罰則付き回帰とデータ解析環境R

荒木　孝治

1996 年，Tibshirani は，線形回帰分析において最小 2 乗法に ℓ_1 罰則を課す回帰母数の推定法である lasso を提案した．これは，変数選択と回帰母数の推定を同時に行うもので，これに触発され，以降，さまざまな手法が提案されてきた．本稿では，罰則付き回帰の近年の展開を，データ解析環境 R への実装の関連で報告する．

キーワード：データ分析，罰則付き回帰，数理計画

1．はじめに

Hesterberg は，あるとき，統計学における最も重要な問題は何かと Bradley Efron に聞いた．これは，「ブートストラップ法」という返答を半ば予期しての問いかけであったが，返事は，「回帰における変数選択」であった [11].

回帰分析における変数選択は，統計学における非常に重要な問題の 1 つである．特に近年，データ収集のコストが劇的に下がったり収集が自動化されたりしたため，さまざまな分野でデータが膨大となっており，また，データ数 n に比して，変数の数 p が大きくなる状況も生じており，その重要性が増している．例えば，マイクロアレイデータの解析において，$p = 10,000$ に対して $n = 100$ といった状況がある．このとき，説明変数の数 p は大きいが，真に効果のある説明変数の数は少ないというスパース性を考えると，今までにない形での変数選択が必要となる．

1996 年，Tibshirani は，最小 2 乗法に ℓ_1 罰則というものを課すことにより，変数選択と同時に回帰母数を推定する手法を提案した [16]．これに触発され，以降，さまざまな手法が提案されており，新しい局面での活用が期待されている [4]．本稿では，罰則付き回帰の展開を，データ解析環境 R での実装状況との関連で報告する．

以下利用する記号を定義しておく．ベクトル $x = (x_1, x_2, \ldots, x_p)$ に対して，ℓ_q ノルム $\|x\|_q = (\sum_{j=1}^m x_j^q)^{1/q}$ を定める（ℓ_0：$\|x\|_0 = \#\{x_j \neq 0\}$, ℓ_∞：$\|x\|_\infty = \max_j\{|x_j|\}$）．また，$I(x) = 0$ $(x < 0)$, 1 $(x \geq 0)$，$(x)_+ = xI(x)$ とし，$\mathrm{I}_p \in \mathbb{R}^{p\times p}$ を単位行列とする．

2．線形回帰モデル

目的変数を $y \in \mathbb{R}^n$，p 個の説明変数を $x_1, x_2, \ldots, x_p \in \mathbb{R}^n$（デザイン行列 $X \in \mathbb{R}^{n\times p}$），回帰母数を $\beta \in \mathbb{R}^p$ とする線型回帰モデル $y = X\beta + \varepsilon$ を考える．誤差 ε の分布は，$N(\mathbf{0}, \sigma^2\mathrm{I}_n)$ とする．以下，y に関しては中心化，X に関しては列単位で標準化しているものとする．

β の推定には，通常，最小 2 乗法 (ols：ordinary least squares) を用いる．これは，

$$\|y - X\beta\|_2^2 = \sum_{i=1}^n (y_i - \sum_{j=1}^p x_{ij}\beta_j)^2 \tag{2.1}$$

を最小とする β をその推定値 $\widehat{\beta}$ とする方法であり，これを，

$$\widehat{\beta}^{ols} = \underset{\beta}{\operatorname{argmin}} \|y - X\beta\|_2^2 \tag{2.2}$$

と表す．X^TX の逆行列 $(X^TX)^{-1}$ が存在するとき，$\widehat{\beta}^{ols} = (X^TX)^{-1}X^Ty$ となる．$p > n$ のとき，$(X^TX)^{-1}$ は存在しない．

変数選択においては，真に効果のある説明変数のみを残し，モデルを簡単にするということと，将来のデータに対する予測精度の向上という 2 つの観点が重要である．変数選択の伝統的な方法として，総当たり法（すべての組合せのモデルを考える）や逐次選択法がある．また，選択の際の規準としては，Mallows の C_p，赤池情報量規準 AIC，ベイズ情報量規準 BIC などが用いられている．これらの規準は，「モデルの適合性の尺度 + モデルの複雑さに対する罰則項」と考えることができ，罰則項は，モデルに取り込まれた説明変数の数となっている．こうした手法は，β に関する非連続な

あらき　たかはる
関西大学商学部
〒 564–8680 大阪府吹田市山手町 3–3–35

関数の最適化の問題であることから，データの変動に対して選択が不安定となることが知られている [3].

3. 罰則付き回帰の基礎

3.1 リッジ回帰

線型モデルの推定において説明変数間に強い相関があるという多重共線性に対処するための手法として，リッジ (ridge) 回帰がある [12]．リッジ回帰では，通常の最小 2 乗法に対して，回帰係数の 2 次の項に関する制約を式 (3.1) に示す形で課す.

$$\min_{\beta} \|y - X\beta\|_2^2 \quad \text{s.t. } \|\beta\|_2^2 \le s. \tag{3.1}$$

式 (3.1) と同等な表現として式 (3.2) がある [10]．この括弧内の第 2 項を ℓ_2 罰則項という.

$$\widehat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}}\left\{\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2\right\}, \quad \lambda \ge 0. \tag{3.2}$$

λ は罰則の強さを調整する（チューニング）パラメータである．$\lambda = 0$ のとき，最小 2 乗法であり，λ が大きくなると罰則が強くなる．式 (3.2) は陽に解くことができ，$\widehat{\beta}^{ridge} = (X^TX + \lambda \mathrm{I}_p)^{-1}X^Ty$ である．リッジ回帰では β の推定値を縮小させるが，変数選択は行われない．説明変数が多くあったり，それらの相関が強かったりするとき，データに対する過適合の傾向を減少させる機能を持つ.

3.2 lasso 前史

罰則項を一般に $p_\lambda(|\beta|)$ とするとき，罰則付き回帰の問題は,

$$\underset{\beta}{\operatorname{argmin}}\left\{\|y - X\beta\|_2^2 + p_\lambda(|\beta|)\right\} \tag{3.3}$$

と表すことができる．bridge 回帰 [9] は，罰則項を $p_\lambda(|\beta|) = \lambda\sum_{j=1}^p |\beta|_j^q$ とする．これは，$q = 0$ のとき，0 でない回帰母数 β の要素の数が罰則（ℓ_0 罰則）となり，$q = 2$ のときリッジ回帰（ℓ_2 罰則）となる．$0 < q \le 1$ のときに変数選択が行われ，$q \ge 1$ のときに縮小が生じる.

また，nonnegative garrote [2] は，式 (3.4) に示す罰則付き回帰である．これは，非負の因子 $u = (u_1, u_2, \ldots, u_p)$ を利用して，$\widehat{\beta}^{ols} \cdot u = (u_1\widehat{\beta}_1^{ols}, u_2\widehat{\beta}_2^{ols}, \ldots, u_p\widehat{\beta}_p^{ols})^T$ の形で $\widehat{\beta}^{ols}$ の各要素を個別に直接縮小して推定する.

$$\widehat{\beta}^{ng} = \underset{\beta}{\operatorname{argmin}}\left\{\|y - X\widehat{\beta}^{ols} \cdot u\|_2^2 + \lambda\sum_{j=1}^p u_j\right\}, \quad u_j \ge 0. \tag{3.4}$$

3.3 lasso

リッジ回帰では変数選択できない．nonnegative garrote は，$p > n$ の状況では最小 2 乗推定量が存在しないため利用できない．そこで Tibshirani は，罰則を $p_\lambda(|\beta|) = \lambda\sum_{j=1}^p |\beta_j|$（$\ell_1$ 罰則）とする lasso (least absolute selection and shrinkage operator) を考えた [16, 17]．ちなみに，garrote が“絞首刑具”を意味するのに対して，lasso は“投げ縄”を意味する.

説明変数の数 p が 2 の場合，リッジ回帰と lasso とを模式図で表すと図 1 のようになる．リッジ回帰では推定量を縮小させるが，その値は 0 には収束しない．これに対して lasso は，0 に縮小させることができるので，変数選択と縮小を同時に行う．lasso は，C_p や AIC などの ℓ_0 とリッジ回帰の ℓ_2 との中間的罰則を持つ推定法である.

lasso は，変数選択とパラメータの推定を同時に行うという利点を持つが，(a) 推定に偏りが生じ，(b) 説明変数間に相関がある場合，そのうちの少数個のみが選択される傾向にあり，(c) $p > n$ のとき，選択される説明変数の個数は最大 n であるという欠点を持つ [23].

3.4 オラクル性

Fan and Li は，変数選択における望ましい性質として,

- 変数選択の一致性：サンプルサイズ n が大きくなるとき，0 でない係数（$\beta_j \neq 0$）を持つ説明変数が正しく選択される確率が 1 に収束する,
- 漸近正規性：0 でない係数を持つ説明変数に対する推定量は，漸近正規性を持つ,

というオラクル性（Oracle property）を提案した [7].

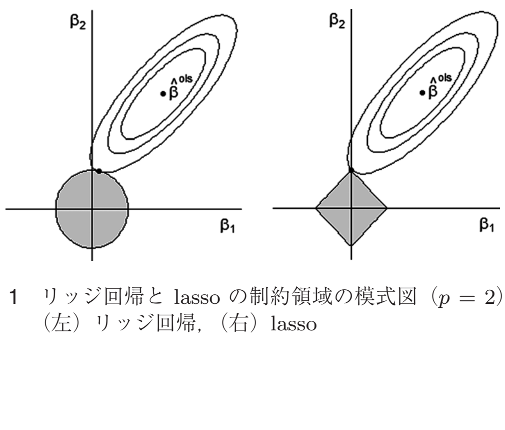

図 1 リッジ回帰と lasso の制約領域の模式図（$p = 2$）：（左）リッジ回帰,（右）lasso

3.5 罰則付き回帰の展開

lasso は，当時の 2 次計画法のプログラムを用いて解かれていた．これはあまり拡張性がなく，計算の効率が良くなかった [17]. lasso のこうしたアルゴリズム上の問題を克服するための画期的な提案が LARS (Least Angle Regression Selection) であった [6]. これは，λ のすべての値に対して lasso の最小化問題を解く手法で，計算効率を大きく向上させることができ，lasso が注目される契機になった．

lasso の欠点は，1 つのチューニングパラメータ λ で調整することから生じている．SCAD (Smoothly Clipped Absolute Deviation) [7] は，式 (3.5) に示す非凸関数の罰則 $p_\lambda(|\beta|)$ を持つ手法である．

$$p'_\lambda(\theta) = \lambda\left\{I(\theta \le \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda)\right\}, \quad \theta > 0, \quad a > 2. \tag{3.5}$$

これは 2 つのチューニングパラメータ λ, a を持ち（$a = 3.7$ が推奨されている），小さな推定値には大きな罰則，大きな推定値には小さな罰則を与える．また，一定の条件のもとで，SCAD はオラクル性を持つ．lasso の罰則関数が凸関数であるのに対して，SCAD では非凸関数となっている．罰則関数が微分不可能な点を持つ非凸関数のとき，得られる罰則付き推定値は，スパース性，連続性，不偏性という良い性質を持つ．しかし，モデルへの当てはめが困難なため，局所 2 次近似や局所線形近似等の計算法が提案されている． [7, 24]

adaptive lasso [22] は，罰則項として $p_\lambda(|\beta|) = \lambda\sum_{j=1}^{p} \widehat{w}_j|\beta_j|$ を持つ．これは，nonnegative garrote 同様の 2 段階推定法である．第 1 段階で β の初期推定量 $\widehat{\beta}^{init}$ を求め，第 2 段階で，これを利用した重み $\widehat{w} = 1/|\widehat{\beta}^{init}|^\gamma, \gamma > 0$, を利用する ℓ_1 罰則付き回帰を行う．これは SCAD と同様，0 に近い推定量に対しては大きな重みを適用して 0 にし，絶対値の大きな推定量に対しては，小さな重みを適用することにより推定の偏りを減少させる．初期推定量 $\widehat{\beta}^{init}$ としては，$\widehat{\beta}^{ols}$, $\widehat{\beta}^{ridge}$ などが提案されている．adaptive lasso もオラクル性を持つが [22]，そのパフォーマンスは初期推定量に影響を受ける．

Dantzig selector [5] は，$\|X^T(y - X\beta)\|_\infty < \lambda$ という制約の下で $\|\beta\|_1$ を最小とする β の推定法である．これは，残差と X との相関を一定の範囲に押さえて，β を推定することに相当する．線形計画法を用いて解くことができ，著者らが論文を準備しているときに，シンプレックス法の開発者である G. B. Dantzig が亡くなったという知らせを受けて命名された．また，MCP (Minimax Concave Plus) [21] は，罰則を $p_\lambda(|\beta|) = \lambda\sum_{j=1}^{p} g_{\lambda,\gamma}(|\beta_j|)$ とする手法である．ここで，$\lambda \ge 0$, $\gamma > 1$ に対して，

$$g_{\lambda,\gamma}(\theta) = \int_0^\theta (1 - x/(\gamma\lambda))_+ dx \tag{3.6}$$

である．SCAD 同様，MCP も非凸の罰則である．MCP は，パラメータ γ により，lasso の解を最小 2 乗推定量に近づける機能を持つ．$\gamma \to \infty$ のとき，MCP と lasso は一致する．MCP は SCAD と比べて構造が簡単で，計算も容易であるという利点を持つ．

罰則付き回帰の重要な展開の方向として，説明変数の選択におけるグルーピングの問題がある．説明変数間にグループや階層的な関係があることがある．例えば，交互作用が選択される場合，交互作用に含まれる主効果も選択したいという状況や，多項式回帰の場合，高次の項が存在するとき，それよりも低次の項も選択したいという状況である．グループが未知の場合，罰則項を $p_\lambda(|\beta|) = \lambda_1\sum_{j=1}^{p}|\beta_j| + \lambda_2\sum_{j=1}^{p}\beta_j^2$ とする elastic net が提案されている [23]. これは，lasso (ℓ_1) とリッジ回帰 (ℓ_2) の罰則を結合したもので，相関が強い変数はグループとして選択され，その係数は同じような値として推定される傾向を持つ．さらに，$p > n$ のとき，n 個以上の説明変数を選択することができる．なお，変数のグループが既知の場合の手法として，group lasso がある [19].

3.6 罰則の機能

デザイン行列 X が $X^TX = \mathrm{I}_n$ を満たす簡単な場合を考える．このとき，lasso, elastic net, SCAD などの罰則付き回帰問題は陽に解くことができ，最小 2 乗推定量 $\widehat{\beta}^{ols}$ との間に，次の関係がある [7, 20].

$$\widehat{\beta}^{lasso} = \mathrm{sign}(\widehat{\beta}^{ols})(|\widehat{\beta}^{ols}| - \lambda)_+, \tag{3.7}$$

$$\widehat{\beta}^{enet} = \mathrm{sign}(\widehat{\beta}^{ols})(|\widehat{\beta}^{ols}| - \lambda_1)_+/(1 + 2\lambda_2), \tag{3.8}$$

$$\widehat{\beta}^{SCAD} = \begin{cases} \mathrm{sign}(\widehat{\beta}^{ols})(|\widehat{\beta}^{ols}| - \lambda)_+ & |\widehat{\beta}^{ols}| \le 2\lambda, \\ \{(a-1)\widehat{\beta}^{ols} - a\lambda\cdot\mathrm{sign}(\widehat{\beta}^{ols})\}/(a-2) & 2\lambda < |\widehat{\beta}^{ols}| \le a\lambda, \\ \widehat{\beta}^{ols} & |\widehat{\beta}^{ols}| > a\lambda. \end{cases} \tag{3.9}$$

これらを図 2 に示す．図より，$\widehat{\beta}^{ols}$ で絶対値が小さなものを 0 としている．また，lasso や elastic net で

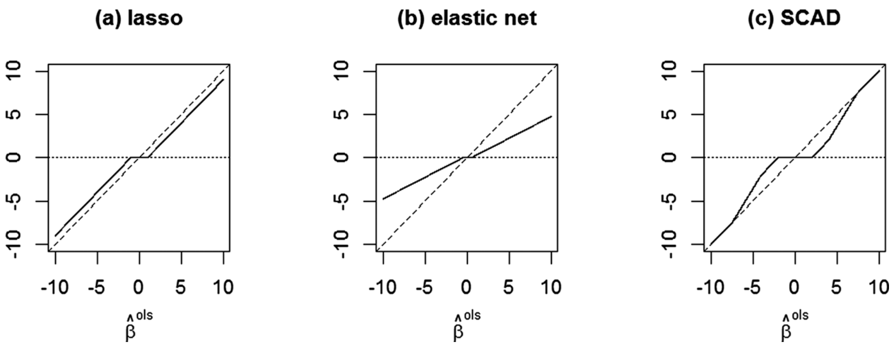

図 2　最小 2 乗推定量 $\widehat{\beta}^{ols}$（破線）と (a) lasso（$\lambda = 1$），(b) elastic net（$\lambda_1 = \lambda_2 = 1/2$），(c) SCAD（$\lambda = 2,\ a = 3.7$）の関係

は，0 に向けて縮小されるために偏りが生じている．しかし SCAD では，絶対値がより大きくなると，$\widehat{\beta}^{ols}$ と一致している．

3.7　ハイブリッド法

罰則付き回帰の発展の方向として，新しい罰則の開発以外に，罰則のハイブリッド化と手法のハイブリッド化の 2 つを考えることができる．

3.7.1　罰則の結合

elastic net は，ℓ_1 と ℓ_2 の 2 つの罰則のハイブリッド法であった．これは，罰則項を組み合わせることにより，1 つの罰則項では実現できない性質の獲得を目的としている．基本的には，縮小する罰則項（例えば，ℓ_2）と，変数選択する罰則項（例えば，ℓ_1 のように微分不可能な点を持つもの）との組合せとなる．SCAD と ℓ_2 を結合する罰則を利用する方法も提案されている [18, 20]．これは，lasso や SCAD，elastic net と比べて，特に説明変数間の関係が強いとき，優れたパフォーマンスを示すと報告されている．

変数選択で，0 でない回帰母数の個数自体を罰則項とする ℓ_0 を用いることは自然である．C_p や AIC，BIC なども，パラメータは固定されているが，ℓ_0 を用いる罰則付き指標と考えることができる．しかし，こうした最小化問題は NP 困難な問題である．そのため例えば，$\ell_0^\varepsilon(\theta) = 1\ (|\theta| \geq \varepsilon)$, $|\theta|/\varepsilon\ (|\theta| < \varepsilon)$ といった形で ℓ_0 を凸化したものを利用する $\alpha\ell_0^\varepsilon(\theta) + (1-\alpha)\ell_1(\theta)$ という罰則が提案されている [13]．

3.7.2　手法の結合

nonnegative garrote や adaptive lasso では，最小 2 乗推定量を求め，それを用いて第 2 段階での推定を罰則付きで行った．relaxed lasso [14] では，パラメータを $\lambda \in [0, \infty)$ とする推定を lasso を用いて行い，その 0 でない推定量に対して $\lambda \in [0, 1)$ とする lasso を再度適用する．なお，lasso による変数選択の後，選択された変数のみを用いて再度最小 2 乗法を適用し，推定の偏りを避けるという lasso と ols のハイブリッド法の提案もある [6]．

超高次元の場合（$p \gg n$），一定のスクリーニングを行うことにより，$p > n$ の状況に説明変数の数を絞ってから罰則付き回帰を行うことが考えられる．そのため，こうしたスクリーニング法の研究も活発になっている．その 1 つに，sure independence screening がある [8]．

4.　チューニングパラメータの選択

チューニングパラメータ λ の選択は，C_p や情報量規準（AIC，BIC 等）を用いて行うことができる．しかし，非凸の罰則に対しては，これらが弱点を持つことが指摘されており [1]，交差検証法 (CV: Cross Validation) や一般化交差検証法 (GCV: Generalized CV) の利用が提案されている．

5.　データ解析環境 R

罰則付き回帰は，現在非常に活発な研究対象となっているが，lasso は，発表後，しばらくのあいだ注目されなかった．Tibshirani は，後にその原因を，1996 年当時のコンピュータの計算速度の問題や，スパース性の統計的・数値的優位性が即座に評価されなかったこと，巨大データ（n または p）の問題が稀だったことに加えて，新しい手法を素早く，容易に共有できる R 言語を研究コミュニティが持っていなかったからではないかと回顧している [17]．

R 言語は，世界中の第一線の研究者や実務家が協力して開発しているオープンソースのデータ解析環境であり [15]，Windows や Linux，Mac OS で利用できる．また，ライセンスは GNU GPL に基づくのでフリーに利用でき，改変や機能を追加した結果を一定の

表 1　罰則付き回帰に関連する R のパッケージ

推定方式	パッケージ
adaptive lasso	`lqa, msgps, parcor, quadrupen`
bayesian lasso	`bayesQR, BLR, fGWAS, monomvn`
bridge	`lqa`
Dantzig selector	`flave, hdlm`
elastic net	`elasticnet, glmnet, lqa, msgps, pensim, quadrupen`
fused lasso	`cghFLasso, FLLat, flsa, genlasso, gvcm.cat, lqa, penalized`
group bridge	`grpreg`
group lasso	`gglasso, grplasso, grpreg, SGL, standGL`
group MCP	`grpreg`
group SCAD	`grpreg`
lasso	`apple, biglars, elasticnet, fastcox, genlasso, glmmlasso, glmnet, hdlm, lars, lassoshooting, lmmlasso, lqa, msgps, penalized, pensim, plus, quadrupen, SGL`
least angle regression	`biglars, lars, parcor`
MCP	`apple, cvplogistic, hdlm, ncvreg, plus, sparsenet`
nonnegative garrote	`lqa`
relaxed lasso	`relaxo`
ridge	`foba, parcor, penalized, pensim, ridge`
SCAD	`cvplogistic, hdlm, lqa, ncvreg, plus, SIS, quantreg`
sure independence screening	`SIS`

ルールの下で自由に配布できる．標準的なさまざまな分析手法を利用できるのみならず，最新の論文で発表された分析手法がパッケージなどの形で公開されることも多く，即座に共有・試行することができる．プログラミング言語でもあるため，目的に応じて自由に機能を変更・拡張できる．

R は，2005 年に公開されたバージョン 2.1.0 より国際化版となった．メッセージが日本語化され，日本語のデータや変数名を取り扱うことができる．R の利用法やプログラミングに関する日本語の本も 100 冊以上出版されている．また，日本語による R に関する豊富な情報が RjpWiki (http://www.okada.jp.org/RWiki/) から入手できる．また，英語環境では，例えば R に関するブログ記事を集めた R Blogger（http://www.r-bloggers.com/）がある．

R は基本パッケージと拡張パッケージから構成される．基本パッケージで，基本的な分析を実行することができ，拡張パッケージを用いると，より目的に特化した，あるいはより高度な分析などが可能となる．パッケージ数は，The Comprehensive R Archive Network (CRAN；http://cran.r-project.org/) で公開されているものだけでも 4,200 を超えている．そのため，パッケージを目的で分類した CRAN Task Views というガイドがウェブ (http://cran.r-project.org/web/views/) で提供されている．また，バイオインフォマティクス関連のパッケージが Bioconductor (http://www.bioconductor.org) で公開されている．

罰則付き回帰に関連する R のパッケージも多くある．それらの一部を表 1 に示す．また，罰則付き回帰の考え方は，他の手法，例えば，一般化線形モデル，判別分析，サポートベクターマシーン，主成分分析，因子分析，偏回帰，グラフィカルモデリング，時系列解析などにも適用されており，現在，活発な研究領域となっている [4, 10]．これらの代表的なものもパッケージとして提供されている．

参考文献

[1] P. Breheny and J. Huang, Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection, *The Annals of Applied Statistics*, **5**, 232–253, 2011.

[2] L. Breiman, Better subset regression using the nonnegative garrote, *Technometrics*, **37**, 373–384, 1995.

[3] L. Breiman, Heuristics of instability and stabilization in model selection, *The Annals of Statistics*, **24**, 2350–2383, 1996.

[4] P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer, 2011.

[5] E. Candes and T. Tao, The Dantzig selector: Statistical estimation when p is much larger than n (with discussion), *The Annals of Statistics*, **35**, 2313–2404, 2007.

[6] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, Least angle regression (with discussion), *The Annals of Statistics*, **32**, 407–499, 2004.

[7] J. Fan and R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348–

1360, 2001.
[8] J. Fan and J. Lv, Sure independence screening for ultrahigh dimensional feature space (with discussion), *Journal of the Royal Statistical Society* B, **70**, 849–911, 2008.
[9] I. E. Frank and J. H. Friedman, A statistical view of some chemometrics regression tools, *Technometrics*, **35**, 109–148, 1993.
[10] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed, Springer, 2009.
[11] T. Hesterberg, N. H. Choi, L. Meier, and C. Fraley, Least angle and ℓ_1 penalized regression: A review, *Statistics Surveys*, **2**, 61–93, 2008.
[12] A. E. Hoerl and R. W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, **12**, 55–67, 1970.
[13] Y. Liu and Y. Wu, Variable selection via a combination of the L_0 and L_1 penalties, *Journal of Computational and Graphical Statistics*, **16**, 782–798, 2007.
[14] N. Meinshausen, Relaxed lasso, *Computational Statistics & Data Analysis*, **52**, 374–393, 2007.
[15] R Development Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2012, ISBN 3-900051-07-0, http://www.R-project.org/.
[16] R. Tibshirani, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society* B, **58**, 267–288, 1996.
[17] R. Tibshirani, Regression shrinkage and selection via the lasso: A retrospective, *Journal of the Royal Statistical Society* B, **73**, 273–282, 2011.
[18] X. Wang, T. Park, and K. C. Carriere, Variable selection via combined penalization for high-dimensional data analysis, *Computational Statistics & Data Analysis*, **54**, 2230–2243, 2010.
[19] M. Yuan and Y. Lin, Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society* B, **68**, 49–67, 2006.
[20] L. Zeng and J. Xie, Group variable selection via SCAD-L_2, *Statistics: A Journal of Theoretical and Applied Statistics*, iFirst, 1–18, 2012.
[21] C.-H. Zhang, Nearly unbiased variable selection under minimax concave penalty, *The Annals of Statistics*, **38**, 894–942, 2010.
[22] H. Zou: The adaptive lasso and its oracle properties, *Journal of the American Statistical Association*, **101**, 1418–1429, 2006.
[23] H. Zou and T. Hastie, Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society* B, **67**, 301–320, 2005.
[24] H. Zou and R. Li, One-step sparse estimates in nonconcave penalized likelihood models (with discussion), *The Annals of Statistics*, **36**, 1509–1566, 2008.

