Adaptive Sensing for Detection and Estimation of Structured Sparse Signals

Rui M. Castro

Collaborators:  
Ervin Tánczos (TU/e)  
Pierre-André Savalle (ENS Cachan)  
Gábor Lugosi (UPF)
Inference of Sparse Signals

Can we reliably detect/identify such sparse patterns?

Setting: we want to make inference about a set

\[ S \subseteq \{1, \ldots, n\}, \quad |S| \ll n \]

E.g.: \( S \) represents the differentially expressed genes or the “anomalous” nodes in a network
The Normal Means Model

Define the signal

\[ x_i = \begin{cases} 
\mu & i \in S \\
0 & i \notin S
\end{cases}, \text{ where } \mu > 0 \]

Observation model:

\[ Y_i = x_i + W_i, \quad i \in \{1, \ldots, n\}, \]

where \( W_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1) \)

How small can \( \mu \) be so that we can still perform reliable inference about the unknown signal?

Let’s try to generalize this sensing model a bit...
An Adaptive Sensing Model

Normal means: \[ Y_i = x_i + W_i, \quad i \in \{1, \ldots, n\}, \]
where \( W_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 1) \)

Measures each entry of \( x \) **exactly once**, with precision **one** (precision = 1/variance)

- Sequentially choose which entry to measure, based on measurements done so far...
- Measure each entry multiple times, with independent noise realizations and different precision levels
An Adaptive Sensing Model

Consider a model allowing multiple observations...

The $k^{th}$ observation ($k \in \{1, 2, \ldots\}$) is given by

$$Y_k = x_{A_k} + (\Gamma_k)^{-1/2} W_k,$$

where $W_k \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$

$A_k$ - entry of $x$ measured at $k$th observation

$\Gamma_k$ - precision of the $k$th measurement

...subject to a total precision budget

$$\mathbb{E} \left( \sum_{k=1}^{\infty} \Gamma_k \right) \leq n$$

Precision = SNR control

The sensing budget is often related to the total time spent collecting measurements.
Adaptive vs. Non-Adaptive Sensing

Non-Adaptive sensing:

\( \{A_k, \Gamma_k\}_{k=1}^{\infty} \) must be chosen prior to the collection of any observations.

Adaptive Sensing:

\( A_k, \Gamma_k \) are chosen sequentially and are functions of

\( \{Y_\ell, A_\ell, \Gamma_\ell\}_{\ell=1}^{k-1} \)

Key Idea: allow future sensing precision and location to depend on past observations !!!
Inference Goals

Assume the set $S$ belongs to a class of subsets $\mathcal{C}$.

**Estimation:**

Construct a set estimator $\hat{S} = \hat{S}(Y, A, \Gamma)$ minimizing:

\[
\text{Expected Hamming Distance: } \max_{S \in \mathcal{C}} \mathbb{E}_S [|\hat{S} \Delta S|]
\]

**Detection:**

\[
H_0 : S = \emptyset \quad \text{vs.} \quad H_1 : S \in \mathcal{C}
\]

(no signal present)  (a signal in the class)

Given a test function $\hat{\Phi} \in \{0, 1\}$ minimize

\[
R_c(\hat{\Phi}) = \mathbb{P}_0(\hat{\Phi} \neq 0) + \max_{S \in \mathcal{C}} \mathbb{P}_S(\hat{\Phi} \neq 1).
\]

Obviously, the difficulty of the problem depends on the class $\mathcal{C}$ under consideration...
Structured and Unstructured Classes

$s$-sets (structureless case):
The class of ALL subsets of $\{1, \ldots, n\}$ with cardinality $s$.

For other classes $\mathcal{C}$ we say the signal support has **structure**:

$s$-intervals:

$\mathcal{C} = \{\{1, \ldots, s\} , \{2, \ldots, s + 1\} , \ldots , \{n-s+1, \ldots, n\}\}$.

$k$ disjoint $s$-intervals:

$s$-stars in a complete graph with $n$ edges:
Structured and Unstructured Classes

Other interesting structured classes:

unions of \( k \) s-stars
(with distinct centers)

\[
\begin{array}{c}
\text{a size } s \text{ submatrix of a} \\
\sqrt{n} \times \sqrt{n} \text{ matrix}
\end{array}
\]

Structure can be very helpful for non-adaptive sensing, both for estimation or detection:

- Butucea, Ingster, “Detection of a sparse submatrix of a high-dimensional noisy matrix”, Bernoulli 2013
Signal Detection

**Theorem: (Non-Adaptive Sens.)** (Addario-Berry et al., 2010) Consider the class of all $s$-sets, where $s \ll \sqrt{n}$. If $R_C(\hat{\Phi}) \leq \epsilon$ necessarily

$$\mu \geq \sqrt{\log \left( \frac{n}{s^2} \right)}.$$ 

**Theorem: (Adaptive Sens.)** (C. '12) If $R_C(\hat{\Phi}) \leq \epsilon$ we have necessarily

$$\mu \geq \sqrt{\frac{2}{s} \log \frac{1}{2\epsilon}}.$$ 

- There is a sensing/detection algorithm achieving this bound
- **Structural assumptions cannot further improve this result!!**
Estimation

**Theorem: (Non-Adaptive Sens.)** Let \( \mathcal{C} \) denote the class of all \( s \)-sets and \( \varepsilon > 0 \). If

\[
\max_{S \in \mathcal{C}} \mathbb{E}_S(|\hat{S} \Delta S|) \leq \varepsilon,
\]

then necessarily \( \mu \geq \sqrt{2 \log n} \).

Furthermore, if signal is sparse \( (s \ll n) \) then the dependence on the extrinsic dimension is always of the form \( \sim \sqrt{\log n} \), regardless of the structure.

**Can adaptive sensing improve performance?**
Simple Thresholding

A simple sequential thresholding procedure

\[ Y_i^{(1)} = x_i + \mathcal{N}(0, 3) \]
Simple Thresholding

For sparse signals we remove about half of the components from further consideration at each step...
Adaptive vs. Non-Adaptive

Requirements to ensure that $\max_{S \in C} \mathbb{E} \left[ |\hat{S} \Delta S| \right] \to 0$, as $n \to \infty$:

**Simple Thresholding:**

$$\mu \geq \sqrt{6 \log s + 6.1 \log \log_2 n}$$

**Best non-adaptive sensing procedure:**

$$\mu \geq \sqrt{2 \log n} + a_n,$$

where $a_n \to \infty$.

Actually, the $\log \log n$ term is an artifact of the simple procedure, and can be removed by either using component-wise SLRT, or using a slightly more involved thresholding procedure (Malloy-Nowak ’12).


These same ideas can be used to deal with structured classes of support sets...
Idea for a General Procedure

**s-sets:** the Simple Thresholding/Multiple SLRT approach is a coordinate-wise query of all \( n \) entries

A general approach for structured cases:

- Devise a **noiseless** support estimation procedure, making the “minimal” number of queries necessary to uniquely identify the support (and exploring the structure of the signal class).

- **Robustify** the noiseless procedure to be able to deal with noisy observations, using SLRTs
Example $s$-intervals

**Search phase:** sequentially sample entries $1, s + 1, 2s + 2, \ldots$ until a significant component is found.

**Refinement:** sample elements to the left of the significant entry until reaching the end of the interval.

Once a noiseless procedure has been chosen, all we need to do is to replace the noiseless queries by SLRTs to ensure

i) the probability of not recovering the support is small

ii) The expected total precision spent satisfies the constraint

$$\mathbb{E} \left( \sum_{k=1}^{\infty} \Gamma_k \right) \leq n$$
Upper and Lower Bounds

With careful calibration of the SLRTs one can significantly improve on non-adaptive sensing. One can also show nearly matching adaptive sensing lower bounds.

Scaling laws necessary and sufficient to ensure $\max_{S \in \mathcal{S}} \mathbb{E}_S[\tilde{S} \Delta S] \to 0$:

<table>
<thead>
<tr>
<th></th>
<th>Non-Adaptive</th>
<th>Adaptive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sim \sqrt{\log n}$</td>
<td>$\sim \sqrt{\log s}$</td>
</tr>
<tr>
<td></td>
<td>$\sim \sqrt{\frac{1}{s} \log n/s}$</td>
<td>$\sim \sqrt{\frac{1}{s} \log ks}$</td>
</tr>
<tr>
<td></td>
<td>$\sim \sqrt{\log \frac{n}{s}}$</td>
<td>$\sim \sqrt{\frac{1}{s} \log^2 ks}$</td>
</tr>
<tr>
<td></td>
<td>$\sim \sqrt{\frac{1}{s} \log n/s}$</td>
<td>$\sim \sqrt{\frac{1}{s} \log^2 s}$</td>
</tr>
</tbody>
</table>

Linear Projections/Compressed Sensing

Instead of point-samples, one can consider the linear projection measurements

\[
\begin{pmatrix}
Y \\
\end{pmatrix} \in \mathbb{R}^\ell \\
\begin{pmatrix}
A \\
\end{pmatrix} \in \mathbb{R}^{\ell \times n} \\
\begin{pmatrix}
W \\
\end{pmatrix} \sim \mathcal{N}(0, I) \\
\begin{pmatrix}
x \\
\end{pmatrix} \in \mathbb{R}^n
\]

For carefully chosen sensing matrices \( A \) signals of sufficient magnitude can be reliably estimated by taking \( \ell \ll n \) projections.

In this setting the precision budget is conveniently casted as the restriction

\[
\mathbb{E}[\|A\|_F^2] \leq n
\]
Detection using Linear Projections

**Theorem:** (Arias-Castro, '12) For reliable detection using adaptive sensing it is necessary and sufficient for the signal magnitude to be of the order

\[ \sim \sqrt{\frac{1}{s^2}}. \]

As in the previous setting, **structure doesn’t help for detection.**

Moreover, we can achieve the above bound using only a **non-adaptive sensing** procedure !!!

For estimation the story is different…
Support Estimation with Linear Proj.

**Non-Adaptive Sensing:** If the number of measurements $\ell$ is large enough it is still necessary that

$$\mu \sim \sqrt{\log n} \quad \text{(see e.g., Wainwright, '09)} .$$

**Theorem:** (Haupt, Baraniuk, C. and Nowak '12) Assume $x$ has $|S| = s$ nonzeros entries, and collect a total of $O(s \max\{\log s \log \log n, \log n\})$ adaptive observations. If the minimum signal amplitude is greater than a constant times

$$\sqrt{\log s + \log \log_2 \log n}$$

we will recover the exact signal support with probability at least $1 - o(1)$.

**Related work ->**

Malloy, Nowak, “Near-optimal Adaptive Compressed Sensing”, 2012
Adaptive CS for Structured Classes

The high-level ideas used for the non-compressed case can be used in this setting as well.

$s$-intervals:

\[ C = \{\{1, \ldots, s\} , \{2, \ldots, s + 1\} , \ldots , \{n - s + 1, \ldots, n\}\} . \]

Scaling laws necessary and sufficient for \( \max_{S \in C} \mathbb{E}_S[\hat{S} \Delta S] \to 0. \)

**Non-Adaptive:** \( \mu \sim \sqrt{\frac{\log n}{s}} \)

**Adaptive:** \( \mu \sim \sqrt{\frac{\log s}{s^2}} \) (Balakrishnan et al, '12)

Related work →

Adaptive Compressed Sensing

Scaling laws necessary and sufficient to ensure \( \max_{S \in C_j} \mathbb{E}_S[\hat{S} \triangle S] \to 0 \):

<table>
<thead>
<tr>
<th></th>
<th>Non-Adaptive</th>
<th>Adaptive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \sim \sqrt{\log n} )</td>
<td>( \sim \sqrt{\log s} ) ( \sim \sqrt{\log s} )</td>
</tr>
<tr>
<td></td>
<td>( \sim \sqrt{\frac{1}{s^2} \log n/s} )</td>
<td>( \sim \sqrt{\frac{1}{s^2} \log ks} ) ( \sim \sqrt{\frac{1}{s^2} \log ks} )</td>
</tr>
<tr>
<td></td>
<td>( \sim \sqrt{\log \frac{\sqrt{n}}{s}} )</td>
<td>( \sim \sqrt{\frac{1}{s^2} \log ks} ) ( \sim \sqrt{\frac{1}{s^2} \log ks} )</td>
</tr>
<tr>
<td></td>
<td>( \sim \sqrt{\frac{1}{s^2} \log n/s} )</td>
<td>( \sim \sqrt{\frac{1}{s^2 \sqrt{s}} \log s} ) ( \sim \sqrt{\frac{1}{s^2} \log s} )</td>
</tr>
</tbody>
</table>

\[ sk \ll \sqrt{n} \]


These limits can be attained taking only \( \ell \sim s \log n \) projections, by the same principles of the CASS\(^{(s)}\) algorithm.

Some Remarks

• The submatrix case is more delicate than presented here. For non-square matrices one encounters different inference regimes...

• Good lower bounds on the sample complexity of adaptive compressive sensing are still needed.

Conjecture: Near the estimation threshold we need $\Omega(s \log n)$ projections (best existing lower bound we are aware of is $\Omega(s)$).

Aksoylar, C., Saligrama, V “Information-theoretic bounds for adaptive sparse recovery” (2014)
Detection of Correlations

Consider the following setting:

\[ U = \begin{bmatrix}
X_{1,1} & X_{1,2} & \cdots & X_{1,n} \\
X_{2,1} & X_{2,2} & \cdots & X_{2,n} \\
\vdots & \vdots & \ddots & \vdots \\
\end{bmatrix} \]

Rows are independent samples from a zero mean Gaussian vector with a covariance matrix \( \Sigma_S \).

e.g., \( \Sigma_S = 
\begin{bmatrix}
1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & 1 & 0 & & & & & & \\
\vdots & \ddots & \ddots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\vdots & \ddots & \ddots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\vdots & \ddots & \ddots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & 1
\end{bmatrix} \)

Let \( S \subseteq \{1, \ldots, n\} \)

\[ \Sigma_S(i,j) = \begin{cases} 
1 & \text{if } i = j \\
\rho & \text{if } i, j \in S, i \neq j \\
0 & \text{otherwise} 
\end{cases}, \quad 0 < \rho < 1. \]
Detection of Correlations

Consider the following setting:

\[
U = \begin{bmatrix}
X_{1,1} & X_{1,2} & \cdots & X_{1,n} \\
X_{2,1} & X_{2,2} & \cdots & X_{2,n} \\
\vdots & \vdots & \ddots & \vdots \\
\end{bmatrix}
\]

Rows are independent samples from a zero mean Gaussian vector with a covariance matrix \(\Sigma\).

Goal: From the observation of only some entries of \(U\) we want to test

\[H_0 : S = \emptyset \quad \text{vs.} \quad H_1 : S \in \mathcal{C},\]

where \(\mathcal{C}\) is some class of sets with cardinality \(s\).

Motivation: detection of anomalies, when anomalous behavior can only be detected by considering signals as a collection (contextual anomaly detection).

Note that \(X_{i,j} \sim \mathcal{N}(0,1)\), so measuring individual entries does not provide any information in isolation...
Detection of Correlations

Consider the following setting:

\[
U = \begin{bmatrix}
X_{1,1} & X_{1,2} & \cdots & X_{1,n} \\
X_{2,1} & X_{2,2} & \cdots & X_{2,n} \\
\vdots & \vdots & \ddots & \vdots \\
\end{bmatrix}
\]

Rows are independent samples from a zero mean Gaussian vector with a covariance matrix \( \Sigma \).

Non-Adaptive sensing:

Observe \( m \times n \) entries chosen prior to any observations.

Adaptive Sensing:

Observe \( m \times n \) entries, chosen in an online and sequential fashion.
Adaptive Sensing

**Open Problem:** is there an adaptive sensing procedure improving the performance for s-sets???

Arias-Castro, Bubeck, Lugosi, “Detecting Positive Correlations in a Multivariate Sample”, 2012