

A DIMENSION REDUCTION SCHEME FOR THE COMPUTATION OF OPTIMAL UNIONS OF SUBSPACES

A. ALDROUBI, M. ANASTASIO, C. CABRELLI, AND U. MOLTER

ABSTRACT. Given a set of points \mathcal{F} in a high dimensional space, the problem of finding a union of subspaces $\cup_i V_i \subseteq \mathbb{R}^N$ that best explains the data \mathcal{F} increases dramatically with the dimension of \mathbb{R}^N . In this article, we study a class of transformations that map the problem into another one in lower dimension. We use the best model in the low dimensional space to approximate the best solution in the original high dimensional space. We then estimate the error produced between this solution and the optimal solution in the high dimensional space.

1. INTRODUCTION

Given a set of vectors (points) $\mathcal{F} = \{f_1, \dots, f_m\}$ in a Hilbert space \mathcal{H} (finite or infinite dimensional), the problem of finding a union of subspaces $\cup_i V_i \subseteq \mathcal{H}$ that best explains the data \mathcal{F} has applications to mathematics and engineering [9, 11, 12, 13, 14, 15, 6, 18]. The subspaces V_i allowed in the model are often constrained. For example the subspaces V_i may be constrained to belong to a family of closed subspaces \mathcal{C} [4]. A typical example for $\mathcal{H} = \mathbb{R}^N$ is when \mathcal{C} is the set of subspaces of dimension $k \ll N$. If \mathcal{C} satisfies the so called Minimum Subspace Approximation Property (MSAP), an optimal solution to the non-linear subspace modeling problem that best fit the data exists, and algorithms to find these subspaces were developed [4]. Necessary and sufficient conditions for \mathcal{C} to satisfy the MSAP are obtained in [5].

In some applications the model is a finite union of subspaces and \mathcal{H} is finite dimensional. Once the model is found, the given data points can be clustered and classified according to their distances from the subspaces, giving rise to the so called *subspace clustering problem* (see e.g., [9] and the references therein). Thus a dual problem is to first find a “best partition” of the data. Once this partition is obtained, the associated optimal subspaces can be

Date: November 21, 2018.

2000 Mathematics Subject Classification. Primary 94A12, 94A20; Secondary 15A52, 65F15, 15A18.

Key words and phrases. Sparsity, projective clustering, dimensionality reduction, random matrices, concentration inequalities.

The research of A. Aldroubi is supported in part by NSF Grant DMS-0807464. M. Anastasio, C. Cabrelli and U. Molter are partially supported by Grants UBACyT X149 and X028 (UBA), PICT 2006-00177 (ANPCyT), and PIP 2008-398 (CONICET)..

easily found. In any case, the search for an optimal partition or optimal subspaces usually involves heavy computations that dramatically increases with the dimensionality of \mathcal{H} . Thus one important feature is to map the data into a lower dimensional space, and solve the transformed problem in this lower dimensional space. If the mapping is chosen appropriately, the original problem can be solved exactly or approximately using the solution of the transformed data.

In this article, we concentrate on the non-linear subspace modeling problem when the model is a finite union of subspaces of \mathbb{R}^N of dimension $k \ll N$. Our goal is to find transformations from a high dimensional space to lower dimensional spaces with the aim of solving the subspace modeling problem using the low dimensional transformed data. We find the optimal data partition for the transformed data and use this partition for the original data to obtain the subspace model associated to this partition. We then estimate the error between the model thus found and the optimal subspaces model for the original data.

2. PRELIMINARIES

Since one of our goals is to model a set of data by a union of subspaces, we first provide a measure of how well a given set of data can be modeled by a union of subspaces.

We will assume in this article that the data belongs to the finite dimensional space \mathbb{R}^N . There is no loss of generality in doing that, since it is easy to see that the subspaces of any optimal solution belong to the span of the data, which is a finite dimensional subspace of our (possible infinite dimensional) Hilbert space. (see [3], Lemma 4.2). So we can assume that the initial Hilbert space is the span of the data.

Definition 2.1. Given a set of vectors $\mathcal{F} = \{f_1, \dots, f_m\}$ in \mathbb{R}^N , a real number $\rho \geq 0$ and positive integers $l, k < N$ we will say that the data \mathcal{F} is (l, k, ρ) -sparse if there exist subspaces V_1, \dots, V_l of \mathbb{R}^N with $\dim(V_i) \leq k$ for $i = 1, \dots, l$, such that

$$e(\mathcal{F}, \{V_1, \dots, V_l\}) = \sum_{i=1}^m \min_{1 \leq j \leq l} d^2(f_i, V_j) \leq \rho,$$

where d stands for the euclidean distance in \mathbb{R}^N .

When \mathcal{F} is $(l, k, 0)$ -sparse, we will simply say that \mathcal{F} is (l, k) -sparse.

Note that if \mathcal{F} is (l, k) -sparse, there exist l subspaces V_1, \dots, V_l of dimension at most k , such that

$$\mathcal{F} \subseteq \cup_{i=1}^l V_i.$$

For the general case $\rho > 0$, the (l, k, ρ) -sparsity of the data implies that \mathcal{F} can be partitioned into a small number of subsets, in such a way that each subset belongs to or is at no more than ρ -distance from a low dimensional

subspace. The collection of these subspaces provides an optimal non-linear sparse model for the data.

Observe that if the data \mathcal{F} is (l, k, ρ) -sparse, a model which verifies Definition 2.1 provides a dictionary of length not bigger than lk (and in most cases much smaller) in which our data can be represented using at most k atoms with an error smaller than ρ .

More precisely, let $\{V_1, \dots, V_l\}$ be a collection of subspaces which satisfies Definition 2.1 and D a set of vectors from $\bigcup_j V_j$ that is minimal with the property that its span contains $\bigcup_j V_j$. Then for each $f \in \mathcal{F}$ there exists $\Lambda \subseteq D$ with $\#\Lambda \leq k$ such that

$$\|f - \sum_{g \in \Lambda} \alpha_g g\|_2^2 \leq \rho, \quad \text{for some scalars } \alpha_g.$$

In [4] the authors studied the problem of finding, for each given set of pairs (l, k) , the minimum ρ -sparsity value of the data. They also provided an algorithm for finding the optimal value of ρ , as well as the optimal subspaces associated with ρ and the corresponding optimal partition of the data. Specifically, denote by \mathcal{B} the collection of *bundles* of subspaces of \mathbb{R}^N ,

$$\mathcal{B} = \{B = \{V_1, \dots, V_l\} : \dim(V_i) \leq k, i = 1, \dots, l\},$$

and for $\mathcal{F} = \{f_1, \dots, f_m\}$ a finite subset of \mathbb{R}^N , define

$$e_0(\mathcal{F}) := \inf\{e(\mathcal{F}, B) : B \in \mathcal{B}\}. \quad (1)$$

As a special case of a general theorem in [4] we obtain the next theorem.

Theorem 2.2. *Let $\mathcal{F} = \{f_1, \dots, f_m\}$ be vectors in \mathbb{R}^N , and let l and k be given ($l < m$, $k < N$), then there exists a bundle $B_0 = \{V_1^0, \dots, V_l^0\} \in \mathcal{B}$ such that*

$$e(\mathcal{F}, B_0) = e_0(\mathcal{F}) = \inf\{e(\mathcal{F}, B) : B \in \mathcal{B}\}. \quad (2)$$

Any bundle $B_0 \in \mathcal{B}$ satisfying (2) will be called an optimal bundle for \mathcal{F} .

The following relations between partitions of the indices $\{1, \dots, m\}$ and bundles will be relevant for our analysis.

We will denote by $\mathbf{\Pi}_l(\{1, \dots, m\})$ the set of all l -sequences $\mathbf{S} = \{S_1, \dots, S_l\}$ of subsets of $\{1, \dots, m\}$ satisfying the property that for all $1 \leq i, j \leq l$,

$$\bigcup_{r=1}^l S_r = \{1, \dots, m\} \quad \text{and} \quad S_i \cap S_j = \emptyset \quad \text{for } i \neq j.$$

We want to emphasize that this definition does not exclude the case when some of the S_i are the empty set. By abuse of notation, we will still call the elements of $\mathbf{\Pi}_l(\{1, \dots, m\})$ *partitions* of $\{1, \dots, m\}$.

Definition 2.3. Given a bundle $B = \{V_1, \dots, V_l\} \in \mathcal{B}$, we can split the set $\{1, \dots, m\}$ into a partition $\mathbf{S} = \{S_1, \dots, S_l\} \in \mathbf{\Pi}_l(\{1, \dots, m\})$ with respect to that bundle, by grouping together into S_i the indices of the vectors in \mathcal{F} that are closer to a given subspace V_i than to any other subspace V_j , $j \neq i$.

Thus, the partitions generated by B are defined by $\mathbf{S} = \{S_1, \dots, S_l\} \in \Pi_l(\{1, \dots, m\})$, where

$$j \in S_i \quad \text{if and only if} \quad d(f_j, V_i) \leq d(f_j, V_h), \quad \forall h = 1, \dots, l.$$

We can also associate to a given partition $\mathbf{S} \in \Pi_l$ the bundles in \mathcal{B} as follows:

Definition 2.4. Given a partition $\mathbf{S} = \{S_1, \dots, S_l\} \in \Pi_l$, a bundle $B = \{V_1, \dots, V_l\} \in \mathcal{B}$ is generated by \mathbf{S} if and only if for every $i = 1, \dots, l$,

$$\sum_{j \in S_i} d^2(f_j, V_i) \leq \sum_{j \in S_i} d^2(f_j, W) \quad \text{for all subspaces } W \text{ such that } \dim(W) \leq k.$$

In this way, for a given data set \mathcal{F} , every bundle has a set of associated partitions (those that are generated by the bundle) and every partition has a set of associated bundles (those that are generated by the partition). Note however, that the fact that \mathbf{S} is generated by B does not imply that B is generated by \mathbf{S} , and vice versa. However, if B_0 is an optimal bundle that solves the problem for the data \mathcal{F} as in Theorem 2.2, then in this case, the partition \mathbf{S}_0 generated by B_0 also generates B_0 . On the other hand not every pair (B, \mathbf{S}) with this property produces the minimal error $e_0(\mathcal{F})$.

Here and subsequently, the partition \mathbf{S}_0 generated by the optimal bundle B_0 will be called an optimal partition for \mathcal{F} .

If M is a set of data and V is a subspace of \mathbb{R}^N , we will denote by $E(M, V)$ the mean square error of the data M to the subspace V , i.e.

$$E(M, V) = \sum_{f \in M} d^2(f, V). \quad (3)$$

3. MAIN RESULTS

The problem of finding the optimal union of subspaces that best models a given set of data \mathcal{F} when the dimension of the ambient space N is large is computationally expensive. When the dimension k of the subspaces is considerably smaller than N , it is natural to map the data onto a lower-dimensional subspace, solve an associated problem in the lower dimensional space and map the solution back into the original space. Specifically, given the data set $\mathcal{F} = \{f_1, \dots, f_m\} \subseteq \mathbb{R}^N$ which is (l, k, ρ) -sparse and a sampling matrix $A \in \mathbb{R}^{r \times N}$, with $r \ll N$, find the optimal partition of the sampled data $\mathcal{F}' := A(\mathcal{F}) = \{Af_1, \dots, Af_m\} \subseteq \mathbb{R}^r$, and use this partition to find an approximate solution to the optimal model for \mathcal{F} .

3.1. Dimensionality reduction: The ideal case $\rho = 0$. In this section we will assume that the data $\mathcal{F} = \{f_1, \dots, f_m\} \subseteq \mathbb{R}^N$ is (l, k) -sparse, i.e., there exist l subspaces of dimension at most k such that \mathcal{F} lies in the union of these subspaces. For this ideal case, we will show that we can always recover the optimal solution to the original problem from the optimal solution to the problem in the low dimensional space as long as the low dimensional space has dimension $r > k$.

We will begin with the proof that for any sampling matrix $A \in \mathbb{R}^{r \times N}$, the measurements $\mathcal{F}' = A(\mathcal{F})$ are (l, k) -sparse in \mathbb{R}^r .

Lemma 3.1. *Assume the data $\mathcal{F} = \{f_1, \dots, f_m\} \subseteq \mathbb{R}^N$ is (l, k) -sparse and let $A \in \mathbb{R}^{r \times N}$. Then $\mathcal{F}' := A(\mathcal{F}) = \{Af_1, \dots, Af_m\} \subseteq \mathbb{R}^r$ is (l, k) -sparse.*

Proof. Let V_1^0, \dots, V_l^0 be optimal spaces for \mathcal{F} . Since

$$\dim(A(V_i^0)) \leq \dim(V_i^0) \leq k \quad \forall 1 \leq i \leq l,$$

and

$$\mathcal{F}' \subseteq \bigcup_{i=1}^l A(V_i^0),$$

it follows that $W := \{A(V_1^0), \dots, A(V_l^0)\}$ is an optimal bundle for \mathcal{F}' and $e(\mathcal{F}', W) = 0$. □

Let $\mathcal{F} = \{f_1, \dots, f_m\} \subseteq \mathbb{R}^N$ be (l, k) -sparse and $A \in \mathbb{R}^{r \times N}$. By Lemma 3.1, \mathcal{F}' is (l, k) -sparse. Thus, there exists an optimal partition $\mathbf{S} = \{S_1, \dots, S_l\}$ for \mathcal{F}' in $\mathbf{\Pi}_l(\{1, \dots, m\})$, such that

$$\mathcal{F}' \subseteq \bigcup_{i=1}^l W_i,$$

where $W_i := \text{span}\{Af_j\}_{j \in S_i}$ and $\dim(W_i) \leq k$. Note that $\{W_1, \dots, W_l\}$ is an optimal bundle for \mathcal{F}' .

We can define the bundle $B_{\mathbf{S}} = \{V_1, \dots, V_l\}$ by

$$V_i := \text{span}\{f_j\}_{j \in S_i}, \quad \forall 1 \leq i \leq l. \quad (4)$$

Since $\mathbf{S} \in \mathbf{\Pi}_l(\{1, \dots, m\})$, we have that

$$\mathcal{F} \subseteq \bigcup_{i=1}^l V_i.$$

Thus, the bundle $B_{\mathbf{S}}$ will be optimal for \mathcal{F} if $\dim(V_i) \leq k$, $\forall 1 \leq i \leq l$. The above discussion suggests the following definition:

Definition 3.2. Let $\mathcal{F} = \{f_1, \dots, f_m\} \subseteq \mathbb{R}^N$ be (l, k) -sparse. We will call a matrix $A \in \mathbb{R}^{r \times N}$ *admissible* for \mathcal{F} if for every optimal partition \mathbf{S} for \mathcal{F}' , the bundle $B_{\mathbf{S}}$ defined by (4) is optimal for \mathcal{F} .

The next proposition states that almost all $A \in \mathbb{R}^{r \times N}$ are admissible for \mathcal{F} .

The Lebesgue measure of a set $E \subseteq \mathbb{R}^q$ will be denoted by $|E|$.

Proposition 3.3. *Assume the data $\mathcal{F} = \{f_1, \dots, f_m\} \subseteq \mathbb{R}^N$ is (l, k) -sparse and let $r > k$. Then, almost all $A \in \mathbb{R}^{r \times N}$ are admissible for \mathcal{F} .*

Proof. If a matrix $A \in \mathbb{R}^{r \times N}$ is not admissible, there exists an optimal partition $\mathbf{S} \in \Pi_l$ for \mathcal{F}' such that the bundle $B_{\mathbf{S}} = \{V_1, \dots, V_l\}$ is not optimal for \mathcal{F} .

Let \mathcal{D}_k be the set of all the subspaces V in \mathbb{R}^N of dimension bigger than k , such that $V = \text{span}\{f_j\}_{j \in S}$ with $S \subseteq \{1, \dots, m\}$.

Thus, we have that the set of all the matrices of $\mathbb{R}^{r \times N}$ which are not admissible for \mathcal{F} is contained in the set

$$\bigcup_{V \in \mathcal{D}_k} \{A \in \mathbb{R}^{r \times N} : \dim(A(V)) \leq k\}.$$

Note that the set \mathcal{D}_k is finite, since there are finitely many subsets of $\{1, \dots, m\}$. Therefore, the proof of the proposition is complete by showing that for a fixed subspace $V \subseteq \mathbb{R}^N$, such that $\dim(V) > k$, it is true that

$$|\{A \in \mathbb{R}^{r \times N} : \dim(A(V)) \leq k\}| = 0. \quad (5)$$

Let then V be a subspace such that $\dim(V) = t > k$. Given $\{v_1, \dots, v_t\}$ a basis for V , by abuse of notation, we continue to write V for the matrix in $\mathbb{R}^{N \times t}$ with vectors v_i as columns. Thus, proving (5) is equivalent to proving that

$$|\{A \in \mathbb{R}^{r \times N} : \text{rank}(AV) \leq k\}| = 0. \quad (6)$$

As $\min\{r, t\} > k$, the set $\{A \in \mathbb{R}^{r \times N} : \text{rank}(AV) \leq k\}$ is included in

$$\{A \in \mathbb{R}^{r \times N} : \det(V^* A^* AV) = 0\}. \quad (7)$$

Since $\det(V^* A^* AV)$ is a non-trivial polynomial in the $r \times N$ coefficients of A , the set (7) has Lebesgue measure zero. Hence, (6) follows. \square

3.2. Dimensionality reduction: The non-ideal case $\rho > 0$. Even if a set of data is drawn from a union of subspaces, in practice it is often corrupted by noise. Thus, in general $\rho > 0$, and our goal is to estimate the error produced when we solve the associated problem in the lower dimensional space and map the solution back into the original space.

Intuitively, if $A \in \mathbb{R}^{r \times N}$ is an arbitrary matrix, the set $\mathcal{F}' = A\mathcal{F}$ will preserve the original sparsity only if the matrix A does not change the geometry of the data in an essential way. One can think that in the *ideal* case, since the data is sparse, it actually lies in an union of low dimensional subspaces (which is a very thin set in the ambient space).

However, when the data is not 0-sparse, but only ρ -sparse with $\rho > 0$, the optimal subspaces plus the data do not lie in a thin set. This is the main obstacle in order to obtain an analogous result as in the ideal case.

Far from having the result that for *almost any* matrix A the geometry of the data will be preserved, we have the Johnson-Lindenstrauss lemma, that guaranties - for a given data set - the existence of *one* such matrix A .

In what follows, we will use random matrices to obtain positive results for the $\rho > 0$ case.

Let (Ω, \Pr) be a probability measure space. Given $r, N \in \mathbb{N}$, a random matrix $A_\omega \in \mathbb{R}^{r \times N}$ is a matrix with entries $(A_\omega)_{i,j} = a_{i,j}(\omega)$, where $\{a_{i,j}\}$ are independent and identically distributed random variables for every $1 \leq i \leq r$ and $1 \leq j \leq N$.

Definition 3.4. We say that a random matrix $A_\omega \in \mathbb{R}^{r \times N}$ satisfies the concentration inequality if for every $0 < \varepsilon < 1$, there exists $c_0 = c_0(\varepsilon) > 0$ (independent of r, N) such that for any $x \in \mathbb{R}^N$,

$$\Pr\left((1 - \varepsilon)\|x\|_2^2 \leq \|A_\omega x\|_2^2 \leq (1 + \varepsilon)\|x\|_2^2\right) \geq 1 - 2e^{-rc_0} \quad (8)$$

Such matrices are easy to come by as the next proposition shows [1].

Proposition 3.5. Let $A_\omega \in \mathbb{R}^{r \times N}$ be a random matrix whose entries are chosen independently from either $\mathcal{N}(0, \frac{1}{r})$ or $\{\frac{-1}{\sqrt{r}}, \frac{1}{\sqrt{r}}\}$ Bernoulli. Then A_ω satisfies (8) with $c_0(\varepsilon) = \frac{\varepsilon^2}{4} - \frac{\varepsilon^3}{6}$.

By using random matrices A_ω satisfying (8) to produce the lower dimensional data set \mathcal{F}' , we will be able to recover with high probability an optimal partition for \mathcal{F} using the optimal partition of \mathcal{F}' .

Below we will state the main results of Section 3.2 and we will give their proofs in Section 4.

Note that by Lemma 3.1, if $\mathcal{F} = \{f_1, \dots, f_m\} \subseteq \mathbb{R}^N$ is $(l, k, 0)$ -sparse, then $A_\omega(\mathcal{F})$ is $(l, k, 0)$ -sparse for all $\omega \in \Omega$. The following proposition is a generalization of Lemma 3.1 to the case where \mathcal{F} is (l, k, ρ) -sparse with $\rho > 0$.

Proposition 3.6. Assume the data $\mathcal{F} = \{f_1, \dots, f_m\} \subseteq \mathbb{R}^N$ is (l, k, ρ) -sparse with $\rho > 0$. If $A_\omega \in \mathbb{R}^{r \times N}$ is a random matrix which satisfies (8), then $A_\omega \mathcal{F}$ is $(l, k, (1 + \varepsilon)\rho)$ -sparse with probability at least $1 - 2me^{-rc_0}$.

Hence if the data is mapped with a random matrix which satisfies the concentration inequality, then with high probability, the sparsity of the transformed data is close to the sparsity of the original data. Further, as the following theorem shows, we obtain an estimation for the error between \mathcal{F} and the bundle generated by the optimal partition for $\mathcal{F}' = A_\omega \mathcal{F}$.

Note that, given a constant $\alpha > 0$, the scaled data $\alpha \mathcal{F} = \{\alpha f_1, \dots, \alpha f_m\}$ satisfies that $e(\alpha \mathcal{F}, B) = \alpha^2 e(\mathcal{F}, B)$ for any bundle B . So, an optimal bundle for \mathcal{F} is optimal for $\alpha \mathcal{F}$, and vice versa. Therefore, we can assume that the data $\mathcal{F} = \{f_1, \dots, f_m\}$ is *normalized*, that is, the matrix $M \in \mathbb{R}^{N \times m}$ which has the vectors $\{f_1, \dots, f_m\}$ as columns has unitary Frobenius norm. Recall that the Frobenius norm of a matrix $M \in \mathbb{R}^{N \times m}$ is defined by

$$\|M\|^2 := \sum_{i=1}^N \sum_{j=1}^m M_{i,j}^2, \quad (9)$$

where $M_{i,j}$ are the coefficients of M .

Theorem 3.7. *Let $\mathcal{F} = \{f_1, \dots, f_m\} \subseteq \mathbb{R}^N$ be a normalized data set and $0 < \varepsilon < 1$. Assume that $A_\omega \in \mathbb{R}^{r \times N}$ is a random matrix satisfying (8) and \mathbf{S}_ω is an optimal partition for $\mathcal{F}' = A_\omega \mathcal{F}$ in \mathbb{R}^r . If B_ω is a bundle generated by the partition \mathbf{S}_ω and the data \mathcal{F} in \mathbb{R}^N as in Definition 2.3, then with probability exceeding $1 - (2m^2 + 4m)e^{-rc_0}$, we have*

$$e(\mathcal{F}, B_\omega) \leq (1 + \varepsilon)e_0(\mathcal{F}) + \varepsilon c_1, \quad (10)$$

where $c_1 = (l(d - k))^{1/2}$ and $d = \text{rank}(\mathcal{F})$.

Finally, we can use this theorem to show that the set of matrices which are η -admissible (see definition below) is large.

The following definition generalizes Definition 3.2 to the ρ -sparse setting, with $\rho > 0$.

Definition 3.8. Assume $\mathcal{F} = \{f_1, \dots, f_m\} \subseteq \mathbb{R}^N$ is (l, k, ρ) -sparse and let $0 < \eta < 1$. We will say that a matrix $A \in \mathbb{R}^{r \times N}$ is η -admissible for \mathcal{F} if for any optimal partition \mathbf{S} for $\mathcal{F}' = A\mathcal{F}$ in \mathbb{R}^r , the bundle $B_{\mathbf{S}}$ generated by \mathbf{S} and \mathcal{F} in \mathbb{R}^N , satisfies

$$e(\mathcal{F}, B_{\mathbf{S}}) \leq \rho + \eta.$$

We have the following generalization of Proposition 3.3, which provides an estimate on the size of the set of η -admissible matrices.

Corollary 3.9. *Let $\mathcal{F} = \{f_1, \dots, f_m\} \subseteq \mathbb{R}^N$ be a normalized data set and $0 < \eta < 1$. Assume that $A_\omega \in \mathbb{R}^{r \times N}$ is a random matrix which satisfies property (8) for $\varepsilon = \eta (1 + \sqrt{l(d - k)})^{-1}$. Then A_ω is η -admissible for \mathcal{F} with probability at least $1 - (2m^2 + 4m)e^{-rc_0(\varepsilon)}$.*

Proof. Using the fact that $e_0(\mathcal{F}) \leq E(\mathcal{F}, \{0\}) = \|\mathcal{F}\|^2 = 1$, we conclude from Theorem 3.7 that

$$\Pr\left(e(\mathcal{F}, B_\omega) \leq e_0(\mathcal{F}) + \varepsilon(1 + c_1)\right) \geq 1 - c_2 e^{-rc_0(\varepsilon)}, \quad (11)$$

where $c_1 = (l(d - k))^{1/2}$, $d = \text{rank}(\mathcal{F})$, and $c_2 = 2m^2 + 4m$. That is,

$$\Pr\left(e(\mathcal{F}, B_\omega) \leq e_0(\mathcal{F}) + \eta\right) \geq 1 - (2m^2 + 4m)e^{-rc_0(\varepsilon)}.$$

□

As a consequence of the previous corollary, we have a bound on the dimension of the lower dimensional space to obtain a bundle which produces an error at η -distance of the minimal error with high probability.

Now, using that $c_0(\varepsilon) \geq \frac{\varepsilon^2}{12}$ for random matrices with gaussian or Bernoulli entries (see Proposition 3.5), from Theorem 3.7 we obtain the following corollary.

Corollary 3.10. *Let $\eta, \delta \in (0, 1)$, be given. Assume that $A_\omega \in \mathbb{R}^{r \times N}$ is a random matrix whose entries are as in Proposition 3.5.*

Then for every r satisfying,

$$r \geq \frac{12(1 + \sqrt{l(d-k)})^2}{\eta^2} \ln \left(\frac{2m^2 + 4m}{\delta} \right)$$

with probability at least $1 - \delta$ we have that

$$e(\mathcal{F}, B_\omega) \leq e_0(\mathcal{F}) + \eta.$$

We want to remark here that the results of subsection 3.2 are valid for any probability distribution that satisfies the concentration inequality (8). The bound on the error is still valid for $\rho = 0$. However in that case we were able to obtain sharp results.

4. PROOFS

4.1. Background and supporting results. Before proving the results of the previous section we need several known theorems, lemmas, and propositions below.

Given $M \in \mathbb{R}^{m \times m}$ a Hermitian matrix, let $\lambda_1(M) \geq \lambda_2(M) \geq \dots \geq \lambda_m(M)$ be its eigenvalues and $s_1(M) \geq s_2(M) \geq \dots \geq s_m(M) \geq 0$ be its singular values.

Recall that the Frobenius norm defined in (9) satisfies that

$$\|M\|^2 = \sum_{1 \leq i, j \leq m} M_{i,j}^2 = \sum_{i=1}^m s_i^2(M),$$

where $M_{i,j}$ are the coefficients of M .

Given $x \in \mathbb{R}^N$, we write $\|x\|_2$ for the ℓ^2 norm of x in \mathbb{R}^N .

Theorem 4.1. [8, Theorem III.4.1]

Let $A, B \in \mathbb{R}^{m \times m}$ be Hermitian matrices. Then for any choice of indices $1 \leq i_1 < i_2 < \dots < i_k \leq m$,

$$\sum_{j=1}^k (\lambda_{i_j}(A) - \lambda_{i_j}(B)) \leq \sum_{j=1}^k \lambda_j(A - B).$$

Corollary 4.2. Let $A, B \in \mathbb{R}^{m \times m}$ be Hermitian matrices. Assume k and d are two integers which satisfy $0 \leq k \leq d \leq m$, then

$$\left| \sum_{j=k+1}^d (\lambda_j(A) - \lambda_j(B)) \right| \leq (d - k)^{1/2} \|A - B\|.$$

Proof. Since $A - B$ is Hermitian, it follows that for each $1 \leq j \leq m$ there exists $1 \leq i_j \leq m$ such that

$$|\lambda_j(A - B)| = s_{i_j}(A - B).$$

From this and Theorem 4.1 we have

$$\begin{aligned}
\sum_{j=k+1}^d (\lambda_j(A) - \lambda_j(B)) &\leq \sum_{j=1}^{d-k} \lambda_j(A - B) \leq \sum_{j=1}^{d-k} s_{i_j}(A - B) \\
&\leq \sum_{j=1}^{d-k} s_j(A - B) \leq (d - k)^{1/2} \left(\sum_{j=1}^{d-k} s_j^2(A - B) \right)^{1/2} \\
&\leq (d - k)^{1/2} \|A - B\|.
\end{aligned}$$

□

Remark 4.3. Note that the bound of the previous corollary is sharp. Indeed, let $A \in \mathbb{R}^{m \times m}$ be the diagonal matrix with coefficients $a_{ii} = 2$ for $1 \leq i \leq d$, and $a_{ii} = 0$ otherwise. Let $B \in \mathbb{R}^{m \times m}$ be the diagonal matrix with coefficients $b_{ii} = 2$ for $1 \leq i \leq k$, $b_{ii} = 1$ for $k + 1 \leq i \leq d$, and $b_{ii} = 0$ otherwise. Thus,

$$\left| \sum_{j=k+1}^d (\lambda_j(A) - \lambda_j(B)) \right| = \left| \sum_{j=k+1}^d (2 - 1) \right| = d - k.$$

Further $\|A - B\| = (d - k)^{1/2}$, and therefore

$$\left| \sum_{j=k+1}^d (\lambda_j(A) - \lambda_j(B)) \right| = (d - k)^{1/2} \|A - B\|.$$

Lemma 4.4. [7] *Suppose that $A_\omega \in \mathbb{R}^{r \times N}$ is a random matrix which satisfies (8) and $u, v \in \mathbb{R}^N$, then*

$$|\langle u, v \rangle - \langle A_\omega u, A_\omega v \rangle| \leq \varepsilon \|u\|_2 \|v\|_2,$$

with probability at least $1 - 4e^{-rc_0}$.

The following proposition was proved in [16], but we include its proof for the sake of completeness.

Proposition 4.5. *Let $A_\omega \in \mathbb{R}^{r \times N}$ be a random matrix which satisfies (8) and $M \in \mathbb{R}^{N \times m}$ be a matrix. Then, we have*

$$\|M^* M - M^* A_\omega^* A_\omega M\| \leq \varepsilon \|M\|^2,$$

with probability at least $1 - 2(m^2 + m)e^{-rc_0}$.

Proof. Set $Y_{i,j}(\omega) = (M^* M - M^* A_\omega^* A_\omega M)_{i,j} = \langle f_i, f_j \rangle - \langle A_\omega f_i, A_\omega f_j \rangle$. By Lemma 4.4 with probability at least $1 - 4e^{-rc_0}$ we have that

$$|Y_{i,j}(\omega)| \leq \varepsilon \|f_i\|_2 \|f_j\|_2 \tag{12}$$

Note that if (12) holds for all $1 \leq i \leq j \leq m$, then

$$\begin{aligned} \|M^*M - M^*A_\omega^*A_\omega M\|^2 &= \sum_{1 \leq i, j \leq m} Y_{i,j}(\omega)^2 \\ &\leq \varepsilon^2 \sum_{1 \leq i, j \leq m} \|f_i\|_2^2 \|f_j\|_2^2 = \varepsilon^2 \|M\|^4. \end{aligned}$$

Thus, by the union bound, we obtain

$$\begin{aligned} &\Pr\left(\|M^*M - M^*A_\omega^*A_\omega M\| \leq \varepsilon \|M\|^2\right) \\ &\geq \Pr\left(|Y_{i,j}(\omega)| \leq \varepsilon \|f_i\|_2 \|f_j\|_2 \quad \forall 1 \leq i \leq j \leq m\right) \\ &\geq 1 - \sum_{1 \leq i \leq j \leq m} 4e^{-rc_0} = 1 - 2(m^2 + m)e^{-rc_0}. \end{aligned}$$

□

4.2. New results and proof of Theorem 3.7. Given $M \in \mathbb{R}^{N \times m}$ with columns $\{f_1, \dots, f_m\}$ and a subspace $V \subseteq \mathbb{R}^N$, let $E(M, V)$ be as in (3), that is

$$E(M, V) = \sum_{i=1}^m d^2(f_i, V).$$

We will denote the k -minimal error associated with M by

$$E_k(M) := \min_{V: \dim(V) \leq k} E(M, V).$$

Let $d := \text{rank}(M)$. Eckart-Young's Theorem (see [17]) states that

$$E_k(M) = \sum_{j=k+1}^d \lambda_j(M^*M), \quad (13)$$

where $\lambda_1(M^*M) \geq \dots \geq \lambda_d(M^*M) > 0$ are the positive eigenvalues of M^*M .

Lemma 4.6. *Assume that $M \in \mathbb{R}^{N \times m}$ and $A \in \mathbb{R}^{r \times N}$ are arbitrary matrices. Let $S \in \mathbb{R}^{N \times s}$ be a submatrix of M . If $d := \text{rank}(M)$ is such that $0 \leq k \leq d$, then*

$$|E_k(S) - E_k(AS)| \leq (d - k)^{1/2} \|S^*S - S^*A^*AS\|.$$

Proof. Let $d_s := \text{rank}(S)$. We have $\text{rank}(AS) \leq d_s$. If $d_s \leq k$, the result is trivial. Otherwise by (13) and Corollary 4.2, we obtain

$$\begin{aligned} |E_k(S) - E_k(AS)| &= \left| \sum_{j=k+1}^{d_s} (\lambda_j(S^*S) - \lambda_j(S^*A^*AS)) \right| \\ &\leq (d_s - k)^{1/2} \|S^*S - S^*A^*AS\|. \end{aligned}$$

As S is a submatrix of M , we have that $(d_s - k)^{1/2} \leq (d - k)^{1/2}$, which proves the lemma.

□

Recall that $e_0(\mathcal{F})$ is the optimal value for the data \mathcal{F} , and $e_0(A_\omega \mathcal{F})$ is the optimal value for the data $\mathcal{F}' = A_\omega \mathcal{F}$ (See (1)). A relation between these two values is given by the following lemma.

Lemma 4.7. *Let $\mathcal{F} = \{f_1, \dots, f_m\} \subseteq \mathbb{R}^N$ and $0 < \varepsilon < 1$. If $A_\omega \in \mathbb{R}^{r \times N}$ is a random matrix which satisfies (8), then with probability exceeding $1 - 2me^{-rc_0}$, we have*

$$e_0(A_\omega \mathcal{F}) \leq (1 + \varepsilon)e_0(\mathcal{F}).$$

Proof. Let $V \subseteq \mathbb{R}^N$ be a subspace and $M \in \mathbb{R}^{N \times m}$ be a matrix. Using (8) and the union bound, with probability at least $1 - 2me^{-rc_0}$ we have that

$$\begin{aligned} E(A_\omega M, A_\omega V) &= \sum_{i=1}^m d^2(A_\omega f_i, A_\omega V) \leq \sum_{i=1}^m \|A_\omega f_i - A_\omega(P_V f_i)\|_2^2 \\ &\leq (1 + \varepsilon) \sum_{i=1}^m \|f_i - P_V f_i\|_2^2 = (1 + \varepsilon)E(M, V), \end{aligned}$$

where P_V is the orthogonal projection onto V .

Assume that $\mathbf{S} = \{S_1, \dots, S_l\}$ is an optimal partition for \mathcal{F} and $\{V_1, \dots, V_l\}$ is an optimal bundle for \mathcal{F} . Suppose that $m_i = \#(S_i)$ and $M_i \in \mathbb{R}^{N \times m_i}$ are the matrices which have $\{f_j\}_{j \in S_i}$ as columns. From what has been proved above and the union bound, with probability exceeding $1 - \sum_{i=1}^l 2m_i e^{-rc_0} = 1 - 2me^{-rc_0}$, it holds

$$e_0(A_\omega \mathcal{F}) \leq \sum_{i=1}^l E(A_\omega M_i, A_\omega V_i) \leq (1 + \varepsilon) \sum_{i=1}^l E(M_i, V_i) = (1 + \varepsilon)e_0(\mathcal{F}).$$

□

Proof of Proposition 3.6. This is a direct consequence of Lemma 4.7. □

Proof of Theorem 3.7. If $\mathbf{S}_\omega = \{S_\omega^1, \dots, S_\omega^l\}$, and $m_\omega^i = \#(S_\omega^i)$, let $M_\omega^i \in \mathbb{R}^{N \times m_\omega^i}$ be the matrices which have $\{f_j\}_{j \in S_\omega^i}$ as columns. Since $B_\omega = \{V_\omega^1, \dots, V_\omega^l\}$ is generated by \mathbf{S}_ω and \mathcal{F} , it follows that $E(M_\omega^i, V_\omega^i) = E_k(M_\omega^i)$. And as \mathbf{S}_ω is an optimal partition for $A_\omega \mathcal{F}$ in \mathbb{R}^r , we have that $\sum_{i=1}^l E_k(A_\omega M_\omega^i) = e_0(A_\omega \mathcal{F})$.

Hence, using Lemma 4.6, Lemma 4.7, and Proposition 4.5, with high probability it holds that

$$\begin{aligned}
e(\mathcal{F}, B_\omega) &\leq \sum_{i=1}^l E(M_\omega^i, V_\omega^i) = \sum_{i=1}^l E_k(M_\omega^i) \\
&\leq \sum_{i=1}^l E_k(A_\omega M_\omega^i) + (d-k)^{1/2} \sum_{i=1}^l \|M_\omega^{i*} M_\omega^i - M_\omega^{i*} A_\omega^* A_\omega M_\omega^i\| \\
&\leq e_0(A_\omega \mathcal{F}) + (l(d-k))^{1/2} \left(\sum_{i=1}^l \|M_\omega^{i*} M_\omega^i - M_\omega^{i*} A_\omega^* A_\omega M_\omega^i\|^2 \right)^{1/2} \\
&\leq (1+\varepsilon)e_0(\mathcal{F}) + (l(d-k))^{1/2} \|M^* M - M^* A_\omega^* A_\omega M\| \\
&\leq (1+\varepsilon)e_0(\mathcal{F}) + \varepsilon(l(d-k))^{1/2},
\end{aligned}$$

where $M \in \mathbb{R}^{N \times m}$ is the unitary Frobenius norm matrix which has the vectors $\{f_1, \dots, f_m\}$ as columns.

The right side of (10) follows from Proposition 4.5, Lemma 4.7, and the fact that

$$\begin{aligned}
&\Pr\left(e(\mathcal{F}, B_\omega) \leq (1+\varepsilon)e_0(\mathcal{F}) + \varepsilon(l(d-k))^{1/2}\right) \\
&\geq \Pr\left(\|M^* M - M^* A_\omega^* A_\omega M\| \leq \varepsilon \text{ and } e_0(A_\omega \mathcal{F}) \leq (1+\varepsilon)e_0(\mathcal{F})\right) \\
&\geq 1 - (2(m^2 + m)e^{-rc_0} + 2me^{-rc_0}) = 1 - (2m^2 + 4m)e^{-rc_0}.
\end{aligned}$$

□

5. CONCLUSIONS AND RELATED WORK

The existence of optimal union of subspaces models and an algorithm for finding them was obtained in [4]. In the present paper we have focused on the computational complexity of finding these models. More precisely, we studied techniques of dimension reduction for the algorithm proposed in [4]. These techniques can also be used in a wide variety of situations and are not limited to this particular application.

We used random linear transformations to map the data to a lower dimensional space. The “projected” signals were then processed in that space, (i.e. finding the optimal union of subspaces) in order to produce an optimal partition. Then we applied this partition to the original data to obtain the associated model for that partition and obtained a bound for the error.

We have analyzed two situations. First we studied the case when the data belongs to a union of subspaces (ideal case with no noise). In that case we obtained the optimal model using almost any transformation (see Proposition 3.3).

In the presence of noise, the data usually doesn’t belong to a union of low dimensional subspaces. Thus, the distances from the data to an optimal model add up to a positive error. In this case, we needed to restrict the

admissible transformations. We applied recent results on distributions of matrices satisfying concentration inequalities, which also proved to be very useful in the theory of compressed sensing.

We were able to prove that the model obtained by our approach is quasi optimal with a high probability. That is, if we map the data using a random matrix from one of the distributions satisfying the concentration law, then with high probability, the distance of the data to the model is bounded by the optimal distance plus a constant. This constant depends on the parameter of the concentration law, and the parameters of the model (number and dimension of the subspaces allowed in the model).

Let us remark here that the problem of finding the optimal union of subspaces that fit a given data set is also known as “Projective clustering”. Several algorithms have been proposed in the literature to solve this problem. Particularly relevant is [10] (see also references therein) where the authors used results from volume and adaptive sampling to obtain a polynomial-time approximation scheme. See [2] for a related algorithm.

REFERENCES

- [1] D. Achlioptas, Database-friendly random projections: Johnson-Lindenstrauss with binary coins, *Special issue on PODS 2001 (Santa Barbara, CA), J. Comput. System Sci.*, **66**(4), 671–687, 2003.
- [2] P. Agarwal and N. Mustafa, k-means projective clustering, *Proc. Principles of Database Systems (PODS04), ACM Press*, 155–165, 2004.
- [3] A. Aldroubi, C. A. Cabrelli, D. Hardin, and U. M. Molter, Optimal shift invariant spaces and their parseval frame generators, *Applied and Computational Harmonic Analysis*, **23**(2), 273–283, 2007.
- [4] A. Aldroubi, C. Cabrelli, and U. Molter, Optimal non-linear models for sparsity and sampling, *J. Fourier Anal. Appl.*, **14**(5-6), 793–812, 2008.
- [5] A. Aldroubi and R. Tessera, On the existence of optimal subspace clustering models, *Found. Comput. Math.*, to appear, 2010.
- [6] M. Anastasio and C. Cabrelli, Sampling in a Union of Frame Generated Subspaces, *Sampl. Theory Signal Image Process.*, **8**(3), 261–286, 2009.
- [7] R. Arriaga and S. Vempala, *An algorithmic theory of learning: robust concepts and random projection*, 40th Annual Symposium on Foundations of Computer Science (New York, 1999), 616–623, IEEE Computer Soc., Los Alamitos, CA, 1999.
- [8] R. Bhatia, *Matrix analysis*, Graduate Texts in Mathematics, 169, Springer-Verlag, New York, 1997.
- [9] G. Chen and G. Lerman, Foundations of a multi-way spectral clustering framework for hybrid linear modeling, *Found. Comput. Math.*, **9**(5), 517–558, 2009.
- [10] A. Deshpande, L. Rademacher, S. Vempala, and G. Wang, Matrix Approximation and Projective Clustering via Volume Sampling, *Theory of Computing*, **2**, 225–247, 2006.
- [11] Y. C. Eldar and M. Mishali, Robust recovery of signals from a structured union of subspaces, *IEEE Transactions on Information Theory*, **55**(11), 5302–5316, 2009.
- [12] E. Elhamifar and R. Vidal, Sparse subspace clustering, *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2790–2797, 20–25 June 2009.
- [13] K. Kanatani, Motion segmentation by subspace separation and model selection, *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, **2**, 586–591, 2001.

- [14] K. Kanatani and C. Matsunaga, Estimating the number of independent motions for multibody motion segmentation, *5th Asian Conference on Computer Vision*, 7–12, 2002.
- [15] Y. M Lu and M. N. Do, A theory for sampling signals from a union of subspaces, *IEEE Trans. Signal Process.*, **56**(6), 2334–2345, 2008.
- [16] T. Sarlós, Improved approximation algorithms for large matrices via random projection, *Foundations of Computer Science, 2006. FOCS '06. 47th Annual IEEE Symposium on*, 143–152, Oct. 2006.
- [17] E. Schmidt, Zur Theorie der linearen und nichtlinearen Integralgleichungen, (German), *Math. Ann.*, **63**(4), 433–476, 1907.
- [18] R. Vidal, Y. Ma, and S. Sastry, Generalized principal component analysis (GPCA), *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**(12), 1945–1959, 2005.

(Akram Aldroubi) DEPARTMENT OF MATHEMATICS, VANDERBILT UNIVERSITY, 1326 STEVENSON CENTER, NASHVILLE, TN 37240

E-mail address, Akram Aldroubi: akram.aldroubi@vanderbilt.edu

(M. Anastasio, C. Cabrelli and U. Molter) DEPARTAMENTO DE MATEMÁTICA, FACULTAD DE CIENCIAS EXACTAS Y NATURALES, UNIVERSIDAD DE BUENOS AIRES, CIUDAD UNIVERSITARIA, PABELLÓN I, 1428 CAPITAL FEDERAL, ARGENTINA, AND IMAS, UBA-CONICET, ARGENTINA

E-mail address, Magalí Anastasio: manastas@dm.uba.ar

E-mail address, Carlos Cabrelli: cabrelli@dm.uba.ar

E-mail address, Ursula M. Molter: umolter@dm.uba.ar