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

Po-Ling Loh

UC Berkeley

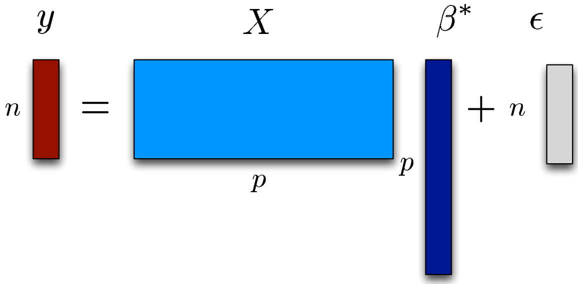
NIPS 2011
December 13, 2011

Joint work with Martin Wainwright

Introduction

- 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 ...
  - $p \approx 10,000, n \approx 100$

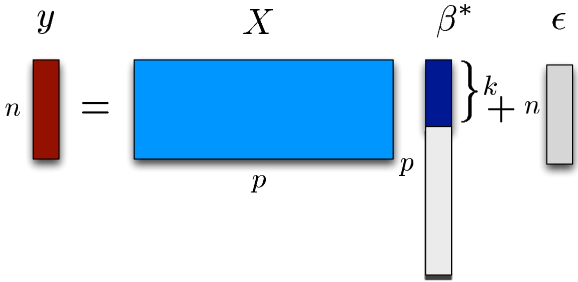
Sparse linear regression



- Linear model:

$$y_i = x_i^T \beta^* + \epsilon_i, \qquad i = 1, \ldots, n$$

Sparse linear regression



- Linear model:

$$y_i = x_i^T \beta^* + \epsilon_i, \qquad i = 1, \ldots, n$$

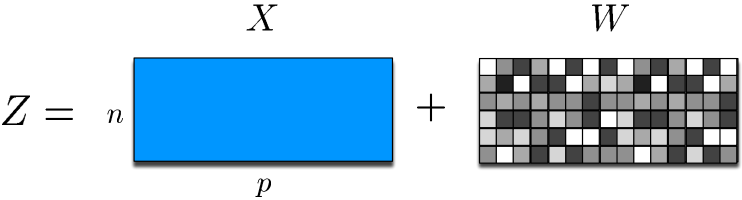
- When $p \gg n$, assume sparsity: $\|\beta^*\|_0 \leq k$

Corrupted variables

- Additional complications when Z observed in place of X

Corrupted variables

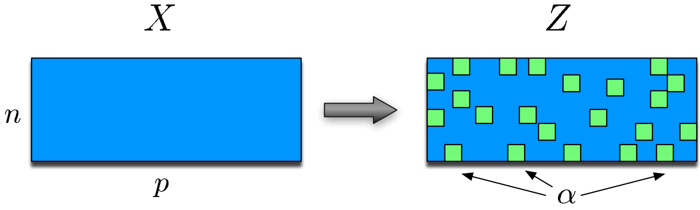
- Additional complications when Z observed in place of X
- **Additive noise:** $Z = X + W$, where $X \perp\!\!\!\perp W$ and Σ_w is known



- **Ex:** Medical or experimental data, portfolio optimization

Corrupted variables

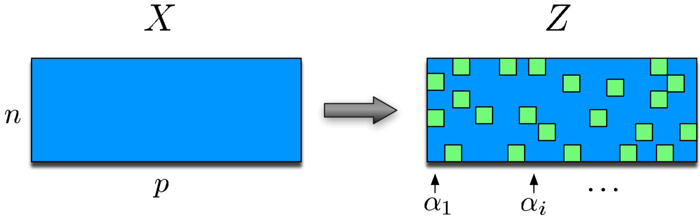
- Additional complications when Z observed in place of X
- **Missing data:** entries of X missing independently with probability α



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

Corrupted variables

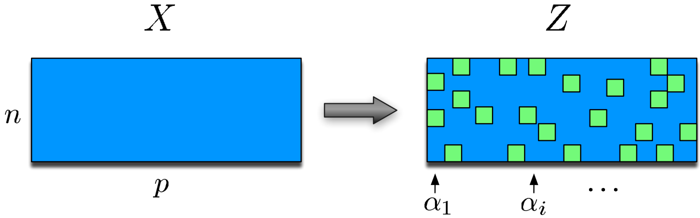
- Additional complications when Z observed in place of X
- **Missing data:** entries of X missing independently with probability α



- Each column may have separate probability α_i of missing entries

Corrupted variables

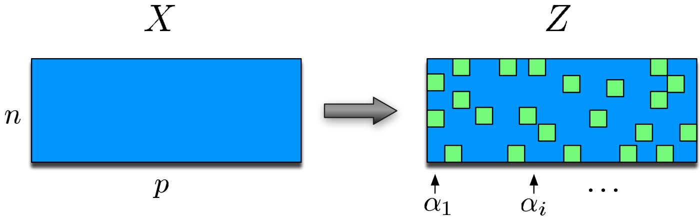
- Additional complications when Z observed in place of X
- **Missing data:** entries of X missing independently with probability α



- **Model:** $$y = X\beta^* + \epsilon$$

Corrupted variables

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



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

Lasso as plug-in estimator

- Note that

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

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):

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

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):

$$\begin{aligned} \widehat{\beta} &\in \arg\min_{\|\beta\|_1 \leq R} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 \right\} \\ &= \arg\min_{\|\beta\|_1 \leq R} \left\{ \frac{1}{2} \beta^T \left(\frac{X^T X}{n} \right) \beta - \frac{y^T X}{n} \beta \right\} \end{aligned}$$

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 $(\widehat{\Gamma}, \widehat{\gamma})$ of $(\Sigma_x, \mathrm{Cov}(X, y))$ based on (y, Z), solve constrained program

$$\boxed{\widehat{\beta} \in \arg\min_{\|\beta\|_1 \le R} \left\{ \frac{1}{2} \beta^T \widehat{\Gamma} \beta - \widehat{\gamma}^T \beta \right\}}$$

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

$$\widehat{\Gamma} = \frac{Z^T Z}{n} - \Sigma_w, \qquad \widehat{\gamma} = \frac{Z^T y}{n}$$

Example: Additive noise

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

$$\widehat{\Gamma} = \frac{Z^T Z}{n} - \Sigma_w, \qquad \widehat{\gamma} = \frac{Z^T y}{n}$$

- Objective:

$$\widehat{\beta} \in \arg \min_{\|\beta\|_1 \le R} \left\{ \frac{1}{2} \beta^T \left(\frac{Z^T Z}{n} - \Sigma_w \right) \beta - \frac{y^T Z}{n} \beta \right\}$$

Example: Missing data

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

Example: Missing data

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

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

- Then

$$\widehat{\Gamma} = \frac{\widehat{Z}^T \widehat{Z}}{n} - \alpha \, \mathrm{diag}\left(\frac{\widehat{Z}^T \widehat{Z}}{n} \right)$$

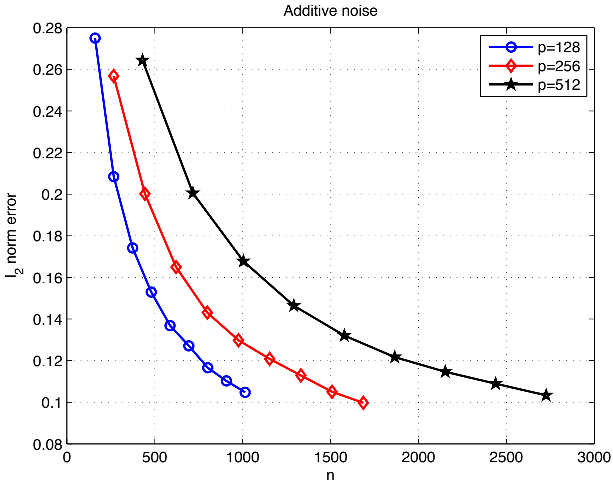
  satisfies $\mathbb{E}(\widehat{\Gamma}) = \Sigma_x$ and $\mathrm{Cov}(\widehat{Z}, y) = \mathrm{Cov}(X, y)$
- Objective:

$$\widehat{\beta} \in \arg \min_{\|\beta\|_1 \leq R} \left\{ \frac{1}{2} \beta^T \widehat{\Gamma} \beta - \frac{y^T \widehat{Z}}{n} \beta \right\}$$

High-dimensional consistency?

High-dimensional consistency?

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



Theoretical guarantees: canonical Lasso

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

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

- 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 $\widehat{\Gamma}$ and deviation conditions on $(\widehat{\gamma}, \widehat{\Gamma})$, **any** *global optimum $\widehat{\beta}$ satisfies*

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

- Deviation conditions:

$$\|\widehat{\gamma} - \mathrm{Cov}(X, y)\|_\infty, \quad \|(\widehat{\Gamma} - \Sigma_x)\beta^*\|_\infty \precsim \varphi(\sigma_\epsilon) \left(\sqrt{\frac{\log p}{n}} \right)$$

Theoretical guarantees: modified Lasso

Theorem (Statistical error)

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

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

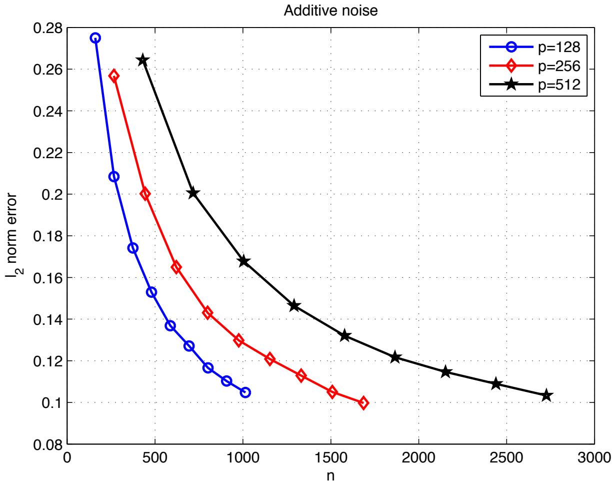
- Deviation conditions:

$$\|\widehat{\gamma} - \mathrm{Cov}(X, y)\|_\infty, \quad \|(\widehat{\Gamma} - \Sigma_x)\beta^*\|_\infty \precsim \varphi(\sigma_\epsilon) \left(\sqrt{\frac{\log p}{n}} \right)$$

- $\varphi(\sigma_\epsilon)$ is a function of corruption pattern and noise variance, decreases with SNR and increases with α

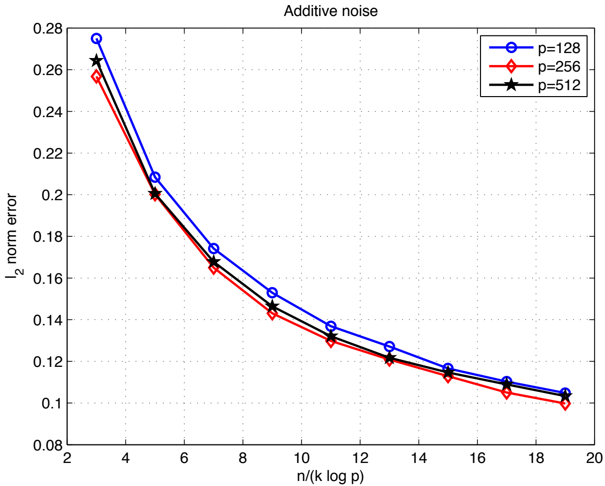
High-dimensional consistency?

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



High-dimensional consistency?

- ℓ_2-error vs. rescaled sample size $n/(k \log p)$
- Curves stack up, verifying theoretical results



Optimization of objective

- Corrected objective is **not** convex

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

- Hessian has at least $p - n$ negative eigenvalues

Optimization of objective

- Corrected objective is **not** convex

$$\widehat{\beta} \in \arg\min_{\|\beta\|_1 \le R} \left\{ \frac{1}{2} \beta^T \left(\frac{Z^T Z}{n} - \Sigma_w \right) \beta - \frac{y^T Z}{n} \beta \right\}$$

- Hessian has at least $p - n$ negative eigenvalues

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

Projected gradient descent

- Solve constrained optimization problem

$$\min_{\beta} \underbrace{\frac{1}{2n}\|y - X\beta\|_2^2}_{\mathcal{L}(\beta)} \quad \text{s.t.} \quad \|\beta\|_1 \leq R$$

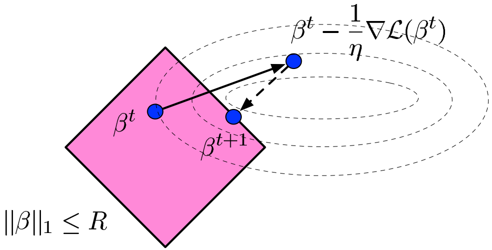
Projected gradient descent

- Solve constrained optimization problem

$$\min_{\beta} \underbrace{\frac{1}{2n}\|y - X\beta\|_2^2}_{\mathcal{L}(\beta)} \quad \text{s.t.} \quad \|\beta\|_1 \leq R$$

- Produces iterates β^t with

$$\beta^{t+1} = \Pi\left(\beta^t - \frac{1}{\eta}\nabla\mathcal{L}(\beta^t)\right), \qquad \text{stepsize} \quad \eta > 0$$



Projected gradient descent

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

$$\|\beta^t - \widehat{\beta}\|_2 \leq \gamma^t \|\beta^0 - \widehat{\beta}\|_2$$

Projected gradient descent

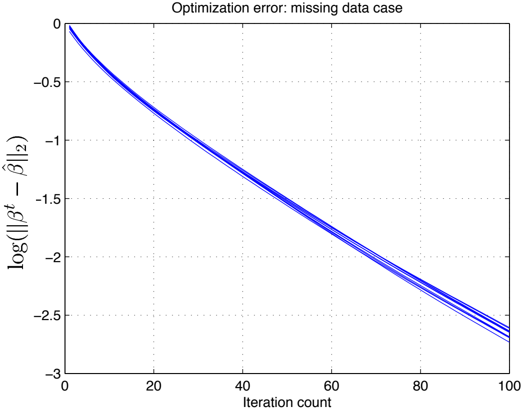
- Linear convergence when $\mathcal{L}$ is smooth and strongly convex (Bertsekas '95):

$$\|\beta^t - \widehat{\beta}\|_2 \leq \gamma^t \|\beta^0 - \widehat{\beta}\|_2$$

- **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 $\|\beta^t - \widehat{\beta}\|_2$
- $p = 512,\ k \approx \sqrt{p},\ n \approx 5k \log p$



Theoretical guarantees: modified Lasso

Theorem (Optimization error)

For the modified Lasso,

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

Theoretical guarantees: modified Lasso

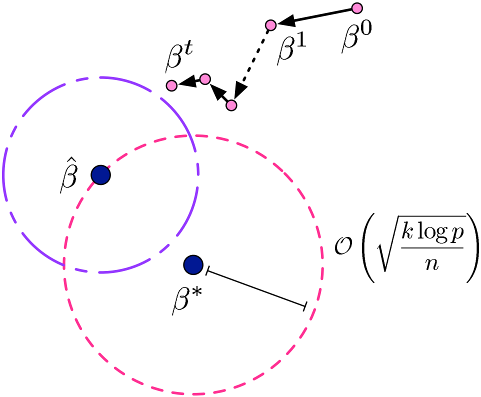
Theorem (Optimization error)

For the modified Lasso,

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

- 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

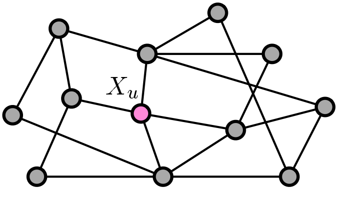


- Statistical error: $\|\widehat{\beta} - \beta^*\|_2 = \mathcal{O}\left(\sqrt{\frac{k \log p}{n}}\right)$
- Optimization error: $\|\beta^t - \widehat{\beta}\|_2 = \gamma^t \|\beta^0 - \widehat{\beta}\|_2 + o\left(\sqrt{\frac{k \log p}{n}}\right)$

Application: Gaussian graphical models

- Conditional independence property for graphical model:

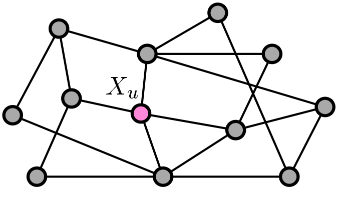
$$X_u \mid X_{V \setminus \{u\}} \stackrel{d}{=} X_u \mid X_{N(u)}$$

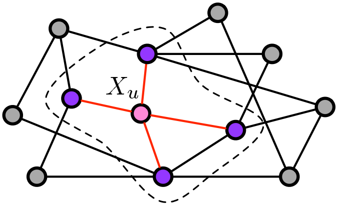


Application: Gaussian graphical models

- Conditional independence property for graphical model:

$$X_u \mid X_{V\setminus\{u\}} \stackrel{d}{=} X_u \mid X_{N(u)}$$





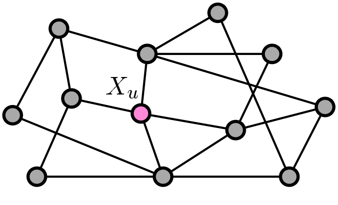
- When $X \sim N(0, \Sigma)$, entries of $\Theta = \Sigma^{-1}$ may be recovered via nodewise linear regression (Meinshausen and Bühlmann '06, Yuan '10)

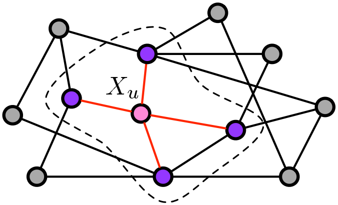
$$\text{sparsity } k \iff \text{max degree of vertex}$$

Application: Gaussian graphical models

- Conditional independence property for graphical model:

$$X_u \mid X_{V \setminus \{u\}} \stackrel{d}{=} X_u \mid X_{N(u)}$$





- 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 Θ**

Gaussian inverse covariance estimation

Theorem (Spectral norm consistency)

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

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

- Matches rates for fully-observed case

Summary

- Provided a Lasso variant based on noisy observations (y, Z), such that

$$\|\widehat{\beta} - \beta^*\|_2 = \mathcal{O}\left(\sqrt{\frac{k \log p}{n}}\right)$$

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

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

- Demonstrated that global minimizers $\widehat{\beta}$ 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 Σ_w
- Other corruption patterns: multiplicative noise, censored data

Form of φ

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

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

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

$$\varphi = \frac{\sigma_\epsilon}{\sigma_x(1-\alpha)} + \frac{1}{(1-\alpha)^2}$$

Estimation of Θ

Algorithm:

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

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

Some references

- A. Agarwal, S. Negahban, and M.J. Wainwright (2011). Fast global convergence of gradient methods for high-dimensional statistical recovery. ArXiv paper.
- P. Loh and M.J. Wainwright (2011). High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. ArXiv paper.
- N. Meinshausen and P. Bühlmann (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics.*
- G. Raskutti, M.J. Wainwright, and B. Yu (2010). Restricted eigenvalue properties for correlated Gaussian random designs. *Journal of Machine Learning Research.*
- M. Rosenbaum and A.B. Tsybakov (2010). Sparse recovery under matrix uncertainty. *Annals of Statistics.*
- R. Tibshirani (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B.*
- M. Yuan (2010). High-dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research.*