Fast and Accurate ℓ_1 Penalized Estimators

Andrew Barron and Cong Huang

YALE UNIVERSITY, DEPARTMENT OF STATISTICS

Presentation at Rutgers University, December 12, 2007

Outline

- Penalization for Least Squares and Log Likelihood Criteria
- The ℓ_1 Penalized Greedy Pursuit (LPGP) algorithm
  - Description of the algorithm
  - Analysis of its performance
- Advantages and Disadvantages of *LPGP*
- Key Ideas of the Proof
- Risk Characterization
  - What forms of penalty permit desirable risk bounds?
- Conclusion

ℓ_1 Penalized Least Squares

- Suppose the data are $(X_i, Y_i)_{i=1}^n$ and a library $\mathcal{H} = \{h\}$ is given. Find a function in the linear span of $\mathcal{H}$ to minimize the following objective function.

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \sum_h \beta_h h(X_i))^2 + \lambda \sum_h |\beta_h|$$

- This optimization is also called the Lasso (Tibshironi 1996) and Basis Pursuit (Chen and Donoho 1996).

ℓ_1 Penalized Log Likelihood

- For an exponential family with statistics taken from a given library $\mathcal{H}$ of functions of the data $\underline{X}$.
- Find the parameters β with which these statistics are to be linearly combined to optimize the objective function

$$-\log \text{likelihood}(\beta) + \lambda \sum_h |\beta_h|$$

- Investigated in various special cases in Park and Hasti (2007), Koh, Kim and Boyd (2007), Banerjee, Ghaoui and dAspermont (2007), and Friedman, Hastie and Tibshirani (2007).

ℓ_1 Penalized Greedy Pursuit (*LPGP*)

First suppose the library $\mathcal{H}$ is normalized in that $\|h\| = 1$ for all $h \in \mathcal{H}$.

- Algorithm

  Initialize $\hat{f}_0 = 0$.

  Then for $m = 1, 2, \ldots$, iteratively, given $\hat{f}_{m-1} = \sum_{j=1}^{m-1} \beta_{j,m-1} h_j$, we seek

  $$\hat{f}_m(x) = \alpha \hat{f}_{m-1}(x) + \beta h(x)$$

  to minimize the objective function over choices of h, α, β,

  $$\frac{1}{n} \sum_{i=1}^{n} \big(Y_i - \alpha f_{m-1}(X_i) - \beta h(X_i)\big)^2 + \lambda\big(|\beta| + \alpha \sum_{j=1}^{m-1} |\beta_{j,m-1}|\big)$$

  yielding $h_m, \alpha_m, \beta_{m,m}$ and $\beta_{j,m} = \alpha_m \beta_{j,m-1}$ for $j = 1, 2, \ldots, m-1$.

ℓ_1 Penalized Greedy Pursuit (*LPGP*)

First suppose the library $\mathcal{H}$ is normalized in that $\|h\| = 1$ for all $h \in \mathcal{H}$.

- Algorithm

  Initialize $\hat{f}_0 = 0$.

  Then for $m = 1, 2, \ldots$, iteratively, given $\hat{f}_{m-1} = \sum_{j=1}^{m-1} \beta_{j,m-1} h_j$, we seek

$$\hat{f}_m(x) = \alpha \hat{f}_{m-1}(x) + \beta h(x)$$

  to minimize the objective function over choices of h, α, β.

$$\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \alpha f_{m-1}(X_i) - \beta h(X_i)\big)^2 + \lambda\Big(|\beta| + \alpha \sum_{j=1}^{m-1} |\beta_{j,m-1}|\Big)$$

  yielding $h_m, \alpha_m, \beta_{m,m}$ and $\beta_{j,m} = \alpha_m \beta_{j,m-1}$ for $j = 1, 2, \ldots, m-1$.

- Key Conclusion

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_m(X_i))^2 + \lambda \sum_{j=1}^{m} |\beta_{j,m}| \le \inf_{\underline{\beta}} \left\{ \frac{1}{n}\sum_{i=1}^{n}(Y_i - f_{\underline{\beta}}(X_i))^2 + \lambda \sum_h |\beta_{f,h}| + \frac{4(\sum_h |\beta_{f,h}|)^2}{m+1} \right\},$$

where $f_{\underline{\beta}} = \sum_h \beta_{f,h} h$.

Advantages and Disadvantages of *LPGP*

- Let p =Card($\mathcal{H}$), typically much larger than the data size n.
- As we shall see, the number of steps m for statistical accurate fit is typically much less than n.
- Advantages

  Computation of ℓ_1 penalized solution with explicit guarantee of accuracy

  Time cost pnm v.s. pn^2 for an alternative strategies (LARS)
- Disadvantages

  Not an exact solution

  Algorithm basically is for the case of fixed λ.

Key Ideas of the Proof

- Assume $\mathcal{H}$ is closed under sign-change (otherwise replaced by $\mathcal{H} \cup -\mathcal{H}$), so the coefficients of linear combination are kept non-negative.
- Denote $v_m = \sum_{j=1}^{m} \beta_{j,m}$ and $v = \sum_h \beta_{f,h}$ for a particular $f_{\underline{\beta}} = \sum_h \beta_{f,h} h_f$. Let

$$e_m^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_m(X_i))^2 - \frac{1}{n}\sum_{i=1}^{n}(Y_i - f_{\underline{\beta}}(X_i))^2 + \lambda v_m.$$

Key Ideas of the Proof

- Assume $\mathcal{H}$ is closed under sign-change (otherwise replaced by $\mathcal{H} \cup -\mathcal{H}$), so the coefficients of linear combination are kept non-negative.
- Denote $v_m = \sum_{j=1}^{m} \beta_{j,m}$ and $v = \sum_h \beta_{f,h}$ for a particular $f_{\underline{\beta}} = \sum_h \beta_{f,h} h_f$. Let
$$e_m^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_m(X_i))^2 - \frac{1}{n}\sum_{i=1}^{n}(Y_i - f_{\underline{\beta}}(X_i))^2 + \lambda v_m.$$
- By the choice of α_m, $\beta_{m,m}$ and h_m, the value is at least as good as if we use $\alpha = 1 - \frac{2}{m+1}$ and $\beta = \bar{\alpha} v$, we have
$$e_m^2 \le \frac{1}{n}\sum_{i=1}^{n}(Y_i - \alpha \hat{f}_{m-1} - \bar{\alpha} v h(X_i))^2 - \frac{1}{n}\sum_{i=1}^{n}(Y_i - f_{\underline{\beta}}(X_i))^2 + \lambda\big[\alpha v_{m-1} + \bar{\alpha} v\big],$$
where $\bar{\alpha} = 1 - \alpha$.

Key Ideas of the Proof

- Assume $\mathcal{H}$ is closed under sign-change (otherwise replaced by $\mathcal{H} \cup -\mathcal{H}$), so the coefficients of linear combination are kept non-negative.
- Denote $v_m = \sum_{j=1}^{m} \beta_{j,m}$ and $v = \sum_h \beta_{f,h}$ for a particular $f_{\underline{\beta}} = \sum_h \beta_{f,h} h_f$. Let

$$e_m^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_m(X_i))^2 - \frac{1}{n}\sum_{i=1}^{n}(Y_i - f_{\underline{\beta}}(X_i))^2 + \lambda v_m.$$

- By the choice of α_m, $\beta_{m,m}$ and h_m, the value is at least as good as if we use $\alpha = 1 - \frac{2}{m+1}$ and $\beta = \bar{\alpha} v$, we have

$$e_m^2 \leq \frac{1}{n}\sum_{i=1}^{n}(Y_i - \alpha \hat{f}_{m-1} - \bar{\alpha} v h(X_i))^2 - \frac{1}{n}\sum_{i=1}^{n}(Y_i - f_{\underline{\beta}}(X_i))^2 + \lambda\big[\alpha v_{m-1} + \bar{\alpha} v\big],$$

where $\bar{\alpha} = 1 - \alpha$. We may rearrange it as

$$\begin{aligned} e_m^2 \ \leq \ & \alpha e_{m-1}^2 + \bar{\alpha}^2 b_h + \bar{\alpha}\lambda v \\ & -\frac{2\alpha\bar{\alpha}}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_{m-1}(X_i))(v h(X_i) - f_{\underline{\beta}}(X_i)) \\ & -\frac{\alpha\bar{\alpha}}{n}\sum_{i=1}^{n}(\hat{f}_{m-1}(X_i) - f_{\underline{\beta}}(X_i))^2, \end{aligned}$$

where $b_h = \frac{1}{n}\sum_{i=1}^{n}(Y_i - v h(X_i))^2 - \frac{1}{n}\sum_{i=1}^{n}(Y_i - f_{\underline{\beta}}(X_i))^2$.

Key Ideas of the Proof

- Since the inequality holds for all h, this e_m^2 is less than the average of the right side for any convenient distribution on the choices of h. We consider the distribution that h is chosen to be h_f with probability $\frac{\beta_{f,h}}{v}$ so that the expectation of $vh(x)$ is $f_{\underline{\beta}}(x)$.
- Then $(Y_i - \hat{f}_{m-1}(X_i))(vh(X_i) - f_{\underline{\beta}}(X_i))$ has expectation 0 and b_h has expectation not more than v^2. Thus
$$e_m^2 \leq \alpha e_{m-1}^2 + \bar{\alpha}^2 v^2 + \lambda \bar{\alpha} v,$$
where $\bar{\alpha} = \frac{2}{m+1}$.

Key Ideas of the Proof

- Since the inequality holds for all h, e_m^2 is less than the average of the right side for any convenient distribution on the choices of h. We consider the distribution that h is chosen to be h_f with probability $\frac{\beta_{f,h}}{v}$ so that the expectation of $vh(x)$ is $f_{\underline{\beta}}(x)$.

- Then $(Y_i - \hat{f}_{m-1}(X_i))(vh(X_i) - f_{\underline{\beta}}(X_i))$ has expectation 0 and b_h has expectation not more than v^2. Thus
$$e_m^2 \le \alpha e_{m-1}^2 + \bar{\alpha}^2 v^2 + \lambda \bar{\alpha} v,$$
where $\bar{\alpha} = \frac{2}{m+1}$.

- Initially $e_0^2 \le v^2 + \lambda v$. By induction assuming that $e_{m-1}^2 \le \frac{4v^2}{m} + \lambda v$, we establish that
$$e_m^2 \le \frac{4v^2}{m+1} + \lambda v.$$

- Thus
$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_m(X_i))^2 + \lambda v_m \le \frac{1}{n}\sum_{i=1}^{n}(Y_i - f_{\underline{\beta}}(X_i))^2 + \lambda v + \frac{4v^2}{m+1}.$$

Results Re-expressed for Un-normalized $\mathcal{H}$

Drop the normalization condition.

- Algorithm

  Initialize $\hat{f}_0 = 0$.

  Then for $m = 1, 2, \ldots$, iteratively, given the terms of $\hat{f}_{m-1} = \sum_{j=1}^{m-1} \beta_{j,m-1} h_j$, we seek $\hat{f}_m = \alpha \hat{f}_{m-1} + \beta h$ to minimize the objective function over choices of h, α, β.

$$\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \alpha f_{m-1}(X_i) - \beta h(X_i)\big)^2 + \lambda\big(|\beta|\|h\| + \alpha\sum_{j=1}^{m-1}|\beta_{j,m-1}|\|h_j\|\big)$$

  yielding $h_m, \alpha_m, \beta_{m,m}$ and $\beta_{j,m} = \alpha_m \beta_{j,m-1}$ for $j = 1, 2, \ldots, m-1$.

- Key Conclusion

  Let $V(f) = \|f\|_{1,\mathcal{H}} = \inf\{\sum_h |\beta_{f,h}|\|h\| : f = \sum_h \beta_{f,h} h\}$. Thus,

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_m(X_i))^2 + \lambda\sum_{j=1}^{m}|\beta_{j,m}|\|h_j\| \le \inf_f\left\{\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2 + \lambda V(f) + \frac{4V^2(f)}{m+1}\right\}.$$

Risk Bounds of the Estimators obtained from *LPGP*

- Suppose $(X_i, Y_i)_{i=1}^n$ are independently drawn from the distribution of (X, Y). The target regression function $f^*(x) = E[Y|X = x]$ is unknown and is to be estimated. The error $\epsilon = Y - f^*(X)$ is assumed to have a conditional distribution given X which satisfies certain moment conditions.
- We work with the set $\mathcal{F}$, the linear span of library $\mathcal{H}$.
- Suppose $\{\hat{f}_m, m = 1, 2, \ldots\}$ is the sequence of estimators formulated from the *LPGP* algorithm.
- Measure of loss is the generalization error for $\mu = P_x$,

$$\|f - f^*\|^2 = \int (f(x) - f^*(x))^2 \mu(dx).$$

Risk Bounds of the Estimators obtained from *LPGP*

- Risk bounds for ℓ_1 penalization

  If $\lambda_n > B\sqrt{\frac{\log p}{n}}$, we may run *LPGP* for many steps to reach an approximation of the Lasso solution $\hat{f}$. It has the following risk bound.

$$\begin{aligned} & \mathbb{E}\|\hat{f} - f^*\|^2 \\ \le\ & (1+\delta) \inf_{f \in \mathcal{F}} \left\{ \|f - f^*\|^2 + \lambda_n V(f) \right\} + \frac{C_\delta}{n}. \end{aligned}$$

Risk Bounds of the Estimators obtained from *LPGP*

- Risk bounds for ℓ_1 penalization

  If $\lambda_n > B\sqrt{\frac{\log p}{n}}$, we may run *LPGP* for many steps to reach an approximation of the Lasso solution $\hat{f}$. It has the following risk bound.

$$\begin{aligned} & \mathbb{E}\|\hat{f} - f^*\|^2 \\ \leq\ & (1+\delta) \inf_{f \in \mathcal{F}} \left\{ \|f - f^*\|^2 + \lambda_n V(f) \right\} + \frac{C_\delta}{n}. \end{aligned}$$

- Risk bounds for model selection

  If λ_n is chosen much smaller (e.g. of the order of $1/n$). We choose $\hat{m}$ to minimize the penalized least squares

$$\frac{1}{n}\sum_{i=1}^{m}(Y_i - \hat{f}_m(X_i))^2 + \lambda_n(\sum_{j=1}^{n}|\beta_{j,m}|\|h_j\|_n) + \frac{\gamma m \log p}{n},$$

where γ is a constant. Then the risk of the estimator $\hat{f}_{\hat{m}}$ is bounded by

$$\begin{aligned} & \mathbb{E}\|\hat{f}_{\hat{m}} - f^*\|^2 \\ \leq\ & (1+\delta) \inf_m \inf_{f \in \mathcal{F}} \left\{ \|f - f^*\|^2 + \lambda_n V(f) + \frac{4V^2(f)}{m} + \frac{\gamma m \log p}{n} \right\} + \frac{C_\delta}{n} \\ \leq\ & (1+\delta) \inf_{f \in \mathcal{F}} \left\{ \|f - f^*\|^2 + \lambda_n' V(f) \right\} + \frac{C_\delta}{n}, \end{aligned}$$

where $\lambda_n' = \lambda_n + B_1\sqrt{\frac{\log p}{n}}$.

Conclusion

- Subset selection procedures may be used in ℓ_1-penalized least squares optimization.
- m-term chosen by relaxed greedy pursuit or by ℓ_1-penalized greedy pursuit provides accuracy within order $V^2(f)/m$ of the minimal objective function.
- Ultimate penalty is

$$\min\left\{\lambda_n V(f), \frac{m\log p}{n}\right\}$$

- Risk of the estimate is captured by the ideal tradeoff between $\|f - f^*\|^2$ and the penalty.