Stable recovery of the factors from a deep matrix product and application to convolutional networks. Focus on sparsity constraints.

François Malgouyres

Institut de Mathématiques de Toulouse ; UMR5219
Université de Toulouse ; CNRS
UPS IMT F-31062 Toulouse Cedex 9, France
malgouyres@math.univ-toulouse.fr

Joseph Landsberg

Department of Mathematics
Mailstop 3368
Texas A& M University
College Station, TX 77843-3368
jml@math.tamu.edu

Abstract

We study a deep matrix factorization problem. It takes as input a matrix X obtained by multiplying K matrices (called factors). Each factor is obtained by applying a fixed linear operator to a vector of parameters satisfying a sparsity constraint. We provide sharp conditions on the structure of the model that guarantee the stable recovery of the factors from the knowledge of X and the model for the factors. This is crucial in order to interpret the factors and the intermediate features obtained when applying a few factors to a datum. When $K = 1$: the paper provides compressed sensing statements; $K = 2$ covers (for instance) Non-negative Matrix Factorization, Dictionary learning, low rank approximation, phase recovery. The particularity of this paper is to extend the study to deep problems. As an illustration, we detail the analysis and provide (entirely computable) guarantees for the stable recovery of a (non-neural) sparse convolutional network.

1 Introduction

Let $K \in \mathbb{N}^*$, $m_1 \dots m_{K+1} \in \mathbb{N}$, write $m_1 = m$, $m_{K+1} = n$. We impose the factors to be structured matrices defined by a (typically small) number S of unknown parameters. More precisely, for $k = 1 \dots K$, let

$$\begin{array}{rcl} M_k : \mathbb{R}^S & \longrightarrow & \mathbb{R}^{m_k \times m_{k+1}}, \\ h & \longmapsto & M_k(h) \end{array}$$

be a linear map. We assume that we know the matrix $X \in \mathbb{R}^{m \times n}$ which is provided by

$$X = M_1(\overline{\mathbf{h}}_1) \cdots M_K(\overline{\mathbf{h}}_K) + e, \tag{1}$$

for an unknown error term e and parameters $\overline{\mathbf{h}} = (\overline{\mathbf{h}}_k)_{1 \leq k \leq K} \in \mathbb{R}^{S \times K}$. Moreover, considering a family of possible supports $\mathcal{M}$ (e.g., all the supports of size S', for a given $S' \leq S$). We assume

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

that the $\overline{\mathbf{h}}$ satisfy a sparsity constraint of the form : there exists $\overline{\mathcal{S}} = (\overline{\mathcal{S}}_k)_{1\leq k\leq K} \in \mathcal{M}$ such that $\text{supp}\left(\overline{\mathbf{h}}\right) \subset \overline{\mathcal{S}}$ (i.e.: $\forall k$, $\text{supp}\left(\overline{\mathbf{h}}_k\right) \subset \overline{\mathcal{S}}_k$)

This work investigates conditions imposed on (1) for which we can (up to obvious scale rearrangement) stably recover the parameters $\overline{\mathbf{h}}$ from X. When $K = 1$, the statements are compressed sensing statements [15]. When $K = 2$ and when extended to other constraints on the parameters $\overline{\mathbf{h}}$, the statements apply to already studied problems such as: low rank approximation [4], Non-negative matrix factorization [11, 8, 10, 2], dictionary learning [9], phase retrieval [4], blind deconvolution [1, 7, 12]. Most of these papers use the same lifting property we are using. They further propose to convexify the problem. A more general bilinear framework is considered in [7]. The main existing statements when $K \geq 3$ is very recent [13]. It is also applied to (non-neural) convolutional networks.

The present work describes an alternative analysis, specialized to sparsity constraints, of the results exposed in [13]. Doing so, we obtain better bounds (defined with an analogue of the lower-RIP) and weaker constraints on the model. Its application to (non-neural) sparse convolutional networks lead to simple necessary and sufficient conditions of stable recovery, for a large class of solvers. The stability inequality (see Theorem 3) only involves explicit and simple ingredients of the problem.

2 Notations and preliminaries on tensors

Set $\mathbb{N}_K = \{1, \ldots, K\}$ and $\mathbb{R}_*^{S\times K} = \{\mathbf{h} \in \mathbb{R}^{S\times K}, \forall k \in \mathbb{N}_K, \|\mathbf{h}_k\| \neq 0\}$. Define an equivalence relation in $\mathbb{R}_*^{S\times K}$: for any $\mathbf{h}$, $\mathbf{g} \in \mathbb{R}^{S\times K}$, $\mathbf{h} \sim \mathbf{g}$ if and only if there exists $(\lambda_k)_{k\in\mathbb{N}_K} \in \mathbb{R}^K$ such that

$$\prod_{k=1}^{K} \lambda_k = 1 \qquad \text{and} \qquad \forall k \in \mathbb{N}_K, \mathbf{h}_k = \lambda_k \mathbf{g}_k.$$

Denote the equivalence class of $\mathbf{h} \in \mathbb{R}_*^{S\times K}$ by $[\mathbf{h}]$. For any $p \in [1, \infty]$, we denote the usual ℓ^p norm by $\|.\|_p$ and define the mapping $d_p : (\mathbb{R}_*^{S\times K}/\sim \times \mathbb{R}_*^{S\times K}/\sim) \to \mathbb{R}$ by

$$d_p([\mathbf{h}], [\mathbf{g}]) = \inf_{\substack{\mathbf{h}' \in [\mathbf{h}] \cap \mathbb{R}_{\text{diag}}^{S\times K} \\ \mathbf{g}' \in [\mathbf{g}] \cap \mathbb{R}_{\text{diag}}^{S\times K}}} \|\mathbf{h}' - \mathbf{g}'\|_p \qquad , \forall \mathbf{h}, \mathbf{g} \in \mathbb{R}_*^{S\times K}, \tag{2}$$

where

$$\mathbb{R}_{\text{diag}}^{S\times K} = \{\mathbf{h} \in \mathbb{R}_*^{S\times K}, \forall k \in \mathbb{N}_K, \|\mathbf{h}_k\|_\infty = \|\mathbf{h}_1\|_\infty\}.$$

It is proved in [13] that d_p is a metric on $\mathbb{R}_*^{S\times K}/\sim$.

The real valued tensors of order K whose axes are of size S are denoted by $T \in \mathbb{R}^{S\times\ldots\times S}$. The space of tensors is abbreviated $\mathbb{R}^{S^K}$. We say that a tensor $T \in \mathbb{R}^{S^K}$ is of *rank* 1 if and only if there exists a collection of vectors $\mathbf{h} \in \mathbb{R}^{S\times K}$ such that T is the outer product of the vectors $\mathbf{h}_k$, for $k \in \mathbb{N}_K$, that is, for any $\mathbf{i} = (i_1, \ldots, i_K) \in \mathbb{N}_S^K$,

$$T_{\mathbf{i}} = \mathbf{h}_{1,i_1} \ldots \mathbf{h}_{K,i_K}.$$

The set of all the tensors of rank 1 is denoted by Σ_1. Moreover, we parametrize $\Sigma_1 \subset \mathbb{R}^{S^K}$ using the Segre embedding

$$\begin{array}{rcl} P : \mathbb{R}^{S\times K} & \longrightarrow & \Sigma_1 \subset \mathbb{R}^{S^K} \\ \mathbf{h} & \longmapsto & (\mathbf{h}_{1,i_1}\mathbf{h}_{2,i_2} \ldots \mathbf{h}_{K,i_K})_{\mathbf{i}\in\mathbb{N}_S^K} \end{array} \tag{3}$$

Theorem 1. Stability of $[\mathbf{h}]$ from $P(\mathbf{h})$, see [13]

Let $\mathbf{h}$ and $\mathbf{g} \in \mathbb{R}_*^{S\times K}$ be such that $\|P(\mathbf{g}) - P(\mathbf{h})\|_\infty \leq \frac{1}{2} \max(\|P(\mathbf{h})\|_\infty, \|P(\mathbf{g})\|_\infty)$. For all $p, q \in [1, \infty]$,

$$d_p([\mathbf{h}], [\mathbf{g}]) \leq 7(KS)^{\frac{1}{p}} \min\left(\|P(\mathbf{h})\|_\infty^{\frac{1}{K}-1}, \|P(\mathbf{g})\|_\infty^{\frac{1}{K}-1}\right) \|P(\mathbf{h}) - P(\mathbf{g})\|_q. \tag{4}$$

It is also explained in [13] that we can *lift* the problem and show that the map

$$(\mathbf{h}_1, \ldots, \mathbf{h}_K) \longmapsto M_1(\mathbf{h}_1)M_2(\mathbf{h}_2) \ldots M_K(\mathbf{h}_K),$$

uniquely determines a linear map

$$\mathcal{A} : \mathbb{R}^{S^K} \longrightarrow \mathbb{R}^{m\times n},$$

such that for all $\mathbf{h} \in \mathbb{R}^{S\times K}$

$$M_1(\mathbf{h}_1)M_2(\mathbf{h}_2) \ldots M_K(\mathbf{h}_K) = \mathcal{A}P(\mathbf{h}). \tag{5}$$

3 General conditions for the stable recovery under sparsity constraint

From now on, the analysis differs from the one presented in [13]. It is dedicated to models that enforce sparsity. In this particular situation, we can indeed have a different view of the geometry of the problem. In order to describe it, we first establish some notation.

We define a support by $\mathcal{S} = (\mathcal{S}_k)_{1\leq k\leq K}$, with $\mathcal{S}_k \subset \mathbb{N}_S$, and denote the set of all supports by $\mathbb{S}$. We have $\mathcal{M} \subset \mathbb{S}$. For a given support $\mathcal{S} \in \mathbb{S}$, we denote

$$\mathbb{R}^{S\times K}_{\mathcal{S}} = \{\mathbf{h} \in \mathbb{R}^{S\times K} \mid \mathbf{h}_{k,i} = 0, \text{ for all } k \in \mathbb{N}_K \text{ and } i \notin \mathcal{S}_k\}$$

(i.e., for all k, $\operatorname{supp}(\mathbf{h}_k) \subset \mathcal{S}_k$) and

$$\mathbb{R}^{S^K}_{\mathcal{S}} = \{T \in \mathbb{R}^{S^K} \mid T_{\mathbf{i}} = 0, \text{ for all } \mathbf{i} \notin \mathcal{S}\}.$$

We also denote by $\mathcal{P}_{\mathcal{S}}$ the orthogonal projection from $\mathbb{R}^{S^K}$ onto $\mathbb{R}^{S^K}_{\mathcal{S}}$. We trivially have for all $T \in \mathbb{R}^{S^K}$ and all $\mathbf{i} \in \mathbb{N}^K_S$

$$(\mathcal{P}_{\mathcal{S}}T)_{\mathbf{i}} = \begin{cases} T_{\mathbf{i}} & \text{, if } \mathbf{i} \in \mathcal{S}, \\ 0 & \text{, otherwise.} \end{cases}$$

As explained in the introduction, we assume that there exists a known family of admissible supports $\mathcal{M} \subset \mathbb{S}$, an unknown support $\overline{\mathcal{S}} \in \mathcal{M}$ and unknown parameters $\overline{\mathbf{h}} \in \mathbb{R}^{S\times K}_{\overline{\mathcal{S}}}$ that we would like to estimate from the noisy matrix product

$$X = M_1(\overline{\mathbf{h}}_1)\ldots M_K(\overline{\mathbf{h}}_K) + e. \tag{6}$$

We assume that there exists $\delta \geq 0$ such that the error satisfies

$$\|e\| \leq \delta. \tag{7}$$

Also, we consider an inexact minimization context and assume that we have a way to find $\mathcal{S}^* \in \mathcal{M}$ and $\mathbf{h}^* \in \mathbb{R}^{S\times K}_{\mathcal{S}^*}$ such that

$$\|M_1(\mathbf{h}^*_1)\ldots M_K(\mathbf{h}^*_K) - X\| \leq \eta, \tag{8}$$

for some $\eta > 0$. Typically, $\mathcal{S}^*$ and $\mathbf{h}^*$ are found by an algorithm (most often a heuristic) that tries to solve

$$\operatorname{argmin}_{\mathcal{S}\in\mathcal{M}, \mathbf{h}\in\mathbb{R}^{S\times K}_{\mathcal{S}}} \|M_1(\mathbf{h}_1)\ldots M_K(\mathbf{h}_K) - X\|^2.$$

However, what matters is not to truly solve the above problem (which might be impossible to truly minimize or might only be accessible up to a generalization error). What matters is to find an input for which its objective function is reasonably small. Notice that, if the minimization is proved to be accurate, we can assume in the bounds that will be presented latter in the paper that $\eta \leq \delta$.

In the geometrical view described in the sequel, we consider different linear operators $\mathcal{A}_{\mathcal{S}}$, with $\mathcal{S} \in \mathbb{S}$, such that for all $\mathbf{h} \in \mathbb{R}^{S\times K}_{\mathcal{S}}$

$$\mathcal{A}_{\mathcal{S}}P(\mathbf{h}) = M_1(\mathbf{h}_1)\ldots M_K(\mathbf{h}_K).$$

In order to achieve that, considering (5), we simply define for any $\mathcal{S} \in \mathbb{S}$

$$\mathcal{A}_{\mathcal{S}} = \mathcal{A}\mathcal{P}_{\mathcal{S}}. \tag{9}$$

Definition 1. Deep-$\mathcal{M}$-Null Space Property

Let $\gamma > 0$, we say that $\mathcal{A}$ satisfies the deep-$\mathcal{M}$-Null Space Property (deep-$\mathcal{M}$-NSP) *with constant γ if there exists $\varepsilon > 0$ such that for all $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$, any $T \in P(\mathbb{R}^{S\times K}_{\mathcal{S}}) - P(\mathbb{R}^{S\times K}_{\mathcal{S}'})$ satisfying $\|\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'}T\| \leq \varepsilon$ and any $T' \in \operatorname{Ker}(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'})$, we have*

$$\|T\| \leq \gamma\|T - \mathcal{P}_{\mathcal{S}\cup\mathcal{S}'}T'\|. \tag{10}$$

An interesting property of the deep-$\mathcal{M}$-NSP is that, as stated in the next proposition, it can be a consequence of simple tests.

Proposition 1. Sufficient condition for deep-$\mathcal{M}$-NSP

If $\mathrm{Ker}\,(\mathcal{A}) \cap \mathbb{R}^{S^K}_{\mathcal{S}\cup\mathcal{S}'} = \{0\}$*, for all* $\mathcal{S}$ *and* $\mathcal{S}' \in \mathcal{M}$*, then* $\mathcal{A}$ *satisfies the deep-*$\mathcal{M}$*-NSP with constant* $\gamma = 1$ *(and any* ε*).*

As a consequence, almost every $\mathcal{A}$ *such that*

$$\mathrm{rk}\,(\mathcal{A}) \geq \max_{\mathcal{S},\mathcal{S}'\in\mathcal{M}} (\dim(\mathbb{R}^{S^K}_{\mathcal{S}\cup\mathcal{S}'}))$$

$\mathcal{A}$ *satisfies the deep-*$\mathcal{M}$*-NSP with constant* $\gamma = 1$ *(and any* ε*).*

Proof. In order to prove the first statement of the proposition, let us consider $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$, $T' \in \mathrm{Ker}\,(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'})$. We have $\mathcal{A}\mathcal{P}_{\mathcal{S}\cup\mathcal{S}'}T' = 0$ and therefore $\mathcal{P}_{\mathcal{S}\cup\mathcal{S}'}T' \in \mathrm{Ker}\,(\mathcal{A})$. Moreover, by definition, $\mathcal{P}_{\mathcal{S}\cup\mathcal{S}'}T' \in \mathbb{R}^{S^K}_{\mathcal{S}\cup\mathcal{S}'}$. Therefore, applying the hypothesis of the proposition, we obtain $\mathcal{P}_{\mathcal{S}\cup\mathcal{S}'}T' = 0$ and (10) always holds when $\gamma = 1$.

The second assertion is a straightforward consequence of the first one. □

For instance, for any given $S' \leq \frac{S}{2}$, if $\mathcal{M}$ contains all the supports $\mathcal{S}$ such that, for all $k \in \mathbb{N}_K$, $|\mathcal{S}_k| = S'$, we have $\max_{\mathcal{S},\mathcal{S}'\in\mathcal{M}}(\dim(\mathbb{R}^{S^K}_{\mathcal{S}\cup\mathcal{S}'})) = (2S')^K$. The proposition guarantees that almost every $\mathcal{A}$ such that $\mathrm{rk}\,(\mathcal{A}) \geq (2S')^K$ satisfies the deep-$\mathcal{M}$-NSP with constant $\gamma = 1$ (and any ε).

Interestingly, in the compressed sensing framework when $K = 1$ and $\mathcal{A}$ is a sampling matrix, the first statement of the above proposition says that any sampling matrix with column rank larger than twice the maximal sparsity allowed by the model satisfies the deep-$\mathcal{M}$-NSP with constant $\gamma = 1$ (and any ε). To the best of our knowledge, this leads, even in the case $K = 1$ to a new stability condition.

Definition 2. Deep-lower-RIP constant

There exists a constant $\sigma_{\mathcal{M}} > 0$ *such that for any* $\mathcal{S}$ *and* $\mathcal{S}' \in \mathcal{M}$ *and any* T *in the orthogonal complement of* $\mathrm{Ker}\,(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'})$

$$\sigma_{\mathcal{M}} \|\mathcal{P}_{\mathcal{S}\cup\mathcal{S}'}T\| \leq \|\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'}T\|. \tag{11}$$

We call $\sigma_{\mathcal{M}}$ *Deep-lower-RIP constant of* $\mathcal{A}$ *with regard to* $\mathcal{M}$*.*

The existence of $\sigma_{\mathcal{M}}$ is a straightforward consequence of the fact that the restriction of $\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'}$ on the orthogonal complement of $\mathrm{Ker}\,(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'})$ is injective. We even have for all T in the orthogonal complement of $\mathrm{Ker}\,(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'})$

$$\|\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'}T\| \geq \sigma_{\mathcal{S}\cup\mathcal{S}'}\|T\| \geq \sigma_{\mathcal{S}\cup\mathcal{S}'}\|\mathcal{P}_{\mathcal{S}\cup\mathcal{S}'}T\|,$$

where $\sigma_{\mathcal{S}\cup\mathcal{S}'} > 0$ is the smallest non-zero singular value of $\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'}$. We obtain the existence of $\sigma_{\mathcal{M}}$ by taking the infimum of the constants $\sigma_{\mathcal{S}\cup\mathcal{S}'}$ over the finite family of $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$.

Theorem 2. Sufficient condition for stable recovery

Assume $\mathcal{A}$ *satisfies the deep-*$\mathcal{M}$*-NSP with the constant* $\gamma > 0$*. For any* $\mathcal{S}^* \in \mathcal{M}$ *and* $\mathbf{h}^* \in \mathbb{R}^{S\times K}_{\mathcal{S}^*}$ *as in* (8) *with* η *and* δ *sufficiently small, we have*

$$\|P(\mathbf{h}^*) - P(\overline{\mathbf{h}})\| \leq \frac{\gamma}{\sigma_{\mathcal{M}}}\,(\delta + \eta),$$

where $\sigma_{\mathcal{M}}$ *is the Deep-lower-RIP constant of* $\mathcal{A}$ *with regard to* $\mathcal{M}$*. Moreover, if* $\overline{\mathbf{h}} \in \mathbb{R}^{S\times K}_{*}$

$$d_p([\mathbf{h}^*], [\overline{\mathbf{h}}]) \leq 7(KS)^{\frac{1}{p}} \min\left(\|P(\overline{\mathbf{h}})\|_\infty^{\frac{1}{K}-1}, \|P(\mathbf{h}^*)\|_\infty^{\frac{1}{K}-1}\right) \frac{\gamma}{\sigma_{\mathcal{M}}}(\delta + \eta).$$

Proof. We have

$$\begin{aligned} \|\mathcal{A}_{\mathcal{S}^*\cup\overline{\mathcal{S}}}(P(\mathbf{h}^*) - P(\overline{\mathbf{h}}))\| &= \|\mathcal{A}_{\mathcal{S}^*}P(\mathbf{h}^*) - \mathcal{A}_{\overline{\mathcal{S}}}P(\overline{\mathbf{h}})\| \\ &\leq \|\mathcal{A}_{\mathcal{S}^*}P(\mathbf{h}^*) - X\| + \|\mathcal{A}_{\overline{\mathcal{S}}}P(\overline{\mathbf{h}}) - X\| \\ &\leq \delta + \eta \end{aligned}$$

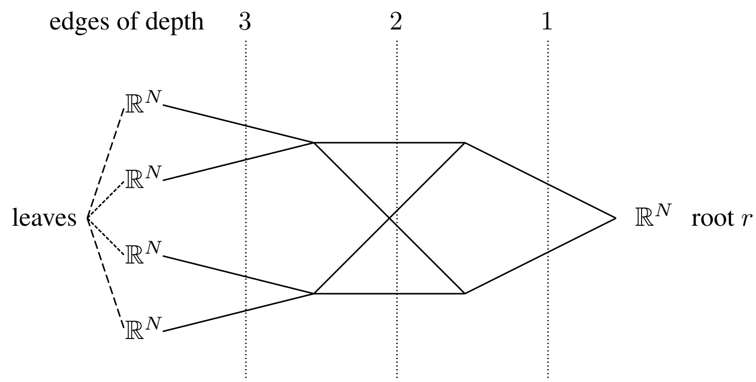

Figure 1: Example of the considered convolutional network. To every edge is attached a convolution kernel. The network does not involve non-linearities or sampling.

If we further decompose (the decomposition is unique)

$$P(\mathbf{h}^*) - P(\overline{\mathbf{h}}) = T + T',$$

where $T' \in \mathrm{Ker}\,(\mathcal{A}_{\mathcal{S}^* \cup \overline{\mathcal{S}}})$ and T is orthogonal to $\mathrm{Ker}\,(\mathcal{A}_{\mathcal{S}^* \cup \overline{\mathcal{S}}})$, we have

$$\|\mathcal{A}_{\mathcal{S}^* \cup \overline{\mathcal{S}}}(P(\mathbf{h}^*) - P(\overline{\mathbf{h}}))\| = \|\mathcal{A}_{\mathcal{S}^* \cup \overline{\mathcal{S}}} T\| \geq \sigma_{\mathcal{M}} \|\mathcal{P}_{\mathcal{S}^* \cup \overline{\mathcal{S}}} T\|,$$

where $\sigma_{\mathcal{M}}$ is the Deep-lower-RIP constant of $\mathcal{A}$ with regard to $\mathcal{M}$. We finally obtain, since $\mathcal{P}_{\mathcal{S}^* \cup \overline{\mathcal{S}}} P(\mathbf{h}^*) = P(\mathbf{h}^*)$ and $\mathcal{P}_{\mathcal{S}^* \cup \overline{\mathcal{S}}} P(\overline{\mathbf{h}}) = P(\overline{\mathbf{h}})$,

$$\|P(\mathbf{h}^*) - P(\overline{\mathbf{h}}) - \mathcal{P}_{\mathcal{S}^* \cup \overline{\mathcal{S}}} T'\| = \|\mathcal{P}_{\mathcal{S}^* \cup \overline{\mathcal{S}}} T\| \leq \frac{\delta + \eta}{\sigma_{\mathcal{M}}}.$$

Since $\mathcal{A}$ satisfies the deep-$\mathcal{M}$-NSP with constant γ, when $\delta + \eta \leq \varepsilon$, the first inequality of the theorem holds:

$$\|P(\mathbf{h}^*) - P(\overline{\mathbf{h}})\| \leq \gamma \, \frac{\delta + \eta}{\sigma_{\mathcal{M}}}.$$

When $\overline{\mathbf{h}} \in \mathbb{R}_*^{S \times K}$, for $\delta + \eta$ small, we can apply Theorem 1 and obtain the second inequality. $\square$

Theorem 2 differ from its analogue in [13]. In particular, it is dedicated to sparsity constraint with much weaker hypotheses on $\mathcal{A}$. The constant of the upper bound is also different.

4 Application to (non neural) sparse convolutional network

We consider a (non neural) sparse convolutional network as depicted in Figure 1. The network typically aims at performing a linear analysis or synthesis of a signal living in $\mathbb{R}^N$. The considered convolutional network is defined from a rooted directed acyclic graph $\mathcal{G}(\mathcal{E}, \mathcal{N})$ composed of nodes $\mathcal{N}$ and edges $\mathcal{E}$. Each edge connects two nodes. The root of the graph is denoted by r and the set containing all its leaves is denoted by $\mathcal{F}$. We denote by $\mathcal{P}$ the set of all paths connecting the leaves and the root. We assume, without loss of generality, that the length of any path between any leaf and the root is independent of the considered leaf and equal to some constant $K \geq 0$. We also assume that, for any edge $e \in \mathcal{E}$, the number of edges separating e and the root is the same for all paths between e and r. This length is called the depth of e. For any $k \in \mathbb{N}_K$, we denote the set containing all the edges of depth k, by $\mathcal{E}(k)$. For $e \in \mathcal{E}(k)$, we also say that e belongs to the layer k.

Moreover, to any edge e is attached a convolution kernel of maximal support $\mathcal{S}_e \subset \mathbb{N}_N$. We assume (without loss of generality) that $\sum_{e \in \mathcal{E}(k)} |\mathcal{S}_e|$ is independent of k ($|\mathcal{S}_e|$ denotes the cardinality of $\mathcal{S}_e$). We take

$$S = \sum_{e \in \mathcal{E}(1)} |\mathcal{S}_e|.$$

For any edge e, we consider the mapping $\mathcal{T}_e : \mathbb{R}^S \longrightarrow \mathbb{R}^N$ that maps any $h \in \mathbb{R}^S$ into the convolution kernel h_e, attached to the edge e, whose support is $\mathcal{S}_e$. It simply writes at the right location (i.e. those in $\mathcal{S}_e$) the entries of h defining the kernel on the edge e. As in the previous section, we assume a sparsity constraint and will only consider a family $\mathcal{M}$ of possible supports $\mathcal{S} \subset \mathbb{N}_S^K$.

At each layer k, the convolutional network computes, for all $e \in \mathcal{E}(k)$, the convolution between the signal at the origin of e; then, it attaches to any ending node the sum of all the convolutions arriving at that node. Examples of such convolutional networks includes wavelets, wavelet packets [14] or the fast transforms optimized in [5, 6]. It is similar to the usual convolutional neural network except that the network does not involve any non-linearity and the supports are not fixed. It is clear that the operation performed at any layer depends linearly on the parameters $h \in \mathbb{R}^S$ and that its results serves as inputs for the next layer. The (non neural) convolutional network therefore depends on parameters $\mathbf{h} \in \mathbb{R}^{S \times K}$ and takes the form

$$X = M_1(\mathbf{h}_1) \dots M_K(\mathbf{h}_K),$$

where the operators M_k satisfy the hypothesis of the present paper.

This section applies the results of the preceding section in order to identify conditions such that any unknown parameters $\overline{\mathbf{h}} \in \mathbb{R}^{S \times K}$ satisfying $\operatorname{supp}\left(\overline{\mathbf{h}}\right) \subset \overline{\mathcal{S}}$, for a given $\overline{\mathcal{S}} \in \mathcal{M}$, can be stably recovered from $X = M_1(\overline{\mathbf{h}}_1) \dots M_K(\overline{\mathbf{h}}_K)$ (possibly corrupted by an error).

In order to do so, let us define a few notations. Notice first that, we apply the convolutional network to an input $x \in \mathbb{R}^{N|\mathcal{F}|}$, where x is the concatenation of the signals $x^f \in \mathbb{R}^N$ for $f \in \mathcal{F}$. Therefore, X is the (horizontal) concatenation of $|\mathcal{F}|$ matrices $X^f \in \mathbb{R}^{N \times N}$ such that

$$Xx = \sum_{f \in \mathcal{F}} X^f x^f \qquad \text{, for all } x \in \mathbb{R}^{N|\mathcal{F}|}.$$

Let us consider the convolutional network defined by $\mathbf{h} \in \mathbb{R}^{S \times K}$ as well as $f \in \mathcal{F}$ and $n \in \mathbb{N}_N$. The column of X corresponding to the entry n in the leaf f is the translation by n of

$$\sum_{p \in \mathcal{P}(f)} \mathcal{T}^p(\mathbf{h}) \tag{12}$$

where $\mathcal{P}(f)$ contains all the paths of $\mathcal{P}$ starting from the leaf f and

$$\mathcal{T}^p(\mathbf{h}) = \mathcal{T}_{e^1}(\mathbf{h}_1) * \dots * \mathcal{T}_{e^K}(\mathbf{h}_K) \qquad \text{, where } p = (e^1, \dots, e^K).$$

Moreover, we define for any $k \in \mathbb{N}_K$ the mapping $\mathbf{e}_k : \mathbb{N}_S \longrightarrow \mathcal{E}(k)$ which provides for any $i \in \mathbb{N}_S$ the unique edge of $\mathcal{E}(k)$ such that the i^{th} entry of $h \in \mathbb{R}^S$ contributes to $\mathcal{T}_{\mathbf{e}_k(i)}(h)$. Also, for any $\mathbf{i} \in \mathbb{N}_S^K$, we denote $\mathbf{p}_\mathbf{i} = (\mathbf{e}_1(\mathbf{i}_1), \dots, \mathbf{e}_K(\mathbf{i}_K))$ and, for any $\mathcal{S} \in \mathcal{M}$,

$$\mathbf{I}_\mathcal{S} = \left\{ \mathbf{i} \in \mathbb{N}_S^K \,|\, \mathbf{i} \in \mathcal{S} \text{ and } \mathbf{p}_\mathbf{i} \in \mathcal{P} \right\}.$$

The latter contains all the indices of $\mathcal{S}$ corresponding to a valid path in the network. For any set of parameters $\mathbf{h} \in \mathbb{R}^{S \times K}$ and any path $\mathbf{p} \in \mathcal{P}$, we also denote by $\mathbf{h}^\mathbf{p}$ the restriction of $\mathbf{h}$ to its indices contributing to the kernels on the path $\mathbf{p}$. We also define, for any $\mathbf{i} \in \mathbb{N}_S^K$, $\mathbf{h}^\mathbf{i} \in \mathbb{R}^{S \times K}$ (beware not to confuse the notations) by

$$\mathbf{h}^\mathbf{i}_{k,j} = \begin{cases} 1 & \text{, if } j = \mathbf{i}_k \\ 0 & \text{otherwise} \end{cases} \qquad \text{, for all } k \in \mathbb{N}_K \text{ and } j \in \mathbb{N}_S. \tag{13}$$

We can deduce from (12) that $\mathcal{A}P(\mathbf{h}^\mathbf{i})$ simply convolves the entries at one leaf with a dirac delta function. Thefore, all the entries of $\mathcal{A}P(\mathbf{h}^\mathbf{i})$ are in $\{0, 1\}$ and we denote $\mathcal{D}_\mathbf{i} = \{(i,j) \in \mathbb{N}_N \times \mathbb{N}_{N|\mathcal{F}|} | \mathcal{A}P(\mathbf{h}^\mathbf{i})_{i,j} = 1\}$.

We also denote $\mathbb{1} \in \mathbb{R}^S$ a vector of size S with all its entries equal to 1. For any edge $e \in \mathcal{E}$, $\mathbb{1}^e \in \mathbb{R}^S$ consists of zeroes except for the entries corresponding to the edge e which are equal to 1. For any $\mathcal{S} \subset \mathbb{N}_S$, we define $\mathbb{1}^\mathcal{S} \in \mathbb{R}^S$ which consists of zeroes except for the entries corresponding to the indexes in $\mathcal{S}$. For any $\mathbf{p} = (e^1, \dots, e^K) \in \mathcal{P}$, the support of $M_1(\mathbb{1}^{e_1}) \dots M_K(\mathbb{1}^{e_K})$ is denoted by $\mathcal{D}^\mathbf{p}$.

Finally, we remind that because of (5), there exists a unique mapping

$$\mathcal{A} : \mathbb{R}^{S^K} \longrightarrow \mathbb{R}^{N\times N|\mathcal{F}|}$$

such that

$$\mathcal{A}P(\mathbf{h}) = M_1(\mathbf{h}_1)\ldots M_K(\mathbf{h}_K) \qquad \text{, for all } \mathbf{h} \in \mathbb{R}^{S\times K},$$

where P is the Segre embedding defined in (3).

Proposition 2. Necessary condition of identifiability of a sparse network

Either $\mathbb{R}^{S\times K}$ is not identifiable or, for any $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$, all the entries of $M_1(\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'})\ldots M_K(\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'})$ belong to $\{0,1\}$. When the latter holds :

1. *For any distinct $\mathbf{p}$ and $\mathbf{p}' \in \mathcal{P}$, we have $\mathcal{D}^{\mathbf{p}} \cap \mathcal{D}^{\mathbf{p}'} = \emptyset$.*
2. $\mathrm{Ker}\,(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'}) = \{T \in \mathbb{R}^{S^K} \,|\, \forall \mathbf{i} \in \mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}, T_{\mathbf{i}} = 0\}$.

Proof. **Let us assume that:** There exist $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$ and an entry of $M_1(\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'})\ldots M_K(\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'})$ that does not belong to $\{0,1\}$.

Using (12), we know that there is $f \in \mathcal{F}$ and $n \in \mathbb{N}_N$ such that

$$\sum_{p\in\mathcal{P}(f)} \mathcal{T}^p(\mathbb{1})_n \geq 2.$$

As a consequence, there is $\mathbf{i}$ and $\mathbf{j} \in \mathcal{S}\cup\mathcal{S}'$ with $\mathbf{i} \neq \mathbf{j}$ and

$$\mathcal{T}^{\mathbf{p_i}}(\mathbf{h}^{\mathbf{i}})_n = \mathcal{T}^{\mathbf{p_j}}(\mathbf{h}^{\mathbf{j}})_n = 1.$$

Therefore,

$$\mathcal{A}P(\mathbf{h}^{\mathbf{i}}) = \mathcal{A}P(\mathbf{h}^{\mathbf{j}})$$

and the network is not identifiable. This proves the first statement.

Let us assume that: For any $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$, all the entries of $M_1(\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'})\ldots M_K(\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'})$ belong to $\{0,1\}$.

We immediately observe that (12) leads to the item 1 of the Proposition.

To prove the second item, we can easily check that $(P(\mathbf{h}^{\mathbf{i}}))_{\mathbf{i}\notin\mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}}$ forms a basis of $\{T \in \mathbb{R}^{S^K} \,|\, \forall \mathbf{i} \in \mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}, T_{\mathbf{i}} = 0\}$. We can also easily check using (12) and (9) that, for any $\mathbf{i} \notin \mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}$,

$$\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'}P(\mathbf{h}^{\mathbf{i}}) = \begin{cases} 0 & \text{, if } \mathbf{i} \notin \mathcal{S}\cup\mathcal{S}' \\ M_1(\mathbf{h}^{\mathbf{i}}_1)\ldots M_K(\mathbf{h}^{\mathbf{i}}_K) = 0 & \text{, if } \mathbf{i} \in \mathcal{S}\cup\mathcal{S}' \text{ and } \mathbf{p_i} \notin \mathcal{P} \end{cases}$$

As a consequence, $\{T \in \mathbb{R}^{S^K} \,|\, \forall \mathbf{i} \in \mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}, T_{\mathbf{i}} = 0\} \subset \mathrm{Ker}\,(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'})$.

To prove the converse inclusion, we observe that for any distinct $\mathbf{i}$ and $\mathbf{j} \in \mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}$, we have $\mathcal{D}_{\mathbf{i}}\cap\mathcal{D}_{\mathbf{j}} = \emptyset$. This implies that

$$\mathrm{rk}\,(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'}) \geq |\mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}| = S^K - \dim(\{T \in \mathbb{R}^{S^K} \,|\, \forall \mathbf{i} \in \mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}, T_{\mathbf{i}} = 0\}).$$

Finally, we deduce that $\dim(\mathrm{Ker}\,(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'})) \leq \dim(\{T \in \mathbb{R}^{S^K} \,|\, \forall \mathbf{i} \in \mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}, T_{\mathbf{i}} = 0\})$ and therefore

$$\mathrm{Ker}\,(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'}) = \{T \in \mathbb{R}^{S^K} \,|\, \forall \mathbf{i} \in \mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}, T_{\mathbf{i}} = 0\}.$$

□

Proposition 2 extends Proposition 8 (of [13]) by considering several possible supports. Said differently, the latter proposition corresponds to Proposition 2 when $\mathcal{M} = \{\mathbb{N}_S^K\}$.

The interest of the condition in Proposition 2 is that it can easily be computed by applying the network to dirac delta functions, when $|\mathcal{M}|$ is not too large. Notice that, beside the known examples in blind-deconvolution (i.e. when $K = 2$ and $|\mathcal{P}| = 1$) [1, 3], there are known (truly deep) convolutional networks that satisfy the condition of the first statement of Proposition 2. For instance, the convolutional network corresponding to the un-decimated Haar (wavelet)[1] transform is a tree and for any of its leaves $f \in \mathcal{F}$, $|\mathcal{P}(f)| = 1$. Moreover, the support of the kernel living on the edge e, of depth k, on this path is $\{0, 2^k\}$. The first condition of Proposition 2 therefore holds. However, it is clear that the necessary condition will be rarely satisfied.

Proposition 3. *If $|\mathcal{P}| = 1$ and, for any $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$, all the entries of $M_1(\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'}) \ldots M_K(\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'})$ belong to $\{0,1\}$, then $\mathrm{Ker}\,(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'})$ is the orthogonal complement of $\mathbb{R}^{S^K}_{\mathcal{S}\cup\mathcal{S}'}$ and $\mathcal{A}$ satisfies the deep-$\mathcal{M}$-NSP with constant $\gamma = 1$ (and any ε). Moreover, the deep-lower-RIP of $\mathcal{A}$ with regard to $\mathcal{M}$ is $\sigma_{\mathcal{M}} = \sqrt{N}$.*

Proof. The fact that, $\mathrm{Ker}\,(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'})$ is the orthogonal complement of $\mathbb{R}^{S^K}_{\mathcal{S}\cup\mathcal{S}'}$ is a direct consequence of Proposition 2 and the fact that, when $|\mathcal{P}| = 1$, $\mathbf{I}_{\mathcal{S}\cup\mathcal{S}'} = \mathcal{S}\cup\mathcal{S}'$. We then trivialy deduce that, for any $T' \in \mathrm{Ker}\,(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'})$, $\mathcal{P}_{\mathcal{S}\cup\mathcal{S}'}T' = 0$. A straightforward consequence is that $\mathcal{A}$ satisfies the deep-$\mathcal{M}$-NSP with constant $\gamma = 1$ (and any ε).

To calculate $\sigma_{\mathcal{M}}$, let us consider $\mathcal{S}, \mathcal{S}' \in \mathcal{M}$ and T in the orthogonal complement of $\mathrm{Ker}\,(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'})$. We express T under the form $T = \sum_{\mathbf{i}\in\mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}} T_{\mathbf{i}} P(\mathbf{h}^{\mathbf{i}})$, where $\mathbf{h}^{\mathbf{i}}$ is defined (13). Let us also remind that, applying Proposition 2, the supports of $\mathcal{A}P(\mathbf{h}^{\mathbf{i}})$ and $\mathcal{A}P(\mathbf{h}^{\mathbf{j}})$ are disjoint, when $\mathbf{i} \neq \mathbf{j}$. Let us finally add that, since $\mathcal{A}P(\mathbf{h}^{\mathbf{j}})$ is the matrix of a convolution with a Dirac mass, we have $|\mathcal{D}_{\mathbf{j}}| = N$. We finally have

$$\begin{aligned}\|\mathcal{A}T\|^2 &= \|\sum_{\mathbf{i}\in\mathbf{I}} T_{\mathbf{i}}\mathcal{A}P(\mathbf{h}^{\mathbf{i}})\|^2, \\ &= N\sum_{\mathbf{i}\in\mathbf{I}} T_{\mathbf{i}}^2 = N\|T\|^2,\end{aligned}$$

from which we deduce the value of $\sigma_{\mathcal{M}}$. □

In the sequel, we establish stability results for a convolutional network estimator. In order to do so, we consider a convolutional network of known structure $\mathcal{G}(\mathcal{E},\mathcal{N})$, $(\mathcal{S}_e)_{e\in\mathcal{E}}$ and $\mathcal{M}$. The convolutional network is defined by unknown parameters $\overline{\mathbf{h}} \in \mathbb{R}^{S\times K}$ satisfying a constraint $\mathrm{supp}\left(\overline{\mathbf{h}}\right) \subset \overline{\mathcal{S}}$ for an unknown support $\overline{\mathcal{S}} \in \mathcal{M}$. We consider the noisy situation where

$$X = M_1(\overline{\mathbf{h}}_1)\ldots M_K(\overline{\mathbf{h}}_K) + e,$$

with $\|e\| \leq \delta$ and an estimate $\mathbf{h}^* \in \mathbb{R}^{S\times K}$ such that

$$\|M_1(\mathbf{h}_1^*)\ldots M_K(\mathbf{h}_K^*) - X\| \leq \eta.$$

We say that two networks sharing the same structure and defined by $\mathbf{h}$ and $\mathbf{g} \in \mathbb{R}^{S\times K}$ are equivalent if and only if

$$\forall \mathbf{p} \in \mathcal{P}, \exists (\lambda_e)_{e\in\mathbf{p}} \in \mathbb{R}^{\mathbf{p}}, \text{ such that } \prod_{e\in\mathbf{p}} \lambda_e = 1 \text{ and } \forall e \in \mathbf{p}, \mathcal{T}_e(\mathbf{g}) = \lambda_e \mathcal{T}_e(\mathbf{h}).$$

We trivially observe that applying the networks defind by equivalent parameters lead to the same result. The equivalence class of $\mathbf{h} \in \mathbb{R}^{S\times K}$ is denoted by $\{\mathbf{h}\}$. For any $p \in [1, +\infty]$, we define

$$\delta_p(\{\mathbf{h}\}, \{\mathbf{g}\}) = \left(\sum_{\mathbf{p}\in\mathcal{P}} d_p([\mathbf{h}^{\mathbf{p}}], [\mathbf{g}^{\mathbf{p}}])^p\right)^{\frac{1}{p}},$$

where we remind that $\mathbf{h}^{\mathbf{p}}$ (resp $\mathbf{g}^{\mathbf{p}}$) denotes the restriction of $\mathbf{h}$ (resp $\mathbf{g}$) to the path $\mathbf{p}$ and d_p is defined in (2). Since d_p is a metric, we easily prove that δ_p is a metric between network classes.

Theorem 3. *If for any $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$, all the entries of $M_1(\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'}) \ldots M_K(\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'})$ belong to $\{0,1\}$ and if there exists $\varepsilon > 0$ such that for all $e \in \mathcal{E}$, $\|\mathcal{T}_e(\overline{\mathbf{h}})\|_\infty \geq \varepsilon$ then*

$$\delta_p(\{\mathbf{h}^*\}, \{\overline{\mathbf{h}}\}) \leq 7\frac{(KS)^{\frac{1}{p}}}{\sqrt{N}}\varepsilon^{1-K}\,(\delta + \eta).$$

[1] Un-decimated means computed with the "Algorithme à trous", [14], Section 5.5.2 and 6.3.2. The Haar wavelet is described in [14], Section 7.2.2, p. 247 and Example 7.7, p. 235

Proof. Let us consider a path $\mathbf{p} \in \mathcal{P}$, using (12), since all the entries of $M_1(\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'}) \ldots M_K(\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'})$ belong to $\{0, 1\}$, the restriction of the network to $\mathbf{p}$ satisfy the same property. Therefore, we can apply Proposition 3 and Theorem 2 to the restriction of the convolutional network to $\mathbf{p}$ and obtain for any $p \in [1, \infty]$

$$d_p([(\mathbf{h}^*)^{\mathbf{p}}], [\overline{\mathbf{h}}^{\mathbf{p}}]) \leq 7\frac{(KS)^{\frac{1}{p}}}{\sqrt{N}}\varepsilon^{1-K}(\delta^{\mathbf{p}} + \eta^{\mathbf{p}}),$$

where $\delta^{\mathbf{p}}$ and $\eta^{\mathbf{p}}$ are the restrictions of the errors on $\mathcal{D}^{\mathbf{p}}$. Finally, using item 1 of Proposition 2

$$\begin{aligned} \delta_p(\{\mathbf{h}^*\}, \{\overline{\mathbf{h}}\}) &\leq 7\frac{(KS)^{\frac{1}{p}}}{\sqrt{N}}\varepsilon^{1-K}\left(\sum_{\mathbf{p}\in\mathcal{P}}(\delta^{\mathbf{p}} + \eta^{\mathbf{p}})^p\right)^{\frac{1}{p}}, \\ &\leq 7\frac{(KS)^{\frac{1}{p}}}{\sqrt{N}}\varepsilon^{1-K}\,(\delta + \eta). \end{aligned}$$

□

References

[1] Arif Ahmed, Benjamin Recht, and Justin Romberg. Blind deconvolution using convex programming. *IEEE Transactions on Information Theory*, 60(3):1711–1732, 2014.

[2] Sanjeev Arora, Rong Ge, Ravindran Kannan, and Ankur Moitra. Computing a nonnegative matrix factorization–provably. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 145–162. ACM, 2012.

[3] Sohail Bahmani and Justin Romberg. Lifting for blind deconvolution in random mask imaging: Identifiability and convex relaxation. *SIAM Journal on Imaging Sciences*, 8(4):2203–2238, 2015.

[4] Emmanuel J Candes, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[5] Olivier Chabiron, François Malgouyres, Jean-Yves Tourneret, and Nicolas Dobigeon. Toward fast transform learning. *International Journal of Computer Vision*, pages 1–22, 2014.

[6] Olivier Chabiron, François Malgouyres, Herwig Wendt, and Jean-Yves Tourneret. Optimization of a fast transform structured as a convolutional tree. *preprint HAL*, (hal-01258514), 2016.

[7] Sunav Choudhary and Urbashi Mitra. Identifiability scaling laws in bilinear inverse problems. *arXiv preprint arXiv:1402.2637*, 2014.

[8] David L Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? 2003.

[9] Rodolphe Jenatton, Rémi Gribonval, and Francis Bach. Local stability and robustness of sparse dictionary learning in the presence of noise. arxiv, 2012.

[10] Hans Laurberg, Mads Græsbøll Christensen, Mark D Plumbley, Lars Kai Hansen, and Søren Holdt Jensen. Theorems on positive data: On the uniqueness of nmf. *Computational intelligence and neuroscience*, 2008, 2008.

[11] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[12] Xiaodong Li, Shuyang Ling, Thomas Strohmer, and Ke Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *CoRR*, abs/1606.04933, 2016.

[13] François Malgouyres and Joseph Landsberg. Stable recovery of the factors from a deep matrix product and application to convolutional network. *arXiv preprint arXiv:1703.08044*, 2017.

[14] Stéphane Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, Boston, 1998.

[15] Igor R. Shafarevich. *Basic algebraic geometry. 1*. Springer, 2010.