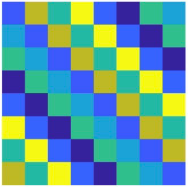
Computational Principles for High-dim Data Analysis

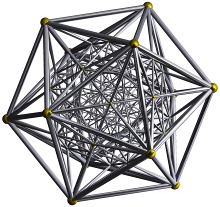
(Lecture Nineteen)

Yi Ma

EECS Department, UC Berkeley

November 4, 2021





Structured Nonlinear Low-Dimensional Models
Sparsity in Convolution and Deconvolution

1. Convolution for Image Modeling
2. Convolution and Circulant Matrix
3. The Blind Short-and-Sparse Deconvolution

"The mathematical sciences particularly exhibit order, symmetry, and limitations; and these are the greatest forms of the beautiful."
– *Aristotle*, Metaphysica

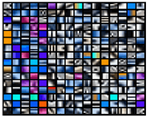
Importance of Mathematical Modeling

If you formulate a problem correctly,
you are more than halfway solved it!

Sparsity in Appearance of Image Patches

Patch-level image modeling
(e.g. denoising or super-resolution)
with a sparsifying dictionary:

$$I_{\text{patch}} = \underset{\text{dictionary}}{\boldsymbol{A}} \times \underset{\text{sparse}}{\boldsymbol{x}} + \underset{\text{noise}}{\boldsymbol{z}}. \quad (1)$$



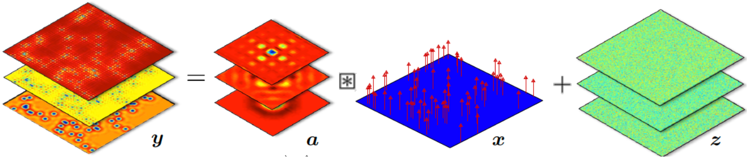
Dictionary learning: the motifs or atoms of the dictionary are unknown:

$$\underset{\textbf{data}}{\boldsymbol{Y}} = \underset{\textbf{dictionary}}{\boldsymbol{A}} \; \underset{\textbf{sparse}}{\boldsymbol{X}}. \quad (2)$$

- Band-limited signals: $\boldsymbol{A} = \boldsymbol{F}$, the Fourier transform (JPEG);
- Piecewise smooth: $\boldsymbol{A} = \boldsymbol{W}$, the wavelet transforms (JPEG2000);
- For natural images $\boldsymbol{A}$ can be **learned** from patch samples $\boldsymbol{Y}$.

Sparsity in Occurrence of Patch Motif(s)

The same motif $\boldsymbol{a} \in \boldsymbol{A}$ occurs at a sparse number of locations $(i_1, j_1), \ldots, (i_k, j_k)$ in space:



The overall observation $\boldsymbol{y}$ can be modeled as a superposition of translated versions of the motif $\boldsymbol{a}$, one for each of locations (i_ℓ, j_ℓ):

$$\underbrace{\boldsymbol{y}(i,j,e)}_{\textbf{data}} = \sum_{\ell=1}^{k} \underbrace{\boldsymbol{a}(i - i_\ell, j - j_\ell, e)}_{\textbf{translated motif}} \quad + \quad \underbrace{\boldsymbol{z}(i,j,e)}_{\textbf{noise}}. \tag{3}$$

One could generalize this to multiple motifs.

Modeling Translational Occurrence by Convolution

Define a two-dimensional sparse signal $\boldsymbol{x} \in \mathbb{R}^{w \times h}$, which takes on value 1 at locations (i_ℓ, j_ℓ) and zero elsewhere:

$$\boldsymbol{y}(\cdot,\cdot,e) = \boldsymbol{a}(\cdot,\cdot,e) * \boldsymbol{x} \quad + \quad \boldsymbol{z}(\cdot,\cdot,e). \tag{4}$$

Combining these equations for all energy levels e, the observed data $\boldsymbol{y}$ is a convolution of the motif $\boldsymbol{a}$ and a field $\boldsymbol{x}$ of sparse spikes:

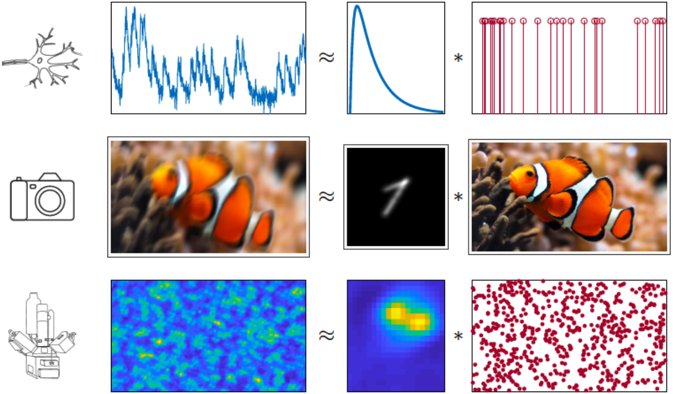
$$\underbrace{\boldsymbol{y}}_{\text{data}} = \underbrace{\boldsymbol{a}}_{\text{motif}} * \underbrace{\boldsymbol{x}}_{\text{sparse spikes}} + \underbrace{\boldsymbol{z}}_{\text{noise}}, \tag{5}$$

$\boldsymbol{x}$ could also take different values other than 1 to model the intensity or weight of the motif at each location.

The sparse occurrence/convolution model does generalize to other transformation groups, such as rotation etc.

Modeling Translational Occurrence by Convolution

Examples: Neuron, Camera, and Microscopy



The sparse occurrence/convolution model does generalize to other transformation groups, such as rotation etc.

Background: Convolution and Circulant Matrix

Given a vector $\boldsymbol{a} = [a_0, a_1, \dots, a_{n-1}]^* \in \mathbb{R}^n$, we may arrange all its circularly shifted versions in a circulant matrix form as

$$\boldsymbol{A} \quad \doteq \quad \mathrm{circ}(\mathbf{a}) = \begin{bmatrix} a_0 & a_{n-1} & \dots & a_2 & a_1 \\ a_1 & a_0 & a_{n-1} & \cdots & a_2 \\ \vdots & a_1 & a_0 & \ddots & \vdots \\ a_{n-2} & \vdots & \ddots & \ddots & a_{n-1} \\ a_{n-1} & a_{n-2} & \dots & a_1 & a_0 \end{bmatrix} \quad \in \mathbb{R}^{n\times n}. \quad (6)$$

It is easy to see that the multiplication of such a circulant matrix $\boldsymbol{A}$ with a vector $\boldsymbol{x}$ gives a (circular) convolution $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{a} \circledast \boldsymbol{x}$ with:

$$(\boldsymbol{a} \circledast \boldsymbol{x})_i = \sum_{j=0}^{n-1} x_j a_{i+n-j \bmod n}. \quad (7)$$

Fact: all circulant matrices share the same set of eigenvectors!

Background: Eigenvectors of Circulant Matrices

Let $\mathfrak{i} = \sqrt{-1}$ and $\omega_n := \exp(-\frac{2\pi\mathfrak{i}}{n})$ and we define the matrix:

$$\boldsymbol{F}_n \doteq \frac{1}{\sqrt{n}} \begin{bmatrix} \omega_n^0 & \omega_n^0 & \cdots & \omega_n^0 & \omega_n^0 \\ \omega_n^0 & \omega_n^1 & \cdots & \omega_n^{n-2} & \omega_n^{n-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \omega_n^0 & \omega_n^{n-2} & \cdots & \omega_n^{(n-2)^2} & \omega_n^{(n-2)(n-1)} \\ \omega_n^0 & \omega_n^{n-1} & \cdots & \omega_n^{(n-2)(n-1)} & \omega_n^{(n-1)^2} \end{bmatrix} \in \mathbb{C}^{n\times n}. \tag{8}$$

$\boldsymbol{F}_n$ is known as the *discrete Fourier transform* (DFT), with $\boldsymbol{F}_n\boldsymbol{F}_n^* = \boldsymbol{I}$.

Theorem (Eigenvectors of Circulant Matrix)

An $n \times n$ matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ is a circulant matrix if and only if it is diagonalizable by the unitary matrix $\boldsymbol{F}_n$:

$$\boldsymbol{F}_n^*\boldsymbol{A}\boldsymbol{F}_n = \boldsymbol{D}_{\boldsymbol{a}} \quad \text{or} \quad \boldsymbol{A} = \boldsymbol{F}_n\boldsymbol{D}_{\boldsymbol{a}}\boldsymbol{F}_n^*, \tag{9}$$

where $\boldsymbol{D}_{\boldsymbol{a}}$ is a diagonal matrix of (possibly complex) eigenvalues.

Probably the reason why our brain computes in spectral domain.

The Blind Deconvolution Problem

Problem: how to recover both the motif $\boldsymbol{a}$ and sparse spikes $\boldsymbol{x}$ from the observed data $\boldsymbol{y}$:

$$\underbrace{\boldsymbol{y}}_{\textbf{data}} = \underbrace{\boldsymbol{a}}_{\textbf{motif}} * \underbrace{\boldsymbol{x}}_{\textbf{sparse spikes}} + \underbrace{\boldsymbol{z}}_{\textbf{noise}}, \tag{10}$$

This problem is underdetermined (Why?).

We need to leverage low-dimensional structure in both $\boldsymbol{a}$ and $\boldsymbol{x}$ by assuming a **short-and-sparse** model (studied in the 90's):

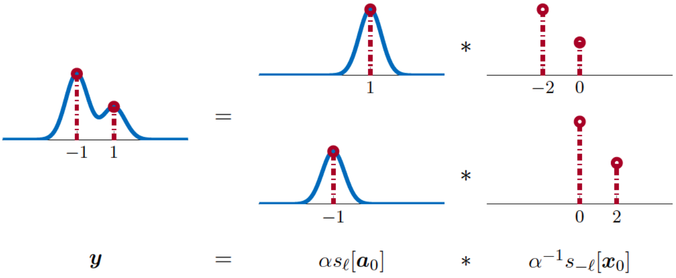
1. $\boldsymbol{a}$ is spatially localized, i.e., it is a *short* signal, whose spatial extent is small compared to that of $\boldsymbol{y}$;
2. $\boldsymbol{x}$ is *sparse*, since it contains only one nonzero entry for each instance of the motif in $\boldsymbol{y}$. (Why not dense?)

Solution by Optimization

Simultaneously recover both $\boldsymbol{a}$ and $\boldsymbol{x}$ by the bilinear Lasso (BL):

$$\min_{\boldsymbol{a},\boldsymbol{x}} \varphi_{\mathsf{BL}}(\boldsymbol{a},\boldsymbol{x}) \doteq \underbrace{\tfrac{1}{2}\|\boldsymbol{y}-\boldsymbol{a}*\boldsymbol{x}\|_F^2}_{\text{data fidelity}} + \underbrace{\lambda\|\boldsymbol{x}\|_1}_{\boldsymbol{x}\text{ sparse}} \quad \text{such that} \quad \underbrace{\boldsymbol{a}\in\mathcal{A}.}_{\boldsymbol{a}\text{ short}} \tag{11}$$

Ambiguity due to a **scaling-shift symmetry**:

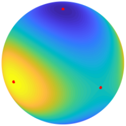


Taxonomy of Symmetric Nonconvex Problems

Nonconvex Problems with Discrete Symmetries

Eigenvector Computation

Maximize a quadratic form over the sphere.



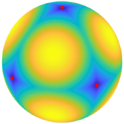
$\max_{\boldsymbol{x}\in\mathbb{S}^{n-1}} \frac{1}{2}\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x}.$

Symmetry: $\boldsymbol{x} \mapsto -\boldsymbol{x}$

$\mathbb{G} = \{\pm 1\}$

Dictionary Learning

Approximate a given matrix $\boldsymbol{Y}$ *as* $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$, *with* $\boldsymbol{X}$ *sparse*



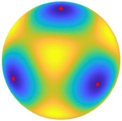
$\min_{\boldsymbol{A}\in\mathcal{A},\boldsymbol{X}} \frac{1}{2}\|\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{X}\|_F^2 + \lambda\|\boldsymbol{X}\|_1.$

Symmetry: $(\boldsymbol{A},\boldsymbol{X}) \mapsto (\boldsymbol{A}\boldsymbol{\Gamma}, \boldsymbol{X}\boldsymbol{\Gamma}^*)$

$\mathbb{G} = \mathsf{SP}(n)$

Tensor Decomposition

Determine components $\boldsymbol{a}_i$ *of an orthogonal decomposable tensor* $\boldsymbol{T} = \sum_i \boldsymbol{a}_i \otimes \boldsymbol{a}_i \otimes \boldsymbol{a}_i \otimes \boldsymbol{a}_i$



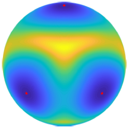
$\max_{\boldsymbol{X}\in\mathsf{O}(n)} \sum_i \boldsymbol{T}(\boldsymbol{x}_i,\boldsymbol{x}_i,\boldsymbol{x}_i,\boldsymbol{x}_i).$

Symmetry: $\boldsymbol{X} \mapsto \boldsymbol{X}\boldsymbol{\Gamma}$

$\mathbb{G} = \mathsf{P}(n)$

Short-and-Sparse Deconvolution

Recover a short $\boldsymbol{a}$ *and a sparse* $\boldsymbol{x}$ *from their convolution* $\boldsymbol{y} = \boldsymbol{a} \circledast \boldsymbol{x}$.



$\min_{\boldsymbol{a},\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{a}\circledast\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_1.$

Symmetry: $(\boldsymbol{a},\boldsymbol{x}) \mapsto (\alpha s_\tau[\boldsymbol{a}], \alpha^{-1}s_{-\tau}[\boldsymbol{x}])$

$\mathbb{G} = \mathbb{Z}_n \times \mathbb{R}_*$ or $\mathbb{G} = \mathbb{Z}_n \times \{\pm 1\}$

Symmetry in Short-and-Sparse Deconvolution

Letting s_τ denote a shift by τ pixels, we have

$$\boldsymbol{y} = s_\tau[\boldsymbol{a}] * s_{-\tau}[\boldsymbol{x}] = \boldsymbol{a} * \boldsymbol{x}, \quad \text{with} \quad \underbrace{\|\boldsymbol{a}\|_F = 1}_{\text{normalization}}. \tag{12}$$

If $\boldsymbol{a}$ is *shift incoherent*:

$$\underbrace{\mu_s = \max_{\tau \neq 0} |\langle \boldsymbol{a}, s_\tau[\boldsymbol{a}] \rangle| \approx 0,}_{\text{incoherence}} \quad \text{or} \quad \underbrace{\text{circ}(\boldsymbol{a}) \approx \boldsymbol{I}}_{\text{isometry}}.$$

the bilinear Lasso loss in (11) can be approximated as

$$\begin{aligned} \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{a} * \boldsymbol{x}\|_F^2 &= \tfrac{1}{2}\|\boldsymbol{y}\|_F^2 + \tfrac{1}{2}\|\boldsymbol{a} * \boldsymbol{x}\|_F^2 - \langle \boldsymbol{y}, \boldsymbol{a} * \boldsymbol{x} \rangle \\ &\approx \tfrac{1}{2}\|\boldsymbol{y}\|_F^2 + \tfrac{1}{2}\|\boldsymbol{x}\|_F^2 - \langle \boldsymbol{y}, \boldsymbol{a} * \boldsymbol{x} \rangle. \end{aligned} \tag{13}$$

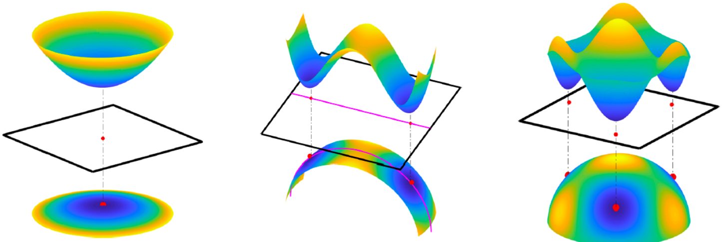
This gives:

$$\varphi_{\text{ABL}}(\boldsymbol{a}, \boldsymbol{x}) \doteq \tfrac{1}{2}\|\boldsymbol{y}\|_F^2 + \tfrac{1}{2}\|\boldsymbol{x}\|_F^2 - \langle \boldsymbol{y}, \boldsymbol{a} * \boldsymbol{x} \rangle + \lambda \|\boldsymbol{x}\|_1, \quad \|\boldsymbol{a}\|_F = 1. \tag{14}$$

Landscape of the Objective Function

Geometry of the approximate bilinear Lasso (ABL) objective:

$$\varphi_{\mathsf{ABL}}(\boldsymbol{a},\boldsymbol{x}) \doteq \frac{1}{2}\|\boldsymbol{y}\|_F^2 + \frac{1}{2}\|\boldsymbol{x}\|_F^2 - \langle \boldsymbol{y}, \boldsymbol{a} * \boldsymbol{x}\rangle + \lambda\|\boldsymbol{x}\|_1, \quad \boldsymbol{a} \in \mathcal{A}. \quad (15)$$



(a) a single shift $s_{\ell_1}[\boldsymbol{a}_0]$ (b) two shifts $s_{\ell_1}[\boldsymbol{a}_0]$, $s_{\ell_2}[\boldsymbol{a}_0]$ (c) multiple shifts

Notice: equivalent (symmetric) solutions are local minimizers, and there is negative curvature in symmetry breaking directions.

Sparsity and Shift-Coherence Tradeoff

Solving the sparse-and-short deconvolution (SaSD) from:

$$\min_{\boldsymbol{a},\boldsymbol{x}} \varphi_{\mathsf{BL}}(\boldsymbol{a},\boldsymbol{x}) \doteq \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{a}*\boldsymbol{x}\|_F^2+\lambda\|\boldsymbol{x}\|_1 \quad \text{such that} \quad \boldsymbol{a}\in\mathcal{A}. \tag{16}$$

A sparsity-coherence tradeoff: Smaller $\mu_s(\boldsymbol{a}_0)$ allows higher $\theta(\boldsymbol{x})$.

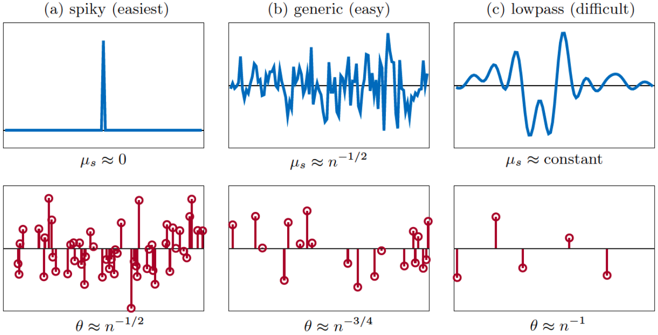


Figure: In order of increasing difficulty: (a) when $\boldsymbol{a}_0$ is a Dirac delta function, $\mu_s(\boldsymbol{a}_0)=0$; (b) when $\boldsymbol{a}_0$ is uniform on the sphere $\mathbb{S}^{n-1}$, its shift-coherence is roughly $\mu_s(\boldsymbol{a}_0)\approx n^{-1/2}$; (c) when $\boldsymbol{a}_0$ is low-pass, $\mu_s(\boldsymbol{a}_0)\to \text{const.}$ as n grows.

Alternating Descent Algorithm for SaSD

Solving the sparse-and-short deconvolution (SaSD) from:

$$\min_{\boldsymbol{a},\boldsymbol{x}} \varphi_{\mathsf{BL}}(\boldsymbol{a},\boldsymbol{x}) \doteq \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{a}*\boldsymbol{x}\|_F^2+\lambda\|\boldsymbol{x}\|_1 \quad \text{such that} \quad \boldsymbol{a}\in\mathcal{A}. \tag{17}$$

Fix a and take a proximal gradient step on x.

$$\text{Gradient w.r.t. } \boldsymbol{x}: \quad \nabla_{\boldsymbol{x}}\psi(\boldsymbol{a},\boldsymbol{x}) = \boldsymbol{\iota}_{\boldsymbol{x}}^*\,\check{\boldsymbol{a}}*(\boldsymbol{a}*\boldsymbol{x}-\boldsymbol{y}). \tag{18}$$

$$\text{Proximal gradient:} \quad \boldsymbol{x}_{k+1} = \operatorname{prox}_{t\lambda g}\left[\boldsymbol{x}_k - t\nabla_{\boldsymbol{x}}\psi(\boldsymbol{a}_k,\boldsymbol{x}_k)\right]. \tag{19}$$

Fix x and take a projected gradient step on $a\in\mathcal{A}$ and $\|a\|_2=1$.

$$\text{Gradient w.r.t. } \boldsymbol{a}: \quad \nabla_{\boldsymbol{a}}\psi(\boldsymbol{a},\boldsymbol{x}) = \boldsymbol{\iota}_{\boldsymbol{a}}^*\check{\boldsymbol{x}}*(\boldsymbol{a}*\boldsymbol{x}-\boldsymbol{y}). \tag{20}$$

$$\text{Proximal gradient:} \quad \boldsymbol{a}_{k+1} = \mathcal{P}_{\mathcal{A}}\left[\boldsymbol{a}_k - \tau_k\nabla_{\boldsymbol{a}}\psi\left(\boldsymbol{a}_k,\boldsymbol{x}_{k+1}\right)\right]. \tag{21}$$

Additional Heuristics

In practice, the kernel $\boldsymbol{a}$ might not be so shift incoherent.

Better Optimization Algorithm: Momentum Acceleration

$$\boldsymbol{w}_k = \boldsymbol{x}_k + \beta \cdot \underbrace{(\boldsymbol{x}_k - \boldsymbol{x}_{k-1})}_{\text{inertial term}}, \tag{22}$$

$$\boldsymbol{x}_{k+1} = \operatorname{prox}_{t_k g}\left[\boldsymbol{w}_k - t_k \nabla_{\boldsymbol{x}} \psi(\boldsymbol{a}_k, \boldsymbol{w}_k)\right]. \tag{23}$$

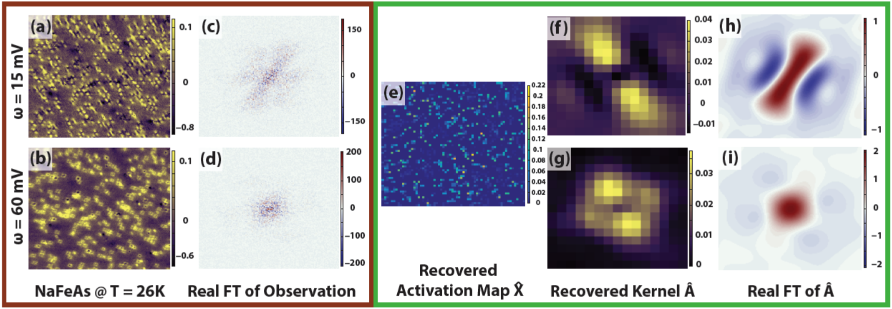
Better Optimization Strategy: Homotopy Continuation
Gradually decreasing λ_n to produce the *solution path* $\left\{(\hat{\boldsymbol{a}}_n, \hat{\boldsymbol{x}}_n; \lambda_n)\right\}$. By ensuring that $\boldsymbol{x}$ remains sparse along the solution path.

Better Initialization: from the Data
Small pieces of $\boldsymbol{y}$ are superpositions of a few shifted copies of $\boldsymbol{a}_0$. One could select a small window of $\boldsymbol{y}$ and then normalizes it to initialize $\boldsymbol{a}$.

An Example of Scanning Tunneling Microscopy

Short and Sparse Deconvolution on Real NaFeAs Data[1]

This dataset $\boldsymbol{y}$ consists of measurements across a $100 \times 100nm^2$ area at $E = 41$ different bias voltages.



[1] Dictionary learning in Fourier-transform scanning tunneling spectroscopy, Sky Cheung et. al., Nature Communications, 2020.

Assignments

- Reading: Section 7.3.3 and Chapter 12.
- Written Homework #4.