Linear Model and Least Squares

Data: \((Y_i, X_i)\), \(X_i = (X_{i1}, \ldots, X_{ip})'\), \(i = 1, \ldots, n\).
\(Y_i\): continuous

LM: \(Y_i = \beta_0 + \sum_{j=1}^{p} X_{ij} \beta_j + \epsilon_i\),
\(\epsilon_i\)'s iid with \(E(\epsilon_i) = 0\) and \(\text{Var}(\epsilon_i) = \sigma^2\).

\(\text{RSS}(\beta) = \sum_{i=1}^{n} (Y_i - \beta_0 - \sum_{j=1}^{p} X_{ij} \beta_j)^2 = ||Y - X\beta||^2_2\).

LSE (OLSE): \(\hat{\beta} = \text{arg min}_\beta \text{RSS}(\beta) = (X'X)^{-1}X'Y\).

Nice properties: Under true model,
\(E(\hat{\beta}) = \beta\),
\(\text{Var}(\hat{\beta}) = \sigma^2 (X'X)^{-1}\),
\(\hat{\beta} \sim N(\beta, \text{Var}(\hat{\beta}))\),
Gauss-Markov Theorem: \(\hat{\beta}\) has min var among all linear unbiased estimates.
Some questions:
\[ \hat{\sigma}^2 = \frac{RSS(\hat{\beta})}{(n - p - 1)} \].
Q: what happens if the denominator is \( n \)?
Q: what happens if \( X'X \) is (nearly) singular?

What if \( p \) is large relative to \( n \)?

Variable selection:
forward, backward, stepwise: fast, but may miss good ones;
best-subset: too time consuming.
FIGURE 3.6. Comparison of four subset-selection techniques on a simulated linear regression problem $Y = X^T \beta + \varepsilon$. There are $N = 300$ observations on $p = 31$ standard Gaussian variables, with pairwise correlations all equal to 0.85. For 10 of the variables, the coefficients are drawn at random from a $N(0, 0.4)$ distribution; the rest are zero. The noise
Shrinkage or regularization methods

- Use regularized or penalized RSS:
  \[
  PRSS(\beta) = RSS(\beta) + \lambda J(\beta).
  \]
  \(\lambda\): penalization parameter to be determined; 
  (thinking about the p-value threshold in stepwise selection, or 
  subset size in best-subset selection.)

- Ridge: 
  \[J(\beta) = \sum_{j=1}^{p} \beta_j^2; \text{ prior: } \beta_j \sim N(0, \tau^2).\]
  \[
  \hat{\beta}^R = (X'X + \lambda I)^{-1}X'Y.
  \]

- Properties: biased but small variances,
  \[
  E(\hat{\beta}^R) = (X'X + \lambda I)^{-1}X'X \beta,
  \]
  \[
  \text{Var}(\hat{\beta}^R) = \sigma^2(X'X + \lambda I)^{-1}X'X(X'X + \lambda I)^{-1} \leq \text{Var}(\hat{\beta}),
  \]
  \[
  df(\lambda) = tr[X(X'X + \lambda I)^{-1}X'] \leq df(0) = tr(X(X'X)^{-1}X') = 
  tr((X'X)^{-1}X'X) = p.
  \]
Lasso: $J(\beta) = \sum_{j=1}^{p} |\beta_j|$.  
Prior: $\beta_j$ Laplace or DE(0, $\tau^2$);  
No closed form for $\hat{\beta}^L$.  

Properties: biased but small variances,  
$df(\hat{\beta}^L) = \#$ of non-zero $\hat{\beta}_j^L$’s (Zou et al.).  

Special case: for $X'X = I$, or simple regression ($p = 1$),  
$\hat{\beta}_j^L = ST(\hat{\beta}_j, \lambda) = \text{sign}(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda)_+$,  
compared to:  
$\hat{\beta}_j^R = \hat{\beta}_j/(1 + \lambda)$,  
$\hat{\beta}_j^B = HT(\hat{\beta}_j, M) = \hat{\beta}_j I(\text{rank}(\hat{\beta}_j) \leq M)$.  

A key property of Lasso: $\hat{\beta}_j^L = 0$ for large $\lambda$, but not $\hat{\beta}_j^R$.  
—simultaneous parameter estimation and selection.
Note: for a convex $J(\beta)$ (as for Lasso and Ridge), min PRSS is equivalent to:
\[
\min \text{RSS}(\beta) \quad \text{s.t.} \quad J(\beta) \leq t.
\]
Offer an intuitive explanation on why we can have $\hat{\beta}_j^L = 0$; see Fig 3.11.
Theory: $|\beta_j|$ is singular at 0; Fan and Li (2001).
How to choose $\lambda$?
obtain a solution path $\hat{\beta}(\lambda)$, then, as before, use tuning data or CV or model selection criterion (e.g. AIC or BIC).
Example: R code ex3.1.r
**FIGURE 3.11.** Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the red ellipses are the contours of the least squares error function.
FIGURE 3.8. Profiles of ridge coefficients for the prostate cancer example, as the tuning parameter $\lambda$ is varied. Coefficients are plotted versus $df(\lambda)$, the effective degrees of freedom. A vertical line is drawn at $df = 5.0$, the value chosen by cross-validation.
FIGURE 3.10. Profiles of lasso coefficients, as the tuning parameter \( t \) is varied. Coefficients are plotted versus \( s = t / p \). A vertical line is drawn at \( s = 0 \), the value chosen by cross-validation. Compare Figure 3.8 on page 9; the lasso profiles hit zero, while those for ridge do not. The profiles are piece-wise linear, and so are computed only at the points displayed.
Lasso: biased estimates; alternatives:

Relaxed lasso: 1) use Lasso for VS; 2) then use LSE or MLE on the selected model.

Use a non-convex penalty:
SCAD: eq (3.82) on p.92;
Bridge $J(\beta) = \sum_j |\beta_j|^q$ with $0 < q < 1$;
Adaptive Lasso (Zou 2006): $J(\beta) = \sum_j |\beta_j|/|\tilde{\beta}_j,0|$;
Truncated Lasso Penalty (Shen, Pan & Zhu 2012, JASA): $J(\beta; \tau) = \sum_j \min(|\beta_j|, \tau)$, or $J(\beta; \tau) = \sum_j \min(|\beta_j|/\tau, 1)$.

Choice b/w Lasso and Ridge: bet on a sparse model? risk prediction for GWAS (Austin, Pan & Shen 2013, SADM).

Elastic net (Zou & Hastie 2005):

$$J(\beta) = \sum_j \alpha |\beta_j| + (1 - \alpha)\beta_j^2$$

may select correlated $X_j$'s.
FIGURE 3.20. The lasso and two alternative non-convex penalties designed to penalize large coefficients less. For SCAD we use $\lambda = 1$ and $a = 4$, and $\nu = \frac{1}{2}$ in the last panel.
Group Lasso: a group of variables are to be 0 (or not) at the same time,
\[ J(\beta) = ||\beta||_2, \]
i.e. use \( L_2 \)-norm, not \( L_1 \)-norm for Lasso or squared \( L_2 \)-norm for Ridge.
better in VS (but worse for parameter estimation?)
Grouping/fusion penalties: encouraging equalities b/w \( \beta_j \)'s (or \( |\beta_j| \)'s).
  - Fused Lasso: \( J(\beta) = \sum_{j=1}^{p-1} |\beta_j - \beta_{j+1}| \)
  - Ridge penalty: grouping implicitly, why?
  - (8000) Grouping pursuit (Shen & Huang 2010, JASA):
\[ J(\beta; \tau) = \sum_{j=1}^{p-1} TLP(\beta_j - \beta_{j+1}; \tau) \]
Grouping penalties:

1. **(8000) Zhu, Shen & Pan (2013, JASA):**

   \[
   J_2(\beta; \tau) = \sum_{j=1}^{p-1} TLP(|\beta_j| - |\beta_{j+1}|; \tau);
   \]

   \[
   J(\beta; \tau_1, \tau_2) = \sum_{j=1}^{p} TLP(\beta_j; \tau_1) + J_2(\beta; \tau_2);
   \]

2. **(8000) Kim, Pan & Shen (2013, Biometrics):**

   \[
   J'_2(\beta) = \sum_{j \sim k} |I(\beta_j \neq 0) - I(\beta_k \neq 0)|;
   \]

   \[
   J_2(\beta; \tau) = \sum_{j \sim k} |TLP(\beta_j; \tau) - TLP(\beta_k; \tau)|;
   \]

3. **(8000) Dantzig Selector (§3.8).**

4. **(8000) Theory (§3.8.5); Greenshtein & Ritov (2004) (persistence); Zou 2006 (non-consistency) ...**
R packages for penalized GLMs (and Cox PHM)

- glmnet: Ridge, Lasso and Elastic net.
- ncvreg: SCAD, MCP
- TLP: [https://github.com/ChongWu-Biostat/glmtlp](https://github.com/ChongWu-Biostat/glmtlp)
  Vignette: [http://www.tc.umn.edu/~wuxx0845/glmtlp](http://www.tc.umn.edu/~wuxx0845/glmtlp)
- FGSG: grouping/fusion penalties (based on Lasso, TLP, etc) for LMs
- More general convex programming: Matlab CVX package.
Computational Algorithms for Lasso

- Quadratic programming: the original; slow.
- LARS (§3.8): the solution path is piece-wise linear; at a cost of fitting a single LM; not general?
- Incremental Forward Stagewise Regression (§3.8): approx; related to boosting.
- A simple (and general) way: $|\beta_j| = \frac{\beta_j^2}{|\hat{\beta}_j^{(r)}|}$; truncate a current estimate $|\hat{\beta}_j^{(r)}| \approx 0$ at a small $\epsilon$.
- Coordinate-descent algorithm (§3.8.6): update each $\beta_j$ while fixing others at the current estimates—recall we have a closed-form solution for a single $\beta_j$!
- Simple and general but not applicable to grouping penalties.
- ADMM (Boyd et al 2011).
  http://stanford.edu/~boyd/admm.html
Sure Independence Screening (SIS)

- Q: penalized (or stepwise ...) regression can do automatic VS; just do it?
- Key: there is a cost/limit in performance/speed/theory.
- Q2: some methods (e.g. LDA/QDA/RDA) do not have VS, then what?
- Going back to basics: first conduct marginal VS,
  1) $Y \sim X_1$, $Y \sim X_2$, ..., $Y \sim X_p$;
  2) choose a few top ones, say $p_1$;
  $p_1$ can be chosen somewhat arbitrarily, or treated as a tuning parameter
  3) then apply penalized reg (or other VS) to the selected $p_1$ variables.
- Called SIS with theory (Fan & Lv, 2008, JRSS-B).
  R package SIS;
  iterative SIS (ISIS); why? a limitation of SIS ...
Using Derived Input Directions

- PCR: PCA on $X$, then use the first few PCs as predictors. Use a few top PCs explaining a majority (e.g. 85% or 95%) of total variance;
  ∆ number of components: a tuning parameter; use (genuine) CV;
  Used in genetic association studies, even for $p < n$ to improve power.
  +: simple;
  -: PCs may not be related to $Y$. 
Partial least squares (PLS): multiple versions; see Alg 3.3.
Main idea:
1) regress $Y$ on each $X_j$ univariately to obtain coef est $\phi_{1j}$;
2) first component is $Z_1 = \sum_j \phi_{1j}X_j$;
3) regress $X_j$ on $Z_1$ and use the residuals as new $X_j$;
4) repeat the above process to obtain $Z_2$, ...;
5) Regress $Y$ on $Z_1$, $Z_2$, ...

Choice of # components: tuning data or CV (or AIC/BIC?)

Contrast PCR and PLS:
PCA: $\max_{\alpha} \text{Var}(X_\alpha)$ s.t. ....;
PLS: $\max_{\alpha} \text{Cov}(Y, X_\alpha)$ s.t. ...;
Continuum regression (Stone & Brooks 1990, JRSS-B)

Penalized PCA (...) and Penalized PLS (Huang et al 2004, BI; Chun & Keles 2012, JRSS-B; R packages ppls, spls).

Example code: ex3.2.r
FIGURE 3.7. Estimated prediction error curves and their standard errors for the various selection and shrinkage methods. Each curve is plotted as a function of the corresponding complexity parameter for that.