ℓ_1 Penalized Likelihood: Fast Algorithms and Risk Bounds

Xi (Rossi) Luo

(with Andrew Barron)
Department of Statistics
Yale University

Innovation and Inventiveness in Statistics Methodologies
In Honor of John Hartigan
Yale University
May 16, 2009

Problem

- Data $X_1, \ldots, X_n$ be *i.i.d.* in $\mathbb{R}^p$ distributed as

$$p_f(x) = \frac{e^{f(x)} p_0(x)}{C_f}$$

where reference $p_0(x)$ known and we are interested in estimating $f(x)$.

- Consider estimator $\hat{f}(x) = f_{\hat{\beta}}(x) = \sum_{h \in \mathcal{H}} \hat{\beta}_h h(x)$ that minimizes

$$\frac{1}{n} \sum_{i=1}^{n} \log(1/p_{f_\beta}(X_i)) + \lambda \sum_h |\beta_h|.$$

- Special case $\mathcal{H} = \{x_1, \ldots, x_p\} \bigcup \{x_1 x_2, x_1 x_3, \ldots, x_{p-1} x_p\}$ and may also use polynomials, trigonometric terms, splines, and wavelets.

ℓ_1 penalty is risk valid for λ_n of order $1/\sqrt{n}$

- Log-density estimator: $p_{f_\beta}(x) = e^{f_\beta(x)} p_0(x)/C_{f_\beta}$

Theorem

The ℓ_1 penalized likelihood estimator $\hat{f}(x) = f_{\hat{\beta}}(x) = \sum_{h \in \mathcal{H}} \hat{\beta}_h h(x)$ achieving

$$\min_{\beta} \{ \frac{1}{n} \log \frac{1}{p_{f_\beta}(\underline{x})} + \lambda_n ||\beta||_1 \}$$

has the following risk bound

$$\mathbb{E} d(p_{f^*}, p_{f_{\hat{\beta}}}) \le \inf_{\beta} \left\{ \underbrace{D(p_{f^*} || p_{f_\beta})}_{approximation} + \underbrace{\lambda_n ||\beta||_1}_{complexity} \right\} + \frac{4 \log(2M)}{n}$$

for every sample size provided that $\lambda_n \ge \sqrt{\frac{2 \log(2M)}{n}}$, where $M = Card(\mathcal{H})$ $(= p)$.

Adaptive ℓ_1 Penalized Regression and Risk Bounds

- Regression model: $Y = f^*(X) + \sigma N(0,1)$

Theorem

ℓ_1 penalized least squares estimator $\hat{f}(x) = f_{\hat{\beta}}(x) = \sum_{h \in \mathcal{H}} \hat{\beta}_h h(x)$ achieving

$$\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f_\beta(x_i))^2 + 2\sigma\lambda_n ||\beta||_1 \right\}$$

has the following risk bounds

$$\mathbb{E}||f^* - f_{\hat{\beta}}||_n^2 \leq 2 \underbrace{\inf_{\beta} \left\{ ||f^* - f_\beta||_n^2 + 2\sigma\lambda_n ||\beta||_1 \right\}}_{\leq 0 + 2\sigma\lambda_n ||\beta^*||_1 \quad \textit{replacing } \beta = \beta^* \textit{ if } f^* = f_{\beta^*}} + \frac{8\sigma^2 \log(2M)}{n}$$

for every sample size provided that $\lambda_n \geq \sqrt{\frac{2\log(2M)}{n}}$.

- Estimate unknown $\sigma = \frac{1}{2}\lambda_n ||\beta||_1 + \sqrt{\left[\frac{1}{2}\lambda_n ||\beta||_1\right]^2 + \frac{1}{n}\sum_{i=1}^{n}(Y_i - f_\beta(x_i))^2}$ and similar result holds (Proc. WITMSE '08, Luo with Barron).

Thank You!

Relaxed Greedy Pursuit

- Initialize with $\hat{f}^{(0)}(x) = 0$. Given $\hat{f}^{(k-1)}(x)$, iteratively set

$$\hat{f}^{(k)}(x) = \alpha \hat{f}^{(k-1)}(x) + \gamma h(x)$$

with $\alpha = \alpha^{(k)}$, $\gamma = \gamma^{(k)}$ and $h = h^{(k)}$ chosen by

$$\arg\min_{\alpha,\,\gamma,h} \left\{ L(\alpha \hat{f}^{(k-1)} + \gamma h) + \lambda[\alpha v^{(k-1)} + |\gamma|] \right\}$$

where $L(f) = \frac{1}{n}\sum_{i=1}^{n} \log(1/p_f(X_i))$, $v^{(k-1)} = \sum_{j=1}^{M} |\beta_j^{(k-1)}|$ for $\hat{f}^{(k-1)} = \sum_{j=1}^{M} \beta_j^{(k-1)} h_j$, and $M = Card(\mathcal{H})$.

- Iterate until desired accuracy.

Computational Accuracy

Suppose $\|h(x)\|_\infty \le C$ for all $h(x) \in \mathcal{H}$.

Theorem

The k step RGP estimator $\hat{f}^{(k)}(x) = \sum_{j=1}^{M} \beta_j^{(k)} h_j(x)$ has the following computational accuracy bound valid for all X, $\lambda \ge 0$

$$\frac{1}{n}\sum_{i=1}^{n} \log(1/p_{\hat{f}^{(k)}}(X_i)) + \lambda||\beta^{(k)}||_1$$

$$\le \inf_{f_\beta} \left\{ \frac{1}{n}\sum_{i=1}^{n} \log(1/p_{f_\beta}(X_i)) + \lambda||\beta||_1 + \frac{2C^2||\beta||_1^2}{k+1} \right\}$$

where $||\beta||_1 = \sum_{j=1}^{M} |\beta_j|$ *and* $f_\beta = \sum_{j=1}^{M} \beta_j h_j$.

Similar conclusion for unbounded multivariate Gaussians as arise in Gaussian inverse covariance matrix estimation.