On the Global-Local Dichotomy in Sparsity Modeling

Chapter · February 2017

DOI: 10.1007/978-3-319-69802-1_1

On the global-local dichotomy in sparsity modeling

Dmitry Batenkov, Yaniv Romano and Michael Elad

Abstract The traditional sparse modeling approach, when applied to inverse problems with large data such as images, essentially assumes a sparse model for small overlapping data patches. While producing state-of-the-art results, this methodology is suboptimal, as it does not attempt to model the entire global signal in any meaningful way – a nontrivial task by itself. In this paper we propose a way to bridge this theoretical gap by constructing a global model from the bottom up. Given local sparsity assumptions in a dictionary, we show that the global signal representation must satisfy a constrained underdetermined system of linear equations, which can be solved efficiently by modern optimization methods such as Alternating Direction Method of Multipliers (ADMM). We investigate conditions for unique and stable recovery, and provide numerical evidence corroborating the theory.

Key words: sparse representations, inverse problems, convolutional sparse coding

Dmitry Batenkov
Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. e-mail: batenkov@mit.edu

Yaniv Romano
Department of Electrical Engineering, Technion - Israel Institute of Technology, 32000 Haifa, Israel. e-mail: yromano@tx.technion.ac.il

Michael Elad
Department of Computer Science, Technion - Israel Institute of Technology, 32000 Haifa, Israel. e-mail: elad@cs.technion.ac.il

1 Introduction

1.1 The need for a new local-global sparsity theory

The sparse representation model [19] provides a powerful approach to various inverse problems in image and signal processing such as denoising [20, 34], deblurring [52, 16] and super-resolution [51, 44], to name a few [33]. This model assumes that a signal can be represented as a sparse linear combination of a few columns (called atoms) taken from a matrix termed dictionary. Consecutively, given a signal, the sparse recovery of its representation over a dictionary is called sparse-coding or pursuit. Due to computational and theoretical aspects, when treating high dimensional data most of the existing sparsity-inspired methods utilize local-patched-based representations rather than the global ones, i.e. they divide a signal into small overlapping blocks (patches), reconstruct these patches using standard sparse recovery techniques, and subsequently average the overlapping regions [13, 19]. While this approach leads to highly efficient algorithms producing state-of-the-art results, it is fundamentally limited because the basic sparse model applies to patches only, and does not take into account the dependencies between them.

As an attempt to tackle this flaw, methods based on the notion of *structured sparsity* [21, 29, 28, 31, 50] started to appear; for example, in [34, 16, 44] the observation that a patch may have similar neighbors in its surroundings (often termed the self-similarity property) is injected to the pursuit, leading to improved local estimations. Another possibility to consider the dependencies between patches is to exploit the multi-scale nature of the signals [35, 48, 36]. A different direction is suggested by the EPLL [55, 47, 36], which encourages the patches of the final estimate (i.e., after the application of the averaging step) to comply with the local prior. Also, a related work [43, 42] suggests promoting the local estimations to agree on their shared content (the overlap) as a way to achieve a coherent reconstruction of the signal.

Recently, an alternative to the traditional patch-based prior was suggested in the form of the convolutional, or shift-invariant, sparse coding (CSC) model [24, 12, 27, 26, 49, 45]. Rather than dividing the image into local patches and process each of these independently, this approach imposes a specific structure on the global dictionary – a concatenation of banded circulant matrices – and applies a global pursuit. A thorough theoretical analysis of this model was proposed very recently in [38, 39, 37], providing a clear understanding of its success.

The empirical success of the above algorithms indicates the great potential of reducing the inherent gap that exists between the independent local processing of patches and the global nature of the signal at hand. However, a key and highly desirable part is still missing – a theory which would suggest how to modify the basic sparse model to take into account mutual dependence between the patches, what approximation methods to use, and how to efficiently design and learn the corresponding structured dictionary.

1.2 Content and organization of the paper

In this paper we propose a systematic investigation of the signals which are implicitly defined by local sparsity assumptions. A major theme in what follows is that the presence of patch overlaps reduces the number of degrees of freedom, which, in turn, has theoretical and practical implications. In particular, this allows more accurate estimates for uniqueness and stability of local sparse representations, as well as better bounds on performance of existing sparse approximation algorithms. Moreover, the global point of view allows for development of new pursuit algorithms, which consist of local operation on one hand, while also taking into account the patch overlaps on the other hand. Some aspects of the offered theory are still incomplete, and several exciting research direction emerge as well.

The paper is organized as follows. In Section 2 we develop the basic framework for signals which are patch-sparse, building the global model from the "bottom up", and discuss some theoretical properties of the resulting model. In Section 3 we consider the questions of reconstructing the representation vector, and of denoising a signal in this new framework. We describe "globalized" greedy pursuit algorithms [40] for these tasks, where the patch disagreements play a major role. We show that the frequently used Local Patch Averaging (LPA) approach is in fact suboptimal in this case. In Section 4 we describe several instances/classes of the local-global model in some detail, exemplifying the preceding definitions and results. The examples include piecewise-constant signals, signature-type (periodic) signals, and more general bottom-up models. In Section 5 we present results of extensive numerical experiments, where in particular we show that one of the new globalized pursuits, based on the ADMM algorithm, turns out to have superior performance in all the cases considered. We conclude the paper in Section 6 by discussing possible research directions.

2 Local-global sparsity

We start with the local sparsity assumptions for every patch, and subsequently provide two complimentary characterizations of the resulting global signal space. On one hand, we show that the signals of interest admit a global "sparse-like" representation with a dictionary of convolutional type, and with additional linear constraints on the representation vector. On the other hand, the signal space is in fact a union of linear subspaces, where each subspace is a kernel of a certain linear map. Finally we connect the two points of view by showing that the original local dictionary must carry a combinatorial structure. Concluding this section, we provide some theoretical analysis of the properties of the resulting model, in particular uniqueness and stability of representation. For this task, we define certain measures of the dictionary, similar to the classical spark, coherence function, and the Restricted Isometry Property, which take the additional dictionary structure into account.

2.1 Preliminaries

Definition 1 (Spark of a matrix). Given a dictionary $D \in \mathbb{R}^{n \times m}$, the *spark* of D is defined as the minimal number of columns which are linearly dependent:

$$\sigma(D) := \min\{j : \exists s \subset [1, \ldots, m], |s| = j, \operatorname{rank} D_s < j\}. \tag{1}$$

Clearly $\sigma(D) \leqslant n+1$.

Definition 2. Given a vector $\alpha \in \mathbb{R}^m$, the ℓ_0 pseudo-norm is the number of nonzero elements in α:

$$\|\alpha\|_0 := \#\{j : \alpha_j \neq 0\}.$$

Definition 3. Let $D \in \mathbb{R}^{n \times m}$ be a dictionary with normalized atoms. The μ_1 coherence function (Tropp's Babel's function) is defined as

$$\mu_1(s) := \max_{i \in [1,\ldots,m]} \max_{S \subset [1,\ldots m] \setminus \{i\},\, |S| = s} \sum_{j \in S} \left|\langle d_i, d_j \rangle\right|.$$

Definition 4. Given a dictionary D as above, the Restricted Isometry constant of order k is the smallest number δ_k such that

$$(1-\delta_k)\|\alpha\|_2^2 \leqslant \|D\alpha\|_2^2 \leqslant (1+\delta_k)\|\alpha\|_2^2$$

for every $\alpha \in \mathbb{R}^m$ with $\|\alpha\|_0 \leqslant k$.

For any matrix M, we denote by $\mathscr{R}(M)$ the column space (range) of M.

2.2 Globalized local model

In what follows we treat one-dimensional signals $x \in \mathbb{R}^N$ of length N, divided into $P = N$ overlapping patches of equal size n (so that the original signal is thought to be periodically extended). The other natural choice is $P = N - n + 1$, but for simplicity of derivations we consider only the periodic case.

So we define for each $i = 1, \ldots P$

$$R_i := \left[\mathbf{0} \ldots \mathbf{0}\; Id_{n \times n}\; \mathbf{0} \ldots \mathbf{0}\right] \in \mathbb{R}^{n \times N}, \tag{2}$$

the operator extracting i-th patch from the signal.

Definition 5. Given local dictionary $D \in \mathbb{R}^{n\times m}$, sparsity level $s < n$, signal length N, and the number of overlapping patches P, the *globalized local-sparse* model is the set

$$\mathscr{M} = \mathscr{M}(D,s,P,N) := \left\{ x \in \mathbb{R}^N,\ R_i x = D\alpha_i,\ \|\alpha_i\|_0 \leqslant s\ \forall i = 1,\dots,P \right\}. \quad (3)$$

This model suggests that each patch, $R_i x$ is assumed to have an s-sparse representation α_i, and this way we have charactatized the global x by describing the local nature of its patches.

Next we derive a "global" characterization of $\mathscr{M}$. Starting with the equations

$$R_i x = D\alpha_i, \qquad i = 1,\dots,P,$$

and using the equality $Id = \frac{1}{n}\sum_{i=1}^{P} R_i^T R_i$, we have a representation

$$x = \frac{1}{n}\sum_{i=1}^{P} R_i^T R_i x = \sum_{i=1}^{P}\left(\frac{1}{n} R_i^T D\right)\alpha_i.$$

Let the global "convolutional" dictionary D_G be defined as the horizontal concatenation of the (vertically) shifted versions of $\frac{1}{n}D$, i.e.

$$D_G := \left[\left(\frac{1}{n} R_i^T D \right) \right]_{i=1,\dots P} \in \mathbb{R}^{N\times mP}. \quad (4)$$

Let $\Gamma \in \mathbb{R}^{mP}$ denote the concatenation of the local sparse codes, i.e.

$$\Gamma := \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_P \end{bmatrix}.$$

Given a vector Γ as above, we will denote by $\tilde{R}_i$ the operator of extracting its i-th portion[1], i.e. $\tilde{R}_i \Gamma \equiv \alpha_i$.

Summarizing the above developments, we have the global convolutional representation for our signal as follows:

$$x = D_G \Gamma. \quad (5)$$

Next, applying R_i to both sides of (5) and using (3), we obtain

$$D\alpha_i = R_i x = R_i D_G \Gamma. \quad (6)$$

Let $\Omega_i := R_i D_G$ denote the i-th stripe from the global convolutional dictionary D_G. Thus (6) can be rewritten as

[1] Notice that while R_i extracts the i-th patch from the signal x, the operator $\tilde{R}_i$ extracts the representation α_i of $R_i x$ from Γ.

$$\underbrace{\left[\mathbf{0}\ \ldots\ \mathbf{0}\ D\ \mathbf{0}\ \ldots\ \mathbf{0}\right]}_{:=Q_i} \Gamma = \Omega_i \Gamma, \tag{7}$$

or $(Q_i - \Omega_i)\Gamma = 0$. Since this is true for all $i = 1, \ldots, P$, we have shown that the vector Γ satisfies

$$\underbrace{\begin{bmatrix} Q_1 - \Omega_1 \\ \vdots \\ Q_P - \Omega_P \end{bmatrix}}_{:=M \in \mathbb{R}^{nP \times mP}} \Gamma = 0.$$

Thus, the condition that the patches $R_i x$ agree on the overlaps is equivalent to the global representation vector Γ residing in the null-space of the matrix M.

An easy computation provides the dimension of this null-space (see proof in Appendix A: Proof of Lemma 1), or in other words the overall number of degrees of freedom of admissible Γ.

Lemma 1. *For any frame $D \in \mathbb{R}^{n \times m}$ (i.e. a full rank dictionary), we have*

$$\dim \ker M = N(m - n + 1).$$

Note that in particular for $m = n$ we have $\dim \ker M = N$, i.e. every signal admits a unique representation $x = D_G \Gamma$ where $\Gamma = \left(D^{-1} R_1 x, \ldots, D^{-1} R_P x\right)^T$.

Definition 6. Given $\Gamma = [\alpha_1, \ldots, \alpha_P]^T \in \mathbb{R}^{mP}$, the $\|\cdot\|_{0,\infty}$ pseudo-norm is defined by

$$\|\Gamma\|_{0,\infty} := \max_{i=1,\ldots,P} \|\alpha_i\|_0.$$

Thus, every signal complying with the patch-sparse model, with sparsity s for each patch, admits the following representation.

Theorem 1. *Given D, s, P, and N, the globalized local-sparse model (3) is equivalent to*

$$\mathscr{M} = \left\{x \in \mathbb{R}^N :\ x = D_G \Gamma,\ M\Gamma = 0,\ \|\Gamma\|_{0,\infty} \leqslant s\right\}. \tag{8}$$

Proof. If $x \in \mathscr{M}$ (according to (3)), then by the above construction x belongs to the set defined by the RHS of (8) (let's call it $\mathscr{M}^*$ for the purposes of this proof only). In the other direction, assume that $x \in \mathscr{M}^*$. Now $R_i x = R_i D_G \Gamma = \Omega_i \Gamma$, and since $M\Gamma = 0$, we have $R_i x = Q_i \Gamma = D\tilde{R}_i \Gamma$. Denote $\alpha_i := \tilde{R}_i \Gamma$, and so we have that $R_i x = D\alpha_i$ with $\|\alpha_i\|_0 \leqslant s$, i.e. $x \in \mathscr{M}$ by definition. □

What are the values of s we are interested in? In addition to the natural requirement that $s < n$ (as in Definition 5), we would like to have uniqueness of sparse representations. We say that α_i is a *minimal* representation of x_i if $x_i = D\alpha_i$ such that the matrix $D_{\operatorname{supp} \alpha_i}$ has full rank – and therefore the atoms participating in the representation are linearly independent. While we treat uniqueness in more detail in Subsection 2.4, at this point we would like to restrict the discussion to minimal patch representations. Notice that α_i might be a minimal representation but not a unique one with minimal sparsity.

Definition 7. Given a signal $x \in \mathscr{M}$, let us denote by $\rho(x)$ the set of all locally sparse *and minimal* representations of x:

$$\rho(x) := \left\{ \Gamma \in \mathbb{R}^{mP} : \|\Gamma\|_{0,\infty} \leqslant s, \; x = D_G \Gamma, \; M\Gamma = 0, \; D_{\operatorname{supp} \tilde{R}_i \Gamma} \text{ is full rank.} \right\}$$

Let us now go back to the definition (3). Consider a signal $x \in \mathscr{M}$, and let $\Gamma \in \rho(x)$. Denote $s_i := \operatorname{supp} \tilde{R}_i \Gamma$. Then we have $R_i x \in \mathscr{R}(D_{s_i})$ and therefore we can write $R_i x = P_{s_i} R_i x$, where P_{s_i} is the orthogonal projection operator onto $\mathscr{R}(D_{s_i})$. In fact, since D_{s_i} is full rank, we have $P_{s_i} = D_{s_i} D_{s_i}^{\dagger}$ where $D_{s_i}^{\dagger} = \left(D_{s_i}^T D_{s_i}\right)^{-1} D_{s_i}^T$ is the Moore-Penrose pseudoinverse of D_{s_i}.

Definition 8. Given a support sequence $\mathscr{S} = (s_1, \ldots, s_P)$, define the matrix $A_{\mathscr{S}}$ as follows:

$$A_{\mathscr{S}} := \begin{bmatrix} (I_n - P_{s_1}) R_1 \\ (I_n - P_{s_2}) R_2 \\ \vdots \\ (I_n - P_{s_P}) R_P \end{bmatrix} \in \mathbb{R}^{nP \times N}.$$

The map $A_{\mathscr{S}}$ measures the local patch discrepancies, i.e. how "far" is each local patch from the range of a particular subset of the columns of D.

Definition 9. Given a model $\mathscr{M}$, denote by $\Sigma_{\mathscr{M}}$ the set of all valid supports, i.e.

$$\Sigma_{\mathscr{M}} := \left\{ (s_1, \ldots, s_P) : \; \exists x \in \mathscr{M}, \; \Gamma \in \rho(x) \; s.t. \; \forall i = 1, \ldots, P : \; s_i = \operatorname{supp} \tilde{R}_i \Gamma \right\}.$$

With this notation in place, it is immediate to see that the global signal model is a union of subspaces.

Theorem 2. *The global model is equivalent to the union of subspaces*

$$\mathscr{M} = \bigcup_{\mathscr{S} \in \Sigma_{\mathscr{M}}} \ker A_{\mathscr{S}}.$$

Remark 1. Contrary to the well-known Union of Subspaces model [9, 32], the subspaces $\{\ker A_{\mathscr{S}}\}$ do not have in general a sparse joint basis, and therefore our model is distinctly different from the well-known block-sparsity model [21, 22].

An important question of interest is to estimate $\dim \ker A_{\mathscr{S}}$ for a given $\mathscr{S} \in \Sigma_{\mathscr{M}}$. One possible solution is to investigate the "global" structure of the corresponding signals (as is done in Subsection 4.1 and Subsection 4.2), while another option is to utilize information about "local connections" (Subsection 2.3 and Subsection 4.4).

2.3 Local support dependencies

In this section we highlight the importance of the local connections (briefly mentioned above) between the neighboring patches of the signal, and therefore between the corresponding subspaces containing those patches. This in turn allows to characterize $\Sigma_{\mathscr{M}}$ as the set of all "realizable" paths in a certain dependency graph derived from the dictionary D. This point of view allows to describe the model $\mathscr{M}$ using only the intrinsic properties of the dictionary, in contrast to Theorem 2.

First we show the equivalence of the condition $M\Gamma = 0$ to equality of pairwise overlaps.

Definition 10. Define the "extract from top/bottom" operators $S_T \in \mathbb{R}^{(n-1)\times n}$ and $S_B \in \mathbb{R}^{(n-1)\times n}$:

$$S_{T(op)} = \begin{bmatrix} I_{n-1} & \mathbf{0} \end{bmatrix}, \quad S_{B(ottom)} = \begin{bmatrix} \mathbf{0} & I_{n-1} \end{bmatrix}.$$

The following result is proved in Appendix B: Proof of Lemma 2.

Lemma 2. *Let $\Gamma = [\alpha_1, \dots, \alpha_P]^T$. Under the above definitions, the following are equivalent:*

1. $M\Gamma = 0$;
2. *For each $i = 1, \dots, P$, we have $S_B D\alpha_i = S_T D\alpha_{i+1}$.*

Definition 11. Let the matrix $M_* \in \mathbb{R}^{(n-1)P\times mP}$ be defined as

$$M_* := \begin{bmatrix} S_B D & -S_T D & & \\ & S_B D & -S_T D & \\ & & \ddots & \ddots \\ & & & \end{bmatrix}.$$

Corollary 1. *The global model is equivalent to*

$$\mathscr{M} = \left\{ x \in \mathbb{R}^N : \; x = D_G \Gamma, \; M_* \Gamma = 0, \; \|\Gamma\|_{0,\infty} \leqslant s \right\}.$$

Proposition 1. *Let $0 \neq x \in \mathscr{M}$ and $\Gamma \in \rho(x)$ with $\operatorname{supp}\Gamma = (s_1, \dots, s_P)$. Then for $i = 1, \dots, P$*

$$\operatorname{rank}\begin{bmatrix} S_B D_{s_i} & -S_T D_{s_{i+1}} \end{bmatrix} < |s_i| + |s_{i+1}| \leqslant 2s, \tag{9}$$

where by definition $\operatorname{rank}\emptyset = -\infty$.

Proof. $x \in \mathscr{M}$ implies by Lemma 2 that for every $i = 1, \dots P$

$$\begin{bmatrix} S_B D & -S_T D \end{bmatrix} \begin{bmatrix} \alpha_i \\ \alpha_{i+1} \end{bmatrix} = 0.$$

But

$$\begin{bmatrix} S_B D & -S_T D \end{bmatrix} \begin{bmatrix} \alpha_i \\ \alpha_{i+1} \end{bmatrix} = \begin{bmatrix} S_B D_{s_i} & -S_T D_{s_{i+1}} \end{bmatrix} \begin{bmatrix} \alpha_i|s_i \\ \alpha_{i+1}|s_{i+1} \end{bmatrix} = 0,$$

and therefore the matrix $\begin{bmatrix} S_B D_{s_i} & -S_T D_{s_{i+1}} \end{bmatrix}$ must be rank-deficient. Note in particular that the conclusion still holds if one (or both) of the $\{s_i, s_{i+1}\}$ is empty. □

The preceding result suggests a way to describe all the supports in $\Sigma_{\mathscr{M}}$.

Definition 12. Given a dictionary D, we define an abstract directed graph $\mathscr{G}_{D,s} = (V, E)$, with the vertex set

$$V = \left\{ (i_1, \dots, i_k) \subset [1, \dots, m] : \quad \operatorname{rank} D_{i_1,\dots,i_k} = k < n \right\},$$

and the edge set

$$E = \left\{ (s_1, s_2) \in V \times V : \quad \operatorname{rank} \begin{bmatrix} S_B D_{s_1} & -S_T D_{s_2} \end{bmatrix} < \min\{n-1, |s_1| + |s_2|\} \right\}.$$

In particular, $\emptyset \in V$ and $(\emptyset, \emptyset) \in E$ with $\operatorname{rank}[\emptyset] := -\infty$.

Remark 2. It might be impossible to compute $\mathscr{G}_{D,s}$ in practice. However we set this issue aside for now, and only explore the theoretical ramifications of its properties.

Definition 13. The set of all directed paths of length P in $\mathscr{G}_{D,s}$, not including the self-loop $\underbrace{(\emptyset, \emptyset, \dots \emptyset)}_{\times P}$, is denoted by $\mathscr{C}_{\mathscr{G}}(P)$.

Definition 14. A path $\mathscr{S} \in \mathscr{C}_{\mathscr{G}}(P)$ is called *realizable* if $\dim \ker A_{\mathscr{S}} > 0$. The set of all realizable paths in $\mathscr{C}_{\mathscr{G}}(P)$ is denoted by $\mathscr{R}_{\mathscr{G}}(P)$.

Thus we have the following result.

Theorem 3. *Suppose* $0 \neq x \in \mathscr{M}$. *Then*

1. *Every representation* $\Gamma = (\alpha_i)_{i=1}^P \in \rho(x)$ *satisfies* $\operatorname{supp} \Gamma \in \mathscr{C}_{\mathscr{G}}(P)$, *and therefore*

$$\Sigma_{\mathscr{M}} \subseteq \mathscr{R}_{\mathscr{G}}(P). \tag{10}$$

2. *The model* $\mathscr{M}$ *can be characterized "intrinsically" by the dictionary as follows:*

$$\mathscr{M} = \bigcup_{\mathscr{S} \in \mathscr{R}_{\mathscr{G}}(P)} \ker A_{\mathscr{S}}. \tag{11}$$

Proof. Let $\operatorname{supp} \Gamma = (s_1, \dots, s_P)$ with $s_i = \operatorname{supp} \alpha_i$ if $\alpha_i \neq \mathbf{0}$, and $s_i = \emptyset$ if $\alpha_i = \mathbf{0}$. Then by Proposition 1 we must have that

Algorithm 0.1 Constructing a signal from $\mathscr{M}$ via $\mathscr{G}$

1. Construct a path $\mathscr{S} \in \mathscr{C}_{\mathscr{G}}(P)$.
2. Construct the matrix $A_{\mathscr{S}}$.
3. Find a nonzero vector in $\ker A_{\mathscr{S}}$.

$$\operatorname{rank}\left[S_B D_{s_i} \; -S_T D_{s_{i+1}}\right] < |s_i| + |s_{i+1}| \leqslant 2s.$$

Furthermore, since $\Gamma \in \rho(x)$ we must have that D_{s_i} is full rank for each $i = 1, \dots, P$. Thus $(s_i, s_{i+1}) \in \mathscr{G}_{D,s}$, and so $\operatorname{supp}\Gamma \in \mathscr{R}_{\mathscr{G}}(P)$. Since by assumption $\operatorname{supp}\Gamma \in \Sigma_{\mathscr{M}}$, this proves (10).

To show (11), notice that if $\operatorname{supp}\Gamma \in \mathscr{R}_{\mathscr{G}}(P)$, then for every $x \in \ker A_{\operatorname{supp}\Gamma}$ we have $R_i x = P_{s_i} R_i x$, i.e. $R_i x = D\alpha_i$ for some α_i with $\operatorname{supp}\alpha_i \subseteq s_i$. Clearly in this case $|\operatorname{supp}\alpha_i| \leqslant s$ and therefore $x \in \mathscr{M}$. The other direction of (11) follows immediately from the definitions. □

Definition 15. The dictionary D is called "(s,P)-good" if

$$|\mathscr{R}_{\mathscr{G}}(P)| > 0.$$

Theorem 4. *The set of "(s,P)-good" dictionaries has measure zero in the space of all $n \times m$ matrices.*

Proof. Every low rank condition defines a finite number of algebraic equations on the entries of D (given by the vanishing of all the $2s \times 2s$ minors of $\left[S_B D_{s_i},\, S_T D_{s_j}\right]$). Since the number of possible graphs is finite (given fixed n, m and s), the resulting solution set is a finite union of semi-algebraic sets of low dimension, and hence has measure zero. □

The above considerations suggest that the good dictionaries are hard to come by; we provide explicit constructions in Section 4.

Now suppose the graph $\mathscr{G}$ is known (or can be easily constructed). Then this gives a simple procedure to generate signals from $\mathscr{M}$, presented in Algorithm 0.1.

An interesting question arises: given $\mathscr{S} \in \mathscr{C}_{\mathscr{G}}(P)$, can we say something about $\dim\ker A_{\mathscr{S}}$? In particular, when is it strictly positive (i.e. when $\mathscr{S} \in \mathscr{R}_{\mathscr{G}}(P)$?) While in general the question seems to be difficult, in some special cases this number can be estimated using only the properties of the local connections (s_i, s_{i+1}), by essentially counting the additional "degrees of freedom" when moving from patch i to patch $i+1$. We discuss this in more details in Subsection 4.4 (in particular see Proposition 11), while here we show the following easy result.

Proposition 2. *For every $\mathscr{S} \in \mathscr{R}_{\mathscr{G}}(P)$ we have*

$$\dim\ker A_{\mathscr{S}} = \dim\ker M_*^{(\mathscr{S})}.$$

Proof. Notice that

$$\ker A_{\mathscr{S}} = \left\{ D_G^{(\mathscr{S})} \Gamma_{\mathscr{S}},\, M_*^{(\mathscr{S})} \Gamma_{\mathscr{S}} = 0 \right\} = im \left(D_G^{(\mathscr{S})} |_{\ker M_*^{(\mathscr{S})}} \right),$$

and therefore $\dim \ker A_{\mathscr{S}} \leqslant \dim \ker M_*^{(\mathscr{S})}$. Furthermore, the map $D_G^{(\mathscr{S})}|_{\ker M_*^{(\mathscr{S})}}$ is injective, because if $D_G^{(\mathscr{S})} \Gamma_{\mathscr{S}} = 0$ and $M_*^{(\mathscr{S})} \Gamma_{\mathscr{S}} = 0$, we must have that $D_{s_i} \alpha_i |_{s_i} = 0$ and, since D_{s_i} has full rank, also $\alpha_i = 0$. The conclusion follows. □

2.4 Uniqueness and stability

Given a signal $x \in \mathscr{M}$, it has a globalized representation $\Gamma \in \rho(x)$ according to Theorem 1. When is such a representation unique, and under what conditions can it be recovered when the signal is corrupted with noise?

In other words, we study the problem

$$\min \|\Gamma\|_{0,\infty} \qquad s.t.\, D_G \Gamma = D_G \Gamma_0,\, M\Gamma = 0 \qquad (P_{0,\infty})$$

and its noisy version

$$\min \|\Gamma\|_{0,\infty} \qquad s.t.\, \|D_G \Gamma - D_G \Gamma_0\| \leqslant \varepsilon,\, M\Gamma = 0 \qquad \left(P_{0,\infty}^{\varepsilon}\right).$$

For this task, we define certain measures of the dictionary, similar to the classical spark, coherence function, and the Restricted Isometry Property, which take the additional dictionary structure into account. In general, the additional structure implies *possibly* better uniqueness as well as stability to perturbations, however it is an open question to show they are *provably* better in certain cases.

The key observation is that the global model $\mathscr{M}$ imposes a constraint on the allowed local supports.

Definition 16. Denote the set of allowed local supports by

$$\mathscr{T} := \{\tau :\ \exists (s_1, \ldots, \tau, \ldots, s_P) \in \Sigma_{\mathscr{M}}\}.$$

Recall the definition of the spark (1). Clearly $\sigma(D)$ can be equivalently rewritten as

$$\sigma(D) = \min\{j :\ \exists s_1, s_2 \subset [1, \ldots, m],\ |s_1 \cup s_2| = j,\ \operatorname{rank} D_{s_1 \cup s_2} < j\}. \qquad (12)$$

Definition 17. The *globalized spark* $\sigma^*(D)$ is

$$\sigma^*(D) := \min\{j :\ \exists s_1, s_2 \in \mathscr{T},\ |s_1 \cup s_2| = j,\ \operatorname{rank} D_{s_1 \cup s_2} < j\}. \qquad (13)$$

The following proposition is immediate by comparing (12) with (13).

Proposition 3. $\sigma^*(D) \geqslant \sigma(D)$.

The globalized spark provides a uniqueness result in the spirit of [17].

Theorem 5 (Uniqueness). *Let $x \in \mathscr{M}(D,s,N,P)$. If $\exists \Gamma \in \rho(x)$ for which $\|\Gamma\|_{0,\infty} < \frac{1}{2}\sigma^*(D)$ (i.e. it is a sufficiently sparse solution of $P_{0,\infty}$), then it is the unique solution (and so $\rho(x) = \{\Gamma\}$).*

Proof. Suppose that $\exists \Gamma_0 \in \rho(x)$ which is different from Γ. Put $\Gamma_1 := \Gamma - \Gamma_0$, then $\|\Gamma_1\|_{0,\infty} < \sigma^*(D)$, while $D_G\Gamma_1 = 0$ and $M\Gamma_1 = 0$. Denote $\beta_j := \tilde{R}_j\Gamma_1$. By assumption, there exists an index i for which $\beta_i \neq 0$, but we must have $D\beta_j = 0$ for every j, and therefore $D_{\operatorname{supp}\beta_i}$ must be rank deficient – contradicting the fact that $\|\beta_i\| < \sigma^*(D)$. □

In classical sparsity, we have the bound

$$\sigma(D) \geqslant \min\{s:\ \mu_1(s-1) \geqslant 1\}, \tag{14}$$

where μ_1 is given by Definition 3. In a similar fashion, the globalized spark σ^* can be bounded by an appropriate analog of "coherence" – however, computing this new coherence appears to be in general intractable.

Definition 18. Given the model $\mathscr{M}$, we define the following globalized coherence function

$$\mu_1^*(s) := \max_{S \in \mathscr{T} \cup \mathscr{T}, |S|=s} \max_{j \in S} \sum_{k \in S \setminus \{j\}} \left|\langle d_j, d_k \rangle\right|,$$

where $\mathscr{T} \cup \mathscr{T} := \{s_1 \cup s_2 :\ s_1, s_2 \in \mathscr{T}\}$.

Theorem 6. *The globalized spark σ^* can be bounded by the globalized coherence as follows[2]:*

$$\sigma^*(D) \geqslant \min\{s:\ \mu_1^*(s) \geqslant 1\}.$$

Proof. Following closely the corresponding proof in [17], assume by contradiction that

$$\sigma^*(D) < \min\{s:\ \mu_1^*(s) \geqslant 1\}.$$

Let $s^* \in \mathscr{T} \cup \mathscr{T}$ with $|s^*| = \sigma^*(D)$ for which D_{s^*} is rank-deficient. Then the restricted Gram matrix $G := D_{s^*}^T D_{s^*}$ must be singular. On the other hand, $\mu_1^*(|s^*|) < 1$, and so in particular

$$\max_{j \in s^*} \sum_{k \in s^* \setminus \{j\}} \left|\langle d_j, d_k \rangle\right| < 1.$$

But that means that G is diagonally dominant and therefore $\det G \neq 0$, a contradiction. □

[2] In general $\min\{s:\ \mu_1^*(s-1) \geqslant 1\} \neq \max\{s:\ \mu_1^*(s) < 1\}$ because the function μ_1^* need not be monotonic.

We see that $\mu_1^*(s+1) \leqslant \mu_1(s)$ since the outer maximization is done on a smaller set. Therefore, in general the bound of Theorem 6 appears to be sharper than (14).

A notion of globalized RIP can also be defined as follows.

Definition 19. The globalized RIP constant of order k associated to the model $\mathscr{M}$ is the smallest number $\delta_{k,\mathscr{M}}$ such that

$$\left(1-\delta_{k,\mathscr{M}}\right)\|\alpha\|_2^2 \leqslant \|D\alpha\|_2^2 \leqslant \left(1+\delta_{k,\mathscr{M}}\right)\|\alpha\|_2^2$$

for every $\alpha \in \mathbb{R}^m$ with $\operatorname{supp}\alpha \in \mathscr{T}$.

Immediately one can see the following (recall Definition 4).

Proposition 4. *The globalized RIP constant is upper bounded by the standard RIP constant:*

$$\delta_{k,\mathscr{M}} \leqslant \delta_k.$$

Definition 20. The generalized RIP constant of order k associated to signals of length N is the smallest number $\delta_k^{(N)}$ such that

$$\left(1-\delta_k^{(N)}\right)\|\Gamma\|_2^2 \leqslant \|D_G\Gamma\|_2^2 \leqslant \left(1+\delta_k^{(N)}\right)\|\Gamma\|_2^2$$

for every $\Gamma \in \mathbb{R}^{mN}$ satisfying $M\Gamma = 0,\ \|\Gamma\|_{0,\infty} \leqslant k$.

Proposition 5. *We have*

$$\delta_k^{(N)} \leqslant \frac{\delta_{k,\mathscr{M}}+(n-1)}{n} \leqslant \frac{\delta_k+(n-1)}{n}.$$

Proof. Obviously it is enough to show only the leftmost inequality. If $\Gamma = (\alpha_i)_{i=1}^N$ and $\|\Gamma\|_{0,\infty} \leqslant k$, this gives $\|\alpha_i\|_0 \leqslant k$ for all $i = 1,\dots,P$. Further, setting $x := D_G\Gamma$ we clearly have $\Gamma \in \rho(x)$ and so $\operatorname{supp}\Gamma \in \Sigma_{\mathscr{M}}$. Thus $\operatorname{supp}\alpha_i \in \mathscr{T}$, and therefore

$$\left(1-\delta_{k,\mathscr{M}}\right)\|\alpha_i\|_2^2 \leqslant \|D\alpha_i\|_2^2 \leqslant \left(1+\delta_{k,\mathscr{M}}\right)\|\alpha_i\|_2^2.$$

By Corollary 4 we know that for every Γ satisfying $M\Gamma = 0$, we have

$$\|D_G\Gamma\|_2^2 = \frac{1}{n}\sum_{i=1}^N \|D\alpha_i\|_2^2.$$

Now for the lower bound,

$$\begin{aligned}\|D_G\Gamma\|_2^2 &\geqslant \frac{1-\delta_{k,\mathscr{M}}}{n}\sum_{i=1}^N \|\alpha_i\|_2^2 = \left(1-1+\frac{1-\delta_{k,\mathscr{M}}}{n}\right)\|\Gamma\|_2^2 \\ &= \left(1-\frac{\delta_{k,\mathscr{M}}+(n-1)}{n}\right)\|\Gamma\|_2^2.\end{aligned}$$

For the upper bound,

$$\begin{aligned}\|D_G \Gamma\|_2^2 &\leqslant \frac{1+\delta_{k,\mathscr{M}}}{n} \sum_{i=1}^{N} \|\alpha_i\|_2^2 < \left(1 + \frac{\delta_{k,\mathscr{M}}+1}{n}\right) \|\Gamma\|_2^2 \\ &\leqslant \left(1 + \frac{\delta_{k,\mathscr{M}} + (n-1)}{n}\right) \|\Gamma\|_2^2.\end{aligned}$$

□

Theorem 7 (Uniqueness and stability of $P_{0,\infty}$ via RIP). *Suppose that $\delta_{2s}^{(N)} < 1$, and suppose further that $x = D_G \Gamma_0$ with $\|\Gamma_0\|_{0,\infty} = s$ and $\|D_G \Gamma_0 - x\|_2 \leqslant \varepsilon$. Then every solution $\hat{\Gamma}$ of the noise-constrained $P_{0,\infty}^{\varepsilon}$ problem*

$$\hat{\Gamma} \leftarrow \arg\min_{\Gamma} \|\Gamma\|_{0,\infty} \; s.t. \; \|D_G \Gamma - x\| \leqslant \varepsilon, \; M\Gamma = 0$$

satisfies

$$\|\hat{\Gamma} - \Gamma_0\|_2^2 \leqslant \frac{4\varepsilon^2}{1 - \delta_{2s}^{(N)}}.$$

In particular, Γ_0 is the unique solution of the noiseless $P_{0,\infty}$ problem.

Proof. Immediate using the definition of the globalized RIP:

$$\begin{aligned}\|\hat{\Gamma} - \Gamma_0\|_2^2 &< \frac{1}{1 - \delta_{2s}^{(N)}} \|D_G \left(\hat{\Gamma} - \Gamma_0\right)\|_2^2 \leqslant \frac{1}{1 - \delta_{2s}^{(N)}} \left(\|D_G \hat{\Gamma} - x\|_2 + \|D_G \Gamma_0 - x\|_2\right)^2 \\ &\leqslant \frac{4\varepsilon^2}{1 - \delta_{2s}^{(N)}}.\end{aligned}$$

□

3 Pursuit algorithms

In this section we consider the problem of efficient projection onto the model $\mathscr{M}$. First we treat the "oracle" setting, i.e. when the supports of the local patches (and therefore of the global vector Γ) is known. We show that the patch averaging (LPA) method is not a good projector, however repeated application of it does achieve the desired result.

For the non-oracle setting, we consider "local" and "globalized" pursuits. The former type does not use any dependencies between the patches, and tries to reconstruct the supports α_i completely locally, using standard methods such as OMP – and as we demonstrate, it can be guaranteed to succeed in more cases than the standard analysis would imply. However a possibly better alternative exists, namely a "globalized" approach with the patch disagreements as a major driving force.

3.1 Global (oracle) projection, local patch averaging (LPA) and the local-global gap

Here we briefly consider the question of efficient projection onto the subspace $\ker A_{\mathscr{S}}$, given $\mathscr{S}$.

As customary in the literature [14], the projector onto $\ker A_{\mathscr{S}}$ can be called *an oracle*. In effect, we would like to compute

$$x_G(y,\mathscr{S}) := \arg\min_x \|y-x\|_2^2 \qquad s.t.\, A_{\mathscr{S}}x = 0, \tag{15}$$

given $y \in \mathbb{R}^N$.

To make things concrete, let us assume the standard Gaussian noise model:

$$y = x + \mathscr{N}\left(0,\sigma^2 I\right). \tag{16}$$

The following is well-known.

Proposition 6. *In the Gaussian noise model* (16)*, the performance of the oracle estimator* (15) *is*

$$MSE(x_G) = (\dim\ker A_{\mathscr{S}})\,\sigma^2.$$

Let us now turn to the LPA method. The (linear part of) LPA is the solution to the minimization problem:

$$\hat{x} = \arg\min_x \sum_{i=1}^{P} \|R_i x - P_{s_i} R_i y\|_2^2,$$

where y is the noisy signal. This has a closed-form solution

$$\hat{x}_{LPA} = \left(\sum_i R_i^T R_i\right)^{-1}\left(\sum_i R_i^T P_{s_i} R_i\right) y = \underbrace{\left(\frac{1}{n}\sum_i R_i^T P_{s_i} R_i\right)}_{:=M_A} y. \tag{17}$$

Again, the following fact is well-established.

Proposition 7. *In the Gaussian noise model* (16)*, the performance of the averaging estimator* (17) *is*

$$MSE(\hat{x}_{LPA}) = \sigma^2 \sum_{i=1}^{N} \lambda_i,$$

where $\{\lambda_1,\dots,\lambda_N\}$ are the eigenvalues of $M_A M_A^T$.

Thus, there exists a *local-global gap* in the oracle setting, illustrated in Figure 1 on page 16. In Subsection 4.1 we estimate this gap for a specific case of piecewise-constant signals.

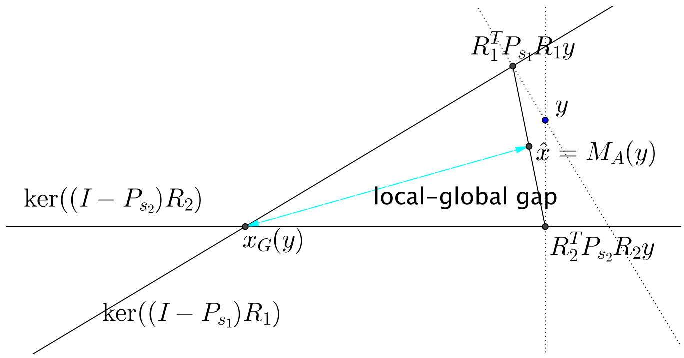


Fig. 1: The local-global gap, oracle setting. Illustration for the case $P = 2$.

The following result is proved in Appendix C: Proof of Theorem 8.

Theorem 8. *Iterations of* (17) *converge to* x_G.

To conclude, *the iterated LPA algorithm provides an efficient method for computing the global oracle projection* x_G.

3.2 Local pursuit guarantees

Now we turn to the question of projection onto the model $\mathscr{M}$ when the support of Γ is not known.

Here we show that the local OMP [15, 40] in fact succeeds in more cases than can be predicted by the classical unconstrained sparse model for each patch. We use the modified coherence function (which is unfortunately intractable to compute)

$$\eta_1^*(s) := \max_{S \in \mathscr{T}} \left(\max_{j \in S} \sum_{k \in S \setminus \{j\}} \left| \langle d_k, d_j \rangle \right| + \max_{j \notin S} \sum_{k \in S} \left| \langle d_k, d_j \rangle \right| \right).$$

The proof of the following theorem is very similar to proving the guarantee for the standard OMP via the Babel function (Definition 3), see e.g. [23, Theorem 5.14] – and therefore we do not reproduce it here.

Theorem 9. *If* $\eta_1^*(s) < 1$ *then the local OMP will recover the supports of any* $x \in \mathscr{M}$.

Since the modified coherence function takes the allowed local supports into consideration, one can readily conclude that

$$\eta_1^*(s) \leqslant \mu_1(s) + \mu_1(s-1),$$

and therefore Theorem 9 gives in general a better guarantee than the one based on μ_1.

3.3 Globalized pursuits

We now turn to consider several pursuit algorithms, aiming at solving the $P_{0,\infty}/P_{0,\infty}^{\varepsilon}$ problems, in the globalized model. The main question is how to project the patch supports onto the nonconvex set $\Sigma_{\mathscr{M}}$.

The core idea is to relax the constraint $M_* \Gamma = 0,\ \|\Gamma\|_{0,\infty} \leqslant s$, and allow for some patch disagreements, so that the term $\|M_* \Gamma_k\|$ is not exactly zero. Intuitive explanation is as follows: the disagreement term "drives" the pursuit, and the probability of success is higher because we only need to "jump-start" it with the first patch, and then by strengthening the weight of the penalty related to this constraint the supports will "align" themselves correctly. Justifying this intuition, at least in some cases, is a future research goal.

3.3.1 Q-OMP

Given $\beta > 0$, we define

$$Q_\beta := \begin{bmatrix} D_G \\ \beta M_* \end{bmatrix}.$$

The main idea of the Q-OMP algorithm is to substitute the matrix Q_β as a proxy for the constraint $M_* \Gamma = 0$, by plugging it as a dictionary to the OMP algorithm. Then, given the obtained support $\mathscr{S}$, as a way to ensure that this constraint is met, one can construct the matrix $A_{\mathscr{S}}$ and project the signal onto the subspace $\ker A_{\mathscr{S}}$ (in Subsection 3.1 we show how such a projection can be done efficiently). The Q-OMP algorithm is detailed in Algorithm 0.2. Let us re-emphasize the point that various values of β correspond to different weightings of the model constraint $M_* \Gamma = 0$, and this might possibly become useful when considering relaxed models (see Section 6).

3.3.2 ADMM-based approach

In what follows we extend the above idea and develop an ADMM-type pursuit [11]. We start with the following global objective:

$$\hat{x} \leftarrow \arg\min_x \|y - x\|_2^2 \quad \text{s.t. } x = D_G \Gamma, M_* \Gamma = 0,\ \|\Gamma\|_{0,\infty} < K.$$

Algorithm 0.2 The Q-OMP algorithm – A globalized pursuit

Given: noisy signal y, dictionary D, local sparsity s, parameter $\beta > 0$.

1. Construct the matrix Q_β.
2. Run the OMP algorithm on the vector $Y := \begin{bmatrix} y \\ \mathbf{0} \end{bmatrix}$, with the dictionary Q_β and sparsity sN. Obtain the global support vector $\hat{\Gamma}$ with $\operatorname{supp}\hat{\Gamma} = \hat{\mathscr{S}}$.
3. Construct the matrix $A_{\hat{\mathscr{S}}}$ and project y onto $\ker A_{\hat{\mathscr{S}}}$.

Clearly, it is equivalent to $\hat{x} = D_G\hat{\Gamma}$, where

$$\hat{\Gamma} \leftarrow \arg\min_{\Gamma} \|y - D_G\Gamma\|_2^2 \quad s.t.\, M_*\Gamma = 0,\ \|\Gamma\|_{0,\infty} < K. \tag{18}$$

Applying Corollary 4, we have the following result.

Proposition 8. *The following problem is equivalent to (18):*

$$\begin{aligned} \hat{\Gamma} &\leftarrow \arg\min_{\{\alpha_i\}} \sum_{i=1}^{P} \|R_i y - D\alpha_i\|_2^2 \\ s.t.\ S_B D\alpha_i &= S_T D\alpha_{i+1} \text{ and } \|\alpha_i\|_0 < K \text{ for } i = 1,\ldots,P. \end{aligned} \tag{19}$$

We propose to approximate solution of the nonconvex problem (19) as follows. Define new variables z_i (which we would like to be equal to α_i eventually), and rewrite the problem in ADMM form (here Z is the concatenation of all the z_i's):

$$\{\hat{\Gamma}, \hat{Z}\} \leftarrow \arg\min_{\Gamma, Z} \sum_{i=1}^{P} \|R_i y - D\alpha_i\|_2^2 \quad s.t.\, S_B D\alpha_i = S_T D z_{i+1},\ \alpha_i = z_i,\ \|\alpha_i\|_0 < K.$$

The constraints can be written in concise form

$$\underbrace{\begin{bmatrix} I \\ S_B D \end{bmatrix}}_{:=A} \alpha_i = \underbrace{\begin{bmatrix} I & 0 \\ 0 & S_T D \end{bmatrix}}_{:=B} \begin{pmatrix} z_i \\ z_{i+1} \end{pmatrix},$$

and so globally we would have the following structure (for $N = 3$)

$$\underbrace{\begin{bmatrix} A & & \\ & A & \\ & & A \end{bmatrix}}_{:=\tilde{A}} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = \underbrace{\begin{bmatrix} I & & \\ & S_T D & \\ & I & \\ & & S_T D \\ & & I \\ S_T D & & \end{bmatrix}}_{:=\tilde{B}} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix}$$

Algorithm 0.3 The ADMM-based pursuit for $P_{0,\infty}$

Given: noisy signal y, dictionary D, local sparsity s, parameter $\rho > 0$. The augmented Lagrangian is

$$L_\rho\left(\{\alpha_i\},\{z_i\},\{u_i\}\right) = \sum_{i=1}^{P} \|R_i y - D\alpha_i\|_2^2 + \frac{\rho}{2}\sum_{i=1}^{P} \|A\alpha_i - B\begin{pmatrix} z_i \\ z_{i+1}\end{pmatrix} + u_i\|_2^2.$$

1. Repeat until convergence:

   a. Minimization wrt $\{\alpha_i\}$ is a batch-OMP:

$$\alpha_i^{k+1} \leftarrow \arg\min_{\alpha_i} \|R_i y - D\alpha_i\|_2^2 + \frac{\rho}{2}\|A\alpha_i - B\begin{pmatrix} z_i^k \\ z_{i+1}^k\end{pmatrix} + u_i^k\|_2^2, \quad s.t. \|\alpha_i\|_0 < K$$

$$\alpha_i^{k+1} \leftarrow OMP\left(\tilde{D} = \begin{bmatrix} D \\ \sqrt{\frac{\rho}{2}}A \end{bmatrix}, \tilde{y}_i^k = \begin{pmatrix} R_i y \\ \sqrt{\frac{\rho}{2}}\left(B\begin{pmatrix} z_i^k \\ z_{i+1}^k\end{pmatrix} - u_i^k\right)\end{pmatrix}, K\right).$$

   b. Minimization wrt z is a least squares problem with a sparse matrix, which can be implemented efficiently:

$$Z^{k+1} \leftarrow \arg\min_Z \|\tilde{A}\Gamma^{k+1} + U^k - \tilde{B}Z\|_2^2$$

   c. Dual update:

$$U^{k+1} \leftarrow \tilde{A}\Gamma^{k+1} - \tilde{B}Z + U^k.$$

2. Compute $\hat{y} := D_G\hat{\Gamma}$.

Our ADMM-based method is defined in Algorithm 0.3.

4 Examples

We now turn to present several classes of signals that belong to the proposed globalized model, where each of these is obtained by imposing a special structure on the local dictionary. Then, we demonstrate how one can sample from $\mathscr{M}$ and generate such signals.

4.1 *Piecewise constant signals*

The (unnormalized) Heaviside $n \times n$ dictionary H_n is the upper triangular matrix with 1's in the upper part (see Figure 2 on page 20). Formally, each local atom d_i of length n is expressed as a step function, given by $d_i^T = [\mathbf{1_i}\ ,\ \mathbf{0_{n-i}}]^T$, $1 \leq i \leq n$, where $\mathbf{1_i}$ is a vector of ones of length i. Similarly, $\mathbf{0}_{n-i}$ is a zero-vector of length $n-i$. The following property is verified by noticing that H_n^{-1} is the discrete difference operator.

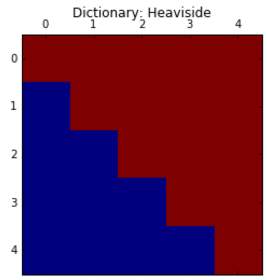


Fig. 2: Heaviside dictionary H_4. Red is 1, blue is 0.

Proposition 9. *If a patch $x_i \in \mathbb{R}^n$ has $L-1$ steps, then its (unique) representation in the Heaviside dictionary H_n has at most L nonzeros.*

Corollary 2. *Let $x \in \mathbb{R}^N$ be a piecewise-constant signal with at most $L-1$ steps per each segment of length n(in the periodic sense). Then*

$$x \in \mathscr{M}\left(H_n, L, N, P=N\right).$$

Remark 3. The model $\mathscr{M}\left(H_n, L, N, P=N\right)$ contains also some signals having exactly L steps in a particular patch, but those patches must have their last segment with zero height.

As an example, one might synthesize signals with sparsity $\|\Gamma\|_{0,\infty} \leq 2$ according to the following scheme:

1. Draw at random the support of Γ with the requirement that the distance between the jumps within the signal will be at least the length of a patch (this allows at most 2 non-zeros per patch, one for the step and the second for the bias/DC).
2. Multiply each step by a random number.

The global subspace $A_{\mathscr{S}}$ and the corresponding global oracle denoiser x_G (15) in the PWC model can be explicitly described.

Proposition 10. *Let $x \in \mathbb{R}^N$ consist of s constant segments with lengths ℓ_r, $r = 1, \ldots, s$, and let Γ be the (unique) global representation of x in $\mathscr{M}$ (i.e. $\rho(x) = \{\Gamma\}$). Then*

$$\ker A_{\operatorname{supp}\Gamma} = \ker\left(I_N - \operatorname{diag}\left(B_r\right)_{r=1}^{s}\right), \tag{20}$$

where $B_r = \frac{1}{\ell_r}\mathbf{1}_{\ell_r \times \ell_r}$. Therefore, $\dim\ker A_{\operatorname{supp}\Gamma} = s$ and $MSE\left(\hat{x}_G\right) = s\sigma^2$ under the Gaussian noise model (16).

Proof. Every signal $y \in \ker A_{\operatorname{supp}\Gamma}$ has the same "local jump pattern" as x, and therefore it also has the same *global* jump pattern. That is, every such y consists of s constant segments with lengths ℓ_r. It is an easy observation that such signals satisfy

$$y = \operatorname{diag}(B_r)_{r=1}^{s} y.$$

This proves (20). It is easy to see that $\dim\ker(I_{\ell_r} - B_r) = 1$, and therefore

$$\ker(I_N - \operatorname{diag}(B_r)_{r=1}^{s}) = s.$$

The proof is finished by invoking Proposition 6. □

In other words, the global oracle is the averaging operator within the constant segments of the signal, which is quite intuitive.

It turns out that the LPA performance (and the local-global gap) can be accurately described by the following result. We provide an outline of proof in 6.3.

Theorem 10. *Let $x \in \mathbb{R}^N$ consist of s constant segments with lengths ℓ_r, $r = 1, \dots, s$, and assume the Gaussian noise model (16). Then*

1. *There exists a function $R(n,\alpha) : \mathbb{N} \times \mathbb{N} \to \mathbb{R}^+$, with $R(n,\alpha) > 1$, such that*

$$MSE(\hat{x}_{LPA}) = \sigma^2 \sum_{r=1}^{s} R(n, \ell_r).$$

2. *The function $R(n,\alpha)$ satisfies:*
   a. $R(n,\alpha) = 1 + \frac{\alpha\left(2\alpha H_\alpha^{(2)} - 3\alpha + 2\right) - 1}{n^2}$ *if $n \geqslant \alpha$, where $H_\alpha^{(2)} = \sum_{k=1}^{\alpha} \frac{1}{k^2}$;*
   b. $R(n,\alpha) = \frac{11}{18} + \frac{2\alpha}{3n} + \frac{6\alpha - 11}{18n^2}$ *if $n \leqslant \frac{\alpha}{2}$.*

Corollary 3. *The function $R(n,\alpha)$ is monotonically increasing in α (with n fixed) and monotonically decreasing in n (with α fixed). Furthermore,*

1. $\lim_{n\to\infty} R(n,n) = \frac{\pi^2}{3} - 2 \approx 1.28968$;
2. $\lim_{n\to\infty} R(n,2n) = \frac{35}{18} \approx 1.9444$.

Thus, for reasonable choices of the patch size, the local-global gap is roughly a constant multiple of the number of segments, reflecting the global complexity of the signal.

For numerical examples of reconstructing the PWC signals using our local-global framework, see Subsection 5.2.

4.2 Signature-type dictionaries

Another type of signals that comply with our model are those represented via a signature dictionary, which has been shown to be effective for image restoration [5].

Algorithm 0.4 Constructing the signature dictionary

1. Choose the base signal $x \in \mathbb{R}^m$.
2. Compute $D(x) = [R_1 x, R_2 x, ..., R_m x]$, where R_i extracts the i-th patch of size n in a cyclic fashion.
3. Normalization: $\widetilde{D}(x) = [d_1, \ldots, d_m]$, where $d_i = \frac{R_i x}{\|R_i x\|_2}$.

This dictionary is constructed from a small signal, $x \in \mathbb{R}^m$, such that its every patch (in varying location, extracted in a cyclic fashion), $R_i x \in \mathbb{R}^n$, is a possible atom in the representation, namely $d_i = R_i x$. As such, every consecutive pair of atoms $(i, i+1)$ is essentially a pair of overlapping patches that satisfy $S_B d_i = S_T d_{i+1}$ (before normalization). The complete algorithm is presented for convenience in Algorithm 0.4.

Given D as above, one can generate signals $y \in \mathbb{R}^N$, where N is an integer multiple of m, with s non-zeros per patch, by the easy procedure outlined below.

1. Init: Construct a base signal $b \in \mathbb{R}^N$ by replicating $x \in \mathbb{R}^m$ N/m times (note that b is therefore periodic). Set $y = 0$.
2. Repeat for $j = 1, \ldots, s$:
   a. Shift: Circularly shift the base signal by t_j positions, denoted by $\text{shift}(b, t_j)$, for some $t_j = 0, 1, \ldots, m-1$ (drawn at random).
   b. Aggregate: $y = y + \omega_j \cdot \text{shift}(b, t_j)$, where ω is an arbitrary random scalar.

Notice that a signal constructed in this way must be periodic, as it is easily seen that

$$\ker A_{\mathscr{S}} = \text{span}\left\{\text{shift}\left(b, t_i\right)\right\}_{i=1}^{s},$$

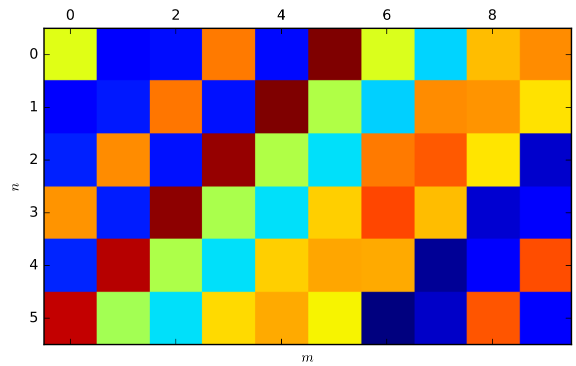
while the support sequence $\mathscr{S}$ is

$$\mathscr{S} = \left(\left[t_1, t_2, \ldots, t_s\right], \left[t_1, t_2, \ldots, t_s\right] + 1, \ldots, \left[t_1, t_2, \ldots, t_s\right] + N\right) \pmod{m}.$$

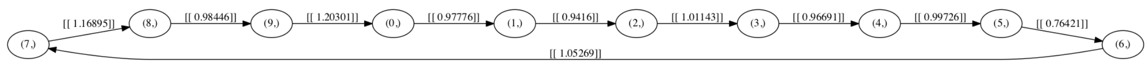
Assuming that there are no additional relations between the single atoms of D except those from the above construction, the dependency graph $\mathscr{G}_{D,1}$ of the resulting dictionary is easily seen to be *cyclic,* and all $\mathscr{S} \in \Sigma_{\mathscr{M}}$ are of the above form.

In Figure 3 on page 23 we give an example of a signature-type dictionary D for $(n, m) = (6, 10)$, its dependency graph $\mathscr{G}_D$, and a signal x with $N = P = 30$ together with its corresponding sparse representation Γ.

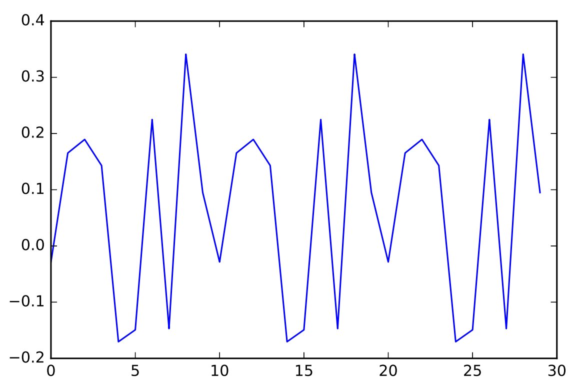
Remark 4. It might seem that every $n \times m$ Hankel matrix such as the one shown in Figure 3 on page 23 produces a signature-type dictionary with a nonempty signal space $\mathscr{M}$. However this is not the case, because such a dictionary will usually fail to generate signals of length larger than $n + m - 1$, as its dependency graph will not be cyclic (but rather consist of a single chain of nodes).



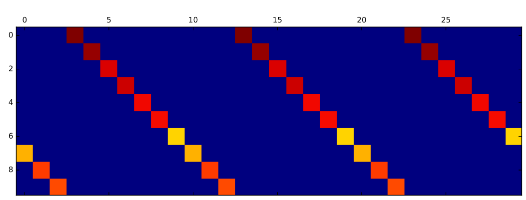
(a) The dictionary matrix D



(b) The dependency graph $\mathscr{G}_{D,1}$. The numerical values above the edges are the transfer matrices (scalars) $C_{i,j}$, satisfying $S_B d_i = C_{i,j} S_T d_j$ (see Subsection 4.4).



(c) The signal $x \in \ker A_{\mathscr{S}}$ for $\mathscr{S}$ generated by $t_1 = 6$ and $s = 1$, with $P = N = 30$.



(d) The coefficient matrix Γ corresponding to the signal x in (c)

Fig. 3: An example of the signature dictionary with $n = 6$, $m = 10$. See Remark 4.

Algorithm 0.5 Constructing the multi-signature dictionary

1. Input: n, m, s such that s divides m. Put $r := \frac{m}{s}$.
2. Select a signal basis matrix $X \in \mathbb{R}^{r \times s}$ and r nonsingular transfer matrices $M_i \in \mathbb{R}^{s \times s}$, $i = 1, \dots, r$.
3. Repeat for $i = 1, \dots, r$:
   a. Let $Y_i = [y_{i,1}, \dots, y_{i,s}] \in \mathbb{R}^{n \times s}$, where each $y_{i,j}$ is the i-th patch (of length n) of the signal x_j.
   b. Put the s-tuple $[d_{i,1}, \dots, d_{i,s}] = Y_i \times M_i$ as the next s atoms in D.

4.2.1 Multi-signature dictionaries

One can generalize the construction of Subsection 4.2 and consider s-tuples of initial base signals $x_i, \dots, x_s$, instead of a single x. The desired dictionary D will consist of corresponding s-tuples of atoms, which are constructed from those base signals. In order to avoid ending up with the same structure as the case $s = 1$, we also require a "mixing" of the atoms. The complete procedure is outlined in Algorithm 0.5.

In order to generate a signal of length N from $\mathscr{M}$, one can follow these steps (again we assume that m divides N):

1. Create a base signal matrix $X^G \in \mathbb{R}^{N \times s}$ by stacking $s\frac{N}{m}$ copies of the original basis matrix X. Set $y = 0$.
2. Repeat for $j = 1, \dots, k$:
   a. Select a base signal $b_j \in \mathscr{R}\left(X^G\right)$ and shift it (in a circular fashion) by some $t_j = 0, 1, \dots, R-1$.
   b. Aggregate: $y = y + \text{shift}(b_j, t_j)$ (note that here we do not need to multiply by a random scalar).

This procedure will produce a signal y of local sparsity $k \cdot s$. The corresponding support sequence can be written as

$$\mathscr{S} = (s_1, s_2, \dots, s_N),$$

where $s_i = s_1 + i \pmod{m}$ and

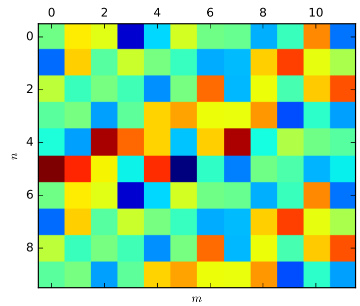
$$s_1 = \left[(t_1, 1), (t_1, 2), \dots, (t_1, s), \dots, (t_k, 1), (t_k, 2), \dots, (t_k, s)\right].$$

Here (t_j, i) denotes the atom $d_{t_j, i}$ in the notation of Algorithm 0.5. The corresponding signal space is

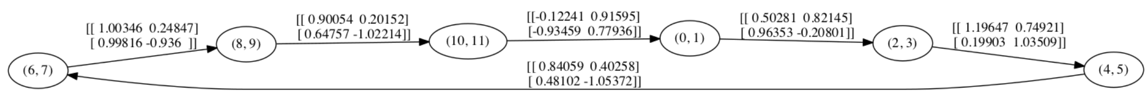
$$\ker A_{\mathscr{S}} = \text{span}\left\{\text{shift}\left(X^G, t_j\right)\right\}_{j=1}^{k},$$

and it is of dimension $k \cdot s$.

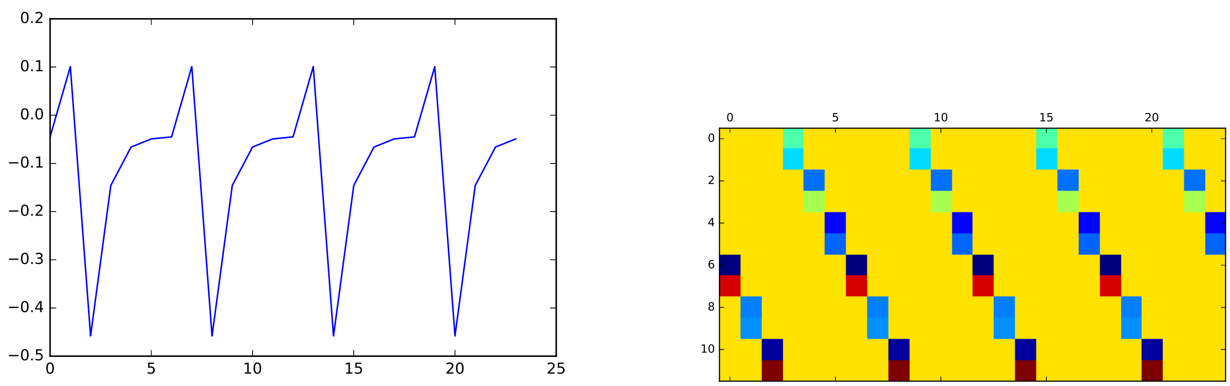
An example of a multi-signature dictionary and corresponding signals may be seen in Figure 4 on page 25.



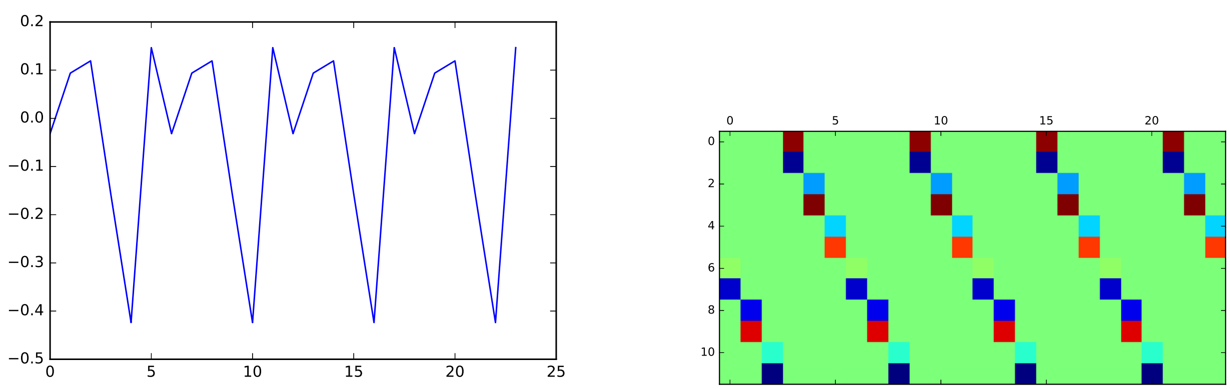
(a) The dictionary D



(b) The dependency graph $\mathscr{G}_{D,2}$. The numerical values above the edges are the transfer matrices $C_{i,j}$, explained below in Subsection 4.4.



(c) The first signal and its sparse representation in $\ker A_{\mathscr{S}}$ with $N = 24$, $k = 1$ and $t_1 = 5$.



(d) The second signal and its sparse representation in $\ker A_{\mathscr{S}}$.

Fig. 4: Example of multi-signature dictionary with $n = 10,\ m = 12,\ s = 2$.

4.3 Convolutional dictionaries

An important class of signals is the *sparse convolution model,* where each signal $x \in \mathbb{R}^N$ can be written as a linear combination of shifted "waveforms" $\boldsymbol{d}_i \in \mathbb{R}^n$, each $\boldsymbol{d}_i$ being a column in the local dictionary $D' \in \mathbb{R}^{n \times m}$. More conveniently, any such x can be represented as a circular convolution of $\boldsymbol{d}_i$ with a (sparse) "feature map" $\boldsymbol{\psi}_i \in \mathbb{R}^N$:

$$x = \sum_{i=1}^{m} \boldsymbol{d}_i *_N \boldsymbol{\psi}_i. \tag{21}$$

Such signals arise in various applications, such as audio classification [8, 25, 46], neural coding [18, 41], mid-level image representation and denoising [30, 54, 53].

Formally, the convolutional class can be re-cast into the patch-sparse model of this paper as follows. First, we can rewrite (21) as

$$x = \underbrace{\left[\boldsymbol{C}_1 \; \boldsymbol{C}_2 \; \dots \; \boldsymbol{C}_m\right]}_{:=\boldsymbol{E}} \boldsymbol{\Psi},$$

where each $\boldsymbol{C}_i \in \mathbb{R}^{N \times N}$ is a banded circulant matrix with its first column being equal to $\boldsymbol{d}_i$, and $\boldsymbol{\Psi} \in \mathbb{R}^{Nm}$ is the concatenation of the $\boldsymbol{\psi}_i$'s. It is easy to see that by permuting the columns of $\boldsymbol{E}$ one obtains precisely the global convolutional dictionary nD_G based on the local dictionary D' (recall (4)). Therefore we obtain

$$x = \underbrace{D_G\left(D'\right)}_{:=D'_G} \boldsymbol{\Gamma'}. \tag{22}$$

While it is tempting to conclude from comparing (22) and (5) that the convolutional model is equivalent to the patch-sparse model, an essential ingredient is missing, namely the requirement of equality on overlaps, $M\boldsymbol{\Gamma'} = 0$. Indeed, nothing in the definition of the convolutional model restricts the representation $\boldsymbol{\Psi}$ (and therefore $\boldsymbol{\Gamma'}$), therefore in principle the number of degrees of freedom remains Nm, as compared to $N(m-n+1)$ from Proposition 2.

To fix this, following [38, 39] we apply R_i to (22) and obtain $R_i x = R_i D'_G \boldsymbol{\Gamma'}$. The "stripe" $\Omega'_i = R_i D'_G$ has only $(2n-1)m$ nonzero consecutive columns, and in fact the nonzero portion of Ω'_i is equal for all i. This implies that every x_i has a representation $x_i = \Theta \boldsymbol{\gamma}_i$ in the "pseudo-local" dictionary

$$\Theta\left(D'\right) := \left[Z_B^{(n-1)} D' \; \dots \; D' \; \dots \; Z_T^{(n-1)} D'\right] \in \mathbb{R}^{n \times (2n-1)m},$$

where the operators $Z_B^{(k)}$ and $Z_T^{(k)}$ are given by Definition 10 in Appendix B: Proof of Lemma 2. If we now assume that our convolutional signals satisfy

$$\|\boldsymbol{\gamma}_i\|_0 \leqslant s \qquad \forall i,$$

then we have shown that they belong to $\mathscr{M}\left(\Theta\left(D'\right), s, P, N\right)$, and thus can be formally treated by the framework we have developed.

It turns out that this direct approach is quite naive, as the dictionary $\Theta\left(D'\right)$ is extremely bad-equipped for sparse reconstruction (for example it has repeated atoms, and therefore $\mu\left(\Theta\left(D'\right)\right) = 1$). To tackle this problem, a convolutional sparse coding framework was recently developed in [38, 39], where the explicit dependencies between the sparse representation vectors $\boldsymbol{\gamma}_i$ (and therefore the special structure of the corresponding constraint $M\left(D'\right)\boldsymbol{\Gamma'} = 0$) were exploited quite extensively, resulting in efficient recovery algorithms and nontrivial theoretical guarantees. We refer the reader to [38, 39] for further details and examples.

4.4 Arbitrary dependency graphs

The examples considered in the previous sections are somewhat special. For the most general case, one can define an abstract graph $\mathscr{G}$ with some desirable properties, and subsequently look for a nontrivial realization D of the graph, so that in addition $\mathscr{R}_{\mathscr{G}} \neq \emptyset$. Let us therefore discuss each one of those steps, along with a specific example.

4.4.1 Defining an abstract $\mathscr{G}$ with desirable properties

In this context, we would want $\mathscr{G}$ to contain *sufficiently many different long cycles*, which would correspond to long signals and a rich resulting model $\mathscr{M}$. In contrast with the models from Subsection 4.2, one therefore should allow for some branching mechanism. An example of a possible $\mathscr{G}$ is given in Figure 5 on page 27. It differs only slightly from the example in Figure 3 on page 23. Notice that due to the structure of $\mathscr{G}$ there are many possible paths in $\mathscr{C}_{\mathscr{G}}(P)$. In fact, a direct search algorithm yields $\left|\mathscr{C}_{\mathscr{G}}(70)\right| = 37614$.

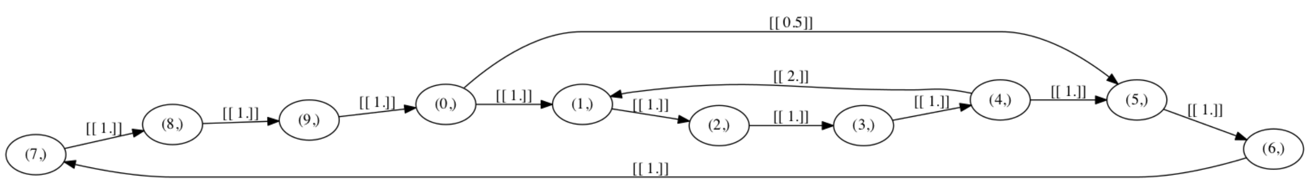


Fig. 5: A possible dependency graph $\mathscr{G}$ with $m = 10$. In this example, $\left|\mathscr{C}_{\mathscr{G}}(70)\right| = 37614$.

4.4.2 Finding a dictionary D which has $\mathscr{G}$ as its dependency graph

Every edge in $\mathscr{G}$ corresponds to a conditions of the form (9) imposed on the entries of D. As discussed in Theorem 4, this in turn translates to a set of algebraic equations. So the natural idea would be to write out the large system of such equations and look for a solution over the field $\mathbb{R}$ by well-known algorithms in numerical algebraic geometry [7]. However, this approach is highly impractical because these algorithms have (single or double) exponential running time. We consequently propose a simplified, more direct approach to the problem.

In detail, we replace the low-rank conditions (9) with more explicit and restrictive ones below.

Assumptions(*) For each $(s_i, s_j) \in \mathscr{G}$ we have $|s_i| = |s_j| = k$. We require that $\operatorname{span} S_B D_{s_i} = \operatorname{span} S_T D_{s_j} = \Lambda_{i,j}$ with $\dim \Lambda_{i,j} = k$. Thus there exists a non-singular transfer matrix $C_{i,j} \in \mathbb{R}^{k \times k}$ such that

$$S_B D_{s_i} = C_{i,j} S_T D_{s_j}. \tag{23}$$

In other words, every column in $S_B D_{s_i}$ must be a specific linear combination of the columns in $S_T D_{s_j}$. This is much more restrictive than the low-rank condition, but on the other hand, given the matrix $C_{i,j}$, it defines a set of linear constraints on D. To summarize, the final algorithm is presented in Algorithm 0.6. In general, nothing guarantees that for a particular choice of $\mathscr{G}$ and the transfer matrices, there is a nontrivial solution D, however in practice we do find such solutions. For example, taking the graph from Figure 5 on page 27 and augmenting it with the matrices $C_{i,j}$ (scalars in this case), we obtain a solution over $\mathbb{R}^6$ which is shown in Figure 6 on page 28. Notice that while the resulting dictionary has a Hankel-type structure similar to what we have seen previously, the additional dependencies between the atoms produce a rich signal space structure, as we shall demonstrate in the following section.

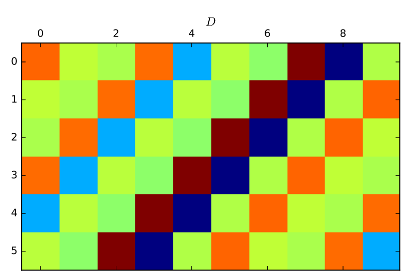


Fig. 6: A realization $D \in \mathbb{R}^{6\times 10}$ of $\mathscr{G}$ from Figure 5 on page 27.

Algorithm 0.6 Finding a realization D of the graph $\mathscr{G}$

1. Input: a graph $\mathscr{G}$ satisfying the **Assumptions(*)** above, and the dimension n of the realization space $\mathbb{R}^n$.
2. Augment the edges of $\mathscr{G}$ with arbitrary nonsingular transfer matrices $C_{i,j}$.
3. Construct the system of linear equations given by (23).
4. Find a nonzero D solving the system above over $\mathbb{R}^n$.

4.4.3 The resulting signal space

Given $\mathscr{G}_D$ and the signal length $N = P$, the signals x can be generated according to Algorithm 0.1 on page 10. Not all paths in $\mathscr{C}_\mathscr{G}$ are realizable, but it turns out that in our example we have $|\mathscr{R}_\mathscr{G}(70)| = 17160$. Three different signals and their supports $\mathscr{S}$ are shown in Figure 7 on page 30. As can be seen from these examples, the resulting model $\mathscr{M}$ is indeed much richer than the signature-type construction from Subsection 4.2.

Using the restricted construction of this section, the following estimate can be easily shown.

Proposition 11. *Assume that the model satisfies **Assumptions(*)** above. Then for every $\mathscr{S} \in \mathscr{R}_\mathscr{G}(P)$*

$$\dim\ker A_\mathscr{S} \leqslant k.$$

Proof. The idea is to construct a spanning set for $\ker M_*^{(\mathscr{S})}$ and invoke Proposition 2. Let us relabel the nodes along $\mathscr{S}$ to be $1, 2, \ldots, P$. Starting from an arbitrary α_1 with support $|s_1| = k$, we use (23) to obtain, for $i = 1, 2, \ldots, P-1$, a formula for the next portion of the global representation vector Γ

$$\alpha_{i+1} = C_{i,i+1}^{-1}\alpha_i. \tag{24}$$

This gives a set Δ consisting of overall k linearly independent vectors Γ_i with $\operatorname{supp}\Gamma_i = \mathscr{S}$. It may happen that equation (24) is not satisfied for $i = P$. However, every Γ with $\operatorname{supp}\Gamma = \mathscr{S}$ and $M_*^{(\mathscr{S})}\Gamma_\mathscr{S} = 0$ must belong to $\operatorname{span}\Delta$, and therefore

$$\dim\ker M_*^{(\mathscr{S})} \leqslant \dim\operatorname{span}\Delta = k.$$

□

We believe that Proposition 11 can be extended to more general graphs, not necessarily satisfying **Assumptions(*)**. In particular, the following estimate appears to hold for a general model $\mathscr{M}$ and $\mathscr{S} \in \mathscr{R}_\mathscr{G}(P)$:

$$\dim\ker A_\mathscr{S} \leqslant |s_1| + \sum_i \left(|s_{i+1}| - \operatorname{rank}\left[S_B D_{s_i}\ S_T D_{s_{i+1}}\right]\right).$$

We leave the rigorous proof of this result to a future work.

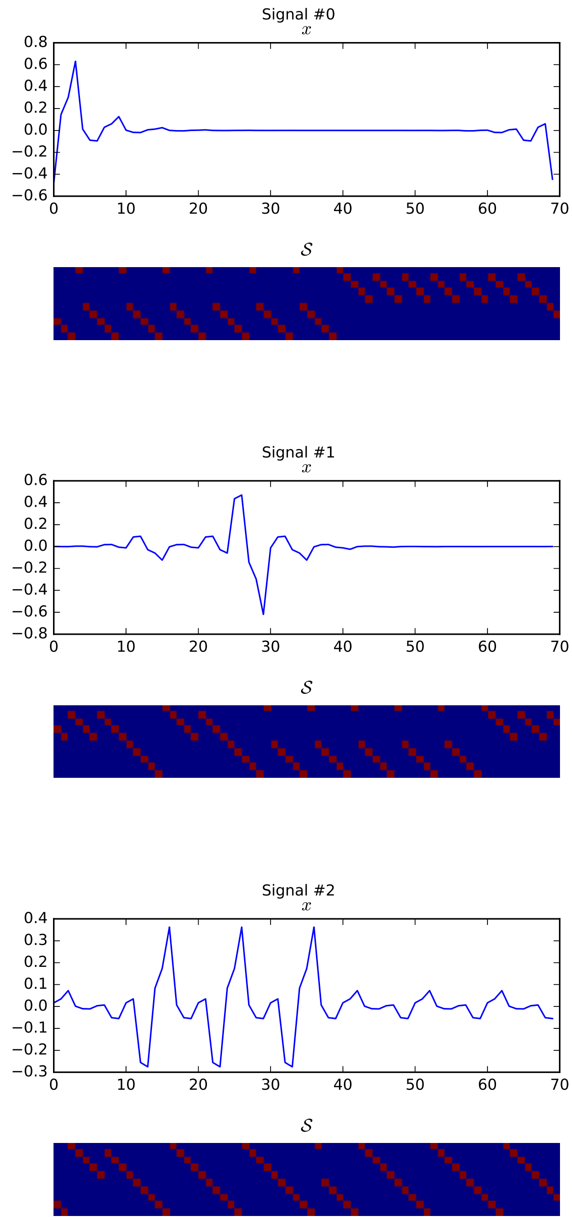


Fig. 7: Examples of signals from $\mathscr{M}$ and the corresponding supports $\mathscr{S}$

4.4.4 Further remarks

While the presented model is the hardest to analyze theoretically, even in the restricted case of **Assumptions(*)** (when does a nontrivial realization of a given $\mathscr{G}$ exist? how does the answer depend on n? When $\mathscr{R}_{\mathscr{G}}(P) \neq \emptyset$? etc?), we hope that this construction will be most useful in applications such as denoising of natural signals.

5 Numerical experiments

In this section, we test the effectiveness of the globalized model for recovering the signals from Section 4, both in the noiseless and noisy cases. These results are also compared to the classical LPA approach.

5.1 Signature-type signals

In this section we investigate the performance of the pursuit algorithms on signals complying with the signature dictionary model elaborated in Subsection 4.2, constructed from one base signal and allowing for varying values of s.

5.1.1 Constructing the dictionary

In the context of the LPA algorithm, the condition for its success in recovering the representation is a function of the mutual coherence of the local dictionary – the smaller this measure the larger the number of non-zeros that are guaranteed to be recovered. Leveraging this, we aim at constructing $D \in \mathbb{R}^{n \times m}$ of a signature type that has a small coherence. This can be cast as an optimization problem

$$x_0 = \arg\min_{x \in \mathbb{R}^m} \mu\left(\tilde{D}(x)\right), \; D = D(x_0),$$

where $\tilde{D}(x)$ is computed by Algorithm 0.4 and μ is the (normalized) coherence function.

In our experiments, we choose $n = 15$ and $m = 20$, and minimize the above loss function via gradient descent, resulting in $\mu(D(x)) = 0.26$. We used the TensorFlow open source package [3]. As a comparison, the coherence of a random signature dictionary is about 0.5.

5.1.2 Noiseless case

In this setting, we test the ability of the globalized OMP to perfectly recover the sparse representation of clean signature-type signals. Figure 8 compares the proposed algorithm (for different choices of $\beta \in \{0.25, 0.5, 1, 2, 5\}$) with the LPA one by providing their probability of success in recovering the true sparse vectors, averaged over 10^3 randomly generated signals of length $N = 100$.

From a theoretical perspective, since $\mu(D) = 0.26$, the LPA algorithm is guaranteed to recover the representation when $\|\Gamma\|_{0,\infty} \leq 2$, while as can be seen in practice it successfully recover these for $\|\Gamma\|_{0,\infty} \leq 3$. Comparing the LPA approach to the globalized OMP, one can observe that for $\beta \geq 1$ the latter consistently outperforms the former, having a perfect recovery for $\|\Gamma\|_{0,\infty} \leq 4$. Another interesting insight of this experiment is the effect of β on the performance; roughly speaking, a relatively large value of this parameter results in a better success-rate than the very small ones, thereby emphasizing importance of the constraint $M_* \Gamma = 0$. On the other hand, β should not be too large since then the importance of the signal is reduced compared to the constraint, which might lead to deterioration in the success-rate (see the curve that corresponds to$\beta = 5$ in Figure 8).

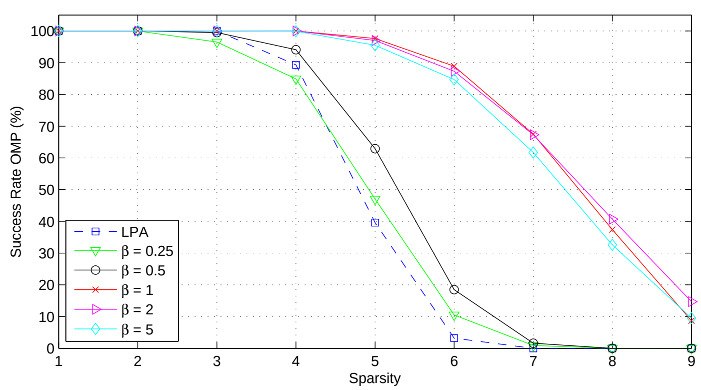


Fig. 8: Probability of the success (%) of the globalized OMP (as a function of β) and the LPA algorithms to perfectly recover the sparse representations of test signals from the signature dictionary model, averaged over 10^3 realizations.

5.1.3 Noisy case

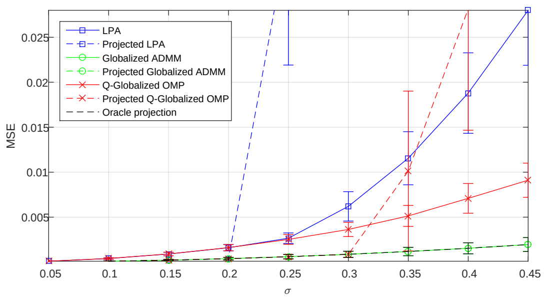
In what follows, the stability of the proposed globalized OMP and the ADMM-pursuit are tested and compared to the traditional LPA algorithm. In addition, we provide the projected versions of these algorithms, which by definition satisfy the constraint of our model. More specifically, given the estimated support $\hat{S}$ of each pursuit, we ensure that the constraint $M_*\hat{\Gamma} = 0$ is met by constructing the matrix $A_{\hat{S}}$ and then projecting the signal onto the subspace $\ker A_{\hat{\mathscr{S}}}$. In addition to the above, we provide the restoration performance of the oracle estimator, serving as an indication for the best possible denoising that can be achieved. In this case, the oracle projection matrix A_S is constructed according to the ground-truth support S.

Per each local cardinality $1 \leq \|\Gamma\|_{0,\infty} \leq 4$ we generate 10 random signature-type signals, where each of these is corrupted by white additive Gaussian noise with standard deviation σ, ranging from 0.05 up-to 0.5. The global number of non-zeros is injected to the globalized OMP, and the information regarding the local sparsity is utilized both by the LPA algorithm and our ADMM-pursuit (which is based on local sparse recovery operations). Following Figure 9, which plots the Mean Squared Error (MSE) of the estimation as a function of the noise level, the ADMM-pursuit achieves the best denoising performance, having similar results to the oracle estimator for all noise-levels and sparsity factors. The source of superiority of the ADMM pursuit might be its inherent ability to obtain an estimation that perfectly fits to the globalized model, i.e., a reconstruction that is identical to its projected version. The second best algorithm being the globalized OMP; for relatively small noise levels, its projected version performs as good as the oracle one, indicating that it successfully recovers the true supports. For large noise levels, however, this algorithm tends to err and results in local estimations that do not "agree" with each other on the overlaps. Yet, the denoising performance of the globalized OMP is better than the one of the LPA algorithm. Notice that the latter performs similarly to the oracle estimator only for very low noise levels and relatively small sparsity factors. This sheds light on the difficulty of finding the true supports, the non trivial solution of this problem, and the great advantage of the proposed globalized model.

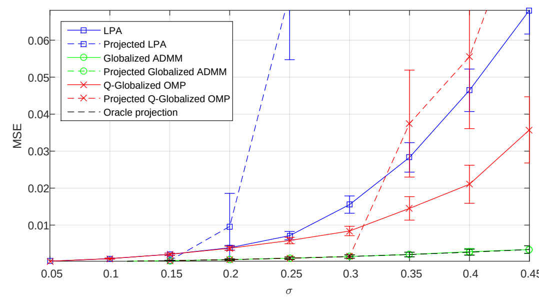
Similar conclusion holds for the stable recovery of the sparse representations. Per each pursuit algorithm, Figure 10 illustrates the ℓ_2 distance between the original sparse vector Γ and its estimation $\hat{\Gamma}$, averaged over the different noise realizations. As can be seen, the ADMM-pursuit achieves the most stable recovery, the globalized OMP is slightly behind it, and both of them outperform the LPA algorithm especially in the challenging cases of high noise levels and/or large sparsity factors.

5.2 Denoising PWC Signals

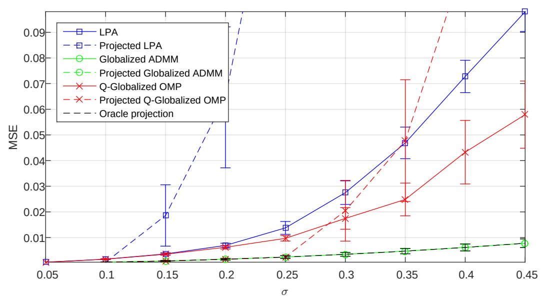
In this scenario, we test the ability of the globalized ADMM-pursuit to restore corrupted PWC signals, and compare these to the outcome of the LPA algorithm. Similarly to the previous subsection, the projected versions of the two pursuit algorithms are provided along with the one of the oracle estimator. Following the description in Section 4.1, we generate a signal of length $N = 200$, composed of patches of size $n = m = 20$ with a local sparsity of at most 2 non-zeros in the $\ell_{0,\infty}$-sense. These sig-



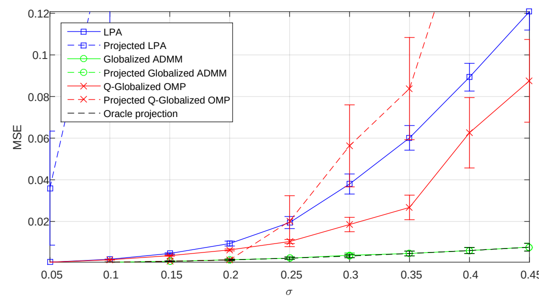
(a) $\|\Gamma\|_{0,\infty} \leq 1$



(b) $\|\Gamma\|_{0,\infty} \leq 2$



(c) $\|\Gamma\|_{0,\infty} \leq 3$

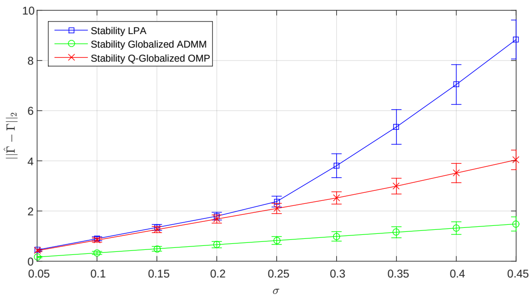


(d) $\|\Gamma\|_{0,\infty} \leq 4$

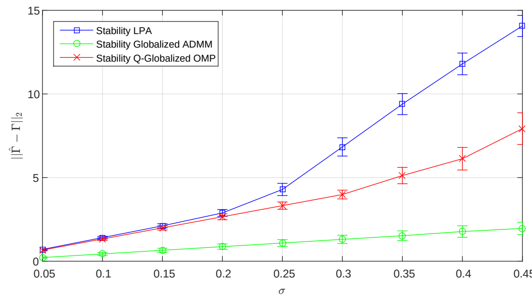
Fig. 9: Denoising performance of the globalized OMP, ADMM-pursuit and LPA algorithm along with their projected versions for various noise levels and sparsity factors. The projected version of the oracle estimator is provided as well, demonstrating the best possible restoration that can be achieved. The signals are drawn from the signature dictionary model.

nals are then contaminated by a white additive Gaussian noise with σ in the range of 0.1 to 0.9.

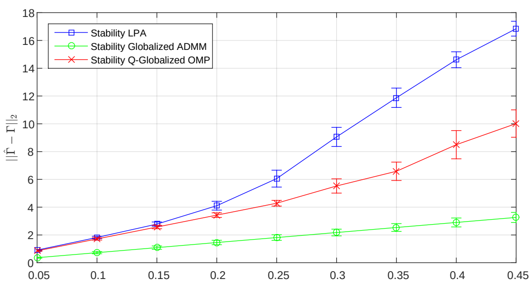
The restoration performance (in terms of MSE) of the above-mentioned algorithms is illustrated in Figure 11 and the stability of the estimates is demonstrated in Figure 12, where the results are averaged over 10 noise realizations. As can be seen, the globalized approach significantly outperforms the LPA algorithm for all noise levels. Furthermore, when $\sigma \leq 0.5$, the ADMM-pursuit performs similarly to the oracle estimator. One can also notice that, similarly to the previous subsection, the ADMM-pursuit and its projected version resulting in the very same estimation, i.e. this algorithm forces the signal to conform with the patch-sparse model globally. On the other hand, following the visual illustration given in Figure 13, the projected version of the LPA algorithm has zero segments, which are the consequence of a complete disagreement in the support (local inconsistency). This is also reflected in



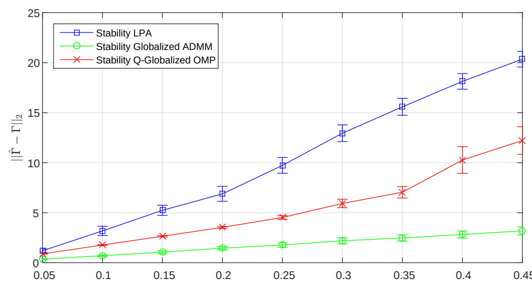
(a) $\|\Gamma\|_{0,\infty} \leq 1$



(b) $\|\Gamma\|_{0,\infty} \leq 2$



(c) $\|\Gamma\|_{0,\infty} \leq 3$



(d) $\|\Gamma\|_{0,\infty} \leq 4$

Fig. 10: Stability of the globalized OMP, ADMM-pursuit and LPA algorithm for various noise levels and sparsity factors. The signals are drawn from the signature dictionary model.

Figure 11, illustrating that even for a very small noise level ($\sigma = 0.1$) the projected version of the LPA algorithm has a very large estimation error (MSE ≈ 0.18) compared to the one of the ADMM-pursuit (MSE ≈ 0.0004), indicating that the former fails in obtaining a consistent representation of the signal.

6 Discussion

In this work we have presented an extension of the classical theory of sparse representations to signals which are locally sparse, together with novel pursuit algorithms. We envision several promising research directions which might emerge from this work.

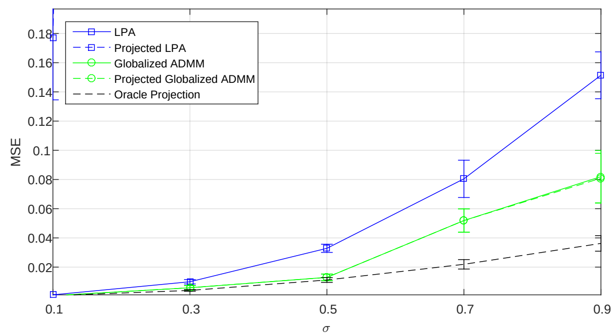


Fig. 11: Denoising performance of the ADMM-pursuit and the LPA algorithm for various noise levels, tested for signals from the piecewise-constant model with $\|\Gamma\|_{0,\infty} \leq 2$.

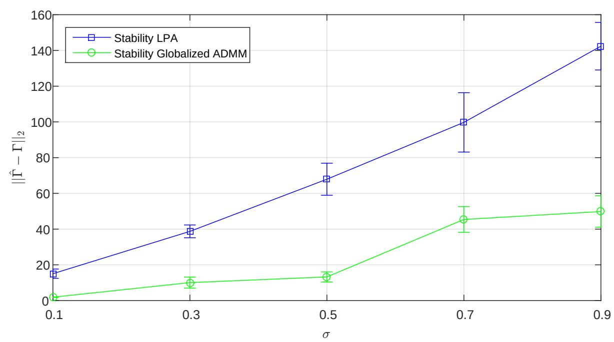


Fig. 12: Stability of the ADMM-pursuit and the LPA algorithm for various noise levels, tested for signals from the piecewise-constant model with $\|\Gamma\|_{0,\infty} \leq 2$.

6.1 Relation to other models

Viewed globally, the resulting signal model can be considered a sort of "structured sparse" model, however, in contrast to other such constructions ([50, 29, 28, 31] and others), our model incorporates both structure in the representation coefficients and a structured dictionary.

The recently developed framework of Convolutional Sparse Coding (CSC) [38, 39, 37] bears some similarities to our work, in that it, too, has a convolutional representation of the signal via a dictionary identical in structure to D_G. However, the underlying local sparsity assumptions are drastically different in the two models, resulting in very different guarantees and algorithms. That said, we believe that it would be important to provide precise connections between the results, possibly

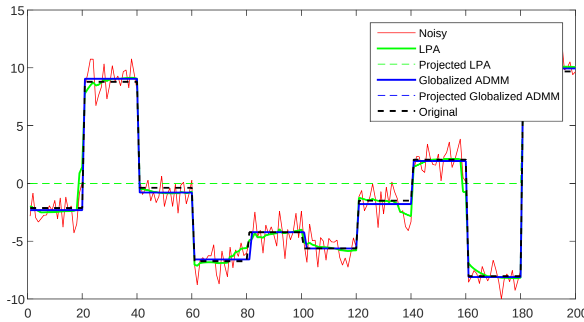


Fig. 13: Denoising of a PWC signal contaminated with additive Gaussian noise ($\sigma = 1.1$) via several pursuit algorithms: Input noisy signal (MSE = 1.173), LPA algorithm (MSE = 0.328), projected LPA (MSE = 24.672), ADMM-pursuit (MSE = 0.086), projected ADMM-pursuit (MSE = 0.086), and oracle estimator (MSE = 0.047).

leading to their deeper understanding. First steps in this direction are outlined in Subsection 4.3.

6.2 Further extensions

The decomposition of the global signal $x \in \mathbb{R}^N$ into its patches:

$$x \mapsto (R_i x)_{i=1}^P, \tag{25}$$

is a special case of a more general decomposition, namely

$$x \mapsto (w_i \mathscr{P}_i x)_{i=1}^P, \tag{26}$$

where $\mathscr{P}_i$ is the (orthogonal) projection onto a subspace W_i of $\mathbb{R}^N$, and w_i are some weights. This observation naturally places our theory, at least partially, into the framework of *fusion frames*, a topic which is generating much interest recently in the applied harmonic analysis community [1, Chapter 13]. In fusion frame theory, which is motivated by applications such as distributed sensor networks, the starting point is precisely the decomposition (26). Instead of the reconstruction formula $x = \sum_i \frac{1}{n} R_i^T R_i x$, in fusion frame theory we have

$$x = \sum_i w_i^2 S_{\mathscr{W}}^{-1} (\mathscr{P}_i x),$$

where $S_{\mathscr{W}}$ is the associated *fusion frame operator*. The natural extension of our work to this setting would seek to enforce some sparsity of the projections $\mathscr{P}_i x$. Perhaps the most immediate variant of (25) in this respect would be to drop the periodicity requirement, resulting in a slightly modified R_i operators near the endpoints of the signal. We would like to mention some recent works which investigate different notions of fusion frame sparsity [10, 4, 6].

Another intriguing possible extension of our work is to relax the complete overlap requirement between patches and consider an "approximate patch sparsity" model, where the patch disagreement vector $M\Gamma$ is not zero but "small". In some sense, one can imagine a full "spectrum" of such models, ranging from a complete agreement (this work) to an arbitrary disagreement (such as in the CSC framework mentioned above).

6.3 Learning models from data

The last point above brings us to the question of how to obtain "good" models, reflecting the structure of the signals at hand (such as speech/images etc.) We hope that one might use the ideas presented here in order to create novel learning algorithms. In this regard, the main difficulty is how to parametrize the space of allowed models in an efficient way. While we presented some initial ideas in Section 4, in the most general case (incorporating the approximate sparsity direction above) the problem remains widely open.

Acknowledgements The research leading to these results has received funding from the European Research Council under European Union's Seventh Framework Programme, ERC Grant agreement no. 320649. The authors would also like to thank Jeremias Sulam, Vardan Papyan, Raja Giryes and Gitta Kutinyok for inspiring discussions.

Appendix A: Proof of Lemma 1

Proof. Denote $Z := \ker M$ and consider the linear map $A : Z \to \mathbb{R}^N$ given by the restriction of the "averaging map" $D_G : \mathbb{R}^{mP} \to \mathbb{R}^N$ to Z.

1. Let us see first that $im(A) = \mathbb{R}^N$. Indeed, for every $x \in \mathbb{R}^N$ consider its patches $x_i = R_i x$. Since D is full rank, there exist $\{\alpha_i\}$ for which $D\alpha_i = x_i$. Then setting $\Gamma := (\alpha_1, \dots, \alpha_P)$ we have both $D_G \Gamma = x$ and $M\Gamma = 0$ (by construction, see Section 2), i.e. $\Gamma \in Z$ and the claim follows.
2. Define
$$J := \ker D \times \ker D \times \dots \ker D \subset \mathbb{R}^{mP}.$$
We claim that $J = \ker A$.

a. In one direction, let $\Gamma = (\alpha_1, \ldots, \alpha_P) \in \ker A$, i.e. $M\Gamma = 0$ and $D_G\Gamma = 0$. Immediately we see that $\frac{1}{n}D\alpha_i = 0$ for all i, and therefore $\alpha_i \in \ker D$ for all i, thus $\Gamma \in J$.

b. In the other direction, let $\Gamma = (\alpha_1, \ldots, \alpha_P) \in J$, i.e. $D\alpha_i = 0$. Then the local representations agree, i.e. $M\Gamma = 0$, thus $\Gamma \in Z$. Furthermore, $D_G\Gamma = 0$ and therefore $\Gamma \in \ker A$.

3. By the fundamental theorem of linear algebra we conclude

$$\begin{aligned}\dim Z &= \dim im\,(A) + \dim\ker A = N + \dim J\\ &= N + (m-n)N = N(m-n+1).\end{aligned}$$

□

Appendix B: Proof of Lemma 2

We start with an easy observation.

Proposition 12. *For any vector $\rho \in \mathbb{R}^N$, we have*

$$\|\rho\|_2^2 = \frac{1}{n}\sum_{j=1}^{N}\|R_j\rho\|_2^2.$$

Proof. Since

$$\|\rho\|_2^2 = \sum_{j=1}^{N}\rho_j^2 = \frac{1}{n}\sum_{j=1}^{N}n\rho_j^2 = \frac{1}{n}\sum_{j=1}^{N}\sum_{k=1}^{n}\rho_j^2,$$

we can rearrange the sum and get

$$\begin{aligned}\|\rho\|_2^2 &= \frac{1}{n}\sum_{k=1}^{n}\sum_{j=1}^{N}\rho_j^2 = \frac{1}{n}\sum_{k=1}^{n}\sum_{j=1}^{N}\rho_{(j+k) \mod N}^2 = \frac{1}{n}\sum_{j=1}^{N}\sum_{k=1}^{n}\rho_{(j+k) \mod N}^2\\ &= \frac{1}{n}\sum_{j=1}^{N}\|R_j\rho\|_2^2.\end{aligned}$$

□

Corollary 4. *Given $M\Gamma = 0$, we have*

$$\|y - D_G\Gamma\|_2^2 = \frac{1}{n}\sum_{j=1}^{N}\|R_jy - D\alpha_j\|_2^2.$$

Proof. Using Proposition 12, we get

$$\|y - D_G \Gamma\|_2^2 = \frac{1}{n}\sum_{j=1}^{N} \|R_j y - R_j D_G \Gamma\|_2^2 = \frac{1}{n}\sum_{j=1}^{N} \|R_j y - \Omega_j \Gamma\|_2^2.$$

Now since $M\Gamma = 0$, then by definition of M we have $\Omega_j \Gamma = D\alpha_j$ (see (7)), and this completes the proof. □

Recall Definition 10 and (2). Multiplying the corresponding matrices gives

Proposition 13. *We have the following equality for all $i = 1, \dots P$:*

$$S_B R_i = S_T R_{i+1}. \tag{27}$$

To facilitate the proof, we introduce extension of Definition 10 to multiple shifts as follows.

Definition 21. Let n be fixed. For $k = 0, \dots, n-1$ let

1. $S_T^{(k)} := \begin{bmatrix} I_{n-k} & \mathbf{0} \end{bmatrix}$ and $S_B^{(k)} := \begin{bmatrix} \mathbf{0} & I_{n-k} \end{bmatrix}$ denote the operators extracting the top (resp. bottom) $n-k$ entries from a vector of length n; the matrices have dimension $(n-k) \times n$.
2. $Z_B^{(k)} := \begin{bmatrix} S_B^{(k)} \\ \mathbf{0}_{k \times n} \end{bmatrix}$ and $Z_T^{(k)} := \begin{bmatrix} \mathbf{0}_{k \times n} \\ S_T^{(k)} \end{bmatrix}$.
3. $W_B^{(k)} := \begin{bmatrix} \mathbf{0}_{k \times n} \\ S_B^{(k)} \end{bmatrix}$ and $W_T^{(k)} := \begin{bmatrix} S_T^{(k)} \\ \mathbf{0}_{k \times n} \end{bmatrix}$.

Note that $S_B = S_B^{(1)}$ and $S_T = S_T^{(1)}$. We have several useful consequences of the above definitions. The proofs are carried out via elementary matrix identities and are left to the reader.

Proposition 14. *For any $n \in \mathbb{N}$ the following hold:*

1. $Z_T^{(k)} = \left(Z_T^{(1)}\right)^k$ *and* $Z_B^{(k)} = \left(Z_B^{(1)}\right)^k$ *for* $k = 0, \dots, n-1$*;*
2. $W_T^{(k)} W_T^{(k)} = W_T^{(k)}$ *and* $W_B^{(k)} W_B^{(k)} = W_B^{(k)}$ *for* $k = 0, \dots, n-1$*;*
3. $W_T^{(k)} W_B^{(j)} = W_B^{(j)} W_T^{(k)}$ *for* $j, k = 0, \dots, n-1$*;*
4. $Z_B^{(k)} = Z_B^{(k)} W_B^{(k)}$ *and* $Z_T^{(k)} = Z_T^{(k)} W_T^{(k)}$ *for* $k = 0, \dots, n-1$*;*
5. $W_B^{(k)} = Z_T^{(1)} W_B^{(k-1)} Z_B^{(1)}$ *and* $W_T^{(k)} = Z_B^{(1)} W_T^{(k-1)} Z_T$ *for* $k = 1, \dots, n-1$*;*
6. $Z_B^{(k)} Z_T^{(k)} = W_T^{(k)}$ *and* $Z_T^{(k)} Z_B^{(k)} = W_B^{(k)}$ *for* $k = 0, \dots, n-1$*;*
7. $(n-1) I_{n \times n} = \sum_{k=1}^{n-1} \left(W_B^{(k)} + W_T^{(k)}\right)$.

Proposition 15. *If the vectors $u_1,\dots,u_N \in \mathbb{R}^n$ satisfy pairwise*

$$S_B u_i = S_T u_{i+1},$$

then they also satisfy for each $k=0,\dots,n-1$ the following:

$$W_B^{(k)} u_i = Z_T^{(k)} u_{i+k}, \tag{28}$$

$$Z_B^{(k)} u_i = W_T^{(k)} u_{i+k}. \tag{29}$$

Proof. It is easy to see that the condition $S_B u_i = S_T u_{i+1}$ directly implies

$$Z_B^{(1)} u_i = W_T^{(1)} u_{i+1}, \quad W_B^{(1)} u_i = Z_T^{(1)} u_{i+1} \quad \forall i. \tag{30}$$

Let us first prove (28) by induction on k. The base case $k=1$ is precisely (30). Assuming validity for $k-1$ and $\forall i$, we have

$$\begin{aligned}
W_B^{(k)} u_i &= Z_T^{(1)} W_B^{(k-1)} Z_B^{(1)} u_i && \text{(by Proposition 14, item 5)} \\
&= Z_T^{(1)} W_B^{(k-1)} W_T^{(1)} u_{i+1} && \text{(by (30))} \\
&= Z_T^{(1)} W_T^{(1)} W_B^{(k-1)} u_{i+1} && \text{(by Proposition 14, item 3)} \\
&= Z_T^{(1)} W_T^{(1)} Z_T^{(k-1)} u_{i+k} && \text{(by the induction hypothesis)} \\
&= Z_T^{(1)} Z_T^{(k-1)} u_{i+k} && \text{(by Proposition 14, item 4)} \\
&= Z_T^{(k)} u_{i+k}. && \text{(by Proposition 14, item 1)}
\end{aligned}$$

To prove (29) we proceed as follows:

$$\begin{aligned}
Z_B^{(k)} u_i &= Z_B^{(k)} W_B^{(k)} u_i && \text{(by Proposition 14, item 4)} \\
&= Z_B^{(k)} Z_T^{(k)} u_{i+k} && \text{(by (28) which is already proved)} \\
&= W_T^{(k)} u_{i+k}. && \text{(by Proposition 14, item 6)}
\end{aligned}$$

This finishes the proof of Proposition 15.

□

Example 1. To help the reader understand the claim of Proposition 15, consider the case $k=2$, and take some three vectors u_i, u_{i+1}, u_{i+2}. We have $S_B u_i = S_T u_{i+1}$ and also $S_B u_{i+1} = S_T u_{i+2}$. Then clearly $S_B^{(2)} u_i = S_T^{(2)} u_{i+2}$ (see Figure 14 on page 42) and therefore $W_B^{(2)} u_i = Z_T^{(2)} u_{i+2}$.

Let us now present the proof of Lemma 2.

Proof. We show equivalence in two directions.

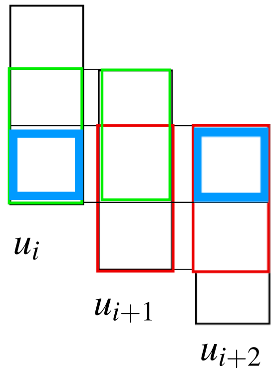


Fig. 14: Illustration to the proof of Proposition 15. The green pair is equal, as well as the red pair. It follows that the blue elements are equal as well.

- (1) $\Longrightarrow$ (2): Let $M\Gamma = 0$. Define $x := D_G\Gamma$, and then further denote $x_i := R_i x$. Then on the one hand:

$$\begin{aligned} x_i &= R_i D_G \Gamma \\ &= \Omega_i \Gamma && \text{(definition of } \Omega_i) \\ &= D\alpha_i. && (M\Gamma = 0) \end{aligned}$$

On the other hand, because of (27) we have $S_B R_i x = S_T R_{i+1} x$, and by combining the two we conclude that $S_B D\alpha_i = S_T D\alpha_{i+1}$.

- (2) $\Longrightarrow$ (1): In the other direction, suppose that $S_B D\alpha_i = S_T D\alpha_{i+1}$. Denote $u_i := D\alpha_i$. Now consider the product $\Omega_i\Gamma$ where $\Omega_i = R_i D_G$. One can easily be convinced that in fact

$$\Omega_i \Gamma = \frac{1}{n}\left(\sum_{k=1}^{n-1}\left(Z_B^{(k)} u_{i-k} + Z_T^{(k)} u_{i+k}\right) + u_i\right).$$

Therefore

$$\begin{aligned} (\Omega_i - Q_i)\Gamma &= \frac{1}{n}\left(u_i + \sum_{k=1}^{n-1}\left(Z_B^{(k)} u_{i-k} + Z_T^{(k)} u_{i+k}\right)\right) - u_i \\ &= \frac{1}{n}\left(\sum_{k=1}^{n-1}\left(W_T^{(k)} u_i + W_B^{k} u_i\right) - (n-1)\, u_i\right) && \text{(by Proposition 15)} \\ &= 0. && \text{(by Proposition 14, item 7)} \end{aligned}$$

Since this holds for all i, we have shown that $M\Gamma = 0$.

□

Appendix C: Proof of Theorem 8

Recall that $M_A = \frac{1}{n}\sum_i R_i^T P_{s_i} R_i$. We first show that M_A is a contraction.

Proposition 16. $\|M_A\|_2 \leqslant 1$.

Proof. Closely following a similar proof in [42], divide the index set $\{1,\dots,N\}$ into n groups representing *non-overlapping* patches: for $i=1,\dots,n$ let

$$K(i) := \left\{ i, i+n, \dots, i+\left(\left\lfloor \frac{N}{n} \right\rfloor - 1\right) n \right\} \mod N.$$

Now

$$\begin{aligned}
\|M_A x\|_2 &= \frac{1}{n}\left\| \sum_{i=1}^{N} R_i^T P_{s_i} R_i x \right\|_2 \\
&= \frac{1}{n}\left\| \sum_{i=1}^{n} \sum_{j\in K(i)} R_j^T P_{s_j} R_j x \right\|_2 \\
&\leqslant \frac{1}{n}\sum_{i=1}^{n}\left\| \sum_{j\in K(i)} R_j^T P_j R_j x \right\|_2 .
\end{aligned}$$

By construction, $R_j R_k^T = \mathbf{0}_{n\times n}$ for $j,k \in K(i)$ and $j \neq k$. Therefore for all $i=1,\dots,n$ we have

$$\begin{aligned}
\left\| \sum_{j\in K(i)} R_j^T P_{s_j} R_j x \right\|_2^2 &= \sum_{j\in K(i)} \left\| R_j^T P_{s_j} R_j x \right\|_2^2 \\
&\leqslant \sum_{j\in K(i)} \left\| R_j x \right\|_2^2 \leqslant \|x\|_2^2 .
\end{aligned}$$

Substituting in back into the preceding inequality finally gives

$$\|M_A x\|_2 \leqslant \frac{1}{n}\sum_{i=1}^{n} \|x\|_2 = \|x\|_2 .$$

□

Now let us move on to prove Theorem 8.

Proof. Define

$$\hat{P}_i := (I - P_{s_i}) R_i .$$

It is easy to see that

$$\sum_i \hat{P}_i^T \hat{P}_i = A_{\mathscr{S}}^T A_{\mathscr{S}} .$$

Let the SVD of $A_{\mathscr{S}}$ be

$$A_{\mathscr{S}} = U\Sigma V^T.$$

Now

$$\begin{aligned} V\Sigma^2 V^T = A_{\mathscr{S}}^T A_{\mathscr{S}} = \sum_i \hat{P}_i^T \hat{P}_i &= \sum_i R_i^T R_i - \underbrace{\sum_i R_i^T P_{s_i} R_i}_{:=T} \\ &= nI - T. \end{aligned}$$

Therefore $T = nI - V\Sigma V^T$, and

$$M_A = \frac{1}{n}T = I - \frac{1}{n}V\Sigma^2 V^T = V\left(I - \frac{\Sigma^2}{n}\right)V^T.$$

This shows that the eigenvalues of M_A are $\tau_i = 1 - \frac{\sigma_i^2}{n}$ where $\{\sigma_i\}$ are the singular values of $A_{\mathscr{S}}$. Thus we obtain

$$M_A^k = V \operatorname{diag}\left\{\tau_i^k\right\} V^T.$$

If $\sigma_i = 0$ then $\tau_i = 1$, and in any case, by Proposition 16 we have $|\tau_i| \leqslant 1$. Let the columns of the matrix W consist of the singular vectors of $A_{\mathscr{S}}$ corresponding to $\sigma_i = 0$ (and so $\operatorname{span} W = \mathscr{N}(A_{\mathscr{S}})$), then

$$\lim_{k\to\infty} M_A^k = WW^T.$$

Thus, as $k \to \infty$, M_A^k tends to the orthogonal projector onto $\mathscr{N}(A_{\mathscr{S}})$. □

Appendix D: Proof of Theorem 10

Recall that the signal consists of s constant segments of corresponding lengths $\ell_1, \dots, \ell_s$. We would like to compute the MSE for every pixel within every such segment of length $\alpha := \ell_r$. For each patch, the oracle provides the locations of the jump points within the patch.

Let us calculate the MSE for pixel with index 0 inside a constant (**nonzero**) segment $[-k, \alpha - k - 1]$ with value v (Figure 15 on page 45 might be useful). The oracle estimator has the explicit formula

$$\hat{x}_A^{r,k} = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{b_j - a_j + 1}\sum_{i=a_j}^{b_j}(v + z_i), \tag{31}$$

where $j = 1, \dots, n$ corresponds to the index of the overlapping patch containing the pixel, intersecting the constant segment on $[a_j, b_j]$, so that

$$a_j = -\min(k, n-j),$$
$$b_j = \min(\alpha - k - 1, j-1).$$

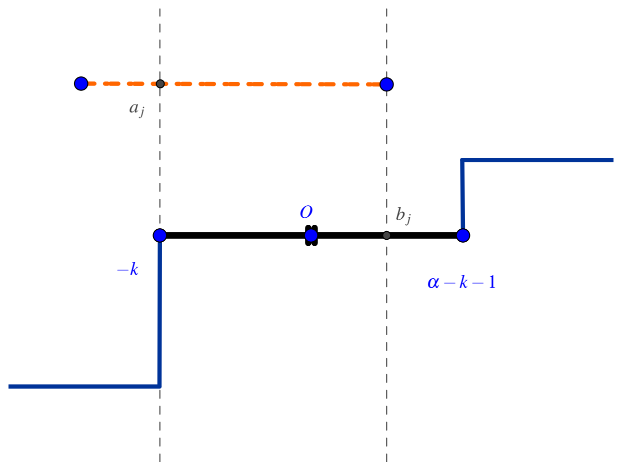


Fig. 15: The oracle estimator for the pixel O in the segment (black). The orange line is patch number $j = 1, \ldots, n$, and the relevant pixels are between a_j and b_j. The signal itself is shown to extend beyond the segment (blue line).

Now, the oracle error for the pixel is

$$\hat{x}_A^{r,k} - v = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{b_j - a_j + 1} \sum_{i=a_j}^{b_j} z_i$$
$$= \sum_{i=-k}^{\alpha-k-1} c_{i,\alpha,n,k} z_i,$$

where the coefficients $c_{i,\alpha,n,k}$ are some *positive* rational numbers depending only on i, α, n and k. It is easy to check by rearranging the above expression that

$$\sum_{i=-k}^{\alpha-k-1} c_{i,\alpha,n,k} = 1, \tag{32}$$

and furthermore, denoting $d_i := c_{i,\alpha,n,k}$ for fixed α, n, k, we also have that

$$d_{-k} < d_{-k+1} < \ldots d_0 > d_1 > \ldots d_{\alpha-k-1}. \tag{33}$$

Example 2. $n = 4,\ \alpha = 3$

- For $k = 1$:

$$\begin{aligned}\hat{x}_A^{r,k}-v &= \frac{1}{4}\left(\frac{1}{2}+\frac{1}{3}+\frac{1}{3}+\frac{1}{2}\right)z_0+\frac{1}{4}\left(\frac{1}{2}+\frac{1}{3}+\frac{1}{3}\right)z_{-1}+\frac{1}{4}\left(\frac{1}{3}+\frac{1}{3}+\frac{1}{2}\right)z_1\\ &= \underbrace{\frac{7}{24}}_{d_{-1}}z_{-1}+\underbrace{\frac{5}{12}}_{d_0}z_0+\underbrace{\frac{7}{24}}_{d_1}z_1\end{aligned}$$

- For $k=2$:

$$\begin{aligned}\hat{x}_A^{r,k}-v &= \frac{1}{4}\left(\frac{1}{3}+\frac{1}{3}+\frac{1}{2}+1\right)z_0+\frac{1}{4}\left(\frac{1}{3}+\frac{1}{3}+\frac{1}{2}\right)z_{-1}+\frac{1}{4}\left(\frac{1}{3}+\frac{1}{3}\right)z_{-2}\\ &= \frac{13}{24}z_0+\frac{7}{24}z_{-1}+\frac{1}{6}z_{-2}\end{aligned}$$

Now consider the optimization problem

$$\min_{c\in\mathbb{R}^{\alpha}} c^T c \quad \text{s.t } \mathbf{1}^T c = 1.$$

It can be easily verified that it has the optimal value $\frac{1}{\alpha}$, attained at $c^*=\alpha\mathbf{1}$. From this, (32) and (33) it follows that

$$\sum_{i=-k}^{\alpha-k-1} c_{i,\alpha,n,k}^2 > \frac{1}{\alpha}.$$

Since the z_i are i.i.d., we have

$$\mathbb{E}\left(\hat{x}_A^{r,k}-v\right)^2=\sigma^2\sum_{i=-k}^{\alpha-k-1}c_{i,\alpha,n,k}^2,$$

while for the entire nonzero segment of length $\alpha=\ell_r$

$$E_r:=\mathbb{E}\left(\sum_{k=0}^{\alpha-1}\left(\hat{x}_A^{r,k}-v\right)^2\right)=\sum_{k=0}^{\alpha-1}\mathbb{E}\left(\hat{x}_A^{r,k}-v\right)^2=\sigma^2\sum_{k=0}^{\alpha-1}\sum_{i=-k}^{\alpha-k-1}c_{i,\alpha,n,k}^2.$$

Defining

$$R(n,\alpha):=\sum_{k=0}^{\alpha-1}\sum_{i=-k}^{\alpha-k-1}c_{i,\alpha,n,k}^2,$$

we obtain that $R(n,\alpha)>1$ and furthermore

$$\mathbb{E}\|\hat{x}_A-x\|^2=\sum_{r=1}^{s}E_r=\sigma^2\sum_{r=1}^{s}R(n,\ell_r)>s\sigma^2.$$

This proves item (1) of Theorem 10. For showing the explicit formulas for $R(n,\alpha)$ in item (2), we have used automatic symbolic simplification software MAPLE [2].

By construction (31), it is not difficult to see that if $n\geqslant\alpha$ then

$$R(n,\alpha) = \frac{1}{n^2}\sum_{k=0}^{\alpha-1}\Big(\sum_{j=0}^{k}\big(2H_{\alpha-1} - H_k + \frac{n-\alpha+1}{\alpha} - H_{\alpha-1-j}\big)^2 + \sum_{j=k+1}^{\alpha-1}\big(2H_{\alpha-1} - H_{\alpha-k-1} + \frac{n-\alpha+1}{\alpha} - H_j\big)^2\Big),$$

where $H_k := \sum_{i=1}^{k}\frac{1}{i}$ is the k-th harmonic number. This simplifies to

$$R(n,\alpha) = 1 + \frac{\alpha(2\alpha H_\alpha^{(2)} + 2 - 3\alpha) - 1}{n^2},$$

where $H_k^{(2)} = \sum_{i=1}^{k}\frac{1}{i^2}$ is the k-th harmonic number of the second kind.
On the other hand, for $n \leqslant \frac{\alpha}{2}$ we have

$$R(n,\alpha) = \sum_{k=0}^{n-2} c_{n,k}^{(1)} + \sum_{k=n-1}^{\alpha-n} c_{n,k}^{(2)} + \sum_{k=\alpha-n+1}^{\alpha-1} c_{n,\alpha-1-k}^{(1)},$$

where

$$c_{n,k}^{(1)} = \frac{1}{n^2}\left(\sum_{j=k}^{n-1}(H_{n-1} - H_j + \frac{k+1}{n})^2 + \sum_{i=n-k}^{n-1}(\frac{n-i}{n})^2 + \sum_{i=0}^{k-1}(H_{n-1} - H_k + \frac{k-i}{n})^2\right)$$

and

$$c_{n,k}^{(2)} = \frac{1}{n^2}\left(\sum_{j=k-n+1}^{k}\left(\frac{j-k+n}{n}\right)^2 + \sum_{j=k+1}^{k+n-1}\left(\frac{k+n-j}{n}\right)^2\right).$$

Automatic symbolic simplification of the above gives

$$R(n,\alpha) = \frac{11}{18} + \frac{2\alpha}{3n} - \frac{5}{18n^2} + \frac{\alpha-1}{3n^3}.$$

References

1. Finite Frames - Theory and Applications. URL http://www.springer.com/birkhauser/mathematics/book/978-0-8176-8372-6 37
2. Maplesoft, a division of Waterloo Maple Inc. http://www.maplesoft.com 46
3. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015). URL http://tensorflow.org/. Software available from tensorflow.org 31
4. Aceska, R., Bouchot, J.L., Li, S.: Local sparsity and recovery of fusion frames structured signals. preprint (2015). URL http://www.mathc.rwth-aachen.de/~bouchot/

files/pubs/FusionCSfinal.pdf 38
5. Aharon, M., Elad, M.: Sparse and redundant modeling of image content using an image-signature-dictionary. SIAM Journal on Imaging Sciences **1**(3), 228–247 (2008) 21
6. Ayaz, U., Dirksen, S., Rauhut, H.: Uniform recovery of fusion frame structured sparse signals. Applied and Computational Harmonic Analysis **41**(2), 341–361 (2016). DOI 10.1016/j.acha.2016.03.006. URL http://www.sciencedirect.com/science/article/pii/S1063520316000294 38
7. Basu, S., Pollack, R., Roy, M.F.: Algorithms in real algebraic geometry, 2nd ed edn. No. v. 10 in Algorithms and computation in mathematics. Springer, Berlin ; New York (2006) 28
8. Blumensath, T., Davies, M.: Sparse and shift-invariant representations of music. Audio, Speech, and Language Processing, IEEE Transactions on **14**(1), 50–57 (2006). URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1561263 26
9. Blumensath, T., Davies, M.E.: Sampling theorems for signals from the union of finite-dimensional linear subspaces. Information Theory, IEEE Transactions on **55**(4), 1872–1882 (2009). URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4802322 7
10. Boufounos, P., Kutyniok, G., Rauhut, H.: Sparse Recovery From Combined Fusion Frame Measurements. IEEE Transactions on Information Theory **57**(6), 3864–3876 (2011). DOI 10.1109/TIT.2011.2143890 38
11. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2011). DOI 10.1561/2200000016. URL http://dx.doi.org/10.1561/2200000016 17
12. Bristow, H., Eriksson, A., Lucey, S.: Fast convolutional sparse coding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 391–398 (2013) 2
13. Bruckstein, A.M., Donoho, D.L., Elad, M.: From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM review **51**(1), 34–81 (2009). URL http://epubs.siam.org/doi/abs/10.1137/060657704 2
14. Candes, E.J.: Modern statistical estimation via oracle inequalities. Acta numerica **15**, 257–325 (2006). URL http://journals.cambridge.org/abstract_S0962492906230010 15
15. Chen, S., Billings, S.A., Luo, W.: Orthogonal least squares methods and their application to non-linear system identification. International Journal of control **50**(5), 1873–1896 (1989) 16
16. Dong, W., Zhang, L., Shi, G., Li, X.: Nonlocally centralized sparse representation for image restoration. IEEE Transactions on Image Processing **22**(4), 1620–1630 (2013) 2
17. Donoho, D.L., Elad, M.: Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization. Proceedings of the National Academy of Sciences **100**(5), 2197–2202 (2003). DOI 10.1073/pnas.0437847100. URL http://www.pnas.org/content/100/5/2197 12
18. Ekanadham, C., Tranchina, D., Simoncelli, E.P.: A unified framework and method for automatic neural spike identification. Journal of Neuroscience Methods **222**, 47–55 (2014). DOI 10.1016/j.jneumeth.2013.10.001. URL http://www.sciencedirect.com/science/article/pii/S0165027013003415 26
19. Elad, M.: Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing. Springer Science & Business Media (2010) 2
20. Elad, M., Aharon, M.: Image denoising via sparse and redundant representations over learned dictionaries. IEEE Transactions on Image processing **15**(12), 3736–3745 (2006) 2
21. Eldar, Y.C., Mishali, M.: Block sparsity and sampling over a union of subspaces. In: Digital Signal Processing, 2009 16th International Conference on, pp. 1–8. IEEE (2009). URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5201211 2, 7
22. Eldar, Y.C., Mishali, M.: Robust recovery of signals from a structured union of subspaces. Information Theory, IEEE Transactions on **55**(11), 5302–5316 (2009) 7
23. Foucart, S., Rauhut, H.: A mathematical introduction to compressive sensing. Springer (2013). URL http://link.springer.com/content/pdf/10.1007/978-0-8176-4948-7.pdf 16

24. Grosse, R., Raina, R., Kwong, H., Ng, A.Y.: Shift-invariance sparse coding for audio classification. arXiv preprint arXiv:1206.5241 (2012) 2
25. Grosse, R., Raina, R., Kwong, H., Ng, A.Y.: Shift-invariance sparse coding for audio classification. arXiv:1206.5241 [cs, stat] (2012). URL http://arxiv.org/abs/1206.5241. arXiv: 1206.5241 26
26. Gu, S., Zuo, W., Xie, Q., Meng, D., Feng, X., Zhang, L.: Convolutional sparse coding for image super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1823–1831 (2015) 2
27. Heide, F., Heidrich, W., Wetzstein, G.: Fast and flexible convolutional sparse coding. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5135–5143. IEEE (2015) 2
28. Huang, J., Zhang, T., Metaxas, D.: Learning with structured sparsity. Journal of Machine Learning Research **12**(Nov), 3371–3412 (2011) 2, 36
29. Huang, J., Zhang, T., et al.: The benefit of group sparsity. The Annals of Statistics **38**(4), 1978–2004 (2010) 2, 36
30. Kavukcuoglu, K., Sermanet, P., Boureau, Y.l., Gregor, K., Mathieu, M., Cun, Y.L.: Learning convolutional feature hierarchies for visual recognition. In: J.D. Lafferty, C.K.I. Williams, J. Shawe-Taylor, R.S. Zemel, A. Culotta (eds.) Advances in Neural Information Processing Systems 23, pp. 1090–1098. Curran Associates, Inc. (2010). URL http://papers.nips.cc/paper/4133-learning-convolutional-feature-hierarchies-for-visual-recognition.pdf 26
31. Kyrillidis, A., Baldassarre, L., Halabi, M.E., Tran-Dinh, Q., Cevher, V.: Structured Sparsity: Discrete and Convex Approaches. In: H. Boche, R. Calderbank, G. Kutyniok, J. Vybíral (eds.) Compressed Sensing and its Applications, Applied and Numerical Harmonic Analysis, pp. 341–387. Springer International Publishing (2015). URL http://link.springer.com/chapter/10.1007/978-3-319-16042-9_12. DOI: 10.1007/978-3-319-16042-9_12 2, 36
32. Lu, Y.M., Do, M.N.: A theory for sampling signals from a union of subspaces. IEEE Transactions on Signal Processing (2007) 7
33. Mairal, J., Bach, F., Ponce, J.: Sparse modeling for image and vision processing. arXiv preprint arXiv:1411.3230 (2014) 2
34. Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A.: Non-local sparse models for image restoration. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 2272–2279. IEEE (2009) 2
35. Mairal, J., Sapiro, G., Elad, M.: Learning multiscale sparse representations for image and video restoration. Multiscale Modeling & Simulation **7**(1), 214–241 (2008) 2
36. Papyan, V., Elad, M.: Multi-Scale Patch-Based Image Restoration. IEEE Transactions on Image Processing **25**(1), 249–261 (2016). DOI 10.1109/TIP.2015.2499698 2
37. Papyan, V., Romano, Y., Elad, M.: Convolutional Neural Networks Analyzed via Convolutional Sparse Coding. arXiv:1607.08194 [cs, stat] (2016). URL http://arxiv.org/abs/1607.08194. ArXiv: 1607.08194 2, 36
38. Papyan, V., Sulam, J., Elad, M.: Working locally thinking globally - Part I: Theoretical guarantees for convolutional sparse coding. arXiv preprint arXiv:1607.02005 (2016) 2, 26, 27, 36
39. Papyan, V., Sulam, J., Elad, M.: Working locally thinking globally - Part II: Stability and algorithms for convolutional sparse coding. arXiv preprint arXiv:1607.02009 (2016) 2, 26, 27, 36
40. Pati, Y.C., Rezaiifar, R., Krishnaprasad, P.: Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In: Asilomar Conference on Signals, Systems and Computers, pp. 40–44. IEEE (1993) 3, 16
41. Quiroga, R.: Spike sorting. Scholarpedia **2**(12), 3583 (2007). DOI 10.4249/scholarpedia.3583 26
42. Romano, Y., Elad, M.: Boosting of image denoising algorithms. SIAM Journal on Imaging Sciences **8**(2), 1187–1219 (2015) 2, 43

43. Romano, Y., Elad, M.: Patch-disagreement as away to improve k-svd denoising. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1280–1284. IEEE (2015) 2
44. Romano, Y., Protter, M., Elad, M.: Single image interpolation via adaptive nonlocal sparsity-based modeling. IEEE Transactions on Image Processing **23**(7), 3085–3098 (2014) 2
45. Rusu, C., Dumitrescu, B., Tsaftaris, S.: Explicit shift-invariant dictionary learning. IEEE Signal Process. Lett **21**, 6–9 (2014). URL http://www.schur.pub.ro/Idei2011/Articole/SPL_2014_shifts.pdf 2
46. Smith, E., Lewicki, M.S.: Efficient coding of time-relative structure using spikes. Neural Computation **17**(1), 19–45 (2005). URL http://dl.acm.org/citation.cfm?id=1119614 26
47. Sulam, J., Elad, M.: Expected patch log likelihood with a sparse prior. In: International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition, pp. 99–111. Springer (2015) 2
48. Sulam, J., Ophir, B., Elad, M.: Image denoising through multi-scale learnt dictionaries. In: 2014 IEEE International Conference on Image Processing (ICIP), pp. 808–812. IEEE (2014) 2
49. Thiagarajan, J.J., Ramamurthy, K.N., Spanias, A.: Shift-invariant sparse representation of images using learned dictionaries. In: IEEE Workshop on Machine Learning for Signal Processing, 2008. MLSP 2008, pp. 145–150 (2008). DOI 10.1109/MLSP.2008.4685470 2
50. Tropp, J.A., Gilbert, A.C., Strauss, M.J.: Algorithms for simultaneous sparse approximation. part i: Greedy pursuit. Signal Processing **86**(3), 572–588 (2006) 2, 36
51. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. IEEE transactions on image processing **19**(11), 2861–2873 (2010) 2
52. Yu, G., Sapiro, G., Mallat, S.: Solving inverse problems with piecewise linear estimators: From gaussian mixture models to structured sparsity. IEEE Transactions on Image Processing **21**(5), 2481–2499 (2012). DOI 10.1109/TIP.2011.2176743 2
53. Zeiler, M., Taylor, G., Fergus, R.: Adaptive deconvolutional networks for mid and high level feature learning. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 2018–2025 (2011). DOI 10.1109/ICCV.2011.6126474 26
54. Zeiler, M.D., Krishnan, D., Taylor, G.W., Fergus, R.: Deconvolutional networks. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp. 2528–2535. IEEE (2010). URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5539957 26
55. Zoran, D., Weiss, Y.: From learning models of natural image patches to whole image restoration. In: Computer Vision (ICCV), 2011 IEEE International Conference on, p. 479–486. IEEE (2011). URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6126278 2