Mixed Regression: Minimax Optimal Rates

Constantine Caramanis

The University of Texas at Austin
constantine@utexas.edu

Joint work with Yudong Chen and Xinyang Yi

June 20, 2014

a simple problem

$$y_i = \langle \mathbf{x}_i, \beta^* \rangle + e_i, \quad i = 1, \ldots, n,$$

a simple problem

$$y_i = \langle \mathbf{x}_i, \beta^* \rangle + e_i, \quad i = 1, \ldots, n,$$

- $\beta^* \in \mathbb{R}^p$

a simple problem

$$y_i = \langle \mathbf{x}_i, \beta^* \rangle + e_i, \quad i = 1, \ldots, n,$$

- $\beta^* \in \mathbb{R}^p$
- statistics: $n \geq p$, error $\sim \sigma\sqrt{p/n}$.

a simple problem

$$y_i = \langle \mathbf{x}_i, \beta^* \rangle + e_i, \quad i = 1, \ldots, n,$$

- $\beta^* \in \mathbb{R}^p$
- statistics: $n \geq p$, error $\sim \sigma\sqrt{p/n}$.
- computation: $\min : \|\mathbf{y} - X\beta\|_2$.

a simple problem

- sparse version: $\beta^* \in \mathbb{R}^p$, sparse.
- low-rank version: $\beta^* \in \mathbb{R}^{p \times p}$, low-rank.
- low-rank plus sparse: $\beta^* \in \mathbb{R}^{p \times p}$, $\beta^* = L + S$.
- low-rank plus sparse plus column sparse: $\beta^* \in \mathbb{R}^{p \times p}$, $\beta^* = L + S + C$.
- etc.

a simple problem?

- sparse version: $\beta^* \in \mathbb{R}^p$, sparse.
- low-rank version: $\beta^* \in \mathbb{R}^{p \times p}$, low-rank.
- low-rank plus sparse: $\beta^* \in \mathbb{R}^{p \times p}$, $\beta^* = L + S$.
- low-rank plus sparse plus column sparse: $\beta^* \in \mathbb{R}^{p \times p}$, $\beta^* = L + S + C$.
- etc.

- mixture: $\beta^* = \beta_1^*$ or $\beta^* = \beta_2^*$?

a simple problem?

- sparse version: $\beta^* \in \mathbb{R}^p$, sparse.
- low-rank version: $\beta^* \in \mathbb{R}^{p \times p}$, low-rank.
- low-rank plus sparse: $\beta^* \in \mathbb{R}^{p \times p}$, $\beta^* = L + S$.
- low-rank plus sparse plus column sparse: $\beta^* \in \mathbb{R}^{p \times p}$, $\beta^* = L + S + C$.
- etc.

- mixture: $\beta^* = \beta_1^*$ or $\beta^* = \beta_2^*$?

a mixture problem

$$y_i = z_i \cdot \langle \mathbf{x}_i, \beta_1^* \rangle + (1 - z_i) \cdot \langle \mathbf{x}_i, \beta_2^* \rangle + e_i, \quad i = 1, \ldots, n,$$
$$\beta_1^*, \beta_2^* \in \mathbb{R}^p, \quad z_i \in \{0, 1\}.$$

mixture problems: applications

- why: superpositions of simple processes.

mixture problems: applications

- why: superpositions of simple processes.
- this problem: mixed populations, etc.
- other problems: subspace clustering, topic modeling

computation and statistics

- if we don't care about computational complexity, (often) it's easy.
- if we don't care about sample complexity, (sometimes) it's easy.

computation and statistics

- if we don't care about computational complexity, (often) it's easy.
- if we don't care about sample complexity, (sometimes) it's easy.
- if we care about both...

hardness, past approaches

- exact solution seems to be hard (SUBSET-SUM).

hardness, past approaches

- exact solution seems to be hard (SUBSET-SUM).
- classical: expectation maximization – guess labels, find (β_1^*, β_2^*), repeat.

hardness, past approaches

- exact solution seems to be hard (Subset-Sum).
- classical: expectation maximization – guess labels, find (β_1^*, β_2^*), repeat.
- tensor approach.

expectation maximization (EM)

- easy computation.
- noisy case: no global convergence guarantees.
- noiseless case – mixed linear equations: guaranteed convergence w/ optimal rates (joint work with Xinyang Yi and Sujay Sanghavi).

tensor-based approach

$$y_i = z_i \cdot \langle \mathbf{x}_i, \beta_1^* \rangle + (1 - z_i) \cdot \langle \mathbf{x}_i, \beta_2^* \rangle + e_i, \quad i = 1, \ldots, n.$$

- regress y_i against $\mathbf{x}_i$: get $\lambda \beta_1^* + (1 - \lambda) \beta_2^*$.
- regress y_i^2 against $\mathbf{x}_i^{\otimes 2}$: get $\lambda (\beta_1^*)^{\otimes 2} + (1 - \lambda)(\beta_2^*)^{\otimes 2}$.
- regress y_i^3 against $\mathbf{x}_i^{\otimes 3}$: get $\lambda (\beta_1^*)^{\otimes 3} + (1 - \lambda)(\beta_2^*)^{\otimes 3}$.
- (Chaganty & Liang, 2013).

tensor-based approach

$$y_i = z_i \cdot \langle \mathbf{x}_i, \beta_1^* \rangle + (1 - z_i) \cdot \langle \mathbf{x}_i, \beta_2^* \rangle + e_i, \quad i = 1, \ldots, n.$$

- regress y_i against $\mathbf{x}_i$: get $\lambda \beta_1^* + (1-\lambda)\beta_2^*$.
- regress y_i^2 against $\mathbf{x}_i^{\otimes 2}$: get $\lambda (\beta_1^*)^{\otimes 2} + (1-\lambda)(\beta_2^*)^{\otimes 2}$.
- regress y_i^3 against $\mathbf{x}_i^{\otimes 3}$: get $\lambda (\beta_1^*)^{\otimes 3} + (1-\lambda)(\beta_2^*)^{\otimes 3}$.
- (Chaganty & Liang, 2013).

- advantage: can handle k-mixtures: $\{\beta_1^*, \ldots, \beta_k^*\}$.
- issues: sample complexity $O(p^6)$

our approach

optimization & concentration inequality machinery for matrices.

this talk: optimal rates

$$y_i = z_i \cdot \langle \mathbf{x}_i, \beta_1^* \rangle + (1 - z_i) \cdot \langle \mathbf{x}_i, \beta_2^* \rangle + e_i, \quad i = 1, \ldots, n.$$

a convex formulation such that: if $\mathbf{x}_i$ independent, sub-Gaussian,

- minimax-optimal rates when $\{e_i\}$ arbitrary norm-bounded.
- minimax-optimal rates when $\{e_i\}$ sub-Gaussian, and balanced mixture.

a convex formulation

$$\begin{aligned} K^* &= \frac{1}{2}(\beta_1^* \beta_2^{*\top} + \beta_2^* \beta_1^{*\top}) \\ \mathbf{g}^* &= \frac{1}{2}(\beta_1^* + \beta_2^*). \end{aligned}$$

given $(K^*, \mathbf{g}^*)$,

$$J^* = \mathbf{g}^* \mathbf{g}^{*\top} - K^* = \frac{1}{4}(\beta_1^* - \beta_2^*)(\beta_1^* - \beta_2^*)^\top.$$

arbitrary noise: a convex formulation

$$\begin{aligned} \min_{K,\mathbf{g}} \quad & \|K\|_* \\ \text{subject to} \quad & \sum_{i=1}^{n} \left| -\left\langle \mathbf{x}_i \mathbf{x}_i^\top, K \right\rangle + 2y_i \left\langle \mathbf{x}_i, \mathbf{g} \right\rangle - y_i^2 \right| \leq \eta. \end{aligned}$$

arbitrary noise: a convex formulation

$$
\begin{aligned}
\min_{K,\mathbf{g}} \quad & \|K\|_* \\
\text{subject to} \quad & \sum_{i=1}^{n} \left| -\left\langle \mathbf{x}_i \mathbf{x}_i^\top, K \right\rangle + 2y_i \left\langle \mathbf{x}_i, \mathbf{g} \right\rangle - y_i^2 \right| \leq \eta.
\end{aligned}
$$

- $\hat{J} = \hat{\mathbf{g}}\hat{\mathbf{g}}^\top - \hat{K}$
- $\hat{\beta}_1, \hat{\beta}_2 = \hat{\mathbf{g}} \pm \sqrt{\hat{\lambda}}\hat{\mathbf{v}}$.

stochastic noise: a convex formulation

$$\min_{K,\mathbf{g}} : \sum_{i=1}^{n} \left(-\left\langle \mathbf{x}_i \mathbf{x}_i^\top, K \right\rangle + 2y_i \left\langle \mathbf{x}_i, \mathbf{g} \right\rangle - y_i^2 + \sigma^2 \right)^2 + \mu \left\| K \right\|_* .$$

- $\hat{J} = \hat{\mathbf{g}}\hat{\mathbf{g}}^\top - \hat{K}$
- $\hat{\beta}_1, \hat{\beta}_2 = \hat{\mathbf{g}} \pm \sqrt{\hat{\lambda}}\hat{\mathbf{v}}$.

outline from here

- upper bounds: sample complexity and rates of convergence
- lower bounds (algorithm free) on rates of convergence
- some proof ideas

arbitrary noise: upper bounds

theorem. suppose $n_1, n_2 \geq c_0 \cdot p$ and $\|\beta_1 - \beta_2\|$ and $\|\beta_i\|$ are bounded below. then w.h.p.,

$$\begin{aligned} \|\hat{K} - K^*\|_F &\leq c_1 \frac{\|\mathbf{e}\|_2}{\sqrt{n}} \\ \|\hat{\mathbf{g}} - \mathbf{g}^*\|_2 &\leq c_2 \frac{\|\mathbf{e}\|_2}{\sqrt{n}} \\ \|\hat{\beta}_i - \hat{\beta}_i^*\|_2 &\leq c_3 \frac{\|\mathbf{e}\|_2}{\sqrt{n}}, \quad i = 1, 2. \end{aligned}$$

corollary. in the noiseless case, we have exact recovery with $O(p)$ samples.

arbitrary noise: lower bounds

for any estimator $\hat{\theta} = (\hat{\beta}_1, \hat{\beta}_2)$ that is a measurable function of the data, there exists (β_1^*, β_2^*) and noise $\mathbf{e}$, with expected loss bounded below.

arbitrary noise: lower bounds

theorem. let $\Theta(\gamma)$ be the set of (β_1, β_2) with γ-bounded norm and separation. if $n \geq c_1 \cdot p$, then for any labeling, w.h.p.,

$$\inf_{\hat{\theta}} \sup_{\theta^* \in \Theta(\gamma)} \sup_{\|\mathbf{e}\| \leq \varepsilon} \|\hat{\beta}_i - \beta_i^*\| \geq c_2 \frac{\varepsilon}{\sqrt{n}}.$$

therefore: the rate $\|\mathbf{e}\|/\sqrt{n}$ is minimax optimal.

stochastic noise: upper bounds

theorem. suppose $n_1, n_2 \geq c_0 \cdot p \log^8 p$, and are balanced, and $\|\beta_1 - \beta_2\|$ and $\|\beta_i\|$ are bounded below. then w.h.p.,

$$\|\hat{\beta}_i - \beta_i^*\| \leq \underbrace{\sigma \sqrt{\frac{p}{n}} \log^4 n}_{(a)} + \min \left\{ \underbrace{\frac{\sigma^2}{\|\beta_1^*\| + \|\beta_2^*\|} \sqrt{\frac{p}{n}}}_{(b)}, \underbrace{\sigma \left(\frac{p}{n}\right)^{1/4}}_{(c)} \right\} \log^4 n.$$

stochastic noise: upper bounds

theorem. suppose $n_1, n_2 \geq c_0 \cdot p \log^8 p$, and are balanced, and $\|\beta_1 - \beta_2\|$ and $\|\beta_i\|$ are bounded below. then w.h.p.,

$$\|\hat{\beta}_i - \beta_i^*\| \leq \underbrace{\sigma \sqrt{\frac{p}{n}} \log^4 n}_{(a)} + \min \left\{ \underbrace{\frac{\sigma^2}{\|\beta_1^*\| + \|\beta_2^*\|} \sqrt{\frac{p}{n}}}_{(b)}, \underbrace{\sigma \left(\frac{p}{n}\right)^{1/4}}_{(c)} \right\} \log^4 n.$$

- (a): high SNR, (b): medium SNR, (c): low SNR.

stochastic noise: lower bounds

for any estimator $\hat{\theta} = (\hat{\beta}_1, \hat{\beta}_2)$ that is a measurable function of the data, there exists (β_1^*, β_2^*) with expected loss bounded below.

stochastic noise: lower bounds

theorem. let $\Theta(\gamma)$ be the set of (β_1, β_2) with γ-bounded norm and separation. if $n \geq c_1 \cdot p$, $\mathbf{e} \sim N(0, \sigma^2 I)$, $z_i \sim Ber(1/2)$, then w.h.p.

(a) $\gamma > \sigma$:

$$\inf_{\hat{\theta}} \sup_{\theta^* \in \Theta(\gamma)} \mathbb{E}_{X,\mathbf{z},\mathbf{e}} \|\hat{\beta}_i - \beta_i^*\| \geq c\sigma \sqrt{\frac{p}{n}}.$$

(b) $\sigma(\frac{p}{n})^{\frac{1}{4}} \leq \gamma \leq \sigma$:

$$\inf_{\hat{\theta}} \sup_{\theta^* \in \Theta(\gamma)} \mathbb{E}_{X,\mathbf{z},\mathbf{e}} \|\hat{\beta}_i - \beta_i^*\| \geq c \frac{\sigma^2}{\gamma} \sqrt{\frac{p}{n}}.$$

(c) $0 < \gamma < \sigma(\frac{p}{n})^{\frac{1}{4}}$:

$$\inf_{\hat{\theta}} \sup_{\theta^* \in \Theta(\gamma)} \mathbb{E}_{X,\mathbf{z},\mathbf{e}} \|\hat{\beta}_i - \beta_i^*\| \geq c\sigma \left(\frac{p}{n}\right)^{\frac{1}{4}}.$$

some key ideas: upper bounds

$$\|\hat{\beta}_1 - \beta_1^*\| + \|\hat{\beta}_2 - \beta_2^*\|$$

convex optimization: show there is enough curvature *in the right directions* near desired solution, so that finite data + noise has bounded effect.

some key ideas: upper bounds

directions away from $(K^*, \mathbf{g}^*)$:

some key ideas: upper bounds

directions away from $(K^*, \mathbf{g}^*)$:

- concentration inequalities show that $(K^*, \mathbf{g}^*)$ feasible w.h.p.

some key ideas: upper bounds

directions away from $(K^*, \mathbf{g}^*)$:

- concentration inequalities show that $(K^*, \mathbf{g}^*)$ feasible w.h.p.
- $(\hat{K}, \hat{\mathbf{g}}) = (K^* + \hat{H}, \mathbf{g}^* + \hat{\mathbf{h}})$: $\|\hat{H}_T^\top\|_* \leq \|\hat{H}_T\|_*$.
- $T = \{Z + Y\}$, where Z (resp. Y) – a matrix with column (row) space $\mathrm{span}\{\beta_1^*, \beta_2^*\}$.

some key ideas: upper bounds

- for $b = 1, 2$ define:

$$
\begin{aligned}
B_{b,j} &= \mathbf{x}_{b,2j}\mathbf{x}_{b,2j}^{\top} - \mathbf{x}_{b,2j-1}\mathbf{x}_{b,2j-1}^{\top}, \\
(\mathscr{B}_b Z)_j &= \frac{1}{n_b/2}\langle B_{b,j}, Z\rangle.
\end{aligned}
$$

some key ideas: upper bounds

- for $b = 1, 2$ define:

$$
\begin{aligned}
B_{b,j} &= \mathbf{x}_{b,2j}\mathbf{x}_{b,2j}^\top - \mathbf{x}_{b,2j-1}\mathbf{x}_{b,2j-1}^\top, \\
(\mathscr{B}_b Z)_j &= \frac{1}{n_b/2}\langle B_{b,j}, Z\rangle.
\end{aligned}
$$

- feasibility of $(K^* + \hat{H}, \mathbf{g}^* + \hat{\mathbf{h}})$ implies

$$
\sum_b n \underbrace{\|\mathscr{B}_b(-\hat{H} + 2\beta_b^* \hat{\mathbf{h}}^\top)\|_1}_{\star} - c \sum_b \sqrt{n}\|\mathbf{e}\|_2 \|\hat{\mathbf{h}}\|_2 \leq 2\eta.
$$

some key ideas: upper bounds

- for $b = 1, 2$ define:

$$\begin{aligned} B_{b,j} &= \mathbf{x}_{b,2j}\mathbf{x}_{b,2j}^\top - \mathbf{x}_{b,2j-1}\mathbf{x}_{b,2j-1}^\top, \\ (\mathscr{B}_b Z)_j &= \frac{1}{n_b/2}\langle B_{b,j}, Z\rangle. \end{aligned}$$

- feasibility of $(K^* + \hat{H}, \mathbf{g}^* + \hat{\mathbf{h}})$ implies

$$\sum_b n \underbrace{\|\mathscr{B}_b(-\hat{H} + 2\beta_b^* \hat{\mathbf{h}}^\top)\|_1}_{\star} - c\sum_b \sqrt{n}\|\mathbf{e}\|_2\|\hat{\mathbf{h}}\|_2 \le 2\eta.$$

- curvature: show $(\star)$ is related to the norms $\|\hat{H}\|_F$ and $\|\hat{\mathbf{h}}\|_2$.

some key ideas: upper bounds

- finally: given $(\hat{K}, \hat{\mathbf{g}})$ close to $(K^*, \mathbf{g}^*)$, need to show $\hat{\beta}_i$ close to β_i^*.

some key ideas: upper bounds

- finally: given $(\hat{K}, \hat{\mathbf{g}})$ close to $(K^*, \mathbf{g}^*)$, need to show $\hat{\beta}_i$ close to β_i^*.
- perturbation bounds: Weyl's inequality, and Davis-Kahan's sine theorem.

lower bounds: framework

information-theoretic setup:

- nature sends $\theta^* = (\beta_1^*, \beta_2^*) \in \Theta$.
- we receive $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$.
- capacity of channel: sample size vs resolution (error).

lower bounds: Fano's inequality

- let $\Theta(\gamma)$ be pairs (β_1, β_2) with norm and separation lower bounded by γ.

lower bounds: Fano's inequality

- let $\Theta(\gamma)$ be pairs (β_1, β_2) with norm and separation lower bounded by γ.
- let $\Theta_{\text{disc}} = \{\theta_1, \ldots, \theta_M\} \subseteq \Theta(\gamma)$ be a δ-packing.

lower bounds: Fano's inequality

- let $\Theta(\gamma)$ be pairs (β_1, β_2) with norm and separation lower bounded by γ.
- let $\Theta_{\text{disc}} = \{\theta_1, \ldots, \theta_M\} \subseteq \Theta(\gamma)$ be a δ-packing.
- $\inf_{\hat{\theta}} \sup_{\theta^* \in \Theta(\gamma)} \mathbb{E}[d(\hat{\theta}, \theta^*)] \geq \inf_{\hat{\theta}} \sup_{\theta^* \in \Theta_{\text{disc}}} \mathbb{E}[d(\hat{\theta}, \theta^*)]$

lower bounds: Fano's inequality

- let $\Theta(\gamma)$ be pairs (β_1, β_2) with norm and separation lower bounded by γ.

- let $\Theta_{\text{disc}} = \{\theta_1, \ldots, \theta_M\} \subseteq \Theta(\gamma)$ be a δ-packing.

- $\inf_{\hat{\theta}} \sup_{\theta^* \in \Theta(\gamma)} \mathbb{E}[d(\hat{\theta}, \theta^*)] \geq \inf_{\hat{\theta}} \sup_{\theta^* \in \Theta_{\text{disc}}} \mathbb{E}[d(\hat{\theta}, \theta^*)]$

- $\inf_{\hat{\theta}} \sup_{\theta^* \in \Theta_{\text{disc}}} \mathbb{E}[d(\hat{\theta}, \theta^*)] \geq \delta \inf_{\tilde{\theta}} \mathbb{P}(\tilde{\theta} \neq \theta^*)$.

lower bounds: Fano's inequality

- (Fano). for any estimator $\hat{X}$ of X, with $X \to Y \to \hat{X}$,

$$\mathbf{P}(\hat{X} \neq X) \geq \frac{H(X|Y) - 1}{\log|\mathscr{X}|}.$$

lower bounds: Fano's inequality

- (Fano). for any estimator $\hat{X}$ of X, with $X \to Y \to \hat{X}$,

$$\mathbf{P}(\hat{X} \neq X) \geq \frac{H(X|Y) - 1}{\log|\mathscr{X}|}.$$

- rearranging and using our notation:

$$\mathbb{P}(\tilde{\theta} \neq \theta^*) \geq 1 - \frac{I((\mathbf{y}, X); \theta^*) + \log 2}{\log M}.$$

lower bounds: Fano's inequality

- (Fano). for any estimator $\hat{X}$ of X, with $X \to Y \to \hat{X}$,

$$\mathbf{P}(\hat{X} \neq X) \geq \frac{H(X|Y) - 1}{\log|\mathscr{X}|}.$$

- rearranging and using our notation:

$$\mathbb{P}(\tilde{\theta} \neq \theta^*) \geq 1 - \frac{I((\mathbf{y}, X); \theta^*) + \log 2}{\log M}.$$

- construction: want M big, and $I((\mathbf{y}, X); \theta^*)$ small.

lower bounds: Fano's inequality

how δ-packing constructions work.

- typically non-constructive
- randomized algorithm guarantees
- here: use a Varshamov-Gilbert bound result:

$$\exists \{\xi_1, \ldots, \xi_{M_0}\} \subset \{0,1\}^p,$$

$M_0 \geq 2^{p/8}$, $\|\xi_i - \xi_j\|_0 \geq p/8$.

lower bounds: Fano's inequality

how computations work:

- $\mathbb{P}_i$ conditional distribution of $(X, \mathbf{y})$ on $\beta^* = \beta_i$.
- then by definition and convexity (Jensen's) of mutual information:

$$\begin{aligned} I(\beta^*; (X, \mathbf{y})) &= \frac{1}{M} \sum_i D(\mathbb{P}_i \| (1/M) \sum_j \mathbb{P}_j) \\ &\leq \frac{1}{M^2} \sum_{i,j} D(\mathbb{P}_i \| \mathbb{P}_j). \end{aligned}$$

lower bounds: Fano's inequality

high SNR vs low SNR

- high SNR: enough to obtain lower bound for regression, $\theta_i = \xi_i$, and $\mathbb{P}_i$ Gaussian. hence:

$$D(\mathbb{P}_i \| \mathbb{P}_j) = \mathbb{E}_X \frac{\|X\beta_i - X\beta_j\|^2}{2\sigma^2}$$

- low SNR: $\theta_i = (\xi_i, -\xi_i)$, and $\mathbb{P}_i$ is mixture of two Gaussians.

some other things we know

stochastic setting for unbalanced labels:

$$
\begin{aligned}
\min: \quad & \sum_{i=1}^{n} \left(-\langle \mathbf{x}_i \mathbf{x}_i^\top, K\rangle + 2y_i\langle \mathbf{x}_i, \mathbf{g}\rangle - y_i^2 + \sigma^2\right)^2 - 4\sigma^2\left(y_i - \langle \mathbf{x}_i, \mathbf{g}\rangle\right)^2 \\
\text{s.t.}: \quad & \|K\|_* \le \|K^*\|_*.
\end{aligned}
$$

- non-convex
- can show gradient descent converges to optimal solution.
- again key is showing curvature in some directions from optimal solution.

some other things we don't know

- relaxing assumptions.
- more than two components in mixture.
- dealing with a "garbage" component.

conclusion

find out more from:

`http://users.ece.utexas.edu/∼cmcaram/`

or e-mail:

`constantine@utexas.edu`