Elementary Estimators for High-Dimensional Linear Regression

Eunho Yang, Aurélie C. Lozano, Pradeep Ravikumar.
ICML-2014

Zoltán Szabó

Machine Learning Journal Club, Gatsby

November 24, 2014

Problem domain: linear regression model, highD

Observation:

$$y_i = \mathbf{x}_i^T \boldsymbol{\theta}^* + w_i \quad (i = 1, \ldots, n) \quad \Leftrightarrow \quad \mathbf{y} = \mathbf{X}\boldsymbol{\theta}^* + \mathbf{w}.$$

Basic assumptions:

- Given input-output: $\mathbf{x}_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$; $\mathbf{X} = \left[\mathbf{x}_1^T; \ldots; \mathbf{x}_n^T\right]$.
- Observation noise: $w_i \sim N(0, \sigma^2)$, i.i.d.
- Fixed, unknown parameter of interest: $\boldsymbol{\theta}^* \in \mathbb{R}^p$.
- High-dimensional setting: $n \ll p$.

Previous methods

- (Structured) sparse, low-rank solvers: iterative methods, no analytical formula.
- Examples:
  - Lasso (ℓ_1-regularized least squares):

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda_n \|\boldsymbol{\theta}\|_1 .$$

  - Dantzig estimator (linear program):

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^p : \frac{1}{n}\|\mathbf{X}^T(\mathbf{X}\boldsymbol{\theta}-\mathbf{y})\|_\infty \le \lambda_n} \|\boldsymbol{\theta}\|_1 .$$

Goal + idea

- OLS ($n > p$): $\hat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{y})$.
- Ridge solution ($\epsilon > 0$):

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X} + \epsilon\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \epsilon \|\boldsymbol{\theta}\|_2^2.$$

- Idea:
  - Task: highD linear regression with structural constraints.
  - Suggested solvers: Dantzig-type estimators (structured).
- Result: analytical solution + theoretical guarantees.

- R: regularizer 'compatible' with our structural constraint.
- $R^*(\mathbf{u}) = \sup_{\boldsymbol{\theta}\in\mathbb{R}^p\setminus\{\mathbf{0}\}} \frac{\mathbf{u}^T\boldsymbol{\theta}}{R(\boldsymbol{\theta})} = \sup_{\boldsymbol{\theta}\in\mathbb{R}^p:R(\boldsymbol{\theta})\le 1} \langle\boldsymbol{\theta},\mathbf{u}\rangle$: dual norm.
- Elem-solvers:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^p:R^*\left(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}}\right)\le\lambda_n} R(\boldsymbol{\theta}),$$

where

$$\bar{\boldsymbol{\theta}}_{\text{OLS}} := \left[T_\nu\left(\frac{\mathbf{X}^T\mathbf{X}}{n}\right)\right]^{-1}\frac{\mathbf{X}^T\mathbf{y}}{n}, \quad \bar{\boldsymbol{\theta}}_{\text{RIDGE}} := (\mathbf{X}^T\mathbf{X}+\epsilon\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y},$$

Diagonal dominizer with ν: $T_\nu(\mathbf{A}) = \begin{cases} A_{ii}+\nu & i=j, \\ sign(A_{ij})(|A_{ij}|-\nu) & i\neq j. \end{cases}$

Structural constraint: subspace pair $(\mathcal{M}, \bar{\mathcal{M}}^{\perp})$

- True parameter $\boldsymbol{\theta}^* \in \mathcal{M} \subseteq \bar{\mathcal{M}}$.
- $\mathcal{M}$: model subspace, typically low-dimensional.
- $\bar{\mathcal{M}}^{\perp}$: perturbation subspace, perturbations from $\mathcal{M}$.
- Examples (details soon):
    - sparse/structured-sparse vectors ($\mathcal{M} = \bar{\mathcal{M}}$),
    - low-rank matrices.

Decomposable regularizer: R

- R is decomposable w.r.t. $(\mathcal{M}, \bar{\mathcal{M}}^{\perp})$ if

$$R(\mathbf{u}+\mathbf{v}) = R(\mathbf{u}) + R(\mathbf{v}), \quad \forall \mathbf{u} \in \mathcal{M}, \mathbf{v} \in \bar{\mathcal{M}}^{\perp}.$$

- Meaning:
  - For a norm R: l.h.s. $\le$ r.h.s.
  - $\mathbf{u}+\mathbf{v}$: perturbation of the model vector $\mathbf{u}$ from $\mathcal{M}$.
  - Decomposable R:
    - penalizes deviations as much as possible,
    - l.h.s.=r.h.s.

Example-1: $\boldsymbol{\theta}^* \in \mathbb{R}^p$ is sparse (or group-sparse)

- $S = supp(\boldsymbol{\theta}^*) \subseteq \{1, \ldots, p\}$.
- $\mathcal{M} = \mathcal{M}(S) := \{\boldsymbol{\theta} \in \mathbb{R}^p : supp(\boldsymbol{\theta}) \subseteq S\}$
- $\mathcal{M} = \bar{\mathcal{M}}$, $\bar{\mathcal{M}}^\perp = \{\boldsymbol{\theta} \in \mathbb{R}^p : supp(\boldsymbol{\theta}) \subseteq S^c\}$.
- $R(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1$ is decomposable w.r.t. $(\mathcal{M}, \bar{\mathcal{M}}^\perp)$:

$$\mathbf{u} = (\mathbf{u}_S, \mathbf{0}_{S^c}) \in \mathcal{M}, \quad \mathbf{v} = (\mathbf{0}_S, \mathbf{v}_{S^c}) \in \bar{\mathcal{M}}^\perp,$$
$$\|\mathbf{u} + \mathbf{v}\|_1 = \|(\mathbf{u}_S, \mathbf{0}_{S^c})\|_1 + \|(\mathbf{0}_S, \mathbf{v}_{S^c})\|_1 = \|\mathbf{u}\|_1 + \|\mathbf{v}\|_1 .$$

Example-2: $\mathbf{\Theta}^* \in \mathbb{R}^{p_1 \times p_2}$ is low-rank

- Nuclear norm:

$$\|\mathbf{\Theta}\|_* = \|\boldsymbol{\sigma}(\mathbf{\Theta})\|_1 .$$

- Subspace pair: $\mathbf{\Theta}^* \Rightarrow U = col(\mathbf{\Theta}^*),\ V = row(\mathbf{\Theta}^*)$,

$$\mathcal{M}(U,V) := \{\mathbf{\Theta} \in \mathbb{R}^{p_1 \times p_2} : col(\mathbf{\Theta}) \subseteq U, row(\mathbf{\Theta}) \subseteq V\},$$
$$\bar{\mathcal{M}}^{\perp}(U,V) := \{\mathbf{\Theta} \in \mathbb{R}^{p_1 \times p_2} : col(\mathbf{\Theta}) \subseteq U^{\perp}, row(\mathbf{\Theta}) \subseteq V^{\perp}\}.$$

Example-2: continued

- $\mathcal{M} \subseteq \bar{\mathcal{M}}$:

$$\mathcal{M} \ni \mathbf{A} = \mathbf{U}_1\mathbf{D}_1\mathbf{V}_1^T \Rightarrow \mathbf{U}_1 \subseteq U, \mathbf{V}_1 \subseteq V$$
$$\bar{\mathcal{M}}^\perp \ni \mathbf{B} = \mathbf{U}_2\mathbf{D}_2\mathbf{V}_2^T \Rightarrow \mathbf{U}_2 \subseteq U^\perp, \mathbf{V}_2 \subseteq V^\perp,$$
$$\mathbf{A}^T\mathbf{B} = \mathbf{V}_1\mathbf{D}_1^T\mathbf{U}_1^T\mathbf{U}_2\mathbf{D}_2\mathbf{V}_2^T = \mathbf{V}_1\mathbf{D}_1^T\mathbf{0}\mathbf{D}_2\mathbf{V}_2^T = \mathbf{0} \Rightarrow$$
$$\langle \mathbf{A}, \mathbf{B} \rangle = tr\left(\mathbf{A}^T\mathbf{B}\right) = 0 \Rightarrow \mathcal{M} \subseteq \left(\bar{\mathcal{M}}^\perp\right)^\perp = \bar{\mathcal{M}}.$$

- Any $(\mathbf{A}, \mathbf{B}) \in (\mathcal{M}, \bar{\mathcal{M}}^\perp)$ have orthogonal row and column spaces $\Rightarrow \|\mathbf{A} + \mathbf{B}\|_* = \|\mathbf{A}\|_* + \|\mathbf{B}\|_*$.

Dual norm: R^*

$G_1, \ldots, G_N$: partition of $\{1, \ldots, p\}$.

$$R(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1, \qquad R^*(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_\infty.$$

$$R(\boldsymbol{\theta}) = \sum_{n=1}^{N} \|\boldsymbol{\theta}_{G_n}\|_{a_n}, \qquad (l_1/l_{\mathbf{a}} - \text{norm}, a_n \in [2, \infty])$$

$$R^*(\boldsymbol{\theta}) = \max_{n=1,\ldots,N} \|\boldsymbol{\theta}_{G_n}\|_{a_n^*}, \qquad (l_\infty/l_{\mathbf{a}^*} - \text{norm}, \frac{1}{a_n} + \frac{1}{a_n^*} = 1).$$

$$R(\boldsymbol{\Theta}) = \|\boldsymbol{\Theta}\|_* = \|\boldsymbol{\sigma}(\boldsymbol{\Theta})\|_1, \qquad R^*(\boldsymbol{\Theta}) = \|\boldsymbol{\sigma}(\boldsymbol{\Theta})\|_\infty.$$

Analytical solution: (group-)sparse example

Regularizers:

$$R(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1, \qquad R(\boldsymbol{\theta}) = \sum_{n=1}^{N} \|\boldsymbol{\theta}_{G_n}\|_{a_n}.$$

Coordinate/group-decomposable tasks ($\Leftarrow$ constraint: $\|\cdot\|_\infty$); explicit solutions:

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= S_{\lambda_n}(\bar{\boldsymbol{\theta}}),\\
[S_\lambda(\mathbf{u})]_i &= sign(u_i)\max(|u_i| - \lambda, 0), \quad \text{(soft-thresholding)}\\
[S_\lambda(\mathbf{u})]_{G_i} &= \frac{\mathbf{u}_{G_i}}{\|\mathbf{u}_{G_i}\|_{a_i}}\max(\|\mathbf{u}_{G_i}\|_{a_i} - \lambda, 0), \quad \text{(block soft-thresholding).}
\end{aligned}$$

+Def: subspace compatibility constant (Ψ), projection

- It measures the relation between R and $\|\cdot\|_2$:

$$\Psi(\mathcal{M}) := \sup_{\mathbf{u}\in\mathcal{M}\setminus\{\mathbf{0}\}} \frac{R(\mathbf{u})}{\|\mathbf{u}\|_2}.$$

  Example (sparse): $\|\boldsymbol{\theta}^*\|_0 = k \to \mathcal{M} \to \Psi(\mathcal{M}) = \sqrt{k}$.
- Projection to a subspace S:

$$\Pi_S(\mathbf{u}) := \arg\min_{\mathbf{v}\in S} \|\mathbf{u}-\mathbf{v}\|_2 .$$

Theoretical guarantee: deterministic bound

If

- R is decomposable w.r.t. $(\mathcal{M}, \bar{\mathcal{M}}^{\perp})$, $\boldsymbol{\theta}^* \in \mathcal{M}$,
- $\lambda_n \geq R^*(\boldsymbol{\theta}^* - \bar{\boldsymbol{\theta}})$,

then the elem-estimators $(\hat{\boldsymbol{\theta}})$ satisfy the error bounds

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

Note: for Elem-OLS λ_n can be chosen "better".

Proof:

$$\begin{aligned}
\boldsymbol{\Delta} &:= \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*, \\
R^*(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) &\le \lambda_n \quad \text{(feasibility of } \hat{\boldsymbol{\theta}}\text{)}, \\
R^*(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) &\le \lambda_n \quad \text{(our assumption)}, \\
R^*(\boldsymbol{\Delta}) &\overset{\text{(i)}}{=} R^*(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}} + \bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \overset{\text{(ii)}}{\le} R^*(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) + R^*(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \overset{\text{(iii)}}{\le} 2\lambda_n.
\end{aligned}$$

Reasoning: (i) $\pm\bar{\boldsymbol{\theta}}$, (ii) triangle ineq. for R^*, (iii) see above.

Notation: $(S, S^c) := (\mathcal{M}, \bar{\mathcal{M}}^{\perp})$, $\boldsymbol{\Delta}_S := \Pi_S(\boldsymbol{\Delta})$.

$$\begin{aligned}R(\boldsymbol{\theta}^*) &\overset{\text{(i)}}{=} R(\boldsymbol{\theta}^*) + R(\boldsymbol{\Delta}_{S^c}) - R(\boldsymbol{\Delta}_{S^c}) \overset{\text{(ii)}}{=} R(\boldsymbol{\theta}^* + \boldsymbol{\Delta}_{S^c}) - R(\boldsymbol{\Delta}_{S^c})\\
&\overset{\text{(iii)}}{\le} R(\boldsymbol{\theta}^* + \boldsymbol{\Delta}_{S^c} + \boldsymbol{\Delta}_S) + R(\boldsymbol{\Delta}_S) - R(\boldsymbol{\Delta}_{S^c})\\
&\overset{\text{(iv)}}{=} R(\boldsymbol{\theta}^* + \boldsymbol{\Delta}) + R(\boldsymbol{\Delta}_S) - R(\boldsymbol{\Delta}_{S^c}).\end{aligned}$$

Reasoning: (i) $\pm R(\boldsymbol{\Delta}_{S^c})$, (ii) R decomposable, $\boldsymbol{\theta}^* \in \mathcal{M}$, (iii) reverse triangle ineq. for R, (iv) $\boldsymbol{\Delta}_{S^c} + \boldsymbol{\Delta}_S \overset{?}{=} \boldsymbol{\Delta}$ ($\mathcal{M} = \bar{\mathcal{M}}$: OK).

$$R(\boldsymbol{\theta}^* + \boldsymbol{\Delta}) \overset{(i)}{=} R(\hat{\boldsymbol{\theta}}) \overset{(ii)}{\leq} R(\boldsymbol{\theta}^*) \quad \overset{(iii)}{\Rightarrow} \quad 0 \leq R(\boldsymbol{\Delta}_S) - R(\boldsymbol{\Delta}_{S^c}),$$

$$\|\boldsymbol{\Delta}\|_2^2 = \langle \boldsymbol{\Delta}, \boldsymbol{\Delta} \rangle \overset{(iv)}{\leq} R^*(\boldsymbol{\Delta}) R(\boldsymbol{\Delta}) \overset{(v)}{\leq} R^*(\boldsymbol{\Delta})[R(\boldsymbol{\Delta}_S) + R(\boldsymbol{\Delta}_{S^c})]$$

$$\overset{(vi)}{\leq} 2R^*(\boldsymbol{\Delta}) R(\boldsymbol{\Delta}_S) \overset{(vii)}{\leq} 4\Psi(S)\lambda_n \|\boldsymbol{\Delta}_S\|_2 \overset{(viii)}{\Rightarrow}$$

$$\|\boldsymbol{\Delta}_S\|_2 \leq 4\Psi(S)\lambda_n.$$

Reasoning: (i) $\boldsymbol{\Delta}$ definition, (ii) objective function of $\hat{\boldsymbol{\theta}}$, (iii) combination with the previous result, (iv) generalized CBS, (v) triangle ineq. for R with $\boldsymbol{\Delta} \overset{?}{=} \boldsymbol{\Delta}_S + \boldsymbol{\Delta}_{S^c}$, (vi) end of first row, (vii) $R^*(\boldsymbol{\Delta}) \leq 2\lambda_n$ bound; Ψ def. $\Rightarrow R(\boldsymbol{\Delta}_S) \leq \Psi(S) \|\boldsymbol{\Delta}_S\|_2$, (viii) $\|\boldsymbol{\Delta}_S\|_2 \leq \|\boldsymbol{\Delta}\|_2$: proj. is non-expansive.

Proof: error bound in R

$$R(\mathbf{\Delta}) \overset{\text{(i)}}{\leq} R(\mathbf{\Delta}_S) + R(\mathbf{\Delta}_{S^c}) \overset{\text{(ii)}}{\leq} 2R(\mathbf{\Delta}_S) \overset{\text{(iii)}}{\leq} 2\Psi(S)\,\|\mathbf{\Delta}_S\|_2$$
$$\overset{\text{(iv)}}{\leq} 8[\Psi(S)]^2\lambda_n.$$

Reasoning: (i) triangle ineq. for R with $\mathbf{\Delta} \overset{?}{=} \mathbf{\Delta}_S + \mathbf{\Delta}_{S^c}$, (ii) previous $R(\mathbf{\Delta}_{S^c}) \leq R(\mathbf{\Delta}_S)$ bound, (iii) Ψ def.
$\Rightarrow R(\mathbf{\Delta}_S) \leq \Psi(S)\,\|\mathbf{\Delta}_S\|_2$, (iv) previous $\|\mathbf{\Delta}_S\|_2 \leq 4\Psi(S)\lambda_n$ bound.

Numerical experiments

- Elem-OLS is
  - superior to Elem-ridge, especially in highD ($n \ll p$),
  - comparable/superior to alternative (iterative) solvers.
- Gene expression analysis: Elem-OLS
  - beats Lasso (cross-validated performance),
  - finds a biologically motivated gene (not selected by Lasso).

Summary

- Task: highD linear regression with structural constraint.
- Proposed technique:
  - closed-form solution,
  - theoretical guarantees.
- Nice numerical properties.

Thank you for the attention!