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

\[ y_i = \langle x_i, \beta^* \rangle + e_i, \quad i = 1, \ldots, n, \]
a simple problem

\[ y_i = \langle x_i, \beta^* \rangle + e_i, \quad i = 1, \ldots, n, \]

\[ \beta^* \in \mathbb{R}^p \]
a simple problem

\[ y_i = \langle x_i, \beta^* \rangle + e_i, \quad i = 1, \ldots, n, \]

- \( \beta^* \in \mathbb{R}^p \)
- statistics: \( n \geq p \), error \( \sim \sigma \sqrt{p/n} \).
a simple problem

\[ y_i = \langle x_i, \beta^* \rangle + e_i, \quad i = 1, \ldots, n, \]

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

\[ y_i = z_i \cdot \langle x_i, \beta_1^* \rangle + (1 - z_i) \cdot \langle x_i, \beta_2^* \rangle + e_i, \quad i = 1, \ldots, n, \]
\[ \beta_1^*, \beta_2^* \in \mathbb{R}^p, \quad z_i \in \{0, 1\}. \]
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
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 ($\text{Subset-Sum}$).
hardness, past approaches

- exact solution seems to be hard (\textsc{Subset-Sum}).

- classical: expectation maximization – guess labels, find \((\beta_1^*, \beta_2^*)\), repeat.
hardness, past approaches

- exact solution seems to be hard (Subset-Sum).

- classical: expectation maximization – guess labels, find $(\beta_1^*, \beta_2^*)$, 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

\[ y_i = z_i \cdot \langle x_i, \beta_1^* \rangle + (1 - z_i) \cdot \langle x_i, \beta_2^* \rangle + e_i, \quad i = 1, \ldots, n. \]

- regress \( y_i \) against \( x_i \): get \( \lambda \beta_1^* + (1 - \lambda)\beta_2^* \).
- regress \( y_i^2 \) against \( x_i^\otimes 2 \): get \( \lambda(\beta_1^*)^\otimes 2 + (1 - \lambda)(\beta_2^*)^\otimes 2 \).
- regress \( y_i^3 \) against \( x_i^\otimes 3 \): get \( \lambda(\beta_1^*)^\otimes 3 + (1 - \lambda)(\beta_2^*)^\otimes 3 \).
- (Chaganty & Liang, 2013).
\[ y_i = z_i \cdot \langle x_i, \beta_1^* \rangle + (1 - z_i) \cdot \langle x_i, \beta_2^* \rangle + e_i, \quad i = 1, \ldots, n. \]

- regress \( y_i \) against \( x_i \): get \( \lambda \beta_1^* + (1 - \lambda) \beta_2^* \).
- regress \( y_i^2 \) against \( x_i \otimes 2 \): get \( \lambda (\beta_1^*) \otimes 2 + (1 - \lambda)(\beta_2^*) \otimes 2 \).
- regress \( y_i^3 \) against \( x_i \otimes 3 \): 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

\[ y_i = z_i \cdot \langle x_i, \beta^*_1 \rangle + (1 - z_i) \cdot \langle x_i, \beta^*_2 \rangle + e_i, \quad i = 1, \ldots, n. \]

a convex formulation such that: if \( x_i \) 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

\[ K^* = \frac{1}{2} (\beta_1^* \beta_2^* + \beta_2^* \beta_1^*) \]

\[ g^* = \frac{1}{2} (\beta_1^* + \beta_2^*) . \]

given \( (K^*, g^*) \),

\[ J^* = g^* g^* - K^* = \frac{1}{4} (\beta_1^* - \beta_2^*)(\beta_1^* - \beta_2^*)^\top . \]
arbitrary noise: a convex formulation

\[
\min_{K, g} \| K \|_* \\
\text{subject to } \sum_{i=1}^n \left| -\langle x_i x_i^\top, K \rangle + 2y_i \langle x_i, g \rangle - y_i^2 \right| \leq \eta.
\]
arbitrary noise: a convex formulation

\[
\begin{align*}
\min_{K, g} & \quad \|K\|_* \\
\text{subject to} & \quad \sum_{i=1}^{n} \left| -\langle x_i x_i^\top, K \rangle + 2y_i \langle x_i, g \rangle - y_i^2 \right| \leq \eta.
\end{align*}
\]

- \( \hat{J} = \hat{g}\hat{g}^\top - \hat{K} \)
- \( \hat{\beta}_1, \hat{\beta}_2 = \hat{g} \pm \sqrt{\hat{\lambda}} \hat{\nu} \).
stochastic noise: a convex formulation

\[
\min_{K,g} \sum_{i=1}^{n} \left( -\langle x_i x_i^\top, K \rangle + 2y_i \langle x_i, g \rangle - y_i^2 + \sigma^2 \right)^2 + \mu \| K \|_* .
\]

- \( \hat{J} = \hat{g} \hat{g}^\top - \hat{K} \)
- \( \hat{\beta}_1, \hat{\beta}_2 = \hat{g} \pm \sqrt{\hat{\lambda}} \hat{\nu} . \)
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.,

$$
\|\hat{K} - K^*\|_F \leq c_1 \frac{\|e\|_2}{\sqrt{n}}
$$

$$
\|\hat{g} - g^*\|_2 \leq c_2 \frac{\|e\|_2}{\sqrt{n}}
$$

$$
\|\hat{\beta}_i - \hat{\beta}_i^*\|_2 \leq c_3 \frac{\|e\|_2}{\sqrt{n}}, \quad i = 1, 2.
$$

**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 \((\beta_1^*, \beta_2^*)\) and noise \(e\), with expected loss bounded below.
arbitrary noise: lower bounds

**Theorem.** Let $\Theta(\gamma)$ be the set of $(\beta_1, \beta_2)$ with $\gamma$-bounded norm and separation. If $n \geq c_1 \cdot p$, then for any labeling, w.h.p.,

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

Therefore: the rate $\|e\|/\sqrt{n}$ 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.,

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

\[
\| \hat{\beta}_i - \beta_i^\ast \| \leq \sigma \sqrt{\frac{p}{n} \log^4 n} + \min \left\{ \sigma^2 \frac{\sqrt{p}}{n}, \sigma \left( \frac{p}{n} \right)^{1/4}, \frac{\sigma^2}{\| \beta_1^\ast \| + \| \beta_2^\ast \|} \sqrt{\frac{p}{n}} \right\} \log^4 n.
\]

- (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 \( (\beta_1^*, \beta_2^*) \) with expected loss bounded below.
stochastic noise: lower bounds

**Theorem.** Let \( \Theta(\gamma) \) be the set of \((\beta_1, \beta_2)\) with \(\gamma\)-bounded norm and separation. If \( n \geq c_1 \cdot p \), \( e \sim N(0, \sigma^2 I) \), \( z_i \sim Ber(1/2) \), then w.h.p.

(a) \( \gamma > \sigma \):

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

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

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

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

\[
\inf_{\hat{\theta}} \sup_{\theta^* \in \Theta(\gamma)} \mathbb{E}_{X,z,e} \| \hat{\beta}_i - \beta^*_i \| \geq c \sigma \left( \frac{p}{n} \right)^{1/4}.
\]
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 \((K^*, g^*)\):
some key ideas: upper bounds

directions away from \((K^*, g^*)\):

- concentration inequalities show that \((K^*, g^*)\) feasible w.h.p.
some key ideas: upper bounds

directions away from \((K^*,g^*)\):

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

- \((\hat{K},\hat{g}) = (K^* + \hat{H},g^* + \hat{h}): \|\hat{H}^T\|_* \leq \|\hat{H}_T\|_*\). 

- \(T = \{Z + Y\}\), where \(Z\) (resp. \(Y\)) – a matrix with column (row) space span\(\{\beta_1^*, \beta_2^*\}\).
some key ideas: upper bounds

for $b = 1, 2$ define:

$$B_{b,j} = x_{b,2j}x_{b,2j}^\top - x_{b,2j-1}x_{b,2j-1}^\top,$$

$$(B_b Z)_j = \frac{1}{n_b/2} \langle B_{b,j}, Z \rangle.$$
some key ideas: upper bounds

- for $b = 1, 2$ define:

$$B_{b,j} = x_{b,2j}x_{b,2j}^\top - x_{b,2j-1}x_{b,2j-1}^\top,$$

$$(\mathcal{B}_b Z)_j = \frac{1}{n_b/2} \langle B_{b,j}, Z \rangle.$$

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

$$\sum_b n \| \mathcal{B}_b (-\hat{H} + 2\beta^*_b \hat{h}^\top) \|_1 - c \sum_b \sqrt{n} \| e \|_2 \| \hat{h} \|_2 \leq 2\eta.$$
some key ideas: upper bounds

- for $b = 1, 2$ define:

  $$B_{b,j} = x_{b,2j}x_{b,2j}^\top - x_{b,2j-1}x_{b,2j-1}^\top,$$

  $$\left(\mathcal{B}_b Z\right)_j = \frac{1}{n_b/2} \langle B_{b,j}, Z \rangle.$$

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

  $$\sum_b n \left\| \mathcal{B}_b (-\hat{H} + 2\beta_b^* \hat{h}^\top) \right\|_1 - c \sum_b \sqrt{n} \left\| e \right\|_2 \left\| \hat{h} \right\|_2 \leq 2\eta.$$

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

Constantine Caramanis

Mixed Regression: Minimax Optimal Rates
some key ideas: upper bounds

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

perturbation bounds: Weyl's inequality, and Davis-Kahan's sine theorem.
some key ideas: upper bounds

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

- perturbation bounds: Weyl's inequality, and Davis-Kahan’s sine theorem.
Information-theoretic setup:

- nature sends $\theta^* = (\beta_1^*, \beta_2^*) \in \Theta$.
- we receive $\{(y_i, x_i)\}_{i=1}^n$.
- capacity of channel: sample size vs resolution (error).
lower bounds: Fano’s inequality

- let $\Theta(\gamma)$ be pairs $(\beta_1, \beta_2)$ with norm and separation lower bounded by $\gamma$. 
lower bounds: Fano’s inequality

- Let $\Theta(\gamma)$ be pairs $(\beta_1, \beta_2)$ with norm and separation lower bounded by $\gamma$.

- Let $\Theta_{\text{disc}} = \{\theta_1, \ldots, \theta_M\} \subseteq \Theta(\gamma)$ be a $\delta$-packing.
lower bounds: Fano’s inequality

- let $\Theta(\gamma)$ be pairs $(\beta_1, \beta_2)$ with norm and separation lower bounded by $\gamma$.

- let $\Theta_{\text{disc}} = \{\theta_1, \ldots, \theta_M\} \subseteq \Theta(\gamma)$ be a $\delta$-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 $(\beta_1, \beta_2)$ with norm and separation lower bounded by $\gamma$.

- Let $\Theta_{\text{disc}} = \{\theta_1, \ldots, \theta_M\} \subseteq \Theta(\gamma)$ be a $\delta$-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 $X \rightarrow Y \rightarrow \hat{X}$,

$$P(\hat{X} \neq X) \geq \frac{H(X|Y) - 1}{\log |\mathcal{X}|}.$$
lower bounds: Fano’s inequality

(Fano). For any estimator $\hat{X}$ of $X$, with $X \rightarrow Y \rightarrow \hat{X}$,

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

Rearranging and using our notation:

$$\mathbb{P}(\hat{\theta} \neq \theta^*) \geq 1 - \frac{I((y, X); \theta^*) + \log 2}{\log M}.$$
lower bounds: Fano’s inequality

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

\[
P(\hat{X} \neq X) \geq \frac{H(X|Y) - 1}{\log |\mathcal{X}|}.
\]

- rearranging and using our notation:

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

- construction: want \( M \) big, and \( I((y, X); \theta^*) \) small.
lower bounds: Fano’s inequality

how $\delta$-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 \((X, y)\) on \( \beta^* = \beta_i \).

- then by definition and convexity (Jensen’s) of mutual information:

\[
I(\beta^*; (X, y)) = \frac{1}{M} \sum_i D(\mathbb{P}_i \| (1/M) \sum_j \mathbb{P}_j) \leq \frac{1}{M^2} \sum_{i,j} D(\mathbb{P}_i \| \mathbb{P}_j).
\]
lower bounds: Fano’s inequality

high SNR vs low SNR

- high SNR: enough to obtain lower bound for regression, $\theta_i = \xi_i$, and $P_i$ Gaussian. hence:

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

- low SNR: $\theta_i = (\xi_i, -\xi_i)$, and $P_i$ is mixture of two Gaussians.
stochastic setting for unbalanced labels:

$$\min \sum_{i=1}^{n} \left( -\langle x_i; x_i^\top, K \rangle + 2y_i \langle x_i, g \rangle - y_i^2 + \sigma^2 \right)^2 - 4\sigma^2 (y_i - \langle x_i, g \rangle)^2$$

s.t. : \[ \|K\|_* \leq \|K^*\|_* . \]

- 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