so as to satisfy a condition defined next.

Definition 2 *A matrix $A \in \mathbb{R}^{m\times n}$ is said to satisfy the **Restricted Isometry Property (RIP)** of order k with constant δ_k if*

$$(1 - \delta_k)\|u\|_2^2 \leq \|Au\|_2^2 \leq (1 + \delta_k)\|u\|_2^2, \ \forall u \in \Sigma_k. \tag{8}$$

Starting with (Candès and Tao, 2005), several papers have derived sufficient conditions that the RIP constant of the matrix A must satisfy in order for ℓ_1-norm minimization to achieve robust sparse recovery. Recently, the "best possible" bound has been proved in (Cai and Zhang, 2014). These results are stated here for the convenience of the reader.

Theorem 3 *(See (Cai and Zhang, 2014, Theorem 2.1)) Suppose A satisfies the RIP of order tk for some number $t \geq 4/3$ such that tk is an integer, with $\delta_{tk} < \sqrt{(t-1)/t}$. Then the recovery procedure in* (7) *achieves robust sparse recovery of order k.*

Theorem 4 *(See (Cai and Zhang, 2014, Theorem 2.2)) Let $t \geq 4/3$. For all $\gamma > 0$ and all $k \geq 5/\gamma$, there exists a matrix A satisfying the RIP of order tk with constant $\delta_{tk} \leq \sqrt{(t-1)/t} + \gamma$ such that the recovery procedure in* (7) *fails for some k-sparse vector.*

Observe that the Lagrangian formulation of the LASSO approach is

$$\hat{x} := \operatorname*{argmin}_{z} [\|Az - y\|_2^2 + \lambda\|z\|_1],$$

whereas the Lagrangian formulation of (7) is

$$\hat{x} := \operatorname*{argmin}_{z} [\|z\|_1 + \beta\|Az - y\|_2],$$

which is essentially the same as the Lagrangian formulation of

$$\hat{x} = \operatorname*{argmin}_{z} \|Az - y\|_2 \text{ s.t. } \|z\|_1 \leq \gamma.$$

This last formulation of sparse regression is known as "square-root LASSO" (Belloni et al., 2014). Therefore the community refers to the approach to compressed sensing given in (7) as the LASSO, though this may not be strictly accurate.

1.3 Compressed Sensing with Group Sparsity

Over the years some variants of LASSO have been proposed for compressed sensing, such as the Group LASSO (GL) (Yuan and Lin, 2006) and the Sparse Group LASSO (SGL) (Simon et al., 2013). In the GL formulation, the index set $\{1, \ldots, n\}$ is partitioned into g disjoint sets $G_1, \ldots, G_g$, and the associated norm is defined as

$$\|z\|_{\mathrm{GL}} := \sum_{i=1}^{g} \|z_{G_i}\|_2, \tag{9}$$

where z_{G_i} denotes the projection of the vector z onto the components in G_i. The notation is intended to remind us that the norm depends on the specific partitioning $\mathcal{G}$. Some authors divide the term $\|z_{G_i}\|_2$ by $|G_i|$, but we do not do that. A further refinement of GL is the sparse group LASSO (SGL), in which the group structure is as before, but the norm is now defined as

$$\|z\|_{\mathrm{SGL},\mu} := \sum_{i=1}^{g} (1-\mu)\|z_{G_i}\|_1 + \mu\|z_{G_i}\|_2, \tag{10}$$

where as before $\mu \in [0, 1]$. If $x \in \mathbb{R}^n$ is an unknown vector, then recovery of x is attempted via

$$\hat{x} = \operatorname*{argmin}_{z} \|z\|_{\mathrm{GL}} \text{ s.t. } \|Az - y\|_2 \leq \epsilon \tag{11}$$

in Group LASSO, and via

$$\hat{x} = \operatorname*{argmin}_{z} \|z\|_{\mathrm{SGL},\mu} \text{ s.t. } \|Az - y\|_2 \leq \epsilon \tag{12}$$

in Sparse Group LASSO.

The main idea behind GL is that one is less concerned about the number of nonzero components of $\hat{x}$, and more concerned about the number of distinct groups containing these nonzero components. Therefore GL attempts to choose an estimate $\hat{x}$ that has nonzero entries in as few distinct sets as possible. In principle, SGL tries to choose an estimate $\hat{x}$ that not only has nonzero components within as few groups as possible, but within those groups, has as few nonzero components as possible. Note that if $\mu = 0$, then SGL reduces to LASSO (because of the summability of the ℓ_1-norm), whereas if $\mu = 1$, then SGL reduces to GL. Note too that if $g = n$ and every set G_i is a singleton $\{i\}$, then GL reduces to LASSO.

1.4 Motivation and Contributions of the Paper

Now we come to the motivation and contributions of the present paper. The LASSO formulation is well-suited for compressed sensing (see Theorem 3), but not so well-suited for sparse regression, because it lacks the grouping effect. The EN formulation is well-suited for sparse regression as it exhibits the grouping effect, but it is not known whether it can achieve compressed sensing.

The first result presented in the paper is that if the EN regularizer of (4) is used instead of the ℓ_1-norm in (7), then the resulting approach *does not achieve* robust sparse recovery unless $m \geq n/4$, that is, the number of measurements grows linearly with respect to the size of the vector. This would not be considered "compressed" sensing. This led us to formulate another regularizer, namely

$$\|z\|_{C,\mu} = (1-\mu)\|z\|_1 + \mu\|z\|_2. \tag{13}$$

Note that, while the EN regularizer in (4) is a convex function, it is not a norm. In contrast, $\| \cdot \|_{C,\mu}$ is not just convex but is also a norm. Also, the EN regularizer in its original form in (3) is intended to have *two* adjustable parameters. Our intent is that, in compressed sensing applications, the constant μ in (13) is a *fixed constant*, and not intended to be varied. Therefore, if the ℓ_1-norm in (7) is replaced by $\| \cdot \|_{C,\mu}$, then there is only one adjustable parameter, namely the Lagrange multiplier associated with the constraint. The same remark applies also to GL and SGL, that is, (11) and (12) respectively. We refer to $\| \cdot \|_{C,\mu}$ as the CLOT norm, with CLOT standing for Combined L-One and Two. It is shown that the CLOT norm combines the best features of both LASSO and EN, in that

- When the CLOT norm is used as the regularizer in sparse regression, the resulting solution exhibits the grouping effect.
- When the ℓ_1-norm is replaced by the CLOT norm in (7), the resulting solution achieves robust sparse recovery if the matrix A satisfies the RIP.
- Moreover, if μ in CLOT is set to zero so that CLOT becomes LASSO, the bound on the RIP constant reduces to the "best possible" bound in Theorem 3.

Clearly the CLOT norm is a special case of the SGL norm with the entire index set $[n]$ being taken as a single group (though the adjective "sparse" is no longer appropriate). This led us to explore whether the SGL norm achieves either grouping effect or robust sparse recovery. We are able to show that SGL does indeed achieve both.

Now we place these contributions in perspective. There is empirical evidence to support the belief that both the GL and the SGL formulations work well for compressed sensing. However, until the publication of a companion paper by a subset of the present authors (Ahsen and Vidyasagar, 2016), there were no proofs that either of these formulations achieved robust sparse recovery. In (Ahsen and Vidyasagar, 2016), it is shown that both the GL and SGL formulations achieve robust sparse recovery *provided the group sizes are sufficiently small.* This restriction on group sizes is removed in the present paper. Moreover, so far as the authors are aware, until now there are no results on the grouping effect for either of these formulations. In the present paper, it is shown that if two columns of the measurement matrix A *that belong to the same group* are highly correlated, then the corresponding components of the estimate $\hat{x}$ have nearly equal values. However, if two columns that belong to different groups are highly correlated, then their coefficients need not be nearly equal. From the standpoint of applications, this is a highly desirable property. To illustrate, suppose the groups represent biological pathways. Then one would wish to assign roughly similar weights to genes in the same pathway, but not necessarily to those in disjoint pathways.

Thus the contributions of the present paper are:

- To show that the EN *does not achieve* robust sparse recovery.
- To show that both the CLOT and SGL formulations achieve both robust sparse recovery as well as the grouping effect.

- To derive a condition under which CLOT achieves robust sparse recovery, which reduces to the "best possible" condition in Theorem 3 when μ is set to zero, so that CLOT becomes LASSO.

Taken together, these results might indicate that CLOT and SGL are attractive alternatives to the LASSO and EN formulations.

2. Main Theoretical Results

This section contains the main contributions of the paper. We begin by showing in Section 2.1 that the solution paths of EN and CLOT are identical if both λ_1 and λ_2 are treated as adjustable parameters. Therefore further research would be needed to establish whether CLOT offers any advantages over EN in numerical performance in sparse regression.

Then we present several theoretical advantages of CLOT over both EN and LASSO. First it is shown in Section 2.2 that the EN approach does not achieve robust sparse recovery, and is therefore not suitable for compressed sensing applications. Next, it is shown in Section 2.3 that the SGL formulation assigns nearly equal weights to highly correlated features *within the same group*, though not necessarily to highly correlated features from different groups. It follows as a corollary that CLOT assigns nearly equal weights to highly correlated features. Then it is shown in Section 2.4 that the SGL formulation achieves robust sparse recovery. The contents of a companion paper by a subset of the present authors Ahsen and Vidyasagar (2016) establish that SGL achieves robust sparse recovery of order k *provided that each group size is smaller than* k. There is no such restriction here. It follows as a corollary that CLOT also achieves robust sparse recovery.

2.1 Relationship Between Solution Paths of EN and CLOT

In this subsection, it is shown that if both μ and λ are tuned via cross-validation in (3), then the solution paths of CLOT are identical to those of EN when both λ_1 and λ_2 are tuned. However, it is shown via an example that if μ is kept fixed and *only* λ is tuned in (3), then CLOT and EN have different solution paths.

Towards this end, we rewrite the CLOT formulation with both μ and λ being tuned in the form

$$\hat{x}_{\rm CLOT} := \operatorname*{argmin}_{z} [\|y - Ax\|_2^2 + \lambda_1 \|z\|_1 + \lambda_2 \|z\|_2]. \tag{14}$$

It is easy to see that the transformation

$$\mu = \frac{\lambda_2}{\lambda_1 + \lambda_2}, \lambda = \lambda_1 + \lambda_2$$

maps (14) into (3). In the other direction, we would define

$$\lambda_1 = (1 - \mu)\lambda, \lambda_2 = \mu\lambda.$$

We are grateful to one of the reviewers for pointing out the result as described in Theorem 5, and providing a proof.

Theorem 5 *Given $y \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m \times n}$, define two vectors:*

$$\hat{x}_{\rm CLOT}(\lambda_1, \tilde{\lambda}_2) = \operatorname*{argmin}_z [\|y - Az\|_2^2 + \lambda_1 \|z\|_1 + \tilde{\lambda}_2 \|z\|_2],$$

$$\hat{x}_{\rm EN}(\lambda_1, \hat{\lambda}_2) = \operatorname*{argmin}_z [\|y - Az\|_2^2 + \lambda_1 \|z\|_1 + \hat{\lambda}_2 \|z\|_2^2].$$

Then for each fixed $\lambda_1 > 0$ and each $\tilde{\lambda}_2 > 0$, there exists a $\hat{\lambda}_2 > 0$ such that

$$\hat{x}_{\rm CLOT}(\lambda_1, \tilde{\lambda}_2) = \hat{x}_{\rm EN}(\lambda_1, \hat{\lambda}_2),$$

and vice versa.

Proof We begin with the following rather obvious observation. Suppose $f(\cdot)$ and $g(\cdot)$ are convex functions, and consider two problems:

$$\text{(P1)} \quad \hat{x}_1(\lambda) = \operatorname*{argmin}_z [f(z) + \lambda g(z)],$$

$$\text{(P2)} \quad \hat{x}_2(c) = \operatorname*{argmin}_z f(z) \text{ s.t. } g(z) \leq c.$$

Then for each λ there exists a c such that $\hat{x}_1(\lambda) = \hat{x}_2(c)$, and vice versa. To establish this, write down the optimality conditions for the two problems, with $\partial f(\cdot)$, $\partial g(\cdot)$ denoting the subgradient sets of $f(\cdot)$, $g(\cdot)$ respectively. Then a necessary and sufficient condition for $\hat{x}_1(\lambda)$ to be the solution of (P1) is:

$$\mathbf{0} \in \partial f(\hat{x}_1(\lambda)) + \lambda \partial g(\hat{x}_1(\lambda)), \tag{15}$$

where $\mathbf{0}$ denotes the zero vector. Similarly, for (P2) the necessary and sufficient conditions are the existence of a constant λ^* such that

$$\mathbf{0} \in \partial f(\hat{x}_1(\lambda^*)) + \lambda \partial g(\hat{x}_1(\lambda^*)), \text{ and } \lambda^*(g(\hat{x}_1(\lambda^*)) - c) = 0. \tag{16}$$

Suppose (15) holds; then (16) holds with $c = g(\hat{x}_1(\lambda))$. Conversely, suppose (16) holds; then (15) holds with $\lambda = \lambda^*$.

Now apply this reasoning with $f(z) = \|y - Az\|_2^2 + \lambda_1 \|z\|_1$, $g_1(z) = \|z\|_2$, $g_2(z) = \|z\|_2^2$. Then each $\hat{x}_{\rm CLOT}(\lambda_1, \tilde{\lambda}_2)$ equals the minimizer of $f(z)$ subject to $\|z\|_2 \leq c$ for some c, while each $\hat{x}_{\rm EN}(\lambda_1, \hat{\lambda}_2)$ equals the minimizer of $f(z)$ subject to $\|z\|_2^2 \leq c'$ for some c'. However, it is obvious that

$$[\|z\|_2 \leq c] \iff [\|z\|_2^2 \leq c^2].$$

Therefore the theorem is proved. $\square$ $\blacksquare$

2.2 Lack of Robust Sparse Recovery of the Elastic Net Formulation

The first result of this section shows that EN formulation does not achieve robust sparse recovery, and therefore is not suitable for compressed sensing applications.

Theorem 6 *Suppose a matrix $A \in \mathbb{R}^{m \times n}$ has the following property: There exist constants C and D such that, whenever $y = Ax + \eta$ for some $x \in R^n$ and $\eta \in \mathbb{R}^m$ with $\|\eta\|_2 \leq \epsilon$, the solution*

$$\hat{x}_{\rm EN} := \operatorname*{argmin}_{z \in \mathbb{R}^n} [(1-\mu)\|z\|_1 + \mu \|z\|_2^2] \text{ } s.t. \text{ } \|y - Az\|_2 \leq \epsilon$$

satisfies

$$\|\hat{x}_{\rm EN} - x\|_2 \leq C \sigma_k(x, \|\cdot\|_1) + D\epsilon. \tag{17}$$

Then

$$m \geq n/4. \tag{18}$$

Proof: Let $\mathcal{N}(A)$ denote the null space of the matrix A, that is, the set of all $h \in \mathbb{R}^n$ such that $Ah = 0$. Let $h \in \mathcal{N}(A)$ be arbitrary, and let $\Lambda \subseteq \{1, \ldots, n\}$ denote the index set of the k largest components of h by magnitude. Therefore

$$\|h_{\Lambda^c}\|_2 = \sigma_k(h, \|\cdot\|_2).$$

Next, (17) implies that, if $\eta = 0$, then $\hat{x}_{\rm EN} = x$ for all $x \in \Sigma_k$. In other words,

$$x = \operatorname*{argmin}_{z \in \mathbb{R}^n} [(1-\mu)\|z\|_1 + \mu \|z\|_2^2] \text{ s.t. } Az = Ax,$$

or equivalently,

$$(1-\mu)\|x\|_1 + \mu\|x\|_2^2 \leq (1-\mu)\|z\|_1 + \mu\|z\|_2^2, \ \forall z \in A^{-1}(\{Ax\}), \ \forall x \in \Sigma_k. \tag{19}$$

Now observe that, because $h \in \mathcal{N}(A)$, we have that

$$Ah_\Lambda = -Ah_{\Lambda^c},$$

and more generally,

$$A(\beta h_\Lambda) = A(-\beta h_{\Lambda^c}), \ \forall \beta > 0.$$

Apply (19) with $x = \beta h_\Lambda \in \Sigma_k$ and $z = -A\beta h_{\Lambda^c} \in A^{-1}(\{Ax\})$. This leads to

$$(1-\mu)\beta\|h_\Lambda\|_1 + \beta^2\mu\|h_\Lambda\|_2^2 \leq (1-\mu)\beta\|h_{\Lambda^c}\|_1 + \beta^2\mu\|h_{\Lambda^c}\|_2^2.$$

Now divide both sides by $\beta^2\mu$, and observe that, for each fixed $\mu > 0$,

$$\frac{1-\mu}{\beta\mu} \to 0 \text{ as } \beta \to \infty.$$

Therefore

$$\|h_\Lambda\|_2^2 \leq \|h_{\Lambda^c}\|_2^2, \text{ or } \|h_\Lambda\|_2 \leq \|h_{\Lambda^c}\|_2, \ \forall h \in \mathcal{N}(A).$$

Next

$$\|h\|_2 \leq \|h_\Lambda\|_2 + \|h_{\Lambda^c}\|_2 \leq 2\|h_{\Lambda^c}\|_2, \ \forall h \in \mathcal{N}(A).$$

Equivalently

$$\|h\|_2 \leq 2\sigma_k(h, \|\cdot\|_2), \ \forall h \in \mathcal{N}(A).$$

This is Equation (5.2) of Cohen-Dahmen-Devore (2009) with $C_0 = 2$. As shown in Theorem 5.1 of that paper, this implies that $m \geq n/C_0^2 = n/4$, which is the desired conclusion. $\square$

Note that the proof of Theorem 6 remains valid even if we were to allow the constant μ to be "tuned", provided that it is bounded away from zero. In other words, the proof does not make use of the fact that μ is a fixed constant. Therefore even in the "naive" version of EN, in which the regularizer is defined as in (3), and both constants α_1 and α_2 are adjusted, robust sparse recovery requires that $m \geq n/4$ provided only that the ratio α_1/α_2 remains bounded away from zero as both parameters are tuned.

2.3 Grouping Property of the SGL and CLOT Formulations

One advantage of the EN over LASSO is that the former assigns roughly equal weights to highly correlated features, as shown in (Zou and Hastie, 2005, Theorem 1) and referred to as the grouping effect. In contrast, if LASSO chooses one feature among a set of highly correlated features, then generically it assigns a zero weight to all the rest. To illustrate, if two columns of A are identical, then in principle LASSO could assign nonzero weights to both columns; however, the slightest perturbation in the data would cause one or the other weight to become zero. The drawback of this is that the finally selected feature set is very sensitive to noise in the measurements. In this section we prove an analog of (Zou and Hastie, 2005, Theorem 1) for SGL formulation. Our result states that if two highly correlated features *within the same group* are chosen by SGL, then they will have roughly similar weights. Since CLOT is a special case of SGL with the entire feature set treated as one group, it follows that CLOT assigns roughly similar weights to highly correlated features in the entire set of features. As a result, the final feature sets obtained using SGL or CLOT are less sensitive to noise in measurements than the ones obtained using LASSO.

Theorem 7 *Let $y \in \mathbb{R}^m, A \in \mathbb{R}^{m \times n}$ be some vector and matrix respectively. Without loss of generality, suppose that y is centered, i.e. $y^t \mathbf{e}_m = 0$, where $\mathbf{e}_m$ denotes a column vector consisting of m ones, and that A is standardized, i.e. $\|a_j\|_2 = 1$ where a_j denotes the j-th column of A. Suppose $\mu > 0$, and let $\mathcal{G}$ denote a partition of $\{1, \ldots, n\}$ into g disjoint subsets. Define*

$$\hat{x} := \operatorname*{argmin}_z \lambda \|y - Az\|_2^2 + \|z\|_{\mathrm{SGL},\mu}, \tag{20}$$

where $\lambda > 0$ is a Lagrange multiplier. Suppose that, for two indices i, j belonging to the same group G_s, we have that $\hat{x}_i \hat{x}_j \neq 0$, where $\hat{x}_i, \hat{x}_j$ denote the components of the vector $\hat{x}$. By changing the sign of one of the columns of A if necessary, it can be assumed that $\hat{x}_i \hat{x}_j > 0$. Define

$$d(i,j) := \frac{|\hat{x}_i - \hat{x}_j|}{2\lambda \|y\|_2}, \rho(i,j) = a_i^t a_j.$$

Then

$$d(i,j) \leq \frac{\sqrt{2(1 - \rho(i,j))} \|\hat{x}^s\|_2}{\mu}, \tag{21}$$

where $\hat{x}^s$ is shorthand for $\hat{x}_{G_s}$.

Proof: Define

$$\begin{aligned} L(z,\mu) &:= \lambda \|y - Az\|_2^2 + \|z\|_{\mathrm{SGL},\mu} \\ &= \lambda \|y - Az\|_2^2 + (1-\mu)\|z\|_1 + \mu \sum_{s=1}^{g} \|z^s\|_2, \end{aligned}$$

where, as above, z^s denotes Z_{G_s}. Then L is differentiable with respect to z_i whenever $z_i \neq 0$. In particular, since both $\hat{x}_i$ and $\hat{x}_j$ are nonzero by assumption, it follows that

$$\left.\frac{\partial L}{\partial z_i}\right|_{z=\hat{x}} = \left.\frac{\partial L}{\partial z_j}\right|_{z=\hat{x}} = 0.$$

Expanding the partial derivatives leads to

$$-2\lambda a_i^t (y - A\hat{x}) + (1-\mu)\mathrm{sign}(\hat{x}_i) + \mu \frac{\hat{x}_i}{\|\hat{x}^s\|_2} = 0,$$

$$-2\lambda a_j^t (y - A\hat{x}) + (1-\mu)\mathrm{sign}(\hat{x}_j) + \mu \frac{\hat{x}_j}{\|\hat{x}^s\|_2} = 0.$$

Subtracting one equation from the other gives

$$2\lambda (a_j^t - a_i^t)(y - A\hat{x}) + \mu \frac{\hat{x}_i - \hat{x}_j}{\|\hat{x}^s\|_2} = 0.$$

Hence

$$\begin{aligned} \frac{|\hat{x}_i - \hat{x}_j|}{\|\hat{x}^s\|_2} &= \frac{2\lambda}{\mu} |(a_j^t - a_i^t)(y - A\hat{x})| \\ &\leq \frac{2\lambda}{\mu} \|a_j - a_i\|_2 \|y - A\hat{x}\|_2 \\ &\leq \frac{2\lambda}{\mu} \sqrt{2(1-\rho(i,j))} \|y\|_2. \end{aligned}$$

In the last step, we use the fact that

$$\|y - A\hat{x}\|_2 \leq \|y - A\mathbf{0}\|_2 = \|y\|_2.$$

Rearranging gives

$$\frac{1}{2\lambda} \frac{|\hat{x}_i - \hat{x}_j|}{\|y\|_2} \leq \frac{\sqrt{2(1-\rho(i,j))}\|\hat{x}^s\|_2}{\mu},$$

which is the desired conclusion. □

Let us illustrate the above result using the CLOT formulation. In the case of CLOT formulation we have $g = 1$, $\mathcal{G} = \{G_1\}$, $G_1 = \{1, \cdots, n\}$, and the inequality in (21) becomes

$$d(i,j) \leq \frac{\sqrt{2(1-\rho(i,j))}\|\hat{x}\|_2}{\mu}, \tag{22}$$

where $\hat{x}$ is the solution of the CLOT formulation, and

$$d(i,j) = \frac{|\hat{x}_i - \hat{x}_j|}{2\lambda \|y\|_2}. \tag{23}$$

Now suppose that two indices i and j are highly correlated such that $\rho(i,j) \approx 1$, so that the right hand side of the inequality in (22) is almost equal to zero. Combining this with (23) we can conclude $\hat{x}_i \approx \hat{x}_j$, so CLOT assigns similar weights to highly correlated variables. □

Though the focus of the present paper is not on the GL formulation, we digress briefly to discuss the implications of Theorem 7 for GL. This theorem also implies that the GL formulation exhibits the grouping effect, because GL is a special case of SGL with $\mu = 1$. Indeed, it can be observed from (21) that the bound on the right side is minimized by setting $\mu = 1$, that is, using GL instead of SGL. This is not surprising, because SGL not only tries to minimize the number of distinct groups containing the support of $\hat{x}$, but within each group, tries to choose as few elements as possible. Thus, within each group, SGL inherits the weaknesses of LASSO. Thus one would expect that, within each group, the feature set chosen would become more sensitive as we decrease μ.

2.4 Robust Sparse Recovery of the SGL and CLOT Formulations

In this subsection, we present some sufficient conditions for the SGL and CLOT formulations to achieve robust sparse recovery. When CLOT is specialized to LASSO by setting $\mu = 0$, the sufficient condition reduces to the "tight" bound given in Theorem 3.

Recall the definitions. The CLOT norm with parameter μ is given by

$$\|z\|_{C,\mu} := (1-\mu)\|z\|_1 + \mu \|z\|_2,$$

while the SGL norm is given by

$$\|z\|_{SGL,\mu} := \sum_{i=1}^{g} [(1-\mu)\|z_{G_i}\|_1 + \mu \|z_{G_i}\|_2]$$

Recall also the problem set-up. The measurement vector y equals $Ax + \eta$ where $\|\eta\|_2 \leq \epsilon$, a known upper bound. The recovered vector $\hat{x}$ is defined as

$$\hat{x} = \operatorname*{argmin}_{z} \|z\|_{\mathrm{SGL},\mu} \text{ s.t. } \|Az - y\|_2 \leq \epsilon \tag{24}$$

if SGL is used, and as

$$\hat{x} = \operatorname*{argmin}_{z} \|z\|_{C,\mu} \text{ s.t. } \|Az - y\|_2 \leq \epsilon \tag{25}$$

if CLOT is used.

Definition 8 *A matrix $A \in \mathbb{R}^{m \times n}$ is said to satisfy the ℓ_2* ***robust null space property (RNSP)*** *if there exist constants $\rho \in (0,1)$ and $\tau \in \mathbb{R}_+$ such that, for all sets $S \subseteq [n]$ with $|S| \leq k$, we have*

$$\|h_S\|_2 \leq \frac{\rho}{\sqrt{k}} \|h_{S^c}\|_1 + \frac{\tau}{\sqrt{k}} \|Ah\|_2. \tag{26}$$

This property was apparently first introduced in (Foucart and Rauhut, 2013, Definition 4.21). Note that the definition in Foucart and Rauhut (2013) has just τ in place of $\tau/\sqrt{k}$. It is easy to show that, if (26) holds, then

$$\|h_S\|_1 \leq \rho \|h_{S^c}\|_1 + \tau \|Ah\|_2.$$

The following result is established in Ranjan and Vidyasagar (2016) in the context of group sparsity, but is new even for conventional sparsity. The reader is directed to that source for the proof.

Theorem 9 *((Ranjan and Vidyasagar, 2016, Theorem 5)) Suppose that, for some number $t > 1$, the matrix A satisfies the RIP of order tk with constant $\delta_{tk} < \sqrt{(t-1)/t}$. Let δ be an abbreviation for δ_{tk}, and define the constants*

$$\nu := \sqrt{t(t-1)} - (t-1) \in (0, 0.5), \tag{27}$$

$$a := [\nu(1-\nu) - \delta(0.5 - \nu + \nu^2)]^{1/2}, \tag{28}$$

$$b := \nu(1-\nu)\sqrt{1+\delta}, \tag{29}$$

$$c := \left[\frac{\delta \nu^2}{2(t-1)} \right]^{1/2}. \tag{30}$$

Then A satisfies the ℓ_2-robust null space property with

$$\rho = \frac{c}{a} < 1, \tau = \frac{b\sqrt{k}}{a^2}. \tag{31}$$

In Theorem 10 below, it is assumed that the matrix A satisfies the RIP of order tk with $\delta_{tk} < \sqrt{(t-1)/t}$, in accordance with Theorem 3. With this assumption, we prove bounds on the residual error $\|\hat{x} - x\|_1$ with SGL; the bounds for CLOT can be obtained simply by setting $g = 1$ in the SGL bounds. Note that, once bounds for $\|\hat{x} - x\|_1$ are proved, it is possible to extend the bounds to $\|\hat{x} - x\|_p$ for all $p \in [1, 2]$; see (Foucart and Rauhut, 2013, Theorem 4.22).

Theorem 10 *Suppose $x \in \mathbb{R}^n$ and that $A \in \mathbb{R}^{m \times n}$ satisfies the RIP of order tk with constant $\delta = \delta_{tk} < \sqrt{(t-1)/t}$, and define constants ρ, τ as in* (31)*. Suppose that*

$$\mu < \frac{1-\rho}{\sqrt{g}(1+\rho)}. \tag{32}$$

Define

$$\gamma = \frac{\mu\sqrt{g}}{1-\mu}. \tag{33}$$

With these assumptions,

$$\|\hat{x} - x\|_1 \leq C\sigma_k(x, \|\cdot\|_1) + D\epsilon, \tag{34}$$

$$\|\hat{x} - x\|_p \leq \frac{1}{k^{1-1/p}}[(1+\rho)C\sigma_k(x, \|\cdot\|_1) + ((1+\rho)D + 2\tau)\epsilon], \ \forall p \in (1, 2], \tag{35}$$

where

$$C = \frac{2(1+\rho)}{(1-\gamma) - (1+\gamma)\rho}, D = \frac{4\tau}{(1-\gamma) - (1+\gamma)\rho}. \tag{36}$$

The proof of the above theorem is presented in an appendix, due to its length.

In the above theorem, we started with the restricted isometry constant δ and computed an upper bound on μ in order for SGL and CLOT to achieve robust sparse recovery. As δ_{tk} gets closer to the limit $\delta_{tk} < \sqrt{(t-1)/t}$ (which is known to be the best possible in view of Theorem 4), the limit on μ would approach zero. It is also possible to start with μ and find an upper bound on δ, by rearranging the inequalities. As this involves just routine algebra, we simply present the final bound. Given the number $t \geq 4/3$, define ν as in (27), and define

$$\theta_1 := \nu(1-\nu) \in (0, 0.25), \theta_2 := 0.5 - \theta_1 \in (0.25, 0.5), \tag{37}$$

$$\theta_3 := \frac{\nu^2}{2(t-1)}. \tag{38}$$

Given $\mu > 0$, define γ as in (33), and define

$$\bar{\rho} := \frac{1-\gamma}{1+\gamma}.$$

If the matrix A satisfies the RIP of order tk with a constant $\delta = \delta_{tk}$, then SGL achieves robust sparse recovery of order k provided

$$\delta < \bar{\rho}^2 \frac{\theta_1}{\theta_3 + \bar{\rho}^2\theta_2}, \tag{39}$$

3. Numerical Examples

In this section we present two simulation studies, to demonstrate the the grouping effect of CLOT and the lack of robust sparse recovery of EN, respectively.

3.1 Grouping Effect of CLOT

To illustrate that CLOT demonstrates the grouping effect as does EN (see Theorem 7), we ran the same example as at the end of (Zou and Hastie, 2005, Section 5). Specifically, we chose Z_1 and Z_2 to be two independent $U(0, 20)$ random variables, and the observation y as $N(Z_1 + 0.1Z_2, 1)$. The six observations were

$$x_1 = Z_1 + \epsilon_1, x_2 = -Z_1 + \epsilon_2, x_3 = Z_1 + \epsilon_3, x_4 = Z_2 + \epsilon_4, x_5 = -Z_2 + \epsilon_5, x_6 = Z_2 + \epsilon_6,$$

where the ϵ_i are i.i.d. $N(0, 1/16)$. The objective was to express y as a linear combination of x_1 through x_6. In other words, we wished to express $y = X\boldsymbol{\beta}$, where X is the matrix with the x_i as columns, and $\boldsymbol{\beta}$ is a 6×1 vector. Ideally the outcome should be to assign

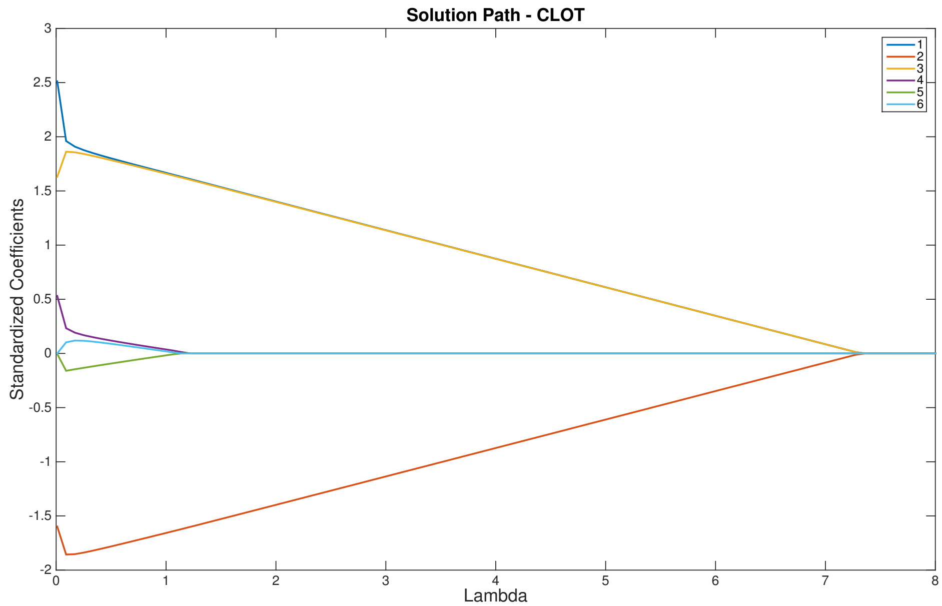

Figure 1: Solution Paths for CLOT

high weights to the correlated group x_1, x_2, x_3 and low weights to x_4, x_5, x_6. Therefore, if $\boldsymbol{\beta}$ denotes the six-dimensional coefficient vector, we should have that

$$\beta_1 \approx -\beta_2 \approx \beta_3,$$

and that β_4 through β_6 are much smaller than β_1 through β_3.

The three algorithms LASSO, EN, and CLOT were implemented via the Lagrangian formulation in (2), with λ being increased. Clearly, when λ is sufficiently large, the optimal value of $\boldsymbol{\beta}$ is the zero vector. The "sufficiently large" value of λ varies from one algorithm to the next. Figures 1 through 3 show the solution trajectories of the three algorithms as functions of λ, with μ set equal to 0.5. From Figures 1 and 2, it is clear that both CLOT and EN very quickly reach the correct proportionalities between the large coefficient values, which eventually become smaller and go to zero as λ becomes larger. In contrast, the LASSO solutions are quite inaccurate.

3.2 Lack of Robust Sparse Recovery by EN

In this subsection we illustrate Theorem 6. In this set-up, $n = 4,000$ and $k = 3$. The first three components of the vector x are assigned values at random using the `Matlab rand` function, which resulted in $[\ 0.8147 \quad 0.9058 \quad 0.1270\ ]$. The remaining components of x were set equal to zero. We could have chosen not just the values but also the location of the nonzero components at random; but this would have been just a permuation of the above example.

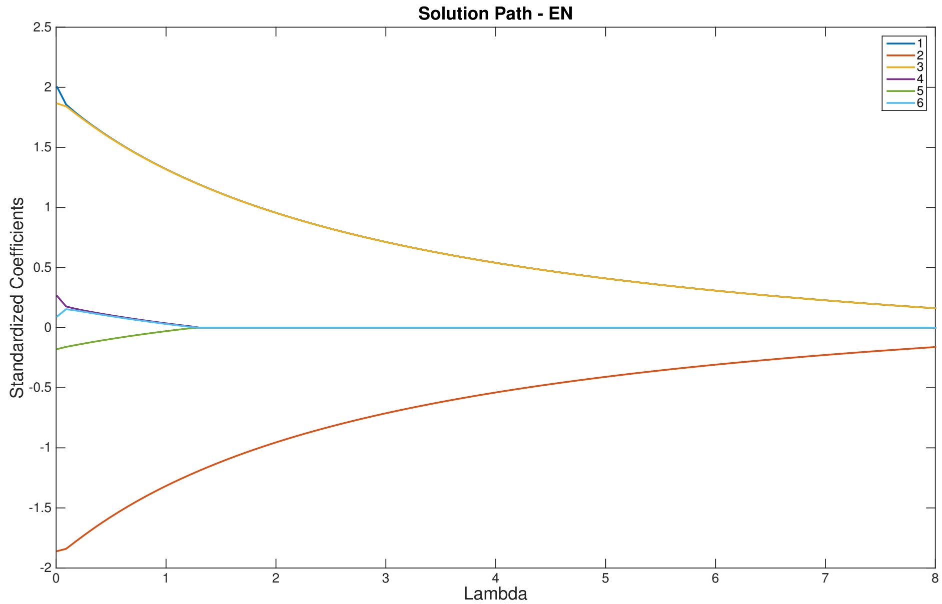

Figure 2: Solution Path for EN

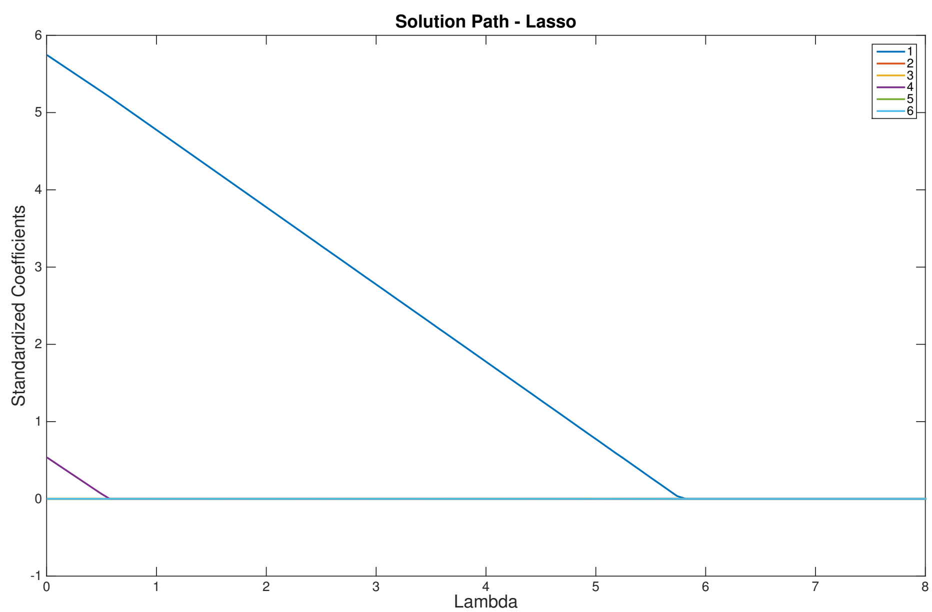

Figure 3: Solution Path for LASSO

Next, in order to achieve robust sparse recovery of order k, following Theorem 3 we needed to choose a $t \geq 4/3$ and then choose a matrix A such that A satisfied the RIP of order $\lceil tk \rceil$ with constant $\delta < \sqrt{(t-1)/t}$. We chose $t = 1.5$, which resulted in $\lceil tk \rceil = 5$, and $\delta_5 < \sqrt{1/3} \approx 0.5774$. We chose $\delta_5 = 0.4$. Therefore we had to choose an integer m and a matrix $A \in \mathbb{R}^{m \times n}$ such that A satisfied the RIP of order 5 with constant $\delta_5 \leq 0.4$. Such a matrix was constructed using the deterministic procedure suggested in (DeVore, 2007). This required the choice of an integer $r \geq 2$ and a prime number p such that

$$p \geq \max\left\{ \frac{(\lceil tk \rceil - 1)r}{\delta}, n^{1/(r+1)} \right\} = \max\{20, 15.87\} = 20.$$

and led to a binary matrix of dimensions $p^2 \times p^{r+1}$. The smallest prime number greater than 20 is $p = 23$, and $m = p^2 = 529$. Note that $m < n/4$. The main advantages of the construction in (DeVore, 2007) are that (i) the construction is deterministic, (ii) the matrix A is binary, and (iii) only a fraction $1/p$ of the elements of A are equal to 1, and the rest are equal to zero. This makes computation very fast.

Once the matrix A was chosen, we defined the measured vector as $y = Ax$; that is, we did not introduce any measurement noise. Then we computed estimates of x using both CLOT and EN. With $\delta = 0.4$, the constant ρ defined in (31) becomes 0.6551, and the bound in (32), after substituting $g = 1$, becomes $\mu < 0.2084$. Therefore we chose $\mu = 0.2$, and defined

$$\hat{x}_{\text{CLOT}} = \operatorname*{argmin}_z \|z\|_{C,\mu} \text{ s.t. } Az = y,$$

$$\hat{x}_{\text{EN}} = \operatorname*{argmin}_z (1-\mu)\|z\|_1 + \mu\|z\|_2^2 \text{ s.t. } Az = y,$$

with the same value of $\mu = 0.2$ in both cases. Then we replaced x by $10^c x$ for $c = 1, 2, 3, 4$ and recomputed the estimates. Because $m < n/4$, according to Theorem 7, EN should fail as the norm of the vector x is increased.

The above constrained optimization problems were solved using the cvx package in `Matlab`, on an HP Pavilion laptop. The actual results are shown in Table 1. For compactness only the first three components of $\hat{x}$ are shown. As can be seen, when $c = 0$ or $c = 1$, both CLOT and EN give the correct answer. However, for larger values of c, the CLOT estimate simply got multiplied by the same scale factor, whereas the EN estimate started diverging from the true value with $c = 2$; the algorithm failed to converge with $c = 4$.

4. Discussion and Concluding Remarks

In this paper we have introduced a new optimization formulation called CLOT (Combined L-One and Two), wherein the regularizer is a convex combination of the ℓ_1- and ℓ_2-norms. This formulation differs from the Elastic Net (EN) formulation, in which the regularizer is a convex combination of the ℓ_1- and ℓ_2-norm *squared*. This seemingly simple modification has fairly significant consequences. In particular, it is shown in this paper that the EN formulation *does not achieve* robust recovery of sparse vectors in the context of compressed sensing, whereas the new CLOT formulation does so. Also, like EN but unlike LASSO, the CLOT formulation achieves the grouping effect, wherein coefficients of highly correlated columns of the measurement (or design) matrix are assigned roughly comparable values. It

c	$\hat{x}_{\text{CLOT}}$	$\hat{x}_{\text{EN}}$
0	0.8147	0.8147
	0.9058	0.9058
	0.1270	0.1270
1	8.1472	8.1472
	9.0579	9.0579
	1.2699	1.2699
2	81.4724	24.1502
	90.5792	31.8304
	12.6987	4.7944
3	814.7236	111.2433
	905.7918	132.8940
	126.9868	15.3818
4	8.1472×10^3	NaN
	9.0579×10^3	NaN
	1.2699×10^3	NaN

Table 1: Comparison of solutions of the CLOT and EN algorithms as the unknown vector is scaled by successive powers of 10. The CLOT output simply scales by the same factor, whereas the EN output begins to be incorrect already when the scale factor is 100. For a scale factor of 10^4, the algorithm fails.

is noteworthy that LASSO does not have the grouping effect and EN (as shown here) does not achieve robust sparse recovery. Therefore the CLOT formulation combines the best features of both LASSO (robust sparse recovery) and EN (grouping effect).

The CLOT formulation is a special case of another one called SGL (Sparse Group LASSO) which was introduced into the literature previously, but without any analysis of either the grouping effect or robust sparse recovery (Simon et al., 2013). It is shown here that SGL achieves robust sparse recovery, and also achieves a version of the grouping effect in that coefficients of highly correlated columns of the measurement (or design) matrix are assigned roughly comparable values, *if the columns belong to the same group*.

There are several papers in the literature that discuss LASSO-like formulations for group sparsity; some of these are discussed here. First, there is a companion paper by a subset of the present authors (Ahsen and Vidyasagar, 2016), which studies the problem of robust sparse recovery with SGL-like formulations, *but with restrictions on the group size*. In contrast, in the present paper, there is no such restriction, which is why the results derived here for the SGL formulation can be directly applied to the CLOT formulation. Second, for the case where several columns of the matrix A are highly correlated, (Bühlmann et al., 2013) suggests a two-stage process whereby first correlated columns are clustered, and second, a variant of LASSO is applied. In a discussion of this paper, namely (Bien and Wegkamp, 2013), all the LASSO variants together with EN are run on various test data. For the purposes of the present discussion, the salient observation is that the EN formulation performed roughly as well – no better and no worse – compared almost all the

LASSO variants. Finally, in (van de Geer, 2014), a general theory is presented whereby the fully decomposable ℓ_1-norm is replaced by a weakly decomposable norm, and oracle bounds are derived for $\|\hat{x} - x\|$ provided that the index set $[n]$ is divided into an "allowed set" and its complement. There is also some discussion of overlapping group decompositions. Specifically, when an element of the index set $[n]$ appears in two groups, the corresponding column of A is simply replicated to remove the overlap. However, if two columns of A are identical (and normalized), then the RIP constant δ_2 would equal zero, as would δ_k for $k \geq 2$. Therefore, if the case of overlapping groups is handled in this manner, then any analysis based on RIP would be infructuous. The above discussion is quite cursory, and the reader may consult these references for fuller details.

It would be worthwhile to study the behavior of SGL with overlapping groups. There are variants of SGL with overlapping groups, provided they satisfy some additional constraints; see (Jenetton et al., 2011; Obozinski et al., 2011) for example. However, in a companion paper (Ahsen and Vidyasagar, 2016), it is shown that the assumptions of (Jenetton et al., 2011; Obozinski et al., 2011) still enforce a nonoverlap constraint, but in a nonobvious fashion. As pointed out in the previous paragraph, the approach of introducing duplicate columns into A to eliminate overlap would render any analysis based on RIP impossible. Thus a suitable approach remains to be discovered.

Acknowledgments

The authors thank two anonymous reviewers and the Handling Editor for their careful reading of earlier versions of this paper and many helpful comments. Theorem 5 and its proof were supplied by one of the reviewers. This research was supported by the National Science Foundation under Award #ECCS-1306630 and by the Cancer Prevention and Research Institute of Texas (CPRIT) under Award No. RP140517, and by the Department of Science and Technology, Government of India.

Appendix A: Proof of Theorem 10

Proof Hereafter we write z^i instead of z_{G_i} in the interests of brevity.

Define $h = \hat{x} - x \in \mathbb{R}^n$. From the definition of the estimate, we have that

$$\|x\|_{SGL,\mu} \geq \|\hat{x}\|_{SGL,\mu} = \|x + h\|_{SGL,\mu}$$

From the definition of the SGL norm, this expands to

$$(1-\mu)\|x\|_1 + \mu \sum_{j=1}^{g} \|x^j\|_2 \geq (1-\mu)\|x+h\|_1 + \mu \sum_{j=1}^{g} \|x^j + h^j\|_2.$$

This can be rearranged as

$$(1-\mu)(\|x\|_1 - \|x+h\|_1) + \mu \sum_{j=1}^{g} (\|x^j\|_2 - \|(x^j + h^j\|_2) \geq 0. \tag{40}$$

We will work separately on each of the two terms separately. First, by the triangle inequality, we have that

$$\|x^j + h^j\|_2 \geq \|x^j\|_2 - \|h^j\|_2, \text{ or } \|h^j\|_2 \geq \|x^j\|_2 - \|x^j + h^j\|_2.$$

As a consequence,

$$\sum_{j=1}^{g} \|h^j\|_2 \geq \sum_{j=1}^{g} [\|x^j\|_2 - \|x^j + h^j\|_2]$$

From Schwarz' inequality, we get

$$\sum_{j=1}^{g} \|h^j\|_2 \leq \sqrt{g}\|h\|_2 \leq \sqrt{g}\|h\|_1 \leq \sqrt{g}(\|h_S\|_1 + \|h_{S^c}\|_1),$$

for any $S \subseteq [n]$. Combining everything gives

$$\sum_{j=1}^{g} [\|x^j\|_2 - \|x^j + h^j\|_2] \leq \sqrt{g}(\|h_S\|_1 + \|h_{S^c}\|_1), \tag{41}$$

for any subset $S \subseteq [n]$. Second, for any subset $S \subseteq [n]$, the decomposability of $\| \cdot \|_1$ implies that

$$\|x\|_1 = \|x_S\| + \|x_{S^c}\|_1,$$

while the triangle inequality implies that

$$\begin{aligned} \|x + h\|_1 &= \|x_S + h_S\|_1 + \|x_{S^c} + h_{S^c}\|_1 \\ &\geq \|x_S\|_1 - \|h_S\|_1 + \|h_{S^c}\|_1 - \|x_{S^c}\|_1. \end{aligned}$$

Therefore

$$\|x\|_1 - \|x + h\|_1 \leq \|h_S\|_1 - \|h_{S^c}\|_1 + 2\|x_{S^c}\|_1. \tag{42}$$

If we now choose S to be the set corresponding to the k largest elements of x by magnitude, then

$$\|x_{S^c}\|_1 = \sigma_k(x, \| \cdot \|_1) =: \sigma_k.$$

With this choice of S, (42) becomes

$$\|x\|_1 - \|x + h\|_1 \leq \|h_S\|_1 - \|h_{S^c}\|_1 + 2\sigma_k. \tag{43}$$

Substituting the bounds (41) and (42) into (40) gives

$$0 \leq \mu\sqrt{g}(\|h_S\|_1 + \|h_{S^c}\|_1) + (1 - \mu)(\|h_S\|_1 - \|h_{S^c}\|_1 + 2\sigma_k).$$

Now recall the definition of the constant γ from (33). Using this definition, the above inequality can be rearranged as

$$0 \leq \gamma(\|h_S\|_1 + \|h_{S^c}\|_1) + (\|h_S\|_1 - \|h_{S^c}\|_1 + 2\sigma_k).$$

and equivalently as

$$(1 - \gamma)\|h_{S^c}\|_1 - (1 + \gamma)\|h_S\|_1 \leq 2\sigma_k. \tag{44}$$

This is the first of two equations that we need.

Now we derive the second equation. From Theorem 9, we know that the matrix A satisfies the l_2-robust null space property, namely (26). An application of Schwarz' inequality shows that

$$\|h_S\|_1 \leq \rho \|h_{S^c}\|_1 + \tau \|Ah\|_2.$$

However, because both x and $\hat{x}$ are feasible for the optimization problem in (24), we get

$$\|Ah\|_2 = \|A\hat{x} + \eta - (Ax + \eta)\|_2 \leq 2\epsilon.$$

Substituting this bound for $\|Ah\|_2$ gives us the second equation we need, namely

$$\|h_S\|_1 \leq \rho \|h_{S^c}\|_1 + 2\tau\epsilon,$$

or equivalently

$$-\rho \|h_{S^c}\|_1 + \|h_S\|_1 \leq 2\tau\epsilon. \tag{45}$$

The two inequalities (44) and (45) can be written compactly as

$$M \begin{bmatrix} \|h_{S^c}\|_1 \\ \|h_S\|_1 \end{bmatrix} \leq \left\{ \begin{bmatrix} 2 \\ 0 \end{bmatrix} \sigma_k + \begin{bmatrix} 0 \\ 2\tau \end{bmatrix} \epsilon \right\}, \tag{46}$$

where the coefficient matrix M is given by

$$M = \begin{bmatrix} (1-\gamma) & -(1+\gamma) \\ -\rho & 1 \end{bmatrix}$$

The matrix M has positive diagonal elements (recall that $\gamma < 1$), and negative off-diagonal elements. Therefore, if $\det(M) > 0$, then every element of M^{-1} is positive, in which we can multiply both sides of (46) by M^{-1}. Now

$$\det(M) = 1 - \gamma - \rho(1+\gamma) > 0 \iff \rho < \frac{1-\gamma}{1+\gamma}.$$

Recall the definition of γ from (33). Now routine algebra shows that

$$\rho < \frac{1-\gamma}{1+\gamma} \iff \rho < \frac{1-\mu\sqrt{g}}{1+\mu\sqrt{g}} \iff \mu < \frac{1-\rho}{\sqrt{g}(1+\rho)},$$

which is precisely (32). Thus we can multiply both sides of (46) by M^{-1}, which gives

$$\begin{bmatrix} \|h_{S^c}\|_1 \\ \|h_S\|_1 \end{bmatrix} \leq \frac{1}{\det(M)} \begin{bmatrix} 1 & (1+\gamma) \\ \\ \rho & (1-\gamma) \end{bmatrix} \left\{ \begin{bmatrix} 2 \\ 0 \end{bmatrix} \sigma_k + \begin{bmatrix} 0 \\ 2\tau \end{bmatrix} \epsilon \right\}.$$

Clearing out the matrix multiplication gives

$$\begin{bmatrix} \|h_{S^c}\|_1 \\ \|h_S\|_1 \end{bmatrix} \leq \frac{1}{\det(M)} \left\{ \begin{bmatrix} 2 \\ 2\rho \end{bmatrix} \sigma_k + \begin{bmatrix} 2\tau(1+\gamma) \\ 2\tau(1-\gamma) \end{bmatrix} \epsilon \right\}. \tag{47}$$

Now the triangle inequality states that

$$\|h\|_1 \leq \|h_{S^c}\|_1 + \|h_S\|_1 = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \|h_{S^c}\|_1 \\ \|h_S\|_1 \end{bmatrix}$$

Substituting from (47) gives

$$\|h\|_1 \leq \frac{1}{\det(M)} [2(1+\rho)\sigma_k + 4\tau\epsilon].$$

By substituting that $\det(M) = (1-\gamma) - (1+\gamma)\rho$, we get (34) with the constants as defined in (36).

To prove (35), suppose $p \in (1, 2]$. This part of the proof closely follows that of (Foucart and Rauhut, 2013, Theorem 4.22), except that we provide explicit values for the constants. Let Λ_0 denote the index set of the k largest components of h by magnitude. Then

$$\|h\|_p \leq \|h_{\Lambda_0}\|_p + \|h_{\Lambda_0^c}\|_p.$$

We will bound each term separately. First, by (Foucart and Rauhut, 2013, Theorem 2.5) and (34), we get

$$\|h_{\Lambda_0^c}\|_p \leq \frac{1}{k^{1-1/p}} \|h\|_1 \leq \frac{1}{k^{1-1/p}} (C\sigma_k + D\epsilon). \tag{48}$$

Now we apply in succession Hölder's inequality, the robust null space property, the fact that $\|Ah\|_2 \leq 2\epsilon$, and (34). This gives

$$\begin{aligned} \|h_{\Lambda_0}\|_p &\leq k^{1/p-1/2} \|h_{\Lambda_0}\|_2 \\ &\leq \frac{k^{1/p-1/2}}{\sqrt{k}} [\rho \|h_{\Lambda_0^c}\|_1 + \tau \|Ah\|_2] \\ &\leq \frac{1}{k^{1-1/p}} [\rho(C\sigma_k + D\epsilon) + 2\tau\epsilon] \\ &= \frac{1}{k^{1-1/p}} [\rho C\sigma_k + (\rho D + 2\tau)\epsilon]. \end{aligned} \tag{49}$$

Adding (48) and (49) gives (35).

■

References

Mehmet Eren Ahsen and Mathukumalli Vidyasagar. Error bounds for compressed sensing algorithms with group sparsity: A unified approach. *Applied and Computational Harmonic Analysis*, page to appear, 2016.

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Pivotal estimation via square-root Lasso in nonparametric regression. *The Annals of Statistics*, 42(2):757–788, 2014.

Jacob Bien and Marten Wegkamp. Discussion of correlated variables in regression: Clustering and sparse estimation". *Journal of Statistical Planning and Inference*, 143(11): 1859–1862, November 2013.

Peter Bühlmann, Philipp Rütimann, Sara van de Geer, and Cun-Hui Zhang. Correlated variables in regression: Clustering and sparse estimation. *Journal of Statistical Planning and Inference*, 143(11):1835–1858, November 2013.

T. Cai and A. Zhang. Sparse representation of a polytope and recovery of sparse signals and low-rank matrices. *IEEE Transactions on Information Theory*, 60(1):122–132, 2014.

Emmanuel Candès. The restricted isometry property and its implications for compresed sensing. *Comptes rendus de l'Académie des Sciences, Série I*, 346:589–592, 2008.

Emmanuel J. Candès and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, December 2005.

Emmanuel J. Candès, J. Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications in Pure and Applied Mathematics*, 59(8): 1207–1223, August 2006.

Albert Cohen, Wolfgang, Dahmen, and Ronald DeVore. Compressed sensing and best k-term approximation. *Journal of the American Mathematical Society*, 22(1):211–231, January 2009.

Mark A. Davenport, Marco F. Duarte, Yonina C. Eldar, and Gitta Kutyniok. Introduction to compressed sensing. In Yonina C. Eldar and Gitta Kutyniok, editors, *Compressed Sensing: Theory and Applications*, pages 1–68. Cambridge University Press, Cambridge, UK, 2012.

Ronald DeVore. Deterministic construction of compressed sensing matrices. *Journal of Complexity*, 23:918–925, 2007.

D. L. Donoho. For most large underdetermined systems of linear equations, the minimal ℓ_1-norm solution is also the sparsest solution. *Communications in Pure and Applied Mathematics*, 59(6):797–829, 2006.

Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing.* Springer-Verlag, 2013.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

R. Jenetton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for hierarchical sparse coding. *Journal of Machine Learning Research*, 12:2297–2334, 2011.

B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24:227–234, 1995.

S. Negabhan, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, December 2012.

Guillaume Obozinski, Laurent Jacob, and Jean-Philippe Vert. Group lasso with overlaps: The latest group lasso approach. *arxiv*, page 1110.0413, 2011.

Michael R. Osborne, Brett Presnell, and Berwin A. Turlach. On the LASSO and its dual. *Journal of Computational and Graphical Statistics*, 9:319–337, 2000.

Shashank Ranjan and Mathukumalli Vidyasagar. Tight performance bounds for compressed sensing with group sparsity. *arXiv*, page 1606.05889v1, 2016.

Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. A sparse group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1), 1996.

A. N. Tikhonov. On the stability of inverse problems. *Doklady Akademii Nauk SSSR*, 39(5): 195–198, 1943.

Sara van de Geer. Weakly decomposable regularization penalties and structured sparsity. *Scandinavian Journal of Statistics*, 41:72–86, 2014.

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

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