Density Corrected Sparse Recovery when R.I.P. Condition is Broken

Ming Lin, Zhengzhong Lan, Alexander G. Hauptmann
Carnegie Mellon University
Pittsburgh, PA, USA

Abstract

The Restricted Isometric Property (R.I.P.) is a very important condition for recovering sparse vectors from high dimensional space. Traditional methods often rely on R.I.P or its relaxed variants. However, in real applications, features are often correlated to each other, which makes these assumptions too strong to be useful. In this paper, we study the sparse recovery problem in which the feature matrix is strictly non-R.I.P. . We prove that when features exhibit cluster structures, which often happens in real applications, we are able to recover the sparse vector consistently. The consistency comes from our proposed density correction algorithm, which removes the variance of estimated cluster centers using cluster density. The proposed algorithm converges geometrically, achieves nearly optimal recovery bound $O(s^2 \log(d))$ where s is the sparsity and d is the nominal dimension.

1 Introduction

In high dimensional statistics, an important problem is to recover a sparse vector from a small number of observations whose number is usually much smaller than the nominal dimension of the observation matrix. [Candes *et al.*, 2006] proves that sparse recovery is possible as long as the feature matrix satisfies the *Restricted Isometric Property* (R.I.P.) condition. Informally, the R.I.P. condition requires any set of columns of the feature matrix with size less than the sparsity to be linearly independent. More likely than not, the R.I.P. condition is too strong to be useful in real world applications. Many R.I.P.-like relaxations have been proposed in the past decades [Van De Geer and Bhlmann, 2009; Foucart, 2012]. These alternatives usually have better constants, or more general forms of the restricted isometric inequalities. See [Van De Geer and Bhlmann, 2009] for a nice survey.

Although there is a vast literature studying relaxed R.I.P.-like conditions, their core requirement is that the feature matrix must be well-conditioned on the sparse subspace. In other words, their largest and smallest singular values must be close to unit. For many real applications, such assumptions are still too strong to be useful. Considering a simple case where we duplicate some columns of the feature matrix, these R.I.P.-like conditions are no longer true because we can find a subset of columns whose smallest singular value is exactly zero.

In this paper we consider a prevalent non-R.I.P. setting in which the features form cluster structures, as can be seen in many machine learning [Lehiste, 1976] and computer vision problems [Lan *et al.*, 2013; Lowe, 2004]. Due to the fact that many features extractors are similar to each others and they reflect the characteristics of the same image, vision features are often correlated and have cluster structures. This correlation is even stronger in those systems that have thousands to millions of features [Lan *et al.*, 2013; Gan *et al.*, 2015a; 2015b].

Under this setting, instead of requiring the observed feature matrix to be R.I.P., we only have the same requirement on the cluster center matrix, which is much easier to satisfy. A trivial solution to perform sparse recovery under this setting is by first clustering the features and then applying the conventional sparse recovery methods on the clustered feature matrix. However, due to the random perturbation, the estimated cluster center is the biased version of the real cluster centers. Unlike instance clustering where we can improve the estimation accuracy by having more samples, the bias in feature clustering process cannot be asymptotically removed even if infinite training instances were given. This inconsistency happens in all conventional methods, including convex/non-convex sparse regularizers , greedy methods and their stochastic or adaptive variants [Agarwal *et al.*, 2012; Shalev-Shwartz and Tewari, 2011; Ghadimi and Lan, 2013; Ji Liu, 2013; Jin *et al.*, 2013; Zhaoran Wang, 2013; Lin *et al.*, 2014b; Lin and Xiao, 2014; Yang *et al.*, 2014; Lin *et al.*, 2014a].

In this paper, under cluster assumption, we develop a consistent sparse recovery method, called *density correction*. The key idea of density correction is that although we cannot eliminate the bias in feature clustering, we are able to estimate the variance of this bias via the cluster density. We can correct the bias in the gradient of the loss function with cluster density in the sparse recovery process. When combining the density correction with hard iterative thresholding, we obtain a consistent sparse estimator.

The remaining paper is organized as following. In section 2 we briefly review several closely related works. We introduce our notations and backgrounds of our method in section 3.

Section 4 gives the details of our algorithm. In section 5 we prove the geometrical convergence rate and the consistency of our method. We conclude our paper with discussions in section 6.

2 Related Work

In this section, we briefly review several closely related works. Our review is by no means to be comprehensive but to capture the backgrounds of our approaches.

Sparse recovery methods can be roughly grouped into three categories: convex regularizer, non-convex regularizer and greedy methods. The convex regularizer mainly bases on ℓ_1-norm LASSO problems [Shah, 2012; Xu *et al.*, 2010; Tibshirani, 1996]. The non-convex regularizer is proved to be more accurate than convex methods in several cases [Zhaoran Wang, 2013; Xiang *et al.*, 2013; Loh and Wainwright, 2013; Gong *et al.*, 2013; Zhang, 2012]. Recently, greedy methods are rediscovered in the sparse community due to its simplicity and efficiency [Xiaotong Yuan, 2014; Ji Liu, 2013; Blumensath, 2012; Foucart, 2011]. However, all these studies are based on R.I.P.-like assumptions therefore fail to cover the non-R.I.P. problems discussed in this paper.

The gradient correction technique dates back to the corrected least square regression [Markovsky and Van Huffel, 2007]. The same technique is considered in sparse regression under feature corruption settings [Chen *et al.*, 2013; Loh and Wainwright, 2012]. These methods require that the feature matrix to be R.I.P. when noise is small. In our setting, the observed feature matrix becomes non-R.I.P. when the perturbation is small. Also, traditional methods require the noise level to be known while our method not.

3 Notations and Backgrounds

The sparse recovery problem aims to recover a sparse vector $\bar{\mathbf{w}}_* \in \mathbb{R}^d$ from feature matrix $\bar{X} \in \mathbb{R}^{n\times d}$ and label $\mathbf{y} \in \mathbb{R}^n$. The sparsity of $\bar{\mathbf{w}}_*$ is denoted by s and its support set is denoted as S_* . We assume that $\mathbf{y}$ is generated by a linear model

$$\mathbf{y} = \bar{X}\bar{\mathbf{w}}_* + \boldsymbol{\xi} \,, \tag{1}$$

where $\boldsymbol{\xi}$ is additive subgaussian noise with noise level

$$\|\frac{1}{\sqrt{n}}\bar{X}^{\mathrm{T}}\boldsymbol{\xi}\|_\infty \le \xi \,.$$

We denote $\bar{X}_s$ the submatrix of $\bar{X}$ whose columns are indexed by set S. we denote $[\mathbf{w}]_F$ the truncation of $\mathbf{w}$ whose elements are zero outside F. For simplicity, let $\nabla_F \ell(\mathbf{w}) \triangleq [\nabla\ell(\mathbf{w})]_F$ be the truncated gradient. $\nabla_s\ell(\mathbf{w})$ is the truncated gradient with largest s elements in absolute value. In this paper we assume that both S_* and $\bar{X}$ are unknown. We can only observe feature matrix X whose columns are randomly sampled from a clustering model with cluster center $\bar{X}$. That is,

$$X = [\bar{X}_1, \bar{X}_1, \bar{X}_2, \bar{X}_2, \bar{X}_2 \cdots \bar{X}_d] + \boldsymbol{\epsilon} \triangleq \bar{X}P + \boldsymbol{\epsilon} \,, \tag{2}$$

where P is called duplication matrix. The random noise $\boldsymbol{\epsilon}$ is subgaussian satisfying

$$\mathrm{E}\{\frac{1}{n}\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\epsilon}\} = D([\sigma_1, \sigma_2, \cdots, \sigma_d]) \,.$$

$D(\cdot)$ is the diagonal function. In the i-th cluster, there are k_i features. We denote $\sigma = \max_i \sigma_i$.

The feature cluster model defined in Eq. (2) coincides with many real world applications. For example, in computer vision, we usually need to encode low level features by clustering, which is called "feature encoding" in literature. Some features are built on other features. For example, semantic features are built from low-level features hence are correlated to low-level features. In those scenarios, for set S with size s, the submatrix X_S is usually ill-conditioned.

The conventional sparse recovery methods assume the feature matrix X to satisfy the R.I.P. [Candes and Plan, 2011] or its variants [Trzasko and Manduca, 2009]. Roughly speaking it requires X_S to be well-conditioned for $|S| \le s$. Then we can estimate $\bar{\mathbf{w}}_*$ via the following optimization problem:

$$\begin{aligned} \mathbb{P}_1 \quad & \mathbf{w}_*^{(1)} = \arg\min_{\mathbf{w}} \ell(\mathbf{w}) \triangleq \frac{1}{2n}\|X\mathbf{w} - \mathbf{y}\|^2 \\ \text{s.t.} \quad & \|\mathbf{w}\|_0 \le \sum_{i\in S_*} k_i \,. \end{aligned}$$

To recover $\bar{\mathbf{w}}_*$,

$$\bar{\mathbf{w}}_* \approx P\mathbf{w}_*^{(1)} \,.$$

However in Eq. (2), when $\boldsymbol{\epsilon} \to \mathbf{0}$, the R.I.P.-like condition clearly does not hold hence conventional methods $\mathbb{P}_1$ are no longer applicable. To address this issue, a naive approach is to cluster the feature matrix and then apply $\mathbb{P}_1$ on the clustered features. Denote $\hat{X}$ the clustered feature matrix. The naive approach can be formulated as

$$\begin{aligned} \mathbb{P}_2 \quad & \mathbf{w}_*^{(2)} = \arg\min_{\mathbf{w}} \hat{\ell}(\mathbf{w}) \triangleq \frac{1}{2n}\left\|\hat{X}^{\mathrm{T}}\mathbf{w} - \mathbf{y}\right\|^2 \\ \text{s.t.} \quad & \|\mathbf{w}\|_0 \le s \,. \end{aligned}$$

After clustering, $\hat{X}$ is expected to be R.I.P. if $\bar{X}$ is R.I.P. . A main drawback of $\mathbb{P}_2$ is that $\mathbf{w}_*^{(2)}$ is not a consistent estimator. The inconsistency of $\mathbf{w}_*^{(2)}$ comes from the uncertainty of $\hat{X}$. Recall that $\hat{X}$ is an estimation of $\bar{X}$ by matrix column clustering, from concentration inequality, it is easy to see that with high probability,

$$\left\|\hat{X}_i - \bar{X}_i\right\| \le O(\frac{\sigma_i}{\sqrt{k_i}}) \,. \tag{3}$$

Therefore, $\hat{X}$ is a biased estimation of $\bar{X}$. This bias cannot be eliminated by more training instances since the right side of Eq. (3) does not depend on n. As a consequence, $\mathbb{P}_2$ which is based on $\hat{X}$ is biased. Even if we already know the support set S_* , $\mathbf{w}_*^{(2)}$ is still biased due to the uncertainty of $\hat{X}$. This is known as the least square inconsistency when we are given the support set S_* . In short, we have the following proposition. We omit the proof since it is trivial once we realize the uncertainty in Eq. (3).

Proposition 1. *No matter how many training instances were given, $\mathbb{P}_1$ and $\mathbb{P}_2$ cannot estimate $\bar{\mathbf{w}}_*$ consistently even if we know the support set S_* and the duplication matrix P.*

Proposition 1 shows that the conventional sparse recovery methods are far from optimal because they are all inconsistent even in the ideal case. Note that their suboptimality does

not inherit from the way they solve the problem $\mathbb{P}_1$ or $\mathbb{P}_2$. We can use any algorithm to solve the above two problems. For example, directly optimize the ℓ_0 norm constraint by greedy methods [Tropp and Gilbert, 2007; Blumensath and Davies, 2008; Xiaotong Yuan, 2014], or use convex [Tibshirani, 1996] or non-convex relaxation methods [Zhang, 2012; Zhaoran Wang, 2013; Lin *et al.*, 2014a]. All these methods cannot eliminate the uncertainty in $\hat{X}$. By proposition 1, no consistent estimation is possible by these methods.

A direct corollary of proposition 1 is that it is impossible to consistently estimate $\bar{X}$ by any other method. Based on Eq. (3), even when we know the duplication matrix P, the best possible accuracy in estimating $\bar{X}_i$ is $O(\sigma_i/\sqrt{k_i})$. This upper bound is based on the concentration of mean value of k_i random variables. Because k_i is finite, it is impossible to improve this estimation unless we find a better way to estimate the mean of random variables.

In next section, we will show that by using the proposed density correction, we can design a consistent sparse estimator under model Eq. (2) even if we cannot eliminate the uncertainty in $\hat{X}$.

4 Density Corrected Sparse Recovery

From Eq. (3), it is impossible to eliminate the uncertainty in $\hat{X}$. To motivate an alternative approach, we examine the gradient of loss function in $\mathbb{P}_2$,

$$n\nabla_{\mathbf{w}}\hat{\ell}(\mathbf{w}) = \hat{X}^{\mathrm{T}}\hat{X}\mathbf{w} - \hat{X}^{\mathrm{T}}\mathbf{y} \,.$$

Denote $\hat{X} = \bar{X} + \theta$, we have

$$\begin{aligned} n\nabla_{\mathbf{w}}\hat{\ell}(\mathbf{w}) =& (\bar{X}+\theta)^{\mathrm{T}}(\bar{X}+\theta)\mathbf{w} - (\bar{X}+\theta)^{\mathrm{T}}\mathbf{y} \\ =& \bar{X}^{\mathrm{T}}\bar{X}\mathbf{w} + \underbrace{\theta^{\mathrm{T}}\theta\mathbf{w}}_{E_1} \\ &+ \underbrace{(\theta^{\mathrm{T}}\bar{X} + \bar{X}^{\mathrm{T}}\theta)\mathbf{w} - (\bar{X}+\theta)^{\mathrm{T}}\mathbf{y}}_{E_2} \,. \end{aligned}$$

Therefore, the bias in the gradient comes from E_1 and E_2. Since the perturbation $\boldsymbol{\epsilon}$ is zero mean subgaussian random variable, θ is also zero mean subgaussian. When n is large enough, E_2/n will converge to $\bar{X}^{\mathrm{T}}\mathbf{y}/n$ thus is consistent. However, E_1 will not converge to zero. This inspires us to eliminate E_1 without eliminating θ itself. In other words, although it is impossible to eliminate the bias term θ, we can still get a consistent estimator if we are able to estimate the variance of this bias term. Based on this intuition, we propose our *Density Corrected Sparse Recovery* (DCSR) in Algorithm 1.

Algorithm 1 is based on Hard Iterative Thresholding (HIT) [Blumensath and Davies, 2009]. It first clusters features into r groups. r could be potentially much larger than the sparsity s. Although we can use any clustering algorithm to cluster features, we prove that the pairwise tree clustering is sufficient to recover the cluster center $\bar{X}$ with the optimal bias $O(\sigma_i/\sqrt{k_i})$ up to constants. For the i-th cluster, its cluster density is stored in ν_i. Actually, ν_i is an unbiased estimation of the variance of perturbation term $\boldsymbol{\xi}$ in model Eq. (2). Then the algorithm does a gradient descent with step size η. The range of η is given in the next section. The key difference is that the gradient is corrected by the cluster density ν_i in Eq. (7). Eq. (7) can be reformulated as a gradient descent step for the following objective function

$$\mathbb{P}_3 \quad \min_{\mathbf{w}_{t-1}} \hat{\ell}(\mathbf{w}_{t-1}) - \frac{1}{2}\|D^{-1/2}(\boldsymbol{\nu})\mathbf{w}_{t-1}\|^2 \,. \tag{10}$$

This is similar to Elastic-Net [Zou and Hastie, 2005]. In Elastic-Net, an ℓ_2-norm regularizer is added in the sparse regression objective function to ensure a faster convergence rate. While in density correction, we subtract an ℓ_2-norm regularizer adaptively to remove the bias in the gradient.

After gradient descent, the algorithm truncates $\hat{\mathbf{w}}_t$ greedily. Note that the density correction is a principled method. We can apply density correction in various sparse regression methods such as LASSO and adaptive non-convex regularizer. We choose HIT because it is recently proven to be efficient in sparse recovery [Xiaotong Yuan, 2014]. In each iteration, the intermediate solution $\mathbf{w}_t$ is always s sparse.

Algorithm 1 Density Corrected Sparse Recovery (DCSR)

1: **Input:** X is the feature matrix . r is the number of columns of $\bar{X}$. s is the sparsity of $\bar{\mathbf{w}}_*$. T is the total iteration number. γ is the step size in gradient descent.
2: Use pairwise tree clustering to cluster columns of X into r groups. The distance between i-th and j-th column is defined by

$$d(i,j) \triangleq \|X_i - X_j\| \,, \tag{4}$$

Denote the index set of i-th cluster as C_i , $i = 1, 2, \cdots r$.
3: Denote $\hat{X}_i$ the cluster center and ν_i the density of C_i. For $|C_i| = k_i$, $\hat{X}_i$ and ν_i is computed by

$$\hat{X}_i = \frac{1}{k_i}\sum_{j\in C_i} X_j \,. \tag{5}$$

$$\nu_i = \frac{1}{nk_i(k_i-1)}\sum_{j\in C_i}\left\|X_j - \hat{X}_i\right\|^2 \,. \tag{6}$$

4: $\boldsymbol{\nu} = [\nu_1, \nu_2, \cdots, \nu_r]^{\mathrm{T}}$.
5: $\mathbf{w}_0 = \mathbf{0}$.
6: **for** $t = 1$ to T **do**
7: Gradient descent with density correction:

$$\hat{\mathbf{w}}_t = \mathbf{w}_{t-1} - \eta\nabla_{\mathbf{w}}\hat{\ell}(\mathbf{w}_{t-1}) + \eta D(\boldsymbol{\nu})\mathbf{w}_{t-1} \tag{7}$$

$$= \mathbf{w}_{t-1} - \eta\frac{1}{n}\hat{X}^{\mathrm{T}}\hat{X}\mathbf{w}_{t-1} \tag{8}$$

$$+ \eta\frac{1}{n}\hat{X}^{\mathrm{T}}\mathbf{y} + \eta D(\boldsymbol{\nu})\mathbf{w}_{t-1} \,. \tag{9}$$

8: Hard iterative sparse thresholding: let S_t be the set of indexes of elements in $\hat{\mathbf{w}}_t$ of s largest amplitude ,

$$\{\mathbf{w}_t\}_i = \begin{cases} \{\hat{\mathbf{w}}_t\}_i & i \in S_t \\ 0 & \text{otherwise} \end{cases} \,.$$

9: **end for**
10: **Output:** Cluster index C_i and sparse vector $\mathbf{w}_T$.

5 Theoretical Analysis

In this section, we give theoretical guarantees for Algorithm 1. Our analysis relies on the following matrix concentration facts [Tropp, 2012].

Lemma 1. *Let $\boldsymbol{\epsilon} \in \mathbb{R}^{n\times d}$ be a subgaussian random matrix with covariance*

$$\mathrm{E}\{\frac{1}{n}\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\epsilon}\} = D([\sigma_1^2, \sigma_2^2, \cdots, \sigma_d^2]) .$$

The eigenvalue of $\frac{1}{n}AA^{\mathrm{T}}$ is bounded by $[\mu, \beta]$. Then with probability at least $1-\delta$, denote $\hat{A} = A + \epsilon$,

$$\lambda_{\max}\{\frac{1}{n}\hat{A}^{\mathrm{T}}\hat{A}\} \leq \hat{\beta} \quad \lambda_{\min}\{\frac{1}{n}\hat{A}^{\mathrm{T}}\hat{A}\} \geq \hat{\mu} ,$$

where c is at least $1/4$,

$$\begin{aligned}\Delta \triangleq& \sqrt{\frac{(d+1)\log(2d/\delta)}{nc}}(\beta + \max_i\{\sigma_i^2\}) \\ \hat{\beta} \triangleq& (\beta + \max_i\{\sigma_i^2\}) + \Delta \\ \hat{\mu} \triangleq& (\mu - \max_i\{\sigma_i^2\}) - \Delta .\end{aligned}$$

We assume the feature cluster center matrix $\bar{X}$ to be τ-distinguishable:

Definition 1. *Feature matrix $\bar{X}$ is τ-distinguishable, if for a positive constant τ ,*

$$\min_{i\neq j}\frac{1}{n}\|\bar{X}_i - \bar{X}_j\| \geq \tau .$$

Clearly, feature clustering is possible if and only if $\bar{X}$ is τ-distinguishable for τ large enough. In the following theorem, we prove that the pairwise clustering step correctly groups features in X if $\bar{X}$ is sufficiently τ-distinguishable.

Theorem 1. *When $\bar{X}$ is at least τ-distinguishable with*

$$\begin{aligned}\tau \geq& 2\max_t\{\sigma_t^2\}(2 + 2\sqrt{\frac{\log(2k_t/\delta)}{nc}} \\ &+ \sqrt{\frac{2\log(2k_t(d-k_t)/\delta)}{nc}}) ,\end{aligned}$$

*then with probability at least $1 - 2r\delta$, $\forall i \in \{1, 2, \cdots, r\}$, the cluster center $\hat{X}_i$ is an **inconsistent** estimator of $\bar{X}_i$,*

$$\frac{1}{n}\left\|\hat{X}_i - \bar{X}_i\right\|^2 \leq \frac{1}{k_i}(\frac{1}{k_i}\sum_{t\in C_i}\sigma_t^2 + \sqrt{\frac{2\log(2/\delta)}{nc}}\max_{t\in C_i}\sigma_t^2) . \tag{11}$$

Proof. First we prove that the columns in X can be separated when noise $\boldsymbol{\epsilon}$ is not too large. For $\bar{X}_t$ and its cluster index set C_t, we require

$$\max_{i,j\in C_t}\frac{1}{n}\|X_i - X_j\|^2 \leq \min_{i\in C_t, k\notin C_t}\frac{1}{n}\|X_i - X_k\|^2 . \tag{12}$$

With probability at least $1 - 2r\delta$,

$$\begin{aligned}&\max_{i,j\in C_t}\frac{1}{n}\|X_i - X_j\|^2 = \max_{i,j\in C_t}\frac{1}{n}\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j\|^2 \\ \leq& 2\max_{t\in C_i}\{\sigma_t^2\}(1 + 2\sqrt{\frac{\log(2k_t/\delta)}{nc}}) .\end{aligned}$$

And

$$\begin{aligned}&\min_{i\in C_t, k\notin C_t}\frac{1}{n}\|X_i - X_k\|^2 \\ =& \min_{i\in C_t, k\notin C_t}\frac{1}{n}\left\|\bar{X}_t - \bar{X}_k + \boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_k\right\|^2 \\ \geq& \min_{i\in C_t, k\in C_q\neq C_t}\frac{1}{n}\left\|\bar{X}_t - \bar{X}_q\right\|^2 - \max_{i\in C_t, k\notin C_t}\frac{1}{n}\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_k\|^2 \\ \geq& \gamma - 2\max_t\{\sigma_t^2\}(1 + \sqrt{\frac{2\log(2k_t(d-k_t)/\delta)}{nc}}) .\end{aligned}$$

Therefore, Eq. (12) is satisfied when

$$\begin{aligned}\gamma \geq& 2\max_t\{\sigma_t^2\}(2 + 2\sqrt{\frac{\log(2k_t/\delta)}{nc}} + \\ &\sqrt{\frac{2\log(2k_t(d-k_t)/\delta)}{nc}}) .\end{aligned}$$

From concentration,

$$\begin{aligned}&\frac{1}{n}\left\|\hat{X}_i - \bar{X}_i\right\|^2 = \frac{1}{n}\left\|\frac{1}{k_i}\sum_{i=1}^{k_i}\epsilon_{i,:}\right\|^2 \\ \leq& \frac{1}{k_i}(\frac{1}{k_i}\sum_{t\in C_i}\sigma_t^2 + \sqrt{\frac{2\log(2/\delta)}{nc}}\max_{t\in C_i}\sigma_t^2) .\end{aligned}$$

□

Theorem 1 claims that when different features in $\bar{X}$ is dissimilar to each other, we can correctly cluster the feature matrix X. However, the estimated cluster center is not consistent. The squared bias of the cluster center estimation is on order of $O(\sigma_i^2/k_i + \sigma^2/\sqrt{n})$. Even n is infinity, this bias will never converge to zero. Theorem 1 shows that it is impossible to consistently estimate $\bar{X}$, no matter how many training instances are given.

As conventional sparse analysis, we assume that $\bar{X}$ satisfies R.I.P. condition. Note that our definition is slightly different from conventional one in order to simplify the notation.

Definition 2 (δ_s-RIP)**.** *A matrix $\bar{X}$ is δ_s-R.I.P. if there is a constant $0 < \delta_s < 1$, for any s sparse vector $\mathbf{w}$ whose support set is F,*

$$(1-\delta_s)\|\mathbf{w}\| \leq \|\frac{1}{n}\bar{X}_F{}^{\mathrm{T}}\bar{X}\mathbf{w}\| \leq (1+\delta_s)\|\mathbf{w}\| .$$

Remark 1 If $\bar{X}$ is δ_s-R.I.P. , it must be τ-distinguishable. To see this, take any two 1-sparse vector $\mathbf{e}_i, \mathbf{e}_j$ where $\mathbf{e}_i$ is the unit vector along the i-th coordinate. When $s \geq 2$, we have

$$\frac{1}{n}\|\bar{X}_i - \bar{X}_j\| = \|\frac{1}{n}\bar{X}(\mathbf{e}_i - \mathbf{e}_j)\| \geq 2\sqrt{1-\delta_s} .$$

Therefore $\bar{X}$ is at least $2\sqrt{1-\delta_s}$ distinguishable. The converse is not true. Clearly $\bar{X}_F$ can be well separated but is low rank. Here we want to emphasize that we should treat the constants δ_s and τ independently. This is because although $\bar{X}$ is at least $2\sqrt{1-\delta_s}$ distinguishable, its optimal τ may be significantly larger than $2\sqrt{1-\delta_s}$.

Following Definition 2, it is easy to check the next lemma. We omit the proof to save space.

Lemma 2. *Assume that* $\mathbf{u}$, $\mathbf{v}$ *are* s *sparse vectors and* $\bar{X}$ *is* δ_{2s}*-R.I.P. . Denote set* $F = \mathrm{supp}(\mathbf{u}) \cup \mathrm{supp}(\mathbf{v})$*, then for any positive* $\eta \leq 1/(1-\delta_{2s})$*,*

$$\|\mathbf{u} - \eta\frac{1}{n}X_F{}^{\mathrm{T}}X\mathbf{u} - (\mathbf{v} - \eta\frac{1}{n}X_F{}^{\mathrm{T}}X\mathbf{v})\| \leq \rho\|\mathbf{u}-\mathbf{v}\| \ . \quad (13)$$

where $\rho = 1 - \eta(1-\delta_{2s})$*.*

Lemma 2 claims that the gradient descent operator is restricted isometric when $\bar{X}$ is R.I.P. . Lemma 2 can be extended to a more generalized version of R.I.P. condition, as shown in [Xiaotong Yuan, 2014]. Lemma 2 is critical in the convergence analysis because when $\rho < 1$, Eq. (13) is a contraction map which indicates a geometrical convergence rate.

With the help of the above discussion, we give the convergence rate of Algorithm 1.

Theorem 2. *Under the same assumptions in Theorem 1 and* $\bar{X}$ *is* δ_{3s}*-R.I.P. . Algorithm 1 converges to* $\bar{\mathbf{w}}_*$ *geometrically,*

$$\|\mathbf{w}_t - \bar{\mathbf{w}}_*\| \leq (\|\bar{\mathbf{w}}_*\| + b)\rho^t + b \ ,$$

where

$$\begin{aligned}
\bar{\sigma} =& \max_i\{\frac{1}{k_i^2}\sum_{t\in C_i}\sigma_i^2\} \\
\Delta_{3s} =& (\|\mathbf{w}_*\|(\bar{\sigma}+1+\delta_{3s}) + \bar{\sigma})\sqrt{\frac{(3s+1)3s\log(6d/\delta)}{nc}} \\
&+ \|\mathbf{w}_*\|\sqrt{\frac{6s\log(6d/\delta)}{nc}}\bar{\sigma} \\
\rho =& 2(1-\eta(1-\delta_{3s}-\Delta_{3s})) \\
b =& \frac{2\eta}{1-\rho}(\frac{3s}{\sqrt{n}}\xi + \Delta_{3s}\|\bar{\mathbf{w}}_*\|) \ ,
\end{aligned}$$

provide that $0 \leq \rho < 1$ *.*

Theorem 2 claims that the convergence rate of Algorithm 1 is linear. The ρ controls the exponential convergence rate which depends on the step size and R.I.P. condition. Δ_{3s} is the bias term due to clustering which converges to zero at speed $O(1/\sqrt{n})$. b is the estimation error due to the noise in label $\mathbf{y}$ and the uncertainty in the clustering step. We see that when the noise level ξ is zero and n is sufficiently large, the estimation error b converges to zero, which indicates the consistency of Algorithm 1.

The novelty of Theorem 2 is its consistency. It is easy to verify that without density correction, sparse recovery methods based on $\mathbb{P}_1$ and $\mathbb{P}_2$ can only estimate $\bar{\mathbf{w}}_*$ up to accuracy $O(\sigma_i/\sqrt{k_i})$, no matter how many training instances are given. By density correction, we improve the accuracy to $O(\sigma_i/\sqrt{nk_i})$. This is significantly better than traditional methods when n is large enough. However, this power does not come free. In R.I.P. sparse recovery, the sampling complexity is $O(s\log d)$. From Theorem 2, in non-R.I.P. sparse recovery, our sampling complexity is $O(s^2\log d)$, which is slightly larger than the R.I.P. one. This extra s in the sampling complexity comes from the uncertainty of clustering step. In our proof, we prove that with a fix support set we can bound the gradient after density correction and this bound holds for any set no larger than s. This step brings us an extra s in sampling complexity. At the time when we write this paper, we are not aware of any published results showing a sharper bound.

We give a sketch of our proof of Theorem 2. First we show that in Eq. (7) the gradient descent after correction is consistent.

Lemma 3. *For any* $\mathbf{w}$ *supported on set* F*,* $|F| \leq s$*, with probability at least 1-*δ*,*

$$\|\nabla_F\hat{\ell}(\mathbf{w}) - [D(\boldsymbol{\nu})]_F\mathbf{w} - \nabla_F\bar{\ell}(\mathbf{w})\| \leq \Delta_s\|\mathbf{w}\|$$

where

$$\begin{aligned}
\Delta_s =& \|\mathbf{w}\|\sqrt{\frac{2s\log(6d/\delta)}{nc}}\max_i\{\frac{1}{k_i^2}\sum_{t\in C_i}\sigma_i^2\} + \\
&\|\mathbf{w}\|\sqrt{\frac{(s+1)s\log(6d/\delta)}{nc}}(\max_i\{\frac{1}{k_i^2}\sum_{t\in C_i}\sigma_i^2\} + 1 + \delta_s) \\
&+ \sqrt{\frac{(s+1)s\log(6d/\delta)}{nc}}\max_i\{\frac{1}{k_i^2}\sum_{t\in C_i}\sigma_i^2\} \ . \quad (14)
\end{aligned}$$

Proof. For a fix set F, similar to standard online learning analysis,

$$\begin{aligned}
&\|\nabla_F\hat{\ell}(\mathbf{w}) - [D(\boldsymbol{\nu})]_F\mathbf{w} - \nabla_F\bar{\ell}(\mathbf{w})\| \\
\leq& \underbrace{\|\frac{1}{n}(\hat{X}_F{}^{\mathrm{T}}\hat{X}_F - \bar{X}_F{}^{\mathrm{T}}\bar{X}_F) - \mathrm{E}\{D(\boldsymbol{\nu})\}}_{E_1} \\
&+ \underbrace{\mathrm{E}\{D(\boldsymbol{\nu})\} - D(\boldsymbol{\nu})}_{E_2})\|_2\|\mathbf{w}\| + \underbrace{\|\frac{1}{n}(\hat{X}_F{}^{\mathrm{T}}\mathbf{y} - \bar{X}_F{}^{\mathrm{T}}\mathbf{y})\|}_{E_3} \ .
\end{aligned}$$

From matrix concentration Lemma 1, with probability at least $1-\delta$,

$$\begin{aligned}
\|E_1\|_2 \leq& \sqrt{\frac{(s+1)\log(2s/\delta)}{nc}}(\max_i\{\frac{1}{k_i^2}\sum_{t\in C_i}\sigma_i^2\} + 1 + \delta_s) \\
\|E_2\|_2 \leq& \sqrt{\frac{2\log(2s/\delta)}{nc}}\max_i\{\frac{1}{k_i^2}\sum_{t\in C_i}\sigma_i^2\} \\
\|E_3\| \leq& \sqrt{\frac{(s+1)\log(2s/\delta)}{nc}}\max_i\{\frac{1}{k_i^2}\sum_{t\in C_i}\sigma_i^2\} \ .
\end{aligned}$$

Now we consider the all possible set $|F| \leq s$ in d dimensional space. Clearly there are no more than $\binom{d}{s} \leq (ed/s)^s$ such sets. We have

$$\log(2s(\frac{ed}{s})^s/\delta) \leq s\log(6d/\delta) \ .$$

Combining all the above together, the proof is done. □

In Lemma 3, Δ_s is the bias of the corrected gradient $\nabla_F\hat{\ell}(\mathbf{w}) + [D(\boldsymbol{\nu})]_F\mathbf{w}$. Δ_s converges to zero at speed $O(1/\sqrt{n})$. The complete proof of Theorem 1 follows immediately. Our proof is based on the proof given in [Yuan *et al.*, 2013, Theorem 2 part b]. It is worth to mention that their model cannot deal with non-R.I.P. feature matrix.

proof of Theorem 2. Denote $F = S^{(t)} \cup S^{(t-1)} \cup S_*$. So $|F| \leq 3s$. We try to bound estimation accuracy of the $3s$ sparse vector $\mathbf{u} = [\hat{\mathbf{w}}_t]_F$. We have

$$\begin{aligned}
&\|\mathbf{u} - \bar{\mathbf{w}}_*\| = \|[\hat{\mathbf{w}}_t]_F - \bar{\mathbf{w}}_*\| \\
\leq&\|\mathbf{w}_{t-1} - \eta(\nabla_F \hat{\ell}(\mathbf{w}_{t-1}) - [D(\boldsymbol{\nu})]_F \mathbf{w}_{t-1}) \\
&- (\bar{\mathbf{w}}_* - \eta \nabla_F \bar{\ell}(\bar{\mathbf{w}}_*))\| + \eta \|\nabla_F \bar{\ell}(\bar{\mathbf{w}}_*)\| \\
\leq&\|\mathbf{w}_{t-1} - \eta \nabla_F \bar{\ell}(\mathbf{w}_{t-1}) - (\bar{\mathbf{w}}_* - \eta \nabla_F \bar{\ell}(\bar{\mathbf{w}}_*))\| \\
&+ \eta \|\nabla_F \bar{\ell}(\bar{\mathbf{w}}_*)\| \\
&+ \eta \|\nabla_F \hat{\ell}(\mathbf{w}_{t-1}) - \eta [D(\boldsymbol{\nu})]_F \mathbf{w}_{t-1} - \nabla_F \bar{\ell}(\mathbf{w}_{t-1})\| \\
\leq&(1 - \eta(1 - \delta_{3s}))\|\mathbf{w}_{t-1} - \bar{\mathbf{w}}_*\| + \eta \|\nabla_F \bar{\ell}(\bar{\mathbf{w}}_*)\| \\
&+ \eta \|\nabla_F \hat{\ell}(\mathbf{w}_{t-1}) - [D(\boldsymbol{\nu})]_F \mathbf{w}_{t-1} - \nabla_F \bar{\ell}(\mathbf{w}_{t-1})\| \\
\leq&(1 - \eta(1 - \delta_{3s}))\|\mathbf{w}_{t-1} - \bar{\mathbf{w}}_*\| \\
&+ \eta \|\nabla_F \bar{\ell}(\bar{\mathbf{w}}_*)\| + \eta \Delta_{3s} \|\mathbf{w}_{t-1}\| \\
\leq&(1 - \eta(1 - \delta_{3s}) + \eta\Delta)\|\mathbf{w}_{t-1} - \bar{\mathbf{w}}_*\| \\
&+ \eta \|\nabla_{3s} \bar{\ell}(\bar{\mathbf{w}}_*)\| + \eta \Delta_{3s} \|\mathbf{w}_*\| \,.
\end{aligned}$$

Fro R.I.P. condition,

$$\begin{aligned}
&\|\mathbf{u} - \bar{\mathbf{w}}_*\| \\
\leq&(1 - \eta(1 - \delta_{3s}))\|\mathbf{w}_{t-1} - \bar{\mathbf{w}}_*\| + \eta \|\nabla_F \bar{\ell}(\bar{\mathbf{w}}_*)\| \\
&+ \eta \|\nabla_F \hat{\ell}(\mathbf{w}_{t-1}) - [D(\boldsymbol{\nu})]_F \mathbf{w}_{t-1} - \nabla_F \bar{\ell}(\mathbf{w}_{t-1})\| \\
\overset{(1)}{\leq}&(1 - \eta(1 - \delta_{3s}))\|\mathbf{w}_{t-1} - \bar{\mathbf{w}}_*\| \\
&+ \eta \|\nabla_F \bar{\ell}(\bar{\mathbf{w}}_*)\| + \eta \Delta_{3s} \|\mathbf{w}_{t-1}\| \\
\leq&(1 - \eta(1 - \delta_{3s}) + \eta\Delta)\|\mathbf{w}_{t-1} - \bar{\mathbf{w}}_*\| \\
&+ \eta \|\nabla_{3s} \bar{\ell}(\bar{\mathbf{w}}_*)\| + \eta \Delta_{3s} \|\mathbf{w}_*\| \,.
\end{aligned}$$

(1) comes from Lemma 3. Because S_t is the largest s elements in $\hat{\mathbf{w}}_t$, we have

$$\|\mathbf{w}_t - \mathbf{w}_*\| \leq 2\|\mathbf{u}_t - \mathbf{w}_*\| \,.$$

Then we get the recursive inequality about $\mathbf{w}_t$. Therefore we have

$$\|\mathbf{w}_t - \bar{\mathbf{w}}_*\| \leq (\|\bar{\mathbf{w}}_*\| + b)\rho^t + b \,,$$

where

$$\begin{aligned}
\rho =& 2(1 - \eta(1 - \delta_{3s} - \Delta_{3s})) \\
b =& \frac{2\eta}{(1 - \rho)}(\|\nabla_{3s} \bar{\ell}(\bar{\mathbf{w}}_*)\| + \Delta_{3s} \|\bar{\mathbf{w}}_*\|) \,.
\end{aligned}$$

To ensure the convergence, we require $\rho < 1$.

Since $\mathbf{y}$ is generated by model Eq. (1),

$$\|\nabla_{3s} \bar{\ell}(\bar{\mathbf{w}}_*)\| = \|[\frac{1}{n} \bar{X}^{\mathrm{T}} \boldsymbol{\xi}]_{3s}\| \leq \frac{3s}{\sqrt{n}} \xi \,.$$

□

6 Conclusion

We propose a density correction sparse recovery algorithm for non-R.I.P. problems. We first show that the conventional methods cannot consistently recover the sparse vector due to feature correlation and cluster center uncertainty. Then we propose the density correction that combines with hard iterative thresholding to recover the sparse vector consistently. The proposed algorithm has geometrical convergence rate. It adaptively removes the variance of the bias term in the clustered feature matrix.

The density correction has potentially more applications other than sparse recovery. It shows that directly learning on the clustered features may not be optimal. The cluster center is usually biased therefore the estimation is usually inconsistent. The density correction improves the estimation to be consistent without estimating the bias term itself, which could be rather difficult in many applications. Instead it exploits the cluster structure to estimate the second order statistics: the variance of the bias term. Considering the fact that clustering is widely used in machine learning, the density correction might be applied in various circumstances.

We mainly focus on least square loss function in this paper. For general convex loss function, it is much more difficult to derive the density correction. A key difficulty is that the expectation of the bias term in the gradient is no longer the cluster density. We need more sophisticated statistics to estimate the variance of this bias term. For example, recent development in convex total least square [Malioutov and Slavov, 2014] extends the total least square estimation to general convex loss functions. We open this topic for future research.

Acknowledgement

This paper was partially supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20068 and by the National Science Foundation under Grant No. IIS-1251187. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/NBC, the National Science Foundation or the U.S. Government.

References

[Agarwal *et al.*, 2012] Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Stochastic optimization and sparse statistical recovery: An optimal algorithm for high dimensions. *arXiv preprint arXiv:1207.4421*, 2012.

[Blumensath and Davies, 2008] Thomas Blumensath and Mike E Davies. Iterative thresholding for sparse approximations. *Journal of Fourier Analysis and Applications*, 14(5-6):629–654, 2008.

[Blumensath and Davies, 2009] Thomas Blumensath and Mike E Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009.

[Blumensath, 2012] Thomas Blumensath. Accelerated iterative hard thresholding. *Signal Processing*, 92(3):752–756, 2012.

[Candes and Plan, 2011] Emmanuel J Candes and Yaniv Plan. A probabilistic and RIPless theory of compressed sensing. *IEEE Transactions on Information Theory*, 57(11):7235–7254, 2011.

[Candes *et al.*, 2006] Emmanuel J Candes, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[Chen *et al.*, 2013] Yudong Chen, Constantine Caramanis, and Shie Mannor. Robust sparse regression under adversarial corruption. In *ICML*, 2013.

[Foucart, 2011] Simon Foucart. Hard thresholding pursuit: an algorithm for compressive sensing. *SIAM Journal on Numerical Analysis*, 49(6):2543–2563, 2011.

[Foucart, 2012] Simon Foucart. Sparse recovery algorithms: sufficient conditions in terms of restricted isometry constants. In *Approximation Theory XIII: San Antonio 2010*, pages 65–77. Springer, 2012.

[Gan *et al.*, 2015a] Chuang Gan, Ming Lin, Yi Yang, Yueting Zhuang, and Alexander G Hauptmann. Exploring semantic interclass relationships (SIR) for zero-shot action recognition. In *AAAI*, 2015.

[Gan *et al.*, 2015b] Chuang Gan, Naiyan Wang, Yi Yang, Dit-Yan Yeung, and Alexander G Hauptmann. Devnet: A deep event network for multimedia event detection and evidence recounting. In *CVPR*, 2015.

[Ghadimi and Lan, 2013] Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23(4):2061–2089, 2013.

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

[Ji Liu, 2013] Jieping Ye Ji Liu, Ryohei Fujimaki. Forward-backward greedy algorithms for general convex smooth functions over a cardinality constraint. *arXiv:1401.0086*, 2013.

[Jin *et al.*, 2013] Rong Jin, Tianbao Yang, and Shenghuo Zhu. A new analysis of compressive sensing by stochastic proximal gradient descent. *arXiv preprint arXiv:1304.4680*, 2013.

[Lan *et al.*, 2013] Zhen-Zhong Lan, Lu Jiang, Shoou-I Yu, Shourabh Rawat, Yang Cai, Chenqiang Gao, Shicheng Xu, Haoquan Shen, Xuanchong Li, Yipei Wang, et al. Cmu-informedia at trecvid 2013 multimedia event detection. In *TRECVID 2013 Workshop*, volume 1, page 5, 2013.

[Lehiste, 1976] Ilse Lehiste. Suprasegmental features of speech. *Contemporary issues in experimental phonetics*, 225:239, 1976.

[Lin and Xiao, 2014] Qihang Lin and Lin Xiao. An adaptive accelerated proximal gradient method and its homotopy continuation for sparse optimization. In *ICML*, 2014.

[Lin *et al.*, 2014a] Ming Lin, Rong Jin, and Changshui Zhang. Efficient sparse recovery via adaptive non-convex regularizers with oracle property. In *UAI*, 2014.

[Lin *et al.*, 2014b] Qihang Lin, Xi Chen, and Javier Pea. A sparsity preserving stochastic gradient methods for sparse regression. *Computational Optimization and Applications*, pages 1–28, 2014.

[Loh and Wainwright, 2012] Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664, 2012.

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

[Lowe, 2004] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[Malioutov and Slavov, 2014] Dmitry Malioutov and Nikolai Slavov. Convex total least squares. *arXiv:1406.0189 [cs, q-bio, stat]*, June 2014. arXiv: 1406.0189.

[Markovsky and Van Huffel, 2007] Ivan Markovsky and Sabine Van Huffel. Overview of total least-squares methods. *Signal processing*, 87(10):2283–2302, 2007.

[Shah, 2012] Rajen Shah. The lasso: Variable selection, prediction and estimation. Technical report, 2012.

[Shalev-Shwartz and Tewari, 2011] Shai Shalev-Shwartz and Ambuj Tewari. Stochastic methods for l1-regularized loss minimization. *JMLR*, 12:1865–1892, 2011.

[Tibshirani, 1996] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[Tropp and Gilbert, 2007] Joel A Tropp and Anna C Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007.

[Tropp, 2012] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

[Trzasko and Manduca, 2009] Joshua Trzasko and Armando Manduca. Relaxed conditions for sparse signal recovery with general concave priors. *TSP*, 57(11):4347–4354, 2009.

[Van De Geer and Bhlmann, 2009] Sara A Van De Geer and Peter Bhlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[Xiang *et al.*, 2013] Shuo Xiang, Xiaotong Shen, and Jieping Ye. Efficient sparse group feature selection via nonconvex optimization. In *ICML*, 2013.

[Xiaotong Yuan, 2014] Tong Zhang Xiaotong Yuan, Ping Li. Gradient hard thresholding pursuit for sparsity-constrained optimization. In *ICML*, 2014.

[Xu *et al.*, 2010] Huan Xu, Constantine Caramanis, and Shie Mannor. Robust regression and lasso. *IEEE Transactions on Information Theory*, 56(7):3561–3574, 2010.

[Yang *et al.*, 2014] Eunho Yang, Aurelie Lozano, and Pradeep Ravikumar. Elementary estimators for high-dimensional linear regression. In *ICML*, 2014.

[Yuan *et al.*, 2013] Xiao-Tong Yuan, Ping Li, and Tong Zhang. Gradient hard thresholding pursuit for sparsity-constrained optimization. *arXiv:1311.5750 [cs, stat]*, November 2013. arXiv: 1311.5750.

[Zhang, 2012] Tong Zhang. Multi-stage convex relaxation for feature selection. *Bernoulli*, 2012.

[Zhaoran Wang, 2013] Tong Zhang Zhaoran Wang, Han Liu. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *arXiv:1306.4960*, 2013.

[Zou and Hastie, 2005] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.