Prior-Less Compressible Structure from Motion

Chen Kong Simon Lucey
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA, US
{chenk, slucey}@cs.cmu.edu

Abstract

Many non-rigid 3D structures are not modelled well through a low-rank subspace assumption. This is problematic when it comes to their reconstruction through Structure from Motion (SfM). We argue in this paper that a more expressive and general assumption can be made around compressible 3D structures. The vision community, however, has hitherto struggled to formulate effective strategies for recovering such structures after projection without the aid of additional priors (e.g. temporal ordering, rigid substructures, etc.). In this paper we present a "prior-less" approach to solve compressible SfM. Specifically, we demonstrate how the problem of SfM - assuming compressible 3D structures - can be theoretically characterized as a block sparse dictionary learning problem. We validate our approach experimentally by demonstrating reconstructions of 3D structures that are intractable using current state-of-the-art low-rank SfM approaches.

1. Introduction

A 3D shape can often be well-approximated as a linear combination of just a few elements from a set of basis or dictionary. When this approximation is exact we say that the 3D shape is sparse. In reality, few real-world 3D shapes are truly sparse, rather they are "compressible" [7], meaning they can be well-approximated by a sparse 3D shape. A scenario often entertained in S*f*M literature [5, 8] is that the same few K dictionary bases can be employed for approximating a set of 3D shape instances (see Figure 1(a)) - these fixed set of basis vectors form the canonical low-rank 3D assumption. We shall refer herein to the finite set of 3D shape instances as the 3D structure.

In compressible S*f*M we make a similar assumption to classical low-rank S*f*M [5, 8] where we assume each 3D shape instance can be described using only K dictionary bases, but a different set of K basis vectors can be employed for each shape instance (see Figure 1(b)). These set of 3D shape instances do not form a single linear subspace, they

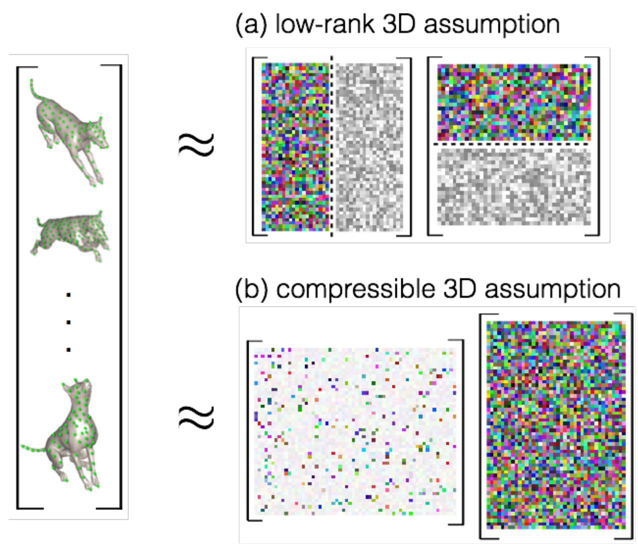


Figure 1: In this paper we explore the problem of "compressible" structure from motion (S*f*M). In this approach we assume a set of 3D shapes, stemming from a non-rigid 3D structure, can be well approximated by a few (*i.e.* K) examples of elements from an unknown basis or dictionary. Classical low-rank S*f*M makes a similar assumption but assumes that the *same* K elements within the dictionary will be used to approximate all 3D shape instances - see (a). Our approach differs in this regard, where we allow for the employment of *different* K elements within the dictionary for each 3D instance - see (b). We describe a novel algorithm, based on block sparse dictionary learning, for obtaining S*f*M reconstructions that were previously deemed intractable using the low-rank assumption employed by current state-of-the-art methods [8].

can instead be thought of as existing in a union of $\binom{L}{K}$ subspaces where L is the total number of basis vectors available. An obvious advantage of this compressible 3D structure assumption is the ability to model a much broader set of 3D structures. A drawback to the assumption, however, is discovering which of the potentially very large number of $\binom{L}{K}$ subspaces best describes the actual 3D shape instance - solely from its 2D projection. It is this dilemma which is at the heart of our paper.

Contributions: In this paper we make the following contributions:

- We demonstrate that a compressible 3D structure under weak perspective projection is 2×3 block-compressible. Based on this insight, we re-interpret compressible S*f*M as a block sparse dictionary learning problem.
- We theoretically characterize the *uniqueness* of block sparse dictionary learning. Further, we show how the uniqueness of block sparse dictionary learning can be utilized to efficiently recover the camera motion and 3D structures. We also propose the employment of dictionary coherence as a measure of reconstructibility of the 3D structures without ground truth.
- Finally, we show empirically the utility of our approach for reconstructing compressible 3D structures using a block-sparse adaptation of the K-SVD algorithm [18]. Impressive reconstruction results are reported on both synthetic and real-world compressible 3D structures.

2. Background

Weak perspective S*f*M deals with the problem of factorizing an image measurement matrix $\boldsymbol{W}$ as the product of camera motion (projection) matrix $\boldsymbol{M}$ and a shape $\boldsymbol{S}$, such that,

$$\boldsymbol{W} = \boldsymbol{M}\boldsymbol{S} \tag{1}$$

where $\boldsymbol{S}$ is the 3D structure consisting of P points deforming over F frames, resulting in a $3F \times P$ concatenated matrix of points[5].

Throughout this paper we consider weak perspective cameras, which is a reasonable assumption for objects whose variation in depth is small compared to their distance from the camera. We further assume that the measurement matrix is already centered, so the camera matrix reduces to a $2F \times 3F$ block diagonal matrix whose blocks $\boldsymbol{M}_1 \dots \boldsymbol{M}_F$ are each 2×3 matrices. The weak perspective camera assumption implies an orthonormal constraint such that $\boldsymbol{M}_f\boldsymbol{M}_f^T = \sigma^2\boldsymbol{I}_2$. A common assumption [5, 23] is to constrain the reconstructed 3D set to have fixed rank K such that $\boldsymbol{S}^\sharp = \boldsymbol{C}^\sharp\boldsymbol{B}^\sharp$ where $\boldsymbol{B}^\sharp \in \mathbb{R}^{K\times 3P}$, $\boldsymbol{C}^\sharp \in \mathbb{R}^{F\times K}$ and $\boldsymbol{S}^\sharp$ is a $F \times 3P$ reshape of $\boldsymbol{S}$ such that

$$\boldsymbol{W} = \boldsymbol{M}(\boldsymbol{C}^\sharp \otimes \boldsymbol{I}_3)\boldsymbol{B} = \boldsymbol{\Pi}\boldsymbol{B} \tag{2}$$

where $\boldsymbol{I}_3$ is a 3×3 identity matrix, $\boldsymbol{B}$ is the $3K \times P$ reshape of the matrix $\boldsymbol{B}^\sharp$ and $\boldsymbol{\Pi} = \boldsymbol{M}(\boldsymbol{C}^\sharp \otimes \boldsymbol{I}_3)$.

Rigid S*f*M: For a rigid 3D structure,

$$\boldsymbol{S}^\sharp = \begin{bmatrix} \boldsymbol{s}_x^T, & \boldsymbol{s}_y^T, & \boldsymbol{s}_z^T \\ \vdots & \vdots & \vdots \\ \boldsymbol{s}_x^T, & \boldsymbol{s}_y^T, & \boldsymbol{s}_z^T \end{bmatrix}$$
$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{s}_x, \boldsymbol{s}_y, \boldsymbol{s}_z & \dots & \boldsymbol{s}_x, \boldsymbol{s}_y, \boldsymbol{s}_z \end{bmatrix}^T \tag{3}$$

it is clear that the rank of $\boldsymbol{S}^\sharp$ *must* be one ($K = 1$) where $\boldsymbol{s}_x$, $\boldsymbol{s}_y$ and $\boldsymbol{s}_z$ are the P dimensional components of the $x-$, $y-$ and $z-$ coordinates of the rigid 3D structure. From Equation 3 this implies that $\boldsymbol{S}$ must have a rank of less than or equal to three due to the reshaping operation on $\boldsymbol{S}^\sharp$. This insight was used to great effect through the seminal work of Tomasi & Kanade [19] who demonstrated that one can compute the decomposition $\boldsymbol{W} = \hat{\boldsymbol{\Pi}}\hat{\boldsymbol{B}}$ via an SVD by preserving the first 3 modes of variation. Tomasi & Kanade also noted that the decomposition is non-unique, such that any nonsingular $\boldsymbol{G}$ can be inserted to form a valid factorization $\boldsymbol{W} = \hat{\boldsymbol{\Pi}}\hat{\boldsymbol{B}} = \hat{\boldsymbol{\Pi}}\boldsymbol{G}\boldsymbol{G}^{-1}\hat{\boldsymbol{B}} = \boldsymbol{\Pi}\boldsymbol{B}$. The matrix $\boldsymbol{G}$ is referred to in literature as the corrective transformation [23].

Low-Rank S*f*M: Bregler *et al.* [5] extended the work of Tomasi & Kanade by assuming that $\boldsymbol{S}^\sharp$ must be of fixed rank $K > 1$ for non-rigid structure. Numerous innovations have followed, most of them centered around introducing additional "priors" to make the non-rigid S*f*M problem less ambiguous. Notable examples of additional priors include: basis [23], temporal [3, 20, 25], articulation [15, 22], and camera motion [13] constraints. These priors, although useful for making the low-rank S*f*M problem tractable, considerably limit its applicability to scenarios where these constraints do not hold.

3. Related Work

"Prior-less" S*f*M: Of particular interest in this paper is the recent work of Dai *et al.* [8], who asked the question: what are the minimal set of constraints/priors required to find a unique solution to the problem of low-rank S*f*M? The authors proposed an approach to whose only prior was to assume that the non-rigid 3D structure could be represented by a linear subspace of known rank K. In this work Dai *et al.* proposed a strategy for estimating the corrective transformation matrix $\boldsymbol{G}$ whereby both the camera motion $\boldsymbol{M}$ and the 3D structure $\boldsymbol{S}$ can be obtained. This approach offered a practical breakthrough to the problem of low-rank S*f*M, which had previously been touted [23] as being theoretically impossible to solve without additional prior/constraints. In this paper we want to ask a similar question: what are the minimal set of constraints/priors required to find a unique solution to the problem of compressible S*f*M?

Manifold S*f*M: Notable efforts have been previously undertaken in literature to replace the low-rank linear subspace assumption with a manifold learning [16, 13]. Most notable is the recent work of Gotardo and Martinez [13] who demonstrated how the "kernel trick" could be employed to model 3D shape as a non-linear subspace. A drawback to this approach, however, was its reliance on additional basis constraints limiting the approach's applicability to the prior-less S*f*M problems of interest within this paper.

There is some overlap between our proposed method here, as it has been demonstrated [9] that the field of manifold learning has a strong link to the recovery of compressed signals. Specifically, it has been demonstrated that a set of K sparse signals forms a K-dimensional Riemannian manifold. Further, it can be shown [9] that many manifold models can be expressed as an infinite union of subspaces.

Union of subspaces S*f*M: Recently, Zhu *et al.* [25] demonstrated a strategy for utilizing a union of local subspaces assumption within S*f*M. Specifically, the authors utilized an adaptation of Dai *et al.* [8] S*f*M approach - which simultaneously reconstruct the 3D structure and affinity matrix. The affinity matrix is of importance as it naturally encodes the cluster/subspace membership of each projected shape sample. Although exhibiting superior performance to Dai *et al.*'s approach for 3D structures that do not adhere to the low-rank assumption, the approach requires prior knowledge of the camera motion. Our proposed approach shares some commonalities with Zhu *et al.*'s work in that we are also making a union of subspaces assumption. An important difference, however, in our proposed approach is that we do not rely on any additional priors other than that the 3D structure is compressible. Specifically, we are able to automatically recover the camera matrices along with the 3D structure from a set of 2D projections.

Compressible S*f*M: The assumption that a 3D structure is compressible has been previously explored in S*f*M literature [26, 24]. Of particular note here is the work of Zhu and Lucey [26] where the authors: (i) assumed that the 3D structure is in a known temporal order, (ii) the camera motions are known, and (iii) the sparse basis is known a priori. Although sharing a similar assumption of the 3D structure being compressible, our work differs considerably to this work in that we employ no prior/constraints.

4. Uniqueness of SDL

An important component of our paper is associated with the uniqueness of Sparse Dictionary Learning (SDL) as it is sometimes known in literature [14]. In general terms the problem of SDL can be described as

$$\arg\min_{\mathbf{D},\boldsymbol{Z}} \|\boldsymbol{X} - \mathbf{D}\boldsymbol{Z}\|_F^2 \quad \text{s.t. } \|\mathbf{z}_i\|_0 = K, \quad i = 1:N \quad (4)$$

where we are trying to recover the concatenation of a sparse coefficient matrix $\boldsymbol{Z}$ and dictionary basis $\mathbf{D}$ from a known set of signals in $\boldsymbol{X} \in \mathbb{R}^{D\times N}$. Specifically, the sparse coefficient matrix is the concatenation of $K-$sparse coefficient vectors $\boldsymbol{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_D]$, and concatenation of $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_M]$ dictionary basis vectors. An important question to ask in the context of applying SDL to S*f*M is how unique is the solution to Equation 4?

Hillar *et al.* [14] recently characterized theoretically the answer to this question. The authors define that if any valid solution $\{\hat{\mathbf{D}}, \hat{\boldsymbol{Z}}\}$ to the SDL objective in Equation 4 is ambiguous up to a $M \times M$ permutation matrix $\boldsymbol{P}$ and a diagonal invertible weighting matrix $\boldsymbol{\Lambda}$ such that $\hat{\mathbf{D}} = \mathbf{D}\boldsymbol{P}\boldsymbol{\Lambda}$, and $\hat{\boldsymbol{Z}} = \boldsymbol{\Lambda}^{-1}\boldsymbol{P}^T\boldsymbol{Z}$, they say that $\boldsymbol{X}$ has a *unique SDL*. Moreover, they proved theoretically that, given large enough N[1], the uniqueness of SDL is achieved if and only if the dictionary $\mathbf{D}$ satisfies the spark condition: [2]:

$$\mathbf{D}\mathbf{z}_1 = \mathbf{D}\mathbf{z}_2 \quad \text{for } K\text{-sparse } \mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^M \Rightarrow \mathbf{z}_1 = \mathbf{z}_2. \quad (5)$$

Coherence as a proxy: The spark condition provides a complete characterization on the uniqueness of SDL. However, verifying whether a matrix $\mathbf{D}$ satisfies the spark condition is an NP-hard problem, which has to visit all $\binom{M}{K}$ subspaces. It is preferable in practice to use properties of $\mathbf{D}$ that are easily computable such as coherence - which measures the largest absolute inner product between any two column vectors in the matrix - and with high probability is indicative of the spark condition of the matrix. In the experimental portion of this paper we shall demonstrate how the coherence of a matrix can be utilized to predict the reconstructibility of a 3D structure solely from its 2D projections.

Block SDL: As we will discuss in the next section, there is a strong connection between compressible S*f*M and Block SDL (BSDL). BSDL is a generalization of the SDL objective in Equation 4:

$$\arg\min_{\mathbf{D},\boldsymbol{Z}} ||\boldsymbol{X} - \mathbf{D}\boldsymbol{Z}||_F^2 \quad \text{s.t.} ||\boldsymbol{Z}_i||_{0,\alpha} = K, \quad i = 1:N/\beta, \quad (6)$$

where $\boldsymbol{Z}_i \in \mathbb{R}^{D\times\beta}$ is a submatrix of $\boldsymbol{Z}$, *i.e.* $\boldsymbol{Z} = [\boldsymbol{Z}_1, ..., \boldsymbol{Z}_{N/\beta}]$. Each $\boldsymbol{Z}_i$ is divided into $\boldsymbol{M}/\alpha$ blocks of size $\alpha \times \beta$ and $\|\boldsymbol{Z}_i\|_{0,\alpha}$ counts the number of blocks of which at least one element is non-zero. α and β need to be chosen such that D and M are perfectly divisible. One of particular importance in our compressible S*f*M problem is 3×2 block-sparsity which we will describe in more detail in the next section on compressible S*f*M.

Definition 1. *If any valid solution* $\{\hat{\mathbf{D}}, \hat{\boldsymbol{Z}}\}$ *to the objective in Equation 6 is ambiguous only up to a* $M \times M$ *block permutation matrix* $\boldsymbol{P}_\alpha$ *and a block-diagonal invertible weighting matrix* $\boldsymbol{\Lambda}_\alpha$ *such that* $\hat{\mathbf{D}} = \mathbf{D}\boldsymbol{P}_\alpha\boldsymbol{\Lambda}_\alpha$, *and* $\hat{\boldsymbol{Z}} = \boldsymbol{\Lambda}_\alpha^{-1}\boldsymbol{P}_\alpha^T\boldsymbol{Z}$, *we say* $\boldsymbol{X}$ *has a unique BSDL.*

[1] Hillar *et al.* offer a lower bound for N, refer to [14] for full discussion
[2] Refer to [14] for the proof.

The block permutation matrix is actually defined as $\boldsymbol{P}_\alpha = \boldsymbol{P} \otimes \boldsymbol{I}_\alpha$ where $\boldsymbol{P}$ is an arbitrary $(M/\alpha) \times (M/\alpha)$ permutation matrix and $\boldsymbol{I}_\alpha$ is a $\alpha \times \alpha$ identity matrix. The block-diagonal invertible weighting matrix $\boldsymbol{\Lambda}_\alpha$ has a $\alpha \times \alpha$ block structure. We now ask the same question: what is the sufficient and necessary condition for the uniqueness of BSDL?

Theorem 1. *Given large enough N[3], the uniqueness of BSDL holds if and only if the matrix $\mathbf{D}$ satisfies the block spark condition:*

$$\mathbf{D}\boldsymbol{Z}_1 = \mathbf{D}\boldsymbol{Z}_2 \quad \text{for } K \text{ block-sparse } \boldsymbol{Z}_1, \boldsymbol{Z}_2 \in \mathbb{R}^{M\times\beta} \Rightarrow \boldsymbol{Z}_1 = \boldsymbol{Z}_2. \tag{7}$$

The complete mathematical proof of Theorem 1 is offered in supplementary material.

5. "Prior-Less" Compressible S*f*M

One can view much of the literature of low-rank S*f*M drawing heavily upon the fact that one can obtain a solution to the rank constrained factorization problem

$$\arg\min_{\boldsymbol{\Pi},\boldsymbol{B}} ||\boldsymbol{W} - \boldsymbol{\Pi}\boldsymbol{B}||_F^2, \quad \text{s.t. } \operatorname{rank}(\boldsymbol{\Pi}) = 3K \tag{8}$$

through an SVD. Even though the SVD returns a unique solution $\{\hat{\boldsymbol{\Pi}}, \hat{\boldsymbol{B}}\}$ it is easy to demonstrate that this solution is just one of many possible solutions to $\boldsymbol{W} = \hat{\boldsymbol{\Pi}}\hat{\boldsymbol{B}} = \hat{\boldsymbol{\Pi}}\boldsymbol{G}\boldsymbol{G}^{-1}\hat{\boldsymbol{B}} = \boldsymbol{\Pi}\boldsymbol{B}$ where the corrective matrix $\boldsymbol{G}$ is any non-singular matrix. The ambiguity of this factorization is problematic for S*f*M problems as additional constraints are required to obtain a unique solution.

For rigid S*f*M (*i.e.* $K = 1$) the application of camera constraints [19] is typically sufficient in order to find a correction matrix $\boldsymbol{G}$ that gives a unique solution. Xiao *et al*. [23] famously demonstrated for $K > 1$ that one cannot determine a unique $\boldsymbol{G}$ since the space of solutions lies in a nullspace of rank $2K^2 - K$. Akhter *et al*. [2] additionally demonstrated that even though $\boldsymbol{G}$ is not unique, any solution to $\boldsymbol{G}$ that satisfies the camera constraints returns a valid 3D shape and camera motion pair. In this paper we want to explore whether moving away from canonical rank constraints and instead assuming that $\boldsymbol{\Pi}$ is block-sparse could result in a far less ambiguous factorization thus resulting in an S*f*M algorithm that can circumvent current theoretical and practical limitations.

Why block-sparse?: Let us assume that the unknown 3D structures $\boldsymbol{S}^\sharp$ are compressible, that is, the 3D structure in each frame (each row of $\boldsymbol{S}^\sharp$) can be approximated by only K basis shapes (K rows of $\boldsymbol{B}^\sharp$.) Therefore, the factorization $\boldsymbol{S}^\sharp = \boldsymbol{C}^\sharp\boldsymbol{B}^\sharp$ results in a set of coefficients $\boldsymbol{C}^\sharp \in \mathbb{R}^{F\times L}$ whose rows are each K-sparse. As pointed out in the previous section, one never has access to the 3D structure $\boldsymbol{S}^\sharp$ a priori only the 2D projections $\boldsymbol{W}$. Interestingly, however, if we know $\boldsymbol{S}^\sharp$ is compressible then from Equation 2 (*i.e.* $\boldsymbol{\Pi} = \boldsymbol{M}(\boldsymbol{C}^\sharp \otimes \boldsymbol{I}_3)$) $\boldsymbol{\Pi}$ *must* be 2×3 block sparse as the camera matrix $\boldsymbol{M}$ is 2×3 block-diagonal. It is this insight that forms the crucial component of our algorithm. From a known measurement matrix $\boldsymbol{W}$ and desired K, L, one can factorize $\boldsymbol{W}^T$ through a 3×2 BSDL process. Note: for S*f*M $\boldsymbol{W} = \boldsymbol{\Pi}\boldsymbol{B}$, whereas for BSDL this would be expressed as $\boldsymbol{W}^T = \boldsymbol{B}^T\boldsymbol{\Pi}^T$ where $\boldsymbol{X} = \boldsymbol{W}^T, \mathbf{D} = \boldsymbol{B}^T$, and $\boldsymbol{Z} = \boldsymbol{\Pi}^T$.

Theorem 2. *If one can recover $\hat{\boldsymbol{B}}$ using a 3×2 BSDL such that $\mathbf{D} = \hat{\boldsymbol{B}}^T$ satisfies the block spark condition, then it can be shown that the transpose of $\hat{\boldsymbol{B}}^\sharp$ satisfies the canonical spark condition, where $\hat{\boldsymbol{B}}^\sharp$ is an $L \times 3P$ reshape of $\hat{\boldsymbol{B}}$. Further, for such BSDL to be unique, K must be less than or equal to $P/3 - 1$.*

Proof. Suppose two K-sparse vectors $\mathbf{z}_1$ and $\mathbf{z}_2$ such that $(\hat{\boldsymbol{B}}^\sharp)^T\mathbf{z}_1 = (\hat{\boldsymbol{B}}^\sharp)^T\mathbf{z}_2$. Then from the reshape, it follows that $\hat{\boldsymbol{B}}^T(\mathbf{z}_1 \otimes \boldsymbol{I}_3) = \hat{\boldsymbol{B}}^T(\mathbf{z}_2 \otimes \boldsymbol{I}_3)$. As $\hat{\boldsymbol{B}}^T$ satisfies the block spark condition, it follows that $\mathbf{z}_1 = \mathbf{z}_2$, therefore, $(\boldsymbol{B}^\sharp)^T$ satisfies the canonical spark condition. Further, the uniqueness of the BSDL factorization requires $\hat{\boldsymbol{B}}^T$ to satisfy the block spark condition. This implies that any $P\times3(K+1)$ submatrices generated by concatenating $K+1$ block columns of $\hat{\boldsymbol{B}}^T$ needs to be full column rank[4]. Therefore K need to be less than or equal to $P/3 - 1$. □

Theorem 2 actually tells us that the uniqueness of the BSDL factorization on 2D projections automatically guarantees the uniqueness of the SDL factorization on the unknown 3D structures. Interestingly, the converse is not always true. This result highlights a drawback in our proposed approach, that is, we cannot recover all compressible structures but the subsets where $\hat{\boldsymbol{\Pi}}$ is sufficiently sparse ($K \leq P/3 - 1$) and $\hat{\boldsymbol{B}}$ satisfies the block spark condition. In the experiments section, we show a strategy that can be utilized in practice to improve the incoherence of $\hat{\boldsymbol{B}}$ and push it to satisfy the block spark condition.

6. Camera and Structure Recovery

As the scale of camera and size of structures are inherently relative, we simply set the camera scale σ to unity, such that $\boldsymbol{M}_f\boldsymbol{M}_f^T = \boldsymbol{I}_2$. Assuming that $\boldsymbol{W} = \hat{\boldsymbol{\Pi}}\hat{\boldsymbol{B}}$ has a unique BSDL, from Definition 1, the corrective matrix $\boldsymbol{G}$ must be of form $\boldsymbol{G} = (\boldsymbol{P} \otimes \boldsymbol{I}_3)\boldsymbol{\Lambda}$. As the permutation ambiguity has no bearing on camera motion and 3D structure, we set P to identity, therefore $\boldsymbol{G} = \boldsymbol{\Lambda}$.

[3] A lower bound of N can be established for BSDL, please refer to the supplementary material for full discussion.

[4] A counterexample for contradiction: if $K = 2$, and $\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3$ are 3 linear dependent block columns of $\hat{\boldsymbol{B}}^T$. In addition, suppose any 2 of them are linear independent. Then subspace spanned by $\{\boldsymbol{b}_1, \boldsymbol{b}_2\}$ are identical to one by $\{\boldsymbol{b}_1, \boldsymbol{b}_3\}$, which breaks the block spark condition.

Denote $\boldsymbol{G}_j$ as j-th block on diagonal of $\boldsymbol{G}$, and $\hat{\boldsymbol{\Pi}}_j, \boldsymbol{\Pi}_j \in \mathbb{R}^{2F\times 3}$ as the j-th coloumn-triplet of $\hat{\boldsymbol{\Pi}}, \boldsymbol{\Pi}$ respectively. From the structure of corrective matrix, it follows that $\boldsymbol{\Pi}_j = \hat{\boldsymbol{\Pi}}_j \boldsymbol{G}_j$, for $j = 1, ..., L$. Define Ω_j as the set of indices pointing to the block $\hat{\boldsymbol{\Pi}}_{ij} \in \mathbb{R}^{2\times 3}$ that is active, *i.e.* $\Omega_j = \text{supp}(\hat{\boldsymbol{\Pi}}_j) = \{i | 1 \le i \le F, \hat{\boldsymbol{\Pi}}_{ij} \neq \boldsymbol{0}\}$. If a certain Ω_j is empty, it is implied that the corresponding atom in the dictionary has never been used. We can then decrease L, and re-learn the dictionary so that Ω_j is never empty.

From Equation 2 (*i.e.* $\boldsymbol{W} = \boldsymbol{M}(\boldsymbol{C}^\sharp \otimes \boldsymbol{I}_3)\boldsymbol{B} = \boldsymbol{\Pi}\boldsymbol{B}$), it is known that $\boldsymbol{\Pi}_{ij} = c_{ij}\boldsymbol{M}_i$, where c_{ij} is $ij-$th elements of $\boldsymbol{C}^\sharp$. Thus, since Ω_j can never be empty, $\hat{\boldsymbol{\Pi}}_{ij}\boldsymbol{G}_j = \boldsymbol{\Pi}_{ij} = c_{ij}\boldsymbol{M}_i$, for each $i \in \Omega_j$. From camera constraints, it follows that

$$\hat{\boldsymbol{\Pi}}_{ij}\boldsymbol{G}_j\boldsymbol{G}_j^T\hat{\boldsymbol{\Pi}}_{ij}^T = c_{ij}^2\boldsymbol{M}_i\boldsymbol{M}_i^T = c_{ij}^2\boldsymbol{I}_2, \quad i \in \Omega_j, \tag{9}$$

and for convenience, let $\boldsymbol{Q}_j = \boldsymbol{G}_j\boldsymbol{G}_j^T$. Since c_{ij} is unknown, let us eliminate it and rewrite Equation 9 as

$$(\hat{\boldsymbol{\Pi}}_{ij}\boldsymbol{Q}_j\hat{\boldsymbol{\Pi}}_{ij}^T)_{11} = (\hat{\boldsymbol{\Pi}}_{ij}\boldsymbol{Q}_j\hat{\boldsymbol{\Pi}}_{ij}^T)_{22}, (\hat{\boldsymbol{\Pi}}_{ij}\boldsymbol{Q}_j\hat{\boldsymbol{\Pi}}_{ij}^T)_{12} = 0, \tag{10}$$

where $(\cdot)_{ij}$ denotes the (i, j)-th elements. Now, denote $\boldsymbol{q}_j = \text{vec}(\boldsymbol{Q}_j)$ as the vectorization of $\boldsymbol{Q}_j$. Let us rewrite Equation 10 in a compact way with the fact that $\text{vec}(\hat{\boldsymbol{\Pi}}_{ij}\boldsymbol{Q}_j\hat{\boldsymbol{\Pi}}_{ij}) = (\hat{\boldsymbol{\Pi}}_{ij} \otimes \hat{\boldsymbol{\Pi}}_{ij})\boldsymbol{q}_j$:

$$\begin{bmatrix} \hat{\boldsymbol{\Pi}}_{ij} \otimes \hat{\boldsymbol{\Pi}}_{ij}(1,:) - \hat{\boldsymbol{\Pi}}_{ij} \otimes \hat{\boldsymbol{\Pi}}_{ij}(4,:) \\ \hat{\boldsymbol{\Pi}}_{ij} \otimes \hat{\boldsymbol{\Pi}}_{ij}(2,:) \end{bmatrix} \boldsymbol{q}_j = \boldsymbol{A}_{ij}\boldsymbol{q}_j = 0, \tag{11}$$

where $\hat{\boldsymbol{\Pi}}_{ij} \otimes \hat{\boldsymbol{\Pi}}_{ij}(k,:)$ denotes k-th row of $\hat{\boldsymbol{\Pi}}_{ij} \otimes \hat{\boldsymbol{\Pi}}_{ij}$. Stacking all such equations for all $i \in \Omega_j$, we obtain

$$\boldsymbol{A}_j\boldsymbol{q}_j = 0. \tag{12}$$

Circumventing the nullspace: One benefit of Equation 12 is that $\boldsymbol{A}_j \in \mathbb{R}^{2|\Omega_j|\times 9}$, where $|\Omega_j|$ is the number of elements in set Ω_j, with high possibility will be overcomplete as $F \gg L$. This result is important as it circumvents the nullspace issue faced by low-rank S*f*M. This null space issue can be problematic in many practical scenarios due to its sensitivity to noise. Similar to Tomasi-Kanade's method [19], we simply pick up the eigenvector corresponding to the least eigenvalue of $\boldsymbol{A}_j^T\boldsymbol{A}_j$ and then $\boldsymbol{Q}_k \in \mathbb{S}_+^3$ holds automatically.

Once $\boldsymbol{Q}_j$ is estimated, the absolute value of c_{ij} can be computed by Equation 9. The sign of c_{ij}, however, is not able to be determined, which actually is an inherent ambiguity without assuming any temporal prior of camera or structures. Considering equation $\boldsymbol{W} = \boldsymbol{M}\boldsymbol{S}$, any block diagonal matrix $\text{blkdiag}(\pm\boldsymbol{I}_3)$ can be inserted between $\boldsymbol{M}$ and $\boldsymbol{S}$, but the compressibility assumption and camera constraint still hold. Dai *et al.* [8] *breaks* their "prior-free" assertion by restricting the camera movement between frames to at most $\pm 90°$ to determine the sign of c_{ij}. In our paper, however, we claim that the absolute sign of c_{ij} cannot be determined by current assumption, but the relative sign in each column can. Thus, the camera matrix and structures can be recovered but up to a sign ambiguity.

Enforcing camera consistency: Let us consider the submatrix $\boldsymbol{G}_j$ in isolation,

$$\hat{\boldsymbol{\Pi}}_{ij}\boldsymbol{G}_j = c_{ij}\boldsymbol{M}_i, \quad \text{for } i \in \Omega_j. \tag{13}$$

One can recover the camera matrices $\{\boldsymbol{M}_i\}_{i\in\Omega_j}$ by solving the system of equations above. Further, if one was to then choose another $\boldsymbol{G}_k$ where $j \neq k$, such that one or more indexes in Ω_j are shared with Ω_k, one can equally recover the camera matrices $\{\boldsymbol{M}_i^*\}_{i\in\Omega_k}$. An inconsistency arises, however, such that we cannot guarantee that

$$\boldsymbol{M}_i^* = \boldsymbol{M}_i, \quad \text{for} \quad i \in \Omega_j \cap \Omega_k. \tag{14}$$

This inconsistency does not just occur across pairs of submatrices within $\boldsymbol{G}$, but actually across all possible submatrices of $\boldsymbol{G}$ with overlapping active blocks. We attempt to resolve this inconsistency in a recursive manner by solving for an orthonormal matrix $\boldsymbol{H}_k$ such that $\boldsymbol{M}_i^*\boldsymbol{H}_k = \boldsymbol{M}_i$. First, we choose an arbitrary $\boldsymbol{G}_j$ (typically the one with most active blocks) and solve for the cameras $\{\boldsymbol{M}_i\}_{i\in\Gamma}$, where we initially set $\Gamma = \Omega_j$. Then we choose a $\boldsymbol{G}_k$ whose $|\Gamma \cap \Omega_k|$ is largest. We solve for the cameras $\{\boldsymbol{M}_i^*\}_{i\in\Omega_k}$, and then find an orthonormal $\boldsymbol{H}_k$ such that,

$$\begin{aligned} \arg\min_{\boldsymbol{H}_k,\boldsymbol{\eta}} \quad & \sum_{i\in\Gamma\cap\Omega_k} \|\boldsymbol{M}_i - \eta_i\boldsymbol{M}_i^*\boldsymbol{H}_k\|_F \\ \text{s.t.} \quad & \boldsymbol{H}_k^T\boldsymbol{H}_k = \boldsymbol{I}, \eta_i = \{+1, -1\}, \end{aligned} \tag{15}$$

where η_i contains the relative sign of elements in $\boldsymbol{C}^\sharp$ for Γ. For the element in $\boldsymbol{C}^\sharp$ that are not explicitly defined through $\boldsymbol{\eta}$, we set them arbitrarily to be positive. We then update $\Gamma \leftarrow \Gamma \cup \Omega_k$ and repeat the process until all cameras and relative signs in $\boldsymbol{C}^\sharp$ are known.

The structure matrix $\boldsymbol{S}$ is then recovered by $(\boldsymbol{C}^\sharp \otimes \boldsymbol{I}_3)\boldsymbol{H}^{-1}\boldsymbol{G}^{-1}\boldsymbol{B}$, where $\boldsymbol{H}$ is a matrix with $\boldsymbol{H}_1, ..., \boldsymbol{H}_L$ on main diagonal.

7. BSDL Algorithm

In this section, we describe our BSDL algorithm that adapts K-SVD [18], OMP [21] and FOCUSS [11] to block sparse situation respectively. However, any valid BSDL method can be employed here as long as it returns a valid factorization $\boldsymbol{W} = \hat{\boldsymbol{\Pi}}\hat{\boldsymbol{B}}$.

Block K-SVD: Similar to regular K-SVD, block K-SVD is an iterative algorithm with 2 steps: 1) Fixing dictionary, solve block-sparse representation by block OMP or block

FOCUSS, and 2) Fixing block-sparse pattern, update dictionary by SVD. The only alternation from regular K-SVD is to keep the first α singular values instead of one when updating each block columns of the dictionary. For compressible SfM $\alpha = 3$. The techniques to get rid of local minimal reported in [18] are also valid and serve in block K-SVD.

Block OMP: To solve block sparse approximation problem, we extend regular Orthogonal Matching Pursuit (OMP) [21] to block OMP. Both of them are greedy algorithms picking the first K atoms in dictionary describing the signal best. Specifically, in each iteration, block OMP computes the inner product of residual and each dictionary atoms left, and picked the atoms corresponding to least inner product value. Then it computes coefficients, associates with chosen atoms, updates residual and repeats until the number of chosen atoms hits the known number K. Block OMP is efficient compared to block FOCUSS, but it succeeds only when the dictionary is sufficient incoherent.

Block FOCUSS: Serving the same function as block OMP, we adapt FOcal Underdetermined System Solver (FOCUSS) [11] to block FOCUSS to estimate the block sparse approximation. Block FOCUSS and FOCUSS are iterative algorithms solving the ℓ_p-norm ($p < 1$) relaxation of block sparse approximation and regular sparse approximation respectively. The only difference between them is the design of the weight matrix $\boldsymbol{W}_{pk}$ (refer to [11] for more detail.) Other than letting $\boldsymbol{W}_{pk} = \text{diag}(\boldsymbol{x}^{k-1})$, block FOCUSS updates $\boldsymbol{W}_{pk}$ by the Frobenius norm of each block in $\boldsymbol{x}^{k-1}$, which promotes elements in one block to be either all active or all zeros. A regularization technique [17] serves also in block FOCUSS balancing the approximation error and sparsity of estimated coefficients. Block FOCUSS can often achieve successful block-sparse estimation even in circumstances where block OMP fails. One drawback, however, is its speed as it is dramatically slower than block OMP.

Initialization: The BSDL factorization itself is inherently an NP-hard problem, therefore it is important to have a good initialization. We relax the BSDL objective using a block ℓ_1-norm, and solve the relaxed problem by Alternating Direction Method of Multipliers (ADMM) [4, 1, 10, 6]. Even though the relaxed problem is not convex either, ADMM splits the objective into several small *convex* sub-problems by introducing several auxiliary variables. A stationary point can be achieved for our ADMM initialization through the judicious choice of parameters [6].

8. Experiments

MATLAB code has been released, check ci2cv.net/paper for further information.

Compressibility: Our first experiment explores the compressibility of real 3D structures from the CMU Motion Capture dataset, where we learned various dictionaries with different dictionary size L and sparsity level K. Figure 2 clearly shows that the real 3D structures are modelled well by our compressibility assumption and the coherence of the learned dictionary is being controlled by balancing the approximation error. This result offers a strategy to achieve a unique BSDL factorization at the cost of approximating structures less precisely, which extends the application of our method.

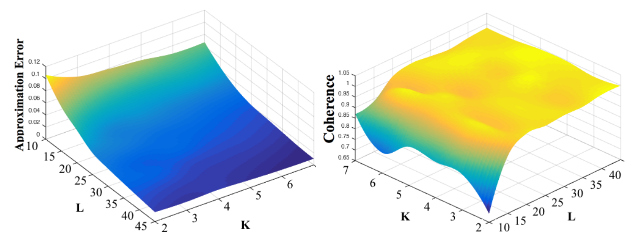


Figure 2: The results of SDL factorization for Motion-4 by Subject-5 in CMU Motion Capture. **Left:** The approximation error. **Right:** The coherence of learned dictionary. With the decrease of K and L, the coherence of learned dictionary becomes better at the cost of approximating structures less precisely.

Recovering temporal order: In Figure 4 we demonstrated that the sparse codes recovered using our method have a natural temporal coherence. This indicates our prior-less approach could be useful for the recovery of the temporal order of 3D structures in future applications. The full analysis of this phenomena is outsize of the scope of this paper.

High-rank performance: To verify the performance of the proposed method on high-rank and full-rank structures, we conducted experiments with synthetic data where the rank of structures is easily controlled. We utilized Dai *et al.*'s work[5] as a baseline, which demonstrated that it outperforms other low-rank SfM methods in [8].

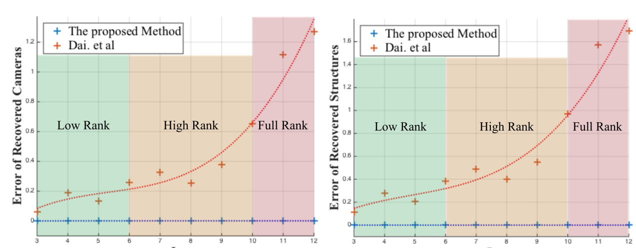


Figure 3: **Left:** The error of estimated camera matrix. **Right:** The error of estimated structures. The error matrices follows [3, 12, 8]. Our methods obtained nearly perfectly results irrespective to rank of structures.

[5]In all our experiments, we visit all possible rank k in [8] to get a final estimation

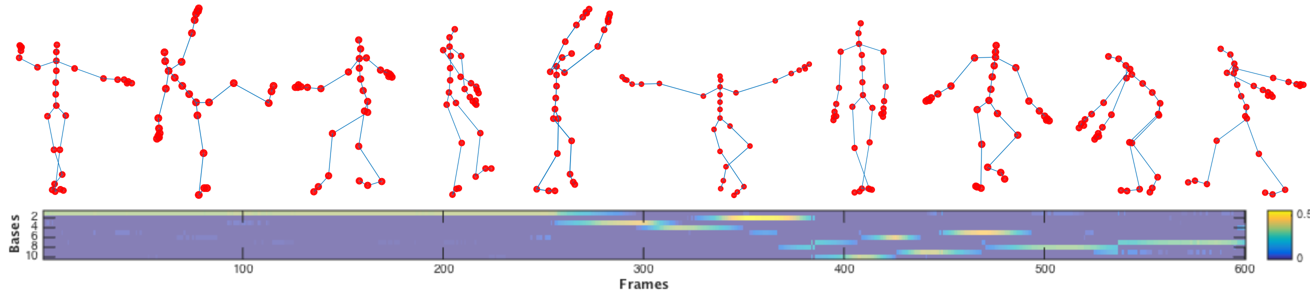


Figure 4: **Top:** 10 learned basis structures for Motion-4 by Subject-5 in CMU Motion Capture when $K = 2, L = 10$. These bases are learned from 3D shape sequences and identical to those learned from 2D image sequences, due to the uniqueness of BSDL. **Bottom:** The visualization of coefficients. The coefficients of each atoms varies gradually in a shape of Gaussian distribution, which reveals the temporal information of video sequence. It is not used in S*f*M, but may be useful for recovering the temporal order of 3D structure in future applications.

The compressible structure $\boldsymbol{S}$, with 100 frames and 30 points in each frames, are generated by random dictionary of size L, such that $\text{rank}(\boldsymbol{S}) = 3L$. We repeat the proposed method as well as Dai *et al.*'s method 50 times for each L from 3 to 12. The results are summarized in Figure 3. It is seen that our method works perfectly and robustly on structures with any rank, while the low-rank S*f*M fails in high-rank and full-rank situations. Moreover, even in low-rank situation, the proposed method outperforms the Dai *et al.*'s method.

Noise performance: To evaluate the performance under noise, we repeat the experiments on low-rank structures (with $L = 5$) at different noise ratios, defined as $\frac{\|W - W_0\|_F}{\|W_0\|_F}$. The Figure 5 demonstrates that our method is sensitive to noise. However, it still works no worse than Dai *et al.*'s method even at high noise ratios.

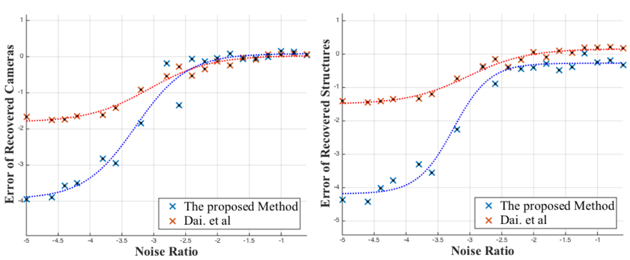


Figure 5: **Left:** The error of estimated camera matrix. **Right:** The error of estimated structures. Both x- and y-axis are in logarithm space. Our method is sensitive to noise, while it still works no worse than the baseline even at high noise.

Practical performance: The proposed method is evaluated on real compressible structures: Motion-4, -5, -6, -7, -8 by Subject-5, and Motion-2, -4 by Subject-1, Motion-5 by Subject-2, Motion-3, -4 by Subject-3 and Motion-13 by Subject-6 in CMU Motion Captures, and a Shark sequence in [20]. The visual evaluation is summarized in Figure 7,

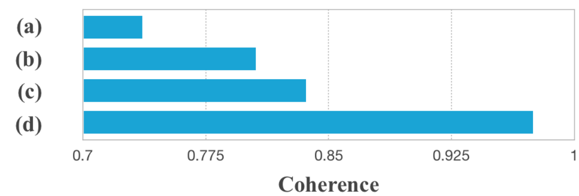
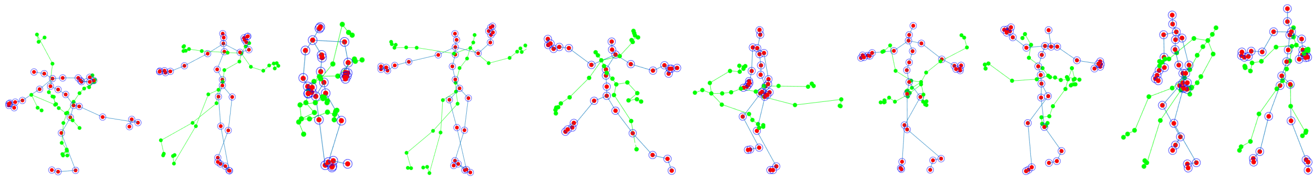


Figure 6: Coherence of learned dictionaries corresponding to the 4 sequences in Figure 7. The coherence for (d) is too poor to guarantee the uniqueness of BSDL. In addition, it also shows that checking the coherence of learned dictionary beforehand is a good way to predict the performance especially when ground truth are not accessible.

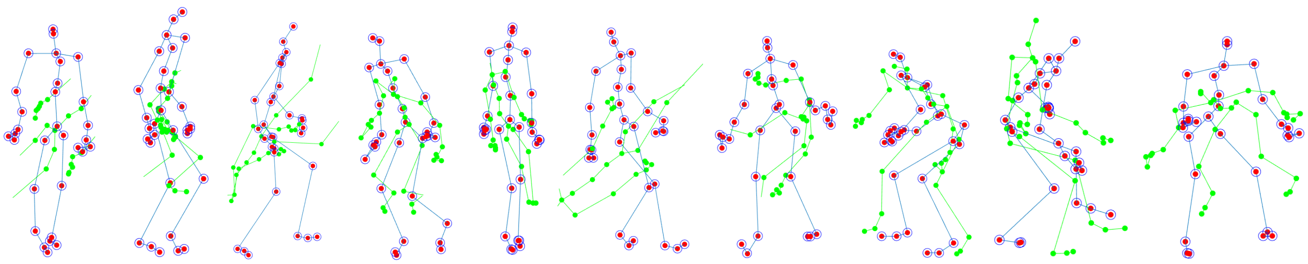
which shows that our method obtains impressive results in Figure 7a, 7b, 7c, while it fails in Figure 7d. Actually, this failure is able to be forecast even without ground truth. As shown in Figure 6, the coherence of the learned dictionary for sequence Shark is too poor to guarantee the uniqueness of the BSDL factorization. This insight offers an effective way to predict the reconstructibility of 3D structure when the ground truth structure are not available in practice.

9. Conclusion

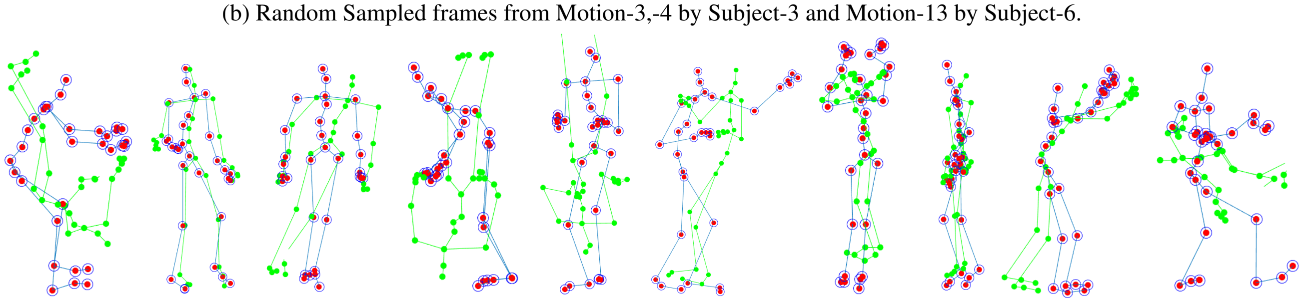
In this paper, we demonstrated that a compressible 3D structure under weak perspective projection is 2×3 block-compressible. Moreover, if a 2×3 unique BSDL factorization can be obtained (of the 2D projections), we showed that the compressible 3D structure and camera motion can be recovered solely by the assumption of compressibility. Superior reconstruction results using our method are achieved in comparison to Dai *et al.*'s low-rank S*f*M method. Impressive results were demonstrated on both synthetic and real-world compressible 3D structures. Finally, we proposed the use of dictionary coherence as a measure of reconstructibility of the projected 3D structures without ground truth -



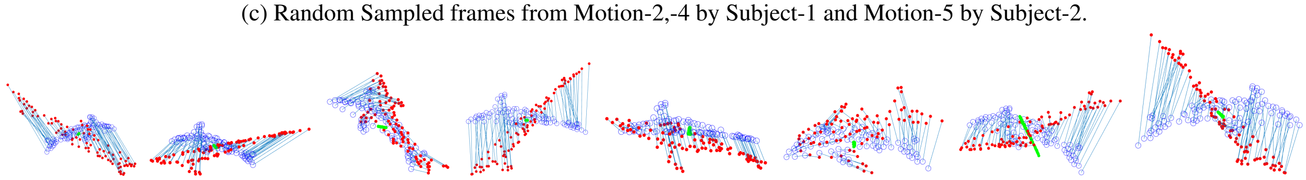
(a) Random Sampled frames from Motion-4,-5,-6 by Subject-5.



(b) Random Sampled frames from Motion-3,-4 by Subject-3 and Motion-13 by Subject-6.



(c) Random Sampled frames from Motion-2,-4 by Subject-1 and Motion-5 by Subject-2.



(d) Random Sampled frames from Shark sequence.

Figure 7: Visual evaluation of estimated structures. **Red dots:** The proposed method. **Blue circles:** Ground Truth. **Green dots:** Dai *et al.*'s method. The proposed method obtained an impressive performance for compressible structures. However it fails in Shark sequences due to the poor coherence of learned dictionary.

which adds practical utility of our approach.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No.1526033.

References

[1] A. Agarwal, A. Anandkumar, P. Jain, and P. Netrapalli. Learning sparsely used overcomplete dictionaries via alternating minimization. *arXiv preprint arXiv:1310.7991*, 2013. 6

[2] I. Akhter, Y. Sheikh, and S. Khan. In defense of orthonormality constraints for nonrigid structure from motion. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1534–1541. IEEE, 2009. 4

[3] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade. Nonrigid structure from motion in trajectory space. In *Advances in neural information processing systems*, pages 41–48, 2009. 2, 6

[4] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011. 6

[5] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 690–696. IEEE, 2000. 1, 2

[6] H. Bristow, A. Eriksson, and S. Lucey. Fast convolutional sparse coding. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 391–398. IEEE, 2013. 6

[7] E. J. Candè and M. B. Wakin. An introduction to com-

pressive sampling. *Signal Processing Magazine, IEEE*, 25(2):21–30, 2008. 1

[8] Y. Dai, H. Li, and M. He. A simple prior-free method for non-rigid structure-from-motion factorization. *International Journal of Computer Vision*, 107(2):101–122, 2014. 1, 2, 3, 5, 6

[9] M. A. Davenport, M. F. Duarte, Y. C. Eldar, and G. Kutyniok. Introduction to compressed sensing. *Preprint*, 93:1–64, 2011. 3

[10] W. Deng, W. Yin, and Y. Zhang. Group sparse optimization by alternating direction method. In *SPIE Optical Engineering+ Applications*, pages 88580R–88580R. International Society for Optics and Photonics, 2013. 6

[11] I. F. Gorodnitsky and B. D. Rao. Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm. *Signal Processing, IEEE Transactions on*, 45(3):600–616, 1997. 5, 6

[12] P. F. Gotardo and A. M. Martinez. Computing smooth time trajectories for camera and deformable shape in structure from motion with occlusion. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(10):2051–2065, 2011. 6

[13] P. F. Gotardo and A. M. Martinez. Kernel non-rigid structure from motion. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 802–809. IEEE, 2011. 2, 3

[14] C. Hillar and F. T. Sommer. When can dictionary learning uniquely recover sparse data from subsamples? In *IEEE Transactions on Information Theory*, 2015. 3

[15] H. S. Park and Y. Sheikh. 3d reconstruction of a smooth articulated trajectory from a monocular image sequence. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 201–208. IEEE, 2011. 2

[16] V. Rabaud and S. Belongie. Re-thinking non-rigid structure from motion. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 3

[17] B. Rao and K. Kreutz-Delgado. Basis selection in the presence of noise. In *Signals, Systems & Computers, 1998. Conference Record of the Thirty-Second Asilomar Conference on*, volume 1, pages 752–756. IEEE, 1998. 6

[18] R. Rubinstein, T. Peleg, and M. Elad. Analysis k-svd: A dictionary-learning algorithm for the analysis sparse model. *Signal Processing, IEEE Transactions on*, 61(3):661–677, 2013. 2, 5, 6

[19] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 9(2):137–154, 1992. 2, 4, 5

[20] L. Torresani, A. Hertzmann, and C. Bregler. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(5):878–892, 2008. 2, 7

[21] J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *Information Theory, IEEE Transactions on*, 53(12):4655–4666, 2007. 5, 6

[22] J. Valmadre, Y. Zhu, S. Sridharan, and S. Lucey. Efficient articulated trajectory reconstruction using dynamic programming and filters. In *Computer Vision–ECCV 2012*, pages 72–85. Springer, 2012. 2

[23] J. Xiao, J. Chai, and T. Kanade. A closed-form solution to non-rigid shape and motion recovery. *International Journal of Computer Vision*, 67(2):233–246, 2006. 2, 4

[24] Y. Zhu, M. Cox, and S. Lucey. 3d motion reconstruction for real-world camera motion. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1–8. IEEE, 2011. 3

[25] Y. Zhu, D. Huang, F. D. L. Torre, and S. Lucey. Complex non-rigid motion 3d reconstruction by union of subspaces. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1542–1549. IEEE, 2014. 2, 3

[26] Y. Zhu and S. Lucey. Convolutional sparse coding for trajectory reconstruction. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(3):529–540, 2015. 3

Prior-Less Compressible Structure from Motion: Supplementary Material

Chen Kong Simon Lucey
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA, US
{chenk, slucey}@cs.cmu.edu

Recall the objective of Block Sparse Dictionary Learning (BSDL) is

$$\underset{\mathbf{D},\boldsymbol{Z}}{\operatorname{argmin}} ||\boldsymbol{X}-\mathbf{D}\boldsymbol{Z}||_F^2 \quad \text{s.t.} ||\boldsymbol{Z}_i||_{0,\alpha}=K, \quad i=1:N/\beta, \tag{1}$$

where $\boldsymbol{Z}_i \in \mathbb{R}^{D\times\beta}$ is a submatrix of $\boldsymbol{Z}$, *i.e.* $\boldsymbol{Z} = [\boldsymbol{Z}_1, ..., \boldsymbol{Z}_{N/\beta}]$. Each $\boldsymbol{Z}_i$ is divided into $\boldsymbol{M}/\alpha$ blocks of size $\alpha\times\beta$ and $\|\boldsymbol{Z}_i\|_{0,\alpha}$ counts the number of blocks of which at least one element is non-zero. α and β need to be chosen such that D and M are perfectly divisible.

Definition 1. *If any valid solution $\{\hat{\mathbf{D}},\hat{\boldsymbol{Z}}\}$ to the objective in Equation 1 is ambiguous only up to a $M\times M$ block permutation matrix $\boldsymbol{P}_\alpha$ and a block-diagonal invertible weighting matrix $\boldsymbol{\Lambda}_\alpha$ such that $\hat{\mathbf{D}} = \mathbf{D}\boldsymbol{P}_\alpha\boldsymbol{\Lambda}_\alpha$, and $\hat{\boldsymbol{Z}} = \boldsymbol{\Lambda}_\alpha^{-1}\boldsymbol{P}_\alpha^T\boldsymbol{Z}$, we say $\boldsymbol{X}$ has a unique BSDL.*

The block permutation matrix is actually defined as $\boldsymbol{P}_\alpha = \boldsymbol{P}\otimes\boldsymbol{I}_\alpha$ where $\boldsymbol{P}$ is an arbitrary $(M/\alpha)\times(M/\alpha)$ permutation matrix and $\boldsymbol{I}_\alpha$ is a $\alpha\times\alpha$ identity matrix. The block-diagonal invertible weighting matrix $\boldsymbol{\Lambda}_\alpha$ has a $\alpha\times\alpha$ block structure. We now ask the same question: what is the sufficient and necessary condition for the uniqueness of BSDL?

Theorem 1. *There exist $K\binom{M/\alpha}{K}^2$ K-block-sparse vectors $\boldsymbol{Z}_1, ..., \boldsymbol{Z}_{N/\beta}$,* i.e. *$N = \beta K\binom{M/\alpha}{K}^2$, such that the uniqueness of BSDL holds if and only if the matrix $\mathbf{D}$ satisfies the block spark condition:*

$$\mathbf{D}\boldsymbol{Z}_1 = \mathbf{D}\boldsymbol{Z}_2 \quad \textit{for } K\textit{-block-sparse } \boldsymbol{Z}_1,\boldsymbol{Z}_2\in\mathbb{R}^{M\times\beta} \Rightarrow \boldsymbol{Z}_1=\boldsymbol{Z}_2. \tag{2}$$

Let's first prove Theorem 1 in the case when $\beta=1$ and once it is proven, the general case $\beta>1$ is simple to handle: We can split sparse causes $\boldsymbol{Z}^i$ into $[\mathbf{z}_1^i,\dots,\mathbf{z}_\beta^i]$, where $\mathbf{z}_j^i\in\mathbb{R}^{D\times 1}$ and then $\mathbf{D}\boldsymbol{Z}^i = \mathbf{D}[\mathbf{z}_1^i,\dots,\mathbf{z}_\beta^i] = \hat{\mathbf{D}}\hat{\boldsymbol{Z}}^i = \hat{\mathbf{D}}[\hat{\mathbf{z}}_1^i,\dots,\hat{\mathbf{z}}_\beta^i]$ is equivalent to $\mathbf{D}\mathbf{z}_j^i = \hat{\mathbf{D}}\mathbf{z}_j^i$, which degenerates to the situation where $\beta=1$.

A simple case when $K=1$: To better understand Theorem 1 and prepare for the proof in full generality, let us start from a simple case when $K=1$. Denote $\mathbf{e}_i^L$ as a L-dimensional column vector that has one in its i-th coordinate and zeros elsewhere. For convenience, let $L=M/\alpha$. Now let us produce M block vectors

$$\mathbf{z}_j^i = (\mathbf{e}_i^L\otimes\mathbf{e}_j^\alpha), \quad i=1,\dots,L, \quad j=1,\dots,\alpha, \tag{3}$$

which denotes that its j-th coordinate in i-th block is one and zeros elsewhere, and $L\binom{\alpha}{2}$ block vectors $\mathbf{z}_{jk}^i = \mathbf{z}_{jk}^i + \mathbf{z}_{jk}^i$, for any i and $j\neq k$.

Now we claim that the uniqueness of BSDL in this simple case can be achieved by these $M+L\binom{\alpha}{2}$ block vectors, which is less than $K\binom{M/\alpha}{K}^2$ assuming $M\gg\alpha$.

Proof. There exists a matrix $\hat{\mathbf{D}}$ and 1-block-sparse vector $\hat{\mathbf{z}}_j^i = (\mathbf{e}_{\pi(i,j)}^L\otimes\boldsymbol{I}_\alpha)\boldsymbol{\lambda}_{ij}$, for some mapping $\pi:\{1,...,L\}\times\{1,...,\alpha\}\to\{1,...,L\}$ and $\boldsymbol{\lambda}_{ij}\in\mathbb{R}^\alpha$, such that

$$\mathbf{D}\mathbf{z}_j^i = \mathbf{D}(\mathbf{e}_i^L\otimes\mathbf{e}_j^\alpha) = \hat{\mathbf{D}}\hat{\mathbf{z}}_j^i = \hat{\mathbf{D}}(\mathbf{e}_{\pi(i,j)}^L\otimes I_\alpha)\boldsymbol{\lambda}_{ij}, \tag{4}$$

We claim that $\pi(i,j)$ is only dependent on i, not j. From Equation 4, we know that for any $j\neq k$, $\mathbf{D}\mathbf{z}_{jk}^i = \mathbf{D}(\mathbf{z}_j^i+\mathbf{z}_k^i) = \mathbf{D}\mathbf{z}_j^i+\mathbf{D}\mathbf{z}_k^i = \hat{\mathbf{D}}\hat{\mathbf{z}}_j^i+\hat{\mathbf{D}}\hat{\mathbf{z}}_k^i = \hat{\mathbf{D}}(\hat{\mathbf{z}}_j^i+\hat{\mathbf{z}}_k^i)$. Since $\mathbf{z}_{jk}^i$ is 1-block-sparse, this implies that $\hat{\mathbf{z}}_j^i+\hat{\mathbf{z}}_k^i$ should also be 1-block-sparse. Therefore $\pi(i,j)=\pi(i,k)$, that is, $\pi:\{1,...,L\}\to\{1,...,L\}$.

$$\mathbf{D}(\mathbf{e}_i^L\otimes\mathbf{e}_j^\alpha) = \hat{\mathbf{D}}(\mathbf{e}_{\pi(i)}^L\otimes\boldsymbol{I}_\alpha)\boldsymbol{\lambda}_{ij}. \tag{5}$$

Let us now prove that $\boldsymbol{\Lambda}_i = [\boldsymbol{\lambda}_{i1},\dots,\boldsymbol{\lambda}_{i\alpha}]$ is invertible. Let $\boldsymbol{Z}^i = [\mathbf{z}_1^i,\dots,\mathbf{z}_\alpha^i]$ and $\hat{\boldsymbol{Z}}^i = [\hat{\mathbf{z}}_1^i,\dots,\hat{\mathbf{z}}_\alpha^i]$. From Equation 5, it follows that $\mathbf{D}\boldsymbol{Z}^i = \mathbf{D}[\mathbf{z}_1^i,\dots,\mathbf{z}_\alpha^i] = \mathbf{D}[(\mathbf{e}_i^L\otimes\mathbf{e}_1^\alpha),...,(\mathbf{e}_i^L\otimes\mathbf{e}_\alpha^\alpha)] = \mathbf{D}(\mathbf{e}_i^L\otimes\boldsymbol{I}_\alpha)$, and $\mathbf{D}\boldsymbol{Z}^i = \hat{\mathbf{D}}\hat{\boldsymbol{Z}}^i = \hat{\mathbf{D}}(\mathbf{e}_{\pi(i)}^L\otimes\boldsymbol{I}_\alpha)\left[\boldsymbol{\lambda}_{i1},...,\boldsymbol{\lambda}_{i\alpha}\right] = \hat{\mathbf{D}}(\mathbf{e}_{\pi(i)}^L\otimes\boldsymbol{I}_\alpha)\boldsymbol{\Lambda}_i$. Therefore,

$$\mathbf{D}(\mathbf{e}_i^L\otimes\boldsymbol{I}_\alpha) = \hat{\mathbf{D}}(\mathbf{e}_{\pi(i)}^L\otimes\boldsymbol{I}_\alpha)\boldsymbol{\Lambda}_i. \tag{6}$$

Due to the fact that $\mathbf{D}$ satisfies the block spark condition, $\operatorname{rank}(\mathbf{D}(\mathbf{e}_i^L \otimes \boldsymbol{I}_\alpha)) = \alpha$. From Equation 6, $\operatorname{rank}(\hat{\mathbf{D}}(\mathbf{e}_{\pi(i)}^L \otimes \boldsymbol{I}_\alpha)\mathbf{\Lambda}_i) = \alpha$. We know that $\operatorname{rank}(\boldsymbol{X}\boldsymbol{Y}) \leq \min(\operatorname{rank}(\boldsymbol{X}), \operatorname{rank}(\boldsymbol{Y}))$, for any matrix $\boldsymbol{X}, \boldsymbol{Y}$. So $\operatorname{rank}(\mathbf{\Lambda}_i) \geq \alpha$. As $\mathbf{\Lambda}_i \in \mathbb{R}^{\alpha\times\alpha}$, $\operatorname{rank}(\mathbf{\Lambda}_i) = \alpha$.

Now, let us show π is necessarily injective. Suppose $\pi(i) = \pi(j)$, with $i \neq j$, then from Equation 6, $\mathbf{D}(\mathbf{e}_i^L \otimes \boldsymbol{I}_\alpha) = \hat{\mathbf{D}}(\mathbf{e}_{\pi(i)}^L \otimes \boldsymbol{I}_\alpha)\mathbf{\Lambda}_i = \hat{\mathbf{D}}(\mathbf{e}_{\pi(j)}^L \otimes \boldsymbol{I}_\alpha)\mathbf{\Lambda}_j\mathbf{\Lambda}_j^{-1}\mathbf{\Lambda}_i = \mathbf{D}(\mathbf{e}_j^L \otimes \boldsymbol{I}_\alpha)\mathbf{\Lambda}_j^{-1}\mathbf{\Lambda}_i$. Since $\mathbf{D}$ satisfies the block spark condition, which implies $\mathbf{D}$ can never map two different 1-block-sparse vectors to the same measurement, this is possible only if $i = j$. Thus, π is injective.

Let $\boldsymbol{P}_\pi$ and $\mathbf{D}$ be generated by

$$\boldsymbol{P}_\pi = \begin{bmatrix} \mathbf{e}_{\pi(1)}^L & \cdots & \mathbf{e}_{\pi(K)}^L \end{bmatrix}, \mathbf{\Lambda} = \begin{bmatrix} \mathbf{\Lambda}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{\Lambda}_L \end{bmatrix}. \tag{7}$$

Since π is injective, $\boldsymbol{P}_\pi$ is a permutation matrix. Let us stack Equation 6 from left-to-right on both sides, and it follows that on left sides,

$$[\mathbf{D}(\mathbf{e}_1^L \otimes \boldsymbol{I}_\alpha), \ldots, \mathbf{D}(\mathbf{e}_L^L \otimes \boldsymbol{I}_\alpha)] = \mathbf{D}, \tag{8}$$

and on right sides,

$$[\hat{\mathbf{D}}(\mathbf{e}_{\pi(1)}^L \otimes \boldsymbol{I}_\alpha)\mathbf{\Lambda}_1, \ldots, \hat{\mathbf{D}}(\mathbf{e}_{\pi(L)}^L \otimes \boldsymbol{I}_\alpha)\mathbf{\Lambda}_L] = \hat{\mathbf{D}}(\boldsymbol{P}_\pi \otimes \boldsymbol{I}_\alpha)\mathbf{\Lambda}. \tag{9}$$

Hence, we proved Theorem 1 for the simple case, where $K = 1$. □

Preparation: We use the same notation reported in [2]: Denote $[L]$ as the set $\{1, \ldots, L\}$ and $\binom{[L]}{K}$ as the K-element subset of $[L]$. Moreover, let the dictionary $\mathbf{D} = [\mathbf{D}_1, \ldots, \mathbf{D}_L]$ with $\mathbf{D}_i \in \mathbb{R}^{D\times\alpha}$, and denote $\operatorname{span}\{\mathbf{D}_\mathcal{S}\}$ as a subspace expanded by $\mathbf{D}_i$, $i \in \mathcal{S}$.

To prove Theorem 1 in general situations, we offer a lemma at first.

Lemma 1. *Suppose that $\mathbf{D}$ satisfies the block spark condition and*

$$\kappa : \binom{[L]}{K} \to \binom{[L]}{K} \tag{10}$$

is a mapping with the following property: for all $\mathcal{S} \in \binom{[L]}{K}$,

$$\operatorname{span}\{\mathbf{D}_\mathcal{S}\} = \operatorname{span}\{\hat{\mathbf{D}}_{\kappa(\mathcal{S})}\}. \tag{11}$$

Then, there exist a permutation matrix $\boldsymbol{P}_\kappa \in \mathbb{R}^{L\times L}$ and an invertible block diagonal matrix $\mathbf{\Lambda} \in \mathbb{R}^{M\times M}$ such that $\mathbf{D} = \hat{\mathbf{D}}(\boldsymbol{P}_\kappa \otimes \boldsymbol{I}_\alpha)\mathbf{\Lambda}$.

Proof. Here we demonstrate, through induction, that if our $K = 1$ case holds, therefore, $K > 1$ case should also hold. First, let us show function κ is injective. Suppose that $\mathcal{S}, \mathcal{S}' \in \binom{[L]}{K}$ are different and $\kappa(\mathcal{S}) = \kappa(\mathcal{S}')$ holds. Then by Equation 11, $\operatorname{span}\{\mathbf{D}_\mathcal{S}\} = \operatorname{span}\{\hat{\mathbf{D}}_{\kappa(\mathcal{S})}\} = \operatorname{span}\{\hat{\mathbf{D}}_{\kappa(\mathcal{S}')}\} = \operatorname{span}\{\mathbf{D}_{\mathcal{S}'}\}$. As $\mathbf{D}$ satisfies the block spark condition, every $K+1$ block columns of $\mathbf{D}$ are linearly independent. From Lemma 2 (see below), it turns out that $\mathcal{S} = \mathcal{S}'$, which implies κ is injective.

Denote $\eta = \kappa^{-1}$ as the inverse of κ. Fix $\mathcal{S} = \{i_1, ..., i_{K-1}\} \in \binom{[L]}{K-1}$, and set $\mathcal{S}_1 = \mathcal{S} \cup \{p\}$ and $\mathcal{S}_2 = \mathcal{S} \cup \{q\}$ for some fixed $p, q \notin \mathcal{S}$ with $p \neq q$. Since $K < L$, $L - (K-1) > 1$, thus, it is always possible to find such p and q. From Equation 11, we obtain:

$$\operatorname{span}\{\mathbf{D}_{\eta(\mathcal{S}_1)}\} = \operatorname{span}\{\hat{\mathbf{D}}_{\mathcal{S}_1}\}, \tag{12}$$

$$\operatorname{span}\{\mathbf{D}_{\eta(\mathcal{S}_2)}\} = \operatorname{span}\{\hat{\mathbf{D}}_{\mathcal{S}_2}\}. \tag{13}$$

Let us intersect Equation 12 and Equation 13, and from Lemma 3 (see below) it follows that $\operatorname{span}\{\hat{\mathbf{D}}_{\mathcal{S}_1}\} \cap \operatorname{span}\{\hat{\mathbf{D}}_{\mathcal{S}_2}\} = \operatorname{span}\{\mathbf{D}_{\eta(\mathcal{S}_1)\cap\eta(\mathcal{S}_2)}\}$. Since $\operatorname{span}\{\hat{\mathbf{D}}_\mathcal{S}\} \subseteq \operatorname{span}\{\hat{\mathbf{D}}_{\mathcal{S}_1}\} \cap \operatorname{span}\{\hat{\mathbf{D}}_{\mathcal{S}_2}\}$, it follows that $\operatorname{span}\{\hat{\mathbf{D}}_\mathcal{S}\} \subseteq \operatorname{span}\{\mathbf{D}_{\eta(\mathcal{S}_1)\cap\eta(\mathcal{S}_2)}\}$. The number of the elements in $\eta(\mathcal{S}_1) \cap \eta(\mathcal{S}_2)$ is $K-1$, since $\eta(p) \neq \eta(q)$, with $p \neq q$, by injectivity of η. Moreover the number of the elements in $\mathcal{S}$ is also $K-1$, which implies that

$$\operatorname{span}\{\hat{\mathbf{D}}_\mathcal{S}\} = \operatorname{span}\{\mathbf{D}_{\eta(\mathcal{S}_1)\cap\eta(\mathcal{S}_2)}\}. \tag{14}$$

The association $\mathcal{S} \to \eta(\mathcal{S}_1) \cap \eta(\mathcal{S}_2)$ from Equation 14 defines a function $\sigma : \binom{[L]}{K-1} \to \binom{[L]}{K-1}$, with property that $\operatorname{span}\{\hat{\mathbf{D}}_\mathcal{S}\} = \operatorname{span}\{\mathbf{D}_{\sigma(\mathcal{S})}\}$.

Finally, let's show that σ is injective. Suppose $\mathcal{S}, \mathcal{S}' \in \binom{[L]}{K-1}$, and $\sigma(\mathcal{S}) = \sigma(\mathcal{S}')$, it follows that $\operatorname{span}\{\hat{\mathbf{D}}_\mathcal{S}\} = \operatorname{span}\{\mathbf{D}_{\sigma(\mathcal{S})}\} = \operatorname{span}\{\mathbf{D}_{\sigma(\mathcal{S}')}\} = \operatorname{span}\{\hat{\mathbf{D}}_{\mathcal{S}'}\}$. As every K block columns of $\mathbf{D}$ are linear independent, and κ is injective, every K block columns of $\hat{\mathbf{D}}$ are also linear independent. From Lemma 2, it follows that $\mathcal{S} = \mathcal{S}'$, which implies σ is injective. Hence, let $\xi = \sigma^{-1}$, with properties: for all $\mathcal{S} \in \binom{[L]}{K-1}$, $\operatorname{span}\{\mathbf{D}_\mathcal{S}\} = \operatorname{span}\{\hat{\mathbf{D}}_{\xi(\mathcal{S})}\}$. □

Lemma 2. *If any set of $K+1$ block columns of matrix $\mathbf{D} = [\mathbf{D}_1, \ldots, \mathbf{D}_L]$ are linear independent, then for $\mathcal{S}, \mathcal{S}' \in \binom{[L]}{K}$,*

$$\operatorname{span}\{\mathbf{D}_\mathcal{S}\} = \operatorname{span}\{\mathbf{D}_{\mathcal{S}'}\} \quad \Rightarrow \quad \mathcal{S} = \mathcal{S}'. \tag{15}$$

Proof. Suppose that $\mathcal{S} \neq \mathcal{S}' \in \binom{[L]}{K}$ satisfying $\operatorname{span}\{\mathbf{D}_\mathcal{S}\} = \operatorname{span}\{\mathbf{D}_{\mathcal{S}'}\}$. Then without loss of generality, there is an $i \in \mathcal{S}$ with $i \notin \mathcal{S}'$, but atoms $\mathbf{D}_i \in \operatorname{span}\{\mathbf{D}_{\mathcal{S}'}\}$, which implies that the $K+1$ block columns indexed by $\mathcal{S}' \cup \{i\}$ are not linear independent, a contradiction to the assumption. □

Lemma 3. *If matrix $\mathbf{D}$ satisfies the block spark condition, then for $\mathcal{S}, \mathcal{S}' \in \binom{[L]}{K}$,*

$$\operatorname{span}\{\mathbf{D}_{\mathcal{S}\cap\mathcal{S}'}\} = \operatorname{span}\{\mathbf{D}_\mathcal{S}\} \cap \operatorname{span}\{\mathbf{D}_{\mathcal{S}'}\}. \tag{16}$$

Proof. The inclusion "$\subseteq$" is trivial, so let us prove "$\supseteq$". Suppose a block vector $\mathbf{x} \in \mathrm{span}\{\mathbf{D}_\mathcal{S}\} \cap \mathrm{span}\{\mathbf{D}_{\mathcal{S}_2}\}$. Express $\mathbf{x}$ as a linear combination of K atoms of $\mathbf{D}$ indexed by $\mathcal{S}$ and, separately, as a combination of K atoms of $\mathbf{D}$ indexed by $\mathcal{S}'$. By the block spark condition, these linear combinations must be identical. In particular, $\mathbf{x}$ was expressed as a linear combination of atoms of $\mathbf{D}$ indexed by $\mathcal{S} \cap \mathcal{S}'$, and thus is in $\mathrm{span}\{\mathbf{D}_{\mathcal{S}\cap\mathcal{S}'}\}$ □

Proof of Theorem 1 when $\beta = 1$: First, we produce a set of $N = K\binom{M/\alpha}{K}^2$ vectors $\boldsymbol{s}_i \in \mathbb{R}^{\alpha K}$ in general linear position (*i.e.* any subset of K of them are linearly independent). One possible strategy is to produce a "Vandermonde" matrix [3]. Next, we form K-block-sparse vectors $\mathbf{z}_1, ..., \mathbf{z}_N$ by taking $\boldsymbol{s}_i$ for the support value of $\mathbf{z}_i$ where each possible support set is represented $K\binom{M/\alpha}{K}$ times. We claim that these $\mathbf{z}_i$ always guarantee the uniqueness of BSDL.

Proof. Suppose there exists an alternate dictionary $\hat{\mathbf{D}}$ and a set of K-block-sparse vectors $\hat{\mathbf{z}}_1, ..., \hat{\mathbf{z}}_N$ such that $\mathbf{D}\mathbf{z}_i = \mathbf{x}_i = \hat{\mathbf{D}}\hat{\mathbf{z}}_i$. As there are $K\binom{M/\alpha}{K}$ $\mathbf{x}_i$ for each support indexed by $\mathcal{S}$, the "pigeon-hole principle"[1] implies that there are at least K vectors $\hat{\mathbf{z}}_{i_1}, ..., \hat{\mathbf{z}}_{i_K}$ using the same support $\mathcal{S}'$. Thus, $\mathrm{span}\{\mathbf{x}_{i_1}, ..., \mathbf{x}_{i_K}\} \subseteq \mathrm{span}\{\hat{\mathbf{D}}_{\mathcal{S}'}\}$. By the general linear position and the block spark condition, $\mathrm{span}\{\mathbf{x}_{i_1}, ..., \mathbf{x}_{i_K}\} = \mathrm{span}\{\mathbf{D}_\mathcal{S}\}$. Therefore $\mathrm{span}\{\mathbf{D}_\mathcal{S}\} \subseteq \mathrm{span}\{\hat{\mathbf{D}}_{\mathcal{S}'}\}$. As the dimension of $\mathrm{span}\{\hat{\mathbf{D}}_{\mathcal{S}'}\}$ is less and equal to K, $\mathrm{span}\{\mathbf{D}_\mathcal{S}\} = \mathrm{span}\{\hat{\mathbf{D}}_{\mathcal{S}'}\}$.

By Lemma 1, Theorem 1 is proved. □

Discussion: A lower $N = (K+1)\binom{M}{K}$ is offered by Hillar *et al.*'s probabilistic theorems in [2] saying that if $K+1$ K-sparse vector $\mathbf{z}_i$ are randomly drawn from each support set, and $\mathbf{D}$ satisfies the spark condition, then $\boldsymbol{X}$ has a unique SDL with probability one. We hypothesize that a lower $N = (K+1)\binom{M/\alpha}{K}$ is also enough for the uniqueness of BSDL to hold with probability one, which will be a focus of future work.

References

[1] R. A. Brualdi. *Introductory combinatorics*. New York, 1992. 3

[2] C. Hillar and F. T. Sommer. When can dictionary learning uniquely recover sparse data from subsamples? In *IEEE Transactions on Information Theory*, 2015. 2, 3

[3] L. R. Turner. Inverse of the vandermonde matrix with applications. 1966. 3

[1] The pigeon-hole principle states that if n items are put into m containers, with $n > m$, then at least one container must contain more than one item [1].