1

Lasso i lineære modeller

Betragt n observationer $\{x_i, y_i\}_{i=1}^n$, hvor $x_i = (x_{i1}, \ldots, x_{ip})$ er en p dimension vektor af forklarende variable eller prediktorer og $y_i \in \mathbb{R}$ er den tilhørende respons variabel.

1.1 Mindste kvadraters metode

Den velkendte estimator for mindste kvadraters metode for (β_0, β) findes udfra

$$\min_{\beta_0,\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\}$$

Løsningen hertil er givet ved

$$\hat{\beta}^{\text{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

Som bekendt er estimatoren unbiased, men ofte har den en høj varians. GRUNDE TIL AT PRØVE ANDET END OLS

1.2 Ridge regression

Ridge regression estimatoren findes udfra

$$\min_{\beta_0,\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\}, \quad \text{underlagt at } \sum_{j=1}^{p} \beta_j^2 \leq t, \tag{1.1}$$

som kan omskrives til et Lagrange problem

$$\min_{\beta_0,\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \tag{1.2}$$

hvor $\lambda \geq 0$ er en såkaldt strafparameter, som bestemmes separat. Der er en en-til-en korrespondance mellem det betingede problem (1.1) og Lagrange problemet (1.2). Første led i (1.2) svarer til OLS, som finder de estimerede koefficienter ved at minimere SSR, mens sidste led mindsker de estimerede koefficienter . På matrix-vektor form er løsningen af ridge regression givet ved

$$\hat{\beta}^R = (\mathbf{X}^T\mathbf{X} + \lambda I_p)^{-1}\mathbf{X}^t\mathbf{y}$$

1.3 Lasso

Lasso finder løsningen til optimerings problemet

$$\min_{\beta_0,\beta}\left\{\frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2\right\},\quad ubb\sum_{j=1}^{p}|\beta_j|\le t \tag{1.3}$$

Betingelsen $\sum_{j=1}^{p}|\beta_j|\le t$ kan skrives mere kompakt ved $\|\beta\|_1\le t$. Dette kan udtrykkes på matrix-vektor notation. Lad $\mathbf{y}=(y_1,\ldots,y_n)$ være en n dimensional vektor med responsvariable og $\mathbf{X}$ være en $n\times p$ matrix med $x_i\in\mathbb{R}^p$ som den i'te række, da kan (1.3) omskrives til

$$\min_{\beta_0,\beta}\left\{\frac{1}{2n}\|\mathbf{y}-\beta_0\mathbf{1}-\mathbf{X}\beta\|_2^2\right\},\quad s.t.\|\beta\|_1\le t,$$

hvor $\mathbf{1}$ er en n dimensionel vektor bestående af 1 og $\|\cdot\|_2$ betegner den Euklidiske norm af vektorer.

Grænsen t begrænser summen af de absolutte værdier af parameter estimaterne. Denne skal specificeres ved en ekstern procedure kaldet *kryds validering*, som vil blive diskuteret i kap –.

Ofte standardiseres prediktorerne $\mathbf{X}$ således at kolonnerne er centeret og har varians 1. Dvs $\frac{1}{n}\sum_{i=1}^{n}x_{ij}=0$ og $\frac{1}{n}\sum_{i=1}^{n}x_{ij}^2=1$. Hvis ikke prediktorerne standardiseres da vil lasso estimaterne afhænge af enhederne. Hvis prediktorerne er målt i samme enhed, da vil vi typisk ikke standardisere. For fuldstændigheden, antager vi også at responsvariablen y_i er centeret, dvs $\frac{1}{n}\sum_{i=1}^{n}y_i=0$. Når data er centreret da kan vi se bort fra skæringen β_0 i lasso optimeringen. Given en optimal lasso løsning $\hat{\beta}$ på det centreret data, kan vi finde løsningen for det ikke-centreret data. Der gælder at

$$\hat{\beta}^{\text{ikke-centreret}}=\hat{\beta}^{\text{centreret}}$$
$$\hat{\beta}_0^{\text{ikke-centreret}}=\bar{y}-\sum_{j=1}^{p}\bar{x}_j\hat{\beta}_j$$

Derfor ser vi bort fra skæringen resten af kapitlet.

Vi kan omskrive lasso problemet til Lagrange form

$$\min_{\beta}\left\{\frac{1}{2n}\|\mathbf{y}-\mathbf{X}\beta\|_2^2+\lambda\|\beta\|_1\right\} \tag{1.4}$$

for $\lambda\ge 0$. Af Lagrange dualiteten er der en bijektion mellem (1.3) og (1.4): for hver værdi af t hvor $\|\beta\|_1\le t$ er opfyldt, da findes en tilhørende værdi af λ som giver den samme løsning for (1.4). Mens løsningen $\hat{\beta}_\lambda$ til (1.4) løser grænse problemet med $t=\|\hat{\beta}_\lambda\|_1$

Variabel udvægelsen for ridge regression og lasso illustreres på figur 1.1.

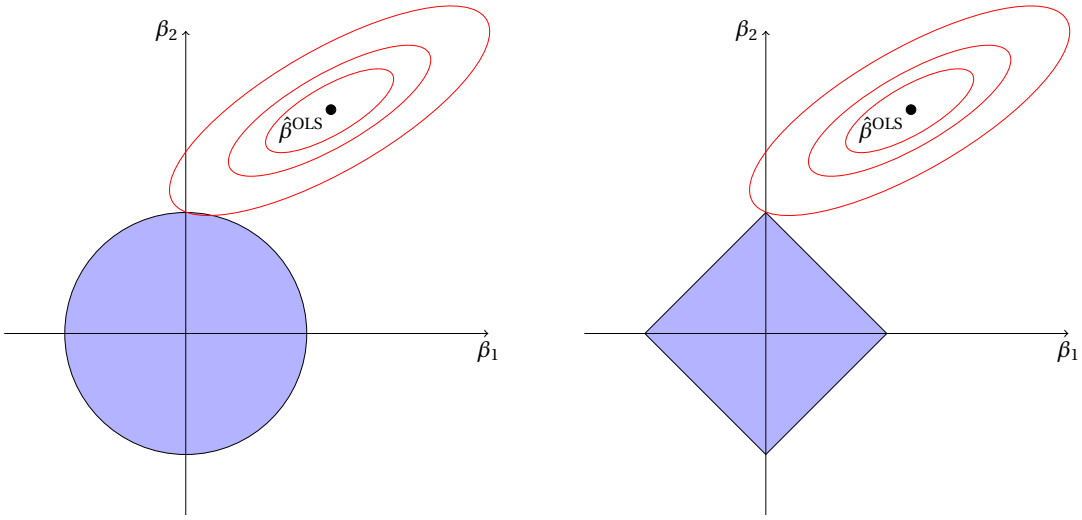

Figure 1.1: *Konturer for SSR og betingelsesområderne for ridge regression (venstre) og lasso (højre). De blå arealer er betingelsesområderne* $|\beta_1| + |\beta_2| \leq t$ *og* $\beta_1^2 + \beta_2^2 \leq t^2$*, mens de røde ellipser er konturkurver for SSR. Konturkurverne har centrum i OLS estimatoren,* $\hat{\beta}^{OLS}$*.*

For $p = 2$ underligges OLS betingelsen $\beta_1^2 + \beta_2^2 \leq t^2$ for ridge regression og betingelsen $|\beta_1| + |\beta_2| \leq t$ for lasso. Ellipserne omkring $\hat{\beta}^{\text{OLS}}$ er konturkurverne for SSR, dvs. SSR er konstant i en given ellipse. Værdien af SSR stiger, som ellipsen udvides fra $\hat{\beta}^{\text{OLS}}$. Ligningerne - og - indikerer at løsningen for ridge regression og lasso er givet ved det første punkt, hvor konturkurverne rammer betingelsesområdet. Siden ridge regression har et cirkulært betingelsesområde, vil skæringen med konturkurverne generelt ikke forekomme direkte på en akse. Modsat ridge regression har lassos betingelses område hjørner i hver akse, hvilket betyder, at hvis løsningen forekommer i et hjørne, da vil en af parametrene β_j være lig 0.

Hvis t er tilstrækkelig stor, da vil betingelsesområderne indeholde $\hat{\beta}^{\text{OLS}}$ og derfor vil ridge regression og lasso estimatorerne være lig OLS estimatoren.

På figur 1.1 har vi blot betragtet det simple tilfælde hvor $p = 2$. Når $p = 3$ vil betingelsesområdet for ridge regression være en kugle, mens betingelsesområdet for lasso vil være en polydron.

Da lasso penalty ikke er differentialbel, findes der ikke en eksplicit løsning til lasso problemet. Vi antager at responsvariablerne y_i og prediktorerne x_{ij} er standardiseret således at $\frac{1}{n}\sum_{i=1}^n y_i = 0$, $\frac{1}{n}\sum_{i=1}^n x_{ij} = 0$ og $\frac{1}{n}\sum_{i=1}^n x_{ij}^2 = 1$. Da kan vi ser bort fra skæringen β_0. Lagrange formen er nyttig for numerisk udregning af løsningen som findes vha en simpel procedure kaldet *coordinate descent*.

Vi kan opskrive objektfunktionen i (1.4) som

$$\frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \sum_{k\neq j} x_{ik}\beta_k - x_{ij}\beta_j\right)^2 + \lambda\sum_{k\neq j}|\beta_k| + \lambda|\beta_j|$$

Vi kan se at løsningen for hver β_j kan udtrykkes ved den partial residual $r_i^{(j)} = y_i - \sum_{k\neq j} x_{ik}\hat{\beta}_k$,

som fjerner ... Da er den j'te koefficient opdateret ved

$$\hat{\beta}_j = S_\lambda\left(\frac{1}{n}\left\langle \mathbf{x}_j, \mathbf{r}^{(j)}\right\rangle\right), \tag{1.5}$$

hvor $r_i = y_i - \sum_{j=1}^{p} x_{ij}\hat{\beta}_j$ er de fulde residualer. Den beskrevne algoritme svarer til metoden *cyclical coordinate descent*, som minimerer en konveks objektfunktion langs hver koordinat af gangen. Under milde regularitets betingelser, konvergerer løsningen til et global optimum. Fra opdateringen (1.5) ser vi at algoritmen foretager en univariat regression af den partial residual på hver prediktor, cycling gennem prediktorerne indtil konvergens. *pathwise coordinate descent*

Coordinate descent er særlig hurtig til at løse lasso problemet da ...

Homotopy metoder er en alternativ teknisk til at løse lasso problemet. Disse producerer en helt sti af løsninger i en frekventiel sekvens, ved at starte med nul. Denne sti er faktisk piecewise lineær. Algoritmen kaldet *least angle regression* (LARS) er en homotopy metode som effektivt konstruerer piecewise lineære stier. En mere teoretisk gennemgang af coordinate descent og LARS algoritmen er givet i kapitel –.

1.4 Grouped lasso

For mange regressions problemer har kovariaterne en naturlig grupperet struktur, og da foretrækkes det at alle koefficienter indenfor en gruppe er ikke-nul (eller nul) samtidig. Betragt en lineær regressions model som har J grupper af kovariater, hvor vektoren $Z_j \in \mathbb{R}^{p_j}$ for $j = 1,\ldots,J$ repræsenterer kovariaterne i gruppe j. Formålet er da at prædiktere responsvariablen $Y \in \mathbb{R}$ baseret på en samling af kovariater $(Z_1,\ldots,Z_J)$. En lineær model for regressions funktionen $\mathbb{E}\left[Y|Z\right]$ er givet ved $\theta_0 + \sum_{j=1}^{J} Z_j^T\theta_j$, hvor $\theta_j \in \mathbb{R}^{p_j}$ repræsenterer en gruppe af p_j regressions koefficienter.

Given en samling af n samples $\{(y_i, z_{i,1}, z_{i,2},\ldots,z_{i,J})\}_{i=1}^{n}$ løser group lasso følgende konveks problem

$$\min_{\theta_0\in\mathbb{R},\ \theta_j\in\mathbb{R}^{p_j}} \left\{ \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \theta_0 - \sum_{j=1}^{J} z_{ij}^T\theta_j\right)^2 + \lambda\sum_{j=1}^{J}\|\theta_j\|_2 \right\}, \tag{1.6}$$

hvor $\|\theta_j\|_2$ er den euklidiske norm af vektoren θ_j. Dette er en grupperet generalisering af lasso, som har følgende egenskaber:

- Afhængig af λ, vil enten alle indgange i vektoren $\hat{\theta}_j$ være nul eller ikke-nul
- Når $p_j = 1$, da har vi at $\|\theta_j\|_2 = |\theta_j|$, således at alle grupper er singletons, dermed reduceres optimerings problemet (1.6) til lasso problemet.

På figur – sammenlignes betingelsesområdet for den grupperet lasso med lasso for tre variable. Vi ser at den grupperet lasso deler egenskaber med både ℓ_1 og ℓ_2 kuglen.

I (1.6), straffes alle grupper ligeligt, hvilket betyder at større grupper vil have en tendens til at blive valgt.

1.5 Elastic net

Lasso er ikke godt til at håndtere højt korreleret variabler. Koefficient stierne har en tendens til at være uregelmæssige.

Det elastiske net er et kompromis mellem strafleddet af ridge og lasso, og løser det konvekse problem

$$\min_{\beta_0,\beta}\left\{\frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\left[\frac{1}{2}(1-\alpha)\|\beta\|_2^2+\alpha\|\beta\|_1\right]\right\}, \tag{1.7}$$

hvor $\alpha\in[0,1]$ er en parameter som kan varieres.

Hvis $\alpha=1$, da reduceres strafleddet til ℓ_1-normen eller strafleddet for lasso og hvis $\alpha=0$ reduceres det til den kvadrerede ℓ_2-norm, svarende til strafleddet for ridge regression.

Coordinate descent opdateringen for j'te koefficient er givet ved

$$\hat{\beta}_j=\frac{S_{\lambda\alpha}\left(\sum_{i=1}^{n}r_{ij}x_{ij}\right)}{\sum_{i=1}^{n}x_{ij}^2+\lambda(1-\alpha)},$$

hvor $S_\mu(z)=\text{sign}(z)(z-\mu)_+$ er soft-thresholding operatoren og $r_{ij}=y_i-\hat{\beta}_0-\sum_{k\neq j}x_{ik}\hat{\beta}_k$ er den partial residual.