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

Yuejie Chi° and Ali Pezeshkić
°ECE and BMI, The Ohio State University
ćECE 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 $y \sim p(y; \theta)$, whose distribution is parameterized by unknown parameters $\theta$.

- **Inversion**: Estimate $\theta$, given a set of samples of $y$.
  - 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:** Superposition of modes:

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

- \( p = 2k \) unknown parameters: \( \theta = [\nu_1, \ldots, \nu_k, \alpha_1, \ldots, \alpha_k]^T \)

- 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

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

- Typically, \( t_i \)'s are uniformly spaced and almost always \( m \geq p \).
Parameter Estimation or Image Inversion

- **Canonical Model:**

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

- **DOA estimation and spectrum analysis:**

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

  where \(\nu\) 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

Review of Classical Parameter Estimation

Review of Compressive Sensing

Fundamental Limits of Subsampling on Parameter Estimation

Sensitivity of Basis Mismatch and Heuristic Remedies

Going off the Grid
   Atomic Norm Minimization
   Enhanced Matrix Completion
Outline

Review of Classical Parameter Estimation

Review of Compressive Sensing

Fundamental Limits of Subsampling on Parameter Estimation

Sensitivity of Basis Mismatch and Heuristic Remedies

Going off the Grid

Atomic Norm Minimization

Enhanced Matrix Completion
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 $\Delta \nu$.

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

- Peak locations are taken as estimates of $\nu_i$ 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.

ML Estimation in Separable Nonlinear Models

- Low-order separable modal representation for the image:

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

Parameters \( \nu \) in \( \Psi \) are nonlinear parameters (like frequency, delay, and Doppler) and \( \alpha \) are linear parameters (complex 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 \( \ell_2 \) minimization.
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 \( \omega_i \).

Uniformly sampled measurement model:

\[ y = \Psi(\nu) \alpha \]

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

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

\[
\min_{\nu, \alpha} \| y - \Psi(\nu) \alpha \|_2^2
\]

\[
\hat{\alpha}_{ML} = \Psi(\nu)^\dagger y
\]

\[
\hat{\nu}_{ML} = \arg\min y^H P_{A(\nu)} y; \quad A^H \Psi = 0
\]

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.


Example:

- Actual modes
- Linear Prediction
- Linear Prediction with Rank Reduction
Classical Parameter Estimation or Image Inversion

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

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

Outline

Review of Classical Parameter Estimation

Review of Compressive Sensing

Fundamental Limits of Subsampling on Parameter Estimation

Sensitivity of Basis Mismatch and Heuristic Remedies

Going off the Grid
  Atomic Norm Minimization
  Enhanced Matrix Completion
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.
Sparse Representation

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

$$s = \Psi x;$$

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

- Images are sparse in the wavelet domain.
Compressed Sensing aims to characterize attributes of a signal with a small number of measurements.

- **Incoherence Sampling:** the linear measurement $y \in \mathbb{C}^m$ is obtained via an *incoherent* matrix $\Phi \in \mathbb{C}^{m \times n}$, as
  $$y = \Phi s + n,$$
  where $m \ll n$. — subsampling.

- The goal is thus to recover $x$ from $y$. 

Compression on the Fly
Uniqueness of Sparse Recovery

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

$$(P0:) \min_x \|x\|_0 \text{ subject to } y = Ax.$$ 

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

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

Theorem (Uniqueness, Donoho and Elad 2002)

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

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

Sparse Recovery via \( \ell_1 \) Minimization

- The above \( \ell_0 \) minimization is NP-hard. A convex relaxation leads to the \( \ell_1 \) minimization, or basis pursuit:

\[
(P1:) \quad \min_x \|x\|_1 \quad \text{subject to} \quad y = Ax.
\]

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

Theorem (Equivalence, Donoho and Elad 2002)

A representation \( y = Ax \) is the unique solution to \((P1)\) if

\[
\|x\|_0 < \frac{1}{2} + \frac{1}{2\mu(A)}.
\]

---

When \( \|n\|_2 \leq \epsilon \), we incorporate this into the basis pursuit:

\[
(P1:) \quad x^* = \arg \min_x \|x\|_1 \quad \text{s.t.} \quad \|y - Ax\|_2 \leq \epsilon
\]

**Restricted Isometry Property:** If \( A \) satisfies the restricted isometry property (RIP) with \( \delta_{2k} \), then for any two \( k \)-sparse vectors \( x_1 \) and \( x_2 \):

\[
1 - \delta_{2k} \leq \frac{\|A(x_1 - x_2)\|_2^2}{\|x_1 - x_2\|_2^2} \leq 1 + \delta_{2k}.
\]

---


Picture Credit: Mostafa Mohsenvand.
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

Review of Classical Parameter Estimation

Review of Compressive Sensing

**Fundamental Limits of Subsampling on Parameter Estimation**

Sensitivity of Basis Mismatch and Heuristic Remedies

**Going off the Grid**
  Atomic Norm Minimization
  Enhanced Matrix Completion
CS and Fundamental Estimation Bounds

- Canonical model before compression:

\[ y = \Psi(\nu)\alpha + n = s(\theta) + n \]

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

- Canonical model after compression:

\[ \Phi y = \Phi(\Psi(\nu)\alpha + n) = \Phi(s(\theta) + n) \]

where \( \Phi \in \mathbb{C}^{m \times n}, m \ll n \), 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: \( y \sim p(y; \theta) \)
- Fisher Score: Sensitivity of log-likelihood function to the parameter vector
  \[
  \frac{\partial}{\partial \theta_i} \log p(y; \theta)
  \]
- Fisher information matrix: Covariance of Fisher score
  \[
  \{J(\theta)\}_{i,j} = E \left[ \left( \frac{\partial}{\partial \theta_i} \log p(y; \nu) \right) \left( \frac{\partial}{\partial \theta_j} \log p(y; \theta) \right) | \theta \right]
  = -E \left[ \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p(y; \theta) | \theta \right]
  \]

Measures the amount of information that the measurement vector \( y \) carries about the parameter vector \( \theta \).
Cramér-Rao Lower Bound (CRB)

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

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

- The $i$th diagonal element of $J^{-1}(\theta)$ lower bounds the MSE of any unbiased estimator $T_i(y)$ of the $i$th parameter $\theta_i$ from $y$.

- Volume of error concentration ellipse:

$$\text{det}[\text{cov}_\theta(T(y))] \geq \text{det}[J^{-1}(\theta)]$$
Complex Normal model (Canonical model):

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

Fisher information matrix:

\[ J(\theta) = G^H(\theta)R^{-1}G(\theta) \]

\[ = \frac{1}{\sigma^2} G^H(\theta)G(\theta), \quad \text{when} \quad R = \sigma^2 I \]

\[ G(\theta) = [g_1(\theta), \ldots, g_k(\theta)]; \quad g_i(\theta) = \frac{\partial s(\theta)}{\partial \theta_i} \]

Cramér-Rao lower bound:

\[ (J^{-1}(\theta))_{ii} = \sigma^2 (g_i^H(\theta)(I - P_{G_i(\theta)})g_i(\theta))^{-1} \]

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

Compressive measurement (canonical model):

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

Fisher information matrix:

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

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

Cramer-Rao lower bound:

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

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

Question: What is the impact of compressive sampling on the Fisher information matrix and the Cramér-Rao bound (CRB) for estimating parameters?
Theorem (Pakrooh, Pezeshki, Scharf, Chi ’13)

(a) For any compression matrix, we have

\[
(J^{-1}(\theta))_{ii} \leq (\hat{J}^{-1}(\theta))_{ii} \leq \frac{1}{\lambda_{\text{min}}(G^T(\theta)P\Phi^T G(\theta))}
\]

(b) For a random compression matrix, we have

\[
(\hat{J}^{-1}(\theta))_{ii} \leq \frac{\lambda_{\text{max}}(J^{-1}(\theta))}{C(1-\epsilon)}
\]

with probability at least \(1 - \delta - \delta'\).

Remarks:

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

\[
Pr \left( \forall q \in \langle G(\theta) \rangle : (1 - \epsilon)\|q\|_2^2 \leq \|\Phi q\|_2^2 \leq (1 + \epsilon)\|q\|_2^2 \right) \geq 1 - \delta.
\]

\( 1 - \delta' \) is the probability that \( \lambda_{\min}(\Phi \Phi^T)^{-1} \) is larger than \( C \).

If entries of \( \Phi_{m \times n} \) are i.i.d. \( \mathcal{N}(0, 1/m) \), then

\[
\delta \leq \left\lceil (2\sqrt{p'/\epsilon'})^p \right\rceil e^{-m(\epsilon^2/4 - \epsilon^3/6)}, \text{ where}
\]

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

\( \delta' \) is determined from the distribution of the largest eigenvalue of a Wishart matrix, and the value of \( C \), from a hypergeometric function.

---

Example: Estimating the DOA of a point source at boresight $\theta_1 = 0$ in the presence of a point interferer at electrical angle $\theta_2$.

The LHS figure shows the after compression CRB (red) for estimating $\theta_1 = 0$ as $\theta_2$ 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 $\delta$ 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

\[
D(p(y; \theta) \| p(y; \theta')) = \int_y p(y; \theta) \log \frac{p(y; \theta)}{p(y; \theta')} dy
\]

---

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

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

- After compression with $\Phi$:

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

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

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

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

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

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


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.

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.

---

**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.

---

Signal Model: Mean Case

- **Before compression:**
  \[ y : \mathcal{CN}_n[\Psi \alpha, \sigma^2 I]; \quad \Psi \in \mathbb{C}^{n \times k} \]

- **After compression with compressive sensing matrix \( \Phi_{cs} \in \mathbb{C}^{m \times n}, m < n \):**
  \[ y_{cs} : \mathcal{CN}_m[\Phi_{cs} \Psi \alpha, \sigma^2 \Phi_{cs} \Phi_{cs}^H] \]

or equivalently (with some abuse of notation):

\[ y_{cs} : \mathcal{CN}_m[\Phi \Psi \alpha, \sigma^2 I], \quad \Phi = (\Phi_{cs} \Phi_{cs}^H)^{-1/2} \Phi_{cs} \]
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 $\langle \Psi \rangle$. 

![Diagram showing the relationship between $\langle \mathbf{H} \rangle$, $\langle \Psi \rangle$, $\mathbb{C}^n$, $\mathbf{P}_{\mathbf{H}\Psi\mathbf{y}}$, and $\mathbf{P}_{\Psi \mathbf{y}}$.]
Subspace Swap Events

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

\[
\min_i |\psi_i^H y|^2 \leq \frac{1}{k} y^H P_{\Psi} y < \frac{1}{n - k} y^H P_H y \leq \max_i |h_i^H y|^2
\]

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

\[
|\psi_{min}^H y|^2 < \frac{1}{n - k} y^H P_H y \leq \max_i |h_i^H y|^2.
\]
Theorem (Pakrooh, Pezeshki, Scharf (GlobalSIP’13))

(a) Before compression:

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

\[ \|\Psi \alpha\|^2_2 / \sigma^2 \text{ is the SNR before compression.} \]

(b) After compression:

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

\[ \|\Phi \Psi \alpha\|^2_2 / \sigma^2 \text{ is the SNR after compression.} \]
Analytical lower bounds for the probability of subspace swap. Array size: \( n = 188 \) elements; Compressed array \( m = 28 \) elements.

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

\[
\text{MSE} \approx P_{ss}\sigma_0^2 + (1 - P_{ss})\sigma_{CR}^2
\]
Analytical lower bounds for the probability of subspace swap for different compression ratios \( n/m \). The before compression (BC) dimension is \( n = 188 \).
Before compression:

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

After compression with compressive sensing matrix \( \Phi_{cs} \in \mathbb{C}^{m \times n}, m < n \):

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

or equivalently (with some abuse of notation):

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

Assume data consists of \( L \) iid realizations of \( y \) arranged as \( Y = [y_1, y_2, \cdots, y_L] \).
Theorem (Pakrooh, Pezeshki, Scharf (GlobalSIP’13))

(a) Before compression:

\[
P_{ss} \geq 1 - P\left[\frac{\text{tr}(\mathbf{Y}^H \mathbf{P} \mathbf{Y} / kL)}{\text{tr}(\mathbf{Y}^H \mathbf{P}_H \mathbf{Y} / (n - k)L)} > 1\right]
\]

\[
= 1 - P\left[F_{2kL, 2(n-k)L} > \frac{1}{1 + \lambda_k / \sigma^2}\right].
\]

\[
\lambda_k = \text{ev}_{\text{min}}(\mathbf{ΨR}_\alpha \alpha \mathbf{Ψ}^H)
\]

\[
\lambda_k / \sigma^2: \text{Effective SNR before compression}
\]

(b) After compression:

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

\[
\lambda'_k = \text{ev}_{\text{min}}(\mathbf{ΦΨR}_\alpha \alpha \mathbf{Ψ}^H \mathbf{Φ}^H)
\]

\[
\lambda'_k / \sigma^2: \text{Effective SNR after compression}
\]
Sensor Array Processing–Covariance Case

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

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

\[
\text{MSE} \approx P_{ss} \sigma_0^2 + (1 - P_{ss}) \sigma_{CR}^2
\]


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

Review of Classical Parameter Estimation

Review of Compressive Sensing

Fundamental Limits of Subsampling on Parameter Estimation

Sensitivity of Basis Mismatch and Heuristic Remedies

Going off the Grid
  Atomic Norm Minimization
  Enhanced Matrix Completion
Convert the nonlinear problem into a linear system via discretization of the parameter space at desired resolution:

\[ s(\theta) = \sum_{i=0}^{k-1} \psi(\nu_i)\alpha_i = \Psi_{ph}\alpha \]

**Over-determined & nonlinear**

**Under-determined linear & sparse**

The set of candidate \( \nu_i \in \Omega \) is quantized to \( \tilde{\Omega} = \{\omega_1, \cdots, \omega_n\} \), \( n > m \); \( \Psi_{ph} \) unknown and \( \Psi_{cs} \) assumed known.
Mathematical (CS) model:

\[ s = \Psi_{cs} x \]

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

Physical (true) model:

\[ s = \Psi_{ph} \alpha \]

The basis \( \Psi_{ph} \) is unknown, and is determined by a point spread function, a Green’s function, or an impulse response, and \( \alpha \) is \( k \)-sparse and unknown.

**Key transformation:**

\[ x = \Psi_{mis} \alpha = \Psi_{cs}^{-1} \Psi_{ph} \alpha \]

\( x \) is sparse in the unknown \( \Psi_{mis} \) basis, not in the identity basis.
DFT Grid Mismatch:

\[
\Psi_{\text{mis}} = \Psi_{cs}^{-1} \Psi_{\text{ph}} = \begin{bmatrix}
L(\Delta \theta_0 - 0) & L(\Delta \theta_1 - \frac{2\pi(n-1)}{n}) & \cdots & L(\Delta \theta_{n-1} - \frac{2\pi}{n}) \\
L(\Delta \theta_0 - \frac{2\pi}{n}) & L(\Delta \theta_1 - 0) & \cdots & L(\Delta \theta_{n-1} - \frac{2\pi}{n}) \\
\vdots & \vdots & \ddots & \vdots \\
L(\Delta \theta_0 - \frac{2\pi(n-1)}{n}) & L(\Delta \theta_1 - \frac{2\pi(n-2)}{n}) & \cdots & L(\Delta \theta_{n-1} - 0)
\end{bmatrix}
\]

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

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

Slow decay of the Dirichlet kernel means that the presumably sparse vector \( x = \Psi_{\text{mis}} \alpha \) is in fact incompressible.
Question: What is the consequence of assuming that $x$ is $k$-sparse in $I$, when in fact it is only $k$-sparse in an unknown basis $\Psi_{mis}$, which is determined by the mismatch between $\Psi_{cs}$ and $\Psi_{ph}$?

Physical Model

$s = \Psi_{ph}\alpha$

CS Sampler

$y = \Phi s$

CS Inverter

$\min \|x\|_1$

s.t. $y = \Phi \Psi_{cs} x$
Sensitivity to Basis Mismatch

- **CS Inverter**: Basis pursuit solution satisfies

\[
\begin{align*}
\text{Noise-free: } & \quad \|x^* - x\|_1 \leq C_0 \|x - x_k\|_1 \\
\text{Noisy: } & \quad \|x^* - x\|_2 \leq C_0 k^{-1/2} \|x - x_k\|_1 + C_1 \epsilon
\end{align*}
\]

where \(x_k\) is the best \(k\)-term approximation to \(x\).

- Similar bounds CoSaMP and ROMP.
- Where does mismatch enter? \(k\)-term approximation error.

\[
x = \Psi_{mis}\alpha = \Psi_{cs}^{-1}\Psi_{ph}\alpha
\]

- **Key**: Analyze the sensitivity of \(\|x - x_k\|_1\) to basis mismatch.
Degeneration of Best \( k \)-Term Approximation

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

Let \( \Psi_{mis} = \Psi_{cs}^{-1} \Psi_{ph} = I + E \), where \( x = \Psi_{mis} \alpha \). Let \( 1 \leq p, q \leq \infty \) and \( 1/p + 1/q = 1 \).

- If the rows \( e^T_\ell \in \mathbb{C}^{1 \times n} \) of \( E \) are bounded as \( \|e_\ell\|_p \leq \beta \), then

\[
\|x - x_k\|_1 \leq \|\alpha - \alpha_k\|_1 + (n - k) \beta \|\alpha\|_q.
\]

- The bound is achieved when the entries of \( E \) satisfy

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

Bounds on Image Inversion Error

**Theorem (inversion error)**

Let \( A = \Phi \Psi_{mis} \) satisfy \( \delta_{2k}^A < \sqrt{2} - 1 \) and \( 1/p + 1/q = 1 \). If the rows of \( E \) satisfy \( \|e_m\|_p \leq \beta \), then

\[
\|x - x^*\|_1 \leq C_0(n - k)\beta\|\alpha\|_q. \quad \text{(noise-free)}
\]

\[
\|x - x^*\|_2 \leq C_0(n - k)k^{-1/2}\beta\|\alpha\|_q + C_1\epsilon. \quad \text{(noisy)}
\]

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

Mismatch in Modal Analysis

Actual modes

Conventional FFT

Compressed sensing

Linear Prediction

Frequency mismatch
Mismatch in Modal Analysis

Actual modes

Conventional FFT

Compressed sensing

Linear Prediction

Damping mismatch
Mismatch in Modal Analysis

Actual modes

Conventional FFT

Compressed sensing

Linear Prediction with Rank Reduction

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 \( s = \Psi cs x \), where \( \Psi cs \) is the \( m \times n \), with \( n = mL \), “DFT” frame that is over-resolved to \( \Delta f = 1/mL \).

\[
\Psi cs = \frac{1}{\sqrt{m}} \begin{bmatrix}
1 & 1 & \cdots & 1 \\
1 & e^{2\pi i/mL} & \cdots & e^{2\pi i(1/mL-1)/mL} \\
\vdots & \vdots & \ddots & \vdots \\
1 & e^{2\pi i(m-1)/mL} & \cdots & e^{2\pi i(m-1)(mL-1)/mL}
\end{bmatrix}
\]

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

---

Noise Limited, Quantization Limited, or Null Space Limited

\[ \ell_1 \text{ inversions for } L = 2, 4, 6, 8 \]

- 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.

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.

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.

![Scatter plots for OMP Estimates](image)

(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

These approaches still assume a grid.


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

Review of Classical Parameter Estimation

Review of Compressive Sensing

Fundamental Limits of Subsampling on Parameter Estimation

Sensitivity of Basis Mismatch and Heuristic Remedies

Going off the Grid
  Atomic Norm Minimization
  Enhanced Matrix Completion
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.

Outline

Review of Classical Parameter Estimation

Review of Compressive Sensing

Fundamental Limits of Subsampling on Parameter Estimation

Sensitivity of Basis Mismatch and Heuristic Remedies

Going off the Grid
  Atomic Norm Minimization
  Enhanced Matrix Completion
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:
  
  $$\|x\|_\mathcal{A} = \inf \left\{ t > 0 : x \in t \text{conv}(\mathcal{A}) \right\} = \inf \left\{ \sum_{i} c_i \left| \begin{array}{l} x = \sum_{i} c_i a_i, \ a_i \in \mathcal{A}, \ c_i > 0 \end{array} \right. \right\}.$$ 

- **Step 3**: formulate a convex program to minimize the atomic norm with respect to the measurements.

---

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 $\ell_1$ 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 $\ell_1$ norm (b) unit ball of nuclear norm

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 \( x \): \( x = \sum_{i=1}^{r} d_i a(f_i) \), \( d_i \in \mathbb{C} \) where \( a(f) \) is the atom defined as
  \[
  a(f) = \frac{1}{\sqrt{n}} \begin{bmatrix} 1 & e^{j2\pi f} & \ldots & e^{j2\pi f(n-1)} \end{bmatrix}.
  \]

- Atomic norm:
  \[
  \| x \|_{A} = \inf \left\{ \sum_{k} |c_k| \mid x = \sum_{k} c_k a(f_k), f_k \in [0, 1) \right\}
  = \inf_{u, t} \left\{ \frac{1}{2} \text{Tr}(\text{toep}(u)) + \frac{1}{2} t \left[ \begin{bmatrix} \text{toep}(u) & x^* \end{bmatrix} x \right] \succ 0 \right\}.
  \]
  which can be equivalently given in an SDP form.

---

Random Subsampling: We observe a subset $\Omega$ of entries of $x$ uniformly at random:

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

Atomic Norm Minimization:

$$\min_s \|s\|_A \quad \text{subject to} \quad s_\Omega = x_\Omega,$$

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

(Primal:) $$\min_{u,s,t} \frac{1}{2} \text{Tr}(\text{toep}(u)) + \frac{1}{2} t$$

subject to

$$\begin{bmatrix} \text{toep}(u) & s \\ s^* & t \end{bmatrix} \succeq 0,$$

$$s_\Omega = x_\Omega.$$
The dual problem can be used to recover the frequencies:

\[(\text{Dual:}) \max_q \langle q_\Omega, x_\Omega \rangle_{\mathbb{R}} \text{ subject to } \|q\|_A^* \leq 1, \quad q_{\Omega^c} = 0.\]

where

\[\|q\|_A^* = \sup_{f \in [0,1)} |\langle q, a(f) \rangle| := \sup_{f \in [0,1)} Q(f)\]

The primal problem is optimal when the dual polynomial \(Q(f)\) satisfies:

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

This is also stable under noise.

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.

---

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

Extension for MMV Models

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

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

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

\[
\|X\|_A = \inf \left\{ \sum_k c_k \left| X = \sum_k c_k A(f_k, b_k) \right. , c_k \geq 0 \right\}
\]

\[
= \inf_{u, W} \left\{ \frac{1}{2} \text{Tr}(\text{toep}(u)) + \frac{1}{2} \text{Tr}(W) \left| \begin{bmatrix} \text{toep}(u) & X \\ X^* & W \end{bmatrix} \right. > 0 \right\}.
\]

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.

Two-Dimensional Frequency Model

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

\[
X = Y \cdot D \cdot Z^T.
\]

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

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

Vandemonde matrix

Vandemonde matrix

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

- **Goal**: We observe a *random subset of entries* of $X$, 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].

---


Consider a fine discretization of the parameter space

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

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

(Primal-discrete:) \( \min_{c_m} \|c_m\|_1 \) s.t. \( A^m_{\Omega} c_m = x_{\Omega} \)

(Dual-discrete:) \( \max_q \langle q_{\Omega}, x_{\Omega} \rangle_{\mathbb{R}} \) s.t. \( |\langle q, a(\omega_m) \rangle| \leq 1, i = 1, \ldots, m; \)

\( q_{\Omega^c} = 0. \)

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

---

Outline

Review of Classical Parameter Estimation

Review of Compressive Sensing

Fundamental Limits of Subsampling on Parameter Estimation

Sensitivity of Basis Mismatch and Heuristic Remedies

Going off the Grid
  Atomic Norm Minimization
  Enhanced Matrix Completion
Matrix Completion?

recall that \( X = \underbrace{\mathbf{Y} \cdot \mathbf{D}}_{\text{Vandemonde}} \cdot \underbrace{\mathbf{Z}^T}_{\text{Vandemonde}} \).

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

\[
\mathbf{Y} := \begin{bmatrix}
1 & 1 & \cdots & 1 \\
y_1 & y_2 & \cdots & y_r \\
\vdots & \vdots & \ddots & \vdots \\
y_1^{m_1-1} & y_2^{m_1-1} & \cdots & y_r^{m_1-1}
\end{bmatrix},
\mathbf{Z} := \begin{bmatrix}
1 & 1 & \cdots & 1 \\
z_1 & z_2 & \cdots & z_r \\
\vdots & \vdots & \ddots & \vdots \\
z_1^{n_2-1} & z_2^{n_2-1} & \cdots & z_r^{n_2-1}
\end{bmatrix}
\]

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

\[
\begin{bmatrix}
\checkmark & ? & ? & \checkmark & \checkmark \\
? & \checkmark & ? & \checkmark & \checkmark \\
? & ? & \checkmark & \checkmark & ? \\
\checkmark & \checkmark & \checkmark & ? & ? \\
\checkmark & \checkmark & ? & ? & \checkmark
\end{bmatrix}
\]
Matrix Completion

- Matrix Completion can be thought as an extension of CS to low-rank matrices.

- **The Netflix problem:** Let $X \in \mathbb{R}^{n_1 \times n_2}$ satisfying $\text{rank}(X) = r$.

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

$$
\begin{align*}
\text{minimize} & \quad \text{rank}(\mathbf{M}) \\
\text{subject to} & \quad \mathcal{P}_\Omega(\mathbf{M}) = \mathcal{P}_\Omega(\mathbf{X}).
\end{align*}
$$
Solve MC via Nuclear Norm Minimization

- Relax the rank minimization problem to a convex optimization:

\[
\begin{align*}
\text{minimize} & \quad \|M\|_* \\
\text{subject to} & \quad \mathcal{P}_\Omega (M) = \mathcal{P}_\Omega (X),
\end{align*}
\]

where \(\|M\|_* = \text{sum}(\sigma_i(M))\).

- Coherence measure: Let the SVD of \(X = U\Lambda V^T\). Define the coherence measure

\[
\begin{align*}
\max_{1 \leq i \leq n_1} \|U^T e_i\|_2 & \leq \sqrt{\frac{\mu n_1}{r}}, \\
\max_{1 \leq i \leq n_2} \|V^T e_i\|_2 & \leq \sqrt{\frac{\mu n_2}{r}}.
\end{align*}
\]

where \(e_i\)'s are standard basis vectors.

- \(1 \leq \mu \leq \max(n_1, n_2)/r\);
- subspace with low coherence: e.g. all one vectors;
- subspace with high coherence: e.g. standard basis vectors.

---


Performance Guarantees and Its Implication

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

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

\[
m > C \mu r n \log^2 n
\]

where \( C \) is some universal constant.

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

---


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

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

$$
X_e = 
\begin{bmatrix}
X_0 & X_1 & \cdots & X_{n_1-k_1} \\
X_1 & X_2 & \cdots & X_{n_1-k_1+1} \\
\vdots & \vdots & \ddots & \vdots \\
X_{k_1-1} & X_k & \cdots & X_{n_1-1}
\end{bmatrix},
$$

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

$$
X_l = 
\begin{bmatrix}
X_{l,0} & X_{l,1} & \cdots & X_{l,n_2-k_2} \\
X_{l,1} & X_{l,2} & \cdots & X_{l,n_2-k_2+1} \\
\vdots & \vdots & \ddots & \vdots \\
X_{l,k_2-1} & X_{l,k_2} & \cdots & X_{l,n_2-1}
\end{bmatrix}.
$$
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:

$$X_e = \begin{bmatrix}
Z_L \\
Z_L Y_d \\
\vdots \\
Z_L Y_d^{k_1-1}
\end{bmatrix}
D \begin{bmatrix}
Z_R, Y_d Z_R, \cdots, Y_d^{n_1-k_1} Z_R
\end{bmatrix},$$

- $Z_L$ and $Z_R$ are Vandermonde matrices specified by $z_1, \ldots, z_r$.
- $Y_d = \text{diag} [y_1, y_2, \cdots, y_r]$.

The enhanced form $X_e$ is low-rank.

- $\text{rank}(X_e) \leq r$
- Spectral Sparsity $\Rightarrow$ Low Rankness

holds even with damping modes.

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

\[
\begin{align*}
\text{minimize} & \quad \text{rank} \left( \mathbf{M}_e \right) \\
\text{subject to} & \quad \mathbf{M}_{i,j} = \mathbf{X}_{i,j}, \forall (i,j) \in \Omega
\end{align*}
\]

where \(\Omega\) denotes the sampling set.

Motivated by Matrix Completion, we will solve its convex relaxation:

\[
\begin{align*}
(\text{EMaC}) : \quad & \text{minimize} \quad \| \mathbf{M}_e \|_* \\
& \text{subject to} \quad \mathbf{M}_{i,j} = \mathbf{X}_{i,j}, \forall (i,j) \in \Omega
\end{align*}
\]

where \(\| \cdot \|_*\) denotes the nuclear norm.

The algorithm is referred to as Enhanced Matrix Completion (EMaC).
Enhanced Matrix Completion (EMaC)

\[(EMaC): \text{minimize} \quad \| M_e \|_* \quad \text{subject to} \quad M_{i,j} = X_{i,j}, \forall (i,j) \in \Omega\]

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

\[
\begin{bmatrix}
\checkmark & ? & ? & \checkmark & \checkmark \\
? & \checkmark & ? & \checkmark & \checkmark \\
? & ? & \checkmark & \checkmark & ? \\
\checkmark & \checkmark & \checkmark & \checkmark & ? \\
\checkmark & \checkmark & ? & ? & \checkmark
\end{bmatrix}
\]
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 \( G_L \) and \( G_R \) as \( r \times r \) Gram matrices such that

\[
(G_L)_{i,l} = \mathcal{K}(k_1, k_2, f_{1i} - f_{1l}, f_{2i} - f_{2l}),
\]

\[
(G_R)_{i,l} = \mathcal{K}(n_1 - k_1 + 1, n_2 - k_2 + 1, f_{1i} - f_{1l}, f_{2i} - f_{2l}).
\]
Introduce Coherence Measure

- **Incoherence condition** holds w.r.t. $\mu$ if
  \[
  \sigma_{\text{min}}(G_L) \geq \frac{1}{\mu}, \quad \sigma_{\text{min}}(G_R) \geq \frac{1}{\mu}.
  \]

- $\mu = \Theta(1)$ 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].

---

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 $\mu$ 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.

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:

\[
\text{(EMaC-Noisy)} : \quad \text{minimize} \quad \| \mathbf{M}_e \|_* \\
\text{subject to} \quad \| \mathcal{P}_\Omega (\mathbf{M} - \mathbf{X}) \|_F \leq \delta,
\]

**Theorem (Chen and Chi, 2013)**

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

\[
\| \hat{\mathbf{X}}_e - \mathbf{X}_e \|_F \leq \left\{ 2\sqrt{n} + 8n + \frac{8\sqrt{2}n^2}{m} \right\} \delta
\]

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.

---

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 $M_0 = X_e$ and $t = 0$.  
2: repeat  
3: 1) $Q_t \leftarrow D_{\tau_t}(M_t)$ (singular-value thresholding)  
4: 2) $M_t \leftarrow \text{Hankel}_{X_0}(Q_t)$ (projection onto a Hankel matrix consistent with observation)  
5: 3) $t \leftarrow t + 1$  
6: until convergence

Figure: dimension: $101 \times 101$, $r = 30$, $\frac{m}{n_1n_2} = 5.8\%$, SNR = 10dB.
Robustness to Sparse Outliers

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

\[ X_{i,l}^{\text{corrupted}} = X_{i,l} + S_{i,l} \]

where \( S_{i,l} \) is of arbitrary amplitude.

▶ Reminiscent of the robust PCA approach [Candes et. al. 2011, Chandrasekaran et. al. 2011]

▶ Solve the following algorithm:

\[(\text{RobustEMaC}) : \minimize_{M,S} \\| M_e \|_* + \lambda \| S_e \|_1 \]

subject to \( (M + S)_{i,l} = X_{i,l}^{\text{corrupted}}, \forall (i, l) \in \Omega \)
Theoretical Guarantees for Robust Recovery

\[(\text{RobustEMaC}) : \quad \text{minimize} \quad \|M_e\|_* + \lambda \|S_e\|_1 \quad \text{subject to} \quad (M + S)_{i,l} = X^\text{corrupted}_{i,l}, \quad \forall (i, l) \in \Omega.\]

**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 \(X\) 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: \(s \sim \Theta(1)\)

---

Robustness to Sparse Corruptions

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

<table>
<thead>
<tr>
<th></th>
<th>EMaC</th>
<th>Atomic Norm</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Signal model</strong></td>
<td>Deterministic</td>
<td>Random</td>
</tr>
<tr>
<td><strong>Observation model</strong></td>
<td>Random</td>
<td>Random</td>
</tr>
<tr>
<td><strong>Success Condition</strong></td>
<td>Coherence</td>
<td>Separation condition</td>
</tr>
<tr>
<td><strong>Amplitudes</strong></td>
<td>No condition</td>
<td>Randomly generated</td>
</tr>
<tr>
<td><strong>Sample Complexity</strong></td>
<td>$\Theta(r \log^3 n)$</td>
<td>$\Theta(r \log r \log n)$</td>
</tr>
<tr>
<td><strong>Bounded Noise</strong></td>
<td>Yes</td>
<td>Not shown</td>
</tr>
<tr>
<td><strong>Sparse Corruptions</strong></td>
<td>Yes</td>
<td>Not shown</td>
</tr>
</tbody>
</table>
Comparisons of EMaC and Atomic Norm Minimization

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

- without separation

- with 1.5 RL separation
References

- 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.
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.


Concluding Remarks

- Compression, whether by linear maps (e.g., Gaussian) or by subsampling, has performance consequences for parameter estimation. Fisher information decreases, CRB increases, and the onset of breakdown threshold increases.

- Model mismatch is inevitable and can result in considerable performance degradation, 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?