> Given predictors $X_{n \times p}$ and response $y_{n \times 1}$,

> Linear model $y = X \beta + \varepsilon$, with $\varepsilon \sim N(0, \sigma^2)$

> Estimate $\beta$ with $(X'X)^{-1}X'y$

> Widely used in a huge amount of empirical statistical research.
Developing Trend

> Classical model requires $p < n$, but recent developments have pushed people beyond the classical model, to $p \gg n$. 
New Data Types

> **MicroArray Data**: $p$ is number of genes, $n$ is number of patients

> **Financial Data**: $p$ is number of stocks, prices, etc, $n$ is number of time points

> **Data Mining**: automated data collection can imply large numbers of variables

> **Texture Classification in Images** (eg. satellite): $p$ is number of pixels, $n$ is number of images
Estimating the model

> Can we find an estimate for $\beta$ when $p \gg n$?

> George Box (1986) *Effect-Sparsity*: the vast majority of factors have zero effect, only a small fraction actually affect the response.

> $y = X\beta + \epsilon$ can still be modeled but now $\beta$ must be *sparse*, containing a few nonzero elements, the remaining elements zero.
Commonly Used Strategies for Sparse Modeling

1. All Subsets Regression
   • Fit all possible linear models for all levels of sparsity.

2. Forward Stepwise Regression
   • Greedy approach that chooses each variable in the model sequentially by significance level.

   • ‘shrinks’ some coefficient estimates to zero.
> LASSO solves: $\min_\beta \| y - X \beta \|^2_2$ s.t. $\| \beta \|_1 \leq t$

for a choice of $t$.

> LARS: a stepwise approximation to LASSO
  • Advantage: guaranteed to stop in $n$ steps

> Allowing variables to be removed from the current set of variables, gives LASSO solution.
A New Perspective

> Up until now we’ve described the statistical view of the problem when $p \gg n$.

> Now introduce ideas from Signal Processing and a new tool for understanding regression when $p > n$, in the case of $n$ large.

> **Claim:** This will allow us to see that, for certain problems, statistical solutions such as LASSO, LARS, are just as good as all subsets regression.
Background from Signal Processing

> There exists a signal \( y \), and several ortho-bases (eg. sinusoids, wavelets, gabor).
> Concatenation of several ortho-bases is a dictionary.
> Postulate that the signal is sparsely representable, i.e. made up from few components of the dictionary.

> Motivation:
  • Image = Texture + Cartoon
  • Signal = Sinusoids + Spikes
  • Signal = CDMA + TDMA + FM + …
Overcomplete Dictionaries

Canonical Basis
- $n$ orthogonal columns

\[
\begin{pmatrix}
1 & 0 & \ldots & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & \ldots & 0 & 1 \\
\end{pmatrix}
\]

Standard Fourier Basis
- where $\omega_k = 2\pi k$, $k = 0, \ldots, n/2$
- 0,1 indicates cosine, sine
- $n$ orthogonal columns

\[
\begin{pmatrix}
\phi_{(\omega,0)} & \phi_{(\omega,1)} \\
\vdots & \vdots \\
\vdots & \vdots \\
\end{pmatrix}
\]

\[A = \begin{bmatrix} B_C \mid B_F \end{bmatrix}_{n \times 2n}\] is an overcomplete dictionary
Example: Image = Texture + Cartoon
(Elad and Starck 2003)

Original Image
Example: Image = Texture + Cartoon
(Elad and Starck 2003)

Cartoon (Curvelets)  Texture (local sinusoids)
Formal Signal Processing Problem Description

Signal decomposition: \( y = Ax \)

With a noise term: \( y = Ax + z, \quad z \sim N(0, \sigma^2) \)

<table>
<thead>
<tr>
<th>Regression</th>
<th>Decomposition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal Matrix</td>
<td>( y ) ( y )</td>
</tr>
<tr>
<td>Coefficients</td>
<td>( X ) ( A )</td>
</tr>
<tr>
<td>Noise</td>
<td>( \beta ) ( x )</td>
</tr>
<tr>
<td>( n ) observations</td>
<td>( \varepsilon ) ( z )</td>
</tr>
<tr>
<td>( p ) predictors</td>
<td>signal length</td>
</tr>
<tr>
<td>( p / n = #bases )</td>
<td></td>
</tr>
</tbody>
</table>

If \( \#bases > 1, \Rightarrow p > n. \)
Signal Processing Solutions

1. **Matching Pursuit** (Mallat, Zhang 1993)
   - Forward Stepwise Regression

2. **Basis Pursuit** (Chen, Donoho 1994)
   - Simple global optimization criteria:

   $$(P_1) \quad \min_x \|x\|_1 \text{ s.t. } y = Ax$$

3. **Maximally Sparse Solution**:
   - Intuitively most compelling but not feasible!

   $$(P_0) \quad \min_x \|x\|_0 \text{ s.t. } y = Ax$$
We can’t hope to do an all subsets search, but we are lucky!

$(P_1)$ is a convex problem, and it can sometimes solve $(P_0)$.
$(l_1, l_0)$ Equivalence

> Signal processing results show $(P_1)$ solves $(P_0)$ for certain problems.

> Donoho, Huo (IEEE IT, 2001)
> Donoho, Elad (PNAS, 2003)
> Tropp (IEEE IT, 2004)
> Gribonval (IEEE IT, 2004)
> Candès, Romberg, Tao (IEEE IT, to appear)
Phase Transition in Random Matrix Model

\[ A_{n \times p}, \quad A_{i,j} \sim N(0,1) \]

\[ y = Ax, \text{ where } x \text{ has } k \text{ random nonzeros, positions random.} \]

Phase Plane \((\delta, \rho)\)

- \(\rho = k / n\) : degree of sparsity
- \(\delta = n / p\) : degree of underdetermination

**Theorem** (DLD 2005) There exists a critical \(\rho_w(\delta)\) such that, for every \(\rho < \rho_w\), for the overwhelming majority of \((y, A)\) pairs, if \(\rho < \rho_w\), \((P_1)\) solves \((P_0)\).
Phase Transition: \((l_1, l_0)\) equivalence

\[ \rho = \frac{k}{n} \]

\[ \delta = \frac{n}{p} \]

Combinatorial Search!

\(P_1\) solves \(P_0\)
Paradigm for study

> $P$ is a property of an algorithm,
> $(y, X)$ is a random ensemble,
> Find the Phase Transitions for property $P$.

Approach pioneered by Donoho, Drori, and Tsaig:
1. Generate $y = X\beta$, where $\beta$ sparse.
2. Run full solution path to find solution $\hat{\beta}$.
3. Property $P : \frac{\|\hat{\beta} - \beta\|_2}{\|\beta\|_2} \leq \varepsilon$
This implies a statistics question!

> Could this paradigm be used for linear regression with noisy data?

> For example, when are LASSO, LARS, Forward Stepwise just as good as all subsets regression?

> Reformulate problems with Noise:

\[
(P_0, \lambda) \quad \min_\beta \ |y - X \beta|^2 + \lambda \| \beta \|_0
\]

\[
(P_1, \lambda) \quad \min_\beta \ |y - X \beta|^2 + \lambda \| \beta \|_1
\]
Experiment Setup

$X_{n \times p}$, with random entries generated from $N(0,1)$, and normalized columns.

$\beta$ is a $p$-vector with the first $k$ entries drawn from $U(0,100)$ remaining entries 0.

$\varepsilon \sim N(0,16)$ $n$-vector.

Create $y = X\beta + \varepsilon$

We find the solution $\hat{\beta}$ using an algorithm (LASSO, LARS, Forward Stepwise) with $y$ and $X$ as inputs.
Questions

> Will there be any phase transition?

> Can we learn something about the properties of these algorithms from the Phase Diagram?
LASSO, LARS Phase Transitions for Noisy Model

LASSO, $z \sim N(0, 16)$

LARS, $z \sim N(0, 16)$
Aside: Stepwise Thresholding

> Stepwise Algorithm – typical implementation:
  - Add the variable with the highest t-statistic to the model, if that t-statistic is greater than $\sqrt{2 \log(p)}$, (Bonferroni).

> Stepwise Algorithm: False Discovery Rate (FDR) Threshold:
  - Add the variable with the highest t-statistic to the model, if that t-statistic’s p-value is less than the FDR statistic.
  - $FDR_{stat} = \frac{q \times k}{p}$, where $q = E_{\#\text{falseDiscoveries}}/\#\text{totalDiscoveries}$ (the FDR parameter), $k$ is the number of variables in the current model, and $p$ is the potential number of variables.
Stepwise Phase Transitions for Noisy Model

Stepwise $\sqrt{2\log(p)}$, $z \sim N(0,16)$

Stepwise FDR, $z \sim N(0,16)$
Phase Transition Surprises

> **Surprise:** LASSO finds underlying model, for $\rho < \rho_{\text{LASSO}}$

> **Hoped for:** LARS finds underlying model, for $\rho < \rho_{\text{LARS}}$.

> **Surprise:** Stepwise only successful for $\rho \ll c \ll \rho_{\text{LASSO}}$. 
Error Analysis: Classical Forward Stepwise

> With increased noise levels, at what sparsity levels does the Stepwise Algorithm continue to recover the correct underlying model, if at all?

> How do these normalized errors compare to the ordinary least squares solution, if you had oracle knowledge of the true underlying model?

> We fix $\delta = .5$ and examine a “slice” of the phase transition diagram.
Stepwise Noise Sensitivity Analysis

Normalized $L_2$ Error for $\delta = .5$, $p=200$. Stepwise with $\sqrt{2\log(p)}$ threshold

Stepwise $\sqrt{2\log(p)}$, $\delta = .5$
Ratio of Stepwise MSE to Oracle MSE

Ratio of Pre-Breakdown MSE to Oracle MSE, with $\delta = .5$, $p=200$

Stepwise $\sqrt{2 \log(p)} \cdot \delta = .5$
Experiences with Noisy Case

> Phase Diagrams revealing, stimulating.
> Stepwise Regression falls apart at a critical sparsity level (why?)
> LARS in same cases works very well!
> Suggests other interesting properties to study.
> Other algorithms: Forward Stagewise, Backward Elimination, Stochastic Search Variable Selection, ...
Introducing SparseLab!

http://sparselab.stanford.edu

> Matlab toolbox that makes software solutions for sparse systems available.
> Growing research on sparsity, variable selection issues – could advance the research community if they have standard tools.
> SparseLab is a system to do this.
SparseLab in Depth

> Reproducible Research: SparseLab makes available the code to reproduce figures in published papers.

> Some papers currently included:
  
  - “Model Selection When the Number of Variables Exceeds the Number of Observations” (Donoho, Stodden 2006)
  - “Extensions of Compressed Sensing” (Tsaig, Donoho 2005)
  - “Neighborliness of Randomly-Projected Simplices in High Dimensions” (Donoho, Tanner 2005)
  - “High-Dimensional Centrally-Symmetric Polytopes With Neighborliness Proportional to Dimension” (Donoho 2005)

> All open source!
Acknowledgments

Iddo Drori
Joshua Sweetkind-Singer
Yaakov Tsaig