ETH zürich

Automatic Control Laboratory

A few notes on l_1-norm regularization

Mingzhou Yin

March 16, 2023, IfA Coffee Talk

The variable selection problem

Consider linear regression model

$$y_i = \sum_{j=1}^{p} \beta_j x_i^j + \epsilon_i = \mathbf{x}_i \beta + \epsilon_i$$

- Data: $(Z_i)_{i=1}^n = \left(\mathbf{x}_i \in \mathbb{R}^{1\times p}, y_i \in \mathbb{R}\right)_{i=1}^n \to \mathbf{y} = \Phi\beta = \sum_{j=1}^p \beta_j \phi_j$
- High-dimensional regime: $p \gg n$
- Sparse problem: only a few covariates are relevant, i.e., β is sparse.

$$S = \{j \,|\, \beta_j \neq 0\} \sim \text{active set}, \quad \log p \cdot |S| \ll n$$

- Sparsity: $\|\beta\|_0 := |S|$ (pseudo-norm)

Examples

- Pole location identification:

$$\mathbf{y} = \sum_{j=1}^{p} c_j \left(G_j * \mathbf{u}\right) + \epsilon_i, \quad c_j\text{: sparse}, \quad G_j\text{: first-order systems}$$

- Network identification:

$$\mathbf{x}^{t+1} = \sum_{j=1}^{p} \left(A_j x_j^t + B_j u_j^t\right), \quad A_j, B_j\text{: sparse}$$

- Switched system identification:

$$\mathbf{x}^{t+1} = A_t \mathbf{x}^t + B_t \mathbf{u}^t, \quad \left(A_{t+1} - A_t\right), \left(B_{t+1} - B_t\right)\text{: sparse}$$

Variable selection as convex optimization

Sparsity-constrained least squares problem:

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}} \quad \sum_{i=1}^{n}\left(y_i-\mathbf{x}_i\beta\right)^2 \quad \text{subject to} \quad \|\beta\|_0 \le m$$

- NP-hard combinatorial problem ($\beta = V\nu$, $\nu \in \mathbb{R}^m$, V: binary matrix)
- The best convex surrogate of the sparsity function: $\|\beta\|_0 \le m \rightarrow \|\beta\|_1 \le \zeta$
- ζ loses its physical meaning $\rightarrow$ equivalent to the Lagrangian form:

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}} \quad J(\beta)=\sum_{i=1}^{n}\left(y_i-\mathbf{x}_i\beta\right)^2+\lambda(\zeta)\,\|\beta\|_1 \qquad \text{(LASSO)}$$

- **L**east **A**bsolute **S**hrinkage and **S**election **O**perator

Regression for different regimes

Classical statistics ($p \ll n$)

- Theory: maximum likelihood estimation
- Main issue: modeling

Non-parametric statistics ($p \approx n$)

- Prior assumption: β is smooth
- Theory: reproducing kernel Hilbert space, Gaussian process
- Main issue: kernel design, hyperparameter selection

High-dimensional statistics ($p \gg n$)

- Prior assumption: β is sparse
- Theory: lasso, compressive sensing
- Main issue: non-convexity, variable selection

Note 1

Lasso shrinks too much — almost always use the adaptive lasso

Least **A**bsolute **S**hrinkage and **S**election **O**perator

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \quad \sum_{i=1}^{n} (y_i - \mathbf{x}_i \beta)^2 + \lambda \|\beta\|_1$$

The bias problem

- The ‘best’ convex surrogate?
- ... is still quite bad

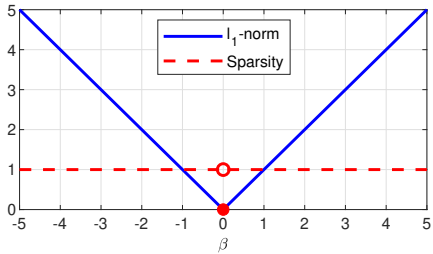
A trivial example:

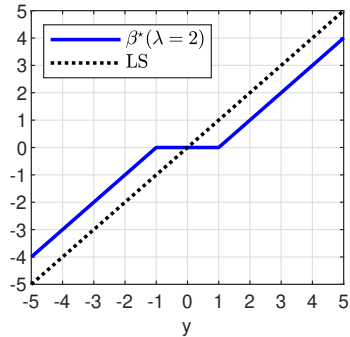
Consider identity regressor:

$$\Phi = \begin{bmatrix} \mathbf{x}_1^\top & \mathbf{x}_2^\top & \dots & \mathbf{x}_n^\top \end{bmatrix}^\top = \mathbb{I}_n,\ n = p$$

The optimal solution is soft thresholding:

$$\beta_j^\star = \mathsf{sgn}(y_j)(|y_j| - \lambda/2)_+$$





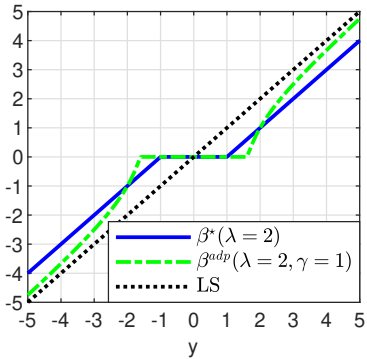
The adaptive lasso

Intuition: Penalize less for large coe's
$\rightarrow$ find large coe's from an initial estimate

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \quad \sum_{i=1}^{n} (y_i - \mathbf{x}_i \beta)^2 + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\beta_j^\star|^\gamma + \epsilon}$$

$\beta_j^\star$: initial estimate from ordinary lasso, $\gamma > 0$

- Weighted lasso with $\lambda_j = \dfrac{\lambda}{|\beta_j^\star|^\gamma + \epsilon}$



An optimization PoV

Consider non-convex problem

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}} \quad \sum_{i=1}^{n} (y_i - \mathbf{x}_i\beta)^2 + \lambda \cdot g_q(\beta)$$

with pseudo-norm

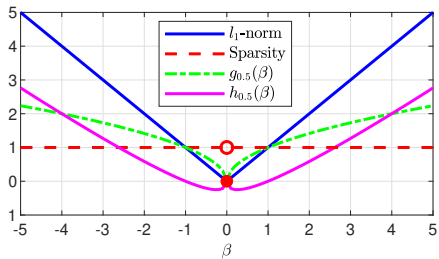
$$g_q(\beta) = \begin{cases} \sum_{j=1}^{p} |\beta_j|^q, & 0 < q < 1 \\ \sum_{j=1}^{p} \ln \dfrac{|\beta_j| + \epsilon}{\epsilon}, & q = 0 \end{cases}$$

Let $g_q(\beta) = \|\beta\|_1 - h_q(\beta)$

$$\Rightarrow \underset{\beta\in\mathbb{R}^p}{\text{minimize}} \quad J(\beta) - \lambda \cdot h_q(\beta)$$

$h_q(\beta)$: convex on $(-\infty, 0)$ and $(0, \infty)$

$\sim$ Difference of convex programming

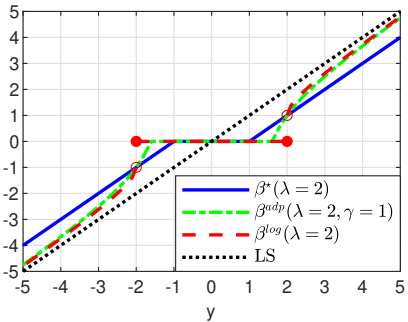


Skipping the details...

Apply the DC algorithm on the DCP (by writing $\beta = \beta_+ - \beta_-$):

$$\beta^{k+1} = \arg\min_{\beta \in \mathbb{R}^p} \quad \sum_{i=1}^{n} (y_i - \mathbf{x}_i \beta)^2 + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{\left|\beta_j^k\right|^{1-q} + \epsilon}$$

- **Adaptive lasso**: 2-step DCA initialized at $\beta^0 = \mathbf{1}_p$ with $\gamma = 1 - q$
- Converging solution is discontinuous w.r.t. $y \sim$ so not necessarily good



Note 2

Lasso can't select stably —
use subsampling when selection is desired

Least **A**bsolute **S**hrinkage and **S**election **O**perator

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \quad \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i \beta\right)^2 + \lambda \left\|\beta\right\|_1$$

Lasso theory (very informal...)

By choosing $\lambda = O(\log p)$,

- under mild conditions, prediction error $\frac{\left\|\hat{\mathbf{y}} - \mathbf{y}^0\right\|_2^2}{n} \to 0$ at rate $\frac{\log p \cdot |S|}{n}$
- ... estimation error $\left\|\hat{\beta} - \beta\right\|_1 \to 0$ at rate $\frac{\sqrt{\log p} \cdot |S|}{\sqrt{n}}$
- if non-zero β_j's are significant: $\min_{j \in S} |\beta_j| \gg \frac{\sqrt{\log p} \cdot |S|}{\sqrt{n}}$, there are no false negatives asymptotically: $\mathbb{P}\left(\hat{S} \supseteq S\right) \to 1$
- very hard to control false positives: $\hat{S} \neq S$ in general

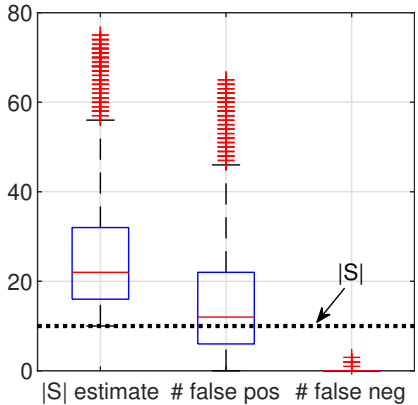
Finite-sample simulation

$$n = 80, \quad p = 1000$$

$$|S| = 10, \quad \beta = \begin{bmatrix} \mathbf{1}_{10} \\ \mathbf{0}_{990} \end{bmatrix}$$

$$x_i^j \sim \mathcal{N}(0,1), \quad \epsilon_i \sim \mathcal{N}(0, 0.1)$$

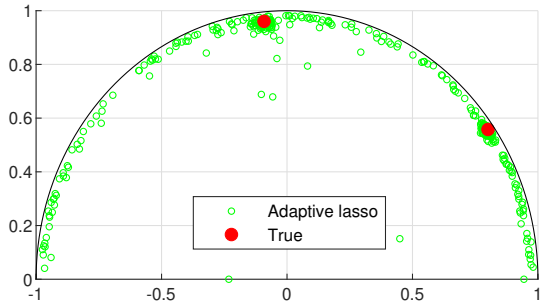
- 1000 simulations
- λ selected by cross-validation



Pole location identification example

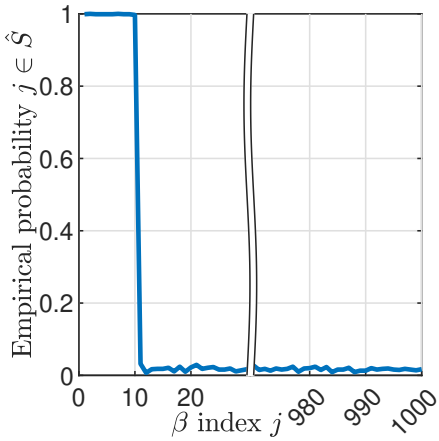
$$n = 100, \quad p = 500$$

- 4th-order systems ($|S| = 4$)
- Unit Gaussian input design
- 20 dB SNR, 100 simulations
- λ selected by cross-validation



If we have more experiments...

- Calculate the empirical probability of $j \in \hat{S}$ from 1000 experiments
- Active set is very clear from the empirical probability



Stability selection

- Generate more experiments artificially by subsampling

for $i = 1, \ldots, n_s$ **do**

Generate a random subsample with $\lfloor n/2 \rfloor$ elements $B_i \subset \{1, 2, \ldots, n\}$

Estimate active set $\hat{S}_{B_i}$ by applying adaptive lasso on subsample B_i

end for

$\hat{S} \leftarrow \left\{ k \,\middle|\, \frac{1}{n_s} \sum_{i=1}^{n_s} \left(\mathbb{1}_{\hat{S}_{B_i}}(k) \right) \geq \tau \right\}$, $\mathbb{1}$: indicator function.

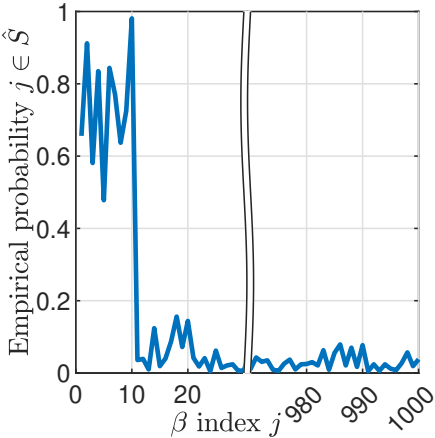
Output: $\hat{S}$

- Number of false positives V is controlled for $\tau > 0.5$

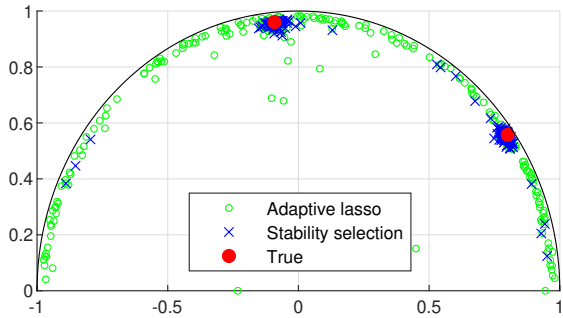
$$\mathbb{E}(V) \leq \frac{\mathbb{E}^2 \left(\left| \hat{S}_{B_i} \right| \right)}{(2\tau - 1)p}$$

Stability selection examples

$n_s = 1000, \quad \mathbb{E}\left(\left|\hat{S}_{B_i}\right|\right) = 75$



$n_s = 50, \quad \tau = 0.9, \quad$ fixed λ



Note 3

Don’t solve lasso as an optimization problem — least angle regression is more efficient and useful

Least **A**bsolute **S**hrinkage and **S**election **O**perator

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \quad \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i \beta\right)^2 + \lambda \left\|\beta\right\|_1$$

Solving lasso

- As one of the simplest non-differentiable convex problems, many algorithms derived in literature
- **Coordinate descent**

$$\hat{\beta}_j^{k+1} = \begin{cases} 0, & \left|2\phi_j^\top \left(\mathbf{y} - \Phi\alpha^j(0)\right)\right| \leq \lambda \\ \underset{\beta_j}{\text{argmin}}\ J\left(\alpha^j(\beta_j)\right), & \text{otherwise} \end{cases}, \quad \alpha_i^j(x) = \begin{cases} \hat{\beta}_i^k, & i \neq j \\ x, & i = j \end{cases}$$

- **ADMM**

$$\begin{cases} \hat{\beta}^{k+1} &= \left(\Phi^\top\Phi + \rho\mathbb{I}\right)^{-1}\left(\Phi^\top\mathbf{y} + \rho\left(z^k - u^k\right)\right) \\ z^{k+1} &= S_{\lambda/\rho}\left(\hat{\beta}^{k+1} + u^k\right) \\ u^{k+1} &= u^k + \hat{\beta}^{k+1} - z^{k+1} \end{cases}, \quad S_\kappa(\cdot)\text{: soft thresholding fun.}$$

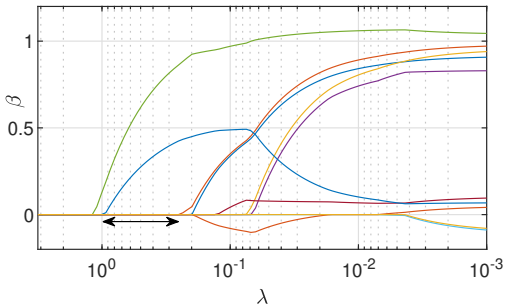
... not just ONE optimization problem

- $\hat{\beta}$ is a function of λ: $\hat{\beta} = \hat{\beta}(\lambda)$
- Theories on optimal λ are typically asymptotic with ambiguous constants

- In practice: solve lasso on a grid of λ & tune by cross-validation
- Trade-off between λ selection accuracy and computational complexity

More importantly...

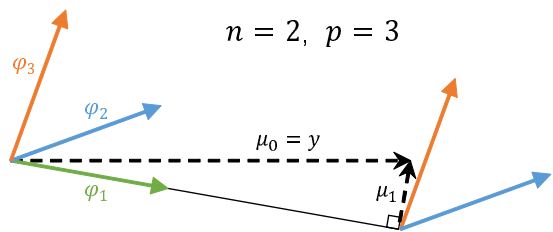
- Not all $\lambda : \hat{\beta}(\lambda)$ are useful
- Only critical points with a sparsity change are of interest
- ... intermediate points only induce unnecessary bias



One algorithm for all

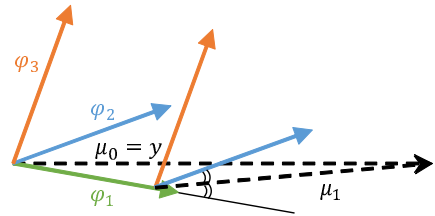
- *What if...* an algorithm automatically detects all the critical λ's and solves lasso for these λ's in one go $\rightarrow$ **Least angle regression** (LARS)
- (away from optimization) The initial idea: **forward selection**
- Iterates between 1) select ϕ_j that correlates the most with the model residual & 2) solves the least-squares problem with selected ϕ_j's

- LS coe's are often too greedy
- Select ϕ_2 instead of ϕ_3 should be more reasonable



Reducing the step size

- Stop when a new covariate correlates with the residual as much as selected covariates (graphically, equiangular)



$\hat{\mathbf{y}}^i$: prediction, S^i: active set

for $i = 0, \dots, n-1$ **do**

Correlations on the residual: $\mathbf{c} = \Phi^\top (\mathbf{y} - \hat{\mathbf{y}}^i)$

Equiangular vector: $\mathbf{u} = \Phi_{S^i} \left(\Phi_{S^i}^\top \Phi_{S^i} \right)^{-1} \mathbf{1}$, $\Phi_{S^i} = (\text{sgn}(c_j)\phi_j)_{j \in S^i}$

Next covariate: $j^+ = \underset{j \in \bar{S}^i}{\text{argmin}}{}^+ \dfrac{\max(|\mathbf{c}|) \pm c_j}{1 \pm a_j}$, $\mathbf{a} = \Phi^\top \mathbf{u}$, η: minimum value

$S^{i+1} = S^i \cup \{j^+\}$, $\hat{\mathbf{y}}^{i+1} = \hat{\mathbf{y}}^i + \eta \mathbf{u}$

end for

Magically close to lasso

- The LARS solution path almost gives all critical lasso solutions $\left\{ \hat{\beta}(\lambda) \,\middle|\, \text{sparsity changes at } \lambda \right\}$
- The only modification: anytime coefficients change sign, remove it from the active set

Remarks:

- The whole LARS-lasso algorithm up to $\left|\hat{S}\right| = m$ is $O(m^3 + nm^2)$, as fast as least-squares on $\Phi \in \mathbb{R}^{n \times m}$
- Do we need the lasso modification? $\left|\hat{S}\right|$ is not monotonic along the regularization path
- The critical λ-values not obtained (do we need them?)
- Trivial extension to adaptive lasso (scaling ϕ_j); harder to extend to group lasso

- Lasso shrinks too much — almost always use the adaptive lasso
- Lasso can't select stably — use subsampling when selection is desired
- Don't solve lasso as an optimization problem — least angle regression is more efficient and useful

References:

[1] Bühlmann, P., & Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer.

[2] Gasso, G., Rakotomamonjy, A., & Canu, S. (2009). Recovering sparse signals with a certain family of nonconvex penalties and DC programming. *IEEE Transactions on Signal Processing*, 57(12), 4686-4698.

[3] Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2), 407-451.