A UNIFIED APPROACH TO MODEL SELECTION AND SPARSE RECOVERY USING REGULARIZED LEAST SQUARES

by Jinchi Lv and Yingying Fan

The annals of Statistics (2009)

presented by Quoc Tran

Mar. 19 . 2010

Outline

1. Goals
2. Backgrounds
   - Sideline information
   - Notations
3. Results
   - Penalty Family
   - Regularized least squares
   - Sparse Recovery
   - Model Selection

A unified approach

- Consider both problems of model selection and sparse recovery in the unified framework of regularized least squares with concave penalties:

$$\min_{\beta \in R^p} \{2^{-1} \|X - \beta\|_2^2 + \Lambda_n \sum_{j=1}^{p} \rho_{\lambda_n}(|\beta_j|)\}$$

- Consider a family of penalty functions that give a smooth homotopy between L_0 and L_1 penalties for both problems. This family includes Lasso [Tibshirani (1996)] and has similar properties as SCAD [Fan (1997) and MCP [Zhang (2007)]:

$$\rho_a(t) = \frac{(a+1)t}{a+t} = \frac{t}{a+t} I\{t \neq 0\} + (\frac{a}{a+t})t$$

Main achievements

- CONDITION 1: $\rho(t)$ is increasing and concave in $t \in [0, \infty)$, and has a continuous derivative $\rho'(t)$ with $\rho'(0^+) \in (0, \infty)$. If $\rho(t)$ is dependent on λ, $\rho'(t; \lambda)$ is increasing in $\lambda \in (0, \infty)$ and $\rho'(0^+)$ is independent of λ.
  - Penalties satisfying Condition 1 and $\lim_{t \to \infty} \rho'(t) = 0$ enjoy the unbiasedness and sparsity. However, the continuity does not generally hold for all penalties in this class.
  - $\rho_a(t)$ provided before satisfies Condition 1 and three properties simultaneously, and share the same spirit as SCAD and MCP.
  - Under some conditions we can obtain optimal $\rho_a(t)$ for the two previous mention problems.

Main achievements(cont)

- For model selection, under some conditions, they can optain weak oracle property, where the dimensionality can grow exponentially with sample size.
- For sparse recovery, they present a sufficient conditions that ensures the recoverability of the sparsest solution.

Sideline information

- About authors: this Fan (Fan, Yingying) is not the famous Fan (Fan, Jianqing) in Princeton. They are both students of Fan, Jianqing. They follow a branch of research developed by Fan, Jianqing:
  - Fan, J. and Li, R. (2001)
  - Fan, J. and Li, R. (2006)
  - Fan, J. and Peng, H. (2004) ....
- This is another effort to provide penalty function, as SCAD and MCP to overcome Lasso weakness.
- This paper is a good survey of the methods so far.
- About result: this is a more equipped but direct generalization of Liu and Wu (2007)

Model Selection and Sparse Recovery

- Sparse Recovery:

$$\min \sum_{j=1}^{p} \rho(|\beta_j|) \text{ subject to } \mathbf{y} = \mathbf{X}\boldsymbol{\beta}, \tag{1}$$

where $\rho(.)$ is a penalty function and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$. The target penalty function is L_0: $\rho(t) = I(t \neq 0)$

- Model selection:

$$\min_{\beta \in R^p} \{2^{-1} \|X - \beta\|_2^2 + \Lambda_n \sum_{j=1}^{p} \rho_{\lambda_n}(|\beta_j|)\} \tag{2}$$

where $\Lambda_n \in (0, \infty)$ is scale parameter and $\lambda_n \in [0, \infty)$ is a regularization parameter indexed by sample size n.

Concavity

- Maximum Concavity:

$$\kappa(\rho) = \sup_{t_1, t_2 \in (0, \infty), t_1 < t_2} - \frac{\rho'(t_2) - \rho'(t_1)}{t_2 - t_1} \tag{3}$$

- Local Concavity at $\mathbf{b} = (b_1, ..., b_q)^T \in \mathbf{R}^q$ with $\|\mathbf{b}\|_0 = q$:

$$\kappa(\rho; \mathbf{b}) = \lim_{\epsilon \to \mathbf{0}^+} \max_{\mathbf{1} \leq \mathbf{j} \leq \mathbf{q}} \sup_{\mathbf{t_1}, \mathbf{t_2} \in (|\mathbf{b_j}| - \epsilon, \mathbf{b_j}| + \epsilon), \mathbf{t_1} < \mathbf{t_2}} - \frac{\rho'(\mathbf{t_2}) - \rho'(\mathbf{t_1})}{\mathbf{t_2} - \mathbf{t_1}} \tag{4}$$

Penalty Family

- Condition 1 provides a general family.
- $\rho_a(t)$ provided above satisfies Condition 1 and three properties.

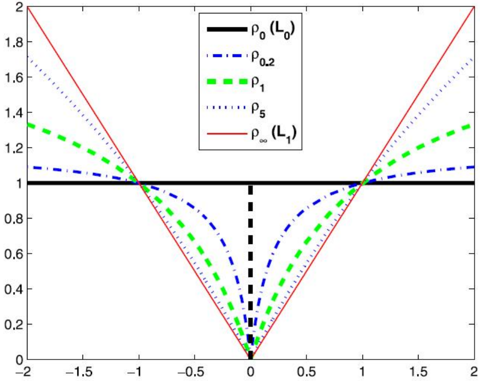

FIG. 1. *Plot of penalty functions ρ_0 (L_0, thick solid), $\rho_{0.2}$ (dash-dot), ρ_1 (dashed), ρ_5 (dotted),*

Regularized least squares

THEOREM 1 (Regularized least squares). *Assume that* p_λ *satisfies Condition* 1 *and* $\widehat{\boldsymbol\beta}^\lambda \in \mathbf{R}^p$ *with* $\mathbf{Q} = \mathbf{X}^T_{\widehat{\mathfrak{M}}_\lambda} \mathbf{X}_{\widehat{\mathfrak{M}}_\lambda}$ *nonsingular, where* $\lambda \in (0, \infty)$ *and* $\widehat{\mathfrak{M}}_\lambda = \operatorname{supp}(\widehat{\boldsymbol\beta}^\lambda)$. *Then* $\widehat{\boldsymbol\beta}^\lambda$ *is a strict local minimizer of* (2) *with* $\lambda_n = \lambda$ *if*

$$\widehat{\boldsymbol\beta}^\lambda_{\widehat{\mathfrak{M}}_\lambda} = \mathbf{Q}^{-1}\mathbf{X}^T_{\widehat{\mathfrak{M}}_\lambda}\mathbf{y} - \Lambda_n \lambda \mathbf{Q}^{-1} \bar\rho(\widehat{\boldsymbol\beta}^\lambda_{\widehat{\mathfrak{M}}_\lambda}), \tag{18}$$

$$\|\mathbf{z}_{\widehat{\mathfrak{M}}^c_\lambda}\|_\infty < \rho'(0+), \tag{19}$$

$$\lambda_{\min}(\mathbf{Q}) > \Lambda_n \lambda \kappa(\rho; \widehat{\boldsymbol\beta}^\lambda_{\widehat{\mathfrak{M}}_\lambda}), \tag{20}$$

where $\mathbf{z} = (\Lambda_n\lambda)^{-1}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol\beta}^\lambda)$, $\lambda_{\min}(\cdot)$ *denotes the smallest eigenvalue of a given symmetric matrix*

Sparse Recovery

THEOREM 2 (Sparse recovery). *Assume that ρ satisfies Condition 1 with $\kappa(\rho) \in [0, \infty)$, $\mathbf{Q} = \mathbf{X}_{\mathfrak{M}_0}^T \mathbf{X}_{\mathfrak{M}_0}$ is nonsingular with $\mathfrak{M}_0 = \operatorname{supp}(\boldsymbol{\beta}_0)$, and $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$. Then $\boldsymbol{\beta}_0$ is a local minimizer of (1) if there exists some $\epsilon \in (0, \min_{j \in \mathfrak{M}_0} |\beta_{0,j}|)$ such that*

$$\max_{j \in \mathfrak{M}_0^c} \max_{\mathbf{u} \in \mathcal{U}_\epsilon} |\langle \mathbf{x}_j, \mathbf{u} \rangle| < \rho'(0+), \tag{22}$$

where $\mathcal{U}_\epsilon = \{\mathbf{X}_{\mathfrak{M}_0} \mathbf{Q}^{-1} \bar{\rho}(\mathbf{v}) : \mathbf{v} \in \mathcal{V}_\epsilon\}$ and $\mathcal{V}_\epsilon = \prod_{j \in \mathfrak{M}_0} \{t : |t - \beta_{0,j}| \le \epsilon\}$.

Optimal ρ_a

THEOREM 3 (Optimal ρ_a penalty for sparse recovery). *Assume that* $\mathbf{Q} = \mathbf{X}_{\mathfrak{M}_0}^T \mathbf{X}_{\mathfrak{M}_0}$ *is nonsingular with* $\mathfrak{M}_0 = \operatorname{supp}(\boldsymbol{\beta}_0)$ *and* $\epsilon \in (0, \min_{j \in \mathfrak{M}_0} |\beta_{0,j}|)$. *Then the optimal penalty* $\rho_{a_{\mathrm{opt}}(\epsilon)}$ *satisfies:*

(a) $a_{\mathrm{opt}}(\epsilon) \in (0, \infty]$ *and is the largest* $a \in (0, \infty]$ *such that*

$$\max_{j \in \mathfrak{M}_0^c} \max_{\mathbf{u} \in \mathcal{U}_\epsilon} |\langle \mathbf{x}_j, \mathbf{u} \rangle| \le 1 + a^{-1}, \tag{26}$$

where $\mathcal{U}_\epsilon = \{\mathbf{X}_{\mathfrak{M}_0} \mathbf{Q}^{-1} \bar{\rho}(\mathbf{v}) : \mathbf{v} \in \mathcal{V}_\epsilon\}$ *and* $\mathcal{V}_\epsilon = \prod_{j \in \mathfrak{M}_0} \{t : |t - \beta_{0,j}| \le \epsilon\}$.

(b) $a_{\mathrm{opt}}(\epsilon) = \infty$ *if and only if*

$$\max_{j \in \mathfrak{M}_0^c} |\langle \mathbf{x}_j, \mathbf{u}_0 \rangle| \le 1, \tag{27}$$

where $\mathbf{u}_0 = \mathbf{X}_{\mathfrak{M}_0} \mathbf{Q}^{-1} \operatorname{sgn}(\boldsymbol{\beta}_{0,\mathfrak{M}_0})$.

Conditions

CONDITION 2. $\mathbf{X}$ satisfies

$$\|(\mathbf{X}_{\mathfrak{M}_0}^T \mathbf{X}_{\mathfrak{M}_0})^{-1}\|_\infty \le C_{1n}, \tag{34}$$

$$\|\mathbf{X}_{\mathfrak{M}_0^c}^T \mathbf{X}_{\mathfrak{M}_0}(\mathbf{X}_{\mathfrak{M}_0}^T \mathbf{X}_{\mathfrak{M}_0})^{-1}\|_\infty \le C_{2n}, \tag{35}$$

where $\mathfrak{M}_0 = \operatorname{supp}(\boldsymbol{\beta}_0)$, $C_{1n} \in (0, \infty)$, $C_{2n} \in [0, C\frac{\rho'(0+)}{\rho'(c_0 b_0)}]$ for some $C, c_0 \in (0, 1)$, $b_0 = \min_{j\in\mathfrak{M}_0} |\beta_{0,j}|$, and $\|\cdot\|_\infty$ denotes the matrix ∞-norm.

Here and below, ρ is associated with regularization parameter $\underline{\lambda}_n$ defined in (38) unless specified otherwise.

CONDITION 3. $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$ for some $\sigma > 0$.

Conditions(cont)

CONDITION 4. There exists some $\gamma \in (0, \frac{1}{2}]$ such that

$$\left[D_{1n} + \frac{\rho'(c_0 b_0)}{\rho'(0+)} D_{2n}\right] C_{1n} = O(n^{-\gamma}), \tag{36}$$

where $D_{1n} = \max_{j \in \mathfrak{M}_0} \|\mathbf{x}_j\|_2$, $D_{2n} = \max_{j \in \mathfrak{M}_0^c} \|\mathbf{x}_j\|_2$ and $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$. Let $u_n \in (0, \infty)$ satisfy $\lim_{n\to\infty} u_n = \infty$, $\underline{\lambda}_n \le \overline{\lambda}_n$, and

$$u_n \le [\kappa_0(C_{2n} D_{1n} + D_{2n})]^{-1} \lambda_{\min}(\mathbf{X}_{\mathfrak{M}_0}^T \mathbf{X}_{\mathfrak{M}_0})(1 - C)\rho'(0+)\sigma^{-1}, \tag{37}$$

where

$$\underline{\lambda}_n = \Lambda_n^{-1} \frac{(C_{2n} D_{1n} + D_{2n}) u_n \sigma}{\rho'(0+) - C_{2n} \rho'(c_0 b_0)} \quad \text{and} \quad \overline{\lambda}_n = \frac{C_{1n}^{-1}(1 - c_0) b_0 - u_n D_{1n} \sigma}{\Lambda_n \rho'(c_0 b_0; \overline{\lambda}_n)}, \tag{38}$$

$C, c_0 \in (0, 1)$ are given in Condition 2, and $\kappa_0 = \max\{\kappa(\rho; \mathbf{b}) : \|\mathbf{b} - \boldsymbol{\beta}_{0,\mathfrak{M}_0}\|_\infty \le (1 - c_0) b_0\}$

Weak Oracle Property

THEOREM 4 (Weak oracle property). *Assume that p_λ in (4) satisfies Condition 1, Conditions 2–4 hold and $p = o(u_n e^{u_n^2/2})$. Then there exists a regularized least squares estimator $\widehat{\boldsymbol{\beta}}^{\lambda_n}$ with regularization parameter $\lambda_n = \underline{\lambda}_n$ defined in (38) such that with probability at least $1 - \frac{2}{\sqrt{\pi}} p u_n^{-1} e^{-u_n^2/2}$, $\widehat{\boldsymbol{\beta}}^{\lambda_n}$ satisfies:*

(a) (Sparsity) $\widehat{\boldsymbol{\beta}}^{\lambda_n}_{\mathfrak{M}_0^c} = \mathbf{0}$;

(b) (L_∞ loss) $\|\widehat{\boldsymbol{\beta}}^{\lambda_n}_{\mathfrak{M}_0} - \boldsymbol{\beta}_{0,\mathfrak{M}_0}\|_\infty \le h = O(n^{-\gamma} u_n)$,

where $\mathfrak{M}_0 = \operatorname{supp}(\boldsymbol{\beta}_0)$ and $h = [D_{1n} + \frac{\rho'(c_0 b_0)}{\rho'(0+)} D_{2n}] C_{1n} u_n (1 - C)^{-1} \sigma$. As a consequence, $\|\widehat{\boldsymbol{\beta}}^{\lambda_n} - \boldsymbol{\beta}_0\|_2 = O_P(\sqrt{s} n^{-\gamma} u_n)$, where $s = \|\boldsymbol{\beta}_0\|_0$.

Simulation Result for large p

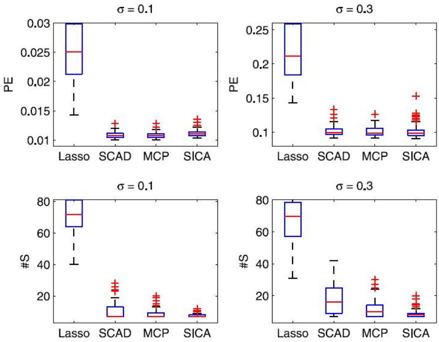

FIG. 4. *Boxplots of PE and #S over 100 simulations for all methods in Simulation 3, where* $p = 600$ *and the rows of* $\mathbf{X}$ *are i.i.d. copies from* $N(\mathbf{0}, \Sigma_0)$. *The x-axis represents different methods. Top panel is for PE and bottom panel is for #S.*