High-dimensional Statistical Models

Pradeep Ravikumar
UT Austin
MLSS 2014
Curse of Dimensionality

**Statistical Learning:** Given $n$ observations from $p(X; \theta^*)$, where $\theta^* \in \mathbb{R}^p$, recover signal/parameter $\theta^*$.

For reliable statistical learning, no. of observations $n$ should scale exponentially with the dimension of data $p$.

What if we do not have these many observations?

What if the dimension of data $p$ scales exponentially with the number of observations $n$ instead?
High-dim. Data: Imaging

Tens of thousands of “voxels” in each 3D image.

Don’t want to spend hundreds of thousands of minutes inside machine in the name of curse of dim.!
High-dim. Data: Gene (Microarray) Experiments

Tens of thousands of genes

Each experiment costs money (so no access to “exponentially” more observations)
High-dim. Data: Social Networks

Millions/Billions of nodes/parameters

Fewer observations
Linear Regression

\[ Y_i = X_i^T \theta^* + \epsilon_i, \ i = 1, \ldots, n \]

\[ Y : \text{real-valued response} \]

\[ X : \text{“covariates/features” in } \mathbb{R}^p \]

**Examples:**

**Finance:** Modeling Investment risk, Spending, Demand, etc. (responses) given market conditions (features)

**Biology:** Linking Tobacco Smoking (feature) to Mortality (response)
Linear Regression

\[ Y_i = X_i^T \theta^* + \epsilon_i, \ i = 1, \ldots, n \]

What if \( p \gg n \)?

Hope for consistent estimation even for such a high-dimensional model, if there is \textit{some} low-dimensional structure!

Sparsity: Only a few entries are non-zero
Sparse Linear Regression

\[ y = X \theta^* + w \]

Set-up: noisy observations \( y = X \theta^* + w \) with sparse \( \theta^* \)

Estimator: Lasso program

\[
\hat{\theta} \in \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \left( y_i - x_i^T \theta \right)^2 + \lambda \frac{1}{n} \sum_{j=1}^{p} |\theta_j|
\]

Some past work:
- Tibshirani, 1996
- Chen et al., 1998
- Donoho/Xue, 2001
- Tropp, 2004
- Fuchs, 2004
- Meinshausen/Buhlmann, 2005
- Candes/Tao, 2005
- Donoho, 2005
- Haupt & Nowak, 2006
- Zhao/Yu, 2006
- Wainwright, 2006
- Zou, 2006
- Koltchinskii, 2007
- Meinshausen/Yu, 2007
- Tsybakov et al., 2008

\[ \|\theta^*\|_0 = |\{ j \in \{1, \ldots, p\} : \theta^*_j \neq 0 \}| \text{ is small} \]

Estimate a sparse linear model:

\[
\min_{\theta} \| y - X \theta \|_2^2 \\
\text{s.t. } \|\theta\|_0 \leq k.
\]

\( \ell_0 \) constrained linear regression!


**Note:** The estimation problem is non-convex
$\ell_1$ Regularization

$\ell_0$ quasi-norm    $\ell_1$ norm    $\ell_2$ norm

Source: Tropp 06

$\ell_1$ norm is the closest “convex” norm to the $\ell_0$ penalty.
\( \ell_1 \) Regularization

**Estimator:** Lasso program

\[
\hat{\theta} \in \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \theta)^2 + \lambda_n \sum_{j=1}^{p} |\theta_j|
\]

Some past work: Tibshirani, 1996; Chen et al., 1998; Donoho/Xuo, 2001; Tropp, 2004; Fuchs, 2004; Meinshausen/Buhlmann, 2005; Candes/Tao, 2005; Donoho, 2005; Haupt & Nowak, 2006; Zhao/Yu, 2006; Wainwright, 2006; Zou, 2006; Koltchinskii, 2007; Meinshausen/Yu, 2007; Tsybakov et al., 2008

**Equivalent:**

\[
\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \theta)^2
\]

s.t. \( \|\theta\|_1 \leq C \).
Group-Sparsity

Parameters in groups: \( \theta = (\theta_{G_1}, \ldots, \theta_{|G_1|}, \ldots, \theta_{p-|G_m|+1}, \ldots, \theta_p) \)

A group analog of sparsity: \( \theta^* = (\star, \ldots, \star, 0, \ldots, 0, \ldots) \)

Only a few groups are active; rest are zero.
Handwriting Recognition

Data: Digit “Two” from multiple writers; Task: Recognize Digit given a new image

Could model digit recognition for each writer separately, or mix all digits for training.

Alternative: Use group sparsity. Model digit recognition for each writer, but make the models share relevant features. (Each image is represented as a vector of features)
Group-sparse Multiple Linear Regression

$m$ Response Variables:

\[ Y_{i}^{(l)} = X_{i}^{T} \Theta^{(l)} + w_{i}^{(l)}, \ i = 1, \ldots, n. \]

Collate into matrices $Y \in \mathbb{R}^{n \times m}$, $X \in \mathbb{R}^{n \times p}$ and $\Theta \in \mathbb{R}^{m \times p}$:

Multiple Linear Regression: $Y = X \Theta + W$. 
Estimate a group-sparse model where rows (groups) of \( \Theta^* \) are sparse:

\[
|\{j \in \{1, \ldots, p\} : \Theta^*_j \neq 0\}| \text{ is small.}
\]
Group Lasso

\[
\min_{\Theta} \left\{ \sum_{l=1}^{m} \sum_{i=1}^{n} (Y_{i}^{(l)} - X_{i}^{T} \Theta_{.l})^2 + \lambda \sum_{j=1}^{p} \| \Theta_{j} \|_{q} \right\}.
\]

Group analog of Lasso.

Lasso: \( \| \theta \|_{0} \rightarrow \sum_{j=1}^{p} | \theta_{j} |. \)

Group Lasso: \( \|(\| \Theta_{.j} \|_{q})\|_{0} \rightarrow \sum_{j=1}^{p} \| \Theta_{j} \|_{q} \)

(Obozinski et al; Negahban et al; Huang et al; ...)

15
Low Rank

Matrix-structured observations: $X \in \mathbb{R}^{k \times m}$, $Y \in \mathbb{R}$.

Parameters are matrices: $\Theta \in \mathbb{R}^{k \times m}$

Linear Model: $Y_i = \text{tr}(X_i \Theta) + W_i, \ i = 1, \ldots, n$.

Applications: Analysis of fMRI image data, EEG data decoding, neural response modeling, financial data.

Also arise in collaborative filtering: predicting user preferences for items (such as movies) based on their and other users’ ratings of related items.
Low Rank

Set-up: Matrix $\Theta^* \in \mathbb{R}^{k \times m}$ with rank $r \ll \min\{k, m\}$.

Estimator:

$$\hat{\Theta} \in \arg\min_{\Theta} \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle X_i, \Theta \rangle)^2 + \lambda_n \min\{k,m\} \sum_{j=1}^{\min\{k,m\}} \sigma_j(\Theta)$$

Some past work: Frieze et al., 1998; Achlioptas & McSherry, 2001; Srebro et al., 2004; Drineas et al., 2005; Rudelson & Vershynin, 2006; Recht et al., 2007; Bach, 2008; Meka et al., 2009; Candes & Tao, 2009; Keshavan et al., 2009
Nuclear Norm

Singular Values of $A \in \mathbb{R}^{k \times m}$: Square-roots of non-zero eigenvalues of $A^T A$.

Matrix Decomposition: $A = \sum_{i=1}^{r} \sigma_i u_i (v_i)^T$.

Rank of Matrix $A = |\{i \in \{1, \ldots, \min\{k,m\} \} : \sigma_i \neq 0\}|$.

Nuclear Norm is the low-rank analog of Lasso:

$$\|A\|_* = \sum_{i=1}^{\min\{k,m\}} \sigma_i.$$
High-dimensional Statistical Analysis

Typical Statistical Consistency Analysis: Holding model size ($p$) fixed, as number of samples goes to infinity, estimated parameter $\hat{\theta}$ approaches the true parameter $\theta^*$.

Meaningless in finite sample cases where $p \gg n$!

Need a new breed of modern statistical analysis: both model size $p$ and sample size $n$ go to infinity!

Typical Statistical Guarantees of Interest for an estimate $\hat{\theta}$:

- Structure Recovery e.g. is sparsity pattern of $\hat{\theta}$ same as of $\theta^*$?
- Parameter Bounds: $\|\hat{\theta} - \theta^*\|$ (e.g. $\ell_2$ error bounds)
- Risk (Loss) Bounds: difference in expected loss
Recall: Lasso

**Estimator:** Lasso program

\[
\hat{\theta} \in \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \theta)^2 + \lambda_n \sum_{j=1}^{p} |\theta_j|
\]

Some past work: Tibshirani, 1996; Chen et al., 1998; Donoho/Xuo, 2001; Tropp, 2004; Fuchs, 2004; Meinshausen/Buhlmann, 2005; Candes/Tao, 2005; Donoho, 2005; Haupt & Nowak, 2006; Zhao/Yu, 2006; Wainwright, 2006; Zou, 2006; Koltchinskii, 2007; Meinshausen/Yu, 2007; Tsybakov et al., 2008

**Statistical Assumption:** \((x_i, y_i)\) from Linear Model:

\[y_i = x_i^T \theta^* + w_i, \text{ with } w_i \sim N(0, \sigma^2).\]
Sparsistency

**Theorem.** Suppose the design matrix $X$ satisfies some conditions (to be specified later), and suppose we solve the Lasso problem with regularization penalty

$$\lambda_n > \frac{2}{\gamma} \sqrt{\frac{2\sigma^2 \log p}{n}}.$$

Then for some $c_1 > 0$, the following properties hold with probability at least $1 - 4 \exp(-c_1 n \lambda_n^2) \to 1$:

- The Lasso problem has unique solution $\hat{\theta}$ with support contained with the true support: $S(\hat{\theta}) \subseteq S(\theta^*)$.
- If $\theta_{\min}^* = \min_{j \in S(\theta^*)} |\theta_j^*| > c_2 \lambda_n$ for some $c_2 > 0$, then $S(\hat{\theta}) = S(\theta^*)$.

(Wainwright 2008; Zhao and Yu, 2006;...)