Elementary Estimators for High-Dimensional Statistical Models

Pradeep Ravikumar

Joint work with Eunho Yang and Aurélie C. Lozano

University of Texas, Austin

Jun. 26, 2014

Background - High-Dimensional Statistics

- When the ambient dimension p is **larger than** sample size n
- Need structural constraints on high-dimensional statistical models
  - Sparsity: only a small number of entries are non-zeros
  - Group sparsity: only small number of groups are non-zeros
  - Low rank: when the parameters are matrix-structured

$$\vdots$$

Background - High-Dimensional Statistics

- When the ambient dimension p is **larger than** sample size n

- Need structural constraints on high-dimensional statistical models
  - Sparsity: only a small number of entries are non-zeros
  - Group sparsity: only small number of groups are non-zeros
  - Low rank: when the parameters are matrix-structured

$$\vdots$$

- ex) Linear models, $y = X\theta^* + w$:
  - minimize $\frac{1}{2n}\|X\theta - y\|_2^2 + \lambda_n\|\theta\|_1$

Background - High-Dimensional Statistics

- When the ambient dimension p is **larger than** sample size n

- Need structural constraints on high-dimensional statistical models
  - Sparsity: only a small number of entries are non-zeros
  - Group sparsity: only small number of groups are non-zeros
  - Low rank: when the parameters are matrix-structured

$$\vdots$$

- ex) Linear models, $y = X\theta^* + w$:
  - minimize $\frac{1}{2n}\|X\theta - y\|_2^2 + \lambda_n\|\theta\|_1$
- ex) Gaussian graphical models, $P(y;\Theta^*) \propto \exp\left\{-\frac{1}{2}\langle\langle yy^\top, \Theta^*\rangle\rangle - A(\Theta^*)\right\}$:
  - minimize $\langle\langle S, \Theta\rangle\rangle - \operatorname{logdet}\Theta + \lambda_n\|\Theta\|_1$

    where $S := \frac{1}{n}\sum_{i=1}^n (x^{(i)} - \bar{x})(x^{(i)} - \bar{x})^\top$, and $\bar{x} := \frac{1}{n}\sum_{i=1}^n x^{(i)}$.

Background - High-Dimensional Statistics

- When the ambient dimension p is **larger than** sample size n
- Surge of Recent Work:
  - (Linear models:) Tibshirani (1996); van de Geer and Buhlmann (2009); Meinshausen and Yu (2009); Candes and Tao (2006); Meinshausen and Bühlmann (2006); Wainwright (2009); Zhao and Yu (2006); Tropp et al. (2006); Zhao et al. (2009); Yuan and Lin (2006); Jacob et al. (2009); Lounici et al. (2009); Baraniuk et al. (2008); Recht et al. (2010); Bach (2008); Negahban et al. (2012) . . .
  - (Inverse covariance estimation:) Yuan and Lin (2007); Friedman et al. (2007); Bannerjee et al. (2008); Ravikumar et al. (2011); Boyd and Vandenberghe (2004); Friedman et al. (2007); Bannerjee et al. (2008); Meinshausen and Bühlmann (2006); Cai et al. (2011) . . .

$$\vdots$$

Background - High-Dimensional Statistics

- When the ambient dimension p is **larger than** sample size n
- Surge of Recent Work:
  - (Linear models:) Tibshirani (1996); van de Geer and Buhlmann (2009); Meinshausen and Yu (2009); Candes and Tao (2006); Meinshausen and Bühlmann (2006); Wainwright (2009); Zhao and Yu (2006); Tropp et al. (2006); Zhao et al. (2009); Yuan and Lin (2006); Jacob et al. (2009); Lounici et al. (2009); Baraniuk et al. (2008); Recht et al. (2010); Bach (2008); Negahban et al. (2012) . . .
  - (Inverse covariance estimation:) Yuan and Lin (2007); Friedman et al. (2007); Bannerjee et al. (2008); Ravikumar et al. (2011); Boyd and Vandenberghe (2004); Friedman et al. (2007); Bannerjee et al. (2008); Meinshausen and Bühlmann (2006); Cai et al. (2011) . . .

$$\vdots$$

Still expensive for very-large scale problems!

Main Question: If we restrict to **closed-form** estimators; can we nonetheless obtain **consistent estimators** with sharp convergence rates?

Why Closed-Form Estimators?

- Current approach to structurally constrained statistical model estimation is two-staged:
  - **Statistical:** Devise regularized likelihood-based statistical estimators
  - **Computational:** Devise efficient optimization methods, allied with parallel/distributed frameworks, to solve these estimators — increasingly important in modern Big Data settings
- **Comptastical** Approach: devise statistical estimators with computational constraints in mind
  - Closed-form estimators are a particularly stringent class of computational constraints
  - As we will show, they can nonetheless enjoy strong statistical guarantees!

1. Elementary Estimators for General Moment Parameters

2. Elementary Estimators for Linear Models

3. Elementary Estimators for Gaussian Graphical Models

Moment Parameter Estimation

- $X \in \mathbb{R}^p$: Random vector with distribution $\mathbb{P}$,
- $\{X_i\}_{i=1}^n$: n i.i.d. observations drawn from $\mathbb{P}$.

Goal: Estimating moment parameter $\mu^* := \mathbb{E}[\phi(X)]$, where $\phi : \mathbb{R}^p \mapsto \mathbb{R}^m$ is vector-valued feature function

WHY NOT Regularized Likelihood-Based Estimators?

- A natural distributional setting: **Exponential family**, with sufficient statistics set $\phi(X)$:

$$\mathbb{P}(X;\theta) = \exp\left\{\langle\theta, \phi(X)\rangle - A(\theta)\right\}$$

- A natural estimator is the ℓ_1-regularized MLE:

$$\underset{\mu}{\text{minimize}}\left\{ \underbrace{-\langle\theta(\mu), \widehat{\mu}_n\rangle + A\big(\theta(\mu)\big)}_{\text{Negative log-likelihood } \mathcal{L}(\mu)} + \|\mu\|_1 \right\}$$

where $\widehat{\mu}_n$ is the *sample* moment: $\frac{1}{n}\sum_{i=1}^{n}\phi(X_i)$

WHY NOT Regularized Likelihood-Based Estimators?

- Let us derive a "Dantzig variant" in this general setting. We have:

$$\nabla \mathcal{L}(\mu) = -\nabla^2 A^*(\mu)\, \widehat{\mu}_n + \nabla^2 A^*(\mu)\, \nabla A\big(\theta(\mu)\big) = \nabla^2 A^*(\mu)\big(-\widehat{\mu}_n + \mu\big).$$

- Then the "Dantzig variant" of the structured moment estimator:

$$\begin{aligned} &\underset{\mu}{\text{minimize}}\ \|\mu\|_1 \\ &\qquad \text{s.t.}\ \Big\|\nabla^2 A^*(\mu)(\mu - \widehat{\mu}_n)\Big\|_\infty \leq \lambda_n, \end{aligned}$$

Proposition: *The estimation problems above are both* **non-convex** *for general exponential families!*

General Structured Moment Estimation

- *Our estimator for general structurally constrained moment parameters:*

$$\begin{aligned} &\underset{\mu}{\text{minimize}}\ \mathcal{R}(\mu) \\ &\qquad \text{s.t.}\ \mathcal{R}^*\left(\mu - \widehat{\mu}_n\right) \leq \lambda_n, \end{aligned}$$

where $\mathcal{R}^*(a) = \sup_{b:\mathcal{R}(b)\neq 0} \frac{\langle a,b\rangle}{\mathcal{R}(b)}$.

- The optimal solution $\widehat{\mu}$ has a **closed-form** solution! (Provided $\mathcal{R}(\cdot)$ is atomic norm (Chandrasekaran et al., 2010))

Statistical Guarantees for General Structures

- Our estimator for general structure:

$$\begin{aligned} &\text{minimize } \mathcal{R}(\mu) \\ &\qquad \text{s.t. } \mathcal{R}^*\big(\mu - \widehat{\mu}_n\big) \le \lambda_n \end{aligned}$$

Theorem

Suppose that the population mean parameter μ^* lies in some low dimensional space $\mathcal{M}$, and that $\mathcal{R}(\cdot)$ is decomposable w.r.t. $\mathcal{M}$. Also suppose that we set $\lambda_n \ge \mathcal{R}^*(\mu^* - \widehat{\mu}_n)$. Then,

$$\begin{aligned} \mathcal{R}^*(\widehat{\mu} - \mu^*) &\le \ 2\lambda_n \ , \\ \|\widehat{\mu} - \mu^*\|_2 &\le \ 4\lambda_n \Psi \ , \\ \mathcal{R}(\widehat{\mu} - \mu^*) &\le \ 8\lambda_n \Psi^2 \end{aligned}$$

where $\Psi := \sup_{u \in \mathcal{M}\setminus\{0\}} \frac{\mathcal{R}(u)}{\|u\|}$.

General Moment Estimation - Sparsity Case

- Our estimator for arbitrary moment parameters: given empirical moment $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n}\phi(X_i)$,

$$\begin{aligned} \underset{\mu}{\text{minimize}} \ & \|\mu\|_1 \\ \text{s. t.} \ & \left\|\mu - \widehat{\mu}_n\right\|_\infty \leq \lambda_n, \end{aligned}$$

Statistical Guarantees - Sparsity Case

- Our estimator for sparsity case:

$$\begin{aligned} &\underset{\mu}{\text{minimize}}\ \|\mu\|_1 \\ &\qquad \text{s. t. } \left\|\mu - \widehat{\mu}_n\right\|_\infty \leq \lambda_n, \end{aligned}$$

Theorem

Suppose that μ^* has at most s **non-zero elements**. *Also suppose that we set $\lambda_n \geq \left\|\mu^* - \widehat{\mu}_n\right\|_\infty$. We then have:*

$$\begin{aligned} \|\widehat{\mu} - \mu^*\|_\infty &\leq\ 2\lambda_n, \\ \|\widehat{\mu} - \mu^*\|_2 &\leq\ 4\sqrt{s}\lambda_n, \\ \|\widehat{\mu} - \mu^*\|_1 &\leq\ 8\, s\lambda_n. \end{aligned}$$

Example: Estimating Covariance

- Special case: Estimating covariance matrix:

$$\Sigma^* = \mathbb{E}[(X - \mathbb{E}(X))(X - \mathbb{E}(X))^\top]$$

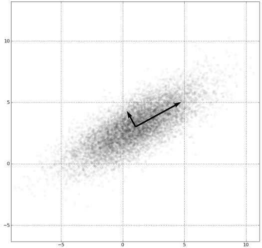


Figure : Principal component analysis, source: Wikipedia

Special Case: Sparse Covariance Estimation

- Our estimator for covariance estimation:

$$\begin{aligned} &\underset{\Sigma}{\text{minimize}} \ \|\Sigma\|_1 \\ &\qquad \text{s. t. } \|S - \Sigma\|_\infty \leq \lambda_n \end{aligned} \tag{1}$$

where $S = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)^\top$, and $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Special Case: Sparse Covariance Estimation

- Decomposable into **element-wise problems**

$$\begin{aligned} \underset{\Sigma_{st}}{\text{minimize}} \; & \|\Sigma_{st}\|_1 \\ \text{s. t.} \; & \|S_{st} - \Sigma_{st}\|_\infty \leq \lambda_n \end{aligned}$$

- The optimal solution $\widehat{\Sigma}$ of (1) is simply $\mathcal{S}_{\lambda_n}(S)$ where $[\mathcal{S}_\lambda(u)]_i = \text{sign}(u_i) \max(|u_i| - \lambda, 0)$
    - Covariance estimation by **element-wise soft-thresholding**: Rothman et al. (2009); Bickel and Levina (2008) analyzed it is **consistent in terms of operator norm**.

Statistical Guarantees

- Our estimator for covariance estimation:

$$\begin{aligned} &\text{minimize } \|\Sigma\|_1 \\ &\qquad \text{s. t. } \|S - \Sigma\|_\infty \leq \lambda_n \end{aligned}$$

Theorem

Suppose that Σ^* of Gaussian has s **non-zero elements** *at most. Also suppose that $\lambda_n = c_1\sqrt{\frac{\log p}{n}}$. Then, with high probability,*

$$\|\widehat{\Sigma} - \Sigma^*\|_\infty \leq \; 2c_1\sqrt{\frac{\log p}{n}}$$

$$\|\widehat{\Sigma} - \Sigma^*\|_F \leq \; 4c_1\sqrt{\frac{s\log p}{n}} \quad \textit{cf. } \textbf{Tighter than} \textit{ previous result: } O\left(\sqrt{\frac{ps\log p}{n}}\right)$$

$$\|\widehat{\Sigma} - \Sigma^*\|_1 \leq \; 8c_1\, s\sqrt{\frac{\log p}{n}}$$

Extension to Superposition Structures

- $\mu^* = \sum_{\alpha \in I} \mu^*_\alpha$, where μ^*_α is a "clean" structured parameter.
- Ex: Robust PCA where Σ^* is the sum of low rank Θ^* and sparse Γ^*

"Elem-Super-Moment" estimators:

$$\begin{aligned} \underset{\mu_1,\mu_2,\ldots,\mu_{|I|}}{\text{minimize}} \; & \sum_{\alpha \in I} \lambda_\alpha \mathcal{R}_\alpha(\mu_\alpha) \\ \text{s.t. } & \mathcal{R}^*_\alpha\Big(\widehat{\mu}_n - \sum_{\alpha \in I} \mu_\alpha\Big) \le \lambda_\alpha \quad \textit{for } \forall \alpha \in I. \end{aligned}$$

Statistical Guarantees for General Structures

- Elem-Super-Moment estimators:

$$\underset{\mu_1,\mu_2,\ldots,\mu_{|I|}}{\text{minimize}} \sum_{\alpha\in I} \lambda_\alpha \mathcal{R}_\alpha(\mu_\alpha)$$

$$\text{s. t. } \mathcal{R}^*_\alpha\Big(\widehat{\mu}_n - \sum_{\alpha\in I}\mu_\alpha\Big) \le \lambda_\alpha \quad \text{for } \forall\alpha\in I.$$

Theorem

Suppose that $\mu^* = \sum_{\alpha\in I}\mu^*_\alpha$, where each μ^*_α lies in a low dimensional subspace $\mathcal{M}_\alpha$, and that each $\mathcal{R}_\alpha(\cdot)$ is decomposable w.r.t. corresp. $\mathcal{M}_\alpha$. Also suppose that we set $\lambda_\alpha \ge \mathcal{R}^*_\alpha(\mu^* - \widehat{\mu})$. We then have:

$$\mathcal{R}^*_\alpha(\widehat{\mu} - \mu^*) \le \; 2\lambda_\alpha \;,$$

$$\mathcal{R}_\alpha(\widehat{\mu}_\alpha - \mu^*_\alpha) \le \; \frac{16|I|}{\lambda_\alpha}\Big(\max_{\alpha\in I}\lambda_\alpha\Psi(\mathcal{M}_\alpha)\Big)^2 \;,$$

$$\|\widehat{\mu} - \mu^*\|_F \le \; 4\sqrt{2|I|}\max_{\alpha\in I}\lambda_\alpha\Psi(\mathcal{M}_\alpha) \;.$$

Experiments - Simulations

- $\Sigma^* = \Sigma_1^* + \Sigma_2^*$
  - where $\Sigma_1^* = 0.5(1_p 1_p^T)$ and $\Sigma_2^* = I_{p/5} \otimes (0.2(1_5 1_5^T) + 0.2 I_5)$

	Method	Spectral	Frobenius	Nuclear	Matrix 1-norm
	Elem-Super-Moment	**7.10 (0.15)**	**8.56(0.18)**	**35.87 (0.43)**	**11.65 (0.12)**
n=100,p=200	Thresholding	8.30 (0.17)	10.43 (0.11)	45.84 (0.39)	19.85 (0.21)
	Well-conditioned	12.22 (0.12)	13.19 (0.17)	48.11 (0.45)	23.89(0.18)
	Elem-Super-Moment	**25.63 (0.54)**	**26.67 (0.49)**	**198.76 (1.31)**	**50.77 (0.72)**
n=100,p=400	Thresholding	33.55 (0.49)	41.91(0.60)	331.41 (2.05)	67.64 (0.73)
	Well-conditioned	35.71 (0.50)	34.83 (0.46)	207.97(2.27)	93.60 (0.91)

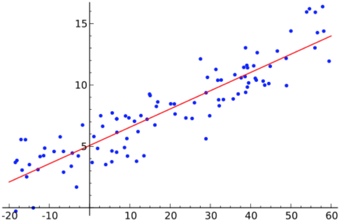
1. Elementary Estimators for General Moment Parameters
2. Elementary Estimators for Linear Models
3. Elementary Estimators for Gaussian Graphical Models

Background - Linear Regression

- Consider the **linear regression model**:

$$y_i = x_i^\top \theta^* + w_i, \quad i = 1, \ldots, n,$$

  - $\theta^* \in \mathbb{R}^p$: fixed unknown regression parameter of interest
  - $y_i \in \mathbb{R}$: real-valued response
  - $x_i \in \mathbb{R}^p$: known observation vector
  - $w_i \in \mathcal{N}(0, \sigma^2)$: independent zero-mean Gaussian noise
  - Collate n independent observations: $y = X\theta^* + w$

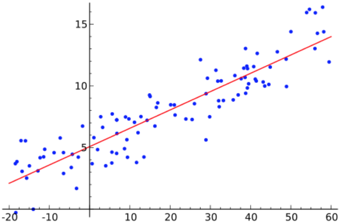


Source: Wikipedia

Background - Linear Regression

- Consider the **linear regression model**:

$$y_i = x_i^\top \theta^* + w_i, \quad i = 1, \ldots, n$$



Source: Wikipedia

- **Used extensively** in practical applications.
  - **Finance**: Modeling Investment risk, Spending, Demand, etc. (responses) given market conditions (features)
  - **Epidemiology**: Linking Tobacco Smoking (feature) to Mortality (response)

Classical Closed-Form Estimators - OLS

- When $p < n$ (and $X^\top X$ is full-rank),
  - Ordinary least squares (OLS) estimator: $(X^\top X)^{-1} X^\top y$
- When $p > n$, $X^\top X$ cannot be full rank
  - The OLS estimator is no longer well-defined.

Classical Closed-Form Estimators - Ridge

- Ridge regularized least squares estimators:

$$\widehat{\theta} = \arg\min_{\theta} \left\{ \|y - X\theta\|_2^2 + \epsilon \|\theta\|_2^2 \right\}.$$

where $\widehat{\theta} = (X^\top X + \epsilon I)^{-1} X^\top y$.

Classical Closed-Form Estimators - Ridge

- Ridge regularized least squares estimators:

$$\widehat{\theta} = \arg\min_{\theta} \left\{ \|y - X\theta\|_2^2 + \epsilon\|\theta\|_2^2 \right\}.$$

where $\widehat{\theta} = (X^\top X + \epsilon I)^{-1} X^\top y$.

Ridge estimators are **not consistent** in high-dimensional sampling regimes!

Variants of Ridge and OLS Closed-Form Estimators

We derived **variants of ridge and OLS closed-form estimators** for **general structurally constrained** linear regression models

The Elem-OLS Estimator

Recall ordinary least squares: $(X^\top X)^{-1} X^\top y$.

The Elem-OLS Estimator

Recall ordinary least squares: $(X^\top X)^{-1} X^\top y$.

For any matrix A, we define element-wise operator T_ν:

$$\left[T_\nu(A)\right]_{ij} = \begin{cases} A_{ii} + \nu & \text{if } i = j \\ \text{sign}(A_{ij})(|A_{ij}| - \nu) & \text{otherwise, } i \neq j \end{cases}$$

$\Rightarrow$ Instead of $X^\top X$, apply T_ν to obtain $T_\nu\left(\frac{X^\top X}{n}\right)$

The Elem-OLS Estimator

Recall ordinary least squares: $(X^\top X)^{-1} X^\top y$.

For any matrix A, we define element-wise operator T_ν:

$$\left[T_\nu(A)\right]_{ij} = \begin{cases} A_{ii} + \nu & \text{if } i = j \\ \text{sign}(A_{ij})(|A_{ij}| - \nu) & \text{otherwise, } i \neq j \end{cases}$$

$\Rightarrow$ Instead of $X^\top X$, apply T_ν to obtain $T_\nu\left(\frac{X^\top X}{n}\right)$

- Each row of X is i.i.d. sampled from $N(0, \Sigma)$
- The design matrix X is column normalized
- The covariance Σ is strictly diagonally dominant

Proposition: *For any* $\nu \geq 8(\max_i \Sigma_{ii})\sqrt{\frac{10\tau \log p'}{n}}$, *the matrix* $T_\nu\left(\frac{X^\top X}{n}\right)$ **is invertible** *with probability at least* $1 - 4/p'^{\tau-2}$ *for* $p' := \max\{n, p\}$ *and any constant* $\tau > 2$.

The Elem-OLS Estimator for General Structure

- *Our Elem-OLS estimator for general structurally constrained linear models:*

$$\begin{aligned} &\underset{\theta}{\text{minimize}}\ \mathcal{R}(\theta) \\ &\text{s. t. } \mathcal{R}^*\left(\theta - \left[T_\nu\Big(\frac{X^\top X}{n}\Big)\right]^{-1}\frac{X^\top y}{n}\right) \le \lambda_n. \end{aligned} \tag{2}$$

The Elem-OLS Estimator - Sparsity Case

- Our Elem-OLS estimator for sparsity case:

$$\widehat{\theta} = S_{\lambda_n}\left(\left[T_\nu\left(\frac{X^\top X}{n}\right)\right]^{-1}\frac{X^\top y}{n}\right)$$

where $[S_\lambda(u)]_i = \mathrm{sign}(u_i)\max(|u_i| - \lambda, 0)$ is the soft-thresholding

Statistical Guarantees of Elem-OLS Estimator

- Our Elem-OLS estimator for general structurally constrained linear models:

$$\begin{aligned} &\underset{\theta}{\text{minimize}}\ \mathcal{R}(\theta) \\ &\text{s. t. } \mathcal{R}^*\left(\theta - \left[T_\nu\left(\frac{X^\top X}{n}\right)\right]^{-1}\frac{X^\top y}{n}\right) \leq \lambda_n. \end{aligned}$$

Theorem

Suppose that the true parameter θ^* lies in some low dimensional space $\mathcal{M}$, and that $\mathcal{R}(\cdot)$ is decomposable w.r.t. $\mathcal{M}$. Denote $\Psi(\mathcal{M}) := \sup_{u \in \mathcal{M}\setminus\{0\}} \left(\mathcal{R}(u)/\|u\|\right)$. Suppose also that we set $\lambda_n \geq \mathcal{R}^*\left(\theta^* - (X^\top X + \epsilon I)^{-1}X^\top y\right)$. We then have:

$$\begin{aligned} &\mathcal{R}^*\left(\widehat{\theta} - \theta^*\right) \leq 2\lambda_n\ , \\ &\|\widehat{\theta} - \theta^*\|_2 \leq 4\Psi(\mathcal{M})\lambda_n\ , \\ &\mathcal{R}\left(\widehat{\theta} - \theta^*\right) \leq 8[\Psi(\mathcal{M})]^2\lambda_n\ . \end{aligned}$$

Statistical Guarantees of Elem-OLS Estimator - Sparsity

- θ^* is sparse with k non-zero entries

Corollary

Suppose $\nu := 8(\max_i \Sigma_{ii})\sqrt{\frac{10\tau \log p}{n}}$ *and* $\lambda_n := \frac{1}{\delta_{\min}}\left(2\sigma\sqrt{\frac{\log p'}{n}} + c\sqrt{\frac{\log p'}{n}}\|\theta^*\|_1\right)$. *Then, any optimal solution of Elem-OLS estimator satisfies*

$$\|\widehat{\theta} - \theta^*\|_\infty \leq \frac{2}{\delta_{\min}}\left(2\sigma\sqrt{\frac{\log p}{n}} + c\sqrt{\frac{\log p}{n}}\|\theta^*\|_1\right),$$

$$\|\widehat{\theta} - \theta^*\|_2 \leq \frac{4}{\delta_{\min}}\left(2\sigma\sqrt{\frac{k\log p}{n}} + c\sqrt{\frac{k\log p}{n}}\|\theta^*\|_1\right),$$

$$\|\widehat{\theta} - \theta^*\|_1 \leq \frac{8}{\delta_{\min}}\left(2\sigma k\sqrt{\frac{\log p}{n}} + ck\sqrt{\frac{\log p}{n}}\|\theta^*\|_1\right)$$

with probability at least $1 - c_1\exp(-c_2 p)$.

Cf.) Similar to rates of standard LASSO: $\|\widehat{\theta}_{LASSO} - \theta^*\|_2 \leq O\left(\sqrt{\frac{k\log p}{n}}\right)$

Experiments - Simulated Data

- $y_i = x_i^\top \theta^* + w_i, \quad i = 1, \ldots, n$:
  - $X \sim N(0, \Sigma)$ where $\Sigma_{i,j} = 0.5^{|i-j|}$
  - $w \sim N(0, 1)$.
  - $k := \|\theta\|_0 = 10$,
  - Non-zero element of θ^* chosen independently and uniformly in $(1, 3)$

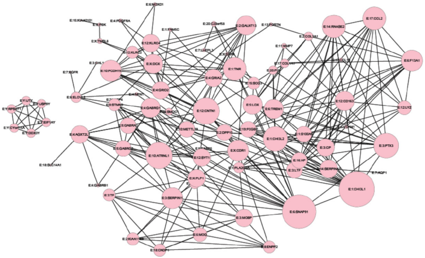
Table : Average performance measure and standard deviation in parenthesis for ℓ_1-penalized comparison methods on simulated data for sparse linear models.

	Method	TP	FP	ℓ_2	ℓ_∞
n=1000,p=1000	Elem-OLS	**100.00 (0.00)**	**2.05 (1.15)**	**0.551 (0.071)**	**0.255 (0.041)**
	Elem-Ridge	**100.00 (0.00)**	2.44 (2.12)	0.741 (0.411)	0.435 (0.064)
	LASSO	**100.00 (0.00)**	9.84 (2.45)	0.563 (0.067)	0.270 (0.039)
	Thr-LASSO	**100.00 (0.00)**	8.33 (1.14)	0.560 (0.066)	0.274 (0.071)
	OMP	98.24 (0.64)	3.20 (1.38)	0.559 (0.113)	0.282 (0.055)
n=1000,p=2000	Elem-OLS	**100.00 (0.00)**	**2.22 (2.02)**	**0.656 (0.111)**	**0.314(0.071)**
	Elem-Ridge	**100.00 (0.00)**	11.94 (4.48)	3.8834 (0.411)	1.678 (0.349)
	LASSO	**100.00 (0.00)**	18.88 (6.93)	0.657(0.110)	0.316(0.075)
	Thr-LASSO	99.59(0.36)	14.35(2.66)	**0.656 (0.099)**	0.315(0.052)
	OMP	96.36(1.00)	10.25 (4.24)	0.735(0.222)	0.536(0.136)

Background - Gaussian Graphical Models

- Consider $X = (X_1, \ldots, X_p)$ with Gaussian distribution $\mathcal{N}(X|\mu, \Sigma)$:

$$\mathbb{P}(X|\theta, \Theta) = \exp\Big(-\frac{1}{2}\langle\langle \Theta, XX^\top \rangle\rangle + \langle \theta, X \rangle - A(\Theta, \theta) \Big)$$



- Θ^{-1} corresponds to the set of edges in Gaussian Markov random fields

- ℓ_1 regularized maximum likelihood estimator:

$$\underset{\Theta \succ 0}{\text{minimize}}\ \langle\langle S, \Theta \rangle\rangle - \log\det\Theta + \lambda_n \|\Theta\|_{1,\text{off}}$$

The Elementary Estimator for Gaussian Graphical Models

- *Our Elem-GM estimator for general structurally constrained Gaussian graphical models:*

$$\begin{aligned} &\underset{\Theta}{\text{minimize}} \ \mathcal{R}(\Theta) \\ &\text{s. t. } \mathcal{R}^*\left(\Theta - [T_\nu(S)]^{-1}\right) \leq \lambda_n \end{aligned}$$

The Elementary Estimator for Gaussian Graphical Models - Sparsity Case

- Our Elem-GM estimator for sparsity case:

$$\widehat{\Theta} = S_{\lambda_n}\Big([T_\nu(S)]^{-1}\Big)$$

where $[S_\lambda(u)]_i = \mathrm{sign}(u_i)\max(|u_i| - \lambda, 0)$ is the soft-thresholding

Statistical Guarantees of Elem-GM Estimator

- Our Elem-GM estimator for general structurally constrained Gaussian graphical models:

$$\begin{aligned} &\underset{\Theta}{\text{minimize}}\ \mathcal{R}(\Theta) \\ &\qquad \text{s.t. } \mathcal{R}^*\Big(\Theta - [T_\nu(S)]^{-1}\Big) \le \lambda_n \end{aligned}$$

Theorem

Suppose that the true parameter Θ^* lies in some low dimensional space $\mathcal{M}$, and that $\mathcal{R}(\cdot)$ is decomposable w.r.t. $\mathcal{M}$. Denote $\Psi(\mathcal{M}) := \sup_{u \in \mathcal{M}\setminus\{0\}} \big(\mathcal{R}(u)/\|u\|\big)$. Suppose also that we set $\lambda_n \ge \mathcal{R}^*\big(\Theta^* - [T_\nu(S)]^{-1}\big)$. We then have:

$$\begin{aligned} &\mathcal{R}^*\big(\widehat{\Theta} - \Theta^*\big) \le 2\lambda_n\ , \\ &\|\widehat{\Theta} - \Theta^*\|_2 \le 4\Psi(\mathcal{M})\lambda_n\ , \\ &\mathcal{R}\big(\widehat{\Theta} - \Theta^*\big) \le 8[\Psi(\mathcal{M})]^2\lambda_n\ . \end{aligned}$$

Statistical Guarantees of Elem-GM Estimator - Sparsity

- Θ^* is sparse with k non-zero entries

Corollary

Suppose $\nu := 8(\max_i \Sigma_{ii})\sqrt{\frac{10\tau \log p}{n}}$ *and* $\lambda_n := O\left(\sqrt{\frac{\log p}{n}}\right)$. *Then, any optimal solution of elementary Gaussian estimator satisfies*

$$\|\widehat{\Theta} - \Theta^*\|_\infty \le O\left(\sqrt{\frac{\log p}{n}}\right), \quad \|\widehat{\Theta} - \Theta^*\|_F \le O\left(\sqrt{k\frac{\log p}{n}}\right)$$

$$\|\widehat{\Theta} - \Theta^*\|_1 \le O\left(k\sqrt{\frac{\log p}{n}}\right)$$

with probability at least $1 - c_1 \exp(-c_2 p)$.

Cf.) Asymp. equivalent to rates of standard ℓ_1 regularized MLE:
$\|\widehat{\Theta}_{\ell_1 MLE} - \Theta^*\|_F \le O\left(\sqrt{\frac{k \log p}{n}}\right)$

Experiments

- Approximately $10p$ non-zero entries in Θ^* (random structure)
- $\lambda_n := K\sqrt{\frac{\log p}{n}}$
- $(n, p) = (800, 1600)$

Table : Performance comparisons our closed-form estimators against state of the art QUIC algorithm (Hsieh et al., 2011) solving ℓ_1 MLE

	K	Time(sec)	ℓ_F (off)	ℓ_∞ (off)	FPR	TPR
Elem-GM	0.01	< 1	6.36	0.1616	0.48	0.99
	0.02	< 1	6.19	0.1880	0.24	0.99
	0.05	< 1	5.91	0.1655	0.06	0.99
	0.1	< 1	6	0.1703	0.01	0.97
QUIC	0.5	2575.5	12.74	0.11	0.52	1.00
	1	1009	7.30	0.13	0.35	0.99
	2	272.1	6.33	0.18	0.16	0.99
	3	78.1	6.97	0.21	0.07	0.94
	4	28.7	7.68	0.23	0.02	0.86

Experiments

- Approximately $10p$ non-zero entries in Θ^* (random structure)
- $\lambda_n := K\sqrt{\frac{\log p}{n}}$
- $(n, p) = (800, 1600)$

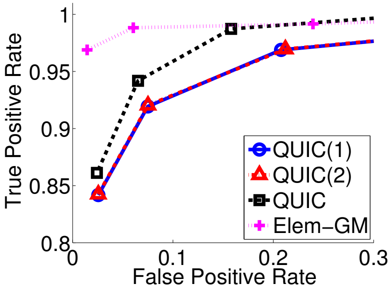


Figure : Receiver operator curves for support set recovery task.

Experiments

- Approximately $10p$ non-zero entries in Θ^* (random structure)
- $\lambda_n := K\sqrt{\frac{\log p}{n}}$
- $(n, p) = (5000, 10000)$

Table : Performance comparisons our closed-form estimators against state of the art QUIC algorithm (Hsieh et al., 2011) solving ℓ_1 MLE

	K	Time(sec)	ℓ_F (off)	ℓ_∞ (off)	FPR	TPR
Elem-GM	0.05	47.3	11.73	0.1501	0.13	1.00
	0.1	46.3	8.91	0.1479	0.03	1.00
	0.5	45.8	5.66	0.1308	0.0	1.00
	1	46.2	8.63	0.1111	0.0	0.99
QUIC	2	*	*	*	*	*
	2.5	*	*	*	*	*
	3	4.8×10^4	9.85	0.1083	0.06	1.00
	3.5	2.7×10^4	10.51	0.1111	0.04	0.99

Experiments

- Approximately $10p$ non-zero entries in Θ^* (random structure)
- $\lambda_n := K\sqrt{\frac{\log p}{n}}$
- $(n, p) = (5000, 10000)$

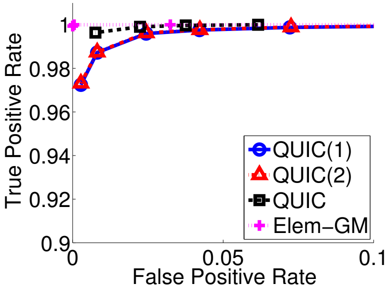


Figure : Receiver operator curves for support set recovery task.

Conclusion

- We propose a case of elementary convex estimators for estimating general statistical models
  - Available in **closed-form** in many cases
  - Provide a unified statistical analysis for general structure
- Future work
  - Develop this closed form estimation framework for more general high-dimensional problems
  - Extend the framework to non-convex penalty functions

Thank you!

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

S. van de Geer and P. Buhlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37(1):246–270, 2009.

E. Candes and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 2006.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436–1462, 2006.

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Information Theory*, 55:2183–2202, May 2009.

P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2006.

J. A. Tropp, A. C. Gilbert, and M. J. Strauss. Algorithms for simultaneous sparse approximation. *Signal Processing*, 86:572–602, April 2006. Special issue on "Sparse approximations in signal and image processing".

P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*, 37(6A):3468–3497, 2009.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B*, 1(68):49, 2006.

L. Jacob, G. Obozinski, and J. P. Vert. Group Lasso with Overlap and Graph Lasso. In *International Conference on Machine Learning (ICML)*, pages 433–440, 2009.

K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer. Taking advantage of sparsity in multi-task learning. Technical Report arXiv:0903.1468, ETH Zurich, March 2009.

R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde. Model-based compressive sensing. Technical report, Rice University, 2008. Available at arxiv:0808.3572.

B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, Vol 52(3): 471–501, 2010.

F. Bach. Consistency of trace norm minimization. *Journal of Machine Learning Research*, 9:1019–1048, June 2008.

S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, 2007.

O. Bannerjee, , L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Jour. Mach. Lear. Res.*, 9:485–516, March 2008.

P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, UK, 2004.

T. Cai, W. Liu, and X. Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. In *48th Annual Allerton Conference on Communication, Control and Computing*, 2010.

A. J. Rothman, E. Levina, and J. Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association (Theory and Methods)*, 104:177–186, 2009.

P. J. Bickel and E. Levina. Covariance regularization by thresholding. *Annals of Statistics*, 36(6):2577–2604, 2008.

C.-J. Hsieh, M. Sustik, I. Dhillon, and P. Ravikumar. Sparse inverse covariance matrix estimation using quadratic approximation. In *Neur. Info. Proc. Sys. (NIPS)*, 24, 2011.