To Adapt or Not To Adapt
The Power and Limits of Adaptivity for Sparse Estimation

Mark A. Davenport
Stanford University
Department of Statistics
Sensor Explosion
Data Deluge

The data deluge

The Economist

Overload
Global information created and available storage
Exabytes

Source: IDC

Information created
Available storage

FORECAST

2005 06 07 08 09 10 11
0 250 500 750 1,000 1,250 1,500 1,750 2,000
“Paper became so cheap, and printers so numerous, that a deluge of authors covered the land”

Alexander Pope, 1728
Large Hadron Collider at CERN

Compact Muon Solenoid detector

320 terabits per second raw data

Stop-gap: perform ad-hoc triage to 800 Gbps, recording only “interesting events”
Data Deluge Challenges

How can we get our hands on as much data as possible?

How can we extract as much information as possible from a limited amount of data?

How can we avoid having to acquire so much data to begin with?

How can we extract any information at all from a massive amount of high-dimensional data?
Low-Dimensional Structure

How can we exploit low-dimensional structure to address the challenges posed by the “data deluge”?

- Visualization
- Feature extraction/selection
- Compression
- Regularization of ill-posed inverse problems
- Underpins compressive sensing
Compressive Sensing

\[ y = X \theta + z \]

- \( y \): Measurements
- \( X \): Measurement matrix
- \( \theta \): Parameter vector
- \( z \): Noise vector

\( n \times p \)
\( n \ll p \)
\( p \times 1 \)
\( k \) nonzeros

When (and how well) can we estimate \( \theta \) from the measurements \( y \)?
How Well Can We Estimate $\theta$?

- What do we know via compressive sensing?
  - feasible *nonadaptive* schemes with known performance guarantees

- Can we improve upon compressive sensing?
  - lower bound on the performance of *any* nonadaptive scheme

- What are the benefits of adaptivity?
  - lower bound on the performance of *any adaptive* scheme
  - practical implications
Compressive Sensing

- How should we design $X$ to ensure that $y$ contains as much information about $\theta$ as possible?

- What algorithms do we have for recovering $\theta$ from $y$?

[Candès, Romberg, and Tao; Donoho - 2005]
How To Design $X$?

Prototypical sensing model:

$$y = X\theta + z \quad z \sim \mathcal{N}(0, \sigma^2 I)$$

- Constrain $X$ to have unit-norm rows
- Pick $X$ at *random!*
  - i.i.d. Gaussian entries (with variance $1/p$)
  - random rows from a unitary matrix
- As long as $n = O(k \log(p/k))$, with high probability a random $X$ will satisfy the *restricted isometry property*
Restricted Isometry Property (RIP)

\[ \frac{\|X\theta_1 - X\theta_2\|_2^2}{\|\theta_1 - \theta_2\|_2^2} \approx \frac{n}{p} \quad \|\theta_1\|_0, \|\theta_2\|_0 \leq k \]
How To Design $X$?

Prototypical sensing model:

$$y = X\theta + z \quad z \sim \mathcal{N}(0, \sigma^2 I)$$

- Constrain $X$ to have unit-norm rows
- Pick $X$ at random!
  - i.i.d. Gaussian entries (with variance $1/p$)
  - random rows from a unitary matrix
- As long as $n = O(k \log(p/k))$, with high probability a random $X$ will satisfy the restricted isometry property
- Deep connections with Johnson-Lindenstrauss Lemma
  - see Baraniuk, Davenport, DeVore, and Wakin (2008)
How To Recover $\theta$?

- Lots and lots of algorithms
  - $\ell_1$-minimization (Lasso, Dantzig selector)
  - greedy algorithms (matching pursuit, forward selection)

If $X$ satisfies the RIP, $\|\theta\|_0 \leq k$, and $y = X\theta + z$ with $z \sim \mathcal{N}(0, \sigma^2 I)$, then

$$\hat{\theta} = \arg\min_{\theta' \in \mathbb{R}^p} \|\theta'\|_1$$

s.t. $\|X^*(y - X\theta')\|_\infty \leq c\sqrt{\log p}\sigma$

satisfies

$$\mathbb{E} \|\hat{\theta} - \theta\|^2_2 \leq C\frac{p}{n}k\sigma^2 \log p.$$  

[Candès and Tao - 2005]
How Well Can We Estimate $\theta$?

- What do we know via compressive sensing?
  
  For any $\theta$ we can achieve $\mathbb{E} \| \hat{\theta} - \theta \|_2^2 \leq C \frac{p}{n} k \sigma^2 \log p$

- Can we improve upon compressive sensing?

- What are the benefits of adaptivity?
Room For Improvement?

Let $x_i$ denote the $i^{th}$ row of $X$

$$y_i = \langle x_i, \theta \rangle + z_i$$

$x_i$ and $\theta$ are almost orthogonal

- We are using most of our “sensing power” to sense entries that aren’t even there!
- Tremendous loss in signal-to-noise ratio (SNR)
- It’s hard to imagine any way to avoid this...
Minimax Lower Bounds

- There exists matrices $X$ such that for any (sparse) $\theta$ we have
  $$\mathbb{E} \|\hat{\theta} - \theta\|_2^2 \leq C \frac{p}{n} k \sigma^2 \log p.$$

- We would like to know if there exists any $X$ or any recovery algorithm that can do much better for all $\theta$

- **Minimax lower bound:** For any $X$ and any $\hat{\theta}$, there exists a $\theta$ such that
  $$\mathbb{E} \|\hat{\theta} - \theta\|_2^2 \geq ?$$

- The bound will be determined by the **worst-case** $\theta$
Theorem

For any matrix $X$ (with unit-norm rows) and any recovery procedure $\hat{\theta}$, there exists a $\theta$ with $\|\theta\|_0 \leq k$ such that if $y = X\theta + z$ with $z \sim \mathcal{N}(0, \sigma^2 I)$, then

$$\mathbb{E} \|\hat{\theta}(y) - \theta\|_2^2 \geq C'' \frac{p}{n} k\sigma^2 \log(p/k).$$

Compressive sensing is already operating at the limit

[Candès and Davenport - 2011]
Suppose that $y = \theta + z$ with $z \sim \mathcal{N}(0, I)$ and that $k = 1$

$$\mathbb{E} \|\hat{\theta}(y) - \theta\|_2^2 \geq C' \log p$$

\[
\sqrt{\log p} \quad \|z\|_\infty \approx \sqrt{\log p}
\]
Proof Recipe

• Construct a set $\Theta$ of $k$-sparse vectors such that
  - $|\Theta| = \left( \frac{p}{k} \right)^{k/4}$
  - $\|\theta_i - \theta_j\|_2 \geq \frac{1}{2}$ for all $\theta_i, \theta_j \in \Theta$
  - $\frac{1}{|\Theta|} \sum_i \theta_i \theta_i^* \approx \frac{1}{p} I$

• Scale this set to the worst-case amplitude and use Fano’s Inequality to show that if $\theta$ is selected uniformly at random from $\Theta$, then the Bayes risk is large

• $\Theta$ can be constructed simply by picking $k$-sparse vectors at random
How Well Can We Estimate $\theta$?

- What do we know via compressive sensing?

  For any $\theta$ we can achieve
  \[ \mathbb{E} \| \hat{\theta} - \theta \|_2^2 \leq C \frac{p}{n} k \sigma^2 \log p \]

- Can we improve upon compressive sensing?

  There exist $\theta$ such that
  \[ \mathbb{E} \| \hat{\theta} - \theta \|_2^2 \geq C' \frac{p}{n} k \sigma^2 \log(\frac{p}{k}) \]

- What are the benefits of adaptivity?
Adaptivity to the Rescue?

Think of sensing as a game of 20 questions

Simple strategy: Use \( n/2 \) measurements to find the support, and the remainder to estimate the values.
Thought Experiment

Suppose that after $n/2$ measurements we have perfectly estimated the support.

\[
\mathbb{E} (\hat{\theta}_i - \theta_i)^2 = \frac{2k}{n} \sigma^2
\]

\[
\mathbb{E} \|\hat{\theta} - \theta\|_2^2 = \frac{2k}{n} k \sigma^2 \ll \frac{p}{n} k \sigma^2 \log p
\]
Does Adaptivity Really Help?

Sometimes...

- **Noise-free measurements, but non-sparse signal**
  - adaptivity doesn’t help if you want a uniform guarantee
  - probabilistic adaptive algorithms can reduce the required number of measurements from $O(k \log(p/k))$ to $O(k \log \log(p/k))$ [Indyk et al. - 2011]

- **Noisy setting**
  - distilled sensing [Haupt et al. - 2007, 2010]
  - adaptivity can reduce the estimation error to

\[
\mathbb{E} \left\| \hat{\theta} - \theta \right\|_2^2 = \frac{p}{n} k \sigma^2
\]

\[
\mathbb{E} \left\| \hat{\theta} - \theta \right\|_2^2 = \frac{k}{n} k \sigma^2
\]

Which is it?
Suppose we have a budget of \( n \) measurements of the form

\[
y_i = \langle x_i, \theta \rangle + z_i \quad \text{where} \quad \|x_i\|_2 = 1 \quad \text{and} \quad z_i \sim \mathcal{N}(0, \sigma^2)
\]

The vector \( x_i \) can have an arbitrary dependence on the measurement history, i.e., \((x_1, y_1), \ldots, (x_{i-1}, y_{i-1})\)

**Theorem**

There exist \( \theta \) with \( \|\theta\|_0 \leq k \) such that for any adaptive measurement strategy and any recovery procedure \( \hat{\theta} \),

\[
\mathbb{E} \|\hat{\theta}(y) - \theta\|_2^2 \geq C \frac{p}{n} k \sigma^2.
\]

Thus, in general, adaptivity does not significantly help!

[Arias-Castro, Candès, and Davenport - 2011]
Proof Strategy

Step 1: Consider sparse signals with nonzeros of amplitude
\[ \mu \approx \sigma \sqrt{p/n} \]

Step 2: Show that if given a budget of \( n \) measurements, you cannot detect the support very well.

Step 3: Immediately translate this into a lower bound on the MSE.

To make things simpler, we will consider a Bernoulli prior \( \pi(\theta) \) instead of a uniform \( k \)-sparse prior:

\[
\theta_j = \begin{cases} 
0 & \text{with probability } 1 - k/p \\
\mu > 0 & \text{with probability } k/p
\end{cases}
\]
Proof of Main Result

Let $S = \{j : \theta_j \neq 0\}$ and set $\sigma^2 = 1$

For any estimator $\hat{\theta}$, define $\hat{S} := \{j : |\hat{\theta}_j| \geq \mu/2\}$

Whenever $j \in S \setminus \hat{S}$ or $j \in \hat{S} \setminus S$, $|\hat{\theta}_j - \theta_j| \geq \frac{\mu^2}{4}$

$$||\hat{\theta} - \theta||_2^2 \geq \frac{\mu^2}{4} |S \setminus \hat{S}| + \frac{\mu^2}{4} |\hat{S} \setminus S| = \frac{\mu^2}{4} |\hat{S} \Delta S|$$

$$\mathbb{E} ||\hat{\theta} - \theta||_2^2 \geq \frac{\mu^2}{4} \mathbb{E} |\hat{S} \Delta S|$$
Proof of Main Result

Lemma
Under the Bernoulli prior, *any* estimate $\hat{S}$ satisfies

$$\mathbb{E} |\hat{S} \Delta S| \geq k \left( 1 - \frac{\mu}{2} \sqrt{n} \right).$$

Thus,

$$\mathbb{E} \|\hat{\theta} - \theta\|_2^2 \geq \frac{\mu^2}{4} \mathbb{E} |\hat{S} \Delta S| \geq k \cdot \frac{\mu^2}{4} \left( 1 - \frac{\mu}{2} \sqrt{\frac{n}{p}} \right)$$

Plug in $\mu = \frac{8}{3} \sqrt{\frac{p}{n}}$ and this reduces to

$$\mathbb{E} \|\hat{\theta} - \theta\|_2^2 \geq \frac{4}{27} \cdot \frac{k p}{n} \geq \frac{1}{7} \cdot \frac{k p}{n}$$
Key Ideas in Proof of Lemma

\[
P_{0,j}(y_1, \ldots, y_n) = \mathbb{P}(y_1, \ldots, y_n | \theta_j = 0)
\]

\[
P_{1,j}(y_1, \ldots, y_n) = \mathbb{P}(y_1, \ldots, y_n | \theta_j = \mu)
\]

\[
\mathbb{E} |\hat{S}\Delta S| \geq \frac{k}{p} \sum_j (1 - ||P_{1,j} - P_{0,j}||_{TV})
\]

\[
\geq k - \frac{k}{\sqrt{p}} \sqrt{\sum_j ||P_{1,j} - P_{0,j}||_{TV}^2}
\]

\[
\sum_j ||P_{1,j} - P_{0,j}||_{TV}^2 \leq \frac{\mu^2}{4} n \quad \Rightarrow \quad \mathbb{E} |\hat{S}\Delta S| \geq k \left(1 - \frac{\mu}{2} \sqrt{\frac{n}{p}}\right)
\]
Key Ideas in Proof of Lemma

Pinsker’s Inequality

\[ \| \mathbb{P} - \mathbb{Q} \|_{TV} \leq \sqrt{K(\mathbb{P}, \mathbb{Q})/2} \]

\[ \| \mathbb{P}_{1,j} - \mathbb{P}_{0,j} \|_{TV}^2 \leq \frac{\pi_0}{2} K(\mathbb{P}_{0,j}, \mathbb{P}_{1,j}) + \frac{\pi_1}{2} K(\mathbb{P}_{1,j}, \mathbb{P}_{0,j}) \]

\[ \leq \frac{\mu^2}{4} \sum_i \mathbb{E} x_{i,j}^2 \]

\[ \sum_j \| \mathbb{P}_{1,j} - \mathbb{P}_{0,j} \|_{TV}^2 \leq \frac{\mu^2}{4} \sum_{i,j} \mathbb{E} x_{i,j}^2 = \frac{\mu^2}{4} n \]
How Well Can We Estimate $\theta$?

- What do we know via compressive sensing?

  For any $\theta$ we can achieve $\mathbb{E}\|\hat{\theta} - \theta\|_2^2 \leq C \frac{p}{n} k\sigma^2 \log p$

- Can we improve upon compressive sensing?

  There exist $\theta$ such that $\mathbb{E}\|\hat{\theta} - \theta\|_2^2 \geq C' \frac{p}{n} k\sigma^2 \log(p/k)$

- What are the benefits of adaptivity?

  Minimal?
Suppose that $k = 1$ and that $\theta_{j*} = \mu$

Recursive Bisection [Iwen and Tewfik - 2011]
- split measurements into $\log p$ stages
- in each stage, use measurements to decide if the nonzero is in the left or right half of the “active set”
- after subdividing $\log p$ times, return support
Adaptivity In Practice

Suppose that \( k = 1 \) and that \( \theta_j = \mu \)

Recursive Bisection [Iwen and Tewfik - 2011]
- split measurements into \( \log p \) stages
- in each stage, use measurements to decide if the nonzero is in the left or right half of the “active set”
- after subdividing \( \log p \) times, return support
Suppose that $k = 1$ and that $\theta_{j^*} = \mu$

Recursive Bisection [Iwen and Tewfik - 2011]

- split measurements into $\log p$ stages
- in each stage, use measurements to decide if the nonzero is in the left or right half of the “active set”
- after subdividing $\log p$ times, return support
Suppose that $k = 1$ and that $\theta_{j^*} = \mu$

Recursive Bisection [Iwen and Tewfik - 2011]
- split measurements into $\log p$ stages
- in each stage, use measurements to decide if the nonzero is in the left or right half of the “active set”
- after subdividing $\log p$ times, return support
Experimental Results

[Arias-Castro, Candès, and Davenport - 2011]
Looking Forward
Adaptivity in Practice

- Sharp bounds to differentiate the regions where adaptivity helps and where it doesn’t

- Practical algorithms that work well for all values of $\mu$

- New theory for restricted adaptive measurements
  - single-pixel camera: 0/1 measurements
  - magnetic resonance imaging (MRI): Fourier measurements
  - analog-to-digital converters: linear filter measurements

- New sensors and architectures that can actually acquire adaptive measurements
Beyond Recovery

When and how can we directly solve inference problems directly from measurements?

- “Compressive signal processing”
- Links with machine learning
  - Johnson-Lindenstrauss lemma and geometry preservation
  - quantized compressive sensing and logistic regression
Beyond Sparsity

- Learned dictionaries, structured sparsity, models for continuous-time signals
- Multi-signal models
  - e.g., sensor networks/arrays, multi-modal data, ...
- Low-rank matrix models
- Manifold/parametric models

Acquisition
- how to design $X$
- practical devices
- adaptivity

Recovery
- practical algorithms
- robust
- stable

Inference
- classification
- estimation
- learning
Acknowledgements

- Richard Baraniuk (Rice)
- Emmanuel Candès (Stanford)
- Ery Arias-Castro (UC San Diego)
- Ronald DeVore (Texas A&M)
- Marco Duarte (UMass Amherst)
- Kevin Kelly (Rice)
- Michael Wakin (Colorado School of Mines)
http://stat.stanford.edu/~markad
markad@stanford.edu