Compressive Parameter Estimation: The Good, The Bad, and The Ugly

Yuejie Chi[o] and Ali Pezeshki[c]
[o]ECE and BMI, The Ohio State University
[c]ECE and Mathematics, Colorado State University

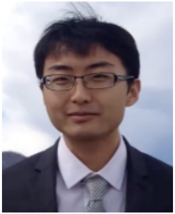
Statistical Signal Processing Workshop 2014
Gold Coast, Australia

Jun. 29, 2014

Acknowledgements

Yuxin Chen

Pooria Pakrooh



Wenbing Dang

Louis Scharf

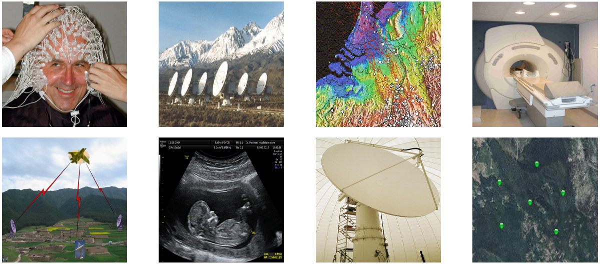
Robert Calderbank

Edwin Chong

Supported by NSF under grants CCF-1018472 and CCF-1017431

Parameter Estimation or Image Inversion

- **Image:** Observable image $\mathbf{y} \sim p(\mathbf{y}; \boldsymbol{\theta})$, whose distribution is parameterized by unknown parameters $\boldsymbol{\theta}$.



- **Inversion:** Estimate $\boldsymbol{\theta}$, given a set of samples of $\mathbf{y}$.
  - Source location estimation in MRI and EEG
  - DOA estimation in sensor array processing
  - Frequency and amplitude estimation in spectrum analysis
  - Range, Doppler, and azimuth estimation in radar/sonar

Parameter Estimation or Image Inversion

- **Canonical Model:** Supperposition of modes:

$$y(t) = \sum_{i=0}^{k-1} \psi(t; \nu_i)\alpha_i + n(t)$$

  - $p = 2k$ unknown parameters: $\boldsymbol{\theta} = [\nu_1, \ldots, \nu_k, \alpha_1, \ldots, \alpha_k]^T$
  - Parameterized modal function: $\psi(t; \nu)$
  - Additive noise: $n(t)$
- **After Sampling:**

$$\begin{bmatrix} y(t_0) \\ y(t_1) \\ \vdots \\ y(t_{m-1}) \end{bmatrix} = \sum_{i=0}^{k-1} \begin{bmatrix} \psi(t_0; \nu_i) \\ \psi(t_1; \nu_i) \\ \vdots \\ \psi(t_{m-1}; \nu_i) \end{bmatrix} \alpha_i + \begin{bmatrix} n(t_0) \\ n(t_1) \\ \vdots \\ n(t_{m-1}) \end{bmatrix}$$

or

$$\mathbf{y} = \boldsymbol{\Psi}(\boldsymbol{\nu})\boldsymbol{\alpha} + \mathbf{n} = \sum_{i=0}^{k-1} \boldsymbol{\psi}(\nu_i)\alpha_i + \mathbf{n}$$

- Typically, t_i's are uniformly spaced and almost always $m > p$.

Parameter Estimation or Image Inversion

- **Canonical Model:**

$$\mathbf{y} = \boldsymbol{\Psi}(\boldsymbol{\nu})\boldsymbol{\alpha} + \mathbf{n} = \sum_{i=0}^{k-1} \psi(\nu_i)\alpha_i + \mathbf{n}$$

  - DOA estimation and spectrum analysis:

$$\psi(\nu) = [e^{jt_0\nu}, e^{jt_1\nu}, \ldots, e^{jt_{m-1}\nu}]^T$$

    where ν is the DOA (electrical angle) of a radiating point source.
  - Radar and sonar:

$$\psi(\nu) = [w(t_0 - \tau)e^{j\omega t_0}, w(t_1 - \tau)e^{j\omega t_1}, \ldots, w(t_{m-1} - \tau)e^{j\omega t_{m-1}}]^T$$

    where $w(t)$ is the transmit waveform and $\nu = (\tau, \omega)$ are delay and Doppler coordinates of a point scatterer.

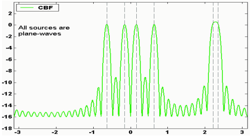
Outline

Outline

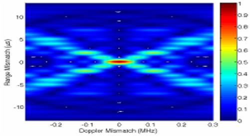
Classical Parameter Estimation or Image Inversion

- **Matched filtering**
  - Sequence of rank-one subspaces, or 1D test images, is matched to the measured image by filtering, correlating, or phasing.
  - Test images are generated by scanning a prototype image (e.g., a waveform or a steering vector) through frequency, wavenumber, doppler, and/or delay at some desired resolution $\Delta\nu$.

$$P(\ell) = \|\psi(\ell\Delta\nu)^H \mathbf{y}\|_2^2$$



Bearing Response



Cross-Ambiguity

  - Peak locations are taken as estimates of ν_i and peak values are taken as estimates of source powers $|\alpha_i|^2$.
  - Resolution: Rayleigh Limit (RL), inversely proportional to the number of measurements

- **Matched filtering (Cont.)**
  - Extends to subspace matching for those cases in which the model for the image is comprised of several dominant modes.
  - Extends to whitened matched filter, or minimum variance unbiased (MVUB) filter, or generalized sidelobe canceller.

H. L. Van Trees, "Detection, Estimation, and Modulation Theory: Part I",

D. J. Thomson, "Spectrum estimation and harmonic analysis," *Proc. IEEE,* vol. 70, pp. 10551096, Sep. 1982.

T.-C.Lui and B. D. Van Veen, "Multiple window based minimum variance spectrum estimation for multidimensional random fields," *IEEE Trans. Signal Process.*, vol. 40, no. 3, pp. 578–589, Mar. 1992.

L. L. Scharf and B. Friedlander, "Matched subspace detectors," IEEE Trans. Signal Process., vol. 42, no. 8, pp. 21462157, Aug. 1994.

A. Pezeshki, B. D. Van Veen, L. L. Scharf, H. Cox, and M. Lundberg, "Eigenvalue beamforming using a multi-rank MVDR beamformer and subspace selection," *IEEE Trans. Signal Processing,* vol. 56, no. 5, pp. 1954–1967, May 2008.

Classical Parameter Estimation or Image Inversion

- **ML Estimation in Separable Nonlinear Models**
  - Low-order separable modal representation for the image:

    $$\mathbf{y} = \boldsymbol{\Psi}(\boldsymbol{\nu})\boldsymbol{\alpha} + \mathbf{n} = \sum_{i=0}^{k-1} \psi(\nu_i)\alpha_i + \mathbf{n}$$

    Parameters $\boldsymbol{\nu}$ in $\boldsymbol{\Psi}$ are nonlinear parameters (like frequency, delay, and Doppler) and $\boldsymbol{\alpha}$ are linear parameters (comples amplitudes).
  - Estimates of linear parameters (complex amplitudes of modes) and nonlinear mode parameters (frequency, wavenumber, delay, and/or doppler) are extracted, usually based on maximum likelihood (ML), or some variation on linear prediction, using ℓ_2 minimization.

Classical Parameter Estimation or Image Inversion

- **Estimation of Complex Exponential Modes**
  - Physical model:

$$y(t) = \sum_{i=0}^{k-1} \nu_i^t \alpha_i + n(t); \quad \psi(t;\nu_i) = \nu_i^t$$

    where $\nu_i = e^{d_i + j\omega_i}$ is a complex exponential mode, with damping d_i and frequency ω_i.
  - Uniformly sampled measurement model:

$$\mathbf{y} = \boldsymbol{\Psi}(\boldsymbol{\nu})\boldsymbol{\alpha}$$

$$\boldsymbol{\Psi}(\boldsymbol{\nu}) = \begin{bmatrix} \nu_0^0 & \nu_1^0 & \cdots & \nu_{k-1}^0 \\ \nu_0^1 & \nu_1^1 & \cdots & \nu_{k-1}^1 \\ \nu_0^2 & \nu_1^2 & \cdots & \nu_{k-1}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \nu_0^{m-1} & \nu_1^{m-1} & \cdots & \nu_{k-1}^{m-1} \end{bmatrix}.$$

    Here, without loss of generality, we have taken the samples at $t = \ell t_0$, for $\ell = 0, 1, \ldots, m-1$, with $t_0 = 1$.

Classical Parameter Estimation or Image Inversion

ML Estimation of Complex Exponential Modes

$$\min_{\boldsymbol{\nu},\boldsymbol{\alpha}} \|\mathbf{y} - \boldsymbol{\Psi}(\boldsymbol{\nu})\boldsymbol{\alpha}\|_2^2$$

$$\hat{\boldsymbol{\alpha}}_{ML} = \boldsymbol{\Psi}(\boldsymbol{\nu})^{\dagger}\mathbf{y}$$

$$\hat{\boldsymbol{\nu}}_{ML} = \text{argmin } \mathbf{y}^H \mathbf{P}_{\mathbf{A}(\boldsymbol{\nu})}\mathbf{y}; \quad \mathbf{A}^H\boldsymbol{\Psi} = \mathbf{0}$$

Prony's method (1795), modified least squares, linear prediction, and Iterative Quadratic Maximum Likelihood (IQML) are used to solve exact ML or its modifications. Rank-reduction is used to combat noise.

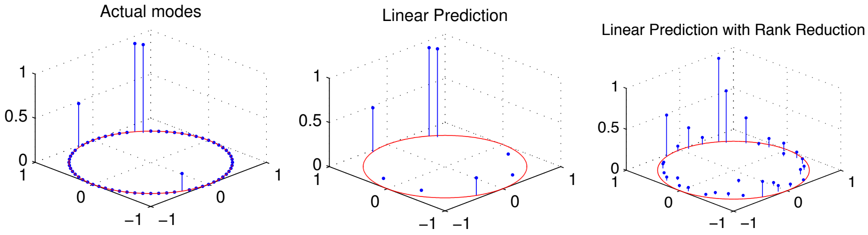
D. W. Tufts and R. Kumaresan, "Singular value decomposition and improved frequency estimation using linear prediction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 30, no. 4, pp. 671675, Aug. 1982.

D. W. Tufts and R. Kumaresan, "Estimation of frequencies of mul- tiple sinusoids: Making linear prediction perform like maximum likelihood," *Proc. IEEE.*, vol. 70, pp. 975989, 1982.

L. L. Scharf "Statistical Signal Processing," Prentice Hall, 1991.

Classical Parameter Estimation or Image Inversion

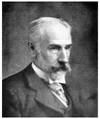
- **Example**:



Classical Parameter Estimation or Image Inversion

- Fundamental limits and performance bounds:
  - Fisher Information
  - Kullback-Leibler divergence
  - Cramér-Rao bounds
  - Ziv-Zakai bound
  - SNR Thresholds

Fisher



Edgeworth

Kullback

Leibler

Cramér

Rao

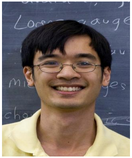
- **Key fact:** Any subsampling of the measured image has consequences for resolution (or bias) and for variability (or variance) in parameter estimation.

L. L. Scharf "Statistical Signal Processing," Prentice Hall, 1991.

Outline

Review of Compressed Sensing

- Compressed Sensing [Name coined by David Donoho] was pioneered by Donoho and Candès, Tao and Romberg in 2004.

- There is now a vast literature on this topic since the last decade.

Compressed sensing
DL Donoho - Information Theory, IEEE Transactions on, 2006 - ieeexplore.ieee.org
Abstract—Suppose is an unknown vector in(a digital image or signal); we plan to measure general linear functionals of and then reconstruct. If is known to be compressible by transform coding with a known transform, and we reconstruct via the nonlinear procedure ...
Cited by 9878 Related articles All 31 versions Cite Save More

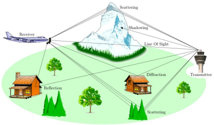
Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information
EJ **Candès**, J Romberg, T **Tao** - Information Theory, IEEE ..., 2006 - ieeexplore.ieee.org
Abstract—This paper considers the model problem of recon-structing an object from incomplete frequency samples. Consider a discrete-time signal and a randomly chosen set of frequencies. Is it possible to reconstruct from the partial knowledge of its Fourier ...
Cited by 6692 Related articles All 38 versions Cite Save

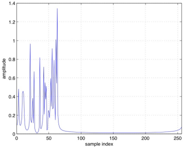
Sparse Representation

Sparsity: Many real world signals admit sparse representation. The signal $\mathbf{s} \in \mathbb{C}^n$ is sparse in a basis $\mathbf{\Psi} \in \mathbb{C}^{n \times n}$, as

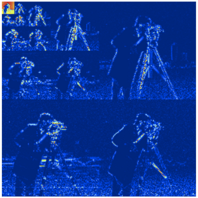
$$\mathbf{s} = \mathbf{\Psi}\mathbf{x};$$

- Multipath channels are sparse in the number of strong paths.



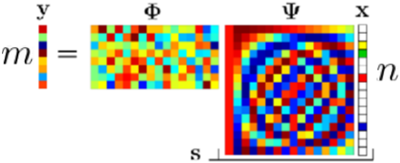


- Images are sparse in the wavelet domain.



Compression on the Fly

Compressed Sensing aims to characterize attributes of a signal with a small number of measurements.



- **Incoherence Sampling:** the linear measurement $\mathbf{y} \in \mathbb{C}^m$ is obtained via an *incoherent* matrix $\mathbf{\Phi} \in \mathbb{C}^{m \times n}$, as

$$\mathbf{y} = \mathbf{\Phi}\mathbf{s} + \mathbf{n},$$

  where $m \ll n$. — *subsampling*.

- The goal is thus to recover $\mathbf{x}$ from $\mathbf{y}$.

Uniqueness of Sparse Recovery

- Let $\mathbf{A} = \mathbf{\Phi\Psi} \in \mathbb{C}^{m\times n}$. We seek the sparsest signal satisfying the observation:

$$(\text{P0:}) \quad \min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{subject to} \quad \mathbf{y} = \mathbf{A}\mathbf{x}.$$

  where $\|\cdot\|_0$ counts the number of nonzero entries.

- Spark: Let Spark($\mathbf{A}$) be the size of the smallest linearly dependent subset of columns of $\mathbf{A}$.

Theorem (Uniqueness, Donoho and Elad 2002)

A representation $\mathbf{y} = \mathbf{A}\mathbf{x}$ *is necessarily the sparsest possible if* $\|\mathbf{x}\|_0 < \mathit{Spark}(\mathbf{A})/2$.

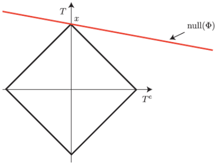
Proof: If $\mathbf{x}$ and $\mathbf{x}'$ satisfy $\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{x}'$, $\|\mathbf{x}'\|_0 \leq \|\mathbf{x}\|_0$, then $\mathbf{A}(\mathbf{x} - \mathbf{x}') = 0$ for $\|\mathbf{x} - \mathbf{x}'\|_0 < \text{Spark}(\mathbf{A})$ implies $\mathbf{x} = \mathbf{x}'$.

D. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," PNAS 100.5 (2003): 2197-2202.

Sparse Recovery via ℓ_1 Minimization

- The above ℓ_0 minimization is NP-hard. A convex relaxation leads to the ℓ_1 minimization, or basis pursuit:

$$\text{(P1:)} \quad \min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{subject to} \quad \mathbf{y} = \mathbf{A}\mathbf{x}.$$



- Mutual Coherence: Let $\mu(\mathbf{A}) = \max_{i \neq j} |\langle \mathbf{a}_i, \mathbf{a}_j \rangle|$, where $\mathbf{a}_i$ and $\mathbf{a}_j$ are normalized columns of $\mathbf{A}$.
  - Spark$(\mathbf{A}) > 1/\mu(\mathbf{A})$.

Theorem (Equivalence, Donoho and Elad 2002)

A representation $\mathbf{y} = \mathbf{A}\mathbf{x}$ *is the unique solution to (P1) if* $\|\mathbf{x}\|_0 < \frac{1}{2} + \frac{1}{2\mu(\mathbf{A})}$.

S. Chen, D. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," SIAM journal on scientific computing 20, no. 1 (1998): 33-61.

D. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," PNAS 100.5 (2003): 2197-2202.

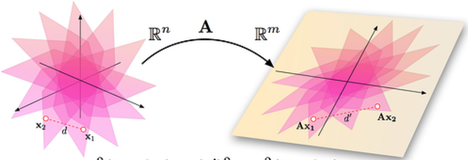
Stable Sparse Recovery via Convex Relaxation

- When $\|\mathbf{n}\|_2 \leq \epsilon$, we incorporate this into the basis pursuit:

$$\text{(P1:)} \quad \mathbf{x}^* = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \leq \epsilon$$

- Restricted Isometry Property: If $\mathbf{A}$ satisfies the restricted isometry property (RIP) with δ_{2k}, then for any two k-sparse vectors $\mathbf{x}_1$ and $\mathbf{x}_2$:

$$1 - \delta_{2k} \leq \frac{\|\mathbf{A}(\mathbf{x}_1 - \mathbf{x}_2)\|_2^2}{\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2} \leq 1 + \delta_{2k}.$$



$$d^2(1 - \delta_{2k}) \leq (d')^2 \leq d^2(1 + \delta_{2k})$$

E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," Information Theory, IEEE Transactions on 52.2 (2006): 489-509.
Picture Credit: Mostafa Mohsenvand.

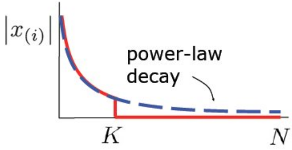
Performance Guarantees Via RIP

Theorem (Candès, Tao, Romberg, 2006)

If $\delta_{2k} < \sqrt{2} - 1$, *then for any vector* $\mathbf{x}$, *the solution to basis pursuit satisfies*

$$\|\mathbf{x}^* - \mathbf{x}\|_2 \leq C_0 k^{-1/2} \|\mathbf{x} - \mathbf{x}_k\|_1 + C_1 \epsilon.$$

where $\mathbf{x}_k$ *is the best* k*-term approximation of* $\mathbf{x}$.



- exact recovery if $\mathbf{x}$ is k-sparse and $\epsilon = 0$.
- stable recovery if $\mathbf{A}$ preserves the isometry between sparse vectors.

- Many random ensembles (e.g. Gaussian, sub-Gaussian, partial DFT) satisfies the RIP as soon as

$$m \sim \Theta(k \log(n/k))$$

Extensions

- **Recovery algorithms:** Orthogonal Matching Pursuit (OMP), CoSaMP, Subspace Pursuit, Iterative Hard Thresholding, Bayesian inference, approximate message passing, etc...
- **Refined signal models:** tree sparsity, group sparsity, multiple measurements, etc...
- **Measurement schemes:** deterministic sensing matrices, structured random matrices, adaptive measurements, etc...

References

- The CS repository hosted by the DSP group at Rice University: `http://dsp.rice.edu/cs`
- The Nuite Blanche blog, maintained by Igor Carron, has a well-maintained list: `https://sites.google.com/site/igorcarron2/cs`
- Check the Nuite Blanche blog for recent updates: `http://nuit-blanche.blogspot.com/`

Outline

CS and Fundamental Estimation Bounds

- Canonical model before compression:

$$\mathbf{y} = \boldsymbol{\Psi}(\boldsymbol{\nu})\boldsymbol{\alpha} + \mathbf{n} = \mathbf{s}(\boldsymbol{\theta}) + \mathbf{n}$$

  where $\boldsymbol{\theta}^T = [\boldsymbol{\nu}^T, \boldsymbol{\alpha}^T] \in \mathbb{C}^p$ and $\mathbf{s}(\boldsymbol{\theta}) = \boldsymbol{\Psi}(\nu)\boldsymbol{\alpha} \in \mathbb{C}^n$.

- Canonical model after compression:

$$\boldsymbol{\Phi}\mathbf{y} = \boldsymbol{\Phi}(\boldsymbol{\Psi}(\boldsymbol{\nu})\boldsymbol{\alpha} + \mathbf{n}) = \boldsymbol{\Phi}(\mathbf{s}(\boldsymbol{\theta}) + \mathbf{n})$$

  where $\boldsymbol{\Phi} \in \mathbb{C}^{m \times n}$, $m \ll n$, is a compressive sensing matrix.

- Question: How are fundamental limits for parameter estimation (i.e., Fisher Information, CRB, KL divergence, etc.) affected by compressively sensing the data?

Fisher Information

- Observable: $\mathbf{y} \sim p(\mathbf{y}; \boldsymbol{\theta})$
- Fisher Score: Sensitivity of log-likelihood function to the parameter vector

$$\frac{\partial}{\partial \theta_i} \log p(\mathbf{y}; \boldsymbol{\theta})$$

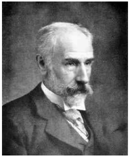
- Fisher information matrix: Covariance of Fisher score

$$\{\mathbf{J}(\boldsymbol{\theta})\}_{i,j} = E\left[\left(\frac{\partial}{\partial \theta_i} \log p(\mathbf{y}; \boldsymbol{\nu})\right)\left(\frac{\partial}{\partial \theta_j} \log p(\mathbf{y}; \boldsymbol{\theta})\right) |\boldsymbol{\theta}\right]$$

$$= -E\left[\frac{\partial^2}{\partial^2 \theta_i \theta_j} \log p(\mathbf{y}; \boldsymbol{\theta}) | \boldsymbol{\theta}\right]$$

Measures the amount of information that the measurement vector $\mathbf{y}$ carries about the parameter vector $\boldsymbol{\theta}$.

Fisher



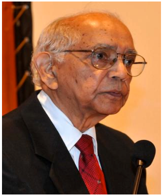
Edgeworth

Cramér-Rao Lower Bound (CRB)

- Cramér-Rao lower bound: Lower bounds the error covariance of any unbiased estimatior $T(\mathbf{y})$ of the parameter vector $\boldsymbol{\theta}$ from measurement $\mathbf{y}$.

$$\mathrm{tr}[\mathrm{cov}_{\boldsymbol{\theta}}(T(\mathbf{y}))] \geq \mathrm{tr}[\mathbf{J}^{-1}(\boldsymbol{\theta})]$$


Cramér


Rao

- The ith diagonal element of $\mathbf{J}^{-1}(\boldsymbol{\theta})$ lower bounds the MSE of any unbiased estimator $T_i(\mathbf{y})$ of the ith parameter θ_i from $\mathbf{y}$.
- Volume of error concentration ellipse:

$$\det[\mathrm{cov}_{\boldsymbol{\theta}}(T(\mathbf{y}))] \geq \det[\mathbf{J}^{-1}(\boldsymbol{\theta})]$$

CS, Fisher Information, and CRB

- Complex Normal model (Canonical model):

$$\mathbf{y} = \mathbf{s}(\boldsymbol{\theta}) + \mathbf{n} \in \mathbb{C}^n; \quad \mathbf{y} = \mathcal{CN}_n[\mathbf{s}(\theta), \mathbf{R}]$$

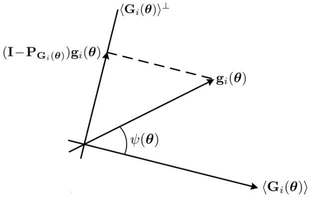
- Fisher information matrix:

$$\mathbf{J}(\boldsymbol{\theta}) = \mathbf{G}^H(\boldsymbol{\theta})\mathbf{R}^{-1}\mathbf{G}(\boldsymbol{\theta})$$
$$= \frac{1}{\sigma^2}\mathbf{G}^H(\boldsymbol{\theta})\mathbf{G}(\boldsymbol{\theta}), \quad \text{when} \quad \mathbf{R} = \sigma^2\mathbf{I}$$

$$\mathbf{G}(\boldsymbol{\theta}) = [\mathbf{g}_1(\boldsymbol{\theta}), \ldots, \mathbf{g}_k(\boldsymbol{\theta})]; \quad \mathbf{g}_i(\boldsymbol{\theta}) = \frac{\partial \mathbf{s}(\boldsymbol{\theta})}{\partial \theta_i}$$

- Cramér-Rao lower bound:

$$(\mathbf{J}^{-1}(\boldsymbol{\theta}))_{ii} = \sigma^2(\mathbf{g}_i^H(\boldsymbol{\theta})(\mathbf{I} - \mathbf{P}_{\mathbf{G_i}(\boldsymbol{\theta})})\mathbf{g}_i(\boldsymbol{\theta}))^{-1}$$



When one sensitivity looks likes a linear combination of others, performance is poor.

L. L. Scharf and L. T. McWhorter, "Geometry of the Cramer-Rao bound," *Signal Process.*, vol. 31, no. 3, pp. 301–311, Apr. 1993.

CS, Fisher Information, and CRB

- Compressive measurement (canonical model):

$$\mathbf{z} = \boldsymbol{\Phi}\mathbf{y} = \boldsymbol{\Phi}[\mathbf{s}(\boldsymbol{\theta}) + \mathbf{n}] \in \mathbb{C}^m;$$

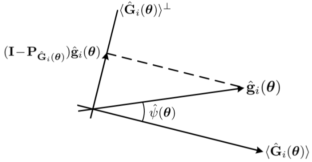
- Fisher information matrix:

$$\hat{\mathbf{J}}(\boldsymbol{\theta}) = \frac{1}{\sigma^2}\mathbf{G}^H(\boldsymbol{\theta})\mathbf{P}_{\boldsymbol{\Phi}^H}\mathbf{G}(\boldsymbol{\nu}) = \hat{\mathbf{G}}^H(\boldsymbol{\theta})\hat{\mathbf{G}}(\boldsymbol{\theta})$$

$$\hat{\mathbf{G}}(\boldsymbol{\theta}) = [\hat{\mathbf{g}}_1(\boldsymbol{\theta}), \ldots, \hat{\mathbf{g}}_k(\boldsymbol{\theta})]; \quad \hat{\mathbf{g}}_i(\boldsymbol{\theta}) = \mathbf{P}_{\boldsymbol{\Phi}^H}\frac{\partial \mathbf{s}(\boldsymbol{\theta})}{\partial \theta_i}$$

- Cramer-Rao lower bound:

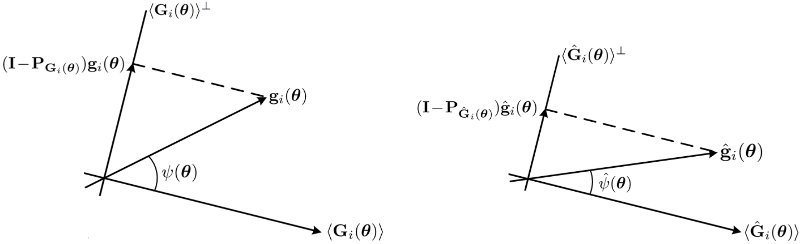
$$(\hat{\mathbf{J}}^{-1}(\boldsymbol{\theta}))_{ii} = \sigma^2(\hat{\mathbf{g}}_i^H(\boldsymbol{\theta})(\mathbf{I}-\mathbf{P}_{\hat{\mathbf{G}}_i(\boldsymbol{\theta})})\hat{\mathbf{g}}_i(\boldsymbol{\theta}))^{-1}$$



Compressive measurement reduces the distance between subspaces: loss of information.

L. L. Scharf and L. T. McWhorter, "Geometry of the Cramer-Rao bound," *Signal Process.*, vol. 31, no. 3, pp. 301–311, Apr. 1993.

CS, Fisher Information, and CRB



- **Question:** What is the impact of compressive sampling on the Fisher information matrix and the Cramér-Rao bound (CRB) for estimating parameters?

CS, Fisher Information, and CRB

Theorem (Pakrooh, Pezeshki, Scharf, Chi '13)

(a) For any compression matrix, we have

$$(\mathbf{J}^{-1}(\boldsymbol{\theta}))_{ii} \leq (\hat{\mathbf{J}}^{-1}(\boldsymbol{\theta}))_{ii} \leq 1/\lambda_{min}(\mathbf{G}^T(\boldsymbol{\theta})\mathbf{P}_{\boldsymbol{\Phi}^T}\mathbf{G}(\boldsymbol{\theta}))$$

(b) For a random compression matrix, we have

$$(\hat{\mathbf{J}}^{-1}(\boldsymbol{\theta}))_{ii} \leq \frac{\lambda_{max}(\mathbf{J}^{-1}(\boldsymbol{\theta}))}{C(1-\epsilon)}$$

with probability at least $1-\delta-\delta'$.

Remarks:

- $(\hat{\mathbf{J}}^{-1})_{ii}$ is the CRB in estimating the ith parameter θ_i.
- CRB always gets worse after compressive sampling.
- Theorem gives a confidence interval and a confidence level for the increase in CRB after random compression.

CS, Fisher Information, and CRB

- δ satisfies

$$Pr\left(\forall \mathbf{q} \in \langle \mathbf{G}(\boldsymbol{\theta}) \rangle : (1-\epsilon)\|\mathbf{q}\|_2^2 \leq \|\boldsymbol{\Phi}\mathbf{q}\|_2^2 \leq (1+\epsilon)\|\mathbf{q}\|_2^2\right) \geq 1-\delta.$$

- $1-\delta'$ is the probability that $\lambda_{min}((\boldsymbol{\Phi}\boldsymbol{\Phi}^T)^{-1})$ is larger than C.
- If entries of $\boldsymbol{\Phi}_{m\times n}$ are i.i.d. $\mathcal{N}(0, 1/m)$, then
  - $\delta \leq \lceil (2\sqrt{p}/\epsilon')^p \rceil e^{-m(\epsilon^2/4-\epsilon^3/6)}$, where

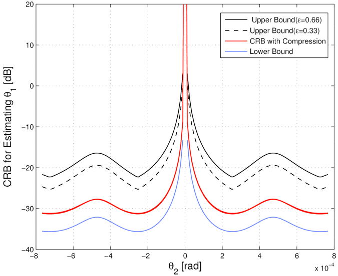
$$\left(\frac{3\epsilon'}{1-\epsilon'}\right)^2 + 2\left(\frac{3\epsilon'}{1-\epsilon'}\right) = \epsilon.$$

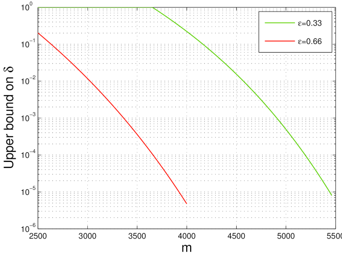
  - δ' is determined from the distribution of the largest eigenvalue of a Wishart matrix, and the value of C, from a hypergeometric function.

P. Pakrooh, L. L. Scharf, A. Pezeshki and Y. Chi, "Analysis of Fisher information and the Cramer-Rao bound for nonlinear parameter estimation after compressed sensing", *in Proc. 2013 IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP),* Vancouver May 26-31, 2013.

CRB after Compression

Example: Estimating the DOA of a point source at boresight $\theta_1 = 0$ in the presence of a point interferer at electrical angle θ_2.





- The LHS figure shows the after compassion CRB (red) for estimating $\theta_1 = 0$ as θ_2 is varied inside the $(-2\pi/n, 2\pi/n]$ interval. Gaussian compression is done from dimension $n = 8192$ to $m = 3000$.
- Bounds on the after compression CRB are shown in blue and black. The upper bounds in black hold with probability at least $1 - \delta - \delta'$, where $\delta' = 0.05$. An upper bound for δ versus the dimension of the compression matrix is plotted on the RHS.

Kullback-Leibler (KL) Divergence

KL divergence: A non-symmetric measure of the difference between two probability distributions

$$D(p(\mathbf{y};\boldsymbol{\theta})||p(\mathbf{y};\boldsymbol{\theta}')) = \int_{\mathbf{y}} p(\mathbf{y};\boldsymbol{\theta}) \log \frac{p(\mathbf{y};\boldsymbol{\theta})}{p(\mathbf{y};\boldsymbol{\theta}')} d\mathbf{y}$$

Kullback

Leilber

S. Kullback and R. A. Leibler, "On Information and Sufficiency," *Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

CS and KL Divergence

KL divergence between $\mathcal{CN}(\mathbf{s}(\boldsymbol{\theta}), \mathbf{R})$ and $\mathcal{CN}(\mathbf{s}(\boldsymbol{\theta}'), \mathbf{R})$:

$$D(\boldsymbol{\theta}, \boldsymbol{\theta}') = \frac{1}{2}[(\mathbf{s}(\boldsymbol{\theta}) - \mathbf{s}(\boldsymbol{\theta}'))^H \mathbf{R}^{-1} (\mathbf{s}(\boldsymbol{\theta}) - \mathbf{s}(\boldsymbol{\theta}'))].$$

- After compression with $\boldsymbol{\Phi}$:

$$\hat{D}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \frac{1}{2}[(\mathbf{s}(\boldsymbol{\theta}) - \mathbf{s}(\boldsymbol{\theta}'))^H \boldsymbol{\Phi}^H (\boldsymbol{\Phi}\mathbf{R}\boldsymbol{\Phi}^H)^{-1} \boldsymbol{\Phi} (\mathbf{s}(\boldsymbol{\theta}) - \mathbf{s}(\boldsymbol{\theta}'))].$$

- With white noise $\mathbf{R} = \sigma^2 \mathbf{I}$:

$$\hat{D}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \frac{1}{2\sigma^2}[(\mathbf{s}(\boldsymbol{\theta}) - \mathbf{s}(\boldsymbol{\theta}'))^H \mathbf{P}_{\boldsymbol{\Phi}^H} (\mathbf{s}(\boldsymbol{\theta}) - \mathbf{s}(\boldsymbol{\theta}'))].$$

Theorem (Pakrooh, Pezeshki, Scharf, and Chi (ICASSP'13))

$$C(1-\epsilon) D(\boldsymbol{\theta}, \boldsymbol{\theta}') \leq \hat{D}(\boldsymbol{\theta}, \boldsymbol{\theta}') \leq D(\boldsymbol{\theta}, \boldsymbol{\theta}')$$

with probability at least $1 - \delta - \delta'$, *where* δ, δ'.

References on CS, Fisher Infromation, and CRB

- L. L. Scharf, E. K. P. Chong, A. Pezeshki, and J. R. Luo, "Compressive sensing and sparse inversion in signal processing: Cautionary notes," *in Proc. 7th Workshop on Defence Applications of Signal Processing (DASP), Coolum, Queensland, Australia, Jul. 10-14, 2011.*

- L. L. Scharf, E. K. P. Chong, A. Pezeshki, and J. R. Luo, "Sensitivity considerations in compressed sensing," *in Conf. Rec. 45th Annual Asilomar Conf. Signals, Systs., Computs., Pacific Grove, CA,, Nov. 2011, pp. 744–748.*

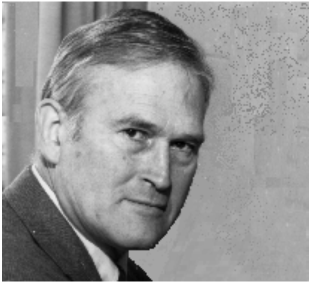
- P. Pakrooh, L. L. Scharf, A. Pezeshki and Y. Chi, "Analysis of Fisher information and the Cramer-Rao bound for nonlinear parameter estimation after compressed sensing", *in Proc. 2013 IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP), Vancouver May 26-31, 2013.*

Work by Others

- Nielsen, Christensen, and Jensen (ICASSP'12): Bounds on mean value of Fisher Information after random compression.
- Ramasamy, Venkateswaran, and Madhow (Asilomar'12): Bounds on Fisher information after compression in a different noisy model.
- Babadi, Kalouptsidis, and Tarokh (TSP 2009): Existence of an estimator ("Joint Typicality Estimator") that asymptotically achieves the CRB in linear parameter estimation with random Gaussian compression matrices.

Breakdown Threshold and Subspace Swaps

- Threshold effect: Sharp deviation of Mean Squared Error (MSE) performance from Cramer-Rao Bound (CRB).
- Breakdown threshold: SNR at which a threshold effect occurs with non-negligible probability.

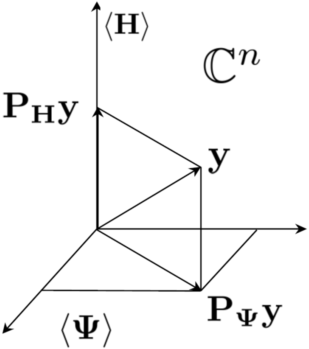


Donald W. Tufts (1933-2012)

D. W. Tufts, A. C. Kot, and R. J. Vacarro, "The threshold effect in signal processing algorithms which use an estimated subspace," in *SVD and Signal Processing, II,* R. J. Vaccaro (Ed), New York: Elsevier, 1991.

Breakdown Threshold and Subspace Swaps

- Subspace Swap: Event in which measured data is more accurately resolved by one or more modes of an *orthogonal subspace* to the signal subspace.
- Cares only about what the data itself is saying.
- Bound probability of a subspace swap to predict breakdown SNRs.



J. K. Thomas, L. L. Scharf, and D. W. Tufts, "Probability of a Subspace Swap in the SVD," *IEEE Trans Signal Proc.*, vol 43, no 3, pp 730-736, Mar. 1995.

Signal Model: Mean Case

- Before compression:

$$\mathbf{y} : \mathcal{CN}_n[\boldsymbol{\Psi}\boldsymbol{\alpha}, \sigma^2\mathbf{I}]; \quad \boldsymbol{\Psi} \in \mathbb{C}^{n\times k}$$

- After compression with compressive sensing matrix $\boldsymbol{\Phi}_{cs} \in \mathbb{C}^{m\times n}, m < n$:

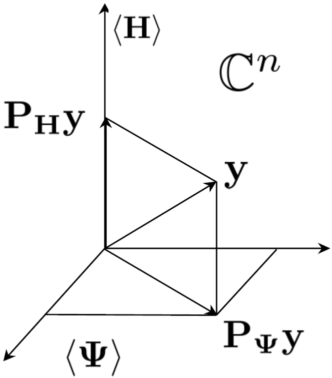
$$\mathbf{y}_{cs} : \mathcal{CN}_m[\boldsymbol{\Phi}_{cs}\boldsymbol{\Psi}\boldsymbol{\alpha}, \sigma^2\boldsymbol{\Phi}_{cs}\boldsymbol{\Phi}_{cs}^H]$$

or equivalently (with some abuse of notation):

$$\mathbf{y}_{cs} : \mathcal{CN}_m[\boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{\alpha}, \sigma^2\mathbf{I}], \quad \boldsymbol{\Phi} = (\boldsymbol{\Phi}_{cs}\boldsymbol{\Phi}_{cs}^H)^{-1/2}\boldsymbol{\Phi}_{cs}$$

Subspace Swap Events

- Subspace Swap Event E: One or more modes of the orthogonal subspace $\langle \mathbf{H} \rangle$ resolves more energy than one or more modes of the noise-free signal subspace $\langle \mathbf{\Psi} \rangle$.



Subspace Swap Events

- Subevent F: Average energy resolved in the orthogonal subspace $\langle \mathbf{H} \rangle = \langle \mathbf{\Psi} \rangle^{\perp}$ is greater than the average energy resolved in the noise-free signal subspace $\langle \mathbf{\Psi} \rangle$.

$$\min_i |\psi_i^H \mathbf{y}|^2 \leq \frac{1}{k} \mathbf{y}^H \mathbf{P}_{\mathbf{\Psi}} \mathbf{y} < \frac{1}{n-k} \mathbf{y}^H \mathbf{P}_{\mathbf{H}} \mathbf{y} \leq \max_i |\mathbf{h}_i^H \mathbf{y}|^2$$

- Subevent G: Energy resolved in the apriori minimum mode ψ_{min} of the noise-free signal subspace $\langle \mathbf{\Psi} \rangle$ is smaller than the average energy resolved in the orthogonal subspace $\langle \mathbf{H} \rangle$.

$$|\psi_{min}^H \mathbf{y}|^2 < \frac{1}{n-k} \mathbf{y}^H \mathbf{P}_{\mathbf{H}} \mathbf{y} \leq \max_i |\mathbf{h}_i^H \mathbf{y}|^2.$$

Probability of Subspace Swap: Mean Case

Theorem (Pakrooh, Pezeshki, Scharf (GlobalSIP'13))

(a) Before compression:

$$P_{ss} \geq 1 - P[\frac{\mathbf{y}^H \mathbf{P}_{\boldsymbol{\Psi}} \mathbf{y}/k}{\mathbf{y}^H \mathbf{P}_{\mathbf{H}} \mathbf{y}/(n-k)} > 1]$$
$$= 1 - P[F_{2k,2(n-k)}(\|\boldsymbol{\Psi}\boldsymbol{\alpha}\|_2^2/\sigma^2) > 1]$$

$\|\boldsymbol{\Psi}\boldsymbol{\alpha}\|_2^2/\sigma^2$ *is the SNR before compression.*

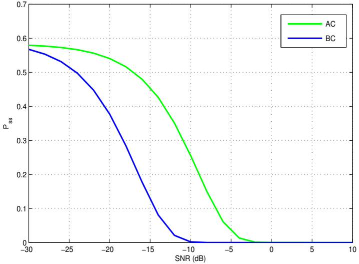
(b) After compression:

$$P_{ss} \geq 1 - P[F_{2k,2(m-k)}(\|\boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{\alpha}\|_2^2/\sigma^2) > 1]$$

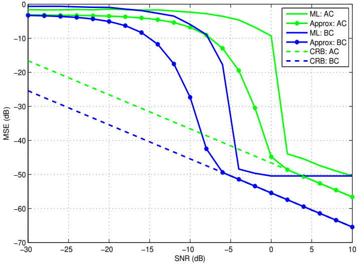
$\|\boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{\alpha}\|_2^2/\sigma^2$ *is the SNR after compression.*

P. Pakrooh, A. Pezeshki, and L. L. Scharf, "Threshold effects in parameter estimation from compressed data," *Proc. 1st IEEE Global Conference on Signal and Information Processing,* Austin, TX, Dec. 2013.

Sensor Array Processing–Mean Case



Analytical lower bounds for the probability of subspace swap. Array size: $n = 188$ elements; Compressed array $m = 28$ elements.

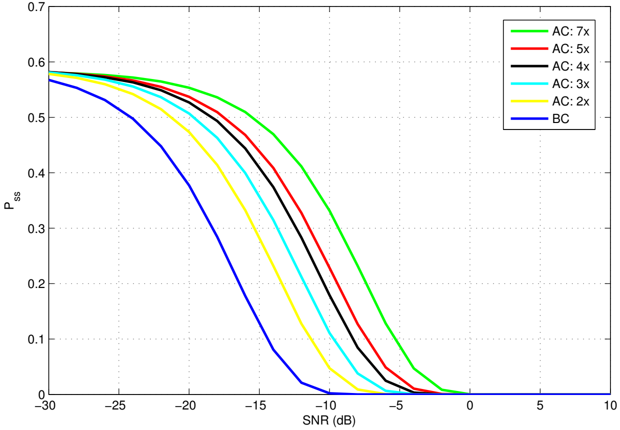


Empirical MSE (average over 200 trials) and MSE bounds for estimating $\theta_1 = 0$; Interfering source at $\theta_2 = \pi/188$; Array size: $n = 188$ elements; Compressed array $m = 28$ elements.

- Bounds on subspace swap probabilities and SNR treshold are poredictive of MSE performance loss.
- ML Approximation: Method of intervals

$$\mathrm{MSE} \approx P_{ss}\sigma_0^2 + (1 - P_{ss})\sigma_{CR}^2$$

Sensor Array Processing–Mean Case



Analytical lower bounds for the probability of subspace swap for different compression ratios n/m. The before compression (BC) dimension is $n = 188$.

Signal Model: Covariance Case

- Before compression:

$$\mathbf{y} : \mathcal{CN}_n[\mathbf{0}, \mathbf{\Psi}\mathbf{R}_{\alpha\alpha}\mathbf{\Psi}^H + \sigma^2\mathbf{I}]; \quad \mathbf{\Psi} \in \mathbb{C}^{n\times k}$$

- After compression with compressive sensing matrix $\mathbf{\Phi}_{cs} \in \mathbb{C}^{m\times n}$, $m < n$:

$$\mathbf{y}_{cs} : \mathcal{CN}_m[\mathbf{0}, \mathbf{\Phi}_{cs}\mathbf{\Psi}\mathbf{R}_{\alpha\alpha}\mathbf{\Psi}^H\mathbf{\Phi}_{cs}^H + \sigma^2\mathbf{\Phi}_{cs}\mathbf{\Phi}_{cs}^H]$$

  or equivalently (with some abuse of notation):

$$\mathbf{y}_{cs} : \mathcal{CN}_m[\mathbf{0}, \mathbf{\Phi}\mathbf{\Psi}\mathbf{R}_{\alpha\alpha}\mathbf{\Phi}\mathbf{\Psi}^H + \sigma^2\mathbf{I}], \quad \mathbf{\Phi} = (\mathbf{\Phi}_{cs}\mathbf{\Phi}_{cs}^H)^{-1/2}\mathbf{\Phi}_{cs}.$$

- Assume data consists of L iid realizations of $\mathbf{y}$ arranged as $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_L]$.

Probablity of Subspace Swap: Covariance Case

Theorem (Pakrooh, Pezeshki, Scharf (GlobalSIP'13))

(a) Before compression:

$$P_{ss} \geq 1 - P[\frac{tr(\mathbf{Y}^H \mathbf{P}_{\boldsymbol{\Psi}} \mathbf{Y}/kL)}{tr(\mathbf{Y}^H \mathbf{P}_{\mathbf{H}} \mathbf{Y}/(n-k)L)} > 1]$$
$$= 1 - P[F_{2kL,2(n-k)L} > \frac{1}{1+\lambda_k/\sigma^2}].$$

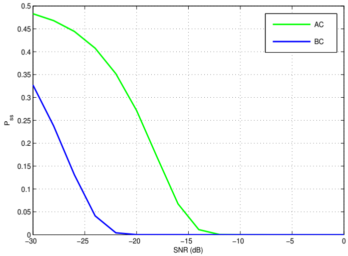
$\lambda_k = ev_{min}(\boldsymbol{\Psi}\mathbf{R}_{\alpha\alpha}\boldsymbol{\Psi}^H)$
λ_k/σ^2: *Effective SNR before compression*

(b) After compression:

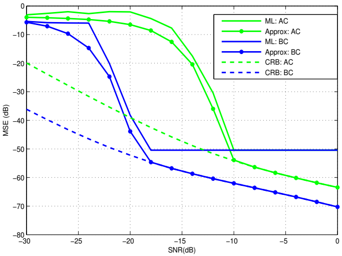
$$P_{ss} \geq 1 - P[F_{2kL,2(m-k)L} > \frac{1}{1+\lambda'_k/\sigma^2}].$$

$\lambda'_k = ev_{min}(\boldsymbol{\Phi}\boldsymbol{\Psi}\mathbf{R}_{\alpha\alpha}\boldsymbol{\Psi}^H\boldsymbol{\Phi}^H)$
λ'_k/σ^2: *Effective SNR after compression*

Sensor Array Processing–Covariance Case



Analytical lower bounds for the probability of subspace swap; $n = 188$ and $m = 28$



Empirical MSE and MSE bounds; Interfering source at $\theta_2 = \pi/188$; 200 snapshots; Averaged over 500 trials; $n = 188$ and $m = 28$.

- Bounds on subspace swap probabilities and SNR treshold are poredictive of MSE performance loss.
- ML Approximation: Method of intervals

$$\mathrm{MSE} \approx P_{ss}\sigma_0^2 + (1 - P_{ss})\sigma_{CR}^2$$

P. Pakrooh, A. Pezeshki, and L. L. Scharf, “Threshold effects in parameter estimation from compressed data,” *Proc. 1st IEEE Global Conference on Signal and Information Processing*, Austin, TX, Dec. 2013.

References on Breakdown Thresholds

- P. Pakrooh, A. Pezeshki, and L. L. Scharf, "Threshold effects in parameter estimation from compressed data," *Proc. 1st IEEE Global Conference on Signal and Information Processing,* Austin, TX, Dec. 2013.
- D. Tufts, A. Kot, and R. Vaccaro, The threshold effect in signal processing algorithms which use an estimated subspace, *SVD and Signal Processing II: Algorithms, Analysis and Applications,* New York: Elsevier, 1991, pp. 301320.
- J. K. Thomas, L. L. Scharf, and D. W. Tufts, The probability of a subspace swap in the SVD, *IEEE Transactions on Signal Processing,* vol. 43, no. 3, pp. 730736, Mar. 1995.
- B. A. Johnson, Y. I. Abramovich, and X. Mestre, MUSIC, G-MUSIC, and maximum-likelihood performance breakdown, *IEEE Transactions on Signal Processing,* vol. 56, no. 8, pp. 3944-3958, Aug. 2008.

Intermediate Recap: Fundamental Limits

- Compression (even with Gaussian or similar random matrices) has performance consequences.
- The CR bound increases and the onset of threshold SNR increases. These increases may be quantified to determine where compressive sampling is viable.

Outline

Basis Mismatch: NonLin. Overdet. vs. Lin. Underdet.

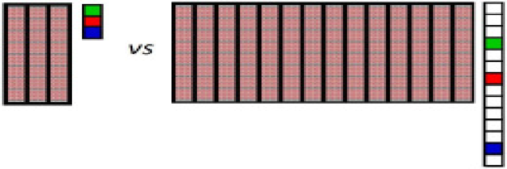
- Convert the nonlinear problem into a linear system via discretization of the parameter space at desired resolution:

$$\mathbf{s}(\boldsymbol{\theta}) = \sum_{i=0}^{k-1} \psi(\boldsymbol{\nu}_i)\alpha_i = \boldsymbol{\Psi}_{ph}\boldsymbol{\alpha}$$

Over-determined & nonlinear

$$\mathbf{s} \approx [\psi(\omega_1), \cdots, \psi(\omega_n)] \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \boldsymbol{\Psi}_{cs}\mathbf{x}$$

Under-determined linear & sparse



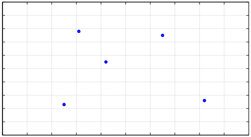
- The set of candidate $\nu_i \in \Omega$ is quantized to $\tilde{\Omega} = \{\omega_1, \cdots, \omega_n\}$, $n > m$; $\boldsymbol{\Psi}_{ph}$ unknown and $\boldsymbol{\Psi}_{cs}$ assumed known.

Basis Mismatch: A Tale of Two Models

Mathematical (CS) model:

$$\mathbf{s} = \boldsymbol{\Psi}_{cs}\mathbf{x}$$

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



Physical (true) model:

$$\mathbf{s} = \boldsymbol{\Psi}_{ph}\boldsymbol{\alpha}$$

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 k-sparse and unknown.

Key transformation:

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

$\mathbf{x}$ is sparse in the unknown $\boldsymbol{\Psi}_{mis}$ basis, not in the identity basis.

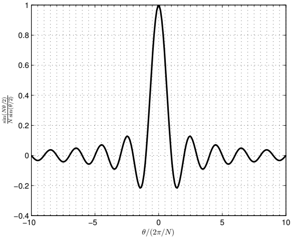
Basis Mismatch: From Sparse to Incompressible

DFT Grid Mismatch:

$$\boldsymbol{\Psi}_{mis} = \boldsymbol{\Psi}_{cs}^{-1}\boldsymbol{\Psi}_{ph} = \begin{bmatrix} L(\Delta\theta_0 - 0) & L(\Delta\theta_1 - \frac{2\pi(n-1)}{n}) & \cdots & L(\Delta\theta_{n-1} - \frac{2\pi}{n}) \\ L(\Delta\theta_0 - \frac{2\pi}{n}) & L(\Delta\theta_1 - 0) & \cdots & L(\Delta\theta_{n-1} - \frac{2\pi\cdot 2}{n}) \\ \vdots & \vdots & \ddots & \vdots \\ L(\Delta\theta_0 - \frac{2\pi(n-1)}{n}) & L(\Delta\theta_1 - \frac{2\pi(n-2)}{n}) & \cdots & L(\Delta\theta_{n-1} - 0) \end{bmatrix}$$

where $L(\theta)$ is the Dirichlet kernel:

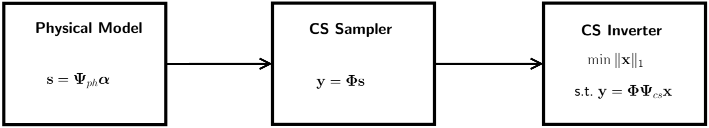
$$L(\theta) = \frac{1}{n}\sum_{\ell=0}^{n-1} e^{j\ell\theta} = \frac{1}{n}e^{j\frac{\theta(n-1)}{2}}\frac{\sin(\theta n/2)}{\sin(\theta/2)}.$$



Slow decay of the Dirichlet kernel means that the presumably sparse vector $\mathbf{x} = \boldsymbol{\Psi}_{mis}\boldsymbol{\alpha}$ is in fact incompressible.

Basis Mismatch: Fundamental Question

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



Sensitivity to Basis Mismatch

- CS Inverter: Basis pursuit solution satisfies

$$\begin{aligned} \text{Noise-free:} \quad & \|\mathbf{x}^* - \mathbf{x}\|_1 \leq C_0 \|\mathbf{x} - \mathbf{x}_k\|_1 \\ \text{Noisy:} \quad & \|\mathbf{x}^* - \mathbf{x}\|_2 \leq C_0 k^{-1/2} \|\mathbf{x} - \mathbf{x}_k\|_1 + C_1 \epsilon \end{aligned}$$

  where $\mathbf{x}_k$ is the best k-term approximation to $\mathbf{x}$.
- Similar bounds CoSaMP and ROMP.
- Where does mismatch enter? k-term approximation error.

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

- **Key:** Analyze the sensitivity of $\|\mathbf{x} - \mathbf{x}_k\|_1$ to basis mismatch.

Degeneration of Best $k-$Term Approximation

Theorem (Chi, Scharf, P., Calderbank (TSP 2011))

Let $\boldsymbol{\Psi}_{mis} = \boldsymbol{\Psi}_{cs}^{-1}\boldsymbol{\Psi}_{ph} = \mathbf{I} + \mathbf{E}$*, where* $\mathbf{x} = \boldsymbol{\Psi}_{mis}\boldsymbol{\alpha}$*. Let* $1 \leq p, q \leq \infty$ *and* $1/p + 1/q = 1$*.*

- *If the rows* $\mathbf{e}_\ell^T \in \mathbb{C}^{1\times n}$ *of* $\mathbf{E}$ *are bounded as* $\|\mathbf{e}_\ell\|_p \leq \beta$*, then*

$$\|\mathbf{x} - \mathbf{x}_k\|_1 \leq \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_k\|_1 + (n-k)\beta\|\boldsymbol{\alpha}\|_q.$$

- *The bound is achieved when the entries of* $\mathbf{E}$ *satisfy*

$$e_{mn} = \pm\beta \cdot e^{j(\arg(\alpha_m)-\arg(\alpha_n))} \cdot (|\alpha_n|/\|\boldsymbol{\alpha}\|_q)^{q/p}.$$

Y. Chi, L.L. Scharf, A. Pezeshki, and A.R. Calderbank, "Sensitivity to basis mismatch in compressed sensing," *IEEE Transactions on Signal Processing,* vol. 59, no. 5, pp. 2182–2195, May 2011.

Bounds on Image Inversion Error

Theorem (inversion error)

Let $\mathbf{A} = \mathbf{\Phi\Psi}_{mis}$ *satisfy* $\delta_{2k}^{\mathbf{A}} < \sqrt{2} - 1$ *and* $1/p + 1/q = 1$. *If the rows of* $\mathbf{E}$ *satisfy* $\|\mathbf{e}_m\|_p \leq \beta$, *then*

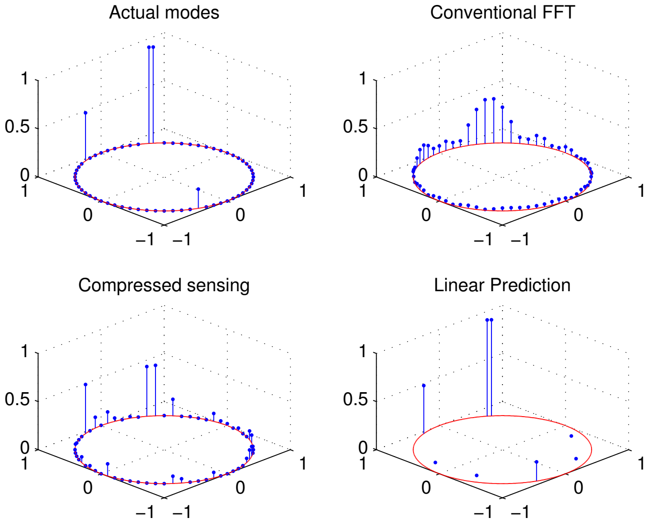
$$\|\mathbf{x} - \mathbf{x}^*\|_1 \leq C_0(n-k)\beta\|\boldsymbol{\alpha}\|_q. \quad \textit{(noise-free)}$$

$$\|\mathbf{x} - \mathbf{x}^*\|_2 \leq C_0(n-k)k^{-1/2}\beta\|\boldsymbol{\alpha}\|_q + C_1\epsilon. \quad \textit{(noisy)}$$

- Message: In the presence of basis mismatch, *exact or near-exact sparse recovery cannot be guaranteed*. Recovery *may* suffer large errors.

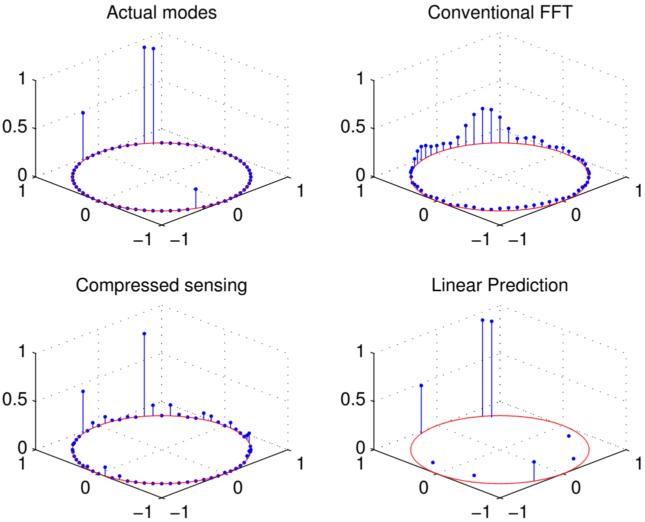
Y. Chi, L.L. Scharf, A. Pezeshki, and A.R. Calderbank, "Sensitivity to basis mismatch in compressed sensing," *IEEE Transactions on Signal Processing*, vol. 59, no. 5, pp. 2182–2195, May 2011.

Mismatch in Modal Analysis



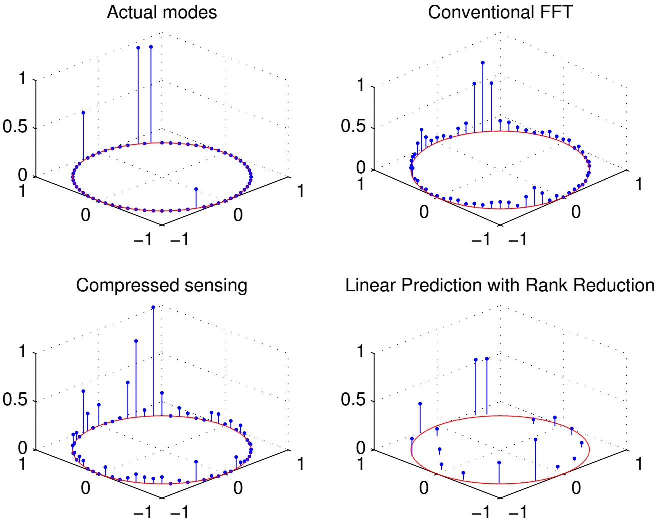
Frequency mismatch

Mismatch in Modal Analysis



Damping mismatch

Mismatch in Modal Analysis



Frequency mismatch–noisy measurements

Mismatch in Modal Analysis

But what if we make the grid finer and finer?

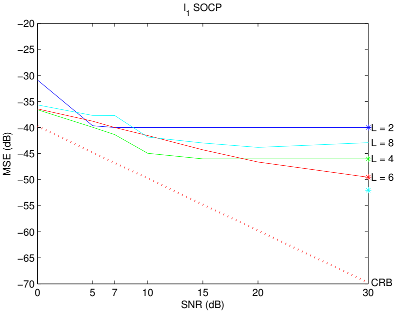
- Over-resolution experiment:
  - $m = 25$ samples
  - Equal amplitude complex tones at $f_1 = 0.5$ Hz and $f_2 = 0.52$ Hz (half the Rayleigh limit apart), mismatched to mathematical basis.
  - Mathematical model is $\mathbf{s} = \boldsymbol{\Psi}_{cs}\mathbf{x}$, where $\boldsymbol{\Psi}_{cs}$ is the $m \times n$, with $n = mL$, "DFT" frame that is over-resolved to $\Delta f = 1/mL$.

$$\boldsymbol{\Psi}_{cs} = \frac{1}{\sqrt{m}} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & e^{j\frac{2\pi}{mL}} & \cdots & e^{j\frac{2\pi(mL-1)}{mL}} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & e^{j\frac{2\pi(m-1)}{mL}} & \cdots & e^{j\frac{2\pi(m-1)(mL-1)}{mL}} \end{bmatrix}.$$

- What we will see:
  - MSE of inversion is noise-defeated, noise-limited, quantization limited, or null-space limited—depending on SNR.

L. L. Scharf, E. K. P. Chong, A. Pezeshki, and J. R. Luo, "Sensitivity considerations in compressed sensing," *in Conf. Rec. Asilomar'11,* Pacific Grove, CA,, Nov. 2011, pp. 744–748.

Noise Limited, Quantization Limited, or Null Space Limited

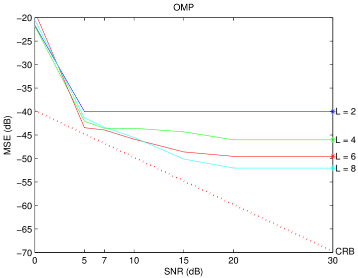


ℓ_1 inversions for $L = 2, 4, 6, 8$

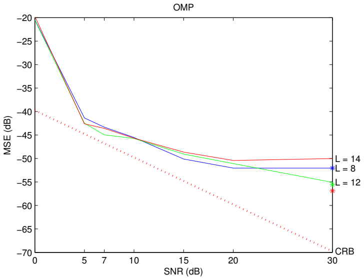
- From noise-defeated to noise-limited to quantization-limited to null-space limited.
- Results are actually too optimistic. For a weak mode in the presence of a strong interfering mode, the results are worse.

L. L. Scharf, E. K. P. Chong, A. Pezeshki, and J. R. Luo, "Sensitivity considerations in compressed sensing," *in Conf. Rec. Asilomar'11,* Pacific Grove, CA,, Nov. 2011, pp. 744–748.

Noise Limited, Quantization Limited, or Null Space Limited



(a) OMP for $L = 2, 4, 6, 8$



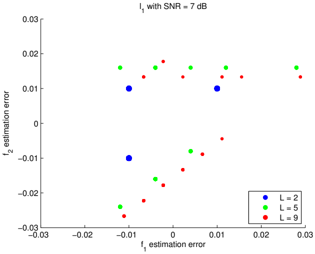
(b) OMP for $L = 8, 12, 14$

- Again, from noise-defeated to noise-limited to quantization-limited to null-space limited.
- Again, for a weak mode in the presence of a strong interfering mode, the results are much worse.

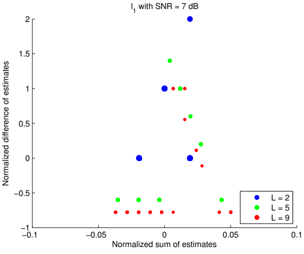
L. L. Scharf, E. K. P. Chong, A. Pezeshki, and J. R. Luo, "Sensitivity considerations in compressed sensing," *in Conf. Rec. Asilomar'11,* Pacific Grove, CA,, Nov. 2011, pp. 744–748.

Scatter Plots for BPDN Estimates

- Scatter plots for the normalized errors in estimating the sum and difference frequencies using BPDN.



(a) $(f_1 + f_2)$

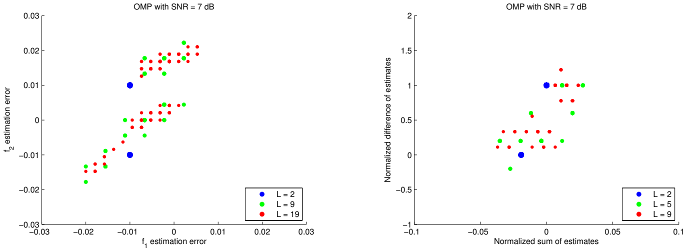


(b) $(f_1 - f_2)$

- At $L = 2$ mean-squared error is essentially bias-squared, whereas for $L = 9$ it is essentially variance.
- Average frequency is easy to estimate, but the difference frequency is hard to estimate. (Vertical scale is nearly 10 times the horizontal scale.)
- BPDN favors large negative differences over large positive differences (better estimates the mode at f_1 than it estimates the mode at f_2).

Scatter Plots for OMP Estimates

- Scatter plots for the normalized errors in estimating the sum and difference frequencies using BPDN.



(a) $(f_1 + f_2)$ (b) $(f_1 - f_2)$

- Preference for large negative errors in estimating the difference frequency disappears.
- Correlation between sum and difference errors reflects the fact that a large error in extracting the first mode will produce a large error in extracting the second.

References on Model Mismatch in CS

- Y. Chi, A. Pezeshki, L. L. Scharf, and R. Calderbank, "Sensitivity to basis mismatch in compressed sensing," *in Proc. ICASSP'10,* Dallas, TX, Mar. 2010, pp. 3930 –3933.
- Y. Chi, L.L. Scharf, A. Pezeshki, and A.R. Calderbank, "Sensitivity to basis mismatch in compressed sensing," *IEEE Transactions on Signal Processing,* vol. 59, no. 5, pp. 2182–2195, May 2011.
- L. L. Scharf, E. K. P. Chong, A. Pezeshki, and J. R. Luo, "Compressive sensing and sparse inversion in signal processing: Cautionary notes," *in Proc. DASP'11,* Coolum, Queensland, Australia, Jul. 10-14, 2011.
- L. L. Scharf, E. K. P. Chong, A. Pezeshki, and J. R. Luo, "Sensitivity considerations in compressed sensing," *in Conf. Rec. Asilomar'11,* Pacific Grove, CA,, Nov. 2011, pp. 744–748.
- M. A. Herman and T. Strohmer, "General deviants: An analysis of perturbations in compressed sensing," *IEEE J. Selected Topics in Signal Processing,* vol. 4, no. 2, pp. 342349, Apr. 2010.
- D. H Chae, P. Sadeghi, and R. A. Kennedy, "Effects of basis-mismatch in compressive sampling of continuous sinusoidal signals," *Proc. Int. Conf. on Future Computer and Commun.*, Wuhan, China, May 2010.

Remedies to Basis Mismatch : A Partial List

These approaches still assume a grid.

- H. Zhu, G. Leus, and G. B. Giannakis, "Sparsity-cognizant total least-squares for perturbed compressive sampling," *IEEE Transactions on Signal Processing*, vol. 59, May 2011.
- M. F. Duarte and R. G. Baraniuk, "Spectral compressive sensing," *Applied and Computational Harmonic Analysis*, Vol. 35, No. 1, pp. 111-129, 2013.
- A. Fannjiang and W. Liao, "Coherence-Pattern Guided Compressive Sensing with Unresolved Grids," *SIAM Journal of Imaging Sciences*, Vol. 5, No. 1, pp. 179-202, 2012.

Intermediate Recap: Sensitivity of CS to Basis Mismatch

- Basis mismatch is inevitable and sensitivities of CS to basis mismatch need to be fully understood. No matter how finely we grid the parameter space, the actual modes almost never lie on the grid.

- The consequence of over-resolution (very fine gridding) is that performance follows the Cramer-Rao bound more closely at low SNR, but at high SNR it departs more dramatically from the Cramer-Rao bound.

- This matches intuition that has been gained from more conventional modal analysis where there is a qualitatively similar trade-off between bias and variance. That is, bias may be reduced with frame expansion (over-resolution), but there is a penalty to be paid in variance.

Outline

Going Off the Grid

We will discuss two recent approaches that allow for parameter estimation without discretization with theoretical guarantees.

- Atomic Norm Minimization by [Tang et. al., 2012]:
  - Tightest convex relaxation to recover the spectral sparse signals;
  - $\Theta(r \log r \log n)$ samples are sufficient to guarantee exact recovery if the frequencies are well-separated by about 4RL;
  - Extensions to multi-dimensional frequencies and multiple measurement vector models.
- Enhanced Matrix Completion by [Chen and Chi, 2013]:
  - take advantage of shift invariance of harmonics and reformulate the problem into completion of a matrix pencil.
  - $\Theta(r \log^3 n)$ samples are sufficient to guarantee exact recovery if the Gram matrix formed by sampling the Dirichlet kernel at pairwise frequency separations are well-conditioned.
  - work with multi-dimensional frequencies.

G. Tang; Bhaskar, B.N.; Shah, P.; Recht, B., "Compressed Sensing Off the Grid," Information Theory, IEEE Transactions on , vol.59, no.11, pp.7465,7490, Nov. 2013.

Y. Chen and Y. Chi, "Robust Spectral Compressed Sensing via Structured Matrix Completion," IEEE Trans. on Information Theory, Apr. 2013, in revision.

Outline

The Atomic Norm Approach

The **atomic norm** is proposed to find tightest convex relaxations of general parsimonious models including sparse signals as a special case. The prescribed recipe is:

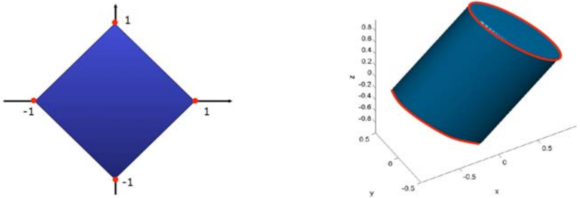
- Step 1: assume the signal of interest can be written as a superposition of small numbers of *atoms* in $\mathcal{A}$:
$$x = \sum_{i=1}^{r} c_i a_i, \quad a_i \in \mathcal{A}, \ c_i > 0.$$
- Step 2: define the atomic norm of the signal as:
$$\|x\|_{\mathcal{A}} = \inf \{t > 0 : x \in t\mathrm{conv}(\mathcal{A})\}$$
$$= \inf \left\{ \sum_i c_i \middle| x = \sum_i c_i a_i, \ a_i \in \mathcal{A}, c_i > 0 \right\}.$$
- Step 3: formulate a convex program to minimize the atomic norm with respect to the measurements.

Chandrasekaran, V., B. Recht, P. A. Parrilo, and A. S. Willsky. "The convex geometry of linear inverse problems." Foundations of Computational Mathematics 12, no. 6 (2012): 805-849.

Examples: The Atomic Norm Approach

Several popular approaches become special cases of the atomic norm minimization framework.

- Sparse signals: *an atom* for *sparse signals* is a normalized vector of sparsity one, and the atomic norm is ℓ_1 norm;
- Low-rank matrices: *an atom* for *low-rank matrices* is a normalized rank-one matrix; and the atomic norm is nuclear norm;



(a) unit ball of ℓ_1 norm (b) unit ball of nuclear norm

Chandrasekaran, V., B. Recht, P. A. Parrilo, and A. S. Willsky. "The convex geometry of linear inverse problems." Foundations of Computational Mathematics 12, no. 6 (2012): 805-849.

Atomic Norm For Spectrally-Sparse Signals

- Let $x(t) = \sum_{i=1}^{r} d_i e^{j2\pi f_i t}$, $f_i \in [0,1)$, $t = 0, \ldots, n-1$. Denote $\mathcal{F} = \{f_i\}_{i=1}^{r}$.
- Stack $x(t)$ into a vector $\mathbf{x}$: $\mathbf{x} = \sum_{i=1}^{r} d_i \mathbf{a}(f_i)$, $d_i \in \mathbb{C}$ where $\mathbf{a}(f)$ is the atom defined as

$$\mathbf{a}(f) = \frac{1}{\sqrt{n}} \begin{bmatrix} 1 & e^{j2\pi f} & \ldots & e^{j2\pi f(n-1)} \end{bmatrix}.$$

- Atomic norm:

$$\begin{aligned} \|\mathbf{x}\|_{\mathcal{A}} &= \inf \left\{ \sum_k |c_k| \,\middle|\, \mathbf{x} = \sum_k c_k \mathbf{a}(f_k), f_k \in [0,1) \right\} \\ &= \inf_{\mathbf{u},t} \left\{ \frac{1}{2}\mathrm{Tr}(\mathrm{toep}(\mathbf{u})) + \frac{1}{2} t \,\middle|\, \begin{bmatrix} \mathrm{toep}(\mathbf{u}) & \mathbf{x} \\ \mathbf{x}^* & t \end{bmatrix} \succ \mathbf{0} \right\}. \end{aligned}$$

which can be equivalently given in an SDP form.

G. Tang; Bhaskar, B.N.; Shah, P.; Recht, B., "Compressed Sensing Off the Grid," Information Theory, IEEE Transactions on , vol.59, no.11, pp.7465,7490, Nov. 2013.

Spectral Compressed Sensing with Atomic Norm

- Random Subsampling: We observe a subset Ω of entries of $\mathbf{x}$ *uniformly at random*:

$$\mathbf{x}_\Omega = \mathcal{P}_\Omega(\mathbf{x}).$$

- Atomic Norm Minimization:

$$\min_{\mathbf{s}} \ \|\mathbf{s}\|_{\mathcal{A}} \quad \text{subject to} \quad \mathbf{s}_\Omega = \mathbf{x}_\Omega,$$

- It can be solved efficiently using off-the-shelf SDP solvers.

$$\begin{aligned} \text{(Primal:)} \quad & \min_{\mathbf{u},\mathbf{s},t} \ \frac{1}{2}\mathrm{Tr}(\mathrm{toep}(\mathbf{u})) + \frac{1}{2}t \\ \text{subject to} \quad & \begin{bmatrix} \mathrm{toep}(\mathbf{u}) & \mathbf{s} \\ \mathbf{s}^* & t \end{bmatrix} \succeq \mathbf{0}, \\ & \mathbf{s}_\Omega = \mathbf{x}_\Omega. \end{aligned}$$

G. Tang; Bhaskar, B.N.; Shah, P.; Recht, B., "Compressed Sensing Off the Grid," Information Theory, IEEE Transactions on , vol.59, no.11, pp.7465,7490, Nov. 2013.

Recovering the Frequencies via Dual Polynomial

- The dual problem can be used to recover the frequencies:

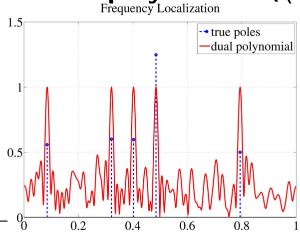
$$\text{(Dual:)} \quad \max_{\mathbf{q}} \langle \mathbf{q}_\Omega, \mathbf{x}_\Omega \rangle_{\mathbb{R}} \quad \text{subject to} \quad \|\mathbf{q}\|_{\mathcal{A}}^* \leq 1, \quad \mathbf{q}_{\Omega^c} = \mathbf{0}.$$

where

$$\|\mathbf{q}\|_{\mathcal{A}}^* = \sup_{f \in [0,1)} |\langle \mathbf{q}, \mathbf{a}(f) \rangle| := \sup_{f \in [0,1)} Q(f)$$

- The primal problem is optimal when the **dual polynomial** $Q(f)$ satisfies:

$$\begin{cases} Q(f_i) = \text{sign}(d_i), & f_i \in \mathcal{F} \\ |Q(f)| < 1, & f \notin \mathcal{F} \\ \mathbf{q}_{\Omega^c} = \mathbf{0} \end{cases}$$



- This is also stable under noise.

G. Tang; Bhaskar, B.N.; Shah, P.; Recht, B., "Compressed Sensing Off the Grid," Information Theory, IEEE Transactions on , vol.59, no.11, pp.7465,7490, Nov. 2013.

Performance Guarantees for Noiseless Recovery

Theorem (Tang et. al., 2012)

Suppose a subset of m entries are observed uniformly at random. Additionally, assume the phase of the coefficients are drawn i.i.d. from the uniform distribution on the complex unit circle and

$$\Delta = \min_{f_j \neq f_k} |f_j - f_k| \geq \frac{1}{\lfloor (n-1)/4 \rfloor},$$

then $m \geq C \left\{ \log^2 \frac{n}{\delta}, r \log \frac{n}{\delta} \log \frac{r}{\delta} \right\}$ is sufficient to guarantee exact recovery with probability at least $1-\delta$ with respect to the random samples and signs, where C is some numerical constant.

- Random data model, and random observation model.
- $m = \Theta(r \log n \log r)$ samples suffice if a separation condition of about 4RL is satisfied.

G. Tang; Bhaskar, B.N.; Shah, P.; Recht, B., "Compressed Sensing Off the Grid," Information Theory, IEEE Transactions on , vol.59, no.11, pp.7465,7490, Nov. 2013.

Phase Transition

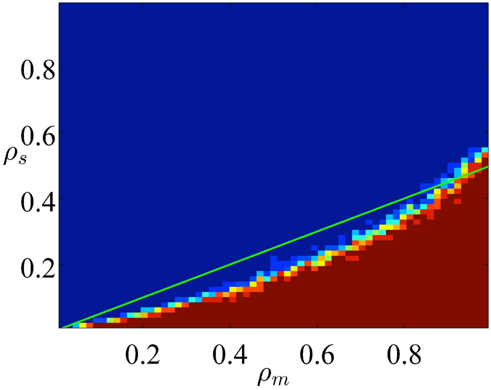


Figure : Phase transition diagrams when $n = 128$ and the separation is set to be 1.5 RL. Both signs and magnitudes of the coeffients are random.

G. Tang; Bhaskar, B.N.; Shah, P.; Recht, B., "Compressed Sensing Off the Grid," Information Theory, IEEE Transactions on , vol.59, no.11, pp.7465,7490, Nov. 2013.

Extension for MMV Models

- For multiple spectrally-sparse signals $\mathbf{X} = \sum_{i=1}^{r} \mathbf{a}(f_i)\mathbf{b}_i^* \in \mathbb{C}^{n\times L}$, we define the atomic set $\mathcal{A}$ composed of atoms as

$$\mathbf{A}(f, \mathbf{b}) = \mathbf{a}(f)\mathbf{b}^* \in \mathbb{C}^{n\times L}, \quad \|\mathbf{b}\|_2 = 1.$$

- The atomic norm is defined and computed as [Chi, 2013]

$$\begin{aligned}\|\mathbf{X}\|_{\mathcal{A}} &= \inf \left\{ \sum_k c_k \middle| \mathbf{X} = \sum_k c_k \mathbf{A}(f_k, \mathbf{b}_k), c_k \geq 0 \right\} \\ &= \inf_{\mathbf{u},\mathbf{W}} \left\{ \frac{1}{2}\mathrm{Tr}(\mathrm{toep}(\mathbf{u})) + \frac{1}{2}\mathrm{Tr}(\mathbf{W}) \middle| \begin{bmatrix} \mathrm{toep}(\mathbf{u}) & \mathbf{X} \\ \mathbf{X}^* & \mathbf{W} \end{bmatrix} \succ \mathbf{0} \right\}.\end{aligned}$$

- The single vector case becomes a special case when $L = 1$. The algorithm is tractable however the complexity might become high when L is large.

Y. Chi, "Joint Sparsity Recovery for Spectral Compressed Sensing", in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, May 2014.

Two-Dimensional Frequency Model

- Stack $x(\mathbf{t}) = \sum_{i=1}^{r} d_i e^{j2\pi\langle \mathbf{t}, \mathbf{f}_i \rangle}$ into a matrix $\mathbf{X} \in \mathbb{C}^{n_1 \times n_2}$.
- The matrix $\mathbf{X}$ has the following **Vandermonde decomposition**:

$$\mathbf{X} = \mathbf{Y} \cdot \underbrace{\mathbf{D}}_{\text{diagonal matrix}} \cdot \mathbf{Z}^T.$$

Here, $\mathbf{D} := \text{diag}\{d_1, \cdots, d_r\}$ and

$$\mathbf{Y} := \underbrace{\begin{bmatrix} 1 & 1 & \cdots & 1 \\ y_1 & y_2 & \cdots & y_r \\ \vdots & \vdots & \vdots & \vdots \\ y_1^{n_1-1} & y_2^{n_1-1} & \cdots & y_r^{n_1-1} \end{bmatrix}}_{\text{Vandemonde matrix}}, \mathbf{Z} := \underbrace{\begin{bmatrix} 1 & 1 & \cdots & 1 \\ z_1 & z_2 & \cdots & z_r \\ \vdots & \vdots & \vdots & \vdots \\ z_1^{n_2-1} & z_2^{n_2-1} & \cdots & z_r^{n_2-1} \end{bmatrix}}_{\text{Vandemonde matrix}}$$

where $y_i = \exp(j2\pi f_{1i})$, $z_i = \exp(j2\pi f_{2i})$, $\mathbf{f}_i = (f_{1i}, f_{2i})$.

- Goal: We observe a *random subset of entries* of $\mathbf{X}$, and wish to recover the missing entries.

Extension for Two-dimensional Frequencies

- The atomic norm can be similarly defined for two-dimensional frequencies and similar sample complexity holds [Chi and Chen, 2013].
- However, the atomic norm doesn't have a simple equivalent SDP form as in 1D since the Vandermonde decomposition lemma doesn't hold for two-dimensional frequencies.
- The exact SDP characterization is studied by [Xu et. al., 2014].

Y. Chi and Y. Chen, "Compressive Recovery of 2-D Off-Grid Frequencies," in *Asilomar Conference on Signals, Systems, and Computers (Asilomar)*, Pacific Grove, CA, Nov. 2013.

Xu, Weiyu, et al. "Precise semidefinite programming formulation of atomic norm minimization for recovering d-dimensional ($D \geq 2$) off-the-grid frequencies." Information Theory and Applications Workshop (ITA), 2014.

Just Discretize?

- Consider a fine discretization of the parameter space

$$\mathcal{F}_m = \{\omega_1, \ldots, \omega_m\} \subset [0, 1)$$

and $\mathbf{A}^m = [\mathbf{a}(\omega_1), \ldots, \mathbf{a}(\omega_m)]$, a discrete approximation of the atomic norm minimization is

$$\text{(Primal-discrete:)} \min_{\mathbf{c}_m} \|\mathbf{c}_m\|_1 \quad \text{s.t. } \mathbf{A}_\Omega^m \mathbf{c}_m = \mathbf{x}_\Omega$$

$$\text{(Dual-discrete:)} \max_{\mathbf{q}} \langle \mathbf{q}_\Omega, \mathbf{x}_\Omega \rangle_{\mathbb{R}} \quad \text{s.t. } |\langle \mathbf{q}, \mathbf{a}(\omega_m) \rangle| \leq 1, i = 1, \ldots, m;$$

$$\mathbf{q}_{\Omega^c} = \mathbf{0}.$$

- Under mild technical conditions, the solution to the discrete approximation converges to that of the atomic norm minimization as the discretization gets finer.

Tang, G., B. N. Bhaskar, and B. Recht. "Sparse recovery over continuous dictionaries-just discretize." Signals, Systems and Computers, 2013 Asilomar Conference on. IEEE, 2013..

Outline

Matrix Completion?

recall that $\mathbf{X} = \underbrace{\mathbf{Y}}_{\textbf{Vandemonde}} \cdot \underbrace{\mathbf{D}}_{\textbf{diagonal}} \cdot \underbrace{\mathbf{Z}^T}_{\textbf{Vandemonde}}$.

where $\mathbf{D} := \text{diag}\{d_1, \cdots, d_r\}$, and

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

- Quick observation: $\mathbf{X}$ is a low-rank matrix with $\text{rank}(\mathbf{X}) = r$.
- Quick idea: can we apply *Matrix Completion* algorithms on $\mathbf{X}$?

$$\begin{bmatrix} \surd & ? & ? & \surd & \surd \\ ? & \surd & ? & \surd & \surd \\ ? & ? & \surd & \surd & ? \\ \surd & \surd & \surd & \surd & ? \\ \surd & \surd & ? & ? & \surd \end{bmatrix}$$

Matrix Completion

- Matrix Completion can be thought as an extension of CS to low-rank matrices.
- The Netflix problem: Let $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$ satisfying $\text{rank}(\mathbf{X}) = r$.

$$\text{users} \quad \overset{\text{movies}}{\begin{bmatrix} \surd & ? & ? & \surd & \surd \\ ? & \surd & ? & \surd & \surd \\ ? & ? & \surd & \surd & ? \\ \surd & \surd & \surd & \surd & ? \\ \surd & \surd & ? & ? & \surd \end{bmatrix}}$$

- Given the set of observations $\mathcal{P}_\Omega(\mathbf{X})$, find the matrix with the smallest rank that satisfies the observations:

$$\begin{aligned} \underset{\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}}{\text{minimize}} \quad & \text{rank}(\mathbf{M}) \\ \text{subject to} \quad & \mathcal{P}_\Omega\left(\mathbf{M}\right) = \mathcal{P}_\Omega\left(\mathbf{X}\right), \end{aligned}$$

Solve MC via Nuclear Norm Minimization

- Relax the rank minimization problem to a convex optimization:

$$\underset{\mathbf{M}\in\mathbb{R}^{n_1\times n_2}}{\text{minimize}} \quad \|\mathbf{M}\|_*$$
$$\text{subject to} \quad \mathcal{P}_\Omega(\mathbf{M}) = \mathcal{P}_\Omega(\mathbf{X}),$$

  where $\|\mathbf{M}\|_* = \text{sum}(\sigma_i(\mathbf{M}))$.
- Coherence measure: Let the SVD of $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$. Define the coherence measure

$$\max_{1\le i\le n_1} \|\mathbf{U}^T\mathbf{e}_i\|_2 \le \sqrt{\frac{\mu n_1}{r}}, \quad \max_{1\le i\le n_2} \|\mathbf{V}^T\mathbf{e}_i\|_2 \le \sqrt{\frac{\mu n_2}{r}}.$$

  where $\mathbf{e}_i$'s are standard basis vectors.
  - $1 \le \mu \le \max(n_1, n_2)/r$;
  - subspace with low coherence: e.g. all one vectors;
  - subspace with high coherence: e.g. standard basis vectors.

Candès, E. J., and B. Recht. "Exact matrix completion via convex optimization." Foundations of Computational mathematics 9.6 (2009): 717-772.
Candès, E. J., and T. Tao. "The power of convex relaxation: Near-optimal matrix completion." Information Theory, IEEE Transactions on 56.5 (2010): 2053-2080.

Performance Guarantees and Its Implication

Theorem (Candès and Recht 2009, Gross 2010, Chen 2013)

Assume we collect $m = |\Omega|$ *samples of* $\mathbf{X}$ *uniformly at random. Let* $n = \max\{n_1, n_2\}$. *Then the nuclear norm minimization algorithm recovers* $\mathbf{X}$ *exactly with high probability if*

$$m > C\mu rn \log^2 n$$

where C *is some universal constant.*

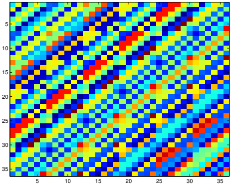
- **Implication on our problem:** Can we apply *Matrix Completion* algorithms on the two-dimensional frequency data matrix $\mathbf{X}$?
  - Yes, but it yields sub-optimal performance. It requires at least $r \max\{n_1, n_2\}$ samples.
  - No, $\mathbf{X}$ is no longer low-rank if $r > \min(n_1, n_2)$. Note that r can be as large as $n_1 n_2$

Candès, E. J., and B. Recht. "Exact matrix completion via convex optimization." Foundations of Computational mathematics 9.6 (2009): 717-772.

Gross, David. "Recovering low-rank matrices from few coefficients in any basis." Information Theory, IEEE Transactions on 57, no. 3 (2011): 1548-1566.

Chen, Yudong. "Incoherence-Optimal Matrix Completion." arXiv preprint arXiv:1310.0154 (2013).

Revisiting Matrix Pencil: Matrix Enhancement



Given a data matrix $\mathbf{X}$, Hua proposed the following matrix enhancement for two-dimensional frequency models:

- Choose two pencil parameters k_1 and k_2;
- An **enhanced form** $\mathbf{X}_e$ is an $k_1 \times (n_1 - k_1 + 1)$ *block Hankel matrix* :

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

where each block is a $k_2 \times (n_2 - k_2 + 1)$ *Hankel* matrix as follows

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

Low Rankness of the Enhanced Matrix

- Choose pencil parameters $k_1 = \Theta(n_1)$ and $k_2 = \Theta(n_2)$, the dimensionality of X_e is proportional to $n_1 n_2 \times n_1 n_2$.
- The enhanced matrix can be decomposed as follows:

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

  - $\mathbf{Z}_\mathrm{L}$ and $\mathbf{Z}_\mathrm{R}$ are Vandermonde matrices specified by $z_1, \ldots, z_r$,
  - $\mathbf{Y}_\mathrm{d} = \mathrm{diag}\left[y_1, y_2, \cdots, y_r\right]$.
- The enhanced form $\mathbf{X}_\mathrm{e}$ is low-rank.
  - $\mathrm{rank}\left(\mathbf{X}_\mathrm{e}\right) \leq r$
  - Spectral Sparsity $\Rightarrow$ Low Rankness
- holds even with damping modes.

Hua, Yingbo. "Estimating two-dimensional frequencies by matrix enhancement and matrix pencil." Signal Processing, IEEE Transactions on 40, no. 9 (1992): 2267-2280.

Enhanced Matrix Completion (EMaC)

- The natural algorithm is to find the enhanced matrix with the minimal rank satisfying the measurements:

$$\begin{aligned} \underset{\mathbf{M}\in\mathbb{C}^{n_1\times n_2}}{\text{minimize}} \quad & \text{rank}\left(\mathbf{M}_e\right) \\ \text{subject to} \quad & \mathbf{M}_{i,j} = \mathbf{X}_{i,j}, \forall (i,j)\in\Omega \end{aligned}$$

  where Ω denotes the sampling set.
- Motivated by Matrix Completion, we will solve its convex relaxation:

$$\text{(EMaC)}: \quad \begin{aligned} \underset{\mathbf{M}\in\mathbb{C}^{n_1\times n_2}}{\text{minimize}} \quad & \|\mathbf{M}_e\|_* \\ \text{subject to} \quad & \mathbf{M}_{i,j} = \mathbf{X}_{i,j}, \forall (i,j)\in\Omega \end{aligned}$$

  where $\|\cdot\|_*$ denotes the nuclear norm.
- The algorithm is referred to as *Enhanced Matrix Completion (EMaC)*.

Enhanced Matrix Completion (EMaC)

$$
\text{(EMaC)}: \quad \underset{\mathbf{M}\in\mathbb{C}^{n_1\times n_2}}{\text{minimize}} \quad \|\mathbf{M}_{\mathrm{e}}\|_* \qquad \text{subject to} \quad \mathbf{M}_{i,j} = \mathbf{X}_{i,j}, \forall (i,j)\in\Omega
$$

- existing MC result won't apply – requires at least $\mathcal{O}(nr)$ samples
- **Question**: How many samples do we need?

$$
\begin{bmatrix}
\surd & ? & ? & \surd & \surd \\
? & \surd & ? & \surd & \surd \\
? & ? & \surd & \surd & ? \\
\surd & \surd & \surd & \surd & ? \\
\surd & \surd & ? & ? & \surd
\end{bmatrix}
\qquad
\begin{bmatrix}
? & \surd & \surd & ? & \surd & ? & \surd & \surd & ? & ? & \surd & \surd \\
\surd & \surd & ? & ? & ? & \surd & \surd & ? & ? & \surd & \surd & ? \\
\surd & ? & ? & \surd & \surd & \surd & ? & \surd & \surd & \surd & ? & \surd \\
? & ? & \surd & ? & \surd & ? & \surd & ? & \surd & ? & \surd & \surd \\
\surd & ? & \surd & \surd & ? & ? & \surd & \surd & \surd & ? & \surd & \surd \\
? & \surd & \surd & ? & ? & \surd & \surd & ? & ? & \surd & \surd & ? \\
\surd & \surd & ? & \surd & \surd & \surd & ? & \surd & \surd & \surd & ? & \surd \\
\surd & ? & \surd & ? & \surd & ? & \surd & \surd & \surd & ? & \surd & ? \\
? & ? & \surd & \surd & \surd & ? & \surd & \surd & ? & ? & \surd & ? \\
? & \surd & \surd & ? & ? & \surd & \surd & ? & ? & \surd & ? & ? \\
\surd & \surd & ? & \surd & \surd & \surd & ? & \surd & \surd & ? & ? & \surd \\
\surd & ? & \surd & \surd & \surd & ? & \surd & ? & ? & ? & \surd & ?
\end{bmatrix}
$$

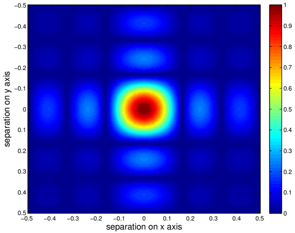
Introduce Coherence Measure

- Define the 2-D Dirichlet kernel:

$$\mathcal{K}(k_1, k_2, f_1, f_2) := \frac{1}{k_1 k_2}\left(\frac{1-e^{-j2\pi k_1 f_1}}{1-e^{-j2\pi f_1}}\right)\left(\frac{1-e^{-j2\pi k_2 f_2}}{1-e^{-j2\pi f_2}}\right),$$

- Define $\mathbf{G}_{\mathrm{L}}$ and $\mathbf{G}_{\mathrm{R}}$ as $r \times r$ Gram matrices such that

$$(\mathbf{G}_{\mathrm{L}})_{i,l} = \mathcal{K}(k_1, k_2, f_{1i} - f_{1l}, f_{2i} - f_{2l}),$$
$$(\mathbf{G}_{\mathrm{R}})_{i,l} = \mathcal{K}(n_1 - k_1 + 1, n_2 - k_2 + 1, f_{1i} - f_{1l}, f_{2i} - f_{2l}).$$

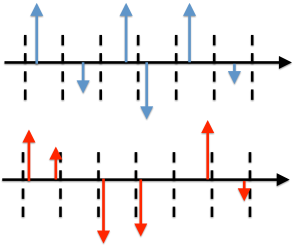


Introduce Coherence Measure

- **Incoherence condition** holds w.r.t. μ if

$$\sigma_{\min}\left(\mathbf{G}_L\right) \geq \frac{1}{\mu}, \quad \sigma_{\min}\left(\mathbf{G}_R\right) \geq \frac{1}{\mu}.$$

- $\mu = \Theta(1)$ holds under many scenarios:
  - Randomly generated frequencies;
  - Mild perturbation of grid points;
  - In 1D, well-separated frequencies by 2 times RL [Liao and Fannjiang, 2014].



W. Liao and A. Fannjiang. "MUSIC for Single-Snapshot Spectral Estimation: Stability and Super-resolution." arXiv preprint arXiv:1404.1484 (2014).

Theoretical Guarantees for Noiseless Case

Theorem (Chen and Chi, 2013)

Let $n = n_1 n_2$. If all measurements are noiseless, then EMaC recovers $\mathbf{X}$ *perfectly with high probability if*

$$m > C\mu r \log^3 n.$$

where C is some universal constant.

- deterministic signal model, random observation;
- coherence condition μ only depends on the frequencies but the amplitudes.
- near-optimal within logarithmic factors: $\Theta(r \log^3 n)$.
- general theoretical guarantees for Hankel (Toeplitz) matrix completion, which are useful for applications in control, MRI, natural language processing, etc.

Y. Chen and Y. Chi, "Robust Spectral Compressed Sensing via Structured Matrix Completion," IEEE Trans. on Information Theory, Apr. 2013, in revision.

Phase Transition

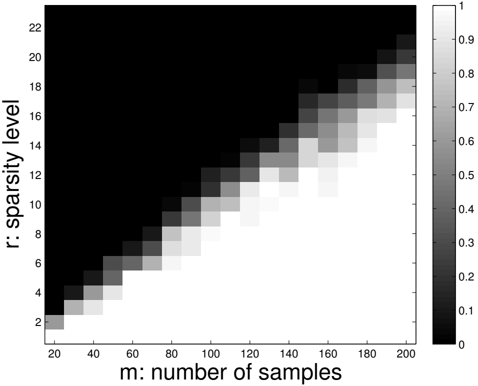


Figure : Phase transition diagrams where spike locations are randomly generated. The results are shown for the case where $n_1 = n_2 = 15$.

Robustness to Bounded Noise

Assume the samples are noisy $\mathbf{X} = \mathbf{X}^o + \mathbf{N}$, where $\mathbf{N}$ is bounded noise:

$$\text{(EMaC-Noisy)}: \quad \underset{\mathbf{M}\in\mathbb{C}^{n_1\times n_2}}{\text{minimize}} \quad \|\mathbf{M}_\mathrm{e}\|_* \qquad \text{subject to} \quad \|\mathcal{P}_\Omega(\mathbf{M}-\mathbf{X})\|_\mathrm{F} \le \delta,$$

Theorem (Chen and Chi, 2013)

Suppose $\mathbf{X}^\mathrm{o}$ *satisfies* $\|\mathcal{P}_\Omega(\mathbf{X}-\mathbf{X}^\mathrm{o})\|_F \le \delta$. *Under the conditions of Theorem 1, the solution to* **EMaC-Noisy** *satisfies*

$$\|\hat{\mathbf{X}}_\mathrm{e} - \mathbf{X}_\mathrm{e}\|_F \le \left\{ 2\sqrt{n} + 8n + \frac{8\sqrt{2}n^2}{m} \right\} \delta$$

with probability exceeding $1 - n^{-2}$.

- The average entry inaccuracy is bounded above by $\mathcal{O}(\frac{n}{m}\delta)$. In practice, EMaC-Noisy usually yields better estimate.

Y. Chen and Y. Chi, "Robust Spectral Compressed Sensing via Structured Matrix Completion," IEEE Trans. on Information Theory, Apr. 2013, in revision.

Singular Value Thresholding (Noisy Case)

- Several optimized solvers for Hankel matrix completion exist, for example [Fazel et. al. 2013, Liu and Vandenberghe 2009]

Algorithm 1 Singular Value Thresholding for EMaC

1: **initialize** Set $\mathbf{M}_0 = \mathbf{X}_\mathrm{e}$ and $t = 0$.
2: **repeat**
3: 1) $\mathbf{Q}_t \leftarrow \mathcal{D}_{\tau_t}(\mathbf{M}_t)$ (*singular-value thresholding*)
4: 2) $\mathbf{M}_t \leftarrow \mathrm{Hankel}_{\mathbf{X}_0}(\mathbf{Q}_t)$ (*projection onto a Hankel matrix consistent with observation*)
5: 3) $t \leftarrow t + 1$
6: **until** convergence

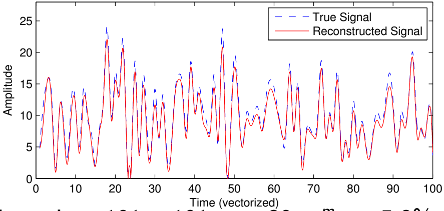
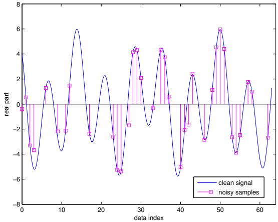


Figure : dimension: 101×101, $r = 30$, $\frac{m}{n_1 n_2} = 5.8\%$, SNR $= 10$dB.

- What if a constant portion of measurements are arbitrarily corrupted?

$$\mathbf{X}_{i,l}^{\text{corrupted}} = \mathbf{X}_{i,l} + \mathbf{S}_{i,l}$$



where $\mathbf{S}_{i,l}$ is of arbitrary amplitude.

- Reminiscent of the robust PCA approach [Candes et. al. 2011, Chandrasekaran et. al. 2011]
- Solve the following algorithm:

$$(\textbf{RobustEMaC}): \quad \underset{\mathbf{M},\mathbf{S}\in\mathbb{C}^{n_1\times n_2}}{\text{minimize}} \quad \|\mathbf{M}_e\|_* + \lambda\|\mathbf{S}_e\|_1$$
$$\text{subject to} \quad (\mathbf{M}+\mathbf{S})_{i,l} = \mathbf{X}_{i,l}^{\text{corrupted}}, \ \forall(i,l)\in\Omega$$

Theoretical Guarantees for Robust Recovery

$$(\textbf{RobustEMaC}): \quad \underset{\mathbf{M},\mathbf{S}\in\mathbb{C}^{n_1\times n_2}}{\text{minimize}} \quad \|\mathbf{M}_e\|_* + \lambda\|\mathbf{S}_e\|_1$$
$$\text{subject to} \quad (\mathbf{M}+\mathbf{S})_{i,l} = \mathbf{X}^{\text{corrupted}}_{i,l}, \ \forall (i,l)\in\Omega.$$

Theorem (Chen and Chi, 2013)

Assume the percent of corrupted entries is s is a small constant. Set $n = n_1 n_2$ and $\lambda = \frac{1}{\sqrt{m\log n}}$. Then RobustEMaC recovers $\mathbf{X}$ with high probability if

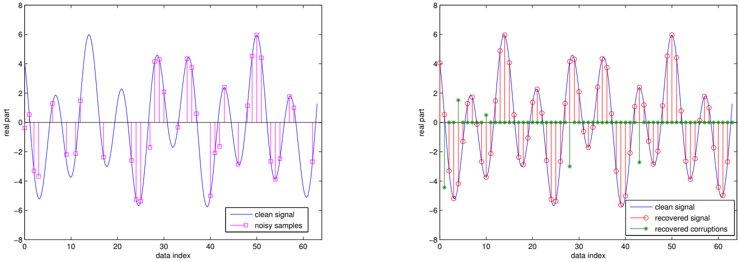
$$m > C\mu r^2 \log^3 n,$$

where C is some universal constant.

- Sample complexity: $m \sim \Theta(r^2\log^3 n)$, slight loss than the previous case;
- Robust to a constant portion of outliers: $s \sim \Theta(1)$

Y. Chen and Y. Chi, "Robust Spectral Compressed Sensing via Structured Matrix Completion," IEEE Trans. on Information Theory, Apr. 2013, in revision.

Robustness to Sparse Corruptions



(a) Observation (b) Recovery

Figure : Robustness to sparse corruptions: (a) Clean signal and its corrupted subsampled samples; (b) recovered signal and the sparse corruptions.

Phase Transition for Line Spectrum Estimation

Fix the amount of corruption as 10% of the total number of samples:

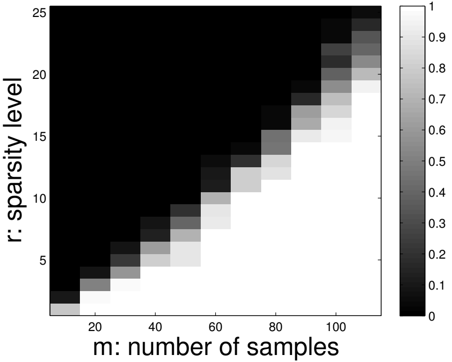


Figure : Phase transition diagrams where spike locations are randomly generated. The results are shown for the case where $n = 125$.

Comparisons between the two Approaches

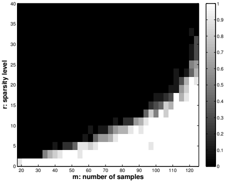
	EMaC	Atomic Norm
Signal model	Deterministic	Random
Observation model	Random	Random
Success Condition	Coherence	Separation condition
Amplitudes	No condition	Randomly generated
Sample Complexity	$\Theta(r \log^3 n)$	$\Theta(r \log r \log n)$
Bounded Noise	Yes	Not shown
Sparse Corruptions	Yes	Not shown

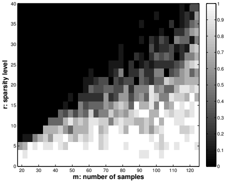
Comparisons of EMaC and Atomic Norm Minimization

Phase transition for line spectrum estimation: numerically, the EMaC approach seems less sensitive to the separation condition.

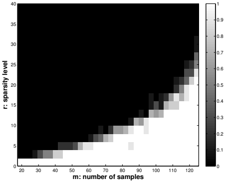
EMaC | Atomic Norm

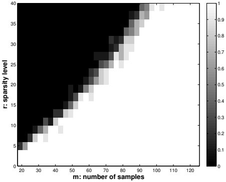
- without separation





- with 1.5 RL separation





References

- Chandrasekaran, Recht, Parrilo, Willsky (2010): general framework of atomic norm minimization.
- Tang, Bhaskar, Shah, Recht (2012): line spectrum estimation using atomic norm minimization with random sampling.
- Bhaskar, Tang, Recht (2012): line spectrum denoising using atomic norm minimization with consecutive samples.
- Candès and Fernandez-Granda (2012): Super-resolution using total variation minimization (equivalent to atomic norm) from low-pass samples.
- Chi (2013): line spectrum estimation using atomic norm minimization with multiple measurement vectors.
- Xu et. al. (2014): atomic norm minimization with prior information.
- Chen and Chi (2013): multi-dimensional frequency estimation via enhanced matrix completion.

References

- Xu et. al. (2013): exact SDP characterization of atomic norm minimization for high-dimensional frequencies.
- Tang et. al. (2013): near minimax line spectrum denoising via atomic norm minimization.
- Chi and Chen (2013): higher dimensional spectrum estimation using atomic norm minimization with random sampling.
- Hua (1992): matrix pencil formulation for multi-dimensional frequencies.
- Liao and Fannjiang (2014): analysis of the MUSIC algorithm with separation conditions.

Concluding Remarks

- Compression, whether by linear maps (e.g, Gaussian) or by subsampling, has performance consequences for parameter estimtion. Fisher information decreases, CRB increases, and the onset of breakdown threshold increases.
- Model mismatch is inevitable and can result in considerable performamce degredation, and therefore sensitivities of CS to model mismatch need to be fully understood.
- Recent off-the-grid methods provide a way forward for a class of problems, where modes to be estimated respect certain separation or coherence conditions. But sub-Rayleigh resolution still eludes us!

Thank You! Questions?