High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity

Po-Ling Loh

UC Berkeley

NIPS 2011
December 13, 2011

Joint work with Martin Wainwright
High-dimensional problems: \# parameters $p \gg \#$ observations $n$

Numerous applications in science and engineering

- DNA microarray analysis
- Health studies, longitudinal analysis
- Portfolio optimization
- Compressed sensing, MRI/fMRI
- Face recognition, spam filtering, astronomy, climatology . . .
- $p \approx 10,000$, $n \approx 100$
Sparse linear regression

- Linear model:
  \[ y_i = x_i^T \beta^* + \epsilon_i, \quad i = 1, \ldots, n \]
Sparse linear regression

Linear model:

\[ y_i = x_i^T \beta^* + \epsilon_i, \quad i = 1, \ldots, n \]

When \( p \gg n \), assume sparsity:

\[ \| \beta^* \|_0 \leq k \]
Additional complications when $Z$ observed in place of $X$
Corrupted variables

- Additional complications when $Z$ observed in place of $X$
- **Additive noise:** $Z = X + W$, where $X \perp \perp W$ and $\Sigma_w$ is known

Ex: Medical or experimental data, portfolio optimization
Corrupted variables

- Additional complications when $Z$ observed in place of $X$
- **Missing data:** entries of $X$ missing independently with probability $\alpha$

- **Ex:** Voting records, survey data, broken sensor arrays
Corrupted variables

- Additional complications when $Z$ observed in place of $X$
- **Missing data**: entries of $X$ missing independently with probability $\alpha$

Each column may have separate probability $\alpha_i$ of missing entries
Corrupted variables

- Additional complications when $Z$ observed in place of $X$
- **Missing data:** entries of $X$ missing independently with probability $\alpha$

\[ y = X\beta^* + \epsilon \]
Corrupted variables

- Additional complications when $Z$ observed in place of $X$
- **Missing data:** entries of $X$ missing independently with probability $\alpha$

Unlike EM methods, our method converges to a near-global optimum despite non-convexity
Note that

\[ \hat{\beta} \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \beta^T \Sigma_x \beta - \beta^* \Sigma_x \beta \right\} \]
Lasso as plug-in estimator

- Note that

\[ \beta^* \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \beta^T \Sigma_x \beta - \beta^*^T \Sigma_x \beta \right\} \]

- Compare to Lasso (Tibshirani '96):

\[ \hat{\beta} \in \arg \min_{\|\beta\|_1 \leq R} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 \right\} \]
Lasso as plug-in estimator

- Note that

\[ \beta^* \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \beta^T \Sigma_x \beta - \beta^*^T \Sigma_x \beta \right\} \]

- Compare to Lasso (Tibshirani ’96):

\[ \hat{\beta} \in \arg \min_{\|\beta\|_1 \leq R} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 \right\} \]

\[ = \arg \min_{\|\beta\|_1 \leq R} \left\{ \frac{1}{2} \beta^T \left( \frac{X^T X}{n} \right) \beta - \frac{y^T X}{n} \beta \right\} \]
Lasso as plug-in estimator

- Note that

\[ \beta^* \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \beta^T \Sigma_x \beta - \beta^* T \Sigma_x \beta \right\} \]

- **Idea:** form unbiased estimators \((\hat{\Gamma}, \hat{\gamma})\) of \((\Sigma_x, \text{Cov}(X, y))\) based on \((y, Z)\), solve constrained program

\[ \hat{\beta} \in \arg \min_{\|\beta\|_1 \leq R} \left\{ \frac{1}{2} \beta^T \hat{\Gamma} \beta - \hat{\gamma}^T \beta \right\} \]
Example: Additive noise

Since $Z = X + W$ and $X \perp \perp W$, we have $\Sigma_z = \Sigma_x + \Sigma_w$ and $\text{Cov}(y, X) = \text{Cov}(y, Z)$.
Example: Additive noise

Since $Z = X + W$ and $X \perp W$, we have $\Sigma_z = \Sigma_x + \Sigma_w$ and $\text{Cov}(y, X) = \text{Cov}(y, Z)$

Use

$$\hat{\Gamma} = \frac{Z^T Z}{n} - \Sigma_w, \quad \hat{\gamma} = \frac{Z^T y}{n}$$
Example: Additive noise

Since $Z = X + W$ and $X \perp \! \! \! \perp W$, we have $\Sigma_z = \Sigma_x + \Sigma_w$ and $\text{Cov}(y, X) = \text{Cov}(y, Z)$

Use

$$\hat{r} = \frac{Z^T Z}{n} - \Sigma_w, \quad \hat{\gamma} = \frac{Z^T y}{n}$$

Objective:

$$\hat{\beta} \in \arg \min_{\|\beta\|_1 \leq R} \left\{ \frac{1}{2} \beta^T \left( \frac{Z^T Z}{n} - \Sigma_w \right) \beta - \frac{y^T Z}{n} \beta \right\}$$
Example: Missing data

- $X \sim N(0, \Sigma_x)$, $Z \in \mathbb{R}^{n \times p}$ is observed data with probability $\alpha$ of missing values.
Example: Missing data

- \( X \sim N(0, \Sigma_x) \), \( Z \in \mathbb{R}^{n \times p} \) is observed data with probability \( \alpha \) of missing values

- Let

\[
\hat{Z}_{ij} = \begin{cases} 
\frac{Z_{ij}}{1-\alpha} & \text{if } Z_{ij} \text{ is observed} \\
0 & \text{otherwise}
\end{cases}
\]

- Then

\[
\hat{\Gamma} = \frac{\hat{Z}^T \hat{Z}}{n} - \alpha \text{diag} \left( \frac{\hat{Z}^T \hat{Z}}{n} \right)
\]

satisfies \( \mathbb{E}(\hat{\Gamma}) = \Sigma_x \) and \( \text{Cov}(\hat{Z}, y) = \text{Cov}(X, y) \)

- Objective:

\[
\hat{\beta} \in \arg \min_{\|\beta\|_1 \leq R} \left\{ \frac{1}{2} \beta^T \hat{\Gamma} \beta - \frac{y^T \hat{Z}}{n} \beta \right\}
\]
High-dimensional consistency?

Additive noise

\[ \| \hat{\beta} - \beta^* \|_2 \to 0 \quad \text{as} \quad n \to \infty \]
High-dimensional consistency?

- Modified Lasso with additive noise, \( k \approx \sqrt{p} \)
- Consistency: \( \| \hat{\beta} - \beta^* \|_2 \rightarrow 0 \) as \( n \rightarrow \infty \)
Theoretical guarantees: canonical Lasso

- Under restricted eigenvalue conditions on $X$ (Bickel, Ritov & Tsybakov '08, van de Geer & Bühlmann '09),

$$
\| \hat{\beta} - \beta^* \|_1 = O \left( k \sqrt{\frac{\log p}{n}} \right), \quad \| \hat{\beta} - \beta^* \|_2 = O \left( \sqrt{\frac{k \log p}{n}} \right)
$$

- RE conditions hold w.h.p. when $X$ is a random matrix with rows sampled i.i.d. from a (sub)-Gaussian distribution (Raskutti et al. '09)
Theoretical guarantees: modified Lasso

Theorem (Statistical error)

Under modified RE condition $\hat{\Gamma}$ and deviation conditions on $(\hat{\gamma}, \hat{\Gamma})$, any global optimum $\hat{\beta}$ satisfies

$$\|\hat{\beta} - \beta^*\|_1 \lesssim \varphi(\sigma_\epsilon) \left( k \sqrt{\frac{\log p}{n}} \right), \quad \|\hat{\beta} - \beta^*\|_2 \lesssim \varphi(\sigma_\epsilon) \left( \sqrt{\frac{k \log p}{n}} \right)$$

Deviation conditions:

$$\|\hat{\gamma} - \text{Cov}(X, y)\|_\infty, \quad \|\hat{\Gamma} - \Sigma_X\beta^*\|_\infty \lesssim \varphi(\sigma_\epsilon) \left( \sqrt{\frac{\log p}{n}} \right)$$
Theoretical guarantees: modified Lasso

**Theorem (Statistical error)**

*Under modified RE condition $\hat{\Gamma}$ and deviation conditions on $(\hat{\gamma}, \hat{\Gamma})$, any global optimum $\hat{\beta}$ satisfies*

\[
\|\hat{\beta} - \beta^*\|_1 \lesssim \varphi(\sigma_\epsilon) \left( k \sqrt{\frac{\log p}{n}} \right), \quad \|\hat{\beta} - \beta^*\|_2 \lesssim \varphi(\sigma_\epsilon) \left( \sqrt{\frac{k \log p}{n}} \right)
\]

- **Deviation conditions:**

\[\|\hat{\gamma} - \text{Cov}(X, y)\|_\infty, \quad \|\hat{\Gamma} - \Sigma_X\beta^*\|_\infty \lesssim \varphi(\sigma_\epsilon) \left( \sqrt{\frac{\log p}{n}} \right)\]

- $\varphi(\sigma_\epsilon)$ is a function of corruption pattern and noise variance, decreases with SNR and increases with $\alpha$
High-dimensional consistency?

- Modified Lasso with additive noise, $k \approx \sqrt{p}$
- Consistency: $\|\hat{\beta} - \beta^*\|_2 \to 0$ as $n \to \infty$
High-dimensional consistency?

- $\ell_2$-error vs. rescaled sample size $n/(k \log p)$
- Curves stack up, verifying theoretical results
Optimization of objective

- Corrected objective is **not** convex

\[ \hat{\beta} \in \arg \min_{\|\beta\|_1 \leq R} \left\{ \frac{1}{2} \beta^T \left( \frac{Z^T Z}{n} - \Sigma_w \right) \beta - \frac{y^T Z}{n} \beta \right\} \]

- Hessian has at least \( p - n \) negative eigenvalues
Corrected objective is **not** convex

\[ \hat{\beta} \in \arg \min_{\|\beta\|_1 \leq R} \left\{ \frac{1}{2} \beta^T \left( \frac{Z^T Z}{n} - \Sigma_w \right) \beta - \frac{y^T Z}{n} \beta \right\} \]

- Hessian has at least \( p - n \) negative eigenvalues

- Pretend objective is convex, apply projected gradient descent algorithm
Projected gradient descent

- Solve constrained optimization problem

$$\min_{\beta} \frac{1}{2n} \|y - X\beta\|_2^2 \quad \text{s.t.} \quad \|\beta\|_1 \leq R$$
Projected gradient descent

- Solve constrained optimization problem

\[
\min_{\beta} \frac{1}{2n} \|y - X\beta\|^2_2 \quad \text{s.t.} \quad \|\beta\|_1 \leq R
\]

- Produces iterates \( \beta^t \) with

\[
\beta^{t+1} = \Pi \left( \beta^t - \frac{1}{\eta} \nabla \mathcal{L}(\beta^t) \right), \quad \text{stepsize} \quad \eta > 0
\]
Projected gradient descent

- Linear convergence when $\mathcal{L}$ is smooth and strongly convex (Bertsekas '95):

\[ \|\beta^t - \hat{\beta}\|_2 \leq \gamma^t \|\beta^0 - \hat{\beta}\|_2 \]
Projected gradient descent

- Linear convergence when $\mathcal{L}$ is smooth and strongly convex (Bertsekas '95):
  \[ \| \beta^t - \hat{\beta} \|_2 \leq \gamma^t \| \beta^0 - \hat{\beta} \|_2 \]

- When $\mathcal{L}$ non-convex, projected gradient descent may not converge, or converge at slower rates
Global linear convergence observed in practice

- For fixed problem instance, 10 runs of projected gradient descent, plotted optimization error $\| \beta^t - \hat{\beta} \|_2$
- $p = 512$, $k \approx \sqrt{p}$, $n \approx 5k \log p$
Theoretical guarantees: modified Lasso

Theorem (Optimization error)

For the modified Lasso,

$$\|\beta^t - \hat{\beta}\|_2 \leq \gamma^t \|\beta^0 - \hat{\beta}\|_2 + o\left(\sqrt{\frac{k \log p}{n}}\right)$$
Theoretical guarantees: modified Lasso

Theorem (Optimization error)

For the modified Lasso,

$$
\|\beta^t - \hat{\beta}\|_2 \leq \gamma^t \|\beta^0 - \hat{\beta}\|_2 + o\left(\sqrt{\frac{k \log p}{n}}\right)
$$

- Use results from Agarwal, Negahban & Wainwright (NIPS ’10), applied to non-convex objective
- Requires restricted strong convexity (RSC) and restricted smoothness (RSM), holding w.h.p. in settings of interest
Illustration of statistical and optimization error

Statistical error: \[ \| \hat{\beta} - \beta^* \|_2 = \mathcal{O} \left( \sqrt{\frac{k \log p}{n}} \right) \]

Optimization error: \[ \| \beta^t - \hat{\beta} \|_2 = \gamma^t \| \beta^0 - \hat{\beta} \|_2 + o \left( \sqrt{\frac{k \log p}{n}} \right) \]
Conditional independence property for graphical model:

\[ X_u \mid X_{V \backslash \{u\}} \overset{d}{=} X_u \mid X_{N(u)} \]
Application: Gaussian graphical models

- Conditional independence property for graphical model:
  \[ X_u \mid X_{V \setminus \{u\}} \overset{d}{=} X_u \mid X_{N(u)} \]

- When \( X \sim N(0, \Sigma) \), entries of \( \Theta = \Sigma^{-1} \) may be recovered via nodewise linear regression (Meinshausen and Bühlmann ’06, Yuan ’10)

  \[
  \text{sparsity } k \iff \text{max degree of vertex}
  \]
Application: Gaussian graphical models

- Conditional independence property for graphical model:

\[ X_u \mid X_{V \setminus \{u\}} \overset{d}{=} X_u \mid X_{N(u)} \]

- When \( X \sim N(0, \Sigma) \), entries of \( \Theta = \Sigma^{-1} \) may be recovered via nodewise linear regression (Meinshausen and Bühlmann ’06, Yuan ’10)

\[ \text{sparsity } k \iff \text{max degree of vertex} \]

- For corrupted observations, use noisy regression to recover \( \Theta \)
Gaussian inverse covariance estimation

Theorem (Spectral norm consistency)

For estimate $\hat{\Theta}$ based on corrupted observations of a Gaussian graphical model,

$$\| \hat{\Theta} - \Theta \|_{op} = O\left( k \sqrt{\frac{\log p}{n}} \right)$$

- Matches rates for fully-observed case
• Provided a Lasso variant based on noisy observations \((y, Z)\), such that

\[
\|\hat{\beta} - \beta^*\|_2 = O \left( \sqrt{\frac{k \log p}{n}} \right)
\]

• Derived an estimator for the inverse covariance matrix of a (noisy) Gaussian graphical model, such that

\[
\|\hat{\Theta} - \Theta\|_{op} = O \left( k \sqrt{\frac{\log p}{n}} \right)
\]

• Demonstrated that global minimizers \(\hat{\beta}\) for the non-convex objective can be obtained via projected gradient descent
Open questions

- Support recovery for corrupted observations
- Minimax lower bounds
- Additive noise model with unknown $\Sigma_w$
- Other corruption patterns: multiplicative noise, censored data
Form of $\varphi$

- **Additive noise:** $X_i \sim N(0, \sigma_x^2 I), W_i \sim N(0, \sigma_w^2 I)$:

  $$\varphi = \sqrt{1 + \frac{\sigma_w^2}{\sigma_x^2}} \sqrt{\frac{\sigma_w^2}{\sigma_x^2} + \frac{\sigma_x^2}{\sigma_x^2}}$$

- **Missing data:** $X_i \sim N(0, \sigma_x^2 I)$,

  $$\varphi = \frac{\sigma_x}{\sigma_x(1 - \alpha)} + \frac{1}{(1 - \alpha)^2}$$
Estimation of $\Theta$

Algorithm:

- Perform $p$ linear regressions of the variables $Z^i$ upon the remaining variables $Z^{-i}$, using the modified Lasso program with estimators $(\hat{\Gamma}(i), \hat{\gamma}(i))$
- Estimate scalars $a_i$ using plug-in estimator $\hat{a}_i = - (\hat{\Gamma}_{ii} - \hat{\Gamma}_{i,-i} \hat{\theta}^i)^{-1}$
- Form the matrix $\tilde{\Theta}$ with $\tilde{\Theta}_{i,-i} = \hat{a}_i \hat{\theta}^i$ and $\tilde{\Theta}_{ii} = -\hat{a}_i$
- Symmetrize: $\hat{\Theta} \in \operatorname{arg min}_{\Theta \in S^p} \|\Theta - \hat{\Theta}\|_{\ell_1 \to \ell_1}$

- Last step is an LP, can be optimized with standard techniques
Some references

