1   Data Separation by Sparse Representations

Gitta Kutyniok

Universität Osnabrück

Modern data are often composed of two or more morphologically distinct constituents, and one typical goal is the extraction of those components. Recently, sparsity methodologies have been successfully utilized to solve this problem, both theoretically as well as empirically. The key idea is to choose a deliberately overcomplete representation made of several frames each one providing a sparse expansion of one of the components to be extracted. The morphological difference between the components is then encoded as incoherence conditions of those frames. The decomposition principle is to minimize the ℓ_1 norm of the frame coefficients. This chapter shall serve as an introduction to and a survey about this exciting area of research as well as a reference for the state-of-the-art of this research field.

1.1   Introduction

Over the last years, scientists face an ever growing deluge of data, which needs to be transmitted, analyzed, and stored. A close analysis reveals that most of these data might be classified as multimodal data, i.e., being composed of distinct subcomponents. Prominent examples are audio data, which might consist of a superposition of the sounds of different instruments, or imaging data from neurobiology, which is typically a composition of the soma of a neuron, its dendrites, and its spines. In both these exemplary situations, the data has to be separated into appropriate single components for further analysis. In the first case, separating the audio signal into the signals of the different instruments is a first step to enable the audio technician to obtain a musical score from a recording. In the second case, the neurobiologist might aim to analyze the structure of dendrites and spines separately for the study of Alzheimer specific characteristics. Thus data separation is often a crucial step in the analysis of data.

As a scientist, three fundamental problems immediately come to one's mind:

(P1) What is a mathematically precise meaning of the vague term 'distinct components'?

(P2) How do we separate data algorithmically?

(P3) When is separation possible at all?

To answer those questions, we need to first understand the key problem in data separation. In a very simplistic view, the essence of the problem is as follows: Given a composed signal x of the form $x = x_1 + x_2$, we aim to extract the unknown components x_1 and x_2 from it. Having one known data and two unknowns obviously makes this problem underdetermined. Thus, the novel paradigm of sparsity – appropriately utilized – seems a perfect fit for attacking data separation, and this chapter shall serve as both an introduction into this intriguing application of sparse representations as well as a reference for the state-of-the-art of this research area.

1.1.1   Morphological Component Analysis

Intriguingly, when considering the history of Compressed Sensing, the first mathematically precise result on recovery of sparse vectors by ℓ_1 minimization is related to a data separation problem: The separation of sinusoids and spikes in [16, 11]. Thus it might be considered a milestone in the development of Compressed Sensing. In addition, it reveals a surprising connection with uncertainty principles.

The general idea allowing separation in [16, 11] was to choose two bases or frames Φ_1 and Φ_2 adapted to the two components to be separated in such a way that Φ_1 and Φ_2 provide a sparse representation for x_1 and x_2, respectively. Searching for the sparsest representation of the signal in the combined (highly overcomplete) dictionary $[\,\Phi_1\,|\,\Phi_2\,]$ should then intuitively enforce separation provided that x_1 does not have a sparse representation in Φ_2 and that x_2 does not have a sparse representation in Φ_1. This general concept was later – in the context of image separation, but the term seems to be fitting in general – coined *Morphological Component Analysis* [36].

This viewpoint now measures the morphological difference between components in terms of the incoherence of suitable sparsifying bases or frames Φ_i, thereby giving one possible answer to (P1); see also the respective chapters in the novel book [33]. As already debated in the introduction [40], one possibility for measuring incoherence is the *mutual coherence*. We will however see in the sequel that there exist even more appropriate coherence notions, which provide a much more refined measurement of incoherence specifically adapted to measuring morphological difference.

1.1.2   Separation Algorithms

Going again back in time, we observe that far before [11], Coifman, Wickerhauser, and co-workers already presented very inspiring empirical results on the separation of image components using the idea of Morphological Component Analysis, see [7]. After this, several techniques to actually compute the sparsest expansion in a composed dictionary $[\,\Phi_1\,|\,\Phi_2\,]$ were introduced. In [31], Mallat and Zhang developed *Matching Pursuit* as one possible methodology. The study by Chen,

Donoho, and Saunders in [6] then revealed that the ℓ_1 norm has a tendency to find sparse solutions when they exist, and coined this method *Basis Pursuit.*

As explained before, data separation by Morphological Component Analysis – when suitably applied – can be reduced to a sparse recovery problem. To solve this problem, there nowadays already exist a variety of utilizable algorithmic approaches; thereby providing a general answer to (P2). Such approaches include, for instance, a canon of greedy-type algorithms, and we refer to the introduction [40] and chapter [42] for further details. Most of the theoretical separation results however consider ℓ_1 minimization as the main separation technique, which is what we will also mainly focus on in this chapter.

1.1.3   Separation Results

As already mentioned, the first mathematically precise result was derived in [11] and solved the problem of separation of sinusoids and spikes. After this 'birth of sparse data separation', a deluge of very exciting results started. One direction of research are general results on sparse recovery and Compressed Sensing; here we would like to cite the excellent survey paper [4] and the introduction [40].

Another direction continued the idea of sparse data separation initiated in [11]. In this realm, the most significant theoretical results might be considered firstly the series of papers [19, 10], in which the initial results from [11] are extended to general composed dictionaries, secondly the paper [23], which also extends results from [11] though with a different perspective, and thirdly the papers [3] and [14], which explore the clustering of the sparse coefficients and the morphological difference of the components encoded in it.

We also wish to mention the abundance of empirical work showing that utilizing the idea of sparse data separation often gives very compelling results in practice, as examples, we refer to the series of papers on applications to astronomical data [2, 36, 34], to general imaging data [32, 20, 35], and to audio data [22, 25].

Let us remark that also the classical problem of denoising can be regarded as a separation problem, since we aim to separate a signal from noise by utilizing the characteristics of the signal family and the noise. However, as opposed to the separation problems discussed in this chapter, denoising is not a 'symmetric' separation task, since the characterization of the signal and the noise are very different.

1.1.4   Design of Sparse Dictionaries

For satisfactorily answering (P3), one must also raise the question of how to find suitable sparsifying bases or frames for given components. This search for 'good' systems in the sense of sparse dictionaries can be attacked in two ways, either non-adaptively or adaptively.

The first path explores the structure of the component one would like to extract, for instance, it could be periodic such as sinusoids or anisotropic such as edges in images. This typically allows one to find a suitable system among the already very well explored representation systems such as the Fourier basis, wavelets, or shearlets, to name a few. The advantage of this approach is the already explored structure of the system, which can hence be exploited for deriving theoretical results on the accuracy of separation, and the speed of associated transforms.

The second path uses a training set of data similar to the to-be-extracted component, and 'learns' a system which best sparsifies this data set. Using this approach customarily referred to as *dictionary learning*, we obtain a system extremely well adapted to the data at hand; as the state-of-the-art we would like to mention the K-SVD algorithm introduced by Aahron, Elad, and Bruckstein in [1]; see also [17] for a 'Compressed Sensing' perspective to K-SVD. Another appealing dictionary training algorithm, which should be cited is the method of optimal directions (MOD) by Engan et al. [21]. The downside however is the lack of a mathematically exploitable structure, which makes a theoretical analysis of the accuracy of separation using such a system very hard.

1.1.5   Outline

In Section 1.2, we discuss the formal mathematical setting of the problem, present the nowadays already considered classical separation results, and then discuss more recent results exploiting the clustering of significant coefficients in the expansions of the components as a means to measure their morphological difference. We conclude this section by revealing a close link of data separation to uncertainty principles. Section 1.3 is then devoted to both theoretical results as well as applications for separation of 1D signals, elaborating, in particular, on the separation of sinusoids and spikes. Finally, Section 1.4 focuses on diverse questions concerning separation of 2D signals, i.e., images, such as the separation of point- and curvelike objects, again presenting both application aspects as well as theoretical results.

1.2   **Separation Estimates**

As already mentioned in the introduction, data separation can be regarded within the framework of underdetermined problems. In this section, we make this link mathematically precise. Then we discuss general estimates on the separability of composed data, firstly without any knowledge of the geometric structure of sparsity patterns, and secondly, by taking known geometric information into account. A revelation of the close relation with uncertainty principles concludes the section.

In Sections 1.3 and 1.4, we will then see the presented general results and uncertainty principles in action, i.e., applied to real-world separation problems.

1.2.1   Relation with Underdetermined Problems

Let x be our signal of interest, which we for now consider as belonging to some Hilbert space $\mathcal{H}$, and assume that

$$x = x_1^0 + x_2^0.$$

Certainly, real data are typically composed of multiple components, hence not only the situation of two components, but three or more is of interest. We will however focus on the two-component situation to clarify the fundamental principles behind the success of separating those by sparsity methodologies. It should be mentioned though that, in fact, most of the presented theoretical results can be extended to the multiple component situation in a more or less straightforward manner.

To extract the two components from x, we need to assume that – although we are not given x_1^0 and x_2^0 – certain 'characteristics' of those components are known to us. Such 'characteristics' might be, for instance, the pointlike structure of stars and the curvelike structure of filaments in astronomical imaging. This knowledge now enables us to choose two representation systems, Φ_1 and Φ_2, say, which allow sparse expansions of x_1^0 and x_2^0, respectively. Such representation systems might be chosen from the collection of well-known systems such as wavelets. A different possibility is to choose adaptively the systems via dictionary learning procedures. This approach however requires training data sets for the two components x_1^0 and x_2^0 as discussed in Subsection 1.1.4.

Given now two such representation systems Φ_1 and Φ_2, we can write x as

$$x = x_1^0 + x_2^0 = \Phi_1 c_1^0 + \Phi_2 c_2^0 = [\, \Phi_1 \mid \Phi_2 \,] \begin{bmatrix} c_1^0 \\ c_2^0 \end{bmatrix}$$

with $\|c_1^0\|_0$ and $\|c_2^0\|_0$ 'sufficiently small'. Thus, the data separation problem has been reduced to solving the underdetermined linear system

$$x = [\, \Phi_1 \mid \Phi_2 \,] \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} \tag{1.1}$$

for $[c_1, c_2]^T$. Unique recovery of the original vector $[c_1^0, c_2^0]^T$ automatically extracts the correct two components x_1^0 and x_2^0 from x, since

$$x_1^0 = \Phi_1 c_1^0 \quad \text{and} \quad x_2^0 = \Phi_2 c_2^0.$$

Ideally, one might want to solve

$$\min_{c_1, c_2} \|c_1\|_0 + \|c_2\|_0 \quad \text{s.t.} \quad x = [\, \Phi_1 \mid \Phi_2 \,] \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}, \tag{1.2}$$

which however is an NP-hard problem. As already discussed in the introduction [40], instead one aims to solve the ℓ_1 minimization problem

$$(\mathrm{Sep}_s) \qquad \min_{c_1,c_2} \|c_1\|_1 + \|c_2\|_1 \quad \text{s.t.} \quad x = [\,\Phi_1 \mid \Phi_2\,] \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}. \qquad (1.3)$$

The lower case 's' in Sep_s indicates that the ℓ_1 norm is placed on the synthesis side. Other choices for separation are, for instance, greedy-type algorithms – and we refer the reader to the introduction [40] and chapter [42]. In this chapter we will focus on ℓ_1 minimization as the separation technique, consistent with most known separation results from the literature.

Before discussing conditions on $[c_1^0, c_2^0]^T$ and $[\,\Phi_1 \mid \Phi_2\,]$, which guarantee unique solvability of (1.1), let us for a moment debate whether uniqueness is necessary at all. If Φ_1 and Φ_2 form bases, it is certainly essential to recover $[c_1^0, c_2^0]^T$ uniquely from (1.1). However, some well-known representation systems are in fact redundant and typically constitute Parseval frames such as curvelets or shearlets. Also, systems generated by dictionary learning are normally highly redundant. In this situation, for each possible separation

$$x = x_1 + x_2, \qquad (1.4)$$

there exist infinitely many coefficient sequences $[c_1, c_2]^T$ satisfying

$$x_1 = \Phi_1 c_1 \quad \text{and} \quad x_2 = \Phi_2 c_2. \qquad (1.5)$$

Since we are *only* interested in the correct separation and *not* in computing the sparsest expansion, we can circumvent presumably arising numerical instabilities when solving the minimization problem (1.3) by selecting a particular coefficient sequence for each separation. Assuming Φ_1 and Φ_2 are Parseval frames, we can exploit this structure and rewrite (1.5) as

$$x_1 = \Phi_1(\Phi_1^T x_1) \quad \text{and} \quad x_2 = \Phi_2(\Phi_2^T x_2).$$

Thus, for each separation (1.4), we choose a *specific* coefficient sequence when expanding the components in the Parseval frames, in fact, we choose the *analysis sequence*. This leads to the following different ℓ_1 minimization problem in which the ℓ_1 norm is placed on the *analysis* rather than the *synthesis* side:

$$(\mathrm{Sep}_a) \qquad \min_{x_1,x_2} \|\Phi_1^T x_1\|_1 + \|\Phi_2^T x_2\|_1 \quad \text{s.t.} \quad x = x_1 + x_2. \qquad (1.6)$$

This new minimization problem can be also regarded as a mixed ℓ_1-ℓ_2 problem, since the analysis coefficient sequence is exactly the coefficient sequence which is minimal in the ℓ_2 norm. For more information, we refer to chapter [41].

1.2.2   General Separation Estimates

Let us now discuss the main results of successful data separation, i.e., stating conditions on $[c_1^0, c_2^0]^T$ and $[\,\Phi_1 \mid \Phi_2\,]$ for extracting x_1^0 and x_2^0 from x. The strongest known general result was derived in 2003 by Donoho and Elad [10] and used the

notion of mutual coherence. Recall that, for a normalized frame $\Phi = (\varphi_i)_{i \in I}$, the *mutual coherence* of Φ is defined by

$$\mu(\Phi) = \max_{i,j \in I, i \neq j} |\langle \varphi_i, \varphi_j \rangle|. \tag{1.7}$$

We remark that the result by Donoho and Elad was already stated in the introduction [40], however without a proof, which we will present in this section. For the convenience of the reader, we recall the relevant result.

Theorem 1.1 ([10]). *Let Φ_1 and Φ_2 be two frames for a Hilbert space $\mathcal{H}$, and let $x \in \mathcal{H}$, $x \neq 0$. If $x = [\Phi_1|\Phi_2]c$ and*

$$\|c\|_0 < \frac{1}{2} \left(1 + \frac{1}{\mu([\Phi_1|\Phi_2])} \right),$$

then the solution of the ℓ_1 minimization problem (Sep_s) stated in (1.3) coincides with the solution of the ℓ_0 minimization problem stated in (1.2).

Before presenting the proof, we require some prerequisites. Firstly, we need to introduce a slightly stronger version of the null space property than the one discussed in the introduction [40].

Definition 1.1. *Let $\Phi = (\varphi_i)_{i \in I}$ be a frame for a Hilbert space $\mathcal{H}$, and let $\mathcal{N}(\Phi)$ denote the null space of Φ. Then Φ is said to have the* null space property of order k *if*

$$\|1_\Lambda d\|_1 < \frac{1}{2} \|d\|_1$$

for all $d \in \mathcal{N}(\Phi) \setminus \{0\}$ and for all sets $\Lambda \subseteq I$ with $|\Lambda| \leq k$.

This notion provides a very useful characterization of the existence of unique sparse solutions of the ℓ_1 minimization problem (Sep_s) stated in (1.3).

Lemma 1.1. *Let $\Phi = (\varphi_i)_{i \in I}$ be a frame for a Hilbert space $\mathcal{H}$, let $x \in \mathcal{H}$, and let $k \in \mathbb{N}$. Then the following conditions are equivalent.*

(i) All vectors c with $\|c\|_0 \leq k$ are unique solutions of the ℓ_1 minimization problem (Sep_s) stated in (1.3) (with Φ instead of $[\Phi_1|\Phi_2]$).
(ii) Φ satisfies the null space property of order k.

Proof. First, assume that (i) holds. Let $d \in \mathcal{N}(\Phi) \setminus \{0\}$ and $\Lambda \subseteq I$ with $|\Lambda| \leq k$ be arbitrary. Then, by (i), the sparse vector $1_\Lambda d$ is the unique minimizer of $\|c\|_1$ subject to $\Phi c = \Phi(1_\Lambda d)$. Further, since $d \in \mathcal{N}(\Phi) \setminus \{0\}$,

$$\Phi(-1_{\Lambda^c} d) = \Phi(1_\Lambda d).$$

Hence

$$\|1_\Lambda d\|_1 < \|1_{\Lambda^c} d\|_1,$$

or, in other words,

$$\|1_\Lambda d\|_1 < \frac{1}{2}\|d\|_1,$$

which implies (ii), since d and Λ were chosen arbitrarily.

Secondly, assume that (ii) holds, and let c_1 be a vector satisfying $x = \Phi c_1$ with $\|c_1\|_0 \leq k$ and support denoted by Λ. Further, let c_2 be an arbitrary solution of $x = \Phi c$, and set

$$d = c_2 - c_1.$$

Then

$$\|c_2\|_1 - \|c_1\|_1 = \|1_{\Lambda^c} c_2\|_1 + \|1_\Lambda c_2\|_1 - \|1_\Lambda c_1\|_1 \geq \|1_{\Lambda^c} d\|_1 - \|1_\Lambda d\|_1.$$

This term is greater than zero for any $d \neq 0$ if

$$\|1_{\Lambda^c} d\|_1 > \|1_\Lambda d\|_1,$$

or

$$\frac{1}{2}\|d\|_1 > \|1_\Lambda d\|_1.$$

This is ensured by (ii). Hence $\|c_2\|_1 > \|c_1\|_1$, and thus c_1 is the unique solution of (Sep_s). This implies (i). $\qquad\blacksquare$

Using this result, we next prove that a solution satisfying $\|c\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\Phi)}\right)$ is the unique solution of the ℓ_1 minimization problem (Sep_s).

Lemma 1.2. *Let $\Phi = (\varphi_i)_{i \in I}$ be a frame for a Hilbert space $\mathcal{H}$, and let $x \in \mathcal{H}$. If c is a solution of the ℓ_1 minimization problem (Sep_s) stated in (1.3) (with Φ instead of $[\Phi_1|\Phi_2]$) and satisfies*

$$\|c\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\Phi)}\right),$$

then it is the unique solution.

Proof. Let $d \in \mathcal{N}(\Phi) \setminus \{0\}$, hence, in particular,

$$\Phi d = 0;$$

thus also

$$\Phi^\star \Phi d = 0. \qquad\qquad (1.8)$$

Without loss of generality, we now assume that the vectors in Φ are normalized. Then, (1.8) implies that, for all $i \in I$,

$$d_i = -\sum_{j \neq i} \langle \varphi_i, \varphi_j \rangle d_j.$$

Using the definition of mutual coherence $\mu(\Phi)$ (cf. [40] or (1.7)), we obtain

$$|d_i| \leq \sum_{j \neq i} |\langle \varphi_i, \varphi_j \rangle| \cdot |d_j| \leq \mu(\Phi)(\|d\|_1 - |d_i|),$$

and hence

$$|d_i| \leq \left(1 + \frac{1}{\mu(\Phi)}\right)^{-1} \|d\|_1.$$

Thus, by the hypothesis on $\|c\|_0$ and for any $\Lambda \subseteq I$ with $|\Lambda| = \|c\|_0$, we have

$$\|1_\Lambda d\|_1 \leq |\Lambda| \cdot \left(1 + \frac{1}{\mu(\Phi)}\right)^{-1} \|d\|_1 = \|c\|_0 \cdot \left(1 + \frac{1}{\mu(\Phi)}\right)^{-1} \|d\|_1 < \frac{1}{2}\|d\|_1.$$

This shows that Φ satisfies the null space property of order $\|c\|_0$, which, by Lemma 1.1, implies that c is the unique solution of (Sep_s). $\square$

We further prove that a solution satisfying $\|c\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\Phi)}\right)$ is also the unique solution of the ℓ_0-minimization problem.

Lemma 1.3. *Let $\Phi = (\varphi_i)_{i \in I}$ be a frame for a Hilbert space $\mathcal{H}$, and let $x \in \mathcal{H}$. If c is a solution of the ℓ_0 minimization problem stated in (1.2) (with Φ instead of $[\Phi_1|\Phi_2]$) and satisfies*

$$\|c\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\Phi)}\right),$$

then it is the unique solution.

Proof. By Lemma 1.2, the hypotheses imply that c is the unique solution of the ℓ_1 minimization problem (Sep_s). Now, towards a contradiction, assume that there exists some $\tilde{c}$ satisfying $x = \Phi\tilde{c}$ with $\|\tilde{c}\|_0 \leq \|c\|_0$. Then $\tilde{c}$ must satisfy

$$\|\tilde{c}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\Phi)}\right).$$

Again, by Lemma 1.2, $\tilde{c}$ is the unique solution of the ℓ_1 minimization problem (Sep_s), a contradiction. $\square$

These lemmata now immediately imply Theorem 1.1.

Proof of Theorem 1.1. Theorem 1.1 follows from Lemmata 1.2 and 1.3. $\square$

Interestingly, in the situation of Φ_1 and Φ_2 being two orthonormal bases the bound can be slightly strengthened. For the proof of this result, we refer the reader to [19].

Theorem 1.2 ([19])**.** *Let Φ_1 and Φ_2 be two orthonormal bases for a Hilbert space $\mathcal{H}$, and let $x \in \mathcal{H}$. If $x = [\Phi_1|\Phi_2]c$ and*

$$\|c\|_0 < \frac{\sqrt{2} - 0.5}{\mu([\Phi_1|\Phi_2])},$$

then the solution of the ℓ_1 minimization problem (Sep_s) stated in (1.3) coincides with the solution of the ℓ_0 minimization problem stated in (1.2).

This shows that in the special situation of two orthonormal bases, the bound is nearly a factor of 2 stronger than in the general situation of Theorem 1.1.

1.2.3   Clustered Sparsity as a Novel Viewpoint

In a concrete situation, we often have more information on the geometry of the to-be-separated components x_1^0 and x_2^0. This information is typically encoded in a particular clustering of the non-zero coefficients if a suitable basis or frame for the expansion of x_1^0 or x_2^0 is chosen. Think, for instance, of the tree clustering of wavelet coefficients of a point singularity. Thus, it seems conceivable that the morphological difference is encoded not only in the incoherence of the two chosen bases or frames adapted to x_1^0 and x_2^0, but in the interaction of the elements of those bases or frames associated with the clusters of significant coefficients. This should intuitively allow for weaker sufficient conditions for separation.

One possibility for a notion capturing this idea is the so-called *joint concentration* which was introduced in [14] with concepts going back to [16], and was in between again revived in [11]. To provide some intuition for this notion, let Λ_1 and Λ_2 be subsets of indexing sets of two Parseval frames. Then the joint concentration measures the maximal fraction of the total ℓ_1 norm which can be concentrated on the index set $\Lambda_1 \cup \Lambda_2$ of the combined dictionary.

Definition 1.2. *Let $\Phi_1 = (\varphi_{1i})_{i \in I}$ and $\Phi_2 = (\varphi_{2j})_{j \in J}$ be two Parseval frames for a Hilbert space $\mathcal{H}$. Further, let $\Lambda_1 \subseteq I$ and $\Lambda_2 \subseteq J$. Then the* joint concentration *$\kappa = \kappa(\Lambda_1, \Phi_1; \Lambda_2, \Phi_2)$ is defined by*

$$\kappa(\Lambda_1, \Phi_1; \Lambda_2, \Phi_2) = \sup_x \frac{\|1_{\Lambda_1} \Phi_1^T x\|_1 + \|1_{\Lambda_2} \Phi_2^T x\|_1}{\|\Phi_1^T x\|_1 + \|\Phi_2^T x\|_1}.$$

One might ask how the notion of joint concentration relates to the widely exploited, and for the previous result utilized mutual coherence. For this, we first briefly discuss some derivations of mutual coherence. A first variant better adapted to clustering of coefficients was the *Babel function* first introduced in [10] and later in [37] under the label *cumulative coherence function*, which, for a normalized frame $\Phi = (\varphi_i)_{i \in I}$ and some $m \in \{1, \dots, |I|\}$ is defined by

$$\mu_B(m, \Phi) = \max_{\Lambda \subset I, |\Lambda| = m} \max_{j \notin \Lambda} \sum_{i \in I} |\langle \varphi_i, \varphi_j \rangle|.$$

This notion was later refined in [3] by considering the so-called *structured p-Babel function*, defined for some family $\mathcal{S}$ of subsets of I and some $1 \le p < \infty$ by

$$\mu_{sB}(\mathcal{S}, \Phi) = \max_{\Lambda \in \mathcal{S}} \left(\max_{j \notin \Lambda} \sum_{i \in I} |\langle \varphi_i, \varphi_j \rangle|^p \right)^{1/p}.$$

Another variant, better adapted to data separation, is the *cluster coherence* introduced in [14], whose definition we now formally state. Notice that we do not assume that the vectors are normalized.

Definition 1.3. *Let $\Phi_1 = (\varphi_{1i})_{i \in I}$ and $\Phi_2 = (\varphi_{2j})_{j \in J}$ be two Parseval frames for a Hilbert space $\mathcal{H}$, let $\Lambda_1 \subseteq I$, and let $\Lambda_2 \subseteq J$. Then the* cluster coherence $\mu_c(\Lambda_1, \Phi_1; \Phi_2)$ *of Φ_1 and Φ_2 with respect to Λ_1 is defined by*

$$\mu_c(\Lambda_1, \Phi_1; \Phi_2) = \max_{j \in J} \sum_{i \in \Lambda_1} |\langle \varphi_{1i}, \varphi_{2j} \rangle|,$$

and the cluster coherence $\mu_c(\Phi_1; \Lambda_2, \Phi_2)$ *of Φ_1 and Φ_2 with respect to Λ_2 is defined by*

$$\mu_c(\Phi_1; \Lambda_2, \Phi_2) = \max_{i \in I} \sum_{j \in \Lambda_2} |\langle \varphi_{1i}, \varphi_{2j} \rangle|.$$

The relation between joint concentration and cluster coherence is made precise in the following result from [14].

Proposition 1.1 ([14]). *Let $\Phi_1 = (\varphi_{1i})_{i \in I}$ and $\Phi_2 = (\varphi_{2j})_{j \in J}$ be two Parseval frames for a Hilbert space $\mathcal{H}$, and let $\Lambda_1 \subseteq I$ and $\Lambda_2 \subseteq J$. Then*

$$\kappa(\Lambda_1, \Phi_1; \Lambda_2, \Phi_2) \leq \max\{\mu_c(\Lambda_1, \Phi_1; \Phi_2), \mu_c(\Phi_1; \Lambda_2, \Phi_2)\}.$$

Proof. Let $x \in \mathcal{H}$. We now choose coefficient sequences c_1 and c_2 such that

$$x = \Phi_1 c_1 = \Phi_2 c_2$$

and, for $i = 1, 2$,

$$\|c_i\|_1 \leq \|d_i\|_1 \quad \text{for all } d_i \text{ with } x = \Phi_i d_i. \tag{1.9}$$

This implies that

$$\|1_{\Lambda_1} \Phi_1^T x\|_1 + \|1_{\Lambda_2} \Phi_2^T x\|_1$$

$$= \|1_{\Lambda_1} \Phi_1^T \Phi_2 c_2\|_1 + \|1_{\Lambda_2} \Phi_2^T \Phi_1 c_1\|_1$$

$$\leq \sum_{i \in \Lambda_1} \left(\sum_{j \in J} |\langle \varphi_{1i}, \varphi_{2j} \rangle| |c_{2j}| \right) + \sum_{j \in \Lambda_2} \left(\sum_{i \in I} |\langle \varphi_{1i}, \varphi_{2j} \rangle| |c_{1i}| \right)$$

$$= \sum_{j \in J} \left(\sum_{i \in \Lambda_1} |\langle \varphi_{1i}, \varphi_{2j} \rangle| \right) |c_{2j}| + \sum_{i \in I} \left(\sum_{j \in \Lambda_2} |\langle \varphi_{1i}, \varphi_{2j} \rangle| \right) |c_{1i}|$$

$$\leq \mu_c(\Lambda_1, \Phi_1; \Phi_2) \|c_2\|_1 + \mu_c(\Phi_1; \Lambda_2, \Phi_2) \|c_1\|_1$$

$$\leq \max\{\mu_c(\Lambda_1, \Phi_1; \Phi_2), \mu_c(\Phi_1; \Lambda_2, \Phi_2)\}(\|c_1\|_1 + \|c_2\|_1).$$

Since Φ_1 and Φ_2 are Parseval frames, we have

$$x = \Phi_i(\Phi_i^T \Phi_i c_i) \quad \text{for } i = 1, 2.$$

Hence, by exploiting (1.9),

$$\|1_{\Lambda_1}\Phi_1^T x\|_1 + \|1_{\Lambda_2}\Phi_2^T x\|_1$$

$$\leq \max\{\mu_c(\Lambda_1, \Phi_1; \Phi_2), \mu_c(\Phi_1; \Lambda_2, \Phi_2)\}(\|\Phi_1^T\Phi_1 c_1\|_1 + \|\Phi_2^T\Phi_2 c_2\|_1)$$

$$= \max\{\mu_c(\Lambda_1, \Phi_1; \Phi_2), \mu_c(\Phi_1; \Lambda_2, \Phi_2)\}(\|\Phi_1^T x\|_1 + \|\Phi_2^T x\|_1).$$

$\square$

Before stating the data separation estimate which uses joint concentration, we need to discuss the conditions on sparsity of the components in the two Parseval frames. Since for real data 'true sparsity' is unrealistic, a weaker condition will be imposed. In the introduction [40], compressibility was already introduced. For the next result, yet a different notion invoking the clustering of the significant coefficients will be required. This notion, first utilized in [9], is defined for our data separation problem as follows.

Definition 1.4. *Let* $\Phi_1 = (\varphi_{1i})_{i\in I}$ *and* $\Phi_2 = (\varphi_{2j})_{j\in J}$ *be two Parseval frames for a Hilbert space* $\mathcal{H}$, *and let* $\Lambda_1 \subseteq I$ *and* $\Lambda_2 \subseteq J$. *Further, suppose that* $x \in \mathcal{H}$ *can be decomposed as* $x = x_1^0 + x_2^0$. *Then the components* x_1^0 *and* x_2^0 *are called* δ-*relatively sparse in* Φ_1 *and* Φ_2 *with respect to* Λ_1 *and* Λ_2, *if*

$$\|1_{\Lambda_1^c}\Phi_1^T x_1^0\|_1 + \|1_{\Lambda_2^c}\Phi_2^T x_2^0\|_1 \leq \delta.$$

We now have all ingredients to state the data separation result from [14], which – as compared to Theorem 1.1 – now invokes information about the clustering of coefficients.

Theorem 1.3 ([14]). *Let* $\Phi_1 = (\varphi_{1i})_{i\in I}$ *and* $\Phi_2 = (\varphi_{2j})_{j\in J}$ *be two Parseval frames for a Hilbert space* $\mathcal{H}$, *and suppose that* $x \in \mathcal{H}$ *can be decomposed as* $x = x_1^0 + x_2^0$. *Further, let* $\Lambda_1 \subseteq I$ *and* $\Lambda_2 \subseteq J$ *be chosen such that* x_1^0 *and* x_2^0 *are* δ-*relatively sparse in* Φ_1 *and* Φ_2 *with respect to* Λ_1 *and* Λ_2. *Then the solution* $(x_1^\star, x_2^\star)$ *of the* ℓ_1 *minimization problem* (Sep_a) *stated in* (1.6) *satisfies*

$$\|x_1^\star - x_1^0\|_2 + \|x_2^\star - x_2^0\|_2 \leq \frac{2\delta}{1 - 2\kappa}.$$

Proof. First, using the fact that Φ_1 and Φ_2 are Parseval frames,

$$\|x_1^\star - x_1^0\|_2 + \|x_2^\star - x_2^0\|_2 = \|\Phi_1^T(x_1^\star - x_1^0)\|_2 + \|\Phi_2^T(x_2^\star - x_2^0)\|_2$$
$$\leq \|\Phi_1^T(x_1^\star - x_1^0)\|_1 + \|\Phi_2^T(x_2^\star - x_2^0)\|_1.$$

The decomposition $x_1^0 + x_2^0 = x = x_1^\star + x_2^\star$ implies

$$x_2^\star - x_2^0 = -(x_1^\star - x_1^0),$$

which allows us to conclude that

$$\|x_1^\star - x_1^0\|_2 + \|x_2^\star - x_2^0\|_2 \leq \|\Phi_1^T(x_1^\star - x_1^0)\|_1 + \|\Phi_2^T(x_1^\star - x_1^0)\|_1. \qquad (1.10)$$

By the definition of κ,

$$\|\Phi_1^T(x_1^\star - x_1^0)\|_1 + \|\Phi_2^T(x_1^\star - x_1^0)\|_1$$
$$= (\|1_{\Lambda_1}\Phi_1^T(x_1^\star - x_1^0)\|_1 + \|1_{\Lambda_2}\Phi_2^T(x_1^\star - x_1^0)\|_1) + \|1_{\Lambda_1^c}\Phi_1^T(x_1^\star - x_1^0)\|_1$$
$$+ \|1_{\Lambda_2^c}\Phi_2^T(x_2^\star - x_2^0)\|_1$$
$$\leq \kappa \cdot \left(\|\Phi_1^T(x_1^\star - x_1^0)\|_1 + \|\Phi_2^T(x_1^\star - x_1^0)\|_1 \right) + \|1_{\Lambda_1^c}\Phi_1^T(x_1^\star - x_1^0)\|_1$$
$$+ \|1_{\Lambda_2^c}\Phi_2^T(x_2^\star - x_2^0)\|_1,$$

which yields

$$\|\Phi_1^T(x_1^\star - x_1^0)\|_1 + \|\Phi_2^T(x_1^\star - x_1^0)\|_1$$
$$\leq \frac{1}{1 - \kappa}(\|1_{\Lambda_1^c}\Phi_1^T(x_1^\star - x_1^0)\|_1 + \|1_{\Lambda_2^c}\Phi_2^T(x_2^\star - x_2^0)\|_1)$$
$$\leq \frac{1}{1 - \kappa}(\|1_{\Lambda_1^c}\Phi_1^T x_1^\star\|_1 + \|1_{\Lambda_1^c}\Phi_1^T x_1^0\|_1 + \|1_{\Lambda_2^c}\Phi_2^T x_2^\star\|_1 + \|1_{\Lambda_2^c}\Phi_2^T x_2^0\|_1).$$

Now using the relative sparsity of x_1^0 and x_2^0 in Φ_1 and Φ_2 with respect to Λ_1 and Λ_2, we obtain

$$\|\Phi_1^T(x_1^\star - x_1^0)\|_1 + \|\Phi_2^T(x_1^\star - x_1^0)\|_1 \leq \frac{1}{1 - \kappa} \left(\|1_{\Lambda_1^c}\Phi_1^T x_1^\star\|_1 + \|1_{\Lambda_2^c}\Phi_2^T x_2^\star\|_1 + \delta \right). \tag{1.11}$$

By the minimality of $x_1^\star$ and $x_2^\star$ as solutions of (Sep_a) implying that

$$\sum_{i=1}^{2} \left(\|1_{\Lambda_i^c}\Phi_i^T x_i^\star\|_1 + \|1_{\Lambda_i}\Phi_i^T x_i^\star\|_1 \right) = \|\Phi_1^T x_1^\star\|_1 + \|\Phi_2^T x_2^\star\|_1$$
$$\leq \|\Phi_1^T x_1^0\|_1 + \|\Phi_2^T x_2^0\|_1,$$

we have

$$\|1_{\Lambda_1^c}\Phi_1^T x_1^\star\|_1 + \|1_{\Lambda_2^c}\Phi_2^T x_2^\star\|_1$$
$$\leq \|\Phi_1^T x_1^0\|_1 + \|\Phi_2^T x_2^0\|_1 - \|1_{\Lambda_1}\Phi_1^T x_1^\star\|_1 - \|1_{\Lambda_2}\Phi_2^T x_2^\star\|_1$$
$$\leq \|\Phi_1^T x_1^0\|_1 + \|\Phi_2^T x_2^0\|_1 + \|1_{\Lambda_1}\Phi_1^T(x_1^\star - x_1^0)\|_1 - \|1_{\Lambda_1}\Phi_1^T x_1^0\|_1$$
$$+ \|1_{\Lambda_2}\Phi_2^T(x_2^\star - x_2^0)\|_1 - \|1_{\Lambda_2}\Phi_2^T x_2^0\|_1.$$

Again exploiting relative sparsity leads to

$$\|1_{\Lambda_1^c}\Phi_1^T x_1^\star\|_1 + \|1_{\Lambda_2^c}\Phi_2^T x_2^\star\|_1 \leq \|1_{\Lambda_1}\Phi_1^T(x_1^\star - x_1^0)\|_1 + \|1_{\Lambda_2}\Phi_2^T(x_2^\star - x_2^0)\|_1 + \delta. \tag{1.12}$$

Combining (1.11) and (1.12) and again using joint concentration,

$$\|\Phi_1^T(x_1^\star - x_1^0)\|_1 + \|\Phi_2^T(x_1^\star - x_1^0)\|_1$$
$$\leq \frac{1}{1 - \kappa} \left[\|1_{\Lambda_1}\Phi_1^T(x_1^\star - x_1^0)\|_1 + \|1_{\Lambda_2}\Phi_2^T(x_1^\star - x_1^0)\|_1 + 2\delta \right]$$
$$\leq \frac{1}{1 - \kappa} \left[\kappa \cdot \left(\|\Phi_1^T(x_1^\star - x_1^0)\|_1 + \|\Phi_2^T(x_1^\star - x_1^0)\|_1 \right) + 2\delta \right].$$

Thus, by (1.10), we finally obtain

$$\|x_1^\star - x_1^0\|_2 + \|x_2^\star - x_2^0\|_2 \leq \left(1 - \frac{\kappa}{1-\kappa}\right)^{-1} \cdot \frac{2\delta}{1-\kappa} = \frac{2\delta}{1-2\kappa}.$$

$\square$

Using Proposition 1.1, this result can also be stated in terms of cluster coherence, which on the one hand provides an easier accessible estimate and allows a better comparison with results using mutual coherence, but on the other hand poses a slightly weaker estimate.

Theorem 1.4 ([14]). *Let $\Phi_1 = (\varphi_{1i})_{i \in I}$ and $\Phi_2 = (\varphi_{2j})_{j \in J}$ be two Parseval frames for a Hilbert space $\mathcal{H}$, and suppose that $x \in \mathcal{H}$ can be decomposed as $x = x_1^0 + x_2^0$. Further, let $\Lambda_1 \subseteq I$ and $\Lambda_2 \subseteq J$ be chosen such that x_1^0 and x_2^0 are δ-relatively sparse in Φ_1 and Φ_2 with respect to Λ_1 and Λ_2. Then the solution $(x_1^\star, x_2^\star)$ of the minimization problem (Sep_a) stated in (1.6) satisfies*

$$\|x_1^\star - x_1^0\|_2 + \|x_2^\star - x_2^0\|_2 \leq \frac{2\delta}{1 - 2\mu_c},$$

with

$$\mu_c = \max\{\mu_c(\Lambda_1, \Phi_1; \Phi_2), \mu_c(\Phi_1; \Lambda_2, \Phi_2)\}.$$

To thoroughly understand this estimate, it is important to notice that both relative sparsity δ as well as cluster coherence μ_c depend heavily on the choice of the sets of significant coefficients Λ_1 and Λ_2. Choosing those sets too large allows for a very small δ, however μ_c might not be less than $\frac{1}{2}$ anymore, thereby making the estimate useless. Choosing those sets too small will force μ_c to become simultaneously small, in particular, smaller than $\frac{1}{2}$, with the downside that δ might be large.

It is also essential to realize that the sets Λ_1 and Λ_2 are a mere analysis tool; they do not appear in the minimization problem (Sep_a). This means that the algorithm does not care about this choice at all, however the estimate for accuracy of separation does.

Also note that this result can be easily generalized to general frames instead of Parseval frames, which then changes the separation estimate by invoking the lower frame bound. In addition, a version including noise was derived in [14].

1.2.4   Relation with Uncertainty Principles

Intriguingly, there exists a very close connection between uncertainty principles and data separation problems. Given a signal $x \in \mathcal{H}$ and two bases or frames Φ_1 and Φ_2, loosely speaking, an uncertainty principle states that x cannot be sparsely represented by Φ_1 and Φ_2 simultaneously; one of the expansions is

always not sparse unless $x = 0$. For the relation to the 'classical' uncertainty principle, we refer to Subsection 1.3.1.

The first result making this uncertainty viewpoint precise was proven in [19] with ideas already lurking in [16] and [11]. Again, it turns out that the mutual coherence is an appropriate measure for allowed sparsity, here serving as a lower bound for the simultaneously achievable sparsity of two expansions.

Theorem 1.5 ([19])**.** *Let Φ_1 and Φ_2 be two orthonormal bases for a Hilbert space $\mathcal{H}$, and let $x \in \mathcal{H}$, $x \neq 0$. Then*

$$\|\Phi_1^T x\|_0 + \|\Phi_2^T x\|_0 \geq \frac{2}{\mu([\Phi_1|\Phi_2])}.$$

Proof. First, let $\Phi_1 = (\varphi_{1i})_{i \in I}$ and $\Phi_2 = (\varphi_{2j})_{j \in J}$. Further, let $\Lambda_1 \subseteq I$ and $\Lambda_2 \subset J$ denote the support of $\Phi_1^T x$ and $\Phi_2^T x$, respectively. Since $x = \Phi_1 \Phi_1^T x$, for each $j \in J$,

$$|(\Phi_2^T x)_j| = \left| \sum_{i \in \Lambda_1} (\Phi_1^T x)_i \langle \varphi_{1i}, \varphi_{2j} \rangle \right|. \tag{1.13}$$

Since Φ_1 and Φ_2 are orthonormal bases, we have

$$\|x\|_2 = \|\Phi_1^T x\|_2 = \|\Phi_2^T x\|_2. \tag{1.14}$$

Using in addition the Cauchy-Schwarz inequality, we can continue (1.13) by

$$|(\Phi_2^T x)_j|^2 \leq \|\Phi_1^T x\|_2^2 \cdot \left| \sum_{i \in \Lambda_1} |\langle \varphi_{1i}, \varphi_{2j} \rangle|^2 \right| \leq \|x\|_2^2 \cdot |\Lambda_1| \cdot \mu([\Phi_1|\Phi_2])^2.$$

This implies

$$\|\Phi_2^T x\|_2 = \left(\sum_{j \in \Lambda_2} |(\Phi_2^T x)_j|^2 \right)^{1/2} \leq \|x\|_2 \cdot \sqrt{|\Lambda_1| \cdot |\Lambda_2|} \cdot \mu([\Phi_1|\Phi_2]).$$

Since $|\Lambda_i| = \|\Phi_i^T x\|_0$, $i = 1, 2$, and again using (1.14), we obtain

$$\sqrt{\|\Phi_1^T x\|_0 \cdot \|\Phi_2^T x\|_0} \geq \frac{1}{\mu([\Phi_1|\Phi_2])}.$$

Using the geometric-algebraic relationship,

$$\frac{1}{2}(\|\Phi_1^T x\|_0 + \|\Phi_2^T x\|_0) \geq \sqrt{\|\Phi_1^T x\|_0 \cdot \|\Phi_2^T x\|_0} \geq \frac{1}{\mu([\Phi_1|\Phi_2])},$$

which proves the claim. □

This result can be easily connected to the problem of simultaneously sparse expansions. The following version was first explicitly stated in [4].

Theorem 1.6 ([4]). *Let Φ_1 and Φ_2 be two orthonormal bases for a Hilbert space $\mathcal{H}$, and let $x \in \mathcal{H}$, $x \neq 0$. Then, for any two distinct coefficient sequences c_i satisfying $x = [\Phi_1|\Phi_2]c_i$, $i = 1, 2$, we have*

$$\|c_1\|_0 + \|c_2\|_0 \geq \frac{2}{\mu([\Phi_1|\Phi_2])},$$

Proof. First, set $d = c_1 - c_2$ and partition d into $[d_{\Phi_1}, d_{\Phi_2}]^T$ such that

$$0 = [\Phi_1|\Phi_2]d = \Phi_1 d_{\Phi_1} + \Phi_2 d_{\Phi_2}.$$

Since Φ_1 and Φ_2 are bases and $d \neq 0$, the vector y defined by

$$y = \Phi_1 d_{\Phi_1} = -\Phi_2 d_{\Phi_2}$$

is non-zero. Applying Theorem 1.5, we obtain

$$\|d\|_0 = \|d_{\Phi_1}\|_0 + \|d_{\Phi_2}\|_0 \geq \frac{2}{\mu([\Phi_1|\Phi_2])}.$$

Since $d = c_1 - c_2$, we have

$$\|c_1\|_0 + \|c_2\|_0 \geq \|d\|_0 \geq \frac{2}{\mu([\Phi_1|\Phi_2])}.$$

$\square$

We would also like to mention the very recent paper [39] by Tropp, in which he studies uncertainty principles for random sparse signals over an incoherent dictionary. He, in particular, shows that the coefficient sequence of each non-optimal expansion of a signal contains far more non-zero entries than the one of the sparsest expansion.

1.3   Signal Separation

In this section, we study the special situation of signal separation, where we refer to 1D signals as opposed to images, etc. For this, we start with the most prominent example of separating sinusoids from spikes, and then discuss further problem classes.

1.3.1   Separation of Sinusoids and Spikes

Sinusoidal and spike components are intuitively the morphologically most distinct features of a signal, since one is periodic and the other transient. Thus, it seems natural that the first results using sparsity and ℓ_1 minimization for data separation were proven for this situation. Certainly, real-world signals are never a pristine combination of sinusoids and spikes. However, thinking of audio data from a recording of musical instruments, these components are indeed an essential part of such signals.

The separation problem can be generally stated in the following way: Let the vector $x \in \mathbb{R}^n$ consist of n samples of a continuum domain signal at times $t \in \{0, \ldots, n-1\}$. We assume that x can be decomposed into

$$x = x_1^0 + x_2^0.$$

Here x_1^0 shall consist of n samples – at the same points in time as x – of a continuum domain signal of the form

$$\frac{1}{\sqrt{n}} \sum_{\omega=0}^{n-1} c_{1\omega}^0 e^{2\pi i \omega t/n}, \quad t \in \mathbb{R}.$$

Thus, by letting $\Phi_1 = (\varphi_{1\omega})_{0 \leq \omega \leq n-1}$ denote the Fourier basis, i.e.,

$$\varphi_{1\omega} = \left(\tfrac{1}{\sqrt{n}} e^{2\pi i \omega t/n} \right)_{0 \leq t \leq n-1},$$

the discrete signal x_1^0 can be written as

$$x_1^0 = \Phi_1 c_1^0 \quad \text{with } c_1^0 = (c_{1\omega}^0)_{0 \leq \omega \leq n-1}.$$

If x_1^0 is now the superposition of very few sinusoids, then the coefficient vector c_1^0 is sparse.

Further, consider a continuum domain signal which has a few spikes. Sampling this signal at n samples at times $t \in \{0, \ldots, n-1\}$ leads to a discrete signal $x_2^0 \in \mathbb{R}^n$ which has very few non-zero entries. In order to expand x_2^0 in terms of a suitable representation system, we let Φ_2 denote the Dirac basis, i.e., Φ_2 is simply the identity matrix, and write

$$x_2^0 = \Phi_2 c_2^0,$$

where c_2^0 is then a sparse coefficient vector.

The task now consists in extracting x_1^0 and x_2^0 from the known signal x, which is illustrated in Figure 1.1. It will be illuminating to detect the dependence on the number of sampling points of the bound for the sparsity of c_1^0 and c_2^0 which still allows for separation via ℓ_1 minimization.
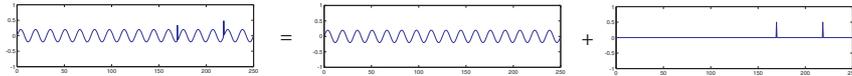
Figure 1.1 Separation of artificial audio data into sinusoids and spikes.

The intuition that – from a morphological standpoint – this situation is extreme, can be seen by computing the mutual coherence between the Fourier basis Φ_1 and the Dirac basis Φ_2. For this, we obtain

$$\mu([\Phi_1 | \Phi_2]) = \frac{1}{\sqrt{n}}, \tag{1.15}$$

and, in fact, $1/\sqrt{n}$ is the minimal possible value. This can be easily seen: If Φ_1 and Φ_2 are two general orthonormal bases of $\mathbb{R}^n$, then $\Phi_1^T \Phi_2$ is an orthonormal

matrix. Hence the sum of squares of its entries equals n, which implies that all entries can not be less than $1/\sqrt{n}$.

The following result from [19] makes this dependence precise. We wish to mention that the first answer to this question was derived in [11]. In this paper the slightly weaker bound of $(1 + \sqrt{n})/2$ for $\|c_1^0\|_0 + \|c_2^0\|_0$ was proven by using the general result in Theorem 1.1 instead of the more specialized Theorem 1.2 exploited to derive the result from [19] stated below.

Theorem 1.7 ([19]). *Let Φ_1 be the Fourier basis for $\mathbb{R}^n$ and let Φ_2 be the Dirac basis for $\mathbb{R}^n$. Further, let $x \in \mathbb{R}^n$ be the signal*

$$x = x_1^0 + x_2^0, \quad \text{where } x_1^0 = \Phi_1 c_1^0 \text{ and } x_2^0 = \Phi_2 c_2^0,$$

with coefficient vectors $c_i^0 \in \mathbb{R}^n$, $i = 1, 2$. If

$$\|c_1^0\|_0 + \|c_2^0\|_0 < (\sqrt{2} - 0.5)\sqrt{n},$$

then the ℓ_1 minimization problem (Sep_s) stated in (1.3) recovers c_1^0 and c_2^0 uniquely, and hence extracts x_1^0 and x_2^0 from x precisely.

Proof. Recall that we have (cf. (1.15))

$$\mu([\Phi_1|\Phi_2]) = \frac{1}{\sqrt{n}}.$$

Hence, by Theorem 1.2, the ℓ_1 minimization problem (Sep_s) recovers c_1^0 and c_2^0 uniquely, provided that

$$\|c_1^0\|_0 + \|c_2^0\|_0 < \frac{\sqrt{2} - 0.5}{\mu([\Phi_1|\Phi_2])} = (\sqrt{2} - 0.5)\sqrt{n}.$$

The theorem is proved.   □

The classical uncertainty principle states that, roughly speaking, a function cannot both be localized in time as well as in frequency domain. A discrete version of this fundamental principle was – besides the by now well-known continuum domain Donoho-Stark uncertainty principle – derived in [16]. It showed that a discrete signal and its Fourier transform cannot both be highly localized in the sense of having 'very few' non-zero entries. We will now show that this result – as it was done in [11] – can be interpreted as a corollary from data separation results.

Theorem 1.8 ([16]). *Let $x \in \mathbb{R}^n$, and denote its Fourier transform by $\hat{x}$. Then*

$$\|x\|_0 + \|\hat{x}\|_0 \geq 2\sqrt{n}.$$

Proof. For the proof, we intend to use Theorem 1.5. First, we note that by letting Φ_1 denote the Dirac basis, we trivially have

$$\|\Phi_1^T x\|_0 = \|x\|_0.$$

Secondly, letting Φ_2 denote the Fourier basis, we obtain

$$\hat{x} = \Phi_2^T x.$$

Now recalling that, by (1.15),

$$\mu([\Phi_1|\Phi_2]) = \frac{1}{\sqrt{n}},$$

we can conclude from Theorem 1.5 that

$$\|x\|_0 + \|\hat{x}\|_0 = \|\Phi_1^T x\|_0 + \|\Phi_2^T x\|_0 \geq \frac{2}{\mu([\Phi_1|\Phi_2])} = 2\sqrt{n}.$$

This finishes the proof. $\qquad\square$

As an excellent survey about sparsity of expansions of signals in the Fourier and Dirac basis, data separation, and related uncertainty principles as well as on very recent results using random signals, we refer to [38].

1.3.2 Further Variations

Let us briefly mention the variety of modifications of the previously discussed setting, most of them empirical analyses, which were developed during the last few years.

The most common variation of the sinusoid and spike setting is the consideration of a more general periodic component, which is then considered to be sparse in a Gabor system, superimposed by a second component, which is considered to be sparse in a system sensitive to spike-like structures similar to wavelets. This is, for instance, the situation considered in [22]. An example for a different setting is the substitution of a Gabor system by a Wilson basis, analyzed in [3]. In this paper, as already mentioned in Subsection 1.2.3, the clustering of coefficients already plays an essential role. It should also be mentioned that a specifically adapted norm, namely the mixed $\ell_{1,2}$ or $\ell_{2,1}$ norm, is used in [25] to take advantage of this clustering, and various numerical experiments show successful separation.

1.4 Image Separation

This section is devoted to discuss results on image separation exploiting Morphological Component Analysis, first focussing on empirical studies and secondly on theoretical results.

1.4.1 Empirical Results

In practice, the observed signal x is often contaminated by noise, i.e., $x = x_1^0 + x_2^0 + n$ containing the to-be-extracted components x_1^0 and x_2^0 and some noise

n. This requires an adaption of the ℓ_1 minimization problem. As proposed in numerous publications, one typically considers a modified optimization problem – so-called *Basis Pursuit Denoising* – which can be obtained by relaxing the constraint in order to deal with noisy observed signals. The ℓ_1 minimization problem (Sep_s) stated in (1.3), which places the ℓ_1 norm on the *synthesis* side then takes the form:

$$\min_{c_1,c_2} \|c_1\|_1 + \|c_2\|_1 + \lambda\|x - \Phi_1 c_1 - \Phi_2 c_2\|_2^2$$

with appropriately chosen regularization parameter $\lambda > 0$. Similarly, we can consider the relaxed form of the ℓ_1 minimization problem (Sep_a) stated in (1.6), which places the ℓ_1 norm on the *analysis* side:

$$\min_{x_1,x_2} \|\Phi_1^T x_1\|_1 + \|\Phi_2^T x_2\|_1 + \lambda\|x - x_1 - x_2\|_2^2.$$

In these new forms, the additional content in the image – the noise –, characterized by the property that it can not be represented sparsely by either one of the two systems Φ_1 and Φ_2, will be allocated to the residual $x - \Phi_1 c_1 - \Phi_2 c_2$ or $x - x_1 - x_2$ depending on which of the two minimization problems stated above is chosen. Hence, performing this minimization, we not only separate the data, but also succeed in removing an additive noise component as a by-product.

There exist by now a variety of algorithms which numerically solve such minimization problems. One large class are, for instance, iterative shrinkage algorithms; and we refer to the beautiful new book [18] by Elad for an overview. It should be mentioned that it is also possible to perform these separation procedures locally, thus enabling parallel processing, and again we refer to [18] for further details.

Let us now delve into more concrete situations. One prominent class of empirical studies concerns the separation of point- and curvelike structures. This type of problem arises, for instance, in astronomical imaging, where astronomers would like to separate stars (pointlike structures) from filaments (curvelike structures). Another area in which the separation of points from curves is essential is neurobiological imaging. In particular, for Alzheimer research, neurobiologists analyze images of neurons, which – considered in 2D – are a composition of the dendrites (curvelike structures) of the neuron and the attached spines (pointlike structures). For further analysis of the shape of these components, dendrites and spines need to be separated.

From a mathematical perspective, pointlike structures are generally speaking 0D structures whereas curvelike structures are 1D structures, which reveals their morphological difference. Thus it seems conceivable that separation using the idea of Morphological Component Analysis can be achieved, and the empirical results presented in the sequel as well as the theoretical results discussed in Subsection 1.4.2 give evidence to this claim.

To set up the minimization problem properly, the question arises which systems adapted to the point- and curvelike objects to use. For extracting pointlike

structures, wavelets seem to be optimal, since they provide optimally sparse approximations of smooth functions with finitely many point singularities. As a sparsifying system for curvelike structures, two different possibilities were explored so far. From a historical perspective, the first system to be utilized were *curvelets* [5], which provide optimally sparse approximations of smooth functions exhibiting curvilinear singularities. The composed dictionary of wavelets-curvelets is used in MCALab[1], and implementation details are provided in the by now considered fundamental paper [35]. A few years later *shearlets* were developed, see [24] or the survey paper [27], which deal with curvilinear singularities in a similarly favorable way as curvelets (cf. [28]), but have, for instance, the advantage of providing a unified treatment of the continuum and digital realm and being associated with a fast transform. Separation using the resulting dictionary of wavelets-shearlets is implemented and publicly available in ShearLab[2]. For a close comparison between both approaches we refer to [29] – in this paper the separation algorithm using wavelets and shearlets is also detailed –, where a numerical comparison shows that ShearLab provides a faster as well as more precise separation.

For illustrative purposes, Figure 1.2 shows the separation of an artificial image composed of points, lines, and a circle as well as added noise into the point-like structures (points) and the curvelike structures (lines and the circle), while removing the noise simultaneously. The only visible artifacts can be seen at the intersections of the curvelike structures, which is not surprising since it is even justifiable to label these intersections as 'points'. As an example using real data, we present in Figure 1.3 the separation of a neuron image into dendrites and spines again using ShearLab.

Another widely explored category of image separation is the separation of cartoons and texture. Here, the term cartoon typically refers to a piecewise smooth part in the image, and texture means a periodic structure. A mathematical model for a *cartoon* was first introduced in [8] as a C^2 function containing a C^2 discontinuity. In contrast to this, the term *texture* is a widely open expression, and people have debated for years over an appropriate model for the texture content of an image. A viewpoint from applied harmonic analysis characterizes texture as a structure which exhibits a sparse expansion in a Gabor system. As a side remark, the reader should be aware that periodizing a cartoon part of an image produces a texture component, thereby revealing the very fine line between cartoons and texture, illustrated in Figure 1.4.

As sparsifying systems, again curvelets or shearlets are suitable for the cartoon part, whereas discrete cosines or a Gabor system can be used for the texture part. MCALab uses for this separation task a dictionary composed of curvelets and discrete cosines, see [35]. For illustrative purposes, we display in Figure 1.5 the

---

[1] MCALab (Version 120) is available from `http://jstarck.free.fr/jstarck/Home.html`.
[2] ShearLab (Version 1.1) is available from `http://www.shearlab.org`.

(a) Original image

(b) Noisy image

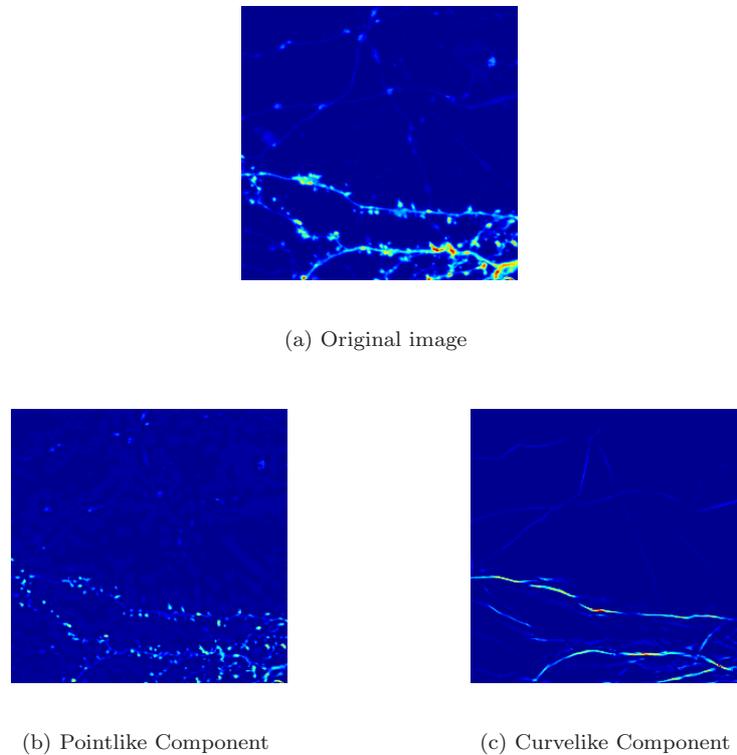(c) Pointlike Component

(d) Curvelike Component
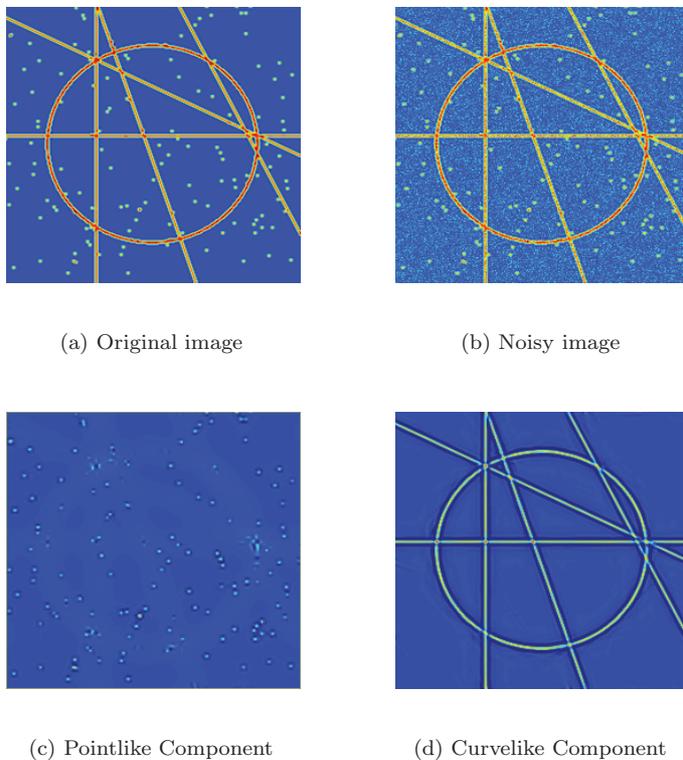
Figure 1.2 Separation of an artificial image composed of points, lines, and a circle into point- and curvelike components using ShearLab.

separation of the Barbara image into cartoon and texture component performed by MCALab. As can be seen, all periodic structure is captured in the texture part, leaving the remainder to the cartoon component.

1.4.2   Theoretical Results

The first theoretical result explaining the successful empirical performance of Morphological Component Analysis was derived in [14] by considering the separation of point- and curvelike features in images coined the *Geometric Separation Problem*. The analysis in this paper has three interesting features. Firstly, it introduces the notion of cluster coherence (cf. Definition 1.3) as a measure for the geometric arrangements of the significant coefficients and hence the encoding of the morphological difference of the components. It also initiates the study of ℓ_1 minimization in frame settings, in particular those where singleton coherence within one frame may be high. Secondly, it provides the first analysis of a continuum model in contrast to the previously studied discrete models which obscure

(a) Original image

(b) Pointlike Component

(c) Curvelike Component

Figure 1.3 Separation of a neuron image into point- and curvelike components using ShearLab.
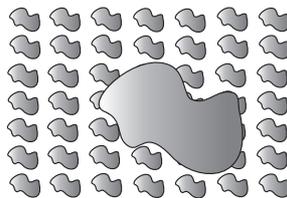
Figure 1.4 Periodic small cartoons versus one large cartoon.

continuum elements of geometry. And thirdly, it explores microlocal analysis to understand heuristically why separation might be possible and to organize a rigorous analysis. This general approach applies in particular to two variants of geometric separation algorithms. One is based on tight frames of radial wavelets and curvelets and the other uses orthonormal wavelets and shearlets.

These results are today the only results providing a theoretical foundation to image separation using ideas from sparsity methodologies. The same situation – separating point- and curvelike objects – is also considered in [13] however using

(a) Barbara image
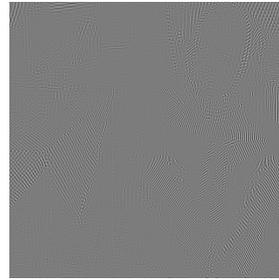
(b) Cartoon Component   (c) Texture Component

Figure 1.5 Separation of the Barbara image into cartoon and texture using MCALab.

thresholding as a separation technique. Finally, we wish to mention that some initial theoretical results on the separation of cartoon and texture in images are contained in [15].

Let us now dive into the analysis of [14]. As a mathematical model for a composition of point- and curvelike structures, the following two components are considered: The function $\mathcal{P}$ on $\mathbb{R}^2$, which is smooth except for point singularities and defined by

$$\mathcal{P} = \sum_{i=1}^{P} |x - x_i|^{-3/2},$$

serves as a model for the pointlike objects, and the distribution $\mathcal{C}$ with singularity along a closed curve $\tau : [0,1] \to \mathbb{R}^2$ defined by

$$\mathcal{C} = \int \delta_{\tau(t)} dt,$$

models the curvelike objects. The general model for the considered situation is then the sum of both, i.e.,

$$f = \mathcal{P} + \mathcal{C}, \tag{1.16}$$

and the *Geometric Separation Problem* consists of recovering $\mathcal{P}$ and $\mathcal{C}$ from the observed signal f.

As discussed before, one possibility is to set up the minimization problem using an overcomplete system composed of wavelets and curvelets. For the analysis, radial wavelets are used due to the fact that they provide the same subbands as curvelets. To be more precise, let W be an appropriate window function. Then *radial wavelets* at scale j and spatial position $k = (k_1, k_2)$ are defined by the Fourier transforms

$$\hat{\psi}_\lambda(\xi) = 2^{-j} \cdot W(|\xi|/2^j) \cdot e^{i\langle k, \xi/2^j \rangle},$$

where $\lambda = (j, k)$ indexes scale and position. For the same window function W and a 'bump function' V, *curvelets* at scale j, orientation ℓ, and spatial position $k = (k_1, k_2)$ are defined by the Fourier transforms

$$\hat{\gamma}_\eta(\xi) = 2^{-j\frac{3}{4}} \cdot W(|\xi|/2^j) V((\omega - \theta_{j,\ell}) 2^{j/2}) \cdot e^{i(R_{\theta_{j,\ell}} A_{2^{-j}} k)' \xi},$$

where $\theta_{j,\ell} = 2\pi\ell/2^{j/2}$, R_θ is planar rotation by $-\theta$ radians, A_a is anisotropic scaling with diagonal $(a, \sqrt{a})$, and we let $\eta = (j, \ell, k)$ index scale, orientation, and scale; see [5] for more details. The tiling of the frequency domain generated by these two systems is illustrated in Figure 1.6.

(a) Radial wavelets   (b) Curvelets
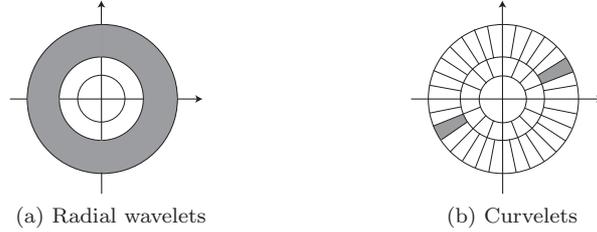
Figure 1.6 Tiling of the frequency domain by radial wavelets and curvelets.

By using again the window W, we define the family of filters F_j by their transfer functions

$$\hat{F}_j(\xi) = W(|\xi|/2^j), \qquad \xi \in \mathbb{R}^2.$$

These filters provide a decomposition of any distribution g into pieces g_j with different scales, the piece g_j at subband j generated by filtering g using F_j:

$$g_j = F_j \star g.$$

A proper choice of W then enables reconstruction of g from these pieces using the formula

$$g = \sum_j F_j \star g_j.$$

Application of this filtering procedure to the model image f from (1.16) yields the decompositions

$$f_j = F_j \star f = F_j \star (\mathcal{P} + \mathcal{C}) = \mathcal{P}_j + \mathcal{C}_j,$$

where $(f_j)_j$ is known, and we aim to extract $(\mathcal{P}_j)_j$ and $(\mathcal{C}_j)_j$. We should mention at this point that, in fact, the pair $(\mathcal{P}, \mathcal{C})$ was chosen in such a way that $\mathcal{P}_j$ and $\mathcal{C}_j$ have the same energy for each j, thereby making the components comparable as we go to finer scales and the separation challenging at *each* scale.

Let now Φ_1 and Φ_2 be the tight frame of radial wavelets and curvelets, respectively. Then, for each scale j, we consider the ℓ_1 minimization problem (Sep_a) stated in (1.6), which now reads:

$$\min_{P_j, C_j} \|\Phi_1^T P_j\|_1 + \|\Phi_2^T C_j\|_1 \quad \text{s.t.} \quad f_j = P_j + C_j. \qquad (1.17)$$

Notice that we use the 'analysis version' of the minimization problem, since both radial wavelets as well as curvelets are overcomplete systems.

The theoretical result of the precision of separation of f_j via (1.17) proved in [14] can now be stated in the following way:

Theorem 1.9 ([14])**.** *Let $\hat{P}_j$ and $\hat{C}_j$ be the solutions to the optimization problem (1.17) for each scale j. Then we have*

$$\frac{\|\mathcal{P}_j - \hat{P}_j\|_2 + \|\mathcal{C}_j - \hat{C}_j\|_2}{\|\mathcal{P}_j\|_2 + \|\mathcal{C}_j\|_2} \to 0, \quad j \to \infty.$$

This result shows that the components $\mathcal{P}_j$ and $\mathcal{C}_j$ are recovered with asymptotically arbitrarily high precision at very fine scales. The energy in the pointlike component is completely captured by the wavelet coefficients, and the curvelike component is completely contained in the curvelet coefficients. Thus, the theory evidences that the Geometric Separation Problem can be satisfactorily solved by using a combined dictionary of wavelets and curvelets and an appropriate ℓ_1 minimization problem, as already the empirical results indicate.

We next provide a sketch of proof and refer to [14] for the complete proof.

Sketch of proof of Theorem 1.9. The main goal will be to apply Theorem 1.4 to each scale and monitor the sequence of bounds $\frac{2\delta}{1 - 2\mu_c}$ now dependent on scale. For this, let j be arbitrarily fixed, and apply Theorem 1.4 in the following way:

- x: Filtered signal f_j $(= \mathcal{P}_j + \mathcal{C}_j)$.
- Φ_1: Wavelets filtered with F_j.
- Φ_2: Curvelets filtered with F_j.
- Λ_1: Significant wavelet coefficients of $\mathcal{P}_j$.
- Λ_2: Significant curvelet coefficients of $\mathcal{C}_j$.
- δ_j: Degree of approximation by significant coefficients.
- $(\mu_c)_j$: Cluster coherence of wavelets-curvelets.

If

$$\frac{2\delta_j}{1 - 2(\mu_c)_j} = o(\|\mathcal{P}_j\|_2 + \|\mathcal{C}_j\|_2) \quad \text{as } j \to \infty \tag{1.18}$$

can be then shown, the theorem is proved.

One main problem to overcome is the highly delicate choice of Λ_1 and Λ_2. It would be ideal to define those sets in such a way that

$$\delta_j = o(\|\mathcal{P}_j\|_2 + \|\mathcal{C}_j\|_2) \quad \text{as } j \to \infty \tag{1.19}$$

and

$$(\mu_c)_j \to 0 \quad \text{as } j \to \infty \tag{1.20}$$

are true. This would then imply (1.18), hence finish the proof.

A microlocal analysis viewpoint now provides insight into how to suitably choose Λ_1 and Λ_2 by considering the wavefront sets of $\mathcal{P}$ and $\mathcal{C}$ in phase space $\mathbb{R}^2 \times [0, 2\pi)$, i.e.,

$$WF(\mathcal{P}) = \{x_i\}_{i=1}^P \times [0, 2\pi)$$

and

$$WF(\mathcal{C}) = \{(\tau(t), \theta(t)) : t \in [0, L(\tau)]\},$$

where $\tau(t)$ is a unit-speed parametrization of τ and $\theta(t)$ is the normal direction to τ at $\tau(t)$. Heuristically, the significant wavelet coefficients should be associated with wavelets whose index set is 'close' to $WF(\mathcal{P})$ in phase space and, similarly, the significant curvelet coefficients should be associated with curvelets whose index set is 'close' to $WF(\mathcal{C})$. Thus, using Hart Smith's phase space metric,

$$d_{HS}((b, \theta); (b', \theta')) = |\langle e_\theta, b - b' \rangle| + |\langle e_{\theta'}, b - b' \rangle| + |b - b'|^2 + |\theta - \theta'|^2,$$

where $e_\theta = (\cos(\theta), \sin(\theta))$, an 'approximate' form of sets of significant wavelet coefficients is

$$\Lambda_{1,j} = \{\text{wavelet lattice}\} \cap \{(b, \theta) : d_{HS}((b, \theta); WF(\mathcal{P})) \leq \eta_j 2^{-j}\},$$

and an 'approximate' form of sets of significant curvelet coefficients is

$$\Lambda_{2,j} = \{\text{curvelet lattice}\} \cap \{(b, \theta) : d_{HS}((b, \theta); WF(\mathcal{C})) \leq \eta_j 2^{-j}\}$$

with a suitable choice of the distance parameters $(\eta_j)_j$. In the proof of Theorem 1.9, the definition of $(\Lambda_{1,j})_j$ and $(\Lambda_{2,j})_j$ is much more delicate, but follows this intuition. Lengthy and technical estimates then lead to (1.19) and (1.20), which – as mentioned before – completes the proof. $\qquad\square$

Since it was already mentioned in Subsection 1.4.1 that a combined dictionary of wavelets and shearlets might be preferable, the reader will wonder whether the just discussed theoretical results can be transferred to this setting. In fact, this is proven in [26], see also [12]. It should be mentioned that one further advantage

of this setting is the fact that now a basis of wavelets can be utilized in contrast
to the tight frame of radial wavelets explored before.

As a wavelet basis, we now choose *orthonormal Meyer wavelets*, and refer to
[30] for the definition. For the definition of shearlets, for $j \geq 0$ and $k \in \mathbb{Z}$, let
– the notion A_{2^j} was already introduced in the definition of curvelets – $\tilde{A}_{2^j}$ and
S_k be defined by

$$\tilde{A}_{2^j} = \begin{pmatrix} 2^{j/2} & 0 \\ 0 & 2^j \end{pmatrix} \quad \text{and} \quad S_k = \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix}.$$

For $\phi, \psi, \tilde{\psi} \in L^2(\mathbb{R}^2)$, the *cone-adapted discrete shearlet system* is then the union
of

$$\{\phi(\cdot - m) : m \in \mathbb{Z}^2\},$$

$$\{2^{\frac{3}{4}j}\psi(S_k A_{2^j} \cdot - m) : j \geq 0, -\lceil 2^{j/2} \rceil \leq k \leq \lceil 2^{j/2} \rceil, m \in \mathbb{Z}^2\},$$

and

$$\{2^{\frac{3}{4}j}\tilde{\psi}(S_k^T \tilde{A}_{2^j} \cdot - m) : j \geq 0, -\lceil 2^{j/2} \rceil \leq k \leq \lceil 2^{j/2} \rceil, m \in \mathbb{Z}^2\}.$$

The term 'cone-adapted' originates from the fact that these systems tile the
frequency domain in a cone-like fashion; see Figure 1.7b.

(a) Wavelets   (b) Shearlets
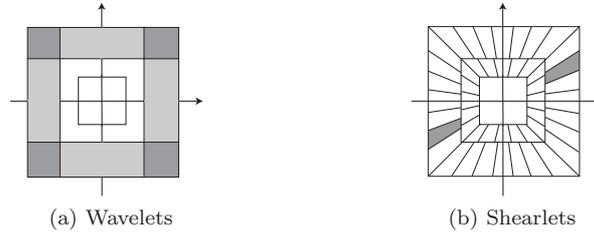
Figure 1.7 Tiling of the frequency domain by orthonormal Meyer wavelets and
shearlets.

As can be seen from Figure 1.7, the subbands associated with orthonormal
Meyer wavelets and shearlets are the same. Hence a similar filtering into scaling
subbands can be performed as for radial wavelets and curvelets.

Adapting the optimization problem (1.17) by using wavelets and shearlets
instead of radial wavelets and curvelets generates purported point- and curvelike
objects $\hat{W}_j$ and $\hat{S}_j$, say, for each scale j. Then the following result, which shows
similarly successful separation as Theorem 1.9, was derived in [26] with the new
concept of sparsity equivalence, here between shearlets and curvelets, introduced
in the same paper as main ingredient.

Theorem 1.10 ([26]). *We have*

$$\frac{\|\mathcal{P}_j - \hat{W}_j\|_2 + \|\mathcal{C}_j - \hat{S}_j\|_2}{\|\mathcal{P}_j\|_2 + \|\mathcal{C}_j\|_2} \to 0, \quad j \to \infty.$$

Acknowledgements

References

[1] Aharon, M., Elad, M., and Bruckstein, A.M. (2006). The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation, *IEEE Trans. Signal Proc.*, **54(11)**, 4311–4322.

[2] Bobin, J., Starck, J.-L., Fadili, M.J., Moudden, Y., and Donoho, D.L. (2007). Morphological component analysis: An adaptive thresholding strategy, *IEEE Trans. Image Proc.*, **16(11)**, 2675–2681.

[3] Borup, L., Gribonval, R., and Nielsen, M. (2008). Beyond coherence: Recovering structured time-frequency representations, *Appl. Comput. Harmon. Anal.*, **24(1)**, 120–128.

[4] Bruckstein, A.M., Donoho, D.L., and Elad, M. (2009). From sparse solutions of systems of equations to sparse modeling of signals and images, SIAM Review, **51(1)**, 34–81.

[5] Candès, E.J. and Donoho, D.L. (2005). Continuous curvelet transform: II. Discretization of frames, *Appl. Comput. Harmon. Anal.*, **19(2)**, 198–222.

[6] Chen, S. S., Donoho, D. L., and Saunders, M. A. (1998). Atomic decomposition by basis pursuit, *SIAM J. Sci. Comput.*, **20(1)**, 33–61.

[7] Coifman, R.R. and Wickerhauser, M.V. (1993). Wavelets and adapted waveform analysis. A toolkit for signal processing and numerical analysis, *Different perspectives on wavelets* (San Antonio, TX, 1993), 119–153, Proc. Sympos. Appl. Math., **47**, Amer. Math. Soc., Providence, RI.

[8] Donoho, D.L. (2001). Sparse components of images and optimal atomic decomposition, *Constr. Approx.*, **17(3)**, 353–382.

[9] Donoho, D.L. (2006). Compressed sensing, *IEEE Trans. Inform. Theory*, **52(4)**, 1289–1306.

[10] Donoho, D.L. and Elad, M. (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization, *Proc. Natl. Acad. Sci. USA*, **100(5)**, 2197–2202.

[11] Donoho, D.L. and Huo, X. (2001). Uncertainty principles and ideal atomic decomposition, *IEEE Trans. Inform. Theory*, **47(7)**, 2845–2862.

[12] Donoho, D.L. and Kutyniok, G. (2009). Geometric separation using a wavelet-shearlet dictionary, *SampTA'09* (Marseille, France, 2009), Proc., 2009.

[13] Donoho, D.L. and Kutyniok, G. (2010). Geometric separation by single-pass alternating thresholding, preprint.

[14] Donoho, D.L. and Kutyniok, G. (2010). Microlocal analysis of the geometric separation problem, preprint.

[15] Donoho, D.L. and Kutyniok, G. (2011). Geometric separation of cartoons and texture via ℓ_1 minimization, preprint.

[16] Donoho, D.L. and Stark, P.B. (1989). Uncertainty principles and signal recovery, *SIAM J. Appl. Math.*, **49(3)**, 906–931.

[17] Duarte-Carvajalino, J.M. and Sapiro, G. (2009). Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization, *IEEE Trans. Image Proc.*, **18(7)**, 1395–1408.

[18] Elad, M. (2010). *Sparse and redundant representations*, Springer, New York.

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

[20] Elad, M., Starck, J.-L., Querre, P., and Donoho, D.L. (2005). Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA), *Appl. Comput. Harmon. Anal.*, **19(3)**, 340–358.

[21] Engan, K., Aase, S.O., and Hakon-Husoy, J.H. (1999). Method of optimal directions for frame design, *IEEE Int. Conf. Acoust., Speech, Signal Process.*, **5**, 2443–2446.

[22] Gribonval, R. and Bacry, E. (2003). Harmonic decomposition of audio signals with matching pursuit, *IEEE Trans. Signal Proc.*, **51(1)**, 101–111.

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

[24] Guo, K., Kutyniok, G., and Labate, D. (2006). Sparse multidimensional representations using anisotropic dilation and shear operators, *Wavelets and Splines* (Athens, GA, 2005), Nashboro Press, Nashville, TN, 2006, 189–201.

[25] Kowalski, M. and Torrésani, B. (2010). Sparsity and persistence: Mixed norms provide simple signal models with dependent coefficients, *Signal, Image and Video Proc.*, to appear.

[26] Kutyniok, G. (2010). Sparsity equivalence of anisotropic decompositions, preprint.

[27] Kutyniok, G., Lemvig, J., and Lim, W.-Q (2010). Compactly supported shearlets, *Approximation Theory XIII* (San Antonio, TX, 2010), Springer, to appear.

[28] Kutyniok, G. and Lim, W.-Q (2010). Compactly supported shearlets are optimally sparse, preprint.

[29] Kutyniok, G. and Lim, W.-Q (2010). Image separation using shearlets, preprint.

[30] Mallat, S.G. (1998). *A wavelet tour of signal processing*, Academic Press, Inc., San Diego, CA.

[31] Mallat, S.G. and Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries, *IEEE Trans. Signal Proc.*, **41(12)**, 3397–3415.

[32] Meyer, F.G., Averbuch, A., and Coifman, R.R. (2002). Multi-layered image representation: Application to image compression, *IEEE Trans. Image Proc.*, **11(9)**, 1072–1080.

[33] Starck, J.-L., Murtagh, F., and Fadili, J.M. (2010). *Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity*, Cambridge University Press, New York, NY.

[34] Starck, J.-L., Elad, M., and Donoho, D.L. (2005). Redundant multiscale transforms and their application for morphological component analysis, *Adv. Imag. Electr. Phys.*, **132**, 287–348.

[35] Starck, J.-L., Elad, M., and Donoho, D.L. (2005). Image decomposition via the combination of sparse representations and a variational approach, *IEEE Trans. Image Proc.*, **14(10)**, 1570–1582.

[36] Starck, J.-L., Moudden, Y., Bobin, J., Elad, M., and Donoho, D.L. (2005). Morphological component analysis, *Wavelets XI* (San Diego, CA, 2005), SPIE Proc. 5914, SPIE, Bellingham, WA.

[37] Tropp, J.A. (2004). Greed is good: Algorithmic results for sparse approximation, *IEEE Trans. Inform. Theory*, **50(10)**, 2231–2242.

[38] Tropp, J.A. (2008). On the linear independence of spikes and sines, *J. Fourier Anal. Appl.*, **14(5-6)**, 838–858.

[39] Tropp, J.A. (2010). The sparsity gap: Uncertainty principles proportional to dimension, *Proc. 44th IEEE Conf. Information Sciences and Systems* (CISS), 1–6, Princeton, NJ.

[40] Introduction Chapter.

[41] Structured Sparsity Chapter.

[42] Greedy Chapter.

Index