Construction of Incoherent Dictionaries via Direct Babel Function Minimization

Editors: Jun Zhu and Ichiro Takeuchi

Abstract

Highly incoherent dictionaries have broad applications in machine learning. Minimizing the mutual coherence is a common intuition to construct incoherent dictionaries in the previous methods. However, as pointed out by Tropp (2004), mutual coherence does not offer a very subtle description and Babel function, as a generalization of mutual coherence, is a more attractive alternative. However, it is much more challenging to optimize. In this work, we minimize the Babel function directly to construct incoherent dictionaries. As far as we know, this is the first work to optimize the Babel function. We propose an augmented Lagrange multiplier based algorithm to solve this nonconvex and nonsmooth problem with the convergence guarantee that every accumulation point is a KKT point. We define a new norm $\|\mathbf{X}\|_{\infty,max_p}$ and propose an efficient method to compute its proximal operation with $O(n^2 \log n)$ complexity, which dominates the running time of our algorithm, where max_p means the sum of the largest p elements and n is the number of the atoms. Numerical experiments testify to the advantage of our method.

Keywords: Incoherent dictionaries, Babel function, mutual coherence, optimization algorithm.

1. Introduction

Highly incoherent dictionaries are widely used with great success in compressed sensing (Candès et al., 2006; Donoho, 2006), sparse representation (Bruckstein et al., 2009) and dictionary learning (Donoho and Huo, 2001; Donoho and Elad, 2003; Gribonval and Nielsen, 2003). Specific applications include feature selection (Bajwa et al., 2010), network anomaly detection (Andrysiak and Saganowski, 2015) and incoherent subspaces learning for classification (Barchiesi and Plumbley, 2015) in machine learning, denoising (Wang et al., 2014), compression (Sezer et al., 2008) and inpainting (Elad et al., 2005) in image processing and channel estimation (Li et al., 2016) in signal processing. Other applications include the coding theory and communications (Strohmer and Heath, 2003), robust transmission (Fickus and Mixon, 2012) and quantum computing (Eldar and Forney, 2002).

A dictionary in a Hilbert space is a finite redundant collection of unit-norm vectors which spans the whole space. We use $\mathbf{D} = [\mathbf{d}_i, \cdots, \mathbf{d}_n]$ as the matrix form of the dictionary and $\mathbf{d}_i$ as an atom of $\mathbf{D}$. We say that a dictionary is incoherent when the atoms have a low dependency on each other. Incoherence plays an important role for the stable signal recovery in compressed sensing, sparse representation and dictionary learning (Tropp, 2004; Arora et al., 2014). Mutual coherence is a simple way to characterize the incoherence, defined as the maximum absolute inner product between two distinct atoms:

$$\mu(\mathbf{D}) = \max_{1 \le i,j \le n, i \ne j} \frac{|\langle \mathbf{d}_i, \mathbf{d}_j \rangle|}{\|\mathbf{d}_i\| \|\mathbf{d}_j\|}.$$

Minimizing the mutual coherence is a straightforward intuition for the construction of incoherent dictionaries (Tsiligianni et al., 2014; Rusu and Gonzálezprelcic, 2015; Lin et al., 2018). However, as mentioned in (Tropp, 2004), mutual coherence does not offer a very subtle description of incoherence since it only reflects the most extreme correlations. When most of the inner products are tiny, mutual coherence can be downright misleading. Babel function (Tropp, 2004), as a generalization of mutual coherence, can avoid this disadvantage. It measures the maximal total coherence between an atom and a collection of other atoms:

$$B(p) = \max_{\Lambda,|\Lambda|=p} \max_{i\notin\Lambda} \sum_{j\in\Lambda} \frac{|\langle \mathbf{d}_i, \mathbf{d}_j\rangle|}{\|\mathbf{d}_i\|\|\mathbf{d}_j\|}.$$

This motivates us to minimize the Babel function for the construction of incoherent dictionaries. The following example further verifies the advantage of minimizing the Babel function with some theoretical guarantee.

1.1. Compressed Sensing: An Example

Compressed sensing merges the sampling and compression by exploiting the sparsity. Consider a signal $\mathbf{x} \in \mathbb{R}^d$, which can be sparsely represented over a redundant bases $\mathbf{\Phi} \in \mathbb{R}^{d\times n}$, i.e., $\mathbf{x} = \mathbf{\Phi}\alpha$ with $\|\alpha\|_0 \ll n$. Given a sensing matrix $\mathbf{\Psi} \in \mathbb{R}^{m\times d}$ with $\|\alpha\|_0 < m \ll n$, compressed sensing suggests to represent $\mathbf{x}$ by m scalars given by $\mathbf{y} = \mathbf{\Psi}\mathbf{x}$. The original signal $\mathbf{x}$ can be recovered from $\mathbf{y}$ by exploiting its sparse representation, i.e., solving the following problem

$$\min_{\alpha} \|\alpha\|_0 \quad s.t. \quad \mathbf{y} = \mathbf{D}\alpha, \tag{1}$$

where $\mathbf{D} = \mathbf{\Psi}\mathbf{\Phi}$ is referred to as the effective dictionary.

It is known that solving problem (1) is NP-hard (Natarajan, 1995). Thus a few approximate strategies are proposed such as the Basic Pursuit (BP) (Chen et al., 1998) and Orthogonal Matching Pursuit (OMP) (Pati et al., 1993). A fundamental question is that under what conditions the solutions of these approximate strategies are identical to the solution of problem (1). Mutual coherence and Babel function can be used to characterize the conditions of successful recovery.

Theorem 1 *(Tropp, 2004) For problem (1), if a feasible point α satisfies*

$$\|\alpha\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D})}\right), \tag{2}$$

or

$$B(\|\alpha\|_0 - 1) + B(\|\alpha\|_0) < 1, \tag{3}$$

then α is the solution of problem (1) and can be obtained by OMP and BP. ■

Since $B(p)$ is a non-decreasing function on p, condition (3) is equivalent to $\|\alpha\|_0 < \max\{p+1 : B(p-1)+B(p) < 1\}$. It can be easily checked that $\frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D})}\right) \leq \max\{p+1 :$

$B(p-1)+B(p)<1\}$ via $B(p)\leq p\mu$. Thus condition (2) imposes more restrictions on the sparsity than condition (3), which verifies the superiority of Babel function to mutual coherence. Moreover, Tropp (2004) constructed an example to further explain this comment. In infinite-dimensional Hilbert space, for the i-th element of atom $\mathbf{d}_k$, let $\mathbf{d}_k(i)=\begin{cases}0, & i<k,\\ r^{i-k}\sqrt{1-r^2} & k\leq i.\end{cases}$ Then $\mu=r$ and $B(p)\leq\frac{2r}{1-r}$. If $r=0.2$, then condition (2) requires $\|\alpha\|_0<3$ while (3) holds for all $\|\alpha\|_0$.

Theorem 1 demonstrates that in order to recover a unique sparse representation, we need to construct highly incoherent $\mathbf{D}$. That is to say, small mutual coherence or Babel function. Since condition (3) provides a broader boundary of recovery guarantee than condition (2), minimizing the Babel function is superior to minimizing the mutual coherence. We follow (Elad, 2007; Tsiligianni et al., 2014; Xu et al., 2010) to consider the optimization of the sampling process, i.e., optimize $\mathbf{\Psi}$ with given $\mathbf{\Phi}$.

1.2. Previous Work

The alternating projection method (Tropp et al., 2005; Elad, 2007; Tsiligianni et al., 2014; Xu et al., 2010) is one of the most influential work on the design of incoherent dictionaries. The procedure can be described as the following iterative steps: 1. Normalize the columns of $\mathbf{\Psi\Phi}$ and compute the Gram matrix $\mathbf{G}=(\mathbf{\Psi\Phi})^T\mathbf{\Psi\Phi}$. 2. Project $\mathbf{G}$ onto set $\mathbf{S}_1=\{\mathbf{G}:|\mathbf{G}_{i,j}|\leq t, i\neq j\}$, where t is some threshold such as the Welch bound $\sqrt{\frac{n-m}{m(n-1)}}$ (Welch, 1974). 3. Project $\mathbf{G}$ onto set $\mathbf{S}_2=\{\mathbf{G}:\text{rank}(\mathbf{G})\leq m, \mathbf{G}\succeq 0\}$ and obtain $\mathbf{D}\in\mathbf{R}^{m\times n}$ where $\mathbf{G}=\mathbf{D}^T\mathbf{D}$. 4. Form the new sensing matrix $\mathbf{\Psi}$ via solving a least square problem: $\min_{\mathbf{\Psi}}\|\mathbf{D}-\mathbf{\Psi\Phi}\|_F^2$.

The alternating projection method can only make the off-diagonal values of $\mathbf{G}$ lower than the parameter t. It will get a sub-optimal solution when t is larger than the true mutual coherence. Otherwise, sets $\mathbf{S}_1$ and $\mathbf{S}_2$ have no intersections. Note that the Welch bound is not tight when $n>m(m+1)/2$. Moreover, the convergence is not proved in (Elad, 2007; Tsiligianni et al., 2014; Xu et al., 2010). The least square step makes it difficult to apply the convergence result of the standard alternating projection method (Lewis et al., 2009).

There are some other strategies besides the projection based methods. Lin et al. (2018) optimized a smoothed approximation of the mutual coherence $\|\mathbf{G}-\mathbf{I}\|_\infty$ directly and Duarte-Carvajalino and Sapiro (2009) minimized a square loss rather than the l_∞ loss of $\mathbf{G}-\mathbf{I}$ for ease of calculation in applications of image processing. Rusu (2013) and Rusu and Gonzálezprelcic (2015) solved a sequence of convex optimization problems. To simplify the algorithm, in this paper we only consider to construct data-independent dictionaries with high incoherence, rather than learning a dictionary via fitting the data. So we do not review the methods in the dictionary learning community.

1.3. Contributions

The above methods all base on the intuition of minimizing the mutual coherence. As far as we know, there is no literature focusing on minimizing the Babel function. In this paper, we directly minimize the Babel function to construct incoherent dictionaries. In summary, our contributions include:

1. We propose an augmented Lagrange multiplier based method to minimize the Babel function directly with the convergence guarantee that every accumulation point is a KKT point. To the best of our knowledge, we are the first to optimize the Babel function. Due to its property of *direct* minimization, our method can obtain higher incoherence measured by both mutual coherence and Babel function. Our method can also be used to minimize the mutual coherence directly as a special case.

2. We define a new norm $\|\mathbf{X}\|_{\infty,max_p}$, which has a strong relationship with the Babel function. We propose an efficient method to compute its proximal operation with $O(n^2 \log n)$ complexity, which is required for the minimization of the Babel function in each iteration, and thus dominates the running time of our algorithm. Besides minimizing the Babel function, we expect that this norm can also be used in other applications with a requirement of regularizing the largest p elements.

2. Minimizing the Babel Function

In this section, we discuss how to minimize the Babel function directly. We first define two norms. For a vector $\mathbf{x} \in \mathbb{R}^n$ and a matrix $\mathbf{X} \in \mathbb{R}^{n\times n}$, define

$$\|\mathbf{x}\|_{max_p} = \sum_{j=1}^{p} |\mathbf{x}_{\delta(j)}|,$$
$$\|\mathbf{X}\|_{\infty,max_p} = \max_{1\le i\le n} \sum_{j=1}^{p} |\mathbf{X}_{i,\delta(j)}|,$$

where $\mathbf{x}_{\delta(j)}$ is the j-th largest entry of $\mathbf{x}$ in absolute value and $\mathbf{X}_{i,\delta(j)}$ is the j-th largest entry of the i-th row of $\mathbf{X}$ in absolute value. It can be easily checked that $\|\mathbf{x}\|_{max_p}$ and $\|\mathbf{X}\|_{\infty,max_p}$ are norms. $\|\mathbf{X}\|_{\infty,max_p}$ could be considered to be between the $l_{\infty,1}$ norm and $l_{\infty,\infty}$ norm and can be used to regularize the largest p elements of each row of $\mathbf{X}$. From the definition of Babel function, we have

$$B(p) = \left\|\mathbf{D}^T\mathbf{D} - \mathbf{I}\right\|_{\infty,max_p},$$

with $\|\mathbf{D}_i\| = 1, i = 1, \cdots, n$. We use $\mathbf{D}_i \in \mathbf{R}^{m,1}$ as the i-th column of $\mathbf{D}$ and $\mathbf{D}_{i,:} \in \mathbf{R}^{1,n}$ as the i-th row. $\mathbf{I}$ is the identity matrix. So we can solve the following problem to minimize the Babel function directly.

$$\begin{aligned} &\min_{\mathbf{D}\in\mathbf{R}^{m\times n}} \quad \left\|\mathbf{D}^T\mathbf{D} - \mathbf{I}\right\|_{\infty,max_p}, \\ &s.t.\ \ \mathbf{D}^T \in \mathrm{Span}(\mathbf{\Phi}^T), \quad \|\mathbf{D}_i\| = 1, i = 1, \cdots, n, \end{aligned} \tag{4}$$

where $\mathbf{D}^T \in \mathrm{Span}(\mathbf{\Phi}^T)$ comes from $\mathbf{D} = \mathbf{\Psi}\mathbf{\Phi}$ for the sake of applying our method to compressed sensing. Note that we consider the problem with given $\mathbf{\Phi} \in \mathbb{R}^{d\times n}$ and unknown $\mathbf{\Psi} \in \mathbb{R}^{m\times d}$ in this paper. Let $r = \mathrm{rank}(\mathbf{\Phi})$ and $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{\Phi}$ be the compact SVD of $\mathbf{\Phi}$, then $\mathbf{D} = \mathbf{\Psi}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{R}\mathbf{V}^T$ where $\mathbf{R} \equiv \mathbf{\Psi}\mathbf{U}\mathbf{\Sigma} \in \mathbb{R}^{m\times r}$. Thus the solution of problem (4) is $\mathbf{D} = \mathbf{R}\mathbf{V}^T$, where $\mathbf{R}$ is the solution of

$$\begin{aligned} &\min_{\mathbf{R}\in\mathbf{R}^{m\times r}} \quad \left\|\mathbf{V}\mathbf{R}^T\mathbf{R}\mathbf{V}^T - \mathbf{I}\right\|_{\infty,max_p}, \\ &s.t. \qquad \mathbf{V}_{i,:}\mathbf{R}^T\mathbf{R}\mathbf{V}_{i,:}^T = 1, i = 1, \cdots, n. \end{aligned}$$

Algorithm 1 Augmented Lagrangian Multiplier method for direct Babel Function minimization (ALM-BF)

Initialize $0 < \varpi < 1$, $\gamma > 1$, $\tau > 0$, $\underline{\Lambda} < \overline{\Lambda}$, ρ^0, $\mathbf{X}^0$, $\mathbf{Y}^0$, $\mathbf{W}^0$, $\mathbf{\Lambda}_1^0$ and $\mathbf{\Lambda}_2^0$.
for $k = 0, 1, 2, \cdots$ **do**
  Step 1: Let $(\mathbf{X}^{k,0}, \mathbf{Y}^{k,0}, \mathbf{W}^{k,0}) = (\mathbf{X}^k, \mathbf{Y}^k, \mathbf{W}^k)$
  repeat
    $\mathbf{X}^{k,t+1} = \mathrm{Prox}_{\frac{1}{\rho^k}\|\cdot\|_{\infty,max_p}}\left((\rho^k\mathbf{Y}^{k,t} - \mathbf{\Lambda}_1^k + \tau\mathbf{X}^{k,t})/(\rho^k+\tau)\right)$.
    $\mathbf{Y}^{k,t+1} = \mathrm{Proj}_{\Pi}\left((\rho^k\mathbf{X}^{k,t+1} + \mathbf{\Lambda}_1^k + \rho^k\mathbf{V}\mathbf{W}^{k,t}\mathbf{V}^T - \rho^k\mathbf{I} - \mathbf{\Lambda}_2^k + \tau\mathbf{Y}^{k,t})/(2\rho^k+\tau)\right)$.
    $\mathbf{W}^{k,t+1} = \mathrm{Proj}_{\Omega}\left((\mathbf{V}^T(\rho^k\mathbf{Y}^{k,t+1} + \rho^k\mathbf{I} + \mathbf{\Lambda}_2^k)\mathbf{V} + \tau\mathbf{W}^{k,t})/(\rho^k+\tau)\right)$.
    Let $(\boldsymbol{\sigma}_1^{k,t+1}, \boldsymbol{\sigma}_2^{k,t+1}, \boldsymbol{\sigma}_3^{k,t+1}) \in \partial_{\mathbf{X},\mathbf{Y},\mathbf{W}} L(\mathbf{X}^{k,t+1}, \mathbf{Y}^{k,t+1}, \mathbf{W}^{k,t+1}, \mathbf{\Lambda}_1^k, \mathbf{\Lambda}_2^k)$.
  until $\|(\boldsymbol{\sigma}_1^{k,t+1}, \boldsymbol{\sigma}_2^{k,t+1}, \boldsymbol{\sigma}_3^{k,t+1})\|_F \le \epsilon_k$.
  Let $(\mathbf{X}^{k+1}, \mathbf{Y}^{k+1}, \mathbf{W}^{k+1}) = (\mathbf{X}^{k,t+1}, \mathbf{Y}^{k,t+1}, \mathbf{W}^{k,t+1})$.
  Step 2: $\hat{\mathbf{\Lambda}}_1^{k+1} = \mathbf{\Lambda}_1^k + \rho^k(\mathbf{X}^{k+1} - \mathbf{Y}^{k+1})$, $\mathbf{\Lambda}_1^{k+1} = \mathrm{Proj}_{[\underline{\Lambda},\overline{\Lambda}]}(\hat{\mathbf{\Lambda}}_1^{k+1})$.
    $\hat{\mathbf{\Lambda}}_2^{k+1} = \mathbf{\Lambda}_2^k + \rho^k(\mathbf{Y}^{k+1} - \mathbf{V}\mathbf{W}^{k+1}\mathbf{V}^T + \mathbf{I})$, $\mathbf{\Lambda}_2^{k+1} = \mathrm{Proj}_{[\underline{\Lambda},\overline{\Lambda}]}(\hat{\mathbf{\Lambda}}_2^{k+1})$.
  Step 3: $\rho^{k+1} = \gamma\rho^k$ if
    $\|\mathbf{X}^{k+1} - \mathbf{Y}^{k+1}\|_F > \varpi\|\mathbf{X}^k - \mathbf{Y}^k\|_F$ or $\|\mathbf{Y}^{k+1} - \mathbf{V}\mathbf{W}^{k+1}\mathbf{V}^T + \mathbf{I}\|_F > \varpi\|\mathbf{Y}^k - \mathbf{V}\mathbf{W}^k\mathbf{V}^T + \mathbf{I}\|_F$.
    else $\rho^{k+1} = \rho^k$.
end for
Let $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{\Phi} \in \mathbb{R}^{d\times n}$ be its compact SVD. Find $\mathbf{R} \in \mathbb{R}^{m\times r}$ such that $\mathbf{R}^T\mathbf{R} = \mathbf{W}^k$. Find a solution $\mathbf{\Psi}$ (must exist since $r \le d$) of $\mathbf{R} = \mathbf{\Psi}\mathbf{U}\mathbf{\Sigma}$. Output $\mathbf{\Psi} \in \mathbb{R}^{m\times d}$ and $\mathbf{D} = \mathbf{R}\mathbf{V}^T \in \mathbb{R}^{m\times n}$.

Let $\mathbf{W} = \mathbf{R}^T\mathbf{R}$ and introduce auxiliary variables $\mathbf{X}$ and $\mathbf{Y}$, we can solve the following problem

$$\begin{aligned} \min_{\mathbf{X}\in\mathbf{R}^{n\times n}, \mathbf{Y}\in\Pi, \mathbf{W}\in\Omega} \quad & f(\mathbf{X}) = \|\mathbf{X}\|_{\infty,max_p}, \\ s.t. \quad & \mathbf{X} = \mathbf{Y}, \quad \mathbf{Y} = \mathbf{V}\mathbf{W}\mathbf{V}^T - \mathbf{I}, \end{aligned} \tag{5}$$

with given $\mathbf{V} \in \mathbf{R}^{n\times r}$, where $\Pi = \{\mathbf{Y} \in \mathbf{R}^{n\times n} : \mathbf{Y}_{i,i} = 0, i = 1, \cdots, n\}$ and $\Omega = \{\mathbf{W} \in \mathbf{R}^{r\times r} : \mathbf{W} = \mathbf{W}^T, \mathbf{W} \succeq 0, \mathrm{rank}(\mathbf{W}) \le m\}$. Let δ_Π and $\delta_\Omega(\mathbf{W})$ be the indicator functions of Π and Ω, respectively. The augmented Lagrangian function is

$$\begin{aligned} L(\mathbf{X}, \mathbf{Y}, \mathbf{W}, \mathbf{\Lambda_1}, \mathbf{\Lambda_2}) = f(\mathbf{X}) + \delta_\Pi(\mathbf{Y}) + \delta_\Omega(\mathbf{W}) + \langle \mathbf{\Lambda_1}, \mathbf{X} - \mathbf{Y}\rangle + \left\langle \mathbf{\Lambda_2}, \mathbf{Y} - \mathbf{V}\mathbf{W}\mathbf{V}^T + \mathbf{I}\right\rangle \\ + \frac{\rho}{2}\|\mathbf{X} - \mathbf{Y}\|_F^2 + \frac{\rho}{2}\|\mathbf{Y} - \mathbf{V}\mathbf{W}\mathbf{V}^T + \mathbf{I}\|_F^2. \end{aligned}$$

We can solve problem (5) using the augmented Lagrange multiplier method, which is described in Algorithm 1. Proj_Ω means the projection onto set Ω and $\mathrm{Proj}_{[\underline{\Lambda},\overline{\Lambda}]}(\hat{\mathbf{\Lambda}})$ means projecting each element of $\hat{\mathbf{\Lambda}}$ such that $\underline{\Lambda} \le \hat{\mathbf{\Lambda}}_{i,j} \le \overline{\Lambda}, \forall i, j$ In Algorithm 1, the most challenging step is to find an approximate critical point of the following problem in step 1

$$\min_{\mathbf{X},\mathbf{Y},\mathbf{W}} \hat{L}(\mathbf{X}, \mathbf{Y}, \mathbf{W}) = L(\mathbf{X}, \mathbf{Y}, \mathbf{W}, \mathbf{\Lambda}_1^k, \mathbf{\Lambda}_2^k) \tag{6}$$

We can use the Proximal Alternating Minimization method (Bolte et al., 2014) to solve it, i.e., alternately solve

$$\begin{aligned}
\mathbf{X}^{k,t+1} &= \underset{\mathbf{X}}{\operatorname{argmin}}\, \hat{L}(\mathbf{X}, \mathbf{Y}^{k,t}, \mathbf{W}^{k,t}) + \frac{\tau}{2}\|\mathbf{X} - \mathbf{X}^{k,t}\|_F^2, \\
\mathbf{Y}^{k,t+1} &= \underset{\mathbf{Y}}{\operatorname{argmin}}\, \hat{L}(\mathbf{X}^{k,t+1}, \mathbf{Y}, \mathbf{W}^{k,t}) + \frac{\tau}{2}\|\mathbf{Y} - \mathbf{Y}^{k,t}\|_F^2, \\
\mathbf{W}^{k,t+1} &= \underset{\mathbf{W}}{\operatorname{argmin}}\, \hat{L}(\mathbf{X}^{k,t+1}, \mathbf{Y}^{k,t+1}, \mathbf{W}) + \frac{\tau}{2}\|\mathbf{W} - \mathbf{W}^{k,t}\|_F^2.
\end{aligned}$$

In Algorithm 1, projections onto sets Ω and Π have closed form solutions. The proximal operation of $\|\mathbf{X}\|_{\infty,max_p}$, defined as

$$\operatorname{Prox}_{\frac{1}{\rho}\|\cdot\|_{\infty,max_p}}(\mathbf{Y}) = \underset{\mathbf{X}\in\mathbf{R}^{n\times n}}{\operatorname{argmin}}\, \|\mathbf{X}\|_{\infty,max_p} + \frac{\rho}{2}\|\mathbf{X} - \mathbf{Y}\|_F^2, \tag{7}$$

can be computed exactly with $O(n^2 \log n)$ complexity, which is described in Sections 3, 4 and 5. $\hat{L}(\mathbf{X}, \mathbf{Y}, \mathbf{W})$ satisfies the KL condition (Bolte et al., 2014) and thus $\{\mathbf{X}^{k,t}, \mathbf{Y}^{k,t}, \mathbf{W}^{k,t}\}$ is guaranteed to converge to a critical point of problem (6) when $t \to \infty$. So the requirement of $\|(\boldsymbol{\sigma}_1^{k,t+1}, \boldsymbol{\sigma}_2^{k,t+1}, \boldsymbol{\sigma}_3^{k,t+1})\|_F \le \epsilon_k$ in step 1 can be satisfied with arbitrarily ϵ_k. It should be mentioned that in Step 1 of Algorithm 1 we only need $(\mathbf{X}^{k,t+1}, \mathbf{Y}^{k,t+1}, \mathbf{W}^{k,t+1})$ to be an inexact critical point of $\min_{\mathbf{X},\mathbf{Y},\mathbf{W}} L(\mathbf{X}, \mathbf{Y}, \mathbf{W}, \boldsymbol{\Lambda}_1^k, \boldsymbol{\Lambda}_2^k)$, the global minimum is not needed.

The convergence of the augmented Lagrange multiplier method is proved in (Conn et al., 1991; Andreani et al., 2008). However, the standard analysis considers the general nonlinear programming with constraints of $h_i(\mathbf{x}) = 0$ and $h_j(\mathbf{x}) \le 0$, where h_i and h_j are required to be differential. Thus it cannot be applied directly to our problem with a positive semidefinite and rank constraint. By exploiting the property of the normal cone of Ω and Π, we can have the convergence result established in Theorem 2. Some literatures use the KL condition to obtain that the sequence globally converges to a KKT point, e.g., (Wang et al., 2015). However, they require strong assumptions which are not satisfied by problem (5).

Theorem 2 *Assume that* $\{\mathbf{X}^k, \mathbf{Y}^k, \mathbf{W}^k\}$ *is bounded,* $\epsilon_k \to 0$ *and* $\mathbf{W}^*\mathbf{V}_{i,:}^T\mathbf{V}_{i,:}$, $i = 1, \cdots, n$, *are linearly independent. Let* $(\mathbf{X}^*, \mathbf{Y}^*, \mathbf{W}^*)$ *be an accumulation point of* $(\mathbf{X}^k, \mathbf{Y}^k, \mathbf{W}^k)$, *then* $(\mathbf{X}^*, \mathbf{Y}^*, \mathbf{W}^*)$ *is a KKT point of problem (5).*

The assumption of linear independence is a standard assumption in the convergence analysis of the augmented Lagrange multiplier method (Conn et al., 1991; Andreani et al., 2008). We consider the special case of $r = n$ to verify it. Then we have that $\mathbf{V}_{i,:}$ is orthogonal to $\mathbf{V}_{j,:}$ if $i \ne j$. So it is equivalent to say that there does not exist $i \in \{1, 2, \cdots, n\}$ such that $\mathbf{W}^*\mathbf{V}_{i,:}^T = \mathbf{0}$, which means that $\mathbf{W}^*\mathbf{V}^T$ does not contain columns with all 0. In the standard compressed sensing scenario with $\boldsymbol{\Phi} = \mathbf{I}$ and $\mathbf{V} = \mathbf{I}$, it is equivalent to say that $\mathbf{W}^*$ does not contain columns with all 0.

2.1. Initialize with Alternating Projection

Initialization is very important for nonconvex programming. In this subsection, we discuss how to give a suitable initializer for ALM-BF. In experiments, we find that ALM-BF is more easily to get stuck at a bad saddle point than the alternating projection method. Intuitively

Algorithm 2 Alternating Projection

Initialize $\mathbf{X}^0$, $\mathbf{W}^0$, τ, t.
for $k = 0, 1, 2, \cdots$ **do**
  $\mathbf{X}^{k+1} = \mathrm{Proj}_{\Theta}\left((\mathbf{V}\mathbf{W}^k\mathbf{V}^T - \mathbf{I} + \tau\mathbf{X}^k)/(1+\tau)\right)$,
  $\mathbf{W}^{k+1} = \mathrm{Proj}_{\Omega}\left((\mathbf{V}^T(\mathbf{X}^{k+1} + \mathbf{I})\mathbf{V} + \tau\mathbf{W}^k)/(1+\tau)\right)$.
end for
Output $\mathbf{X}^k$ and $\mathbf{W}^k$.

speaking, The alternating projection method projects all the elements of the Gram matrix below a threshold at each iteration while ALM-BF only decreases the largest few ones.

Based on this intuition, we can initialize ALM-BF via a projection based procedure. We formulate the problem as

$$\min_{\mathbf{X}\in\Theta, \mathbf{W}\in\Omega} \quad \|\mathbf{X} - \mathbf{V}\mathbf{W}\mathbf{V}^T + \mathbf{I}\|_F^2, \tag{8}$$

where $\Theta = \{\mathbf{X} \in \mathbf{R}^{n\times n} : \mathbf{X}_{i,i} = 0, -t \le \mathbf{X}_{i,j} \le t, \forall i, j = 1, \cdots, n\}$, $\Omega = \{\mathbf{W} \in \mathbf{R}^{r\times r} : \mathbf{W} = \mathbf{W}^T, \mathbf{W} \succeq 0, \mathrm{rank}(\mathbf{W}) \le m\}$ and t is an estimation of the mutual coherence. We have no prior knowledge on t except $0 \le t \le 1$. In the alternating projection method, it is not easy to tune this parameter. However, in our method, we only use it as a initialization and thus it can be set conservatively or progressively. We can use the Proximal Alternating Minimization method (Bolte et al., 2014) to solve it, which consists of two steps at each iteration:

$$\mathbf{X}^{k,t+1} = \operatorname*{argmin}_{\mathbf{X}\in\Theta} \frac{1}{2}\|\mathbf{X} - \mathbf{V}\mathbf{W}^k\mathbf{V}^T + \mathbf{I}\|_F^2 + \frac{\tau}{2}\|\mathbf{X} - \mathbf{X}^k\|_F^2,$$
$$\mathbf{W}^{k,t+1} = \operatorname*{argmin}_{\mathbf{W}\in\Omega} \frac{1}{2}\|\mathbf{W} - \mathbf{V}^T(\mathbf{X}^{k+1} + \mathbf{I})\mathbf{V}\|_F^2 + \frac{\tau}{2}\|\mathbf{W} - \mathbf{W}^k\|_F^2.$$

The difference between this projection procedure and the methods in (Elad, 2007; Tsiligianni et al., 2014; Xu et al., 2010) is that we make use of matrix $\mathbf{V}$, the singular vectors of $\mathbf{\Phi}$, and thus avoid the least square step in (Elad, 2007; Tsiligianni et al., 2014; Xu et al., 2010). This minor change ensures the convergence of this projection procedure. Applying the convergence result in (Bolte et al., 2014) directly, we can have the following convergence theorem. We describe the method in Algorithm 2.

Theorem 3 *If $\{\mathbf{X}^k\}$ and $\{\mathbf{W}^k\}$ are bounded, $0 < \tau < 1$, then $\{\mathbf{X}^k, \mathbf{W}^k\}$ generated by Algorithm 2 converges to a critical point of problem (8).*

3. Proximal Operation of $\|\mathbf{X}\|_{\infty,max_p}$

Until now, all steps in Algorithm 1 are computable except the update of $\mathbf{X}$, which requires to compute the proximal operation of $\|\mathbf{X}\|_{\infty,max_p}$, defined as (7). Theoretically, it cannot be simply computed as the shrinkage of the largest p elements of each row of $\mathbf{Y}$. For example, let $\mathbf{Y} = [2, 1.1, 0.9]$, $\rho = 1$ and $p = 2$. Then the simple shrinkage leads to $\mathbf{X} = [1, 0.1, 0.9]$ with the objective in (7) of 2.9. However, $\mathbf{X} = [1, 0.5, 0.5]$ leads to a lower objective of

2.26. From now on, we focus on how to compute it efficiently. Let $\|\mathbf{X}\|^*_{\infty,max_p}$ be the Fenchel dual norm of $\|\mathbf{X}\|_{\infty,max_p}$. The following lemma reduces the proximal operation of $\text{Prox}_{\frac{1}{\rho}\|\cdot\|_{\infty,max_p}}(\mathbf{Y})$ to the projection operation of $\text{Proj}_{\|\cdot\|^*_{\infty,max_p}\leq 1}(\rho\mathbf{Y})$.

Lemma 4 *Let* $\mathbf{W}^* = Proj_{\|\cdot\|^*_{\infty,max_p}\leq 1}(\rho\mathbf{Y})$*, then we have* $Prox_{\frac{1}{\rho}\|\cdot\|_{\infty,max_p}}(\mathbf{Y}) = \mathbf{Y} - \frac{\mathbf{W}^*}{\rho}$.

The following theorem gives an explicit expression of $\|\mathbf{X}\|^*_{\infty,max_p}$.

Theorem 5 *Let* $\|\mathbf{x}\|^*_{max_p}$ *and* $\|\mathbf{X}\|^*_{\infty,max_p}$ *be the Fenchel dual norm of* $\|\mathbf{x}\|_{max_p}$ *and* $\|\mathbf{X}\|_{\infty,max_p}$*, respectively, then*

$$\|\mathbf{x}\|^*_{max_p} = \max\left\{\|\mathbf{x}\|_\infty, \frac{1}{p}\|\mathbf{x}\|_1\right\} \equiv \|\mathbf{x}\|_{max\left\{l_\infty,\frac{1}{p}l_1\right\}},$$
$$\|\mathbf{X}\|^*_{\infty,max_p} = \sum_{i=1}^{n}\|\mathbf{X}_{i,:}\|^*_{max_p} \equiv \|\mathbf{X}\|_{1,max\left\{l_\infty,\frac{1}{p}l_1\right\}}.$$

4. Projection onto the $\|\mathbf{x}\|_{max\left\{l_\infty,\frac{1}{p}l_1\right\}}$ Ball

As will be shown in Section 5, the projection of $\mathbf{Z}$ onto the $\|\mathbf{X}\|_{1,max\left\{l_\infty,\frac{1}{p}l_1\right\}}$ ball can be solved by projecting each row of $\mathbf{Z}$ onto the $\|\mathbf{x}\|_{max\left\{l_\infty,\frac{1}{p}l_1\right\}}$ ball. Thus in this section we will give an efficient method to solve this subproblem. Formulate the problem as:

$$\min_{\mathbf{x}}\frac{1}{2}\|\mathbf{x}-\mathbf{z}\|^2, \quad s.t. \quad \|\mathbf{x}\|_{max\left\{l_\infty,\frac{1}{p}l_1\right\}} \leq t. \tag{10}$$

For simplicity, we let $\mathbf{z}_1 \geq \mathbf{z}_2 \geq \cdots \geq \mathbf{z}_n \geq 0$. Then the optimum solution must satisfy $\mathbf{x}_1 \geq \mathbf{x}_2 \geq \cdots \geq \mathbf{x}_n \geq 0$. This assumption imposes no limitation but simplifies our analysis. For the general case, we can recover the true solution by the sign and location of each element of $\mathbf{z}$. The problem can be reformulated as:

$$\min_{\mathbf{x}}\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i-\mathbf{z}_i)^2, \quad s.t. \quad \mathbf{x}_i \leq t, \forall i, \quad \frac{1}{p}\sum_{i=1}^{n}\mathbf{x}_i \leq t, \quad \mathbf{x}_i \geq 0, \forall i.$$

The Lagrangian function is

$$L(\mathbf{x},\alpha,\theta,\beta) = \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i-\mathbf{z}_i)^2 + \sum_{i=1}^{n}\langle\alpha_i,\mathbf{x}_i-t\rangle + \left\langle\theta,\sum_{i=1}^{n}\mathbf{x}_i - pt\right\rangle - \sum_{i=1}^{n}\langle\beta_i,\mathbf{x}_i\rangle.$$

By analyzing the KKT conditions, we have the following theorem to characterize the optimum solution. We use $\text{num}(\mathbf{z}_i \geq t)$ to count the number of the elements of $\mathbf{z}$ satisfying $\mathbf{z}_i \geq t$.

Theorem 6 *Let* $\{\mathbf{x},\alpha,\theta,\beta\}$ *be the KKT point,* $s = num(\mathbf{z}_i \geq t)$*, then we have*

1. *If* $\|\mathbf{z}\|_\infty \leq t$ *and* $\|\mathbf{z}\|_1 \leq pt$*, then* $\mathbf{x} = \mathbf{z}$.
2. *If* $\|\mathbf{z}\|_\infty > t$ *and* $\|\mathbf{z}\|_1 \leq pt$*, then* $\mathbf{x}_j = t$ *if* $\mathbf{z}_j > t$*;* $\mathbf{x}_j = \mathbf{z}_i$ *if* $\mathbf{z}_j \leq t$.

3. *If* $\|\mathbf{z}\|_\infty \le t$ *and* $\|\mathbf{z}\|_1 > pt$, *then* $\mathbf{x}_j = \mathbf{z}_j - \theta$ *if* $\mathbf{z}_j > \theta$; $\mathbf{x}_j = 0$ *if* $\mathbf{z}_j \le \theta$. *Moreover,* $\sum_{\mathbf{z}_j > \theta}(\mathbf{z}_j - \theta) = pt$.

4. *If* $\|\mathbf{z}\|_\infty > t$ *and* $\|\mathbf{z}\|_1 > pt$, *then* $\mathbf{x}_j = t$ *if* $\mathbf{z}_j - \theta \ge t$; $\mathbf{x}_j = \mathbf{z}_j - \theta$ *if* $0 < \mathbf{z}_j - \theta < t$; $\mathbf{x}_j = 0$ *if* $\mathbf{z}_j \le \theta$. *Specially,*

   (a) *If* $\mathbf{z}_p - \mathbf{z}_{p+1} \ge t$, *then* $\mathbf{x}_j = t, \forall j \in [1, p]$; $\mathbf{x}_j = 0, \forall j \in [p+1, n]$.

   (b) *If* $\mathbf{z}_p - \mathbf{z}_{p+1} < t$ *and* $st + \sum_{\mathbf{z}_i < t} \mathbf{z}_i \le pt$, *then* $\theta = 0$.

   (c) *If* $\mathbf{z}_p - \mathbf{z}_{p+1} < t$ *and* $st + \sum_{\mathbf{z}_i < t} \mathbf{z}_i > pt$, *then* $\theta > 0$ *and moreover,* $num(\mathbf{z}_i - \theta \ge t) \times t + \sum_{0 < \mathbf{z}_i - \theta < t}(\mathbf{z}_i - \theta) = pt$.

We can give an intuitive explanation of Theorem 6. If $\|\mathbf{z}\|_\infty \le t$ and $\|\mathbf{z}\|_1 \le pt$, then $\mathbf{z}$ is feasible and $\mathbf{x} = \mathbf{z}$. If $\|\mathbf{z}\|_\infty > t$ and $\|\mathbf{z}\|_1 \le pt$, then we only need to project $\mathbf{z}$ onto the l_∞ ball, which is a truncation operation. If $\|\mathbf{z}\|_\infty \le t$ and $\|\mathbf{z}\|_1 > pt$, then the problem reduces to the projection onto the l_1 ball, which is a shrinkage operation. If $\|\mathbf{z}\|_\infty > t$ and $\|\mathbf{z}\|_1 > pt$, we should combine the truncation and shrinkage operation.

In cases 1, 2, 4.(a) and 4.(b), $\mathbf{x}$ can be computed directly. The remaining problem is to find θ in cases 3 and 4.(c). In case 3, θ can be obtained by the method in (Duchi et al., 2008). We leave the details in the supplementary material.

To find θ in case 4.(c), our strategy is to construct a continuous, piecewise linear and decreasing function $h(\theta) = \text{num}(\mathbf{z}_i - \theta \ge t) \times t + \sum_{0 < \mathbf{z}_i - \theta < t}(\mathbf{z}_i - \theta)$ and find θ via $h(\theta) = pt$. The critical problem is to find the piecewise linear intervals and then express $h(\theta)$ explicitly in each interval. Lemma 7 gives a dynamic procedure to sequentially find these intervals. For some r and d, let $r + j = \max\{i : \mathbf{z}_i = \mathbf{z}_{r+1}\}$ and $d + k = \max\{i : \mathbf{z}_i = \mathbf{z}_{d+1}\}$. Specially, let $k^* = \max\{i : \mathbf{z}_i = \mathbf{z}_1\}$. This allows the repetition in $\mathbf{z}$.

Lemma 7 *Let* $\mathbf{z}_{n+1} = 0$ *and* $\mathbf{z}_0 = \infty$. *Define interval*

$$S(r, d) = (\max\{\mathbf{z}_{d+1}, \mathbf{z}_{r+1} - t\}, \min\{\mathbf{z}_d, \mathbf{z}_r - t\}].$$

Move left from nonempty interval $S(0, k^*) = (\max\{\mathbf{z}_{k^*+1}, \mathbf{z}_1 - t\}, \mathbf{z}_1]$ *and end when* $S(r, d)$ *reaches 0. For nonempty interval* $S(r, d)$,

1. *If* $\mathbf{z}_{d+1} < \mathbf{z}_{r+1} - t < \min\{\mathbf{z}_d, \mathbf{z}_r - t\}$, *then interval* $S(r + j, d)$ *is on the left of* $S(r, d)$ *and* $S(r + j, d)$ *is nonempty.*

2. *If* $\mathbf{z}_{r+1} - t < \mathbf{z}_{d+1} < \min\{\mathbf{z}_d, \mathbf{z}_r - t\}$, *then interval* $S(r, d + k)$ *is on the left of* $S(r, d)$ *and* $S(r, d + k)$ *is nonempty.*

3. *If* $\mathbf{z}_{r+1} - t = \mathbf{z}_{d+1} < \min\{\mathbf{z}_d, \mathbf{z}_r - t\}$, *then interval* $S(r + j, d + k)$ *is on the left of* $S(r, d)$ *and* $S(r + j, d + k)$ *is nonempty.*

The union of these disjoint intervals is $[0, \mathbf{z}_1]$.

In Lemma 7, the main intuition of defining $S(r, d)$ in such way is that if $\theta \in S(r, d)$, then $\mathbf{z}_i - \theta \ge t, \forall i \in [1, r]$ and $0 \le \mathbf{z}_i - \theta < t, \forall i \in [r + 1, d]$, which can be used to derive the expression of $h(\theta)$ in the following Lemma.

Algorithm 3 Projection onto the $\|\mathbf{x}\|_{max\left\{\infty, \frac{1}{p}l_1 \le t\right\}}$ Ball

Input $\mathbf{z}$ and t.
Let $\mathbf{z}_1 \ge \mathbf{z}_2 \cdots \ge \mathbf{z}_n$, $\mathbf{z}_{n+1} = 0$ and $s = num(\mathbf{z}_i \ge t)$.
if $\|\mathbf{z}\|_\infty \le t$ and $\|\mathbf{z}\|_1 \le pt$ **then**
  $\mathbf{x} = \mathbf{z}$.
else if $\|\mathbf{z}\|_\infty > t$ and $\|\mathbf{z}\|_1 \le pt$ **then**
  $\mathbf{x}_i = t$ for $\mathbf{z}_i > t$ and $\mathbf{x}_i = \mathbf{z}_i$ for $\mathbf{z}_i \le t$.
else if $\|\mathbf{z}\|_\infty \le t$ and $\|\mathbf{z}\|_1 > pt$ **then**
  for $d = 1, \cdots, n$ **do**
    if $\sum_{i=1}^{d+1} \mathbf{z}_i - (d+1)\mathbf{z}_{d+1} \ge pt \ge \sum_{i=1}^{d} \mathbf{z}_i - d\mathbf{z}_d$ **then**
      $\theta = \frac{\sum_{i=1}^{d} \mathbf{z}_i - pt}{d}$, $\mathbf{x}_i = \mathbf{z}_i - \theta$ for $i \in [1, d]$ and $\mathbf{x}_i = 0$ for $i \in [d+1, n]$. Terminate.
    end if
  end for
else if $\mathbf{z}_p - \mathbf{z}_{p+1} \ge t$ **then**
  $\mathbf{x}_i = t$ for $i \in [1, p]$; $\mathbf{x}_i = 0$ for $i \in [p+1, n]$.
else if $st + \sum_{\mathbf{z}_i < t} \mathbf{z}_i \le pt$ **then**
  $\mathbf{x}_i = t$ for $\mathbf{z}_i \ge t$ and $\mathbf{x}_i = \mathbf{z}_i$, for $\mathbf{z}_i < t$.
else
  for each interval $S(r, d)$ constructed in Lemma 7 **do**
    Let $a = \max\{\mathbf{z}_{d+1}, \mathbf{z}_{r+1} - t\}$, $b = \min\{\mathbf{z}_d, \mathbf{z}_r - t\}$.
    if $rt + \sum_{i=r+1}^{d} \mathbf{z}_i - (d-r)b \le pt \le rt + \sum_{i=r+1}^{d} \mathbf{z}_i - (d-r)a$ **then**
      $\theta = \frac{\sum_{i=r+1}^{d} \mathbf{z}_i + (r-p)t}{d-r}$, $\mathbf{x}_i = t$ for $i \in [1, r]$, $\mathbf{x}_i = \mathbf{z}_i - \theta$ for $i \in [r+1, d]$ and $\mathbf{x}_i = 0$ for $i \in [d+1, n]$. Terminate.
    end if
  end for
end if

Lemma 8 *In case 4.(c), let* $h(\theta) = num(\mathbf{z}_i - \theta \ge t) \times t + \sum_{0 < \mathbf{z}_i - \theta < t} (\mathbf{z}_i - \theta)$. *Consider* $S(r, d)$ *constructed in Lemma 7, then*

$$h(\theta) = rt + \sum_{i=r+1}^{d} \mathbf{z}_i - (d-r)\theta, \quad \theta \in S(r, d).$$

$h(\theta)$ *with* $\theta \in [0, \mathbf{z}_1]$ *is continuous, piecewise linear, non-increasing and there is a unique solution for* $h(\theta) = pt$.

We describe the projection method in Algorithm 3. The complexity of Algorithm 3 is $O(n\log n)$: first sort $\mathbf{z}$ with $O(n\log n)$ complexity, then go through the disjoint intervals (if needed) and obtain each $\mathbf{x}_i$ with $O(n)$ complexity, where n is the length of $\mathbf{z}$.

5. Projection onto the $\|\mathbf{X}\|_{1,max\left\{l_\infty,\frac{1}{p}l_1\right\}}$ Ball

In this section, we consider the projection of $\mathbf{Z}$ onto the $\|\mathbf{X}\|_{1,max\left\{l_\infty,\frac{1}{p}l_1\right\}}$ ball. Formulate the problem as:

$$\min_{\mathbf{X}} \frac{1}{2}\|\mathbf{Z}-\mathbf{X}\|_F^2, \quad s.t. \quad \|\mathbf{X}\|_{1,max\left\{l_\infty,\frac{1}{p}l_1\right\}} \leq T.$$

Let $\mathbf{Z}_{i,1} \geq \mathbf{Z}_{i,2} \geq \cdots \geq \mathbf{Z}_{i,n} \geq 0, \forall i$, then the optimum solution must satisfy $\mathbf{X}_{i,1} \geq \mathbf{X}_{i,2} \geq \cdots \geq \mathbf{X}_{i,n} \geq 0, \forall i$. If $\|\mathbf{Z}\|_{1,max\left\{\infty,\frac{1}{p}l_1\right\}} \leq T$, then $\mathbf{Z}$ is the optimum solution. Otherwise, the optimum solution must be on the boundary of the constraint. So we can reformulate the problem as

$$\min_{\mathbf{X},\mathbf{g}} \frac{1}{2}\sum_{i,j}(\mathbf{Z}_{i,j}-\mathbf{X}_{i,j})^2, \quad s.t. \quad \mathbf{X}_{i,j}\leq \mathbf{g}_i, \mathbf{X}_{i,j}\geq 0, \forall i,j. \quad \frac{1}{p}\sum_{j=1}^{n}\mathbf{X}_{i,j}\leq \mathbf{g}_i, \forall i. \quad \sum_{i=1}^{n}\mathbf{g}_i = T. \tag{11}$$

The Lagrangian function is:

$$L(\mathbf{X},\mathbf{g},\alpha,\theta,\beta,\lambda)=\frac{1}{2}\sum_{i,j}(\mathbf{Z}_{i,j}-\mathbf{X}_{i,j})^2+\sum_{i,j}\langle\alpha_{i,j},\mathbf{X}_{i,j}-\mathbf{g}_i\rangle+\sum_{i=1}^{n}\left\langle\theta_i,\sum_{j=1}^{n}\mathbf{X}_{i,j}-p\mathbf{g}_i\right\rangle+\left\langle\lambda,\sum_{i=1}^{n}\mathbf{g}_i-T\right\rangle-\sum_{i,j}\langle\beta_{i,j},\mathbf{X}_{i,j}\rangle\,.$$

Quattoni et al. (2009) considered the problem of projecting $\mathbf{Z}$ onto the $l_{1,\infty}$ ball by transforming it to projecting each row of $\mathbf{Z}$ onto the l_∞ ball. We borrow their idea and project each row of $\mathbf{Z}$ onto the $\|\mathbf{x}\|_{max\left\{l_\infty,\frac{1}{p}l_1\right\}}$ ball. Thus we should first find each $\mathbf{g}_i$ in problem (11), which plays the role of t in problem (10). By analyzing the KKT conditions, we can have the following Lemma, which gives us a direction to find $\mathbf{g}_i$.

Lemma 9 *At the optimum solution, either (1)* $\mathbf{g}_i > 0$ *and* $\sum_{j=1}^{p}(\mathbf{Z}_{i,j}-\mathbf{X}_{i,j}) = \lambda$*; or (2)* $\mathbf{g}_i = 0$ *and* $\sum_{j=1}^{p}\mathbf{Z}_{i,j} \leq \lambda$.

From Lemma 9 we know that for the rows of $\mathbf{Z}$ whose sum of the largest p elements is less than λ, the projection is 0. Otherwise, we should use $\sum_{j=1}^{p}(\mathbf{Z}_{i,j}-\mathbf{X}_{i,j}) = \lambda$ and $\sum_{i=1}^{n}\mathbf{g}_i = T$ to compute $\mathbf{g}_i$. Then use $\mathbf{g}_i$ to compute the projection of $\mathbf{Z}_{i,:}$. As will be proved in our supplementary material, function $g_i(\mathbf{g}_i) = \sum_{j=1}^{p}(\mathbf{Z}_{i,j}-\mathbf{X}_{i,j})$ is continuous, piecewise linear and strictly decreasing, where $\mathbf{X}_{i,:} = \mathrm{Proj}_{\|\cdot\|_{max\left\{\infty,\frac{1}{p}l_1\right\}}\leq \mathbf{g}_i}(\mathbf{Z}_{i,:})$ only depends on $\mathbf{g}_i$. Let $g_i^{-1}(\lambda)$ with $\lambda \in [0, \sum_{j=1}^{p}\mathbf{Z}_{i,j}]$ be the inverse function of $g_i(\mathbf{g}_i)$ and we can have $g_i^{-1}(\sum_{j=1}^{p}\mathbf{Z}_{i,j}) = 0$. Define

$$G_i^{-1}(\lambda) = \begin{cases} 0, & \lambda \geq \sum_{j=1}^{p}\mathbf{Z}_{i,j}, \\ g_i^{-1}(\lambda), & \lambda < \sum_{j=1}^{p}\mathbf{Z}_{i,j}, \end{cases} \quad \text{and } G^{-1}(\lambda) = \sum_{i=1}^{n}G_i^{-1}(\lambda), \quad \lambda \in \left[0, \max_i \sum_{j=1}^{p}\mathbf{Z}_{i,j}\right].$$

Then $G^{-1}(\lambda)$ is also continuous, piecewise linear and strictly decreasing. Moreover, we can have $G^{-1}(\lambda) \in \left[0, \|\mathbf{Z}\|_{1,max\left\{\infty,\frac{1}{p}l_1\right\}}\right]$. So there is a unique solution λ^* for $G^{-1}(\lambda) = T$.

Algorithm 4 Projection onto the $\|\mathbf{X}\|_{1,max\left\{\infty,\frac{1}{p}l_1\right\}}$ Ball

Input $\mathbf{Z}$, T, p.
Get the piecewise linear intervals of $G^{-1}(\lambda)$: $[\lambda_1,\lambda_2],[\lambda_2,\lambda_3],\cdots,[\lambda_{q-1},\lambda_q]$ with $\lambda_1<\cdots<\lambda_q$.
if $\|\mathbf{Z}\|_{1,max\left\{\infty,\frac{1}{p}l_1\right\}}\le T$ **then**
  $\mathbf{X}=\mathbf{Z}$.
else
  $l=1$, $r=q$.
  while 1 **do**
    if $r-l=1$ **then**
      find $\lambda^*\in[\lambda_l,\lambda_r]$ such that $G^{-1}(\lambda^*)=T$ and let $\mathbf{g}_i^*=G_i^{-1}(\lambda^*),\forall i\in[1,n]$. Break.
    end if
    $v=[(l+r)/2]$.
    if $G^{-1}(\lambda_v)=T$ **then**
      $\mathbf{g}_i^*=G_i^{-1}(\lambda_v),\forall i\in[1,n]$. Break.
    else if $G^{-1}(\lambda_v)>T$ **then**
      $l=v$.
    else
      $r=v$.
    end if
  end while
end if
$\mathbf{X}_i=\mathrm{Proj}_{\|\cdot\|_{max\left\{\infty,\frac{1}{p}l_1\right\}}\le\mathbf{g}_i^*}(\mathbf{Z}_i),\forall i\in[1,n]$.

We can find it efficiently by bisearch. Let $\mathbf{g}_i^*=G_i^{-1}(\lambda^*)$, then we can get $\mathbf{X}$ by $\mathbf{X}_{i,:}=\mathrm{Proj}_{\|\cdot\|_{max\left\{\infty,\frac{1}{p}l_1\right\}}\le\mathbf{g}_i^*}(\mathbf{Z}_{i,:})$.

We describe the method in Algorithm 4 with $\mathrm{O}(n^2\log n)$ complexity: getting the intervals of $G^{-1}(\lambda)$, finding λ^* and projecting the rows of $\mathbf{Z}$ all have a complexity of $O(n^2\log n)$, where $n\times n$ is the size of $\mathbf{Z}$. We leave some computational details of Algorithm 4 in our supplementary material.

6. Numerical Experiments

In this section, we verify the convergence of our methods in Section 6.1 and test the performance for the construction of incoherent dictionaries in Section 6.2.

6.1. Convergence

We first verify the convergence of the proposed methods: the Augmented Lagrangian Multiplier method with direct Babel Function minimization (ALM-BF) and the Alternating Projection method (APM, Algorithm 2). We take Φ to be a $d\times n$ random Gaussian matrix and test on three settings with varying sizes of Φ: (1) $d=400$, $n=500$; (2) $d=800$, $n=1000$; (3) $d=1200$, $n=1500$. We fix $m=50$ and $p=20$ in model (7). Thus the

redundancy of the effective dictionary $\mathbf{D}$, n/m, varies on the three settings. In ALM-BF we set $\gamma = 1.2$, $\varpi = 0.9$, $\underline{\Lambda} = 10^{-20}$, $\overline{\Lambda} = 10^{20}$ and $\tau = 10^{-5}$. We run the inner loop of ALM-BF for 10 iterations and 100 iterations respectively and note the method as ALM-BF-5 and ALM-BF-100. We set the threshold t as the Welch bound $\sqrt{\frac{n-m}{m(n-1)}}$ in Algorithm 2. Figure 1 plot the curves of the mutual coherence $\max_{1\leq i,j\leq n} \frac{|\langle \mathbf{d}_i, \mathbf{d}_j\rangle|}{\|\mathbf{d}_i\|_2\|\mathbf{d}_j\|_2}$, Babel function $\max_{\mathbf{\Lambda},|\mathbf{\Lambda}|=p} \max_{j\notin\mathbf{\Lambda}} \sum_{i\in\mathbf{\Lambda}} \frac{|\langle \mathbf{d}_i, \mathbf{d}_j\rangle|}{\|\mathbf{d}_i\|_2\|\mathbf{d}_j\|_2}$, constraint violations $\|\mathbf{X}-\mathbf{Y}\|_F^2$ and $\|\mathbf{Y}-\mathbf{V}\mathbf{W}\mathbf{V}^T+\mathbf{I}\|_F^2$ vs. iteration respectively for ALM-BF-10, ALM-BF-100 and APM. We run Algorithm 2 for 50 (100; 200) iterations as the initialization procedure for ALM-BF on the setting of $d = 400$, $n = 500$ ($d = 800$, $n = 1000$; $d = 1200$, $n = 1500$). We can see that both ALM-BF and APM converge well. Since ALM-BF minimizes the Babel function directly while APM only uses an approximated threshold, ALM-BF produces a solution with much lower mutual coherence and Babel function. ALM-BF-5 performs a little worse than ALM-BF-100. In applications with large size matrix $\mathbf{D}$, too many inner iterations are not affordable and we can still obtain a good solution with only a few inner iterations. We should mention that the initialization is critical for ALM-BF. Otherwise, it may get stuck at a bad saddle point or local minimum, especially when d and n are large.

6.2. Comparison on the Babel function and mutual coherence

In this section, we test the performance of ALM-BF for the construction of incoherent dictionaries. We take $\mathbf{\Phi} \in \mathbf{R}^{d\times n}$ to be a random Gaussian matrix and construct incoherent $\mathbf{D} \in \mathbf{R}^{m\times n}$ satisfying $\mathbf{D}^T \in \text{Span}(\mathbf{\Phi}^T)$. We compare ALM-BF with the Alternating Projection Method (APM, Algorithm 1 in the supplementary material. We propose it for the initialization), the method of Elad's (Elad, 2007), Duarte's (Duarte-Carvajalino and Sapiro, 2009), Xu's (Xu et al., 2010), Tsiligianni's (Tsiligianni et al., 2014), Lin's (Lin et al., 2018) and random dictionary. We also compare ALM-BF with its specialization of ALM-MC (ALM with direct Mutual Coherence minimization) by setting $p = 1$ in model (5). We do not compare with the method in (Rusu, 2013) since they do not consider the constraint $\mathbf{D}^T \in \text{Span}(\mathbf{\Phi}^T)$. We also do not compare with the learning based methods, such as the projection and rotation method (Barchiesi and Plumbley, 2013) and K-SVD (Aharon et al., 2006). However, our method can be easily extended to learn a incoherent dictionary based on the data by using the mutual coherence or Babel function as a regularization (Bao et al., 2016). In fact, the projection step in (Barchiesi and Plumbley, 2013) used the alternating projection method (Tropp et al., 2005) and the regularizer in (Bao et al., 2016) is the square loss, which is similar to (Duarte-Carvajalino and Sapiro, 2009). In Algorithm 1 we set $\gamma = 1.2$, $\varpi = 0.9$, $\rho^0 = 0.01$, $\tau = 10^{-5}$, $\underline{\Lambda} = 10^{-20}$ and $\overline{\Lambda} = 10^{20}$. We set the parameters of the compared methods following the corresponding literatures. We test on $d = 400, n = 500$, $d = 800, n = 1000$ and $d = 1200, n = 1500$ with fixed $p = 20$. We take the outer and inner iteration number as 50 and 10 for ALM-BF and ALM-MC with additional 50 (100,200) iterations of Algorithm 2 for the initialization procedure on $d = 400, n = 500$ ($d = 800, n = 1000$, $d = 1200, n = 1500$). We run all the other methods for 550 (600,700) iterations for a fair comparison. The complexity in each iteration is $\mathrm{O}(n^3)$ for all the compared methods: In ALM-BF, ALM-MC and Lin's method, the projection onto the $\{\mathbf{X} : \|\mathbf{X}\|_{1,max\left\{\infty,\frac{1}{p}l_1\right\}} \leq 1\}$ ball or l_1 ball needs $\mathrm{O}(n^2 \log n)$ complexity and the

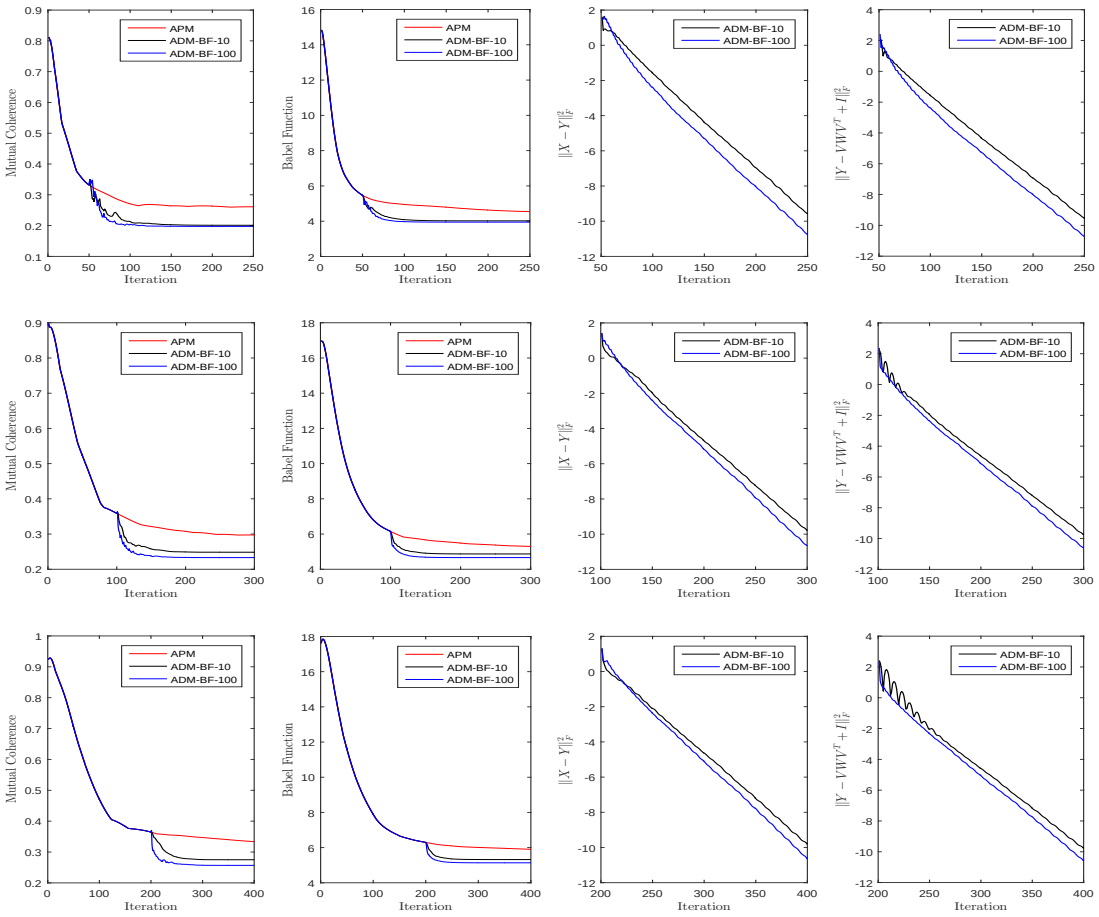
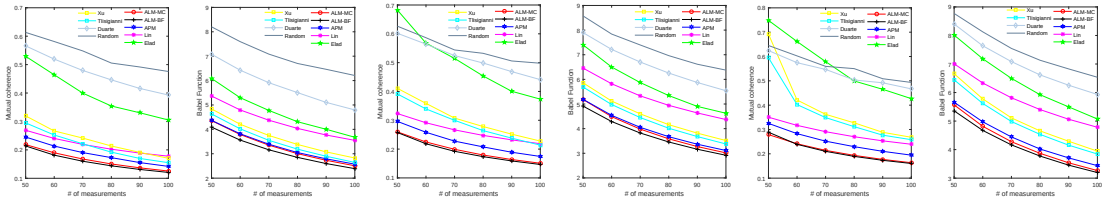


Figure 1: The mutual coherence and Babel function of ADM-BF and APM. The constraint violations of ADM-BF. Top: $d = 400$, $n = 500$. Middle: $d = 800$, $n = 1000$. Bottom: $d = 1200$, $n = 1500$



(a) Mutual Coherence (b) Babel Function (c) Mutual Coherence (d) Babel Function (e) Mutual Coherence (f) Babel Function

(a,b) $d = 400$, $n = 500$ (c,d) $d = 800$, $n = 1000$ (e,f) $d = 1200$, $n = 1500$

Figure 2: Compare ALM-BF, ALM-MC and APM with the method of Elad's, Xu's, Tlisigianni's, Duarte's, Lin's and random matrix.

matrix multiplications need $O(n^3)$ complexity. In the method of Elad's, Xu's, Tsiligianni's and Duarte's, eigenvalue decomposition and several matrix multiplications are needed.

Figure 2 shows the averaged Babel function and mutual coherence of $\mathbf{D}$ as a function of measurement m over 10 runnings. We can see that APM performs superior to the other alternating projection methods since it avoids the least square step. ALM-BF and ALM-MC obtains the lowest Babel function and mutual coherence due to their property of direct minimization. ALM-BF and ALM-MC performs similar on the characterization of mutual coherence, but ALM-BF produces smaller Babel function than ALM-MC. This verifies that minimizing the Babel function can reduce not only the top coherence but also the total coherence.

References

M. Aharon, M. Elad, and A. Bruckstein. K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.

R. Andreani, E. Birgin, J. Martínez, and M. Schuverdt. Augmented Lagrangian methods under the constant positive linear dependence constraint qualification. *Mathematical Programming*, 111(1):5–32, 2008.

T. Andrysiak and L. Saganowski. Incoherent dictionary learning for sparse representation in network anomaly detection. *Schedae Informaticae*, 24:63–71, 2015.

S. Arora, R. Ge, and A. Moitra. New algorithms for learning incoherent and overcomplete dictionaries. In *COLT*, 2014.

W. Bajwa, R. Calderbank, and S. Jafarpour. Why Gabor frames? two fundamental measures of coherence and their role in model selection. *Journal of Communications & Networks*, 12(4):289–307, 2010.

C. Bao, H. Ji, Y. Quan, and Z. Shen. Dictionary learning for sparse coding: Algorithms and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(7):1356–1369, 2016.

D. Barchiesi and M. Plumbley. Learning incoherent dictionaries for sparse approximation using iterative projections and rotations. *IEEE Transactions on Signal Processing*, 61(8):2055–2065, 2013.

D. Barchiesi and M. Plumbley. Learning incoherent subspaces: Classification via incoherent dictionary learning. *Journal of Signal Processing Systems*, 79(2):189–199, 2015.

J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014.

A. Bruckstein, D. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review*, 51:34–81, 2009.

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

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

A. Conn, N. Gould, and P. Toint. A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds. *SIAM Journal on Numerical Analysis*, 28:545–572, 1991.

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

D. Donoho and M. Elad. Optimally sparse representation in general (non-orthogonal) dictionaries via l_1-minimization. In *PNAS*, 2003.

D. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, 2001.

J. Duarte-Carvajalino and G. Sapiro. Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization. *IEEE Transactions on Image Processing*, 18(7):1395–1408, 2009.

J. Duchi, S. Shai, Y. Singer, and T. Chandra. Efficient projections onto the l_1-ball for learning in high dimensions. In *ICML*, 2008.

M. Elad. Optimized projections for compressed sensing. *IEEE Transactions on Signal Processing*, 55(12):5695–5702, 2007.

M. Elad, J. Starck, P. Querre, and D. Donoho. Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA). *Applied & Computational Harmonic Analysis*, 19(3):340–358, 2005.

Y. Eldar and G. Forney. Optimal tight frames and quantum measurement. *IEEE Transactions on Information Theory*, 48(3):599–610, 2002.

M. Fickus and D. Mixon. Numerically erasure-robust frames. *Linear Algebra and its Applications*, 437(6):1394–1407, 2012.

R. Gribonval and M. Nielsen. Sparse representations in unions of bases. *IEEE Transactions on Information Theory*, 49(12):3320–3325, 2003.

A. Lewis, D. Luke, and J. Malick. Local linear convergence for alternating and averaged nonconvex projections. *Foundations of Computational Mathematics*, 9(4):485–513, 2009.

N. Li, M. Osborna, G. Wang, and M. Sawana. A digital multichannel neural signal processing system using compressed sensing. *Digital Signal Processing*, 55:64–77, 2016.

Z. Lin, C. Lu, and H. Li. Optimized projections for compressed sensing via direct mutual coherence minimization. *Signal Processing*, 151:45–55, 2018.

B. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

Y. Pati, R. Rezaiifar, and P. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Asilomar Conference on Signals, Systems and Computers*, pages 40–44, 1993.

A. Quattoni, X. Carreras, M. Collins, and T. Darrell. An efficient projection for $l_{1,\infty}$ regularization. In *ICML*, 2009.

C. Rusu. Design of incoherent frames via convex optimization. *IEEE Signal Processing Letters*, 20(7):673–676, 2013.

C. Rusu and N. Gonzálezprelcic. Optimized compressed sensing via incoherent frames designed by convex optimization. *Mathematics*, 2015.

O. Sezer, O. Harmanci, and O. Guleryuz. Sparse orthonormal transforms for image compression. In *ICIP*, 2008.

T. Strohmer and R. Heath. Grassmannian frames with applications to coding and communication. *Applied and Computational Harmonic Analysis*, 14(3):257–275, 2003.

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

J. Tropp, I. Dhillon, R. Heath, and T. Strohmer. Designing structured tight frames via an alternating projection method. *IEEE Transactions on Information Theory*, 51(1):188–209, 2005.

E. Tsiligianni, L. Kondi, and A. Katsaggelos. Construction of incoherent unit norm tight frames with application to compressed sensing. *IEEE Transactions on Information Theory*, 60(4):2319–2330, 2014.

J. Wang, J. Cai, Y. Shi, and B. Yin. Incoherent dictionary learning for sparse representation based image denoising. In *ICIP*, 2014.

Y. Wang, W. Yin, and J. Zeng. Global convergence of ADMM in nonconvex nonsmooth optimization. *arXiv:1511.06324*, 2015.

L. Welch. Lower bounds on the maximum cross correlation of signals (corresp.). *IEEE Transactions on Information theory*, pages 397–399, 1974.

J. Xu, Y. Pi, and Z. Cao. Optimized projection matrix for compressive sensing. *EURASIP Journal on Advances in Signal Processing*, 2010:43, 2010.