J Fourier Anal Appl (2008) 14: 793–812
DOI 10.1007/s00041-008-9040-2

Optimal Non-Linear Models for Sparsity and Sampling

Akram Aldroubi · Carlos Cabrelli · Ursula Molter

Received: 13 July 2007 / Published online: 3 October 2008

Abstract Given a set of vectors (the data) in a Hilbert space $\mathcal{H}$, we prove the existence of an optimal collection of subspaces minimizing the sum of the square of the distances between each vector and its closest subspace in the collection. This collection of subspaces gives the best sparse representation for the given data, in a sense defined in the paper, and provides an optimal model for sampling in union of subspaces. The results are proved in a general setting and then applied to the case of low dimensional subspaces of $\mathbb{R}^N$ and to infinite dimensional shift-invariant spaces in $L^2(\mathbb{R}^d)$. We also present an iterative search algorithm for finding the solution subspaces. These results are tightly connected to the new emergent theories of compressed sensing and dictionary design, signal models for signals with finite rate of innovation, and the subspace segmentation problem.

Communicated by Michael Elad.

The research of Akram Aldroubi is supported in part by NSF Grant DMS-0504788. The research of Carlos Cabrelli and Ursula Molter is partially supported by Grants: PICT 15033, CONICET, PIP 5650, UBACyT X058 and X108.

A. Aldroubi (✉)
Department of Mathematics, Vanderbilt University, 1326 Stevenson Center, Nashville, TN 37240, USA
e-mail: akram.aldroubi@vanderbilt.edu

C. Cabrelli · U. Molter
Departamento de Matemática, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Pabellón I, 1428 Capital Federal, Buenos Aires, Argentina

C. Cabrelli
e-mail: cabrelli@dm.uba.ar

U. Molter
e-mail: umolter@dm.uba.ar

C. Cabrelli · U. Molter
Conicet, Buenos Aires, Argentina

Keywords Sampling · Sparsity · Compressed sensing · Frames

Mathematics Subject Classification (2000) Primary 41A65 · 42C15 · Secondary 68P30 · 94A20

1 Introduction

A new paradigm for signal sampling and reconstruction recently developed by Lu and Do [21] starts from the point of view that signals live in some union of subspaces $\mathcal{M} = \bigcup_{i\in I} V_i$, instead of a single vector space $\mathcal{M} = V$ such as the space of band-limited functions also known as the Paley-Wiener space. This new paradigm is general and includes (when $\mathcal{M} = V$) the classical Shannon sampling theory and its extensions [3], as well as sampling of signal with finite rate of innovation (see e.g., [15, 24]). In the new framework, when we have more than one subspace, the signal space model $\mathcal{M} = \bigcup_{i\in I} V_i$ is non-linear and the techniques for reconstructing a signal $f \in \bigcup_{i\in I} V_i$ from its samples $\{f(x_j)\}_j$ are involved and the reconstruction operators are non-linear.

Since for each class of signals the starting point of this new theory is the knowledge of the signal space $\mathcal{M} = \bigcup_{i\in I} V_i$, the first step for implementing the theory is to find an appropriate signal model $\mathcal{M} = \bigcup_{i\in I} V_i$ from a set of observed data $\mathcal{F} = \{f_1, \ldots, f_m\}$. For the classical sampling theory, the problem of finding the shift-invariant space model $\mathcal{M} = V$ from a set of observed data has been studied and solved in [4, 5]. For the new sampling paradigm, the problem consists in proving the existence and finding subspaces $V_1, \ldots, V_l$, of some Hilbert space $\mathcal{H}$ that minimize the expression

$$e(\mathcal{F}, \{V_1, \ldots, V_l\}) = \sum_{i=1}^{m} \min_{1\le j\le l} d^2(f_i, V_j), \tag{1.1}$$

over all possible choices of l subspaces belonging to an appropriate class of subspaces of $\mathcal{H}$. Here $\mathcal{F} = \{f_1, \ldots, f_m\} \subset \mathcal{H}$ is a set of observed data and d is the distance function in $\mathcal{H}$.

It is well known that the problem of sampling and reconstruction of signals with finite rate of innovation is closely related to the developing theory of compressed sensing (see e.g., [8–10, 13, 14, 25] and the references therein). Compressed sensing proposes to find a vector $x \in \mathbb{R}^N$ from the knowledge of the values, when applied to x, of a relatively small set of functionals $\{\psi_k : k = 1, \ldots, p\}$ (where $p \ll N$). Obviously, the problem of finding x from the set $\{y_k = \langle x, \psi_k\rangle : k = 1, \ldots, p\}$ is ill-posed. However, it becomes meaningful if x is assumed to be sufficiently sparse.

A typical assumption of sparsity is that x has at most n non-zero components ($\|x\|_0 \le n$), where $n \le 2p \ll N$. As a consequence of this assumption of sparsity, the vector x belongs to some union of subspaces, each of which is generated by exactly n vectors from the canonical basis of $\mathbb{R}^N$. In matrix formulation this problem can be stated as follows: find $x \in \mathbb{R}^N$ with $\|x\|_0 \le n$ from the matrix equation $y = Ax$ where A is a $p \times N$ matrix and y is a given vector in $\mathbb{R}^p$.

A related problem consists in finding an approximation to the vector y using a sparse vector x. Formally, this problem can be stated as follows: find $\min_x \|x\|_0$ subject to the constraint $\|Ax - y\|_2 \leq \varepsilon$ for some given ε. The above two problems, their analysis, extensions, and efficient algorithms for finding their solutions can be found in [1, 2, 6, 8–10, 13, 14, 17, 27] and the references therein.

If in the above problems the matrix A is also an unknown to be found together with the set of unknown vectors $\{x_i : i = 1, \ldots, m\} \subset \mathbb{R}^N$, then these problems become the problems of finding a *dictionary* A from the data $\{y_i : i = 1, \ldots, m\} \subset \mathbb{R}^p$ obtained by sampling the sparse vectors $\{x_i : i = 1, \ldots, m\} \subset \mathbb{R}^N$ see e.g., [1, 2, 17]. In this context, the columns of A are called *atoms* of A. Under appropriate assumptions on the data and dictionary, the problem has a unique solution up to a permutation of the columns of A [1, 2]. Finding the solution to this problem by exhaustive methods is computationally intractable, but the K-SVD algorithm described in [1] provides a computationally effective search algorithm.

The problem of finding the signal model for signals with finite rate of innovation consists of finding a set $\mathcal{M} = \bigcup_{i \in I} V_i$, formed by subspaces V_i that are infinite dimensional, in general, but usually structured, e.g., each V_i is a shift-invariant space. However, the signal modeling problem as described by (1.1) is closely related to the dictionary design problem for sparse data, described in the previous paragraph.

To see this relation, let us formulate the dictionary design problem as follows: given a class of signals, determine if there exists a dictionary of small size, such that each of the signals can be represented with minimal sparsity.

More precisely, assume that we have a class of m signals, where m is a very large number. We want to know whether there exists a dictionary, such that every signal in the class is a linear combination of at most n atoms in the dictionary. Clearly, to make the problem meaningful and realistic the length of the dictionary should be small compared with m.

It follows, that if for a given set of data such a dictionary exists, then the data can be partitioned into subsets each of which belongs to a subspace of dimension at most n (i.e. to the subspace generated by the atoms that the signal uses in its representation). That is, each subset of the partition can be associated to a low dimensional subspace.

Conversely, if our class of signals can be partitioned into l subsets, such that the signals in each subset belong to a subspace of dimension no bigger than n, then by choosing a set of generators from each of the subspaces, we can construct a dictionary of length at most ln with the property that each of the signals can be represented using at most n atoms in the dictionary.

This suggests that the problem of finding a dictionary where the signals have sparse representation can be solved by finding a small collection of low dimensional subspaces containing our signals, and viceversa.

So, we will say that the class of signals is (l, n)-*sparse* (see also Definition 4.3) if there exist l subspaces of dimension at most n, such that the signals in our class belong to the union of these l subspaces. From the above discussion, it is clear that if our data is (l, n)-sparse then there exists a dictionary of length at most ln.

A related problem is the subspace segmentation problem for a set of signals in $\mathbb{R}^N$ (see for example [22, 23]). This problem occurs in the context of segmentation clustering and classification, and consists in finding whether there exist l subspaces

of dimension at most n, such that the signals in the class belong to the union of these l subspaces. The subspace segmentation problem has important applications in computer vision, image processing and other areas of engineering, and it has recently been solved using algebraic methods and algebraic geometrical tools [28]. The method for solving it (known as the Generalized Principle Component Analysis (GPCA)) has also been extended to deal with moderate noise in the data [28]. Moreover, the uniqueness problem has been addressed in [23].

Now assume that for a given l and n our data is not (l, n)-sparse. In that case we prove that there still exists a collection of optimal subspaces providing the needed sparsity. More precisely, if $\varepsilon > 0$ is given, we determine that there exists a collection of l subspaces of dimension at most n such that the sum of the squares of the distance of each signal to the union of the subspaces (i.e., the *total error*) is not larger than ε, (see formula (1.1)). In that case we will say that our data is (l, n, ε)-*sparse*.

As before it is clear that if our data is (l, n, ε)-sparse, then a dictionary of length at most ln exists such that every signal in our class can be approximated using a linear combination of at most n atoms from the dictionary, with total error not larger than ε.

Note that this definition of sparsity is an intrinsic property of the data and the space where they belong to, and does not depend on any fixed dictionary.

A relevant and important question is then, given a class of signals and a small number n, which is the minimun possible ε such the data is (l, n, ε)-sparse?

In this paper we present a general scheme that allows us to solve the problem described in (1.1), thereby finding the signal model for the new signal sampling paradigm described in [21], finding a new method for solving the segmentation subspace problem that is optimal in the presence of noise [28], and solving the (l, n, ε)-sparsity problem (in the sense defined above) for a given set of data, in different contexts. Specifically, given a set $\mathcal{F}$ of m vectors and numbers l, n such that $n, l < m$, we prove the existence of no more than l subspaces of dimension no bigger than n that provide the minimum ε such that the vectors in $\mathcal{F}$ are (l, n, ε)-sparse. When the minimum ε is zero, the data is (l, n)-sparse. We also give an iterative search algorithm to find the solution subspaces.

It is important to remark here that an optimal solution can have less than l subspaces, and the dimensions of the subspaces can be less than n. Since the minimization we consider is over unions of no more than l subspaces, where the dimension of the subspaces is no bigger than n, some of the optimal solutions for a given (l, n) (that is, some of the solutions that give the smallest ε) will yield the minimum $l_0 \leq l$ such that the data is (l_0, n, ε)-sparse, that is l is set to be just an upper bound for the number of allowable subspaces. Furthermore, the number n constraining the dimension of the subspaces is also only an upper bound, that is, an optimal solution can have subspaces of dimension strictly less than n.

1.1 Organization and Contribution

In this paper we solve the abstract problem described in (1.1). Unlike prior work in the subspace segmentation problem (see e.g., [28] and the references therein), it does not assume that the data $\mathcal{F}$ comes from union of subspaces, but instead it finds the best union of subspaces that matches the data, and therefore it is well adapted for

subspace segmentation in the presence of noise, and for the problem of sparsity and dictionary design in compressed sensing. Moreover, the setting includes finite and infinite dimensional spaces, and therefore can be used to solve the signal modeling problem described in [21]. The subspaces that are sought are not restricted to be orthogonal, or with equal dimensions or with trivial intersection, and there can be any number of subspaces up to a prescribed number l.

In Sect. 2 we formally state as Problem 1 the question described by (1.1) and introduce a general abstract scheme for solving this specific problem, together with a lemma that will provide the tool for an algorithm described in a later section.

In Sect. 3 we consider the case of the Hilbert space $L^2(\mathbb{R}^d)$ and where the infinite dimensional subspaces are shift-invariant. We show that the general theory in Sect. 2 applies to this case and thereby we solve (1.1) in this situation. As a consequence we show how the signal modeling problem is solved in Sect. 3.4.

In Sect. 4 we consider the finite dimensional case $\mathbb{R}^N$ and particularize the solution found in Sect. 2 to this case. This allows us to tie our method to the problem of finding sparsity models and the problem of finding optimal dictionaries as described in Sect. 4.2.

In Sect. 5 we present an iterative search algorithm for finding the solution to Problem (1.1). We prove that the algorithm terminates in finitely many steps. The algorithm is iterative and switches between subspace estimation and data segmentation in a way that is similar to the subspace segmentation methods and the K-SVD method described in [1, 2, 23, 28] and the references therein.

2 Abstract Hilbert Space Case

In this section we will introduce an abstract scheme that, in particular, contains the problems mentioned in the introduction. This scheme is much more general and can be used in many other situations.

In this theoretical setting we will prove the existence of optimal solutions and provide the mathematical background for the algorithms to find these solutions.

We will start by describing the basic ingredients for that setting and introducing some required notation. First we will define the class of subspaces that we will use for the minimization.

Let $\mathcal{H}$ be a Hilbert space. For $x, y \in \mathcal{H}$ let us denote by $d(x, y) = \|x - y\|_{\mathcal{H}}$. Given a finite subset $\mathcal{F} \subset \mathcal{H}$ and a closed subspace V of $\mathcal{H}$, we denote by $E(\mathcal{F}, V)$ the total distance of the data set $\mathcal{F}$ to the subspace V, i.e.

$$E(\mathcal{F}, V) = \sum_{f \in \mathcal{F}} d^2(f, V). \tag{2.1}$$

We set $E(\mathcal{F}, V) = 0$ for $\mathcal{F} = \emptyset$ and any subspace V of $\mathcal{H}$.

Let $\mathcal{C}$ be a family of closed subspaces of $\mathcal{H}$ containing the zero subspace. We will say that $\mathcal{C}$ has the Minimal Approximation Property (MAP) if for any finite set $\mathcal{F}$ of vectors in $\mathcal{H}$ there exist a subspace $V_0 \in \mathcal{C}$ that minimizes $E(\mathcal{F}, V)$ over all the subspaces $V \in \mathcal{C}$. That is,

$$E(\mathcal{F}, V_0) = \min_{V \in \mathcal{C}} E(\mathcal{F}, V) \leq E(\mathcal{F}, V), \quad \forall V \in \mathcal{C}. \tag{2.2}$$

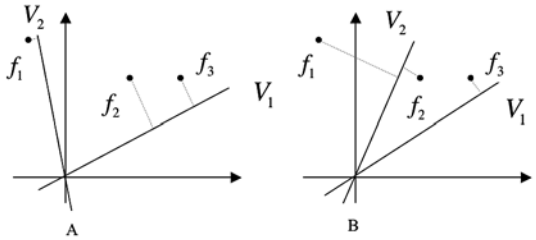

Fig. 1 Illustration for the objective function in Problem 1: A data set consists of three points $\mathcal{F} = \{f_1, f_2, f_3\}$ in $\mathbb{R}^2$. (**A**) Value of the objective function is $e = d^2(f_1, V_2) + d^2(f_2, V_1) + d^2(f_3, V_1)$; and (**B**) Value of the objective function is $e = d^2(f_1, V_2) + d^2(f_2, V_2) + d^2(f_3, V_1)$. Note that the configuration of V_1, V_2 in Panel **A** forced a partition of the data into $P_1 = \{f_1\}$ and $P_2 = \{f_2, f_3\}$, while the configuration in **B** forced the partition $P_1 = \{f_1, f_2\}$ and $P_2 = \{f_3\}$ for the same data

Any subspace $V_0 \in \mathcal{C}$ satisfying (2.2) will be called an *optimal subspace for* $\mathcal{F}$. Note that if $\mathcal{F} = \emptyset$ then every subspace in $\mathcal{C}$ is optimal. We will choose the zero subspace in that case. For the rest of this section we will assume that the class $\mathcal{C}$ has the Minimal Approximation Property.

Next, since we are interested in models that are union of subspaces, we will arrange the subspaces in finite bundles that will be our main objects, and define the distance (error) between a bundle and a set of vectors.

To do this, let us fix $m, l \in \mathbb{N}$ with $1 \le l \le m$ and let $\mathcal{F} = \{f_1, \ldots, f_m\}$ be a finite set of vectors in $\mathcal{H}$.

Define $\mathfrak{B}$ to be the set of sequences of elements in $\mathcal{C}$ of length l, i.e.

$$\mathfrak{B} = \mathfrak{B}(l) = \big\{\mathbf{V} = \{V_1, \ldots, V_l\} : V_i \in \mathcal{C}, 1 \le i \le l\big\}.$$

We will call these finite sequences *bundles*. For $\mathbf{V} \in \mathfrak{B}$ with $\mathbf{V} = \{V_1, \ldots, V_l\}$, we define,

$$e(\mathcal{F}, \mathbf{V}) = \sum_{f \in \mathcal{F}} \min_{1 \le j \le l} d^2(f, V_j). \tag{2.3}$$

Remark Note that $e(\mathcal{F}, \mathbf{V})$ is computed as follows: For each $f \in \mathcal{F}$ find the space $V_{j(f)}$ in $\mathbf{V}$ closest to f, compute $d^2(f, V_{j(f)})$, and then sum over all values found by letting f run through $\mathcal{F}$ (see Fig. 1). Also note that $e(\mathcal{F}, \mathbf{V})$ is a non-linear function of $\mathcal{F}$.

Hence, for the problems described in the introduction, what we want is to minimize e over all possible bundles of subspaces. This is formulated in the following problem.

Problem 1

(1) *Given a finite set* $\mathcal{F} \subset \mathcal{H}$, *minimize* $e(\mathcal{F}, \mathbf{V})$ *over* $\mathbf{V} \in \mathfrak{B}$. *That is, find*

$$\inf\{e(\mathcal{F}, \mathbf{V}) : \mathbf{V} \in \mathfrak{B}\}. \tag{2.4}$$

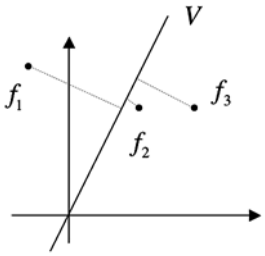

Fig. 2 Illustration for the objective function in Problem 1: Same data set $\mathcal{F} = \{f_1, f_2, f_3\}$ as in Fig. 1, but for a single subspace V. This objective function is the classical least squares cost function

(2) *Find a bundle* $\mathbf{V}_0 \in \mathfrak{B}$ (*if it exists*) *such that*

$$e(\mathcal{F}, \mathbf{V}_0) = \inf\{e(\mathcal{F}, \mathbf{V}) : \mathbf{V} \in \mathfrak{B}\}. \tag{2.5}$$

Any $\mathbf{V}_0 \in \mathfrak{B}$ *that satisfies* (2.5) *will be called a* solution *to Problem* 1.

We will show (Theorem 2.2) that Problem 1 can be solved, i.e. for a given data set $\mathcal{F}$, there does exist a bundle $\mathbf{V}_0$ that minimizes $e(\mathcal{F}, \mathbf{V})$. Moreover, we propose an algorithm to find this optimal bundle.

If we set $l = 1$, $\mathcal{H} = \mathbb{R}^d$, and $\mathcal{C} = \mathfrak{L}_n$ to be the set of all subspaces of dimension smaller (or equal) than n, then Problem 1 reduces to the classical least squares problem. This last problem has been studied extensively (see Fig. 2 for an illustration in $\mathbb{R}^2$), and it can be solved using the well-known Singular Value Decomposition (SVD) (see e.g., [16, 26]).

Before stating the main results, we need to give some definitions and set some notation.

We will denote by $\Pi = \Pi_l$ the set of all l-sequences $P = \{\mathcal{F}_1, \ldots, \mathcal{F}_l\}$ of subsets of $\mathcal{F}$ satisfying the property that for all $1 \le i, j \le l$,

$$\mathcal{F}_i \subset \mathcal{F}, \qquad \mathcal{F} = \bigcup_{s=1}^{l} \mathcal{F}_s, \quad \text{and} \quad \mathcal{F}_i \cap \mathcal{F}_j = \emptyset \quad \text{for } i \neq j.$$

Note that we allow some of the elements of $P \in \Pi$ to be the empty set. By abuse of language we will still call the elements of Π_l *partitions* (of $\mathcal{F}$).

For $P \in \Pi_l$, $P = \{\mathcal{F}_1, \ldots, \mathcal{F}_l\}$ and $\mathbf{V} \in \mathfrak{B}$, $\mathbf{V} = \{V_1, \ldots, V_l\}$ we define,

$$\Gamma(P, \mathbf{V}) = \sum_{i=1}^{l} E(\mathcal{F}_i, V_i). \tag{2.6}$$

So Γ measures the error between a fixed partition P and a fixed bundle $\mathbf{V}$.

The following relations between partitions in Π_l and bundles of length l in $\mathcal{C}$ will be relevant for our analysis.

Given a bundle $\mathbf{V} \in \mathfrak{B}$, $\mathbf{V} = \{V_1, \ldots, V_l\} \subset \mathcal{C}$, we can partition the set $\mathcal{F}$ into a *best partition* $P = \{\mathcal{F}_1, \ldots, \mathcal{F}_l\}$, by grouping together into $\mathcal{F}_i$, the vectors in $\mathcal{F}$ that are closer to a given subspace V_i than to any other subspace V_j, $j \neq i$ (see Fig. 1). However, there are situations in which a vector $f \in \mathcal{F}$ is at equal distance from two or more subspaces from the bundle. Hence, there may be more than one partition

associated to each *bundle*, and there is a subset $\Omega_l(\mathbf{V}) \subset \Pi_l$ of *best partitions* in Π_l naturally associated to $\mathbf{V}$ defined by

$$P = \{\mathcal{F}_1, \ldots, \mathcal{F}_l\} \in \Pi_l \text{ is a member of } \Omega_l(\mathbf{V}) \text{ if it satisfies}$$
$$f \in \mathcal{F}_j \text{ implies that } d(f, V_j) \le d(f, V_h), h = 1, \ldots, l.$$

Conversely, since $\mathcal{C}$ has the MAP, given a partition $P = \{\mathcal{F}_1, \ldots, \mathcal{F}_l\}$, we can define a *best bundle* $\mathbf{V}_P = \{V_1, \ldots, V_l\} \in \mathfrak{B}$ by finding (for each i) the space V_i that minimizes (2.1) for the given $\mathcal{F}_i$. However, there are situations in which, for some $\mathcal{F}_i$, there are more than one subspace V_i that minimizes (2.1). Hence, there is a subset $\mathcal{W}(P) \subset \mathfrak{B}$ of *best bundles* associated to $\mathcal{F}$ defined by

$$\mathbf{V}_P = \{V_1, \ldots, V_l\} \in \mathfrak{B} \text{ is a member of } \mathcal{W}(P) \text{ if } V_i \text{ is an optimal}$$
$$\text{subspace for } \mathcal{F}_i \text{ (in the sense of (2.2)) for each } i = 1, \ldots, l.$$

In what follows when we refer to a best partition associated to a bundle $\mathbf{V}$ we will mean, any element in $\Omega_l(\mathbf{V})$. Similarly, when we talk of a best bundle associated to a partition P, this will mean an element in $\mathcal{W}(P)$.

We also consider the set of all pairs $(P, \mathbf{V}_P)$, where $P \in \Pi_l$ and $\mathbf{V}_P \in \mathcal{W}(P)$. We will say that a pair $(P_0, \mathbf{V}_{P_0})$ is Γ-*minimal* if

$$\Gamma(P_0, \mathbf{V}_{P_0}) \le \Gamma(P, \mathbf{V}_P) \tag{2.7}$$

for all such pairs.

Note that when trying to compute $e(\mathcal{F}, \mathbf{V})$, for each $f \in \mathcal{F}$ we first have to find the subspace $V_{j(f)}$ in $\mathbf{V}$ that is closest to f and then compute $d^2(f, V_{j(f)})$ (see remark after the definition of $e(\mathcal{F}, \mathbf{V})$ and Fig. 1). While for Γ, a partition is given and we just compute the distance of each function to its corresponding space (not the closest one necessarily). The surprising fact is that e and Γ can indeed be compared, as the following lemma shows. In addition, this result will later give us the key to obtain an algorithm for Problem 1.

Lemma 2.1 *Let $(P_0, \mathbf{V}_{P_0})$ be a Γ-minimal pair. Then we have*

$$e(\mathcal{F}, \mathbf{V}_{P_0}) = \Gamma(P_0, \mathbf{V}_{P_0}). \tag{2.8}$$

Proof It is clear that $e(\mathcal{F}, \mathbf{V}_{P_0}) \le \Gamma(P_0, \mathbf{V}_{P_0})$.

For the other inequality, if $\mathbf{V}_{P_0} = \{V_1, \ldots, V_l\}$ then for any $P \in \Omega_l(\mathbf{V}_{P_0})$ we have

$$e(\mathcal{F}, \mathbf{V}_{P_0}) = \sum_{i=1}^{m} \min_{1 \le j \le l} d^2(f_i, V_j) = \Gamma(P, V_{P_0}). \tag{2.9}$$

In addition, $\Gamma(P, V_{P_0}) \ge \Gamma(P, \mathbf{V}_P)$, with $\mathbf{V}_P \in \mathcal{W}(P)$. But by the minimality of $\Gamma(P_0, V_{P_0})$ given by hypothesis, we have that $\Gamma(P, \mathbf{V}_P) \ge \Gamma(P_0, \mathbf{V}_{P_0})$, and the lemma follows. □

We are now ready to prove the following theorem which shows that we can solve Problem 1.

Theorem 2.2 *Let $\mathcal{H}$ be a Hilbert space, m, l positive integers with $l \leq m$ and $\mathcal{F} = \{f_1, \ldots, f_m\}$ a set of vectors in $\mathcal{H}$. Then*

(1) *There exists a bundle $\mathbf{V}_0 \in \mathfrak{B}$ that solves Problem 1 for the data $\mathcal{F}$, that is,*

$$e(\mathcal{F}, \mathbf{V}_0) = \inf\{e(\mathcal{F}, \mathbf{V}) : \mathbf{V} \in \mathfrak{B}\}.$$

(2) *If $(P_0, \mathbf{V}_{P_0})$ is a Γ-minimal pair, then all the elements of $\mathcal{W}(P_0)$, are solutions to Problem 1.*
(3) *Furthermore, if $\mathbf{V}_0$ is a solution to Problem 1, then there exists $P_0 \in \Pi_l$ such that $\mathbf{V}_0 \in \mathcal{W}(P_0)$, i.e. $(P_0, \mathbf{V}_0)$ is a Γ-minimal pair.*

In other words, Theorem 2.2 states that Problem 1 has a solution for every finite set of vectors $\mathcal{F} = \{f_1, \ldots, f_m\} \subset \mathcal{H}$ and every $l \geq 1$ if and only if $\mathcal{C}$ has the MAP property. One direction of the theorem is trivial. The interesting implication is that if Problem 1 can be solved for any $\mathcal{F}$ and $l = 1$ then it can be solved for any $\mathcal{F}$ and any $l \geq 1$.

Proof We will prove that if $(P_0, \mathbf{V}_{P_0})$ is a Γ-minimal pair, then

$$e(\mathcal{F}, \mathbf{V}_{P_0}) \leq e(\mathcal{F}, \mathbf{V}), \quad \forall \mathbf{V} \in \mathfrak{B}.$$

For this, let us choose an arbitrary $\mathbf{V} \in \mathfrak{B}$. We have that for each $P \in \Omega_l(\mathbf{V})$

$$\Gamma(P, \mathbf{V}) = e(\mathcal{F}, \mathbf{V}).$$

Clearly $\Gamma(P, \mathbf{V}_P) \leq \Gamma(P, \mathbf{V})$, for each $\mathbf{V}_P \in \mathcal{W}(P)$.
Because of the minimality of $\Gamma(P_0, \mathbf{V}_{P_0})$, we have

$$\Gamma(P_0, \mathbf{V}_{P_0}) \leq \Gamma(P, \mathbf{V}_P).$$

As a consequence of Lemma 2.1 we know that then

$$\Gamma(P_0, \mathbf{V}_{P_0}) = e(\mathcal{F}, \mathbf{V}_{P_0}),$$

which proves,

$$e(\mathcal{F}, \mathbf{V}_{P_0}) \leq e(\mathcal{F}, \mathbf{V}).$$

This shows that if $(P_0, \mathbf{V}_{P_0})$ is a Γ-minimal pair, then each bundle $\mathbf{V}_{P_0}$ solves Problem 1 for the data $\mathcal{F}$. Since the total number of pairs is finite, then there exist minimal pairs. This proves parts (1) and (2) of the Theorem.

For part (3) let $\mathbf{V}_0 \in \mathfrak{B}$ be a solution to Problem 1, i.e. $e(\mathcal{F}, \mathbf{V}_0) \leq e(\mathcal{F}, \mathbf{V}),\ \forall\, \mathbf{V} \in \mathfrak{B}$. Consider $P_0 \in \Omega_l(\mathbf{V}_0)$ and let $\mathbf{V}_{P_0} \in \mathcal{W}(P_0)$. Then, since $P_0 \in \Omega_l(\mathbf{V}_0)$ and by the minimality of $\mathbf{V}_0$ we have

$$\Gamma(P_0, \mathbf{V}_0) = e(\mathcal{F}, \mathbf{V}_0) \leq e(\mathcal{F}, \mathbf{V}_{P_0}) \leq \Gamma(P_0, \mathbf{V}_{P_0}).$$

Therefore, $\Gamma(P_0, \mathbf{V}_0) \leq \Gamma(P_0, \mathbf{V}_{P_0})$, but by definition of Γ, $\Gamma(P_0, \mathbf{V}_{P_0}) \leq \Gamma(P_0, \mathbf{V})$ for any $\mathbf{V} \in \mathfrak{B}$. So,

$$\Gamma(P_0, \mathbf{V}_0) = \Gamma(P_0, \mathbf{V}_{P_0}), \quad \text{and} \quad \mathbf{V}_0 \in \mathcal{W}(P_0).$$

Moreover, $(P_0, \mathbf{V}_0)$ is Γ-minimal since

$$\Gamma(P_0, \mathbf{V}_0) = e(\mathcal{F}, \mathbf{V}_0) \leq e(\mathcal{F}, \mathbf{V}_P) \leq \Gamma(P, \mathbf{V}_P).$$

This completes the proof of the Theorem. □

Remark If $0 < l_1 < l_2$, then for any $\mathbf{V} \in \mathfrak{B}(l_1)$, $\mathbf{V} = \{V_1, \ldots, V_{l_1}\}$ the bundle $\mathbf{V}' = \{V_1, \ldots, V_{l_1}, \{0\}, \ldots, \{0\}\}$ belongs to $\mathfrak{B}(l_2)$ and therefore, we have

$$e(\mathcal{F}, \mathbf{V}_{P_0}(l_1)) \geq e(\mathcal{F}, \mathbf{V}_{P_0}(l_2)).$$

So the error decreases (or at least does not increase) when l (the number of subspaces) increases. Note that in case that the number of subspaces equals the number of data, the error is zero, since we can pick for each data signal the subspace spanned by itself.

It is important to remark here that optimal bundles can have the zero subspace as some of its components. So, if l_0 is the number of subspaces that have dimension greater than zero, in some optimal bundle $\mathbf{V}_0$, then the bundle with l_0 components obtained after the $l - l_0$ zero components are removed from $\mathbf{V}_0$, is also an optimal bundle for the Problem 1 when $\mathfrak{B}(l)$ is replaced by $\mathfrak{B}(l_0)$. Thus as mentioned in the introduction, the number l is simply a set to be an a priori upper bound on the number of subspaces, and the optimal solution(s) can have any number of subspaces $l_0 \leq l$.

3 The Shift-Invariant Space Case and Optimal Nonlinear Signal Models

In this section we will apply the theory of Sect. 2 to the Hilbert space $L^2(\mathbb{R}^d)$. In order to do that we will select a family of subspaces with the Minimal Approximation Property. We will describe in what follows the necessary setting.

We begin by recalling the definition of frames and some of their properties (see for example [11, 12, 18, 19]).

Let $\mathcal{H}$ be a Hilbert space and $\{u_i\}_{i \in I}$ a countable subset of $\mathcal{H}$. The set $\{u_i\}_{i \in I}$ is said to form a *frame* for $\mathcal{H}$ if there exist $q, Q > 0$ such that

$$q\|f\|^2 \leq \sum_{i \in I} |\langle f, u_i \rangle|^2 \leq Q\|f\|^2, \quad \forall f \in \mathcal{H}.$$

If $q = Q$, then $\{u_i\}_{i \in I}$ is called a *tight frame*, and it is called a *Parseval frame* if $q = Q = 1$.

If $\{u_i\}_{i \in I}$ is a Parseval frame for a subspace W of a Hilbert space $\mathcal{H}$, and if $a \in \mathcal{H}$, then the orthogonal projection of a onto W is given by:

$$\mathcal{P}_W(a) = \sum_{i \in I} \langle a, u_i \rangle u_i. \tag{3.1}$$

Thus, a Parseval frames acts as if it were an orthonormal basis of W, even though it may not be one.

3.1 Shift-Invariant Spaces

In this paper, a shift-invariant space will be a subspace of $L^2(\mathbb{R}^d)$ of the form:

$$S(\Phi) := \text{closure}_{L_2}\ \text{span}\{\varphi_i(x-k) :\ i=1,\dots,n,\ k\in\mathbb{Z}^d\}, \tag{3.2}$$

where $\Phi=\{\varphi_1,\dots,\varphi_n\}$ is a set of functions in $L^2(\mathbb{R}^d)$. The functions $\varphi_1,\varphi_2,\dots,\varphi_n$ are called a *set of generators* for the space $S=S(\Phi)=S(\varphi_1,\dots,\varphi_n)$ and any such space S is called a *finitely generated shift-invariant space* (*FSIS*) (see e.g., [7]). These spaces are often used as standard signal and image models. For example, if $n=1,\ d=1$ and $\phi(x)=\text{sinc}(x)$, then the underlying space is the space of band-limited functions (often used in communications).

Finitely generated shift-invariant spaces, can have different sets of generators. The *length* of an FSIS S is,

$$l(S)=\min\{\ell\in\mathbb{N} : \exists\,\varphi_1,\dots,\varphi_\ell\in S \text{ with } S=S(\varphi_1,\dots,\varphi_\ell)\}.$$

If $S=\{0\}$, we set $l(S)=0$. We will denote by $\mathcal{L}_n$ the set of all shift-invariant spaces with length less than or equal to n. That is, an element in $\mathcal{L}_n$ is a shift-invariant space that has a set of s generators with $s\le n$.

3.2 The Minimal Approximation Property for SIS

In [4] it was proven that $\mathcal{L}_n$ has the MAP. More precisely,

Theorem 3.1 *Let $\mathcal{F}=\{f_1,\dots,f_m\}$ be a set of functions in $L^2(\mathbb{R}^d)$. Then there exists $V\in\mathcal{L}_n$ such that*

$$\sum_{i=1}^m \|f_i-\mathcal{P}_V f_i\|^2 \le \sum_{i=1}^m \|f_i-\mathcal{P}_{V'} f_i\|^2,\quad \forall V'\in\mathcal{L}_n. \tag{3.3}$$

Here $\mathcal{P}_V$ denote the orthogonal projection onto the subspace V.

Furthermore, an explicit description of an optimal space (that is not necessarily unique) and an estimation of the error, was obtained in [4], as is described below. Let us call,

$$\mathcal{E}(\mathcal{F},n)=\min_{V'\in\mathcal{L}_n}\sum_{i=1}^m \|f_i-\mathcal{P}_{V'} f_i\|^2. \tag{3.4}$$

To compute the error $\mathcal{E}(\mathcal{F},n)$ we need to consider the Gramian matrix $G_{\mathcal{F}}$ of $\mathcal{F}=\{f_1,\dots,f_m\}$. Specifically, the *Gramian* G_Φ of a set of functions $\Phi=\{\varphi_1,\dots,\varphi_n\}$ with elements in $L^2(\mathbb{R}^d)$ is defined to be the $n\times n$ matrix of $\mathbb{Z}^d$-periodic functions

$$[G_\Phi(\omega)]_{i,j}=\sum_{k\in\mathbb{Z}^d}\widehat{\varphi_i}(\omega+k)\overline{\widehat{\varphi_j}(\omega+k)},\quad \omega\in\mathbb{R}^d, \tag{3.5}$$

where $\widehat{\varphi_i}$ denotes the Fourier transform of φ_i, and $\overline{\widehat{\varphi_i}}$ denotes the complex conjugate of $\widehat{\varphi_i}$.

The next theorem produces a set of generators for an optimal space $V \in \mathcal{L}_n$ and provides a formula for the exact value of the error.

Theorem 3.2 ([4]) *Under the same assumptions as in Theorem 3.1, let* $\lambda_1(\omega) \geq \lambda_2(\omega) \geq \cdots \geq \lambda_m(\omega)$ *be the eigenvalues of the Gramian* $G_{\mathcal{F}}(\omega)$. *Then*

(1) *The eigenvalues* $\lambda_i(\omega)$, $1 \leq i \leq m$ *are* $\mathbb{Z}^d$*-periodic, measurable functions in* $L^2([0,1]^d)$ *and*

$$\mathcal{E}(\mathcal{F}, n) = \sum_{i=n+1}^{m} \int_{[0,1]^d} \lambda_i(\omega) d\omega. \tag{3.6}$$

(2) *Let* $E_i := \{\omega : \lambda_i(\omega) \neq 0\}$, *and define* $\tilde{\sigma}_i(\omega) = \lambda_i^{-1/2}(\omega)$ *on* E_i *and* $\tilde{\sigma}_i(\omega) = 0$ *on* E_i^c. *Then, there exists a choice of measurable left eigenvectors* $y_1(\omega), \ldots, y_n(\omega)$ *with* $y_i = (y_{i1}, \ldots, y_{im})^t$, $i = 1, \ldots, n$, *associated with the first n largest eigenvalues of* $G_{\mathcal{F}}(\omega)$ *such that the functions defined by*

$$\hat{\varphi}_i(\omega) = \tilde{\sigma}_i(\omega) \sum_{j=1}^{m} y_{ij}(\omega) \hat{f}_j(\omega), \quad i = 1, \ldots, n, \ \omega \in \mathbb{R}^d \tag{3.7}$$

are in $L^2(\mathbb{R}^d)$.

Furthermore, the corresponding set of functions $\Phi = \{\varphi_1, \ldots, \varphi_n\}$ *is a set of generators for an optimal space V and the set* $\{\varphi_i(\cdot - k), k \in \mathbb{Z}^d, i = 1, \ldots, n\}$ *is a Parseval frame for V*.

Note that (3.7) says that in particular the generators of the optimal space are $l_2(\mathbb{Z})$-linear combinations of the integer translates of the data $\mathcal{F}$.

3.3 Best Approximation by Bundles of SIS

Let $\mathcal{F} = \{f_1, \ldots, f_m\}$ be functions in $L^2(\mathbb{R}^d)$ and n a positive integer smaller than m.

The result in Theorem 3.1 says that the class $\mathcal{L}_n$ has the Minimal Approximation Property.

Define $\mathcal{S} = \{\{S_1, \ldots, S_l\} : S_i \in \mathcal{L}_n\}$ to be the set of bundles of SIS in $\mathcal{L}_n$. Now we can apply Theorem 2.2 to conclude that

Theorem 3.3 *Let* $\mathcal{F} = \{f_1, \ldots, f_m\}$ *vectors in* $L^2(\mathbb{R}^d)$, *then there exist a bundle* $\mathbf{S}_0 = \{S_1^0, \ldots, S_l^0\} \in \mathcal{S}$ *such that*

$$e(\mathcal{F}, \mathbf{S}_0) = \sum_{i=1}^{m} \min_{1 \leq j \leq l} d_2^2(f_i, S_j^0) \leq \sum_{i=1}^{m} \min_{1 \leq j \leq l} d_2^2(f_i, S_j) \tag{3.8}$$

over all bundles $\mathbf{S} = \{S_1, \ldots, S_l\} \in \mathcal{S}$.

Let $P_0 = \{\mathcal{F}_1^0, \ldots, \mathcal{F}_l^0\}$ be a best partition of $\mathcal{F}$ associated to the optimal bundle $\mathbf{S}_0 = \{S_1^0, \ldots, S_l^0\}$ (i.e. $P_0 = \{\mathcal{F}_1^0, \ldots, \mathcal{F}_l^0\} \in \Omega_l(\mathbf{S}_0)$, is such that $f_j \in \mathcal{F}_i^0$ implies $d(f_j, S_i^0) \leq d(f_j, S_k^0)$, $k = 1, \ldots, l$).

Using Theorem 3.2 for each $h = 1, \ldots, l$ and such that $\mathcal{F}_h^0 \neq \emptyset$, a set of generators forming a Parseval frame can be obtained for the optimal space S_h^0 in terms of the singular values and singular vectors of the gramian $G_{\mathcal{F}_h}$ associated to the subset $\mathcal{F}_h^0$. Furthermore a formula for the minimum error $E(\mathcal{F}_h^0, S_h^0)$ is given in terms of the singular values of $G_{\mathcal{F}_h^0}$. Therefore, thanks to Lemma 2.1 and the definition of Γ (see (2.6)), $e(\mathcal{F}, \mathbf{S}_0)$ can be computed exactly.

3.4 Optimal Signal Models

If it is known a priori that a class of signals belong to a union of shift-invariant spaces $\bigcup_{i=1}^l S_i$ each of length no larger than n, then combining Theorems 3.2 and 3.3, we can find the signal model $\bigcup_{i=1}^l S_i$ exactly and the generators of each space S_i. This solution includes the case where the signal class consists of a single shift-invariant space solved in [5].

When the data $\mathcal{F}$ is corrupted by noise, or if the true signals are not from a union of shift-invariant spaces $\bigcup_{i=1}^l S_i$ of length no larger than n, but we still wish to model the class of signals by such a union, then we can use the bundle $S_0 = \{S_1^0, \ldots, S_l^0\}$ found in Theorem 3.3 to obtain an optimal signal model $\bigcup_i S_i^0$ compatible with the observed data. The optimal signal model $\bigcup_i S_i^0$ is a union of infinite dimensional spaces S_i^0, but each S_i^0 is a shift-invariant space that can be generated by at most n frame generators $\Phi_0^i = \{\varphi_{0,1}^i, \ldots, \varphi_{0,s_i}^i\}$, $s_i \leq n$, $i = 1, \ldots, q$, $q \leq l$ (we only use the spaces S_i^0 that have length larger than 0). Each signal $f_j \in \mathcal{F}$ can now be modeled by its orthogonal projection $f_j^a = \mathcal{P}_{S_i^0} f_j$ onto its closest space S_i^0. which consists of countably many but generally infinite linear combinations of atoms, i.e., $f_j^a = \sum_{p=1}^{s_i} \sum_k c_k^p \varphi_{0,p}^i(. - k)$.

Note that if the data $\mathcal{F}$ is corrupted by noise, then the optimal model can be used as a denoising method. Other applications of the optimal signal model are those of learning, data segmentation, and classification.

4 The Finite Dimensional Case $\mathbb{R}^N$, Sparse Representations, and Optimal Dictionaries

In this section we will consider Problem 1 for the case in which the Hilbert space is $\mathbb{R}^N$ (in applications, usually one thinks of N as being very large). In this case, our data $\mathcal{F} = \{f_1, \ldots, f_m\}$ are vectors in $\mathbb{R}^N$. Let us denote by $\mathfrak{L}_n$ the set of all subspaces of dimension smaller (or equal) than n. To see that $\mathfrak{L}_n$ has the MAP, we will recourse to some well-known results about Singular Value Decomposition (SVD) and in particular, the Eckart-Young Theorem.

Let us briefly recall the SVD decomposition (for a detailed treatment see for example [20], or the Appendix of [4]). Let $A \in \mathbb{R}^{N \times m}$ with columns $\{a_1, \ldots, a_m\}$, and let r be the rank of A. One can obtain its SVD as follows. Consider the matrix $A^t A \in \mathbb{R}^{m \times m}$. Since $A^t A$ is self-adjoint and positive semi-definite, its eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ are nonnegative and the associated eigenvectors $y_1, \ldots, y_m$ can

be chosen to form an orthonormal basis of $\mathbb{R}^m$. Note that the rank r of A corresponds to the largest index i such that $\lambda_i > 0$. The left singular vectors $u_1, \ldots, u_r$ can then be obtained from

$$\sqrt{\lambda_i} u_i = A y_i, \quad \text{that is} \quad u_i = \lambda_i^{-1/2} \sum_{j=1}^{m} y_{ij} a_j \quad (1 \le i \le r).$$

Here $y_i = (y_{i1}, \ldots, y_{im})^t$. The remaining left singular vectors $u_{r+1}, \ldots, u_m$ can be chosen to be any orthonormal collection of $m-r$ vectors in $\mathbb{R}^N$ that are perpendicular to span $\{a_1, \ldots, a_m\}$. One may then readily verify that

$$A = \sum_{k=1}^{m} \sqrt{\lambda_k} u_k y_k^t = U \Lambda^{1/2} Y^t, \tag{4.1}$$

where $U \in \mathbb{R}^{N \times m}$ is the matrix $U = \{u_1, \ldots, u_m\}$, $\Lambda^{1/2} = \operatorname{diag}(\lambda_1^{1/2}, \ldots, \lambda_m^{1/2})$, and $Y = \{y_1, \ldots, y_m\} \in \mathbb{R}^{m \times m}$ with $U^t U = I_m = Y^t Y = Y Y^t$.

The following theorem of Schmidt (cf. [26]) (usually coined as Eckart-Young Theorem [16]) shows that our set $\mathfrak{L}_n$ has the MAP. (We again denote by $\mathcal{P}_V$ the orthogonal projection onto the space V.)

Theorem 4.1 (Eckart-Young) *Let $\{f_1, \ldots, f_m\}$ be a set of vectors in $\mathbb{R}^N$ and $r = \dim(\operatorname{span}\{f_1, \ldots, f_m\})$. Suppose that the associated matrix $A = \{f_1, \ldots, f_m\}$, has SVD $A = U \Lambda^{1/2} Y^t$ and that $0 < n \le r$. If $W = \operatorname{span}\{u_1, \ldots, u_n\}$, then*

$$\{\mathcal{P}_W f_1, \ldots, \mathcal{P}_W f_m\} = \sum_{i=1}^{n} \sqrt{\lambda_i} u_i y_i^t = A_n$$

and

$$\sum_{i=1}^{m} \|a_i - \mathcal{P}_W a_i\|_2^2 \le \sum_{i=1}^{m} \|a_i - \mathcal{P}_V a_i\|_2^2, \quad \forall V \in \mathfrak{L}_n. \tag{4.2}$$

Furthermore, the space W is unique if $\lambda_{n+1} \ne \lambda_n$. In addition,

$$E(\mathcal{F}, W) = \min_{V \in \mathfrak{L}_n} \sum_{i=1}^{m} \|f_i - \mathcal{P}_V f_i\|^2 = \sum_{j=n+1}^{r} \lambda_j. \tag{4.3}$$

4.1 Best Non-Linear Approximation by Bundles of Subspaces in $\mathbb{R}^N$

Let $\mathcal{F} = \{f_1, \ldots, f_m\}$ be a set of vectors in $\mathbb{R}^N$ and $n \le m$. As indicated before, Theorem 4.1 states precisely that $\mathfrak{L}_n$ has the MAP.

Define again $\mathfrak{B} = \mathfrak{B}(l)$ to be the set of non-empty bundles of length l in $\mathfrak{L}_n$, i.e.

$$\mathfrak{B} = \{\{V_1, \ldots, V_l\} : V_i \in \mathfrak{L}_n,\ i = 1, \ldots, l\}.$$

We then have the following theorem.

Theorem 4.2 *Let $\mathcal{F} = \{f_1, \ldots, f_m\}$ be vectors in $\mathbb{R}^N$, and let l and n be given ($l < m$, $n < N$), then there exist a bundle $\mathbf{V}_0 = \{V_1^0, \ldots, V_l^0\} \in \mathfrak{B}$, such that*

$$e(\mathcal{F}, \mathbf{V}_0) = \sum_{i=1}^{n} \min_{1 \le j \le l} d_2^2(f_i, V_j^0) = \inf\{e(\mathcal{F}, \mathbf{V}) : \mathbf{V} \in \mathfrak{B}\}.$$

Let $P_0 = \{\mathcal{F}_1^0, \ldots, \mathcal{F}_l^0\}$ be the best partition of $\mathcal{F}$ associated to the optimal bundle $\mathbf{V}_0 = \{V_1^0, \ldots, V_l^0\}$ (i.e. $P_0 = \{\mathcal{F}_1^0, \ldots, \mathcal{F}_l^0\} \in \Omega_l(\mathbf{V}_0)$, is such that $f_j \in \mathcal{F}_i^0$ implies $d(f_j, V_i^0) \le d(f_j, V_k^0)$, $k = 1, \ldots, l$).

Now, using Theorem 4.1 for each $h = 1, \ldots, l$ and such that $\mathcal{F}_h^0 \neq \emptyset$, a set of generators forming an orthonormal base can be obtained for the optimal space V_h^0 in terms of the singular values and singular vectors of the matrix A_h associated to the subset $\mathcal{F}_h^0$. Furthermore, by (4.3), a formula for the minimum error $E(\mathcal{F}_h^0, V_h^0)$, and therefore for $e(\mathcal{F}, \mathbf{V}_0)$, is given in terms of the singular values of A_h.

Remark Theorem 4.2 remains true if we replace the set $\mathfrak{L}_n$ by the set $\mathfrak{L}_{(n_1,\ldots,n_l)}$ of bundles $\{V_1, \ldots, V_l\}$ such that $\dim V_i \le n_i$, for $i = 1, \ldots, l$.

4.2 Sparsity and Optimal Dictionaries

Now, we will describe the relation between the solution to Problem 1 for $\mathbb{R}^N$, as described in Section 4.1, with the problem of dictionary finding and sparsity. Let us introduce the following problem. See for example [1].

Problem 2 *Given data $\mathcal{F} = \{f_1, \ldots, f_m\}$ in $\mathbb{R}^N$ and positive integers n and d, find a dictionary D (i.e. a set of vectors in $\mathbb{R}^N$) of length at most d, such that each f_i can be written as a linear combination of at most n atoms in D.*

That is, (in matrix notation) find a $N \times r$ matrix D, with columns $a_1, \ldots, a_r$ in $\mathbb{R}^N$ and $r \le d$, such that there exist an $r \times m$ matrix X, with columns $x_1, \ldots, x_m$ in $\mathbb{R}^r$, such that $\mathcal{F} = DX$ with $\|x_i\|_0 \le n$, for $i = 1, \ldots, m$.

Now we will introduce a definition of sparsity and show its connection with Problem 2.

Definition 4.3 Let n, l, m be positive integers, with $n, l < m$.

Given a set of vectors $\mathcal{F} = \{f_1, \ldots, f_m\}$ in $\mathbb{R}^N$ and a real number $\varepsilon \ge 0$, we will say that the data $\mathcal{F}$ is (l, n, ε)-*sparse* if there exist subspaces $V_1, \ldots, V_l$, of $\mathbb{R}^N$ with $\dim(V_i) \le n$ for $i = 1, \ldots, l$, such that

$$e(\mathcal{F}, \{V_1, \ldots, V_l\}) = \sum_{i=1}^{m} \min_{1 \le j \le l} d^2(f_i, V_j) \le \varepsilon. \tag{4.4}$$

When $\mathcal{F}$ is $(l, n, 0)$-sparse, we will simply say that $\mathcal{F}$ is (l, n)-sparse. We will also say that the data is ε-sparse if the values of l and n are clear from the context.

Note that if $\mathcal{F}$ is (l, n, ε)-sparse, then it is also (l, n, η)-sparse for every $\eta \geq \varepsilon$. So, usually it is interesting to know the minimun ε such that the data is ε-sparse. The above definition of sparsity is an intrinsic property of the data and the Hilbert space in which the data lives, and does not depend on any specific dictionary.

4.2.1 The case $\varepsilon = 0$

Let us now consider the case $\varepsilon = 0$.

If the data $\mathcal{F}$ is (l, n)-sparse and $V_1^0, \ldots, V_l^0$ are optimal spaces (that is when $e(\mathcal{F}, \{V_1^0, \ldots, V_l^0\}) = 0$ and $\dim(V_i^0) \leq n$ for $i = 1, \ldots, l$) then each $f \in \mathcal{F}$ belongs to some of the spaces $\{V_i^0\}_{i=1,\ldots,l}$.

For each $i = 1, \ldots, l$, let us call $r_i = \dim(V_i^0)$ and let $\{w_{i1}, \ldots, w_{ir_i}\}$ be an orthonormal basis of V_i^0. Define

$$D = \bigcup_{i=1}^{l} \{w_{ij} : 1 \leq j \leq r_i\} \subset \mathbb{R}^N.$$

The vectors in D have the property that each $f \in \mathcal{F}$ can be written as a linear combination of at most n elements in D. In other words, we have found a dictionary D such that it solves Problem 2, for the data $\mathcal{F}$ and length s with $s = r_1 + \cdots + r_l \leq ln$.

So Theorem 4.2 provides a solution to Problem 2 with a dictionary of length at most ln, in case that the data is (l, n)-sparse. We will see below that if the data $\mathcal{F}$ is not (l, n)-sparse, then Theorem 4.2, provides the minimum ε such that the data is (l, n, ε)-sparse.

Note that if the basis of each V_i^0 is properly chosen then in many cases the number of atoms in the dictionary can be reduced, due to the fact that the subspaces can have non-trivial intersections.

So, as before, let $V_1^0, \ldots, V_l^0$ be optimal spaces, and let $\mathcal{U} = \{u_1, \ldots, u_s\}$ be a set of vectors with the property that for each $i \in \{1, \ldots, l\}$ there is a subset $\mathcal{U}_i \subset \mathcal{U}$ such that $\operatorname{span}(\mathcal{U}_i) = V_i^0$. Set $D_0 = \{w_1, \ldots, w_{s_0}\}$ to be a minimal set with this property.

Then Theorem 4.2 implies that D_0 is a dictionary that solves Problem 2 for data $\mathcal{F}$ and positive integers n and $d = ln$. We want to remark here that a minimal set is not a linearly independent set in general. It is not difficult to see that considering all possible intersections of the subspaces V_i^0, a minimal set can be constructed.

4.2.2 The case $\varepsilon > 0$

If the data is not (l, n)-sparse, then Theorem 4.2 implies that there is no dictionary D of length $d = ln$ or smaller that solves Problem 2, for the data $\mathcal{F}$.

If we still want to find a dictionary of length no larger than ln with $\|x_i\|_0 \leq n$ for $i = 1, \ldots, n$, then the question is: what error do we have to allow in order to have a solution? In other words, what is the minimum ε such that the data $\mathcal{F}$ is (l, n, ε)-sparse? This question gives rise to the following extension of Problem 2.

Problem 3 *Let $\mathcal{F} = \{f_1, \ldots, f_m\}$ in $\mathbb{R}^N$, and n, d positive integers. With the same notation as in Problem 2, given $\varepsilon \geq 0$, find a dictionary D with no more than d atoms*

and a matrix X such that

$$\|\mathcal{F} - DX\| \leq \varepsilon$$

with $\|x_i\|_0 \leq n$, *for* $i = 1, \ldots, m$.

Theorem 4.2 provides in this case the solution for the minimum possible error and establishes the exact value of the error. More precisely, let $V_1^0, \ldots, V_l^0$ be a bundle of optimal subspaces and let $\varepsilon = e(\mathcal{F}, \{V_1^0, \ldots, V_l^0\})$. Let us choose as before a minimal set $D_0 = \{w_1, \ldots, w_{s_0}\}$ for that solution, then we have that there exist vectors $x_1, \ldots, x_m$ in R^{s_0} such that

$$\|\mathcal{F} - D_0 X\| = \varepsilon$$

with $\|x_i\|_0 \leq n$, for $i = 1, \ldots, m$. Furthermore, given any $N \times r$ matrix D with $r \leq s_0$ and any matrix X with columns $x_1, \ldots, x_m$ in $\mathbb{R}^r$ and $\|x_i\|_0 \leq n$ for $i = 1, \ldots, m$, we have

$$\|\mathcal{F} - DX\| \geq \varepsilon.$$

So, a solution of Problem 1 gives an optimal solution for Problem 3 and finds the exact (optimal) ε-sparsity.

Note that when n or l increase then in general the minimum error will decrease. It is also important to emphasize here that some subspaces from an optimal bundle for the data $\mathcal{F}$ and (l, n) can have dimension zero, so in applications these subspaces can be removed and the non-trivial subspaces will produce the same error. Thus, the subspaces that are found are not restricted to be orthogonal, or with equal dimensions or with trivial intersection, and there can be any number of subspaces up to a prescribed number l, and each subspace can be of any dimension up to a prescribed number n.

5 Search Algorithm

Although Theorem 2.2 establishes the existence of a global minimizer solution to Problem 1, an exhaustive search over all possible partitions is not feasible in practice and a search algorithm is needed. Lemma 2.1 used in the proof of Theorem 2.2 suggests an iterative search algorithm that we will present in this section, and we will show that the algorithm always terminates in finitely many steps. The search algorithm for finding the solution to Problem 1 is given in the program below, with the notation of Sect. 2. It uses two choice functions G, H, where G is a choice function assigning $P \longmapsto \mathbf{V}_P \in \mathcal{W}(P)$, and H is a choice function assigning $\mathbf{V} \longmapsto P \in \Omega_l(\mathbf{V})$.

Algorithm

(1) Pick any partition $P_1 \in \Pi_l$;
(2) Find and choose $\mathbf{V}_{P_1} = \mathrm{G}(P_1) \in \mathcal{W}(P_1) \subset \mathfrak{B}$ by minimizing $\Gamma(P_1, \mathbf{V})$ over $\mathbf{V} \in \mathfrak{B}$;
(3) Set $j = 1$;
(4) **While** $\Gamma(P_j, \mathbf{V}_{P_j}) > e(\mathcal{F}, \mathbf{V}_{P_j})$;
(5) Choose a new partition $P_{j+1} = \mathrm{H}(\mathbf{V}_{P_j}) \in \Omega_l(\mathbf{V}_{P_j})$ associated to $\mathbf{V}_{P_j}$;

(6) Find and choose $\mathbf{V}_{P_{j+1}} = \mathrm{G}(P_j) \in \mathcal{W}(P_j)$, by minimizing $\Gamma(P_{j+1}, \mathbf{V})$ over $\mathbf{V} \in \mathfrak{B}$;
(7) Increase j by 1, i.e., $j \to j+1$;
(8) **End while**

Note that this algorithm, starting from a bundle $\mathbf{V}_{P_1}$ in step (2), produces a sequence of bundles $\mathbf{V}_{P_1}, \mathbf{V}_{P_2}, \mathbf{V}_{P_3}, \ldots$ with the property that $e(\mathcal{F}, \mathbf{V}_{P_1}) \geq e(\mathcal{F}, \mathbf{V}_{P_2}) \geq e(\mathcal{F}, \mathbf{V}_{P_3}) \geq \cdots$. The algorithm stops precisely when for some $j \geq 1$ $e(\mathcal{F}, \mathbf{V}_{P_j}) = e(\mathcal{F}, \mathbf{V}_{P_{j+1}})$. We will now see that the algorithm terminates in finitely many steps.

Proof We first note that if $\Gamma(P_j, \mathbf{V}_{P_j}) > e(\mathcal{F}, \mathbf{V}_{P_j})$, then $P_{j+1} \notin \Omega_l(V_{P_j})$. To see this, we argue by contradiction: if $P_{j+1} \in \Omega_l(V_{P_j})$, then $\Gamma(P_{j+1}, \mathbf{V}_{P_{j+1}}) = \Gamma(P_j, \mathbf{V}_{P_j})$. But we have that

$$\Gamma(P_{j+1}, \mathbf{V}_{P_{j+1}}) \leq e(\mathcal{F}, \mathbf{V}_{P_j}) < \Gamma(P_j, \mathbf{V}_{P_j}),$$

which is a contradiction.

Therefore, since the set of partitions is finite, the algorithm must stop in at most $\#\Pi_l$ steps. The algorithm terminates when $\Gamma(P_{end}, \mathbf{V}_{P_{end}}) = e(\mathcal{F}, \mathbf{V}_{P_{end}})$. □

Note 1 A partition P_m such that $\Gamma(P_m, \mathbf{V}_{P_m}) = e(\mathcal{F}, \mathbf{V}_{P_m})$ will be called a *minimal partition* (see remark below).

Remark

(1) The algorithm can be formulated as a search for minimal partitions in the partially ordered set $(\Pi_l, \preceq)$, where the order of the elements in Π_l depends on the specific choice functions G, H in (2), (5) and (6) of the algorithm. Specifically, $P \preceq Q$ if there exists an integer $s \geq 1$ such that $P = (\mathrm{H}\,\mathrm{G})^s Q$. Since $(\Pi_l, \preceq)$ is a partially ordered set with finitely many elements, a nonempty set of minimal partitions $\mathcal{M} \subset \Pi_l$ exists.
(2) The algorithm will always terminate in finitely many steps but the bundle $\mathbf{V}_{P_m}$ associated to the final minimal partition P_m may not be the global minimizer of Problem 1. The algorithm can be viewed as a search in a directed graph whose vertices are the partitions. Each iteration moves from one partition to the next via a directed edge. If the graph has a single component, then the algorithm will always end at a partition whose associated bundle is a global minimizer. However, if the graph has more than one component, then the algorithm will end up at a partition, whose associated bundle is not necessarily the global minimizer.
(3) In the search algorithm, steps (2) and (6) must be implemented by some other minimizing algorithms. For the two cases that we studied in this paper, a space $\mathbf{V}_{P_{j+1}}$ that minimizes $\Gamma(P_{j+1}, \mathbf{V})$ over $\mathbf{V} \in \mathfrak{B}$ can be explicitly found and computed, by the Eckhard-Young Theorem for subspaces of $\mathcal{H} = \mathbb{R}^N$, and by Theorem 2.1 ([4]) for shift-invariant spaces of $\mathcal{H} = L^2(\mathbb{R}^d)$. Both methods are based on the Singular Value Decomposition, they are easily implemented, and all the approximation errors can be computed exactly.

(4) In searching for a dictionary with d atoms such that each data point $f \in \mathbb{R}^N$ from a set of data $\mathcal{F} = \{f_1, \ldots, f_m\}$ is approximated by a single atom, our method coincides with the K-SVD algorithm proposed in ([1, 2]) and produces the same dictionary if steps (2) and (6) are implemented using the SVD. However, for the case where each data point is approximated by a linear combination of $n > 1$ atoms, the two methods are not comparable, even if steps (2) and (6) are implemented using the SVD.

6 Conclusions

Theorem 2.2 can be viewed as a way of finding an optimal (generally non-linear) signal model of the form $\bigcup_{i=1}^{l} S_i$ from some observed data. For example, the application of Theorem 2.2 to shift-invariant spaces in Sect. 3.1 produced Theorem 3.3, which gives the optimal signal model of at most l shift-invariant spaces compatible with the observed data. The resulting solution is an optimal bundle $S_0 = \{S_1^0, \ldots, S_l^0\}$ that consists of a finite sequence of infinite dimensional spaces, such that each space S_i^0 of this sequence is generated by the integer translates of at most n generators. This type of best signal model $\bigcup_{i \in I} S_i^0$ derived from a set of observed data $\mathcal{F} = \{f_1, \ldots, f_m\} \subset L^2$ may be used for example in sampling and reconstruction as well as other applications.

If applied to $\mathbb{R}^N$, Theorem 2.2, can be viewed as a way of finding an optimal sparse representation of the data $\mathcal{F} \subset \mathbb{R}^N$, optimal dictionaries, and subspace segmentation in the presence of noise as discussed in Sect. 4.2. Specifically, the application of Theorem 2.2 to the finite dimensional case $\mathbb{R}^N$ produces Theorem 4.2 which gives the optimal ε-sparse representation of the data as discussed in Sect. 4.2. In particular, Theorem 4.2 proves the existence of a dictionary with minimal error and minimal length for sparse data representation.

One contribution of this work is that it unifies and complements some of the new non-linear techniques used in sampling theory, the Generalized Principle Components Analysis, and the dictionary design problem. However, there are still many questions that need to be addressed before this methodology becomes applicable. For example, the algorithm proposed in the last section may end up at a local minimum which is not a global minimum. The termination of the algorithm depends on the initial condition. Thus, an interesting question is to estimate the number of minima in terms of some characteristics of $\mathcal{C}$ and $\mathcal{F}$. Another interesting question is to estimate the speed at which the algorithm converges in terms of some characteristics of $\mathcal{C}$ and $\mathcal{F}$. Testing the dependence of algorithm on the initial partition, the data, and noise level, for the case $\mathcal{H} = \mathbb{R}^N$ and $\mathcal{C} = \mathfrak{L}_n$ using an SVD implementation for steps (2) and (6) is also important for future research, and applications.

References

1. Aharon, M., Elad, M., Bruckstein, A.M.: The k-svd: An algorithm for designing of overcomplete dictionaries for sparse representation. IEEE Trans. Signal Process. **54**(11), 4311–4322 (2006)

2. Aharon, M., Elad, M., Bruckstein, A.M.: On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them. Linear Algebra Appl. **416**(1), 48–67 (2006). MR2232919 (2007a:94026)
3. Aldroubi, A., Gröchenig, K.-H.: Non-uniform sampling in shift-invariant space. SIAM Rev. **43**(4), 585–620 (2001)
4. Aldroubi, A., Cabrelli, C.A., Hardin, D., Molter, U.M.: Optimal shift invariant spaces and their parseval frame generators. Appl. Comput. Harmon. Anal. **23**, 273–283 (2007)
5. Aldroubi, A., Cabrelli, C., Hardin, D.P., Molter, U., Rodado, E.: Determining sets of shift invariant spaces. In: Proceedings of ICWA (Chenai, India), 2003
6. Baraniuk, R., Davenport, M., DeVore, R.A., Wakin, M.: A simple proof of the restricted isometry property for random matrices. Constr. Approx. (2007). doi:10.1007/s00365-007-9003-x
7. Bownik, M.: The structure of shift-invariant subspaces of $L^2(\mathbb{R}^n)$. J. Funct. Anal. **177**, 282–309 (2000)
8. Candès, E., Romberg, J.: Quantitative robust uncertainty principles and optimally sparse decompositions. Found. Comput. Math. **6**, 227–254 (2006)
9. Candès, E., Tao, T.: Near optimal signal recovery from random projections: Universal encoding strategies. IEEE Trans. Inf. Theory **52**, 5406–5425 (2006)
10. Candès, E., Romberg, J., Tao, T.: Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory **52**, 489–509 (2006)
11. Casazza, P.G.: The art of frame theory. Taiwan. J. Math. **4**(2), 129–201 (2000)
12. Christensen, O.: An Introduction to Frames and Riesz Basis. Applied and Numerical Harmonic Analysis. Birkhäuser, Basel (2003)
13. DeVore, R.A.: Deterministic constructions of compressed sensing matrices. J. Complex. **23**, 918–925 (2007)
14. Donoho, D.L.: Compressed sensing. IEEE Trans. Inf. Theory **52**, 1289–1306 (2006)
15. Dragotti, P.L., Vetterli, M., Blu, T.: Sampling moments and reconstructing signals of finite rate of innovation: Shannon meets strang-fix. IEEE Trans. Signal Process. **55**, 1741–1757 (2007)
16. Eckart, C., Young, G.: The approximation of one matrix by another of lower rank. Psychometrica **1**, 211–218 (1936)
17. Gribonval, R., Nielsen, M.: Sparse decompositions in unions of bases. IEEE Trans. Inf. Theory **49**, 3320–3325 (2003)
18. Gröchenig, K.: Foundations of Time-Frequency Analysis. Appl. Numer. Harmon. Anal. Birkhäuser, Basel (2001)
19. Hernández, E., Weiss, G.: A First Course on Wavelets. CRC Press, Boca Raton (1996)
20. Horn, R., Johnson, C.: Matrix Analysis. Cambridge University Press, Cambridge (1985)
21. Lu, Y., Do, M.N.: A theory for sampling signals from a union of subspaces. IEEE Trans. Signal Process. **56**, 2334–2345 (2008)
22. Ma, Y., Derksen, H., Hong, W., Wright, J.: Segmentation of multivariate mixed data via lossy coding and compression. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **29**(9), 1546–1562 (2007)
23. Ma, Y., Yang, A., Derksen, H., Fossum, R.: Estimation of subspace arrangements with applications in modeling and segmenting mixed data. SIAM Rev. **50**, 413–458 (2008)
24. Maravic, I., Vetterli, M.: Sampling and reconstruction of signals with finite rate of innovation in the presence of noise. IEEE Trans. Signal Process. **53**, 2788–2805 (2005)
25. Rauhut, H., Schass, K., Vandergheynst, P.: Compressed sensing and redundant dictionaries. IEEE Trans. Inf. Theory **4**, 2210–2219 (2008)
26. Schmidt, E.: Zur theorie der linearen und nichtlinearen integralgleichungen. i teil. entwicklung willkürlichen funktionen nach system vorgeschriebener. Math. Ann. **63**, 433–476 (1907)
27. Tropp, J.A.: Greed is good: Algorithmic results for sparse approximation. IEEE Trans. Inf. Theory **50**, 2231–2242 (2004)
28. Vidal, R., Ma, Y., Sastry, S.: Generalized principal component analysis (gpca). IEEE Trans. Pattern Anal. Mach. Intell. **27**, 1–15 (2005)