Analysis Operator Learning for Overcomplete Cosparse Representations

Mehrdad Yaghoobi, Sangnam Nam, Rémi Gribonval, Mike E. Davies

ANALYSIS OPERATOR LEARNING FOR OVERCOMPLETE COSPARSE REPRESENTATIONS

Mehrdad Yaghoobi[†]*, Sangnam Nam*[‡]*, Remi Gribonval*[‡]*, and Mike E. Davies*[†]

[†] Institute for Digital Communications (IDCom), the University of Edinburgh, EH9 3JL, UK
[‡] INRIA, Centre Inria Rennes - Bretagne Atlantique, 35042 Rennes Cedex, France
yaghoobi@ieee.org, remi.gribonval@inria.fr, sangnam.nam@inria.fr, mike.davies@ed.ac.uk

ABSTRACT

We consider the problem of learning a low-dimensional signal model from a collection of training samples. The mainstream approach would be to learn an overcomplete *dictionary* to provide good approximations of the training samples using sparse synthesis coefficients. This famous sparse model has a less well known counterpart, in analysis form, called the cosparse analysis model. In this new model, signals are characterized by their parsimony in a transformed domain using an overcomplete *analysis operator*. We propose to learn an analysis operator from a training corpus using a constrained optimization program based on L1 optimization. We derive a practical learning algorithm, based on projected subgradients, and demonstrate its ability to robustly recover a ground truth analysis operator, provided the training set is of sufficient size. A local optimality condition is derived, providing preliminary theoretical support for the well-posedness of the learning problem under appropriate conditions.

1. INTRODUCTION

Sparse signal models, associated to redundant signal dictionaries, are widely used in all areas of signal processing. Traditionally, sparsity is considered using a synthesis model, where high-dimensional and complex data vectors $\mathbf{y} \in \mathbb{R}^m$ are approximated using linear combinations of few elements or *atoms* from an overcomplete collection called *dictionary*:

$$\mathbf{y} \approx \sum x_k \phi_k = \Phi \mathbf{x}, \tag{1}$$

where $\Phi \in \mathbb{R}^{m \times q}$, $m \leq q$, is the dictionary and its columns ϕ_k are the atoms. A plethora of algorithms have been derived that provide sparse representations of a given input in this synthesis approach, or more generally solve linear inverse problems of the type $\mathbf{y} = \Phi \mathbf{x}$ where $\mathbf{x}$ is sparse. One of the most celebrated approaches is based on ℓ_1 minimization

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \text{ s.t. } \mathbf{y} = \Phi \mathbf{x}. \tag{2}$$

1.1 Dictionary Learning

Many classes of natural data (audio, images) are associated to known dictionaries (Gabor, wavelets, ...) which provide somewhat accurate approximations of any signal in the class using sparse expansions. When no specific dictionary is available, techniques have been developed to learn a dictionary from a collection of training samples, $\{\mathbf{y}_i\}_{i \in \mathscr{I}}$ [1, 11, 12, 14, 15, 19, 20]. There has been much practical success in this direction, and some success guarantees have been achieved [6, 8] for the learning problem formulated in the form of the following optimization problem,

$$\min_{\mathbf{X}, \Phi \in \mathscr{D}} \|\mathbf{X}\|_1 \text{ s.t. } \mathbf{Y} = \Phi \mathbf{X}, \tag{3}$$

where $\mathbf{Y} \in \mathbb{R}^{m \times l} = [\mathbf{y}_i]_{i \in \mathscr{I}}$, $\mathbf{X} \in \mathbb{R}^{q \times l}$ and $\mathscr{D}$ are respectively the training signal matrix, the coefficient matrix and an admissible set of dictionaries.

1.2 Cosparsity and Cosparse Analysis Model

However, the famous sparse synthesis model has a less well-known counterpart, in analysis form, which has received much less attention in terms of algorithms, and virtually no attention at all in terms of model learning. The main contributions of the current paper are to explore this new terrain.

In this new model, called the cosparse analysis model [13], signals are characterized by their parsimony in a transformed domain, using a given overcomplete *analysis operator* $\Omega \in \mathbb{R}^{n \times m}$:

$$\mathbf{z} = \Omega \mathbf{y} \tag{4}$$

where $\mathbf{z}$ has a minimal representation and Ω is the analysis operator. In this setting, the concept of sparsity in (4) is slightly different to the standard definition of sparsity in (1), as the number of zero elements in $\mathbf{z}$, $p = n - \|\mathbf{z}\|_0$ has a more important role in analyzing the model (4). It has thus been named *cosparsity* [3, 13], to prevent any possible confusions. In the context of linear inverse problems, it has been shown that the cosparse analysis model can lead to uniqueness guarantees that mimic that of the sparse model [3, 13], and new cosparse recovery algorithms have been designed [13] .

1.3 Analysis Operator Learning

When a set of samples $\mathbf{Y} = [\mathbf{y}_i]_{i \in \mathscr{I}}$, is given, a question is how can we choose a suitable Ω, which provides the highest cosparsity for $\mathbf{Y}$? This is the central problem considered in this paper. Specifically, our main contributions are

- a constrained optimization program for Analysis Operator Learning (AOL) based on ℓ_1 minimization, in section 2;
- a practical learning algorithm, based on projected subgradient, in section 3;
- empirical results suggesting that the approach works reasonably well, even when the algorithm is initialized with a random analysis operator, provided the size of the training set is sufficient, in section 4;
- a local optimality condition for testing the optimality of an analysis operator with respect to the proposed program, in sections 5 and 6.

2. CONSTRAINED ANALYSIS OPERATOR LEARNING

The standard approach for many similar model adaptation problems, is to define a relevant optimization problem such that its optimal solution promotes maximal sparsity of $\mathbf{Z} := \Omega \mathbf{Y}$. A convex sparsity promoting penalty $f(\Omega)$ is the sum of absolute values of $\mathbf{Z}$, *i.e.* $f(\Omega) = \|\Omega \mathbf{Y}\|_1$. Unconstrained minimization of $f(\Omega)$ has

This work is supported by EU FP7, FET-Open grant number 225913 and EPSRC grant EP/F039697/1. MED acknowledges support of his position from the Scottish Funding Council and their support of the Joint Research Institute with the Heriot-Watt University as a component part of the Edinburgh Research Partnership.

some trivial solutions. A solution for such a minimization problem is $\Omega = \mathbf{0}$! A suggestion to exclude such trivial solutions is to restrict the solution set to an admissible set $\mathscr{C}$. We here start by investigating the problem of using a closed admissible set. AOL can now be formulated as,

$$\min_{\Omega} \|\Omega \mathbf{Y}\|_1 \text{ s.t. } \Omega \in \mathscr{C} \tag{5}$$

We here initially propose some constraints for the problem (5) and explain why some of them can not individually exclude uninteresting solutions. We finally propose a combined constraint $\mathscr{C}$, which is the Uniform Normalized Tight Frame (UNTF). The proposed constraint is smooth and differentiable on its boundary.

2.1 Row norm constraints are insufficient

The first constraint is on the norms of rows of Ω, *i.e.* $\|\omega_i\| = c$ for the i^{th} row. By applying this row norm constraint on Ω, we find the best $\omega^* \in \mathbb{R}^m$, *i.e.* minimizer of $\|\omega^T \mathbf{Y}\|_1$, which repeats to generate Ω, *i.e.* $\Omega_1^* := [\omega_i = \omega]_{i\in[1,m]}^T$, and thus $\text{Rank}\{\Omega_1^*\} = 1$. Such a solution is not interesting to us as we are looking for a full rank overcomplete operator. If we want to use this constraint individually and still get a reasonable solution, we need to change the objective of (5) to consider the interrelation of the rows of Ω [17], which is out of the scope of this paper.

2.2 Row norm + full rank constraints are insufficient

A full rank fixed row norm constraint $\mathscr{C}_F$ on Ω provides solutions with very small condition numbers. To illustrate this, denote $\mathscr{P}_{\mathscr{C}_F}$ the orthogonal projection onto $\mathscr{C}_F$, and consider $\mathbf{A}$ and ε respectively a random Gaussian matrix and a very small constant. The projection $\mathscr{P}_{\mathscr{C}_F}\{\varepsilon\mathbf{A} + \Omega_1^*\}$ has a low objective value in (5). We thus need a geometrical constraint on Ω to not allow ω_i's get arbitrary close to each other, *i.e.* $\left|\langle \omega_i, \omega_j \rangle\right| \approx \|\omega_i\|.\|\omega_j\|$.

2.3 Tight frame constraints are insufficient

In a complete setting $m = n$, an orthonormality constraint can resolve the ill-conditioning of the problem. The rows of Ω are geometrically as separated as possible. Letting $n > m$, the orthonormality constraint is not further applicable. An alternative is the orthonormality constraint in the ambient space, $\forall i \neq j,\ \omega^i \perp \omega^j$ and $\|\omega^i\|_2 = \|\omega^j\|_2 = 1$, where ω^i and $\omega^j \in \mathbb{R}^n$ are respectively the i^{th} and j^{th} columns of Ω. The admissible set of this constraint is the set of tight frames in $\mathbb{R}^{n\times m}$, *i.e.* $\Omega^T\Omega = \mathbf{I}$, where $\mathbf{I}$ is the identity operator in $\mathbb{R}^m$. The admissible set $\mathscr{C} = \{\Omega \in \mathbb{R}^{n\times m} : \Omega^T\Omega = \mathbf{I}\}$ is smooth and differentiable on its boundary.

Although the tight frame constraint could have seemed appropriate to avoid "trivial" solutions to (5), preliminary empirical and theoretical investigations indicate that when the training set is full rank, the analysis operator minimizing (5) using this constraint is always an orthonormal basis completed by zero columns [4].

2.4 Proposed constraint: Uniform Normalized Tight Frame

This motivates us to apply an extra constraint. Here, we choose to combine the unit row norm and the tight frame constraints, yielding the UNTF constraint set. UNTF frames have many interesting properties, but to the authors' knowledge, there is no analytical method to find the projection of a point onto this set. Many attempts have been done to find such an (approximate) projection, see for example [18] for an alternating projection based method.

2.5 Algorithms ?

Such a constraint $\mathscr{C}$ is not convex, hence the optimization problem (5) is non-convex. In section 3 we propose a projected subgradient solvers, which is likely to find a local minimum. When the ground truth analysis operator Ω_0 used to generate the training corpus $\mathbf{Y}$ is a local minimum of the optimization problem, we expect to identify it using the projected subgradient algorithm, provided we start from a point close enough to Ω_0. However, this can only happen if there are not "too many" spurious local minima. The experimental results, in section 4 will show that this seems to be the case, and the underlying analysis operator is reliably recovered even when the algorithm is started far from it, when the size of training set $\mathscr{I}$ is large enough.

Algorithm 1 Projected Subgradient Method for Analysis Operator Learning

initialization: $k = 1$, K_{max}, $\Omega^{[0]} = \mathbf{0}$, $\Omega^{[1]} = \Omega_{in}$, γ, $\varepsilon \ll 1$
while $\varepsilon \leq \|\Omega^{[k]} - \Omega^{[k-1]}\|_F$ and $k \leq K_{max}$ **do**
  $\Omega_G = \partial f(\Omega^{[k]})$
  $\Omega^{[k+1]} = \mathscr{P}_{UN}\left\{\mathscr{P}_{TF}\left\{\Omega^{[k]} - \gamma\Omega_G\right\}\right\}$
  $k = k + 1$
end while
output: $\Omega_{out} = \Omega^{[k-1]}$

3. PROJECTED SUBGRADIENT ALGORITHM FOR AOL

Subgradient methods have often been used to minimize convex objectives, when the solution is sought only with a few significant figures. These methods are generally slow to find exact solutions, as they converge linearly. In the AOL, we also need to solve (5) to find a solution with a reasonable precision. As the problem is constrained, we use the projected subgradient method. The subgradient of the objective is simply $\partial f(\Omega) = \mathbf{Y}\overline{\text{sgn}}(\Omega^T\mathbf{Y})^T$, where $\overline{\text{sgn}}$ is the extended sign function defined as follows,

$$\{\overline{\text{sgn}}(\mathbf{A})\}_{ij} = \overline{\text{sgn}}(\mathbf{A}_{ij})$$
$$\overline{\text{sgn}}(\mathbf{a}) = \begin{cases} 1 & \mathbf{a} > 0, \\ [-1,1] & \mathbf{a} = 0, \\ -1 & \mathbf{a} < 0. \end{cases} \tag{6}$$

In the projected subgradient methods, we have to choose a value in the set of subgradients. We randomly choose a value in $[-1, 1]$, when the corresponding element is zero.

Projection of an operator, with non-zero rows, onto the space of fixed row norm frames is easy and can be done by normalizing each row to have c norm, we use $\mathscr{P}_{UN}$ to denote this projection. If a row is zero, any normalized vector has the same distance to the zero vector, and we thus choose a normalized random vector. The projection can be found by,

$$\mathscr{P}_{UN}\{\Omega\} = [\mathscr{P}_{UN}\{\omega_i\}]_i, \tag{7}$$
$$\mathscr{P}_{UN}\{\omega\} := \begin{cases} \frac{\omega}{\|\omega\|_2} & \|\omega\|_2 \neq 0 \\ \nu & \text{otherwise,} \end{cases}$$

where ν is a random vector on the unit sphere.

Projection of a full rank matrix onto the tight frame manifold is also easy and can be done using a singular value decomposition of the linear operator [18]. Let $\mathbf{A} \in \mathbb{R}^{n\times m}$ be the given point and $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T$ be a singular value decomposition of $\mathbf{A}$ and $\mathbf{I}_{n\times m}$ be a diagonal matrix with identity on the main diagonal. The projection of $\mathbf{A}$ can be found using,

$$\mathscr{P}_{TF}(\mathbf{A}) = \mathbf{U}\mathbf{I}_{n\times m}\mathbf{V}^T. \tag{8}$$

As mentioned in section 1, a point on the intersection of the uniformly normalized set and the set of tight-frames can often be found by alternatingly projecting onto these sets. Note that, there is no guarantee for convergence to an UNTF using this method, but this technique practically works very well [18]. As the projected subgradient continuously changes the current point, which needs to be projected onto the UNTF's, we only use a single pair of projections at each iteration of the algorithm. In practice we found that the

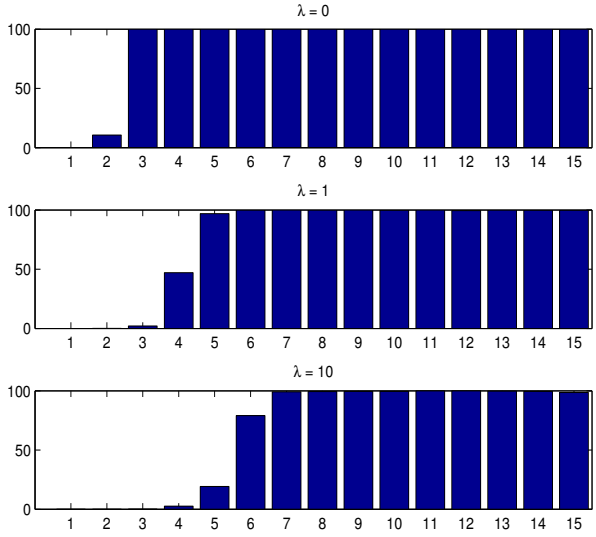

Figure 1: The average percentage of operator recovery for different λ's, where λ controls how far is the starting point Ω_{in} form Ω_0. The x-axis presents the cosparsity of the synthetic data.

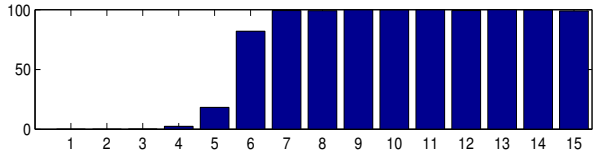

Figure 2: The average percentage of operator recovery with the random starting point. The x-axis presents the cosparsity of the signals.

solutions converge to UNTF's. A pseudocode of this algorithm is presented in Algorithm 1. As the projection onto the admissible set is not exact, although we have not observed instability by choosing a small gradient step size, convergence of this algorithm needs to be investigated in the future work.

4. EMPIRICAL EVIDENCE

In this section we present some simulation results to empirically show the local optimality of synthetic generative operators. We also empirically demonstrate the convergence of the proposed algorithm and the global recovery of the operator, when the size of the training set is large.

A random operator $\Omega_{0^-} \in \mathbb{R}^{24\times 16}$ was generated using *i.i.d.* zero mean, unit variance normal random variables[1]. The generative analysis operator Ω_0 is made by alternatingly projecting Ω_{0^-} onto the sets of UN's and TF's. A set of training samples was generated, with different cosparsities, by randomly selecting a normal vector in the orthogonal complement space of a randomly selected p rows of Ω_0. Such a vector $\mathbf{y}_i$ has (at least) p zero components in $\Omega_0\mathbf{y}_i$, and it thus is p cosparse. To initialize the proposed algorithm, we used a linear model to generate the initial Ω by combining the generative operator Ω_0 and a normalized random matrix $\mathbf{N}$, *i.e.* $\Omega_{in} = \Omega_0 + \lambda\,\mathbf{N}$, and then alternatingly projecting onto UN and TF. It is clear that when λ is zero, we actually initialize Ω with the generative model Ω_0 and when $\lambda \to \infty$, the initial Ω_{in} will be random.

[1] Ω_{0^-} is not necessarily a UNTF and needs to be projected onto the set of UNTF's.

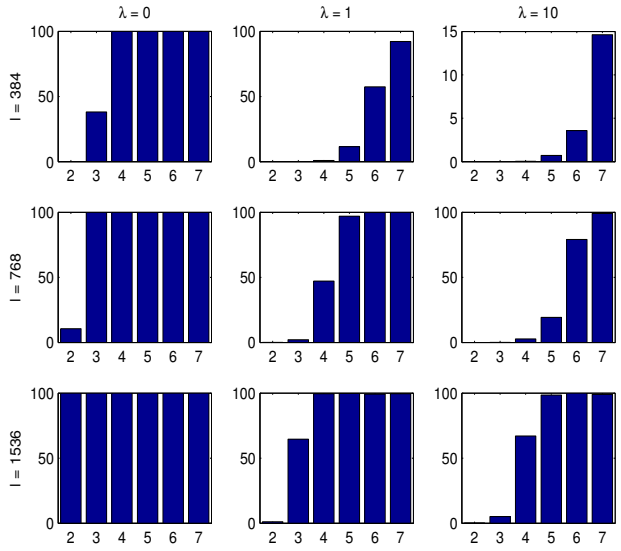

Figure 3: The average percentage of operator recovery with different training set population size l. The x-axis presents the cosparsity of the signals.

In the first experiment, we chose a set of size $l = 768$ of such training corpus. The projected subgradient method was iterated 50000 times. To check the local optimality of the operator and the size of basin of attraction, we chose $\lambda = 0, 1,$ and 10. The average percentage of operator (rows) recovery, *i.e.* the maximum ℓ_2 distance of any recovered row and the closest row of the generative operator, is not more than $\sqrt{.001}$, for different cosparsity and 100 trials, are plotted in Figure 1. We practically observe that the operator is the local optimum even when the cosparsity of the signal is as low as 3. We also see that the average recovery reduces by starting from a point far from the the actual generative operator. When λ becomes large, we do not further degrade the average recovery as the initial point is close to random, projected onto the admissible set. We also repeated the simulations with a (pseudo-) random admissible Ω_{in}. The result is shown in Figure 2. The average recovery is very similar to the previous experiment, when λ is large, which confirms our expectation.

We now investigate the role of l on the average operator recovery by some simulations. We kept using previous experiment settings and repeated the simulations for two new training sets, with populations of $l = 384$ and 1536, which are $1/2$ and 2 times of the population in the previous experiments. We show the average operator recovery for $\lambda = 0, 1, 10$ in Figure 3. The simulation results show not only that Ω_0 can be locally identified with even less cosparse signals, smaller p, but the basin of attraction is also extended and now the generative operator can be recovered by starting from a distant initial point, even using 2 cosparse signals.

The promising results in robust recovery of the generating operator encourage us to investigate the proposed framework by characterizing the local optima of (5) in a general setting. This helps us to check, if a given Ω_0 is a local minimum, and thus locally identifiable using a gradient based solver. We briefly introduce such a qualification for a local optimality in the next section. It is derived using the convexity of the objective and the smoothness of the boundary of $\mathscr{C}$. The local optimality, for the UNTF admissible set, is also investigated as a special case and a straightforward qualification is derived. The result has a flavor similar to those obtained in the context of dictionary learning for the synthesis model [8], as well as in hybrid synthesis/analysis framework [10].

5. LOCAL OPTIMALITY OF AN OPERATOR

The set of matrices in $\mathbb{R}^{n\times m}$ equipped with the Hilbert-Schmidt inner-product, defined by $\langle \mathbf{A},\mathbf{B}\rangle := \mathrm{tr}\{\mathbf{A}^T\mathbf{B}\}$, is an inner-product space. The ℓ_1 function can now be reformulated as $\|\mathbf{A}\|_1 = \langle \mathbf{A}, \mathrm{sgn}(\mathbf{A})\rangle$, where $\mathrm{sgn}(\cdot)$ is the element-wise sign operator.

The set of zeros of $\mathbf{Z}_{n\times l} = \Omega\mathbf{Y}$, has an important role in variational analysis of the ℓ_1 objective. Let Λ be the index set of non-zero elements of $\mathbf{Z}$ and $\overline{\Lambda}$ be Λ's complement. Note that the i^{th} column of $\mathbf{Z}$, $\mathbf{z}_i$, is found by $\Omega\mathbf{y}_i$. $\mathbf{y}_i$ is cosparse, when $\mathbf{z}_i$ has some zero elements. The cosparsity of the matrix $\mathbf{Y}$ can similarly be found by $|\overline{\Lambda}|$. We define $\mathbf{X}_\Lambda$ as the matrix which has the components of $\mathbf{X}$ on the support Λ, and the rest is zero.

***Theorem* 1 (Optimality Condition)** *Let $\mathscr{C}$ be a compact set and $\mathscr{T}_{\mathscr{C}}(\Omega)$ be the tangent cone at a boundary point Ω. If $\mathbf{Z} = \Omega\mathbf{Y}$, the necessary and sufficient condition for a boundary point Ω being a local optimum of (5) is to satisfy the following inequality for any non-zero $\Theta \in \mathscr{T}_{\mathscr{C}}(\Omega)$,*

$$|\langle \Theta\mathbf{Y}, \mathrm{sgn}(\mathbf{Z})\rangle| < \left\|(\Theta\mathbf{Y})_{\overline{\Lambda}}\right\|_1, \tag{9}$$

where $\overline{\Lambda}$ is the index set of zeros of $\mathbf{Z}$.

A complete proof of this theorem will be shown in [4]. We here present a sketch of the proof. The objective of (5) is convex and $\mathscr{C}$ is smooth on boundary. Using the variational analysis, a boundary point Ω with this setting is a local optimum, if the directional derivative of the objective in Ω is positive for any (non-zero) direction in the tangent space of $\mathscr{C}$ at Ω, see [16, Theorem 13.24 and Example 13.25].

This theorem can be seen as an extension of Theorem 1 in [10], where it is derived in a vector space, under unit norm constraint. The new theorem is more general, and we only need to check (9) for any non-zero vectors in $\mathscr{T}_{\mathscr{C}}(\Omega)$. Unfortunately this is not an easy task in general. In the next section we reformulate (5), with a UNTF admissible set. The tangent space of $\mathscr{C}$ can be analytically derived. We thus give an explicit qualification for the optimality.

6. UNIFORM NORMALIZED TIGHT FRAME LEARNING

The set of UNTF's in $\mathbb{R}^{n\times m}$, $m < n$, was defined in the introduction as $\mathscr{C} = \{\forall \Omega \in \mathbb{R}^{n\times m} : \Omega^T\Omega = \mathbf{I},\ \forall i\ \|\omega_i\|_2 = c\}$, where $c = \sqrt{\frac{m}{n}}$ and ω_i is the i^{th} row of Ω. Note that $\mathscr{C}$ is now a manifold and any admissible point is thus on the boundary of $\mathscr{C}$. For a given training $\mathbf{Y}_{m\times l}$, $l \gg m$, the operator learning problem (5) can now be reformulated as,

$$\begin{aligned} \min_{\Omega} \|\Omega\mathbf{Y}\|_1 \ \text{s.t.} \quad & \Omega^T\Omega = \mathbf{I} \\ & \forall i\ \|\omega_i\|_2 = c. \end{aligned} \tag{10}$$

To check the optimality of an admissible point Ω_0, we can slightly deviate Ω_0 in a direction Δ in the tangent space of $\mathscr{C}$ at Ω_0, *i.e.* $\Omega = \Omega_0 + \Delta$. The tangent space of $\mathscr{C}$ is found by letting the directional derivative[2] of $h(\Omega) = \Omega^T\Omega - \mathbf{I}$ and $h(\omega_i) = \|\omega_i\| - c$, for each i, be zero,

$$\partial h(\Omega)(\Delta) = \Delta^T\Omega + \Omega^T\Delta = 0, \tag{11}$$

and

$$\partial h(\omega_i)(\delta_i) = 2\omega_i\delta_i^T = 0, \tag{12}$$

which is simply $\langle \omega_i, \delta_i\rangle = 0$. We can now rewrite (10) using the new variable Δ,

$$\begin{aligned} \min_{\Delta} \|(\Omega_0+\Delta)\mathbf{Y}\|_1 \ \text{s.t.} \quad & \Delta^T\Omega_0 + \Omega_0^T\Delta = 0 \\ & \forall i\ \langle \omega_{0i}, \delta_i\rangle = 0, \end{aligned} \tag{13}$$

where δ^i is the i^{th} row of Δ. (13) is a convex problem with linear constraints. If the new problem has a single solution $\Delta = \mathbf{0}$, Ω_0 would then be a local minimum for (10). To find a qualification that guarantees $\mathbf{0}$ to be the only solution of (13), we reparametrize the problem. Let $\mathbf{Z}_0 := \Omega_0\mathbf{Y}$ and $\Delta_Z := \mathbf{Z} - \mathbf{Z}_0 = \Delta\mathbf{Y}$. As $\mathbf{Y}$ is full rank, we can define $\Theta := \mathbf{Y}^T(\mathbf{Y}\mathbf{Y}^T)^{-1}$ and find Δ, respectively δ_i, using,

$$\Delta = \Delta_Z\Theta \quad , \quad \delta^i = \delta_Z^i\Theta.$$

To reformulate (13) and use a higher dimension problem based on Δ_Z, we should consider the fact that each row of $\mathbf{Z}$, $\mathbf{z}^i$, can only live in the subspace spanned by rows of $\mathbf{Y}$. This means that $\mathbf{z}^i$ does not have any component in the null space of $\mathbf{Y}$. We can now introduce an extra constraint to consider this fact by $\Delta_Z(\mathbf{I} - \Theta\mathbf{Y}) = \mathbf{0}$. If we define $\mathbf{P} := \mathbf{I} - \Theta\mathbf{Y}$, the problem can be reformulated as,

$$\begin{aligned} \min_{\Delta_Z} \|\mathbf{Z}_0 + \Delta_Z\|_1 \ \text{s.t.} \quad & \Theta^T\Delta_Z^T\mathbf{Z}_0\Theta + \Theta^T\mathbf{Z}_0^T\Delta_Z\Theta = 0 \\ & \forall i\ \langle \mathbf{Z}_0^i\Theta, \delta_Z^i\Theta\rangle = 0 \\ & \Delta_Z\mathbf{P} = 0. \end{aligned} \tag{14}$$

Let a vector to matrix operator "vect$\{\cdot\}$" be defined such that,

$$\begin{aligned} &\{\mathrm{vect}\{\mathbf{Z}\}\}_k = \mathbf{Z}_{ij}, \quad i = k \mod l \\ &1 \le k \le nl \qquad\qquad j = \lfloor \frac{k-1}{l} \rfloor + 1, \end{aligned}$$

and the corresponding inverse operator be "mat$\{\cdot\}$". If $\eta := \mathrm{vect}\{\Delta_Z\}$ and $\mathbf{z}_0 := \mathrm{vect}\{\mathbf{Z}_0\}$, the constraint of (14) is *linear* and can be presented as $\Phi\eta = \mathbf{0}$. Appendix A explains how to derive Φ. The last reformulation is to rewrite (14) in a vector form,

$$\min_{\eta} \|\mathbf{z}_0 + \eta\|_1 \ \text{s.t.}\ \Phi\eta = 0. \tag{15}$$

Note that the solution of this problem can be mapped to the matrix form using "mat" operator. (15) is the formulation which has been used to show the optimality of a vector $\mathbf{z}_0$ for the problem (2), *i.e.* an admissible $\mathbf{z_0}$ is the optimal solution of (2) iff $\eta = \mathbf{0}$ is the only solution of (15). Such a qualification has been derived as follows:

$$\|\eta_\lambda\|_1 < \left\|\eta_{\overline{\lambda}}\right\|_1, \qquad \forall \eta \neq \mathbf{0} \in \mathscr{N}_\Phi, \tag{16}$$

is the necessary and sufficient condition for the optimality of $\mathbf{z}_0$, where $\lambda := \sup\{\mathbf{z}_0\}$ and $\mathscr{N}_\Phi$ is the null-space of Φ [5, 7].

***Remark* 1** *The extra constraint $\Delta_Z\mathbf{P} = \mathbf{0}$ in (14) can be interpreted as the orthogonality of each row of Δ_Z to the null-space of $\mathbf{Y}$. Therefore the size of this constraint is actually $n \times (l-m)$, which is smaller than $n \times l$, the size that is related to a full rank $\mathbf{P}$.*

***Remark* 2** *The derived local optimality condition is a deterministic qualification. How to check this qualification, is another important question which we do not answer here in this paper. The authors admit that such a test is not easy, but it might be possible to deliver a probabilistic statement using the new technique recently introduced in the matrix completion context [9].*

***Remark* 3** *For any $1 < p < m$, we can generate an (at least) p cosparse signal with a given cosparsity pattern. It means that, we can choose a signal in the orthogonal complement of the space spanned by the rows of Ω corresponding to the p cosparsity pattern. Note that if (16) is valid for any $|\overline{\lambda}| = p$, it is valid for any larger index sets including $\overline{\lambda}$. This helps us not to worry about checking (16) for the situations that by choosing a vector in the orthogonal complement space of the selected p columns of Ω, other elements also vanish.*

[2] For the definition of directional derivative, see for example [2].

7. CONCLUSIONS

This paper presented a new framework for low-dimensional signal model adaptation. The linear model, which is here called the analysis operator, can be used to sparsify a classes of signals. The new framework helps to apply various constraints to the operators.

A simple, but efficient, algorithm based on the projected subgradient technique, was also presented to recover such operators. The algorithm relies on the projection onto the constraint set. We used the algorithm to first practically show the local optimality of the operator for the proposed optimization problem, which it shows identifiability of the operator using gradient based methods. We then tested the algorithm to recover the synthetic random operators in another experiment, when the algorithm is fed with a point in the neighborhood of the true analysis operator. We demonstrated that the operator is usually recovered when the training corpus is large enough. Even when we do not know a neighborhood of the generative operator, we practically showed that there is still a good chance to recover the operator, when the training signals are enough cosparse.

In the second half of the paper, the local identifiability of such an operator is investigated and a necessary and sufficient qualification was presented. An example of such constraints, *i.e.* UNTF, which was empirically shown to be a reasonable constraint for this problem, was investigated in more detail and a more sensible qualification for the identifiability of operators, was derived. The qualification is deterministic but difficult to check. Checking such a qualification is left for future work.

APPENDIX

A. DERIVING A LINEAR REPRESENTATION OF THE CONSTRAINTS

Let the subscript $\cdot_{ij}$ be the element located in the i^{th} row and j^{th} column of the operand. To show linearity of the constraint, we only need to represent each constraint as a weighted sum of $\delta_{\Z ij}$. Although the correct notation is based on the vector subindexing of vector η, we keep using the matrix form of subindexing as it is easier to understand for the readers. A mapping between two types of subindexing is easy, as $\eta = \mathrm{vect}\{\Delta_\Z\}$. (14) has three constraints, where the last two are clearly linear, and they can be represented as,

$$\begin{aligned} \sum_k \delta_{\Z ik}\, \mathbf{P}_{kj} = 0, \quad \forall i \in [1,n],\ j \in [1,l] \\ \sum_k \delta_{\Z ik}\, \mathbf{Q}_{ki} = 0, \quad \forall i \in [1,n], \end{aligned} \tag{17}$$

where $\mathbf{Q} := \Theta\Theta^T\mathbf{Z}_0^T$. Let $\mathbf{A} := \mathbf{Z}_0\Theta$. The first constraint can now be reformulated as, $\Theta^T\Delta_\Z^T\mathbf{A} + \mathbf{A}^T\Delta_\Z\Theta = \mathbf{0}$. To derive this equation in a similar form to (17), we reformulate $\Delta_\Z\Theta$, then left-multiply the result with $\mathbf{A}^T$.

$$\{\Theta^T\Delta_\Z^T\}_{ij} = \sum_q^{1\le q\le l} \delta_{\Z jq}\,\Theta_{qi} \tag{18}$$

$$\begin{aligned} \{(\Theta^T\Delta_\Z^T)\,\mathbf{A}\}_{ij} &= \sum_k^{1\le k\le n} \left(\sum_q^{1\le q\le l} \delta_{\Z kq}\,\Theta_{qi}\right)\mathbf{A}_{kj} \\ &= \sum_k^{1\le k\le n}\ \sum_q^{1\le q\le l} \Theta_{qi}\,\mathbf{A}_{kj}\,\delta_{\Z kq} \end{aligned} \tag{19}$$

We now reformulate the first constraint as,

$$\sum_k^{1\le k\le n}\ \sum_q^{1\le q\le l} \left(\Theta_{qi}\,\mathbf{A}_{kj} + \Theta_{qj}\,\mathbf{A}_{ki}\right)\delta_{\Z kq} = 0\,,\ \forall i,\ j \in [1,m] \tag{20}$$

We can generate $\Phi_{(nl+m^2+n)\times nl}$ using the weight of (17) and (20), corresponding to the vector η, and make the linear presentation as $\Phi\,\eta = \mathbf{0}$.

REFERENCES

[1] M. Aharon, E. Elad, and A.M. Bruckstein. K-SVD: an algorithm for designining of overcomplete dictionaries for sparse representation. *IEEE Trans. on Signal Processing*, 54(11):4311–4322, 2006.

[2] J. Dattorro. *Convex Optimization and Euclidean Distance Geometry*. Meboo Publishing, 2009. (v2009.06.18), Palo Alto, CA.

[3] M.E. Davies, M. Elad, R. Gribonval, and S. Nam. Sparse analysis models and algorithms. in preparation.

[4] M.E. Davies, R. Gribonval, S. Nam, and M. Yaghoobi. A theorem on analysis operator learning. in preparation.

[5] M. Elad and A.M. Bruckstein. A generalized uncertainty principle and sparse representation in pairs of bases. *IEEE Trans. on Information Theory*, 48(9):2558–2567, 2002.

[6] Q. Geng, H. Wang, and J. Wright. On the local correctness of l1 minimization for dictionary learning. Technical Report abs/1101.5672, CoRR, 2011.

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

[8] R. Gribonval and K. Schnass. Dictionary identification - sparse matrix-factorisation via ℓ_1 minimisation. *IEEE Trans. on Information Theory*, 56(7):3523–3539, July 2010.

[9] D. Gross. Recovering low-rank matrices from few coefficients in any basis. submitted, arXiv:0910.1879, 2009.

[10] F. Jaillet, R. Gribonval, M.D. Plumbley, and H. Zayyani. An L1 criterion for dictionary learning by subspace identification. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010.

[11] M.S. Lewicki and T.J. Sejnowski. Learning overcomplete representations. *Neural Comp*, 12(2):337–365, 2000.

[12] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:19–60, 2010.

[13] S. Nam, M.E. Davies, M. Elad, and R. Gribonval. Cosparse analysis modeling- uniqueness and algorithms. In *IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

[14] B.A. Olshausen and D.J. Field. Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Research*, 37(23):3311–3325, 1997.

[15] M.D. Plumbley. Dictionary learning for L1-exact sparse coding. In *International Conference on Independent Component Analysis and Signal Separation, ICA*, 2007.

[16] R.T. Rockafellar and R.J.-B Wets. *Variational Analysis*, volume 317. Grundlehren der Math. Wissenschaften, Springer, Berlin, 1997.

[17] R. Rubinstein. Training analysis operators. SMALL project, internal technical report, December 2009.

[18] J.A. Tropp, I.S. Dhillon, R.W. Heath Jr., and T. Strohmer. Designing structural tight frames via an alternating projection method. *IEEE Trans. on Information Theory*, 51(1):188–209, 2005.

[19] M. Yaghoobi, T. Blumensath, and M. Davies. Dictionary learning for sparse approximations with the majorization method. *IEEE Trans. on Signal Processing*, 57(6):2178–2191, 2009.

[20] M. Yaghoobi and M. Davies. Dictionary learning for sparse representations: A pareto curve root finding approach. In *Lecture notes in computer science, LVA/ICA*, pages 410–417. Springer-Verlag, 2010.