Cor 26

Suppose the rows of X are mean-zero iid multivariate normal with covariance matrix Σ^0. Suppose diagonal entries of Σ^0 are bounded and its min e-val c_{min} is bounded away from 0. Then $\mathbb{P}(\phi^2_{\hat{S}} \geq c_{min}/2) \to 1$ provided $s\sqrt{\log p/n} \to 0$.

2.2.7 Computation

Coordinate descent gives a way of minimizing a function $f: \mathbb{R}^d \to \mathbb{R}$ of the form

$$f(x) = g(x) + \sum_{j=1}^d h_j(x_j)$$

where g is convex and differentiable

$h_j: \mathbb{R} \to \mathbb{R}$ is convex.

We start with an initial guess of the minimizer $x^{(0)}$ ($x^{(0)} = 0$) and repeat for $m = 1, 2, \ldots$

$$x_1^{(m)} = \arg\min_{x_1 \in \mathbb{R}} f(x_1, x_2^{(m-1)}, \ldots, x_d^{(m-1)})$$

$$x_2^{(m)} = \arg\min_{x_2 \in \mathbb{R}} f(x_1^{(m)}, x_2, x_3^{(m-1)}, \ldots, x_d^{(m-1)})$$

$$\vdots$$

$$x_d^{(m)} = \arg\min_{x_3 \in \mathbb{R}} f(x_1^{(m)}, \ldots, x_{d-1}^{(m)}, x_d)$$

Tseng (2001) proves that provided $A_0 = \{x : f(x) \leq f(x^{(0)})\}$ is a compact set, every converging sequence of the $x^{(m)}$ converges to a minimizer.

Corollary 27 Suppose A_0 compact. Then

i) there exists a minimizer x^* of f, and $f(x^{(m)}) \to f(x^*)$

ii) if x^* is the unique minimizer then $x^{(m)} \to x^*$.

Can replace individual coordinates by blocks of coordinates and the same result holds.

If $x = (x_1, \ldots, x_B)$, $x_j \in \mathbb{R}^{d_j}$, a blockwise coordinate descent can minimize functions of the form

$$f(x) = g(x) + \sum_{j=1}^B h_j(x_j)$$

g: convex, differentiable; $h_j: \mathbb{R}^{d_j} \to \mathbb{R}$ convex

We typically solve the Lasso at a grid of λ values $\lambda_0 > \lambda_1 > \ldots > \lambda_L$. It is helpful to first solve at $\lambda = \lambda_0$ and then use the soln at $\lambda_{\ell-1}$ to initialize minimization of the objective corresponding to λ_ℓ (warm start).

An active set strategy improves speed greatly.

For $l = 1, \dots, L$

1. Initialize $A = \{k : \hat{\beta}_{\lambda_{l-1}, k} \neq 0\}$
2. Perform coordinate descent only on coordinates in A to obtain value $\hat{\beta}$ ($\hat{\beta}_{A^c} = 0$)
3. Set $V = \{k : \frac{1}{n} |X_k^T (Y - X\hat{\beta})| > \lambda_l\}$
4. If V is empty, set $\hat{\beta}^l_{\lambda_L} = \hat{\beta}$, else update $A = A \cup V$ and go to 2.

2.3 Extensions

2.3.1 Structural penalties

Different penalty functions can be used to encourage different types of sparsity. Given a partition $G_1 \cup \dots \cup G_q = \{1, \dots, p\}$, the group lasso uses the penalty

$$\lambda \sum_{j=1}^{q} m_j \|\beta_{G_j}\|_2 .$$

Typically $m_j = \sqrt{|G_j|}$ to balance groups of different sizes. The penalty encourages an entire group G to have $\beta_G = 0$ or $\beta_k \neq 0 \ \forall k \in G$.
This is useful in e.g. when groups occur through coding for categorical predictors.

Reducing bias

Although the Lasso shrinkage gives sparse estimates and reduces variance, it also introduces (unwanted) bias. Various modifications have been proposed.
LARS-OLS hybrid : overestimate β^0_{λ} by OLS regression
Relaxed Lasso : Lasso regression
Adaptive Lasso : given an initial estimate $\hat{\beta}$ e.g. from the Lasso, solve

$$\hat{\beta}^{adapt}_{\lambda} = \underset{\beta : \mathbb{R}^p, \beta_{\hat{S}^c} = 0}{\arg\min} \left\{ \frac{1}{2n} \|Y - X\beta\|_2^2 + \lambda \sum_{k \in \hat{S}} \frac{|\beta_k|}{|\hat{\beta}_k|} \right\}$$

where $\hat{S} = \{k : \hat{\beta}_k \neq 0\}$

~~Non-convex penalty~~

~~Take a ...~~

EoSL 625-627