ECE 18-898G: Special Topics in Signal Processing: Sparsity, Structure, and Inference

Super resolution, atomic norms and structured matrix completion

Yuejie Chi

Department of Electrical and Computer Engineering

Carnegie Mellon University

Spring 2018

Outline

- Parameter estimation, super resolution
- Classical parametric approach
  - Prony's method
  - MUSIC
  - Matrix pencil
- Optimization-based methods
  - Basis mismatch
  - Atomic norm minimization
  - Connections to low-rank matrix completion

Parameter estimation

Model: a signal is mixture of r modes

$$x[t] = \sum\nolimits_{i=1}^{r} d_i \psi(t; \nu_i), \qquad t \in \mathbb{Z}$$

- d_i : amplitudes
- ν_i : modal parameter
- ψ: (known) modal function, e.g. point spread function
- r: model order
- $2r$ unknown parameters: $\{d_i\}$ and $\{\nu_i\}$

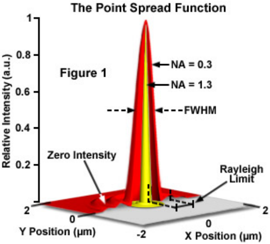
High-resolution source localization

Consider a time signal

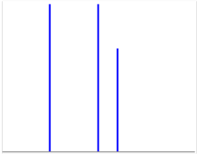
$$z(t) = \sum_{i=1}^{r} d_i \delta(t - t_i)$$

- Resolution is limited by point spread function $h(t)$ of imaging system

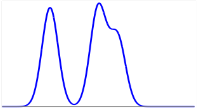
$$x(t) = z(t) * h(t)$$



point spread function $h(t)$



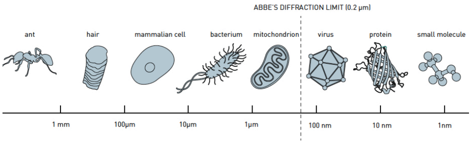
$z(t)$



$x(t)$

Single-molecule fluorescence microscopy

How do we break the diffraction limit of optical microscopy?



The Nobel Prize in Chemistry 2014 "for the development of super-resolved fluorescence microscopy".

E. Betzig

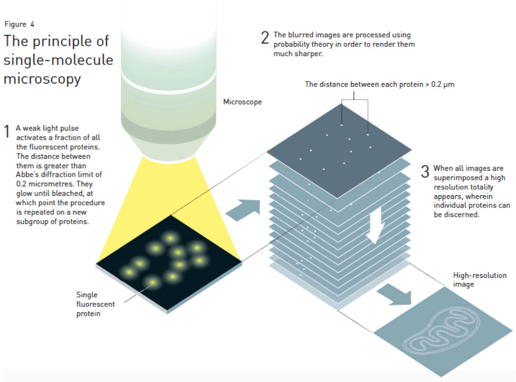
S. W. Hell

W. E. Moerner

Photo credit: `https://www.nobelprize.org/nobel_prizes/chemistry/laureates/2014/`.

Single-molecule fluorescence microscopy

Single-molecule based superresolution techniques achieve nanometer spatial resolution by integrating the temporal information of the switching dynamics of fluorophores (emitters).



High density implies better time resolution.

Figure credit: "The Nobel Prize in Chemistry 2014 - Popular Information".

Spectral-domain viewpoint

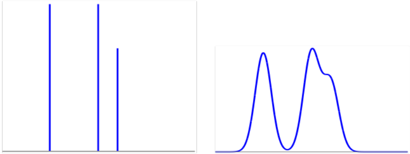
time domain: $$x(t) = z(t) * h(t) = \sum_{i=1}^{r} d_i h(t - t_i)$$

spectral domain: $$\hat{x}(f) = \hat{z}(f)\,\hat{h}(f) = \sum_{i=1}^{r} d_i \underbrace{\hat{h}(f)}_{\text{known}} e^{j2\pi f t_i}$$

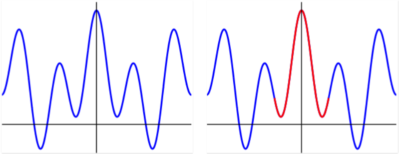
$\Longrightarrow$ observed data: $$\frac{\hat{x}(f)}{\hat{h}(f)} = \sum_{i=1}^{r} d_i \underbrace{e^{j2\pi f t_i}}_{\psi(f;t_i)}, \qquad \forall f : \hat{h}(f) \neq 0$$

$h(t)$ is usually band-limited (suppress high-frequency components)

Application: super-resolution imaging



(a) highly resolved signal $z(t)$; (b) low-pass version $x(t)$



(c) Fourier transform $\hat{z}(f)$; (d) (red) observed spectrum $\hat{x}(f)$

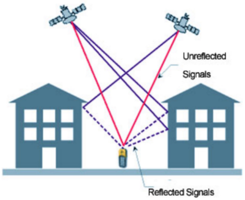
Fig. credit: Candes, Fernandez-Granda '14

Super-resolution: extrapolate high-end spectrum (fine scale details) from low-end spectrum (low-resolution data)

Application: multipath communication channels

In wireless communications, transmitted signals arrive at the receiver by multiple paths, due to reflection from objects (e.g. buildings).

multipath in wireless comm



Suppose $h(t)$ is transmitted signal, then received signal is

$$x(t) = \sum_{i=1}^{r} d_i \, h(t - t_i) \qquad (t_i : \text{ delay in } i^{\text{th}} \text{ path})$$

$\rightarrow$ same as super-resolution model

Basic model

- **Signal model:** a mixture of sinusoids at r distinct frequencies

$$x[t] = \sum_{i=1}^{r} d_i e^{j2\pi t f_i}$$

where $f_i \in [0,1)$: frequencies; d_i : amplitudes

  - *Sparsity in a continuous dictionary*: f_i can assume **ANY** value in $[0,1)$

- **Observed data:**

$$\boldsymbol{x} = \left[x[0], \cdots, x[n-1]\right]^\top$$

or a subsampled version of it in an index set $T \in \{0, 1, \ldots, n-1\}$.

- **Goal:** retrieve the frequencies / recover signal (also called harmonic retrieval)

Matrix / vector representation

Alternatively, the observed data can be written as

$$\boldsymbol{x} = \boldsymbol{V}_{n\times r}\,\boldsymbol{d} \tag{10.1}$$

where $\boldsymbol{d} = [d_1, \cdots, d_r]^\top$;

$$\boldsymbol{V}_{n\times r} := \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ z_1 & z_2 & z_3 & \cdots & z_r \\ z_1^2 & z_2^2 & z_3^2 & \cdots & z_r^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ z_1^{n-1} & z_2^{n-1} & z_3^{n-1} & \cdots & z_r^{n-1} \end{bmatrix} \quad \text{(Vandermonde matrix)}$$

with $z_i = e^{j2\pi f_i}$.

- Basic property of Vandermonde matrix: the columns of $\boldsymbol{V}_{n\times r}$ are *linearly independent* as long as $f_i \neq f_j$, $r \leq n$.

Prony's method

Prony's method

- A *parametric method* proposed by Gaspard Riche de Prony in 1795 based on polynomial interpolation.
- **Key idea:** construct an annihilating filter + polynomial root finding

Annihilating filter

- Define a filter by (Z-transform or characteristic polynomial)

$$G(z) = \sum_{l=0}^{r} g_l z^{-l} = \prod_{l=1}^{r} (1 - z_l z^{-1})$$

whose roots are $\{z_l = e^{j2\pi f_l} \mid 1 \leq l \leq r\}$

- $G(z)$ is called annihilating filter since it annihilates $x[k]$, i.e.

$$q[k] := \underbrace{g_k * x[k]}_{\text{convolution}} = 0 \qquad (10.2)$$

Proof:

$$q[k] = \sum_{i=0}^{r} g_i x[k-i] = \sum_{i=0}^{r} \sum_{l=1}^{r} g_i d_l z_l^{k-i}$$

$$= \sum_{l=1}^{r} d_l z_l^k \left(\underbrace{\sum_{i=0}^{r} g_i z_l^{-i}}_{=0} \right) = 0$$

Annihilating filter

Equivalently, one can write (10.2) as

$$\boldsymbol{X}_{\mathrm{e}}\boldsymbol{g} = \boldsymbol{0}, \tag{10.3}$$

where $\boldsymbol{g} = [g_r, \cdots, g_0]^\top$ and

$$\boldsymbol{X}_{\mathrm{e}} := \underbrace{\begin{pmatrix} x[0] & x[1] & x[2] & \cdots & x[r] \\ x[1] & x[2] & x[3] & \cdots & x[r+1] \\ x[2] & x[3] & x[4] & \cdots & x[r+2] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x[n-r-1] & x[n-r] & \cdots & \cdots & x[n-1] \end{pmatrix}}_{\text{Hankel matrix}} \in \mathbb{C}^{(n-r)\times(r+1)} \tag{10.4}$$

Thus, we can obtain coefficients $\{g_i\}$ (hence the filter $G(z)$) by solving linear system (10.3). Is the solution unique?

$$n - r > r + 1 \Longrightarrow r < (n-1)/2$$

A crucial decomposition

Vandermonde decomposition

$$\boldsymbol{X}_{\mathrm{e}} = \boldsymbol{V}_{(n-r)\times r}\,\mathrm{diag}(\boldsymbol{d})\,\boldsymbol{V}_{(r+1)\times r}^{\top} \tag{10.5}$$

where $\boldsymbol{X}_{\mathrm{e}} \in \mathbb{C}^{(n-r)\times(r+1)}$.

Implications: if $r < (n-1)/2$ and $d_i \neq 0$, then

- rank$(\boldsymbol{X}_{\mathrm{e}}) =$ rank$(\boldsymbol{V}_{(n-r)\times r}) =$ rank$(\boldsymbol{V}_{(r+1)\times r}) = r$
- $\mathrm{null}(\boldsymbol{X}_{\mathrm{e}})$ is 1-dimensional $\iff$ nonzero solution to $\boldsymbol{X}_{\mathrm{e}}\boldsymbol{g} = \boldsymbol{0}$ is unique

A crucial decomposition

Vandermonde decomposition

$$\boldsymbol{X}_{\mathrm{e}} = \boldsymbol{V}_{(n-r)\times r}\,\mathrm{diag}(\boldsymbol{d})\,\boldsymbol{V}_{(r+1)\times r}^{\top} \tag{10.5}$$

where $\boldsymbol{X}_{\mathrm{e}} \in \mathbb{C}^{(n-r)\times(r+1)}$.

Proof: For any i and j,

$$\begin{aligned}[\,\boldsymbol{X}_{\mathrm{e}}]_{i,j} = x[i+j-2] = \sum_{l=1}^{r} d_l z_l^{i+j-2} = \sum_{l=1}^{r} z_l^{i-1} d_l z_l^{j-1} \\ = \left(\boldsymbol{V}_{(n-r)\times r}\right)_{i,:}\mathrm{diag}(\boldsymbol{d})\left(\boldsymbol{V}_{(r+1)\times r}\right)_{j,:}^{\top}\end{aligned}$$

Prony's method

Algorithm 10.1 Prony's method

1. Find $\boldsymbol{g} = [g_r, \cdots, g_0]^\top \neq \mathbf{0}$ that solves $\boldsymbol{X}_\mathrm{e}\boldsymbol{g} = \mathbf{0}$
2. Compute r roots $\{z_l \mid 1 \leq l \leq r\}$ of $G(z) = \sum_{l=0}^{r} g_l z^{-l}$
3. Calculate f_l via $z_l = e^{j2\pi f_l}$

Drawbacks:

- need to estimate the model order
- Root-finding for polynomials becomes difficult for large r
- Numerically unstable in the presence of noise
- don't work with subsampling or missing data

Subspace method: MUSIC

MUltiple SIgnal Classification (MUSIC)

- Let $\boldsymbol{z}(f) := \begin{bmatrix} 1 \\ e^{j2\pi f} \\ \vdots \\ e^{j2\pi rf} \end{bmatrix}$, from the annihilating filter in Prony, $G(e^{j2\pi f_l}) = 0$, we have

$$\boldsymbol{z}(f_l)^\top \boldsymbol{g} = 0,$$

  where $\boldsymbol{g} \in \mathrm{null}(\boldsymbol{X}_\mathrm{e})$.

- Consider a generalized $\boldsymbol{X}_\mathrm{e}$ that has a larger null space, than utilize that subspace for frequency recovery.

MUltiple SIgnal Classification (MUSIC)

Consider a (slightly more general) Hankel matrix

$$\boldsymbol{X}_{\mathrm{e}} = \begin{pmatrix} x[0] & x[1] & x[2] & \cdots & x[k] \\ x[1] & x[2] & x[3] & \cdots & x[k+1] \\ x[2] & x[3] & x[4] & \cdots & x[k+2] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x[n-k-1] & x[n-k] & \cdots & \cdots & x[n-1] \end{pmatrix} \in \mathbb{C}^{(n-k)\times(k+1)}$$

where $r \leq k \leq n-r$ (note that $k = r$ in Prony's method).

- $\mathrm{null}(\boldsymbol{X}_{\mathrm{e}})$ might span multiple dimensions by taking $k > r$

MUltiple SIgnal Classification (MUSIC)

- Generalize Prony's method by computing $\{\boldsymbol{v}_i \mid 1 \le i \le k-r+1\}$ that forms orthonormal basis for $\mathrm{null}(\boldsymbol{X}_\mathrm{e})$, call that subspace $\boldsymbol{V}$

- Let $\boldsymbol{z}(f) := \begin{bmatrix} 1 \\ e^{j2\pi f} \\ \vdots \\ e^{j2\pi k f} \end{bmatrix}$, then it follows from Vandermonde decomposition that

$$\boldsymbol{z}(f_l)^\top \boldsymbol{v}_i = 0, \qquad 1 \le i \le k-r+1,\ 1 \le l \le r$$

- Thus, $\{f_l\}$ are peaks in pseudospectrum

$$S(f) := \frac{1}{\|\boldsymbol{z}(f_l)^\top \boldsymbol{V}\|_2^2} = \frac{1}{\sum_{i=1}^{k-r+1} |\boldsymbol{z}(f)^\top \boldsymbol{v}_i|^2}$$

MUSIC algorithm

Algorithm 10.2 MUSIC

1. Compute orthonormal basis $\{\boldsymbol{v}_i \mid 1 \leq i \leq k - r + 1\}$ for $\mathrm{null}(\boldsymbol{X}_\mathrm{e})$
2. Return r largest peaks of $S(f) := \frac{1}{\sum_{i=1}^{k-r+1} |\boldsymbol{z}(f)^\top \boldsymbol{v}_i|^2}$, where $\boldsymbol{z}(f) := [1, e^{j2\pi f}, \cdots, e^{j2\pi k f}]^\top$

Drawbacks:

- need to estimate the model order
- don't work with subsampling or missing data

Sparse recovery?

Optimization methods for super resolution?

Recall our representation in (10.1):

$$\boldsymbol{x} = \boldsymbol{V}_{n\times r}\,\boldsymbol{d} \qquad (10.6)$$

- **Challenge:** both $\boldsymbol{V}_{n\times r}$ and $\boldsymbol{d}$ are unknown

One can view (10.6) as sparse representation over a continuous dictionary $\{\boldsymbol{z}(f) = [1, e^{j2\pi f}, \cdots, e^{j2\pi(n-1)f}]^\top \mid 0 \le f < 1\}$,

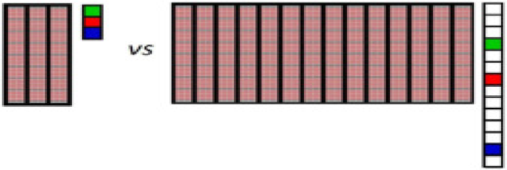
$$\boldsymbol{x} = \sum_{i=1}^{r} d_i \boldsymbol{z}(f_i)$$

Sparse recovery?

Convert nonlinear representation into linear system via discretization at desired resolution:

$$\text{(assume)} \qquad \boldsymbol{x} = \underbrace{\boldsymbol{\Psi}}_{n\times p \text{ overcomplete DFT matrix}} \boldsymbol{\beta}$$

- representation over a discrete frequency set $\{0, \frac{1}{p}, \cdots, \frac{p-1}{p}\}$
- gridding resolution: $1/p$



Over-determined nonlinear versus Under-determined linear and sparse

Sparse recovery via ℓ_1 minimization

Solve ℓ_1 minimization:

$$\text{minimize}_{\boldsymbol{\beta}\in\mathbb{C}^p} \ \|\boldsymbol{\beta}\|_1 \quad \text{s.t. } \boldsymbol{x} = \boldsymbol{\Psi}\boldsymbol{\beta}$$

If $\boldsymbol{\beta}$ is r-sparse, then recovery from $n = O(r \log p)$ samples, and robust against subsampling, noise and outliers enabled by the machinery of **convex optimization**.

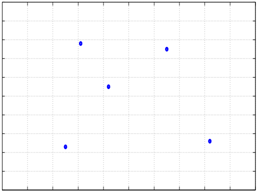
Sparse recovery via ℓ_1 minimization

Solve ℓ_1 minimization:

$$\text{minimize}_{\boldsymbol{\beta} \in \mathbb{C}^p} \ \|\boldsymbol{\beta}\|_1 \quad \text{s.t. } \boldsymbol{x} = \boldsymbol{\Psi}\boldsymbol{\beta}$$

If $\boldsymbol{\beta}$ is r-sparse, then recovery from $n = O(r \log p)$ samples, and robust against subsampling, noise and outliers enabled by the machinery of **convex optimization**.

The issue of being off-the-grid: the point sources / frequencies f_i never lies on the discrete set!

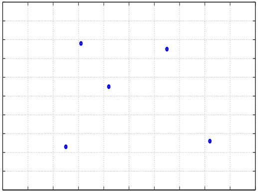


Basis Mismatch: A Tale of Two Models

Mathematical (CS) model:

$$\boldsymbol{x} = \boldsymbol{\Psi}_{cs}\boldsymbol{\beta}$$

The basis $\boldsymbol{\Psi}_{cs}$ is assumed, typically a gridded imaging matrix (e.g., n point DFT matrix or identity matrix), and $\boldsymbol{\beta}$ is presumed to be r-sparse.



Physical (true) model:

$$\boldsymbol{x} = \boldsymbol{\Psi}_{ph}\boldsymbol{\alpha}$$

The basis $\boldsymbol{\Psi}_{ph}$ is unknown, and is determined by a point spread function, a Green's function, or an impulse response, and $\boldsymbol{\alpha}$ is r-sparse and unknown.

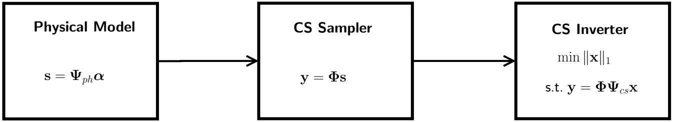
Key transformation:

$$\boldsymbol{\beta} = \boldsymbol{\Psi}_{mis}\boldsymbol{\alpha} = \boldsymbol{\Psi}_{cs}^{-1}\boldsymbol{\Psi}_{ph}\boldsymbol{\alpha}$$

$\boldsymbol{x}$ is sparse in the unknown mismatch $\boldsymbol{\Psi}_{mis}$ basis.

Basis Mismatch: Fundamental Question

Question: What is the consequence of assuming that $\boldsymbol{x}$ is k-sparse in $\boldsymbol{I}$, when in fact it is only k-sparse in an *unknown* basis $\boldsymbol{\Psi}_{mis}$, which is determined by the mismatch between $\boldsymbol{\Psi}_{cs}$ and $\boldsymbol{\Psi}_{ph}$?



Discretization destroys sparsity

Suppose $n = p$ (square case), and recall

$$\boldsymbol{x} = \boldsymbol{\Psi}\boldsymbol{\beta} = \boldsymbol{V}_{n\times r}\boldsymbol{d}$$

$$\Longrightarrow \qquad \boldsymbol{\beta} = \boldsymbol{\Psi}^{-1}\boldsymbol{V}_{n\times r}\boldsymbol{d}$$

Ideally, if $\boldsymbol{\Psi}^{-1}\boldsymbol{V}_{n\times r} \approx$ submatrix of $\boldsymbol{I}$, then sparsity is preserved.

Discretization destroys sparsity

Suppose $n = p$ (square case), and recall

$$\boldsymbol{x} = \boldsymbol{\Psi}\boldsymbol{\beta} = \boldsymbol{V}_{n\times r}\boldsymbol{d}$$

$$\Longrightarrow \qquad \boldsymbol{\beta} = \boldsymbol{\Psi}^{-1}\boldsymbol{V}_{n\times r}\boldsymbol{d}$$

Simple calculation gives

$$\boldsymbol{\Psi}^{-1}\boldsymbol{V}_{n\times r} = \begin{bmatrix} D(\delta_0) & D(\delta_1) & \cdots & D(\delta_r) \\ D(\delta_0 - \frac{1}{p}) & D(\delta_1 - \frac{1}{p}) & \cdots & D(\delta_r - \frac{1}{p}) \\ \vdots & \vdots & \ddots & \vdots \\ D(\delta_0 - \frac{p-1}{p}) & D(\delta_1 - \frac{p-1}{p}) & \cdots & D(\delta_r - \frac{p-1}{p}) \end{bmatrix}$$

where f_i is mismatched to grid $\{0, \frac{1}{p}, \cdots, \frac{p-1}{p}\}$ by δ_i, and

$$D(f) := \frac{1}{p}\sum_{l=0}^{p-1} e^{j2\pi l f} = \frac{1}{p} e^{j\pi f(p-1)} \underbrace{\frac{\sin(\pi f p)}{\sin(\pi f)}}_{\text{heavy tail}} \qquad \text{(Dirichlet kernel)}$$

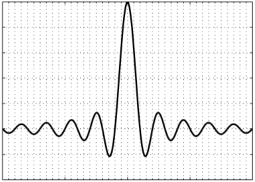
Discretization destroys sparsity

Suppose $n = p$ (square case), and recall

$$\boldsymbol{x} = \boldsymbol{\Psi}\boldsymbol{\beta} = \boldsymbol{V}_{n\times r}\boldsymbol{d}$$

$$\Longrightarrow \qquad \boldsymbol{\beta} = \boldsymbol{\Psi}^{-1}\boldsymbol{V}_{n\times r}\boldsymbol{d}$$

Slow decay / spectral leakage of Dirichlet kernel



If $\delta_i = 0$ (no mismatch), $\boldsymbol{\Psi}^{-1}\boldsymbol{V}_{n\times r}$ = submatrix of $\boldsymbol{I}$

$\Longrightarrow \quad \boldsymbol{\Psi}^{-1}\boldsymbol{V}_{n\times r}\boldsymbol{d}$ is sparse

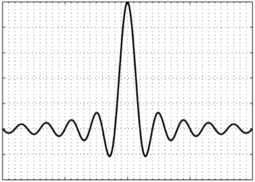
Discretization destroys sparsity

Suppose $n = p$ (square case), and recall

$$\boldsymbol{x} = \boldsymbol{\Psi}\boldsymbol{\beta} = \boldsymbol{V}_{n\times r}\boldsymbol{d}$$

$$\Longrightarrow \qquad \boldsymbol{\beta} = \boldsymbol{\Psi}^{-1}\boldsymbol{V}_{n\times r}\boldsymbol{d}$$

Slow decay / spectral leakage of Dirichlet kernel



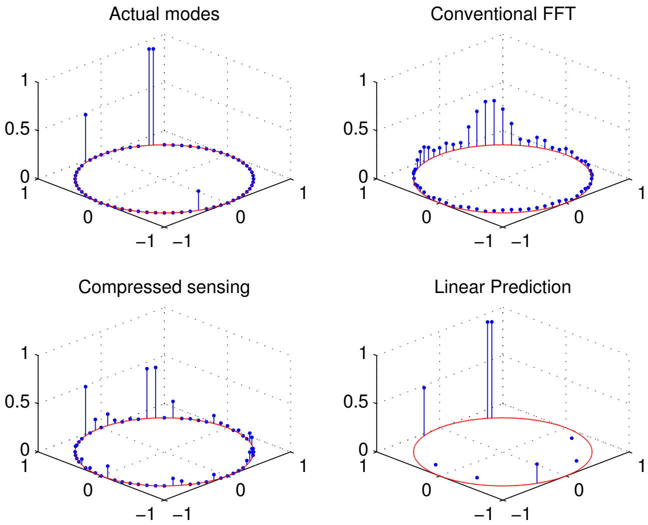
If $\delta_i \neq 0$ (e.g. randomly generated), $\boldsymbol{\Psi}^{-1}\boldsymbol{V}_{n\times r}$ may be far from submatrix of $\boldsymbol{I}$

$\Longrightarrow \quad \boldsymbol{\Psi}^{-1}\boldsymbol{V}_{n\times r}\boldsymbol{d}$ may be **incompressible**

- Finer gridding does not help!

Mismatch of DFT basis

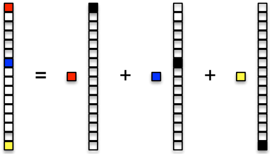
Loss of sparsity after discretization due to basis mismatch

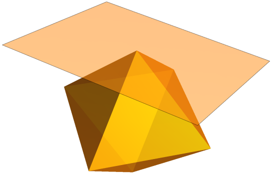


Grid-free methods: atomic norm minimization

Inspirations for Atomic Norm Minimization

- Prior information to exploit: there are only a few active parameters (sparse!), the exact number of which is unknown.
- In compressed sensing, a sparse signal is simple – it is a parsimonious sum of the canonical basis vectors $\{\boldsymbol{e}_k\}$.
- The ℓ_1 norm enforces sparsity w.r.t. the canonical basis vectors.
- The unit ℓ_1 norm ball is $\mathrm{conv}\{\pm\boldsymbol{e}_k\}$, the convex hull of the basis vectors – enforcing sparsity with respect to canonical basis vectors.



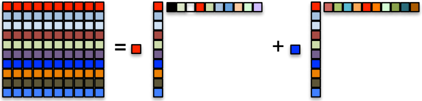


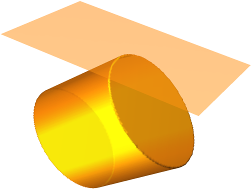
Inspirations for Atomic Norm Minimization

- A low rank matrix has a sparse representation in terms of unit-norm, rank-one matrices.
- The dictionary $D = \{\boldsymbol{u}\boldsymbol{v}^T : \|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2 = 1\}$ is continuously parameterized and has infinite number of primitive signals.
- We enforce low-rankness using the nuclear norm:

$$\|\boldsymbol{X}\|_* = \min\{\|\boldsymbol{\sigma}\|_1 : \boldsymbol{X} = \sum_i \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T\}$$

- The nuclear norm ball is the convex hull of unit-norm, rank-one matrices.
- A hyperplane touches the nuclear norm ball at low-rank solutions.





Atomic Set

- Consider a dictionary or set of atoms $\mathcal{A} = \{\boldsymbol{\psi}(\nu) : \nu \in N\} \subset \mathbb{R}^n$ or $\mathbb{C}^n$.
- The parameter space N can be finite, countably infinite, or continuous.
- The atoms $\{\boldsymbol{\psi}(\nu)\}$ are building blocks for signal representation.
- Examples: canonical basis vectors, rank-one matrices.
- **Line spectral atoms:**

$$\boldsymbol{a}(f, \phi) = e^{j\phi}[1, e^{j2\pi f}, \ldots, e^{j2\pi(n-1)f}]^T : \nu \in [0, 1]$$

Atomic Norms

- Prior information: the signal is simple w.r.t. $\mathcal{A}$— it has a parsimonious decomposition using atoms in $\mathcal{A}$

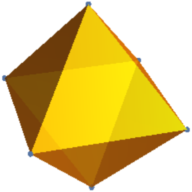
$$\boldsymbol{x} = \sum_{k=1}^{r} \alpha_k \boldsymbol{\psi}(\nu_k)$$

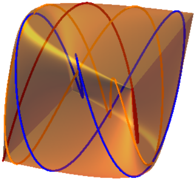
Definition 10.1 (Atomic norm, Chandrasekaran et al. '10)

The atomic norm of any $\boldsymbol{x}$ is defined as

$$\|\boldsymbol{x}\|_{\mathcal{A}} := \inf \left\{ \|\boldsymbol{d}\|_1 : \boldsymbol{x} = \sum_k d_k \psi(\nu_k) \right\} = \inf \left\{ t > 0 : \boldsymbol{x} \in t \operatorname{conv}(\mathcal{A}) \right\}$$

- The unit ball of the atomic norm is the convex hull of $\mathcal{A}$.





Dual norm of atomic norms

- The dual atomic norm is defined as

$$\|\boldsymbol{q}\|_{\mathcal{A}}^{*} := \sup_{\boldsymbol{x}:\|\boldsymbol{x}\|_{\mathcal{A}}\le 1} |\langle \boldsymbol{x}, \boldsymbol{q}\rangle| = \sup_{\boldsymbol{a}\in\mathcal{A}} |\langle \boldsymbol{a}, \boldsymbol{q}\rangle|$$

- For **line spectral atoms**, the dual atomic norm is the maximal magnitude of a complex trigonometric polynomial.

$$\|\boldsymbol{q}\|_{\mathcal{A}}^{*} = \sup_{\boldsymbol{a}\in\mathcal{A}} |\langle \boldsymbol{a}, \boldsymbol{q}\rangle| = \sup_{f\in[0,1]} \left|\sum_{k=0}^{n-1} q_k e^{j2\pi k f}\right|$$

Dual norm of atomic norms

- The dual atomic norm is defined as

$$\|\boldsymbol{q}\|_{\mathcal{A}}^{*} := \sup_{\boldsymbol{x}:\|\boldsymbol{x}\|_{\mathcal{A}}\leq 1} |\langle \boldsymbol{x}, \boldsymbol{q}\rangle| = \sup_{\boldsymbol{a}\in\mathcal{A}} |\langle \boldsymbol{a}, \boldsymbol{q}\rangle|$$

- For **line spectral atoms**, the dual atomic norm is the maximal magnitude of a complex trigonometric polynomial.

$$\|\boldsymbol{q}\|_{\mathcal{A}}^{*} = \sup_{\boldsymbol{a}\in\mathcal{A}} |\langle \boldsymbol{a}, \boldsymbol{q}\rangle| = \sup_{f\in[0,1]} \left| \sum_{k=0}^{n-1} q_k e^{j2\pi k f} \right|$$

Atoms	Atomic Norm	Dual Atomic Norm
canonical basis vectors	ℓ_1 norm	ℓ_∞ norm
finite atoms	$\|\cdot\|_D$	$\|D^\top \boldsymbol{q}\|_\infty$
unit-norm, rank-one matrices	nuclear norm	spectral norm
line spectral atoms	$\|\cdot\|_{\mathcal{A}}$	$\|\cdot\|_{\mathcal{A}}^{*}$

SDP representation of atomic norm

Consider set of line spectral atoms

$$\mathcal{A} := \left\{ \boldsymbol{a}(f,\phi) := e^{j\phi} \cdot [1, e^{j2\pi f}, \cdots, e^{j2\pi(n-1)f}]^\top \;\middle|\; f \in [0,1), \phi \in [0,2\pi) \right\},$$

then

$$\|\boldsymbol{x}\|_{\mathcal{A}} = \inf_{d_k \geq 0,\ \phi_k \in [0,2\pi),\ f_k \in [0,1)} \left\{ \sum\nolimits_k d_k \mid \boldsymbol{x} = \sum\nolimits_k d_k \boldsymbol{a}(f_k, \phi_k) \right\}$$

Lemma 10.2 (Tang, Bhaskar, Shah, Recht '13)

For any $\boldsymbol{x} \in \mathbb{C}^n$,

$$\|\boldsymbol{x}\|_{\mathcal{A}} = \inf \left\{ \frac{1}{2n} \mathrm{Tr}\left(\mathrm{Toeplitz}(\boldsymbol{u})\right) + \frac{1}{2} t \;\middle|\; \begin{bmatrix} \mathrm{Toeplitz}(\boldsymbol{u}) & \boldsymbol{x} \\ \boldsymbol{x}^* & t \end{bmatrix} \succeq \mathbf{0} \right\} \tag{10.7}$$

Caratheodory's decomposition lemma

Lemma 10.3

Any Toeplitz matrix $\boldsymbol{P} \succeq \boldsymbol{0}$ *can be represented as*

$$\boldsymbol{P} = \boldsymbol{V} \mathrm{diag}(\boldsymbol{d}) \boldsymbol{V}^*,$$

where $\boldsymbol{V} := [\boldsymbol{a}(f_1, 0), \cdots, \boldsymbol{a}(f_r, 0)]$, $d_i \geq 0$, *and* $r = \mathrm{rank}(\boldsymbol{P})$.

- Vandermonde decomposition can be computed efficiently via root finding

Proof of Lemma 10.2

Let SDP($\boldsymbol{x}$) be value of RHS of (10.7).

1. **Show that** SDP($\boldsymbol{x}$) $\leq \|\boldsymbol{x}\|_{\mathcal{A}}$**.**

- Suppose $\boldsymbol{x} = \sum_k d_k \boldsymbol{a}(f_k, \phi_k)$ for $d_k \geq 0$. Picking $\boldsymbol{u} = \sum_k d_k \boldsymbol{a}(f_k, 0)$ and $t = \sum_k d_k$ gives (exercise)

$$\mathrm{Toeplitz}(\boldsymbol{u}) = \sum_k d_k \boldsymbol{a}(f_k, 0)\boldsymbol{a}^*(f_k, 0) = \sum_k d_k \boldsymbol{a}(f_k, \phi_k)\boldsymbol{a}^*(f_k, \phi_k)$$

$$\Rightarrow \quad \begin{bmatrix} \mathrm{Toeplitz}(\boldsymbol{u}) & \boldsymbol{x} \\ \boldsymbol{x}^* & t \end{bmatrix} = \sum_k d_k \begin{bmatrix} \boldsymbol{a}(f_k, \phi_k) \\ 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{a}(f_k, \phi_k) \\ 1 \end{bmatrix}^* \succeq \mathbf{0}$$

- Given that $\frac{1}{n}\mathrm{Tr}(\mathrm{Toeplitz}(\boldsymbol{u})) = t = \sum_k d_k$, one has

$$\mathsf{SDP}(\boldsymbol{x}) \leq \sum\nolimits_k d_k.$$

Since this holds for any decomposition of $\boldsymbol{x}$, we conclude this part.

Proof of Lemma 10.2

2. **Show that $\|\boldsymbol{x}\|_{\mathcal{A}} \leq \mathsf{SDP}(\boldsymbol{x})$.**

   i) Suppose for some $\boldsymbol{u}$,

$$\begin{bmatrix} \mathrm{Toeplitz}(\boldsymbol{u}) & \boldsymbol{x} \\ \boldsymbol{x}^* & t \end{bmatrix} \succeq \mathbf{0}. \qquad (10.8)$$

Lemma 10.3 suggests Vandermonde decomposition

$$\mathrm{Toeplitz}(\boldsymbol{u}) = \boldsymbol{V} \mathrm{diag}(\boldsymbol{d}) \boldsymbol{V}^* = \sum_k d_k \boldsymbol{a}(f_k, 0) \boldsymbol{a}^*(f_k, 0).$$

This together with the fact $\|\boldsymbol{a}(f_k, 0)\| = \sqrt{n}$ gives

$$\frac{1}{n} \mathrm{Tr}\left(\mathrm{Toeplitz}(\boldsymbol{u})\right) = \sum_k d_k.$$

Proof of Lemma 10.2

2. **Show that $\|\boldsymbol{x}\|_{\mathcal{A}} \leq \mathsf{SDP}(\boldsymbol{x})$.**

   ii) It follows from (10.8) that $\boldsymbol{x} \in \mathrm{range}(\boldsymbol{V})$, i.e.

$$\boldsymbol{x} = \sum_k w_k \boldsymbol{a}(f_k, 0) = \boldsymbol{V}\boldsymbol{w}$$

   for some $\boldsymbol{w}$. By Schur's complement lemma,

$$\boldsymbol{V}\mathrm{diag}(\boldsymbol{d})\boldsymbol{V}^* \succeq \frac{1}{t}\boldsymbol{x}\boldsymbol{x}^* = \frac{1}{t}\boldsymbol{V}\boldsymbol{w}\boldsymbol{w}^*\boldsymbol{V}^*.$$

   Let $\boldsymbol{q}$ be any vector s.t. $\boldsymbol{V}^*\boldsymbol{q} = \mathrm{sign}(\boldsymbol{w})$. Then

$$\sum\nolimits_k d_k = \boldsymbol{q}^*\boldsymbol{V}\mathrm{diag}(\boldsymbol{d})\boldsymbol{V}^*\boldsymbol{q} \succeq \frac{1}{t}\boldsymbol{q}^*\boldsymbol{V}\boldsymbol{w}\boldsymbol{w}^*\boldsymbol{V}^*\boldsymbol{q} = \frac{1}{t}\left(\sum\nolimits_k |w_k|\right)^2$$

$$\Rightarrow \quad t\sum\nolimits_k d_k \geq \left(\sum\nolimits_k |w_k|\right)^2$$

$$\overset{\text{AM-GM inequality}}{\Longrightarrow} \quad \frac{1}{2n}\mathrm{Tr}\left(\mathrm{Toeplitz}(\boldsymbol{u})\right) + \frac{1}{2}t \geq \sqrt{t\sum\nolimits_k d_k} \geq \sum\nolimits_k |w_k| \geq \|\boldsymbol{x}\|_{\mathcal{A}}$$

Atomic norm minimization

$$\text{minimize}_{\boldsymbol{z}\in\mathbb{C}^n} \quad \|\boldsymbol{z}\|_{\mathcal{A}}$$
$$\text{s.t.} \quad z_i = x_i, \quad i \in T \quad \text{(observation set)}$$

$$\Updownarrow$$

$$\text{minimize}_{\boldsymbol{z}\in\mathbb{C}^n} \quad \frac{1}{2n}\mathrm{Tr}\left(\text{Toeplitz}(\boldsymbol{u})\right) + \frac{1}{2}t$$
$$\text{s.t.} \quad z_i = x_i, \quad i \in T$$
$$\begin{bmatrix} \mathrm{Toeplitz}(\boldsymbol{u}) & \boldsymbol{z} \\ \boldsymbol{z}^* & t \end{bmatrix} \succeq \mathbf{0}$$

Localization via dual solution

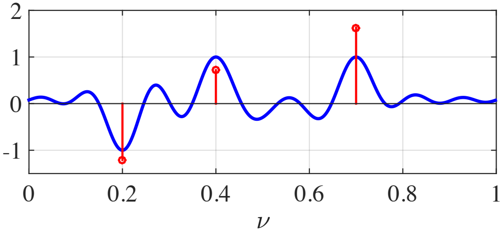
Identify activated atoms (source localization) via the dual solution q:

$$\max \langle \boldsymbol{x}, \boldsymbol{q} \rangle \quad \text{subject to} \quad \|\boldsymbol{q}\|_{\mathcal{A}}^{*} \leq 1$$

- Relaxation is tight (recover the decomposition), when:

$$\text{strict boundeness: } |\langle \boldsymbol{a}(f), \boldsymbol{q} \rangle| < 1, \quad f \in [0,1] \backslash \{f_l\}$$

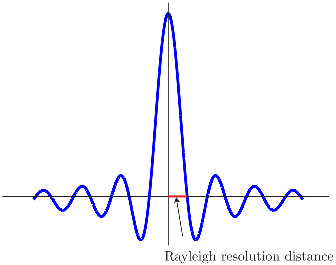
$$\text{interpolation: } \langle \boldsymbol{a}(f_l, 0), \boldsymbol{q} \rangle = \text{sign}(d_l),$$



Key metrics

Minimum separation Δ of $\{f_l \mid 1 \le l \le r\}$ is

$$\Delta := \min_{i \neq l} |f_i - f_l|$$



Rayleigh resolution limit: $\lambda_c = \frac{2}{n-1}$

Performance guarantees for super resolution

Suppose $T = \{-\frac{n-1}{2}, \cdots, \frac{n-1}{2}\}$

Theorem 10.4 (Candes, Fernandez-Granda '14)

Suppose that

- **Separation condition:** $\Delta \geq \frac{4}{n-1} = 2\lambda_{\rm c}$;

Then atomic norm (or total-variation) minimization is exact.

- A deterministic result
- Can recover at most $n/4$ spikes from n consecutive samples
- Does not depend on amplitudes / phases of spikes

Optimality condition

- Define $\mu^\star = \sum_{k=1}^{r} d_k \delta(f - f_k)$.
- Atomic decomposition studies the parameter estimation ability of total variation minimization in the full-data, noise-free case.
- Recall the dual problem:

$$\max \langle \boldsymbol{q}, \boldsymbol{x} \rangle \quad \text{s.t.} \qquad \underbrace{|\langle \boldsymbol{q}, \boldsymbol{a}(f) \rangle| \le 1, \forall f \in [0,1)}_{\|\boldsymbol{q}\|_{\mathcal{A}}^{*} \le 1}$$

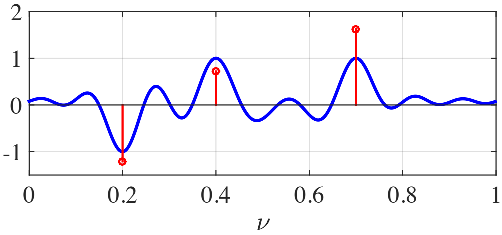
- Define a function $q(f) = \langle \boldsymbol{q}, \boldsymbol{a}(f) \rangle$. $\mu^\star$ is optimal if and only if

$$\text{dual feasibility: } \|q(f)\|_{L_\infty} \le 1$$
$$\text{complementary slackness: } q(f_k) = \mathsf{sign}(d_k), k \in [r]$$

Optimality condition

- To ensure the uniqueness of the optimal solution $\mu^\star$, we strengthen the optimality condition to:

$$\text{strict boundeness: } |q(f)| < 1, \nu \in f \in [0,1)/\{f_k\}$$
$$\text{interpolation: } q(f_k) = \text{sign}(d_k), k \in [r]$$



- **Dual certificate:** constructive proof to design such a dual polynomial.

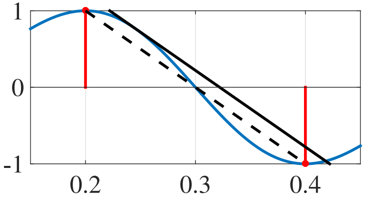
Resolution Limits I

- To simultaneously interpolate sign$(d_i) = +1$ and sign$(d_j) = -1$ at f_i and f_j respectively while remain bounded imposes constraints on the derivative of $q(f)$:

$$\|\nabla q(\hat{f})\|_2 \geq \frac{|q(f_i) - q(f_j)|}{|f_i - f_j|} = \frac{2}{|f_i - f_j|}$$

- By mean-value theorem, there exists $\hat{f} \in (f_i, f_j)$ such that

$$q'(\hat{f}) = \frac{2}{|f_j - f_i|}$$



Resolution Limits II

- For certain classes of functions $\mathcal{F}$, if the function values are uniformly bounded by 1, this limits the maximal achievable derivative, i.e.,

$$\sup_{g \in \mathcal{F}} \frac{\|g'\|_\infty}{\|g\|_\infty} < \infty.$$

- For $\mathcal{F} = \{$trigonometric polynomials of degree at most $n\}$,

$$\|g'(f)\|_\infty \leq 2\pi n \|g(f)\|_\infty.$$

- This is the classical Markov-Bernstein's inequality.
- Resolution limit for line spectral signals: If $\min_{i \neq j} |f_i - f_j| < \frac{1}{\pi n}$, then there is a sign pattern for $\{d_k\}$ such that $\sum_k d_k \boldsymbol{a}(f_k)$ is not an atomic decomposition.

Resolution Limits III

- Using a theorem by Turán about the roots of trigonometric polynomials, Duval and Peyŕe obtained a better critical separation bound

$$\min_{i \neq j} |f_i - f_j| > \frac{1}{n}.$$

- Sign pattern of $\{d_j\}$ plays a big role. There is no resolution limit if, e.g., all d_j are positive ([Schiebinger, Robeva & Recht, 2015]).

Compressed sensing off the grid

Suppose T is random subset of $\{0, \cdots, N-1\}$ of cardinality n

— Extend compressed sensing to continuous domain

Theorem 10.5 (Tang, Bhaskar, Shah, Recht '13)

Suppose that

- **Random sign:** $\text{sign}(d_i)$ *are i.i.d. and random;*
- **Separation condition:** $\Delta \geq \frac{4}{N-1}$*;*
- **Sample size:** $n \gtrsim \max\{r \log r \log N, \log^2 N\}$.

Then atomic norm minimization is exact with high prob.

Connection to low-rank matrix completion

Recall Hankel matrix

$$\boldsymbol{X}_{\mathrm{e}} := \begin{pmatrix} x[0] & x[1] & x[2] & \cdots & x[k] \\ x[1] & x[2] & x[3] & \cdots & x[k+1] \\ x[2] & x[3] & x[4] & \cdots & x[k+2] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x[n-k-1] & x[n-k] & \cdots & \cdots & x[n-1] \end{pmatrix}$$

$$= \boldsymbol{V}_{(n-k)\times r}\,\mathrm{diag}(\boldsymbol{d})\,\boldsymbol{V}_{(k+1)\times r}^{\top} \quad \text{(Vandermonde decomposition)}$$

- rank $(\boldsymbol{X}_{\mathrm{e}}) \leq r$
- Spectral sparsity $\Longleftrightarrow$ low rank

Recovery via Hankel matrix completion

Enhanced Matrix Completion (EMaC):

$$\begin{aligned} \underset{\boldsymbol{z}\in\mathbb{C}^n}{\text{minimize}} \quad & \|\boldsymbol{Z}_{\mathsf{e}}\|_* \\ \text{s.t.} \quad & z_i = x_i, \quad i \in T \end{aligned}$$

When T is random subset of $\{0, \cdots, N-1\}$:

- Coherence measure is closely related to separation condition (Liao & Fannjiang '16)
- Similar performance guarantees as atomic norm minimization (Chen, Chi, Goldsmith '14)

Extension to 2D frequencies

Signal model: a mixture of 2D sinusoids at r distinct frequencies

$$x[\boldsymbol{t}] = \sum_{i=1}^{r} d_i e^{j2\pi\langle \boldsymbol{t}, \boldsymbol{f}_i \rangle}$$

where $\boldsymbol{f}_i \in [0,1)^2$: frequencies; d_i : amplitudes

- Multi-dimensional model: $\boldsymbol{f}_i$ can assume *ANY* value in $[0,1)^2$

Vandermonde decomposition

$$\boldsymbol{X} = [x(t_1, t_2)]_{0 \le t_1 < n_1, 0 \le t_2 < n_2}$$

Vandermonde decomposition:

$$\boldsymbol{X} = \boldsymbol{Y} \cdot \mathrm{diag}(\boldsymbol{d}) \cdot \boldsymbol{Z}^\top.$$

where

$$\boldsymbol{Y} := \begin{bmatrix} 1 & 1 & \cdots & 1 \\ y_1 & y_2 & \cdots & y_r \\ \vdots & \vdots & \vdots & \vdots \\ y_1^{n_1-1} & y_2^{n_1-1} & \dots & y_r^{n_1-1} \end{bmatrix}, \boldsymbol{Z} := \begin{bmatrix} 1 & 1 & \cdots & 1 \\ z_1 & z_2 & \cdots & z_r \\ \vdots & \vdots & \vdots & \vdots \\ z_1^{n_2-1} & z_2^{n_2-1} & \dots & z_r^{n_2-1} \end{bmatrix}$$

with $y_i = \exp(j2\pi f_{1i}), \quad z_i = \exp(j2\pi f_{2i})$.

Multi-fold Hankel matrix (Hua '92)

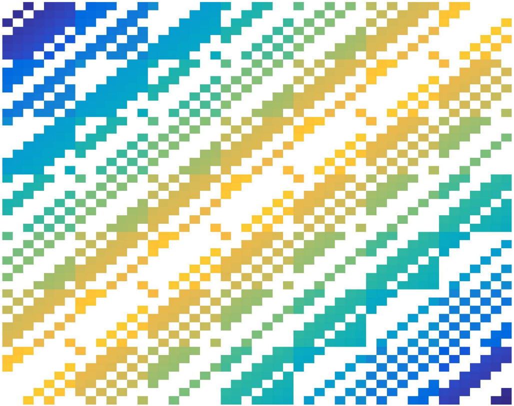
An **enhanced form $\boldsymbol{X}_{\mathsf{e}}$**: $k_1 \times (n_1 - k_1 + 1)$ block Hankel matrix

$$\boldsymbol{X}_{\mathsf{e}} = \begin{bmatrix} \boldsymbol{X}_0 & \boldsymbol{X}_1 & \cdots & \boldsymbol{X}_{n_1-k_1} \\ \boldsymbol{X}_1 & \boldsymbol{X}_2 & \cdots & \boldsymbol{X}_{n_1-k_1+1} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{X}_{k_1-1} & \boldsymbol{X}_{k_1} & \cdots & \boldsymbol{X}_{n_1-1} \end{bmatrix},$$

where each block is $k_2 \times (n_2 - k_2 + 1)$ Hankel matrix:

$$\boldsymbol{X}_l = \begin{bmatrix} x_{l,0} & x_{l,1} & \cdots & x_{l,n_2-k_2} \\ x_{l,1} & x_{l,2} & \cdots & x_{l,n_2-k_2+1} \\ \vdots & \vdots & \vdots & \vdots \\ x_{l,k_2-1} & x_{l,k_2} & \cdots & x_{l,n_2-1} \end{bmatrix}.$$

Multi-fold Hankel matrix (Hua '92)



$\boldsymbol{X}_{\mathrm{e}}$

Low-rank structure of enhanced matrix

- Enhanced matrix can be decomposed as

$$\boldsymbol{X}_{\mathsf{e}} = \begin{bmatrix} \boldsymbol{Z}_{\mathsf{L}} \\ \boldsymbol{Z}_{\mathsf{L}} \boldsymbol{Y}_{\mathsf{d}} \\ \vdots \\ \boldsymbol{Z}_{\mathsf{L}} \boldsymbol{Y}_{\mathsf{d}}^{k_1 - 1} \end{bmatrix} \mathrm{diag}(\boldsymbol{d}) \left[\boldsymbol{Z}_{\mathsf{R}}, \boldsymbol{Y}_{\mathsf{d}} \boldsymbol{Z}_{\mathsf{R}}, \cdots, \boldsymbol{Y}_{\mathsf{d}}^{n_1 - k_1} \boldsymbol{Z}_{\mathsf{R}} \right],$$

  - $\boldsymbol{Z}_{\mathsf{L}}$ and $\boldsymbol{Z}_{\mathsf{R}}$ are Vandermonde matrices specified by $z_1, \ldots, z_r$
  - $\boldsymbol{Y}_{\mathsf{d}} = \mathsf{diag}\left[y_1, y_2, \cdots, y_r \right]$

- Low-rank: $\mathsf{rank}\left(\boldsymbol{X}_{\mathsf{e}} \right) \le r$

Recovery via Hankel matrix completion

Enhanced Matrix Completion (EMaC):

$$\begin{aligned} \underset{\boldsymbol{z}\in\mathbb{C}^n}{\text{minimize}} \quad & \|\boldsymbol{Z}_{\mathsf{e}}\|_* \\ \text{s.t.} \quad & z_{i,j} = x_{i,j}, \quad (i,j) \in T \end{aligned}$$

- Can be easily extended to higher-dimensional frequency models

Reference

[1] ICASSP 2016 Tutorial, "*Convex Optimization Techniques for Super-resolution Parameter Estimation*." Y. Chi and G. Tang.

[2] "*Sampling theory: beyond bandlimited systems*," Y. C. Eldar, *Cambridge University Press*, 2015.

[3] "*Multiple emitter location and signal parameter estimation*," R. Schmidt, *IEEE transactions on antennas and propagation*, 1986.

[4] "*Estimating two-dimensional frequencies by matrix enhancement and matrix pencil*," Y. Hua, *IEEE Transactions on Signal Processing*, 1992.

[5] "*Superresolution via sparsity constraints*," D. Donoho, *SIAM journal on mathematical analysis*, 1992.

[6] "*Sparse nonnegative solution of underdetermined linear equations by linear programming*," D. Donoho, J. Tanner, *Proceedings of the National Academy of Sciences*, 2005.

Reference

[7] "*Sensitivity to basis mismatch in compressed sensing*," Y. Chi, L. Scharf, A. Pezeshki, R. Calderbank, *IEEE Transactions on Signal Processing*, 2011.

[8] "*The convex geometry of linear inverse problems*," V. Chandrasekaran, B. Recht, P. Parrilo, A. Willsky, *Foundations of Computational mathematics*, 2012.

[9] "*Towards a mathematical theory of super-resolution*," E. Candes, C. Fernandez-Granda, *Communications on Pure and Applied Mathematics*, 2014.

[10] "*Compressed sensing off the grid*," G. Tang, B. Bhaskar, P. Shah, B. Recht, *IEEE Transactions on Information Theory*, 2013.

[11] "*Robust spectral compressed sensing via structured matrix completion*," Y. Chen, Y. Chi, *IEEE Transactions on Information Theory*, 2014.