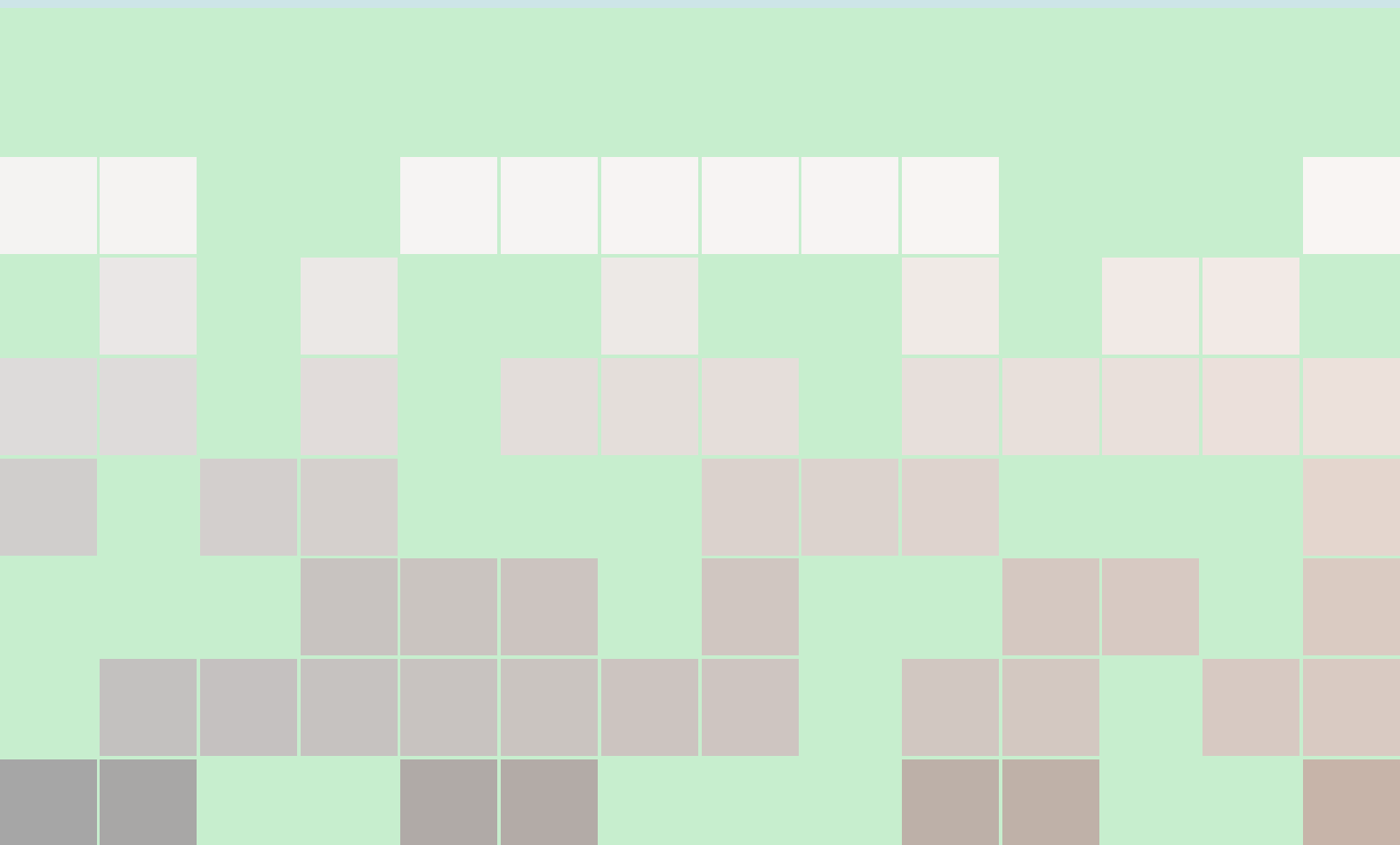



High Dimensional Statistic Papers

Xin Wen





Contents

I

Part One

1	Basic concept	9
2	Variable Selection	11
2.1	介绍各种方法	23
2.1.1	Non-Neg 1995	24
2.1.2	Lasso	25
2.1.3	Relaxed Lasso	28
2.1.4	Adaptive Lasso	29
2.1.5	Group Lasso	30
2.1.6	CAP	34
2.1.7	Sparsity group lasso	34
2.1.8	Overlap group lasso	35
2.1.9	岭回归	36
2.1.10	主成分回归	38
2.1.11	偏最小二乘方法	39
2.1.12	Bridge Estimator	40
2.1.13	Elastic net--未知分组的群组模型选择方法	41
2.1.14	适应性的 elastic net 方法及其大样本性质	45
2.1.15	Mnet	47
2.1.16	Fused Lasso	47

2.1.17	Graph Lasso	53
2.2	介绍求解 LASSO 的算法	53
2.2.1	LARS 算法	54
2.2.2	Coordinate Descent 算法	56
2.2.3	SCAD	59
2.2.4	SIS	63
2.2.5	ISIS	63
2.2.6	HOLP	64
2.2.7	Dantzing selector	67
2.2.8	ADS	82
2.2.9	DASSO	83
2.2.10	Group Bridge	84
2.2.11	Group MCP	84
2.2.12	MCP	85
2.3	Lasso 惩罚的性质	86
3	Multiple and Nonparametric Regression	93
3.0.1	The Gauss-Markov Theorem	94
3.0.2	Weighted Least-Squares	98
3.0.3	Box-Cox Transformation	100
3.0.4	Model Building and Basis Expansions	101
3.0.5	Polynomial Regression	102
3.0.6	Spline Regression	102
3.0.7	Multiple Covariates	106
3.0.8	Ridge Regression	107
3.0.9	ℓ_2 Penalized Least Squares	108
3.0.10	Bayesian Interpretation	109
3.0.11	Ridge Regression Solution Path	110
3.0.12	Regression in Reproducing Kernel Hilbert Space	113
3.1	Introduce to penalized least-squares	120
3.1.1	Classical Variable Selection Criteria	120
3.1.2	Regularization parameters for PLS	159
3.1.3	Degrees of freedom	159
3.1.4	Extensions to Nonparametric Modeling	166
4	Penalized Least Squares: Properties	173
4.0.1	Performance Benchmarks	173

5	论文综述	183
6	NONCONCAVE PENALIZED LIKELIHOOD WITH A DIVERGING NUMBER OF PARAMETERS	185
6.1	Properties of penalized likelihood estimation	185
6.1.1	Regularity conditions	185
6.1.2	Oracle properties	187
6.1.3	Estimation of covariance matrix	189
6.1.4	Likelihood ratio test	190
6.2	Proofs of theorems.	191
7	On model consistency of Lasso	205
7.1	Model Selection Consistency and Irrepresentable Conditions	205
7.1.1	Model Selection Consistency for Small q and p	208
8	A selection overview of variable selection	213

II

Part Four

Bibliography	217
Articles	217
Books	217

Part One

1	Basic concept	9
2	Variable Selection	11
2.1	介绍各种方法	
2.2	介绍求解 LASSO 的算法	
2.3	Lasso 惩罚的性质	
3	Multiple and Nonparametric Regression	93
3.1	Introduce to penalized least-squares	
4	Penalized Least Squares: Properties	173
5	论文综述	183
6	NONCONCAVE PENALIZED LIKELIHOOD WITH A DIVERGING NUMBER OF PARAMETERS	185
6.1	Properties of penalized likelihood estimation	
6.2	Proofs of theorems.	
7	On model consistency of Lasso	205
7.1	Model Selection Consistency and Irrepresentable Conditions	
8	A selection overview of variable selection	213



1. Basic concept



2. Variable Selection

共线性

不妨先假设自变量 x_1, x_2 和因变量 y 都是中心化的，则建立的回归模型为

$$\hat{Y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

令

$$L_{11} = \sum_{i=1}^n x_{i1}^2, \quad L_{12} = \sum_{i=1}^n x_{i1} x_{i2}, \quad L_{22} = \sum_{i=1}^n x_{i2}^2$$

那么可以计算出 x_1 与 x_2 的相关系数 r_{12} 如下

$$r_{12} = \frac{L_{12}}{\sqrt{L_{11}L_{22}}}$$

求得协方差阵为

$$\text{Cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

进而

$$X^T X = \begin{bmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{bmatrix}$$

$$\begin{aligned} (X^T X)^{-1} &= \frac{1}{|X^T X|} \begin{bmatrix} L_{22} & -L_{12} \\ -L_{12} & L_{11} \end{bmatrix} \\ &= \frac{1}{L_{11}L_{22} - L_{12}^2} \begin{bmatrix} L_{22} & -L_{12} \\ L_{12} & L_{11} \end{bmatrix} \\ &= \frac{1}{L_{11}L_{22}(1 - r_{12}^2)} \begin{bmatrix} L_{22} & -L_{12} \\ -L_{12} & L_{11} \end{bmatrix} \end{aligned}$$

最后得到方差

$$\begin{aligned}\text{Var}(\hat{\beta}_1) &= \frac{\sigma^2}{(1 - r_{12}^2) L_{11}} \\ \text{Var}(\hat{\beta}_2) &= \frac{\sigma^2}{(1 - r_{12}^2) L_{22}}\end{aligned}$$

由结果可以看出, $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的方差与 x_1 和 x_2 的相关系数有关, 当 $|r_{12}|$ 增大的时 $\text{Var}(\hat{\beta}_1)$ 与 $\text{Var}(\hat{\beta}_2)$ 也会增大, 而当 $|r_{12}|$ 趋于 1 时, $\text{Var}(\hat{\beta}_1)$ 与 $\text{Var}(\hat{\beta}_2)$ 将趋于无穷大。

多重共线性的度量

首先介绍方差膨胀因子 (Variance Inflation Factor, VIF) 与容忍度 (Tolerance), 不妨假定样本数据是经过中心化和标准化的, 则有

$$X^T X = (r_{ij})$$

令

$$M = (m_{ij}) = (X^T X)^{-1}$$

则记 x_j 的方差膨胀因子为

$$\text{VIF}_j = m_j$$

我们知道最小二乘法得到的估计参数满足

$$\text{Var}(\hat{\beta}^0) = \sigma^2 (X^T X)^{-1}$$

所以有

$$\text{Var}(\hat{\beta}_j) = m_{jj} \sigma^2 / L_{jj}, \quad j = 1, 2, \dots, p$$

其中 L_{jj} 是 x_j 的离差平方和, 所以我们可以看出方差膨胀因子与回归参数估计量的方差是成正比的 [41,42]。

用 R_j^2 表示第 j 个变量在其他变量上回归时的拟合优度, 则容忍度可以表示为

$$\text{TOL}_j = 1 - R_j^2$$

进而通过证明可以得到

$$\text{VIF}_j = m_{jj} = \frac{1}{1 - R_j^2} = \frac{1}{\text{TOL}_j}$$

一般我们认为, 容忍度太小 (比如小于 0.2 或 0.1) 或者方差膨胀因子太大 (比如大于 5 或 10) 则判定有多重共线性问题。然后介绍条件数 (Condition Number), 记 $\lambda_1, \lambda_2, \dots, \lambda_p$ 为 $X^T X$ 的特征值, 其中 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$, 则定义条件数为

$$\kappa = \sqrt{\frac{\lambda_1}{\lambda_p}}$$

当自变量矩阵正交时, 条件数 κ 为 1。一般我们认为当 $\kappa > 15$ 的时候存在共线性问题, 而当 $\kappa > 30$ 的时候说明共线性问题比较严重 [43]。当然, 无论是方差膨胀因子法还是条件数判别法在不同情况下并没有一个确定的判断准则, 所以可能会不是很准确, 但不失为一个参考。

Proposition 2.0.1 — OLS 的性质. 对于一般线性回归方程, $Y = X\beta + \epsilon$, 记 β 的最小二乘估计量 (OLS) 为 $\hat{\beta}^0 = (X^T X)^{-1} X^T Y$, 最小二乘法的优点是简单易得, 无偏, \sqrt{n} 一致, 缺点也很明显:

1. 如果设计矩阵 X 不是列满秩, 则 $\hat{\beta}^0$ 不唯一;
2. 当自变量之间存在高度相关性时, $Var(\hat{\beta}^0) = (X^T X)^{-1} \sigma^2$ 会增大, 从而导致 $MSE(\hat{\beta}^0)$ 增加; 更深入的讨论可见

Proposition 2.0.2 当 $q = 4, 2, 1, 0.5, 0.1$ 时, 考虑约束 $\sum_{j=1}^p \|\beta_j\|_q$ 的可行域; 随着 q 的减少, 可行域逐渐缩小, 当 $q < 1$ 时, 可行域由凸变成非凸: 因此 $q = 1$ (Lasso) 是保证可行域为凸的最小 q 值: 而可行域为凸, 凸优化理论可以保证存在最小值而非若干个极小值。

Proposition 2.0.3 凸惩罚和非凸惩罚的对比:

1. 凸惩罚: Lasso, Adaptive lasso 等, 它们的优点包括凸惩罚可以保证解的唯一性, 因此有高效的算法得到估计里, 估计量具有稳定性和能产生稀疏的系数等;
2. 非凸惩罚: 代表为 SCAD 和 MCP, 优点主要是估计量具有 Oracle 性质的同时能立生稀疏的系数, 缺点包括惩罚函数的非凸性质无法保证解的唯一性, 因此可能产生多个局部最优值, 使得最终的结果稳定性不如凸惩罚, 同时, 非凸惩罚中还增加了一个凹度参数 (如 SADC 和 MCP 中的 γ), 进而增加了计算的复杂度。

Proposition 2.0.4 — 惩罚函数. 分为惩罚最小二乘函数 (Penalized Least Squares Function) 和惩罚似然函数 (Penalized Likelihood Function); 前者用于一般的线性回归模型, 后者用于 GLM

惩罚项一个常用的形式为 $P_\lambda(|\beta|) = \lambda \sum_{j=1}^p |\beta_j|^m, m \geq 0$. 当 $0 \leq m \leq 1$ 时, 由于惩罚项在零点存在奇异性, 即零点导数不存在, 该方法可使模型的部分系数估计正好为 0, 从而实现模型选择 (为什么 singular causes sparse?). 特别的, 当 $m = 0$ 时, 即得 L_0 惩罚项 $\lambda \sum_{j=1}^p I(|\beta_j| \neq 0)$, 这与 AIC 准则、BIC 准则等形式相似。当 $m = 1$ 时, 即得 L_1 惩罚项 (又称 LASSO 惩罚项) $\lambda \sum_{j=1}^p |\beta_j|$

Proposition 2.0.5 — 变量选择的一致性. 1. 对变量选择的一致性定义为:

$$P(i: \beta_i \neq 0 = i: \beta_i \neq 0) \rightarrow 1, \text{ as } n \rightarrow \infty$$

即某变量选择的方法具备变量选择的一致性, 则对于稀疏模型而言, 当样本量 $n \rightarrow \infty$ 时, 通过该方法选择出正确自变量的概率趋近于 1.

2. 在变量选择一致性的基础上, Zhao and Yu(2006) 里提出了符号一致性, 定义为:

$$\hat{\beta} \rightarrow_s \beta, \text{ as } n \rightarrow \infty$$

其中, s 是符号函数 sign 的缩写. 显然, 符号一致性是强于变量选择一致性的, 前者仅要求非 0 系数的位置一致, 而后者要求非 0 系数的位置和符号均一致, 所以若某变量选择方法具备符号一致性, 则其必须具备变量选择一致性, 反之不一定.

3. Zhao and Yu(2006) 就符号一致性进一步细化, 提出了强符号一致 (Strongly Sign Consistent): 存在 $\lambda_n = f(n)$, 即 λ_n 只与样本量 n 有关, 使得

$$\lim_{n \rightarrow \infty} P(\beta^n(\lambda_n) \rightarrow_s \beta^n) = 1$$

和常规符号一致 (General Sign Consistent): 存在 $\lambda \geq 0$, 使得

$$\lim_{n \rightarrow \infty} P(\beta^n \rightarrow_s \beta^n) = 1$$

强符号一致相对于常规符号一致的不同在前者要求 λ 与样本量 n 之间存在一定的关系, 从而可以在已知 n 的前提下, 确定 λ 。

Definition 2.0.1 — Soft and hard thresholding. 软阈值 (soft thresholding) 和硬阈值 (hard thresholding) 最早是由 Donoho and Johnstone(1994) 年提出的, 硬阈值函数的定义为:

$$\eta_H(\omega, \lambda) = \omega I(|\omega| > \lambda)$$

$$\eta_s(\omega, \lambda) = \text{sgn}(\omega)(|\omega| - \lambda)_+$$

其中, $I(|\omega| > \lambda)$ 为示性函数, sgn 为符号函数, 当 $|\omega| > \lambda$ 时, $(|\omega| - \lambda)_+ = |\omega| - \lambda$, 当 $|\omega| < \lambda$ 时, $(|\omega| - \lambda)_+ = 0$ 。因此硬阈值函数等价于

$$\eta_H(\omega, \lambda) = \begin{cases} \omega, & |\omega| > \lambda \\ 0, & |\omega| \leq \lambda \end{cases}$$

软阈值函数等价于

$$\eta_s(\omega, \lambda) = \begin{cases} \omega + \lambda, & \omega < -\lambda \\ 0, & |\omega| \leq \lambda \\ \omega - \lambda, & \omega > \lambda \end{cases}$$

Proposition 2.0.6 — Hard thresholding 的作用. 硬阈值 (HardThresholding) 可以求解如下优化问题:

$$\arg \min_x \|X - B\|_2^2 + \lambda \|X\|_0$$

其中:

$$X = [x_1, x_2, \dots, x_N]^T$$

$$B = [b_1, b_2, \dots, b_N]^T$$

$\|X\|_0$ 是求向量为向量 X 的零范数, 即向量 X 中非零元素的个数。根据范数的定义, 可以将上面优化问题的目标函数拆开:

$$\begin{aligned} F(X) &= \|X - B\|_2^2 + \lambda \|X\|_0 \\ &= \left[(x_1 - b_1)^2 + \lambda \|x_1\|_0 \right] + \left[(x_2 - b_2)^2 + \lambda \|x_2\|_0 \right] + \cdots \left[(x_N - b_N)^2 + \lambda \|x_N\|_0 \right] \end{aligned}$$

其中拆分项中符号 $\|x\|_0$ 的意思是

$$\|x\|_0 = \begin{cases} 1 & x \neq 0 \\ 0 & , x = 0 \end{cases}$$

现在, 我们可以通过求解 N 个独立的形如函数

$$f(x) = (x - b)^2 + \lambda \|x\|_0$$

的优化问题, 来求解这个问题。将 $f(x)$ 进一步写为:

$$f(x) = \begin{cases} (x - b)^2 + \lambda & , x \neq 0 \\ b^2 & , x = 0 \end{cases}$$

对于 $x \neq 0$ 部分, 我们知道它的最小值在 $x = b$ 处取得, 最小值为 λ 。现在的问题是 λ 与 b^2 到底谁更小? 最小者将是函数 $f(x)$ 的最小值。求解不等式 $b^2 > \lambda$ 可得

$$|b| > \sqrt{\lambda}$$

此时最小值在 $x = b$ 处取得; 求解不等式 $b^2 < \lambda$ 可得

$$|b| < \sqrt{\lambda}$$

此时最小值在 $x = 0$ 处取得; 因此

$$\arg \min f(x) = \begin{cases} 0 & , |b| < \sqrt{\lambda} \\ b & , |b| > \sqrt{\lambda} \end{cases}$$

即为硬阈值 (Hard Thresholding) 的公式。至此, 我们可以得到优化问题

$$\arg \min_x \|X - B\|_2^2 + \lambda \|X\|_0$$

的解为

$$h(B, \sqrt{\lambda}) = \eta_H(B, \sqrt{\lambda}) = \begin{cases} 0 & , |B| < \sqrt{\lambda} \\ B & , |B| > \sqrt{\lambda} \end{cases}$$

注: 该式为硬阈值 (Hard Thresholding) 的矩阵形式, 这里的 B 是一个向量, 应该是逐个元素分别执行硬阈值函数。

当优化问题变为

$$\arg \min_x \|X - B\|_2^2 / 2 + \lambda \|X\|_0$$

因为对目标函数乘一个常系数不影响极值点的获得, 所以可等价于优化问题

$$\arg \min_x \|X - B\|_2^2 + 2\lambda \|X\|_0$$

此时的解为 $\text{hard}(B, \sqrt{2\lambda})$

Proposition 2.0.7 — 渐近正态分布. 如果 $\sqrt{n}(\hat{\beta}_n - \beta) \rightarrow^d N(0, \Sigma)$, 其中 Σ 是半正定矩阵, 则称 $\hat{\beta}_n$ 为渐近正态分布 (asymptotically normally distributed), 称 Σ 为渐近方差 (asymptotic variance). 由于 $\sqrt{n}(\hat{\beta}_n - \beta)$ 收敛到一个非退化的分布, 故 $(\hat{\beta} - \beta)$ 收敛到 0 的速度与 $\frac{1}{\sqrt{n}}$ 的速度大致相同, 因此 β_n 也被称为“ \sqrt{n} 收敛估计量”或者“ \sqrt{n} 一致估计量”.

为了进一步阐述该性质, 假设设计阵是列正交的, 惩罚最小二乘的一般表达式可改写为:

$$\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\hat{\beta}^0\|^2 + \|\hat{\beta}^0 - \beta\|^2 + \sum_{j=1}^p P_{\lambda_j}(|\beta_j|) \right\}$$

这也就等价于下述一系列最小化问题

$$\min_{\beta_j \in \mathbb{R}} \left\{ (\hat{\beta}_j^0 - \beta_j)^2 + P_{\lambda_j}(|\beta_j|) \right\}, \quad j = 1, \dots, p$$

对 β_j 进行一阶微分可得

$$\text{sign}(\beta_j) \left(|\beta_j| + P'_{\lambda_j}(|\beta_j|) \right) - \hat{\beta}_j^0, \quad j = 1, \dots, p$$

Proposition 2.0.8 — Oracle Property. Fan 和 Li 推导出惩罚最小二乘估计具备 Oracle 三条性质如下:

1. 若 $\min_{\beta_j \in \mathbb{R}} \left\{ |\beta_j| + P'_{\lambda_j}(|\beta_j|) \right\} > 0$, 则参数估计满足稀疏性。
2. 若对较大的 $|\beta_j|$ 有 $P'_{\lambda_j}(|\beta_j|) = 0$, 则参数估计满足无偏性。
3. 当且仅当 $\arg \min_{\beta_j \in \mathbb{R}} \left\{ |\beta_j| + P'_{\lambda_j}(|\beta_j|) \right\} = 0$ 时, 参数估计对数据有连续性。

因此可以看出, L_p 惩罚项在 $0 \leq p < 1$ 时不满足连续性条件; 在 $p = 1$ 时 (即 L_1 惩罚项) 不满足无偏性条件; 在 $p > 1$ 时, 不满足稀疏性条件。因此, L_p 惩罚项不能同时满足上述三条性质, 即不具备 Oracle 性质。当然, 也有很多惩罚项被证明具备 Oracle 性质。

注意, 并不是所有具有神谕性的估计量都是好的估计量, 神谕性是好估计量的必要条件, 而为充要条件 (满足 Oracle 的估计量不一定是好估计量, 但是不满足的一定不是好估计量)。比如在估计 p 的情况下, 对于某 \sqrt{n} 的一致估计量 $\hat{\beta}$, 其阈值估计量 $\hat{\beta}I(|\hat{\beta}| > n^{-1/4})$ 是具有神谕性的, 但是该阈值估计量并不好, 因为其完全未考虑自变量个数和模型结构, 仅仅是为了构造具有神谕性的估计量而人为构造的。

Proposition 2.0.9 — 交叉验证. K 折交叉验证法 (K-fold Cross Validation): 给定 n 个样本并将其分为 K 组, 第 i 组的 m_i 个样本分别记作 $(y_{i,j}, X_{i,j})$. 记 $\hat{\theta}_{-i}$ 为使用除第 i 组外的样本作为训练集所估计出的 θ . 记 $\alpha(y_1, y_2)$ 为给定的损失函数. $\hat{y}_{i,j|\theta}$ 为使用自变量 $X_{i,j}$ 和参数 θ 作出的对 $y_{i,j}$ 的估计; 则对于模型 A 而言, 其交叉验证的结果如下定义:

$$CV(A) = \frac{1}{K} \sum_{i=1}^K \sum_{j=1}^{m_i} \alpha(\hat{y}_{i,j|\hat{\theta}_{-i}}, y_{i,j})$$

通过寻找 $CV(A_0) = \min CV(A)$ 得到最优模型 A_0 . 一般常用的损失函数为 $\alpha(y_1, y_2) = (y_1 - y_2)^2$. 常用的 K 折交叉验证法有 2 重、5 重、10 重, 其中 2 重计算。复杂程度较低, 速度较快, 但 10 重一般效果更好。

在 CV 的基础上, Golub et al(1979) 证明了对于岭回归 (Ridge Regression), 用广义交叉验证法对调节参数进行选择是一个很好的方法, 因此本文采用在后续调节参数的确定过程中, 采用 GCV 方法, 其定义如下:

$$GCV(A) = \frac{1}{n} \frac{\sum_{i=1}^n (y_i - \hat{y}_i|_{\hat{\theta}})^2}{(1 - \text{tr}(H(A))/n)^2}$$

其中矩阵 $H(A)$ 为估计系数矩阵, $\hat{\beta}$ 为全体样本下的参数估计, h_{ii} 为矩阵 $H(A)$ 中 (i, i) 位置的元素。

$$\hat{\beta}_j^L = \text{sign}(\hat{\beta}_j^0) \left(\left| \hat{\beta}_j^0 \right| - \frac{\lambda}{2} \right)^+$$

我们把式子 (2-10) 改写成矩阵形式

$$\hat{\beta}^L = \left(X^T X - \frac{\lambda}{2} B^{-1} \right)^{-1} X^T Y$$

因此模型中参数的有效数字是

$$d(\lambda) = \text{Tr} \left[X \left(X^T X - \frac{\lambda}{2} B^{-1} \right)^{-1} X^T \right]$$

于是进而可以构造广义交叉验证的统计量

$$GCV(\lambda) = \frac{\sum_{i=1}^n \left(y_i - X_{(i)}^T \hat{\beta}^L \right)^2}{n[1 - d(\lambda)/n]^2}$$

优化 S 时, 还是从 0 到 1 选 40 个值, 若使用广义交叉验证法, 那么只需要进行 40 次 Lasso 计算过程, 相比于 N 折交叉验证法更加高效。当然广义交叉验证法也

具体推导看北交通的论文

介绍 AIC, BIC

1. 残差平方和 RSS 就是这样的统计量, 定义如下:

$$RSS = \|y - X\beta\|_2^2$$

最小化 RSS 则可以得到普通最小二乘估计 (Ordinary Least Squares, OLS) $\hat{\beta}_{OLS} = (X^T X)^{-1} X^T y$. 当满足基本假设条件时, OLS 估计是最好的线性无偏估计 (BLUE) [8]。对于正态线性模型, OLS 估计和极大似然估计 (MLE) 一致。将 OLS 代入 RSS 可得

$$RSS = y^T \left(I - X (X^T X)^{-1} X^T \right) y$$

则对于子模型来说,

$$RSS_q = y^T \left(I - X_q (X_q^T X_q)^{-1} X_q^T \right) y$$

当在子模型中再添加一个自变量 X_q 时, 设计矩阵变为 $X_{q+1} = (X_q, x_q)$, 其中 $x_q = (x_{1q}, \dots, x_{nq})^T$, 此时残差平方和为

$$RSS_{q+1} = y^T \left(I - X_{q+1} (X_{q+1}^T X_{q+1})^{-1} X_{q+1}^T \right) y$$

由分块矩阵求逆公式可得

$$(X_{q+1}^T X_{q+1})^{-1} = \begin{pmatrix} X_q^T X_q & X_q^T x_q \\ x_q^T X_q & x_q^T x_q \end{pmatrix}^{-1} = \begin{pmatrix} (X_q^T X_q)^{-1} + aba^T & c \\ c^T & b \end{pmatrix}$$

其中 $a = (X_q^T X_q)^{-1} X_q^T x_q$, $b^{-1} = x_q^T x_q - x_q^T X_q (X_q^T X_q)^{-1} X_q^T x_q$, $c = -ab$ 则

$$\begin{aligned} & X_{q+1} (X_{q+1}^T X_{q+1})^{-1} X_{q+1}^T \\ &= (X_q, x_q) \begin{pmatrix} X_q^T X_q & X_q^T x_q \\ x_q^T X_q & x_q^T x_q \end{pmatrix}^{-1} \begin{pmatrix} x_q^T \\ x_q^T \end{pmatrix} \\ &= X_q (X_q^T X_q)^{-1} X_q^T + X_q aba^T X_q^T + x_q c X_q^T + X_q c x_q^T + x_q b x_q^T \end{aligned}$$

那么

$$X_{q+1} (X_{q+1}^T X_{q+1})^{-1} X_{q+1}^T - X_q (X_q^T X_q)^{-1} X_q^T = b (X_q a - x_q) (X_q a - x_q)^T \geq 0$$

则由 RSS_{q+1} 和 RSS_q 的计算式可得

$$RSS_{q+1} \leq RSS_q$$

也就是说当协变量子集扩大时, RSS 是逐渐减小的。那么按 RSS 越小越好的准则, 则所有自变量都要被选中, 这是不现实的。因此, 为了防止过多的协变量被选中, 需要给 RSS_q 乘以一个随 q 增大而增大的惩罚因子, 定义平均残差平方和:

$$RMS_q = \frac{1}{n-q} RSS_q$$

当 q 增大时, 由于 $(n-q)^{-1}$ 的惩罚作用, RMS_q 先下降后增加, 按照“ RMS_q 越小越好”的准则来选择子模型, 就称为平均残差平方和准则, 或 RMS_q 准则。

RMS_q 准则的出发点是模型的拟合优劣程度, 而选择回归方程的目的之一便是能更好的预测, 因此也可以考虑哪种模型对因变量预测得更准确, 从这个角度出发, 则可以使用下面的准则。假设子模型 (2.2) 的因变量预测值为 $\hat{y} = X_q \hat{\beta}_q$, 其中 $\hat{\beta}_q$ 为 $\beta_q = (\beta_0, \beta_1, \dots, \beta_{q-1})^T$ 的最小二乘估计, 则预测均方误差 $MSEP$ (Mean Square Error of Prediction) 可以作为评价模型预测精度的指标。定义为

$$MSEP = E \|\hat{y} - y\|^2$$

将其展开则有

$$MSEP = E \left(X_q \hat{\beta}_q - E \left(X_q \hat{\beta}_q \right) + E \left(X_q \beta_q \right) - y \right)^2 = \text{Var} \left(X_q \beta_q \right) + (E\hat{\varepsilon})^2$$

由上式可知预测均方误差由两部分构成，一个是预测值 $X_q \beta_q$ 的方差，另一个是期望误差的平方。按照“MSEP 越小越好”的准则来选择子模型，就称为 C_p 准则，它是由 Mallows [Mallows CL. 1973 提出的，定义为

$$C_p = \frac{RSS_q}{\hat{\sigma}^2} - n + 2q$$

其中 $\hat{\sigma}^2$ 为包含 p 个解释变量的完整模型的误差方差估计， $\hat{\sigma}^2 = \frac{\|y - x\hat{\beta}_{full}\|^2}{n-p}$ ， RSS_q 是包含 q 个解释变量的子集模型的残差平方和。在实际操作中，对于候选的子集模型， C_p 常常被绘制成和 p 有关的图像，从最小化预测值的总误差的意义上， C_p 近似等于 p 的模型被视为可接受的模型 [8]。但 C_p 准则是一个保守的模型选择器，它会倾向于过拟合 [Woodroffe, 1982]。并且在选择模型时，这种方法并不具有一致性，当 $n \rightarrow \infty$ 时，它经常倾向于选择更大的模型。这里，一个模型选择准则满足一致性指的是选出的模型以概率 1 收敛到真实模型。

2. 除了从拟合优度和预测精度两方面出发，还有的准则是基于信息论提出的。例如下面要介绍的 AIC (Akaike Information Criterion) 准则 [1]。

AIC 准则利用了信息论中的 Kullback-Leibler 信息量 [1]。由 Akaike 论文中的定义，设数据的真实密度函数是 $g(u)$ ，模型的密度函数是 $f(u | \theta_q)$ ，则关于 $g(u)$ 的 K-L 信息量是

$$I(g, f) = E_g \left(\ln \frac{g(u)}{f(u | \theta_q)} \right) = E_g(\ln g(u)) - E_g(\ln f(u | \theta_q))$$

其中 E_g 表示在分布 $g(u)$ 下求期望。上式可用于评价真实密度和模型密度的差异，其值越小，表示选用模型与真模型越接近。由于 $E_g(\ln g(u))$ 和备选模型无关，并且未知，因此可以看作常数。令 $R_{KL} = E_y E_g \left(-2 \ln f(u | \hat{\theta}_q(y)) \right)$ 为 K-L 信息下的风险函数，其值越小表示备选模型越好。AIC 准则就是由 R_{KL} 的渐近无偏估计推导出来的。

考虑含 $q \leq p$ 个参数的参数模型。假设它的密度函数是 $f(y | \theta_q)$ ，最大似然函数 (MLE) 是 $f(\hat{\theta}_q | y)$ ，其中 θ_q 是未知参数， $\hat{\theta}_q$ 是它的极大似然估计。由于风险函数越小越好，因此考虑它的相反数，并加上一个惩罚项，就得到了 $\ln f(\hat{\theta}_q | y) - q$ ，考虑使之达到最大值时的子模型。在线性模型中，定义

$$AIC = -2 \ln f(\hat{\theta}_q | y) + 2q$$

准则是选择使 AIC 值达到最小的子模型。对于 (2.2) 式的正态线性子模型，参数 $\theta_q = (\beta_q, \sigma_q^2)$ ，其似然函数为

$$f(\theta_q | y) = (2\pi\sigma_q^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma_q^2} \sum_{i=1}^n \left(y_i - \sum_{j=0}^{q-1} \beta_j x_{ij} \right)^2 \right\}$$

其中 $x_{i0} \equiv 1, i = 1, 2, \dots, n$ 。则对数似然函数为

$$\ln f(\theta_q | y) = -\frac{n}{2} \ln(2\pi\sigma_q^2) - \frac{1}{2\sigma_q^2} \sum_{i=1}^n \left(y_i - \sum_{j=0}^{q-1} \beta_j x_{ij} \right)^2$$

易得 β_q 和 σ_q^2 的极大似然估计为

$$\begin{aligned} \widehat{\beta}_q &= (X_q^T X_q)^{-1} X_q^T y \\ \widehat{\sigma}_q^2 &= \frac{1}{n} y^T \left(I - X_q (X_q^T X_q)^{-1} X_q^T \right) y = \frac{RSS_q}{n} \end{aligned}$$

将其代入 $\ln f(\theta_q | y)$ 可得极大似然函数为

$$\ln f(\widehat{\theta}_q | y) = \left(-\frac{n}{2} + \frac{n}{2} \ln \left(\frac{n}{2\pi} \right) \right) - \frac{n}{2} \ln RSS_q$$

去掉与 q 无关的常数项, 代入 AIC 定义式可得在子模型 (2.2) 下, AIC 变为

$$AIC = n \ln RSS_q + 2q$$

由上述表达式可以看出, 第一项是对模型拟合程度的评价, 其值是越小越好, 这样就容易选出更多的变量; 第二项是所选变量数。如果选入的变趾太多, 该项就会越大, 这要和最小化 AIC 的准则相悖, 因此该项起到控制选择变量个数的作用。从优化的角度思考, 它可以看作是一个惩罚项。和 Mallows's C_p 准则类似, AIC 准则也是一个“保守”的估计器, 因为它会将更多的变量选进来。一旦变量过多就会造成“过拟合”现象的出现。AIC 准则同样不满足模型选择的一致性 [Nishii R. 1984]。当 n 很小时, AIC 准则所依赖的渐近逼近代很差 [Dziak, Li and Collins. 2005]。因此, [Hurvich and Tsai. 1989] 是出了一个小样本修正, 即 AIC_c 统计量, 定义为

$$AIC_c = AIC + \frac{2q(q+1)}{n-q-1}$$

当 n 增大时, AIC_c 收敛于 AIC, 因此无论样本大小如何, AIC_c 都优于 AIC。如果说 AIC 准则的动机是拟合模型和真实模型的 Kullback-Leibler 差异, 那么 BIC (Bayesian Information Criterion) 准则就是 [Schwarz. 1978] 从贝叶斯的角度出发, 通过估计 Bayes 因子渐近展开的主项导出 BIC。具体来说, 贝叶斯方法用于模型选择的出发点是在给定数据 y 的情况下最大化子模型 M_i 的后验概率。利用贝叶斯公式, 在数据 y 已知时子模型的后验概率可表示为

$$P(M_i | y) = \frac{P(y | M_i) P(M_i)}{P(y)}$$

其中 $P(y|M_i)$ 称作子模型 M_i 的边际似然函数。若所有的候选模型的概率均等, 即 $P(M_i)$ 都相等, 则要求 $P(M_i|y)$ 的最大值就等价于求 $P(y|M_i)$ 的最大值, 而

$$P(y | M_i) = \int_{\theta_i} f(y | \theta_i) g_i(\theta_i) d\theta_i = \int \exp(\log(f(y | \theta_i) g_i(\theta_i))) d\theta_i$$

其中 $f(\mathbf{y} | \boldsymbol{\theta}_i)$ 为子模型下给定参数时 \mathbf{y} 的密度函数, $\boldsymbol{\theta}_i$ 为子模型下的参数向量, $g_i(\boldsymbol{\theta}_i)$ 是参数 $\boldsymbol{\theta}_i$ 的概率密度函数。令 $Q = \log(f(\mathbf{y} | \boldsymbol{\theta}) g_i(\boldsymbol{\theta}_i))$, 它在后验众数 $\tilde{\boldsymbol{\theta}}_i$ 处取得最大值, 因此将其在 $\tilde{\boldsymbol{\theta}}_i$ 处展开, 可得到下面的近似:

$$Q \approx \log \left(f(\mathbf{y} | \tilde{\boldsymbol{\theta}}_i) g_i(\tilde{\boldsymbol{\theta}}_i) + (\boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i)^T \nabla_{\boldsymbol{\theta}_i} Q|_{\tilde{\boldsymbol{\theta}}_i} + \frac{1}{2} (\boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i)^T H_{\boldsymbol{\theta}_i} (\boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i) \right)$$

其中 $H_{\boldsymbol{\theta}_i}$ 是一个 $|\boldsymbol{\theta}_i| \times |\boldsymbol{\theta}_i|$ 的矩阵且 $(H)_{mn} = \frac{\partial^2 Q}{\partial \theta_m \partial \theta_n} \Big|_{\tilde{\boldsymbol{\theta}}_i}$ 令 $\tilde{H}_{\boldsymbol{\theta}_i} = -H_{\boldsymbol{\theta}_i}$, 则 $P(\mathbf{y} | \mathbf{M}_i)$ 可以用下面式子来逼近:

$$\begin{aligned} P(\mathbf{y} | M_i) &\approx \int \exp \left\{ Q|_{\tilde{\boldsymbol{\theta}}_i} + (\boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i)^T \nabla_{\boldsymbol{\theta}_i} Q|_{\tilde{\boldsymbol{\theta}}_i} - \frac{1}{2} (\boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i)^T \tilde{H}_{\boldsymbol{\theta}_i} (\boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i) \right\} d\boldsymbol{\theta}_i \\ &= \exp \left(Q|_{\tilde{\boldsymbol{\theta}}_i} \right) \int \exp \left\{ -\frac{1}{2} (\boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i)^T \tilde{H}_{\boldsymbol{\theta}_i} (\boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i) \right\} d\boldsymbol{\theta}_i \\ &= \exp \left(Q|_{\tilde{\boldsymbol{\theta}}_i} \right) \int \exp \left\{ -\frac{1}{2} X^T \tilde{H}_{\boldsymbol{\theta}_i} X \right\} dX \end{aligned}$$

由于 $\tilde{H}_{\boldsymbol{\theta}_i}$ 的定义可知它是对称矩阵, 因此可将其对角化 $\tilde{H}_{\boldsymbol{\theta}_i} = S^T \Lambda S$ 。作代换 $X = S^T U$, Jacobi 矩阵 $J_{mn}(U) = \partial X_m / \partial U_n$, 则 $J(U) = S^T$, 因此 $\det J(U) = 1$ 上式变为

$$\begin{aligned} P(\mathbf{y} | M_i) &\approx \exp \left(Q|_{\tilde{\boldsymbol{\theta}}_i} \right) \int \exp \left\{ -\frac{1}{2} U^T \Lambda U \right\} (\det J(U)) dU \\ &= \exp \left(Q|_{\tilde{\boldsymbol{\theta}}_i} \right) \int \exp \left\{ -\frac{1}{2} \sum_{j=1}^{|\boldsymbol{\theta}_i|} \lambda_j U_j^2 \right\} dU \\ &= \exp \left(Q|_{\tilde{\boldsymbol{\theta}}_i} \right) \prod_{j=1}^{|\boldsymbol{\theta}_i|} \sqrt{\frac{2\pi}{\lambda_j}} \\ &= \exp \left(Q|_{\tilde{\boldsymbol{\theta}}_i} \right) \frac{(2\pi)^{|\boldsymbol{\theta}_i|/2}}{\prod_{j=1}^{|\boldsymbol{\theta}_i|} \lambda_j^{1/2}} \\ &= f(\mathbf{y} | \tilde{\boldsymbol{\theta}}_i) g_i(\tilde{\boldsymbol{\theta}}_i) \frac{(2\pi)^{|\boldsymbol{\theta}_i|/2}}{|\tilde{H}_{\boldsymbol{\theta}_i}|^{1/2}} \end{aligned}$$

其中 λ_j 是矩阵 $\tilde{H}_{\boldsymbol{\theta}_i}$ 的第 j 个特征值。将上式取对数, 可以得到

$$2 \ln P(\mathbf{y} | M_i) = 2 \ln f(\mathbf{y} | \tilde{\boldsymbol{\theta}}_i) + 2 \ln g_i(\tilde{\boldsymbol{\theta}}_i) + |\boldsymbol{\theta}_i| \ln 2\pi + \ln |\tilde{H}_{\boldsymbol{\theta}_i}^{-1}|$$

$\tilde{H}_{\boldsymbol{\theta}_i}$ 也称作观测数据的 Fisher 信息矩阵。令 $g_i(\boldsymbol{\theta}_i) = 1$, 即无信息的均匀分布且 $f(\mathbf{y} | \boldsymbol{\theta}_i)$ 和似然函数 $L(\boldsymbol{\theta}_i | \mathbf{y})$ 相等, 则 $\tilde{H}_{\boldsymbol{\theta}_i}$ 的每个分量可表示为:

$$\begin{aligned} \tilde{H}_{mn} &= - \frac{\partial^2 \ln L(\boldsymbol{\theta}_i | \mathbf{y})}{\partial \theta_m \partial \theta_n} \Big|_{\boldsymbol{\theta}_i = \tilde{\boldsymbol{\theta}}_i} \\ &= - \frac{\partial^2 \sum_{j=1}^n \ln L(\boldsymbol{\theta}_i | y_j)}{\partial \theta_m \partial \theta_n} \Big|_{\boldsymbol{\theta}_i = \tilde{\boldsymbol{\theta}}_i} \\ &= - \frac{\partial^2 \left(\frac{1}{n} \sum_{j=1}^n n \ln L(\boldsymbol{\theta}_i | y_j) \right)}{\partial \theta_m \partial \theta_n} \Big|_{\boldsymbol{\theta}_i = \tilde{\boldsymbol{\theta}}_i} \end{aligned}$$

设随机变量 $X_j = n \ln L(\theta_i | y_j)$ 。假设观测数据 y_1, y_2, \dots, y_n 是独立同分布的, 样本数 n 很大, 则应用弱大数定律可得

$$\frac{1}{n} \sum_{j=1}^n n \ln L(\theta_i | y_j) \xrightarrow{P} E[n \ln L(\theta_i | y_j)]$$

因此

$$\begin{aligned} \tilde{H}_{mn} &= - \left. \frac{\partial^2 E[n \ln L(\theta_i | y_j)]}{\partial \theta_m \partial \theta_n} \right|_{\theta_i = \tilde{\theta}_i} \\ &= - n \left. \frac{\partial^2 E[\ln L(\theta_i | y_j)]}{\partial \theta_m \partial \theta_n} \right|_{\theta_i = \tilde{\theta}_i} \\ &= - n \left. \frac{\partial^2 E[\ln L(\theta_i | y_1)]}{\partial \theta_m \partial \theta_n} \right|_{\theta_i = \tilde{\theta}_i} \\ &= n I_{mn} \end{aligned}$$

则有

$$|\tilde{H}_{\theta_i}| = n^{|\theta_i|} |I_{\theta_i}|$$

其中 I_{θ_i} 表示单个数据点 y_1 的 Fisher 信息矩阵。将上述结果代入 (2.8) 式可得:

$$2 \log P(y | M_i) = 2 \ln L(\hat{\theta}_i | y) + 2 \ln g_i(\tilde{\theta}_i) + |\theta_i| \ln 2\pi - |\theta_i| \ln n - \ln |I_{\theta_i}|$$

对于很大的 n , 保留含 n 的项, 忽略其他的项, 则可得

$$\ln P(y | M_i) \approx \ln L(\hat{\theta}_i | y) - \frac{|\theta_i|}{2} \ln n$$

由此导出了在线性模型下 BIC 的定义:

$$BIC = -2 \ln L(\hat{\theta}_q | y) + q \ln n$$

对于正态线性模型, BIC 则成为

$$BIC = n \ln RSS_q + q \ln n$$

根据寻找后验概率最大的子模型的出发点, BIC 值达到最小的子模型是最优模型。与 AIC 准则相比, BIC 在第二项中的惩罚增加了, 这使得 BIC 在选择变量时不再那么“轻易”而是非常小心翼翼。同时, 与 AIC 倾向于过拟合不同的是, BIC 是一种满足一致性的模型选择技术, 即满足相合性。也就是说当样本大小 n 足够大时, 最小的 BIC 模型将会以概率 1 收敛到真实模型。

3. Akaike (1973, 1974) 由信息论出发, 提出通过最小化拟合模型与真实模型之间的 KL 距离 (Kullback-Leibler divergence) 进行模型选择. 证明了在不考虑常数项的前提下, KL 距离的估计值近似于

$$-L_n(\hat{\theta}(s)) + \dim(\hat{\theta}(s)) = -L_n(\hat{\theta}(s)) + v(s)$$

其中, $L_n(\boldsymbol{\theta})$ 表示对数似然函数, $\hat{\boldsymbol{\theta}}(s)$ 表示 $\boldsymbol{\theta}(s)$ 的极大似然估计, $\dim(\hat{\boldsymbol{\theta}}(s))$ 表示参数的维数, 即模型 s 的大小 $v(s)$. Akaike (1973) 根据上述分析给出模型选择的评价标准 AIC (Akaike information criterion)

$$\text{AIC}(s) = -L_n(\hat{\boldsymbol{\theta}}(s)) + v(s)$$

并选取最小化该目标函数的模型作为最优模型.

4. Schwarz (1978) 利用贝叶斯理论, 通过给定先验分布得到了另一种模型选择评价准则, 称为 BIC (Bayesian information criterion), 其通过最小化如下函数选取最优模型

$$\text{BIC}(s) = -2L_n(\hat{\boldsymbol{\theta}}(s)) + \log n \cdot v(s)$$

与 AIC 相比, BIC 的惩罚项**考虑了样本容量**, 由于当 $n > 7$ 时总有 $\log n > 2$, 因此 BIC 的惩罚项取值大于 AIC, 从而当样本量过大时, BIC 可有效防止模型引入过多参数而导致的模型复杂度过高的问题.

5. Chen & Chen (2008) 对 BIC 进行了推广并提出了 EBIC (extended Bayesian information criterion) 准则, 更加适合高维统计模型. 记 \mathcal{S} 为模型空间, 其可以表示为 $\mathcal{S} = \bigcup_{j=1}^p \mathcal{S}_j$, 其中 \mathcal{S}_j 为两两不交的模型空间的子集, 且每一个 \mathcal{S}_j 中包含的模型具有相同的维度. 对于 $s \in \mathcal{S}_j$ 和参数 $0 \leq \gamma \leq 1$, 定义

$$\text{EBIC}_\gamma(s) = -2L_n(\hat{\boldsymbol{\theta}}(s)) + \log n \cdot v(s) + 2\gamma \log v(\mathcal{S}_j)$$

其中 $v(s)$ 和 $v(\mathcal{S}_j)$ 分别表示模型 s 和集合 \mathcal{S}_j 的维度. 当 $\gamma = 0$ 时, EBIC 即退化为经典的 BIC 准则. EBIC 对 BIC 进行了合理推广, 为高维模型下的模型选择问题提供了可靠的评价准则. Chen & Chen (2008) 在合理的假设下证明了, 当 $p_n = O(n^\kappa)$ 时, 若 $\gamma > 1 - 1/(2\kappa)$, EBIC 能够以趋于 1 的概率选择到真实模型, 即 EBIC 具备相合性. 更具体地, 若 $\gamma = 1$, 则 EBIC 的相合性对任意 $\kappa > 0$ 成立, 若 $\gamma = 0.5$, 则 EBIC 在 $\kappa < 1$ 时具备相合性. 相反地, 当 $p_n > \sqrt{n}$ 时, 传统的 BIC 准则不具备相合性.

6. 这几种判别方法都需要对 $2^p - 1$ 个模型遍历, 计算量大, 故常用向前等贪婪算法.
7. 向前: 计算变量与响应变量的偏相关系数. 由于向前选择法只进行选入而不删除, 并且不考虑已选入变量和待选变量的相关性, 因此可能会出现多重共线性的情况.
8. 向后: 通常是从全模型开始, 比较每个协变量的相关系数的 F 检验统计量值, 选出最小的那个, 若其值同时小于给定的临界值, 则将其剔除. 再考虑剩下的变量, 重复上述过程.
- 9.

$$\begin{aligned} E \|\hat{\boldsymbol{\beta}}_{\text{OLS}} - \boldsymbol{\beta}\|^2 &= \sigma^2 \text{tr}(\mathbf{X}^T \mathbf{X})^{-1} \\ D \|\hat{\boldsymbol{\beta}}_{\text{OLS}} - \boldsymbol{\beta}\|^2 &= 2\sigma^4 \text{tr}(\mathbf{X}^T \mathbf{X})^{-2} \end{aligned}$$

2.1 介绍各种方法

当设计矩阵为正交矩阵时, Lasso 的解的形式:

$$\hat{\beta}_j^{LASSO} = \text{sign}(\hat{\beta}_j^{\text{ols}}) \left(\hat{\beta}_j^{\text{ols}} - \frac{\lambda}{2} \right)_+$$

$$(x)_+ = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

LASSO 的解其实就是对最小二乘的解做了一个压缩：对于较小的那部分系数，直接将其压缩到零；对于较大的系数，作了一个平移压缩，这句话结合图片理解。

R 只有在设计矩阵是正交的前提下，才有显示解。对于设计矩阵不满足该条件，则需要采取特定的算法来求解，例如 LARS。

LASSO 模型罚函数的奇异性像一柄双刃剑：它可以同时实现系数估计和变量选择，但也让基于梯度的算法失效而令其难以有效计算。

Ridge regression 的解的形式

$$\hat{\beta}_j^{\text{Ridge}} = \frac{1}{1 + \lambda} \hat{\beta}_j^{\text{ols}}$$

可以看作是对 OLS 的解整体做了压缩，但是没有变量选择，没有参数的稀疏性。

Table 2.3 Estimators of β_j from (2.21) in the case of an orthonormal model matrix \mathbf{X} . .

q	Estimator	Formula
0	Best subset	$\tilde{\beta}_j \cdot \mathbb{I} \left[\tilde{\beta}_j > \sqrt{2\lambda} \right]$
1	Lasso	$\text{sign}(\tilde{\beta}_j) (\tilde{\beta}_j - \lambda)_+$
2	Ridge	$\tilde{\beta}_j / (1 + \lambda)$

2.1.1 Non-Neg 1995

随后 Breiman(1995) 年提出 non-negative garrotte (简称为 NNG), 其定义如下:

$$\hat{\beta}^{\text{NNG}} = \arg \min \sum_{i=1}^N \left(y_i - \sum_{j=1}^p c_j \hat{\beta}_j^0 x_{ij} \right)^2, \sum_{j=1}^p c_j \leq t, c_j \geq 0$$

随着 t 的减少, c_j 中部分元素被压缩至 0, 从而得到新的 $\hat{\beta}_j^{\text{NNG}} = c_j \hat{\beta}_j^0$. 在设计阵列正交的条件下, NG 方法参数估计有显式表达如下

$$\hat{\beta}_j^{\text{NG}} = \hat{c}_j \hat{\beta}_j^0 = \left(1 - \lambda / \left(\hat{\beta}_j^0 \right)^2 \right)_+, \quad j = 1, \dots, p$$

其中 $(\cdot)_+$ 表示非负项。特别地, \hat{c}_j 可将部分系数压缩到 0, 从而实现模型选择的目的。where λ is chosen so that $\|\hat{c}\|_1 = t$. Hence if the coefficient $\tilde{\beta}_j$ is large, the shrinkage factor will be close to 1 (no shrinkage), but if it is small the estimate will be shrunk toward zero.

NNG 方法的缺点主要包括:

1. 由于 NNG 的解依赖 OLS, 所以当自变量个数大于样本量 ($p > n$) 时, 由于 OLS 估计量不稳定, 故 NNG 估计量不稳定;
2. 在存在 OLS 估计量的前提下, NNG 估计量受 OLS 估计量正负号的影响.
3. 总的来说: 虽然具有变量选择的特点了, 但是由于过度依赖 OLS, 继承了 OLS 的缺点

Following this, Yuan and Lin(2007c) and Zou(2006) have shown that the nonnegative garrote is path-consistent under less stringent conditions than the lasso. This holds if the initial estimates are \sqrt{N} -consistent, for example those based on least squares (when $p < N$), the lasso, or the elastic net. "Pathconsistent" means that the solution path contains the true model somewhere in its path indexed by t or λ . On the other hand, the convergence of the parameter estimates from the nonnegative garrote tends to be slower than that of the initial estimate.

2.1.2 Lasso

Proposition 2.1.1 Lasso 优点

1. 凸函数, 但是原点不可导
2. 岭回归, 最小二乘法, 偏最小二乘法, 主成分法等方法无法达到变量选择的目的, 子集法则在 $p > 30$ 时不适用;
3. 若解释变量中不存在完全线性相关的解释变量, 则既是在 $p > N$ 的情况下 Lasso 惩罚也能产生唯一的解, 并且 Lasso 最多选择 $\min(p, N)$ 个解释变量, 这在 $p \gg N$ 的情况下非常有用, 比如 $p = 40,000$ 而 $N = 100$, 此时采用 Lasso 惩罚函数, 则最多保留 100 个解释变量, 远小于最初的 40,000 个解释变量。
4. 上一条也表达了一个缺点, 即害怕共线性, 同时受限于 (p, N) 两个参数, 没法选择更加少的特征

Proposition 2.1.2 Lasso 缺点

1. 对于 $p > n$ 的情形, 由于凸优化问题的本质, lasso 只能选择至多 n 个变量, 这在很多问题上的应用都受到了限制。
2. 不适用于相关性高的变量组: 对于相关性比较高的变量组, lasso 只是不确定的选择其中一个变量。
3. 对于 $n > p$ 的情形, 变量间存在相关性的情形下, 表现不如岭回归。

Proposition 2.1.3 逐步回归和岭回归的缺点:

1. 对于逐步回归法总是可以选择出便于解释的模型, 但是选择结果不稳定, 数据小的改变就会改变选择结果, 这也会影响其预测精确度。
2. 子集回归法都是根据选择准则选择出变量, 然后在对子模型进行系数估计, 应用性就会减弱很多。
3. 对于岭回归, 只是会使得所有的变量系数都成比例的变小, 但不会使得任何一个变量系数为 0, 因此不会自动选择出便于解释的模型。

LASSO:

$$\min_{\beta} \|y - X\beta\|^2 \quad \text{subject to } \|\beta\|_{l_1} \leq t$$

记 Lasso 惩罚得到的参数估计为 β^L 则

$$\hat{\beta}^L = \arg \min \sum_{i=1}^N \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2, \sum_j |\beta_j| \leq t$$

一般写成拉格朗日形式 (Lagrangian form), 如下所示:

$$\hat{\beta}^L = \arg \min (Y - X\beta)^T (Y - X\beta) + \lambda \|\beta\|_1$$

根据 lagrange 对偶性质 (Lagrangian Duality) 可知, λ 与 t 是一一对应的. 下面简单地证明当 $p = 2$ 时, λ 与 t 一一对应.

λ 和 t 构成以下方程组:

$$\begin{aligned} \hat{\beta}_1^L &= \text{sign}(\hat{\beta}_1^0) \left(\hat{\beta}_1^0 - \frac{\lambda}{2} \right)^+ \\ \sum_j |\hat{\beta}_j^L| &= t \end{aligned}$$

不失一般性, 假设 $\hat{\beta}_j^0$ 均为正数, 则有如下方程组:

$$\begin{aligned} \hat{\beta}_1^L &= \left(\hat{\beta}_1^0 - \frac{\lambda}{2} \right)^+ \quad \hat{\beta}_2^L = \left(\hat{\beta}_2^0 - \frac{\lambda}{2} \right)^+ \\ \hat{\beta}_1^L + \hat{\beta}_2^L &= t \end{aligned}$$

解上式可得

$$\lambda = \hat{\beta}_1^0 + \hat{\beta}_2^0 - t$$

证毕. 一般情况下:

$$t = \sum_{i=1}^p \text{sign}(\hat{\beta}_i^0) \hat{\beta}_i^0 - p \frac{\lambda}{2}$$

若设计阵列正交, LASSO 方法参数估计的显式表达为

$$\hat{\beta}_j^{LASSO} = \text{sign}(\hat{\beta}_j^0) \left(|\hat{\beta}_j^0| - \lambda/2 \right)_+, \quad j = 1, \dots, p$$

我们令

$$G(\beta, X, Y, \lambda) = \sum_i \left(y_i - \sum_j \beta_j x_{ij} \right)^2 + \lambda \sum_j |\beta_j|$$

现在把 G 写成矩阵形式

$$G(\beta, X, Y, \lambda) = (Y - X\beta)^T (Y - X\beta) + \lambda B I_p$$

其中 $\beta = (\beta_1, \beta_2, \dots, \beta_p)^T$, I_p 是 $p \times p$ 的单位矩阵, B 是有 j 个对角元素 $|\beta_j|$ 的对角阵, G 取极小值就得到了 β 的最佳估计 $\hat{\beta}$. 我们可以把上式写作

$$G(\beta, X, Y, \lambda) = Y^T Y - \beta^T X^T Y - Y^T X \beta + \beta^T X^T X \beta + \lambda B$$

再对上式求 β 的偏导得

$$\frac{\partial G(\beta, X, Y, \lambda)}{\partial \beta} = -2Y^T X + 2(X^T X)\beta + \lambda \text{sign}(\beta)$$

通过使得上式为零, 我们可以得到以下的结果

$$\hat{\beta}_j^L = \hat{\beta}_j^0 - \frac{\lambda}{2} \text{sign}(\hat{\beta}_j^L)$$

Lemma 2.1 如果 $y = x - \frac{\lambda}{2} \text{sign}(y)$, 那么 x 和 y 具有相同的符号。注意到

$$y = x - \frac{\lambda}{2} \text{sign}(y) \Leftrightarrow x = y + \frac{\lambda}{2} \text{sign}(y)$$

如果 y 是正的, 那么 $y + \text{sign}(y)$ 也是正的, 所以 x 是正的, 反之, 如果 y 是负的, x 也是负的, 所以 x 和 y 具有相同的符号。■

根据引理可以得出 $\text{sign}(\hat{\beta}_j^L) = \text{sign}(\hat{\beta}_j^0)$, 故

$$\hat{\beta}_j^L = \hat{\beta}_j^0 - \frac{\lambda}{2} \text{sign}(\hat{\beta}_j^0) = \left(\hat{\beta}_j^0 - \frac{\lambda}{2}\right) I_{[\hat{\beta}_j^0 \geq 0]} + \left(\hat{\beta}_j^0 + \frac{\lambda}{2}\right) I_{[\hat{\beta}_j^0 < 0]}$$

又可以导出

$$\hat{\beta}_j^L = \text{sign}(\hat{\beta}_j^0) \left(\left| \hat{\beta}_j^0 \right| - \frac{\lambda}{2} \right)^+$$

其中 z^+ 表示 z 大于 0 时取 z , 否则取 0。我们可以很直观地看出 $\lambda/2$ 可以作为一个判断非零估计的阈值, 也就是说, 如果所对应的 $|\hat{\beta}_j^0|$ 小于这个阈值, 那么 $|\hat{\beta}_j^L|$ 就等于零。通过解决以下问题

$$\begin{cases} \hat{\beta}_j^L = \text{sign}(\hat{\beta}_j^0) \left(\left| \hat{\beta}_j^0 \right| - \frac{\lambda}{2} \right)^+ \\ \sum_j |\hat{\beta}_j^L| = t \end{cases}$$

我们可以计算出参数 λ , 其中我们选择参数 λ 使得

$$\sum_j |\hat{\beta}_j^L| = t$$

因此每一个 λ 的值都对应一个唯一的 t 的值。我们先看 $p = 2$ 时的简单情形, 不失一般性, 我们不妨假定最小二乘估计 $\hat{\beta}_j^0$ 都是正的, 那么问题就变成解方程组:

$$\begin{cases} \hat{\beta}_1^L = \left(\hat{\beta}_1^0 - \frac{\lambda}{2}\right)^+ \\ \hat{\beta}_2^L = \left(\hat{\beta}_2^0 - \frac{\lambda}{2}\right)^+ \\ \hat{\beta}_1^L + \hat{\beta}_2^L = t \end{cases} \quad (2.1)$$

解得

$$\lambda = \hat{\beta}_1^0 + \hat{\beta}_2^0 - t \quad (2.2)$$

然后再把式子 (2.2) 代入到式子 (2-11) 中就可以求得 Lasso 估计 $\hat{\beta}_j^L$. 然后拓展到一般情形

$$t = \sum_j |\hat{\beta}_j^0| - \frac{\lambda}{2} p$$

我们可以看出 λ 与 Lasso 参数 t 之间具有一一对应的关系。通常情况下, Lasso 估计是没有封闭解的.

当设计阵非列正交时, 可给定 λ 然后使用二次算法求解参数估计。更一般地, 可使用最小角回归算法 (LARS) 求解。

lasso 的特点:

1. 零点不可导, 故可以选择变量, 并且同时估计参数 (解释)
2. 其解的形式为 $\hat{\beta}_j = \text{sign}(\hat{\beta}_j^{OLS}) (\hat{\beta}_j^{OLS} - t)$, 这里 $\hat{\beta}_j^{OLS}$ 是普通最小二乘的解。因此对变量系数的估计是有偏的。
3. lasso 变量选择不一定是相合的, 需要不可表示性的条件
4. 不满足 oracle 性质。
5. 相对于岭回归, 压缩程度减小。
6. 计算复杂度小, 并且参数具有连续性。
7. 不具有相合性, 或者说需要条件
8. t 是根据交叉验证得到的

2.1.3 Relaxed Lasso

Relaxed LASSO 是由 Meinshausen 提出的。主要目的是: 减少 Lasso 对于非零变量的惩罚程度, 故将原来 Lasso 选择变量及参数估计一步就可以完成的工作分为两步。它的主要思想为: 先计算 LASSO 在由全路径方法选取的调整参数下的参数估计结果 (调整参数选择将在第五节讨论), 选出合适的变量; 对于选出的变量, 再次应用 LASSO, 但减小或者消除惩罚因子的作用, 因此第二步不进行变量选择, 只进行参数估计。由此, Relaxed LASSO 会得到与普通 LASSO 方法同样的模型, 但是回归参数估计不同, **前者不会过度缩小非零参数**, 因为模型选择和参数估计被分成两个独立的过程。上述方法是基于第一步 LASSO 能够选出真实模型的前提假设的。放松惩罚项可以更准确的估计参数值。若令第二步的惩罚项为零, 则为典型的 LASSO/OLS 方法。一些经验和理论的结果表明, 该方法优于普通的 LASSO 方法。

Meinshausen (2007) 提出了 Relaxed Lasso:

$$\hat{\beta}^{\lambda, \phi} = \arg \min_{\beta} \frac{1}{n} \sum_{i=1}^n (Y_i - X_i^T \beta I_{M_\lambda})^2 + \phi \lambda \|\beta\|_1$$

其中, $\lambda \in [0, \infty), \phi \in (0, 1], I_{M_\lambda}$ 是指示函数, $M_\lambda = \{1 \leq k \leq p \mid \beta_k^L \neq 0\}$ 是指标集, 其效

果是:

$$\{\beta I_{M_\lambda}\}_k = \begin{cases} 0, k \notin M_\lambda \\ \beta_k, k \in M_\lambda \end{cases}$$

对于足够大的 λ , 如 $\lambda > 2 \max_k n^{-1} \sum_{i=1}^n Y_i X_i^k$, M_λ 为空集; 当 $\lambda = 0$ 且自变量个数小于样本量个数时, $M_\lambda = \{1, \dots, p\}$. 根据公式可以看出, Relaxed Lasso 惩罚较 Lasso 惩罚多出了 ϕ 这个参数。惩罚参数 λ 控制变量选择部分, 放松参数 ϕ 的控制参数的压缩速度。当 $\phi = 1$ 时, Relaxed Lasso 和 Lasso 无差别; 当 $\phi < 1$ 时, 压缩参数的速度将低于传统的 Lasso 惩罚: 当 $\phi = 0$ 时, Relaxed Lasso 估计量与 OLS 估计量一致, 类似于 Efronet al. (2004) 提出的 LARS-OLS 混合方法, 即用 LARS 的方法进行变量选择, 再用 OLS 方法对被选入模型的自变量进行系数估计。得到被选入模型的自变量; (2) 在第一步得到的自变量的基础上, 用惩罚参数 $\lambda\phi$ 再次进行 Lasso 估计。

当设计矩阵是正交阵时, Relaxed Lasso 的解为

$$\hat{\beta}_k^{\lambda, \phi} = \begin{cases} \hat{\beta}_k^0 - \phi\lambda, & \hat{\beta}_k^0 > \lambda \\ 0, & |\hat{\beta}_k^0| \leq \lambda \\ \hat{\beta}_k^0 + \phi\lambda, & \hat{\beta}_k^0 < -\lambda \end{cases}$$

从上述解可以看出: 当 $\phi = 0$ 时, 解为硬阈值函数; 当 $\phi = 1$ 时, 解为软阈值函数。当 $0 < \phi < 1$ 时, 解介于硬阈值和软阈值之间, 类似于桥回归中 γ 介于 $[0, 1]$ 的情形。

Meinshausen (2007) 证明了:

1. 当自变量中噪音变量占比较大时 (即 q/p 的比例较小时), Relaxed Lasso 估计量的 MSE 远小于 Lasso 估计量的 MSE, 因为前者在不牺牲预测精度的前提下, 选入模型的变量更少;
2. 在高维数据情况下, Relaxed Lasso 惩罚的收敛速度远高于 Lasso 惩罚, 因为前者的收敛速度受自变量个数的影响程度远小于后者。

2.1.4 Adaptive Lasso

针对 Lasso 的缺点, Zou(2005) 提出了适应性的 lasso 方法,

$$\hat{\beta} = \arg \min_{\beta} \|y - X\beta\|^2 + \lambda \sum_{i=1}^p \hat{w}_i |\beta_i|$$

该方法利用全模型最小乘估计计算不同变量的惩罚项。若某变量最小二乘参数估计值较大, 则其更可能为真实模型中的变量, 因此该变量在惩罚最小二乘估计时惩罚项应较小, 以确保其有更大的概率被选入模型。

它的惩罚项为

$$p_\lambda(|\beta_j|) = \lambda \frac{1}{|\hat{\beta}_j^0|^\gamma} |\beta_j|$$

再假定 X 是正交阵, 即 $X^T X = I$, 自适应 Lasso 估计用 $\hat{\beta}_j^{\text{lasso}}$ 表示为

$$\hat{\beta}_j^{\text{lasso}} = \text{sign}(\hat{\beta}_j^0) \left(|\hat{\beta}_j^0| - \lambda / |\hat{\beta}_j^0|^\gamma \right)^+$$

自适应 Lasso 根据估计参数的大小给予每个惩罚项不同的权重, 这样模型在不影响拟合效果的前提下选择出的变量将多于 Lasso 回归, 能够防止过多地压缩参数。

在 l_1 惩罚中对每个不同的变量给定不同的权重, 通常情况下取 $\hat{w}_j = 1 / |\hat{\beta}_j|^\gamma$, 并且证明了只要权重 \hat{w}_j 是 \sqrt{n} 相合的, 此方法就具有 Oracle 性质。

同 Relaxed LASSO 性质类似, Adaptive LASSO 也可以减弱 LASSO 对非零系数的缩减, 从而减小偏差。但 Adaptive LASSO 更重要的意义在于**当变量个数固定而样本量趋于无穷时, 它具有相合性, 且这些参数估计的分布与事先给定非零变量位置的最小二乘得到的参数估计的分布渐近相同。**

Zou(2006) 提出 Adaptive Lasso(下称自适应 Lasso), 所谓自适应是指惩罚函数会根据估计量的大小而调整, 而非 Lasso 中固定的 λ 惩罚, 其中形式为:

$$\hat{\beta}_n = \arg \min_{\beta} \|Y - X\beta\|_2^2 + \lambda_n \sum_{i=1}^p \hat{w}_i \|\beta_i\|_1$$

其中, λ_n 表示 λ 会随着样本容量 n 的变化而变化; 因为 λ_n 的变化将导致对 β 的估计发生变化, 因此用 $\hat{\beta}_n$ 表示系数的估计量; $\hat{w} = \frac{1}{|\hat{\beta}|^\gamma}$, 其中 $\gamma > 0$ 是一个常数, Zou(2006) 建议选择 \sqrt{n} 收敛估计量 $\hat{\beta}$ 作为初始估计量 (initial estimator), 如可以令 $\hat{\beta} = \hat{\beta}^{OLS}$. 从上述定义中可以看出, 若初始估计量 $\hat{\beta}$ 的某些分量较大 (如 $\hat{\beta}_j$), 则当 $\gamma \geq 1$ 时, 该分量对应的惩罚系数 $\hat{w}_j = \frac{1}{|\hat{\beta}_j|^\gamma}$ 相对于其他分量对应的惩罚系数较小, 因此自适应 Lasso 惩罚函数在保证估计量 (一定的假设前提下) 的同时, 减少了估计量的偏差。

Zou(2006) 证明了当自变量个数固定, $\frac{\lambda_n}{\sqrt{n}} \rightarrow 0$ 且 $\lambda_n n^{\frac{\gamma-1}{2}} \rightarrow \infty$ 时, 自适应 Lasso 估计量具有神谕性, 且其渐近方差为 $\sigma^2 \times C_{11}^{-1}$, 其中 σ^2 是随机误差项的方差, C_{11} 的定义同上. 并且, Zou(2006) 根据 Efron et al.(2004) 的 LARS 算法给出了自适应 Lasso 的算法步骤:

1. 令 $x_j^{**} = x_j / \omega_j$
2. 对于全部的 λ_n , 计算 $\beta^{**} = \arg \min \left\| y - \sum_{j=1}^p x_j^{**} \beta_j \right\|^2 + \lambda_n \sum_{j=1}^p |\beta_j|$
3. $\hat{\beta}^{*(n)} = \hat{\beta}^{**} / \omega_j$

在实践过程中, $\hat{\beta}, \lambda_n, \gamma$ 是需要通过交叉验证法确定的. Zou(2006) 中提出的自适应 Lasso 的有如下缺点: (1) 需要找到一个 \sqrt{n} 一致的估计量作为初始估计量, 而在高维数据乃至超高维数据的情况下, 很难确保找估计量的性质。针对缺点 (1), Huang et al(2008) 提此在高维数据下存在并构造 \sqrt{n} 一致估计量的充分条件;

针对缺点 (2), Huang et al(2008) 证明了在高维数据的情况下, 若初始估计量存在, 则自适应 Lasso 估计具有神谕性; 在超高维数据的情况下, 若设计矩阵满足偏正交条件等前提条件, 并将边际回归估计量作为初始估计量, 则自适应 Lasso 惩罚具有神谕性。

2.1.5 Group Lasso

在**已知变量的分组情况**, 在进行模型选择时, 我们希望能同时保留或删除同一组的变量。Yuan 和 Lin 提出的 Group LASSO, Zhao 等提出的 Composite Absolute Penalty(CAP)

方法都是处理上述问题的方案。

Yuan and Lin 提出了 Group Lasso, 用于保证当自变量中存在组变量时对同一组变量进行同时选入或者剔除, Group LASSO 的基本形式如下: 假设变量分为 J 组, 下标的集合分别表示为 G_1, \dots, G_J 每组的变量数目为 p_1, \dots, p_J . 记 $\beta_{G_j} = (\beta_j)_{j \in G_j}$ 为 β 相应元素构成的子向量.

$$Y = \sum_{j=1}^J X_j \beta_j + \epsilon$$

其中, Y 是 $n \times 1$ 的向量, $\epsilon \sim N_n(0, \sigma^2 I)$, X_j 是 $n \times p_j$ 的矩阵, β_j 是 $p_j \times 1$ 的系数向量. 对于 $\beta \in R^d, d \geq 1$ 和对称的正定阵 $K_{d \times d}$, 我们有如下记号:

$$\|\beta\|_K = \sqrt{(\beta^T K \beta)}$$

Group lasso 估计量是方程的解:

$$\hat{\beta} = \arg \min_{\beta} \frac{1}{2} \left\| Y - \sum_{j=1}^J X_j \beta_j \right\|^2 + \lambda \sum_{j=1}^J \|\beta_j\|_{K_j}$$

其中 $\|\beta_{G_j}\|_{K_j} = (\beta_{G_j}^T K_j \beta_{G_j})^{1/2}$ 是正定矩阵 K_j 决定的椭圆范数, $\lambda \geq 0$. 如果选择二范数, 容易导致大的组容易被选出. 可以看出, 在 Group LASSO 方法中令所有 $p_j = 1, K_j$ 为单位矩阵即得 LASSO 方法. 如上形式的惩罚项鼓励变量以组群为单位进入或剔除于选择模型.

Yuan and lin 根据 Karush-Kuhn-Tucker 条件得到了, 当 $K_j = p_j I_{p_j}$ 时, 上述方程最小解的充分必要条件:

$$\begin{aligned} -X_j^T(Y - X\beta) + \frac{\lambda \beta_j \sqrt{p_j}}{\|\beta_j\|} &= 0, \forall \beta_j \neq 0 \\ \left\| -X_j^T(Y - X\beta) \right\| &\leq \lambda \sqrt{p_j}, \forall \beta_j = 0 \end{aligned}$$

当 $X_j^T X_j = I_{p_j}$ 时, 上述两式的解为:

$$\beta_j = \left(1 - \frac{\lambda \sqrt{p_j}}{\|S_j\|} \right)_+ S_j$$

其中, $S_j = X_j^T (Y - X\beta_{-j})$.

当然, Yuan and Lin 提出了用于解决 Group Lasso 的 group LARS 算法, 其主要思路如下: 记 $\theta(r, X)$ 为向量 r 和矩阵 X 列空间的夹角, 即 $\theta(r, X)$ 为向量 r 与其在矩阵 X 列空间投影的夹角, 则 $\cos^2 \theta(r, X_j)$ 等于当 r 为因变量 Y 和 X_j 为自变量进行线性回归的 R^2 . 最开始令所有系数均为 0, 找到与因变量 Y 的回归系数最大的自变量 (例如 X_1 , 与之对应的 R_1^2), 并使得回归系数继续沿着 Y 在 X_1 列空间的投影方向移动, 在移动的过程中, R_1 逐渐减少, 直至出现第二个自变量 (例如 X_2), 其与 Y 的回归系数 $R_1^2 = R_2^2$, 然后, 系数沿着 Y 在 X_1 和 X_2 共同张成的列空间的投影方向前进, 在移动的过程中, R_2 和 R_1 逐渐减少且保持相等, 直至出现第三个自变量 (例如 X_3) 有 $R_3 = R_2 = R_1$; 此时改变系数方向, 使其沿着 Y 在 $X_1 X_2$ 和 X_3 共同张成的列空间的投影方向前进, 以此类推.

当 K 为常数矩阵时 (即常数乘以单位矩阵), Group Lasso 的惩罚又被称为 L_1/L_2 惩罚, 因为其对每一个组变量的组内系数进行了 L_2 惩罚, 又对每个组整体系数进行 ($p = 1, 2, \dots, \infty$), SCAD/ L_2 惩罚 (Wang et al., 2007; Zeng and Xie, 2014; Guo et al., 2015), MCP/ L_2 惩罚等.

在实际生活中, Group Lasso 的应用场景有很多, 比如: 在处理分类数据及多因子方差分析中, 我们通常其所具有的多个水平值视为多个特征变量, 那么在进行变量选择时, 我们将其归为一个组, 从而保证它们会被同时选进来或剔除; 在基因学中, 将起同一生物功能的多个基因序列归入一个组; 在机器学习领域, 多任务学习中使用同样模型参数的输入变量进行多个目标变量的预测时, 输入变量被归入一个组; 并且随着实际应用的逐渐复杂, 更多的 Group Lasso 模型被开发:

Jacob et al(2009) 考虑了当组变量之间存在重叠的情况, 即不同的组变量可能包含某些相同的变量; Friedman et al(2010) 提出了 Sparse Group Lasso 模型, 既能在组变量之间能产生稀疏模型, 又能在单个组变量内部产生稀疏模型, 其函数形式为:

$$\hat{\beta} = \arg \min_{\beta} \left\| Y - \sum_{j=1}^J X_j \beta_j \right\|^2 + \lambda_1 \sum_{j=1}^J \|\beta_j\|_{K_j} + \lambda_2 \|\beta\|$$

从函数形式看, Sparse Group Lasso 与 Group Lasso 的关系类似于 Elastic Net 与 Ridged 的关系, 前者均是在后者的基础上增加了 Lasso 惩罚项, 不同的 Sparse Group Lasso 得到的组变量内的稀疏性, Elastic Net 得到的群组变量之间的稀疏性.

Zhao et al(2009) 和 Kim and Xing(2010) 提出了 Tree Structured Group Lasso, 其前提是数据之间不仅存在结构, 其结构还满足一定的次序关系, 如在基因领域中, 某基因是否表达取决于另一个基因是否已经表达; 其关系可以用树形象的表示出来, 因此又被称为树状组变量.

Group Lasso 的优点是目标函数是关于未知参数的凸函数, 存在唯一的全局最小值. 许多学者研究了它的性质, Bach 发现对于确定的 p , 在不可表条件的变形条件下, Group Lasso 在随机设计模型中具有组选择一致性; Nardi 和 Rinaldo 研究了不可表条件下 Group Lasso 的一致性, 以及在受限特征根条件下预测和估计误差的界值; Wei 和 Huang 考虑了在稀疏 Riesz 条件下, Group Lasso 预测和估计误差的稀疏性质以及 ℓ_2 界值.

此外 Group Lasso 是 Lasso 在组结构下的扩展, Lasso 惩罚的解析解说明其参数估计值是有偏的, 显然不具有 Oracle 性质, 同理 Group Lasso 不具备该性质. 一个很自然的问题是, 在什么条件下 Group Lasso 比 Lasso 好? Huang 和 Zhang 提出了强群组稀疏性的概念, 表明在强群组稀疏条件和其他某些条件下, Group Lasso 比 Lasso 更优良. Group Lasso 组间是基于 L_1 范数惩罚, 因而具有类似 Lasso 的缺点. Lasso 会过度压缩大系数, GLasso 对系数大的组也会过度压缩. Lasso 往往会选择不重要的变量进入模型从而无法区分系数较小的变量和不重要变量, 导致很高的假阳性, Group Lasso 的参数估计值偏差过大, 也往往会选择过多的组.

$$\text{minimize}_{(\theta_1, \dots, \theta_J)} \left\{ \frac{1}{2} \left\| \mathbf{y} - \sum_{j=1}^J \mathbf{Z}_j \theta_j \right\|_2^2 + \lambda \sum_{j=1}^J \|\theta_j\|_2 \right\}$$

For this problem, the zero subgradient equations, take the form

$$-\mathbf{Z}_j^T \left(\mathbf{y} - \sum_{\ell=1}^J \mathbf{Z}_\ell \hat{\theta}_\ell \right) + \lambda \hat{s}_j = 0, \quad \text{for } j = 1, \dots, J$$

where $\hat{s}_j \in \mathbb{R}^{p_j}$ is an element of the subdifferential of the norm $\|\cdot\|_2$ evaluated at $\hat{\theta}_j$. Whenever $\hat{\theta}_j \neq 0$, then we necessarily have $\hat{s}_j = \hat{\theta}_j / \|\hat{\theta}_j\|_2$, whereas when $\hat{\theta}_j = 0$, then \hat{s}_j is any vector with $\|\hat{s}_j\|_2 \leq 1$.

One method for solving the zero subgradient equations is by holding fixed all block vectors $\{\hat{\theta}_k, k \neq j\}$, and then solving for $\hat{\theta}_j$. Doing so amounts to performing block coordinate descent on the group lasso objective function. Since the problem is convex, and the penalty is block separable, it is guaranteed to converge to an optimal solution. With all $\{\hat{\theta}_k, k \neq j\}$ fixed, we write

$$-\mathbf{Z}_j^T (\mathbf{r}_j - \mathbf{Z}_j \hat{\theta}_j) + \lambda \hat{s}_j = 0$$

where $\mathbf{r}_j = \mathbf{y} - \sum_{k \neq j} \mathbf{Z}_k \hat{\theta}_k$ is the j^{th} partial residual. From the conditions satisfied by the subgradient \hat{s}_j , we must have $\hat{\theta}_j = 0$ if $\|\mathbf{Z}_j^T \mathbf{r}_j\|_2 < \lambda$, and otherwise the minimizer $\hat{\theta}_j$ must satisfy

$$\hat{\theta}_j = \left(\mathbf{Z}_j^T \mathbf{Z}_j + \frac{\lambda}{\|\hat{\theta}_j\|_2} \mathbf{I} \right)^{-1} \mathbf{Z}_j^T \mathbf{r}_j \quad (2.3)$$

This update is similar to the solution of a ridge regression problem, except that the underlying penalty parameter depends on $\|\hat{\theta}_j\|_2$. Unfortunately, Equation (2.3) does not have a closed-form solution for $\hat{\theta}_j$ unless \mathbf{Z}_j is orthonormal. In this special case, we have the simple update

$$\hat{\theta}_j = \left(1 - \frac{\lambda}{\|\mathbf{Z}_j^T \mathbf{r}_j\|_2} \right) \mathbf{Z}_j^T \mathbf{r}_j \quad (2.4)$$

where $(t)_+ := \max\{0, t\}$ is the positive part function.

An alternative approach is to apply the **composite gradient methods** to this problem. Doing so leads to an algorithm that is also iterative within each block; at each iteration the block-optimization problem is approximated by an easier problem, for which an update such as (4.15) is possible. In detail, the updates take the form

$$\begin{aligned} \omega &\leftarrow \theta_j + \nu \cdot \mathbf{Z}_j^T (\mathbf{r}_j - \mathbf{Z}_j \hat{\theta}_j), \text{ and} \\ \hat{\theta}_j &\leftarrow \left(1 - \frac{\nu \lambda}{\|\omega\|_2} \right)_+ \omega \end{aligned}$$

where ν is a step-size parameter.

2.1.6 CAP

Zhao 等提出的 CAP 方法与 Group LASSO 类似, 只是将 K_j 范数替换为 L_{γ_j} 范数 $\left(\|\beta\|_{\gamma_j} = \sqrt[\gamma_j]{\sum |\beta_j|^{\gamma_j}}\right)$, 并用 L_{γ_0} 范数连接每组的惩罚项。特别地, CAP 允许组与组之间有重叠的变量。最小化下述式子即得 CAP 方法的参数估计:

$$\|Y - X\beta\|^2 + \lambda \left\| \left(\|\beta_{G_1}\|_{\gamma_1}, \dots, \|\beta_{G_J}\|_{\gamma_J} \right)^T \right\|_{\gamma_0}$$

其中 $\gamma_j > 1, j = 0, 1, \dots, J$. 上述通过调整惩罚项以实现特定模型选择目的的思想可以推广到更多的方法。例如已知重要的变量可以不加惩罚因子, 而疑似噪声的变量可以配置更大的惩罚项。对不同的变量给予不同的惩罚项可以加入选择的先验信息, 这样惩罚最小二乘估计就会变得更加灵活。

2.1.7 Sparsity group lasso

In order to achieve within-group sparsity, we augment the basic group lasso with an additional ℓ_1 -penalty, leading to the convex program

$$\text{minimize}_{\{\theta_j \in \mathbb{R}^{p_j}\}_{j=1}^J} \left\{ \frac{1}{2} \left\| \mathbf{y} - \sum_{j=1}^J \mathbf{Z}_j \theta_j \right\|_2^2 + \lambda \sum_{j=1}^J \left[(1 - \alpha) \|\theta_j\|_2 + \alpha \|\theta_j\|_1 \right] \right\}$$

with $\alpha \in [0, 1]$. Much like the elastic net, the parameter α creates a bridge between the group lasso ($\alpha = 0$) and the lasso ($\alpha = 1$).

Since the optimization problem is convex, its optima are specified by zero subgradient equations, similar to for the group lasso. More precisely, any optimal solution must satisfy the condition

$$-\mathbf{Z}_j^T \left(\mathbf{y} - \sum_{\ell=1}^J \mathbf{Z}_\ell \hat{\theta}_\ell \right) + \lambda(1 - \alpha) \cdot \hat{s}_j + \lambda \alpha \hat{t}_j = 0, \text{ for } j = 1, \dots, J$$

where $\hat{s}_j \in \mathbb{R}^{p_j}$ belongs to the subdifferential of the Euclidean norm at $\hat{\theta}_j$ and $\hat{t}_j \in \mathbb{R}^{p_j}$ belongs to the subdifferential of the ℓ_1 -norm at $\hat{\theta}_j$; in particular, we have each $\hat{t}_{jk} \in \text{sign}(\theta_{jk})$ as with the usual lasso.

We once again solve these equations via block-wise coordinate descent, although the solution is a bit more complex than before. As in Equation $-\mathbf{Z}_j^T (\mathbf{r}_j - \mathbf{Z}_j \hat{\theta}_j) + \lambda \hat{s}_j = 0$ in group lasso, with \mathbf{r}_j the partial residual in the j^{th} coordinate, it can be seen that $\hat{\theta}_j = 0$ if and only if the equation

$$\mathbf{Z}_j^T \mathbf{r}_j = \lambda(1 - \alpha) \hat{s}_j + \lambda \alpha \hat{t}_j$$

has a solution with $\|\hat{s}_j\|_2 \leq 1$ and $\hat{t}_{jk} \in [-1, 1]$ for $k = 1, \dots, p_j$. Fortunately, this condition is easily checked, and we find that (Exercise 4.12)

$$\hat{\theta}_j = 0 \quad \text{if and only if} \quad \left\| \mathcal{S}_{\lambda\alpha} \left(\mathbf{Z}_j^T \mathbf{r}_j \right) \right\|_2 \leq \lambda(1 - \alpha)$$

where $\mathcal{S}_{\lambda\alpha}(\cdot)$ is the soft-thresholding operator applied here componentwise to its vector argument $\mathbf{Z}_j^T \mathbf{r}_j$. Notice the similarity with the conditions for the group lasso, except here we use the soft-thresholded gradient $\mathcal{S}_{\lambda\alpha}(\mathbf{Z}_j^T \mathbf{r}_j)$. Likewise, if $\mathbf{Z}_j^T \mathbf{Z}_j = \mathbf{I}$, then as shown in Exercise 4.13, we have

$$\hat{\theta}_j = \left(1 - \frac{\lambda(1-\alpha)}{\|\mathcal{S}_{\lambda\alpha}(\mathbf{Z}_j^T \mathbf{r}_j)\|_2}\right)_+ \mathcal{S}_{\lambda\alpha}(\mathbf{Z}_j^T \mathbf{r}_j)$$

In the general case when the \mathbf{Z}_j are not orthonormal and we have checked that $\hat{\theta}_j \neq 0$, finding $\hat{\theta}_j$ amounts to solving the subproblem

$$\text{minimize}_{\theta_j \in \mathbb{R}^{p_j}} \left\{ \frac{1}{2} \|\mathbf{r}_j - \mathbf{Z}_j \theta_j\|_2^2 + \lambda(1-\alpha) \|\theta_j\|_2 + \lambda\alpha \|\theta_j\|_1 \right\}$$

Here we can again use generalized gradient descent (Section (5.3.3)) to produce a simple iterative algorithm to solve each block, as in Equation (4.16a). The algorithm would iterate until convergence the sequence

$$\begin{aligned} \omega &\leftarrow \theta_j + \nu \cdot \mathbf{Z}_j^T (\mathbf{r}_j - \mathbf{Z}_j \theta_j), \text{ and} \\ \theta_j &\leftarrow \left(1 - \frac{\nu\lambda(1-\alpha)}{\|\mathcal{S}_{\lambda\alpha}(\omega)\|_2}\right)_+ \mathcal{S}_{\lambda\alpha}(\omega) \end{aligned}$$

where ν is the step size.

2.1.8 Overlap group lasso

The overlap group lasso is a modification that allows variables to contribute to more than one group. 会有两个问题：

1. 属于多个组的变量的参数是各组上的参数估计的和，故这些变量更加不会为 0
2. 如果一个组的被压缩到零，则重叠的变量在其他组的值也都为零

Jacob, Obozinski and Vert (2009) recognized this problem, and hence proposed the replicated variable approach or overlap group lasso. In general, the sets of possible nonzero coefficients always correspond to groups, or the unions of groups. They also defined an implicit penalty on the original variables that yields the replicated variable approach as its solution, which we now describe.

Denote by $\mathbf{v}_j \in \mathbb{R}^p$ a vector which is zero everywhere except in those positions corresponding to the members of group j , and let $\mathcal{V}_j \subseteq \mathbb{R}^p$ be the subspace of such possible vectors. In terms of the original variables, $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_p)$, the coefficient vector is given by the sum $\beta = \sum_{j=1}^J \mathbf{v}_j$, and hence the overlap group lasso solves the problem

$$\min_{\mathbf{v}_j \in \mathcal{V}_j, j=1, \dots, J} \left\{ \frac{1}{2} \left\| \mathbf{y} - \mathbf{X} \left(\sum_{j=1}^J \mathbf{v}_j \right) \right\|_2^2 + \lambda \sum_{j=1}^J \|\mathbf{v}_j\|_2 \right\}$$

This optimization problem can be re-cast in the terms of the original β variables by defining a suitable penalty function. With

$$\Omega_{\mathcal{V}}(\beta) := \inf_{\substack{v_j \in \mathcal{V}_j \\ \beta = \sum_{j=1}^J v_j}} \sum_{j=1}^J \|v_j\|_2$$

it can then be shown (Jacob et al. 2009) that solving problem is equivalent to solving

$$\text{minimize}_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \Omega_{\mathcal{V}}(\beta) \right\}$$

This equivalence is intuitively obvious, and underscores the mechanism underlying this penalty; the contributions to the coefficient for a variable are distributed among the groups to which it belongs in a norm-efficient manner. There are two rings corresponding to the two groups, with X_2 in both groups.

2.1.9 岭回归

根据 Gauss-Markov 定理可知, 在满足线性回归模型的基本假设条件下, 用最小二乘得到的回归系数估计量是最小方差线性无偏估计。并且存在多重共线性时, 也并不影响这个性质的成立, 所以在有偏估计中寻找方差更加小的估计量。

在自变量存在相关性的情况下, 容易造成 $|X'X| = 0$, 这样的系数的估计 $\hat{\beta} = (X'X)^{-1} X'y$ 会退化, 难以计算 (解释), 并且系数的方差 $\sigma^2 (X'X)^{-1}$ 会变得很大。岭回归一种修正的最小二乘估计方法, 提供了一个比最小二乘更加稳定的估计, 并且回归系数的标准差会变小。

考虑给病态矩阵 $X'X$ 加上一个正常的单位矩阵 kI , 那么 $X'X + kI$ 的奇异程度会小于 $X'X$ 的奇异程度, 这样定义 $\hat{\beta}(k) = (X'X + kI)^{-1} X'y$ 为 β 的岭回归估计, 这里 k 为惩罚参数, 控制惩罚程度以及相关阵的退化程度。实际上这个定义等价于

$$\hat{\beta} = \arg \min_{\beta} \|y - X\beta\|^2 + \lambda \|\beta\|_{l_2}$$

这里 $\|\beta\|_{l_2} = \sum_{i=1}^p \beta_i^2$ 也就是相当于对系数的 l_2 惩罚的似然方法, 对系数的平方进行惩罚。给出等价形式:

$$\beta = \arg \min \sum_{i=1}^N \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \quad \text{s.t.} \quad \sum_j \beta_j^2 \leq t$$

则其惩罚残差平方和 (PRSS) 为:

$$\text{PRSS}(\beta) = \sum (Y - X\beta)^2 + \lambda \|\beta\|_2^2$$

因次 PRSS 是凸函数, 因此其存在全局唯一解 (显示解), 对 PRSS 求关于 β 的导数, 并令函数等于 0:

$$\frac{\partial \text{PRSS}}{\partial \beta} = -2X^T(Y - X\beta) + 2\lambda\beta = 0$$

解得：

$$\hat{\beta}_{\lambda}^R = (X^T X + \lambda I_p)^{-1} X^T Y$$

由上式得出如下结论：

1. $\hat{\beta}_{\lambda}^R$ 是有封闭解的, 而不像 $\hat{\beta}^L$ 只有软阈解. $X^T X + \lambda I_p$ 中包括 λI_p 可以保证即使 $X^T X$ 是不可逆的, $X^T X + \lambda I_p$ 仍可逆, 从而保证解一定存在.
2. $\hat{\beta}_{\lambda}^R$ 是有偏的: 令 $R = X^T X$

$$\begin{aligned}\hat{\beta}_{\lambda}^R &= (X^T X + \lambda I_p)^{-1} X^T Y = (R + \lambda I_p)^{-1} R (R^{-1} X^T Y) \\ &= [R (I_p + \lambda R^{-1})]^{-1} R [(X^T X)^{-1} X^T Y] \\ &= (I_p + \lambda R^{-1})^{-1} R^{-1} \hat{\beta}^0 \\ &= (I_p + \lambda R^{-1}) \hat{\beta}^0\end{aligned}$$

由上式可得

$$E(\hat{\beta}_{\lambda}^R) = E[(I_p + \lambda R^{-1}) \hat{\beta}^0] = (I_p + \lambda R^{-1}) \hat{\beta} \neq \beta (if \lambda \neq 0)$$

3. Ridge 惩罚无法保证稀疏模型: 由 $\hat{\beta}_{\lambda}^R = (X^T X + \lambda I_p)^{-1} X^T Y$ 可见, Ridge Penalty 的解仅是一种特殊情况下的 OLS 解, 无本质区别.

Proposition 2.1.4 — 解释岭回归. 对于惩罚参数所起的作用, 我们从下面可以看出 (如何推导):

$$\hat{\beta}(k) = (X'X + kI)^{-1} X'y = \frac{1}{1+k} \begin{pmatrix} 1 & \frac{\rho_{12}}{1+k} & \frac{\rho_{1p}}{1+k} \\ & 1 & \cdot \\ & & 1 & \frac{\rho_{p-1,p}}{1+k} \\ & & & 1 \end{pmatrix}^{-1} X'y$$

从上面我们可以看出变量之间的相关性都减小了, 压缩了 $\frac{1}{1+k}$, 惩罚参数在一定意义表示的就是**降低相关性的程度**. 我们从一个新的角度来看这个问题:

$$\begin{aligned}L(\lambda) &= \|y - X\beta\|_2^2 + \lambda \|\beta\|_2 \\ &= y'y - 2y'X\beta + \beta'X'X\beta + \lambda\beta'\beta \\ &= y'y - 2y'X\beta + \beta'(X'X + \lambda I_p)\beta \\ &= y^{*'}y^* - 2y^{*'} \begin{pmatrix} X \\ \sqrt{\lambda}I_{p*p} \end{pmatrix} \beta + \beta' \begin{pmatrix} X \\ \sqrt{\lambda}I_{p*p} \end{pmatrix}' \begin{pmatrix} X \\ \sqrt{\lambda}I_{p*p} \end{pmatrix} \beta\end{aligned}$$

这里我们重新定义符号 $y^{*'} = (y', 0, 0 \dots 0)_{(n+p)*1}$, $X^* = \begin{pmatrix} X \\ \sqrt{\lambda}I_{p*p} \end{pmatrix}$ 这样岭回归就可以写成一般最小二乘的形式, 如下:

$$L(\lambda) = \|y^* - X^*\beta\|^2$$

重新定义的符号, X^* 的秩为 p , 是满秩防止了退化的情形 (原来是退化, 加上一个纯量矩阵使得其不是退化的)。所以根据上式, 我们可以理解为岭回归就是增加一些样本点 (原来的 n 个样本, 变成了 $(n+p)$ 个样本), 增加样本点也具有减小方差的作用。

我们以二元自变量为例来看: 假设 x_1, x_2, y 都已经中心化处理, $\text{var}(\hat{\beta}) = \sigma^2 (X'X)^{-1}$, 化简上式的得到二元情形下的方差估计为

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{(1 - r_{12}^2) \sum_{i=1}^n x_{i1}^2}$$

, 其中 r_{12} 为两个自变量之间的相关系数, 从上式可以看出如果变量之间完全相关, 即 $r_{12} = 1$, 方差就会变得无穷大。在相关系数不变的情况下, 我们增大样本容量会使得 $\sum_{i=1}^n x_{i1}^2$ 变大, 从而达到减小方差的目的。从上面可知, 岭回归本质上增加了样本容量, 从变换的形式上可以看出, 增加的样本形式为 $x_j = (0, 0, \dots, 1, 0, \dots, 0), y_j = 0$, 其实可以看出只是增加了单个变量的样本。

从这个角度来看, 我们是等权重增加样本, 我们进一步改善岭回归。由于每个系数对于回归方程所起的作用不同, 我们使用不等权重的方法来增加样本, 即我们增加样本形式为 $x_j = (0, 0, \dots, w_j, 0, \dots, 0), y = 0$, 也就是定义岭回归估计为 $\hat{\beta}(k) = (X'X + kW)^{-1} X'y$, $W = \text{diag}(w_1, \dots, w_p)$ 这里的权重我们可以选择为回归系数的初始最小二乘估计, 或者对于 $p > n$ 的情形我们直接选择为解释变量与因变量的相关系数 $x_j'y$ 作为权重。

Proposition 2.1.5 — 岭回归的变量选择. 对于岭估计解的形式, 可以看出本质上是对每个系数进行了适当的压缩, 但是本质上不能够把不显著的变量易除掉。但是, 我们可以从岭迹 (参数系数估计关于惩罚参数的轨迹) 的角度进行重新看待岭估计的变量选择的作用。

1. 在设计阵已经中心化和标准化情况下, 可以直接比较岭回归系数的大小。可以剔除掉回归系数比较稳定且绝对值很小的自变量。随着惩罚参数的增大, 一定程度上去除了变量之间的相关性, 如果系数比较稳定且很小, 说明变量与其他变量的相关性不大, 并且对因变量的解释程度不显著, 因此可以剔除。
2. 随着惩罚参数的增大, 回归系数比较稳定, 岭迹比较震荡, 如果最后趋近于 0, 说明随着相关程度的减弱, 自变量对于因变量的作用很微弱, 因此要剔除掉这种自变量。

2.1.10 主成分回归

主成分分析 (PCA) 是对原始变量进行线性变换, 得到一组线性无关的潜变量。换言之, 可将主成分分析看做是对坐标系的旋转, 使得数据在第一维坐标方向上含有的信息量最大, 在第二维坐标方向上次之。实际应用中, 可根据需要选取前几个主成分建立模型。

PCR 的一大优势在于其自变量是线性无关的, 避免了线性回归中共线性性的问题。原变量中具有共线性性的变量通过线性变换结合在一起, 使得拟合的模型更容易解释。因此, 当自变量有较强的共线性时, 适宜考虑采用 PCR 建立模型。除此之外, PCR 通过选取少量的潜变量拟合模型, 有效避免了过度拟合的问题, 因而一般来说具有更高的预测准确性。

PCR 中主成分个数的选取与主成分分析有所不同。

1. 主成分分析是以尽可能多地提取变量信息为目的, 多依照主成分解释方差的比例选取变量数目

2. PCR 则是以回归为目的, 多根据舍一验证等方法选取主成分数目。PCR 中主成分个数可以看做调整参数。

具体的计算过程

PCR 主要目的是降维, 利用变量的线性组合来表示成新的变量。如何选取线性组合: 使得新变量含有的信息最多--利用方差来体现, 即寻找一组系数使得新变量的方差最大化。计算第一主成分: $Z_1 = Xa_1$, 这个问题转化为: 求 $a_1 = (a_{11}, a_{21}, \dots, a_{p1})$, 使得在 $a_1' a_1 = 1$ 的约束下, $\text{var}(Z_1)$ 达到最大值, 即包含最多的数据的变异信息。使用 Lagrange 乘子法解决条件极值问题。令

$$\varphi(a_1) = \text{var}(Xa_1) - \lambda_1 (a_1' a_1 - 1) = a_1' \Sigma a_1 - \lambda_1 (a_1' a_1 - 1)$$

分别对 a_1, λ_1 求导可得:

$$\begin{cases} \frac{\partial \varphi}{\partial a_1} = 2(\Sigma - \lambda_1 I) a_1 = 0 \\ \frac{\partial \varphi}{\partial \lambda_1} = a_1' a_1 - 1 = 0 \end{cases}$$

由于 $a_1 \neq 0$, 故 $|\Sigma - \lambda_1 I| = 0$, 所以求解上而方程组其实就是求解 Σ 的特征值与特征向量的问题。 $\text{var}(Z_1) = a_1' \Sigma a_1 = a_1' \lambda_1 a_1 = \lambda_1$, 因此求得 Σ 得最大特征值对应的特征向量即可。

继续求第二主轴 a_2 , 并且与 a_1 标准正交, 包含数据第二大变异方向的成分。这个问题是在 $a_2' a_2 = 1, a_2' a_1 = 0$ 的约束下, 求 $\text{var}(Xa_2)$ 的最大值, 同样使用 Lagrange 乘子法求极值问题, 类似于求解第一主轴的方法, 即 a_2 为 Σ 的特征值对应的特征向量, 由于 $a_1' a_2 = 0$, 这时 λ_2 只能为 Σ 的第二大特征值, 这样求出对应的特征向量即为 a_2 依此类推, 依此可以求出 X 的主成分 Z_1, Z_2, \dots, Z_m , 其中 $Z_k = Xa_k$, 并且 $\text{var}(Z_k) = \lambda_k$, 因此有 $\text{var}(Z_1) \geq \text{var}(Z_2) \geq \dots \geq \text{var}(Z_m)$, 也就是说每个主成分包含的信息逐渐减少。之后对主成分 Z_1, Z_2, \dots, Z_m 建立回归方程, 在用原始变量进行代换。

通常情况下, 由于相关阵的特征值代表了主成分包含信息的程度, 可以使用累计贡献率 $(\lambda_1 + \lambda_2 + \dots + \lambda_m) / \sum_{i=1}^p \lambda_i$ 利用之前选择的阈值来选择主成分的个数。

2.1.11 偏最小二乘方法

偏最小二乘 (PLS) 方法已被广泛应用于候选变量较多的模型选择问题, 具体而言, PLS 方法构建少量与因变量具有很大相关性的正交潜变量, 然后建立因变量和潜变量之间的回归模型。

潜变量 $t^{(k)}$ 是自变量 X_1, \dots, X_p 的线性组合, 在确定了前 k 个潜变量后, $t^{(k+1)}$ 是与之正交且与因变量 Y 有最大相关系数的线性组合。令 X 表示自变量向量 (X_1, \dots, X_p) , Y 表示响应变量, X 表示标准化的设计阵, Y 表示响应变量观测向量。则在正交性约束下, 第 $k+1$ 个潜变量权重系数表达式如下:

$$c^{(k+1)} = \arg \max_{c^T c = 1} \text{cov}(Xc, Y) = \arg \max_{c^T c = 1} \frac{1}{n} c^T X^T Y$$

其中 n 是样本量。上述表达式与主成分分析的求解公式类似。事实上, PLS 方法与 PCR 方法有很多相似之处: 二者都是采用选取少量潜变量代替原始变量, 从而达到降维的目的。

选取潜变量的数目可以看做是调整参数, 随着潜变量数目增加, 数据拟合度相应增加, 但并不意味着预测精度也相应增加。特别地, 当潜变量数目与原变量数目相同时, 选择模型与全模型完全相同。值得注意的是, PLS 的潜变量作为原始变基的线性组合, 其权重不是响应变量和自变量的线性函数, 因而增加了计算的难度。

PLS 和 PCR 的差别在于: 前者选取潜变量是以与响应变量的**相关性**为导向的, 而 PCA 方法则不需要响应变量的信息, 只是选取**方差最大**的线性组合. 从这个角度可以看出, 当降维目的在于提高模型预测的准确性时, 选用 PLS 方法更为明智, 而 PCA 方法则侧重提取自变量的信息。

1. Stone 和 Brooks 将 PLS、PCR 以及 OLS(普通最小二乘) 三种方法统一到同一个理论框架下, 并指出 OLS 方法和 PCR 方法是连续谱的两个极端情形, 而 PLS 介于二者之间。类似于 Bridge 方法, 在该理论框架下, 还可以讨论连续谱对应的其它可能的方法。
2. Datta 等比较了 PLS 方法和 LASSO 方法在高维数据中的应用。他们指出两种方法均适用于变量维数较高的情形, 但当变量中有较多噪音时, LASSO 方法的预测精度较 PLS 更高。

由于主成分回归在提取主成分时, 只是尽可能包含了自变量的信息, 没有考虑到与因变量的解释关系。偏最小二乘回归在这方面改善了主成分回归。在提取主成分的时候, 考虑到了不仅主成分要包含数据的信息, 还要与 y 的相关程度达到最大, 使得主成分对因变量的解释程度达到最大, 与主成分回归中主成分提取的方法有些差别。

对于第一主成分的提取 $Z_1 = Xa_1$, 要在 $a_1'a_1 = 1$ 的约束下, 使得 $\text{cov}(Xa_1, y)$ 达到最大值, 之后类似于主成分回归中主成分的计算方法, 可以知道 $a_1 = \frac{x'y}{\|X'y\|}$, 从这可以看出主成分的系数就是 x_j 与 y 的相关系数的标准化, 也就是相关性越强, 在 Z_1 中的权重就越大, 这也符合实际的解释, 相关性越大, 对于因变量的解释程度越大, 自然权重就越大。之后的主成分提取, 用回归的残差来替代 y 来做回归, 提取剩下的主成分。

至于主成分的提取个数, 观察残差向量在一定程度上就可以看出是否达到我们想要的理想结果。因此偏最小二乘回归也处理了共线性情形下的回归问题, 也达到了降维均效果, 并且对因变量的解释程度达到最大。这样利用与因变量的相关程度达到最大, 也可以有效的去掉那些信息冗余的变量, 去掉了一些相关性强, 但是对于因变量解释比较弱的变量。

2.1.12 Bridge Estimator

Frank and Friedman(1993) 提出 bridge 估计量, 形式为

$$\hat{\beta}_{brdg} = \arg \min_{\beta} \|y - X\beta\|^2 + \lambda \sum_{i=1}^p |\beta_i|^\gamma$$

等价于

$$\hat{\beta}_{brdg} = \arg \min_{\beta} \|y - X\beta\|^2 \text{ subject to } \sum_{i=1}^p |\beta_i|^\gamma \leq t, t \geq 0$$

观察 bridge 估计量的形式我们可以看出当 $\gamma = 2$ 时, 就是岭回归, 当 $\gamma = 1$ 时, 就是 lasso 方法, 这两种经典的方法都是 bridge 估计量的特例。

我们可以计算 bridge 估计量 $\gamma > 1$ 的方差为

$$\text{var}(\hat{\beta}) = (X'X + D(\hat{\beta}))^{-1} X'X (X'X + D(\hat{\beta}))^{-1} \sigma^2 \mathbf{1}$$

其中 $D(\hat{\beta}) = \text{diag}(\lambda\gamma(\gamma-1)|\hat{\beta}|^{\gamma-2}/2)$ 并且可以证明 $\text{var}(\hat{\beta}_{\text{brdg}}) \leq \text{var}(\hat{\beta}_{\text{ols}})$. 从这里可以看出 $\hat{\beta}_{\text{brdg}}$ 的方差要小于普通最小二乘估计量的方差, 这在存在共线性的情况下, 可以起到一定的作用, 通过控制 γ, λ 来达到减小方差的目的, 解决共线性造成的系数估计不稳定的情形。

观察上面的方差形式, 我们也可以看出: 当 $\lambda = 0$ 时, 方差为 $(X'X)^{-1} \sigma^2$, 与普通最小二乘估计的方差形式相同。另外, 当 $\gamma = 2$ 时 $D(\hat{\beta}) = \lambda I$, 方差形式为 $\text{var}(\hat{\beta}) = (X'X + \lambda I)^{-1} X'X (X'X + \lambda I)^{-1} \sigma^2$, 也正好等于岭估计的方差。

Proposition 2.1.6 — 参数选择方法. 为了选择参数 λ, γ , 我们通常选用 GCV(generalized cross validation) 方法, 对于给定的 $\lambda \geq 0, \gamma \geq 1$, 可以通过

$$\left(X'X + \frac{\lambda\gamma}{2} \text{diag}(|\beta_j|^{\gamma-2}) \right) \beta = X'y$$

计算出 GCV 的权重矩阵为 $p(\lambda) = \text{trace} \left(X (X'X + \lambda W)^{-1} X' \right) - p$ 定义

$$\text{GCV} = \frac{\|y - X\hat{\beta}\|^2}{n(1 - p(\lambda)/n)^2}$$

在对 (λ, γ) 网上最小化 GCV 值, 其实也就是最小化预测平方误差。

2.1.13 Elastic net--未知分组的群组模型选择方法

当一组强相关的解释变量同时存在时, 普通的 LASSO 方法倾向于选取其中一个变量。但有的情形下, 我们希望将这一组强相关的变量都选出来。前面提到的 Bridge 方法的惩罚项是严格凸的, 并且具有群组效应, 但是不能实现模型选择。Zou 和 Hastie 结合 LASSO 方法与 Bridge 方法的优点, 提出了既有群组效应又能进行模型选择的 Elastic Net (EN) 方法来解决未知变量分组情况下的组群模型选择。该方法的简单形式如下

$$\|Y - X\beta\|^2 + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2$$

假设有 n 个数据样本, p 个变量, $y = (y_1, y_2, \dots, y_n)'$ 是因变量, X 是设计阵。我们假设因变量以及预测变量中心化, 即

$$\sum_{i=1}^n y_i = 0, \quad \sum_{i=1}^n x_{ij} = 0, \quad \sum_{i=1}^n x_{ij}^2 = 1 \quad \text{for } j = 1, 2, \dots, p$$

实际上, 其也可以看作是惩罚的似然方法, 等价于下面的最优化问题:

$$\hat{\beta} = \arg \min_{\beta} \|y - X\beta\|^2 \text{ subject to } \lambda_1 |\beta|_1 + \lambda_2 |\beta|^2 \leq t$$

惩罚部分是 lasso 与岭回归的线性组合, 是一个严格的凸优化问题。同时施加 L_1 和 L_2 惩罚的最小二乘问题, 并且 Additive lasso 和 elastic net 都具有 oracle 性质。

弹性网的优点表现在:

1. 岭回归部分能很好地处理高度相关的数据，消除变量间的多重共线性；
2. Lasso 部分使得它能够选择变量
3. for any $\alpha < 1$ and $\lambda > 0$, the elastic-net problem $\min_{(\beta_0, \beta) \in \mathbb{R} \times \mathbb{R}^p} \left\{ \frac{1}{2} \sum_{i=1}^N (y_i - \beta_0 - x_i^T \beta)^2 + \lambda \left[\frac{1}{2} (1 - \alpha) \|\beta\|_1 + \frac{\alpha}{2} \|\beta\|_2^2 \right] \right\}$ is strictly convex: a unique solution exists irrespective of the correlations or duplications in the X_j .
4. elastic-net ball shares attributes of the ℓ_2 ball and the ℓ_1 ball: the sharp corners and edges encourage selection, and the curved contours encourage sharing of coefficients.

它改善了 Lasso 存在的三个缺点：

1. Lasso 最终选择的变量个数不会超过样本容量 n ;
2. Lasso 对高度相关的变量进行选择时，只选择其中一个，而不关心是哪一个;
3. 当 $n > p$ 时，若解释变量存在高度相关，经验表明 Lasso 预测性能很差

弹性网的缺点之一是往往选择过多的变量组。由于式中惩罚函数内 Lasso 惩罚的存在，很显然弹性网的参数估计不满足无偏性，因此不具有 Oracle 性质。此外，Jia 和 Yu 研究了 $p \gg n$ 时弹性网的性质，指出在不可表条件以及其他一些复杂条件下，弹性网具有选择一致性和 Oracle 性质。

对数据集 (X, Y) 作如下变换

$$X_{(n+p) \times p}^* = (1 + \lambda_2)^{-\frac{1}{2}} \begin{pmatrix} X \\ \sqrt{\lambda_2} I \end{pmatrix}, \quad Y_{(n+p)}^* = \begin{pmatrix} Y \\ 0 \end{pmatrix}$$

其中 I 是 $p \times p$ 的单位矩阵。记 $\gamma = \frac{\lambda_1}{\sqrt{1+\lambda_2}}, \beta^* = \sqrt{1+\lambda_2} \beta$, 则可将 EN 方法的最优化问题转化为 LASSO 方法的最优化

$$\|Y^* - X^* \beta^*\|^2 + \gamma \sum_{j=1}^p |\beta_j^*|$$

下面我们讨论其解的形式，对于固定的 λ_1, λ_2 我们对目标函数 $L(\lambda_1, \lambda_2, \beta)$ 关于每个 β_j 求导数，对于 $\beta_j \neq 0$ 时，目标函数可导，但是对于 $\beta_j = 0$ 时，目标函数不可导，所以求其左右导数，类似于 lasso 的求法，我们可以得到如下的形式：（为了方便分析，这里我们只给出在自变量正交情形下的解的形式）

$$\hat{\beta}_j(\text{naive}) = \frac{(|\hat{\beta}_j(\text{OLS})| - \lambda_1/2)_+}{1 + \lambda_2} \text{sgn}\{\hat{\beta}_j(\text{OLS})\}$$

这里 $\hat{\beta}_j(\text{OLS}) = X_j^T y$ 。观察岭估计解的形式 $\hat{\beta}_j(\text{ridge}) = \hat{\beta}_j(\text{OLS}) / (1 + \lambda_2)$ 以及 lasso 解的形式 $\hat{\beta}_j(\text{lasso}) = (|\hat{\beta}_j(\text{OLS})| - \lambda_1/2)_+ \text{sgn}(\hat{\beta}_j(\text{OLS}))$ 可以看出 elastic net 的解是 lasso 以及岭回归的综合形式，某种程度上类似于两步估计，先进行 lasso 估计进行变量选择，然后在对其进行岭估计对系数进行压缩，处理共线性。但是模拟表明除非当 λ_1, λ_2 的选择靠近岭估计或者 lasso 估计时，elastic net 的效果并不理想。因此我们需要对其解进行改进。另外，从解本身来看，两者的估计都是有偏估计，elastic net 方法相当于进行了两次有偏

估计, 这样不仅不会帮助减小方差, 反而增大了估计的方差。这样我们试图去恢复一次偏差处理, 重新定义 elastic net 的解为

$$\hat{\beta}_j(\text{elastic net}) = (1 + \lambda_2) \hat{\beta}_j(\text{naive})$$

实际上, 对初始的解恢复了岭估计的偏, 减小了估计的偏差。

上述参数估计可视为 LASSO 估计 (参数为 λ_1) 与岭回归估计 (参数为 λ_2) 的结合, 经历了两次系数缩减的过程。

1. 将 Elastic Net 惩罚转变成 Lasso 惩罚, 因此可以用计算 Lasso 惩罚估计量的方法计算 Elastic Net 惩罚估计量.
2. 克服了 Lasso 至多选择 n 个变量的缺点, X^* 这个样本的容量是 $n + p$, 而 X^* 的秩最多为 p , 因此通过 Lasso 最多可以选出 p 个变量, 这一点在 $p > n$ 的情况下很有用.

Elastic net 与 lasso 的等价性

$$\hat{\beta}^* = \arg \min_{\beta^*} \|y^* - X^* \beta^*\| + \gamma \|\beta^*\|_1$$

这个定理的原理与之前对岭回归的讨论相似, 只不过增加了 l_1 惩罚部分--也就是在变量替换后的岭回归的基础上再加上一个惩罚。从这个定理我们可以看出, l_2 惩罚部分可以看作是一部分增补人造的数据样本, 这样减小了估计的方差 (原来的岭回归的惩罚就是 l_2)。

从另一个角度来看这个问题, 通过数据增补以后, 我们可以看出设计阵变成了 $(n + p) * p$ 维, 并且秩为 p , 这意味着 elastic net 原理上可以选择 p 个变量, 克服了 lasso 的第一个问题 (当 $p > n$ 时, lasso 只能选择 n 个变量)。

对于这里的数据增补问题, 我们也可以把它看作是给定了 β 一些先贝叶斯验信息。我们使用不等权重的方法来增加样本, 即我们增加样本形式为 $x_j = (0, 0, \dots, w_j, 0, \dots, 0)$, $y = 0$, 也就是定义 elastic net 估计为 $\hat{\beta} = \arg \min_{\beta} |y - X\beta|^2 + \lambda_2 W|\beta|^2 + \lambda_1 |\beta|_1$, $W = \text{diag}(w_1, \dots, w_p)$ 。这里的权重我们可以选择为回归系数的初始最小二乘估计, 或者对于 $p > n$ 的情形我们直接选择解释变量与因变量的相关系数 $x_j'y$ 作为权重。也就是相当于对每个系数指定的先验分布不是等方差的。

Elastic Net 可以选出群组变量

一个好的变量选择和参数估计方法对于组效应来说, 对于高度相关的变量的系数估计应该近似相等, 或者在极端的情形下, 对于完全相关总变量的参数估计应该相同。elastic net 本质上是严格的凸优化问题, 可以得到如下引理:

Lemma 2.2 假设 $x_i = x_j$ 对于任意 $i, j \in \{1, \dots, p\}$, 如果优化问题是严格凸的, 那么可以得到 $\hat{\beta}_i = \hat{\beta}_j$, 对于任意的 $\lambda > 0$ ■

下证 Elastic Net 可以选出群组变量: 对于给定的数据集 (Y, X) , 和惩罚系数 (λ_1, λ_2) , 记 $\hat{\beta}(\lambda_1, \lambda_2)$ 为初始 Elastic Net 估计量, 定义路径差异系数如下:

$$D_{\lambda_1, \lambda_2} = \frac{1}{|Y|} |\hat{\beta}_i(\lambda_1, \lambda_2) - \hat{\beta}_j(\lambda_1, \lambda_2)|$$

则当 $\hat{\beta}_i(\lambda_1, \lambda_2) \hat{\beta}_j(\lambda_1, \lambda_2) > 0$ (该假设前提是合适, 因为对于这两个群组变量而言, 其系数必定非 0 且同号), 由 $\frac{\partial PRSS(\beta)}{\partial \lambda_i} = 0, i = 1, 2$, 得到两个方程, 做差可得

$$(x_j^T - x_i^T)(Y - X\beta) + \lambda_2(\hat{\beta}_i - \hat{\beta}_j) = 0$$

上式等价于

$$\hat{\beta}_i - \hat{\beta}_j = \frac{1}{\lambda_2} (x_i^T - x_j^T) \hat{r}(\lambda_1, \lambda_2)$$

其中, $\hat{r}(\lambda_1, \lambda_2) = Y - X\hat{\beta}$ 是残差向量. 因为 X 已经标准化, 因此 $|x_i - x_j|^2 = 2(1 - \rho)$. 根据 β 的定义可知 $PRSS(\hat{\beta}) \leq PRSS(\beta = 0)$, 即 $|\hat{r}|^2 + \lambda_1|\beta| + \lambda_2|\beta|^2 \leq |Y|^2$, 从上式可以得到 $|\hat{r}| \leq |Y|$, 因此可以推导出以下方程

$$D_{\lambda_1, \lambda_2}(i, j) \leq \frac{1}{\lambda_2} \sqrt{2(1 - \rho_{i,j})}$$

从上述公式意味着相关系数越大的两个变量, 其系数路径差异越小, 被同时选出的可能性越大, 因此 Elastic Net 可以选出群组变量. 在得到 $\hat{\beta}^*$ 后, 令 $\hat{\beta} = \frac{1}{\sqrt{1+\lambda_2}} \hat{\beta}^*$, $\beta(\text{Elastic net}) = (1 + \lambda_2) \hat{\beta}^*$.

这里 $\rho = x_i'x_j$, 表示两个解释变量的样本相关系数. $D_{\lambda_1, \lambda_2}(i, j)$ 用来度量两个变量系数估计的距离, 这个定理给出了这个距离的定量的描述, 给出了一个确定的上界, 之间的距离与两个预测变量的相关系数相关, 相关系数越大, 变量之间相关性越大, 那么它们系数的估计相差就越小.

一个特例, 当 $\rho = 1$ 时, $D_{\lambda_1, \lambda_2}(i, j) = 0$ 这与上面的引理的结论一致. 因此, 从这个定理可以看出, elastic net 方法具有组效应, 对于相关的变量给出的系数估计大致相同. 而 lasso 方法并不具有组效应, 由于其并不是严格的凸优化问题, 因此并不具有唯一解, 那么高度相关的变量之间的距离很难度量, 也不会大致相等, 这是 lasso 的弱点, 由以下引理说明.

Lemma 2.3 — Lasso. 假设 $x_i = x_j$ 对于任意 $i, j \in \{1, \dots, p\}$, 优化问题的解 $\hat{\beta} = \arg \min_{\beta} \|y - X\beta\|^2 + \lambda \|\beta\|_1$, 那么 $\hat{\beta}_i \hat{\beta}_j \geq 0$ 并且 $\hat{\beta}^*$ 也是优化问题的解

$$\hat{\beta}^* = \begin{cases} \hat{\beta}_k & \text{if } k \neq i \text{ and } k \neq j \\ (\hat{\beta}_i + \hat{\beta}_j) * s & \text{if } k = i \\ (\hat{\beta}_i + \hat{\beta}_j) * (1 - s) & \text{if } k = j \end{cases}$$

处理共线性问题

类似之前的讨论, elastic net 方法是 lasso 与 ridge 方法的综合, 因此在处理共线性的方法上类似 ridge 回归的方法, 利用惩罚因子来降低变量之间的相关性, 并且在一定程度上增加了些人造样本, 减小了方差, 这在 $p > n$ 情形时体现尤其明显. 之前我们的讨论都是限于 $X'X$ 变量正交的情形, 下面我们给出一般情况下的解.

Theorem 2.1.7 给定数据 (X, y) 以及参数 (λ_1, λ_2) , elastic net 的解为:

$$\hat{\beta} = \arg \min_{\beta} \beta' \left(\frac{X'X + \lambda_2 W}{1 + \lambda_2} \right) \beta - 2y'X\beta + \lambda_1 |\beta|_1$$

这个定理我们可以看出, elastic net 可以看作是 lasso 更加稳健的形式, 降低了共线性造成的影响。注意到去除共线性的样本相关阵为

$$\frac{X'X + \lambda_2 W}{1 + \lambda_2} = (1 - \gamma)\Gamma + \gamma W$$

其中 $\gamma = \lambda_2 / (1 + \lambda_2)$, 通过 λ_2 来控制相关性的处理程度, 可以向 W 靠近。这样通过降低相关性, 也提高了预测的精确度。

Elastic net 求解算法

Zou and Hastie(2005) 提出了适用 Elastic Net 的算法 LARS-EN。Elastic Net 的不足有如下几点:

1. Elastic Net 惩罚不具有神谕性, 因此 Zou and Zhang 提出了自适应 Elastic Net 惩罚 (Adaptive Elastic Net), 该惩罚在保留 Elastic Net 惩罚选择群组变量优势的同时, 具有神谕性, 因此自适应 Elastic Net 惩罚的性质更加优良。
2. Anbari and Mkhadri 通过模拟分析发现, 当群组变量之间的相关程度 $|\rho| \leq 0.95$ 时, Elastic Net 选择群组变量的结果变得不那么可靠。针对这一缺点, Algamil and Lee(2015) 提出了调整 Elastic Net 惩罚 (Adjusted Elastic Net) 和自适应调整 Elastic Net 惩罚 (Adaptive Adjusted Elastic Net), 分别针对较弱相关程度和中等相关程度的群组变量, 并且能保证变量选择的一致性。

Coordinate descent

The coordinate descent update for the j^{th} coefficient takes the form

$$\hat{\beta}_j = \frac{\mathcal{S}_{\lambda\alpha} \left(\sum_{i=1}^N r_{ij} x_{ij} \right)}{\sum_{i=1}^N x_{ij}^2 + \lambda(1 - \alpha)} \quad (2.5)$$

where $\mathcal{S}_{\mu}(z) := \text{sign}(z)(z - \mu)_+$ is the soft-thresholding operator, and $r_{ij} := y_i - \hat{\beta}_0 - \sum_{k \neq j} x_{ik} \hat{\beta}_k$ is the partial residual. We cycle over the updates (4.4) until convergence.

2.1.14 适应性的 elastic net 方法及其大样本性质

对于 elastic net 方法, 也同样遇到 lasso 的问题, 是一种有偏估计, 并且在很情形下并不是估计相合的, 因此 Zou and Zhang(2009) 提出了适应的 elastic net 方法, 在 l_1 惩罚部分对每个参数乘以不同的权重。

我们假设 $\hat{\mathcal{A}}_{en} = \{j : \hat{\beta}_j \neq 0\}$ 表示在 elastic net 估计不为 0 的系数, 定义适应的 elastic net 解形式如下:

$$\hat{\beta}_{\hat{\mathcal{A}}_{en}} = (1 + \lambda_2) \left\{ \arg \min_{\beta} \left\| y - X_{\hat{\mathcal{A}}_{en}} \beta \right\|^2 + \lambda_1^* \sum_{j \in \hat{\mathcal{A}}_{en}} \hat{w}_j |\beta_j| + \lambda_2 \|w^* \beta\|^2 \right\}$$

这样适应的 elastic net 既可以处理共线性问题, 在适当的正则条件下又具备 adaptive lasso 的 oracle 性质。下面我们假设如下正则条件:

1. 设 $\lambda_{\min}, \lambda_{\max}$ 表示矩阵的最小及最大特征值, 假设

$$b \leq \lambda_{\min}(X'X) \leq \lambda_{\max}(X'X) \leq B$$

这里 b, B 是两个正常数。

- 2.

$$\lim_{n \rightarrow \infty} \frac{\max_{i=1,2,\dots,n} \sum_{j=1}^p x_{ij}^2}{n} = 0$$

3. $E[\epsilon^{2+\delta}] < \infty$ 对于某个 $\delta > 0$

4. $\lim_{n \rightarrow \infty} \frac{\log(p)}{\log(n)} = \nu$ 对于某个 $0 \leq \nu < 1$

- 5.

$$\lim_{n \rightarrow \infty} \frac{\lambda_2}{n} = 0, \quad \lim_{n \rightarrow \infty} \frac{\lambda_1}{\sqrt{n}} = 0$$

并且

$$\lim_{n \rightarrow \infty} \frac{\lambda_1^*}{\sqrt{n}} = 0, \quad \lim_{n \rightarrow \infty} \frac{\lambda_1^*}{\sqrt{n}} n^{\frac{(1-\nu)(1+\gamma)-1}{2}} = \infty$$

- 6.

$$\lim_{n \rightarrow \infty} \frac{\lambda_2}{\sqrt{n}} \sqrt{\sum_{j \in \mathcal{A}} \beta_j^{*2}} = 0, \quad \lim_{n \rightarrow \infty} \min \left(\frac{n}{\lambda_1 \sqrt{p}}, \left(\frac{\sqrt{n}}{\sqrt{p} \lambda_1^*} \right)^{\frac{1}{\gamma}} \right) \left(\min_{j \in \mathcal{A}} |\beta_j^*| \right) \rightarrow \infty$$

其中 (a1)(a2) 是假设设计阵的稳定性, (a3) 是为了利用 Lyapunov 中心极限定理的必要条件, (a4) 说明了变量个数可以随着样本个数而增大, 但是速度要受到限制。(a5)(a6) 是建立了一些关于惩罚参数的约束, 也是为了相合性的需要。在上述正则条件的约束下, 我们可以得到适应得 elastic net 的 oracle 性质。

Theorem 2.1.8 在上述的 (a1)-(a6) 正则条件下, 适应的 elastic net 方法具有 oracle 性质, 也就是 $\hat{\beta}(\text{aden})$ 满足

1. 变量选择的相合性: $\Pr(\{j: \hat{\beta}(\text{aden})_j \neq 0\} = \mathcal{A}) \rightarrow 1$
2. 估计的渐进正态性:

$$\alpha' \frac{I + \lambda_2 \Sigma_A^{-1}}{1 + \frac{\lambda_2}{n}} \Sigma_A^{\frac{1}{2}} (\hat{\beta}(\text{aden})_{\mathcal{A}} - \beta_{\mathcal{A}}^*) \rightarrow_d N(0, \sigma^2)$$

这里 $\Sigma_{\mathcal{A}} = X'_{\mathcal{A}} X_{\mathcal{A}}, \|\alpha\|_{l_2} = 1$

这个定理给出了在变量个数随着样本数目发散的情形下的 oracle 性质, 改善了一般的 elastic net 的性质。对于权重的选择, 一般为 $\hat{w}_j = (|\hat{\beta}_j(en)|^{-\gamma}, \gamma > 0$

2.1.15 Mnet

Mnet 是处理高度线性相关数据的另一种惩罚方法，其惩罚估计为

$$\hat{\beta}^{Mnet} = \operatorname{argmin} \{L(\beta | y, X) + P_{Mnet}(\lambda, a, |\beta|)\}$$

$$P_{Mnet}(\lambda, a, |\beta|) = \sum_{j=1}^p f_{\lambda_1, a}^{MCP}(|\beta_j|) + \frac{1}{2} \lambda_2 \sum_{j=1}^2 \beta_j^2$$

其中 MCP 函数形式如下

$$f_{\lambda, a}^{MCP}(\theta) = \begin{cases} \lambda\theta - \frac{\theta^2}{2a}, & \theta \leq a\lambda, \\ \frac{a\lambda^2}{2}, & \theta > a\lambda, \end{cases} \quad f'_{\lambda, a}^{MCP}(\theta) = \begin{cases} \lambda - \frac{\theta}{a}, & \theta \leq a\lambda \\ 0, & \theta > a\lambda \end{cases}$$

$|\beta| < a\lambda$ 时, MCP 函数的一阶导数随着 $|\beta|$ 增大而减小, 即 $|\beta|$ 增大, 惩罚函数上升越缓慢; 当 $|\beta| > a\lambda$ 时, **惩罚函数的一阶导数 0, 即对大的回归系数不惩罚**。这改善了 Lasso 过度惩罚大系数的缺点。Huang 等模拟分析表明, Mnet 方法在处理高度相关问题时比弹性网更具优势, 文中分别研究了 $p > n$ 及 $p < n$ 时 Mnet 方法的性质, 指出在某些合理的条件下 (包括能使 Mnet 正确区分零和非零系数的条件), Mnet 估计表现得与 Oracle 岭估计一样, 且两者的估计值符号相同的概率趋向于 1, 因此得出结论“Mnet 以很高的概率等于 Oracle 岭估计”, 这种意义下 Mnet 能正确选择具有非零系数的解释变量且用岭回归来估计相应的参数, 因此具有 Oracle 性质。

2.1.16 Fused Lasso

Tibshirani(2005) 年提出了 Fused Lasso, 用于处理有序变量数据结构, 其形式如下:

$$\hat{\beta} = \operatorname{argmin}_{\beta} \sum_{i=1}^N \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2, \sum_{j=1}^p |\beta_j| \leq t_1, \sum_{j=2}^p |\beta_j - \beta_{j-1}| \leq t_2 \quad (2.6)$$

上述方程中第一个不等式保证了模型的稀疏性, 第二个不等式保证了系数间的稀疏性, 因为后者是对相邻系数的差距进行惩罚, 如果差异过大, 则说明这两个系数原本并不是相邻变量, 因此该惩罚项使得局部系数平滑。第二个不等式的思路是来自于 Land and Friedman (1996), 后者提出了 $\sum_j |\beta_j - \beta_{j_1}|^\alpha \leq t_2$ 形式的惩罚函数, 并着重讨论了当 $\alpha = 0, 1, 2$ 时的情况, 但是他们没有将 $\sum_j |\beta_j - \beta_{j_1}|$ 和 $\sum_j |\beta_j|$ 同时结合起来考察。We can generalize the notion of neighbors from a linear ordering to more general neighborhoods, for examples adjacent pixels in an image. This leads to a penalty of the form

$$\lambda_2 \sum_{i \sim i'} |\theta_i - \theta_{i'}| \quad (2.7)$$

where we sum over all neighboring pairs $i \sim i'$

Problem (2.6) and its relatives are all convex optimization problems, and so all have well-defined solutions. As in other problems of this kind, here we seek efficient **path algorithms** for finding solutions for a range of values for the tuning parameters. Although coordinate descent is one of our favorite algorithms for lasso-like problems, it

need not work for the fused lasso (2.6), because the **difference penalty is not a separable function of the coordinates**. Consequently, coordinate descent can become "stuck" at a non-optimal point.

We begin by considering the structure of the optimal solution $\hat{\theta}(\lambda_1, \lambda_2)$ of the fused lasso problem (2.6) as a function of the two regularization parameters λ_1 and λ_2 . The following result due to Friedman et al. (2007) provides some useful insight into the behavior of this optimum:

Lemma 2.4 For any $\lambda'_1 > \lambda_1$, we have

$$\hat{\theta}_i(\lambda'_1, \lambda_2) = \mathcal{S}_{\lambda'_1 - \lambda_1}(\hat{\theta}_i(\lambda_1, \lambda_2)) \text{ for each } i = 1, \dots, N$$

where \mathcal{S} is the soft-thresholding operator $\mathcal{S}_\lambda(z) := \text{sign}(z)(|z| - \lambda)_+$ ■

One important special case of Lemma 2.4 is the equality

$$\hat{\theta}_i(\lambda_1, \lambda_2) = \mathcal{S}_{\lambda_1}(\hat{\theta}_i(0, \lambda_2)) \text{ for each } i = 1, \dots, N$$

Consequently, if we solve the fused lasso with $\lambda_1 = 0$, all other solutions can be obtained immediately by soft thresholding. On the basis of Lemma 2.4, it suffices to focus our attention on solving the problem ⁸

$$\underset{\theta \in \mathbb{R}^N}{\text{minimize}} \left\{ \frac{1}{2} \sum_{i=1}^N (y_i - \theta_i)^2 + \lambda \sum_{i=2}^N |\theta_i - \theta_{i-1}| \right\} \quad (2.8)$$

We consider several approaches to solving (2.8).

Reparametrization

One simple approach is to reparametrize problem (2.8) so that the penalty is additive. In detail, suppose that we consider a linear transformation of the form $\gamma = \mathbf{M}\theta$ for an invertible matrix $\mathbf{M} \in \mathbb{R}^{N \times N}$ such that

$$\gamma_1 = \theta_1, \text{ and } \gamma_i = \theta_i - \theta_{i-1} \text{ for } i = 2, \dots, N$$

In these transformed coordinates, the problem (2.8) is equivalent to the ordinary lasso problem

$$\underset{\gamma \in \mathbb{R}^N}{\text{minimize}} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{X}\gamma\|^2 + \lambda \|\gamma\|_1 \right\}, \text{ with } \mathbf{X} = \mathbf{M}^{-1} \quad (2.9)$$

In principle, the reparametrize problem (2.9) can be solved using any efficient algorithm for the lasso, including coordinate descent, projected gradient descent or the LARS procedure. However, \mathbf{X} is a lower-triangular matrix with all nonzero entries equal to 1, and hence has large correlations among the "variables." Neither coordinate-descent nor LARS performs well under these circumstances. So despite the fact that reparametrization appears to solve the problem, it is not recommended, and more efficient algorithms exist, as we now discuss.

A Path Algorithm

The one-dimensional fused lasso (2.8) has an interesting property, namely that as the regularization parameter λ increases, pieces of the optimal solution can only be joined together, not split apart. More precisely, letting $\hat{\theta}(\lambda)$ denote the optimal solution to the convex program (2.8) as a function of λ , we have:

Lemma 2.5 — Monotone fusion. Suppose that for some value of λ and some index $i \in \{1, \dots, N-1\}$, the optimal solution satisfies $\hat{\theta}_i(\lambda) = \hat{\theta}_{i+1}(\lambda)$. Then for all $\lambda' > \lambda$, we also have $\hat{\theta}_i(\lambda') = \hat{\theta}_{i+1}(\lambda')$ ■

Friedman et al. (2007) observed that this fact greatly simplifies the construction of the fused lasso solution path. One starts with $\lambda = 0$, for which there are no fused groups, and then computes the smallest value of λ that causes a fused group to form. The parameter estimates for this group are then fused together (i.e., constrained to be equal) for the remainder of the path. Along the way, a simple formula is available for the estimate within each fused group, so that the resulting procedure is quite fast, requiring $\mathcal{O}(N)$ operations. However, we note that the monotone-fusion property in Lemma 2.5 is special to the one-dimensional fused lasso (2.8). For example, it does not hold for the general fused lasso

$$\begin{aligned} \text{minimize}_{(\beta_0, \beta) \in \mathbb{R} \times \mathbb{R}^p} & \left\{ \frac{1}{2} \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 \right. \\ & \left. + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=2}^p |\beta_j - \beta_{j-1}| \right\} \end{aligned}$$

with a model matrix \mathbf{X} , nor for the two-dimensional fused lasso (2.7). See Friedman et al. (2007) and Hoeffling (2010) for more details on this approach.

A Dual Path Algorithm

Tibshirani and Taylor (2011) take a different approach, and develop path algorithms for the convex duals of fused lasso problems. Here we illustrate their approach on the problem (2.8), but note that their methodology applies to the general problem

$$\begin{aligned} \text{minimize}_{(\beta_0, \beta) \in \mathbb{R} \times \mathbb{R}^p} & \left\{ \frac{1}{2} \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 \right. \\ & \left. + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=2}^p |\beta_j - \beta_{j-1}| \right\} \end{aligned}$$

as well.

We begin by observing that problem (2.8) can be written in an equivalent lifted form

$$\text{minimize}_{(\theta, \mathbf{z}) \in \mathbb{R}^N \times \mathbb{R}^{N-1}} \left\{ \frac{1}{2} \|\mathbf{y} - \boldsymbol{\theta}\|_2^2 + \lambda \|\mathbf{z}\|_2 \right\} \text{ subject to } \mathbf{D}\boldsymbol{\theta} = \mathbf{z} \quad (2.10)$$

where we have introduced a vector $\mathbf{z} \in \mathbb{R}^{N-1}$ of auxiliary variables, and \mathbf{D} is a $(N-1) \times N$ matrix of first differences. Now consider the Lagrangian associated with the lifted

problem, namely

$$L(\boldsymbol{\theta}, \mathbf{z}; \mathbf{u}) := \frac{1}{2} \|\mathbf{y} - \boldsymbol{\theta}\|_2^2 + \lambda \|\mathbf{z}\| + \mathbf{u}^T (\mathbf{D}\boldsymbol{\theta} - \mathbf{z}) \quad (2.11)$$

where $\mathbf{u} \in \mathbb{R}^{N-1}$ is a vector of Lagrange multipliers. A straightforward computation shows that the Lagrangian dual function \mathcal{Q} takes form

$$\mathcal{Q}(\mathbf{u}) := \inf_{(\boldsymbol{\theta}, \mathbf{z}) \in \mathbb{R}^N \times \mathbb{R}^{N-1}} L(\boldsymbol{\theta}, \mathbf{z}; \mathbf{u}) = \begin{cases} -\frac{1}{2} \|\mathbf{y} - \mathbf{D}^T \mathbf{u}\|^2 & \text{if } \|\mathbf{u}\|_\infty \leq \lambda \\ -\infty & \text{otherwise} \end{cases} \quad (2.12)$$

The Lagrangian dual problem is to maximize $\mathcal{Q}(\mathbf{u})$, and given an optimal solution $\hat{\mathbf{u}} = \hat{\mathbf{u}}(\lambda)$, we can recover an optimal solution $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\lambda)$ to the original problem by setting $\hat{\boldsymbol{\theta}} = \mathbf{y} - \mathbf{D}^T \hat{\mathbf{u}}$. See Exercise 4.23 for the details of these duality calculations.

When the regularization parameter λ is sufficiently large, the dual maximization, or equivalently the problem of minimizing $-\mathcal{Q}(\mathbf{u})$, reduces to an unrestricted linear regression problem, with optimal solution

$$\mathbf{u}^* := (\mathbf{D}\mathbf{D}^T)^{-1} \mathbf{D}\mathbf{y} \quad (2.13)$$

The restrictions kick in when λ decreases to the critical level $\|\mathbf{u}^*\|_\infty$. Tibshirani and Taylor (2011) show that as we decrease λ , once elements $\hat{u}_j(\lambda)$ of the optimal solution hit the bound λ , then they are guaranteed to never leave the bound. This property leads to a very straightforward path algorithm, similar in spirit to LARS in Section 5.6; see Figure 4.9 for an illustration of the dual path algorithm in action. Exercise 4.23 explores some of the details.

Dynamic Programming for the Fused Lasso

Dynamic programming is a computational method for solving difficult problems by breaking them down into simpler subproblems. In the case of the one-dimensional fused lasso, the linear ordering of the variables means that fixing any variable breaks down the problem into two separate subproblems to the left and right of the fixed variable. In the "forward pass," we move from left to right, fixing one variable and solving for the variable to its left, as a function of this fixed variable. When we reach the right end, a backward pass then gives the complete solution.

Johnson (2013) proposed this dynamic programming approach to the fused lasso. In more detail, we begin by separating off terms in (2.8) that depend on θ_1 , and rewrite the objective function (2.8) in the form

$$f(\boldsymbol{\theta}) = \underbrace{\frac{1}{2} (y_1 - \theta_1)^2 + \lambda |\theta_2 - \theta_1|}_{g(\theta_1, \theta_2)} + \left\{ \frac{1}{2} \sum_{i=2}^N (y_i - \theta_i)^2 + \lambda \sum_{i=3}^N |\theta_i - \theta_{i-1}| \right\} \quad (2.14)$$

This decomposition shows the subproblem to be solved in the first step of the forward pass: we compute $\hat{\theta}_1(\theta_2) := \arg \min_{\theta_1 \in \mathbb{R}} g(\theta_1, \theta_2)$. We have thus eliminated the first variable, and can now focus on the reduced objective function $f_2 : \mathbb{R}^{N-1} \rightarrow \mathbb{R}$ given by

$$f_2(\theta_2, \dots, \theta_N) = f(\hat{\theta}_1(\theta_2), \theta_2, \dots, \theta_N) \quad (2.15)$$

We can then iterate the procedure, maximizing over θ_2 to obtain $\hat{\theta}_2(\theta_3)$, and so on until we obtain $\hat{\theta}_N$. Then we back-substitute to obtain $\hat{\theta}_{N-1} = \hat{\theta}_{N-1}(\hat{\theta}_N)$ and so on for the sequences $\hat{\theta}_{N-2}, \dots, \hat{\theta}_2, \hat{\theta}_1$. If each parameter θ_i can take only one of K distinct values, then each of the minimizers $\hat{\theta}_j(\theta_{j+1})$ can be easily computed and stored as a $K \times K$ matrix. In the continuous case, the functions to be minimized are piecewise linear and quadratic, and care must be taken to compute and store the relevant information in an efficient manner; see Johnson (2013) for the details. The resulting algorithm is the fastest that we are aware of, requiring just $\mathcal{O}(N)$ operations, and considerably faster than the path algorithm described above. Interestingly, if we change the ℓ_1 difference penalty to an ℓ_0 , this approach can still be applied, despite the fact that the problem is no longer convex.

Trend Filtering

The first-order absolute difference penalty in the fused lasso can be generalized to use a higher-order difference, leading to the problem

$$\text{minimize}_{\theta \in \mathbb{R}^N} \left\{ \frac{1}{2} \sum_{i=1}^N (y_i - \theta_i)^2 + \lambda \cdot \left\| \mathbf{D}^{(k+1)} \theta \right\|_1 \right\} \quad (2.16)$$

This is known as trend filtering. Here $\mathbf{D}^{(k+1)}$ is a matrix of dimension $(N - k - 1) \times N$ that computes discrete differences of order $k + 1$. The fused lasso uses first-order differences ($k = 0$), while higher-order differences encourage higher-order smoothness. In general, trend filtering of order k results in solutions that are piecewise polynomials of degree k . Linear trend filtering ($k = 1$) is especially attractive, leading to piecewise-linear solutions. The knots in the solution need not be specified but fall out of the convex optimization procedure. Kim, Koh, Boyd and Gorinevsky (2009) propose an efficient interior point algorithm for this problem. Tibshirani (2014) proves that the trend filtering estimate adapts to the local level of smoothness much better than smoothing splines, and displays a surprising similarity to locally-adaptive regression splines. Further, he shows that the estimate converges to the true underlying function at the minimax rate for functions whose k^{th} derivative is of bounded variation (a property not shared by linear estimators such as smoothing splines). Furthermore, Tibshirani and Taylor (2011) show that a solution with m knots has estimated degrees of freedom given by $\text{df} = m + k + 1$.⁹ As a comparison, we include the fit of a smoothing spline, with the same effective $\text{df} = 4$. While the fits are similar, it appears that trend filtering has found natural change-points in the data.

In (2.16) it is assumed that the observations occur at evenly spaced positions. The penalty can be modified (Tibshirani 2014) to accommodate arbitrary (ordered) positions x_i as follows:

$$\text{minimize}_{\theta \in \mathbb{R}^N} \left\{ \frac{1}{2} \sum_{i=1}^N (y_i - \theta_i)^2 + \lambda \cdot \sum_{i=1}^{N-2} \left| \frac{\theta_{i+2} - \theta_{i+1}}{x_{i+2} - x_{i+1}} - \frac{\theta_{i+1} - \theta_i}{x_{i+1} - x_i} \right| \right\} \quad (2.17)$$

It compares the empirical slopes for adjacent pairs, and encourages them to be the same.

Nearly Isotonic Regression

Tibshirani 2, Hoefling and Tibshirani (2011) suggest a simple modification of the one-dimensional fused lasso that encourages the solution to be monotone. It is based on a relaxation of isotonic regression. In the classical form of isotonic regression, we estimate $\theta \in \mathbb{R}^N$ by solving the constrained minimization problem

$$\text{minimize}_{\theta \in \mathbb{R}^N} \left\{ \sum_{i=1}^N (y_i - \theta_i)^2 \right\} \text{ subject to } \theta_1 \leq \theta_2 \leq \dots \leq \theta_N \quad (2.18)$$

The resulting solution gives the best monotone (nondecreasing) fit to the data. Monotone nonincreasing solutions can be obtained by first flipping the signs of the data. There is a unique solution to problem (2.18), and it can be obtained using the pool adjacent violators algorithm (Barlow, Bartholomew, Bremner and Brunk 1972), or PAVA for short.

Nearly isotonic regression is a natural relaxation, in which we introduce a regularization parameter $\lambda \geq 0$, and instead solve the penalized problem

$$\text{minimize}_{\theta \in \mathbb{R}^N} \left\{ \frac{1}{2} \sum_{i=1}^N (y_i - \theta_i)^2 + \lambda \sum_{i=1}^{N-1} (\theta_i - \theta_{i+1})_+ \right\} \quad (2.19)$$

The penalty term penalizes adjacent pairs that violate the monotonicity property, that is, having $\theta_i > \theta_{i+1}$. When $\lambda = 0$, the solution interpolates the data, and letting $\lambda \rightarrow \infty$, we recover the solution to the classical isotonic regression problem (2.18). Intermediate values of λ yield nonmonotone solutions that trade off monotonicity with goodness-of-fit; this trade-off allows one to assess the validity of the monotonicity assumption for the given data sequence. The solution to the nearly isotonic problem (2.19) can be obtained from a simple modification of the path algorithm discussed previously, a procedure that is analogous to the PAVA algorithm for problem (2.18); see Tibshirani et al. (2011) for details.

Tibshirani(2005) 根据 Gill et al(1997) 提出的两步有效集 SQOPT 算法 (two-phase active set algorithm SQOPT) Fused Lasso 可以用于处理临近信号具有群组特征的信号, 如对于一维分段常数的信号:

$$\hat{\beta} = \arg \min_{\beta} \sum_i (y_i - x\beta_i)^2 + \lambda \sum_i |\beta_i - \beta_{i-1}|$$

其中, y 是带有噪声的信号, β 是待估计的去噪声的分段常数信号: Tibshirani and Wang(2007) 将 Fused Lasso 用于处理比较基因组杂交 (CGH) 数据中的热点发现问题。Fused Lasso 的泛化形式被称为 Generalized Fused Lasso, 其函数形式是:

$$\hat{\beta} = \arg \min_{\beta} \sum_{i=1}^N \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{(i,j) \in E} |\beta_i - \beta_j|$$

其中, E 是指标集。

2.1.17 Graph Lasso

因为用图形或者网络结术进行信息展示是一种非常普遍的方法, 所以 Li and Li(2008) 提出了 Network-constrained 惩罚, 试图将这种先验信息与数据分析相结合; 首先, 对于带权重的网络结构图而言, 可以记为 $G = (V, E, W)$, 其中 $V = 1, \dots, p$ 是顶点的集合, 对应着 p 个自变量; $E = u \sim v$ 是连接自变量 u 和 v 的边的集合; W 是边的权重的集合, $w(u, v)$ 是顶点 u 和 v 之间的权重, 所谓权重是指, 当顶点之间没有方向时, 即在无向图中, 从某顶点出发, 到达其相邻顶点的概率, 在实践过程中, 这个概率一般是通过大规模试验得到频率进行估计; 作者对 G 定义了归一化后的拉普拉斯算子矩阵 K , 并给出了 Graph Lasso 在一般线性回归模型中的定义:

$$L(\lambda_1, \lambda_2, \beta) = \|Y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \beta^T K \beta$$

上式的第一项惩罚保证了模型的稀疏性, 第二项惩罚保证了估计量的平滑性. Graph Lasso 与 Fused Lasso 在一定程度上存在相似性, 两者都试图获得平滑的系数, 但是 Fused Lasso 没有像 Graph Lasso 那样包含图的先验信息, 且 Fused Lasso 采用的是 L_1 正则化, 因此对于相连接的自变量, 其可能得到相同的估计系数。

2.2 介绍求解 LASSO 的算法

Osborne et al(2000) 提出的 Dual Algorithm, Efron et al(2004) 提出的 LARS 算法, Friedman et al.(2007, 2010) 和 Wu and Lange(2008) 提出的 Coordinate Descent 算法。

Dual Algorithm

Osborne et al(2000) 提出了 Dual Algorithm, 并证明可以用于计算任何情况下的 Lasso 解, 本文仅展示当 X 为正交阵时, 该算法的步骤和结果. 令 $\lambda = 2\gamma$, 上式为:

$$\begin{aligned} \hat{\beta}_j^L &= \text{sign}(\hat{\beta}^0) \left(\hat{\beta}_j^0 - \gamma \right)^+, j = 1, \dots, p \\ \text{记 } \sum_j |\hat{\beta}_j^L| &= t, \quad \sum_j |\hat{\beta}_j^0| = t_0, \text{ 则} \\ t_0 - t &= \sum_j |\hat{\beta}_j^0| - \sum_j \left(|\hat{\beta}_j^0| - \gamma \right)^+ \\ &= \sum_j |\hat{\beta}_j^0| I(|\hat{\beta}_j^0| \leq \gamma) + \gamma \sum_j I(|\hat{\beta}_j^0| > \gamma) \\ &= \sum_j b_j + \gamma(p - K) \end{aligned}$$

其中, p 是自变量的个数, $b_1 \leq b_2 \leq \dots \leq b_p$ 是 $|\hat{\beta}_1^0|, |\hat{\beta}_2^0|, \dots, |\hat{\beta}_p^0|$ 的有序统计 $k = \max j: b_j \leq \gamma$. 由上述定义, 有 $t < t_0, K < p$ 和 $b_K \leq \gamma \leq b_{K+1}$. 该算法的具体步骤如下:

1. 令 $c_0 = 0$
2. j 从 1 到 p 进行迭代, 记 $c_j = \sum_{i=1}^j b_i + b_j(p-j), 0 = c_0 \leq c_1 \leq \dots \leq c_p = t_0$
3. 用 GCV 计算出 t 后, 记 $K = \max \{i: c_i \leq t_0 - t\}$
4. 令 $\gamma = (t_0 - t) - \sum_{i=1}^K b_i / (p - K)$
5. 计算出 γ 后, $\hat{\beta}_j^L = \text{sign}(\hat{\beta}_j^0) (\hat{\beta}_j^0 - \gamma)^+ j = 1, \dots, p$

2.2.1 LARS 算法

Efron et al(2004) 提出了 LARS 算法, 并将 LARS 运用于 Lasso 惩罚中, 其具体步骤如下:

1. 对 Y 去中心化, 对 X 进行标准化, 令所有自变量的系数为 0;
2. 找到与残差相关性最大的自变量 x_j
3. 将 x_j 的系数沿着 OLS 的方向前进, 直到出现另一个新的自变量 x_k , 其与当前残差的相关性等于 x_j 与残差的相关性;
4. 将 x_j 和 x_k 的系数 β_j 和 β_k 一起沿着 x_k 的 OLS 方向前进, 直到再次有新的变量被选入模型;
5. 重复步骤 (2)-(4), 直到残差低于某指定值.

保证了准确性和速度。什么叫 OLS 方向? 该方法同时直观地解释了 LASSO 最多选出 $\min(n, p)$ 个协变量的原因, 这是因为当 $n \ll p$ 时, 假设已经选入了 n 个变量, 若再选入新变量, 则其可以用之前变量线性表示, 无法定义角平分线, 也就无法实现算法。

LARS 与经典的逐步向前变量选择算法 (Forward stepwise selection) 有着密切的联系。向前法的思想是变量由少到多, 每次增加一个, 直至没有变量可以引入为止。但此方法却有一个明显缺点: 由于各自变量之间可能存在着相关关系, 因此后续变量的选入可能会使前面已选入的自变量变得不重要, 而向前法又不考虑从已选变量中剔除不重要的变量, 这样最后得到的最优子集可能包含一些对因变量影响不大的自变量。

逐段向前法 (Forward stagewise) 比逐步向前法更加谨慎, 该算法每次都要在所选变量的对应系数上增加或减小一个微量, 其他系数保持不变。这样的过程可以重复, 直到所有残差都为 0 或者系数等于 0。因此这种算法可能需要上千步才能得出最终的模型。LARS 算法结合了这两种算法的长处, 可以用来计算 Lasso 估计, 并且计算量不大。

LARS 方法的步骤大概如下: 与向前法类似, 先设所有协变量的系数为零, 从中选择一个与响应变量相关性最大的, 以 x_1 为例, 然后沿着 x_1 的方向取最大的步长, 直到另一个变量 (例如 x_2) 与当前的残差有同样多的相关性。接下来, LARS 方法不是沿着 x_1 的方向, 而是沿着这两个向量的等角线向前运动, 直到第三个变量与当前的残差有同样多的相关性, 然后, 沿着与三个向量等角的方向继续下去, 即“最小角方向”, 直到第四个变量进入“最相关集合”, 依次类推。其等角性使得其相对于逐段向前法在计算迭代的步长时变得容易。

下面先给出一些符号, 然后给出该算法的代数表达式。设协变量 x_1, x_2, \dots, x_p 是线性

独立的, A 是下标集 $\{1, 2, \dots, p\}$ 的一个子集, 定义矩阵

$$\mathbf{X}_A = (\cdots, s_j x_j, \cdots)_{j \in A}$$

其中符号 $s_j = \pm 1$ (取法见下). 令

$$\mathcal{G}_A = \mathbf{X}'_A \mathbf{X}_A, \quad A_A = \left(\mathbf{1}'_A \mathcal{G}_A^{-1} \mathbf{1}_A \right)^{-\frac{1}{2}}$$

其中 $\mathbf{1}_A$ 是分量均为 1, 维数与 \mathbf{X}_A 相同的向量, 则等角向量

$$\mathbf{u}_A = \mathbf{X}_A \omega_A$$

是单位向量, 且与 \mathbf{X}_A 各列之间的角度相等 (并小于 90°), 其中 $\omega_A = A_A \mathcal{G}_A^{-1} \mathbf{1}_A$

$$\mathbf{X}'_A \mathbf{u}_A = A_A \mathbf{1}_A, \quad \text{且 } \|\mathbf{u}_A\|^2 = 1$$

记 $\hat{\boldsymbol{\mu}} = \mathbf{X} \hat{\boldsymbol{\beta}}$, 将响应变量中心化, 从 $\hat{\boldsymbol{\mu}}_0 = \mathbf{0}$ 开始逐步建立 $\hat{\boldsymbol{\mu}}$. 设 $\hat{\boldsymbol{\mu}}_A$ 是均值的当前 LARS 估计, 记

$$\hat{\mathbf{c}} = \mathbf{X}'(\mathbf{y} - \hat{\boldsymbol{\mu}}_A) = (\hat{c}_1, \dots, \hat{c}_p)'$$

它是表示当前残差与 \mathbf{X} 相关性的向量. 选取具有最大绝对当前相关的变量, 令其下标构成运动集 A , 即

$$\hat{C} = \max_j \{|\hat{c}_j|\}, \quad A = \{j: |\hat{c}_j| = \hat{C}\}$$

令 $s_j = \text{sign}\{\hat{c}_j\}$, $j \in A$, 按照公式 (15) – (17) 计算 \mathbf{X}_A, A_A 和 \mathbf{u}_A , 以及内积向量

$$\mathbf{a} = \mathbf{X}' \mathbf{u}_A = (a_j)$$

LARS 算法的下一步就是更新 $\hat{\boldsymbol{\mu}}_A$, 即

$$\hat{\boldsymbol{\mu}}_{A+} = \hat{\boldsymbol{\mu}}_A + \hat{\gamma} \mathbf{u}_A$$

其中

$$\hat{\gamma} = \min_{j \in A^c} + \left\{ \frac{\hat{C} - \hat{c}_j}{A_A - a_j}, \frac{\hat{C} + \hat{c}_j}{A_A + a_j} \right\}$$

这里 \min^+ 表示只在正值中取最小值. 从而新的运动集为 $A_+ = A \cup \{j\}$, 新的最大绝对相关为 $\hat{C}_+ = \hat{C} - \hat{\gamma} A_A$. 该算法只需要 p 步 (全模型中的变量个数) 便可结束.

将 LARS 算法进行修改, 可以用来寻找 Lasso 估计. LARS 文献引理 7 证明了均值的 Lasso 估计满足

$$\hat{\boldsymbol{\mu}}(\lambda) = \hat{\boldsymbol{\mu}}(\lambda_0) + A_A(\lambda - \lambda_0) \mathbf{u}_A, \quad \lambda \in \mathcal{T}$$

其中 u_A 是等角向量, T 是 λ 轴的开区间, 起点记作 λ_0 。即 $\hat{\mu}(\lambda)$ 沿着由 A 决定的等角向量做线性运动。上式亦可表述为: $\hat{\beta}$ 的非零分量和 λ 之间是分段线性的关系

$$\hat{\beta}_A(\lambda) = \hat{\beta}_A(\lambda_0) + S_A A_A (\lambda - \lambda_0) \omega_A$$

其中 S_A 是对角元素为 $s_j, j \in A$ 的对角矩阵。另外, LARS 文献指出最大绝对相关 $\hat{C}(\lambda)$ 是参数 λ 逐段线性下降的函数。从而, 任给一个 λ_0 值, 从估计值 $\hat{\beta}(\lambda_0)$ 开始, 沿着 λ 增加的方向, 寻找折线 $\hat{\beta}(\lambda)$ 的转折点, 即可确定运动集 A 。然后再用最小二乘法估计模型的系数。

2.2.2 Coordinate Descent 算法

下面考虑坐标下降法的计算复杂度。在每一轮的迭代中, 由于要更新所有方向, 并且每个方向有若干个样本, 因此模型一共的计算量为 $O(np)$, n 为样本数, p 为所有的方向数。考虑 β 的第 s 个分量 β_s , 先对第一项求偏导,

$$\begin{aligned} \|y - x\beta\|_2^2 &= \sum_{i=1}^n \left(y_i - \sum_{j=1}^p x_{ij}\beta_j \right)^2 \\ &= \sum_{i=1}^n \left(y_i^2 + \left(\sum_{j=1}^p x_{ij}\beta_j \right)^2 - 2y_i \left(\sum_{j=1}^p x_{ij}\beta_j \right) \right) \\ \frac{\partial \|y - x\beta\|_2^2}{\partial \beta_s} &= \sum_{i=1}^n \left(2 \left(\sum_{j=1}^p x_{ij}\beta_j \right) x_{is} - 2y_i x_{is} \right) \\ &= -2 \sum_{i=1}^n x_{is} \left(y_i - \sum_{j=1}^p x_{ij}\beta_j \right) \\ &= -2 \sum_{i=1}^n x_{is} \left(y_i - \sum_{j \neq s} x_{ij}\beta_j - x_{is}\beta_s \right) \\ &= -2 \sum_{i=1}^n x_{is} \left(y_i - \sum_{j \neq s} x_{ij}\beta_j \right) + 2\beta_s \sum_{i=1}^n x_{is}^2 \\ &= -2\rho_s + 2\beta_s z_s \end{aligned}$$

其中 $\rho_s = \sum_{i=1}^n x_{is} (y_i - \sum_{j \neq s} x_{ij}\beta_j)$, $z_s = \sum_{i=1}^n x_{is}^2$ 。当上式为零时, 取得的解即为普通最小二乘估计 $\hat{\beta}_s^{OLS} = \rho_s / z_s$ 第二项在原点为尖点, 因此在原点不可导。故可以使用次梯度方法, 如下:

$$\frac{\partial \lambda \sum_{j=1}^p |\beta_j|}{\partial \beta_s} = \begin{cases} \lambda & \text{若 } \beta_s > 0 \\ [-\lambda, \lambda] & \text{若 } \beta_s = 0 \\ -\lambda & \text{若 } \beta_s < 0 \end{cases}$$

则对目标函数求偏导, 并使其等于 0

$$0 = \frac{\partial \|y - X\beta\|_2^2}{\partial \beta_s} + \frac{\partial \lambda \sum_{j=1}^p |\beta_j|}{\partial \beta_s} = \begin{cases} -2\rho_s + 2\beta_s z_s + \lambda & \text{若 } \beta_s > 0 \\ [-2\rho_s - \lambda, -2\rho_s + \lambda] & \text{若 } \beta_s = 0 \\ -2\rho_s + 2\beta_s z_s - \lambda & \text{若 } \beta_s < 0 \end{cases}$$

可推出解为

$$\beta_s = \begin{cases} (\rho_s - \frac{\lambda}{2}) / z_s & \text{若 } \rho_s > \frac{\lambda}{2} \\ 0 & \text{若 } \rho_s \in [-\frac{\lambda}{2}, \frac{\lambda}{2}] \\ (\rho_s + \frac{\lambda}{2}) / z_s & \text{若 } \rho_s < -\frac{\lambda}{2} \end{cases}$$

类似的, 可以计算岭回归的解为

$$\hat{\beta}_s^{\text{ridge}} = \frac{\rho_s}{z_s + \lambda}$$

考虑第 j 个系数, 我们根据 KKT 条件得到该出的最优解满足 $\hat{\beta}_j = \text{sign}(x_j^T r_j) \frac{(x_j^T r_j - \frac{\lambda}{2})_+}{\|x_j\|_2}$
 $r_j = y - \sum_{k \neq j}^p x_k \hat{\beta}_k$ 于是, 我们便可以利用上式来迭代求解 LASSO 模型。

坐标下降法 (Coordinate Descent) 由 Wu and Lange(2008) 提出, 其本质是一个迭代算法, 单次迭代仅针对一个坐标 (即 β 的一个分量), 如在第 t 次迭代中选择 β 的第 k 个分量进行计算, 则有如下式子:

$$\beta_k^{t+1} = \arg \min_{\beta_k} f(\beta_1^t, \dots, \beta_{k-1}^t, \beta_k, \beta_{k+1}^t, \dots, \beta_p^t)$$

其中对于 $j \neq k$ 的分量, $\beta_j^{t+1} = \beta_j^t$. 这个方法的泛化情形之一便是块坐标下降法 (block coordinate descent), 所谓块是指将自变量分为若干个互不重叠的块, 然后每次迭代仅针对其中一块。

该方法的思想是在每次迭代过程中只优化一个参数, 而保持其他参数的值不变, 循环直到所有的参数收敛到给定精度。该方法特别适合求解一维情况下有解析解但是高维情形下不存在的优化问题。其基本流程如下:

1. 给定初始值 $\beta_0^T = (\beta_0^{(1)T}, \dots, \beta_0^{(J)T})$ 和收敛精度, 记已循环次数 $s = 0$
2. 对每个 $j \in (1, \dots, J)$, 在固定 $\beta_s^{(k)} (k \neq j)$ 的情况下, 求解目标函数得到第 j 组变量的参数估计值, 记为 $\beta_{s+1}^{(j)}$, 令 $\beta_s^{(j)} = \beta_{s+1}^{(j)}$
3. 更新 s 至 $s + 1$, 重复第 (2) 步直到收敛。

Block coordinate descent 法最早出现在文献中, 用于求解线性模型下的 Group Lasso. 对于广义线性模型, Meier 等先将损失函数用二次函数逼近, 再用该算法求解 logistic 回归下的 GroupLasso. 也可采用坐标下降法求解 Mnet。

关于该算法的收敛性, Tseng 指出, 即使目标函数不是严格凸的, 只要其不可微部分是可分的, 那么该算法就会收敛。本文中的方法, 损失函数 $L(\beta | y, X)$ 是严格凸函数, 而惩罚函数 $P_\lambda(|\beta|)$ 不一定可微, 例如 Group Lasso 的惩罚, 但是该惩罚在组组之间是可分的, 即 $P_\lambda(|\beta|) = \sum_{j=1}^J P_\lambda(|\beta^{(j)}|)$, 因此目标函数一定会收敛到全局最优解。

Tseng(1988,2001) 证明了对于任何凸的代价函数 (cost function), 只要函数具有可分结构 (separable structure), 则坐标下降法能保证找到全局最小值。

正交情况下的 Lasso 的解

记

$$g(\beta) = (Y - X\beta)^T(Y - X\beta) + \lambda|\beta| = Y^T Y - \beta^T X^T Y - Y^T X\beta + \beta^T X^T X\beta + \lambda|\beta|$$

将上式对 β 求偏导, 得到

$$\frac{\partial g(\beta)}{\partial \beta} = -2Y^T X + 2(X^T X)\beta + \lambda \text{sign}(\beta)$$

假设 X 为正交矩阵 ($X^T X = I$), 令 $\frac{\partial g(\beta)}{\partial \beta} = 0$, 并且由于 $\hat{\beta}_j^0 = Y^T X$, 则

$$\hat{\beta}_j^L = \hat{\beta}_j^0 - \frac{\lambda}{2} \text{sign}(\hat{\beta}_j^L)$$

显而易见, $\hat{\beta}_j^L$ 与 $\hat{\beta}_j^0$ 同号, 因此 $\text{sign}(\hat{\beta}_j^L) = \text{sign}(\hat{\beta}_j^0)$, 方程变为

$$\hat{\beta}_j^L = \hat{\beta}_j^0 - \frac{\lambda}{2} \text{sign}(\hat{\beta}_j^0)$$

上述方程可以写成软阈值函数 (soft thresholding):

$$\hat{\beta}_j^L = \text{sign}(\hat{\beta}_j^0) \left(\hat{\beta}_j^0 - \frac{\lambda}{2} \right)^+$$

由上式可知, $\frac{\lambda}{2}$ 一个阈值, 对于估计值小于 $\frac{\lambda}{2}$ 的系数, Lasso 直接将其压缩至 0, 该系数对应的参数将直接从模型中被剔除, 从而达到变量选择的目的. 并且从上式可以看出, 变量选择和变量估计是同时完成的, 即在捕选变量的同时, 完成对系数的估计.

上述的推导的法二: Assume $\frac{1}{n} X^T X = 1$, and consider

$$y_i = \beta x_i + \varepsilon_i$$

The solution of Lasso is

$$\hat{\beta} = \arg \min_{\beta} \frac{1}{2n} \|Y - X\beta\|_2^2 + \lambda|\beta|$$

Take the partial derivative of β from the above formula

$$\begin{aligned} -\frac{1}{n} X^T (Y - X\beta) + \lambda \text{sign}(\beta) &= 0 \\ -\frac{1}{n} X^T Y + \frac{1}{n} X^T X\beta + \lambda \text{sign}(\beta) &= 0 \end{aligned}$$

Since $\frac{1}{n} X^T X = 1$ and we define $Z = \frac{1}{n} X^T Y$, we obtain

$$\begin{aligned} \beta + \lambda \text{sign}(\beta) &= \frac{1}{n} X^T Y = Z \\ (|\beta| + \lambda) \text{sign}(\beta) &= |z| \text{sign}(z) \end{aligned}$$

Since $\lambda > 0, |\beta| > 0$, and $|z| > 0$, we have $\text{sign}(\beta) = \text{sign}(z)$. Hence

$$\begin{aligned}\beta &= z - \lambda \text{sign}(\beta) \\ &= z - \lambda \text{sign}(z) \\ &= (|z| - \lambda) \text{sign}(z)\end{aligned}$$

And

$$|\beta| \text{sign}(z) = (|z| - \lambda) \text{sign}(z)$$

So

$$\hat{\beta} = \begin{cases} (|z| - \lambda) \text{sign}(z) & |z| \geq \lambda \\ 0 & |z| < \lambda \end{cases}$$

2.2.3 SCAD

$$p_\lambda(|\beta_j|) = \begin{cases} \lambda |\beta_j|, & 0 \leq |\beta_j| < \lambda \\ -(|\beta_j|^2 - 2a\lambda |\beta_j| + \lambda^2 / [2(a-1)]) , & \lambda \leq |\beta_j| < a\lambda \\ (a+1)\lambda^2 / 2, & |\beta_j| \geq a\lambda \end{cases}$$

惩罚函数 $p_\lambda(|\beta|)$ 对 β 求导得到

$$p'_\lambda(|\beta_j|) = \lambda \left\{ I(\beta_j \leq \lambda) + \frac{(a\lambda - \beta_j)^+}{(a-1)\lambda} I(\beta_j > \lambda) \right\}$$

其中 $a > 2, \beta_j > 0$, 最后我们可以由此得到 SCAD 的估计, 用 $\hat{\beta}_j^{\text{SCAD}}$ 表示为

$$\hat{\beta}_j^{\text{SCAD}} = \begin{cases} \text{sign}(\hat{\beta}_j^0) (|\hat{\beta}_j^0| - \lambda)^+, & |\hat{\beta}_j^0| \leq 2\lambda \\ [(a-1)\hat{\beta}_j^0 - \text{sign}(\hat{\beta}_j^0)a\lambda] / (a-2), & 2\lambda \leq |\hat{\beta}_j^0| < a\lambda \\ \hat{\beta}_j^0, & |\hat{\beta}_j^0| \geq a\lambda \end{cases}$$

Lasso 进行选择变量的时候在面对比较小的系数时也会施加相对于它较大的压缩, 这样会出现一定程度的过度压缩问题, SCAD 的惩罚项是一个分段函数, 所得到的 SCAD 估计也是分段形式的, 一定程度上解决了这个问题。

$$[p(|x|)]' = p'(|x|) \text{sgn}(x) \approx \{p'(|x_0|) / |x_0|\} x \quad (2.20)$$

when $x \neq 0$ (actually it should be that both $x, x_0 \neq 0$, shouldn't it?) and the approximation leads to (2.20) follows from a case analysis, according to whether x is positive or negative: thus

$$\text{sign}(x) = \frac{x}{|x|}$$

if $x \neq 0$. Applying this, we find

$$p'(|x|) \cdot \text{sign}(x) \approx p'(|x|) \times \frac{x}{|x|} \approx p'(|x_0|) \times \frac{x}{|x_0|}$$

which justifies (2.20).

$$p(|x|) \approx p(|x_0|) + \frac{1}{2} \{ p'(|x_0|) / |x_0| \} (x^2 - x_0^2), \text{ for } x \approx x_0$$

Proof. (2) follows from a first-order Taylor series approximation for $x^2 - x_0^2$. Notice that if $x \approx x_0$, we have

$$x^2 - x_0^2 \approx 2(x - x_0)x_0$$

(Why? Because when $\epsilon \approx 0$, we have $(x_0 + \epsilon)^2 \approx x_0^2 + 2\epsilon x_0$, thus $(x_0 + \epsilon)^2 - x_0^2 \approx 2\epsilon x_0$.) In other words,

$$x - x_0 \approx \frac{x^2 - x_0^2}{2x_0} \approx \frac{x^2 - x_0^2}{2x}$$

Now we can use this to derive (2). By a Taylor series expansion, we know

$$p(|x|) \approx p(|x_0|) + (x - x_0) \cdot p'(|x_0|) \cdot \text{sign}(x_0)$$

Plug in (1), and we find

$$p(|x|) \approx p(|x_0|) + (x - x_0) \times p'(|x_0|) \times \frac{x}{|x_0|}$$

Finally, plug in our approximation for $x - x_0$ and find

$$p(|x|) \approx p(|x_0|) + \frac{x^2 - x_0^2}{2x} \times p'(|x_0|) \times \frac{x}{|x_0|}$$

which simplifies to

$$p(|x|) \approx p(|x_0|) + \frac{x^2 - x_0^2}{2|x_0|} \times p'(|x_0|)$$

which matches what you listed in (2). In other words, we're still approximating $p(|x|)$ using only a first-order Taylor series approximation, not a second-order Taylor series approximation. ■

Fan and Li(2001) 想提出一种惩罚函数, 以实现估计量的三种性质:

1. 无偏性: 对系数的惩罚随着系数估计量的增大而减少, 从而保证大系数的近似无偏性;
2. 稀疏性: 估计量的解是阈值形式, 从而保证产生稀疏的模型;
3. 连续性: 估计量的求解过程连续, 从而保证模型的稳定性

于是基于固定 p 和设计矩阵为正交矩阵的假设, 他们做出如下变换和讨论:

$$\frac{1}{2}\|Y - X\beta\|_2 + \sum_{j=1}^p P_\lambda(|\beta_j|) = \frac{1}{2}\|Y - \hat{Y}\|_2 + \frac{1}{2}\|Z - \beta\|_2 + \sum_{j=1}^p P_\lambda(|\beta_j|)$$

其中, $Z = X^T Y$. 上式的最小化等价于其中每一项的最小化, 因此以下关注 $\frac{1}{2}(z - \theta)^2 + p_\lambda(|\theta|)$, 其中 z 为 Z 的某一分量, θ 为 β 的某一分量, 对上式进行求导得到 $\text{sgn}(\theta)|\theta| + P'(|\theta|) - z$, 由导函数可知: 若当 $|\theta|$ 较大时, $P'(|\theta|) = 0$, 则有 $\hat{\theta} = z$, 从而保证大系数的估计量无偏; 为保证估计量是阈值形式, 需要满足 $|\theta| + P(|\theta|)$ 的最小值为正; 为保证估计量的求解过程是连续的, 需要满足 $|\theta| + P(|\theta|)$ 在 $\theta = 0$ 处取到最小值, 因此满足稀疏性和连续性, 惩罚函数需要在原点处可导。当 $p_\lambda(|\theta|)$ 为 L_q 惩罚时, Knight and Fu(2000) 证明了当 $q \leq 1$ 时, 得到的估计量有偏, 当 $p_\lambda(|\theta|)$ 为硬阈值惩罚函数时, 得到的估计量不是连续的。

基于以上推理和惩罚函数的缺点, Fan and Li(2001) 提出了一种连续可微的惩罚函数, 名为光滑衔接绝对偏差惩罚 (Penalty, 以下简称 SCAD), 其具体形式如下:

$$P(x | \lambda, \gamma) = \begin{cases} \lambda|x|, & |x| \leq \lambda \\ \frac{2\gamma\lambda|x| - x^2 - \lambda^2}{2(\gamma-1)}, & \lambda < |x| < \gamma\lambda \\ \frac{\lambda^2(\gamma+1)}{2}, & |x| \geq \gamma\lambda \end{cases}$$

其中 $\gamma > 2$ 。当 $|x| \leq \gamma$ 时, 该惩罚函数与 lasso 惩罚函数一致, 当 $\lambda < |x| \leq \gamma\lambda$ 时, 用一个凹的二次函数进行惩罚 (随着 $|x|$ 的增大, 惩罚力度逐渐减少); 当 $|x| \geq \gamma\lambda$ 时, 用一个常数进行惩罚; 对 SCAD 惩罚函数进行求导, 得到:

$$P'(x | \lambda, \gamma) = \begin{cases} \lambda, & |x| \leq \lambda \\ \frac{\gamma\lambda - |x|}{\gamma-1}, & \lambda < |x| < \gamma\lambda \\ 0, & |x| \geq \gamma\lambda \end{cases}$$

从 SCAD 的导函数更容易看出:

1. 其对系数估计量的惩罚速度随着系数估计量绝对值的增大而逐渐减少, 因此 SCAD 保证大系数的无偏性;
2. 其在原点的导数存在, 因此 SCAD 保证了稀疏性和连续性。
3. 由上式可以看出, 当 β_i 比较小的时候, SCAD 的惩罚等同于 LASSO 回归的惩罚, 都保持恒定的惩罚力度 λ , 压缩效果明显, 能够达到变量选择的目的; 当 $\beta_i > a\lambda$ 时, 导数为零, 代表惩罚力度为零, 此时损失函数只留下 RSS 这一项有影响, 因此估计值类似于最小二乘估计, 具有渐近无偏性。

Fan and Peng(2004) 将固定 p 扩展到了 $p = o(n)$ 的情形。Kim et al(2008) 证明了在 Zhao and Yu(2006) 假设的基础上, 若噪音的任意矩都存在, 则 SCAD 估计量在高维数据下具有神谕性: 并且, 当噪音服从高斯分布时, SCAD 估计量在 $p = O(\exp(cn)), \exists c \in (0, 1)$ 下具有神谕性。Kim et al(2008) 还提出了一种名为 CCCP-SACD 的算法, 该算法是凹凸算法 (concave convex procedure, 一般简称为 CCCP) 的耦合, 较 Fan and Li(2001) 的算法更加稳定和迅速, 且从理论上保证了一定收敛至局部最小值, 具体推导过程请参见该文及其参考文献。

综上所述,在一定假设的前提下,Fan and Li(2001) 和 Huang and Xie(2007) 证明了 N-SCAD 的神谕性; Fan and Peng(2004) 证明了 H-SCAD 的神谕性;Kim et al(2008) 证明了 H-SCAD 的神谕性和部分 UH-SCAD 的神谕性. SCAD 相对于 Adaptive Lasso 的优势在于,它不需要事先得到 \sqrt{n} 一致的估计量.

MCP

Zhang(2010) 提出了 MC+ 惩罚方法,MC 是指最小化最大凹度惩罚 (minimax concave Penalty, 以下简称 MCP),+ 是指对应的算法, 全称为带惩罚的线性无偏选择;MCP 的惩罚函数形式为:

$$P(x | \lambda, \gamma) = \begin{cases} \lambda|x| - \frac{x^2}{2\gamma}, & |x| \leq \gamma\lambda \\ \frac{1}{2}\gamma\lambda^2, & |x| > \gamma\lambda \end{cases}$$

其中 $\gamma > 1$, 上述惩罚函数的导数如下:

$$P'(x | \lambda, \gamma) = \begin{cases} \left(\lambda - \frac{|x|}{\gamma}\right) \text{sign}(x), & |x| \leq \gamma\lambda \\ 0, & |x| > \gamma\lambda \end{cases}$$

从上述导数可以看出,MCP 惩罚函数从一开始便随着系数估计量的增大而逐渐减少对其的惩罚程度. 不过不同于 Fan and Li(2001) 的工作,Zhang(2010) 并未给出该种惩罚函数神谕性的条件及证明, 而是直接通过证明算法的收敛点几乎处处唯一.

以下讨论 Lasso,SCAD 和 MCP 的关系

当 X 为正交阵, 则 SCAD 和 MCP 像 Lasso 一样有封闭解,SCAD 估计量的封闭解为:

$$\hat{\beta}^{SCAD} = \begin{cases} S(z | \lambda), & |z| \leq 2\lambda \\ \frac{\gamma-1}{\gamma-2} S\left(z | \frac{\gamma\lambda}{\gamma-1}\right), & 2\lambda < |z| < \lambda\gamma \\ z, & |z| \geq \lambda\gamma \end{cases}$$

MCP 估计量的封闭解为:

$$\hat{\beta}^{MCP} = \begin{cases} \frac{\gamma}{\gamma-1} S(z | \lambda), & |z| \leq \lambda\gamma \\ z, & |z| \geq \lambda\gamma \end{cases}$$

其中 z 是 β 的最小二乘估计, $S(z|\lambda)$ 是 Lasso 的软阈值解. 对于 MCP 的封闭解, 当 $\gamma \rightarrow \infty$, 该式趋近于软阈值函数; 当 $\gamma \rightarrow 1$, 该式趋近于硬阈值函数; 因此当 γ 变动时,MCP 的解在软阈值函数和硬阈值函数之间变动. 对于 SCAD 的封闭解, 当 $\gamma \rightarrow \infty$, 该式趋近于软阈值函数; 但是, 当 $\gamma \rightarrow 2$, 该式不趋近于硬阈值函数, 反而收敛至

$$\begin{cases} S(z | \lambda), & |z| \leq 2\lambda \\ z, & |z| \geq 2\lambda \end{cases}$$

2.2.4 SIS

在超高维数据的情况下, $\log(p)$ 会变得很大, DS 估计量的 L_2 误差约束将变大; 并且随着 p 的增加, UUP 条件更难以被满足; 因此 Fan and Lv(2008) 提出了确保独立筛选法 (Sure Independence Screening, 以下简称 SIS), 该方法的特点是几乎可以确保那些对因变量有显著影响的自变量被筛选出来, 并且筛选过程是独立的

SIS 方法的定义很简单, 记 $\omega = X^T Y$, 则对于 γ , 将 $|\omega|$ 降序排列, 并取前 $[\gamma n]$ 构成子模型 M_{γ} , 其中 $d = [\gamma n]$ 是 γn 的整数部分, 这个方法直接从 p 个自变量中选出 d 个自变量; 从 ω 的定义可以看出, SIS 筛选变量的根据便是自变量与因变量的相关程度; 当 $\gamma \in (0, 1)$ 时, d 是小于样本量 n , 也就是说, SIS 将数据维度从高位甚至超高维直接降至与样本量持平的程度, 在实际操作一般令 $d = n - 1$ 或者 $d = n / \log n$; 当然, $d \geq n$ 也是可以的, 此时便可以将 SIS 与前述的变量选择方法相结合, 如 SCAD、Dantzing Selector、Lasso 等, 并且 d 越大, M_d 包含正确模型的概率越高。

SIS+Dantzing 可以将风险上限从 $\log p$ 降低至 $\log d_n$.

SIS 的缺点: SIS 并未考虑到自变量之间的相关性, 即它默认自变量之间是互相独立的; 若自变量之间存在相关性, 则可能存在部分自变量与因变量的直接相关程度不高, 但间接解释程度较大的可能, 因此 SIS 存在遗落该部分自变量的可能性, 进而导致模型效果较差。

SIS 的优点: 简单; 计算复杂程度低在高维数据的情况下, 其计算复杂程度为 $O(np)$.

利用预测变量和响应变量的边际相关系数进行排序

Proposition 2.2.1 — Some problems of SIS.

1. False Negative: An important predictor that is marginally uncorrelated but jointly correlated with the response cannot be picked by SIS.
2. False positive: Unimportant predictors that are highly correlated with the important predictors can have higher priority to be selected by SIS than important predictors that are relatively weakly related to the response.
3. The issue of collinearity among the predictors adds difficulty to the problem of variable selection.

2.2.5 ISIS

对残差向量迭代使用 SIS。SIS 的确定筛选性依赖于边际相关性假设, 其要求所有活跃预测变量与响应变量间的边际相关系数均不接近于零。然而, **由于预测变量间的相关性** (利用例子来解释), 该假设在高维模型中时常不成立, 从而导致了 SIS 在此类情形下的不佳表现。

■ **Example 2.1 — Barut et al., 2016--解释 SIS 的缺点.** . 考虑线性模型

$$y = 3x_1 + 3x_2 + 3x_3 + 3x_4 + 3x_5 - 7.5x_6 + \sum_{j=7}^p \beta_j x_j + \epsilon$$

其中每个预测变量均服从标准正态分布且与其他预测变量间的相关系数均为 0.5。除了 x_1, \dots, x_6 , 剩余预测变量的回归系数均为 0。另外, 随机误差变量 ϵ 与预测变量独立且服

从均值为 0, 方差为 σ^2 的正态分布。我们计算得到 x_6 与 y 之间的 Pearson 相关系数为 $0(15*0.5-7.5*1)$, 因此边际相关性假设并不成立. Barut et al. (2016) 中的数值模拟结果表明, 在一共 200 次模拟中, SIS 均无法准确识别预测变量 x_6 。

为了解决该问题, Fan & Lv (2008) 提出了 ISIS (iteratively sure independence screening) 方法, 其通过对残差向量迭代运用 SIS 方法, 降低了预测变量间相关性对筛选结果的影响。

在每一步迭代中, ISIS 首先计算响应变量与已选变量间的回归残差, 再将 SIS 方法应用到残差与剩余预测变量上, 从而削弱已选变量与剩余变量间相关性的影响. 通过计算残差, 与已选活跃预测变量相关的非活跃预测变量被选入的优先级将会降低。相反地, 之前与响应变量边际无关的活跃变量可能与残差边际相关, 因此被选入的可能性会显著提高。

Van-ISIS

Fan et al. (2009) 进一步提出了 Van-ISIS (vanilla ISIS) 方法, 其可看作 ISIS 方法在广义线性模型上的推广与改进. Van-ISIS 方法允许已选变量在之后的迭代过程中被筛选掉。相较于 ISIS, Van-ISIS 方法的一个主要改进是其将施加惩罚的变量选择方法同时运用在已选变量和新增变量上, 因此之前选择的预测变量可能在之后的迭代过程中被筛选掉, 从而进一步提高筛选的准确性。

在经典线性回归理论中, 向前回归方法 FR (forward regression) 将使模型 RSS 降低最多的预测变量逐个加入模型当中, 直至提升不再显著. Wang (2009) 提出 FR 方法在超高维线性模型下同样有效, 并在特定假设下证明了其确定筛选性。

2.2.6 HOLP

针对超高维线性模型, Wang & Leng (2016) 提出了一个高效的特征筛选方法, 称为 HOLP (high dimensional ordinary least squares projection). HOLP 通过计算估计量

$$\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)^\top = X^\top (XX^\top)^{-1} Y$$

并根据 $|\hat{\beta}_i|$ 选择预测变量. 其有效的原因在于矩阵 $X^\top (XX^\top)^{-1} X$ 在特定假设下为对角占优矩阵, 从而 $|\hat{\beta}_i|$ 能够尽可能保留 $|\beta_i|$ 的大小排序. Wang & Leng (2016) 在特定假设下证明了 HOLP 的确定筛选性。

对于广义线性模型, 考虑响应变量 y 关于 $X = x$ 服从指数族分布, 密度函数为

$$f_Y(y | x; \theta) = \exp\{y\theta(x) - b(\theta(x)) + c(x; y)\}$$

其中 $b(\cdot)$ 与 $c(\cdot)$ 为已知函数, θ 为未知参数. Fan & Song (2010) 在此基础上提出了基于极大边际似然估计量 MMLE (maximum marginal likelihood estimator) 的特征筛选方法, 通过计算

$$(\hat{\beta}_{j,0}^M, \hat{\beta}_j^M) = \arg \min_{\beta_0, \beta_j} \mathbb{P}_n \{l(\beta_0 + \beta_j X_j, Y)\}$$

并通过将 $|\hat{\beta}_j^M|$ 由大到小排序选择相应的变量, 其中 $l(\beta_0 + \beta_j X_j, Y)$ 为负对数似然函数且 $\mathbb{P}_n f(X, Y) = \frac{1}{n} \sum_{i=1}^n f(X_i, Y_i)$ 为基于样本的经验表示.

为了有效利用此类信息, Barut et al. (2016) 在广义线性模型下提出了条件简选方法 CSIS (conditional sure independence screening). 假设 x_C 为已知的活跃预测变量且 x_D 表示剩余预测变量. CSIS 首先计算

$$(\hat{\beta}_C, \hat{\beta}_j^D) = \arg \min_{\beta_C, \beta_j} \mathbb{P}_n \left\{ l \left(X_C^\top \beta_C + X_j \beta_j, Y \right) \right\}$$

并通过将 $|\hat{\beta}_j^D|$ 由大到小排序来选择相应的预测变量, 其中 $l(X_C^\top \beta_C + X_j \beta_j, Y)$ 为负对数似然函数且 $\mathbb{P}_n f(X, Y) = \frac{1}{n} \sum_{i=1}^n f(X_i, Y_i)$ 为基于样本的经验表示.

ISIS& van-ISIS 的推导

在上述设定下, 我们定义真实模型为

$$\mathcal{T} = \{1 \leq j \leq p : \beta_j \neq 0\}$$

即包含所有活跃预测变量对应指标的指标集. 我们假设真实模型的大小为 $|\mathcal{T}| = t$. 对于任意指标集 $\mathcal{S} \subset \{1, \dots, p\}$, 令 $\beta_{\mathcal{S}}$ 表示向量 β 中对应于 \mathcal{S} 的元素构成的子向量. 类似地, 令 $X_{\mathcal{S}}$ 表示矩阵 X 中对应于 \mathcal{S} 的列向量构成的子矩阵. 另外, 令 $C(X_{\mathcal{S}})$ 表示矩阵 $X_{\mathcal{S}}$ 的列向量张成的线性空间, 并且 $C(X_{\mathcal{S}})^\perp$ 表示 $C(X_{\mathcal{S}})$ 的正交补空间. 同时, 令 $H_{\mathcal{S}}$ 和 $M_{\mathcal{S}}$ 分别表示 $C(X_{\mathcal{S}})$ 和 $C(X_{\mathcal{S}})$ 上的正交投影矩阵. 若 $X_{\mathcal{S}}$ 列满秩, 则 $H_{\mathcal{S}}$ 可写为

$$H_{\mathcal{S}} = X_{\mathcal{S}} \left(X_{\mathcal{S}}^\top X_{\mathcal{S}} \right)^{-1} X_{\mathcal{S}}^\top$$

且 $M_{\mathcal{S}}$ 可以表示为

$$M_{\mathcal{S}} = I_n - H_{\mathcal{S}} = I_n - X_{\mathcal{S}} \left(X_{\mathcal{S}}^\top X_{\mathcal{S}} \right)^{-1} X_{\mathcal{S}}^\top$$

其中 I_n 表示 $n \times n$ 的单位矩阵.

对于标准化的检测变量, SIS 计算估计量

$$\omega = (\omega_1, \dots, \omega_p)^\top = X^\top Y$$

并通过对 $|\omega_j|$ 排序进行变量筛选. 如前文所述, SIS 方法的确定筛选性依赖于边际相关性假设, 即存在某个正常数 c 使得

$$\min_{i \in \mathcal{T}} \left| \text{cov} \left(\beta_i^{-1} y, x_i \right) \right| \geq c$$

为了缓解违反边际相关性假设所带来的负面影响, Fan & Lv (2008) 提出了迭代筛选方法 ISIS, 其通过对残差向量迭代运用 SIS 方法以减少预测变量间相关性的影响. 其算法可以具体描述如下.

1. 步骤 1: 通过 SIS 筛选出大小为 a_1 的模型 \mathcal{A}_1 , 即 $\mathcal{A}_1 = \{1 \leq i \leq p : |X_i^\top Y| \text{ 属于各估计量的绝对值中最大的 } a_1 \text{ 个}\}$. 通过求解如下惩罚最小二乘问题, 从模型 \mathcal{A}_1 中得到子模型 \mathcal{B}_1 .

$$\min_{\beta \in \mathbb{R}^{a_1}} \left[\|Y - X_{\mathcal{A}_1} \beta\|^2 + n \sum_{j=1}^{a_1} p_{\lambda_n}(|\beta_j|) \right]$$

其中 p_{λ_n} 是参数为 $\lambda_n > 0$ 的惩罚函数. 定义 $\mathcal{S}_1 = \mathcal{B}_1$, 且剩余模型为 $\mathcal{S}_1^c = \{1, \dots, p\} - \mathcal{S}_1$

2. 步骤 2: 基于大小为 s_k 的模型 \mathcal{S}_k , 选择大小为 a_{k+1} 的模型 \mathcal{A}_{k+1} , $\mathcal{A}_{k+1} = \{i \in \mathcal{S}_k^c : |X_i^\top M_{\mathcal{S}_k} Y| \text{ 属于各估计量的绝对值中最大的 } a_{k+1} \text{ 个}\}$, 其中 $M_{\mathcal{S}_k} Y$ 表示在 Y 关于 $X_{\mathcal{S}_k}$ 的线性回归中最小二乘方法得到的残差向量. 通过求解如下惩罚最小二乘问题, 从模型 \mathcal{A}_{k+1} 中得到子模型 \mathcal{B}_{k+1}

$$\min_{\beta \in \mathbb{R}^{a_{k+1}}} \left[\|M_{\mathcal{S}_k} Y - X_{\mathcal{A}_{k+1}} \beta\|^2 + n \sum_{j=1}^{a_{k+1}} p_{\lambda_n}(|\beta_j|) \right]$$

令 $\mathcal{S}_{k+1} = \mathcal{S}_k \cup \mathcal{B}_{k+1}$ 且 $\mathcal{S}_{k+1}^c = \{1, \dots, p\} - \mathcal{S}_{k+1}$

3. 步骤 3: 对步骤 2 进行迭代直到选择出的模型 \mathcal{S}_κ 达到了预先设定的最大迭代步数 κ .

在 ISIS 算法中, 每一步所选的预测变量均会保留在最终选择的模型之中. 然而, 这在 Van-ISIS 算法中并不成立. Van-ISIS 方法可以在迭代过程中删除之前步骤所选的预测变量, 其具体算法如下.

1. 步骤 1: 选择模型大小为 a_1 的模型 \mathcal{A}_1 , $\mathcal{A}_1 = \{1 \leq i \leq p : \|M_i Y\|^2 \text{ 属于所有计算量中最小的 } a_1 \text{ 个}\}$ 即模型 \mathcal{A}_1 包含 a_1 个使得逐项回归中 RSS 达到最小的预测变量. 通过求解如下惩罚最小二乘问题, 从模型 \mathcal{A}_1 中得到子模型 \mathcal{S}_1

$$\min_{\beta \in \mathbb{R}^{a_1}} \left[\|Y - X_{\mathcal{A}_1} \beta\|^2 + n \sum_{j=1}^{a_1} p_{\lambda_n}(|\beta_j|) \right]$$

并记剩余模型为 $\mathcal{S}_1^c = \{1, \dots, p\} - \mathcal{S}_1$

2. 步骤 2: 基于大小为 s_k 的模型 \mathcal{S}_k , 选择大小为 a_{k+1} 的模型 \mathcal{A}_{k+1} , $\mathcal{A}_{k+1} = \{i \in \mathcal{S}_k^c : \|M_{\mathcal{S}_k \cup \{i\}} Y\|^2 \text{ 属于所有计算量中最小的 } a_{k+1} \text{ 个}\}$ 通过求解如下惩罚最小二乘问题, 从模型 $\mathcal{S}_k \cup \mathcal{A}_{k+1}$ 中得到子模型 \mathcal{S}_{k+1} .

$$\min_{\beta \in \mathbb{R}^{s_k + a_{k+1}}} \left[\|Y - X_{\mathcal{S}_k \cup \mathcal{A}_{k+1}} \beta\|^2 + n \sum_{j=1}^{s_k + a_{k+1}} p_{\lambda_n}(|\beta_j|) \right]$$

其中 $X_{\mathcal{S}_k \cup \mathcal{A}_{k+1}} = [X_{\mathcal{S}_k}, X_{\mathcal{A}_{k+1}}]$, 并记剩余模型为 $\mathcal{S}_{k+1}^c = \{1, \dots, p\} - \mathcal{S}_{k+1}$.

3. 步骤 3: 对步骤 2 进行迭代直到选择出的模型 \mathcal{S}_κ 达到了预先设定的最大迭代步数 κ .

到目前为止, 所介绍的迭代筛选方法均可看作一个筛选过程与一个选择过程的组合. 在筛选过程中, 我们根据模型 \mathcal{S}_k 确定新模型 \mathcal{A}_{k+1} . 而在选择过程中, 我们从模型 \mathcal{S}_k 和 \mathcal{A}_{k+1} 选出新的模型 \mathcal{A}_{k+1} .

对于筛选过程, 我们有如下三个准则可以选择.

1. 准则 1: $\mathcal{A}_{k+1} = \{i \in \mathcal{S}_k^c : |X_i^\top M_{\mathcal{S}_k} Y| \text{ 属于各估计量的绝对值中最大的 } a_{k+1} \text{ 个}\}.$
2. 准则 2: $\mathcal{A}_{k+1} = \{i \in \mathcal{S}_k^c : \|M_{\mathcal{S}_k \cup \{i\}} Y\|^2 \text{ 属于所有计算量中最小的 } a_{k+1} \text{ 个}\}.$
3. 准则 3: $\mathcal{A}_{k+1} = \{i \in \mathcal{S}_k^c : |\hat{\beta}_i| \text{ 属于各估计量中绝对值最大的 } a_{k+1} \text{ 个}\}$ 其中估计量 $\hat{\beta}_i$ 可由式所得.

对于选择过程, 同样有如下三个准则供我们选择。

1. 准则 1: $\mathcal{S}_{k+1} = \mathcal{S}_k \cup \mathcal{A}_{k+1}$
2. 准则 2: $\mathcal{S}_{k+1} = \mathcal{S}_k \cup \mathcal{B}_{k+1}$, 其中 $\mathcal{B}_{k+1} \subset \mathcal{A}_{k+1}$ 由求解下述惩罚最小二乘问题得到,

$$\min_{\beta \in \mathbb{R}^{a_{k+1}}} \left[\|M_{\mathcal{S}_k} Y - X_{\mathcal{A}_{k+1}} \beta\|^2 + n \sum_{j=1}^{a_{k+1}} p_{\lambda_n}(|\beta_j|) \right]$$

3. 准则 3: $\mathcal{S}_{k+1} \subset \mathcal{S}_k \cup \mathcal{A}_{k+1}$ 由求解下述惩罚最小二乘问题得到,

$$\min_{\beta \in \mathbb{R}^{s_k + a_{k+1}}} \left[\|Y - X_{\mathcal{S}_k \cup \mathcal{A}_{k+1}} \beta\|^2 + n \sum_{j=1}^{s_k + a_{k+1}} p_{\lambda_n}(|\beta_j|) \right]$$

2.2.7 Dantzing selector

Notion

\bar{E} 是补集的意思。给定一个 $p \times p$ 的矩阵 C 和任意的子集 $T_1, T_2 \subseteq \{1, 2, \dots, p\}$, 我们用 C_{T_1, T_2} 表示由矩阵 C 中行和列分别对应于集合 T_1 和 T_2 的那些元素构成的 $|T_1| \times |T_2|$ 的矩阵。另外, 用 $\text{diag}(\beta)$ 表示对角元是由向量 β 中元素所构成的对角阵。对于 $\beta \in \mathbb{R}^p$, $\text{sign}(\beta) = (\text{sign}(\beta_1), \dots, \text{sign}(\beta_p))^\top$ 表示向量 β 的符号函数

Candes 和 Tao 提出了 Dantzig Selector(DS) 方法。DS 方法的参数估计为下述凸优化问题的解:

$$\hat{\beta}^{DS} = \arg \min_{\beta \in \mathbb{R}^p} \left\{ \sum_{j=1}^p |\beta_j| : \|X^T(Y - X\beta)\|_\infty \leq \lambda(\sigma) \right\} \quad (2.21)$$

其中 $\|X^T(Y - X\beta)\|_\infty := \sup_{1 \leq j \leq p} |[X^T(Y - X\beta)]_j|$ 是 L_∞ 范数, $\lambda(\sigma) > 0$ 是调整参数, σ 是真实模型误差的标准差。 L_1 正则化保证了部分系数估计量被压缩至 0, 类似 Lasso 的原理, 从而可以达到变量选择的目的。在实际问题中, 由于真实模型误差的标准差 σ 是未知的, 为此 Meinshausen 等提出了两种解决方案:

1. 矩估计的思想: 是利用数据估计 $\hat{\sigma}$ 并代入上述公式得到调整参数;
2. 利用交叉验证等方法动态选取调整参数, 其结论为采用全路径搜索的调整参数较固定形式的调整参数所选择的模型有更高的预测精度。

Candes 和 Tao 表示 DS 方法限制相关残差向量 $X^T(Y - X\hat{\beta})$ 而不是残差 $(Y - X\hat{\beta})$ 主要基于以下几点考虑:

1. 相关残差具有正交变换不变性, 而残差形式没有。假设 U 是正交矩阵, 有 $(UX)^T(UX\hat{\beta} - UY) = X^T(X\hat{\beta} - Y)$.
2. 相关残差有助于选取与响应变量 Y 具有高相关性的变量。有点类似于 ISIS 的方法, 只对残差项来计算相关系数, 减少了共线性的可能。

DS 方法具有较好的理论性质, 即载参数个数比样本量大很多的情况下, 其参数估计于真实参数之间的 L_2 误差也能保持在 $\log(p)$ 之内。用数学语言描述:

Theorem 2.2.2 假设 $\beta \in \mathbb{R}^p$ 是真实的回归系数, 且只有 S 个分量非零。若设计阵满足“一致不确定原则”即 $\delta_S + \theta_{S,2S} < 1$, 模型误差满足正态性假设, 且调整参数 $\lambda(\sigma) = \sqrt{2(1+a)\log p}$, a 为任意非负常数, 那么 DS 方法的参数估计 $\hat{\beta}^{DS}$ 以超过 $1 - \left(p^a \sqrt{\pi \log(p)}\right)^{-1}$ 的概率满足

$$\|\hat{\beta}^{DS} - \beta\|^2 \leq C_1^2 \cdot (2\log p) \cdot S \cdot \sigma^2$$

其中 $C_1 = 4 / (1 - \delta_S - \theta_{S,2S})$ 。

上述“一致不确定原则”包含两条性质:

1. 第一条性质, “ S 限制等距假设”, 意思为每个元素个数小于 S 的设计阵 X 的列向出集都近似为一个标准正交系统。
2. 第二条性质, “ S 限制正交性”, 意思为对不相交的变量子集给予一定的限制, 则子集张成近似正交的子空间。

上述两条性质对应于两个常量, 即等距常数 δ_S 和限制正交系数 $\theta_{S,S'}$ 。

Definition 2.2.1 — 等距常数. 假设 $X_T (T \subset \{1, \dots, p\})$ 表示提取 X 下标在集合 T 中的相应列向量所构成的子矩阵, $\beta_T = (\beta_j)_{j \in T}$ 是真实系数 β 的部分序列则等距常数 δ_S 为满足下式的最小正常数:

$$(1 - \delta_S) \|\beta_T\|^2 \leq \|X_{TC}\|^2 \leq (1 + \delta_S) \|\beta_T\|^2$$

其中 T 为任意元素个数不超过 S 的集合。

Definition 2.2.2 — 限制正交系数. 假设 $S + S' \leq p$, 则限制正交系数 $\theta_{S,S'}$ 为满足下式的最小正常数:

$$|\langle X_{TC}, X_{T'C} \rangle| \leq \theta_{S,S'} \|\beta_T\| \cdot \|\beta_{T'}\|$$

其中 T 和 T' 分别为任意元素个数不超过 S 和 S' 的不交的集合。

Meinshausen 等 [29] 具体讨论了 DS 方法与 LASSO 方法的异同. 为了方便比较, 采用下述等价形式:

$$\begin{aligned} \text{LASSO: } & \min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|^2, \quad \sum_{j=1}^p |\beta_j| \leq \lambda \\ \text{DS: } & \min_{\beta \in \mathbb{R}^p} \|X^T(Y - X\beta)\|_\infty, \quad \sum_{j=1}^p |\beta_j| \leq \lambda \end{aligned}$$

正如上面所说, 在“一致不确定原则”下, DS 方法参数估计的误差有很好的控制, LASSO 方法参数估计的相合性需要“不可表示条件”的支撑。特别地, Meinshausen 和 Yu 推导了在一个对设计阵要求比“一致不确定原则”更宽松的条件下, LASSO 方法参数估计的均方误差以一个更小的概率限制在与上述 DS 方法相同的范围内。文献中没有一个确切的结论说明 DS 与 LASSO 哪种方法更优。另一方面, 二者形式的相近性自然引起人们探索二者

联系的兴趣。事实上, Meinhshausen 等给出了二者结果相同的充分条件。在 $p \leq n$ 的前提下, 记 $M = (X^T X)^{-1}$, 若满足对角线主导条件

$$M_{jj} > \sum_{i \neq j} |M_{ij}|, \quad \forall j = 1, \dots, p$$

LASSO 方法与 DS 方法的结果完全相同。特别地, 当 $p = 2$ 时上述条件总成立, 即二者总相同。

Candes and Tao(2007) 建议 $\lambda(\sigma) = (1 + t^{-1}) \sigma \sqrt{2 \log p}$, $t > 0$, 并证明在该取值下, 若设计矩阵满足一致不确定准则 (Uniform Uncertainty Orinciples, 以下简称为 UUP)

$$\delta_q + \theta_{q,2q} < 1$$

则 $\hat{\beta}^{DS}$ 以高于 $1 - (\sqrt{\pi \log p} \times p^t)^{-1}$ 的概率满足

$$\|\hat{\beta}^{DS} - \beta\|_2 \leq q C_1^2 (2 \log p) \sigma^2$$

其中 $C_1 = 4 / (1 - \delta_q - \theta_{q,2q})$, 当 $\delta_q + \theta_{q,2q}$ 较小时, $C_1 \approx 4$; δ_q 是 q 阶限制同构常数 (q -restricted isometry constant), $\theta_{q,2q}$ 是限制正交常数 (restricted orthogonality constants)。上式的特点是, $\hat{\beta}^{DS}$ 的 L_2 误差被 $\log(p)$ 的函数沿严格控制, 而非渐近控制; DS 估计量不具有相合性, 且该方法对每个估计量的惩罚力度一样, 这一点与 Lasso 类似。

James et al(2009) 为 DS 方法提供了一种高效算法 DASSO, 其主要步骤是 DASSO 算法和 LARS 算法都是采用分段线性步骤的算法, DASSO 算法与 LARS 算法最大的不同在于 DS 方法不能保证选择出正确的模型,

在超高维数据的情况下, $\log(p)$ 会变得很大, DS 估计量的 L_2 误差约束将变大; 并且随着 p 的增加, UUP 条件更难以被满足。

Definition 2.2.3 — 不可表示条件. 假设对于某个 $E \subset \{1, 2, \dots, p\}$ 满足 $|E| = |T^*|$, $C_{T^*,E}$ 是可逆的, 则**不可表示条件**定义为: 不等式

$$\left| C_{T^*,E}^{-1} C_{T^*,E}^{-1} \text{sign}(\beta_{T^*}) \right| \leq 1 \quad (2.22)$$

成立并且存在一个正常数 η 满足

$$\left| C_{E,T^*}^{-1} C_{E,T^*}^{-1} \text{sign} \left(C_{T^*,E}^{-1} \text{sign}(\beta_{T^*}) \right) \right| \leq 1 - \eta \quad (2.23)$$

其中 $\mathbf{1}$ 是一个每个分量都是 1 的 $(p - q) \times 1$ 维向量, 并且 $|\cdot|$ 表示上面的不等式成立指的是不等式的左侧的向量中每个分量的绝对值逐个与右侧的分量比较。

这里的不可表示条件指的是不显著的变量不可能由显著变量表示。不可表示条件对于 Dantzig 选择器的相合性具有非常重要的作用。在不可表示条件成立的条件下, 我们得到了, 无论 p (变量个数) 是固定的, 还是随着 n 增大的, 甚至是关于 n 以指数的速度增长, Dantzig 选择器都是模型选择相合的。这里的相合指的是依概率符号相合, 即

$$P \left(\hat{\beta}^D(\lambda) = {}_s\beta \right) \rightarrow 1, n \rightarrow \infty$$

其中 $\hat{\beta}^D(\lambda)$ 是 Dantzig 选择器的解, λ 是惩罚参数. 我们还研究了变量选择后的传统的参数估计的相合性. 我们得到了如果显著变量的个数满足 $q = o(n)$, 变量选择后的传统的参数估计也是相合的。

只要潜在的真模型满足不可表示条件, Dantzig 选择器就具有相合性, 但是当不可表示条件不成立的时候, 模型选择的相合性就不再满足了. 此外, Dantzig 估计也达不到 Fan and Li (2001) 和 Fan and Peng (2004) 里给出的 oracle 性质。

Definition 2.2.4 弱不可表示条件: 不等式

$$\left| C_{T^*, E} C_{T^*, E}^{-1} \text{sign}(\beta_{T^*}) \right| \leq 1 \quad (2.24)$$

和

$$\left| C_{\bar{E}, T^*} C_{\bar{E}, T^*}^{-1} \text{sign} \left(C_{T^*, E}^{-1} \text{sign}(\beta_{T^*}) \right) \right| < 1 \quad (2.25)$$

都成立, 其中 $\mathbf{1}$ 是一个每个分量都是 1 的 $(p - q) \times 1$ 维向量, $|\cdot|$ 表示上面的不等式成立指的是不等式的左侧的向量中每个分量的绝对值逐个与右侧的分量比较成立。

无论是固定的 p (变量的个数), q (显著变量个数) 还是 p, q 随着样本增长, 甚至 p 随着样本 n 以指数型增长的情况, 在不可表示条件及其它一些适当的条件下, 我们都可以得到 Dantzig selector 估计量 $\hat{\beta}^D(\lambda)$ 是模型选择相合的。

R

1. Dantzig selector 的不可表示条件与 LASSO 的是完全不同的。前者要更复杂一些, 这可能是由于 Dantzig selector 的解没有显式结构所致。
2. 这里我们所给出的不可表示条件与 Dickcr and Lin(2009) 对于随机设计的所给出的不可表示条件也是有区别的。他们所提出的基于总体的不可表示条件不能推广到 $p > n, q$ 也发散的情形。而这里我们考虑的是固定设计, 我们的条件是基于样本的, 我们的条件可以推广到处理 $p > n$ 的情况。

正则条件:

1. $C = C(n) \rightarrow C^*, n \rightarrow \infty$, 其中 C^* 是一个正定矩阵.
2. $\frac{1}{n} \max_{1 \leq i \leq n} X_i^T X_i \rightarrow 0, \quad n \rightarrow \infty$

罗列 Dantzig selection 最重要的五个定理

Theorem 2.2.3 当 p 和 q 固定的时候, 假设正则条件 (a), (b) 和不可表示条件成立, 则对于满足 $\lambda/n \rightarrow 0$ 和 $\lambda/n^{(c+1)/2} \rightarrow \infty, 0 \leq c < 1$ 的正数 λ , Dantzig selector 估计量是强符号相合的, 即:

$$P\left(\hat{\beta}^D(\lambda) =_s \beta\right) = 1 - o\left(e^{-n^c}\right) \rightarrow 1, \quad n \rightarrow \infty$$

Theorem 2.2.4 假设 ε_i 是满足对于某个 $k > 0, E(\varepsilon_i)^{2k} < \infty$ 成立的独立同分布的随机变量。在条件 (C1) – (C5) 下, 若 $p = o\left(n^{(d_2-d_1)k}\right)$ 对于 $d_2 > d_1$ 成立, λ 是满足 $\lambda/\sqrt{n} = o\left(n^{(d_2-d_1)/2}\right), (\lambda/\sqrt{n})^{2k}/p \rightarrow \infty$ 的正数. 则在不可表示条件满足的条件下, 我们可以得到 $\hat{\beta}^D$ 的强符号相合性, 即

$$P\left(\hat{\beta}^D(\lambda) = {}_s\beta\right) \geq 1 - O\left(\frac{pn^k}{\lambda^{2k}}\right) \rightarrow 1, n \rightarrow \infty$$

假设 κ_{n1} 是矩阵 $B = C_{E,T^*}^{-1}C_{E,E}C_{T^*,E}^{-1}$ 的最大特征值, 其中 E 是满足不可表示条件中的一个。又设 τ_{n1} 是半正定阵 $(I - K)(I - K)^\top$ 的最大特征值, 其中 $K = \mathbf{X}_{T^*}(\mathbf{X}_E^\top \mathbf{X}_{T^*})^{-1} \mathbf{X}_E^\top$ 是一个幂等阵。假设存在 $0 \leq d_1 < d_2 \leq 1$ 的 d_1, d_2 使得下面几条成立。

1. (C1) $q = O(n^{d_1})$
2. (C2) $n^{\frac{1-d_2}{2}} \min_{i \in T^*} |\beta_i| \geq M_1 > 0$
3. (C3) $\|C_{E,T^*}\alpha\|_2^2 \geq M_2 > 0$ 对任意的单位向量 α 都成立, 其中 E 满足不可表示条件。
4. (C4) $0 < \kappa_{n1} \leq \kappa_1 < \infty$
5. (C5) $\tau_{n1} \leq \tau_1 < \infty$

Theorem 2.2.5 假设条件 (C1)-(C5) 成立, ε_i 是独立同分布的随机变量, 并且对所有的 $x \geq 0$ 满足 $P(|\varepsilon_i| > x) \leq K \exp(-Cx^d), i = 1, 2, \dots$, 其中 $1 \leq d \leq 2, C > 0$ 和 K 都是给定的常数。 $p = e^{n^{d_3}}, p = e^{n^{d_3}}, \left(\frac{1}{\log n}\right)^{I\{d=1\}} \left(\frac{\lambda}{\sqrt{n}}\right)^d = O(n^{d_4})$, 其中 $0 < d_3 < d_4 < d_2 d/2$, 则在不可表示条件满足的条件下, 我们可得 $\hat{\beta}^D$ 是强符号相合的, 即:

$$P\left(\hat{\beta}^D(\lambda) = {}_s\beta\right) \geq 1 - O\left(e^{-n^\delta}\right) \rightarrow 1, \quad n \rightarrow \infty$$

其 $\delta = \min\{d_4 - d_3, d_2 d/2\}$

我们也证明了弱不可表示条件是 Dantzig selector 模型选择相合的一个必要条件, 将其表述为下面的定理:

Theorem 2.2.6 对于固定的 p 和 q , 在正则条件 (a) 和条件 (b) 下, 如果 Dantzig selector 估计量是弱符号相合的, 则一定存在 $N, \forall n > N$, 使得弱不可表示条件成立。

上述定理告诉我们, 如果弱不可表示条件不满足, 当然更不必说不可表示条件是不满足的了, 则 Dantzig selector 的模型选择就不再相合了。

Dickcr and Lin (2009) 给出了一个 Dantzig selector 有唯一解的充分条件: $\mathbf{X}^\top \mathbf{X}$ 与 ℓ_1 球不平行。这个条件对于研究 Dantzig selector 的解的大样本性质具有非常重要的作用。 $\mathbf{X}^\top \mathbf{X}$ 与 ℓ_1 -球平行的定义为存在两个子集 $S_1, S_2 \subseteq \{1, \dots, p\}$ 和向量 $w \in \mathbf{R}^{|S_1|}$, 使得 $\|\mathbf{X}^\top \mathbf{X}_{S_1} u\|_\infty \leq 1, \mathbf{X}_{S_2}^\top \mathbf{X}_{S_1} w \in \{\pm 1\}^{|S_2|}$ 和 $\dim\left\{\text{null}\left(\mathbf{X}_{S_1}^\top \mathbf{X}_{S_2}\right)\right\} > 0$ 成立。为了研究相合性, 在本章的后面内容中, 我们总假设 $\mathbf{X}^\top \mathbf{X}$ 与 ℓ_1 -球是不平行的, 也就是说 Dantzig selector 的解是唯一的。下面的引理是 Dickcr and Lin (2009) 中的内容, 现重述之。

Lemma 2.6 向量 $\hat{\beta} \in \mathbf{R}^p$ 是 Dantzig selector (2.21) 的一个解当且仅当存在一个 $\hat{\mu} \in \mathbf{R}^p$, 使得

$$\|\mathbf{X}^\tau(y - \mathbf{X}\hat{\beta})\|_\infty \leq \lambda \quad (2.26)$$

$$\|\mathbf{X}^\tau \mathbf{X} \hat{\mu}\|_\infty \leq 1 \quad (2.27)$$

$$\hat{\mu}^\tau \mathbf{X}^\tau \mathbf{X} \hat{\beta} = \|\hat{\beta}\|_1 \quad (2.28)$$

$$\hat{\mu}^\tau \mathbf{X}^\tau(y - \mathbf{X}\hat{\beta}) = \lambda \|\hat{\mu}\|_1 \quad (2.29)$$

Proposition 2.2.7 若不可表示条件成立, 则有

$$P(\hat{\beta}^D =_s \beta) \geq P(A_n \cap B_n)$$

对于

$$A_n = \left\{ \left| Z_E - C_{E,T^*} C_{E,T^*}^{-1} Z_E \right| \leq \frac{\lambda \eta}{\sqrt{n}} \right\}$$

$$B_n = \left\{ \left| DC_{E,T^*}^{-1} Z_E \right| < \sqrt{n} \left\{ |\beta_{T^*}| - \frac{\lambda}{n} \left| DC_{E,T^*}^{-1} \text{sign}(\tilde{\mu}_E) \right| \right\} \right\}$$

其中 $Z_E = \frac{1}{\sqrt{n}} \mathbf{X}_E^\tau \varepsilon$, $Z_E = \frac{1}{\sqrt{n}} \mathbf{X}_E^\tau \varepsilon$, $D = \text{diag}(\text{sign}(\beta_{T^*}))$, $\tilde{\mu}_E = (\mathbf{X}_{T^*}^\tau \mathbf{X}_E)^{-1} \text{sign}(\beta_{T^*})$

Proof. 我们首先定义 $\tilde{\mu}$ 和 $\tilde{\beta}$ 如下. 对于某个满足不可表示条件的子集 $E \subset \{1, \dots, p\}$, 设

$$\tilde{\mu}_E = (\mathbf{X}_{T^*}^\tau \mathbf{X}_E)^{-1} \text{sign}(\beta_{T^*}), \quad \tilde{\mu}_{\bar{E}} = 0 \quad (2.30)$$

和

$$\tilde{\beta}_{T^*} = (\mathbf{X}_E^\tau \mathbf{X}_{T^*})^{-1} \mathbf{X}_E^\tau y - \lambda (\mathbf{X}_E^\tau \mathbf{X}_{T^*})^{-1} \text{sign}(\tilde{\mu}_E), \quad \tilde{\beta}_{\bar{T}^*} = 0 \quad (2.31)$$

由于 $\mathbf{X}^\tau \mathbf{X}$ 与 ℓ_1 -球是不平行的, 因此 Dantzig selector (2.21) 的解是唯一的, 并且事件 $\{\hat{\beta}^D =_s \beta\}$ 包含交集 $\{\tilde{\beta} = \hat{\beta}^D\} \cap \{\tilde{\beta} =_s \beta\}$. 故只要证明 $A_n \cap B_n \subseteq \{\tilde{\beta} = \hat{\beta}^D\} \cap \{\tilde{\beta} =_s \beta\}$ 即可. 实际上, 我们将分别证明 $A_n \cap B_n \subseteq \{\tilde{\beta} = \hat{\beta}^D\}$ 和 $B_n \Leftrightarrow \{\tilde{\beta} =_s \beta\}$ 成立. 首先来证明 $A_n \cap B_n \subseteq \{\tilde{\beta} = \hat{\beta}^D\}$.

根据引理 2.9, 事件 $\{\tilde{\beta} = \hat{\beta}^D\} \supseteq \{(\tilde{\mu}, \tilde{\beta}) : (\tilde{\mu}, \tilde{\beta}) \text{ satisfy (2.26) - (2.29)}\}$. 因此, 我们只需要证明当事件 A_n 和 B_n 同时发生的时候, 由 (2.30) 式和 (2.31) 式定义的 $(\tilde{\mu}, \tilde{\beta})$ 满足 (2.26) 式到 (2.29) 式. 先将 (2.31) 式代入 (2.26) 中, 我们有

$$\mathbf{X}_E^\tau(y - \mathbf{X}\tilde{\beta}) = \mathbf{X}_E^\tau(y - \mathbf{X}_{T^*}\tilde{\beta}_{T^*}) = \lambda \text{sign}(\tilde{\mu}_E) \quad (2.32)$$

同时,

$$\begin{aligned}
A_n &= \left\{ \left| Z_{\bar{E}} - C_{\bar{E}, T^*} C_{E, T^*}^{-1} Z_E \right| \leq \frac{\lambda \eta}{\sqrt{n}} \right\} \\
&\subseteq \left\{ \left\| Z_{\bar{E}} - C_{\bar{E}, T^*} C_{E, T^*}^{-1} Z_E \right\|_{\infty} \leq \frac{\lambda}{\sqrt{n}} \left(1 - \left\| C_{\bar{E}, T^*} C_{E, T^*}^{-1} \text{sign}(\tilde{\mu}_E) \right\|_{\infty} \right) \right\} \\
&=^? \left\{ \left\| \mathbf{X}_{\bar{E}}^{\tau} \left(I - \mathbf{X}_{T^*} (\mathbf{X}_{\bar{E}}^{\tau} \mathbf{X}_{T^*})^{-1} \mathbf{X}_{\bar{E}} \right) \varepsilon \right\|_{\infty} + \lambda \left\| \mathbf{X}_{\bar{E}}^{\tau} \mathbf{X}_{T^*} (\mathbf{X}_{\bar{E}}^{\tau} \mathbf{X}_{T^*})^{-1} \text{sign}(\tilde{\mu}_E) \right\|_{\infty} \leq \lambda \right\} \\
&\subseteq \left\{ \left\| \mathbf{X}_{\bar{E}}^{\tau} \left(I - \mathbf{X}_{T^*} (\mathbf{X}_{\bar{E}}^{\tau} \mathbf{X}_{T^*})^{-1} \mathbf{X}_{\bar{E}} \right) y + \lambda \mathbf{X}_{\bar{E}}^{\tau} \mathbf{X}_{T^*} (\mathbf{X}_{\bar{E}}^{\tau} \mathbf{X}_{T^*})^{-1} \text{sign}(\tilde{\mu}_E) \right\|_{\infty} \leq \lambda \right\} \\
&= \left\{ \left\| \mathbf{X}_{\bar{E}}^{\tau} (y - \mathbf{X}_{T^*} \tilde{\beta}_{T^*}) \right\|_{\infty} \leq \lambda \right\}
\end{aligned} \tag{2.33}$$

所以事件 A_n 可以推得 $\left\{ \left\| \mathbf{X}_{\bar{E}}^{\tau} (y - \mathbf{X}_{T^*} \tilde{\beta}_{T^*}) \right\|_{\infty} \leq \lambda \right\}$ 成立。再将 (2.30) 代入 (2.27), 可得 $\mathbf{X}_{T^*}^{\tau} \mathbf{X} \tilde{\mu} = \text{sign}(\beta_{T^*})$, 而 (2.22) 式可以推出 $\left\| \mathbf{X}_{T^*}^{\tau} \mathbf{X} \tilde{\mu} \right\|_{\infty} \leq 1$. 现在再来考虑 (2.29) 式. 将 $\tilde{\mu}_E = 0$ 和 (2.31) 式代入 (2.29) 式的左边可得

$$\tilde{\mu}^{\tau} \mathbf{X}^{\tau} (y - \mathbf{X} \tilde{\beta}) = \tilde{\mu}_E^{\tau} \mathbf{X}_{\bar{E}}^{\tau} (y - \mathbf{X}_{T^*} \tilde{\beta}_{T^*}) = \lambda \tilde{\mu}_E^{\tau} \text{sign}(\tilde{\mu}_E) = \lambda \|\tilde{\mu}\|_1 \tag{2.34}$$

对于 (2.28) 式, 一方面, 我们有

$$\tilde{\mu}^{\tau} \mathbf{X}^{\tau} \mathbf{X} \tilde{\beta} = \tilde{\mu}_E^{\tau} \mathbf{X}_{\bar{E}}^{\tau} \mathbf{X}_{T^*} \tilde{\beta}_{T^*} = \text{sign}(\beta_{T^*})^{\tau} \tilde{\beta}_{T^*} \tag{2.35}$$

另一方面, 由于 $D\beta_{T^*} = |\beta_{T^*}|$

$$\begin{aligned}
B_n &= \left\{ \left| DC_{E, T^*}^{-1} Z_E \right| < \sqrt{n} \left\{ |\beta_{T^*}| - \frac{\lambda}{n} \left| DC_{E, T^*}^{-1} \text{sign}(\tilde{\mu}_E) \right| \right\} \right\} \\
&\subseteq \left\{ -DC_{E, T^*}^{-1} Z_E < \sqrt{n} \left\{ |\beta_{T^*}| - \frac{\lambda}{n} \left| DC_{E, T^*}^{-1} \text{sign}(\tilde{\mu}_E) \right| \right\} \right\} \\
&\subseteq \left\{ -DC_{E, T^*}^{-1} Z_E < \sqrt{n} \left\{ D\beta_{T^*} - \frac{\lambda}{n} DC_{E, T^*}^{-1} \text{sign}(\tilde{\mu}_E) \right\} \right\} \\
&= \left\{ D(\mathbf{X}_{\bar{E}}^{\tau} \mathbf{X}_{T^*})^{-1} \mathbf{X}_{\bar{E}}^{\tau} \varepsilon + D\beta_{T^*} > \lambda D(\mathbf{X}_{\bar{E}}^{\tau} \mathbf{X}_{T^*})^{-1} \text{sign}(\tilde{\mu}_E) \right\} \\
&= \left\{ D(\mathbf{X}_{\bar{E}}^{\tau} \mathbf{X}_{T^*})^{-1} \mathbf{X}_{\bar{E}}^{\tau} y - \lambda D(\mathbf{X}_{\bar{E}}^{\tau} \mathbf{X}_{T^*})^{-1} \text{sign}(\tilde{\mu}_E) > 0 \right\} \\
&\equiv \{D\tilde{\beta}_{T^*} > 0\}
\end{aligned} \tag{2.36}$$

并且 $\{D\tilde{\beta}_{T^*} > 0\} = \{\text{diag}(\text{sign}(\beta_{T^*})) \tilde{\beta}_{T^*} > 0\} \subseteq \{\text{sign}(\beta_{T^*})^{\tau} \tilde{\beta}_{T^*} = \|\tilde{\beta}_{T^*}\|_1\}$. 此式与 (2.35) 和 (2.36) 一起, 易得事件 B_n 是可以推出 (2.28) 式的。因此, 在不可表示条件下, 当事件 A_n 和事件 B_n 同时发生的时候, 由 (2.30) 和 (2.31) 定义的 $(\tilde{\mu}, \tilde{\beta})$ 满足 (2.26) – (2.29) 即: $A_n \cap B_n \subseteq \{\tilde{\beta} = \hat{\beta}^D\}$. 下面我们再来证明 $B_n \Leftrightarrow \{\tilde{\beta} = {}_s\beta\}$. 对于由 (2.30) 定义的 $\tilde{\beta}$, 事件 $\{\tilde{\beta} = {}_s\beta\} \Leftrightarrow \{\tilde{\beta}_{T^*} = {}_s\beta_{T^*}\} \Leftrightarrow \{D\tilde{\beta}_{T^*} > 0\}$, 这恰恰就是事件 B_n . 综上, 我们有 $\{\tilde{\beta} = \hat{\beta}^D\} \cap \{\tilde{\beta} = {}_s\beta\} \supseteq (A_n \cap B_n) \cap B_n = A_n \cap B_n$. 命题得证. ■

在下一小节中, 我们将会证明不可表示条件是符号相合的几乎充分必要条件。这里的“几乎充分必要条件”指的是不可表示条件是强符号相合的充分条件, 弱不可表示条件是弱符号相合的必要条件。

对于固定的 p 和 q 的符号相合

Dicker and Lin (2009) 给出了一个类似的条件对于 $n \rightarrow \infty$ 时, p 和 q 是固定的时候。为了更加深入了解我们给出的条件对于更一般的情形, 我们也从固定的 p 和 q 的情况入手。此时, 很自然地我们先给出下面的**正则条件**:

1. $C = C(n) \rightarrow C^*, n \rightarrow \infty$, 其中 C^* 是一个正定矩阵.
2. $\frac{1}{n} \max_{1 \leq i \leq n} X_i^\tau X_i \rightarrow 0, n \rightarrow \infty$

如果 $\{X_i\}$'s 是方差有限的独立同分布的随机变量, 则对于固定的 p , 有 $C = C(n) \rightarrow C^*$, 其中 $C^* = EX_1^\tau X_1$, 并且 $\max_{1 \leq i \leq n} X_i^\tau X_i = o_p(n)$ (Owcn, 2001). 因此, 条件 (a) 和 (b) 很容易满足。下面的定理就给出了符号相合性。

Theorem 2.2.8 当 p 和 q 是固定的情况下, 假设条件 (a) 与条件 (b) 成立, 并且不可表示条件成立, 则对于满足 $\lambda/n \rightarrow 0$ 和 $\lambda/n^{(c+1)/2} \rightarrow \infty, 0 \leq c < 1$ 的正数 λ , Dantzig sclccotr 估计是强符号相合的:

$$P\left(\hat{\beta}^D(\lambda) = {}_s\beta\right) = 1 - o\left(e^{-n^c}\right) \rightarrow 1, n \rightarrow \infty$$

定理3.0.1 表明, 对于固定的 p 和 q , 在不可表示条件及别的适当条件下, Dantzig sclccotr 选出真模型的概率按指数速度趋于 1。另一方面, 下面即将给出的定理 3.0.2 表明弱不可表示条件是 Dantzig sclccotr 估计量弱符号相合的必要条件。这就是为什么我们说不可表示条件是一个几乎充分必要条件, 而非充要条件。

Proof. 定理 2.1 的证明. 令 $\zeta = (\zeta_1, \zeta_2, \dots, \zeta_{p-q})^\tau = C_{E,T^*} C_{E,T}^{-1} \cdot Z_E - Z_E, \xi = (\xi_1, \xi_2, \dots, \xi_q)^\tau = DC_{E,T^*}^{-1} Z_E, h = (h_1, h_2, \dots, h_n)^\tau = DC_{E,T}^{-1} \text{sign}(\tilde{\mu}_E)$. 由命题 2.1 可得, $P(\hat{\beta}^D = {}_s\beta) \geq P(A_n \cap B_n)$. 故有

$$\begin{aligned} 1 - P\left(\hat{\beta}^D = {}_s\beta\right) &\leq 1 - P(A_n \cap B_n) \\ &= P(A_n^c \cup B_n^c) \\ &\leq P(A_n^c) + P(B_n^c) \\ &\leq \sum_{i=1}^{p-q} P\left(|\zeta_i| \geq \frac{\lambda}{\sqrt{n}} \eta_i\right) + \sum_{j=1}^q P\left(|\xi_j| \geq \sqrt{n} \left(|\beta_j| - \frac{\lambda}{n} h_j\right)\right) \end{aligned} \quad (2.37)$$

因为 ε 是一个 n 维分量独立同分布的随机向量, 在条件 (a) 和条件 (b) 下, 我们可得

$$\zeta \rightarrow_d N\left(0, \sigma^2 \left(C_{\bar{E}, \bar{T}}^* (C^*)_{E, T^*}^{-1} C_{E, E}^* (C^*)_{T^*, E}^{-1} C_{T^*, \bar{E}}^* - C_{\bar{E}, \bar{T}^*}^* (C^*)_{E, T^*}^{-1} C_{E, \bar{E}}^* - C_{\bar{E}, E}^* (C^*)_{T^*, E}^{-1} C_{T^*, \bar{E}}^* + C_{E, E}^* \right)\right) \quad (2.38)$$

和

$$\xi \rightarrow_d N\left(0, \sigma^2 \left(D (C^*)_{E, T^*}^{-1} C_{E, E}^* (C^*)_{T^*, E}^{-1} D \right)\right)$$

因此 ζ_i 's 和 ξ_j 's 依分布收敛于均值为零, 方差有限的正态分布。假设 $\forall i, j, E(\zeta_i)^2 \leq t_0^2, E(\xi_j)^2 \leq t_0^2$, 常数 $t_0 > 0$. 则对于任意的 $x > 0$, 由高斯分布的尾概率可得

$$P(\zeta_i > x) < x^{-1} e^{-x^2/2}, \quad P(\xi_j > x) < x^{-1} e^{-x^2/2} \quad (2.39)$$

$i = 1, \dots, p - q, j = 1, \dots, q$. 因此, 若 $\lambda/n \rightarrow 0, \lambda/n^{(c+1)/2} \rightarrow \infty$, 其中 $0 \leq c < 1$, 对于固定的 p 和 q , 我们有

$$\begin{aligned} & \sum_{i=1}^{p-q} P\left(|\zeta_i| \geq \frac{\lambda}{\sqrt{n}} \eta_i\right) \\ & \leq 2 \sum_{i=1}^{p-q} \left(\frac{\lambda}{\sqrt{nt_0}} \eta_i\right)^{-1} \exp\left(-\frac{1}{2} \frac{\lambda^2}{nt_0} \eta_i^2\right) \\ & = o\left(e^{-n^c}\right) \end{aligned} \quad (2.40)$$

和

$$\begin{aligned} & \sum_{j=1}^q P\left(|\xi_j| \geq \sqrt{n} \left(|\beta_j| - \frac{\lambda}{n} h_j\right)\right) \\ & = \sum_{j=1}^q P\left(|\xi_j| \geq \sqrt{n} |\beta_j| + o(\sqrt{n} |\beta_j|)\right) \\ & \leq 2 \sum_{j=1}^q \left(\sqrt{n} |\beta_j| (1 + o(1))\right)^{-1} \exp\left(-\frac{1}{2} \left(\sqrt{n} |\beta_j| (1 + o(1))\right)^2\right) \\ & = o\left(e^{-n^c}\right) \end{aligned} \quad (2.41)$$

由 (2.40), (2.41) 和 (2.37), 定理 3.0.1 得证. ■

定理 2.2.

Theorem 2.2.9 对于固定的 p 和 q , 若条件 (a) 和条件 (b) 成立, Dantzig selector 估计量是弱符号相合的, 则 $\exists N, \forall n \geq N$, 弱不可表示条件成立.

Proof. 定理 2.2 的证明. 假设 T^* 是真模型所对应的变量的下标的全体. 考虑事件 $C_1 = \{\exists \lambda \text{ s.t. } \hat{\beta}^D(\lambda) = \beta\}$, 也就是事件 $\text{sign}(\hat{\beta}_{T^*}^D) = \text{sign}(\beta_{T^*})$, 在此事件上 $\hat{\beta}_{T^*}^D = 0$. 则由引理 2.1 可得存在子集 $E \subseteq \{1, \dots, p\}$ 和 $\bar{\mu} \in \mathbb{R}^p$ 满足 $E = \{j : \bar{\mu}_j \neq 0\}$ 使得

$$\mathbf{X}_{T^*}^\tau \mathbf{X}_E \bar{\mu}_E = \text{sign}(\hat{\beta}_{T^*}^D) = \text{sign}(\beta_{T^*}) \quad (2.42)$$

$$\|\mathbf{X}^\tau \mathbf{X}_E \bar{\mu}_E\|_\infty \leq 1 \quad (2.43)$$

$$\mathbf{X}_E^\tau (y - \mathbf{X}_{T^*} \hat{\beta}_{T^*}^D) = \lambda \text{sign}(\bar{\mu}_E) \quad (2.44)$$

$$\left\| \mathbf{X}^\tau (y - \mathbf{X}_{T^*} \hat{\beta}_{T^*}^D) \right\|_\infty \leq \lambda \quad (2.45)$$

以上四式在集合 C_1 上成立. 然后, 我们将方程 (2.42) 和 (2.44) 得到的解 $\bar{\mu}_E$ 和 $\hat{\beta}_{T^*}^D$ 分别代入 (2.43) 式和 (2.45) 式中, 并且注意到 $C = \frac{1}{n} \mathbf{X}^\tau \mathbf{X}, Z_{T^*} = \frac{1}{\sqrt{n}} \mathbf{X}_{T^*}^\tau \varepsilon, Z_{\bar{T}^*} = \frac{1}{\sqrt{n}} \mathbf{X}_{\bar{T}^*}^\tau \varepsilon$, 我们有

$$\begin{aligned} C_1 \subset C_2 &:= \left\{ \left| C_{T^*,E} C_{T^*,E}^{-1} \text{sign}(\beta_{T^*}) \right| \leq 1 \right. \\ & \quad \left. \left| C_{\bar{E},T^*} C_{\bar{E},T^*}^{-1} Z_E - Z_{\bar{E}} - \frac{\lambda}{\sqrt{\eta}} C_{\bar{E},T^*} C_{\bar{E},T^*}^{-1} \text{sign}\left(C_{T^*,E}^{-1} \text{sign}(\beta_{T^*})\right) \right| \leq \frac{\lambda}{\sqrt{n}} \mathbf{1} \right\} \\ & \stackrel{\Delta}{=} H_1 \cap H_2 \end{aligned}$$

将 $H_2 = \left\{ \left| C_{\bar{E},T^*} C_{E,T^*}^{-1} Z_E - Z_{\bar{E}} - \frac{\lambda}{\sqrt{n}} C_{\bar{E},T^*} C_{E,T^*}^{-1} \text{sign} \left(C_{T^*,E}^{-1} \text{sign}(\beta_{T^*}) \right) \right| \leq \frac{\lambda}{\sqrt{n}} 1 \right\}$ 记为

$$\left\{ \frac{\lambda}{\sqrt{n}} L \leq C_{\bar{E},T^*} C_{E,T^*}^{-1} Z_E - Z_{\bar{E}} \leq \frac{\lambda}{\sqrt{n}} R \right\}$$

其中 $L = C_{\bar{E},T^*} C_{E,T^*}^{-1} \text{sign} \left(C_{T^*,E}^{-1} \text{sign}(\beta_{T^*}) \right) - \mathbf{1}$, $R = C_{\bar{E},T^*} C_{E,T^*}^{-1} \text{sign} \left(C_{T^*,E}^{-1} \text{sign}(\beta_{T^*}) \right) + \mathbf{1}$.

为了证明不可表示条件的必要性,下面我们要反证法来证明之。若不可表示条件不成立,则对于任意的正整数 N , 存在 $n, n > N$, 使得向量 $|C_{\bar{E},T^*} C_{E,T^*}^{-1} \text{sign} \left(C_{T^*,E}^{-1} \text{sign}(\beta_{T^*}) \right)|$ 至少有一个分量不比 1 小。不失一般性,不妨设 $C_{\bar{E},T^*} C_{E,T^*}^{-1} \text{sign} \left(C_{T^*,E}^{-1} \text{sign}(\beta_{T^*}) \right)$ 的第一个分量大于或等于 1, 当然同理可分析小于或等于 -1 的情况. 因此我们有

$$\left(C_{\bar{E},T^*} C_{E,T^*}^{-1} Z_E - Z_{\bar{E}} \right)_1 \in \left[\frac{\lambda}{\sqrt{n}} L_1, \frac{\lambda}{\sqrt{n}} R_1 \right] \subseteq [0, \infty)$$

一方面, 根据定理 3.0.1 的证明中的 (2.38) 式, 随着 n 的增大, $\left(C_{\bar{E},T^*} C_{E,T^*}^{-1} Z_E - Z_{\bar{E}} \right)_1$ 的第一个分量是负的概率是大于 0 的, 且事件 C_2 的概率并不趋于 1, 因此我们有

$$\liminf P(C_1) \leq \liminf P(C_2) < 1$$

这与弱符号相合的定义是矛盾的。因此, 不可表示条件是弱符号相的必要条件。定理得证。 ■

当 p 和 $q \rightarrow \infty$ 时的符号相合

现在我们来考虑 p 和 q 随着 n 的增大趋于无穷的情况。随着 n 的增大, 设计阵 C 的定义就没意义了, 所以前面的条件 (a) 和条件 (b) 的假设就不再合理了, 因此我们要考虑给出一些新的条件。假设 κ_{n1} 是矩阵 $B = C_{\bar{E},T^*}^{-1} C_{E,E} C_{T^*,E}^{-1}$ 的最大特征值, 其中 E 是满足不可表示条件中的一个。又设 τ_{n1} 是半正定阵 $(I - K)(I - K)^\top$ 的最大特征值, 其中 $K = \mathbf{X}_{T^*} (\mathbf{X}_E^\top \mathbf{X}_{T^*})^{-1} \mathbf{X}_E^\top$ 是一个幂等阵。假设存在 $0 \leq d_1 < d_2 \leq 1$ 的 d_1, d_2 使得下面几条成立。

1. (C1) $q = O(n^{d_1})$
2. (C2) $n^{\frac{1-d_2}{2}} \min_{i \in T^*} |\beta_i| \geq M_1 > 0$
3. (C3) $\|C_{\bar{E},T^*} \alpha\|_2^2 \geq M_2 > 0$ 对任意的单位向量 α 都成立, 其中 E 满足不可表示条件。
4. (C4) $0 < \kappa_{n1} \leq \kappa_1 < \infty$
5. (C5) $\tau_{n1} \leq \tau_1 < \infty$

在上面的条件和适当的误差阶的条件下, 由 (2.21) 式表示的 Dantzig selector 在不可表示条件下可以一致地选出真模型。我们将这一结论表述如下:

Theorem 2.2.10 假设 ε_i 是满足 $E(\varepsilon_i)^{2k} < \infty$ 对某个 $k > 0$ 成立的独立同分布的随机变量。在条件 (C1) – (C5) 下, 若 $p = o\left(n^{(d_2-d_1)k}\right)$, $d_2 > d_1$, 并且 λ 是满足 $\lambda/\sqrt{n} = o\left(n^{(d_2-d_1)/2}\right)$, $(\lambda/\sqrt{n})^{2k}/p \rightarrow$

∞ 的正数。则在不可表示条件下, 我们有 $\hat{\beta}^D$ 是强符号相合的, 即:

$$P\left(\hat{\beta}^D(\lambda) = {}_s\beta\right) \geq 1 - O\left(\frac{pn^k}{\lambda^{2k}}\right) \rightarrow 1, n \rightarrow \infty$$

定理 3.0.3 得到了与定理 3.0.1 类似的结论, 只是定理 3.0.3 允许 p, q 是可以随着 n 增长的甚至可以比 n 大。这两个定理表示在不可表示条件下及适当的误差阶的条件下, 无论 p, q 是固定的还是随着 n 增长的, Dantzig selector 都是模型选择相合的。

Lemma 2.7 假设 $\theta = (\theta_1, \theta_2, \dots, \theta_n)^\tau$ 是一个各分量间独立同分布的随机向量, 满足 $E(\theta_1)^{2k} < \infty$ 对某 $k > 0$ 成立. 则对于一个常数向量 α , 有下面的不等式成立

$$E(\alpha^\tau \theta)^{2k} \leq (2k-1)!! \|\alpha\|_2^2 E(\theta_1)^{2k}$$

Proof. 定理 2.3 的证明. 由命题 2.1, $P(\hat{\beta}^D = {}_s\beta) \geq P(A_n \cap B_n)$. 则有

$$\begin{aligned} 1 - P(\hat{\beta}^D = {}_s\beta) &\leq 1 - P(A_n \cap B_n) \\ &= P(A_n^c \cup B_n^c) \\ &\leq P(A_n^c) + P(B_n^c) \\ &\leq \sum_{i=1}^{p-q} P\left(|\zeta_i^1| \geq \frac{\lambda}{\sqrt{n}}\eta\right) + \sum_{j=1}^q P\left(|\zeta_j^1| \geq \sqrt{n}\left(|\beta_j| - \frac{\lambda}{n}h_j\right)\right) \end{aligned}$$

其中 $\zeta^1 = (\zeta_1^1, \zeta_2^1, \dots, \zeta_{p-q}^1)^\tau = C_{\bar{E}, T^*} C_{E, T^*}^{-1} Z_E - Z_{\bar{E}}, \zeta^1 = (\zeta_1^1, \zeta_2^1, \dots, \zeta_q^1)^\tau = DC_{E, T^*}^{-1} Z_E$. $h = (h_1, h_2, \dots, h_n)^\tau = DC_{E, T^*}^{-1} \text{sign}(\tilde{\mu}_E)$. 记 $\zeta^1 = G^\tau \varepsilon$, 其中 $G^\tau = (G_1, G_2, \dots, G_{p-q})^\tau = \frac{1}{\sqrt{n}} (C_{\bar{E}, T^*} C_{E, T^*}^{-1} \mathbf{X}_E^\tau - \mathbf{X}_{\bar{E}}^\tau)$. 则有

$$\begin{aligned} G^\tau G &= \frac{1}{n} (C_{\bar{E}, T^*} C_{E, T^*}^{-1} \mathbf{X}_E^\tau - \mathbf{X}_{\bar{E}}^\tau) (\mathbf{X}_E C_{T^*, E}^{-1} C_{T^*, \bar{E}} - \mathbf{X}_{\bar{E}}) \\ &= C_{\bar{E}, T^*} C_{E, T^*}^{-1} C_{E, E} C_{T^*, E}^{-1} C_{T, E} - C_{\bar{E}, T^*} C_{E, T^*}^{-1} C_{E, E} - C_{E, E} C_{T^*, E}^{-1} C_{T^*, \bar{E}} + C_{E, E} \\ &= \frac{1}{n} \mathbf{X}_{\bar{E}}^\tau (I - K - K^\tau + KK^\tau) \mathbf{X}_{\bar{E}} \\ &= \frac{1}{n} \mathbf{X}_{\bar{E}}^\tau (I - K) (I - K)^\tau \mathbf{X}_{\bar{E}} \end{aligned}$$

其中 $K = \mathbf{X}_{T^*} (\mathbf{X}_E^\tau \mathbf{X}_{T^*})^{-1} \mathbf{X}_E^\tau$. 所以, 由条件 (C5) 和 $(X^j)^\tau (X^j) / n = 1$, 我们有

$$\|G_i\|_2^2 = G_i^\tau G_i = e_i^\tau G^\tau G e_i \leq \tau_1 < +\infty \quad (2.46)$$

对任意的 $i = 1, 2, \dots, p - q$ 成立. 同理, 记 $\zeta^1 = H^\tau \varepsilon$, 其中 $H^\tau = (H_1, H_2, \dots, H_q)^\tau = \frac{1}{\sqrt{n}} DC_{E, T^*}^{-1} \mathbf{X}_E^\tau$. 则有

$$H^\tau H = \frac{1}{n} DC_{E, T^*}^{-1} \mathbf{X}_E^\tau \mathbf{X}_E C_{T^*, E}^{-1} D = DC_{E, T^*}^{-1} C_{E, E} C_{T^*, E}^{-1} D$$

由条件 (C4) 和 $D^2 = I$, 我们有 $\xi_j^1 = H_j^\top \varepsilon$, 其中

$$\|H_j\|_2^2 \leq \kappa_1 < +\infty \quad (2.47)$$

有了 (2.46), (2.47) 和条件 $E(\varepsilon_1)^{2k} < \infty$, 由引理 2.7 可得

$$\begin{aligned} E(\xi_i^1)^{2k} &< \infty, \quad i = 1, \dots, p - q \\ E(\xi_j^1)^{2k} &< \infty, \quad j = 1, \dots, q \end{aligned}$$

从而由切比雪夫不等式可得

$$P(|\xi_i^1| > t) = O(t^{-2k}), \quad P(|\xi_j^1| > t) = O(t^{-2k})$$

对所有的 $i = 1, \dots, p - q, j = 1, \dots, q$ 成立. 显然, 由 (??) 的第一个表达式, 我们可得

$$\sum_{i=1}^{p-q} P\left(|\xi_i^1| \geq \frac{\lambda \eta}{\sqrt{n}}\right) = (p - q)O\left(\frac{n^k}{\lambda^{2k}}\right) = O\left(\frac{pn^k}{\lambda^{2k}}\right) \quad (2.48)$$

另外, 由条件 (C3) 可得

$$\left\|\frac{\lambda}{n}h\right\|_\infty = \left\|\frac{\lambda}{n}DC_{E,T^*}^{-1} \text{sign}(\tilde{\mu}_E)\right\|_\infty \leq \left\|\frac{\lambda}{n}DC_{E,T^*}^{-1} \text{sign}(\tilde{\mu}_E)\right\|_2 \leq \frac{\lambda\sqrt{q}}{n\sqrt{M_2}}$$

由 $\lambda/\sqrt{n} = o(n^{(d_2-d_1)/2})$ 及条件 (C1) 可得

$$\left\|\frac{\lambda}{n}h\right\|_\infty = o(n^{(d_2-1)/2}) \quad (2.49)$$

因此, 当 $\lambda/\sqrt{n} = o(n^{(d_2-d_1)/2})$ 时, 且条件 (C1) 与 (C2) 都成立的时候, 根据 (2.49) 和 (??) 的第二个表达式, 我们有

$$\begin{aligned} &\sum_{j=1}^q P\left(|\xi_j^1| \geq \sqrt{n}\left(|\beta_j| - \frac{\lambda}{n}h_j\right)\right) \\ &= \sum_{j=1}^q P\left(|\xi_j^1| \geq \sqrt{n}|\beta_j| + o(\sqrt{n}|\beta_j|)\right) \\ &= qO(n^{-kd_2}) \\ &= o\left(\frac{pn^k}{\lambda^{2k}}\right) \end{aligned} \quad (2.50)$$

当 $p = o(n^{(d_2-d_1)k})$ 且 $(\lambda/\sqrt{n})^{2k}/p \rightarrow \infty$ 时, 结合 (??), (2.50) 和 (??) 三式, 可得

$$P(\hat{\beta}^D =_s \beta) \geq 1 - O\left(\frac{pn^k}{\lambda^{2k}}\right) \quad (2.51)$$

■

若不可表示条件的 E 满足 $E = T^*$ 时, 则条件 (C4) 可以由条件 (C3) 推得, 而条件 (C5) 显然成立, 因为此时 $(I - K)(I - K)^\top$ 是一个幂等阵, 它的特征值不会超过 1, 所以我们有如下的推论:

推论 2.1。

Corollary 2.2.11 假设定理 3.0.3 中的条件除了条件 (C4), (C5) 其余都成立, 则当不可表示条件对于 $E = T^*$ 成立时,

$$P\left(\hat{\beta}^D(\lambda) = {}_s\beta\right) \geq 1 - O\left(\frac{pn^k}{\lambda^{2k}}\right) \rightarrow 1, n \rightarrow \infty$$

$p = \exp(n^c)$ 且 $q \rightarrow \infty$ 时的符号相合

从定理 3.0.3, 我们知道若对某个正整数 $k > 0$, 当误差项有有限的 $2k$ 次阶矩的话, 我们就可以得到 Dantzig selector 的符号相合的结论。但是, 即使误差的所有阶矩都是有限的, 即: 可以取任意大的正整数, 变量个数 p 最多也就是关于 n 成多项式增长速度, 永远达不到关于 n 的指数阶的增长速度。在这一节中, 我们将会给出 p 关于 n 呈指数速度增长时候的 Dantzig selector 的符号相合的结果。得到这样的结果的代价就是, 误差矩有限的条件要替换为更强的误差是次高斯的条件。我们有如下的定理。

Lemma 2.8 — Huang et al(2008). 设 $\varepsilon_1, \dots, \varepsilon_n$ 是均值和方差分别为 $E\varepsilon_i = 0$ 和 $\text{Var}(\varepsilon_i) = \sigma^2$ 的独立同分布的随机变量。又设它们的尾概率满足 $P(|\varepsilon_i| > x) \leq K \exp(-Cx^d), i = 1, 2, \dots$ 对于某正数 C 和正数 K , 以及 $1 \leq d \leq 2$ 成立。则对于所有满足 $\sum_{i=1}^n k_i^2 = 1$ 的常量 k_i , 我们有

$$f_n(t) = \sup_{\sum_{i=1}^n k_i^2 = 1} P\left\{\left|\sum_{i=1}^n k_i \varepsilon_i\right| > t\right\} \leq \begin{cases} \exp(-t^d/M), & 1 < d \leq 2 \\ \exp(-t^d/\{M(1 + \log n)\}), & d = 1 \end{cases}$$

其中 M 是依赖于 $\{d, K, C\}$ 的正数。 ■

Theorem 2.2.12 假设条件 (C1)-(C5) 成立, ε_i 是独立同分布的随机变量, 并且对所有的 $x \geq 0$ 满足 $P(|\varepsilon_i| > x) \leq K \exp(-Cx^d), i = 1, 2, \dots$, 其中 $1 \leq d \leq 2, C > 0$ 和 K 都是给定的常数. $p = e^{n^{d_3}}, \left(\frac{1}{\log n}\right)^{I\{d=1\}} \left(\frac{\lambda}{\sqrt{n}}\right)^d = O(n^{d_4})$, 其中 $0 < d_3 < d_4 < d_2 d/2$, 则在不可表示条件满足的条件下, 我们可得 $\hat{\beta}^D$ 是强符号相合的, 即:

$$P\left(\hat{\beta}^D(\lambda) = {}_s\beta\right) \geq 1 - O\left(e^{-n^\delta}\right) \rightarrow 1, n \rightarrow \infty$$

其中 $\delta = \min\{d_4 - d_3, d_2 d/2\}$

Proof. 定理 3.0.4 的证明. 与定理 3.0.3 的证明类似, 我们有

$$\begin{aligned} 1 - P\left(\hat{\beta}^D = {}_s\beta\right) &\leq 1 - P(A_n \cap B_n) \\ &= P(A_n^c \cup B_n^c) \\ &\leq P(A_n^c) + P(B_n^c) \end{aligned} \tag{2.52}$$

根据定理 3.0.3 的证明, $G^T G = \frac{1}{n} \mathbf{X}_E^T (I - K)(I - K)^T \mathbf{X}_E$ 其中 $K = \mathbf{X}_{T^*} (\mathbf{X}_E^T \mathbf{X}_{T^*})^{-1} \mathbf{X}_E^T$, 则对于 $\zeta_i^1 = G_i^T \varepsilon = e_i^T G^T \varepsilon$, 其中 e_i 是第 i 个元素为 1, 其余全为 0 的单位向量. 由条件 (C5), 我们有 $\|G_i\|_2^2 = G_i^T G_i = e_i^T G^T G e_i \leq \tau_{n1}, i = 1, \dots, p - q$. 因此对于 $p = e^{n^{d_3}} \left(\frac{1}{\log n}\right)^{I\{d=1\}} \left(\frac{\lambda}{\sqrt{n}}\right)^d =$

$O(n^{d_4})$ 和 $0 < d_3 < d_4 < d_2 d/2, 1 \leq d \leq 2$, 由引理 2.8, 可得

$$\begin{aligned}
 P(A_n^c) &= \sum_{i=1}^{p-q} P\left\{\left|\zeta_i^1\right| > \frac{\lambda}{\sqrt{n}}\eta\right\} \\
 &= \sum_{i=1}^{p-q} P\left\{\left|\frac{1}{\sqrt{\tau_{n1}}}G_i^\tau \varepsilon\right| > \frac{\lambda}{\sqrt{n\tau_{n1}}}\eta\right\} \\
 &\leq (p-q)P\left\{\left|\frac{1}{\|G_1\|_2}G_1^\tau \varepsilon\right| > \frac{\lambda}{\sqrt{n\tau_1}}\eta\right\} \\
 &\leq (p-q)f_n\left(\frac{\lambda}{\sqrt{n\tau_1}}\eta\right) = O\left(e^{n^{d_3-d_4}}\right)
 \end{aligned} \tag{2.53}$$

类似地, 由条件 (C4), 我们有 $\|H_j\|_2^2 \leq \kappa_{n1}, j = 1, \dots, q$. 因此在条件 (C1) (C2) 和条件 (C4) 下, 和 $0 < d_3 < d_4 < d_2 d/2$, 由引理 2.3, 我们有

$$\begin{aligned}
 P(B_n^c) &= \sum_{j=1}^q P\left(\left|\xi_j^1\right| \geq \sqrt{n}\left(|\beta_j| - \frac{\lambda}{n}h_j\right)\right) \\
 &= \sum_{j=1}^q P\left(\left|H_j^\top \varepsilon\right| \geq \sqrt{n}|\beta_j|(1+o(1))\right) \\
 &= \sum_{j=1}^q P\left(\left|\frac{1}{\sqrt{\kappa_{n1}}}H_j^\tau \varepsilon\right| \geq \sqrt{\frac{n}{\kappa_{n1}}}|\beta_j|(1+o(1))\right) \\
 &\leq qP\left(\left|\frac{1}{\|H_1\|}H_1^\top \varepsilon\right| \geq \sqrt{\frac{n}{\kappa_1}}|\beta_j|(1+o(1))\right) \\
 &\leq qf_n\left(\frac{M_1 n^{d_2/2}}{\sqrt{\kappa_1}}(1+o(1))\right) = O\left(e^{-n^{d_2 d/2}}\right)
 \end{aligned} \tag{2.54}$$

结合 (2.52)(2.53) 和 (2.54), 以及 $\delta = \min\{d_4 - d_3, d_2 d/2\}$, 我们可得

$$P\left(\hat{\beta}^D =_s \beta\right) \geq 1 - O\left(e^{-n^\delta}\right) \rightarrow 1, n \rightarrow \infty$$

■

当 $p \gg n$ 时, 若不可表示条件中的子集 $E = T^*$ 时, 我们可以得到与推论 2.2.11 平行的一个推论。

Corollary 2.2.13 假设定理 3.0.4 中的条件除了条件 (C4), (C5) 其余都成立, 若不可表示条件中的 $E = T^*$ 时, 则我们有

$$P\left(\hat{\beta}^D(\lambda) =_s \beta\right) = 1 - O\left(e^{-n^\delta}\right) \rightarrow 1, n \rightarrow \infty$$

其中 $\delta = \min\{d_4 - d_3, d_2 d/2\}$

与定理 3.0.3 相比, 定理 3.0.4 及它的推论告诉我们, 当误差项满足次高斯的时候, 即使 p 关于 n 是指数阶增长的, Dantzig selector 的符号相合也是可以达到的。

R 从上面几节的结论, 我们可以清楚地看到, 尽管 Dantzig selector 和 LASSO 得到的估计都是符号相合的, 但是二者所需的不可表示条件是完全不同的。这也再一次的证明这两种方法的估计的相合性是依格于不同的相关结构的。

变量选择后估计的相合性

从前面章节的定理中, 我们可以看到, 所有显著变量都被选入子模型, 而所有不显著变量都未被选入子模型是依概率 1 成立的。在这一节中, 我们来考虑变量选择以后, 常规估计的相合性。我们可以得到只要 $q = o(n)$, 则相合性成立。我们记 $n \times |\hat{T}|$ 维的设计阵为 $\mathbf{X}_{\hat{T}} = (X_{1\hat{T}}, \dots, X_{n\hat{T}})^\tau$. 则 β 的变量选择后的子模型的最小二乘估计 $\hat{\beta}$ 定义为

$$\hat{\beta}_{\hat{T}} = C_{\hat{T}, \hat{T}}^{-1} \left\{ \frac{1}{n} X_{\hat{T}}^{-\tau} Y \right\} = C_{\hat{T}, \hat{T}}^{-1} \left\{ \frac{1}{n} \sum_{i=1}^n X_{i\hat{T}} Y_i \right\}, \hat{\beta}_{\hat{T}^c} = 0$$

Theorem 2.2.14 假设 Dantzig selector 估计量是强符号相合的, 并且 $\max_{1 \leq i \leq n, 1 \leq j \leq q} x_{iT^*, j}^2 < \infty$ 成立, 其中 $X_{iT^*}^\tau = (x_{iT^*, 1}, \dots, x_{iT^*, q})$ 对应于矩阵 X_{T^*} 的第 i 行, 则有

$$\|\hat{\beta} - \beta\|_2 = O_p \left(\sqrt{\frac{q}{n}} \right)$$

上述定理中的相合性结果与普通的最小二乘估计是完全不同的, 这是因为 $X_{\hat{T}}$ 的维数 $|\hat{T}|$ 是一个随机变量而不是一个固定的数

Proof. 记事件 $\Delta_n = \{\text{sign}(\hat{\beta}) = \text{sign}(\beta)\}$. 在事件 Δ_n 上, $\hat{T} = T^*$ 是固定的, 即为非随机的. 则对于任意的 $\tau > 0$, 有

$$\begin{aligned} & P(|\hat{\beta} - \beta| > \tau) \\ &= P(|\hat{\beta}_{\hat{T}} - \beta_{T^*}| > \tau) \\ &\leq P(|\hat{\beta}_{\hat{T}} - \beta_{T^*}| > \tau, \Delta_n) + P(\Delta_n^c) \\ &= P(|\hat{\beta}_{\hat{T}} - \beta_{T^*}| > \tau | \Delta_n) P(\Delta_n) + P(\Delta_n^c) \end{aligned}$$

若符号相合成立, 则概率 $P(\Delta_n^c)$ 随着 n . 趋于无穷, 是逐渐趋于零的. 因此我们只要证明, 在事件 Δ_n 上, $P(|\hat{\beta}_{T^*} - \beta_{T^*}| > \tau)$ 趋于零即可即只要能够证明, 在事件 Δ_n 上, 有

$$\left\| C_{T^*, T^*}^{-1} \left\{ \frac{1}{n} \sum_{i=1}^n X_{iT^*} \varepsilon_i \right\} \right\|_2 = O_p \left(\sqrt{\frac{q}{n}} \right) \quad (2.55)$$

由于 $\max_{1 \leq i \leq n, 1 \leq j \leq q} x_{iT^*, j}^2 < \infty$, 所以

$$\begin{aligned} E \left\| \left\{ \frac{1}{n} \sum_{i=1}^n X_{iT^*} \varepsilon_i \right\} \right\|_2^2 &= \frac{1}{n^2} \sum_{i=1}^n \|X_{iT^*}\|_2^2 E\{\varepsilon_i^2\} \\ &= O\left(\frac{q}{n}\right) \end{aligned}$$

根据马尔可夫不等式, 可得

$$\left\| \left(\frac{1}{n} \sum_{i=1}^n X_{iT} \cdot \varepsilon_i \right) \right\|_2^2 = O_p \left(\frac{q}{n} \right)$$

从而有

$$\begin{aligned} \left\| C_{T^*, T}^{-1} \left\{ \frac{1}{n} \sum_{i=1}^n X_{iT^*} \varepsilon_i \right\} \right\|_2^2 &= \text{trace} \left(\left(\frac{1}{n} \sum_{i=1}^n X_{iT^*} \varepsilon_i \right)^\tau C_{T^*, T^*}^{-1} \cdot C_{T^*, T^*}^{-1} \left(\frac{1}{n} \sum_{i=1}^n X_{iT^*} \cdot \varepsilon_i \right) \right) \\ &= O \left(\left\| \left(\frac{1}{n} \sum_{i=1}^n X_{iT^*} \varepsilon_i \right) \right\|_2^2 \right) = O_p \left(\frac{q}{n} \right) \end{aligned}$$

上面的第二个等号是由条件 (C2) 得到的. 故等式 (2.55) 成立。 ■

尽管 DS 方法具备不少较好的理论性质, 但是也存在不具有相合性, 不完全满足 Oracle 性质等缺陷。此外, DS 法还有一个弊端, 对每个变量系数的压缩力度一样, 容易产生对具有重要影响的变量的系数过度压缩的现象。为了解决这一问题, Dicker 和 Lin [30] 在线性模型框架下提出了 ADS 方法。

2.2.8 ADS

我们发现 Dantzig 选择器的惩罚方式有些不公平, 因为所有大小的系数惩罚的程度都是一样的。因此, 在第三章中, 我们给不同大小的系数施加不同程度的惩罚, 给出了一种加权的 Dantzig 选择器, 这也就是所谓的适应的 Dantzig 选择器。对于适应的 Dantzig 选择器, 我们研究了它在稀疏高维线性模型下的, 对于不同大小的 p 的渐近性质。我们证明了只要能得到一个合理的初始估计, 在适当的条件下, 而无需满足不可表示条件, 适应的 Dantzig 选择器具有 oracle 性质, 不管 p 以多项式的速度还是以指数的速度趋于无穷。即适应的 Dantzig 选择器的解 $\hat{\beta}(ADS)$ 满足下面两条:

$$1. P(\hat{\beta}(ADS) =_s \beta) \rightarrow 1, \quad n \rightarrow +\infty$$

2. $\sqrt{n} s_n^{-1} \phi_n' (\hat{\beta}(ADS)_T - \beta_T) \xrightarrow{D} N(0, 1)$, 其中 $s_n^2 = \sigma^2 \phi_n' C_{T, T}^{-1} \phi_n$, $\phi_n \in \mathbf{R}^q$ 满足 $\|\phi_n\| \leq 1$

适应的 Dantzig selector 是 Dantzig selector 的一种推广, 通过在原有的最优化问题的目标和约束中添加依赖于数据的权重 w_1, \dots, w_p . 它的具体定义为如下的约束问题

$$\hat{\beta}(ADS) = \arg \min_{\beta} \sum_{j=1}^p w_j |\beta_j| \quad \text{s.t.} \quad \left\| X_j^T (y - \mathbf{X}\beta) \right\|_{\infty} \leq \lambda w_j, j = 1, \dots, p$$

其中 $w_1, \dots, w_p, w_i = |\check{\beta}_i|^{-r} > 0, r > 0, \check{\beta}_i$ 是 β_i 的一个初始估计, $i = 1, \dots, p$. 令 $\beta^0 = \mathbf{W}\beta, \mathbf{Z} = \mathbf{XW}^{-1}$. 则上述的适应的 Dantzig selector 最优化问题 (1.1.8) 可以转化为

$$\min_{\beta^0 \in \mathbf{R}^p} \|\beta^0\|_1 \quad \text{subject to} \quad \left\| \mathbf{Z}' (y - \mathbf{Z}\beta^0) \right\|_{\infty} \leq \lambda$$

所以, 求解适应的 Dantzig selector 实际上等价于求解某个 Dantzig selector 问题, 这就为我们的计算提供了极大的便利。在第三章中, 我们证明了不需要不可表示条件成立的假设,

适应的 Dantzig selector 具有 oracle 性质, 这一结论对于不同大小的 p 都成立。我们将主要结果叙述如下:

定理 1.5.

Theorem 2.2.15 (Oracle 性质) 假设条件 (A1) – (A5) 成立, $p = o\left(n^{(c_2-c_1)k}\right)$, 其中 k 是在条件 (A1) 中定义的, λ 是满足 $\lambda/\sqrt{n} = o\left(n^{(c_2-c_1)/2}\right) \cdot (\lambda/\sqrt{n})^{2k}/p \rightarrow \infty$ 的正数. 找我们有

1. $P(\hat{\beta}(\text{ADS}) =_s \beta) \rightarrow 1, \quad n \rightarrow +\infty$
2. 令 $s_n^2 = \sigma^2 \phi_n' C_{T,T}^{-1} \phi_n$, ϕ_n 是 \mathbb{R}^q 中满足 $\|\phi_n\| \leq 1$ 的任意向量.

如果 λ 还满足 $\lambda n^{c_1+n_2-1/2} \rightarrow 0$, 同时条件 (A6) 也成立, 则有

$$\sqrt{n} s_n^{-1} \phi_n' (\hat{\beta}(\text{ADS})_{T^*} - \beta_{T^*}) \xrightarrow{D} N(0, 1)$$

定理 1.6.

Theorem 2.2.16 (Oracle 性质) 假设条件 (A1'), (A2) – (A5) 成立, 并且 $(\log n)^{I\{d=1\}} (\log p)^{\frac{1}{d}} / \left(\lambda n^{(\alpha_1-\frac{1}{2})} \rightarrow 0, \lambda n^{(x_1-\frac{1}{2})} \rightarrow \infty \right)$. 则找我们有

1. $P(\hat{\beta}(\text{ADS}) =_s \beta) \rightarrow 1, \text{ as } n \rightarrow +\infty$
2. 如果, 进一步, $\lambda n^{c_1+\alpha_2-1/2} \rightarrow 0$. 并且条件 (A6) 成立,

则有

$$\sqrt{n} s_n^{-1} \phi_n' (\hat{\beta}(\text{ADS})_{T^*} - \beta_{T^*}) \xrightarrow{D} N(0, 1)$$

其中 ϕ_n, s_n 与定理 1.5 (b) 中的含义相同.

ADS 法对不同的系数 β_j 采用了不同的权重 ω_j , 与 DS 方法类似, ADS 法定义为在约束条件 $\|X^T r\|_\infty \leq \lambda \omega_j$ 下

$$\min \sum_{j=1}^p \omega_j |\hat{\beta}_j|$$

其中, r 仍表示残差向量 $Y - X\tilde{\beta}$. 既然 ADS 法通过引入权重 ω_j 有效解决了 DS 法所有系数压缩力度一样的弊端, 那么如何确定 ω_j 呢? 自然地, 当 β_j 对应的变量影响很大时, ω_j 应取较小值, 使压缩力度较小; 相反若 β_j 对应的变量影响不是很大时, ω_j 应取较大值, 增加压缩力度。这样就能实现对重要变量和不重要变量压缩力度的区别对待。一般的, 取 $\omega_j = f(|\hat{\beta}_1|)$, $\hat{\beta}_1$ 是 β 的 \sqrt{n} 相合估计, $f(\cdot)$ 是恒为正数的应函数, 且 $f(0) = +\infty$. 在 2013 年和 2011 年, Dicker 和 Lin 和盖玉洁各自证明了在不同维数情况下 ADS 法均具有 Oracle 性质, 这里不作详细叙述, 感兴趣读者可自行阅读。

2.2.9 DASSO

DS 法和 ADS 法给解决高维数据结构问题提供了理论支撑, 但是要想得到广泛应用, 还需要高效的算法支撑。如同 LARS 之于 Lasso, James 等人提出了 DASSO 算法, 该方法无论从计算量还是稳定性上都比较完美的解决了 DS 问题。DASSO 算法的主要步骤如下:

1. 第 1 步: 初始化 β^l 为 p 维零向量, 且令 $l = 1$
2. 第 2 步: 令 B 为向量 β^l 中非零系数的索引集, 假设 $c = X^T(Y - X\beta^l)$, 令 A 为 c 中绝对值最大的协变量的有效索引集, s_A 为 A 中协变量没有取绝对值之前的向量集;
3. 第 3 步: 标记每一个加入 B 集或者从 A 集中剔除的协变量。用新的 A 和 B 集计算 $|B|$ 维方向向量 $h_B = (X_A^T X_B)^{-1} s_A$, 令 h 为 p 维向量, 且与 B 对应的地方设为 h_B , 其余部分设为 0;
4. 第 4 步: 计算在方向向量 h 上的距离 γ 直到一个新的变量进入活动集或者一个系数趋于零. 令 $\beta^{l+1} \leftarrow \beta^l + \gamma h, l \leftarrow l + 1$.
5. 第 5 步: 重复第 2-4 步直到 $\|c\|_\infty = 0$

DASSO 与 LARS 算法都是采用分段线性步骤的算法。James, Radchenko 和 Lv 在文章中证明了 DASSO 算法的一些良好性质, 详细介绍了 DASSO 与 LARS 算法的相似点并指出 DS 与 Lasso 法的等价条件。此外, 解决 DS 问题的算法还有 Primal dual pursuit 和非单调梯度法等, 但前者计算量比 DASSO 大, 后者推广性不如 DASSO 算法。

Group Lasso 和 CAP 选择群组变量时具有“all-in-all-out”的特点, 即一组变量要么全被选入要么全被易除, 而无法在组内选择重要的变量。但是在某些应用中, 这类方法并不十分理想, 例如研究某一疾病发病的影响因素, 一个基因由一组变量来描述, 很显然这组变量中并非每一个都会对该病有显著影响。分析这类问题最理想的方法是既能选择重要变量组又能在组内选择重要变量, 因此产生了双层变量选择方法。

2.2.10 Group Bridge

Huang 等提出 Group Bridge, 它是最早的双层变量选择方法。Group Bridge 在组内进行 Lasso 惩罚, 组间进行 Bridge 惩罚, 惩罚估计如下

$$\hat{\beta}^{Gbridge} = \operatorname{argmin} \left\{ L(\beta | y, X) + \sum_{j=1}^J \lambda p_j^\gamma \left\| \beta^{(j)} \right\|_1^\gamma \right\}$$

由于 Lasso 和 Bridge 都具有单个变量选择的效果, 因此 Group Bridge 具有双层选择功能。式中 $0 < \gamma < 1$, Zhou 等 [33] 提出的方法就是 $\gamma = 0.5$ 时的 Group Bridge。

Huang 等证明了当 $p \rightarrow \infty, n \rightarrow \infty$ 但 $p < n$ 时, 在某些正则条件下, Group Bridge ($0 < \gamma < 1$) 具有群组 Oracle 性质, 即正确选择重要组变量的概率收敛到 1。Group Bridge 中组内大系数的存在可能阻止其他同类变量进入模型, 因此尽管具有群组 Oracle 性质, 但是在组内不具备相合性。值得注意的是, Group Bridge 的目标函数是非凸的, 且在 $\beta_j = 0$ 处不可微。

2.2.11 Group MCP

Group Bridge 在某些点不可微性, 这为求解带来困难。因此 Breheny 和 Huang 提出了 Group MCP (或 Composite MCP), 其组内和组间惩罚都是 MCP 函数, 惩罚估计为

$$\hat{\beta}^{GMCP} = \operatorname{argmin} \left\{ L(\beta | y, X) + \sum_{j=1}^J f_{\lambda, b}^{MCP} \left(\sum_{k=1}^{p_j} f_{\lambda, a}^{MCP} \left(|\beta_k^{(j)}| \right) \right) \right\}$$

由于当且仅当组内达到了最大值，组间惩罚达到最大值，因此 Breheny 和 Huang 规定 $b = p_j a \lambda / 2$ 。此外 Group MCP 具有组内和组间的相合性。

2.2.12 MCP

For squared-error loss we pose the (nonconvex) optimization problem

$$\text{minimize}_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \sum_{j=1}^p P_{\lambda, \gamma}(\beta_j) \right\} \quad (2.56)$$

with the MC+ penalty on each coordinate defined by

$$P_{\lambda, \gamma}(\theta) := \int_0^{|\theta|} \left(1 - \frac{x}{\lambda \gamma} \right)_+ dx \quad (2.57)$$

With coordinate descent in mind, we consider solving a one-dimensional version of (2.56) (in standardized form)

$$\text{minimize}_{\beta \in \mathbb{R}^1} \left\{ \frac{1}{2} (\beta - \tilde{\beta})^2 + \lambda \int_0^{|\beta|} \left(1 - \frac{x}{\lambda \gamma} \right)_+ dx \right\} \quad (2.58)$$

The solution is unique for $\gamma > 1$ and is given by

$$\mathcal{S}_{\lambda, \gamma}(\tilde{\beta}) = \begin{cases} 0 & \text{if } |\tilde{\beta}| \leq \lambda \\ \text{sign}(\tilde{\beta}) \left(\frac{|\tilde{\beta}| - \lambda}{1 - \frac{1}{\gamma}} \right) & \text{if } \lambda < |\tilde{\beta}| \leq \lambda \gamma \\ \tilde{\beta} & \text{if } |\tilde{\beta}| > \lambda \gamma \end{cases} \quad (2.59)$$

Large values of $\tilde{\beta}$ are left alone, small values are set to zero, and intermediate values are shrunk. As γ gets smaller, the intermediate zone gets narrower, until eventually it becomes the hard-thresholding function of best subset (orange curve in figure). By contrast, the threshold functions for the ℓ_q family ($q < 1$) are discontinuous in $\tilde{\beta}$.

Mazumder, Friedman and Hastie (2011) exploit the continuity of $\mathcal{S}_{\lambda, \gamma}$ (in both λ and γ) in a coordinate-descent scheme for fitting solution paths for the entire MC+ family. Starting with the lasso solution, their R package sparsenet (Mazumder, Hastie and Friedman 2012) moves down a sequence in γ toward sparser models, and for each fits a regularization path in λ . Although it cannot claim to solve the nonconvex problem (2.56), this approach is both very fast and appears to find good solutions.

MCP 回归的惩罚函数为

$$\begin{cases} P_{\lambda}^{\text{MCP}}(t)' = \frac{(a\lambda - t)_+}{a}, t \geq 0, a > 1 \\ P_{\lambda}^{\text{SCAD}}(0) = 0 \end{cases}$$

由上式可见 MCP 的惩罚力度随 β_i 的增大而减小，是对回归系数进行有差别地惩罚，因此估计结果更精确。当 $\beta_i > a\lambda$ 时为零，同样达到渐近无偏性。在文献中也证明了上述两种方法具有选择一致性。

2.3 Lasso 惩罚的性质

Proposition 2.3.1 — Sparsity. If the budget t is small enough, the lasso yields sparse solution vectors, having only some coordinates that are nonzero.

1. No sparsity: ℓ_q norms with $q > 1$;
2. No convex: for $q < 1$, the solutions are sparse but the problem is not convex and this makes the minimization very challenging computationally.
3. The value $q = 1$ is the smallest value that yields a convex problem.

Proposition 2.3.2 — 完全线性相关下 Lasso 解的不唯一性. The non-full-rank case can occur when $p \leq N$ due to collinearity, and always occurs when $p > N$. In the latter scenario, there are an infinite number of solutions $\hat{\beta}$ that yield a perfect fit with zero training error.

记 $\hat{\beta} = \operatorname{argmin} \|Y - X\beta\|_2 + \lambda|\beta|_1$, 当 $x_i = x_j$ 时, 令定义 β^* 如下:

$$\hat{\beta}_k = \begin{cases} \hat{\beta}_k, & k \neq i, k \neq j \\ (\hat{\beta}_i + \hat{\beta}_j) \times (s), & k = i \\ (\hat{\beta}_i + \hat{\beta}_j) \times (1 - s), & k = j \end{cases}$$

则对于 $\forall s \in [0, 1]$, β^* 均是方程的最小解.

Proof. 先证明 $\hat{\beta}_i \hat{\beta}_j \geq 0$. 若 $\hat{\beta}_i \hat{\beta}_j \leq 0$, 因为 $X\hat{\beta}^* = X\hat{\beta}$, 且 $\|Y - X\hat{\beta}^*\|_2 = \|Y - X\hat{\beta}\|_2$, 令 $s = \frac{1}{2}$ 则 $|\hat{\beta}^*| < |\hat{\beta}|$, 从而导致 $\eta(\hat{\beta}^*) < \eta(\hat{\beta})$, 这与 $\hat{\beta}$ 是方程的解矛盾, 因此 $\hat{\beta}_i \hat{\beta}_j \geq 0$. 在 $\hat{\beta}_i \hat{\beta}_j \geq 0$ 的基础上, 结论显而易见地成立. 该例子说明当自变量中存在完全线性完全的自变量时, Lasso 无法给出一个唯一解, 这也就意味着这两个变量存在多种可能性: 可能同时被选择, 或同时剔除, 或选择其中一个而剔除另一个, 无法保证这两个变量被同时选出或剔除. ■

In general, when $\lambda > 0$, one can show that if the columns of the model matrix X are in **general position**, then the lasso solutions are unique. To be precise, we say the columns $\{\mathbf{x}_j\}_{j=1}^p$ are in general position if any affine subspace $\mathbb{L} \subset \mathbb{R}^N$ of dimension $k < N$ contains at most $k + 1$ elements of the set $\{\pm \mathbf{x}_1, \pm \mathbf{x}_2, \dots, \pm \mathbf{x}_p\}$, excluding antipodal pairs of points (that is, points differing only by a sign flip). We note that the data in the example in the previous paragraph are not in general position. If the X data are drawn from a continuous probability distribution, then with probability one the data are in general position and hence the lasso solutions will be unique. As a result, non-uniqueness of the lasso solutions can only occur with discrete-valued data, such as those arising from dummy-value coding of categorical predictors. These results have appeared in various forms in the literature, with a summary given by Tibshirani (2013)

Proposition 2.3.3 — Lasso 惩罚的变量选择一致性. For MSE consistency, if β^* and $\hat{\beta}$ are the true and lasso estimated parameters, it can be shown that as $p, n \rightarrow \infty$

$$\|X(\hat{\beta} - \beta^*)\|_2^2 / N \leq C \cdot \|\beta^*\|_1 \sqrt{\log(p)/N}$$

with high probability (Bühlmann and van de Geer 2011, Chapter 6). Thus if $\|\beta^*\|_1 = o(\sqrt{N/\log(p)})$ then the lasso is consistent for prediction. This means that the true parameter vector must be sparse relative to the ratio $N/\log(p)$. The result only assumes that the design \mathbf{X} is fixed and has no other conditions on \mathbf{X} . Consistent recovery of the nonzero support set requires more stringent assumptions on the level of cross-correlation between the predictors inside and outside of the support set. Details are given in Chapter 11.

Orthonormal design

It is instructive to consider the orthonormal design where $p = n$ and the design matrix satisfies $n^{-1}\mathbf{X}^T\mathbf{X} = I_{p \times p}$. In this case, the Lasso estimator is the soft-threshold estimator

$$\hat{\beta}_j(\lambda) = \text{sign}(Z_j) (|Z_j| - \lambda/2)_+, Z_j = (\mathbf{X}^T\mathbf{Y})_j / n (j = 1, \dots, p = n) \quad (2.60)$$

Thus, the estimator can be written as

$$\hat{\beta}_j(\lambda) = g_{\text{soft}, \lambda/2}(Z_j)$$

where $g_{\text{soft}, \lambda}(z) = \text{sign}(z)(|z| - \lambda)_+$, is the soft-threshold function. There, we also show for comparison the hard-threshold and the adaptive Lasso estimator for β_j defined by $\hat{\beta}_{\text{hard}, j}(\lambda) = g_{\text{hard}, \lambda/2}(Z_j)$, $g_{\text{hard}, \lambda}(z) = z\mathbf{1}(|z| \leq \lambda)$

$$\hat{\beta}_{\text{adapt}, j}(\lambda) = g_{\text{adapt}, \lambda/2}(Z_j), \quad g_{\text{adapt}, \lambda}(z) = z(1 - \lambda/|z|)_+ = \text{sign}(z)(|z| - \lambda/|z|)_+$$

Prediction

We refer to prediction whenever the goal is estimation of the regression function $\mathbb{E}[Y | X = x] = \sum_{j=1}^p \beta_j x^{(j)}$ in model (3.1). This is also the relevant quantity for predicting a new response.

$$\hat{\beta}(\lambda) = \arg \min_{\beta} (\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 / n + \lambda \|\beta\|_1) \quad (2.61)$$

Lemma 2.9 Denote the gradient of $n^{-1}\|\mathbf{Y} - \mathbf{X}\beta\|_2^2$ by $G(\beta) = -2\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\beta)/n$. Then a necessary and sufficient condition for $\hat{\beta}$ to be a solution of (2.61) is:

$$\begin{aligned} G_j(\hat{\beta}) &= -\text{sign}(\hat{\beta}_j) \lambda \text{ if } \hat{\beta}_j \neq 0 \\ |G_j(\hat{\beta})| &\leq \lambda \text{ if } \hat{\beta}_j = 0 \end{aligned}$$

Moreover, if the solution of (2.61) is not unique (e.g. if $p > n$) and $G_j(\hat{\beta}) < \lambda$ for some solution $\hat{\beta}$, then $\hat{\beta}_j = 0$ for all solutions of (2.61) ■

Proof. For the first statements regarding a necessary and sufficient characterization of the solution, we invoke subdifferential calculus (Bertsekas, 1995), see also Problem 4.2 in Chapter 4. Denote the criterion function by

$$Q_\lambda(\beta) = \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda\|\beta\|_1$$

For a minimizer $\hat{\beta}(\lambda)$ of $Q_\lambda(\cdot)$ it is necessary and sufficient that the subdifferential at $\hat{\beta}(\lambda)$ is zero. If the j th component $\hat{\beta}_j(\lambda) \neq 0$, this means that the ordinary first derivative at $\hat{\beta}(\lambda)$ has to be zero:

$$\left. \frac{\partial Q_\lambda(\beta)}{\partial \beta_j} \right|_{\beta=\hat{\beta}(\lambda)} = -2\mathbf{X}_j^T(\mathbf{Y} - \mathbf{X}\hat{\beta}(\lambda))/n + \lambda \text{sign}(\hat{\beta}_j(\lambda)) = 0$$

where \mathbf{X}_j is the $n \times 1$ vector $(X_1^{(j)}, \dots, X_n^{(j)})^T$. Of course, this is equivalent to

$$G_j(\hat{\beta}(\lambda)) = -2\mathbf{X}_j^T(\mathbf{Y} - \mathbf{X}\hat{\beta}(\lambda))/n = -\lambda \text{sign}(\hat{\beta}_j(\lambda)) \text{ if } \hat{\beta}_j(\lambda) \neq 0$$

On the other hand, if $\hat{\beta}_j(\lambda) = 0$, the subdifferential at $\hat{\beta}(\lambda)$ has to include the zero element (Bertsekas, 1995). That is: if $\hat{\beta}_j(\lambda) = 0 : G_j(\hat{\beta}(\lambda)) + \lambda e = 0$ for some $e \in [-1, 1]$ But this is equivalent to

$$|G_j(\hat{\beta}(\lambda))| \leq \lambda \text{ if } \hat{\beta}_j(\lambda) = 0$$

And this is the second statement in the characterization of the solution of $\hat{\beta}(\lambda)$. Regarding uniqueness of the zeroes among different solutions we argue as follows. Assume that there exist two solutions $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$ such that for a component j we have $\hat{\beta}_j^{(1)} = 0$ with $|G_j(\hat{\beta}^{(1)})| < \lambda$ but $\hat{\beta}_j^{(2)} \neq 0$. Because the set of all solutions is convex,

$$\hat{\beta}_\rho = (1 - \rho)\hat{\beta}^{(1)} + \rho\hat{\beta}^{(2)}$$

is also a minimizer for all $\rho \in [0, 1]$. By assumption and for $0 < \rho < 1$, $\hat{\beta}_{\rho,j} \neq 0$ and hence, by the first statement from the KKT conditions, $|G_j(\hat{\beta}_\rho)| = \lambda$ for all $\rho \in (0, 1)$. Hence, it holds for $g(\rho) = |G_j(\hat{\beta}_\rho)|$ that $g(0) < \lambda$ and $g(\rho) = \lambda$ for all $\rho \in (0, 1)$. But this is a contradiction to the fact that $g(\cdot)$ is continuous. Hence, a non-active (i.e. zero) component j with $|G_j(\hat{\beta})| < \lambda$ can not be active (i.e. non-zero) in any other solution. ■

Ideally, we would like to infer the active set S_0 from data. We will explain in Section 2.6 that the Lasso as used in (2.10) requires fairly strong conditions on the design matrix \mathbf{X} (Theorem 7.1 in Chapter 7 gives the precise statement.)

A less ambitious but still relevant goal in practice is to find at least the covariates whose corresponding absolute values of the regression coefficients $|\beta_j|$ are substantial. More formally, for some $C > 0$, define the substantial (relevant) covariates as

$$S_0^{\text{relevant}(C)} = \left\{ j; |\beta_j^0| \geq C, j = 1, \dots, p \right\}$$

Using the result in (2.9), one can show (Problem 2.2) that for any fixed $0 < C < \infty$: $\mathbf{P} \left[\hat{S}(\lambda) \supset S_0^{\text{relevant}(C)} \right] \rightarrow 1 (n \rightarrow \infty)$ This result can be generalized as follows. Assume that

Leng et al(2006) 以纯代数运算的方式证明了在一般线性模型中, 当自变量矩阵为正交阵时且和 p 固定时, Lasso 选择出正确模型的概率严格小于 1, 且与 n 无关, 证明的大体思路及主要计算是: 假设正确的系数向量为 $\beta = (\beta_1, \dots, \beta_q, 0, \dots, 0)^T$, 其中有 q 个非零系数和 $p - q > 0$ 个 0 系数, 记 OLS 得到的估计量为 $\hat{\beta}^0$, 则将两向量之差记为 $(\delta_1, \dots, \delta_p)^T = \hat{\beta}^0 - \beta$, 且不失一般性地假设 $|\delta_{q+1}| > |\delta_{q+2}| > \dots > |\delta_p|$. 以下考虑在 $\phi = (\delta_1, \dots, \delta_p)^T$: for $j = 1, \dots, q, \delta_j > -\beta_j, \sum_{i=1}^q \delta_i < 0$ 区域中, Lasso 选出正确变量的概率.

Proof. 如果 $\hat{\beta}^0$ 不能满足

$$|\hat{\beta}_j^0| > |\hat{\beta}_k^0| \text{ for } j \in 1, \dots, q, k \in q+1, \dots, p$$

则显然 Lasso 惩罚未能选择出正确的模型, 因此以下讨论的前提为上式成立. Lasso 的解为 $\hat{\beta}_j^L = (\hat{\beta}_j^0 - \gamma)^+$, $j = 1, 2, \dots, p$, 于是 Lasso 选择出正确的模型等价于

$$\min |\hat{\beta}_1^0|, \dots, |\hat{\beta}_q^0| > \gamma \geq |\hat{\beta}_{q+1}^0|$$

记 $\tau(\gamma) = (\hat{\beta}_1^0 - \gamma)x_1 + \dots + (\hat{\beta}_q^0 - \gamma)x_q$, 则 $\tau(\gamma)$ 的残差平方和 $SE(\gamma) = \sum_{i=1}^q (\delta_i - y_i)^2$. 令 $\gamma_1 = |\hat{\beta}_{q+2}^0|$, 则

$$\tau(\gamma_1) = (\hat{\beta}_1^0 - |\hat{\beta}_{q+2}^0|)x_1 + \dots + (\hat{\beta}_q^0 - |\hat{\beta}_{q+2}^0|)x_q + \text{sign}(\hat{\beta}_{q+1}^0) (|\hat{\beta}_{q+1}^0| - |\hat{\beta}_{q+2}^0|)x_{q+2}$$

则 $\tau(\gamma_1)$ 的残差平方和 $SE(\gamma_1) = \sum_{i=1}^q (\delta_i - |\delta_{q+2}|)^2 + (|\delta_{q+1}| - |\delta_{q+2}|)^2$. 下证 $SE(\gamma) > SE(\gamma_1)$:

$$\begin{aligned} SE(\gamma) &= \sum_{i=1}^q (\delta_i - y_i)^2 = \sum_{i=1}^q (\delta_i - |\delta_{q+2}| + |\delta_{q+2}| + y_i)^2 \\ &= \sum_{i=1}^q (\delta_i - |\delta_{q+2}|)^2 + q(\gamma - |\delta_{q+2}|)^2 + 2(\gamma - |\delta_{p_1+2}|) \sum_{i=1}^{p_1} (|\delta_{p_1+2}| - \delta_i) \\ &= SE(\gamma_1) - (|\delta_{q+1}| - |\delta_{q+2}|)^2 + p_1(\gamma - |\delta_{p_1+2}|)^2 + 2(\gamma - |\delta_{q+2}|) \sum_{i=1}^q (|\delta_{q+2}| - \delta_i) \end{aligned}$$

因为 $\gamma \geq |\delta_{q+1}|$ 所以 $p_1(\gamma - |\delta_{q+2}|)^2 - (|\delta_{q+1}| - |\delta_{q+2}|)^2 \geq (q)(\gamma - |\delta_{q+2}|)^2$. 从而 $SE(\gamma) \geq SE(\gamma_1) + (q)(\gamma - |\delta_{q+2}|)^2 + 2(\gamma - |\delta_{q+2}|) \sum_{i=1}^q (|\delta_{q+2}| - \delta_i)$. 在 $\sum_{i=1}^q \delta_i < 0$ 的前提下, $(q-1)(\gamma - |\delta_{p_1+2}|) + 2 \sum_{i=1}^q (|\delta_{q+2}| - \delta_i) = (q+1)|\delta_{q+2}| + (q-1)\gamma - 2 \sum_{i=1}^q \delta_i > 0$, 因此, 当 $\delta \in \phi$ 时, 有 $SE(\gamma) > SE(\gamma_1)$, 这与 γ 是通过最小化预测误差得所得到的估计量这个假设前提相矛盾. 又因为 $(\delta_1, \dots, \delta_p) \sim N(0, I_d)$, 所以 $\Pr(\delta \in \phi) > \Pr(\delta \in (\delta : 0 > \delta_j > -\beta_j, j = 1, \dots, q)) = C$, 其中 C 是一个与 n 无关, 且严格小于 1 的常数, 证毕. ■

不可表示条件

同年,Zhao and Yu(2006) 针对一般线性模型提出了不可表示条件. 为提高可阅读性, 在本小节以下部分, 在符号中加入样本量 n , 以表示 n 的影响. 以下分别讨论常规数据的高维数据情况下的变量选择的一致性情况. 记样本量为 X_n , 不失一般性, 假设 $\beta^n = (\beta_1^n, \dots, \beta_q^n, \beta_{q+1}^n, \dots, \beta_p^n)^T$ 的前 q 个分量次非 0, 后 $p - q$ 个分量为 0, $X_n(1)$ 和 $X_n(2)$ 分别是 X_n 的前 q 列和后 $p - q$ 列, 记 $C^n = \frac{1}{n} X_n^T X_n$, $C_{11}^n = \frac{1}{n} X_n(1)^T X_n(1)$, $C_{22}^n = \frac{1}{n} X_n(2)^T X_n(2)$, $C_{12}^n = \frac{1}{n} X_n(1)^T X_n(2)$, $C_{21}^n = \frac{1}{n} X_n(2)^T X_n(1)$, 则 $C^n = \begin{bmatrix} C_{11}^n & C_{12}^n \\ C_{21}^n & C_{22}^n \end{bmatrix}$

在 C_{11}^n 可逆的前提下,Zhao 和 Yu(2006) 提出了强不可表示条件 (Strong Irrepresentable Condition): 存在 N , 使得当 $n > N$ 时, 对于正向量 η , 有

$$\left| C_{21}^n (C_{11}^n)^{-1} \text{sign}(\beta_{(1)}^n) \right| \leq 1 - \eta$$

和弱不可表示条件 (Weak Irrepresentable Condition): 存在 N , 使得当 $n > N$ 时, 有

$$\left| C_{21}^n (C_{11}^n)^{-1} \text{sign}(\beta_{(1)}^n) \right| \leq 1$$

上述两个条件对于向量中的每一个分量均是成立的. 两个不可表示条件的不同在于, 当 $|C_{21}^n (C_{11}^n)^{-1} \text{sign}(\beta_{(1)}^n)|$ 从左边向 1 趋近时, 弱不可表示条件恒成立, 但是强不可表示条件在极限处不成立.

常规数据情况下, 对于固定的 p :

1. $C^n \rightarrow C$ as $n \rightarrow \infty$, 其中 C 是正定矩阵;
2. $\frac{1}{n} \max_{1 \leq i \leq n} \left((x_i^n)^T x_i^n \right) \rightarrow 0$, as $n \rightarrow \infty$
3. 强不可表示条件

当以上条件成立时, 则 Lasso 惩罚具有强符号一致的性质, 从而确保变量选择的一致性, 对于 $\forall \lambda_n$ 满足

1. $\lambda_n / n \rightarrow 0$ 和
2. $0 \leq \forall c \leq 1, \lambda_n / n^{\frac{1+c}{2}} \rightarrow \infty$,

便可以提前确定 λ_n . 在上述条件中, 若 x_i 是二阶矩有限的 i.i.d 随机变量, 则有

1. $C = E \left((x_i^n)^T x_i^n \right), \frac{1}{n} X_n^T X_n \rightarrow C$,
2. $\max_{1 \leq i \leq n} (x_i^n)^T x_i^n = o_p(n)$,

故第一项条件和第二项条件较为宽松, 但强不可表示条件的成立较为严格.

另外, 前一小节中 Leng et al(2006) 的结论的前提假设是 X 为正交阵, 很显然若 X 为正交阵则 C^n 不满足不可表示条件, 因此 Leng et al(2006) 的结论是包含在 Zhao 和 Yu(2006) 的结论中. 高维数据情况下, 当 q 和 p 随着样本量 n 的增加而变化时, Zhao 和 Yu(2006) 提出了更为严格的前提假设: 假设存在 $0 \leq c_1 \leq c_2 \leq 1, M_1, M_2, M_3 > 0$ 和整数 $k > 0$, 使得

1. $\frac{1}{n} (X_i^n)^T (X_i^n) \leq M_1, \forall i$
2. $\alpha^T C_{11}^n \alpha \geq M_2, \forall \|\alpha\|_2^2 = 1$
3. $q = O(n^{c_1})$
4. $n^{\frac{1-c_2}{n}} \min_{i=1, \dots, q} |\beta_i^n| \geq M_3$

$$5. E(\epsilon_i^n)^{2k} < \infty$$

6. 强不可表示条件成立

其中, k 是大于 0 的整数。当以上条件满足时, 对于 $p = o(n^{(c_2 - c_1)k})$, Lasso 惩罚具有强符号一致的性质, 此时对于 $\forall \lambda_n$, 只要其满足 $\lambda_n / \sqrt{n} = o(n^{\frac{c_2 - c_1}{2}})$ 和 $\frac{1}{p} \left(\frac{\lambda_n}{\sqrt{n}} \right)^{2k} \rightarrow \infty$ 便可提前确定。在上述条件中, 第一项条件是非常宽松的, 只要对设计矩阵 X 进行标准化; 第二项条件是要求相关矩阵的特征值有最小值, 而实际情况中 C_{11}^n 很有可能未知, 但是当全部设计矩阵的特征值有最小值且 $q/n \rightarrow \rho < 1$, 该条件通常也成立; 第三项条件约束了数据的高维情况。

Zhao 和 Yu(2006) 采用模拟的方法测算出强不可表示条件成立的比例, 由上表可以看出: 随着模型稀疏程度的减少, 强不可表示条件成立的比例在逐渐减少, 并且该比例随着自变量个数的增大而急剧降低, 因此不可表示条件在超高维数据中的应用受到很大限制。并且不可表示条件需要知道正确的非 0 变量个数及其对应的系数符号等的先验信息。

特征值条件

Yang and Yang(2015) 提出了特征值条件 (Eigenvalue Condition), 其主要贡献在于:

1. 特征值条件较不可表示条件有所放松;
2. 分别讨论了高维数据和超高维数据下 Lasso 惩罚变量选择一致性的条件。

特征值条件如下: 记 $v_{\min} C_{11}^n$ 为 C_{11}^n 的最小特征值, 对于 $\forall i = 1, \dots, q, \forall j = q + 1, \dots, p$, 存在 $\delta \in (0, 1)$, 有:

$$\frac{1}{n} X_{i,n}^T X_{j,n} < (1 - \delta) v_{\min} C_{11}^n$$

在高维数据的情况下, 若数据满足特征值条件, 则 Lasso 惩罚具有变量选择一致性, 无需额外对回归方程中的噪音部分进行约束; 而在超高维数据的情况下, 作者将噪音部分分为服从高斯分布和不服从高斯分布两种情况进行讨论, 在满足一定的条件, Lasso 惩罚具有变量选择一致性。但是, Lasso 惩罚的不具有变量选择一致性并不意味着 Lasso 惩罚得到的模型估计一定不好。

Meinshausen and Yu(2009) 定义了估计量的 L_2 一致性: 如果某估计量 $\hat{\beta}$ 满足

$$\|\hat{\beta} - \beta\|_2 \rightarrow 0, n \rightarrow \infty$$

则称 $\hat{\beta}$ 是 L_2 一致的。当估计量是 L_2 一致时, 虽然重要的变量和不重要的变量都被选入模型, 但是不重要变量的系数一定非常小, 从而整体估计的效果仍是不错的。Meinshausen and Yu(2009) 提出了非连贯设计条件 (incoherent design condition):

$$\liminf_{n \rightarrow \infty} \frac{e_n \phi_{\min}(e_n^2 s_n)}{\phi_{\max}(s_n + \min n, p_n)} \geq 18$$

其中, $\phi_{\min}(m) = \min_{\beta: \|\beta\|_0 \leq [m]} \frac{\beta^T C \beta}{\beta^T \beta}$, 同理 $\phi_{\max}(m)$ 。Lasso 估计量是具有 L_2 一致性的, 因此可以保证 Lasso 的预测精度; 换言之, 对于 Lasso 而言, 整体估计值的一致性比变量选择一致性更容易达到。

Meinshausen and Bühlmann(2006) 和 Zhao and Yu(2006) 证明了在高维数据的情况下,Lasso 惩罚的变量选择一致性,同时也证明了,若某调节参数保证了变量选择的一致性,则无法保证预测精度的最优化(即预测误差的最小化),也就是说,变量选择正确但系数估计量存在偏差,这意味着在高维数据的情况下,Lasso 惩罚不具有神谕性.

Proposition 2.3.4 — Lasso 估计量的相合性. Knight 和 Fu(2000) 证明了,在自变量个数 p 固定的情况下,Lasso 估计量具有相合性,其结论如下: 记 $\Sigma^\wedge = X^T X/n$, 假设 $\Sigma^\wedge \rightarrow \Sigma$, 其中 Σ 为可逆阵, $\frac{1}{n} \max_{1 \leq i \leq n} X_i^T X_i \rightarrow 0$; 考虑到样本量 n 的逐渐增大, 因此惩罚函数中需要消除量纲的影响, 故惩罚函数及 Lasso 估计量的定义如下:

$$\hat{\beta} = \arg \min_{\beta} \|Y - X\beta\|_2^2 / (n) + \lambda_n \|\beta\|_1, \lambda_n > 0$$

λ_n 表示其值会随着样本量 n 的变化而变化, 但自变量个数 p 和真实的 β 是不变的, 即不会随着样本量 n 的变化而变化. 若 $\lambda_n \rightarrow \lambda_0$, 则 $\hat{\beta} \rightarrow_p \arg \min l(\beta)$, 其中 $l(\beta) = (\beta - \beta_0)^T \Sigma (\beta - \beta_0) + \lambda_0 \|\beta\|_1$. 因此, 若 $\lambda_0 = 0$, 则 $\arg \min l(\beta) = \beta$, 因此 Lasso 的估计量是一致估计量.

Proposition 2.3.5 — Lasso 估计量的渐近分布. Knight and Fu(2000) 推导出当 p 固定时, 在一定条件下,Lasso 估计量的渐近分布, 其结论如下:

除 $\sqrt{n}\lambda_n \rightarrow \lambda_0 \geq 0$ 外, 其余前提假设同上节, 则 $\sqrt{n}(\hat{\beta} - \beta) \rightarrow^d \arg \min V$. 其中

$$V(u) = -2u^T W + u^T \Sigma u + \lambda_0 \sum_{j=1}^p u_j \text{sign}(\beta_j) I(\beta_j \neq 0) + |u_j| I(\beta_j = 0)$$

$$W \sim N(0, \sigma^2 C). \text{ 因此, 当 } \lambda_0 = 0 \text{ 时, } \arg \min V = W \Sigma^{-1} \sim N(0, \sigma^2 \Sigma^{-1})$$

从上述两个极限方程中可以看出, 当 $\lambda \neq 0$ 时 (实际情况中更多的是非 0), Lasso 估计量的极限值和极限分布均没有封闭解, 因此往往无法用 Lasso 估计量的渐进分布构造的置信区间和检验统计量.

有个附录要看



3. Multiple and Nonparametric Regression

Basic introduction to linear model

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon \quad (3.1)$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

The matrix $\mathbf{X}_{n \times p}$ is known as the **design matrix**. The $\text{RSS}(\boldsymbol{\beta})$ can be written as

$$\text{RSS}(\boldsymbol{\beta}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

Differentiating $\text{RSS}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ and setting the gradient vector to zero, we obtain the normal equations

$$\mathbf{X}^T \mathbf{Y} = \mathbf{X}^T \mathbf{X} \boldsymbol{\beta}$$

Here we assume that $p < n$ and \mathbf{X} has rank p . Hence $\mathbf{X}^T \mathbf{X}$ is invertible and the normal equations yield the least-squares estimator of $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \quad (3.2)$$

In this chapter $\mathbf{X}^T \mathbf{X}$ is assumed to be invertible unless specifically mentioned otherwise. The fitted Y value is

$$\hat{Y} = \mathbf{x} \hat{\boldsymbol{\beta}} = \mathbf{x} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

and the regression residual is

$$\hat{\mathbf{r}} = \mathbf{Y} - \hat{\mathbf{Y}} = \left(\mathbf{I} - \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right) \mathbf{Y}$$

Theorem 3.0.1 Define $\mathbf{P} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$. Then we have

$$\begin{aligned} \mathbf{P} \mathbf{X}_j &= \mathbf{X}_j, \quad j = 1, 2, \dots, p \\ \mathbf{P}^2 &= \mathbf{P} \quad \text{or} \quad \mathbf{P} (\mathbf{I}_n - \mathbf{P}) = \mathbf{0} \end{aligned}$$

namely \mathbf{P} is a **projection matrix** onto the space spanned by the columns of \mathbf{X} .

Proof. It follows from the direct calculation that

$$\mathbf{P} \mathbf{X} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} = \mathbf{X}$$

Taking the j column of the above equality, we obtain the first results. Similarly,

$$\mathbf{P} \mathbf{P} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T = \mathbf{P}$$

This completes the proof. ■

By Theorem 3.0.1 we can write

$$\hat{\mathbf{Y}} = \mathbf{P} \mathbf{Y}, \quad \hat{\mathbf{r}} = (\mathbf{I}_n - \mathbf{P}) \mathbf{Y}$$

and we see two simple identities:

$$\mathbf{P} \hat{\mathbf{Y}} = \hat{\mathbf{Y}}, \quad \hat{\mathbf{Y}}^T \hat{\mathbf{r}} = 0$$

This reveals an interesting geometric interpretation of the method of least squares: the least-squares fit amounts to projecting the response vector onto the linear space spanned by the covariates. Geometric view of least-squares: The fitted value is the blue arrow, which is the projection of \mathbf{Y} on the plane spanned by X_1 and X_2

3.0.1 The Gauss-Markov Theorem

We assume the linear regression model (3.1) with

1. exogeneity: $E(\varepsilon | \mathbf{X}) = 0$
2. homoscedasticity: $\text{Var}(\varepsilon | \mathbf{X}) = \sigma^2$

Theorem 3.0.2 Under model (3.1) with exogenous and homoscedastic error, it follows that

1. (unbiasedness) $E(\hat{\boldsymbol{\beta}} | \mathbf{X}) = \boldsymbol{\beta}$
2. (conditional standard errors) $\text{Var}(\hat{\boldsymbol{\beta}} | \mathbf{X}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$
3. (BLUE) The least-squares estimator $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE).

That is, for any given vector \mathbf{a} , $\mathbf{a}^T \hat{\boldsymbol{\beta}}$ is a linear unbiased estimator of the parameter $\theta = \mathbf{a}^T \boldsymbol{\beta}$. Further, for any linear unbiased estimator $\mathbf{b}^T \mathbf{Y}$ of θ , its variance is at least as large as that of $\mathbf{a}^T \hat{\boldsymbol{\beta}}$

Proof. The first property follows directly from $E(\mathbf{Y} | \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$ and

$$E(\hat{\boldsymbol{\beta}} | \mathbf{X}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\beta}$$

To prove the second property, note that for any linear combination $\mathbf{A}\mathbf{Y}$, its variance-covariance matrix is given by

$$\text{Var}(\mathbf{A}\mathbf{Y} | \mathbf{X}) = \mathbf{A} \text{Var}(\mathbf{Y} | \mathbf{X}) \mathbf{A}^T = \sigma^2 \mathbf{A} \mathbf{A}^T$$

Applying this formula to the least-squares estimator with $\mathbf{A} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ we obtain the property (ii).

To prove property (iii), we first notice that $\mathbf{a}^T \hat{\boldsymbol{\beta}}$ is an unbiased estimator of the parameter $\theta = \mathbf{a}^T \boldsymbol{\beta}$, with the variance

$$\text{Var}(\mathbf{a}^T \hat{\boldsymbol{\beta}} | \mathbf{X}) = \mathbf{a}^T \text{Var}(\hat{\boldsymbol{\beta}} | \mathbf{X}) \mathbf{a} = \sigma^2 \mathbf{a}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{a} \quad (3.3)$$

Now, consider any linear unbiased estimator, $\mathbf{b}^T \mathbf{Y}$, of the parameter θ . The unbiasedness requires that

$$\mathbf{b}^T \mathbf{X}\boldsymbol{\beta} = \mathbf{a}^T \boldsymbol{\beta}$$

namely $\mathbf{X}^T \mathbf{b} = \mathbf{a}$. The variance of this linear estimator is

$$\sigma^2 \mathbf{b}^T \mathbf{b}$$

To prove (iii) we need only to show that

$$\mathbf{a}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{a} \leq \mathbf{b}^T \mathbf{b}$$

Note that

$$(\mathbf{X}^T \mathbf{X})^{-1/2} \mathbf{X}^T \mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1/2} \mathbf{a}$$

Hence, by computing their norms, we have

$$\mathbf{a}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{a} = \mathbf{b}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{b} = \mathbf{b}^T \mathbf{P} \mathbf{b}$$

Note that $\mathbf{P} = \mathbf{P}^2$ which means that the eigenvalues of \mathbf{P} are either 1 or 0 and hence $\mathbf{I}_n - \mathbf{P}$ is semi-positive matrix. Hence,

$$\mathbf{b}^T (\mathbf{I}_n - \mathbf{P}) \mathbf{b} \geq 0$$

or equivalently $\mathbf{b}^T \mathbf{b} \geq \mathbf{b}^T \mathbf{P} \mathbf{b}$ ■

Property (ii) of Theorem 3.0.2 gives the variance-covariance matrix of the least-squares estimate. In particular, the conditional standard error of $\hat{\beta}_i$ is simply $\sigma a_{ii}^{1/2}$ and the covariance between $\hat{\beta}_i$ and $\hat{\beta}_j$ is $\sigma^2 a_{ij}$, where a_{ij} is the (i, j) -th element of matrix $(\mathbf{X}^T \mathbf{X})^{-1}$.

In many applications σ^2 is often an unknown parameter of the model in addition to the regression coefficient vector β . In order to use the variance-covariance formula, we first need to find a good estimate of σ^2 . Given the least-squares estimate of β , RSS can be written as

$$\text{RSS} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = (\mathbf{Y} - \hat{\mathbf{Y}})^T (\mathbf{Y} - \hat{\mathbf{Y}}) \quad (3.4)$$

Define

$$\hat{\sigma}^2 = \text{RSS} / (n - p)$$

This can be shown in Theorem 3.0.3 that $\hat{\sigma}^2$ is an unbiased estimator of σ^2 .

Theorem 3.0.3 Under the linear model (3.1) with homoscedastic error, it follows that

$$E(\hat{\sigma}^2 | \mathbf{X}) = \sigma^2$$

Proof. First by Theorem 3.0.1 we have

$$\text{RSS} = \|(\mathbf{I}_n - \mathbf{P}) \mathbf{Y}\|^2 = \|(\mathbf{I}_n - \mathbf{P})(\mathbf{Y} - \mathbf{X}\beta)\|^2 = \varepsilon^T (\mathbf{I}_n - \mathbf{P}) \varepsilon$$

(($\mathbf{I}_n - \mathbf{P}) \cdot \mathbf{X}\beta = 0$, $(\mathbf{I}_n - \mathbf{P})^2 = (\mathbf{I}_n - \mathbf{P})$) Let $\text{tr}(\mathbf{A})$ be the trace of the matrix \mathbf{A} . Using the property that $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$, we have

$$\text{RSS} = \text{tr} \left\{ (\mathbf{I}_n - \mathbf{P}) \varepsilon \varepsilon^T \right\}$$

Hence, according to homoscedasticity,

$$E(\text{RSS} | \mathbf{X}) = \sigma^2 \text{tr}(\mathbf{I}_n - \mathbf{P})$$

Because the eigenvalues of \mathbf{P} are either 1 or 0, its trace is equal to its rank which is p under the assumption that $\mathbf{X}^T \mathbf{X}$ is invertible. Thus,

$$E(\hat{\sigma}^2 | \mathbf{X}) = \sigma^2 (n - p) / (n - p) = \sigma^2$$

This completes the proof. ■

Statistical Tests

After fitting the regression model, we often need to perform some tests on the model parameters. For example, we may be interested in testing whether a particular regression coefficient should be zero, or whether several regression coefficients should be zero at the same time, which is equivalent to asking whether these variables are important in

presence of other covariates. To facilitate the discussion, we focus on the fixed design case where \mathbf{X} is fixed. This is essentially the same as the random design case but conditioning upon the given realization \mathbf{X} .

We assume a homoscedastic model (3.1) with normal error. That is, ε is a Gaussian random variable with zero mean and variance σ^2 , written as $\varepsilon \sim N(0, \sigma^2)$. Note that

$$\hat{\beta} = \beta + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \varepsilon \quad (3.5)$$

Then it is easy to see that

$$\hat{\beta} \sim N\left(\beta, (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2\right) \quad (3.6)$$

If we look at each $\hat{\beta}_j$ marginally, then $\hat{\beta}_j \sim N(\beta_j, v_j \sigma^2)$ where v_j is the j th diagonal element of $(\mathbf{X}^T \mathbf{X})^{-1}$. In addition,

$$(n - p) \hat{\sigma}^2 \sim \sigma^2 \chi_{n-p}^2 \quad (3.7)$$

and $\hat{\sigma}^2$ is independent of $\hat{\beta}$. The latter can easily be shown as follow.

Test on single β_j

By (3.4), $\hat{\sigma}^2$ depends on \mathbf{Y} through $\mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I}_n - \mathbf{P}) \varepsilon$ whereas $\hat{\beta}$ depends on \mathbf{Y} through (3.5) or $\mathbf{X}^T \varepsilon$. Note that both $(\mathbf{I}_n - \mathbf{P}) \varepsilon$ and $\mathbf{X}^T \varepsilon$ are jointly normal because they are linear transforms of normally distributed random variables, and therefore their independence is equivalent to their uncorrelatedness. This can easily be checked by computing their covariance

$$\mathbb{E}[(\mathbf{I}_n - \mathbf{P}) \varepsilon (\mathbf{X}^T \varepsilon)^T] = \mathbb{E}[(\mathbf{I}_n - \mathbf{P}) \varepsilon \varepsilon^T \mathbf{X}] = \sigma^2 (\mathbf{I}_n - \mathbf{P}) \mathbf{X} = 0$$

So $\hat{\sigma}^2$ and $\hat{\beta}$ is independence. If we want to test the hypothesis that $\beta_j = 0$, we can use the following t test statistic

$$t_j = \frac{\hat{\beta}_j}{\sqrt{v_j \hat{\sigma}^2}} \quad (3.8)$$

which follows a t -distribution with $n - p$ degrees of freedom under the null hypothesis $H_0 : \beta_j = 0$. A level α test rejects the null hypothesis if $|t_j| > t_{n-p, 1-\alpha/2}$, where $t_{n-p, 1-\alpha/2}$ denotes the $100(1 - \alpha/2)$ percentile of the t -distribution with $n - p$ degrees of freedom.

Test on full model and reduced model

In many applications the null hypothesis is that a subset of the covariates have zero regression coefficients. That is, this subset of covariates can be deleted from the regression model: they are unrelated to the response variable given the remaining variables. Under such a null hypothesis, we can reduce the model to a smaller model. Suppose that the

reduced model has p_0 many regression coefficients. Let RSS and RSS_0 be the residual sum-of-squares based on the least-squares fit of the full model and the reduced smaller model, respectively. If the null hypothesis is true, then these two quantities should be similar: The RSS reduction by using the full model is small, in relative term. This leads to the F -statistic:

$$F = \frac{(RSS_0 - RSS) / (p - p_0)}{RSS / (n - p)} \quad (3.9)$$

Under the null hypothesis that the reduced model is correct, $F \sim F_{p-p_0, n-p}$. The normal error assumption can be relaxed if the sample size n is large.

Proof. Proof of 3.6 First, we know that $(\mathbf{X}^T \mathbf{X})^{\frac{1}{2}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) / \sigma$ always has zero mean and an identity variance-covariance matrix. On the other hand, (3.5) gives us

$$(\mathbf{X}^T \mathbf{X})^{\frac{1}{2}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) / \sigma = (\mathbf{X}^T \mathbf{X})^{-\frac{1}{2}} \mathbf{X}^T \boldsymbol{\varepsilon} / \sigma$$

Observe that $(\mathbf{X}^T \mathbf{X})^{-\frac{1}{2}} \mathbf{X}^T \boldsymbol{\varepsilon} / \sigma$ is a linear combination of n i.i.d. random variables $\{\varepsilon_i\}_{i=1}^n$ with zero mean and variance 1. Then the central limit theorem implies that under some regularity conditions,

$$\hat{\boldsymbol{\beta}} \xrightarrow{D} N\left(\boldsymbol{\beta}, (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2\right) \quad (3.10)$$

Consequently, when n is large, the distribution of the t test statistic in (3.8) is approximately $N(0, 1)$, and the distribution of the F test statistic in (3.9) is approximately

$$\chi_{p-p_0}^2 / (p - p_0)$$

■

3.0.2 Weighted Least-Squares

The method of least-squares can be further generalized to handle the situations where errors are heteroscedastic or correlated. In the linear regression model, we would like to keep the assumption $E(\varepsilon | \mathbf{X}) = 0$ which means there is no structure information left in the error term. However, the constant variance assumption $\text{Var}(\varepsilon_i | \mathbf{X}_i) = \sigma^2$ may not likely hold in many applications. For example, if y_i is the average response value of the i th subject in a study in which k_i many repeated measurements have been taken, then it would be more reasonable to assume $\text{Var}(\varepsilon_i | \mathbf{X}_i) = \sigma^2 / k_i$.

Let us consider a modification of model (3.1) as follows

$$Y_i = \sum_{j=1}^p X_{ij} \beta_j + \varepsilon_i; \quad \text{Var}(\varepsilon_i | \mathbf{X}_i) = \sigma^2 v_i \quad (3.11)$$

where v_i s are known positive constants but σ^2 remains unknown. One can still use the ordinary least-squares (OLS) estimator $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$. It is easy to show that the OLS estimator is unbiased but no longer BLUE. In fact, the OLS estimator can be improved by using the weighted least-squares method.

Let $Y_i^* = v_i^{-1/2} Y_i$, $X_{ij}^* = v_i^{-1/2} X_{ij}$, $\varepsilon_i^* = v_i^{-1/2} \varepsilon_i$. Then the new model (3.11) can be written as

$$Y_i^* = \sum_{j=1}^p X_{ij}^* \beta_j + \varepsilon_i^* \quad (3.12)$$

with $\text{Var}(\varepsilon_i^* | \mathbf{X}_i^*) = \sigma^2$. Therefore, the working data $\left\{ (X_{i1}^*, \dots, X_{ip}^*, Y_i^*) \right\}_{i=1}^n$ obey the standard homoscedastic linear regression model. Applying the standard least-squares method to the working data, we have

$$\hat{\beta}^{wls} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n \left(Y_i^* - \sum_{j=1}^p X_{ij}^* \beta_j \right)^2 = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n v_i^{-1} \left(Y_i - \sum_{j=1}^p X_{ij} \beta_j \right)^2$$

It follows easily from Theorem 3.0.2 that the weighted least-squares estimator is the BLUE for β .

In model (3.11) the errors are assumed to be uncorrelated. In general, the method of least-squares can be extended to handle heteroscedastic and correlated errors.

Assume that

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

and the variance-covariance matrix of ε is given

$$\text{Var}(\varepsilon | \mathbf{X}) = \sigma^2 \mathbf{W} \quad (3.13)$$

in which \mathbf{W} is a known positive definite matrix. Let $\mathbf{W}^{-1/2}$ be the square root of \mathbf{W}^{-1} , i.e.,

$$\left(\mathbf{W}^{-1/2} \right)^T \mathbf{W}^{-1/2} = \mathbf{W}^{-1}$$

Then

$$\text{Var} \left(\mathbf{W}^{-1/2} \varepsilon \right) = \sigma^2 \mathbf{I}$$

which are homoscedastic and uncorrelated. Define the working data as follows:

$$\mathbf{Y}^* = \mathbf{W}^{-1/2} \mathbf{Y}, \quad \mathbf{X}^* = \mathbf{W}^{-1/2} \mathbf{X}, \quad \varepsilon^* = \mathbf{W}^{-1/2} \varepsilon$$

Then we have

$$\mathbf{Y}^* = \mathbf{X}^* \beta + \varepsilon^* \quad (3.14)$$

Thus, we can apply the standard least-squares to the working data. First, the residual sum-of-squares (RSS) is

$$\text{RSS}(\beta) = \|\mathbf{Y}^* - \mathbf{X}^*\beta\|^2 = (\mathbf{Y} - \mathbf{X}\beta)^T \mathbf{W}^{-1}(\mathbf{Y} - \mathbf{X}\beta) \quad (3.15)$$

Then the general least-squares estimator is defined by

$$\begin{aligned} \hat{\beta} &= \text{argmin}_{\beta} \text{RSS}(\beta) \\ &= \left(\mathbf{X}^{*T} \mathbf{X}^* \right)^{-1} \mathbf{X}^{*T} \mathbf{Y}^* \\ &= \left(\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{W}^{-1} \mathbf{Y} \end{aligned} \quad (3.16)$$

Again, $\hat{\beta}$ is the BLUE according to Theorem 3.0.2. In practice, it is difficult to know precisely the $n \times n$ covariance matrix \mathbf{W} ; the misspecification of \mathbf{W} in the general least-squares seems hard to avoid.

Let us examine the robustness of the general least-squares estimate. Assume that $\text{Var}(\varepsilon) = \sigma^2 \mathbf{W}_0$, where \mathbf{W}_0 is unknown to us, but we employ the general least-squares method (3.16) with the wrong covariance matrix \mathbf{W} . We can see that the general least-square estimator is still unbiased:

$$\mathbb{E}(\hat{\beta} | \mathbf{X}) = \left(\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{W}^{-1} \mathbf{X} \beta = \beta$$

Furthermore, the variance-covariance matrix is given by

$$\text{Var}(\hat{\beta}) = \left(\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X} \right)^{-1} \left(\mathbf{X}^T \mathbf{W}^{-1} \mathbf{W}_0 \mathbf{W}^{-1} \mathbf{X} \right) \left(\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X} \right)^{-1}$$

which is of order $O(n^{-1})$ under some mild conditions. In other words, using wrong covariance matrix would still give us a root- n consistent estimate. So even when errors are heteroscedastic and correlated, the ordinary least-squares estimate with $\mathbf{W} = \mathbf{I}$ and the weighted least-squares estimate with $\mathbf{W} = \text{diag}(\mathbf{W}_0)$ still give us an unbiased and $n^{-1/2}$ consistent estimator. Of course, we still prefer using a working \mathbf{W} matrix that is identical or close to the true \mathbf{W}_0 .

3.0.3 Box-Cox Transformation

In practice we often take a transformation of the response variable before fitting a linear regression model. The idea is that the transformed response variable can be modeled by the set of covariates via the classical multiple linear regression model. For example, in many engineering problems we expect $Y \propto X_1^{\beta_1} X_2^{\beta_2} \cdots X_p^{\beta_p}$ where all variables are positive. Then a linear model seems proper by taking logarithms: $\log(Y) = \sum_{j=1}^p \beta_j X_j + \varepsilon$. If we assume $\varepsilon \sim N(0, \sigma^2)$, then in the original scale the model is $Y = \left(\prod_{j=1}^p X_j^{\beta_j} \right) \varepsilon^*$ where ε^* is a log-normal random variable: $\log \varepsilon^* \sim N(0, \sigma^2)$.

Box and Cox (1964) advocated the variable transformation idea in linear regression and also proposed a systematic way to estimate the transformation function from data. Their method is now known as Box –Cox transform in the literature. Box and Cox (1964) suggested a parametric family for the transformation function. Let $Y^{(\lambda)}$ denote the transformed response where λ parameterizes the transformation function:

$$Y^{(\lambda)} = \begin{cases} \frac{Y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(Y) & \text{if } \lambda = 0 \end{cases}$$

Box-Cox model assumes that

$$Y^{(\lambda)} = \sum_{j=1}^p X_j \beta_j + \varepsilon$$

where $\varepsilon \sim N(0, \sigma^2)$.

The likelihood function of the Box-Cox model is given by

$$L(\lambda, \beta, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n e^{-\frac{1}{2\sigma^2} \|\mathbf{Y}^{(\lambda)} - \mathbf{X}\beta\|^2} \cdot J(\lambda, \mathbf{Y})$$

where $J(\lambda, \mathbf{Y}) = \prod_{i=1}^n \left| \frac{dy_i^{(\lambda)}}{dy_i} \right| = (\prod_{i=1}^n |y_i|)^{\lambda-1}$. Given λ , the maximum likelihood estimators (MLE) of β and σ^2 are obtained by the ordinary least-squares:

$$\hat{\beta}(\lambda) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}^{(\lambda)}, \quad \hat{\sigma}^2(\lambda) = \frac{1}{n} \left\| \mathbf{Y}^{(\lambda)} - \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}^{(\lambda)} \right\|^2$$

Plugging $\hat{\beta}(\lambda), \hat{\sigma}^2(\lambda)$ into $L(\lambda, \beta, \sigma^2)$ yields a likelihood function of λ

$$\log L(\lambda) = (\lambda - 1) \sum_{i=1}^n \log(|y_i|) - \frac{n}{2} \log \hat{\sigma}^2(\lambda) - \frac{n}{2}$$

Then the MLE of λ is

$$\hat{\lambda}_{mle} = \operatorname{argmax}_{\lambda} \log L(\lambda)$$

and the MLE of β and σ^2 are $\hat{\beta}(\hat{\lambda}_{mle})$ and $\hat{\sigma}^2(\hat{\lambda}_{mle})$, respectively.

3.0.4 Model Building and Basis Expansions

Multiple linear regression can be used to produce nonlinear regression and other very complicated models. The key idea is to create new covariates from the original ones by adopting some transformations. We then fit a multiple linear regression model using augmented covariates.

For simplicity, we first illustrate some useful transformations in the case of $p = 1$, which is closely related to the curve fitting problem in nonparametric regression. In a nonparametric regression model

$$Y = f(X) + \varepsilon$$

we do not assume a specific form of the regression function $f(x)$, but assume only some qualitative aspects of the regression function. Examples include that $f(\cdot)$ is continuous with a certain number of derivatives or that $f(\cdot)$ is convex. The aim is to estimate the function $f(x)$ and its derivatives, without a specific parametric form of $f(\cdot)$.

3.0.5 Polynomial Regression

Without loss of generality, assume X is bounded on $[0,1]$ for simplicity. The Weierstrass approximation theorem states that any continuous $f(x)$ can be uniformly approximated by a polynomial function up to any precision factor. Let us approximate the model by

$$Y = \underbrace{\beta_0 + \beta_1 X + \cdots + \beta_d X^d}_{\approx f(X)} + \varepsilon$$

This polynomial regression is a multiple regression problem by setting $X_0 = 1, X_1 = X, \dots, X_d = X^d$. The design matrix now becomes

$$\mathbf{B}_1 = \begin{pmatrix} 1 & x_1 & \cdots & x_1^d \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_n & \cdots & x_n^d \end{pmatrix}$$

We estimate $f(x)$ by

$$\hat{f}(x) = \hat{\beta}_0 + \sum_{m=1}^d \hat{\beta}_m x^m$$

where $\hat{\beta} = (\mathbf{B}_1^T \mathbf{B}_1)^{-1} \mathbf{B}_1^T \mathbf{Y}$ is the least-squares estimate.

Polynomial functions have derivatives everywhere and are global functions. They are not very flexible in approximating functions with local features such as functions with various degrees of smoothness at different locations. Clearly, cubic polynomial does not fit the motorcycle data very well. Increasing the order of polynomial fits will help reduce the bias issue, but will not solve the lack of fit issue. This is because that the underlying function cannot be economically approximated by a polynomial function. It requires high-order polynomials to reduce approximation biases, but this increases both variances and instability of the fits. This leads to the introduction of spline functions that allow for more flexibility in function approximation.

3.0.6 Spline Regression

Let $\tau_0 < \tau_1 < \cdots < \tau_{K+1}$. A spline function of degree d on $[\tau_0, \tau_{K+1}]$ is a piecewise polynomial function of degree d on intervals $[\tau_j, \tau_{j+1})$ ($j = 0, \dots, K$), with continuous first $d - 1$ derivatives. The points where the spline function might not have continuous d^{th}

derivatives are $\{\tau_j\}_{j=1}^K$, which are called **knots**. Thus, a cubic spline function is a piecewise polynomial function with continuous first two derivatives and the points where the third derivative might not exist are called knots of the cubic spline.

All spline functions of degree d form a linear space. Let us determine its basis functions.

Linear Splines: A continuous function on $[0,1]$ can also be approximated by a piecewise constant or linear function. We wish to use a continuous function to approximate $f(x)$. Since a piecewise constant function is not continuous unless the function is a constant in the entire interval, we use a continuous piecewise linear function to fit $f(x)$. Suppose that we split the interval $[0,1]$ into three regions: $[0, \tau_1], [\tau_1, \tau_2], [\tau_2, 1]$ with given knots τ_1, τ_2 . Denote by $l(x)$ the continuous piecewise linear function. In the first interval $[0, \tau_1]$ we write

$$l(x) = \beta_0 + \beta_1 x, x \in [0, \tau_1]$$

as it is linear. Since $l(x)$ must be continuous at τ_1 , the newly added linear function must have an intercept 0 at point τ_1 . Thus, in $[\tau_1, \tau_2]$ we must have

$$l(x) = \beta_0 + \beta_1 x + \beta_2 (x - \tau_1)_+, x \in [\tau_1, \tau_2]$$

where z_+ equals z if $z > 0$ and zero otherwise. The function is linear in $[\tau_1, \tau_2]$ with slope $\beta_1 + \beta_2$. Likewise, in $[\tau_2, 1]$ we write

$$l(x) = \beta_0 + \beta_1 x + \beta_2 (x - \tau_1)_+ + \beta_3 (x - \tau_2)_+, x \in [\tau_2, 1]$$

The function is now clearly a piecewise linear function with possible different slopes on different intervals. Therefore, the basis functions are

$$B_0(x) = 1, B_1(x) = x, B_2(x) = (x - \tau_1)_+, B_3(x) = (x - \tau_2)_+ \quad (3.17)$$

which are called a **linear spline basis**. We then approximate the nonparametric regression model as

$$Y = \underbrace{\beta_0 B_0(X) + \beta_1 B_1(X) + \beta_2 B_2(X) + \beta_3 B_3(X)}_{\approx f(X)} + \varepsilon$$

This is again a multiple regression problem where we set $X_0 = B_0(X), X_1 = B_1(X), X_2 = B_2(X), X_3 = B_3(X)$. The corresponding design matrix becomes

$$\mathbf{B}_2 = \begin{pmatrix} 1 & x_1 & (x_1 - \tau_1)_+ & (x_1 - \tau_2)_+ \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & (x_n - \tau_1)_+ & (x_n - \tau_2)_+ \end{pmatrix}$$

and we estimate $f(x)$ by

$$\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 (x - \tau_1)_+ + \hat{\beta}_3 (x - \tau_2)_+$$

where $\hat{\beta} = (\mathbf{B}_2^T \mathbf{B}_2)^{-1} \mathbf{B}_2^T \mathbf{Y}$. The above method applies more generally to a multiple knot setting for the data on any intervals.

Cubic Splines: We can further consider fitting piecewise polynomials whose derivatives are also continuous. A popular choice is the so-called **cubic spline** that is a piecewise cubic polynomial function with continuous first and second derivatives. Again, we consider two knots and three regions: $[0, \tau_1], [\tau_1, \tau_2], [\tau_2, 1]$. Let $c(x)$ be a cubic spline. In $[0, \tau_1]$ we write

$$c(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3, x \leq \tau_1$$

And $c(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \delta(x)$ in $[\tau_1, \tau_2]$. By definition, $\delta(x)$ is a cubic function in $[\tau_1, \tau_2]$ and its first and second derivatives equal zero at $x = \tau_1$. Then we must have

$$\delta(x) = \beta_4 (x - \tau_1)_+^3, x \in [\tau_1, \tau_2]$$

which means

$$c(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - \tau_1)_+^3, x \in [\tau_1, \tau_2]$$

Likewise, in $[\tau_2, 1]$ we must have

$$c(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - \tau_1)_+^3 + \beta_5 (x - \tau_2)_+^3, x > \tau_2$$

Therefore, the basis functions are

$$\begin{aligned} B_0(x) &= 1, B_1(x) = x, B_2(x) = x^2, B_3(x) = x^3 \\ B_4(x) &= (x - \tau_1)_+^3, B_5(x) = (x - \tau_2)_+^3 \end{aligned}$$

The corresponding transformed design matrix becomes

$$\mathbf{B}_3 = \begin{pmatrix} 1 & x_1 & x_1^2 & x_1^3 & (x_1 - \tau_1)_+^3 & (x_1 - \tau_2)_+^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 & (x_n - \tau_1)_+^3 & (x_n - \tau_2)_+^3 \end{pmatrix}$$

and we estimate $f(x)$ by

$$\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \hat{\beta}_3 x^3 + \hat{\beta}_4 (x - \tau_1)_+^3 + \hat{\beta}_5 (x - \tau_2)_+^3$$

where $\hat{\beta} = (\mathbf{B}_3^T \mathbf{B}_3)^{-1} \mathbf{B}_3^T \mathbf{Y}$ is the least-squares estimate of the coefficients.

In general, if there are K knots $\{\tau_1, \dots, \tau_K\}$, then the basis functions of cubic splines are

$$\begin{aligned} B_0(x) &= 1, B_1(x) = x, B_2(x) = x^2, B_3(x) = x^3 \\ B_4(x) &= (x - \tau_1)_+^3, \dots, B_{K+3}(x) = (x - \tau_K)_+^3 \end{aligned}$$

By approximating the nonparametric function $f(X)$ by the spline function with knots $\{\tau_j\}_{j=1}^K$, we have

$$Y = \underbrace{\beta_0 B_0(X) + \beta_1 B_1(X) + \dots + \beta_{K+3} B_{K+3}(X)}_{\approx f(X)} + \varepsilon \quad (3.18)$$

This spline regression is again a multiple regression problem.

Natural Cubic Splines: Extrapolation is always a serious issue in regression. It is not wise to fit a cubic function to a region where the observations are scarce. If we must, extrapolation with a linear function is preferred. A natural cubic spline is a special cubic spline with additional constraints: the cubic spline must be linear beyond two end knots. Consider a natural cubic spline, $NC(x)$, with knots at $\{\tau_1, \dots, \tau_K\}$. By its cubic spline representation, we can write

$$NC(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{j=1}^K \beta_{3+j} (x - \tau_j)_+^3$$

First, $NC(x)$ is linear for $x < \tau_1$, which implies that

$$\beta_2 = \beta_3 = 0$$

Second, $NC(x)$ is linear for $x > \tau_K$, which means that

$$\sum_{j=1}^K \beta_{3+j} = 0, \quad \sum_{j=1}^K \tau_j \beta_{3+j} = 0$$

corresponding to the coefficients for the cubic and quadratic term of the polynomial $\sum_{j=1}^K \beta_{3+j} (x - \tau_j)_+^3$ for $x > \tau_K$. We solve for β_{K+2}, β_{K+3} from the above equations and then write $NC(x)$ as

$$NC(x) = \sum_{j=0}^{K-1} \beta_j B_j(x)$$

where the natural cubic spline basis functions are given by

$$\begin{aligned} B_0(x) &= 1, B_1(x) = x \\ B_{j+1}(x) &= \frac{(x - \tau_j)_+^3 - (x - \tau_K)_+^3}{\tau_j - \tau_K} - \frac{(x - \tau_{K-1})_+^3 - (x - \tau_K)_+^3}{\tau_{K-1} - \tau_K} \\ &\text{for } j = 1, \dots, K-2 \end{aligned}$$

Again, by approximating the nonparametric function with the natural cubic spline, we have

$$Y = \sum_{j=0}^{K-1} \beta_j B_j(X) + \varepsilon \quad (3.19)$$

which can be solved by using multiple regression techniques.

3.0.7 Multiple Covariates

The concept of polynomial regression extends to multivariate covariates. The simplest example is the bivariate regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_1 X_2 + \beta_5 X_2^2 + \varepsilon$$

The term $X_1 X_2$ is called the interaction, which quantifies how X_1 and X_2 work together to contribute to the response. Often, one introduces interactions without using the quadratic term, leading to a slightly simplified model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

More generally, the multivariate quadratic regression is of the form

$$Y = \sum_{j=1}^p \beta_j X_j + \sum_{j \leq k} \beta_{jk} X_j X_k + \varepsilon \quad (3.20)$$

and the multivariate regression with main effects (the linear terms) and interactions is of the form

$$Y = \sum_{j=1}^p \beta_j X_j + \sum_{j < k} \beta_{jk} X_j X_k + \varepsilon \quad (3.21)$$

This concept can also be extended to the multivariate spline case. The basis function can be the tensor of univariate spline basis function for not only unstructured $f(\mathbf{x})$, but also other basis functions for structured $f(\mathbf{x})$. Unstructured nonparametric functions are not very useful: If each variable uses 100 basis functions, then there are 100^p basis functions in the tensor products, which is prohibitively large for say, $p = 10$. Such an issue is termed the "curse-of-dimensionality" in literature. See Hastie and Tibshirani (1990) and Fan and Gijbels (1996).

On the other hand, for the structured multivariate model, such as the following additive model (Stone, 1985, 1994; Hastie and Tibshirani, 1990),

$$Y = f_1(X_1) + \cdots + f_p(X_p) + \varepsilon \quad (3.22)$$

the basis functions are simply the collection of all univariate basis functions for approximating f_1, \dots, f_p . The total number grows only linearly with p . In general, let $B_m(\mathbf{x})$

be the basis functions $m = 1, \dots, M$. Then, we approximate multivariate nonparametric regression model $Y = f(\mathbf{X}) + \varepsilon$ by

$$Y = \sum_{m=1}^M \beta_j B_j(\mathbf{X}) + \varepsilon \quad (3.23)$$

This can be fit using a multiple regression technique. The new design matrix is

$$\mathbf{B} = \begin{pmatrix} B_1(\mathbf{X}_1) & \cdots & B_M(\mathbf{X}_1) \\ \vdots & \cdots & \vdots \\ B_1(\mathbf{X}_n) & \cdots & B_M(\mathbf{X}_n) \end{pmatrix}$$

and the least-squares estimate is given by

$$\hat{f}(\mathbf{x}) = \sum_{m=1}^M \hat{\beta}_m B_m(\mathbf{x})$$

where

$$\hat{\beta} = (\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}^T \mathbf{Y}$$

The above fitting implicitly assumes that $M \ll n$. This condition in fact can easily be violated in unstructured multivariate nonparametric regression. For the additive model (3.22), in which we assume $f(\mathbf{x}) = \sum_{j=1}^p f_j(x_j)$ where each $f_j(x_j)$ is a smooth univariate function of x_j , the univariate basis expansion ideas can be readily applied to approximation of each $f_j(x_j)$:

$$f_j(x_j) \approx \sum_{m=1}^{M_j} B_{jm}(x_j) \beta_{jm}$$

which implies that the fitted regression function is

$$f(\mathbf{x}) \approx \sum_{j=1}^p \sum_{m=1}^{M_j} B_{jm}(x_j) \beta_{jm}$$

In Section 2.6.5 and Section 2.7 we introduce a fully nonparametric multiple regression technique which can be regarded as a basis expansion method where the basis functions are given by kernel functions.

3.0.8 Ridge Regression

Recall that the ordinary least squares estimate is defined by $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ when \mathbf{X} is of full rank. In practice, we often encounter highly correlated covariates, which is known as the **collinearity issue**. As a result, although $\mathbf{X}^T \mathbf{X}$ is still invertible, its smallest eigenvalue can be very small. Under the homoscedastic error model, the variance-covariance matrix of the OLS estimate is $\text{Var}(\hat{\beta}) = (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2$. Thus, the collinearity issue makes $\text{Var}(\hat{\beta})$ large.

Hoerl and Kennard (1970) introduced the ridge regression estimator as follows:

$$\hat{\beta}_\lambda = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{Y} \quad (3.24)$$

where $\lambda > 0$ is a regularization parameter. In the usual case ($\mathbf{X}^T \mathbf{X}$ is invertible), ridge regression reduces to OLS by setting $\lambda = 0$. However, ridge regression is always well defined even when \mathbf{X} is not full rank. Under the assumption $\text{Var}(\varepsilon) = \sigma^2 \mathbf{I}$, it is easy to show that

$$\text{Var}(\hat{\beta}_\lambda) = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \sigma^2 \quad (3.25)$$

We always have $\text{Var}(\hat{\beta}_\lambda) < (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2$. Ridge regression estimator reduces the estimation variance by paying a price in estimation bias:

$$\mathbb{E}(\hat{\beta}_\lambda) - \beta = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{X} \beta - \beta = -\lambda (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \beta \quad (3.26)$$

The overall estimation accuracy is gauged by the mean squared error (MSE). For $\hat{\beta}_\lambda$ its MSE is given by

$$\text{MSE}(\hat{\beta}_\lambda) = \mathbb{E}(\|\hat{\beta}_\lambda - \beta\|^2) \quad (3.27)$$

By (3.25) and (3.26) we have

$$\begin{aligned} \text{MSE}(\hat{\beta}_\lambda) &= \text{tr} \left((\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \sigma^2 \right) \\ &\quad + \lambda^2 \beta^T (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-2} \beta \\ &= \text{tr} \left((\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-2} [\lambda^2 \beta \beta^T + \sigma^2 \mathbf{X}^T \mathbf{X}] \right) \end{aligned} \quad (3.28)$$

It can be shown that $\left. \frac{d\text{MSE}(\hat{\beta}_\lambda)}{d\lambda} \right|_{\lambda=0} < 0$, which implies that there are some proper λ values by which ridge regression improves OLS.

3.0.9 ℓ_2 Penalized Least Squares

Define a penalized residual sum-of-squares (PRSS) as follows:

$$\text{PRSS}(\beta | \lambda) = \sum_{i=1}^n \left(Y_i - \sum_{j=1}^p X_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 \quad (3.29)$$

Then let

$$\hat{\beta}_\lambda = \text{argmin}_\beta \text{PRSS}(\beta | \lambda) \quad (3.30)$$

Note that we can write it in a matrix form

$$\text{PRSS}(\beta | \lambda) = \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|^2$$

The term $\lambda \|\beta\|^2$ is called the ℓ_2 -**penalty of β** . Taking derivatives with respect to β and setting it to zero, we solve the root of the following equation

$$-\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\beta) + \lambda\beta = 0$$

which yields

$$\hat{\beta}_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1} \mathbf{X}^T\mathbf{Y}$$

The above discussion shows that ridge regression is equivalent to the ℓ_2 penalized least-squares.

We have seen that ridge regression can achieve a smaller MSE than OLS. In other words, the ℓ_2 penalty term helps regularize (reduce) estimation variance and produces a better estimator when the reduction in variance exceeds the induced extra bias. From this perspective, one can also consider a more general ℓ_q penalized least-squares estimate

$$\min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda \sum_{j=1}^p |\beta_j|^q \quad (3.31)$$

where q is a positive constant. This is referred to as the **Bridge estimator** (Frank and Friedman, 1993). The ℓ_q penalty is strictly concave when $0 < q < 1$, and strictly convex when $q > 1$. For $q = 1$, the resulting ℓ_1 penalized leastsquares is also known as the Lasso (Tibshirani, 1996). Among all Bridge estimators only the ridge regression has a nice closed-form solution with a general design matrix.

3.0.10 Bayesian Interpretation

Ridge regression has a neat Bayesian interpretation in the sense that it can be a formal Bayes estimator. We begin with the homoscedastic Gaussian error model:

$$Y_i = \sum_{j=1}^p X_{ij}\beta_j + \varepsilon_i$$

and $\varepsilon_i | \mathbf{X}_i \sim N(0, \sigma^2)$. Now suppose that β_j 's are also independent $N(0, \tau^2)$ variables, which represent our knowledge about the regression coefficients before seeing the data. In Bayesian statistics, $N(0, \tau^2)$ is called the **prior distribution** of β_j . The model and the prior together give us the posterior distribution of β given the data (the conditional distribution of β given \mathbf{Y}, \mathbf{X}). Straightforward calculations yield

$$P(\beta | \mathbf{Y}, \mathbf{X}) \propto \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{Y} - \mathbf{X}\beta\|^2\right) \exp\left(-\frac{1}{2\tau^2} \|\beta\|^2\right) \quad (3.32)$$

A maximum posteriori probability (MAP) estimate is defined as

$$\begin{aligned} \hat{\beta}^{\text{MAP}} &= \arg\max_{\beta} P(\beta | \mathbf{Y}, \mathbf{X}) \\ &= \arg\max_{\beta} \left\{ -\frac{1}{2\sigma^2} \|\mathbf{Y} - \mathbf{X}\beta\|^2 - \frac{1}{2\tau^2} \|\beta\|^2 \right\} \end{aligned} \quad (3.33)$$

It is easy to see that $\hat{\beta}^{\text{MAP}}$ is ridge regression with $\lambda = \frac{\sigma^2}{\tau^2}$. Another popular Bayesian estimate is the posterior mean. In this model, the posterior mean and posterior mode are the same.

From the Bayesian perspective, it is easy to construct a generalized ridge regression estimator. Suppose that the prior distribution for the entire β vector is $N(0, \Sigma)$, where Σ is a general positive definite matrix. Then the posterior distribution is computed as

$$P(\beta | \mathbf{Y}, \mathbf{X}) \propto \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{Y} - \mathbf{X}\beta\|^2\right) \exp\left(-\frac{1}{2}\beta^T \Sigma^{-1} \beta\right) \quad (3.34)$$

The corresponding MAP estimate is

$$\begin{aligned} \hat{\beta}^{\text{MAP}} &= \operatorname{argmax}_{\beta} P(\beta | \mathbf{Y}, \mathbf{X}) \\ &= \operatorname{argmax}_{\beta} \left\{ -\frac{1}{2\sigma^2}\|\mathbf{Y} - \mathbf{X}\beta\|^2 - \frac{1}{2}\beta^T \Sigma^{-1} \beta \right\} \end{aligned} \quad (3.35)$$

It is easy to see that

$$\hat{\beta}^{\text{MAP}} = \left(\mathbf{X}^T \mathbf{X} + \sigma^2 \Sigma^{-1} \right)^{-1} \mathbf{X}^T \mathbf{Y}$$

This generalized ridge regression can take into account different scales of covariates, by an appropriate choice of Σ .

3.0.11 Ridge Regression Solution Path

The performance of ridge regression heavily depends on the choice of λ . In practice we only need to compute ridge regression estimates at a fine grid of λ values and then select the best from these candidate solutions. Although ridge regression is easy to compute for a λ owing to its nice closed-form solution expression, the total cost could be high if the process is repeated many times. Through a more careful analysis, one can see that the solutions of ridge regression at a fine grid of λ values can be computed very efficiently via **singular value decomposition**.

Assume $n > p$ and \mathbf{X} is full rank. The singular value decomposition (SVD) of \mathbf{X} is given by

$$\mathbf{X} = \mathbf{U} \mathbf{D} \mathbf{V}^T$$

where \mathbf{U} is a $n \times p$ orthogonal matrix, \mathbf{V} is a $p \times p$ orthogonal matrix and \mathbf{D} is a $p \times p$ diagonal matrix whose diagonal elements are the ordered (from large to small) singular values of \mathbf{X} . Then

$$\begin{aligned} \mathbf{X}^T \mathbf{X} &= \mathbf{V} \mathbf{D} \mathbf{U}^T \mathbf{U} \mathbf{D} \mathbf{V}^T = \mathbf{V} \mathbf{D}^2 \mathbf{V}^T \\ \mathbf{X}^T \mathbf{X} + \lambda \mathbf{I} &= \mathbf{V} \mathbf{D}^2 \mathbf{V}^T + \lambda \mathbf{I} = \mathbf{V} (\mathbf{D}^2 + \lambda \mathbf{I}) \mathbf{V}^T \\ (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} &= \mathbf{V} (\mathbf{D}^2 + \lambda \mathbf{I})^{-1} \mathbf{V}^T \end{aligned}$$

The ridge regression estimator $\hat{\beta}_\lambda$ can now be written as

$$\begin{aligned}\hat{\beta}_\lambda &= (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{Y} \\ &= \mathbf{V} (\mathbf{D}^2 + \lambda \mathbf{I})^{-1} \mathbf{D} \mathbf{U}^T \mathbf{Y} \\ &= \sum_{j=1}^p \frac{d_j}{d_j^2 + \lambda} \langle \mathbf{U}_j, \mathbf{Y} \rangle \mathbf{V}_j\end{aligned}\tag{3.36}$$

where d_j is the j^{th} diagonal element of \mathbf{D} and $\langle \mathbf{U}_j, \mathbf{Y} \rangle$ is the inner product between \mathbf{U}_j and \mathbf{Y} (\mathbf{U}_j (\mathbf{V}_j are respectively the j^{th} column of \mathbf{U} and \mathbf{V}). In particular, when $\lambda = 0$, ridge regression reduces to OLS and we have

$$\hat{\beta}^{\text{OLS}} = \mathbf{V} \mathbf{D}^{-1} \mathbf{U}^T \mathbf{Y} = \sum_{j=1}^p \frac{1}{d_j} \langle \mathbf{U}_j, \mathbf{Y} \rangle \mathbf{V}_j\tag{3.37}$$

Based on (3.36) we suggest the following procedure to compute ridge regression at a fine grid $\lambda_1, \dots, \lambda_M$:

1. Compute the SVD of \mathbf{X} and save $\mathbf{U}, \mathbf{D}, \mathbf{V}$.
2. Compute $\mathbf{w}_j = \frac{1}{d_j} \langle \mathbf{U}_j, \mathbf{Y} \rangle \mathbf{V}_j$ for $j = 1, \dots, p$ and save \mathbf{w}_j 's.
3. For $m = 1, 2, \dots, M$
 - (a) compute $\gamma_j = \frac{d_j^2}{d_j^2 + \lambda_m}$
 - (b) compute $\hat{\beta}_{\lambda_m} = \sum_{j=1}^p \gamma_j \mathbf{w}_j$

The essence of the above algorithm is to compute the common vectors $\{\mathbf{w}_j\}_{j=1}^p$ first and then utilize (3.36).

Kernel Ridge Regression

In this section we introduce a nonparametric generalization of ridge regression. Our discussion begins with the following theorem.

Theorem 3.0.4 Ridge regression estimator is equal to

$$\hat{\beta}_\lambda = \mathbf{X}^T (\mathbf{X} \mathbf{X}^T + \lambda \mathbf{I})^{-1} \mathbf{Y}\tag{3.38}$$

and the fitted value of Y at \mathbf{x} is

$$\hat{y} = \mathbf{x}^T \hat{\beta}_\lambda = \mathbf{x}^T \mathbf{X}^T (\mathbf{X} \mathbf{X}^T + \lambda \mathbf{I})^{-1} \mathbf{Y}\tag{3.39}$$

Proof. Observe the following identity

$$(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}) \mathbf{X}^T = \mathbf{X}^T \mathbf{X} \mathbf{X}^T + \lambda \mathbf{X}^T = \mathbf{X}^T (\mathbf{X} \mathbf{X}^T + \lambda \mathbf{I})$$

Thus, we have

$$\mathbf{X}^T = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T (\mathbf{X} \mathbf{X}^T + \lambda \mathbf{I})$$

and

$$\mathbf{X}^T (\mathbf{X}\mathbf{X}^T + \lambda \mathbf{I})^{-1} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T$$

Then by using (3.24)

$$\hat{\beta}_\lambda = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{Y}$$

we obtain (3.38) and hence (3.39). It is important to see that $\mathbf{X}\mathbf{X}^T$ not $\mathbf{X}^T \mathbf{X}$ appears in the expression for $\hat{\beta}_\lambda$. Note that $\mathbf{X}\mathbf{X}^T$ is a $n \times n$ matrix and its ij elements is $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$. Similarly, $\mathbf{x}^T \mathbf{X}^T$ is an n -dimensional vector with the i th element being $\langle \mathbf{x}, \mathbf{x}_i \rangle$ $i = 1, \dots, n$. Therefore, the prediction by ridge regression boils down to computing the inner product between p -dimensional covariate vectors. This is the foundation of the so-called "kernel trick". ■

Suppose that we use another "inner product" to replace the usual inner product in Theorem 3.0.4 then we may end up with a new ridge regression estimator. To be more specific, let us replace $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$ with $K(\mathbf{x}_i, \mathbf{x}_j)$ where $K(\cdot, \cdot)$ is a known function:

$$\begin{aligned} \mathbf{x}^T \mathbf{X}^T &\rightarrow (K(\mathbf{x}, \mathbf{X}_1), \dots, K(\mathbf{x}, \mathbf{X}_n)) \\ \mathbf{X}\mathbf{X}^T &\rightarrow \mathbf{K} = (K(\mathbf{X}_i, \mathbf{X}_j))_{1 \leq i, j \leq n} \end{aligned}$$

By doing so, we turn (3.39) into

$$\hat{y} = (K(\mathbf{x}, \mathbf{X}_1), \dots, K(\mathbf{x}, \mathbf{X}_n)) (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{Y} = \sum_{i=1}^n \hat{\alpha}_i K(\mathbf{x}, \mathbf{X}_i) \quad (3.40)$$

where $\hat{\alpha} = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{Y}$. In particular, the fitted \mathbf{Y} vector is

$$\hat{\mathbf{Y}} = \mathbf{K}(\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{Y} \quad (3.41)$$

The above formula gives the so-called **kernel ridge regression**. Because $\mathbf{X}\mathbf{X}^T$ is at least positive semi-definite, it is required that \mathbf{K} is also positive semidefinite. Some widely used kernel functions (Hastie, Tibshirani and Friedman, 2009) include linear kernel: $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$,

1. polynomial kernel: $K(\mathbf{x}_i, \mathbf{x}_j) = (1 + \langle \mathbf{x}_i, \mathbf{x}_j \rangle)^d, d = 2, 3, \dots$,
2. radial basis kernel: $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2}, \gamma > 0$, which is the Gaussian kernel,
3. and $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|}, \gamma > 0$, which is the Laplacian kernel.

To show how we get (3.41) more formally, let us consider approximate the multivariate regression by using the kernel basis functions $\{K(\cdot, \mathbf{x}_j)\}_{j=1}^n$ so that our observed data are now modeled as

$$Y_i = \sum_{j=1}^n \alpha_j K(\mathbf{X}_i, \mathbf{X}_j) + \varepsilon_i$$

or in matrix form $\mathbf{Y} = \mathbf{K}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$. If we apply the ridge regression

$$\|\mathbf{Y} - \mathbf{K}\boldsymbol{\alpha}\|^2 + \frac{\lambda}{2}\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha}$$

the minimizer of the above problem is

$$\hat{\boldsymbol{\alpha}} = \left(\mathbf{K}^T\mathbf{K} + \lambda\mathbf{K}\right)^{-1}\mathbf{K}^T\mathbf{Y} = \{\mathbf{K}(\mathbf{K} + \lambda\mathbf{I})\}^{-1}\mathbf{K}\mathbf{Y}$$

where we use the fact \mathbf{K} is symmetric. Assuming \mathbf{K} is invertible, we easily get (3.41).

So far we have only derived the kernel ridge regression based on heuristics and the kernel trick.

3.0.12 Regression in Reproducing Kernel Hilbert Space

This subsection we will show the kernel ridge regression can be formally derived based on the theory of function estimation in a reproducing kernel Hilbert space. A Hilbert space is an abstract vector space endowed by the structure of an inner product. Let \mathcal{X} be an arbitrary set and \mathcal{H} be a Hilbert space of realvalued functions on \mathcal{X} , endowed by the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. The evaluation functional over the Hilbert space of functions \mathcal{H} is a linear functional that evaluates each function at a point x :

$$L_x : f \rightarrow f(x), \forall f \in \mathcal{H}$$

A Hilbert space \mathcal{H} is called a reproducing kernel Hilbert space (RKHS) if, for all $x \in \mathcal{X}$, the map L_x is continuous at any $f \in \mathcal{H}$, namely, there exists some $C > 0$ such that

$$|L_x(f)| = |f(x)| \leq C\|f\|_{\mathcal{H}}, \quad \forall f \in \mathcal{H}$$

By the Riesz representation theorem, for all $x \in \mathcal{X}$, there exists a unique element $K_x \in \mathcal{H}$ with the reproducing property

$$f(x) = L_x(f) = \langle f, K_x \rangle_{\mathcal{H}}, \quad \forall f \in \mathcal{H}$$

Since K_x is itself a function in \mathcal{H} , it holds that for every $x' \in \mathcal{X}$, there exists a $K_{x'} \in \mathcal{H}$ such that

$$K_x(x') = \langle K_x, K_{x'} \rangle_{\mathcal{H}}$$

This allows us to define the reproducing kernel $K(x, x') = \langle K_x, K_{x'} \rangle_{\mathcal{H}}$.

From the definition, it is easy to see that the reproducing kernel K is a symmetric and semi-positive function:

$$\sum_{i,j=1}^n c_i c_j K(x_i, x_j) = \sum_{i,j=1}^n c_i c_j \langle K_{x_i}, K_{x_j} \rangle_{\mathcal{H}} = \left\| \sum_{i=1}^n c_i K_{x_i} \right\|_{\mathcal{H}}^2 \geq 0$$

for all c 's and x 's. The reproducing Hilbert space is a class of nonparametric functions, satisfying the above properties.

Let \mathcal{H}_K denote the reproducing kernel Hilbert space (RKHS) with kernel $K(\mathbf{x}, \mathbf{x}')$ (Wahba, 1990; Halmos, 2017). Then, the kernel $K(\mathbf{x}, \mathbf{x}')$ admits the eigen-decomposition

$$K(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{\infty} \gamma_j \psi_j(\mathbf{x}) \psi_j(\mathbf{x}') \quad (3.42)$$

where $\gamma_j \geq 0$ are eigen-values and $\sum_{j=1}^{\infty} \gamma_j^2 < \infty$. Let g and g' be any two functions in \mathcal{H}_K with expansions in terms of these eigen-functions

$$g(\mathbf{x}) = \sum_{j=1}^{\infty} \beta_j \psi_j(\mathbf{x}), \quad g'(\mathbf{x}) = \sum_{j=1}^{\infty} \beta'_j \psi_j(\mathbf{x})$$

and their inner product is defined as

$$\langle g, g' \rangle_{\mathcal{H}_K} = \sum_{j=1}^{\infty} \frac{\beta_j \beta'_j}{\gamma_j} \quad (3.43)$$

The functional ℓ_2 norm of $g(\mathbf{x})$ is equal to

$$\|g\|_{\mathcal{H}_K}^2 = \langle g, g \rangle_{\mathcal{H}_K} = \sum_{j=1}^{\infty} \frac{\beta_j^2}{\gamma_j} \quad (3.44)$$

The first property shows the reproducibility of the kernel K .

Theorem 3.0.5 Let g be a function in \mathcal{H}_K . The following identities hold:

1. $\langle K(\cdot, \mathbf{x}'), g \rangle_{\mathcal{H}_K} = g(\mathbf{x}')$
2. $\langle K(\cdot, \mathbf{x}_1), K(\cdot, \mathbf{x}_2) \rangle_{\mathcal{H}_K} = K(\mathbf{x}_1, \mathbf{x}_2)$
3. If $g(\mathbf{x}) = \sum_{i=1}^n \alpha_i K(\mathbf{x}, \mathbf{x}_i)$, then $\|g\|_{\mathcal{H}_K}^2 = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$

Proof. Write $g(\mathbf{x}) = \sum_{j=1}^{\infty} \beta_j \psi_j(\mathbf{x})$, by (3.42) we have $K(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{\infty} (\gamma_j \psi_j(\mathbf{x}')) \psi_j(\mathbf{x})$. Thus

$$\langle K(\cdot, \mathbf{x}'), g \rangle_{\mathcal{H}_K} = \sum_{j=1}^{\infty} \frac{\beta_j \gamma_j \psi_j(\mathbf{x}')}{\gamma_j} = \sum_{j=1}^{\infty} \beta_j \psi_j(\mathbf{x}') = g(\mathbf{x}')$$

This proves part (i). Now we apply part (i) to get part (ii) by letting $g(\mathbf{x}) = K(\mathbf{x}, \mathbf{x}_2)$. For part (iii) we observe that

$$\begin{aligned} \|g\|_{\mathcal{H}_K}^2 &= \left\langle \sum_{i=1}^n \alpha_i K(\mathbf{x}, \mathbf{x}_i), \sum_{j=1}^n \alpha_j K(\mathbf{x}, \mathbf{x}_j) \right\rangle_{\mathcal{H}_K} \\ &= \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \langle K(\mathbf{x}, \mathbf{x}_i), K(\mathbf{x}, \mathbf{x}_j) \rangle_{\mathcal{H}_K} \\ &= \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \end{aligned}$$

where we have used part (ii) in the final step. ■

Consider a general regression model

$$Y = f(\mathbf{X}) + \varepsilon \quad (3.45)$$

where ε is independent of \mathbf{X} and has zero mean and variance σ^2 . Given a realization $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ from the above model, we wish to fit the regression function in \mathcal{H}_K via the following penalized least-squares:

$$\hat{f} = \operatorname{argmin}_{f \in \mathcal{H}_K} \sum_{i=1}^n [Y_i - f(\mathbf{X}_i)]^2 + \lambda \|f\|_{\mathcal{H}_K}^2, \quad \lambda > 0 \quad (3.46)$$

Note that without $\|f\|_{\mathcal{H}_K}^2$ term there are infinite many functions in \mathcal{H}_K that can fit the observations perfectly, i.e., $Y_i = f(\mathbf{X}_i)$ for $i = 1, \dots, n$. By using the eigen-function expansion of f

$$f(\mathbf{x}) = \sum_{j=1}^{\infty} \beta_j \psi_j(\mathbf{x}) \quad (3.47)$$

an equivalent formulation of (3.46) is

$$\min_{\{\beta_j\}_{j=1}^{\infty}} \sum_{i=1}^n \left[Y_i - \sum_{j=1}^{\infty} \beta_j \psi_j(\mathbf{X}_i) \right]^2 + \lambda \sum_{j=1}^{\infty} \frac{1}{\gamma_j} \beta_j^2 \quad (3.48)$$

Define $\beta_j^* = \frac{\beta_j}{\sqrt{\gamma_j}}$ and $\psi_j^* = \sqrt{\gamma_j} \psi_j$ for $j = 1, 2, \dots$. Then (3.48) can be rewritten as

$$\min_{\{\beta_j^*\}_{j=1}^{\infty}} \sum_{i=1}^n \left[Y_i - \sum_{j=1}^{\infty} \beta_j^* \psi_j^*(\mathbf{X}_i) \right]^2 + \lambda \sum_{j=1}^{\infty} (\beta_j^*)^2 \quad (3.49)$$

The above can be seen as a ridge regression estimate in an infinite dimensional space. Symbolically, our covariate vector is now $(\psi_1^*(\mathbf{x}), \psi_2^*(\mathbf{x}), \dots)$ and the enlarged design matrix is

$$\Psi = \begin{pmatrix} \psi_1^*(\mathbf{X}_1) & \cdots & \psi_j^*(\mathbf{X}_1) & \cdots \\ \vdots & \cdots & \vdots & \cdots \\ \psi_1^*(\mathbf{X}_n) & \cdots & \psi_j^*(\mathbf{X}_n) & \cdots \end{pmatrix}$$

Because Theorem 3.0.4 is valid for any finite dimensional covariate space, it is not unreasonable to extrapolate it to the above infinite dimensional setting. The key assumption is that we can compute the inner product in the enlarged space. This is indeed true because

$$\text{inner product} = \sum_{j=1}^{\infty} \psi_j^*(\mathbf{x}_i) \psi_j^*(\mathbf{x}_{i'}) = \sum_{j=1}^{\infty} \gamma_j \psi_j(\mathbf{x}_i) \psi_j(\mathbf{x}_{i'}) = K(\mathbf{x}_i, \mathbf{x}_{i'})$$

Now we can directly apply the kernel ridge regression formula to get

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^n \hat{\alpha}_i K(\mathbf{x}, \mathbf{X}_i) \quad (3.50)$$

where $\mathbf{K} = (K(\mathbf{X}_i, \mathbf{X}_j))_{1 \leq i, j \leq n}$ and

$$\hat{\boldsymbol{\alpha}} = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{Y} \quad (3.51)$$

We have derived (3.50) by extrapolating Theorem 3.0.4 to an infinite dimensional space. Although the idea seems correct, we still need a rigorous proof. Moreover, Theorem 3.0.4 only concerns ridge regression, but it turns out that (3.50) can be made much more general.

Theorem 3.0.6 — representer theorem (Wahba, 1990). Consider a general loss function $L(y, f(\mathbf{x}))$ and let

$$\hat{f} = \operatorname{argmin}_{f \in \mathcal{H}_K} \sum_{i=1}^n L(y_i, f(\mathbf{x}_i)) + P_\lambda(\|f\|_{\mathcal{H}_K}), \quad \lambda > 0$$

where $P_\lambda(t)$ is a strictly increasing function on $[0, \infty)$. Then we must have

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^n \hat{\alpha}_i K(\mathbf{x}, \mathbf{X}_i) \quad (3.52)$$

where $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \dots, \hat{\alpha}_n)$ is the solution to the following problem

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^n L\left(y_i, \sum_{j=1}^n \alpha_j K(\mathbf{x}, \mathbf{x}_j)\right) + P_\lambda\left(\sqrt{\boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha}}\right) \quad (3.53)$$

Proof. Any function f in \mathcal{H}_K can be decomposed as the sum of two functions: one is in the span $\{K(\cdot, \mathbf{X}_1), \dots, K(\cdot, \mathbf{X}_n)\}$ and the other is in the orthogonal complement. In other words, we write

$$f(\mathbf{x}) = \sum_{i=1}^n \alpha_i K(\mathbf{x}, \mathbf{X}_i) + r(\mathbf{x})$$

where $\langle r(\mathbf{x}), K(\mathbf{x}, \mathbf{X}_i) \rangle_{\mathcal{H}_K} = 0$ for all $i = 1, 2, \dots, n$. By part (i) of Theorem 3.0.5 we have

$$r(\mathbf{x}_i) = \langle r, K(\cdot, \mathbf{X}_i) \rangle_{\mathcal{H}_K} = 0, \quad 1 \leq i \leq n$$

Denote by $g(\mathbf{x}) = \sum_{i=1}^n \alpha_i K(\mathbf{x}, \mathbf{X}_i)$. Then we have $g(\mathbf{X}_i) = f(\mathbf{X}_i)$ for all i , which implies

$$\sum_{i=1}^n L(Y_i, f(\mathbf{X}_i)) = \sum_{i=1}^n L(Y_i, g(\mathbf{X}_i)) \quad (3.54)$$

Moreover, we notice

$$\begin{aligned} \|f\|_{\mathcal{H}_K}^2 &= \langle g + r, g + r \rangle_{\mathcal{H}_K} \\ &= \langle g, g \rangle_{\mathcal{H}_K} + \langle r, r \rangle_{\mathcal{H}_K} + 2\langle g, r \rangle_{\mathcal{H}_K} \end{aligned}$$

and

$$\langle g, r \rangle_{\mathcal{H}_K} = \sum_{i=1}^n \alpha_i \langle K(\cdot, \mathbf{X}_i), r \rangle_{\mathcal{H}_K} = 0$$

Table 2.1: A list of Commonly used kernels.

Linear kernel	$K(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$
Polynomial kernel	$K(\mathbf{x}_i, \mathbf{x}_j) = (1 + \langle \mathbf{x}_i, \mathbf{x}_j \rangle)^d$
Gaussian kernel	$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma \ \mathbf{x}_i - \mathbf{x}_j\ ^2}$
Laplacian kernel	$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma \ \mathbf{x}_i - \mathbf{x}_j\ }$

Thus $\|f\|_{\mathcal{H}_K}^2 = \|g\|_{\mathcal{H}_K}^2 + \|r\|_{\mathcal{H}_K}^2$. Because $P_\lambda(\cdot)$ is a strictly increasing function, we then have

$$P_\lambda(\|f\|_{\mathcal{H}_K}) \geq P_\lambda(\|g\|_{\mathcal{H}_K}) \quad (3.55)$$

and the equality holds if and only if $f = g$. Combining (3.54) and (3.55) we prove (3.52). To prove (3.53), we use (3.52) and part (iii) of Theorem 3.0.5 to write

$$\|f\|_{\mathcal{H}_K}^2 = \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} \quad (3.56)$$

Hence $P_\lambda(\|f\|_{\mathcal{H}_K}) = P_\lambda(\sqrt{\boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha}})$ under (3.52) ■

Theorem 3.0.6 shows that for a wide class of statistical estimation problems in a RKHS, although the criterion is defined in an infinite dimensional space, the solution always has a finite dimensional representation based on the kernel functions. This provides a solid mathematical foundation for the kernel trick without resorting to any optimization/computational arguments.

Let the loss function in Theorem 3.0.6 be the squared error loss and $P_\lambda(t) = \lambda t^2$. Then Theorem 3.0.6 handles the problem defined in (3.46) and (3.53) reduces to

$$\min_{\boldsymbol{\alpha}} \|\mathbf{Y} - \mathbf{K} \boldsymbol{\alpha}\|^2 + \lambda \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} \quad (3.57)$$

It is easy to see the solution is

$$\hat{\boldsymbol{\alpha}} = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{Y}$$

which is identical to (3.51). The fitted multivariate nonparametric regression function is given by (3.52). In practice, one takes a kernel function from the list of linear, polynomial, Gaussian or Laplacian kernels given in Table 2.1. It remains to show how to choose the regularization parameter λ (and γ for Gaussian and Laplacian kernels) to optimize the prediction performance. This can be done by cross-validation methods outlined in the next section.

Leave-one-out and Generalized Cross-validation

We have seen that both ridge regression and the kernel ridge regression use a tuning parameter λ . In practice, we would like to use the data to pick a data-driven λ in order

Table 3.1: A list of commonly used regression methods and their \mathbf{S} matrices. d_j s are the singular values of \mathbf{X} and γ_i s are the eigenvalues of \mathbf{K} .

Method	\mathbf{S}	$\text{tr } \mathbf{S}$
Multiple Linear Regression	$\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$	p
Ridge Regression	$\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T$	$\sum_{j=1}^p \frac{d_j^2}{d_j^2 + \lambda}$
Kernel Regression in RKHS	$\mathbf{K}(\mathbf{K} + \lambda\mathbf{I})^{-1}$	$\sum_{i=1}^n \frac{\gamma_i}{\gamma_i + \lambda}$

to achieve the "best" estimation/prediction performance. This problem is often called **tuning parameter** selection and is ubiquitous in modern statistics and machine learning. A general solution is k -fold crossvalidation (CV), such as 10-fold or 5-fold CV. k -fold CV estimates prediction errors as follows.

1. Divide data randomly and evenly into k subsets.
2. Use one of the subsets as the testing set and the remaining $k - 1$ subsets of data as a training set to compute testing errors.
3. Compute testing errors for each of k subsets of data and average these testing errors.

An interesting special case is the n -fold CV, which is also known as the leave-one-out CV.

In this section we focus on regression problems under the squared error loss. Following the above scheme, the leave-one-out CV error, using the quadratic loss, is defined as

$$\text{CV} = \frac{1}{n} \sum_{i=1}^n \left(Y_i - \hat{f}^{(-i)}(\mathbf{x}_i) \right)^2 \quad (3.58)$$

where $\hat{f}^{(-i)}(\mathbf{x}_i)$ is the predicted value at \mathbf{x}_i computed by using all the data except the i th observation. So in principle we need to repeat the same data fitting process n times to compute the leave-one-out CV. Fortunately, we can avoid much computation for many popular regression methods.

Definition 3.0.1 — linear smoother. A fitting method is called a **linear smoother** if we can write

$$\hat{\mathbf{Y}} = \mathbf{S}\mathbf{Y} \quad (3.59)$$

for any dataset $\{(\mathbf{x}_i, Y_i)\}_1^n$ where \mathbf{S} is a $n \times n$ matrix that only depends on \mathbf{X} . Many regression methods are linear smoothers with different \mathbf{S} matrices.

Assume that a linear smoother is fitted on $\{\mathbf{x}_i, Y_i\}_{i=1}^n$. Let \mathbf{x} be a new covariate vector and $\hat{f}(\mathbf{x})$ be its the predicted value by using the linear smoother. We then augment the dataset by including $(\mathbf{x}, \hat{f}(\mathbf{x}))$ and refit the linear smoother on this augmented dataset.

Definition 3.0.2 — self-stable. The linear smoother is said to be self-stable if the fit based on the augmented dataset is identical to the fit based on the original data regardless of x .

It is easy to check that the three linear smoothers in Table 2.2 all have the self-stable property.

Theorem 3.0.7 For a linear smoother $\hat{\mathbf{Y}} = \mathbf{S}\mathbf{Y}$ with the self-stable property, we have

$$Y_i - \hat{f}^{(-i)}(\mathbf{X}_i) = \frac{Y_i - \hat{Y}_i}{1 - S_{ii}} \quad (3.60)$$

and its leave-one-out CV error is equal to $\frac{1}{n} \sum_{i=1}^n \left(\frac{Y_i - \hat{Y}_i}{1 - S_{ii}} \right)^2$

Proof. We first apply the linear smoother to all the data except the i th to compute $\hat{f}^{(-i)}(\mathbf{X}_i)$. Write $\tilde{y}_j = y_j$ for $j \neq i$ and $\tilde{y}_i = \hat{f}^{(-i)}(\mathbf{X}_i)$. Then we apply the linear smoother to the following working dataset:

$$\left\{ (\mathbf{X}_j, Y_j), j \neq i, (\mathbf{X}_i, \tilde{Y}_i) \right\}$$

The self-stable property implies that the fit stays the same. In particular,

$$\tilde{Y}_i = \hat{f}^{(-i)}(\mathbf{X}_i) = (\mathbf{S}\tilde{\mathbf{Y}})_i = S_{ii}\tilde{Y}_i + \sum_{j \neq i} S_{ij}Y_j \quad (3.61)$$

and

$$\hat{Y}_i = (\mathbf{S}\mathbf{Y})_i = S_{ii}Y_i + \sum_{j \neq i} S_{ij}Y_j \quad (3.62)$$

Combining (3.61) and (3.62) yields

$$\tilde{Y}_i = \frac{\hat{Y}_i - S_{ii}Y_i}{1 - S_{ii}}$$

Thus,

$$Y_i - \tilde{Y}_i = Y_i - \frac{\hat{Y}_i - S_{ii}Y_i}{1 - S_{ii}} = \frac{Y_i - \hat{Y}_i}{1 - S_{ii}}$$

The proof is now complete. ■

Theorem 3.0.7 shows a nice shortcut for computing the leave-one-out CV error of a self-stable linear smoother. For some smoothers $\text{tr } \mathbf{S}$ can be computed more easily than its diagonal elements. To take advantage of this, generalized cross-validation (GCV) (Golub, Heath and Wahba, 1979) is a convenient computational approximation to the leave-one-out CV error.

Suppose that we approximate each diagonal elements of \mathbf{S} by their average which equals $\frac{\text{tr} \mathbf{S}}{n}$ then we have

$$\frac{1}{n} \sum_{i=1}^n \left(\frac{Y_i - \hat{Y}_i}{1 - S_{ii}} \right)^2 \approx \frac{1}{n} \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\left(1 - \frac{\text{tr} \mathbf{S}}{n}\right)^2} := \text{GCV}$$

In the literature $\text{tr} \mathbf{S}$ is called the **effective degrees of freedom** of the linear smoother. Its rigorous justification is based on Stein's unbiased risk estimation theory (Stein, 1981; Efron, 1986). In Table 2.2 we list the degrees of freedom of three popular linear smoothers.

Now we are ready to handle the tuning parameter selection issue in the linear smoother. We write $\mathbf{S} = \mathbf{S}_\lambda$ and

$$\text{GCV}(\lambda) = \frac{1}{n} \frac{\mathbf{Y}^T (\mathbf{I} - \mathbf{S}_\lambda)^2 \mathbf{Y}}{\left(1 - \frac{\text{tr} \mathbf{S}_\lambda}{n}\right)^2}$$

According to GCV, the best λ is given by

$$\lambda^{\text{GCV}} = \underset{\lambda}{\text{argmin}} \frac{1}{n} \frac{\mathbf{Y}^T (\mathbf{I} - \mathbf{S}_\lambda)^2 \mathbf{Y}}{\left(1 - \frac{\text{tr} \mathbf{S}_\lambda}{n}\right)^2}$$

3.1 Introduce to penalized least-squares

This chapter introduces penalized least-squares approaches to variable selection problems in multiple regression models. They provide fundamental insights and basis for model selection problems in other more sophisticated models.

3.1.1 Classical Variable Selection Criteria

When the number of predictors p is larger than the sample size n , the model parameters in the linear model (3.1) are not identifiable. What makes them estimable is the **sparsity assumption** on the regression coefficients $\{\beta_j\}_{j=1}^p$: many of them are too small to matter, so they are ideally regarded as zero. Throughout this chapter, we assume the linear model (3.1):

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

unless otherwise stated.

Subset selection

Among all models with m variables, pick the one with the smallest **residual sum of squares**, which is denoted by RSS_m . This is indeed very intuitive: among the models with the same complexity, a better fit is preferable. This creates a sequence of submodels $\{\mathcal{M}_m\}_{m=0}^p$ indexed by the model size m . The choice of the model size m will be further illuminated in Section 3.1.3.

Computation of the best subset method is expensive even when p is moderately large. At each step, we compare the goodness-of-fit among $\binom{p}{m}$ models of size m and there are 2^p submodels in total. Intuitive and greedy algorithms have been introduced to produce a sequence of submodels with different numbers of variables. These include forward selection also called stepwise addition, backward elimination also named stepwise deletion, and stepwise regression.

In classical statistics, one does not produce the full sequence of the models in the forward selection and backward elimination methods. One often sets a very simple stopping criterion such as when all variables are statistically significant (e.g. P-values for each fitted coefficient is smaller than 0.05).

When p is larger than n , backward elimination cannot be applied since we cannot fit the full model. Yet, the forward selection can still be used to select a sequence of submodels. When $p < n$ or when p is relatively large compared to n , backward elimination cannot produce a stable selection process, but forward selection can as long as it is stopped early enough. These are the advantages of the forward selection algorithm.

Other greedy algorithms include matching pursuit (Mallot and Zhang, 1993), which picks the most correlated variable with the residuals from the previous step of fitting, also referred to as partial residuals, and runs the univariate regression to fit the partial residuals. See Section 3.5.11 for additional details.

Relation with penalized regression

Best subset selection can be regarded as penalized least-squares (PLS). Let $\|\beta\|_0$ be the L_0 -norm of the vector β , which counts the number of nonvanishing components of β . Consider the penalized least-squares with L_0 penalty:

$$\|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|_0 \quad (3.63)$$

The procedure is also referred to as complexity or entropy based PLS. Clearly, given the model size $\|\beta\|_0 = m$, the solution to the penalized least-squares (3.63) is the best subset selection. The computational complexity is NP-hard.

The stepwise algorithms in the last subsection can be regarded as greedy (approximation) algorithms for penalized least-squares (3.63).

Recall RSS_m is the smallest residual sum of squares among models with size m . With this definition, minimization of (3.63) can be written as

$$\text{RSS}_m + \lambda m \quad (3.64)$$

The optimal model size is obtained by minimizing (3.64) with respect to m . Clearly, the regularization parameter λ dictates the size of the model. The larger the λ , the larger the penalty on the model complexity m , the smaller the selected model.

Selection of regularization parameters

The best subset technique does not tell us the choice of model size m . The criterion used to compare two models is usually the **prediction error**. For a completely new observation (\mathbf{X}^{*T}, Y^*) , the prediction error of using model \mathcal{M}_m is

$$\text{PE}(\mathcal{M}_m) = \mathbb{E} \left(Y^* - \hat{\beta}_m^T \mathbf{X}_{\mathcal{M}_m}^* \right)^2$$

where $\hat{\beta}_m$ is the fitted regression coefficient vector with m variables, $\mathbf{X}_{\mathcal{M}_m}^*$ is the subvector of \mathbf{X}^* with the selected variables, and the expectation is taken only with respect to the new random variable (\mathbf{X}^{*T}, Y^*) .

An unbiased estimation of the prediction error $n\text{PE}(\mathcal{M}_m)$ was derived by Mallows (1973) (after ignoring a constant; see Section 3.6.1 for a derivation):

$$C_p(m) = \text{RSS}_m + 2\sigma^2 m \quad (3.65)$$

This corresponds to taking $\lambda = 2\sigma^2$ in the penalized least-squares problem (3.63) or (3.64). The parameter m is chosen to minimize (3.65), which is often referred to as Mallows's C_p criterion.

Akaike (1973, 1974) derived an approximately unbiased estimate of the prediction error (in terms of the Kullback-Leibler divergence) in a general likelihood based model. Translating his criterion into the least-squares setting, it becomes

$$\text{AIC}(m) = \log(\text{RSS}_m / n) + 2m / n$$

which is called the **Akaike information criterion (AIC)**. Note that when $\text{RSS}_m / n \approx \sigma^2$, which is correct when \mathcal{M}_m contains the true model, by Taylor's expansion,

$$\begin{aligned} \log(\text{RSS}_m / n) &= \log \sigma^2 + \log(1 + \text{RSS}_m / (n\sigma^2) - 1) \\ &\approx \log \sigma^2 + (\text{RSS}_m / (n\sigma^2) - 1) \end{aligned}$$

Therefore,

$$\text{AIC}(m) \approx [\text{RSS}_m + 2\sigma^2 m] / (n\sigma^2) + \log \sigma^2 - 1$$

which is approximately the same as the C_p criterion (3.65) after ignoring the affine transformation.

Many information criteria have been derived since the pioneering work of Akaike and Mallows. They correspond to different choices of λ in

$$\text{IC}(m) = \log(\text{RSS}_m / n) + \lambda m / n \quad (3.66)$$

Examples include

1. Bayesian information criterion (BIC, Schwarz, 1978) : $\lambda = \log(n)\sigma^2$

2. ϕ -criterion (Hannan and Quinn, 1979; Shibata, 1984): $\lambda = c(\log \log n)\sigma^2$
3. Risk inflation criterion (RIC, Foster and George, 1994): $\lambda = 2\log(p)\sigma^2$.

Using the Taylor expansion above, the information criterion (3.66) is asymptotically equivalent to (3.64). An advantage of using these information criteria over criterion (3.64) is that they do not need to estimate σ^2 . **The disadvantage of AIC: this also creates the bias issue. In particular, when a submodel contains modeling biases, AIC is no longer an approximately unbiased estimator.** The issue of model selection consistency has been thoroughly studied in Shao (1997).

In summary, the best subset method along with an information criterion corresponds to the L_0 -penalized least-squares with penalty parameters λ being a multiple $2, \log(n), c \log \log n$, and $2\log p$ of σ^2 , respectively for the AIC, BIC, ϕ -criterion, and RIC.

CV & GCV

Cross-validation (Allen 1974; Stone, 1974) is a novel and widely applicable idea for **estimating prediction error** of a model. It involves partitioning a sample of data into a training set used to estimate model parameters and a testing set reserved for validating the analysis of the fitted model. In k -fold cross-validation, the original sample is randomly partitioned into k approximately equal-sized subsamples with index sets $\{\mathcal{S}_j\}_{j=1}^k$. Of the k subsamples, a single subsample \mathcal{S}_k is retained as the validation set, and the remaining data $\{\mathcal{S}_j\}_{j \neq k}$ are used as a training set. The cross-validation process is then repeated k times, with each of the k subsamples used exactly once as the validation data. The prediction error of the k -fold cross-validation is computed as

$$CV_k(m) = n^{-1} \sum_{j=1}^k \left\{ \sum_{i \in \mathcal{S}_j} \left(Y_i - \hat{\beta}_{m, -\mathcal{S}_j}^T \mathbf{X}_{i, \mathcal{M}_m} \right)^2 \right\} \quad (3.67)$$

where $\hat{\beta}_{m, -\mathcal{S}_j}$ is the fitted coefficients of the submodel \mathcal{M}_m without using the data indexed in \mathcal{S}_j . The number of fittings is k , which is much smaller than n . In practice, the popular choice of k is 5 or 10. An interesting choice of k is n , which is called the leave-one-out cross-validation. The leave-one-out CV error of the submodel \mathcal{M}_m is

$$CV(m) = n^{-1} \sum_{i=1}^n \left(Y_i - \hat{\beta}_{m, -i}^T \mathbf{X}_{i, \mathcal{M}_m} \right)^2 \quad (3.68)$$

In general, the leave-one-out CV error is expensive to compute. For multiple linear regression and other linear smoothers with self-stable property, there is a neat formula for computing $CV(m)$ without fitting the model n times. See Theorem 3.0.7 in Section 2.8 of Chapter 2. Another simplification of $CV(m)$ is to use generalized cross-validation Generalized Cross-Validation (GCV, Craven and Wahba, 1979), defined by

$$GCV(m) = \frac{RSS_m}{n(1 - m/n)^2} \quad (3.69)$$

By using a simple Taylor expansion,

$$(1 - m/n)^{-2} = 1 + 2m/n + o(m/n)$$

and $\text{RSS}_m/n \approx \sigma^2$, one can easily see that $\text{GCV}(m)$ is approximately the same as $C_p(m)/n$. A classical choice of m is to maximize the adjusted multiple R^2 , defined by

$$R_{adj,m}^2 = 1 - \frac{n-1}{n-m} \frac{\text{RSS}_m}{\text{RSS}_0} \quad (3.70)$$

where RSS_0 is the sample standard deviation of the response variable $\{Y_i\}$. This is equivalent to minimizing $\text{RSS}_m/(n-m)$. Derived the same way as the GCV, it corresponds to approximately $\lambda = \sigma^2 \ln(3.64)$.

Folded-concave Penalized Least Squares

PLS's minimization problem is impossible to carry out when the dimensionality is high. A natural relaxation is to replace the discontinuous L_0 -penalty by more regular functions. This results in penalized least-squares

$$\begin{aligned} Q(\beta) &= \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \sum_{j=1}^p p_\lambda(|\beta_j|) \\ &\equiv \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \|p_\lambda(|\beta|)\|_1 \end{aligned} \quad (3.71)$$

where $p_\lambda(\cdot)$ is a penalty function in which the regularization or penalization parameters λ are the same for convenience of presentation.

A natural choice is $p_\lambda(\theta) = \lambda\theta^2/2$, whose solution is **ridge regression**

$$\hat{\beta}_{\text{ridge}} = (\mathbf{X}^T \mathbf{X} + n\lambda \mathbf{I}_p)^{-1} \mathbf{X}^T \mathbf{Y} \quad (3.72)$$

which is also called the Tikhonov regularization (Tikhonov, 1943).

Proposition 3.1.1 — Properties of Ridge regression. 1. The estimator shrinks all components toward zero, but none of them are actually zero.

2. It does not have a model selection property and creates biases for large parameters.

In order to reduce the bias, Frank and Friedman (1993) propose to use $p_\lambda(\theta) = \lambda|\theta|^q$ for $0 < q < 2$, called the **bridge regression**, which bridges the best subset selection (penalized L_0) and ridge regression (penalized L_2).

Donoho and Johnstone (1994), Tibshirani (1996) and Chen, Donoho and Sanders (1998) observe that penalized L_1 regression leads to a sparse minimizer and hence possesses a variable selection property. The procedure is called **Lasso** by Tibshirani (1996), for 'least absolute shrinkage and selection operator'. Unlike the complexity penalty $p_\lambda(|\theta|) = \lambda I(|\theta| \neq 0)$, Lasso solves a convex optimization problem. This gives the Lasso huge computational advantages.

L_1 -penalty differs substantially from L_0 -penalty. It penalizes the large parameters too much. To further reduce the bias in the estimation, Antoniadis and Fan (2001) and Fan and Li(2001) introduce **folded concave penalized least-squares**, in which $p_\lambda(\theta)$ is symmetric and concave on each side. In particular, the smoothly clipped absolute deviation (SCAD) penalty [see (3.76) below] is introduced to improve the bias property. **SCAD behaves like the L_1 -penalty at the origin in order to keep the variable selection property and acts like the L_0 -penalty at the tails in order to improve the bias property of the L_1 -penalty.** The smoothness of the penalty function is introduced to ensure the continuity of the solution for model stability.

Orthonormal designs

For an orthonormal design in which the design matrix multiplied by $n^{-1/2}$ is orthonormal (i.e., $\mathbf{X}^T \mathbf{X} = nI_p$, which implies $p \leq n$), (3.71) reduces to

$$\frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \frac{1}{2} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 + \|p_\lambda(|\boldsymbol{\beta}|)\|_1 \quad (3.73)$$

where $\hat{\boldsymbol{\beta}} = n^{-1} \mathbf{X}^T \mathbf{Y}$ is the ordinary least-squares estimate. Noticing that the first term is constant, minimizing (3.73) becomes minimizing

$$\sum_{j=1}^p \left\{ \frac{1}{2} (\hat{\beta}_j - \beta_j)^2 + p_\lambda(|\beta_j|) \right\}$$

which is a componentwise regression problem: each component consists of the univariate PLS problem of the form

$$\hat{\theta}(z) = \arg \min_{\theta} \left\{ \frac{1}{2} (z - \theta)^2 + p_\lambda(|\theta|) \right\} \quad (3.74)$$

Fan and Li (2001) advocate penalty functions that give estimators with the following three properties:

1. Sparsity: The resulting estimator automatically sets small estimated coefficients to zero to accomplish variable selection and reduce model complexity.
2. Unbiasedness: The resulting estimator is nearly unbiased, especially when the true coefficient β_j is large, to reduce model bias.
3. Continuity: The resulting estimator is continuous in the data to reduce instability in model prediction (Breiman, 1996).

The third property is nice to have, but not necessarily required. Let $p_\lambda(t)$ be nondecreasing and continuously differentiable on $[0, \infty)$. Assume that the function $-t - p'_\lambda(t)$ is strictly unimodal on $(0, \infty)$ with the convention $p'_\lambda(0) = p'_\lambda(0+)$. Antoniadis and Fan (2001) characterize the properties of $\hat{\theta}(z)$ as follows:

1. Sparsity if $\min_{t \geq 0} \{t + p'_\lambda(t)\} > 0$, which holds if $p'(0+) > 0$
2. Approximate unbiasedness if $p'_\lambda(t) = 0$ for large t (deri stands for change)

3. Continuity if and only if $\arg \min_{t \geq 0} \{t + p'_\lambda(t)\} = 0$

Note that Properties 1) and 2) require the penalty functions to be folded-concave (symmetric and concave on each side). Fan and Li (2001) advocate to use a family of folded-concave penalized likelihoods as a viable variable selection technique. They do not expect the form of the penalty functions to play a particularly important role provided that it satisfies Properties 1) -3). Antoniadis and Fan (2001) show further that $\hat{\theta}(z)$ is an anti-symmetric shrinkage function:

$$|\hat{\theta}(z)| \leq |z|, \quad \text{and} \quad \hat{\theta}(-z) = -\hat{\theta}(z) \quad (3.75)$$

The approximate unbiasedness requires $\theta(z)/z \rightarrow 1$, as $|z| \rightarrow \infty$

Penalty functions

From the above discussion, singularity at the origin (i.e., $p'_\lambda(0+) > 0$) is sufficient for generating sparsity in variable selection and the concavity is needed to reduce the estimation bias. This leads to a family of folded concave penalty functions with singularity at the origin. The L_1 penalty can be regarded as both a concave and convex function. It falls on the boundary of the family of the folded-concave penalty functions.

The L_q penalty with $q > 1$ is convex. It does not satisfy the sparsity condition, whereas L_1 penalty does not satisfy the unbiasedness condition. The L_q penalty with $0 \leq q < 1$ is concave but does not satisfy the continuity condition. In other words, none of the L_q penalties possesses all three aforementioned properties simultaneously. For this reason, Fan (1997) introduces the smoothly clipped absolute deviation (SCAD), whose derivative is given by

$$p'_\lambda(t) = \lambda \left\{ I(t \leq \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda) \right\} \quad \text{for some } a > 2 \quad (3.76)$$

where $p_\lambda(0) = 0$ and often $a = 3.7$ is used (suggested by a Bayesian argument in Fan and Li, 2001). Now this satisfies the aforementioned three properties. Note that when $a = \infty$, SCAD reduces to the L_1 -penalty.

In response to Fan (1997), Antoniadis (1997) proposes the penalty function

$$p_\lambda(t) = \frac{1}{2}\lambda^2 - \frac{1}{2}(\lambda - t)_+^2 \quad (3.77)$$

which results in the hard-thresholding estimator

$$\hat{\theta}_H(z) = zI(|z| > \lambda) \quad (3.78)$$

Fan and Li (2001) refer to this penalty function as the **hard thresholding penalty**, whose derivative function is $p'_\lambda(t)/2 = (\lambda - t)_+$.

An extension of this penalty function, derived by Zhang (2010) from a minimax point of view, is the minimax **concave penalty** (MCP), whose derivative is given by

$$p'_\lambda(t) = (\lambda - t/a)_+ \quad (3.79)$$

Note that the hard thresholding penalty corresponds to $a = 1$ and the MCP does not satisfy the continuity property. But this is not that important as noted before. Figure 3.2 depicts some of those commonly used penalty functions.

Thresholding by SCAD and MCP

We now look at the PLS estimator $\hat{\theta}(z)$ in (3.74) for some penalties. The entropy penalty (L_0 penalty) and the hard thresholding penalty (3.77) yield the hard thresholding rule (3.78) (Donoho and Johnstone, 1994) and the L_1 penalty gives the soft thresholding rule (Bickel, 1983 ; Donoho and Johnstone, 1994)

$$\hat{\theta}_{\text{soft}}(z) = \text{sgn}(z)(|z| - \lambda)_+ \quad (3.80)$$

The SCAD and MCP give rise to analytical solutions to (3.74), each of which is a linear spline in z . For the SCAD penalty, the solution is

$$\hat{\theta}_{\text{SCAD}}(z) = \begin{cases} \text{sgn}(z)(|z| - \lambda)_+, & \text{when } |z| \leq 2\lambda \\ \text{sgn}(z)[(a-1)|z| - a\lambda]/(a-2), & \text{when } 2\lambda < |z| \leq a\lambda \\ z, & \text{when } |z| \geq a\lambda \end{cases} \quad (3.81)$$

See Fan (1997) and Figure 3.1(b). Note that when $a = \infty$, the SCAD estimator becomes the soft-thresholding estimator (3.80). For the MCP with $a \geq 1$, the solution is

$$\hat{\theta}_{\text{MCP}}(z) = \begin{cases} \text{sgn}(z)(|z| - \lambda)_+/(1 - 1/a), & \text{when } |z| < a\lambda \\ z, & \text{when } |z| \geq a\lambda \end{cases} \quad (3.82)$$

It has discontinuity points at $|z| = \lambda$, which can create model instability. In particular, when $a = 1$, the solution is the hard thresholding function $\hat{\theta}_H(z)$ (3.78). When $a = \infty$, it also becomes a soft-thresholding estimator. In summary, SCAD and MCP are folded concave functions. They are generalizations of the soft-thresholding and hard-thresholding estimators. The former is continuous whereas the latter is discontinuous.

Risk properties

We now numerically compare the risk property of several commonly thresholded-shrinkage estimators under the fundamental model $Z \sim N(\theta, 1)$. Let

$$R(\theta) = E(\hat{\theta}(Z) - \theta)^2$$

be the risk function for the estimator $\hat{\theta}(Z)$. Figure 3.3 depicts $R(\theta)$ for some commonly used penalty functions. To make them comparable, we chose $\lambda = 1$ and 2 for the hard thresholding penalty, and for other penalty functions the values of λ are selected to make their risks at $\theta = 3$ the same as that of the hard thresholding estimator $\hat{\theta}_H(z)$.

The PLS estimators improve the ordinary least squares estimator Z in the region where θ is near zero, and have the same risk as the ordinary least squares estimator when

θ is far away from zero (e.g., 4 standard deviations away). An exception to this is the Lasso estimator. The Lasso estimator has a bias approximately of size λ for large θ , and this causes higher risk as shown in Figure 3.3. The better risk property at the origin is the payoff that we earn for exploring sparsity.

When $\lambda_{\text{hard}} = 2$, Lasso has higher risk than the SCAD estimator except in a small region. Lasso prefers smaller λ due to its bias. For $\lambda_{\text{hard}} = 1$, Lasso outperforms other methods near the origin. As a result, when λ is chosen automatically by data, Lasso has to choose a smaller λ in order to have a desired mean squared error (to reduce the modeling bias). Yet, a smaller value of λ yields a more complex model. This explains why Lasso tends to have many false positive variables in selected models.

Characterization of folded-concave PLS

Folded-concave penalized least-squares (3.71) is in general a non-convex function. It is challenging to characterize the global solution so let us first characterize its local minimizers.

From Lv and Fan (2009) and Zhang (2010), the local concavity of the penalty $p_\lambda(\cdot)$ at $\mathbf{v} = (v_1, \dots, v_q)^T$ is defined as

$$\kappa(p_\lambda; \mathbf{v}) = \lim_{\epsilon \rightarrow 0+} \max_{1 \leq j \leq q, t_1 < t_2 \in (|v_j| - \epsilon, |v_j| + \epsilon)} \frac{p'_\lambda(t_2) - p'_\lambda(t_1)}{t_2 - t_1} \quad (3.83)$$

By the concavity of p_λ on $[0, \infty)$, $\kappa(p_\lambda; \mathbf{v}) \geq 0$. It is easy to see by the meanvalue theorem that $\kappa(p_\lambda; \mathbf{v}) = \max_{1 \leq j \leq q} -p''_\lambda(|v_j|)$ when the second derivative of $p_\lambda(\cdot)$ is continuous. For the L_1 penalty, $\kappa(p_\lambda; \mathbf{v}) = 0$ for any \mathbf{v} . For the SCAD penalty, $\kappa(p_\lambda; \mathbf{v}) = 0$ unless some component of $|\mathbf{v}|$ takes values in $[\lambda, a\lambda]$. In the latter case, $\kappa(p_\lambda; \mathbf{v}) = (a - 1)^{-1} \lambda^{-1}$.

Let $\lambda_{\min}(\mathbf{A})$ be the minimum eigenvalue of a symmetric matrix \mathbf{A} and $\|\mathbf{a}\|_\infty = \max_j |a_j|$. Lv and Fan (2009) prove the following result. The gap between the necessary condition for local minimizer and sufficient condition for strict local minimizer is tiny (non-strict versus strict inequalities).

Theorem 3.1.2 — Characterization of PLSE. Assume that $p_\lambda(|\theta|)$ is folded concave. Then a necessary condition for $\hat{\boldsymbol{\beta}} \in R^p$ being a local minimizer of $Q(\boldsymbol{\beta})$ defined by (3.71) is

$$\begin{aligned} n^{-1} \mathbf{X}_1^T (\mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\beta}}) - p'_\lambda(|\hat{\boldsymbol{\beta}}_1|) \text{sgn}(\hat{\boldsymbol{\beta}}_1) &= \mathbf{0} \\ \left\| n^{-1} \mathbf{X}_2^T (\mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\beta}}) \right\|_\infty &\leq p'_\lambda(0+) \\ \lambda_{\min} \left(n^{-1} \mathbf{X}_1^T \mathbf{X}_1 \right) &\geq \kappa(p_\lambda; \hat{\boldsymbol{\beta}}_1) \end{aligned} \quad (3.84)$$

where \mathbf{X}_1 and \mathbf{X}_2 are respectively the submatrices of \mathbf{X} formed by columns indexed by $\text{supp}(\hat{\boldsymbol{\beta}})$ and its complement, and $\hat{\boldsymbol{\beta}}_1$ is a vector of all non-vanishing components $\hat{\boldsymbol{\beta}}$. On the other hand, if (3.84) hold with inequalities replaced by strict inequalities, then

$\hat{\beta}$ is a strict local minimizer of $Q(\beta)$.

Conditions (3.84) can be regarded as the **Karush-Kuhn-Tucker conditions**. They can also be derived by using subgradient calculus. The first and third conditions (3.84) are respectively the first and second order conditions for $(\hat{\beta}_1, \mathbf{0})$ to be the local minimizer of $Q(\beta_1, \mathbf{0})$, the local minimizer on the restricted coordinate subspace. The second condition (3.84) guarantees the local minimizer on the restricted coordinate subspace is also the local minimizer of the whole space R^p .

When $Q(\beta)$ is strictly convex, there exists at most one local minimizer. In this case, the local minimizer is also the unique global minimizer. For a folded concave penalty function, let $\kappa(p_\lambda)$ be the maximum concavity of the penalty function p_λ defined by

$$\kappa(p_\lambda) = \sup_{t_1 < t_2 \in (0, \infty)} -\frac{p'_\lambda(t_2) - p'_\lambda(t_1)}{t_2 - t_1} \quad (3.85)$$

For the L_1 penalty, SCAD and MCP, we have $\kappa(p_\lambda) = 0, (a-1)^{-1}$, and a^{-1} respectively. Thus, the maximum concavity of SCAD and MCP is small when a is large. When

$$\lambda_{\min}(n^{-1}\mathbf{X}^T\mathbf{X}) > \kappa(p_\lambda) \quad (3.86)$$

the function $Q(\beta)$ is strictly convex, as the convexity of the quadratic loss dominates the maximum concavity of the penalty in (3.71). Hence, the global minimum is unique. Note that condition (3.86) requires $p \leq n$

In general, the global minimizer of the folded-concave penalized least-squares is hard to characterize. Fan and Lv (2011) are able to give conditions under which a solution is global optimal on the union of all m -dimensional coordinate subspaces:

$$S_m = \{\beta \in R^p : \|\beta\|_0 \leq m\} \quad (3.87)$$

Lasso and L_1 Regularization

Lasso gains its popularity due to its convexity and computational expedience. The predecessor of Lasso is the negative garrote. The study of Lasso also leads to the Dantzig selector, the adaptive Lasso and the elastic net. This section touches on the basis of these estimators in which the L_1 -norm regularization plays a central role.

Nonnegative garrote

The nonnegative garrote estimator, introduced by Breiman (1995), is the first modern statistical method that uses the L_1 -norm regularization to do variable selection in multiple linear regression. Consider the usual setting with $p < n$ and let $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)^T$ be the OLS estimator. When $p > n$, $\hat{\beta}$ can be the ridge regression estimator (Yuan and Lin, 2005), the main idea stays the same. Then, the fitted model becomes

$$\hat{Y} = \hat{\beta}_1 X_1 + \dots + \hat{\beta}_p X_p$$

The above model uses all variables. To do variable selection, we introduce the **nonnegative shrinkage parameter** $\theta = (\theta_1, \dots, \theta_p)^T$ and regard $\hat{\beta}$ as fixed, a new fitted model becomes

$$\hat{Y} = \theta_1 Z_1 + \dots + \theta_p Z_p, \quad Z_j = \hat{\beta}_j X_j$$

If θ_j is zero, then variable X_j is excluded from the fitted model. The nonnegative garrote estimates θ via the following L_1 regularized least squares:

$$\min_{\theta \geq 0} \frac{1}{2n} \|\mathbf{Y} - \mathbf{Z}\theta\|^2 + \lambda \sum_j \theta_j \quad (3.88)$$

Note that under the nonnegative constraints $\sum_{j=1}^p \theta_j = \|\theta\|_1$. By varying λ , the nonnegative garrote automatically achieves model selection. Many components of the minimizer of (3.88), $\hat{\theta}$, will be zero. This can be easily seen when \mathbf{X} is scaled orthonormal $\mathbf{X}^T \mathbf{X} = n\mathbf{I}_p$, as in Section 3.2.1. In this case, $\{\mathbf{Z}_j\}$ are still orthogonal and $\mathbf{Z}^T \mathbf{Z}$ is diagonal. The ordinary least-squares estimator is given by

$$\hat{\theta}_0 = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{Y} = (\mathbf{Z}_1 / \|\mathbf{Z}_1\|^2, \dots, \mathbf{Z}_p / \|\mathbf{Z}_p\|^2)^T \mathbf{Y}$$

Note that by the orthogonality of the least-squares fit to its residuals,

$$\begin{aligned} \|\mathbf{Y} - \mathbf{Z}\theta\|^2 &= \|\mathbf{Y} - \mathbf{Z}\hat{\theta}_0\|^2 + \|\mathbf{Z}(\theta - \hat{\theta}_0)\|^2 \\ &= \|\mathbf{Y} - \mathbf{Z}\hat{\theta}_0\|^2 + \sum_{j=1}^p \|\mathbf{Z}_j\|^2 (\theta_j - \hat{\theta}_{j0})^2 \end{aligned}$$

where $\hat{\theta}_{j0} = \mathbf{Z}_j^T \mathbf{Y} / \|\mathbf{Z}_j\|^2$ is the j^{th} component of θ_0 and the last equality utilizes the orthogonality of \mathbf{Z}_j . Therefore, problem (3.88) becomes

$$\min_{\theta \geq 0} \frac{1}{2} \sum_{j=1}^p \|\mathbf{Z}_j\|^2 (\theta_j - \hat{\theta}_{j0})^2 + \lambda \sum_{j=1}^p \theta_j$$

This reduces to the componentwise minimization problem

$$\min_{\theta \geq 0} \frac{1}{2} (\theta - \theta_0)^2 + \lambda \theta$$

whose minimizer is clearly $\hat{\theta} = (\theta_0 - \lambda)_+$ by taking the first derivative and setting it to zero. Applying this to our scenario and noticing $\|\mathbf{Z}_j\|^2 = n\hat{\beta}_j^2$ and $\hat{\theta}_{j0} = n^{-1} \mathbf{X}_j^T \mathbf{Y} / \hat{\beta}_j$ we have

$$\hat{\theta}_j = \left(\hat{\theta}_{j0} - \frac{\lambda}{\|\mathbf{Z}_j\|^2} \right)_+ = \left(\frac{\mathbf{X}_j^T \mathbf{Y}}{n\hat{\beta}_j} - \frac{\lambda}{n\hat{\beta}_j^2} \right)_+$$

In particular, if $\hat{\beta} = n^{-1} \mathbf{X}^T \mathbf{Y}$ is the ordinary least-squares estimate, then

$$\hat{\theta}_j = \left(1 - \frac{\lambda}{n\hat{\beta}_j^2} \right)_+$$

Model selection of the negative garrote now becomes clear. When $|\hat{\beta}_j| \leq \sqrt{\lambda/n}$, it is shrunk to zero. The larger the original estimate, the smaller the shrinkage. Furthermore, the shrinkage rule is continuous. This is in the same spirit as the folded concave PLS such as SCAD introduced in the last section.

Lasso

Lasso, the term coined by Tibshirani (1996), estimates the sparse regression coefficient vector β by minimizing

$$\frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|_1 \quad (3.89)$$

This corresponds to (3.71) by taking $p_\lambda(\theta) = \lambda\theta$ for $\theta \geq 0$. Comparing (3.89) and (3.88), we see that the Lasso does not need to use a preliminary estimator of β , although both use the L_1 -norm to achieve variable selection.

Irrepresentable condition

The KKT conditions (3.84) now become

$$n^{-1} \mathbf{X}_1^T (\mathbf{Y} - \mathbf{X}_1 \hat{\beta}_1) - \lambda \operatorname{sgn}(\hat{\beta}_1) = \mathbf{0} \quad (3.90)$$

and

$$\left\| (n\lambda)^{-1} \mathbf{X}_2^T (\mathbf{Y} - \mathbf{X}_1 \hat{\beta}_1) \right\|_\infty \leq 1 \quad (3.91)$$

Since the third condition (3.84) is satisfied automatically. This first condition says that the signs of nonzero components of Lasso are the same as the correlations of the covariates with the current residual. The equations (3.90) and (3.91) imply that

$$\left\| n^{-1} \mathbf{X}^T (\mathbf{Y} - \mathbf{X} \hat{\beta}) \right\|_\infty \leq \lambda \quad (3.92)$$

Note that condition (3.91) holds for $\hat{\beta} = \mathbf{0}$ when

$$\lambda > \left\| n^{-1} \mathbf{X}^T \mathbf{Y} \right\|_\infty \quad (3.93)$$

Since the condition is imposed with a strict inequality, it is a sufficient condition (Theorem 3.1.2). In other words, when $\lambda > \left\| n^{-1} \mathbf{X}^T \mathbf{Y} \right\|_\infty$, $\hat{\beta} = \mathbf{0}$ is the unique solution and hence Lasso selects no variables. Therefore, we need only to consider λ in the interval $[0, \left\| n^{-1} \mathbf{X}^T \mathbf{Y} \right\|_\infty]$.

We now look at the model selection consistency of Lasso. Assuming the invertibility of $\mathbf{X}_1^T \mathbf{X}_1$, solving equation (3.90) gives

$$\hat{\beta}_1 = \left(\mathbf{X}_1^T \mathbf{X}_1 \right)^{-1} \left(\mathbf{X}_1^T \mathbf{Y} - n\lambda \operatorname{sgn}(\hat{\beta}_1) \right) \quad (3.94)$$

and substituting this into equation (3.91) yields

$$\left\| (n\lambda)^{-1} \mathbf{X}_2^T (\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_1}) \mathbf{Y} + \mathbf{X}_2^T \mathbf{X}_1 (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \text{sgn}(\hat{\boldsymbol{\beta}}_1) \right\|_{\infty} \leq 1 \quad (3.95)$$

where $\mathbf{P}_{\mathbf{X}_1} = \mathbf{X}_1 (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T$ is the projection matrix onto the linear space spanned by the columns of \mathbf{X}_1 . For the true parameter, let $\text{supp}(\boldsymbol{\beta}_0) = \mathcal{S}_0$ so that

$$\mathbf{Y} = \mathbf{X}_{\mathcal{S}_0} \boldsymbol{\beta}_0 + \boldsymbol{\varepsilon} \quad (3.96)$$

If $\text{supp}(\hat{\boldsymbol{\beta}}) = \mathcal{S}_0$, i.e., model selection consistency holds, then $\mathbf{X}_{\mathcal{S}_0} = \mathbf{X}_1$ and $(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_1}) \mathbf{X}_{\mathcal{S}_0} = \mathbf{0}$. By substituting (3.96) into (3.95), we have

$$\left\| (n\lambda)^{-1} \mathbf{X}_2^T (\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_1}) \boldsymbol{\varepsilon} + \mathbf{X}_2^T \mathbf{X}_1 (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \text{sgn}(\hat{\boldsymbol{\beta}}_1) \right\|_{\infty} \leq 1 \quad (3.97)$$

Note that this condition is also sufficient if the inequality is replaced with the strict one (see Theorem 3.1.2).

Typically, the first term in (3.97) is negligible. This will be formally shown in Chapter 4. A specific example is the case $\boldsymbol{\varepsilon} = \mathbf{0}$ as in the compressed sensing problem. In this case, condition (3.97) becomes

$$\left\| \mathbf{X}_2^T \mathbf{X}_1 (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \text{sgn}(\hat{\boldsymbol{\beta}}_1) \right\|_{\infty} \leq 1$$

This condition involves $\text{sgn}(\hat{\boldsymbol{\beta}}_1)$. If we require a stronger consistency $\text{sgn}(\hat{\boldsymbol{\beta}}) = \text{sgn}(\boldsymbol{\beta}_0)$, called the **sign consistency**, then the condition becomes

$$\left\| \mathbf{X}_2^T \mathbf{X}_1 (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \text{sgn}(\boldsymbol{\beta}_{\mathcal{S}_0}) \right\|_{\infty} \leq 1 \quad (3.98)$$

The above condition does not depend on λ . It appeared in Zou (2006) and Zhao and Yu (2006) who coined the name the **irrepresentable condition**.

Note that $(\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T \mathbf{X}_2$ in (3.98) is the matrix of the regression coefficients of each 'unimportant' variable X_j ($j \notin \mathcal{S}_0$) regressed on the important variables $\mathbf{X}_1 = \mathbf{X}_{\mathcal{S}_0}$. The irrepresentable condition is a condition on how strongly the important and unimportant variables can be correlated. Condition (3.98) states that the sum of the signed regression coefficients of each unimportant variable X_j for $j \notin \mathcal{S}_0$ on the important variables $\mathbf{X}_{\mathcal{S}_0}$ cannot exceed 1. The more the unimportant variables, the harder the condition is to meet. The irrepresentable condition is in general very restrictive.

Using the regression intuition, one can easily construct an example when it fails. For example, if an unimportant variable is generated by

$$X_j = \rho s^{-1/2} \sum_{k \in \mathcal{S}_0} \text{sgn}(\beta_k) X_k + \sqrt{1 - \rho^2} \varepsilon_k, \quad s = |\mathcal{S}_0|$$

for some given $|\rho| \leq 1$ (all normalization is to make $\text{Var}(X_j) = 1$), where all other random variables are independent and standardized, then the L_1 norm of the signed regression coefficients of this variable is $|\rho|s^{1/2}$, which can easily exceed 1. The larger the 'important variable' set \mathcal{S}_0 , the easier the irrepresentable condition fails. In addition, we need only one such unimportant predictor that has such a non-negligible correlation with important variables to make the condition fail. See also Corollary 1 in Zou(2006) for a counterexample.

The moral of the above story is that Lasso can have sign consistency, but this happens only in very specific cases. The irrepresentable condition (3.98) is independent of λ . When it fails, Lasso does not have sign consistency and this cannot be rescued by using a different value of λ .

Risk property of Lasso

We now look at the risk property of Lasso. It is easier to explain it under the constrained form:

$$\min_{\|\beta\|_1 \leq c} \|\mathbf{Y} - \mathbf{X}\beta\|^2 \quad (3.99)$$

for some constant c , as in Tibshirani (1996). Define the theoretical risk and empirical risk respectively as

$$R(\beta) = E \left(Y - \mathbf{X}^T \beta \right)^2 \quad \text{and} \quad R_n(\beta) = n^{-1} \sum_{i=1}^n \left(Y_i - \mathbf{X}_i^T \beta \right)^2$$

which are prediction errors using the parameter β . The best prediction error is $R(\beta_0)$. Note that

$$R(\beta) = \gamma^T \Sigma^* \gamma \quad \text{and} \quad R_n(\beta) = \gamma^T \mathbf{S}_n^* \gamma$$

where $\gamma = (-1, \beta^T)^T$, $\Sigma^* = \text{Var} \left((Y, \mathbf{X}^T)^T \right)$, and \mathbf{S}_n^* is the sample covariance matrix based on the data $\left\{ (Y_i, \mathbf{X}_i^T)^T \right\}_{i=1}^n$. Thus, for any β , we have the following risk approximation:

$$\begin{aligned} |R(\beta) - R_n(\beta)| &= \left| \gamma^T (\Sigma^* - \mathbf{S}_n^*) \gamma \right| \\ &\leq \|\Sigma^* - \mathbf{S}_n^*\|_\infty \|\gamma\|_1^2 \\ &= (1 + \|\beta\|_1)^2 \|\Sigma^* - \mathbf{S}_n^*\|_\infty \end{aligned} \quad (3.100)$$

On the other hand, if the true parameter β_0 is in the feasible set, namely, $\|\beta_0\|_1 \leq c$, then $R_n(\hat{\beta}) - R_n(\beta_0) \leq 0$. Using this,

$$0 \leq R(\hat{\beta}) - R(\beta_0) \leq \left\{ R(\hat{\beta}) - R_n(\hat{\beta}) \right\} + \left\{ R_n(\beta_0) - R(\beta_0) \right\} \quad (3.101)$$

By (3.100) along with $\|\hat{\beta}\|_1 \leq c$ and $\|\beta_0\|_1 \leq c$, we conclude that

$$\left| R(\hat{\beta}) - R(\beta_0) \right| \leq 2(1 + c)^2 \|\Sigma^* - \mathbf{S}_n^*\|_\infty$$

When $\|\Sigma^* - \mathbf{S}_n^*\|_\infty \rightarrow 0$, the risk converges. Such a property is called **persistence** by Greenshtein and Ritov (2004). Further details on the rates of convergence for estimating large covariance matrices can be found in Chapter 11. The rate is of order $O(\sqrt{(\log p)/n})$ for the data with Gaussian tails. The above discussion also reveals the relationship between covariance matrix estimation and sparse regression. A robust covariance matrix estimation can also reveal a robust sparse regression.

Persistence requires that the risk based on $\hat{\beta}$ is approximately the same as that of the optimal parameter β_0 , i.e.,

$$R(\hat{\beta}) - R(\beta_0) = o_P(1)$$

By (3.101), this requires only β_0 sparse in the sense that $\|\beta_0\|_1$ does not grow too quickly (recalling $\|\beta_0\|_1 \leq c$) and the large covariance matrix Σ^* can be uniformly consistently estimated. For data with Gaussian tails, since $\|\Sigma^* - \mathbf{S}_n^*\|_\infty = O_P(\sqrt{(\log p)/n})$ (see Chapter 11), we require

$$\|\beta_0\|_1 \leq c = o\left((n/\log p)^{1/4}\right)$$

for Lasso to possess persistence. Furthermore, the result (3.101) does not require to have a true underlying linear model. As long as we define

$$\beta_0 = \operatorname{argmin}_{\|\beta\|_1 \leq c} R(\beta)$$

the risk approximation inequality (3.101) holds by using the same argument above. In conclusion, Lasso has a good risk property when β_0 is sufficiently sparse.

Adaptive Lasso

The irrepresentable condition indicates restrictions on the use of the Lasso as a model/variable selection method. Another drawback of the Lasso is its lack of unbiasedness for large coefficients, as explained in Fan and Li (2001). This can be seen from (3.94). Even when the signal is strong so that $\operatorname{supp}(\hat{\beta}) = \mathcal{S}_0$, by substituting (3.96) into (3.94), we have

$$\hat{\beta}_1 = \beta_0 + \left(\mathbf{X}_1^T \mathbf{X}_1\right)^{-1} \mathbf{X}_1^T \varepsilon - n\lambda \left(\mathbf{X}_1^T \mathbf{X}_1\right)^{-1} \operatorname{sgn}(\hat{\beta}_1)$$

The last term is the bias due to the L_1 penalty. Unless λ goes to 0 sufficiently fast, the bias term is not negligible. However, $\lambda \approx \frac{1}{\sqrt{n}}$ is needed in order to make the Lasso estimate root- n consistent under the fixed p large n setting. For $p \gg n$, the Lasso estimator uses $\lambda \approx \sqrt{\log(p)/n}$ to achieve the optimal rate $\sqrt{|\mathcal{S}_0| \log(p)/n}$. See Chapter 4 for more details. So now, it is clear that the optimal Lasso estimator has non-negligible biases.

Is there a nice fix to these two problems? Zou (2006) proposes to use the adaptively weighted L_1 penalty (a.k.a. adaptive Lasso) to replace the L_1 penalty in penalized linear

regression and penalized generalized linear models. With the weighted L_1 penalty, (3.71) becomes

$$\frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^p w_j |\beta_j| \quad (3.102)$$

To keep the convexity property of the Lasso, w_j should be nonnegative. It is important to note that if the weights are deterministic, then they cannot fix the aforementioned two problems of the Lasso. Suppose that some deterministic weights can make the Lasso gain sign consistency. Then no w_j should be zero, otherwise the variable X_j is always included, which will violate the sign consistency of the Lasso if the underlying model does not include X_j , i.e. X_j is not an important variable. Hence, all w_j s are positive. Then we redefine the regressors as $X_j^w = X_j/w_j$ and $\theta_j = w_j\beta_j, 1 \leq j \leq p$. The underlying regression model can be rewritten as

$$Y = \sum_{j=1}^p X_j^w \theta_j + \epsilon$$

and (3.102) becomes

$$\frac{1}{2n} \|\mathbf{Y} - \mathbf{X}^w \boldsymbol{\theta}\|^2 + \lambda \sum_{j=1}^p |\theta_j|$$

Its corresponding irrepresentable condition is

$$\left\| (\mathbf{X}_2^w)^T \mathbf{X}_1^w \left[(\mathbf{X}_1^w)^T \mathbf{X}_1^w \right]^{-1} \text{sgn}(\boldsymbol{\beta}_{S_0}) \right\|_{\infty} \leq 1$$

We write $\mathbf{W} = (w_1, \dots, w_p)^T = (\mathbf{W}_1, \mathbf{W}_2)^T$. and express the irrepresentable conditions using the original variables, we have

$$\left\| \left[\mathbf{X}_2^T \mathbf{X}_1 \left(\mathbf{X}_1^T \mathbf{X}_1 \right)^{-1} \mathbf{W}_1 \circ \text{sgn}(\boldsymbol{\beta}_{S_0}) \right] \circ \mathbf{W}_2^{-1} \right\| \leq 1$$

Observe that if $\max \mathbf{W}_1 / \inf \mathbf{W}_2 \rightarrow 0$, then this representable condition can be satisfied for general $\mathbf{X}_1, \mathbf{X}_2$ and $\text{sgn}(\boldsymbol{\beta}_{S_0})$. This condition can only be achieved using a data-driven scheme, as we do not know the set S_0 .

Zou (2006) proposes to use a preliminary estimate $\hat{\beta}_j$ to construct w_j . For example, $w_j = |\hat{\beta}_j|^{-\gamma}$ for some $\gamma > 0$, and $\gamma = 0.5, 1$ or 2 . In the case of fixed p large n , the preliminary estimate can be the least-squares estimate. When $p \gg n$, the preliminary estimate can be the lasso estimate and $w_j = p'_\lambda \left(|\hat{\beta}_j^{\text{lasso}}| \right) / \lambda$ with a folded concave penalty $p_\lambda(\cdot)$

As will be seen in Section 3.5.5, the adaptive lasso is connected to the penalized least-squares estimator (3.71) via the local linear approximation with $p'_\lambda(\theta) = \lambda\theta^{-\gamma}$ or $L_{1-\gamma}$ penalty. Since the derivative function is decreasing, the spirit of the adaptive Lasso is the same as the folded-concave PLS. Hence, the adaptive Lasso is able to fix the bias caused by the L_1 penalty. In particular, the adaptive lasso estimator for $\boldsymbol{\beta}_{S_0}$ shares the asymptotical normality property of the oracle OLS estimator for $\boldsymbol{\beta}_{S_0}$, i.e.. $\hat{\boldsymbol{\beta}}_{S_0}^{\text{oracle}} = (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T \mathbf{Y}$

Elastic Net

In the early 2000 s, the Lasso was applied to regression with microarrays to do gene selection. The results were concerning because of high variability. This is mainly caused by the spurious correlation in high-dimensional data, as illustrated in Section 1.3.3 of Chapter 1. How to handle the strong (empirical) correlations among high-dimensional variables while keeping the continuous shrinkage and selection property of the Lasso? Zou and Hastie (2005) propose the **elastic net regularization** that uses a convex combination of L_1 and L_2 penalties. For the penalized least squares, the elastic net estimator is defined as

$$\arg \min_{\beta} \left\{ \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda_2 \|\beta\|^2 + \lambda_1 \|\beta\|_1 \right\} \quad (3.103)$$

where $p_{\lambda_1, \lambda_2}(t) = \lambda_1 |t| + \lambda_2 t^2$ is called the elastic net penalty. Another form of the elastic net penalty is

$$p_{\lambda, \alpha}(t) = \lambda J(t) = \lambda [(1 - \alpha)t^2 + \alpha |t|]$$

with $\lambda = \lambda_1 + \lambda_2$ and $\alpha = \frac{\lambda_1}{\lambda_1 + \lambda_2}$. The elastic net is a pure ridge regression when $\alpha = 0$ and a pure Lasso when $\alpha = 1$. The advantage of using (λ, α) parametrization is that α has a natural range $[0, 1]$. In practice, we can use CV to choose α over a grid such as $0.1k, k = 1, \dots, 10$. For the penalized least squares problem, using (λ_2, λ_1) parametrization is interesting because it can be shown that for a fixed λ_2 the solution path is piecewise linear with respect to λ_1 . Zou and Hastie (2005) exploit this property to derive an efficient pathfollowing algorithm named LARS-EN for computing the entire solution path of the Elastic Net penalized least squares (for each fixed λ_2). It is well known that L_2 regularization gracefully handles collinearity and achieves a good bias-variance trade-off for prediction. The Elastic Net inherits the ability to handle collinearity from its L_2 component and keeps the sparsity property of the Lasso through its L_1 component. With high-dimensional data the Elastic Net often generates a more accurate predictive model than the Lasso. Zou and Hastie (2005) also reveal the **group effect** of the Elastic Net in the sense that highly correlated variables tend to enter or exit the model together, while the Lasso tends to randomly pick one variable and ignore the rest.

To help visualize the fundamental differences between Lasso and Elastic Net, let us consider a synthetic model as follows. Let Z_1 and Z_2 be two independent $\text{unif}(0, 20)$ variables. Response \mathbf{Y} is generated by $\mathbf{Y} = Z_1 + 0.1 \cdot Z_2 + \epsilon$ with $\epsilon \sim N(0, 1)$ and the observed regressors are generated by

$$\begin{aligned} X_1 &= Z_1 + \epsilon_1, & X_2 &= -Z_1 + \epsilon_2, & X_3 &= Z_1 + \epsilon_3 \\ X_4 &= Z_2 + \epsilon_4, & X_5 &= -Z_2 + \epsilon_5, & X_6 &= Z_2 + \epsilon_6 \end{aligned}$$

where ϵ_i are iid $N(0, \frac{1}{16})$. X_1, X_2, X_3 form a group whose underlying factor is Z_1 , and X_4, X_5, X_6 form the other group whose underlying factor is Z_2 . The within group correlations are almost 1 and the between group correlations are almost 0. Ideally, we would want to only identify the Z_1 group (X_1, X_2, X_3) as the important variables. We generated two independent datasets with sample size 100 from this model. The two Lasso solution paths are very different, suggesting the high instability of the Lasso under strong correlations. On the other hand, the two Elastic Net solution paths are almost identical. Moreover, the Elastic Net identifies the corrected variables. The Elastic Net relies on its L_1 component for sparsity and variable selection. Similar to the Lasso case, the Elastic Net also requires a restrictive condition on the design matrix for selection consistency (Jia and Yu, 2010). To bypass this restriction, Zou and Zhang (2009) follow the adaptive Lasso idea and introduce the adaptive Elastic Net penalty $p(|\beta_j|) = \lambda_1 w_j |\beta_j| + \lambda_2 |\beta_j|^2$ where $w_j = |\hat{\beta}^{\text{enet}}_j + 1/n|^{-\gamma}$. The numeric studies therein shows the very competitive performance of the adaptive Elastic Net in terms of variable selection and model estimation.

Dantzig selector

The Dantzig selector, introduced by Candès and Tao (2007), is a novel idea of **casting the regularization problem into a linear program**. Recall that Lasso satisfies (3.92), but it might not have the smallest L_1 norm. One can find the estimator to minimize its L_1 -norm:

$$\min_{\beta \in R^p} \|\beta\|_1, \quad \text{subject to} \quad \left\| n^{-1} \mathbf{X}^T (\mathbf{Y} - \mathbf{X}\beta) \right\|_{\infty} \leq \lambda \quad (3.104)$$

The target function and constraints in (3.104) are linear. The problem can be formulated as a linear program by expressing it as

$$\min_{\mathbf{u}} \sum_{i=1}^p u_i, \quad \mathbf{u} \geq 0, \quad -\mathbf{u} \leq \beta \leq \mathbf{u}, \quad -\lambda \mathbf{1} \leq n^{-1} \mathbf{X}^T (\mathbf{Y} - \mathbf{X}\beta) \leq \lambda \mathbf{1}$$

The name "Dantzig selector" was coined by Emmanuel Candès and Terence Tao to pay tribute to George Dantzig, the father of linear programming who passed away while their manuscript was finalized.

Let $\hat{\beta}_{\text{DZ}}$ be the solution. A necessary condition for $\hat{\beta}_{\text{DZ}}$ to have model selection consistency is that β_0 is in the feasible set of (3.104), with probability tending to one. Using model (3.96), this implies that $\lambda \geq n^{-1} \|\mathbf{X}^T \epsilon\|_{\infty}$. For example, in the case when $\epsilon \sim N(0, \sigma^2 I_n)$ and columns $\{\mathbf{X}_j\}_{j=1}^p$ of \mathbf{X} are standardized so that $n^{-1} \|\mathbf{X}_j\|^2 = 1$, then $\mathbf{X}_j^T \epsilon \sim N(0, \sigma^2/n)$. Then, it can easily be shown (see Section 3.3.7) that it suffices to take λ as $\sigma \sqrt{2(1+\delta)n^{-1} \log p}$ for any $\delta \geq 0$, by using the union bound and the tail probability of normal distribution.

The Dantzig selector opens a new chapter for sparse regularization. Since the value

λ is chosen so that the true parameter β_0 falls in the constraint:

$$P \left\{ \left\| n^{-1} \mathbf{X}^T (\mathbf{Y} - \mathbf{X} \beta_0) \right\|_{\infty} \leq \lambda \right\} \rightarrow 1 \quad (3.105)$$

Fan (2014) interprets the set $\{\beta : \|n^{-1} \mathbf{X}^T (\mathbf{Y} - \mathbf{X} \beta)\|_{\infty} \leq \lambda\}$ as the high confidence set and this high confidence set summarizes the information on the parameter β_0 provided by the data. He argues that this set is too big to be useful in high-dimensional spaces and that we need some additional prior about β_0 . If the prior is that β_0 is sparse, one naturally combines these two pieces of information. This leads to finding the sparsest solution in high confidence set as a natural solution to the sparse regulation. This idea applies to quasi-likelihood based models and includes the Dantzig selector as a specific case. See Fan (2014) for details. The idea is reproduced by Fan, Han, and Liu (2014).

To see how norm-minimization plays a role, let us assume that β_0 is in the feasible set by taking a large enough value λ , i.e., $\lambda \geq n^{-1} \|\mathbf{X}^T \varepsilon\|_{\infty}$ as noted above. This is usually achieved by a probabilistic statement. Let $\hat{\Delta} = \hat{\beta}_{\text{DZ}} - \beta_0$. From the norm minimization, we have

$$\|\beta_0\|_1 \geq \|\hat{\beta}_{\text{DZ}}\|_1 = \|\beta_0 + \hat{\Delta}\|_1 \quad (3.106)$$

Noticing $\mathcal{S}_0 = \text{supp}(\beta_0)$, we have

$$\begin{aligned} \|\beta_0 + \hat{\Delta}\|_1 &= \left\| (\beta_0 + \hat{\Delta})_{\mathcal{S}_0} \right\|_1 + \left\| (\mathbf{0} + \hat{\Delta})_{\mathcal{S}_0^c} \right\|_1 \\ &\geq \|\beta_0\|_1 - \|\hat{\Delta}_{\mathcal{S}_0}\|_1 + \|\hat{\Delta}_{\mathcal{S}_0^c}\|_1 \end{aligned} \quad (3.107)$$

This together with (3.106) entails that

$$\|\hat{\Delta}_{\mathcal{S}_0}\|_1 \geq \|\hat{\Delta}_{\mathcal{S}_0^c}\|_1 \quad (3.108)$$

or that $\hat{\Delta}$ is sparse (the L_1 -norm of $\hat{\Delta}$ on a much bigger set is controlled by that on a much smaller set) or ‘restricted’. For example, with $s = |\mathcal{S}_0|$,

$$\|\hat{\Delta}\|_2 \geq \|\hat{\Delta}_{\mathcal{S}_0}\|_2 \geq \|\hat{\Delta}_{\mathcal{S}_0}\|_1 / \sqrt{s} \geq \|\hat{\Delta}\|_1 / (2\sqrt{s}) \quad (3.109)$$

where the last inequality utilizes (3.108). At the same time, since $\hat{\beta}$ and β_0 are in the feasible set (3.104), we have $\|n^{-1} \mathbf{X}^T \mathbf{X} \hat{\Delta}\|_{\infty} \leq 2\lambda$, which implies further that

$$\|\mathbf{X} \hat{\Delta}\|_2^2 = \hat{\Delta}^T (\mathbf{X}^T \mathbf{X} \hat{\Delta}) \leq \|\mathbf{X}^T \mathbf{X} \hat{\Delta}\|_{\infty} \|\hat{\Delta}\|_1 \leq 2n\lambda \|\hat{\Delta}\|_1$$

Using (3.109), we have

$$\|\mathbf{X} \hat{\Delta}\|_2^2 \leq 4n\lambda \sqrt{s} \|\hat{\Delta}\|_2 \quad (3.110)$$

The regularity condition on \mathbf{X} such as the restricted eigenvalue condition (Bickel, Ritov and Tsybakov, 2009)

$$\min_{\|\hat{\Delta}_{S_0}\|_1 \geq \|\hat{\Delta}_{S_0^c}\|_1} n^{-1} \|\mathbf{X}\hat{\Delta}\|_2^2 / \|\hat{\Delta}\|_2^2 \geq a$$

implies a convergence in L_2 . Indeed, from (3.110), we have

$$a \|\hat{\Delta}\|_2^2 \leq 4\lambda\sqrt{s} \|\hat{\Delta}\|_2, \quad \text{or} \quad \|\hat{\Delta}\|_2^2 \leq 16a^{-2}\lambda^2 s$$

which is of order $O(sn^{-1} \log p)$ by choosing the smallest feasible $\lambda = O(\sqrt{2n^{-1} \log p})$ as noted above. Note that the squared error of each nonsparse term is $O(n^{-1})$ and we have to estimate at least s terms of nonsparse parameters. Therefore, $\|\hat{\Delta}\|_2^2$ should be at least of order $O(s/n)$. The price that we pay for searching the unknown locations of nonsparse elements is merely a factor of $\log p$. In addition, Bickel, Ritov and Tsybakov (2009) show that the Dantzig selector and Lasso are asymptotically equivalent. James, Radchenko and Lv (2009) develop the explicit condition under which the Dantzig selector and Lasso will give identical fits.

The restricted eigenvalue condition basically imposes that the condition number (the ratio of the largest to the smallest eigenvalue) is bounded for any matrix $n^{-1}\mathbf{X}_S^T\mathbf{X}_S$ with $|S| = s$. This requires that the variables in \mathbf{X} are weakly correlated. It does not allow covariates to share common factors and can be very restrictive. A method to weaken this requirement and to adjust for latent factors is given by Kneip and Sarda (2011) and Fan, Ke and Wang (2016)

SLOPE and Sorted Penalties

The Sorted L-One (ℓ_1) Penalized Estimation (*SLOPE*) is introduced in Bogdan et al. (2015) to control the false discovery rate in variable selection. Given a sequence of penalty levels $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$, it finds the solution to the sorted ℓ_1 penalized least squares problem

$$\frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^p \lambda_j |\beta|_{(j)} \quad (3.111)$$

where $|\beta|_{(1)} \geq \dots \geq |\beta|_{(p)}$ are the order statistics of $\{|\beta|_j\}_{j=1}^p$, namely the decreasing sequence of $\{|\beta|_j\}_{j=1}^p$

For orthogonal design as in Section 3.2.1 with $\varepsilon \sim N(0, \sigma^2 \mathbf{I}_p)$, Bogdan et al. (2015) show that the false discovery rate for variable selection is controlled at level q if $\lambda_j = \Phi^{-1}(1 - jq/2p)\sigma/\sqrt{n}$, the rescaled critical values used by Benjamini and Hochberg (1995) for multiple testing. They also provide a fast computational algorithm. Su and Candès (2016) demonstrate that it achieves adaptive minimaxity in prediction and coefficient estimation for high-dimensional linear regression. Note that using the tail property of the standard normal distribution (see (3.113)), it is not hard to see that $\lambda_j \approx \sigma\sqrt{(2/n)\log(p/j)}$.

From the bias reduction point of view, the SLOPE is not satisfactory as it is still ℓ_1 -based penalty. This motivates Feng and Zhang (2017) to introduce sorted folded concave penalties that combine the strengths of concave and sorted penalties. Given a family of univariate penalty functions $p_\lambda(t)$ indexed by λ , the associated estimator is defined as

$$\frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^p p_{\lambda_j}(|\beta_{(j)}|)$$

This is a direct extension of (3.111) as it automatically reproduces it with $p_\lambda(t) = \lambda|t|$. The properties of SLOPE and its generalization will be thoroughly investigated in Section 4.5 under a unified framework.

Concentration inequalities and uniform convergence

The uniform convergence appears in a number of occasions for establishing consistency of regularized estimators. See, for example, (3.92), (3.97) and (3.105). It is fundamental to high-dimensional analysis. Let us illustrate the technique to prove (3.105), which is equivalent to showing

$$P \left\{ \left\| n^{-1} \mathbf{X}\boldsymbol{\varepsilon} \right\|_\infty \leq \lambda \right\} = P \left\{ \max_{1 \leq j \leq p} \left| n^{-1} \sum_{i=1}^n X_{ij} \varepsilon_i \right| \geq \lambda \right\} \rightarrow 1 \quad (3.112)$$

If we assume $\varepsilon_i \sim N(0, \sigma^2)$, the conditional distribution of $n^{-1} \sum_{i=1}^n X_{ij} \varepsilon_i \sim N(0, \sigma^2/n)$ under the standardization $n^{-1} \|\mathbf{X}_j\|^2 = 1$. Therefore, for any $t > 0$ we have

$$\begin{aligned} P \left\{ \left| n^{-1} \sum_{i=1}^n X_{ij} \varepsilon_i \right| \geq t\sigma/\sqrt{n} \right\} &= 2 \int_t^\infty \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) dx \\ &\leq \frac{2}{\sqrt{2\pi}} \int_t^\infty \frac{x}{t} \exp(-x^2/2) dx \\ &= \frac{2}{\sqrt{2\pi}} \exp(-t^2/2) / t \end{aligned} \quad (3.113)$$

In other words, the probability of the average of random variables at least t standard deviation from its mean converges to zero as t goes to ∞ exponentially fast. It is highly concentrated, and such a kind of inequality is called a concentration inequality.

Now, by the union bound, (3.112) and (3.113), we have

$$\begin{aligned} P \left\{ \left\| n^{-1} \mathbf{X}\boldsymbol{\varepsilon} \right\|_\infty > \frac{t\sigma}{\sqrt{n}} \right\} &\leq \sum_{j=1}^p P \left\{ \left| n^{-1} \sum_{i=1}^n X_{ij} \varepsilon_i \right| > \frac{t\sigma}{\sqrt{n}} \right\} \\ &\leq p \frac{2}{\sqrt{2\pi}} \exp(-t^2/2) / t \end{aligned}$$

Taking $t = \sqrt{2(1+\delta)\log p}$, the above probability is $o(p^{-\delta})$. In other words, with probability at least $1 - o(p^{-\delta})$,

$$\left\| n^{-1} \mathbf{X}\boldsymbol{\varepsilon} \right\|_\infty \leq \sqrt{2(1+\delta)}\sigma \sqrt{\frac{\log p}{n}} \quad (3.114)$$

The essence of the above proof relies on the concentration inequality (3.113) and the union bound. Note that the concentration inequalities in general hold for sum of independent random variables with *sub*-Gaussian tails or weaker conditions (see Lemma 2.5). They will appear in later chapters. See Boucheron, Lugosi and Massart (2013) and Tropp (2015) for general treatments. Below, we give a few of them so that readers can get an idea on these inequalities. These types of inequality began with Hoeffding's work in 1963.

Theorem 3.1.3 — Concentration inequalities. Assume that Y_1, \dots, Y_n are independent random variables with mean zero (without loss of generality). Let $S_n = \sum_{i=1}^n Y_i$ be the sum of the random variables.

1. Hoeffding's inequality: If $Y_i \in [a_i, b_i]$, then

$$P(|S_n| \geq t) \leq 2 \exp \left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2} \right)$$

2. Bernstein's inequality. If $E|Y_i|^m \leq m! M^{m-2} v_i / 2$ for every $m \geq 2$ and all i and some positive constants M and v_i , then

$$P(|S_n| \geq t) \leq 2 \exp \left(-\frac{t^2}{2(v_1 + \dots + v_n + Mt)} \right)$$

See Lemma ?? of van der Vaart and Wellner (1996).

3. Sub-Gaussian case: If $E \exp(aY_i) \leq \exp(v_i a^2 / 2)$ for all $a > 0$ and some $v_i > 0$, then, for any $t > 0$

$$P(|S_n| \geq t) \leq 2 \exp \left(-\frac{t^2}{2(v_1 + \dots + v_n)} \right)$$

4. Bounded second moment-Truncated loss: Assume that Y_i are i.i.d. with mean μ and variance σ^2 . Let

$$\hat{\mu}_\tau = \operatorname{argmin} \sum_{i=1}^n \rho_\tau(Y_i - \mu), \quad \rho_\tau(x) = \begin{cases} x^2, & \text{if } |x| \leq \tau \\ \tau(2|x| - \tau), & \text{if } |x| > \tau \end{cases}$$

be the adaptive Huber estimator. Then, for $\tau = \sqrt{nc}/t$ with $c \geq SD(Y)$ (standard deviation of Y), we have (Fan, Li, and Wang, 2017)

$$P \left(|\hat{\mu}_\tau - \mu| \geq t \frac{c}{\sqrt{n}} \right) \leq 2 \exp(-t^2/16), \quad \forall t \leq \sqrt{n}/8$$

5. Bounded second monment- Truncated data: Set $\tilde{Y}_i = \operatorname{sgn}(Y_i) \min(|Y_i|, \tau)$. When $\tau \asymp \sqrt{n}\sigma$, then

$$P \left(\left| \frac{1}{n} \sum_{i=1}^n \tilde{Y}_i - \mu \right| \geq t \frac{\sigma}{\sqrt{n}} \right) \leq 2 \exp(-ct^2)$$

for some universal constant c . See Fan, Wang, and Zhu (2016).

Proof. We give a proof of the sub-Gaussian case to illustrate the simple idea. By Makov's inequality, independence, sub-Gaussianity, we have for any $a > 0$

$$P(S_n \geq t) \leq \exp(-at) E \exp(aS_n) \leq \exp(-at) \prod_{i=1}^n \exp(v_i a^2 / 2)$$

By taking the optimal $a = t / (v_1 + \cdots + v_n)$, we obtain

$$P(S_n \geq t) \leq \exp\left(-\frac{t^2}{2(v_1 + \cdots + v_n)}\right)$$

This way of obtaining the inequality is called **Chernoff bound**. Now, applying the above inequality to $\{-Y_i\}$, we obtain that

$$P(S_n \leq -t) \leq \exp\left(-\frac{t^2}{2(v_1 + \cdots + v_n)}\right)$$

Combining the last two-inequalities, we obtain the result. ■

The common theme of the above results is that **the probability of S_n deviating from its mean more than t times of its standard deviation converges to zero in the rate $\exp(-ct^2)$ for some positive constant c** . Theorem 3.1.3(a) is for bounded random variables, whereas Cases b) and c) extend it to the case with sub-Gaussian moments or tails out. They all yield the same rate of convergence. Cases d) and e) extend the results further to the case only with bounded second moment. This line of work began with Catoni (2012). See also Devroye, Lerasle, Lugosi and Oliveira (2016).

A brief history of model selection

Figure 3.6 summarizes the important developments in model selection techniques. Particular emphasis is given to the development of the penalized least-squares methods. The list is by far from complete. For example, Bayesian model selection is not even included. It intends only to give readers a snapshot on the some historical developments. For example, the SCAD penalty function was actually introduced by Fan (1997) but its systematic developments were given by Fan and Li(2001), who studied the properties and computation of the whole class of the folded-concave penalized least-squares, not just SCAD. As discussed in Section 3.1.2, AIC and BIC criteria can be regarded as penalized L_0 regression. The idea of the ridge regression and the subset selection appear long before the 1970 s (see, e.g., Tikhonov, 1943; Hoerl, 1962).

Sure independence screening, which selects variables based on marginal utilities such as their marginal correlations with the response variable, is not a penalized method. It was introduced by Fan and Lv (2008) to reduce the dimensionality for high-dimensional problems with massive data. It will be systematically introduced in Chapter 8. Because

of its importance in analysis of big data and that it can be combined with PLS, we include it here for completeness.

Debiased Lasso was proposed by Zhang and Zhang (2014), which is further extended by van de Geer, Bühlmann, Ritov, and Dezeure (2014) and improved by Javanmard and Montanari (2014). For distributed estimation of high-dimensional problem, see Chen and Xie (2014), Shamir, Srebro and Zhang (2014), Lee, Liu, Sun and Taylor (2017), Battey, Fan, Liu, Lu and Zhu (2018), Jordan, Lee and Yang (2018), among others.

Bayesian Variable Selection

Bayesian view of the PLS

Sparse penalized regression can be put in the Bayesian framework. One can regard the parameters $\{\beta_j\}_{j=1}^p$ as a realization from a prior distribution having a density $\pi(\cdot)$ with the mode at the origin. If the observed data \mathbf{Y} has a density $p_Y(\mathbf{Y} | \mathbf{X}\beta)$ (conditioned on \mathbf{X}), then the joint density of the data and parameters is given by

$$f(\mathbf{Y}; \beta) = p_Y(\mathbf{Y} | \mathbf{X}\beta) \pi(\beta)$$

The posterior distribution of β given \mathbf{Y} is $f(\mathbf{Y}; \beta) / g(\mathbf{Y})$, which is proportional to $f(\mathbf{Y}; \beta)$ as a function of β , where $g(\mathbf{Y})$ is the marginal distribution of \mathbf{Y} . Bayesian inference is based on the posterior distribution of β given \mathbf{X} and \mathbf{Y} . One possible estimator is to use the posterior mean $E(\beta | \mathbf{X}, \mathbf{Y})$ to estimate β . Another is the posterior mode, which finds

$$\hat{\beta} = \operatorname{argmax}_{\beta \in R^p} \log p_Y(\mathbf{Y} | \mathbf{X}\beta) + \log \pi(\beta)$$

It is frequently taken as a Bayesian estimator. In particular, when $\mathbf{Y} \sim N(\mathbf{X}\beta, \mathbf{I}_n)$ (the standard deviation is taken to be one for convenience), then

$$\log p_Y(\mathbf{Y} | \mathbf{X}\beta) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \|\mathbf{Y} - \mathbf{X}\beta\|^2$$

Thus, finding the posterior mode reduces to minimizing

$$\frac{1}{2} \|\mathbf{Y} - \mathbf{X}\beta\|^2 - \log \pi(\beta)$$

Typically, the prior distributions are taken to be independent: $\pi(\beta) = \prod_{j=1}^p \pi_j(\beta_j)$ where $\pi_j(\cdot)$ is the marginal prior for β_j , though this is not mandatory. In this case, the problem becomes the penalized least-squares

$$\frac{1}{2} \|\mathbf{Y} - \mathbf{X}\beta\|^2 - \sum_{j=1}^p \log \pi_j(\beta_j) \quad (3.115)$$

When $\beta_j \sim_{i.i.d} \exp(-p_\lambda(|\beta_j|))$ where we hide the normalization constant, (3.115) becomes the penalized least-squares (3.71). In particular, when $\beta_j \sim_{i.i.d} \lambda \exp(-\lambda |\beta_j|) / 2$, the double exponential distribution with scale parameter λ , the above minimization problem becomes the Lasso problem (3.89). Note that when $p_\lambda(|\beta_j|)$ is flat (constant) at the

tails, the function $\exp(-p_\lambda(|\beta_j|))$ cannot be scaled to be a density function as it is not integrable. Such a prior is called an improper prior, one with very heavy tails. SCAD penalty corresponds to an improper prior.

The prior $\pi_j(\theta)$ typically involves some parameters γ , called hyper paramambution. One can regard them as the parameters generated from some other prior distributions. Such methods are called hierarchical Bayes. They can also be regarded as fixed parameters and are estimated through maximum likelihood, maximizing the marginal density $g(\mathbf{Y})$ of \mathbf{Y} with respect to γ . In other words, one estimates γ by the maximum likelihood and employs a Bayes rule to estimate parameters β for the given estimated γ . This procedure is referred to as **empirical Bayes**. Park and Casella (2008) discuss the use of empirical Bayes in the Bayesian Lasso by exploiting a hierarchical representation of the double exponential distribution as a scale mixture of normals (Andrews and Mallows 1974 :

$$\frac{a}{2}e^{-a|z|} = \int_0^\infty \frac{e^{-z^2/(2s)}}{\sqrt{2\pi s}} \frac{a^2}{2}e^{-a^2s/2}ds, a > 0$$

They develop a nice Gibbs sampler for sampling the posterior distribution. As demonstrated in Efron (2010), the empirical Bayes plays a very prominent role in large-scale statistical inference.

A Bayesian framework for selection

Bayesian inference of β and model selection are related but not identical problems. Bayesian model selection can be more complex. To understand this, let us denote $\{\mathcal{S}\}$ as all possible models, each model has a prior probability $p(\mathcal{S})$. For example, \mathcal{S} is equally likely among models with the same size and assign the probability proportion to $|\mathcal{S}|^{-\gamma}$. We can also assign the prior probability p_j to models with size j such that models of size j are all equally likely. Within each model \mathcal{S} , there is a parameter vector $\beta_{\mathcal{S}}$ with prior $\pi_{\mathcal{S}}(\cdot)$. In this case, the "joint density" of the models, the model parameters, and the data is

$$p(\mathcal{S})\pi_{\mathcal{S}}(\beta_{\mathcal{S}})p_Y(\mathbf{Y} | \mathbf{X}_{\mathcal{S}}\beta_{\mathcal{S}})$$

There is a large amount of literature on Bayesian model selection. The posterior modes are in general computed by using the Markov Chain Monte Carlo. See for example Andrieu, De Freitas, Doucet and Jordan (2003) and Liu (2008). Bayesian model selection techniques are very powerful in many applications, where disciplinary knowledge can be explicitly incorporated into the prior. See, for example, Raftery (1995).

A popular Bayesian idea for variable selection is to introduce p latent binary variables $Z = (z_1, \dots, z_p)$ such that $z_j = 1$ means variable x_j should be included in the model and $z_j = 0$ means excluding x_j . Given $z_j = 1$, the distribution of β_j has a flat tail (slab),

but the distribution of β_j given $z_j = 0$ is concentrated at zero (spike). The marginal distribution of β_j is a spike and slab prior. For example, assume that β_j is generated from a mixture of the point mass at 0 and a distribution $\pi_j(\beta)$ with probability α_j :

$$\beta_j \sim \alpha_j \delta_0 + (1 - \alpha_j) \pi_j(\beta)$$

See Johnstone and Silverman (2005) for an interesting study of this in wavelet regularization. For computation considerations, the spike distribution is often chosen to be a normal distribution with mean zero and a small variance. The slab distribution is another normal distribution with mean zero and a much bigger variance. See, for example, George and McCulloch (1993), Ishwaran and Rao (2005) and Narisetty and He (2014), among others. A working Bayesian model selection model with the Gaussian spike and slab prior is given as follows:

$$\begin{aligned} \mathbf{Y} \mid (\mathbf{X}, \beta, \sigma^2) &\sim N(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n) \\ \beta_j \mid (Z_j = 0, \sigma^2) &\sim N(0, \sigma^2 v_0) \\ \beta_j \mid (Z_j = 1, \sigma^2) &\sim N(0, \sigma^2 v_1) \\ P(Z_j = 1) &= q \\ \sigma^2 &\sim IG(\alpha_1, \alpha_2) \end{aligned}$$

where IG denotes the inverse Gamma distribution. The data generating process is bottom up in the above representation. The joint posterior distribution $P(\beta, Z, \sigma^2 \mid \mathbf{Y}, \mathbf{X})$ can be sampled by a neat Gibbs sampler. Model selection is based on the marginal posterior probabilities $P(Z_j = 1 \mid \mathbf{Y}, \mathbf{X})$. According to Barbieri and Berger (2004), x_j is selected if $P(Z_j = 1 \mid \mathbf{Y}, \mathbf{X}) \geq 0.5$. This selection method leads to the median probability model which is shown to be predictive optimal. For the high-dimension setting $p \gg n$, Narisetty and He (2014) establish the frequentist selection consistency of the Bayesian approach by using dimension-varying prior parameters: $v_0 = v_0(n, p) = o(n^{-1})$, $v_1 = v_1(n, p) = O\left(\frac{p^{2+\delta}}{n}\right)$ and $q = q(n, p) \approx p^{-1}$.

Numerical Algorithms

This section introduces some early developed algorithms to compute the folded-concave penalized least-squares. We first present the algorithms for computing the Lasso as it is more specific. We then develop algorithms for more general folded concave PLS such as SCAD and MCP. In particular, the connections between the folded-concave PLS and iteratively reweighted adaptive Lasso are made. These algorithms provide us not only a way to implement PLS but also statistical insights on the procedures.

In many applications, we are interested in finding the solution to PLS (3.71) for a range of values of λ . The solutions $\hat{\beta}(\lambda)$ to the PLS as a function of λ are called **solution paths** or **coefficient paths** (Efron, Hastie, Johnstone and Tibshirani, 2004). This allows one to examine how the variables enter into the solution as λ decreases.

Quadratic programs

There are several algorithms for computing Lasso: Quadratic programming, least-angle regression, and coordinate descent algorithm. The first two algorithms are introduced in this and next sections, and the last one will be introduced in Section 3.5.6.

First of all, as in Tibshirani (1996), a convenient alternative is to express the penalized L_1 -regression (3.89) into its dual problem (3.99). Each λ determines a constant c and vice versa. The relationship depends on the data (\mathbf{X}, \mathbf{Y}) .

The quadratic program, employed by Tibshirani (1996), is to regard the constraints $\|\beta\|_1 \leq c$ as $2p$ linear constraints $\mathbf{b}_j^T \beta \leq c$ for all p -tuples \mathbf{b}_j of form $(\pm 1, \pm 1, \dots, \pm 1)$. A simple solution is to write (3.99) as

$$\begin{aligned} & \min_{\beta^+, \beta^-} \|\mathbf{Y} - \mathbf{X}(\beta^+ - \beta^-)\|^2 \\ \text{s.t. } & \sum_{i=1}^p \beta_i^+ + \sum_{i=1}^p \beta_i^- \leq c, \quad \beta_i^+ \geq 0, \quad \beta_i^- \geq 0 \end{aligned} \quad (3.116)$$

This is a $2p$ -variable convex optimization problem and the constraints are linear in those variables. Therefore, the standard convex optimization algorithms and solvers (Boyd and Vandenberghe, 2004) can be employed. An alternative expression to the optimization problem (3.99) is

$$\begin{aligned} & \min_{\beta, \gamma} \|\mathbf{Y} - \mathbf{X}\beta\|^2 \\ \text{s.t. } & -\gamma_i \leq \beta_i \leq \gamma_i, \quad \sum_{i=1}^p \gamma_i \leq c, \quad \gamma_i \geq 0 \end{aligned} \quad (3.117)$$

This is again a $2p$ -variable convex optimization problem with linear constraints.

To find the solution paths, one needs to repeatedly solve a quadratic programming problem for a grid values of c . This is very inefficient and does not offer statistical insights. Osborne, Presnell and Turlach (2000) expressed the L_1 constraint as

$$\text{sgn}(\beta)^T \beta \leq c$$

They treated the problem as a quadratic program with $\text{sgn}(\beta)$ taken from the previous step of the iteration and developed a "homotopy method" based on this linearized constraint. Their homotopy method is related to the solutionpath algorithm of Efron et al. (2004)

Least angle regression

Efron et al. (2004) introduce the Least-Angle Regression (LARS) to explain the striking but mysterious similarity between the lasso regression path and the ϵ -boosting linear regression path observed by Hastie, Friedman and Tibshirani in 2001. Efron et al. (2004) show that the lasso regression and ϵ -boosting linear regression are two variants of LARS with different small modifications, thus explaining their similarity and differences. LARS itself is also an interesting procedure for variable selection and model estimation. LARS

is a forward stepwise selection procedure but operates in a less greedy way than the standard forward selection does. Assuming that all variables have been standardized so that they have mean-zero and unit variance, we now describe the LARS algorithm for the constrained least-square problem (3.99).

Let $\mathbf{z} = \mathbf{X}^T \mathbf{Y} / n$ and $\chi_j = \mathbf{X}^T \mathbf{X}_j / n$, where \mathbf{X}_j is the j^{th} column of \mathbf{X} . Then, necessary conditions for minimizing the Lasso problem (3.89) are [see (3.90) and (3.91)]

$$\begin{cases} \tau z_j - \chi_j^T \mathbf{b} = \text{sgn}(b_j) & \text{if } b_j \neq 0 \\ |\tau z_j - \chi_j^T \mathbf{b}| \leq 1 & \text{if } b_j = 0 \end{cases} \quad (3.118)$$

where $\tau = 1/\lambda$ and $\mathbf{b} = \tau \hat{\boldsymbol{\beta}}$. The solution path is given by $\hat{\mathbf{b}}(\tau)$ that solves (3.118). When $\tau \leq 1 / \|\mathbf{n}^{-1} \mathbf{X}^T \mathbf{Y}\|_\infty$, as noted in Section 3.3.2, the solution is $\hat{\mathbf{b}}(\tau) = 0$. We now describe the LARS algorithm for the constrained least-square problem (3.99). First of all, when $c = 0$ in (3.117), no variables are selected. This corresponds to $\hat{\boldsymbol{\beta}}(\lambda) = 0$ for $\lambda > \|\mathbf{n}^{-1} \mathbf{X}^T \mathbf{Y}\|_\infty$, as noted in Section 3.3.2.

As soon as c moves slightly away from zero, one picks only one variable (\mathbf{X}_1 , say) that has the maximum absolute correlation (least angle) with the response variable \mathbf{Y} . Then, $\hat{\boldsymbol{\beta}}_c = (\text{sgn}(r_1)c, 0, \dots, 0)^T$ is the solution to problem (3.99) for sufficiently small c , where r_1 is the correlation between \mathbf{X}_1 and \mathbf{Y} . Now, as c increases, the absolute correlation between the current residual

$$\mathbf{R}_c = \mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\beta}}_c$$

and \mathbf{X}_1 decreases until a (smallest) value c_1 at which there exists a second variable \mathbf{X}_2 , say, that has the same absolute correlation (equal angle) with \mathbf{R}_{c_1} :

$$|\text{cor}(\mathbf{X}_1, \mathbf{R}_{c_1})| = |\text{cor}(\mathbf{X}_2, \mathbf{R}_{c_1})|$$

Then, $\hat{\boldsymbol{\beta}}_c$ is the solution to problem (3.99) for $0 \leq c \leq c_1$ and the value c_1 can easily be determined, as in (3.122) below.

LARS then proceeds equiangularly between \mathbf{X}_1 and \mathbf{X}_2 until a third variable, \mathbf{X}_3 (say), joins the rank of "most correlated variables" with the current residuals. LARS then proceeds equiangularly between $\mathbf{X}_1, \mathbf{X}_2$ and \mathbf{X}_3 and so on. As we proceed down this path, the maximum of the absolute correlation of covariates with the current residual keeps decreasing until it becomes zero.

The equiangular direction of a set of variables \mathbf{X}_S is given by

$$\mathbf{u}_S = \mathbf{X}_S \left(\mathbf{X}_S^T \mathbf{X}_S \right)^{-1} \mathbf{1} / w_S \equiv \mathbf{X}_S \hat{\boldsymbol{\beta}}_S \quad (3.119)$$

where $\mathbf{1}$ is a vector of 1's and $w_S^2 = \mathbf{1}^T (\mathbf{X}_S^T \mathbf{X}_S)^{-1} \mathbf{1}$ is a normalization constant. The equiangular property can easily be seen:

$$\mathbf{X}_S^T \mathbf{u}_S = \mathbf{1} / w_S, \quad \|\mathbf{u}_S\| = 1$$

We now furnish some details of LARS. Assume that \mathbf{X} is of full rank. Start from $\mu_0 = 0, \mathcal{S} = \emptyset$, the empty set, and $\beta_{\mathcal{S}} = 0$. Let \mathcal{S} be the current active set of variables and $\hat{\mu}_{\mathcal{S}} = \mathbf{X}_{\mathcal{S}} \hat{\beta}_{\mathcal{S}}$ be its current fitted value of \mathbf{Y} . Compute the **marginal correlations** of covariates \mathbf{X} with the current residual (except a normalization constant)

$$\hat{\mathbf{c}} = \mathbf{X}^T (\mathbf{Y} - \hat{\mu}_{\mathcal{S}}) \quad (3.120)$$

Define $s_j = \text{sgn}(\hat{c}_j)$ for $j \in \mathcal{S}$. Note that the absolute correlation does not change if the columns of \mathbf{X} are multiplied by ± 1 . Update the active set of variables by taking the most(binggest) correlated set

$$\mathcal{S}_{\text{new}} = \{j : |\hat{c}_j| = \|\hat{\mathbf{c}}\|_{\infty}\} \quad (3.121)$$

and compute the equiangular direction $\mathbf{u}_{\mathcal{S}_{\text{new}}}$ by (3.119) using variables $\{\text{sgn}(\hat{c}_j) \mathbf{X}_j, j \in \mathcal{S}\}$. For the example in the beginning of this section, if $\mathcal{S} = \emptyset$, an empty set, then $\hat{\mu}_{\mathcal{S}} = 0$ and $\mathcal{S}_{\text{new}} = \{1\}$. If $\mathcal{S} = \{1\}$, then $\hat{\mu}_{\mathcal{S}} = \mathbf{X}_{\mathcal{S}} \hat{\beta}_{\mathcal{S}} = \text{sgn}(r_1) c_1 \mathbf{X}_1$ and $\mathcal{S}_{\text{new}} = \{1, 2\}$. Now compute

$$\gamma_{\mathcal{S}} = \min_{j \in \mathcal{S}^c} \left\{ \frac{\|\mathbf{c}\|_{\infty} - \hat{c}_j}{w_{\mathcal{S}}^{-1} - a_j}, \frac{\|\mathbf{c}\|_{\infty} + \hat{c}_j}{w_{\mathcal{S}}^{-1} + a_j} \right\}, \quad \mathbf{a} = \mathbf{X}^T \mathbf{u}_{\mathcal{S}} \quad (3.122)$$

where "min $^{+}$ " is the minimum taken only over positive components. It is not hard to show that this step size $\gamma_{\mathcal{S}}$ is the smallest positive constant γ such that some new indices will join the active set (see Efron et al., 2004). For example, if $\mathcal{S} = \{1\}$, this $\gamma_{\mathcal{S}}$ is c_1 in the first step. Update the fitted value along the equiangular direction by

$$\hat{\mu}_{\mathcal{S}_{\text{new}}} = \hat{\mu}_{\mathcal{S}} + \gamma_{\mathcal{S}_{\text{new}}} \mathbf{u}_{\mathcal{S}_{\text{new}}} \quad (3.123)$$

The solution path for $\gamma \in (0, \gamma_{\mathcal{S}_{\text{new}}})$ is

$$\hat{\mu}_{\mathcal{S}_{\text{new}}, \gamma} = \hat{\mu}_{\mathcal{S}} + \gamma \mathbf{u}_{\mathcal{S}_{\text{new}}}$$

Note that \mathcal{S}_{new} is always a bigger set than \mathcal{S} . Write $\hat{\mu}_{\mathcal{S}} = \mathbf{X}_{\mathcal{S}} \hat{\beta}_{\mathcal{S}}$, in which $\hat{\beta}_{\mathcal{S}}$ has support \mathcal{S} so that $\hat{\mu}_{\mathcal{S}}$ is in the linear space spanned by columns of $\mathbf{X}_{\mathcal{S}}$. By (3.119), we have

$$\hat{\mu}_{\mathcal{S}_{\text{new}}, \gamma} = \mathbf{X} \left(\hat{\beta}_{\mathcal{S}} + \gamma \beta_{\mathcal{S}_{\text{new}}} \right)$$

Note that by (3.119), $\hat{\beta}_{\mathcal{S}} + \gamma \beta_{\mathcal{S}_{\text{new}}}$ has a support \mathcal{S}_{new} . In terms of coefficients, we have updated the coefficients from $\hat{\beta}_{\mathcal{S}}$ for variables $\mathbf{X}_{\mathcal{S}}$ to

$$\hat{\beta}_{\mathcal{S}_{\text{new}}, \gamma} = \hat{\beta}_{\mathcal{S}} + \gamma \beta_{\mathcal{S}_{\text{new}}} \quad (3.124)$$

for variables $\mathbf{X}_{\mathcal{S}_{\text{new}}}$, expressed in R^p . Some modifications of the signs in the second term in (3.124) is needed since we use the variables $\{\text{sgn}(\hat{c}_j) \mathbf{X}_j, j \in \mathcal{S}\}$ rather than $\mathbf{X}_{\mathcal{S}}$ to compute the equiangular direction $\mathbf{u}_{\mathcal{S}_{\text{new}}}$. The LARS algorithm is summarized as follows.

1. Initialization: Set $\mathcal{S} = \phi, \hat{\mu}_{\mathcal{S}} = 0, \hat{\beta}_{\mathcal{S}} = 0$
2. Step 1: Compute the current correlation vector \hat{c} by (3.120), the new subset \mathcal{S}_{new} , the least angular covariates with the current residual $\mathbf{Y} - \hat{\mu}_{\mathcal{S}}$, by (3.121), and the stepsize $\gamma_{\mathcal{S}_{\text{new}}}$, the largest stepsize along the equiangular direction, by (3.122)
3. Step 2: Update \mathcal{S} with \mathcal{S}_{new} , $\hat{\mu}_{\mathcal{S}}$ with $\hat{\mu}_{\mathcal{S}_{\text{new}}}$ in (3.123), and $\hat{\beta}_{\mathcal{S}}$ with $\hat{\beta}_{\mathcal{S}_{\text{new}}}$ in (3.124) with $\gamma = \gamma_{\mathcal{S}_{\text{new}}}$
4. Iterations: Iterate between Steps 1 and 2 until all variables are included in the model and the solution reaches the OLS estimate.

The entire LARS solution path simply connects p -dimensional coefficients linearly at each discrete step above. However, it is not necessarily the solution to the lasso problem (3.99). The LARS model size is enlarged by one after each step, but the Lasso may also drop a variable from the current model as c increases. Technically speaking, (3.90) shows that Lasso and the current correlation must have the same sign, but the LARS solution path does not enforce this. Efron et al. (2004) show that this sign constraint can easily be enforced in the LARS algorithm: during the ongoing LARS update step, if the j th variable in \mathcal{S} has a sign change before the new variable enters \mathcal{S} , stop the ongoing LARS update, drop the j th variable from the model and recalculate the new equiangular direction for doing the LARS update. Efron et al. (2004) prove that the modified LARS path is indeed the Lasso solution path under a "one at a time" condition, which assumes that at most one variable can enter or leave the model at any time.

Other modifications of the LARS algorithm are also possible. For example, by modifying LARS shrinkage, James and Radchenko (2008) introduce variable inclusion and shrinkage algorithms (VISA) that intend to attenuate the over-shrinkage problem of Lasso. James, Radchenko and Lv (2009) develop an algorithm called Dasso that allows one to fit the entire path of regression coefficients for different values of the Dantzig selector tuning parameter.

The key argument in the LARS algorithm is the piecewise linearity property of the Lasso solution path. This property is not unique to the Lasso PLS. Many statistical models can be formulated as $\min\{\text{Loss} + \lambda \text{Penalty}\}$. In Rosset and Zhu (2007) it is shown that if the loss function is almost quadratic and the penalty is L_1 (or piecewise linear), then the solution path is piecewise linear as a function of λ . Examples of such models include the L_1 penalized Huber regression (Rosset and Zhu, 2007) and the L_1 penalized support vector machine (Zhu, Rosset, Hastie and Tibshirani, 2004). Interestingly, if the loss function is L_1 (or piecewise linear) and the penalty function is quadratic, we can switch their roles when computing and the solution path is piecewise linear as a function of $1/\lambda$. See, for example, the solution path algorithms for the support vector machine (Hastie, Rosset, Tibshirani and Zhu, 2003) and support vector regression (Gunter and Zhu, 2007).

Local quadratic approximations

Local quadratic approximation (LQA) was introduced by Fan and Li (2001) before LARS-Lasso or other effective methods that are available in statistics for computing Lasso. It allows statisticians to implement folded concave penalized likelihood and to compute the standard error of estimated nonzero components. Given an initial value β_0 , approximate the function $p_\lambda(|\beta|)$ locally at this point by a quadratic function $q(\beta | \beta_0)$. This quadratic function is required to be symmetric around zero, satisfying two order conditions

$$q(\beta_0 | \beta_0) = p_\lambda(|\beta_0|) \quad \text{and} \quad q'(\beta_0 | \beta_0) = p'_\lambda(|\beta_0|)$$

These three conditions determine uniquely the quadratic function (quadratic function has three parameters, and they can be calculated by these three conditions)

$$q(\beta | \beta_0) = p_\lambda(|\beta_0|) + \frac{1}{2} \frac{p'_\lambda(|\beta_0|)}{|\beta_0|} (\beta^2 - \beta_0^2) \quad (3.125)$$

Given the current estimate β_0 , by approximating each folded concave function in PLS (3.71) by its LQA, our target becomes minimizing

$$Q(\beta | \beta_0) = \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \sum_{j=1}^p q(\beta_j | \beta_{j0}) \quad (3.126)$$

Minimizing (3.126) is the same as minimizing

$$\frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \sum_{j=1}^p \frac{p'_\lambda(|\beta_{j0}|)}{2|\beta_{j0}|} \beta_j^2$$

This is a ridge regression problem with solution computed analytically as

$$\hat{\beta}_{\text{new}} = \left(\mathbf{X}^T \mathbf{X} + n \text{diag} \{ p'_\lambda(|\beta_{j0}|) / |\beta_{j0}| \} \right)^{-1} \mathbf{X}^T \mathbf{Y} \quad (3.127)$$

The LQA is to iteratively use (3.127), starting from an initial value (e.g. univariate marginal regression coefficients). Fan and Li (2001) note that the approximation (3.125) is not good when $|\beta_{j0}| \leq \varepsilon_0$, a tolerance level. When this happens, delete variables from the model before applying (3.127). This speeds up the computation. Furthermore, they proposed to compute the standard error for surviving variables using (3.127), as if $p'_\lambda(\beta_{j0}) / |\beta_{j0}|$ were non-stochastic. They validated the accuracy of the estimated standard error. See Fan and Peng (2004) for a theoretical proof. Does the algorithm converge? and if so, in what sense? Hunter and Li (2005) realized that the local quadratic approximation is a specific case of the majorization-minimization (MM) algorithm (Hunter and Lange, 2000). First of all, as shown in Figure 3.8, thanks to the folded-concaveness,

$$q(\beta | \beta_0) \geq p_\lambda(\beta) \quad \text{and} \quad q(\beta_0 | \beta_0) = p_\lambda(\beta_0)$$

namely $q(\beta | \beta_0)$ is a convex majorant of $p_\lambda(\cdot)$ with $q(\beta_0 | \beta_0) = p_\lambda(|\beta_0|)$. This entails that

$$Q(\beta | \beta_0) \geq Q(\beta) \text{ and } Q(\beta_0 | \beta_0) = Q(\beta_0) \quad (3.128)$$

namely, $Q(\beta | \beta_0)$ is a convex majorization of the folded-concave PLS $Q(\beta)$ defined by (3.71). Let β_{new} minimize $Q(\beta | \beta_0)$. Then, it follows from (3.128) that

$$Q(\beta_{\text{new}}) \leq Q(\beta_{\text{new}} | \beta_0) \leq Q(\beta_0 | \beta_0) = Q(\beta_0) \quad (3.129)$$

where the second inequality follows from the definition of the minimization. In other words, the target function decreases after each iteration and will converge.

Local linear algorithm

With LARS and other efficient algorithms for computing Lasso, the local linear approximation (LLA) approximates $p_\lambda(|\beta|)$ at β_0 by

$$l(\beta | \beta_0) = p_\lambda(|\beta_0|) + p'_\lambda(|\beta_0|)(|\beta| - |\beta_0|)$$

which is the first-order Taylor expansion of $p_\lambda(|\beta|)$ at the point β_0 . Clearly, as shown in Figure 3.8, $l(\beta | \beta_0)$ is a better approximation than LQA $q(\beta | \beta_0)$. Indeed, it is the minimum convex majorant of $p_\lambda(|\beta|)$ with $l(\beta_0 | \beta_0) = p_\lambda(|\beta_0|)$. With the local linear approximation, given the current estimate β_0 , the folded concave penalized least-squares (3.71) now becomes

$$\frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \sum_{j=1}^p p'_\lambda(|\beta_{j0}|) |\beta_j| \quad (3.130)$$

after ignoring the constant term. This is now an adaptively weighted Lasso problem and can be solved by using algorithms in Sections 3.5.1 and 3.5.2. The algorithm was introduced by Zou and Li (2008). As it is also a specific MM algorithm, as shown in (3.129), the LLA algorithm also enjoys a decreasing target value property (3.129).

Unlike LQA, if one component hits zero at a certain step, it will not always be stuck at zero. For example, if $\beta_0 = 0$, (3.130) reduces to Lasso. In this view, even when the initial estimator is very crude, LLA gives a good one-step estimator.

Through LLA approximation (3.130), the folded-concave PLS can be viewed as an iteratively reweighted penalized L_1 regression. The weights depend on where the current estimates are. The larger the magnitude, the smaller the weighted penalty. This reduces the biases for estimating large true coefficients. Lasso is the one-step estimator of the folded concave PLS with initial estimate $\hat{\beta} = \mathbf{0}$. Lasso puts a full stop yet the folded concave PLS iterates further to reduce the bias due to the Lasso shrinkage.

It is now clear that the adaptive Lasso is a specific example of the LLA implementation of the folded-concave PLS with $p'_\lambda(|\beta|) = |\beta|^{-\gamma}$. This function is explosive near 0

and is therefore inappropriate to use in the iterative application of (3.130): once a component β_j hits zero at certain iteration, its weights cannot be computed or the variable X_j is eliminated forever.

The LLA implementation of folded concave PLS has a very nice theoretical property. Fan, Xue and Zou (2014) show that with Lasso as the initial estimator, with high probability, the LLA implementation (3.130) produces the oracle estimator in one step. The result holds in a very general likelihoodbased context under some mild conditions. This gives additional endorsement of the folded-concave PLS implemented by LLA (3.130)

The implementation of LLA is available in the R package called "SIS" (function: `scadglm`), contributed by Fan, Feng, Samworth, and Wu.

Penalized linear unbiased selection

The penalized linear unbiased selection (PLUS) algorithm, introduced by Zhang (2010), finds multiple local minimizers of folded-concave PLS in a branch of the graph (indexed by $\tau = \lambda^{-1}$) of critical points determined by (3.84). The PLUS algorithm deals with the folded concavepenalized functions $p_\lambda(t) = \lambda^2 \rho(t/\lambda)$, in which $\rho(\cdot)$ is a quadratic spline. This includes L_1 , SCAD (3.76), hard-thresholding penalty (3.77) and MCP(3.79) as specific examples. Let $t_1 = 0, \dots, t_m$ be the knots of the quadratic spline $\rho(\cdot)$. Then, the derivative of $\rho(\cdot)$ can be expressed as

$$\rho'(t) = \sum_{i=1}^m (u_i - v_i t) I(t_i < t \leq t_{i+1}) \quad (3.131)$$

for some constants $\{v_i\}_{i=1}^m$, in which $u_1 = 1$ (normalization), $u_m = v_m = 0$ (flat tail), and $t_{m+1} = \infty$. For example, L_1 penalty corresponds to $m = 1$ MCP corresponds to $m = 2$ with $t_2 = a, v_1 = 1/a$; SCAD corresponds to $m = 3$ with $t_2 = 1, t_3 = a, v_1 = 0, u_2 = a/(a-1)$ and $v_2 = 1/(a-1)$.

Let $\mathbf{z} = \mathbf{X}^T \mathbf{Y}/n$ and $\chi_j = \mathbf{X}^T \mathbf{X}_j/n$, where \mathbf{X}_j is the j^{th} column of \mathbf{X} . Then, estimating equations (3.84) can be written as

$$\begin{cases} \tau z_j - \chi_j^T \mathbf{b} = \text{sgn}(b_j) \rho'(|b_j|) & \text{if } b_j \neq 0 \\ \left| \tau z_j - \chi_j^T \mathbf{b} \right| \leq 1 & \text{if } b_j = 0 \end{cases} \quad (3.132)$$

where $\tau = 1/\lambda$ and $\mathbf{b} = \tau \boldsymbol{\beta}$. (3.132) can admit multiple solutions for each given λ . For example, $\mathbf{b} = \mathbf{0}$ is a local solution to (3.132), when $\lambda \geq \|\mathbf{X}^T \mathbf{Y}/n\|_\infty$. See also (3.93). Unlike Lasso, there can be other local solutions to (3.132). PLUS computes the main branch $\hat{\boldsymbol{\beta}}(\tau)$ starting from $\hat{\boldsymbol{\beta}}(\tau) = \mathbf{0}$, where $\tau = 1/\|\mathbf{X}^T \mathbf{Y}/n\|_\infty$. Let us characterize the solution set of (3.72). The component b_j of a solution \mathbf{b} falls in one of the intervals $\{(t_i, t_{i+1}]\}_{i=1}^m$, or 0, or in one of the intervals $[-t_{i+1}, -t_i)\}_{i=1}^m$. Let us use $i_j \in \{-m, \dots, m\}$ to indicate such an interval and $\mathbf{i} \in \{-m, \dots, m\}^p$ be the vector of indicators. Then, by (3.71), (3.72) can

be written as

$$\begin{cases} \tau z_j - \chi_j^T \mathbf{b} = \text{sgn}(i_j) (u_{i_j} - b_j v_{i_j}), & \bar{t}_{i_j} \leq b_j \leq \bar{t}_{i_j+1}, & i_j \neq 0 \\ -1 \leq \tau z_j - \chi_j^T \mathbf{b} \leq 1, & b_j = 0, & i_j = 0 \end{cases} \quad (3.133)$$

where $u_{-k} = u_k, v_{-k} = v_k$, and $\bar{t}_i = t_i$ for $0 < i \leq m+1$ and $-t_{|i|+1}$ for $-m \leq i \leq 0$. Let $\mathcal{S}_\tau(\mathbf{i})$ be the set of $(\tau \mathbf{z}^T, \mathbf{b}^T)^T$ in R^{2p} , whose coordinates satisfy (3.133). Note that the solution \mathbf{b} is piecewise linear τ . Let $H = R^p$ represent the data \mathbf{z} and its dual $H^* = R^p$ represent the solution \mathbf{b} , and $\mathbf{z} \oplus \mathbf{b}$ be members of $H \oplus H^* = R^{2p}$. The set $\mathcal{S}_\tau(\mathbf{i})$ in R^{2p} is more compactly expressed as

$$\mathcal{S}_\tau(\mathbf{i}) = \{\tau \mathbf{z} \oplus \mathbf{b} : \tau \mathbf{z} \text{ and } \mathbf{b} \text{ satisfy (3.133)}\}$$

For each given τ and \mathbf{i} , the set $\mathcal{S}_\tau(\mathbf{i})$ is a parallelepiped in R^{2p} and $\mathcal{S}_\tau = \tau \mathcal{S}_1$.

The solution \mathbf{b} is the projection of $\mathcal{S}_\tau(\mathbf{i})$ onto H^* , denoted by $\mathcal{S}_\tau(\mathbf{i} | \mathbf{z})$. Clearly, all solutions to (3.132) is a p -dimensional set given by

$$\mathcal{S}_\tau(\mathbf{z}) = \cup \{\mathcal{S}_\tau(\mathbf{i} | \mathbf{z}) : \mathbf{i} \in \{-m, -m+1, \dots, m\}^p\} \quad (3.134)$$

Like LARS, the PLUS algorithm computes a solution $\boldsymbol{\beta}(\tau)$ from $\mathcal{S}_\tau(\mathbf{z})$. Starting from $\tau_0 = 1 / \|\mathbf{n}^{-1} \mathbf{X}^T \mathbf{Y}\|_\infty$, $\hat{\boldsymbol{\beta}}(\tau_0) = \mathbf{0}$ and $\mathbf{i} = \mathbf{0}$, PLUS updates the active set of variables as well as the branch \mathbf{i} , determines the step size τ and solution \mathbf{b} . The solutions between two turning points are connected by lines. We refer to Zhang (2010) for additional details. In addition, Zhang (2010) gives the conditions under which the solution becomes the oracle estimator, derives the risk and model selection properties of the PLUS estimators.

Cyclic coordinate descent algorithms

Consider the sparse penalized least squares, the computation difficulty comes from the nonsmoothness of the penalty function. Observe that the penalty function part is the sum of p univariate nonsmooth functions. Then, we can employ cyclic coordinate descent algorithms (Tseng, 2001; Tseng and Yun, 2009) that successively optimize one coefficient (coordinate) at a time. Let

$$L(\beta_1, \dots, \beta_p) = \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^p p_\lambda(|\beta_j|)$$

and the cyclic coordinate descent (CCD) algorithm proceeds as follows:

1. choose an initial value of $\hat{\boldsymbol{\beta}}$
2. for $j = 1, 2, \dots, p, 1, 2, \dots$, update $\hat{\beta}_j$ by solving a univariate optimization problem of β_j :

$$\hat{\beta}_j^{\text{update}} \leftarrow \arg\min_{\beta_j} L(\hat{\beta}_1, \dots, \hat{\beta}_{j-1}, \beta_j, \hat{\beta}_{j+1}, \dots, \hat{\beta}_p) \quad (3.135)$$

3. Repeat (2) till convergence.

Let $\mathbf{R}_j = \mathbf{Y} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}_{-j}$ be the current residual, where \mathbf{X}_{-j} and $\hat{\boldsymbol{\beta}}_{-j}$ are respectively \mathbf{X} and $\hat{\boldsymbol{\beta}}$ with the j^{th} column and j^{th} component removed. Then, the target function in (3.135) becomes, after ignoring a constant,

$$Q_j(\beta_j) \equiv \frac{1}{2n} \|\mathbf{R}_j - \mathbf{X}_j\beta_j\|^2 + p_\lambda(|\beta_j|)$$

Recall that $\|\mathbf{X}_j\|^2 = n$ by standardization and $\hat{c}_j = n^{-1}\mathbf{X}_j^T\mathbf{R}_j$ is the current covariance [c.f. (3.120)]. Then, after ignoring a constant,

$$Q_j(\beta_j) = \frac{1}{2} (\beta_j - \hat{c}_j)^2 + p_\lambda(|\beta_j|) \quad (3.136)$$

This is the same problem as (3.74). For L_1 , SCAD and MCP penalty, (3.136) admits an explicit solution as in (3.80) – (3.82). In this case, the CCD algorithm is simply an iterative thresholding method.

The CCD algorithm for the Lasso regression is the same as the shooting algorithm introduced by Fu (1998). Friedman, Hastie, Höfling and Tibshirani (2007) implement the CCD algorithm by using several tricks such as warm start, active set update, etc. As a result, they were able to show that the coordinate descent algorithm is actually very effective in computing the Lasso solution path, proving to be even faster than the LARS algorithm. Fan and Lv (2011) extend the CCD algorithm to the penalized likelihood.

The user needs to be careful when applying the coordinate descent algorithm to solve the concave penalized problems because the algorithm converges to a local minima but this solution may not be the statistical optimal one. The choice of initial value becomes very important. In Fan, Xue and Zou (2014) there are simulation examples showing that the solution by CCD is suboptimal compared with the LLA solution in SCAD and MCP penalized regression and logistic regression. It is beneficial to try multiple initial values when using CCD to solve nonconvex problems.

Iterative shrinkage-thresholding algorithms

The iterative shrinkage-thresholding algorithm (ISTA, Daubechies et al., 2004) is developed to optimize the functions of form $Q(\boldsymbol{\beta}) = f(\boldsymbol{\beta}) + g(\boldsymbol{\beta})$, in which f is smooth whereas $g(\boldsymbol{\beta})$ is non-smooth. Note that the gradient descent algorithm

$$\boldsymbol{\beta}_k = \boldsymbol{\beta}_{k-1} - s_k f'(\boldsymbol{\beta}_{k-1})$$

for a suitable stepsize s_k is the minimizer to the local isotropic quadratic approximation of f at $\boldsymbol{\beta}_{k-1}$:

$$f_A(\boldsymbol{\beta} | \boldsymbol{\beta}_{k-1}, s_k) = f(\boldsymbol{\beta}_{k-1}) + f'(\boldsymbol{\beta}_{k-1})^T (\boldsymbol{\beta} - \boldsymbol{\beta}_{k-1}) + \frac{1}{2s_k} \|\boldsymbol{\beta} - \boldsymbol{\beta}_{k-1}\|^2 \quad (3.137)$$

The local isotropic approximation avoids computing the Hessian matrix, which is expensive and requires a lot of storage for high-dimensional optimization. Adapting this idea

to minimizing $Q(\cdot)$ yields the algorithm

$$\beta_k = \operatorname{argmin} \{f_A(\beta | \beta_{k-1}, s_k) + g(\beta)\}$$

In particular, when $g(\beta) = \sum_{j=1}^p p_\lambda(|\beta_j|)$, the problem becomes a componentwise optimization after ignoring a constant

$$\beta_k = \operatorname{argmin} \left\{ \frac{1}{2s_k} \|\beta - (\beta_{k-1} - s_k f'(\beta_{k-1}))\|^2 + \sum_{j=1}^p p_\lambda(|\beta_j|) \right\}$$

for each component of the form (3.74). Let us denote

$$\theta_s(z) = \operatorname{argmin}_\theta \left\{ \frac{1}{2}(z - \theta)^2 + s p_\lambda(|\theta|) \right\}$$

Then, ISTA is to iteratively apply

$$\beta_k = \theta_{s_k}(\beta_{k-1} - s_k f'(\beta_{k-1})) \quad (3.138)$$

In particular, for the Lasso problem (3.29), the ISTA becomes

$$\beta_k = \left(\beta_{k-1} - s_k n^{-1} \mathbf{X}^T (\mathbf{Y} - \mathbf{X} \beta_{k-1}) - s_k \lambda \right)_+ \quad (3.139)$$

Similar iterative formulas can be obtained for SCAD and MCP. This kind of algorithm is called a proximal gradient method in the optimization literature. Note that when $\|f'(\beta) - f'(\theta)\| \leq \|\beta - \theta\|/s_k$ for all β and θ , we have $f_A(\beta | \beta_{k-1}, s_k) \geq f(\beta)$. This holds when the largest eigenvalue of the Hessian matrix $f''(\beta)$ is bounded by $1/s_k$. Therefore, the ISTA algorithm is also a specific implementation of the MM algorithm, when the condition is met. The above isotropic quadratic majorization requires strong conditions regarding to the function f . Inspecting the proof in (3.129) for the MM algorithm, we indeed do not require majorization but only the local majorization $Q(\beta_{\text{new}}) \leq Q(\beta_{\text{new}} | \beta_0)$. This can be achieved by using the backtracking rule to choose the step size s_k as follows. Take an initial step size $s_0 > 0, \delta < 1$, and the initial value β_0 . Find the smallest nonnegative integer i_k such that with $s = \delta^{i_k} s_{k-1}$

$$Q(\beta_{k,s}) \leq Q_A(\beta_{k,s}) \equiv f_A(\beta_{k,s} | \beta_{k-1}, s) + g(\beta_{k,s}) \quad (3.140)$$

where $\beta_{k,s} = \theta_s(\beta_{k-1} - s f'(\beta_{k-1}))$ is the same as the above with emphasis on its dependence on s . Set $s_k = \delta^{i_k} s_{k-1}$ and compute

$$\beta_k = \theta_{s_k}(\beta_{k-1} - s_k f'(\beta_{k-1}))$$

Note that the requirement (3.80) is really the local majorization requirement. It can easily hold since $s_k \rightarrow 0$ exponentially fast as $i_k \rightarrow \infty$. According to (3.69), the sequence of objective values $\{Q(\beta_k)\}$ is non-increasing. The above choice of the step size of s_k can be very

small as k gets large. Another possible scheme is to use $s = \delta^{i_k} s_0$ rather than $s = \delta^{i_k} s_{k-1}$ in (3.80) in choosing s_k

The fast iterative shrinkage-thresholding algorithm (FISTA, Beck and Teboulle, 2009) is proposed to improve the convergence rate of ISTA. It employs Nesterov acceleration idea (Nesterov, 1983). The algorithm runs as follows. Input the step size s such that s^{-1} is the upper bound of the Lipchitz constant of $f'(\cdot)$. Take $\mathbf{x}_1 = \boldsymbol{\beta}_0$ and $t_1 = 1$. Compute iteratively for $k \geq 1$

$$\begin{aligned} \boldsymbol{\beta}_k &= \theta_s (\mathbf{x}_k - s f'(\mathbf{x}_k)), \quad t_{k+1} = \left(1 + \sqrt{1 + 4t_k^2}\right) / 2 \\ \mathbf{x}_{k+1} &= \boldsymbol{\beta}_k + \frac{t_k - 1}{t_{k+1}} (\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1}) \end{aligned}$$

The algorithm utilizes a constant "stepsize" s . The backtracking rule can also be employed to make the algorithm more practical. Beck and Teboulle (2009) show that the FISTA has a quadratic convergence rate whereas the ISTA has only linear convergence rate.

Projected proximal gradient method

Agarwal, Negahban and Wainwright (2012) propose a projected proximal gradient descent algorithm to solve the problem

$$\min_{R(\boldsymbol{\beta}) \leq c} \{f(\boldsymbol{\beta}) + g(\boldsymbol{\beta})\} \quad (3.141)$$

Given the current value $\boldsymbol{\beta}_{k-1}$, approximate the smooth function f by isotropic quadratic (3.77). The resulting unconstrained solution is given by (3.78). Now, project $\boldsymbol{\beta}_k$ onto the set $\{\boldsymbol{\beta} : R(\boldsymbol{\beta}) \leq c\}$ and continue with the next iteration by taking the projected value as the initial value. When $\|R(\boldsymbol{\beta})\| = \|\boldsymbol{\beta}\|_1$, the projection admits an analytical solution. If $\|\boldsymbol{\beta}_k\|_1 \leq c$, then the projection is just itself; otherwise, it is the soft-thresholding at level λ_n so that the constraint $\|\boldsymbol{\beta}_k\|_1 = c$. The threshold level λ_n can be computed as follows:

1. sort $\{|\beta_{k,j}|\}_{j=1}^p$ into $b_1 \geq b_2 \geq \dots \geq b_p$;
2. (2) find $J = \max\{1 \leq j \leq p : b_j - \left(\sum_{r=1}^j b_r - c\right) / j > 0\}$ and let $\lambda_n = \left(\sum_{r=1}^J b_r - c\right) / J$

ADMM

The alternating direction method of multipliers (ADMM) (Douglas and Rachford (1956), Eckstein and Bertsekas (1992)) has a number of successful applications in modern statistical machine learning. Boyd et al. (2011) give a comprehensive review on ADMM. Solving the Lasso regression problem is a classical application of ADMM. Consider the Lasso penalized least square

$$\min_{\boldsymbol{\beta}} \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_1$$

which is equivalent to

$$\min_{\boldsymbol{\beta}, \mathbf{z}} \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\mathbf{z}\|_1 \quad \text{subject to} \quad \mathbf{z} = \boldsymbol{\beta}$$

The augmented Lagrangian is

$$\mathcal{L}_\eta(\boldsymbol{\beta}, \mathbf{z}, \boldsymbol{\theta}) = \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\mathbf{z}\|_1 - \boldsymbol{\theta}^T (\mathbf{z} - \boldsymbol{\beta}) + \frac{\eta}{2} \|\mathbf{z} - \boldsymbol{\beta}\|_2^2$$

where η can be a fixed positive constant set by the user, e.g. $\eta = 1$. The term $\boldsymbol{\theta}^T (\mathbf{z} - \boldsymbol{\beta})$ is the Lagrange multiplier and the term $\frac{\eta}{2} \|\mathbf{z} - \boldsymbol{\beta}\|_2^2$ is its augmentation. The choice of η can affect the convergence speed. ADMM is an iterative procedure. Let $(\boldsymbol{\beta}^k, \mathbf{z}^k, \boldsymbol{\theta}^k)$ denote the k th iteration of the ADMM algorithm for $k = 0, 1, 2, \dots$. Then the algorithm proceeds as follows:

$$\begin{aligned} \boldsymbol{\beta}^{k+1} &= \operatorname{argmin}_{\boldsymbol{\beta}} \mathcal{L}_\eta(\boldsymbol{\beta}, \mathbf{z}^k, \boldsymbol{\theta}^k) \\ \mathbf{z}^{k+1} &= \operatorname{argmin}_{\mathbf{z}} \mathcal{L}_\eta(\boldsymbol{\beta}^{k+1}, \mathbf{z}, \boldsymbol{\theta}^k) \\ \boldsymbol{\theta}^{k+1} &= \boldsymbol{\theta}^k - (\mathbf{z}^{k+1} - \boldsymbol{\beta}^{k+1}) \end{aligned}$$

It is easy to see that $\boldsymbol{\beta}^{k+1}$ has a close form expression and \mathbf{z}^{k+1} is obtained by solving p univariate L_1 penalized problems. More specifically, we have

$$\begin{aligned} \boldsymbol{\beta}^{k+1} &= (\mathbf{X}^T \mathbf{X} / n + \eta \mathbf{I})^{-1} (\mathbf{X}^T \mathbf{Y} / n + \eta \mathbf{z}^k - \eta \boldsymbol{\theta}^k) \\ \mathbf{z}_j^{k+1} &= \operatorname{sgn}(\beta_j^{k+1} + \theta_j^k) \left(|\beta_j^{k+1} + \theta_j^k| - \lambda / \eta \right), j = 1, \dots, p \end{aligned}$$

Iterative Local Adaptive Majorization and Minimization

Iterative local adaptive majorization and minimization is an algorithmic approach to solve the folded concave penalized least-squares problem (3.71) or more generally the penalized quasi-likelihood of the form:

$$f(\boldsymbol{\beta}) + \sum_{j=1}^p p_\lambda(|\beta_j|) \quad (3.142)$$

with both algorithmic and statistical guaranteed, proposed and studied by Fan, Liu, Sun, and Zhang (2018). It combines the local linear approximation (3.130) and the proximal gradient method (3.138) to solve the problem (3.142). More specifically, starting from the initial value $\boldsymbol{\beta}^{(0)} = \mathbf{0}$, we use LLA to case problem (3.142) into the sequence of problems:

$$\hat{\boldsymbol{\beta}}^{(1)} = \operatorname{argmin} \left\{ f(\boldsymbol{\beta}) + \sum_{j=1}^d \lambda_j^{(0)} |\beta_j| \right\}, \quad \text{with } \lambda_j^{(0)} = p'_\lambda(|\hat{\boldsymbol{\beta}}_j^{(0)}|) \quad (3.143)$$

$$\dots\dots\dots \quad (3.144)$$

$$\hat{\boldsymbol{\beta}}^{(t)} = \operatorname{argmin} \left\{ f(\boldsymbol{\beta}) + \sum_{j=1}^d \lambda_j^{(t-1)} |\beta_j| \right\}, \quad \text{with } \lambda_j^{(t-1)} = p'_\lambda(|\hat{\boldsymbol{\beta}}_j^{(t-1)}|) \quad (3.145)$$

Within each problem (3.143) above, we apply proximal gradient method. More specifically, by (3.139), starting from the initial value $\hat{\boldsymbol{\beta}}_{t,0} = \hat{\boldsymbol{\beta}}_{t-1}$, the algorithm used to solve (3.143) utilizes the iterations

$$\hat{\boldsymbol{\beta}}_{t,k} = \left(\hat{\boldsymbol{\beta}}_{t,k-1} - s_{t,k} n^{-1} \mathbf{X}^T (\mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\beta}}_{t,k-1}) - s_{t,k} \lambda \right)_+$$

for $k = 1, \dots, k_t$, where the step size is computed by using $s = \delta^{i_k} s_0$ to check (3.140) in choosing s_k . The flowchat of the algorithm can be summarized in Figure 3.9. This algorithmic approach of statistical estimator is called *I LAMM* by Fan et al. (2018).

Note that the problem (3.143) is convex but not strongly convex. It converges only at a sublinear rate. Hence, it takes longer to get to a consistent neighborhood. Once the estimate is in a consistent neighborhood, from step 2 and on, the solutions are sparse and therefore the function (3.143) is strongly convex in this restricted neighborhood and the algorithmic convergence is exponentially fast (at a linear rate). This leads Fan et al. (2018) to take $k_1 \asymp n / \log p$ and $k_2 \asymp \log n$. In addition, they show that when the number of outer loop $T \asymp \log(\log(p))$, the estimator achieves statistical optimal rates and further iteration will not improve nor deteriorate the statistical errors.

Other Methods and Timeline

There are many other algorithms for computing penalized least-squares problem. For example, matching pursuit, introduced by Mallot and Zhang (1993), is similar to the forward selection algorithm for subset selection. As in the forward selection and LARS, the most correlated variable \mathbf{X}_j (say) with the current residual \mathbf{R} is selected and the univariate regression

$$\mathbf{R} = \beta_j \mathbf{X}_j + \varepsilon$$

is fitted. This is an important deviation from the forward selection in highdimensional regression as the matching pursuit does not compute multiple regression. It is similar but more greedy than the coordinate decent algorithm, as only the most correlated coordinate is chosen. With fitted univariate coefficient $\hat{\beta}_j$, we update the current residual by $\mathbf{R} - \hat{\beta}_j \mathbf{X}_j$. The variables selected as well as coefficients used to compute \mathbf{R} can be recorded along the fit.

Iterated SIS(sure independence screening) introduced in Fan and LV(2008) and extended by Fan, Samworth and Wu (2009) can be regarded as another greedy algorithm for computing folded concave PLS. The basic idea is to iteratively use large scale screening (e.g. marginal screening) and moderate scale selection by using the penalized least-squares. Details will be introduced in Chapter 8.

The *DC* algorithm (An and Tao, 1997) is a general algorithm for minimizing the difference of two convex functions. Suppose that $Q(\boldsymbol{\beta}) = Q_1(\boldsymbol{\beta}) - Q_2(\boldsymbol{\beta})$, where Q_1 and Q_2 are convex. Given the current value $\boldsymbol{\beta}_0$, linearize $Q_2(\boldsymbol{\beta})$ by

$$Q_{2,L}(\boldsymbol{\beta}) = Q_2(\boldsymbol{\beta}_0) + Q_2'(\boldsymbol{\beta}_0)^T (\boldsymbol{\beta} - \boldsymbol{\beta}_0)$$

Now update the minimizer by the convex optimization problem

$$\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta}} \{Q_1(\boldsymbol{\beta}) - Q_{2,L}(\boldsymbol{\beta})\}$$

Note that for any convex function

$$Q_2(\beta) \geq Q_{2,L}(\beta) \quad \text{with} \quad Q_2(\beta_0) = Q_{2,L}(\beta_0)$$

Thus, the DC algorithm is a special case of the MM-algorithm. Hence, its target value should be non-increasing $Q(\beta_{\text{new}}) \leq Q(\beta_0)$ [c.f. (3.129)]. The algorithm has been implemented to support vector machine classifications by Liu, Shen and Doss (2005) and Wu and Liu (2007). It was used by Kim, Choi and Oh (2008) to compute SCAD in which the SCAD penalty function is decomposed as

$$p_\lambda(|\beta|) = \lambda|\beta| - [\lambda|\beta| - p_\lambda(|\beta|)]$$

Agarwal, Negahban and Wainwright (2012) propose the composite gradient descent algorithm. Liu, Yao and Li (2016) propose a mixed integer programming-based global optimization (MIPGO) to solve the class of folded concave penalized least-squares that find a provably global optimal solution. Fan, Liu, Sun and Zhang (2018) propose I-LAMM to simultaneously control of algorithmic complexity and statistical error.

3.1.2 Regularization parameters for PLS

In applications of the folded concave PLS (3.71), one needs to determine the regularization parameter λ . The solution paths such as in Figure 3.7 can help us to choose a model. For example, it is not unreasonable to select a model with $1/\lambda$ somewhat larger than 40 in Figure 3.7. After that point, the model complexity increases substantially and there will be no more variables with large coefficients.

In many situations, one would also like to have data-driven choice of λ . The choice of λ for the L_0 -penalty was addressed in Section 3.1.3. The basic idea of choosing regularization parameters to minimize the estimated prediction error continues to apply. For example, one can choose λ in the folded concave PLS by using cross-validation (3.67). However, other criteria such as AIC and BIC.

utilize the model size m that is specific to L_0 -penalty. We need to generalize this concept of model size, which will be called the degrees of freedom.

3.1.3 Degrees of freedom

To help motivate the definition of degrees of freedom, following Efron (1986) and Efron et al. (2004), we assume that given the covariates \mathbf{X} , \mathbf{Y} has the conditional mean vector $\mu(\mathbf{X})$ (also called regression function) that depends on \mathbf{X} and homoscedastic variance σ^2 . The conditional mean vector μ (whose dependence on \mathbf{X} is suppressed) is unknown and estimated by $\hat{\mu}$, a function of the data (\mathbf{X}, \mathbf{Y}) . Note that

$$\|\mu - \hat{\mu}\|^2 = \|\mathbf{Y} - \hat{\mu}\|^2 - \|\mathbf{Y} - \mu\|^2 + 2(\hat{\mu} - \mu)^T(\mathbf{Y} - \mu) \quad (3.146)$$

Thus, we have Stein's identity: the mean squared error

$$E\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|^2 = E\{\|\mathbf{Y} - \hat{\boldsymbol{\mu}}\|^2 - n\sigma^2\} + 2 \sum_{i=1}^n \text{cov}(\hat{\mu}_i, Y_i) \quad (3.147)$$

and the prediction error

$$E\|\mathbf{Y}^{\text{new}} - \hat{\boldsymbol{\mu}}\|^2 = n\sigma^2 + E\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|^2 = E\{\|\mathbf{Y} - \hat{\boldsymbol{\mu}}\|^2\} + 2df_{\hat{\boldsymbol{\mu}}}\sigma^2 \quad (3.148)$$

with

$$df_{\hat{\boldsymbol{\mu}}} = \sigma^{-2} \sum_{i=1}^n \text{cov}(\hat{\mu}_i, Y_i) \quad (3.149)$$

as the degrees of freedom. If $df_{\hat{\boldsymbol{\mu}}}$ is known and σ^2 is given, a C_p -type of unbiased risk estimation is given by

$$C_p(\hat{\boldsymbol{\mu}}) = \|\mathbf{Y} - \hat{\boldsymbol{\mu}}\|^2 + 2\sigma^2 df_{\hat{\boldsymbol{\mu}}} \quad (3.150)$$

The above formula shows that $df_{\hat{\boldsymbol{\mu}}}$ plays the same role as the number of parameters in (3.65).

For many linear smoothers, their degrees of freedom are indeed known quantities. A linear estimator has the form $\hat{\boldsymbol{\mu}} = S\mathbf{Y}$ with S being a smoother matrix that only depends on \mathbf{X} . See examples given in Section 2.8 of Chapter 2. By independence among \mathbf{Y}_i s, $\text{cov}(\hat{\mu}_i, \mathbf{Y}_i) = \mathbf{S}_{ii}\sigma^2$. From (3.86), it follows that

$$df_{\hat{\boldsymbol{\mu}}} = \frac{1}{\sigma^2} \sum_{i=1}^n \mathbf{S}_{ii}\sigma^2 = \text{tr}(\mathbf{S})$$

We mentioned $\text{tr}(\mathbf{S})$ as the degrees of freedom of the linear smoother S in Chapter 2. Here, a formal justification is provided. In particular, when $\mathbf{S} = \mathbf{X}_1 (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T$, the projection matrix using m variables of the full model, we have

$$df_{\hat{\boldsymbol{\mu}}} = \text{tr}\left(\mathbf{X}_1 (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T\right) = m$$

Therefore, the degrees of freedom formula is an extension of the number of variables used in the classical linear model.

Degrees of freedom can be much more complex for nonlinear model fitting procedures. For example, let us consider the best subset selection. For a given subset size m , the final model always has m variables and one may naively think the degrees of freedom is m . This is in general wrong unless $m = 0$ or $m = p$. This is because the final subset is obtained by exclusively searching over $\binom{p}{m}$ many candidate models. We can not ignore the stochastic nature of the search unless $m = 0$ or $m = p$. A simulation study in Lucas, Fithian and Hastie (2015) shows that the degree of freedom is larger than m and can be

even larger than p . Another interesting and counter-intuitive finding is that the degrees of freedom is not a monotonic increasing function of m which again reflects the complexity due to the stochastic search over $\binom{p}{m}$ many submodels. The same phenomenon is also observed for the degrees of freedom of forward selection.

For least angle regression, Efron et al. (2004) show that under the orthogonal design assumption, the degree of freedom in the m^{th} step of the LARS algorithm is m . This matches our intuition, as at the m^{th} step of the LARS algorithm, m variables are effectively recruited. For a general design matrix, let $\hat{\beta}_\lambda^{\text{lasso}}$ be the Lasso penalized least square estimator with penalization parameter λ . Let $df_\lambda^{\text{lasso}}$ denote its degrees of freedom. Zou, Hastie and Tibshirani (2007) prove a surprising result:

$$df_\lambda^{\text{lasso}} = E \left[\left\| \hat{\beta}_\lambda^{\text{lasso}} \right\|_0 \right] \quad (3.151)$$

Therefore, the number of nonzero estimated coefficients is an exact unbiased estimator of the degrees freedom of the Lasso. The estimation consistency is also established. In theory we view the L_1 PLS as a convex relaxation of L_0 PLS, but their degrees of freedom (model complexity) has very different properties. For the L_0 PLS, the number of nonzero estimated coefficients can severely underestimate the true degrees of freedom. The final model of L_1 PLS is also obtained via a stochastic search, but (3.151) implies that on average the complexity due to stochastic search is zero.

The unbiasedness result is good enough for constructing a C_p type statistic for the Lasso:

$$C_p^{\text{lasso}} = \left\| \mathbf{Y} - \mathbf{X} \hat{\beta}_\lambda^{\text{lasso}} \right\|^2 + 2\sigma^2 \left\| \hat{\beta}_\lambda^{\text{lasso}} \right\|_0 \quad (3.152)$$

which is an exact unbiased estimator of the prediction risk of the Lasso.

Extension of information criteria

Suppose that $\hat{\mu}(\lambda)$ is constructed by using a regularization parameter λ . An extension of the C_p criterion (3.3) and the information criterion (3.4) is

$$C_p(\lambda) = \left\| \mathbf{Y} - \hat{\mu}(\lambda) \right\|^2 + \gamma \sigma^2 df_{\hat{\mu}(\lambda)} \quad (3.153)$$

and

$$\text{IC}(\lambda) = \log \left(\left\| \mathbf{Y} - \hat{\mu}(\lambda) \right\|^2 / n \right) + \gamma df_{\hat{\mu}(\lambda)} / n \quad (3.154)$$

As shown in (3.147), $C_p(\lambda)$ with $\gamma = 2$ is an unbiased estimation of the risk $E \left\| \hat{\mu}(\lambda) - \mu \right\|^2$ except a constant term $-n\sigma^2$. When $\gamma = 2, \log(n)$ and $2\log(p)$ the criteria (3.153) will be called respectively the AIC, BIC, and RIC criterion. With an estimate of σ^2 (see Section 3.7), one can choose λ to minimize (3.153) similarly, we can choose λ to minimize (3.154).

Similarly, one can extend the generalized cross-validation criterion (3.69) to this framework by

$$\text{GCV}(\lambda) = \frac{\|\mathbf{Y} - \hat{\boldsymbol{\mu}}(\lambda)\|^2}{\left(1 - df_{\hat{\boldsymbol{\mu}}(\lambda)}/n\right)^2} \quad (3.155)$$

In particular, when the linear estimator $\hat{\boldsymbol{\mu}}(\lambda) = \mathbf{H}(\lambda)\mathbf{Y}$ is used, by (3.150), we have

$$\begin{aligned} C_p(\lambda) &= \|\mathbf{Y} - \hat{\boldsymbol{\mu}}(\lambda)\|^2 + \gamma\sigma^2 \text{tr}(\mathbf{H}(\lambda)) \\ \text{IC}(\lambda) &= \log(\|\mathbf{Y} - \hat{\boldsymbol{\mu}}(\lambda)\|^2/n) + \gamma \text{tr}(\mathbf{H}(\lambda))/n \end{aligned} \quad (3.156)$$

and

$$\text{GCV}(\lambda) = \frac{\|\mathbf{Y} - \hat{\boldsymbol{\mu}}(\lambda)\|^2}{[1 - \text{tr}(\mathbf{H}(\lambda))/n]^2} \quad (3.157)$$

As mentioned in Section 3.1.3, an advantage of the information criterion and GCV is that no estimation of σ^2 is needed, but this can lead to inaccurate estimation of prediction error.

Application to PLS estimators

For PLS estimator (3.71), $\hat{\boldsymbol{\mu}}(\lambda)$ is not linear in \mathbf{Y} . Some approximations are needed. For example, using the LQA approximation, Fan and Li(2001) regard (3.127) as a linear smoother with (recalling $\boldsymbol{\mu}(\lambda) = \mathbf{X}\hat{\boldsymbol{\beta}}(\lambda)$)

$$\mathbf{H}(\lambda) = \mathbf{X} \left(\mathbf{X}^T \mathbf{X} + n \text{diag} \left\{ p'_\lambda \left(\hat{\beta}_j(\lambda) \right) / \left| \hat{\beta}_j(\lambda) \right| \right\} \right)^{-1} \mathbf{X}^T$$

and choose λ by GCV(3.157). For the LARS-Lasso algorithm, as mentioned at the end of Section 3.6.1, Zou, Hastie and Tibshirani (2007) demonstrate that the degree of freedom is the same as the number of variables used in the LARS algorithm. This motivates Wang, Li and Tsai (2007) and Wang and Leng (2007) to use directly $\|\hat{\boldsymbol{\beta}}\|_0$ as the degree of freedom. This leads to the definition of modified information criterion as

$$\text{IC}^*(\lambda) = \log \left(\left\| \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda \right\|^2 / n \right) + \gamma \frac{\|\hat{\boldsymbol{\beta}}_\lambda\|_0}{n} C_n \quad (3.158)$$

for a sequence of constants C_n . It has been shown by Wang, Li and Tsai (2007) that SCAD with AIC ($\gamma = 2, C_n = 1$) yields an inconsistent model (too many false positives) while BIC ($\gamma = \log n, C_n = 1$) yields a consistent estimation of the model when p is fixed. See also Wang and Leng (2007) for similar model selection results. Wang, Li and Leng (2009) show that the modified BIC(3.99) with $\gamma = \log n$ and $C_n \rightarrow \infty$ produces consistent model selection when SCAD is used. For high-dimensional model selection, Chen and Chen (2008) propose an extended BIC, which adds a multiple of the logarithm of the prior probability of a submodel to BIC. Here, they successfully establish its model selection consistency.

Residual variance and refitted cross-validation

Estimation of noise variance σ^2 is fundamental in statistical inference. It is prominently featured in the statistical inference of regression coefficients. It is also important for variable selection using the C_p criterion (??). It provides a benchmark for forecasting error when an oracle actually knows the underlying regression function. It also arises from genomewide association studies (see Fan, Han and Gu, 2012). In the classical linear model as in Chapter 2, the noise variance is estimated by the residual sum of squares divided by $n - p$. This is not applicable to the high-dimensional situations in which $p > n$. In fact, as demonstrated in Section 1.3.3 (see Figure 1.9 there), the impact of spurious correlation on residual variance estimation can be very large. This leads us to introducing the refitted cross-validation.

In this section, we introduce methods for estimating σ^2 in the highdimensional framework. Throughout this section, we assume the linear model (3.1) with homoscedastic variance σ^2 .

Residual variance of Lasso

A natural estimator of σ^2 is the residual variance of penalized least-squares estimators. As demonstrated in Section 3.3.2, Lasso has a good risk property. We therefore examine when its residual variance gives a consistent estimator of σ^2 . Recall that the theoretical risk and empirical risk are defined by

$$R(\beta) = E \left(Y - \mathbf{X}^T \beta \right)^2 \quad \text{and} \quad R_n(\beta) = n^{-1} \sum_{i=1}^n \left(Y_i - \mathbf{x}_i^T \beta \right)^2$$

Let $\hat{\beta}$ be the solution to the Lasso problem (3.99) and c be sufficiently large so that $\|\beta_0\|_1 \leq c$. Then, $R_n(\beta_0) \geq R_n(\hat{\beta})$. Using this, we have

$$\begin{aligned} R(\beta_0) - R_n(\hat{\beta}) &= [R(\beta_0) - R_n(\beta_0)] + [R_n(\beta_0) - R_n(\hat{\beta})] \\ &\geq R(\beta_0) - R_n(\beta_0) \\ &\geq - \sup_{\|\beta\|_1 \leq c} |R(\beta) - R_n(\beta)| \end{aligned}$$

On the other hand, by using $R(\beta_0) \leq R(\hat{\beta})$, we have

$$R(\beta_0) - R_n(\hat{\beta}) \leq R(\hat{\beta}) - R_n(\hat{\beta}) \leq \sup_{\|\beta\|_1 \leq c} |R(\beta) - R_n(\beta)|$$

Therefore,

$$\left| R(\beta_0) - R_n(\hat{\beta}) \right| \leq \sup_{\|\beta\|_1 \leq c} |R(\beta) - R_n(\beta)|$$

By (3.100), we conclude that

$$\left| R(\beta_0) - R_n(\hat{\beta}) \right| \leq (1 + c)^2 \|\Sigma^* - \mathbf{S}_n^*\|_\infty \quad (3.159)$$

provided that $\|\beta_0\|_1 \leq c$. In other words, the average residual sum of squares of Lasso

$$\hat{\sigma}_{\text{Lasso}}^2 = n^{-1} \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2$$

provides a consistent estimation of σ^2 , if the righthand side of (3.159) goes to zero and $\|\beta_0\|_1 \leq c$

$$R(\beta_0) = \sigma^2 \quad \text{and} \quad R_n(\hat{\beta}) = \hat{\sigma}_{\text{Lasso}}^2$$

As shown in Chapter 11, $\|\Sigma - \hat{\Sigma}_n\|_\infty = O_P(\sqrt{(\log p)/n})$ for the data with Gaussian tails. That means that $\hat{\sigma}_{\text{Lasso}}^2$ is consistent when

$$\|\beta_0\|_1 \leq c = o\left((n/\log p)^{1/4}\right) \quad (3.160)$$

Condition (3.160) is actually very restrictive. It requires the number of significantly nonzero components to be an order of magnitude smaller than $(n/\log p)^{1/4}$. Even when that condition holds, $\hat{\sigma}_{\text{Lasso}}^2$ is only a consistent estimator and can be biased or not optimal. This leads us to consider refitted cross-validation.

Refitted cross-validation

Refitted cross-validation (RCV) was introduced by Fan, Guo and Hao (2012) to deal with the spurious correlation induced by data-driven model selection. In high-dimensional regression models, model selection consistency is very hard to achieve. When some important variables are missed in the selected model, they create a non-negligible bias in estimating σ^2 . When spurious variables are selected into the model, they are likely to predict the realized but unobserved noise ε . Hence, the residual variance will seriously underestimate σ^2 as shown in Section 1.3.3. Note that our observed data follow

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \text{ and } \text{Var}(\varepsilon) = \sigma^2 \mathbf{I}_n$$

Even though we only observe (\mathbf{X}, \mathbf{Y}) , ε is a realized vector in R^n . It can have a spurious correlation with a subgroup of variables \mathbf{X}_S , namely, there exists a vector β_S such that $\mathbf{X}_S \beta_S$ and ε are highly correlated. This can occur easily when the number of predictors p is large as shown in Section 1.3.3. In this case, \mathbf{X}_S can be seen by a model selection technique as important variables. A way to validate the model is to collect new data to see whether the variables in the set S still correlated highly with the newly observed Y . But this is infeasible in many studies and is often replaced by the data splitting technique.

RCV splits data evenly at random into two halves, where we use the first half of the data along with a model selection technique to get a submodel. Then, we fit this submodel to the second half of the data using the ordinary least-squares and get the residual variance. Next we switch the role of the first and the second half of the data and take the average of the two residual variance estimates. The idea differs importantly

from cross-validation in that the refitting in the second stage reduces the influence of the spurious variables selected in the first stage.

We now describe the procedure in detail. Let datasets $(\mathbf{Y}^{(1)}, \mathbf{X}^{(1)})$ and $(\mathbf{Y}^{(2)}, \mathbf{X}^{(2)})$ be two randomly split data. Let $\hat{\mathcal{S}}_1$ be the set of selected variables using data $(\mathbf{Y}^{(1)}, \mathbf{X}^{(1)})$. The variance σ^2 is then estimated by the residual variance of the least-squares estimate using the second dataset along with variables in $\hat{\mathcal{S}}_1$ (only the selected model, not the data, from the first stage is carried to the fit in the second stage), namely,

$$\hat{\sigma}_1^2 = \frac{(\mathbf{Y}^{(2)})^T (\mathbf{I}_{n/2} - \mathbf{P}_{\hat{\mathcal{S}}_1}^{(2)}) \mathbf{Y}^{(2)}}{n/2 - |\hat{\mathcal{S}}_1|}, \quad \mathbf{P}_{\hat{\mathcal{S}}_1}^{(2)} = \mathbf{X}_{\hat{\mathcal{S}}_1}^{(2)} (\mathbf{X}_{\hat{\mathcal{S}}_1}^{(2)T} \mathbf{X}_{\hat{\mathcal{S}}_1}^{(2)})^{-1} \mathbf{X}_{\hat{\mathcal{S}}_1}^{(2)T} \quad (3.161)$$

Compare residual variance estimation in (3.4). Switching the role of the first and second half, we get a second estimate

$$\hat{\sigma}_2^2 = \frac{(\mathbf{Y}^{(1)})^T (\mathbf{I}_{n/2} - \mathbf{P}_{\hat{\mathcal{S}}_2}^{(1)}) \mathbf{Y}^{(1)}}{n/2 - |\hat{\mathcal{S}}_2|}$$

We define the final estimator as the simple average

$$\hat{\sigma}_{\text{RCV}}^2 = (\hat{\sigma}_1^2 + \hat{\sigma}_2^2) / 2$$

or the weighted average defined by

$$\hat{\sigma}_{\text{wRCV}}^2 = \frac{(\mathbf{Y}^{(2)})^T (\mathbf{I}_{n/2} - \mathbf{P}_{\hat{\mathcal{S}}_1}^{(2)}) \mathbf{Y}^{(2)} + (\mathbf{Y}^{(1)})^T (\mathbf{I}_{n/2} - \mathbf{P}_{\hat{\mathcal{S}}_2}^{(1)}) \mathbf{Y}^{(1)}}{n - |\hat{\mathcal{S}}_1| - |\hat{\mathcal{S}}_2|} \quad (3.162)$$

The latter takes into account the degrees of freedom used in fitting the linear model in the second stage. We can now randomly divide the data multiple times and take the average of the resulting RCV estimates.

The point of refitting is that even though $\hat{\mathcal{S}}_1$ may contain some unimportant variables that are highly correlated with $\varepsilon^{(1)}$, they play minor roles in estimating σ^2 in the second stage since they are unrelated with the realized noise vector $\varepsilon^{(2)}$ in the second half of data set. Furthermore, even when some important variables are missed in $\hat{\mathcal{S}}_1$, they still have a good chance of being well approximated by the other variables in $\hat{\mathcal{S}}_1$. Thanks to the refitting in the second stage, the best linear approximation of those selected variables is used to reduce the biases in (3.161).

Unlike cross-validation, the second half of data also plays an important role in fitting. Therefore, its size can not be too small. For example, it should be bigger than $|\hat{\mathcal{S}}_1|$. Yet, larger set \mathcal{S}_1 gives a better chance of sure screening (no false negatives) and hence reduces the bias of estimator (3.161). RCV is applicable to any variable selection rule, including the marginal screening procedure in Chapter 8. Fan, Guo and Hao (2012) show

that under some mild conditions, the method yields an asymptotic efficient estimator of $\hat{\sigma}^2$. In particular, it can handle intrinsic model size $s = o(n)$, much higher than (3.160), when folded concave PLS is used. They verified the theoretical results by numerous simulations. See Section 8.7 for further developments.

3.1.4 Extensions to Nonparametric Modeling

The fundamental ideas of the penalized least squares can be easily extended to the more flexible nonparametric models. This section illustrates the versatility of high-dimensional linear techniques.

Structured nonparametric models

A popular modeling strategy is the generalized additive model (GAM):

$$Y = \mu + f_1(X_1) + \cdots + f_p(X_p) + \varepsilon \quad (3.163)$$

This model was introduced by Stone (1985) to deal with the "curse of dimensionality" in multivariate nonparametric modeling and was thoroughly treated in the book by Hastie and Tibshirani (1990). A simple way to fit the additive model (3.163) is to expand the regression function $f_j(x)$ into a basis:

$$f_j(x) = \sum_{k=1}^{K_j} \beta_{jk} B_{jk}(x) \quad (3.164)$$

where $\{B_{jk}(x)\}_{k=1}^{K_j}$ are the basis functions (e.g. a B-spline basis with certain number of knots) for variable X_j . See Section 2.5.2. Substituting the expansion into (3.163) yields

$$Y = \mu + \{\beta_{1,1} B_{1,1}(X_1) + \cdots + \beta_{1,K_1} B_{1,K_1}(X_1)\} + \cdots \quad (3.165)$$

$$+ \{\beta_{p,1} B_{p,1}(X_p) + \cdots + \beta_{p,K_p} B_{p,K_p}(X_p)\} + \varepsilon \quad (3.166)$$

Treating the basis functions $\{B_{j,k}(X_j) : k = 1, \dots, K_j, j = 1, \dots, p\}$ as predictors, (3.165) is a high-dimensional linear model. By imposing a sparsity assumption, we assume that only a few f_j functions actually enter the model. So, many β coefficients are zero. Therefore, we can employ penalized folded concave PLS (3.71) to solve this problem. Another selection method is via the group penalization. See, for example, the PLASM algorithm in Baskin (1999) and the SpAM algorithm in Ravikumar, Liu, Lafferty and Wasserman (2007).

The varying-coefficient model is another widely-used nonparametric extension of the multiple linear regression model. Conditioning on an exposure variable U , the response and covariates follow a linear model. In other words,

$$Y = \beta_0(U) + \beta_1(U)X_1 + \cdots + \beta_p(U)X_p + \varepsilon \quad (3.167)$$

The model allows regression coefficients to vary with the level of exposure U , which is an observed covariate variable such as age, time, or gene expression. For a survey and various applications, see Fan and Zhang (2008). Expanding the coefficient functions similar to (3.164), we can write

$$Y = \sum_{j=0}^p \left\{ \sum_{k=1}^{K_j} \beta_{j,k} B_{j,k}(U) X_j \right\} + \varepsilon \quad (3.168)$$

where $X_0 = 1$. By regarding variables $\{B_{j,k}(U)X_j, k = 1, \dots, K_j, j = 0, \dots, p\}$ as new predictors, model (3.168) is a high-dimensional linear model. The sparsity assumption says that only a few variables should be in the model (3.167) which implies many zero β coefficients in (3.168). Again, we can employ penalized folded concave PLS (3.71) to do variable selection or use the group selection method.

Group penalty

The penalized least-squares estimate to the nonparametric models in Section 3.8 results in term-by-term selection of the basis functions. In theory, when the folded concave penalty is employed, the selection should be fine. On the other hand, the term-by-term selection does not fully utilize the sparsity assumption of the functions. In both additive model and varying coefficient model, a zero function implies that the whole group of its associated coefficients in the basis expansion is zero. Therefore, model selection techniques should ideally keep or kill a group of coefficients at the same time.

Group penalty was proposed in Antoniadis and Fan (2001, page 966) to keep or kill a block of wavelets coefficients. It was employed by Lin and Zhang (2006) for component selection in smoothing spline regression models, including the additive model as a special case. Their COSSO algorithm iterates between a smoothing spline fit and a non-negative garrote shrinkage and selection. A special case of COSSO becomes a more familiar group lasso regression formulation considered in Yuan and Lin (2006) who named the group penalty group-lasso.

Let $\{\mathbf{x}_j\}_{j=1}^p$ be p groups of variables, each consisting of K_j variables. Consider a generic linear model

$$Y = \sum_{j=1}^p \mathbf{x}_j^T \boldsymbol{\beta}_j + \varepsilon \quad (3.169)$$

Two examples of (3.169) are (3.165) and (3.168) in which \mathbf{x}_j represents K_j spline bases and $\boldsymbol{\beta}_j$ represents their associated coefficients. In matrix form, the observed data based on a sample of size n follow the model

$$\mathbf{Y} = \sum_{j=1}^p \mathbf{X}_j \boldsymbol{\beta}_j + \varepsilon \quad (3.170)$$

where \mathbf{X}_j is $n \times K_j$ design matrix of variables \mathbf{x}_j . The group penalized least-squares is to minimize

$$\frac{1}{2n} \left\| \mathbf{Y} - \sum_{j=1}^p \mathbf{X}_j \boldsymbol{\beta}_j \right\|^2 + \sum_{j=1}^p p_\lambda \left(\left\| \boldsymbol{\beta}_j \right\|_{\mathbf{W}_j} \right) \quad (3.171)$$

where $p_\lambda(\cdot)$ is a penalty function and

$$\left\| \boldsymbol{\beta}_j \right\|_{\mathbf{W}_j} = \sqrt{\boldsymbol{\beta}_j^T \mathbf{W}_j \boldsymbol{\beta}_j}$$

is a generalized norm with a semi-definite matrix \mathbf{W}_j . In many applications, one takes $\mathbf{W}_j = \mathbf{I}_{K_j}$, resulting in

$$\frac{1}{2n} \left\| \mathbf{Y} - \sum_{j=1}^p \mathbf{X}_j \boldsymbol{\beta}_j \right\|^2 + \sum_{j=1}^p p_\lambda \left(\left\| \boldsymbol{\beta}_j \right\| \right) \quad (3.172)$$

For example, the group lasso is defined as

$$\frac{1}{2n} \left\| \mathbf{Y} - \sum_{j=1}^p \mathbf{X}_j \boldsymbol{\beta}_j \right\|^2 + \lambda \sum_{j=1}^p K_j^{1/2} \left\| \boldsymbol{\beta}_j \right\| \quad (3.173)$$

The extra factor $K_j^{1/2}$ is included to balance the impact of group size.

The group-lasso (3.173) was proposed by Baskin (1999) for variable selection in the additive model. Turlach, Venables and Wright (2005) also used the group-lasso for simultaneous variable selection in multiple responses linear regression, a example of multi-task learning.

Assuming a group-wise orthogonality condition, that is, $\mathbf{X}_j^T \mathbf{X}_j = n \mathbf{I}_{K_j}$ for all j , Yuan and Lin (2006) used a group descent algorithm to solve (3.172). Similar to coordinate descent, we update the estimate one group at a time. Consider the coefficients of group j while holding all other coefficients fixed. Then, by $\mathbf{X}_j^T \mathbf{X}_j = n \mathbf{I}_{K_j}$ (3.112) can be written as

$$\frac{1}{2n} \left\| \mathbf{Y}_{-j} - \mathbf{X}_j \hat{\boldsymbol{\beta}}_{-j} \right\|^2 + \frac{1}{2} \left\| \hat{\boldsymbol{\beta}}_{-j} - \boldsymbol{\beta}_j \right\|^2 + \sum_{k=1}^p p_\lambda \left(\left\| \boldsymbol{\beta}_k \right\| \right) \quad (3.174)$$

where $\mathbf{Y}_{-j} = \mathbf{Y} - \sum_{k \neq j} \mathbf{X}_k \boldsymbol{\beta}_k$ and $\hat{\boldsymbol{\beta}}_{-j} = n^{-1} \mathbf{X}_j^T \mathbf{Y}_{-j}$. This problem was solved by Antoniadis and Fan (2001, page 966). They observed that

$$\min_{\boldsymbol{\beta}_j} \left\{ \frac{1}{2} \left\| \hat{\boldsymbol{\beta}}_{-j} - \boldsymbol{\beta}_j \right\|^2 + p_\lambda \left(\left\| \boldsymbol{\beta}_j \right\| \right) \right\} = \min_r \left\{ \frac{1}{2} \min_{\left\| \boldsymbol{\beta}_j \right\|=r} \left\| \hat{\boldsymbol{\beta}}_{-j} - \boldsymbol{\beta}_j \right\|^2 + p_\lambda(r) \right\} \quad (3.175)$$

The inner bracket is minimized at $\hat{\boldsymbol{\beta}}_{j,r} = r \hat{\boldsymbol{\beta}}_{-j} / \left\| \hat{\boldsymbol{\beta}}_{-j} \right\|$. Substituting this into (3.175), the problem becomes

$$\min_r \left\{ \frac{1}{2} \left(\left\| \hat{\boldsymbol{\beta}}_{-j} \right\| - r \right)^2 + p_\lambda(r) \right\} \quad (3.176)$$

Problem (3.176) is identical to problem (3.74), whose solution is denoted by $\hat{\theta}(\|\hat{\beta}_{-j}\|)$. For the L_1 , SCAD and MCP, the explicit solutions are given respectively by (3.80) – (3.82). With this notation, we have

$$\hat{\beta}_j = \frac{\hat{\theta}(\|\hat{\beta}_{-j}\|)}{\|\hat{\beta}_{-j}\|} \hat{\beta}_{-j} \quad (3.177)$$

In particular, for the L_1 -penalty,

$$\hat{\beta}_j = \left(1 - \frac{\lambda}{\|\hat{\beta}_{-j}\|}\right)_+ \hat{\beta}_{-j}$$

and for the hard-thresholding penalty

$$\hat{\beta}_j = I(\|\hat{\beta}_{-j}\| \geq \lambda) \hat{\beta}_{-j}$$

These formulas were given by Antoniadis and Fan (2001, page 966). They clearly show that the strength of the group estimates is pulled together to decide whether or not to keep a group of coefficients.

The groupwise orthogonality condition is in fact not natural and necessary to consider. Suppose that the condition holds for the data $(\mathbf{Y}_i, \mathbf{X}_i), 1 \leq i \leq n$. If we bootstrap the data or do cross-validation to selection λ , the groupwise orthogonality condition easily fails on the perturbed dataset. For computational considerations, the groupwise orthogonality condition is not needed for using the group descent algorithm. Several algorithms for solving the Lasso regression, such as ISTA, FISTA and ADMM, can be readily used to solve the group-lasso regression with a general design matrix. We omit the details here.

Applications

We now illustrate high-dimensional statistical modeling using the monthly house price appreciations (HPA) for 352 counties in the United States. The housing price appreciation is computed based on monthly repeated sales. These 352 counties have the largest repeated sales and hence their measurements are more reliable. The spatial correlations of these 352 HPAs, based on the data in the period from January 2000 to December 2009, are presented in Figure 1.4

To take advantage of the spatial correlation in their prediction, Fan, Lv and Qi (2011) utilize the following high-dimensional time-series regression. Let Y_t^i be the HPA in county i at time t and $\mathbf{X}_{i,t}$ be the observable factors

that drive the market. In the application below, $\mathbf{X}_{i,t}$ will be taken as the national HPA, the returns of the national house price index that drives the overall housing markets. They used the following s -period ahead county-level forecast model:

$$Y_{t+s}^i = \sum_{j=1}^p b_{ij} Y_t^j + \mathbf{X}_{i,t}^T \beta_i + \varepsilon_{t+s}^i, \quad i = 1, \dots, p$$

where $p = 352$ and b_{ij} and β_i are regression coefficients. In this model, we allow neighboring HPAs to influence the future housing price, but we do not know which counties have such prediction power. This leads to the following PLS problem: For each given county i

$$\min_{\{b_{ij}, j=1, \dots, p, \beta_i\}} \sum_{t=1}^{T-s} \left(Y_{t+s}^i - \mathbf{x}_{i,t}^T \beta_i - \sum_{j=1}^p b_{ij} Y_t^j \right)^2 + \sum_{j=1}^p w_{ij} p_\lambda (|b_{ij}|)$$

where the weights w_{ij} are chosen according to the geographical distances between counties i and j . The weights are used to discourage HPAs from remote regions from being used in the prediction. The non-vanishing coefficients represent the selected neighbors that are useful for predicting HPA at county i

Monthly HPA data from January 2000 to December 2009 were used to fit model (3.118) for each county with $s = 1$. The top panel of Figure 3.11 highlights the selected neighborhood HPAs used in the prediction. For each county i , only 3 – 4 neighboring counties are chosen on average, which is reasonable. Figure 3.11 (bottom left) presents the spatial correlations of the residuals using model (3.118). No pattern can be found, which indicates the spatial correlations have already been explained by the neighborhood HPAs. In contrast, if we ignore the neighborhood selection (namely, setting $b_{ij} = 0, \forall i \neq j$), which is a lower-dimensional problem and will be referred to as the OLS estimate, the spatial correlations of the residuals are visible (bottom right). This provides additional evidence on the effectiveness of the neighborhood selection by PLS.


We now compare the forecasting power of the PLS with OLS. Training sample covers the data for 2000.1 – 2005.12, and the test period is 2006.1 – 2009.12. Fan, Lv and Qi (2011) carried out prediction throughout next 3 years in the following manner. For the short-term prediction horizons s from 1 to 6 months, each month is predicted separately using model (3.118); for the time horizon of 7 – 36 months, only the average HPA over 6-month periods (e.g. months 7 – 12, 13 – 18, etc.) is predicted. This increases the stability of the prediction. More precisely, for each of the 6 consecutive months (e.g. months 13 – 18), they obtained a forecast of average HPA during the 6 months using PLS with historical 6-month average HPAs as a training sample. They treated the (annualized) 6-month average as forecast of the middle month of the 6-month period (e.g. month 15.5) and linearly interpolated the months in between. The discounted aggregated squared errors were used as a measure of overall performance of the prediction for county i

$$\text{Forecast Error}_i = \sum_{s=1}^{\tau} \rho^s \left(\hat{Y}_{T+s}^i - Y_{T+s}^i \right)^2, \quad \rho = 0.95 \quad (3.178)$$

where τ is the time horizon to be predicted. The results in Figure 3.12 show that over 352 counties, the sparse regression model (??) with neighborhood information performs on average 30% better in terms of prediction error than the model without using the neighborhood information. Figure 1.4 compares forecasts using OLS with only the national

HPA (blue) and PLS with additional neighborhood information (red) for the largest counties with the historical HPAs (black).

How good is a prediction method? The residual standard deviation σ provides a benchmark measure when the ideal prediction rule is used. To illustrate this, we estimate σ for one-step forecast in San Francisco and Los Angeles, using the HPA data from January 1998 to December 2005 (96 months). The RCV estimates, as a function of the selected model size s , are shown in Figure 3.13. The naive estimates, which compute directly the residual variances, decrease with s due to spurious correlation. On the other hand, the RCV gives reasonably stable estimates for a range of selected models. The benchmarks of prediction errors for both San Francisco and Los Angeles regions are about .53%, comparing the standard deviations of month over month variations of HPAs 1.08% and 1.69%, respectively. In contrast, the rolling one-step prediction errors over 12 months in 2006 are .67% and .86% for San Francisco and Los Angeles areas, respectively. They are clearly larger than the benchmark as expected, but considerably smaller than the standard deviations, which used no variables to forecast. They also show that some small room of improvements in the PLS are possible.



4. Penalized Least Squares: Properties

This chapter describes properties of PLS methods in linear regression models with a large number of covariate variables. We will study the performance of such methods in prediction, coefficient estimation, and variable selection under proper regularity conditions on the noise level, the sparsity of regression coefficients, and the covariance of covariate variables. To make reading easier, we defer some lengthier proofs to the end of each section and make this chapter a self-contained chapter, despite some repetition and slightly modified notation.

4.0.1 Performance Benchmarks

This section provides a general description of theoretical objectives of penalized least squares estimation along with some basic concepts and terminologies for studying such methods in high-dimension.

Suppose we observe covariates $X_{ij}, 1 \leq j \leq p$, and a response variable Y_i from the i -th data point in the sample, $i = 1, \dots, n$. As in the linear regression model (3.1), the covariates and response variable satisfy the relationship

$$Y_i = \sum_{j=1}^p X_{ij}\beta_j + \varepsilon_i$$

In vector notation, it is written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{4.1}$$

See Section 2.1. Unless otherwise stated, the design matrix \mathbf{X} is considered as determinis-

tic. Of course, in the case of random design, the noise vector ε is assumed to be independent of \mathbf{X} .

We are interested in the performance of PLS in the case of large p , including $p \gg n$. Thus, unless otherwise stated, $\{p, \mathbf{X}, \boldsymbol{\beta}, \varepsilon\}$ are all allowed to depend on n and $p = p_n \rightarrow \infty$ as $n \rightarrow \infty$. We denote by \mathbf{X}_j the j -th column $(X_{1j}, \dots, X_{nj})^T$ of the design, $\mathbf{X}_A = (\mathbf{X}_j, j \in A)$ the sub-matrix of the design with variables in A for subsets $A \subseteq \{1, \dots, p\}$, $\bar{\mathbf{X}} = \mathbf{X}^T \mathbf{X} / n$ the normalized Gram matrix, $\bar{\mathbf{X}}_{A,B} = \mathbf{X}_A^T \mathbf{X}_B / n$ its subblocks, and \mathbf{P}_A the orthogonal projection to the range of \mathbf{X}_A . Likewise, we denote by $\mathbf{v}_A = (v_j, j \in A)^T$ the subvector of $\mathbf{v} = (v_1, \dots, v_p)^T$ with indices in A . We denote by $\|\mathbf{v}\|_q = \{\sum_{i=1}^n |v_i|^q\}^{1/q}$ the ℓ_q norm for $1 \leq q < \infty$, with the usual extension $\|\mathbf{v}\|_\infty = \max_i |v_i|$, $\text{supp}(\mathbf{v}) = \{j : v_j \neq 0\}$ the support of a vector \mathbf{v} , and $\|\mathbf{v}\|_0 = \#\{j : v_j \neq 0\}$ the size of the support. Denote by $\phi_{\min}(\mathbf{M})$ and $\phi_{\max}(\mathbf{M})$ the smallest and largest singular values of a matrix \mathbf{M} , respectively.

Let $\mathbf{A} = (a_{ij})$ be an $m \times n$ matrix. We use $\|\mathbf{A}\|_{q,r}$ to denote the ℓ_q to ℓ_r operator norm:

$$\|\mathbf{A}\|_{q,r} = \max_{\|\mathbf{u}\|_q=1} \|\mathbf{A}\mathbf{u}\|_r \quad (4.2)$$

When $q = r$, it is denoted as $\|\mathbf{A}\|_q$. In particular, $\|\mathbf{A}\|_{2,2} = \phi_{\max}(\mathbf{A}^T \mathbf{A})^{1/2}$ and will also be denoted as $\|\mathbf{A}\|$ and

$$\|\mathbf{A}\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^m |a_{ij}| \quad \text{and} \quad \|\mathbf{A}\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^n |a_{ij}| \quad (4.3)$$

are the maximum L_1 -norm of columns and rows, respectively. In particular, for a symmetric matrix \mathbf{A} , $\|\mathbf{A}\|_1 = \|\mathbf{A}\|_\infty$. It also holds that

$$\|\mathbf{A}\|_2 \leq \sqrt{\|\mathbf{A}\|_1 \|\mathbf{A}\|_\infty} \quad (4.4)$$

Performance measures

Here we describe performance measures for prediction, coefficient estimation and variable selection. The goal of prediction is to estimate the response Y at a future design point (X_1, \dots, X_p) , where (X_1, \dots, X_p, Y) is assumed to be independent of the data (\mathbf{X}, \mathbf{Y}) in (4.1) but follows the same model,

$$Y = \sum_{j=1}^p X_j \beta_j + \varepsilon, \quad \text{with} \quad E\varepsilon = 0$$

We measure the performance of a predictor $\hat{\boldsymbol{\beta}}$ by the mean squared prediction error,

$$E \left[\left(Y - \sum_{j=1}^p X_j \hat{\beta}_j \right)^2 \mid \hat{\boldsymbol{\beta}} \right]$$

It follows from the independence between the future observations and the current data that this error measure can be decomposed as

$$E \left[\left(Y - \sum_{j=1}^p X_j \hat{\beta}_j \right)^2 \mid \hat{\boldsymbol{\beta}} \right] = E\varepsilon^2 + E \left[\left(\sum_{j=1}^p X_j \hat{\beta}_j - \sum_{j=1}^p X_j \beta_j \right)^2 \mid \hat{\boldsymbol{\beta}} \right]$$

Thus, the true β is an optimal predictor, and minimizing the mean squared prediction error is equivalent to minimizing

$$R_{\text{pred}}(\hat{\beta}, \beta; X_1, \dots, X_p) = \left(\sum_{j=1}^p X_j \hat{\beta}_j - \sum_{j=1}^p X_j \beta_j \right)^2$$

The above quantity can be viewed as prediction regret of not knowing β at design point (X_1, \dots, X_p) , because it is the difference between the mean squared prediction errors of a predictor $\hat{\beta}$ and the unknown optimal predictor β . For simplicity, we typically further assume that the future design point resembles the design points in the current data in the following sense. For deterministic designs, we assume unless otherwise specified that (X_1, \dots, X_p) is equally likely to be any of the n current design points (X_{i1}, \dots, X_{ip}) in (4.1), so that the expected prediction regret given the data (\mathbf{X}, \mathbf{Y}) is

$$R_{\text{pred}}(\hat{\beta}, \beta) = \frac{1}{n} \sum_{i=1}^n R_{\text{pred}}(\hat{\beta}; X_{i1}, \dots, X_{ip}) = \|\mathbf{X}\hat{\beta} - \mathbf{X}\beta\|_2^2 / n \quad (4.5)$$

We will simply call this quantity (4.5) prediction error. For random designs, we assume that \mathbf{X} has independent and identically distributed (*iid*) rows from a population with a second moment structure $\Sigma = (\text{E}X_{ij}X_{ik})_{p \times p}$. In this case, we assume that the future design point comes from the same population, so that the expected prediction regret given (\mathbf{X}, \mathbf{Y}) is

$$\begin{aligned} R_{\text{pred}}(\hat{\beta}, \beta) &= \text{E} \left[R_{\text{pred}}(\hat{\beta}; X_1, \dots, X_p) \mid \hat{\beta} \right] \\ &= (\hat{\beta} - \beta)^T \Sigma (\hat{\beta} - \beta) \end{aligned} \quad (4.6)$$

As in Section 3.3.2, an estimator $\hat{\beta}$ is persistent if (4.5) or (4.6) converges to zero respectively for deterministic and random designs (Greenshtein and Ritov, 2004).

We can also measure estimation performance of $\hat{\beta}$ with the ℓ_q estimation error

$$\|\hat{\beta} - \beta\|_q = \left(\sum_{j=1}^p |\hat{\beta}_j - \beta_j|^q \right)^{1/q}$$

This quantity is closely related to the prediction regret $R_{\text{pred}}(\hat{\beta}, \beta; X_1, \dots, X_p)$ through the duality between the ℓ_q and $\ell_{q/(q-1)}$ norms as

$$\|\hat{\beta} - \beta\|_q^2 = \max \left\{ \left(\sum_{j=1}^p X_j \hat{\beta}_j - \sum_{j=1}^p X_j \beta_j \right)^2 : \sum_{j=1}^p |X_j|^{q/(q-1)} \leq 1 \right\} \quad (4.7)$$

is the maximum prediction regret for a deterministic future design vector in the unit $\ell_{q/(q-1)}$ ball. In this sense, the ℓ_q estimation error is a conservative prediction error without assuming the resemblance of the future and current design points. For random designs, the persistency property with (4.6) is equivalent to the convergence of the ℓ_2 estimation error to zero, $\|\hat{\beta} - \beta\|_2 \rightarrow 0$, when the eigenvalues of the population covariance matrix Σ are uniformly bounded away from zero and infinity.

The problem of variable selection is essentially the estimation of the support set of the true β , or equivalently the true model,

$$\text{supp}(\beta) = \{j : \beta_j \neq 0\}$$

Here are some commonly used performance measures for variable selection.

Definition 4.0.1 Definition 4.1 Loss functions for variable selection.

1. Incorrect sign selection: $I\{\text{sgn}(\hat{\beta}) \neq \text{sgn}(\beta)\}$ with $\text{sgn}(0) = 0$
2. Incorrect selection: $I\{\text{supp}(\hat{\beta}) \neq \text{supp}(\beta)\}$
3. False positive: $\text{FP}(\hat{\beta}) = |\text{supp}(\hat{\beta}) \setminus \text{supp}(\beta)|$
4. False negative: $\text{FN}(\hat{\beta}) = |\text{supp}(\beta) \setminus \text{supp}(\hat{\beta})|$
5. Total miss: $\text{TM}(\hat{\beta}) = \text{FP}(\hat{\beta}) + \text{FN}(\hat{\beta})$
6. Model size: $\|\hat{\beta}\|_0 = |\text{supp}(\hat{\beta})|$
7. Family wise error rate (FWER) : $P\{\text{FP}(\hat{\beta}) > 0\} = P\{\exists j : \hat{\beta}_j \neq 0 = \beta_j\}$
8. Per comparison error rate (PCER) : $E\{\text{FP}(\hat{\beta})\} / p = \sum_{j: \beta_j \neq 0} P\{\hat{\beta}_j \neq 0\} / p$
9. False discovery rate (FDR) : $E\left\{\text{FP}(\hat{\beta}) / \max\left(1, \|\hat{\beta}\|_0\right)\right\}$

The expectation of the false positive, false negative, and total miss can be all expressed as sums of the Type-I or Type-II errors of tests $I\{\hat{\beta}_j \neq 0\}$ for the hypotheses $H_j : \beta_j = 0$. As the error probabilities of these individual tests may not be easy to track, the false positive and false negative are considered instead. Among other performance measures above, the most stringent one is the correct sign selection. It follows easily from their definitions that

$$\begin{aligned} \text{PCER} &\leq \text{FDR} \\ &\leq \text{FWER} \\ &\leq P\{\text{supp}(\hat{\beta}) \neq \text{supp}(\beta)\} \\ &\leq P\{\text{sgn}(\hat{\beta}) \neq \text{sgn}(\beta)\} \end{aligned}$$

As the three smaller error measures only control Type-I errors in testing H_j , we will focus on the two larger error measures. An estimator $\hat{\beta}$ is sign-consistent if

$$\lim_{n \rightarrow \infty} P\{\text{sgn}(\hat{\beta}) = \text{sgn}(\beta)\} = 1 \quad (4.8)$$

A slightly weaker criterion is the variable selection consistency:

$$\lim_{n \rightarrow \infty} P\{\text{supp}(\hat{\beta}) = \text{supp}(\beta)\} = 1 \quad (4.9)$$

A concept closely related to variable selection consistency (4.9) is oracle property. When the true β is sparse with a small support set $S = \text{supp}(\beta)$ the oracle *LSE* (oracle least squares estimator), denoted by $\hat{\beta}^o$, is defined by

$$\hat{\beta}_S^o = \left(\mathbf{X}_S^T \mathbf{X}_S\right)^{-1} \mathbf{X}_S^T \mathbf{Y}, \quad \hat{\beta}_{S^c}^o = 0 \quad (4.10)$$

This "estimator" is constructed with the aim of an oracle expert with the knowledge of \mathcal{S} , but is not available to the statistician (Fan and Li, 2001). We may say that an estimator $\hat{\beta}$ has the oracle property if it is selection consistent, as statistical procedures based on the selected model would have a high probability of being identical to the same procedure in the true model \mathcal{S} . We may also say that an estimator $\hat{\beta}$ has the oracle property if

$$\sup_{\mathbf{a}} P \left\{ \left| \mathbf{a}^T (\hat{\beta} - \hat{\beta}^o) \right|^2 > \epsilon_n^2 \text{Var} \left(\mathbf{a}^T \hat{\beta}^o \right) \right\} \leq \epsilon_n \quad (4.11)$$

for a certain $\epsilon_n \rightarrow 0$. This oracle property, amongst the weakest version of such, implies that for the estimation of linear functions of β , confidence intervals and other commonly used statistical inference procedures based on $\hat{\beta}$ would have nearly the same performance as those based on the oracle estimator $\hat{\beta}^o$.

Impact of model uncertainty

What is the cost of not knowing the true model $\mathcal{S} = \text{supp}(\beta)$? This can be measured by comparing the best possible performance without knowing \mathcal{S} with the performance of the oracle LSE $\hat{\beta}^o$ in (4.10) at a different noise level. If the best possible performance without knowing \mathcal{S} at the true noise level σ is comparable with the performance of the oracle LSE at an inflated noise level σ' , the ratio σ'/σ is often used to measure the cost of not knowing the oracle model \mathcal{S} . This ratio can be called the noise inflation factor. It is convenient to measure the cost of not knowing the true model \mathcal{S} with the noise inflation factor as the performance of the oracle LSE is well understood.

Let $s = |\mathcal{S}|$ be the size of the true model. We will show in this section that to a large extent, the noise inflation factor is at least of the order $\sqrt{\log(p/s)}$ for prediction and coefficient estimation and $\sqrt{\log(p-s)}$ for variable selection in high-dimensional regression models with Gaussian noise. The impact of this noise inflation may decline or diminish when the signal is strong. When \log is of smaller order than $\log p$, the noise inflation factor is of the order $\sqrt{\log p}$. We will justify the above summary statements with lower and upper performance bounds of matching order in prediction, coefficient estimation, and variable selection. The lower performance bounds, presented in the rest of this section, are applicable to all estimators. The upper performance bounds are derived for PLS estimators in other sections of this chapter.

For simplicity, we assume throughout this discussion of lower bounds for the noise inflation factor that the errors are iid $N(0, \sigma^2)$, i.e. $\varepsilon \sim N(0, \sigma^2 \mathbf{I}_{n \times n})$ in (4.1). Moreover, we assume that the true noise level σ is known and positive as lower bounds in the more difficult case of unknown σ are not smaller.

Bayes lower bounds for orthogonal design

We will start with a simpler case of orthogonal design. A linear regression model has an orthogonal design if $\mathbf{X}_j^T \mathbf{X}_k = 0$ for all $j \neq k$. This does allow $p \rightarrow \infty$. However,

since $\text{rank}(\mathbf{X}) = p$ for orthogonal designs, $p \leq n$ is required. For simplicity, we consider here linear regression models with orthonormal designs: $\mathbf{X}^T \mathbf{X}/n = \mathbf{I}_{p \times p}$, or equivalently $\mathbf{X}_j^T \mathbf{X}_k/n = I\{j = k\}$. In this case, $Z_j = \mathbf{X}_j^T \mathbf{Y}/n$ are sufficient statistics as σ is assumed known here. Moreover, these sufficient statistics are independent of each other and

$$Z_j \sim N(\beta_j, \sigma_n^2)$$

where $\sigma_n = \sigma/n^{1/2}$. This is called the Gaussian sequence model. For orthonormal designs, we have

$$\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2/n = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 = \sum_{j=1}^p (\hat{\beta}_j - \beta_j)^2$$

so that the expected prediction regret is identical to the ℓ_2 estimation error. The following theorem provides a lower bound for the maximum ℓ_q estimation risk of the Bayes rule, including the maximum prediction risk via its equivalence to the ℓ_2 estimation risk using (4.7) with $q = 2$, when the regression coefficients, $\beta_j, j = 1, \dots, p$, follow a coin tossing model. **Theorem 4.1** Suppose $\mathbf{X}^T \mathbf{X}/n = \mathbf{I}_{p \times p}$ and $\varepsilon \sim N(0, \sigma^2 \mathbf{I}_{n \times n})$. Let $1 \leq q < \infty, 0 < \pi_0 < \epsilon_0 \leq 1 - \pi_0, \sigma_n = \sigma/\sqrt{n}, \mu_0 = \sqrt{2 \log(1/\pi_0)} - \sqrt{2 \log(1/\epsilon_0)}$ and G_0 be the prior under which β_j are iid random variables with $P_{G_0}\{\beta_j \neq 0\} = P_{G_0}\{\beta_j = \mu_0 \sigma_n\} = \pi_0$. Then

$$\inf_{\hat{\boldsymbol{\beta}}} E_{G_0} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q^q \geq (1 - \epsilon_0/2) (1 - \epsilon_1) (\pi_0 p) (\sigma_n \mu_0)^q$$

where $\epsilon_1 = 1 - 1/\left[\{\epsilon_0/(1 - \pi_0)\}^{1/(q-1)} + 1\right]^{q-1}$ for $q > 1$ and $\epsilon_1 = 0$ for $q = 1$. **Proof.** The proof is based on the following simple lower bound of the posterior Bayes risk in the estimation of a single β_j

Let δ_1 be an unknown $\{0,1\}$ -valued random variable, whose (posterior) mean is π_1 , and $\omega_1 = \pi_1/(1 - \pi_1)$. Let $R_q(\pi_1)$ be the minimum (posterior) risk for the estimation of δ_1 under the ℓ_q loss. Then,

$$R_q(\pi_1) = \min_{0 \leq x \leq 1} \{\pi_1(1 - x)^q + (1 - \pi_1)x^q\}$$

whose minimum is attained when $x = I\{\pi_1 > 1/2\}$ for $q = 1$ and $x/(1 - x) = \omega_1^{1/(q-1)}$ for $q > 1$. Thus

$$R_q(\pi_1) = \{\pi_1 \wedge (1 - \pi_1)\} I_{\{q=1\}} + \left\{ \pi_1 / \left\{ \omega_1^{1/(q-1)} + 1 \right\}^{q-1} \right\} I_{\{q>1\}}$$

Let $\zeta = Z_1/\sigma_n$ and $\delta_1 = I\{\beta_1 \neq 0\}$. Then, $\zeta | \delta_1 \sim N(\mu_0 \delta_1, 1)$ and $\delta_1 \sim \text{Bernoulli}(\pi_0)$ under P_{G_0} . As (Z_j, β_j) are iid under P_{G_0} , we have

$$\begin{aligned} \inf_{\hat{\boldsymbol{\beta}}} E_{G_0} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q^q &= p \inf_{\hat{\beta}_1} E_{G_0} |\hat{\beta}_1 - \beta_1|^q \\ &= p \sigma_n^q \mu_0^q \min_{f(\cdot)} E_{G_0} |f(\zeta) - \delta_1|^q \end{aligned}$$

The posterior probability for $\delta_1 = 1$ is

$$\begin{aligned}\pi_1(\zeta) &= P_{G_0} \{\delta_1 = 1 \mid \zeta\} \\ &= \pi_0 \varphi(\zeta - \mu_0) / \{\pi_0 \varphi(\zeta - \mu_0) + (1 - \pi_0) \varphi(\zeta)\}\end{aligned}$$

where $\varphi(x) = e^{-x^2/2} / \sqrt{2\pi}$ is the standard normal density. Let $\omega_1(\zeta) = \pi_1(\zeta) / \{1 - \pi_1(\zeta)\}$ be the posterior odds. Because $\pi_0 \varphi(\zeta - \mu_0) + (1 - \pi_0) \varphi(\zeta)$ is the density of ζ , the formula for $R_q(\pi_1)$ gives

$$\begin{aligned}\min_{f(\cdot)} E_{G_0} |f(\zeta) - \delta_1|^q &= E_{G_0} R_q(\pi_1(\zeta)) \\ &= \int \frac{\pi_0 \varphi(\zeta - \mu_0)}{\{\omega_1^{1/(q-1)}(\zeta) + 1\}^{q-1}} d\zeta\end{aligned}$$

where $\{\omega_1^{1/(q-1)}(\zeta) + 1\}^{q-1}$ is treated as $\max\{1, \omega_1(\zeta)\}$ for $q = 1$. Let $a_0 = \sqrt{2 \log(1/\epsilon_0)} = \sqrt{2 \log(1/\pi_0)} - \mu_0$ and $\pi_1^* = \epsilon_0 / (1 - \pi_0)$. As the likelihood ratio $\varphi(\zeta - \mu_0) / \varphi(\zeta)$ is increasing in ζ , for $\zeta \leq \mu_0 + a_0$ the posterior odds is bounded by

$$\begin{aligned}\omega_1(\zeta) &= \pi_0 \varphi(\zeta - \mu_0) / \{(1 - \pi_0) \varphi(\zeta)\} \\ &\leq \pi_0 \varphi(a_0) / \{(1 - \pi_0) \varphi(\mu_0 + a_0)\} \\ &= \pi_0 e^{-a_0^2/2 + (a_0 + \mu_0)^2/2} / (1 - \pi_0) \\ &= \epsilon_0 / (1 - \pi_0)\end{aligned}$$

By assumption, we have $\epsilon_0 / (1 - \pi_0) \leq 1$, so that

$$1 / \{\omega_1^{1/(q-1)}(\zeta) + 1\}^{q-1} \geq 1 / \{(\epsilon_0 / (1 - \pi_0))^{1/(q-1)} + 1\}^{q-1} = 1 - \epsilon_1$$

Thus,

$$\begin{aligned}\min_{f(\cdot)} E_{G_0} |f(\zeta) - \delta_1|^q &\geq (1 - \epsilon_1) \int_{-\infty}^{\mu_0 + a_0} \pi_0 \varphi(\zeta - \mu_0) d\zeta \\ &\geq (1 - \epsilon_1) \pi_0 (1 - \epsilon_0/2)\end{aligned}$$

due to $P\{N(\mu_0, 1) > \mu_0 + a_0\} = P\{N(0, 1) > a_0\} \leq e^{-a_0^2/2}/2 = \epsilon_0/2$ [see (3.52)]. The conclusion follows by applying this inequality to the identity at the beginning of the proof.

This theorem is at the heart of the lower bound results in Donoho and Johnstone (1994) where the minimax ℓ_q risk in ℓ_r balls is studied. The interpretation of this theorem is the clearest when both π_0 and ϵ_0 are small and $\log(1/\epsilon_0)$ is of smaller order than $\log(1/\pi_0)$. In this case, Theorem ?? asserts that in the case where the support of β is unknown, the ℓ_q risk is bounded from below by

$$E_{G_0} \|\hat{\beta} - \beta\|_q^q \geq (1 + o(1)) \{E_{G_0} \|\beta\|_0\} \left\{ \sigma_n \sqrt{2 \log(p / (E_{G_0} \|\beta\|_0))} \right\}^q \quad (4.12)$$

noticing $E_{G_0} \|\beta\|_0 = \pi_0 p$. Under the assumptions of Theorem 4.1, the oracle estimator (4.10) becomes

$$\hat{\beta}^o = (Z_j I \{\beta_j \neq 0\}, j \leq p)^T$$

with $Z_j \sim N(\beta_j, \sigma_n^2)$. It then follows that the oracle estimator has the ℓ_q risk $E_{G_0} \|\hat{\beta}^o - \beta\|_q^q = E_{G_0} \sum_{j=1}^p |Z_j - \beta_j|^q I \{\beta_j \neq 0\} = E_{G_0} \|\beta\|_0 \sigma_n^q E|N(0,1)|^q$. Thus, Theorem 4.1 has the interpretation

$$\left(\frac{\inf_{\hat{\beta}} E_{G_0} \|\hat{\beta} - \beta\|_q^q}{E_{G_0} \|\hat{\beta}^o - \beta\|_q^q} \right)^{1/q} \geq (1 + o(1)) \frac{\sqrt{2 \log(p / (E_{G_0} \|\beta\|_0))}}{\{E|N(0,1)|^q\}^{1/q}} \quad (4.13)$$

In particular, for $q = 2$, the above inequality asserts that in a Bayes setting, the Bayes rule, or the best one can do without knowing the support of β , has a mean squared error no smaller than that of the oracle LSE inflated by approximately a factor of $\sqrt{2 \log(p / \|\beta\|_0)}$. This also gives a noise inflation factor for prediction as the prediction risk is identical to the ℓ_2 estimation error.

The lower bound in Theorem 4.1 is obtained by setting the signal strength, or equivalently the magnitude of the nonzero β_j , at approximately the least informative level. The following example demonstrates that the Bayes estimator may outperform the oracle LSE when the signal is strong and a true prior can be specified.

Example 4.1 Let $\{\mathbf{X}, \beta, \varepsilon\}$ be as in Theorem ?? with possibly different positive parameters $\{\pi_0, \mu_0, a_0, \epsilon_0\}$ satisfying $\mu_0^2/2 \geq a_0 \mu_0 + \log\{(1/\pi_0 - 1)(\mu_0^2/\epsilon_0 - 1)\}$, $0 < \pi_0 \leq 1/2$ and $P\{N(0,1) > a_0\} \leq \epsilon_0/\mu_0^2 < 1$. Then

$$E_{G_0} \|\hat{\beta} - \beta\|_2^2 \leq 2\epsilon_0 E_{G_0} \|\hat{\beta}^o - \beta\|_2^2$$

where $\hat{\beta}$ is the Bayes rule under G_0 and $\hat{\beta}^o$ is the oracle LSE in (4.10). This can be seen as follows. Because $\varphi(a_0)/\varphi(\mu_0 - a_0) = e^{\mu_0^2/2 - a_0 \mu_0} \geq (1/\pi_0 - 1)(\mu_0^2/\epsilon_0 - 1)$ and $\pi_1(\zeta)$ is increasing in ζ , $\zeta \geq \mu_0 - a_0$ implies

$$\pi_1(\zeta) = \frac{\varphi(\zeta - \mu_0)/\varphi(\zeta)}{\varphi(\zeta - \mu_0)/\varphi(\zeta) + (1/\pi_0 - 1)} \geq \pi_1(\mu_0 - a_0) \geq 1 - \epsilon_0/\mu_0^2$$

Thus, $\int \{1 - \pi_1(\zeta)\} \varphi(\zeta - \mu_0) d\zeta \leq \epsilon_0/\mu_0^2 + P\{N(\mu_0, 1) < \mu_0 - a_0\}$, which is bounded by $2\epsilon_0/\mu_0^2$. Let $\sigma_n = \sigma/n^{1/2}$. As in the proof of Theorem 4.1, the ℓ_2 Bayes risk of the Bayes estimator, $p\sigma_n^2 \mu_0^2 \int \{1 - \pi_1(\zeta)\} \varphi(\zeta - \mu_0) d\zeta$ is no greater than $2\epsilon_0 p\sigma_n^2 = 2\epsilon_0 E_{G_0} \|\hat{\beta}^o - \beta\|_2^2$. For variable selection, the noise inflation is expressed in terms of required signal strength. Consider testing $H_j: \beta_j = 0$ with a common one-sided rejection rule $Z_j > \lambda$ for a certain constant λ , where $Z_j \sim N(\beta_j, \sigma_n^2)$ are the independent sufficient statistics for β_j . This class of separable tests includes all Bayes rules for maximizing $P_G\{\text{supp}(\hat{\beta}) = \text{supp}(\beta)\}$ when

β_j are iid nonnegative variables *a priori*. The following theorem provides necessary conditions on λ and the minimum signal strength $\beta_{\min} = \min_{\beta_j \neq 0} |\beta_j|$ for the selection consistency of such methods. Let $\Phi^{-1}(t)$ be the standard normal quantile function, or equivalently the inverse function of $\Phi(t) = \int_{-\infty}^t \varphi(z)dz$. It is well known that for all $t > 0$

$$\frac{t\varphi(t)}{1+t^2} \leq \Phi(-t) = e^{-t^2/2} \int_0^\infty e^{-xt} \varphi(x)dx \leq \min \left\{ \varphi(t)/t, e^{-t^2/2}/2 \right\}$$

Theorem 4.0.1 Suppose Z_j are independent $N(\beta_j, \sigma_n^2)$ variables conditionally on $\beta = (\beta_1, \dots, \beta_p)^T$ under a probability measure P_G . Let $1 > \alpha' > \alpha > 0$, $0 < s < p$ and $\beta_* > 0$ be constants satisfying

$$1 - P_G \{ \beta_{\min} \leq \beta_*, \|\beta\|_0 \leq s \} = \alpha' - \alpha$$

Let λ be a positive constant and $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)^T$ be an estimator of β such that $\hat{\beta}_j = 0$ iff $Z_j \leq \lambda$. If $P_G \{ \text{supp}(\hat{\beta}) = \text{supp}(\beta) \} \geq 1 - \alpha$, then

$$\lambda \geq \sigma_n \Phi^{-1} \left\{ (1 - \alpha')^{1/(p-s)} \right\}, \quad \beta_* \geq \lambda + \sigma_n \Phi^{-1} (1 - \alpha')$$

Proof. From the assumption, we have

$$P_G \left\{ \text{supp}(\hat{\beta}) = \text{supp}(\beta), \beta_{\min} \leq \beta_*, \|\beta\|_0 \leq s \right\} \geq 1 - \alpha' > 0$$

Using the independence of Z_j given β , the right-hand side is bounded from above by

$$\begin{aligned} 1 - \alpha' &\leq P \left(|Z_j| \leq \lambda, j \in \text{supp}(\beta), |Z_j| \geq \lambda, j \notin \text{supp}(\beta), \|\beta\|_0 \leq s \right) \\ &\leq \Phi^{p-s}(\lambda/\sigma_n) \Phi((\beta_* - \lambda)/\sigma_n) \end{aligned}$$

$$\leq \min \left\{ \Phi^{p-s}(\lambda/\sigma_n), \Phi((\beta_* - \lambda)/\sigma_n) \right\}$$

Thus, $\lambda/\sigma_n \geq \Phi^{-1} \left\{ (1 - \alpha')^{1/(p-s)} \right\}$ and $(\beta_* - \lambda)/\sigma_n \geq \Phi^{-1}(1 - \alpha')$. The conclusion follows as $(1 - \alpha')^{1/(p-s)} \geq 1 + (p-s)^{-1} \log(1 - \alpha')$ ■

Suppose β_j are iid variables under P_G and $p - s \rightarrow \infty$, where s is the median of $\|\beta\|_0$. Theorem 4.0.1 asserts that the Bayes rule under the loss function $I\{\text{supp}(\hat{\beta}) \neq \text{supp}(\beta)\}$, which is a thresholding rule with a special λ , requires $\lambda \geq \sigma_n \Phi^{-1} \left((1 - \alpha')^{1/(p-s)} \right) \approx \sigma_n \sqrt{2 \log(p-s)}$ and $\beta_{\min} \geq \lambda + O(\sigma_n)$ to achieve selection consistency. In comparison, for testing an individual hypothesis $\beta_j = 0$ against the alternative $\beta_j \geq \beta_{\min}$, both Type I and Type II error probability are bounded by α when $\beta_{\min} \geq 2\sigma_n \Phi^{-1}(1 - \alpha)$. Thus, due to the multiplicity of the testing problem, an increase of the signal strength β_{\min} by a factor of the order $\sqrt{\log(p-s)}$ is required to control the noise. Again, this can be viewed as a noise inflation factor due to model uncertainty.



5. 论文综述



6. NONCONCAVE PENALIZED LIKELIHOOD WITH A D

6.1 Properties of penalized likelihood estimation

6.1.1 Regularity conditions

To this end, we consider the log-likelihood series $\log f_n(V_n, \beta_n)$, where $f_n(V_n, \beta_n)$ is the density of the random variable V_n , all of which relate to the sample size n , and assume without loss of generality that, unknown to us, the first s_n components of β_n , denoted by β_{n1} , do not vanish and the remaining $p_n - s_n$ coefficients, denoted by β_{n2} , are 0. Our objectives in this paper are to investigate the following asymptotic properties of a nonconcave penalized likelihood estimator.

Regularity condition on penalty.

Let $a_n = \max_{1 \leq j \leq p_n} \{p'_{\lambda_n}(|\beta_{n0j}|), \beta_{n0j} \neq 0\}$ and $b_n = \max_{1 \leq j \leq p_n} \{p''_{\lambda_n}(|\beta_{n0j}|), \beta_{n0j} \neq 0\}$. Then we need to place the following conditions on the penalty functions:

1. (A) $\liminf_{n \rightarrow +\infty} \liminf_{\theta \rightarrow 0+} p'_{\lambda_n}(\theta) / \lambda_n > 0$
 2. (B) $a_n = O(n^{-1/2})$: $\frac{a_n}{n^{-1/2}} < \infty$
 3. (B') $a_n = o(1/\sqrt{np_n})$: $\frac{a_n}{1/\sqrt{np_n}} = 0$
 4. (C) $b_n \rightarrow 0$ as $n \rightarrow +\infty$
 5. (C') $b_n = o_p(1/\sqrt{p_n})$: $\frac{b_n}{1/\sqrt{p_n}} = 0$
 6. (D) there are constants C and D such that, when $\theta_1, \theta_2 > C\lambda_n$, $|p''_{\lambda_n}(\theta_1) - p''_{\lambda_n}(\theta_2)| \leq D|\theta_1 - \theta_2|$
1. Condition (A) makes the penalty function **singular at the origin** so that the penalized likelihood estimators possess the sparsity property.
 2. Conditions (B) and (B') ensure the unbiasedness property for large parameters and

the existence of the root- n -consistent penalized likelihood estimator.

3. Conditions (C) and (C') guarantee that the penalty function does not have much more influence than the likelihood function on the penalized likelihood estimators.
4. Condition (D) is a smoothness condition that is imposed on the nonconcave penalty functions.
5. Under the condition (H) all of these conditions are satisfied by the SCAD penalty and the hard thresholding penalty, as $a_n = 0$ and $b_n = 0$ when n is large enough.

Regularity conditions on likelihood functions

Due to the diverging number of parameters, we cannot assume that likelihood functions are invariant in our study. Some conditions have to be strengthened to keep uniform properties for the likelihood functions and sample series. A higher-order moment of the likelihood functions is a simple and direct method to keep uniform properties, as compared to the usual conditions in the asymptotic theory of the likelihood estimate under finite parameters [see, e.g., Lehmann (1983)]. The conditions that are imposed on the likelihood functions are as follows:

1. (E) For every n the observations $\{V_{ni}, i = 1, 2, \dots, n\}$ are independent and identically distributed with the probability density $f_n(V_{n1}, \beta_n)$, which has a common support, and the model is identifiable. Furthermore, the first and second derivatives of the likelihood function satisfy the equations

$$E_{\beta_n} \left\{ \frac{\partial \log f_n(V_{n1}, \beta_n)}{\partial \beta_{nj}} \right\} = 0 \quad \text{for } j = 1, 2, \dots, p_n$$

and

$$E_{\beta_n} \left\{ \frac{\partial \log f_n(V_{n1}, \beta_n)}{\partial \beta_{nj}} \frac{\partial \log f_n(V_{n1}, \beta_n)}{\partial \beta_{nk}} \right\} = -E_{\beta_n} \left\{ \frac{\partial^2 \log f_n(V_{n1}, \beta_n)}{\partial \beta_{nj} \partial \beta_{nk}} \right\}$$

2. (F) The Fisher information matrix

$$I_n(\beta_n) = E \left[\left\{ \frac{\partial \log f_n(V_n, \beta_n)}{\partial \beta_n} \right\} \left\{ \frac{\partial \log f_n(V_n, \beta_n)}{\partial \beta_n} \right\}^T \right]$$

satisfies conditions

$$0 < C_1 < \lambda_{\min} \{I_n(\beta_n)\} \leq \lambda_{\max} \{I_n(\beta_n)\} < C_2 < \infty$$

for all n and, for $j, k = 1, 2, \dots, p_n$

$$E_{\beta_n} \left\{ \frac{\partial \log f_n(V_{n1}, \beta_n)}{\partial \beta_{nj}} \frac{\partial \log f_n(V_{n1}, \beta_n)}{\partial \beta_{nk}} \right\}^2 < C_3 < \infty$$

and

$$E_{\beta_n} \left\{ \frac{\partial^2 \log f_n(V_{n1}, \beta_n)}{\partial \beta_{nj} \partial \beta_{nk}} \right\}^2 < C_4 < \infty$$

3. (G) There is a large enough open subset ω_n of $\Omega_n \in R^{p_n}$ which contains the true parameter point β_n , such that for almost all V_{ni} the density admits all third derivatives $\partial f_n(V_{ni}, \beta_n) / \partial \beta_{nj} \beta_{nk} \beta_{nl}$ for all $\beta_n \in \omega_n$. Furthermore, there are functions $M_{n jkl}$ such that

$$\left| \frac{\partial \log f_n(V_{ni}, \beta_n)}{\partial \beta_{nj} \beta_{nk} \beta_{nl}} \right| \leq M_{n jkl}(V_{ni})$$

for all $\beta_n \in \omega_n$, and

$$E_{\beta_n} \left\{ M_{n jkl}^2(V_{ni}) \right\} < C_5 < \infty$$

for all p, n and j, k, l .

4. (H) Let the values of $\beta_{n01}, \beta_{n02}, \dots, \beta_{n0s_n}$ be nonzero and $\beta_{n0(s_n+1)}, \beta_{n02}, \dots, \beta_{n0p_n}$ be zero. Then $\beta_{n01}, \beta_{n02}, \dots, \beta_{n0s_n}$ satisfy

$$\min_{1 \leq j \leq s_n} |\beta_{n0j}| / \lambda_n \rightarrow \infty \quad \text{as } n \rightarrow \infty$$

Under conditions (F) and (G), the second and fourth moments of the likelihood function are imposed. The information matrix of the likelihood function is assumed to be positive definite, and its eigenvalues are uniformly bounded. These conditions are stronger than those of the usual asymptotic likelihood theory, but they facilitate the technical derivations.

Condition (H) seems artificial, but it is necessary for obtaining the oracle property. In a finite-parameter situation this condition is implicitly assumed, and is in fact stronger than that imposed here. Condition (H) explicitly shows the rate at which the penalized likelihood can distinguish nonvanishing parameters from 0. Its zero component can be relaxed as

$$\max_{s_n+1 \leq j \leq p_n} |\beta_{n0j}| / \lambda_n \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

6.1.2 Oracle properties

Recall that $V_{ni}, i = 1, \dots, n$, are independent and identically distributed random variables with density $f_n(V_n, \beta_{n0})$. Let

$$L_n(\beta_n) = \sum_{i=1}^n \log f_n(V_{ni}, \beta_n)$$

be the log-likelihood function and let

$$Q_n(\beta_n) = L_n(\beta_n) - n \sum_{j=1}^{p_n} p_{\lambda_n}(|\beta_{nj}|)$$

be the **penalized likelihood function**. (Oracle property.) Under certain conditions of the likelihood function and for certain penalty functions (e.g., SCAD), if p_n does not grow too fast, then by the proper choice of λ_n there exists a penalized likelihood estimator such that $\hat{\beta}_{n2} = 0$ and $\hat{\beta}_{n1}$ behaves the same as the case in which $\beta_{n2} = 0$ is known in advance.

Theorem 6.1.1 — Existence of penalized likelihood estimator. Suppose that the density $f_n(V_n, \beta_{n0})$ satisfies conditions (E) – (G), and the penalty function $p_{\lambda_n}(\cdot)$ satisfies conditions (B) – (D). If $p_n^4/n \rightarrow 0$ as $n \rightarrow \infty$, then there is a local maximizer $\hat{\beta}_n$ of $Q(\beta_n)$ such that $\|\hat{\beta}_n - \beta_{n0}\| = O_p\{\sqrt{p_n}(n^{-1/2} + a_n)\}$, (i.e. $\frac{\|\hat{\beta}_n - \beta_{n0}\|}{\sqrt{p_n}(n^{-1/2} + a_n)} < \infty$) where $a_n = \max_{1 \leq j \leq p_n} \{p'_{\lambda_n}(|\beta_{n0j}|), \beta_{n0j} \neq 0\}$

It is easy to see that if a_n satisfies condition (B), that is, $a_n = O(n^{-1/2})$, then there is a root- (n/p_n) -consistent estimator.

1. This consistent rate is the same as the result of the M-estimator that was studied by Huber (1973), in which the number of parameters diverges.
2. Condition (B) is automatically satisfied by SCAD and Hard, however, is not easily for the penalty functions like L_q . The convergence rate of a_n for the usual convex penalties, such as the L_q -penalty with $q \geq 1$, largely depends on the convergence rate of λ_n . As these penalties do not have an unbiasedness property, they require that λ_n satisfy the condition $\lambda_n = O(n^{-1/2})$ in order to have a root- (n/p_n) -consistent estimator for the penalized likelihood estimator. This requirement will make it difficult to choose λ_n for penalized likelihood in practice. However, if the penalty function is a SCAD or hard thresholding penalty, and condition (H) is satisfied by the model, it is clear that $a_n = 0$ when n is large enough. The root- (n/p_n) -consistent penalized likelihood estimator indeed exists with probability tending to 1, and no requirements are imposed on the convergence rate of λ_n .

Denote

$$\Sigma_{\lambda_n} = \text{diag}\{p''_{\lambda_n}(\beta_{n01}), \dots, p''_{\lambda_n}(\beta_{n0s_n})\}$$

and

$$\mathbf{b}_n = \{p'_{\lambda_n}(|\beta_{n01}|) \text{sgn}(\beta_{n01}), \dots, p'_{\lambda_n}(|\beta_{n0s_n}|) \text{sgn}(\beta_{n0s_n})\}^T$$

Theorem 6.1.2 — Oracle property. Under conditions (A) – (H), if $\lambda_n \rightarrow 0$, $\sqrt{n/p_n}\lambda_n \rightarrow \infty$ and $p_n^5/n \rightarrow 0$ as $n \rightarrow \infty$, then, with probability tending to 1 the root- (n/p_n) -consistent local maximizer $\hat{\beta}_n = \begin{pmatrix} \hat{\beta}_{n1} \\ \hat{\beta}_{n2} \end{pmatrix}$ in Theorem 1 must satisfy:

1. (Sparsity) $\hat{\beta}_{n2} = 0$
2. (Asymptotic normality)

$$\begin{aligned} & \sqrt{n} A_n I_n^{-1/2}(\beta_{n01}) \{I_n(\beta_{n01}) + \Sigma_{\lambda_n}\} \\ & \times [\hat{\beta}_{n1} - \beta_{n01} + \{I_n(\beta_{n01}) + \Sigma_{\lambda_n}\}^{-1} \mathbf{b}_n] \xrightarrow{\mathcal{D}} \mathcal{N}(0, G) \end{aligned}$$

($I_n(\beta_{n01}) + \Sigma_{\lambda_n}$ means that only consider the s_n terms) where A_n is a $q \times s_n$ matrix such

that $A_n A_n^T \rightarrow G$, and G is a $q \times q$ nonnegative symmetric matrix.

By Theorem 6.1.2 the sparsity and the asymptotic normality are still valid when the number of parameters diverges. In fact, the oracle property holds for the SCAD and the hard thresholding penalty function. When n is large enough, $\Sigma_{\lambda_n} = 0$ and $\mathbf{b}_n = 0$ for the SCAD and the hard thresholding penalty. Hence, Theorem 6.1.2(ii) becomes

$$\sqrt{n} A_n I_n^{1/2} (\beta_{n01}) (\hat{\beta}_{n1} - \beta_{n01}) \xrightarrow{D} \mathcal{N}(0, G)$$

which has the same efficiency as the maximum likelihood estimator of β_{n01} based on the submodel with $\beta_{n02} = 0$ known in advance. This demonstrates that the nonconcave penalized likelihood estimator is as efficient as the oracle one. Intrinsically, unbiasedness and singularity at the origin of the SCAD and the hard thresholding penalty functions guarantee this sampling property.

The L_q -penalty, $q \geq 1$, cannot simultaneously satisfy the conditions $\lambda_n = O_p(n^{-1/2})$ and $\sqrt{n/p} \lambda_n \rightarrow \infty$ as $n \rightarrow \infty$. These penalty functions cannot produce estimators with the oracle property. The L_q -penalty, $q < 1$, may satisfy these two conditions at same time. As shown by Knight and Fu(2000) in a finiteparameter setting, it might also have sampling properties that are similar to the oracle property when the number of parameters diverges. However, the bias term in Theorem 2(ii) cannot be ignored.

The condition $p_n^4/n \rightarrow 0$ or $p_n^5/n \rightarrow 0$ as $n \rightarrow \infty$ seems somewhat strong. By refining the structure of the log-likelihood function, such as the generalized linear model $\ell(X^T \beta, Y)$ or the M -estimator from $\sum_{i=1}^n \rho(Y_i - X_i^T \beta)$, the condition can be weakened to $p_n^3/n \rightarrow 0$ as $n \rightarrow \infty$. This condition is in line with that of Huber (1973)

6.1.3 Estimation of covariance matrix

As in Fan and Li (2001), by the sandwich formula let

$$\begin{aligned} \hat{\Sigma}_n = & n \{ \nabla^2 L_n(\hat{\beta}_{n1}) - n \Sigma_{\lambda_n}(\hat{\beta}_{n1}) \}^{-1} \\ & \times \widehat{\text{cov}} \{ \nabla L_n(\hat{\beta}_{n1}) \} \{ \nabla^2 L_n(\hat{\beta}_{n1}) - n \Sigma_{\lambda_n}(\hat{\beta}_{n1}) \}^{-1} \end{aligned}$$

be the estimated covariance matrix of $\hat{\beta}_{n1}$, where

$$\begin{aligned} \widehat{\text{cov}} \{ \nabla L_n(\hat{\beta}_{n1}) \} = & \left\{ \frac{1}{n} \sum_{i=1}^n \frac{\partial L_{ni}(\hat{\beta}_{n1})}{\partial \beta_j} \frac{\partial L_{ni}(\hat{\beta}_{n1})}{\partial \beta_k} \right\} \\ & - \left\{ \frac{1}{n} \sum_{i=1}^n \frac{\partial L_{ni}(\hat{\beta}_{n1})}{\partial \beta_j} \right\} \left\{ \frac{1}{n} \sum_{i=1}^n \frac{\partial L_{ni}(\hat{\beta}_{n1})}{\partial \beta_k} \right\} \end{aligned}$$

Denote by

$$\Sigma_n = \{ I_n(\beta_{n01}) + \Sigma_{\lambda_n}(\beta_{n01}) \}^{-1} I_n(\beta_{n01}) \{ I_n(\beta_{n01}) + \Sigma_{\lambda_n}(\beta_{n01}) \}^{-1}$$

the asymptotic variance of $\hat{\beta}_{n1}$ in Theorem 2(ii).

Theorem 6.1.3 — Consistency of the sandwich formula. If conditions (A) – (H) are satisfied and $p_n^5/n \rightarrow 0$ as $n \rightarrow \infty$, then we have

$$A_n \hat{\Sigma}_n A_n^T - A_n \Sigma_n A_n^T \xrightarrow{p} 0 \quad \text{as } n \rightarrow \infty \quad (6.1)$$

for any $q \times s_n$ matrix A_n such that $A_n A_n^T = G$, where q is any fixed integer.

Theorem 6.1.3 not only proves a conjecture of Fan and Li (2001) about the consistency of the sandwich formula for the standard error matrix, but also extends the result to the situation with a growing number of parameters. The consistent result also offers a way to construct a confidence interval for the estimates of parameters. For a review of sandwich covariance matrix estimation, see the paper of Kauermann and Carroll (2001).

6.1.4 Likelihood ratio test

One of the most celebrated methods in statistics is the likelihood ratio test. Can it also be applied to the penalized likelihood context with a diverging number of parameters? To answer this question, consider the problem of testing linear hypotheses:

$$H_0 : A_n \beta_{n01} = 0 \quad \text{vs.} \quad H_1 : A_n \beta_{n01} \neq 0$$

where A_n is a $q \times s_n$ matrix and $A_n A_n^T = I_q$ for a fixed q . This problem includes testing simultaneously the significance of a few covariate variables.

In the penalized likelihood context a natural likelihood ratio test for the problem is

$$T_n = 2 \left\{ \sup_{\Omega_n} Q(\beta_n | \mathbf{V}) - \sup_{\Omega_n, A_n \beta_{n1} = 0} Q(\beta_n | \mathbf{V}) \right\}$$

The following theorem drives the asymptotic null distribution of the test statistic. It shows that the classical likelihood theory continues to hold for the problem with a growing number of parameters in the penalized likelihood context. It enables one to apply the traditional likelihood ratio method for testing linear hypotheses. In particular, it allows one to simultaneously test whether a few variables are statistically significant by taking some specific matrix A_n

Theorem 6.1.4 When conditions (A) – (H), (B') and (C') are satisfied, under H_0 we have

$$T_n \xrightarrow{\mathcal{D}} \chi_q^2 \quad (6.2)$$

provided that $p_n^5/n \rightarrow 0$ as $n \rightarrow \infty$

For the usual likelihood without penalization, Portnoy (1988) and Murphy (1993) showed that the Wilks type of result continues to hold for specific problems. Our results can be regarded as a further generalization of theirs.

6.2 Proofs of theorems.

In this section, we give rigorous proofs of Theorems 1 – 4

PROOF OF THEOREM 1.

Proof. Let $\alpha_n = \sqrt{p_n} (n^{-1/2} + a_n)$ and set $\|\mathbf{u}\| = C$, where C is a large enough constant. Our aim is to show that for any given ε there is a large constant C such that, for large n we have

$$P \left\{ \sup_{\|\mathbf{u}\|=C} Q_n(\beta_{n0} + \alpha_n \mathbf{u}) < Q_n(\beta_{n0}) \right\} \geq 1 - \varepsilon \quad (6.3)$$

This implies that with probability tending to 1 there is a local maximum $\hat{\beta}_n$ in the ball $\{\beta_{n0} + \alpha_n \mathbf{u} : \|\mathbf{u}\| \leq C\}$ such that $\|\hat{\beta}_n - \beta_{n0}\| = O_p(\alpha_n)$. Using $p_{\lambda_n}(0) = 0$, we have

$$\begin{aligned} D_n(\mathbf{u}) &= Q_n(\beta_{n0} + \alpha_n \mathbf{u}) - Q_n(\beta_{n0}) \\ &\leq^{(i)} L_n(\beta_{n0} + \alpha_n \mathbf{u}) - L_n(\beta_{n0}) \\ &\quad - n \sum_{j=1}^{s_n} \{p_{\lambda_n}(|\beta_{n0j} + \alpha_n u_j|) - p_{\lambda_n}(|\beta_{n0j}|)\} \\ &= (I) + (II) \end{aligned}$$

(i) is for the number of sum is s_n . Then by Taylor's expansion we obtain

$$\begin{aligned} (I) &= \alpha_n \nabla^T L_n(\beta_{n0}) \mathbf{u} + \frac{1}{2} \mathbf{u}^T \nabla^2 L_n(\beta_{n0}) \mathbf{u} \alpha_n^2 + \frac{1}{6} \nabla^T \left\{ \mathbf{u}^T \nabla^2 L_n(\beta_n^*) \mathbf{u} \right\} \mathbf{u} \alpha_n^3 \\ &\triangleq I_1 + I_2 + I_3 \end{aligned}$$

where the vector β_n^* lies between β_{n0} and $\beta_{n0} + \alpha_n \mathbf{u}$, and

$$\begin{aligned} (II) &= - \sum_{j=1}^{s_n} \left[n \alpha_n p'_{\lambda_n}(|\beta_{n0j}|) \operatorname{sgn}(\beta_{n0j}) u_j + n \alpha_n^2 p''_{\lambda_n}(\beta_{n0j}) u_j^2 \{1 + o(1)\} \right] \\ &\triangleq I_4 + I_5 \end{aligned}$$

By condition (F),

$$\begin{aligned} |I_1| &= \left| \alpha_n \nabla^T L_n(\beta_{n0}) \mathbf{u} \right| \leq \alpha_n \left\| \nabla^T L_n(\beta_{n0}) \right\| \|\mathbf{u}\| \\ &=^{(i)} O_p(\alpha_n \sqrt{np_n}) \|\mathbf{u}\| =^{(ii)} O_p(\alpha_n^2 n) \|\mathbf{u}\| \end{aligned} \quad (6.4)$$

(i) is for the conditon (F):for $j, k = 1, 2, \dots, p_n, E_{\beta_n} \left\{ \frac{\partial \log f_n(V_{n1}, \beta_n)}{\partial \beta_{nj}} \frac{\partial \log f_n(V_{n1}, \beta_n)}{\partial \beta_{nk}} \right\}^2 < C_3 < \infty$, (ii)

is for the defination of α_n . Next we consider I_2 . An application of Lemma 6.4 $\left\| \frac{1}{n} \nabla^2 L_n(\beta_{n0}) + I_n(\beta_{n0}) \right\| = o_p\left(\frac{1}{p_n}\right)$ yields that

$$\begin{aligned} I_2 &= \frac{1}{2} \mathbf{u}^T \left[\frac{1}{n} \left\{ \nabla^2 L_n(\beta_{n0}) - E \nabla^2 L_n(\beta_{n0}) \right\} \right] \mathbf{u} \cdot n \alpha_n^2 - \frac{1}{2} \mathbf{u}^T I_n(\beta_{n0}) \mathbf{u} \cdot n \alpha_n^2 \\ &= - \frac{n \alpha_n^2}{2} \mathbf{u}^T I_n(\beta_{n0}) \mathbf{u} + o_p(1) \cdot n \alpha_n^2 \|\mathbf{u}\|^2 \end{aligned} \quad (6.5)$$

By the Cauchy-Schwarz inequality and condition (G), we have

$$\begin{aligned} |I_3| &= \left| \frac{1}{6} \sum_{i,j,k=1}^{p_n} \frac{\partial L_n(\beta_n^*)}{\partial \beta_{ni} \partial \beta_{nj} \partial \beta_{nk}} u_i u_j u_k \alpha_n^3 \right| \\ &\leq \frac{1}{6} \sum_{l=1}^n \left\{ \sum_{i,j,k=1}^{p_n} M_{nijk}^2(V_{nl}) \right\}^{1/2} \|\mathbf{u}\|^3 \alpha_n^3 \end{aligned}$$

(不懂) Since $p_n^4/n \rightarrow 0, a_n \sqrt{np_n} \rightarrow 0$, hence $p_n^2 a_n \rightarrow 0$ as $n \rightarrow \infty$, we have

$$\frac{1}{6} \sum_{l=1}^n \left\{ \sum_{i,j,k=1}^{p_n} M_{nijk}^2(V_{nl}) \right\}^{1/2} \|\mathbf{u}\|^3 \alpha_n^3 \stackrel{(i)}{=} O_p(p_n^{3/2} \alpha_n) n \alpha_n^2 \|\mathbf{u}\|^2 \stackrel{(ii)}{=} o_p(n \alpha_n^2) \|\mathbf{u}\|^2$$

(i) is for $\sum_{l=1}^n \left\{ \sum_{i,j,k=1}^{p_n} M_{nijk}^2(V_{nl}) \right\}^{1/2} = O_p(p_n^{1/2})$, $\|\mathbf{u}\| = O_p(p_n)$, (ii) is for $\frac{M}{n \alpha_n^2} \rightarrow 0$.

Thus,

$$I_3 = o_p(n \alpha_n^2) \|\mathbf{u}\|^2 \quad (6.6)$$

The terms I_4 and I_5 can be dealt with as follows. First,

$$|I_4| \leq \sum_{j=1}^{s_n} |n \alpha_n p'_{\lambda_n}(|\beta_{n0j}|) \operatorname{sgn}(\beta_{n0j}) u_j| \leq \sqrt{s_n} \cdot n \alpha_n a_n \|\mathbf{u}\| \leq n \alpha_n^2 \|\mathbf{u}\| \quad (6.7)$$

($s_n < p_n$) and

$$I_5 = \sum_{j=1}^{s_n} n \alpha_n^2 p''_{\lambda_n}(|\beta_{n0j}|) u_j^2 \{1 + o(1)\} \leq 2 \cdot \max_{1 \leq j \leq s_n} p''_{\lambda_n}(|\beta_{n0j}|) \cdot n \alpha_n^2 \|\mathbf{u}\|^2 \quad (6.8)$$

By (6.4) – (6.8) and condition (C), and allowing $\|\mathbf{u}\|$ to be large enough, all terms I_1, I_3, I_4 and I_5 are dominated by I_2 , which is negative. This proves (6.3). ■

To prove Theorem 6.1.2, we first show that the nonconcave penalized estimator possesses the sparsity property $\hat{\beta}_{n2} = 0$ by the following lemma.

Lemma 6.1 Assume that conditions (A) and (E)-(H) are satisfied. If $\lambda_n \rightarrow 0$, $\sqrt{n/p_n} \lambda_n \rightarrow \infty$ and $p_n^5/n \rightarrow 0$ as $n \rightarrow \infty$, then with probability tending to 1, for any given β_{n1} satisfying $\|\beta_{n1} - \beta_{n01}\| = O_p(\sqrt{p_n/n})$ and any constant C ,

$$Q \left\{ \left(\beta_{n1}^T, 0 \right)^T \right\} = \max_{\|\beta_{n2}\| \leq C(p_n/n)^{1/2}} Q \left\{ \left(\beta_{n1}^T, \beta_{n2}^T \right)^T \right\}$$

Proof. Let $\varepsilon_n = C\sqrt{p_n/n}$. It is sufficient to show that with probability tending to 1 as $n \rightarrow \infty$, for any β_{n1} satisfying $\beta_{n1} - \beta_{n01} = O_p(\sqrt{p_n/n})$ we have, for $j = s_n + 1, \dots, p_n$

$$\frac{\partial Q_n(\beta_n)}{\partial \beta_{nj}} < 0 \quad \text{for } 0 < \beta_{nj} < \varepsilon_n \quad (6.9)$$

$$\frac{\partial Q_n(\beta_n)}{\partial \beta_{nj}} > 0 \quad \text{for } -\varepsilon_n < \beta_{nj} < 0 \quad (6.10)$$

(which means the point 0 is the maximum point) By Taylor expansion,

$$\begin{aligned} \frac{\partial Q_n(\beta_n)}{\partial \beta_{nj}} &= \frac{\partial L_n(\beta_n)}{\partial \beta_{nj}} - np'_{\lambda_n}(|\beta_{nj}|) \operatorname{sgn}(\beta_{nj}) \\ &= \frac{\partial L_n(\beta_{n0})}{\partial \beta_{nj}} + \sum_{l=1}^{p_n} \frac{\partial^2 L_n(\beta_{n0})}{\partial \beta_{nj} \partial \beta_{nl}} (\beta_{nl} - \beta_{n0l}) \\ &\quad + \sum_{l,k=1}^{p_n} \frac{\partial^3 L_n(\beta_n^*)}{\partial \beta_{nj} \partial \beta_{nl} \partial \beta_{nk}} (\beta_{nl} - \beta_{n0l}) (\beta_{nk} - \beta_{n0k}) \\ &\quad - np'_{\lambda_n}(|\beta_{nj}|) \operatorname{sgn}(\beta_{nj}) \\ &\triangleq I_1 + I_2 + I_3 + I_4 \end{aligned}$$

where β_n^* lies between β_n and β_{n0} . Next, we consider I_1, I_2 and I_3 . By a standard argument, we have

$$I_1 = O_p(\sqrt{n}) = O_p(\sqrt{np_n}) \quad (6.11)$$

The term I_2 can be written as

$$\begin{aligned} I_2 &= \sum_{l=1}^{p_n} \left\{ \frac{\partial^2 L_n(\beta_{n0})}{\partial \beta_{nj} \partial \beta_{nl}} - E \frac{\partial^2 L_n(\beta_{n0})}{\partial \beta_{nj} \partial \beta_{nl}} \right\} (\beta_{nl} - \beta_{n0l}) \\ &\quad + \sum_{l=1}^{p_n} E \frac{\partial^2 L_n(\beta_{n0})}{\partial \beta_{nj} \partial \beta_{nl}} (\beta_{nl} - \beta_{n0l}) \\ &\triangleq K_1 + K_2 \end{aligned}$$

Using the Cauchy-Schwarz inequality and $\|\beta_n - \beta_{n0}\| = O_p(\sqrt{p_n/n})$, we have

$$\begin{aligned} |K_2| &= \left| n \sum_{l=1}^{p_n} I_n(\beta_{n0})(j, l) (\beta_{nl} - \beta_{n0l}) \right| \\ &\leq n O_p\left(\sqrt{\frac{p_n}{n}}\right) \left\{ \sum_{l=1}^{p_n} I_n^2(\beta_{n0})(j, l) \right\}^{1/2} \end{aligned}$$

(how we obtain the n ?) As the eigenvalues of the information matrix are bounded according to condition (F), we have

$$\sum_{l=1}^{p_n} I_n^2(\beta_{n0})(j, l) = O(1)$$

This entails that

$$K_2 = O_p(\sqrt{np_n}) \quad (6.12)$$

As for the term K_1 , by the Cauchy-Schwarz inequality we have

$$|K_1| \leq \|\beta_n - \beta_{n0}\| \left[\sum_{l=1}^{p_n} \left\{ \frac{\partial^2 L_n(\beta_{n0})}{\partial \beta_{nj} \partial \beta_{nl}} - E \frac{\partial^2 L_n(\beta_{n0})}{\partial \beta_{nj} \partial \beta_{nl}} \right\}^2 \right]^{1/2}$$

Then from condition (F) and lemma 6.4, it is easy to show that

$$\left[\sum_{l=1}^{p_n} \left\{ \frac{\partial^2 L_n(\beta_{n0})}{\partial \beta_{nj} \partial \beta_{nl}} - E \frac{\partial^2 L_n(\beta_{n0})}{\partial \beta_{nj} \partial \beta_{nl}} \right\}^2 \right]^{1/2} = O_p(\sqrt{np_n})$$

By $\|\beta_n - \beta_{n0}\| = O_p(\sqrt{p_n/n})$ it follows that $K_1 = O_p(\sqrt{np_n})$. This, together with (6.12), yields

$$I_2 = O_p(\sqrt{np_n}) \quad (6.13)$$

Next we consider I_3 . We can write I_3 as follows:

$$\begin{aligned} I_3 &= \sum_{l,k=1}^{p_n} \left\{ \frac{\partial^3 L_n(\beta_n^*)}{\partial \beta_{nj} \partial \beta_{nl} \partial \beta_{nk}} - E \frac{\partial^3 L_n(\beta_n^*)}{\partial \beta_{nj} \partial \beta_{nl} \partial \beta_{nk}} \right\} (\beta_{nj} - \beta_{n0j}) (\beta_{nk} - \beta_{n0k}) \\ &\quad + \sum_{l,k=1}^{p_n} E \frac{\partial^3 L_n(\beta_n^*)}{\partial \beta_{nj} \partial \beta_{nl} \partial \beta_{nk}} (\beta_{nj} - \beta_{n0j}) (\beta_{nk} - \beta_{n0k}) \\ &\triangleq K_3 + K_4 \end{aligned}$$

By condition (G),

$$|K_4| \leq C_5^{1/2} \cdot np_n \cdot \|\beta_n - \beta_{n0}\|^2 = O_p(p_n^2) = o_p(\sqrt{np_n}) \quad (6.14)$$

However, by the Cauchy-Schwarz inequality,

$$K_3^2 \leq \sum_{l,k=1}^{p_n} \left\{ \frac{\partial^3 L_n(\beta_n^*)}{\partial \beta_{nj} \partial \beta_{nl} \partial \beta_{nk}} - E \frac{\partial^3 L_n(\beta_n^*)}{\partial \beta_{nj} \partial \beta_{nl} \partial \beta_{nk}} \right\}^2 \|\beta_n - \beta_{n0}\|^4$$

Under conditions (G) and (H), we have

$$K_3 = O_p \left\{ \left(np_n^2 \frac{p_n^2}{n^2} \right)^{1/2} \right\} = o_p(\sqrt{np_n}) \quad (6.15)$$

From (6.11) and (6.13) – (6.15) we have

$$I_1 + I_2 + I_3 = O_p(\sqrt{np_n})$$

Because $\sqrt{p_n/n}/\lambda_n \rightarrow 0$ and $\liminf_{n \rightarrow \infty} \inf_{\theta \rightarrow 0^+} p'_{\lambda_n}(\theta)/\lambda_n > 0$, from

$$\frac{\partial Q_n(\beta_n)}{\partial \beta_{nj}} = n\lambda_n \left\{ -\frac{p'_{\lambda_n}(|\beta_{nj}|)}{\lambda_n} \operatorname{sgn}(\beta_{nj}) + O_p\left(\sqrt{\frac{p_n}{n}}/\lambda_n\right) \right\}$$

it is easy to see that the sign of β_{nj} completely determines the sign of $\partial Q_n(\beta_n)/\partial \beta_{nj}$. Hence, (6.9) and (6.10) follow. ■

PROOF OF THEOREM 2.

Proof. As shown in Theorem 6.1.1, there is a root- (n/p_n) consistent local maximizer $\hat{\beta}_n$ of $Q_n(\beta_n)$. By Lemma 6.1, part (i) holds that $\hat{\beta}_n$ has the form $(\hat{\beta}_{n1}, 0)^T$. We need only prove part (ii), the asymptotic normality of the penalized nonconcave likelihood estimator $\hat{\beta}_{n1}$. If we can show that

$$\{I_n(\beta_{n01}) + \Sigma_{\lambda_n}\}(\hat{\beta}_{n1} - \beta_{n01}) + \mathbf{b}_n = \frac{1}{n} \nabla L_n(\beta_{n01}) + o_p(n^{-1/2})$$

then

$$\begin{aligned} & \sqrt{n} A_n I_n^{-1/2}(\beta_{n01}) \{I_n(\beta_{n01}) + \Sigma_{\lambda_n}\} \left[\hat{\beta}_{n1} - \beta_{n01} + \{I_n(\beta_{n01}) + \Sigma_{\lambda_n}\}^{-1} \mathbf{b}_n \right] \\ &= \frac{1}{\sqrt{n}} A_n I_n^{-1/2}(\beta_{n01}) \nabla L_n(\beta_{n01}) + o_p \left\{ A_n I_n^{-1/2}(\beta_{n01}) \right\} \end{aligned}$$

By the conditions of Theorem 6.1.2, we have the last term of $o_p(1)$. Then let

$$Y_{ni} = \frac{1}{\sqrt{n}} A_n I_n^{-1/2}(\beta_{n01}) \nabla L_{ni}(\beta_{n01}), \quad i = 1, 2, \dots, n$$

It follows that, for any ε

$$\begin{aligned} \sum_{i=1}^n E \|Y_{ni}\|^2 \mathbf{1}\{\|Y_{ni}\| > \varepsilon\} &= n E \|Y_{n1}\|^2 \mathbf{1}\{\|Y_{n1}\| > \varepsilon\} \\ &\leq n \left\{ E \|Y_{n1}\|^4 \right\}^{1/2} \{P(\|Y_{n1}\| > \varepsilon)\}^{1/2} \end{aligned}$$

By condition (F) and $A_n A_n^T \rightarrow G$, we obtain

$$P(\|Y_{n1}\| > \varepsilon) \leq \frac{E \|A_n I_n^{-1/2}(\beta_{n01}) \nabla L_{n1}(\beta_{n01})\|^2}{n\varepsilon} =^{(?)} O(n^{-1})$$

and

$$\begin{aligned} E \|Y_{n1}\|^4 &= \frac{1}{n^2} E \left\| A_n I_n^{-1/2}(\beta_{n01}) \nabla L_{n1}(\beta_{n01}) \right\|^4 \\ &\leq \frac{1}{n^2} \lambda_{\max}(A_n A_n^T) \lambda_{\max}\{I_n(\beta_{n01})\} E \left\| \nabla^T L_{n1}(\beta_{n01}) \nabla L_{n1}(\beta_{n01}) \right\|^2 \\ &= O\left(\frac{p_n^2}{n^2}\right) \end{aligned}$$

Thus, we have

$$\sum_{i=1}^n E \|Y_{ni}\|^2 \mathbf{1}\{\|Y_{ni}\| > \varepsilon\} = O\left(n \frac{p_n}{n} \frac{1}{\sqrt{n}}\right) = o(1)$$

On the other hand, as $A_n A_n^T \rightarrow G$, we have

$$\sum_{i=1}^n \text{cov}(Y_{ni}) = n \text{cov}(Y_{n1}) = \text{cov} \left\{ A_n I_n^{-1/2}(\beta_{n01}) \nabla L_{n1}(\beta_{n01}) \right\} \rightarrow G$$

Thus, Y_{ni} satisfies the conditions of the Lindeberg-Feller central limit theorem [see van der Vaart (1998)]. This also means that $1/\sqrt{n} A_n I_n(\beta_{n01})^{-1/2} \nabla L_n(\beta_{n01})$ has an asymptotic multivariate normal distribution.

With a slight abuse of notation, let $Q_n(\beta_{n1}) = Q_n(\beta_{n1}, 0)$. As $\hat{\beta}_{n1}$ must satisfy the penalized likelihood equation $\nabla Q_n(\hat{\beta}_{n1}) = 0$ (the necessary condition of maximum point), using the Taylor expansion on $\nabla Q_n(\hat{\beta}_{n1})$ at point β_{n01} , we have

$$\begin{aligned} & \frac{1}{n} [\{\nabla^2 L_n(\beta_{n01}) - \nabla^2 P_{\lambda_n}(\beta_{n1}^{**})\}(\hat{\beta}_{n1} - \beta_{n01}) - \nabla P_{\lambda_n}(\beta_{n01})] \\ &= ? - \frac{1}{n} \left[\nabla L_n(\beta_{n01}) + \frac{1}{2} (\hat{\beta}_{n1} - \beta_{n01})^T \nabla^2 \{\nabla L_n(\beta_{n1}^*)\} (\hat{\beta}_{n1} - \beta_{n01}) \right] \end{aligned} \quad (6.16)$$

where β_{n1}^* and β_{n1}^{**} lie between $\hat{\beta}_{n1}$ and β_{n01} . Now we define

$$\mathcal{L} \triangleq \nabla^2 L_n(\beta_{n01}) - \nabla^2 P_{\lambda_n}(\beta_{n1}^{**})$$

and

$$\mathcal{C} \triangleq \frac{1}{2} (\hat{\beta}_{n1} - \beta_{n01})^T \nabla^2 \{\nabla L_n(\beta_{n1}^*)\} (\hat{\beta}_{n1} - \beta_{n01})$$

Now 6.16 turns to

$$\begin{aligned} & \frac{1}{n} [\mathcal{L}(\hat{\beta}_{n1} - \beta_{n01}) - \nabla P_{\lambda_n}(\beta_{n01})] \\ &= -\frac{1}{n} [\nabla L_n(\beta_{n01}) + \mathcal{C}] \end{aligned}$$

Under conditions (G) and (H) and by the Cauchy-Schwarz inequality, we have

$$\begin{aligned} \left\| \frac{1}{n} \mathcal{C} \right\|^2 &\leq \frac{1}{n^2} \sum_{i=1}^n n^2 \|\hat{\beta}_{n1} - \beta_{n01}\|^4 \sum_{j,k,l=1}^{s_n} M_{njkl}^2(V_{ni}) \\ &= O_p\left(\frac{p_n^2}{n^2}\right) O_p(p_n^3) = o_p\left(\frac{1}{n}\right) \end{aligned} \quad (6.17)$$

At the same time, by Lemma 6.4 in the Appendix and because of condition (H), it is easy to show that

$$\lambda_i \left\{ \frac{1}{n} \mathcal{L} + I_n(\beta_{n01}) + \Sigma_{\lambda_n} \right\} = o_p\left(\frac{1}{\sqrt{p_n}}\right), \quad i = 1, 2, \dots, s_n$$

where $\lambda_i(A)$ is the i th eigenvalue of a symmetric matrix A . As $\hat{\beta}_{n1} - \beta_{n01} = O_p(\sqrt{p_n/n})$.

$$\left\{ \frac{1}{n} \mathcal{L} + I_n(\beta_{n01}) + \Sigma_{\lambda_n} \right\} (\hat{\beta}_{n1} - \beta_{n01}) = o_p\left(\frac{1}{\sqrt{n}}\right) \quad (6.18)$$

Finally, from (6.17) and (6.18) we have

$$\{I_n(\beta_{n01}) + \Sigma_{\lambda_n}\}(\hat{\beta}_{n1} - \beta_{n01}) + \mathbf{b}_n = \frac{1}{n} \nabla L_n(\beta_{n01}) + o_p\left(\frac{1}{\sqrt{n}}\right) \quad (6.19)$$

Following (6.19), Theorem 6.1.2 follows. ■

PROOF OF THEOREM 3.

Proof. Let $\mathcal{A}_n = -n^{-1}\nabla^2 L_n(\hat{\beta}_{n1}) + \Sigma_{\lambda_n}(\hat{\beta}_{n1})$, $\mathcal{B}_n = \widehat{\text{cov}}\{\nabla L_n(\hat{\beta}_{n1})\}$, $\mathcal{A} = I_n(\beta_{n01}) + \Sigma_{\lambda_n}$ and $\mathcal{B} = I_n(\beta_{n01})$. Then we have

$$\begin{aligned}\hat{\Sigma}_n - \Sigma_n &= \mathcal{A}_n^{-1}(\mathcal{B}_n - \mathcal{B})\mathcal{A}_n^{-1} + (\mathcal{A}_n^{-1} - \mathcal{A}^{-1})\mathcal{B}\mathcal{A}_n^{-1} + \mathcal{A}^{-1}\mathcal{B}(\mathcal{A}_n^{-1} - \mathcal{A}^{-1}) \\ &= I_1 + I_2 + I_3\end{aligned}$$

and

$$\mathcal{A}_n^{-1} - \mathcal{A}^{-1} = \mathcal{A}_n^{-1}(\mathcal{A} - \mathcal{A}_n)\mathcal{A}^{-1}$$

Let $\lambda_i(\mathcal{A})$ be the i th eigenvalue of a symmetric matrix \mathcal{A} . If we can show that $\lambda_i(\mathcal{A} - \mathcal{A}_n) = o_p(1)$ and $\lambda_i(\mathcal{B}_n - \mathcal{B}) = o_p(1)$, then from the fact that $|\lambda_i(\mathcal{B})|$ and $|\lambda_i(\mathcal{A})|$ are uniformly bounded away from 0 and infinite, we have

$$\lambda_i(\hat{\Sigma}_n - \Sigma_n) = o_p(1)$$

This means that $\hat{\Sigma}_n$ is a weakly consistent estimator of Σ_n . First, let us consider $\mathcal{A} - \mathcal{A}_n$ and decompose it as follows:

$$\mathcal{A} - \mathcal{A}_n = I_n(\beta_{n01}) + \frac{1}{n}\nabla^2 L_n(\hat{\beta}_{n1}) + \Sigma_{\lambda_n}(\beta_{n01}) - \Sigma_{\lambda_n}(\hat{\beta}_{n1}) \triangleq K_1 + K_2$$

It is obvious that

$$\begin{aligned}\lambda_{\min}(K_1) + \lambda_{\min}(K_2) &\leq \lambda_{\min}(K_1 + K_2) \\ &\leq \lambda_{\max}(K_1 + K_2) \leq \lambda_{\max}(K_1) + \lambda_{\max}(K_2)\end{aligned}$$

Thus, we need only consider $\lambda_i(K_1)$ and $\lambda_i(K_2)$ separately. The term K_1 can be expressed as

$$K_1 = I_n(\beta_{n01}) + \frac{1}{n}\nabla^2 L_n(\beta_{n01}) - \frac{1}{n}\nabla^2 L_n(\beta_{n01}) + \frac{1}{n}\nabla^2 L_n(\hat{\beta}_{n1})$$

According to Lemma 6.4 in the Appendix, we have

$$\|I_n(\beta_{n01}) + \frac{1}{n}\nabla^2 L_n(\beta_{n01})\| = o_p(1) \quad (6.20)$$

As shown in Lemma 6.5,

$$\left\| \frac{\nabla^2 L_n(\hat{\beta}_{n1})}{n} - \frac{\nabla^2 L_n(\beta_{n01})}{n} \right\|^2 = O_p\left(\frac{p_n^4}{n}\right) = o_p(1) \quad (6.21)$$

Thus, it follows from (6.20) and (6.21) that $\|K_1\| = o_p(1)$. This also means that we have

$$\lambda_i(K_1) = o_p(1), \quad i = 1, 2, \dots, s_n \quad (6.22)$$

As $\|\hat{\beta}_{n1} - \beta_{n01}\| = O_p(\sqrt{p_n/n})$, by condition (D), $p''_{\lambda_n}(\hat{\beta}_{nj}) - p''_{\lambda_n}(\beta_{n0j}) = o_p(1)$, that is

$$\lambda_i(K_2) = o_p(1), \quad i = 1, 2, \dots, s_n \quad (6.23)$$

Hence, from (6.22) and (6.23) we have shown that

$$\lambda_i(\mathcal{A} - \mathcal{A}_n) = o_p(1), \quad i = 1, 2, \dots, s_n \quad (6.24)$$

Next we consider $\lambda_i(\mathcal{B}_n - \mathcal{B})$. First we express $\mathcal{B}_n - \mathcal{B}$ as the sum of K_3 and K_4 , where

$$K_3 \triangleq \left\{ \frac{1}{n} \sum_{i=1}^n \frac{\partial L_{ni}(\hat{\beta}_{n1})}{\partial \beta_j} \frac{\partial L_{ni}(\hat{\beta}_{n1})}{\partial \beta_k} \right\} - I_n(\beta_{n01})$$

and

$$K_4 \triangleq - \left\{ \frac{1}{n} \sum_{i=1}^n \frac{\partial L_{ni}(\hat{\beta}_{n1})}{\partial \beta_j} \right\} \left\{ \frac{1}{n} \sum_{i=1}^n \frac{\partial L_{ni}(\hat{\beta}_{n1})}{\partial \beta_k} \right\}$$

Using the aforementioned argument, we need only consider K_3 and K_4 separately. Note that

$$\frac{1}{n} \sum_{i=1}^n \frac{\partial L_{ni}(\hat{\beta}_{n1})}{\partial \beta_j} - p'_{\lambda_n}(|\hat{\beta}_{nj}|) = 0, \quad j = 1, 2, \dots, s_n$$

which implies that

$$\|K_4\|^2 = \sum_{j=1}^{s_n} \sum_{k=1}^{s_n} \{p'_{\lambda_n}(|\hat{\beta}_{nj}|)\}^2 \{p'_{\lambda_n}(|\hat{\beta}_{nk}|)\}^2 \quad (6.25)$$

$$= \left\{ \sum_{j=1}^{s_n} p'_{\lambda_n}(|\hat{\beta}_{nj}|)^2 \right\}^2 \quad (6.26)$$

By Taylor expansion,

$$p'_{\lambda_n}(|\hat{\beta}_{nj}|) = p'_{\lambda_n}(|\beta_{n0j}|) + p''_{\lambda_n}(|\beta_{nj}^*|)(\hat{\beta}_{nj} - \beta_{n0j}) \quad (6.27)$$

where β_{nj}^* lies between $\hat{\beta}_{nj}$ and β_{n0j} . From (5.22) and (5.23) we obtain

$$\|K_4\|^2 \leq 4 \left\{ \sum_{j=1}^{s_n} p'_{\lambda_n}(|\hat{\beta}_{n0j}|)^2 + C \|\hat{\beta}_{n1} - \beta_{n0}\|^2 \right\}^2 \quad (6.28)$$

$$\leq 4 \left\{ p_n a_n^2 + O_p\left(\frac{p_n}{n}\right) \right\}^2 = O_p\left(\frac{p_n^2}{n^2}\right) = o_p(1) \quad (6.29)$$

Finally, we consider K_3 . It is easy to see that K_3 can be decomposed as the sum of K_5 and K_6 , where

$$K_5 \triangleq \left\{ \frac{1}{n} \sum_{i=1}^n \frac{\partial L_{ni}(\hat{\beta}_{n1})}{\partial \beta_j} \frac{\partial L_{ni}(\hat{\beta}_{n1})}{\partial \beta_k} - \frac{1}{n} \sum_{i=1}^n \frac{\partial L_{ni}(\beta_{n01})}{\partial \beta_j} \frac{\partial L_{ni}(\beta_{n01})}{\partial \beta_k} \right\}$$

$$K_6 \triangleq \left\{ \frac{1}{n} \sum_{i=1}^n \frac{\partial L_{ni}(\beta_{n01})}{\partial \beta_j} \frac{\partial L_{ni}(\beta_{n01})}{\partial \beta_k} \right\} - I_n(\beta_{n01})$$

As before, following Lemma 6.4, it is easy to demonstrate that

$$\|K_6\| = o_p(1) \quad (6.30)$$

In the Appendix we show that

$$\|K_5\| = o_p(1) \quad (6.31)$$

By (6.28) – (6.31) we have shown that $\|\mathcal{B}_n - \mathcal{B}\| = o_p(1)$ and

$$\lambda_i(\mathcal{B}_n - \mathcal{B}) = o_p(1), \quad i = 1, \dots, s_n \quad (6.32)$$

It follows from (6.24) and (6.32) that

$$\lambda_i(\hat{\Sigma}_n - \Sigma_n) = o_p(1), \quad i = 1, \dots, s_n$$

This completes the proof for the consistency of the sandwich formula. ■

Let B_n be an $(s_n - q) \times s_n$ matrix which satisfies $B_n B_n^T = I_{s_n - q}$ and $A_n B_n^T = 0$. As β_{n1} is in the orthogonal complement to the linear space that is spanned by rows of A_n under the null hypothesis H_0 , it follows that

$$\beta_n = B_n^T \gamma$$

where γ is an $(s_n - q) \times 1$ vector. Then, under H_0 the penalized likelihood estimator is also the local maximizer $\hat{\gamma}_n$ of the problem

$$Q_n(B_n^T \hat{\gamma}_n) = \max_{\gamma_n} Q_n(B_n^T \gamma_n)$$

To prove Theorem 6.1.4 we need the following two lemmas, the proofs of which are given in the Appendix.

Lemma 6.2 Under the conditions of Theorem 4 and the null hypothesis H_0 , we have

$$\begin{aligned} \hat{\beta}_{n1} - \beta_{n01} &= \frac{1}{n} I_n(\beta_{n01})^{-1} \nabla L_n(\beta_{n01}) + o_p(n^{-1/2}) \\ B_n^T(\hat{\gamma}_n - \gamma_{n0}) &= \frac{1}{n} B_n^T \left\{ B_n I_n(\beta_{n01}) B_n^T \right\}^{-1} B_n^T \nabla L_n(\beta_{n01}) + o_p(n^{-1/2}) \end{aligned}$$

Lemma 6.3 Under the conditions of Theorem 4 and the null hypothesis H_0 , we have

$$Q_n(\hat{\beta}_{n1}) - Q_n(B_n^T \hat{\gamma}_n) = \frac{n}{2} (\hat{\beta}_{n1} - B_n^T \hat{\gamma}_n)^T I_n(\beta_{n01}) (\hat{\beta}_{n1} - B_n^T \hat{\gamma}_n) + o_p(1) \quad (6.33)$$

PROOF OF THEOREM 4.

Proof. Let $\Theta_n = I_n(\beta_{n01})$ and $\Phi_n = \frac{1}{n} \nabla L_n(\beta_{n01})$. By Lemma 6.2 we have

$$\begin{aligned} & \hat{\beta}_{n1} - B_n^T \hat{\gamma}_n \\ &= \Theta_n^{-1/2} \left\{ I_n - \Theta_n^{1/2} B_n^T (B_n \Theta_n B_n^T)^{-1} B_n \Theta_n^{1/2} \right\} \Theta_n^{-1/2} \Phi_n + o_p(n^{-1/2}) \end{aligned} \quad (6.34)$$

(Previous equation is equal to $\hat{\beta}_{n1} - \beta_{n01} - B_n^T(\hat{\gamma}_n - \gamma_{n0})$) It is easy to see that $I_n - \Theta_n^{1/2} B_n^T (B_n \Theta_n B_n^T)^{-1} B_n \Theta_n^{1/2}$ is an idempotent matrix ($A^2 = A$) with rank q . Hence, by a standard argument and condition (F),

$$\hat{\beta}_{n1} - B_n^T \hat{\gamma}_n = O_p\left(\sqrt{\frac{q}{n}}\right) \quad (6.35)$$

Substituting (6.34) into (6.33), we obtain

$$\begin{aligned} & Q_n(\hat{\beta}_{n1}) - Q_n(B_n^T \hat{\gamma}_n) \\ &= \frac{n}{2} \Phi_n^T \Theta_n^{-1/2} \left\{ I_n - \Theta_n^{1/2} B_n^T (B_n \Theta_n B_n^T)^{-1} B_n \Theta_n^{1/2} \right\} \Theta_n^{-1/2} \Phi_n + o_p(1) \end{aligned}$$

By the property of the idempotent matrix, $I_n - \Theta_n^{1/2} B_n^T (B_n \Theta_n B_n^T)^{-1} B_n \Theta_n^{1/2}$ can be written as the product form $D_n^T D_n$, where D_n is a $q \times s_n$ matrix that satisfies $D_n D_n^T = I_q$. As in the proof of Theorem 6.1.2, we have shown that $\sqrt{n} D_n \Theta_n^{-1/2} \Phi_n$ has an asymptotic multivariate normal distribution, that is,

$$\sqrt{n} D_n \Theta_n^{-1/2} \Phi_n \xrightarrow{D} \mathcal{N}(0, I_q)$$

Finally, we have

$$2 \left\{ Q_n(\hat{\beta}_{n1}) - Q_n(B_n^T \hat{\gamma}_n) \right\} = n \left(D_n \Theta_n^{-1/2} \Phi_n \right)^T \left(D_n \Theta_n^{-1/2} \Phi_n \right) + o_p(1) \xrightarrow{D} \chi_q^2$$

■

Lemma 6.4 Under the conditions of Theorem 6.1.1, we have

$$\left\| \frac{1}{n} \nabla^2 L_n(\beta_{n0}) + I_n(\beta_{n0}) \right\| = o_p\left(\frac{1}{p_n}\right) \quad (6.36)$$

and

$$\left\| \left\{ \frac{1}{n} \sum_{i=1}^n \frac{\partial L_{ni}(\beta_{n01})}{\partial \beta_j} \frac{\partial L_{ni}(\beta_{n01})}{\partial \beta_k} \right\} - I_n(\beta_{n0}) \right\| = o_p\left(\frac{1}{p_n}\right) \quad (6.37)$$

■

Proof. For any ε , by Chebyshev's inequality,

$$\begin{aligned} P \left(\left\| \frac{1}{n} \nabla^2 L_n(\beta_{n0}) + I_n(\beta_{n0}) \right\| \geq \frac{\varepsilon}{p_n} \right) &\leq \frac{p_n^2}{n^2 \varepsilon^2} E \sum_{i,j=1}^{p_n} \left\{ \frac{\partial L_n(\beta_{n0})}{\partial \beta_{ni} \beta_{nj}} - E \frac{\partial L_n(\beta_{n0})}{\partial \beta_{ni} \beta_{nj}} \right\}^2 \\ &= \frac{p_n^4}{n} = o(1) \end{aligned}$$

Hence (6.36) follows. Similarly, we can prove (6.37). ■

Lemma 6.5 Under the conditions of Theorem 6.1.2, we have

$$\left\| \frac{\nabla^2 L_n(\hat{\beta}_{n1})}{n} - \frac{\nabla^2 L_n(\beta_{n01})}{n} \right\| = o_p\left(\frac{1}{\sqrt{p_n}}\right)$$

Proof. First we expand the left-hand side of the equation above to the third order,

$$\begin{aligned} \left\| \frac{\nabla^2 L_n(\hat{\beta}_{n1})}{n} - \frac{\nabla^2 L_n(\beta_{n01})}{n} \right\|^2 &= \frac{1}{n^2} \sum_{i,j=1}^{s_n} \left\{ \frac{\partial L_n(\hat{\beta}_{n1})}{\partial \beta_i \partial \beta_j} - \frac{\partial L_n(\beta_{n01})}{\partial \beta_i \partial \beta_j} \right\}^2 \\ &=^{(i)} \frac{1}{n^2} \sum_{i,j=1}^{s_n} \left\{ \sum_{k=1}^{s_n} \frac{\partial L_n(\beta_{n1}^*)}{\partial \beta_i \partial \beta_j \partial \beta_k} (\hat{\beta}_{nk} - \beta_{n0k}) \right\}^2 \end{aligned}$$

(i) is for mean-value theorem. Then by condition (G) and the Cauchy-Schwarz inequality,

$$\begin{aligned} &\frac{1}{n^2} \sum_{i,j=1}^{s_n} \left\{ \sum_{k=1}^{s_n} \frac{\partial L_n(\beta_{n1}^*)}{\partial \beta_i \partial \beta_j \partial \beta_k} (\hat{\beta}_{nk} - \beta_{n0k}) \right\}^2 \\ &\leq \frac{1}{n^2} \sum_{i,j=1}^{s_n} \sum_{k=1}^{s_n} \left\{ \frac{\partial L_n(\beta_{n1}^*)}{\partial \beta_i \partial \beta_j \partial \beta_k} \right\}^2 \|\hat{\beta}_{n1} - \beta_{n01}\|^2 \\ &= \frac{1}{n^2} O_p\left(\frac{p_n}{n}\right) \sum_{i,j,k}^{p_n} \left\{ \sum_{l=1}^n M_{nijl}(V_{nl}) \right\}^2 \\ &= \frac{1}{n^2} O_p\left(\frac{p_n}{n}\right) O_p(p_n^3 n^2) = o_p\left(\frac{1}{p_n}\right) \end{aligned}$$

Proof. PROOF OF (6.31) .

$$\left\{ \frac{1}{n} \sum_{i=1}^n \frac{\partial L_{ni}(\hat{\beta}_{n1})}{\partial \beta_j} \frac{\partial L_{ni}(\hat{\beta}_{n1})}{\partial \beta_k} - \frac{1}{n} \sum_{i=1}^n \frac{\partial L_{ni}(\beta_{n01})}{\partial \beta_j} \frac{\partial L_{ni}(\beta_{n01})}{\partial \beta_k} \right\} = o_p(1)$$

According to Taylor's expansion, we have

$$\begin{aligned} \frac{\partial L_{ni}(\hat{\beta}_{n1})}{\partial \beta_j} &= \frac{\partial L_{ni}(\beta_{n01})}{\partial \beta_j} + \nabla^T \frac{\partial L_{ni}(\beta_{n01})}{\partial \beta_j} (\hat{\beta}_{n1} - \beta_{n0}) \\ &\quad + (\hat{\beta}_{n1} - \beta_{n0})^T \nabla^2 \frac{\partial L_{ni}(\beta_{n1}^*)}{\partial \beta_j} (\hat{\beta}_{n1} - \beta_{n0}) \\ &\triangleq a_{ij} + b_{ij} + c_{ij} \end{aligned}$$

The matrix K_5 can then be expressed as a sum of the following form:

$$\begin{aligned} K_5 &= \frac{1}{n} \left(\sum_{i=1}^n a_{ij} b_{ik} \right) + \frac{1}{n} \left(\sum_{i=1}^n a_{ij} c_{ik} \right) + \frac{1}{n} \left(\sum_{i=1}^n b_{ij} a_{ik} \right) + \frac{1}{n} \left(\sum_{i=1}^n c_{ij} a_{ik} \right) \\ &\quad + \frac{1}{n} \left(\sum_{i=1}^n b_{ij} b_{ik} \right) + \frac{1}{n} \left(\sum_{i=1}^n b_{ij} c_{ik} \right) + \frac{1}{n} \left(\sum_{i=1}^n c_{ij} b_{ik} \right) + \frac{1}{n} \left(\sum_{i=1}^n c_{ij} c_{ik} \right) \\ &\triangleq X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 \end{aligned}$$

($a_{ij}a_{ij}$ is be subtracted.) Considering a matrix of the form $n^{-1} (\sum_{i=1}^n x_{ij}y_{ik}) \triangleq F$, we have

$$\begin{aligned} \|F\|^2 &= \frac{1}{n^2} \sum_{j,k=1}^{s_n} \left(\sum_{i=1}^n x_{ij}y_{ik} \right)^2 \leq \frac{1}{n^2} \sum_{j,k=1}^{s_n} \left(\sum_{i=1}^n x_{ij}^2 \right) \left(\sum_{i=1}^n y_{ik}^2 \right) \\ &= \frac{1}{n^2} \left(\sum_{i=1}^n \sum_{j=1}^{s_n} x_{ij}^2 \right) \left(\sum_{i=1}^n \sum_{k=1}^{s_n} y_{ik}^2 \right) \end{aligned}$$

Thus, the order of $\|X_i\|$ can be determined from those of $\sum_{i=1}^n \sum_{j=1}^{s_n} a_{ij}^2$, $\sum_{i=1}^n \sum_{j=1}^{s_n} b_{ij}^2$ and $\sum_{i=1}^n \sum_{j=1}^{s_n} c_{ij}^2$.

Because of condition (F), for any i and j , $Ea_{ij}^2 \leq C$ and

$$E \left\{ \frac{\partial L_{ni}(\beta_{n01})}{\partial \beta_j \partial \beta_k} \right\}^2 \leq C \quad \text{for any } n, j \text{ and } k$$

we obtain

$$\sum_{i=1}^n \sum_{j=1}^{s_n} a_{ij}^2 = O_p(np_n) \quad (6.38)$$

and

$$\sum_{i=1}^n \sum_{j=1}^{s_n} b_{ij}^2 \leq \sum_{i=1}^n \sum_{j=1}^{s_n} \sum_{k=1}^{s_n} \left\{ \frac{\partial L_{ni}(\beta_{n01})}{\partial \beta_j \partial \beta_k} \right\}^2 \|\hat{\beta}_{n1} - \beta_{n01}\|^2 \quad (6.39)$$

$$= O_p(np_n^2) O_p\left(\frac{p_n}{n}\right) = O_p(p_n^3) \quad (6.40)$$

By condition (G) and using the Cauchy-Schwarz inequality, we show that

$$\sum_{i=1}^n \sum_{j=1}^{s_n} c_{ij}^2 \leq \sum_{i=1}^n \sum_{j=1}^{s_n} \sum_{k=1}^{s_n} \sum_{l=1}^{s_n} \left\{ \frac{\partial L_{ni}(\beta_{n1}^*)}{\partial \beta_j \partial \beta_k \partial \beta_l} \right\}^2 \|\hat{\beta}_{n1} - \beta_{n01}\|^4 \quad (6.41)$$

$$= O_p(np_n^3) O_p\left(\frac{p_n^2}{n^2}\right) = O_p\left(\frac{p_n^5}{n}\right) \quad (6.42)$$

From (6.38) – (6.41) we have

$$\begin{aligned} \|K_5\|^2 &\leq 8 \left(\|X_1\|^2 + \cdots + \|X_8\|^2 \right) \\ &\leq 8 \frac{1}{n^2} \left\{ O_p(np_n p_n^3) + O_p\left(np_n \frac{p_n^5}{n}\right) \right. \\ &\quad \left. + O_p\left(p_n^3 \frac{p_n^5}{n}\right) + O_p(p_n^3 p_n^3) + O_p\left(\frac{p_n^5}{n} \frac{p_n^5}{n}\right) \right\} \\ &= O_p\left(\frac{p_n^4}{n}\right) = o_p(1) \end{aligned}$$

This completes the proof. ■

Proof. PROOF OF LEMMA 6.2. We need only prove the second equation. The first equation can be shown in the same manner. Following the steps of the proof of Theorem 6.1.2, it follows that under H_0 , from 6.19

$$B_n (I_n (\beta_{n01}) + \Sigma_{\lambda_n}) B_n^T (\hat{\gamma}_n - \gamma_{n0}) - B_n \mathbf{b}_n = \frac{1}{n} B_n \nabla L_n (\beta_{n01}) + o_p (n^{-1/2})$$

By the conditions $a_n = o_p (1/\sqrt{np_n})$ and $B_n B_n^T = I_{s_n - q}$, we have

$$\|B_n \mathbf{b}_n\| \leq \|\mathbf{b}_n\| \stackrel{(i)}{\leq} \sqrt{p_n} a_n = o_p (n^{-1/2})$$

(i) is for a_n is the maximum term in b_n , and p_n is the number. On the other hand, since $b_n = o_p (1/\sqrt{p_n})$, we obtain

$$\|B_n \Sigma_{\lambda_n} B_n^T (\hat{\gamma}_n - \gamma_{n0})\| \leq^{(i)} \|\hat{\gamma}_n - \gamma_{n0}\| b_n = o_p \left(\frac{1}{\sqrt{p_n}} \right) O_p \left(\sqrt{\frac{p_n}{n}} \right) = o_p \left(\frac{1}{\sqrt{n}} \right)$$

(i) is for $b_n = \max_{1 \leq j \leq p_n} \{p''_{\lambda_n} (|\beta_{n0j}|), \beta_{n0j} \neq 0\}$. Hence, it follows that

$$B_n I_n (\beta_{n01}) B_n^T (\hat{\gamma}_n - \gamma_{n0}) = \frac{1}{n} B_n \nabla L_n (\beta_{n01}) + o_p (n^{-1/2})$$

As $\lambda_i (B_n I_n (\beta_{n01}) B_n^T)$ is uniformly bounded away from 0 and infinity, we have

$$B_n^T (\hat{\gamma}_n - \gamma_{n0}) = B_n^T \left\{ B_n I_n (\beta_{n01}) B_n^T \right\}^{-1} B_n \nabla L_n (\beta_{n01}) + o_p (n^{-1/2})$$

This completes the proof. ■

Proof. PROOF OF LEMMA 6.3. A Taylor's expansion of $Q_n (\hat{\beta}_{n1}) - Q_n (B_n^T \hat{\gamma}_n)$ at the point $\hat{\beta}_{n1}$ yields

$$Q_n (\hat{\beta}_{n1}) - Q_n (B_n^T \hat{\gamma}_n) = T_1 + T_2 + T_3 + T_4$$

where

$$\begin{aligned} T_1 &= \nabla^T Q_n (\hat{\beta}_{n1}) (\hat{\beta}_{n1} - B_n^T \hat{\gamma}_n) \\ T_2 &= -\frac{1}{2} (\hat{\beta}_{n1} - B_n^T \hat{\gamma}_n)^T \nabla^2 L_n (\hat{\beta}_{n1}) (\hat{\beta}_{n1} - B_n^T \hat{\gamma}_n) \\ T_3 &= \frac{1}{6} \nabla^T \left\{ (\hat{\beta}_{n1} - B_n^T \hat{\gamma}_n)^T \nabla^2 L_n (\beta_{n1}^*) (\hat{\beta}_{n1} - B_n^T \hat{\gamma}_n) \right\} (\hat{\beta}_{n1} - B_n^T \hat{\gamma}_n) \\ T_4 &= \frac{1}{2} (\hat{\beta}_{n1} - B_n^T \hat{\gamma}_n)^T \nabla^2 P_{\lambda_n} (\hat{\beta}_{n1}) \{I + o(I)\} (\hat{\beta}_{n1} - B_n^T \hat{\gamma}_n) \end{aligned}$$

Note that $T_1 = 0$ as $\nabla^T Q (\hat{\beta}_{n1}) = 0$. By Lemma 6.2 and (6.34) it follows that 6.35

$$\hat{\beta}_{n1} - B_n^T \hat{\gamma}_n = O_p \left(\sqrt{\frac{q}{n}} \right)$$

By the conditions $b_n = o_p(1/\sqrt{p_n})$ and $q < p_n$, following the proof of I_3 in Theorem 1, we have

$$T_3 = O_p\left(np_n^{3/2}n^{-3/2}q^{3/2}\right) = o_p(1)$$

and

$$T_4 \leq nb_n \left\| \hat{\beta}_{n1} - B_n^T \hat{\gamma}_n \right\|^2 = no_p\left(\frac{1}{\sqrt{p_n}}\right) O_p\left(\frac{q}{n}\right) = o_p(1)$$

Thus,

$$\begin{aligned} Q_n(\hat{\beta}_{n1}) - Q_n(B_n^T \hat{\gamma}_n) &= T_2 + o_p(1) \\ &= -\frac{1}{2} \left(\hat{\beta}_{n1} - B_n^T \hat{\gamma}_n \right)^T \nabla^2 L_n(\hat{\beta}_{n1}) \left(\hat{\beta}_{n1} - B_n^T \hat{\gamma}_n \right) + o_p(1) \end{aligned} \quad (6.43)$$

It follows from Lemmas 6.4 and 6.5 that

$$\left\| \frac{1}{n} \nabla^2 L_n(\hat{\beta}_{n1}) + I_n(\beta_{n01}) \right\| = o_p\left(\frac{1}{\sqrt{p_n}}\right)$$


Hence, we have

$$\begin{aligned} &\frac{1}{2} \left(\hat{\beta}_{n1} - B_n^T \hat{\gamma}_n \right)^T \left\{ \nabla^2 L_n(\hat{\beta}_{n1}) + nI_n(\beta_{n01}) \right\} \left(\hat{\beta}_{n1} - B_n^T \hat{\gamma}_n \right) \\ &\leq o_p\left(n \frac{1}{\sqrt{p_n}}\right) O_p\left(\frac{q}{n}\right) = o_p(1) \end{aligned}$$

The combination of (6.43) and (??) yields (6.33).

$$Q_n(\hat{\beta}_{n1}) - Q_n(B_n^T \hat{\gamma}_n) = \frac{n}{2} \left(\hat{\beta}_{n1} - B_n^T \hat{\gamma}_n \right)^T I_n(\beta_{n01}) \left(\hat{\beta}_{n1} - B_n^T \hat{\gamma}_n \right) + o_p(1)$$

■



7. On model consistency of Lasso

7.1 Model Selection Consistency and Irrepresentable Conditions

An estimate which is consistent in term of parameter estimation does not necessarily consistently select the correct model (or even attempt to do so) where the reverse is also true. The former requires

$$\hat{\beta}^n - \beta^n \rightarrow_p 0, \text{ as } n \rightarrow \infty$$

while the latter requires

$$P(\{i : \hat{\beta}_i^n \neq 0\} = \{i : \beta_i^n \neq 0\}) \rightarrow 1, \text{ as } n \rightarrow \infty$$

In general, we desire our estimate to have both consistencies. However, to separate the selection aspect of the consistency from the parameter estimation aspect, we make the following definitions about Sign Consistency that does not assume the estimates to be estimation consistent.

Definition 7.1.1 An estimate $\hat{\beta}^n$ is equal in sign with the true model β^n which is written

$$\hat{\beta}^n =_s \beta^n$$

if and only if

$$\text{sign}(\hat{\beta}^n) = \text{sign}(\beta^n)$$

where $\text{sign}(\cdot)$ maps positive entry to 1, negative entry to -1 and zero to zero, that is, $\hat{\beta}^n$ matches the zeros and signs of β .

Definition 2

Definition 7.1.2 Lasso is Strongly Sign Consistent if there exists $\lambda_n = f(n)$, that is, a function of n and independent of Y_n or X_n such that

$$\lim_{n \rightarrow \infty} P(\hat{\beta}^n(\lambda_n) =_s \beta^n) = 1$$

Definition 3

Definition 7.1.3 The Lasso is General Sign Consistent if

$$\lim_{n \rightarrow \infty} P(\exists \lambda \geq 0, \hat{\beta}^n(\lambda) =_s \beta^n) = 1$$

Strong Sign Consistency implies one can use a preselected λ to achieve consistent model selection via Lasso. General Sign Consistency means for a random realization there exists a correct amount of regularization that selects the true model. Obviously, strong sign consistency implies general sign consistency. Surprisingly, as implied by our results, the two kinds of sign consistencies are almost equivalent to one condition. To define this condition we need the following notations on the design. Without loss of generality, assume $\beta^n = (\beta_1^n, \dots, \beta_q^n, \beta_{q+1}^n, \dots, \beta_p^n)^T$ where $\beta_j^n \neq 0$ for $j = 1, \dots, q$ and $\beta_j^n = 0$ for $j = q+1, \dots, p$. Let $\beta_{(1)}^n = (\beta_1^n, \dots, \beta_q^n)^T$ and $\beta_{(2)}^n = (\beta_{q+1}^n, \dots, \beta_p^n)^T$. Now write $X_n(1)$ and $X_n(2)$ as the first q and last $p - q$ columns of X_n respectively and let $C^n = \frac{1}{n} X_n^T X_n$. By setting $C_{11}^n = \frac{1}{n} X_n(1)' X_n(1)$, $C_{22}^n = \frac{1}{n} X_n(2)' X_n(2)$, $C_{12}^n = \frac{1}{n} X_n(1)' X_n(2)$ and $C_{21}^n = \frac{1}{n} X_n(2)' X_n(1)$. C^n can then be expressed in a block-wise form as follows:

$$C^n = \begin{pmatrix} C_{11}^n & C_{12}^n \\ C_{21}^n & C_{22}^n \end{pmatrix}$$

Assuming C_{11}^n is invertible, we define the following Irrepresentable Conditions

Definition 7.1.4 Strong Irrepresentable Condition. There exists a positive constant vector η

$$\left| C_{21}^n (C_{11}^n)^{-1} \text{sign}(\beta_{(1)}^n) \right| \leq 1 - \eta$$

where $\mathbf{1}$ is a $p - q$ by 1 vector of 1's and the inequality holds element-wise.

Definition 7.1.5 Weak Irrepresentable Condition.

$$\left| C_{21}^n (C_{11}^n)^{-1} \text{sign}(\beta_{(1)}^n) \right| < \mathbf{1}$$

where the inequality holds element-wise.

Weak Irrepresentable Condition is slightly weaker than Strong Irrepresentable Condition.

C^n can converge in ways that entries of $\left| C_{21}^n (C_{11}^n)^{-1} \text{sign}(\beta_{(1)}^n) \right|$ approach 1 from the below so that Weak Condition holds but the strict inequality fails in the limit. For a fixed p and $\beta^n = \beta$, the distinction disappears for random designs when, for example, x_i^n 's are i.i.d. realizations with covariance matrix C , since then the two conditions are equivalent to $\left| C_{21} (C_{11})^{-1} \text{sign}(\beta(1)) \right| < 1$ almost surely.

The Irrepresentable Conditions closely resembles a regularization constraint on the regression coefficients of the irrelevant covariates ($\mathbf{X}_n(2)$) on the relevant covariates ($\mathbf{X}_n(1)$). In particular, when signs of the true β are unknown, for the Irrepresentable Condition to hold for all possible signs, we need the L_1 norms of the regression coefficients to be smaller than 1. To see this, recall for (2) to hold for all possible $\text{sign}(\beta(1))$, we need

$$\left| \left(\mathbf{X}_n(1)^T \mathbf{X}_n(1) \right)^{-1} \mathbf{X}_n(1)^T \mathbf{X}_n(2) \right| = \left| (C_{11}^n)^{-1} C_{12}^n \right| < 1 - \eta \quad (7.1)$$

that is, the total amount of an irrelevant covariate represented by the covariates in the true model is not to reach 1 (therefore the name "irrepresentable").

As a preparatory result, the following proposition puts a lower bound on the probability of Lasso picking the true model which quantitatively relates the probability of Lasso selecting the correct model and how well Strong Irrepresentable Condition holds: Proposition 1. Assume Strong Irrepresentable Condition holds with a constant $\eta > 0$ then

$$P(\hat{\beta}^n(\lambda_n) = \beta^n) \geq P(A_n \cap B_n)$$

for

$$\begin{aligned} A_n &= \left\{ \left| (C_{11}^n)^{-1} W^n(1) \right| < \sqrt{n} \left(\left| \beta_{(1)}^n \right| - \frac{\lambda_n}{2n} \left| (C_{11}^n)^{-1} \text{sign}(\beta_{(1)}^n) \right| \right) \right\} \\ B_n &= \left\{ \left| C_{21}^n (C_{11}^n)^{-1} W^n(1) - W^n(2) \right| \leq \frac{\lambda_n}{2\sqrt{n}} \eta \right\} \end{aligned}$$

where

$$W^n(1) = \frac{1}{\sqrt{n}} \mathbf{X}_n(1)' \varepsilon_n \text{ and } \frac{1}{\sqrt{n}} W^n(2) = \mathbf{X}_n(2)' \varepsilon_n$$

It can be argued (see the proof of Proposition 1 in the appendix) that A_n implies the signs of those of $\beta_{(1)}^n$ are estimated correctly. And given A_n, B_n further imply $\hat{\beta}_{(2)}^n$ are shrunk to zero. The regularization parameter λ_n trades off the size of these two events. Smaller λ_n leads to larger A_n but smaller B_n which makes it likely to have Lasso pick more irrelevant variables. On the other hand, larger constant η always leads to larger B_n and have no impact on A_n . So when Strong Irrepresentable Condition holds with a larger constant η , it is easier for Lasso to pick up the true model. This is quantitatively illustrated in Simulation 3.2.

Our main results relate Strong and Weak Irrepresentable Conditions with strong and general sign consistency. We describe the results for small q and p case next followed by

results for large q and p in Section 2.2. Then, analysis and sufficient conditions are given in Section 2.3 to achieve a better understanding of the Irrepresentable Conditions and relate to previous works.

7.1.1 Model Selection Consistency for Small q and p

In this section, we work under the classical setting where q, p and β^n are all fixed as $n \rightarrow \infty$. In this setting, it is natural to assume the following regularity conditions:

$$C^n \rightarrow C, \text{ as } n \rightarrow \infty \quad (7.2)$$

where C is a positive definite matrix. And,

$$\frac{1}{n} \max_{1 \leq i \leq n} \left((x_i^n)^T x_i^n \right) \rightarrow 0, \text{ as } n \rightarrow \infty \quad (7.3)$$

In practice, the covariates are usually scaled so that the diagonal elements of C^n are all 1's. The convergence in (7.2) and (7.3) are deterministic. However, the results in this section also holds quite generally for random designs. Specifically, in the case of a random design, X can be conditioned on and the asymptotic results still apply if the probability of the set where (7.2) and (7.3) hold is 1. In general, (7.2) and (7.3) are weak in the sense that if one assumes x_i are i.i.d. with finite second moments then $C = E \left((x_i^n)^T x_i^n \right), \frac{1}{n} X_n^T X_n \rightarrow$ a.s. C and $\max_{1 \leq i \leq n} x_i^T x_i = o_p(n)$, thus (7.2) and (7.3) hold naturally.

Under these conditions we have the following result. Theorem 1.

Theorem 7.1.1 For fixed q, p and $\beta^n = \beta$, under regularity conditions (3) and (4), Lasso is strongly sign consistent if Strong Irrepresentable Condition holds. That is, when Strong Irrepresentable Condition holds, for $\forall \lambda_n$ that satisfies $\lambda_n/n \rightarrow 0$ and $\lambda_n/n^{\frac{1+c}{2}} \rightarrow \infty$ with $0 \leq c < 1$, we have

$$P(\hat{\beta}^n(\lambda_n) =_s \beta^n) = 1 - o(e^{-n^c})$$

A proof of Theorem 1 can be found in the appendix. Theorem 1 states that, if Strong Irrepresentable Condition holds, then the probability of Lasso selecting the true model approaches 1 at an exponential rate while only the finite second moment of the noise terms is assumed. In addition, from Knight and Fu(2000) we know that for $\lambda_n = o(n)$ Lasso also has consistent estimation and asymptotic normality. Therefore Strong Irrepresentable Condition allows for consistent model selection and parameter estimation simultaneously. On the other hand, Theorem 2 shows that Weak Irrepresentable Condition is also necessary even for the weaker general sign consistency.

Theorem 2.

Theorem 7.1.2 For fixed p, q and $\beta_n = \beta$, under regularity conditions (3) and (4), Lasso is general sign consistent only if there exists N so that Weak Irrepresentable Condition holds for $n > N$.

A proof of Theorem 2 can be found in the appendix. Therefore, Strong Irrepresentable Condition implies strong sign consistency implies general sign consistency implies Weak Irrepresentable Condition. So except for the technical difference between the two conditions, Irrepresentable Condition is almost necessary and sufficient for both strong sign consistency and general sign consistency.

Furthermore, under additional regularity conditions on the noise terms ε_i^n , this "small" p result can be extended to the "large" p case. That is, when p also tends to infinity "not too fast" as n tends to infinity, we show that Strong Irrepresentable Condition, again, implies Strong Sign Consistency for Lasso. 2.2 Model Selection Consistency for Large p and q In the large p and q case, we allow the dimension of the designs C^n and model parameters β_n grow as n grows, that is, $p = p_n$ and $q = q_n$ are allowed to grow with n . Consequently, the assumptions and regularity conditions in Section 2.1 becomes inappropriate as C^n do not converge and β^n may change as n grows. Thus we need to control the size of the smallest entry of $\beta_{(1)}^n$, bound the eigenvalues of C_{11}^n and have the design scale properly. Specifically, we assume: There exists $0 \leq c_1 < c_2 \leq 1$ and $M_1, M_2, M_3, M_4 > 0$ so the following holds:

$$\begin{aligned} \frac{1}{n} (X_i^n)' X_i^n &\leq M_1 \text{ for } \forall i \\ \alpha' C_{11}^n \alpha &\geq M_2, \text{ for } \forall \|\alpha\|_2 = 1 \\ q_n &= O(n^{c_1}) \\ n^{\frac{1-c_2}{2}} \min_{i=1, \dots, q} |\beta_i^n| &\geq M_3 \end{aligned}$$

Condition (5) is trivial since it can always be achieved by normalizing the covariates. (6) requires the design of the relevant covariates have eigenvalues bounded from below so that the inverse of C_{11}^n behaves well. For a random design, if the eigenvalues of the population covariance matrix are bounded from below and $q_n/n \rightarrow \rho < 1$ then (6) usually follows Bai (1999)

The main conditions are (7) and (8) which are similar to the ones in Meinshausen (2005) for Gaussian graphical models. (8) requires a gap of size n^{c_2} between the decay rate of $\beta_{(1)}^n$ and $n^{-\frac{1}{2}}$ since the noise terms aggregate at a rate of $n^{-\frac{1}{2}}$, this prevents the estimation to be dominated by the noise terms. Condition (7) is a sparsity assumption which requires square root of the size of the true model $\sqrt{q_n}$ to grow at a rate slower than the rate gap which consequently prevents the estimation bias of the Lasso solutions from dominating the model parameters. Under these conditions, we have the following result: Theorem 3.

Theorem 7.1.3 Assume ε_i^n are i.i.d. random variables with finite $2k$ 'th moment $E(\varepsilon_i^n)^{2k} < \infty$ for an integer $k > 0$. Under conditions (5),(6),(7) and (8), Strong Irrepresentable Condition implies that Lasso has strong sign consistency for $p_n = o\left(n^{(c_2-c_1)k}\right)$. In particular, for $\forall \lambda_n$ that satisfies $\frac{\lambda_n}{\sqrt{n}} = o\left(n^{\frac{c_2-c_1}{2}}\right)$ and $\frac{1}{p_n} \left(\frac{\lambda_n}{\sqrt{n}}\right)^{2k} \rightarrow \infty$, we have

$$P(\hat{\beta}^n(\lambda_n) = s\beta^n) \geq 1 - O\left(\frac{p_n n^k}{\lambda_n^{2k}}\right) \rightarrow 1 \text{ as } n \rightarrow \infty$$

A proof of Theorem 3 can be found in the appendix.

Theorem 3 states that Lasso can select the true model consistently given that Strong Irrepresentable Condition holds and the noise terms have some finite moments. For example, if only the second moment is assumed, p is allowed to grow slower than $n^{c_2-c_1}$. If all moments of the noise exist then, by Theorem 3, p can grow at any polynomial rate and the probability of Lasso selecting the true model converges to 1 at a faster rate than any polynomial rate. In particular, for Gaussian noises, we have: Theorem 4

Theorem 7.1.4 — Gaussian Noise. Assume ε_i^n are i.i.d. Gaussian random variables. Under conditions (5),(6),(7) and (8), if there exists $0 \leq c_3 < c_2 - c_1$ for which $p_n = O(e^{n^{c_3}})$ then strong Irrepresentable Condition implies that Lasso has strong sign consistency. In particular, for $\lambda_n \propto n^{\frac{1+c_4}{2}}$ with $c_3 < c_4 < c_2 - c_1$

$$P(\hat{\beta}^n(\lambda_n) = s\beta^n) \geq 1 - o\left(e^{-n^{c_3}}\right) \rightarrow 1 \text{ as } n \rightarrow \infty$$

A proof of Theorem 4 can be found in the appendix. As discussed in the introduction, this result has also been obtained independently by Meinshausen and Bühlmann (2006) in their study of high dimensional multivariate Gaussian random variables. This result is obtained more directly for linear models and differs from theirs by the use of fixed designs to accommodate non-Gaussian designs. p_n is also allowed to grow slightly faster than the polynomial rates used in that work.

It is an encouraging result that using Lasso we can allow p to grow much faster than n (up to exponentially fast) while still allow for fast convergence of the probability of correct model selection to 1. However, we note that this fast rate is not achievable for all noise distributions. In general, the result of Theorem 3 is tight in the sense that if higher moments of the noise distribution do not exist then the tail probability of the noise terms does not vanish quick enough to allow p to grow at higher degree polynomial rates.

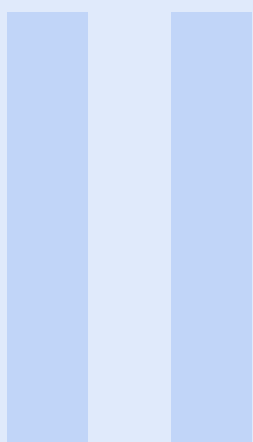
Through Theorem 3 and 4, we have shown, for cases with large p – (polynomial in n given that noise have finite moments, exponential in n for Gaussian noises), Strong Irrepresentable Condition still implies the probability of Lasso selecting the true model converges to 1 at a fast rate. We have found it difficult to show necessariness of Irrepresentable Condition for the large p setting in a meaningful way. This is mainly due to the

technical difficulty that arises from dealing with high dimensional design matrices. However, by the results for the small p case, the necessariness of Irrepresentable Condition is implied to some extent.



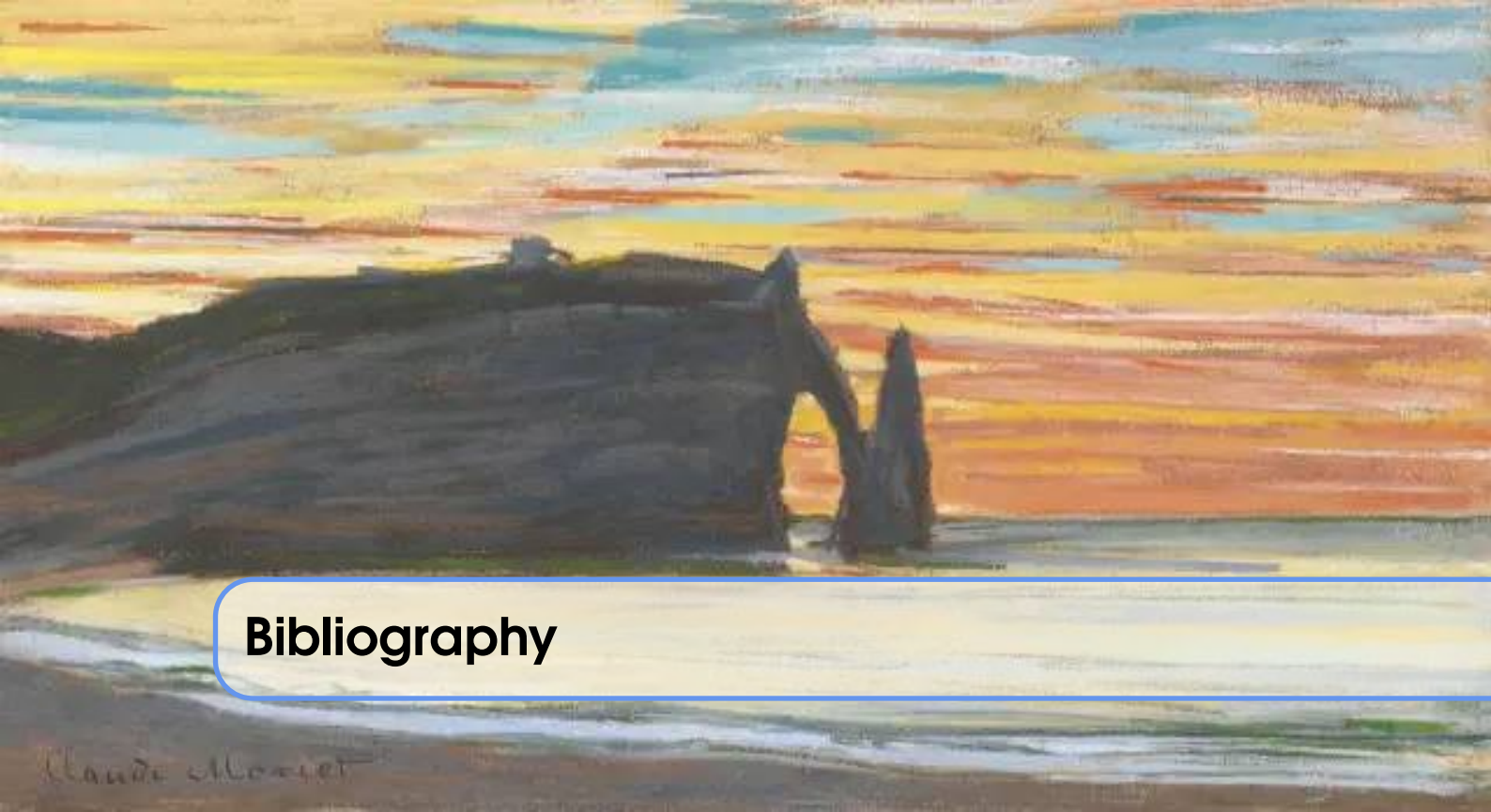
8. A selection overview of variable selection

Claude Monet



Part Four

Bibliography	217
Articles	
Books	



Bibliography

Articles

Books

