Multilinear compressive sensing and an application to convolutional linear networks

François Malgouyres[1] and Joseph Landsberg[2]

[1]Institut de Mathématiques de Toulouse, Université Paul Sabatier
and
[2]Department of Mathematics, Texas A& M University

November 2018

Plan

Statement, without technicality

- $f_{\mathbf{h}}$ a family of functions parameterized by **h** (e.g. linear networks)
- I, X matrix containing input-output pairs

Informal statement

Under a certain **condition** on the family f (e.g. on the topology of the network):
There exists C such that for η small and for any

$$\overline{\mathbf{h}}, \mathbf{h}^* \in \{\mathbf{h} | \; \|f_{\mathbf{h}}(I) - X\| \leq \eta\}$$

we have

$$d(\overline{\mathbf{h}}, \mathbf{h}^*) \leq C\,\eta$$

- If the condition is satisfied we have stably defined features

  $\Rightarrow$ **interpretable learning**

- If the data are known to be generated by a network

  $\Rightarrow$ **control of the risk**

Statement, without technicality

- $f_{\mathbf{h}}$ a family of functions parameterized by **h** (e.g. linear networks)
- I, X matrix containing input-output pairs

Informal statement

Under a certain **condition** on the family f (e.g. on the topology of the network):
There exists C such that for η small and for any

$$\overline{\mathbf{h}}, \mathbf{h}^* \in \{\mathbf{h} | \, \|f_{\mathbf{h}}(I) - X\| \leq \eta\}$$

we have

$$d(\overline{\mathbf{h}}, \mathbf{h}^*) \leq C\,\eta$$

- If the condition is satisfied we have stably defined features

  $\Rightarrow$ **interpretable learning**

- If the data are known to be generated by a network

  $\Rightarrow$ **control of the risk**

Deep structured linear networks

Problem formulation

Let $K \in \mathbb{N}^*$, $m_1 \dots m_{K+1} \in \mathbb{N}$, write $m_1 = m$, $m_{K+1} = n$. We assume that we know the matrix $X \in \mathbb{R}^{m \times n}$ which is (approximatively) the product of factors $X_k \in \mathbb{R}^{m_k \times m_{k+1}}$:

$$X = X_1 \cdots X_K.$$

We investigate models/constraints imposed on the factors X_k for which we can (up to obvious scale rearrangement) stably recover the factors X_k from X.

Deep structured linear networks

Structure of the factors

- For $k = 1 \dots K$, we know

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

- We know models

$$\mathcal{M} = (\mathcal{M}^L)_{L \in \mathbb{N}} \qquad \text{with ,} \qquad \mathcal{M}^L \subset \mathbb{R}^{K \times S}, \forall L.$$

- Assume there exists $\overline{L}$, L^* and $(\overline{\mathbf{h}}_k)_{k=1..K} \in \mathcal{M}^{\overline{L}}$ and $(\mathbf{h}_k^*)_{k=1..K} \in \mathcal{M}^{L^*}$ such that

$$\|M_1(\overline{\mathbf{h}}_1) \cdots M_K(\overline{\mathbf{h}}_K) - X\| \leq \delta,$$

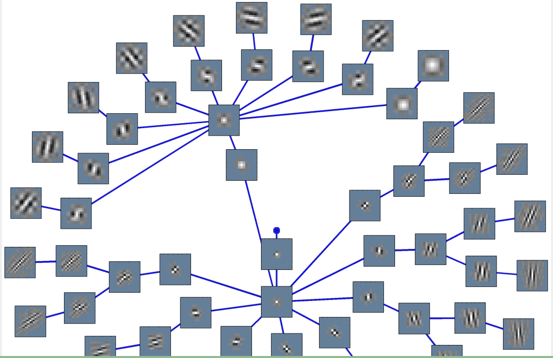
$$\|M_1(\mathbf{h}_1^*) \cdots M_K(\mathbf{h}_K^*) - X\| \leq \eta,$$

Is $(\overline{\mathbf{h}}_k)_{k=1..K}$ close to $(\mathbf{h}_k^*)_{k=1..K}$?

Examples

- $K = 1$: Compressed sensing problem: Recovering $\mathbf{h}_1$ from $M_1(\mathbf{h}_1)$ is linear inverse problem.
- $K = 2$:
  - **Dictionary learning:** $M_1(\mathbf{h}_1)$ is a dictionary of atoms, $M_2(\mathbf{h}_2)$ is sparse
  - **Non-negative matrix factorization:** $M_1(\mathbf{h}_1) \geq 0$ and $M_2(\mathbf{h}_2) \geq 0$
  - **Low rank approximation:** $M_1(\mathbf{h}_1)$ is rectangular "vertical" ($m_1 \gg m_2$), $M_2(\mathbf{h}_2)$ is rectangular "horizontal" ($m_2 \ll m_3$).
  - **Phase recovery:** $M_1(\mathbf{h}_1) = diag(F\mathbf{h}_1)$, $M_2(\mathbf{h}_2) = (F\mathbf{h}_2)^*$, with F the Fourier matrix and $\mathbf{h}_1 = \mathbf{h}_2$.
  - **Blind deconvolution:** $M_1(\mathbf{h}_1)$ is circulant, $M_2(\mathbf{h}_2)$ is a signal
  - **Blind-demixing, self-calibration, Internet of things**...

- *K* large :
  - Fast Fourier, Discrete Cosine, Discrete Wavelet, Jacobi eigenvalue Algorithm
  - Tsiligkaridis, Hero, Zhou: **Kronecker graphical lasso** (IEEE SP 2013)
  - Lyu, Wang: **Multi-layer NMF** (NIPS'13)
  - Kondor, Tevena, Garg: **Multiresolution Matrix fatorization** (ICML 2014)
  - Chabiron, Malgouyres, Wendt, Tourneret: **Fast Transform Learning** (IJCV, 2015)
  - Le Magoarou, Gribonval: **Faust** (IEEE STSP, 2016)
  - Rusu, Thomson: **Transforms based on Householder reflectors** (IEEE SP 2016) and **Givens rotations** (IEEE SP 2017)
  - Sulam, Papyan, Romano, Elad : **Multi-layer Convolutional Sparse Coding** (IEEE SP 2018)



Link with Deep learning

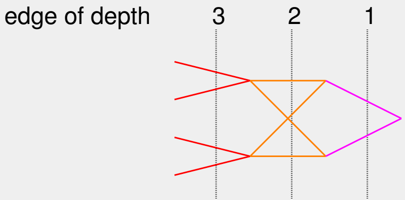


Figure: Deep network

$$\mathcal{N}(\mathbf{h}, I) = U_1 M'_1(\mathbf{h}_1) U_2 M'_2(\mathbf{h}_2) U_3 M'_3(\mathbf{h}_3) I$$

- $M'_k(\mathbf{h}_k)$: is a linear operator, depending linearly on $\mathbf{h}_k$
  - Feed-forward : $M'_3(\mathbf{h}_3) = \begin{pmatrix} \mathbf{h}_{3,1} & \mathbf{h}_{3,2} & 0 & 0 \\ 0 & 0 & \mathbf{h}_{3,3} & \mathbf{h}_{3,4} \end{pmatrix}$
  - Convolutional : $M'_3(\mathbf{h}_3) = \begin{pmatrix} C_1(\mathbf{h}_3) & C_2(\mathbf{h}_3) & 0 & 0 \\ 0 & 0 & C_3(\mathbf{h}_3) & C_4(\mathbf{h}_3) \end{pmatrix}$

    where $C_i(.)$ convolution+sampling matrices.

Link with Deep learning

- With ReLU : $U_k : \mathbb{R}^{n_k \times L} \longmapsto \mathbb{R}^{n_k \times L}$ (where n_k is the size of the layer k) is such that :

$$(U_k M)_{n,l} = a_k(\mathbf{h})_{n,l} M_{n,l} \qquad \text{, with } a_k(\mathbf{h}) \in \{0,1\}^{n_k \times L}.$$

and

$$a_k(\mathbf{h})_{n,l} = \begin{cases} 1 & \text{, if } \left(M'_k(\mathbf{h}_k) U_{k+1} M'_{k+1}(\mathbf{h}_{k+1}) \cdots U_K M'_K(\mathbf{h}_K) X \right)_{n,l} \geq 0 \\ 0 & \text{, otherwise} \end{cases}$$

The function

$$\begin{array}{rcl} a_k : \mathbb{R}^{K \times S} & \longrightarrow & \{0,1\}^{n_k \times L} \\ \mathbf{h} & \longmapsto & a_k(\mathbf{h}) \end{array}$$

is piecewise constant.

As a function of h, the neural network is a piecewise structured linear network

Statement, without technicality

- $f_{\mathbf{h}}$ a family of functions parameterized by **h** (e.g. linear networks)
- I, X matrix containing input-output pairs

Informal statement

Under a certain **condition** on the family f (e.g. on the topology of the network):
There exists C such that for η small and for any

$$\overline{\mathbf{h}}, \mathbf{h}^* \in \{\mathbf{h} | \; \|f_{\mathbf{h}}(I) - X\| \leq \eta\}$$

we have

$$d(\overline{\mathbf{h}}, \mathbf{h}^*) \leq C\,\eta$$

- If the condition is satisfied we have stably defined features

  $\Rightarrow$ **interpretable learning**

- If the data are known to be generated by a network

  $\Rightarrow$ **control of the risk**

Notations

- $\mathbb{N}_k = \{1, \dots, k\}$
- $\mathbf{h} \in \mathbb{R}^{K \times S}$, $\mathbf{h}_k \in \mathbb{R}^S$, $\mathbf{h}_{k,\mathbf{i}_k} \in \mathbb{R}$

Notations

- $\mathbb{N}_k = \{1, \ldots, k\}$
- $\mathbf{h} \in \mathbb{R}^{K \times S}$, $\mathbf{h}_k \in \mathbb{R}^S$, $\mathbf{h}_{k,\mathbf{i}_k} \in \mathbb{R}$
- $\mathbb{R}_*^{K \times S} = \{\mathbf{h} \in \mathbb{R}^{K \times S}, \forall k \in \mathbb{N}_K, \|\mathbf{h}_k\| \neq 0\}$

Notations

- $\mathbb{N}_k = \{1, \dots, k\}$
- $\mathbf{h} \in \mathbb{R}^{K \times S}$, $\mathbf{h}_k \in \mathbb{R}^S$, $\mathbf{h}_{k,\mathbf{i}_k} \in \mathbb{R}$
- $\mathbb{R}_*^{K \times S} = \{\mathbf{h} \in \mathbb{R}^{K \times S}, \forall k \in \mathbb{N}_K, \|\mathbf{h}_k\| \neq 0\}$
- For $\mathbf{h}$ and $\mathbf{g} \in \mathbb{R}_*^{K \times S}$, $\mathbf{h} \sim \mathbf{g}$ if and only if there exists $(\lambda_k)_{k \in \mathbb{N}_K} \in \mathbb{R}^K$ such that

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

  We say $\mathbf{g} \in [\mathbf{h}]$.

Remark

Since for any $\mathbf{g} \in [\overline{\mathbf{h}}]$

$$M_1(\overline{\mathbf{h}}_1) \dots M_K(\overline{\mathbf{h}}_K) = M_1(\mathbf{g}_1) \dots M_K(\mathbf{g}_K)$$

Recovering $[\overline{\mathbf{h}}]$ is the best we can hope for.

- Tensors $T \in \mathbb{R}^{\overbrace{S \times \ldots \times S}^{K \text{ times}}} = \mathbb{R}^{S^K}$

- Tensors $T \in \mathbb{R}^{\overbrace{S \times \ldots \times S}^{K \text{ times}}} = \mathbb{R}^{S^K}$
- Tensor value $T_{i_1,\ldots,i_K}$ or $T_{\mathbf{i}}$, for $\mathbf{i} = (i_1, \ldots, i_K) \in \mathbb{N}_S^K$

- Tensors $T \in \mathbb{R}^{\overbrace{S \times \ldots \times S}^{K \text{ times}}} = \mathbb{R}^{S^K}$
- Tensor value $T_{i_1,\ldots,i_K}$ or $T_{\mathbf{i}}$, for $\mathbf{i} = (i_1, \ldots, i_K) \in \mathbb{N}_S^K$
- $T \in \mathbb{R}^{S^K}$ is of rank 1 if and only if there exists $\mathbf{h} \in \mathbb{R}^{K \times S}$ s.t.:

$$T_{\mathbf{i}} = \mathbf{h}_{1,i_1} \ldots \mathbf{h}_{K,i_K} \qquad , \forall \mathbf{i} \in \mathbb{N}_S^K.$$

We say $T \in \Sigma_1$.

- Tensors $T \in \mathbb{R}^{\overbrace{S \times \ldots \times S}^{K \text{ times}}} = \mathbb{R}^{S^K}$
- Tensor value $T_{i_1,\ldots,i_K}$ or $T_{\mathbf{i}}$, for $\mathbf{i} = (i_1, \ldots, i_K) \in \mathbb{N}_S^K$
- $T \in \mathbb{R}^{S^K}$ is of rank 1 if and only if there exists $\mathbf{h} \in \mathbb{R}^{K \times S}$ s.t.:

$$T_{\mathbf{i}} = \mathbf{h}_{1,i_1} \ldots \mathbf{h}_{K,i_K} \qquad , \forall \mathbf{i} \in \mathbb{N}_S^K.$$

  We say $T \in \Sigma_1$.
- We call $\text{rk}(T)$ the smallest $r \in \mathbb{N}$, there exists $T_1, \ldots, T_r \in \Sigma_1$, s. t.

$$T = T_1 + \ldots + T_r.$$

  We denote $\Sigma_r = \{T \in \mathbb{R}^{S^K} | \text{rk}(T) \leq r\}$.

Facts on Segre embedding and tensors

- $\Sigma_{1,*}$ is a smooth (i.e. C^∞) manifold of dimension $K(S-1)+1$

Facts on Segre embedding and tensors

- $\Sigma_{1,*}$ is a smooth (i.e. C^∞) manifold of dimension $K(S-1)+1$
- Geometry of Σ_2: There exists a closed set $C \subset \Sigma_2$, whose Haudorff measure of dimension $2K(S-1)+2$ (resp. $4(S-1)$) is 0, such that $\Sigma_2 \setminus C$ is a smooth manifold of dimension $2K(S-1)+2$ when $K \geq 3$ (resp. $4(S-1)$, when $K=2$).

Facts on Segre embedding and tensors

- $\Sigma_{1,*}$ is a smooth (i.e. C^∞) manifold of dimension $K(S-1)+1$
- Geometry of Σ_2: There exists a closed set $C \subset \Sigma_2$, whose Haudorff measure of dimension $2K(S-1)+2$ (resp. $4(S-1)$) is 0, such that $\Sigma_2 \setminus C$ is a smooth manifold of dimension $2K(S-1)+2$ when $K \geq 3$ (resp. $4(S-1)$, when $K=2$).
- Segre embedding: Parameterize $\Sigma_1 \subset \mathbb{R}^{S^K}$ by the map

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

Facts on Segre embedding and tensors

- $\Sigma_{1,*}$ is a smooth (i.e. C^∞) manifold of dimension $K(S-1)+1$
- Geometry of Σ_2: There exists a closed set $C \subset \Sigma_2$, whose Haudorff measure of dimension $2K(S-1)+2$ (resp. $4(S-1)$) is 0, such that $\Sigma_2 \setminus C$ is a smooth manifold of dimension $2K(S-1)+2$ when $K \geq 3$ (resp. $4(S-1)$, when $K=2$).
- Segre embedding: Parameterize $\Sigma_1 \subset \mathbb{R}^{S^K}$ by the map

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

Remark

Since for any $\mathbf{g} \in [\overline{\mathbf{h}}]$

$$P(\overline{\mathbf{h}}) = P(\mathbf{g})$$

Recovering $[\overline{\mathbf{h}}]$ from $P(\overline{\mathbf{h}})$ is the best we can hope for.
Recovering $[\overline{\mathbf{h}}]$ from $P(\overline{\mathbf{h}})$ is easy. (By extracting lines in $P(\overline{\mathbf{h}})$.)

Theorem: **Stability of $[\mathbf{h}]$ from $P(\mathbf{h})$**

Let $\mathbf{h}$ and $\mathbf{g} \in \mathbb{R}_*^{K\times S}$ be such that

$$\|P(\mathbf{g}) - P(\mathbf{h})\|_\infty \leq \frac{1}{2} \max\left(\|P(\mathbf{h})\|_\infty, \|P(\mathbf{g})\|_\infty\right).$$

We have for p and $q \in [1,\infty]$,

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

In the theorem, we use the metric

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

where $p > 1$ and

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

Proposition : Sharpness of the bound

There exists $\mathbf{h}$ and $\mathbf{g} \in \mathbb{R}_*^{K\times S}$ such that $\|P(\mathbf{g})\|_\infty \leq \|P(\mathbf{h})\|_\infty$, $\|P(\mathbf{g}) - P(\mathbf{h})\|_\infty \leq \frac{1}{2}\ \|P(\mathbf{h})\|_\infty$ and

$$7(KS)^{\frac{1}{p}} \|P(\mathbf{h})\|_\infty^{\frac{1}{K}-1} \|P(\mathbf{h}) - P(\mathbf{g})\|_q \leq C_q\ d_p([\mathbf{h}],[\mathbf{g}]),$$

where

$$C_q = \begin{cases} 28(KS)^{\frac{1}{q}} & \text{, if } q < +\infty, \\ 28 & \text{, if } q = +\infty. \end{cases}$$

Theorem: Lipschitz continuity of P

We have for any $q \in [1, \infty]$ and any $\mathbf{h}$ and $\mathbf{g} \in \mathbb{R}_*^{K \times S}$,

$$\|P(\mathbf{h}) - P(\mathbf{g})\|_q \leq S^{\frac{K-1}{q}} K^{1-\frac{1}{q}} \max\left(\|P(\mathbf{h})\|_\infty^{1-\frac{1}{K}}, \|P(\mathbf{g})\|_\infty^{1-\frac{1}{K}}\right) d_q([\mathbf{h}], [\mathbf{g}]). \quad (1)$$

The upper bounds in the theorem is tight up to at most a factor K.

The Tensorial Lifting

When $K = 2$: $M_1(\mathbf{h}_1)M_2(\mathbf{h}_2)$ has the form

$$\begin{pmatrix} p_{1,1}(\mathbf{h}_1) & p_{1,2}(\mathbf{h}_1) & \cdots & p_{1,m_2}(\mathbf{h}_1) \\ p_{2,1}(\mathbf{h}_1) & p_{2,2}(\mathbf{h}_1) & \cdots & p_{2,m_2}(\mathbf{h}_1) \\ \vdots & \vdots & \ddots & \vdots \\ p_{m,1}(\mathbf{h}_1) & p_{m,2}(\mathbf{h}_1) & \cdots & p_{m,m_2}(\mathbf{h}_1) \end{pmatrix} \begin{pmatrix} q_{1,1}(\mathbf{h}_2) & \cdots & q_{1,n}(\mathbf{h}_2) \\ q_{2,1}(\mathbf{h}_2) & \cdots & q_{2,n}(\mathbf{h}_2) \\ \vdots & \ddots & \vdots \\ q_{m_2,1}(\mathbf{h}_2) & \cdots & q_{m_2,n}(\mathbf{h}_2) \end{pmatrix}$$

so for i and j

$$(M_1(\mathbf{h}_1)M_2(\mathbf{h}_2))_{i,j} = \begin{pmatrix} p_{i,1}(\mathbf{h}_1) & p_{i,2}(\mathbf{h}_1) & \cdots & p_{i,m_2}(\mathbf{h}_1) \end{pmatrix} \begin{pmatrix} q_{1,j}(\mathbf{h}_2) \\ q_{2,j}(\mathbf{h}_2) \\ \vdots \\ q_{m_2,j}(\mathbf{h}_2) \end{pmatrix}$$

is a polynomial whose monomial are of the form $\mathbf{h}_{1,\mathbf{i}_1}\mathbf{h}_{2,\mathbf{i}_2}$
Ex: $(2\mathbf{h}_{1,3} + 4\mathbf{h}_{1,7})(\mathbf{h}_{2,1} + 5\mathbf{h}_{2,4}) = 2\mathbf{h}_{1,3}\mathbf{h}_{2,1} + 10\mathbf{h}_{1,3}\mathbf{h}_{2,4} + 4\mathbf{h}_{1,7}\mathbf{h}_{2,1} + 20\mathbf{h}_{1,7}\mathbf{h}_{2,4}$

The Tensorial Lifting

Theorem

There exists a unique linear map

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

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

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

- Changing $M_1, M_2, \ldots, M_K$ only modifies $\mathcal{A}$
- The properties of
  - $M_1(\mathbf{h}_1) M_2(\mathbf{h}_2) \cdots M_K(\mathbf{h}_K)$
  - $\mathbf{h} \longmapsto \|M_1(\mathbf{h}_1) M_2(\mathbf{h}_2) \cdots M_K(\mathbf{h}_K) - X\|^2$

  relate to the geometry of $\mathcal{A}$ and Σ_1 (or Σ_2).

- When $K = 1$ and X is vectorized, we simply have $\mathcal{A} = M_1$.
- In most reasonable cases, $\mathcal{A}$ is sparse.
- We can compute $\mathcal{A}P(\mathbf{h})$, whatever $\mathbf{h} \in \mathbb{R}^{K \times S}$, using

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

Proposition

If we consider R independent random collections of vectors $\mathbf{h}^r$, with $r = 1 \dots R$, according to the normal distribution in $\mathbb{R}^{K \times S}$, we have (with probability 1)

$$\dim(\mathrm{Span}((\mathcal{A}P(\mathbf{h}^r))_{r=1..R})) = \begin{cases} R & \text{, if } R \leq \mathrm{rk}(\mathcal{A}) \\ \mathrm{rk}(\mathcal{A}) & \text{, otherwise.} \end{cases} \quad (2)$$

This can be used to compute $\mathrm{rk}(\mathcal{A})$**.**

Identifiability – noise free case

We assume there is $\overline{\mathbf{h}} \in \mathcal{M}^{\overline{L}}$ and

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

There is $\mathbf{h}^* \in \mathcal{M}^{L^*}$

$$X = M_1(\mathbf{h}_1^*)\dots M_K(\mathbf{h}_K^*). \qquad (3)$$

Definition

$[\overline{\mathbf{h}}]$ *is identifiable* iif the elements of $[\overline{\mathbf{h}}]$ are the only solutions of (3).

Theorem : **Necessary and sufficient conditions of identifiability**

1. For any $\overline{L}$ and $\overline{\mathbf{h}} \in \mathcal{M}^{\overline{L}}$: $[\overline{\mathbf{h}}]$ is identifiable if and only if for any $L \in \mathbb{N}$

$$\left(P(\overline{\mathbf{h}}) + \mathrm{Ker}(\mathcal{A})\right) \cap P(\mathcal{M}^L) \subset \{P(\overline{\mathbf{h}})\}.$$

2. $\mathcal{M}$ is identifiable if and only if for any L and $L' \in \mathbb{N}$

$$\mathrm{Ker}(\mathcal{A}) \cap \left(P(\mathcal{M}^L) - P(\mathcal{M}^{L'})\right) \subset \{0\}. \qquad (4)$$

Definition: $\dim_{\min}(\mathcal{M})$

Let $\dim_{\min}(\mathcal{M}) \in \mathbb{N}$ be the largest dimension of the sub-vector spaces V of $\mathbb{R}^{S^K}$ such that there exists a neighborhood $\mathcal{O}$ of the origin, $L \in \mathbb{N}$ and $L' \in \mathbb{N}$ such that

$$(V \cap \mathcal{O}) \subset \left(P(\mathcal{M}^L) - P(\mathcal{M}^{L'})\right).$$

Example : When $K > 1$ and $\mathcal{M} = \mathbb{R}^{K\times S}$, we have $\dim_{\min}(\mathcal{M}) = 2S - 1$.
We always have $\dim_{\min}(\mathcal{M}) \leq 2S - 1$.

Theorem : **Necessary condition of identifiability**

If $\mathrm{rk}(\mathcal{A}) < \dim_{\min}(\mathcal{M})$, then $\mathcal{M}$ is not identifiable.

We assume $P(\mathcal{M})$ is Zariski closed and invariant under rescaling.

Definition: $\dim_{\max}(\mathcal{M})$

$$\dim_{\max}(\mathcal{M}) = \max_{L,L'} \dim \overline{\{sx + ty \mid x \in P(\mathcal{M}^{L'}),\ y \in P(\mathcal{M}^{L}),\ s,t \in \mathbb{R}\}}^{Zar}.$$

We have : $\dim_{\max}(\mathcal{M}) \leq 2\max_L \left(\dim P(\mathcal{M}^L)\right)$.
Example : $\dim_{\max}\left(\mathbb{R}^{K\times S}\right) \leq 2\dim(\Sigma_1) = 2K(S-1)+2$

Theorem : **Almost surely sufficient condition for Identifiability**

For almost every $\mathcal{A}$ such that

$$\mathrm{rk}(\mathcal{A}) \geq \dim_{\max}(\mathcal{M})$$

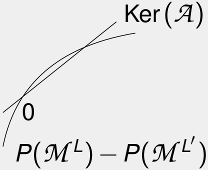
every $\overline{\mathbf{h}} \in \mathbb{R}^{K\times S}$ is identifiable.

Stable recovery – Interpretability

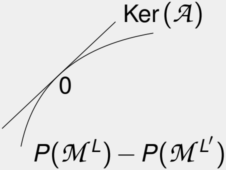
Deep-Null Space Property

Let $\gamma > 0$ and $\rho > 0$, we say that $\text{Ker}(\mathcal{A})$ satisfies the *deep-Null Space Property (deep-NSP)* with respect to the collection of models $\mathcal{M}$ with constants (γ, ρ) if for any L and $L' \in \mathbb{N}$, any $T \in P(\mathcal{M}^L) - P(\mathcal{M}^{L'})$ satisfying $\|\mathcal{A}T\| \leq \rho$ and any $T' \in \text{Ker}(\mathcal{A})$, we have

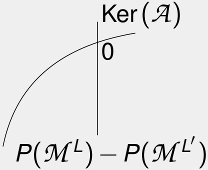
$$\|T\| \leq \gamma \|T - T'\|. \tag{5}$$



NO (deep-NSP)



NO (deep-NSP)



(deep-NSP)

$$\|M_1(\overline{\mathbf{h}}_1)\cdots M_K(\overline{\mathbf{h}}_K)-X\|\leq \delta,$$

and

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

for δ and η small.

Theorem : **Sufficient condition for interpretability**

Assume $\mathrm{Ker}(\mathcal{A})$ satisfies the deep-NSP with respect to the collection of models $\mathcal{M}$ and with the constant (γ,ρ). If $\delta+\eta\leq\rho$, we have

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

where σ_{min} is the smallest non-zero singular value of $\mathcal{A}$. Moreover, if $\overline{\mathbf{h}}\in\mathbb{R}_*^{K\times S}$ and $\frac{\gamma}{\sigma_{min}}(\delta+\eta)\leq\frac{1}{2}\max\left(\|P(\overline{\mathbf{h}})\|_\infty,\|P(\mathbf{h}^*)\|_\infty\right)$ then

$$d_p([\mathbf{h}^*],[\overline{\mathbf{h}}])\leq \frac{7(KS)^{\frac{1}{p}}\gamma}{\sigma_{min}}\min\left(\|P(\overline{\mathbf{h}})\|_\infty^{\frac{1}{K}-1},\|P(\mathbf{h}^*)\|_\infty^{\frac{1}{K}-1}\right)(\delta+\eta). \tag{6}$$

Theorem : **Necessary condition for interpretability**

Assume the interpretability holds: There exists C and $\delta > 0$ such that for any $\overline{L} \in \mathbb{N}$, $\overline{\mathbf{h}} \in \mathcal{M}^{\overline{L}}$, any $X = \mathcal{A}P(\overline{\mathbf{h}}) + e$, with $\|e\| \leq \delta$, any $L^* \in \mathbb{N}$ and any $\mathbf{h}^* \in \mathcal{M}^{L^*}$ such that

$$\|\mathcal{A}P(\mathbf{h}^*) - X\|^2 \leq \|e\|$$

we have

$$d_2([\mathbf{h}^*], [\overline{\mathbf{h}}]) \leq C \min\left(\|P(\overline{\mathbf{h}})\|_\infty^{\frac{1}{K}-1}, \|P(\mathbf{h}^*)\|_\infty^{\frac{1}{K}-1} \right) \|e\|.$$

Then, Ker$(\mathcal{A})$ satisfies the deep-NSP with respect to the collection of models $\mathcal{M}$ with constants

$$(\gamma, \rho) = (CS^{\frac{K-1}{2}} \sqrt{K}\, \sigma_{max}, \delta)$$

where σ_{max} is the spectral radius of $\mathcal{A}$.

Plan

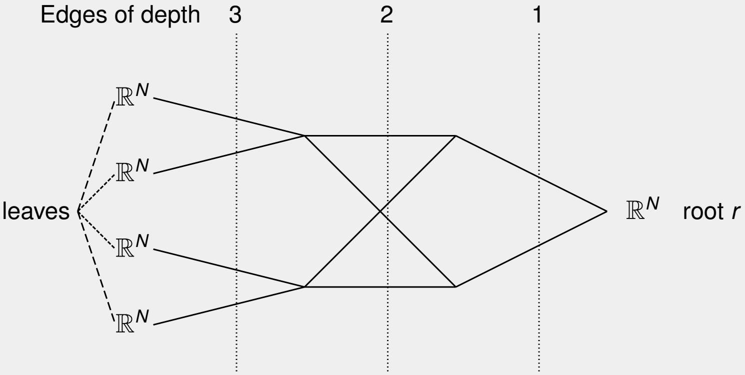


Figure: Example of the **convolutional linear network**. To every edge is attached a convolution kernel. The network does not involve non-linearities or sampling.

$$X = M_1(\mathbf{h}_1)M_2(\mathbf{h}_2)M_3(\mathbf{h}_3) = [X_1 X_2 X_3 X_4] \in \mathbb{R}^{N \times N|\mathcal{F}|}$$

$X_1, ..., X_4$ are convolution matrix

Proposition : Necessary condition of identifiability of a network

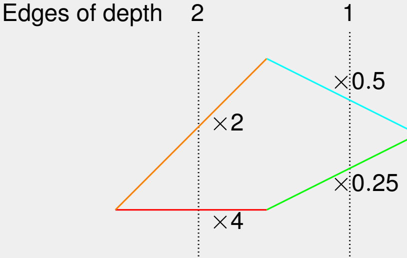
If some of the entries of $M_1(\mathbb{1})\ldots M_K(\mathbb{1})$ do not belong to $\{0,1\}$:

$$\mathbb{R}^{K\times S} \text{ is not identifiable.}$$

The condition "all the entries of $M_1(\mathbb{1})\ldots M_K(\mathbb{1})$ belong to $\{0,1\}$" can be computed by applying the network $|\mathcal{F}|$ times to a dirac delta function.

Proposition

If the network is a branch and all the entries of $M_1(\mathbb{1})\cdots M_K(\mathbb{1})$ belong to $\{0,1\}$, then $\operatorname{Ker}(\mathcal{A}) = \{0\}$ *and* $\operatorname{Ker}(\mathcal{A})$ *satisfies the deep-NSP with respect to any model collection $\mathcal{M}$ with constant $(\gamma,\rho) = (1,+\infty)$. Moreover, we have $\sigma_{min} = \sqrt{N}$.*



h and **g** $\in \mathbb{R}^{K\times S}$ are equivalent if and only if

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

The equivalence class of $\mathbf{h} \in \mathbb{R}^{K\times S}$ 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 $\mathbf{h}^{\mathbf{p}}$ (resp $\mathbf{g}^{\mathbf{p}}$) is the restriction of **h** (resp **g**) to the path **p**.

$$\|M_1(\overline{\mathbf{h}}_1)\cdots M_K(\overline{\mathbf{h}}_K) - X\| \leq \delta,$$

and

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

for δ and η small.

Theorem : **Sufficient condition of interpretability**

If all the entries of $M_1(\mathbb{1})\cdots M_K(\mathbb{1})$ belong to $\{0,1\}$, if there exists $\varepsilon > 0$ such that for all $e \in \mathcal{E}$, $\|\mathcal{T}_e(\overline{\mathbf{h}})\|_\infty \geq \varepsilon$, and if $\delta + \eta \leq \frac{\sqrt{N}\varepsilon^K}{2}$ then

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

where $S' = \max_{e \in \mathcal{E}} |\mathcal{S}_e|$.

Rks :

- The condition "$M_1(\mathbb{1})\cdots M_K(\mathbb{1})$ belong to $\{0,1\}$" is not satisfied by most network structure encountered in practice.
- The action of the activation function favors interpretability.

Thank you for your attention !

paper available on

www.math.univ-toulouse.fr/$\sim$fmalgouy/

or

google: F. Malgouyres