3 Regularisation slide 180

3.1 Basic Notions slide 181

Tall and wide regressions

- So far we have supposed that we have a **tall regression**:
  - the number of units n exceeds the number of variables p,
  - the design matrix X has rank p.
- In many 'modern' settings we instead have a **wide regression**:
  - n and p are comparable, $p > n$, maybe even $p \gg n$;
  - in genomics, for example (typically) $n = O(10^2, 10^3)$, $p = O(10^5, 10^6)$;
  - hence $\text{rank}(X) = \min(n, p) = n$.
- Even tall X may be 'almost singular', making β 'almost inestimable'.
- Solutions:
  - subset selection (drop certain columns of X);
  - seek different good explanations of response variation, not single model;
  - regularisation (often with prediction in mind).
- Certain regularisation methods (e.g., lasso) also perform subset selection.

Different good explanations

- With $p > n$, perhaps $p \gg n$, X is rank-deficient and many β may give $X\beta = y$.
- To find important variables we include intrinsic variables (gender, ...) in all models, and then
  - choose some k (preferably ≤ 15) such that $k < n$ and suppose that $p < k^a$ (let $a = 3$ for easy visualisation);
  - assign each variable to a cell of a hyper-cube with coordinates $\{1, \ldots, k\}^a$;
  - fit a linear model containing each set of k variables corresponding to the ak^{a-1} rows, columns, ... of the cube, so each variable appears in a distinct models;
  - for each such model, retain the two variables that are most significant.
- Iterate the above procedure, retaining only the most significant variables at each stage, aiming for a final set of 10–20 variables, for which a careful analysis is performed, perhaps leading to several different good explanations of the response variation.
- Some cells of the hyper-cube may be empty, and important variables might be assigned to several cells.
- The above design is a form of **balanced incomplete block design (BIBD)** (with k^a treatments and ak^{a-1} blocks).
- See Cox and Battey (2017, PNAS)

Collinearity

- Columns of X **collinear** if there exists a non-zero $v_{p\times 1}$ such that $Xv = 0$, i.e., rank$(X) < p$, so there is no unique $\widehat{\beta}$ minimising $\|y - X\beta\|^2$.
- Software deals with this by dropping columns of X, but it may be better to write $X\beta = XC\gamma$, where XC is full rank and γ has a clear interpretation.
- If X is nearly collinear, its SVD $U_{n\times n}D_{n\times p}V^{\mathrm{T}}_{p\times p}$, with $d_1 \geq \cdots \geq d_p \geq 0$, gives

$$\widehat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y = VD_{-}^{\mathrm{T}}U^{\mathrm{T}}y = \sum_{r=1}^{p}(u_r^{\mathrm{T}}y/d_r)v_r,$$

  so $\widehat{\beta}$ is a linear combination of the vectors v_r with coefficients $u_r^{\mathrm{T}}y/d_r$. As $\mathrm{var}(U^{\mathrm{T}}y) = \sigma^2 I_n$,

$$\mathrm{var}(\widehat{\beta}) = \sigma^2 VD_{-}^{\mathrm{T}}D_{-}V^{\mathrm{T}} = \sigma^2\sum_{r=1}^{p} d_r^{-2}v_r v_r^{\mathrm{T}},$$

  i.e., $\widehat{\beta}$ is unstable in the directions corresponding to the v_r with small singular values d_r.
- In numerical analysis, collinearity often measured using **condition number** $(d_1/d_p)^{1/2}$, but its statistical meaning is unclear.

Regularisation

- Stop $\widehat{\beta}$ from fluctuating too wildly in directions with small eigenvalues d_r, by adding a non-negative penalty $p_\lambda(\beta)$ and choosing β to minimise the **penalised sum of squares**

$$\|y - X\beta\|^2 + p_\lambda(\beta). \tag{16}$$

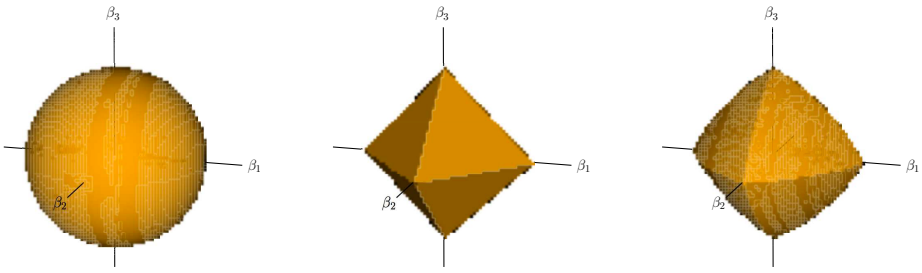
- The strength of the penalty depends on a positive parameter λ that constrains β more as λ increases.
- Often $p_\lambda(\beta) = \lambda p(\beta)$, where, for example,
  - $p(\beta) = \|\beta\|_2^2 = \sum_{r=1}^{p}\beta_r^2$ gives **ridge regression** (aka Tikhonov regularisation);
  - $p(\beta) = \|\beta\|_1 = \sum_{r=1}^{p}|\beta_r|$ gives the **lasso** (aka L_1 regularisation);
  - $p(\beta) = (1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1$ for $0 \leq \alpha \leq 1$ gives the **elastic net**;
  - $p(\beta) = \sum_{g=1}^{G} p_g^{1/2}\|\beta_g\|_2$, with β_g being $p_g \times 1$ sub-vectors of β, gives the **grouped lasso**, which penalises factors with parameters β_g.
- It is useful to see regularisation through the lens of Bayesian inference, with the regularising term equivalent to the prior density.

Bound form

- Equivalently we can take the **bound form** of the minimisation problem, i.e.,

$$\text{minimise}_\beta \quad \|y - X\beta\|_2^2 \quad \text{subject to} \quad p(\beta) \leq t,$$

  for some $t \geq 0$, where setting $t = \infty$ just gives the least squares estimates.
- Below: constraint balls for ridge (left), lasso (centre) and elastic-net (right) regularisation. The sharp corners of the last two allow for variable selection as well as shrinkage.



Bayesian setting

- Treat all unknowns as random variables, and compute conditional distribution of unobserved unknowns conditional on observed unknowns.
- Requires prior density on β, and if σ^2 is known, then a simple combination of **data model** and **prior model** is

$$y \mid \beta, \sigma^2 \sim \mathcal{N}(X\beta, \sigma^2 I_n), \quad \beta \mid \sigma^2 \sim \mathcal{N}_p(\beta_*, \sigma^2 V_*), \tag{17}$$

  where the prior model is determined by β_* and V_*.
- Full specification would require prior on σ^2, but we don't need this.
- Let $\equiv$ mean we have dropped additive constants not involving the argument of a density.
- The log multivariate normal density is

$$\begin{aligned} \log f(x \mid \mu, \Omega) &= -\frac{m}{2}\log 2\pi - \frac{1}{2}\log|\Omega| - \frac{1}{2}(x-\mu)^{\mathrm{T}}\Omega^{-1}(x-\mu) \\ &\equiv x^{\mathrm{T}}\Omega^{-1}\mu - \frac{1}{2}x^{\mathrm{T}}\Omega^{-1}x \\ &\equiv Q(x) = x^{\mathrm{T}}a - \frac{1}{2}x^{\mathrm{T}}Bx, \end{aligned}$$

  say, and as $\exp Q(x)$ is proportional to a unique probability density function,

$$\mathrm{E}(X) = \mu = B^{-1}a, \quad \mathrm{var}(X) = \Omega = B^{-1}, \quad \text{where } B \text{ is the } \textbf{precision matrix}.$$

Bayesian linear model I

- The model (17) gives

$$\begin{aligned}\log f(\beta \mid y, \sigma^2) &= \log\left\{\frac{f(y \mid \beta, \sigma^2) f(\beta \mid \sigma^2)}{f(y \mid \sigma^2)}\right\} \\ &\equiv \log f(y \mid \beta, \sigma^2) + \log f(\beta \mid \sigma^2) \\ &\equiv -\frac{(y - X\beta)^{\mathrm{T}}(y - X\beta)}{2\sigma^2} - \frac{(\beta - \beta_*)^{\mathrm{T}} V_*^{-1} (\beta - \beta_*)}{2\sigma^2} \\ &\propto \|y - X\beta\|_2^2 + (\beta - \beta_*)^{\mathrm{T}} V_*^{-1} (\beta - \beta_*).\end{aligned}$$

- Comparison with (16) shows that $p_\lambda(\beta)$ represents prior beliefs about the likely values of β: before seeing the data, the most plausible value is β_*, with precision V_*^{-1}.
- Dropping more constants,

$$\begin{aligned}\log f(\beta \mid y, \sigma^2) &\equiv \frac{1}{\sigma^2}\left\{\beta^{\mathrm{T}} X^{\mathrm{T}} y - \beta^{\mathrm{T}}(X^{\mathrm{T}} X)\beta/2 + \beta^{\mathrm{T}} V_*^{-1}\beta_* - \beta^{\mathrm{T}} V_*^{-1}\beta/2\right\} \\ &= \frac{1}{2\sigma^2}\left\{2\beta^{\mathrm{T}}(X^{\mathrm{T}} y + V_*^{-1}\beta_*) - \beta^{\mathrm{T}}(X^{\mathrm{T}} X + V_*^{-1})\beta\right\}, \qquad (18)\end{aligned}$$

  which is $Q(x)$ with x, a and B replaced by β, $(X^{\mathrm{T}} y + V_*^{-1}\beta_*)/\sigma^2$ and $(X^{\mathrm{T}} X + V_*^{-1})/\sigma^2$.
- Hence $f(\beta \mid y, \sigma^2)$ is multivariate normal with mean vector and variance matrix

$$\mathrm{E}(\beta \mid y, \sigma^2) = (X^{\mathrm{T}} X + V_*^{-1})^{-1}(X^{\mathrm{T}} y + V_*^{-1}\beta_*), \quad \mathrm{var}(\beta \mid y, \sigma^2) = \sigma^2 (X^{\mathrm{T}} X + V_*^{-1})^{-1}.$$

Bayesian linear model II

- The **maximum a posteriori (MAP) estimator** of β is $\mathrm{E}(\beta \mid y, \sigma^2)$, and the MAP estimator of $A_{q\times p}\beta$ is $A\mathrm{E}(\beta \mid y, \sigma^2)$, which has a posterior normal density.
- When $X^{\mathrm{T}} X$ is invertible,

$$\tilde{\beta} = \mathrm{E}(\beta \mid y, \sigma^2) = (X^{\mathrm{T}} X + V_*^{-1})^{-1}(X^{\mathrm{T}} X\widehat{\beta} + V_*^{-1}\beta_*)$$

  is an average of $\widehat{\beta}$ and β_*, weighted by $X^{\mathrm{T}} X$ and V_*^{-1}.
- The posterior precision matrix

$$\mathrm{var}(\beta \mid y, \sigma^2)^{-1} = X^{\mathrm{T}} X/\sigma^2 + V_*^{-1}/\sigma^2$$

  adds the Fisher information and the prior precision matrix, V_*^{-1}/σ^2.
- High precision corresponds to small variance, and conversely:
  - letting $V_*^{-1} \to 0$ yields an improper prior density; and
  - for large V_*^{-1} the posterior precision is essentially determined by the prior precision.

  Thus the prior density regularises $\widehat{\beta}$ by including β_* and V_*.

Improper prior density

- We only need V_* to add information in directions corresponding to small singular values of X, so we might use an **improper prior** in which V_* is singular:

$$f(\beta \mid \sigma^2) = \frac{1}{(2\pi)^{p/2}|V_*|_+^{1/2}} \exp\left\{-(\beta-\beta_*)^{\mathrm{T}} V_*^{-}(\beta-\beta_*)/(2\sigma^2)\right\}, \tag{19}$$

  where V_* has spectral decomposition ED_*E^{T},
  - $|V_*|_+$ denotes the product of the non-zero elements of D_*, and
  - $V_*^{-} = \sum_{r:d_{*r}>0} e_r e_r^{\mathrm{T}}/d_{*r}$ is a generalized inverse of V_*.
- Below we write V_*^{-} even when V_* is invertible.
- (19) is improper because it is not integrable in the directions of the columns of E for which the corresponding d_r^* equal zero, but we need only that the posterior density of β be proper, i.e., that the posterior precision matrix

$$\mathrm{var}(\beta \mid y, \sigma^2)^{-1} = X^{\mathrm{T}}X/\sigma^2 + V_*^{-1}/\sigma^2$$

  is invertible.

Empirical Bayes

- Use the data to estimate the prior: construct estimators using Bayesian arguments, but assess their properties using classical criteria (bias, MSE, ...)
- The estimator $\tilde{\beta} = \mathrm{E}(\beta \mid y, \sigma^2)$ has mean and variance

$$\begin{aligned} \mathrm{E}(\tilde{\beta} \mid \beta) &= (X^{\mathrm{T}}X + V_*^{-})^{-1}(X^{\mathrm{T}}X\beta + V_*^{-}\beta_*) \\ &= \beta + (X^{\mathrm{T}}X + V_*^{-})^{-1}V_*^{-}(\beta_* - \beta), \\ \mathrm{var}(\tilde{\beta} \mid \beta) &= \sigma^2(X^{\mathrm{T}}X + V_*^{-})^{-1}X^{\mathrm{T}}X(X^{\mathrm{T}}X + V_*^{-})^{-1}. \end{aligned} \tag{20}$$

- Hence $\tilde{\beta}$
  - is biased unless $\beta_* = \beta$,
  - has smaller variance than $\widehat{\beta}$,

  so maybe there is a bias-variance tradeoff when estimating $A\beta$.
- If we write $\mu = \mathrm{E}(\tilde{\beta} \mid \beta)$, then the MSE is

$$\begin{aligned} \mathrm{E}\left(\|A\tilde{\beta} - A\beta\|^2 \mid \beta\right) &= \mathrm{E}\{(\tilde{\beta}-\beta)^{\mathrm{T}}A^{\mathrm{T}}A(\tilde{\beta}-\beta) \mid \beta\} \\ &= \mathrm{E}\left[\mathrm{tr}\left\{A(\tilde{\beta}-\beta)(\tilde{\beta}-\beta)^{\mathrm{T}}A^{\mathrm{T}}\right\} \mid \beta\right] \\ &= \mathrm{tr}\left[\mathrm{E}\left\{A(\tilde{\beta}-\mu+\mu-\beta)(\tilde{\beta}-\mu+\mu-\beta)^{\mathrm{T}}A^{\mathrm{T}} \mid \beta\right\}\right]. \end{aligned}$$

Empirical Bayes II

- The expectation above is

$$A\left\{\mathrm{var}(\tilde{\beta} \mid \beta) + (X^{\mathrm{T}}X + V_*^-)^{-1}V_*^-(\beta - \beta_*)(\beta - \beta_*)^{\mathrm{T}}V_*^-(X^{\mathrm{T}}X + V_*^-)^{-1}\right\}A^{\mathrm{T}},$$

  giving the MSE when estimating a fixed β.
- Taking expectations over the prior model for β gives

$$\mathrm{E}\left(\|A\tilde{\beta} - A\beta\|^2\right) = \sigma^2 \mathrm{tr}\left\{A(X^{\mathrm{T}}X + V_*^-)^{-1}A^{\mathrm{T}}\right\}, \tag{21}$$

  which is larger than $A\mathrm{var}(\tilde{\beta} \mid \beta)A^{\mathrm{T}}$ and does not depend on β_*.
- This computation uses only the mean and variance, so holds under second-order assumptions, but under normal-theory assumptions gives the mean and variance of $\tilde{\beta}$.
- From now on we set $\beta_* = 0$, unless we state otherwise.

Equivalent degrees of freedom

- If we set $\beta_* = 0$, then the fitted values are

$$\tilde{y} = X\tilde{\beta} = X(X^{\mathrm{T}}X + V_*^-)^{-1}X^{\mathrm{T}}y = H_*y,$$

  say.
- We define the **equivalent degrees of freedom** of the fit as

$$\mathrm{edf} = \mathrm{tr}(H_*) = \mathrm{tr}\{X(X^{\mathrm{T}}X + V_*^-)^{-1}X^{\mathrm{T}}\} = p - \mathrm{tr}\{(X^{\mathrm{T}}X + V_*^-)^{-1}V_*^-\},$$

- This is lower than p unless $V_*^- = 0$, so regularisation reduces the degrees of freedom by an amount that depends on V_*.
- The penalised estimate is a linear function of the unpenalised one (if it exists), as we can write

$$\tilde{\beta} = (X^{\mathrm{T}}X + V_*^-)^{-1}X^{\mathrm{T}}X\widehat{\beta} = P_*\widehat{\beta},$$

  say. As

$$\mathrm{edf} = \mathrm{tr}(H_*) = \mathrm{tr}(P_*),$$

  this gives an alternative formula useful in complex models.

How much penalisation?

- ☐ Often V_*^- depends on some $\lambda > 0$ that must be chosen, as well as σ^2, which is usually estimated by a (penalised) residual sum of squares.
- ☐ To estimate λ, we compare y_j with its predicted value $\widehat{y}_{\lambda,j}^- = x_j^{\mathrm{T}}\widehat{\beta}_{\lambda,-j}$, where $\widehat{\beta}_{\lambda,-j}$ is

$$\widehat{\beta}_\lambda = (X^{\mathrm{T}}X + V_*^-)^{-1}X^{\mathrm{T}}y$$

  computed with the jth rows x_j and y_j of X and y omitted.
- ☐ Using Lemma 14, the **leave-one-out cross-validation** sum of squares is then

$$\mathrm{CV}_\lambda = \sum_{j=1}^n (y_j - \widehat{y}_{\lambda,j}^-)^2 = \|y - \widehat{y}_\lambda^-\|^2 = \sum_{j=1}^n \frac{(y_j - \widehat{y}_{\lambda,j})^2}{(1-h_{\lambda,jj})^2},$$

  where $\widehat{y}_{\lambda,j}$ is the jth element of the complete-data fitted value $H_\lambda y$ and $h_{\lambda,jj}$ is the jth diagonal element of $H_\lambda = X(X^{\mathrm{T}}X + V_*^-)^{-1}X^{\mathrm{T}}$ for the overall fit.
- ☐ More often we use the **generalized cross-validation** criterion

$$\mathrm{GCV}_\lambda = \sum_{j=1}^n \frac{(y_j - \widehat{y}_{\lambda,j})^2}{\{1 - \mathrm{tr}(H_\lambda)/n\}^2}.$$

- ☐ Whichever criterion is used, it is typically minimised numerically over a grid of values of λ.

REML

- ☐ Cross-validation makes only second-order assumptions.
- ☐ Under normality, the marginal density of y is $\mathcal{N}\{X\beta_*, \sigma^2(I_n + XV_*X^{\mathrm{T}})\}$, so we could estimate β_*, σ^2 and λ by maximising the corresponding likelihood.
- ☐ If n and p are large, this results in biased estimates of λ and σ^2, so we prefer to eliminate β_*, resulting in a **log restricted likelihood** whose form is given below, with $W_\lambda^{-1} = I_n + XV_*X^{\mathrm{T}}$.

Lemma 31 *In a model in which $y \sim \mathcal{N}(X\beta, \sigma^2 W_\lambda^{-1})$, where W_λ depends on a parameter λ, a log restricted likelihood for σ^2 and λ is*

$$\ell_{\mathrm{REML}}(\sigma^2, \lambda) \equiv \frac{1}{2}\log(|W_\lambda|/|X^{\mathrm{T}}W_\lambda X|) - \frac{n-p}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y - \widehat{y}_\lambda)^{\mathrm{T}}W_\lambda(y - \widehat{y}_\lambda),$$

where $\widehat{\beta}_\lambda = (X^{\mathrm{T}}W_\lambda X)^{-1}X^{\mathrm{T}}W_\lambda y$ and $\widehat{y}_\lambda = X\widehat{\beta}_\lambda$. For fixed λ the restricted maximum likelihood estimator of σ^2 is therefore

$$\widehat{\sigma}_\lambda^2 = \frac{1}{n-p}(y - \widehat{y}_\lambda)^{\mathrm{T}}W_\lambda(y - \widehat{y}_\lambda),$$

and the resulting profile log restricted likelihood for λ is

$$\ell_{\mathrm{p}}(\lambda) \equiv \frac{1}{2}\log(|W_\lambda|/|X^{\mathrm{T}}W_\lambda X|) - \frac{(n-p)}{2}\log\widehat{\sigma}_\lambda^2.$$

Note on Lemma 31

- Suppose that $f(y;\alpha,\beta)$ depends on two parameters, that interest is focused on α, and that for fixed α there is a minimal sufficient statistic s_α for β. Then $f(y;\alpha,\beta) = f(y \mid s_\alpha;\alpha) f(s_\alpha;\alpha,\beta)$, and since the first density on the right is a proper conditional density not depending on β, we can use it for inference on α, in the form

$$\log f(y \mid s_\alpha;\alpha) = \log f(y;\alpha,\beta) - \log f(s_\alpha;\alpha,\beta).$$

  As the left-hand side of this expression does not depend on β, we may be able to simplify the right-hand side by an astute choice of β.

- In the normal model we take $\alpha = (\sigma^2,\lambda)$. If α is fixed, then $s_\alpha = \widehat{\beta}_\alpha = (X^{\mathrm{T}} W_\lambda X)^{-1} X^{\mathrm{T}} W_\lambda y$ is sufficient for β; its distribution is $\mathcal{N}_p\{\beta, \sigma^2 (X^{\mathrm{T}} W_\lambda X)^{-1}\}$. Hence

$$\ell_{\mathrm{REML}}(\sigma^2,\lambda) = \log f(y \mid \widehat{\beta}_\lambda;\sigma^2,\lambda) = \log f(y;\sigma^2,\lambda,\beta) - \log f(\widehat{\beta}_\lambda;\sigma^2,\lambda,\beta)$$

  which equals

$$\begin{aligned} -\frac{n}{2}\log\sigma^2 \quad &+ \quad \tfrac{1}{2}\log|W_\lambda| - \frac{1}{2\sigma^2}(y - X\beta)^{\mathrm{T}} W_\lambda (y - X\beta) \\ &+ \quad \frac{p}{2}\log\sigma^2 - \tfrac{1}{2}\log|X^{\mathrm{T}} W_\lambda X| + \frac{1}{2\sigma^2}(\widehat{\beta}_\lambda - \beta)^{\mathrm{T}} X^{\mathrm{T}} W_\lambda X (\widehat{\beta}_\lambda - \beta), \end{aligned}$$

  or equivalently, on setting $\beta = 0$ and $\widehat{y}_\lambda = X\widehat{\beta}_\lambda$,

$$\tfrac{1}{2}\log(|W_\lambda|/|X^{\mathrm{T}} W_\lambda X|) - \frac{(n-p)}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(y^{\mathrm{T}} W_\lambda y - \widehat{y}_\lambda^{\mathrm{T}} X^{\mathrm{T}} W_\lambda \widehat{y}_\lambda\right).$$

- The last term reduces to the given form because $\widehat{y}_\lambda^{\mathrm{T}} W_\lambda (y - \widehat{y}_\lambda) = 0$, so the term in brackets in the last displayed equation is the residual sum of squares $(y - \widehat{y}_\lambda)^{\mathrm{T}} W_\lambda (y - \widehat{y}_\lambda)$.

- The restricted maximum likelihood estimator $\widehat{\sigma}^2_\lambda$ and the profile log restricted likelihood for λ are obtained by maximising $\ell_{\mathrm{REML}}(\sigma^2,\lambda)$, for fixed λ and then dropping constant terms from $\ell_{\mathrm{REML}}(\widehat{\sigma}^2_\lambda,\lambda)$.

3.2 Simple Applications

Ridge regression

- Used for prediction when X is close to singular.
- If the first column of X is 1_n, we set $\beta_* = 0$ and $V_*^- = \lambda S = \lambda \mathrm{diag}(0, I_{p-1})$, giving

$$\widehat{\beta}_\lambda = (X^{\mathrm{T}} + \lambda S)^{-1} X^{\mathrm{T}} y, \quad \widehat{y}_\lambda = X\widehat{\beta}_\lambda = X(X^{\mathrm{T}} + \lambda S)^{-1} X^{\mathrm{T}} y = H_\lambda y,$$

  and effective degrees of freedom

$$\mathrm{edf}_\lambda = \mathrm{tr}(H_\lambda) = \mathrm{tr}\{(X^{\mathrm{T}}X + \lambda S)^{-1} X^{\mathrm{T}} X\} = \sum_{r=1}^{p} \frac{1}{1 + \lambda \delta_r},$$

  where $\delta_p \geq \cdots \geq \delta_2 > \delta_1 = 0$ are the eigenvalues of $(X^{\mathrm{T}}X)^{-1/2} S (X^{\mathrm{T}}X)^{-1/2}$.
- As λ increases from zero to infinity, edf_λ decreases from $p = \mathrm{rank}(X)$ to 1. The two are equivalent, but edf_λ is more easily interpreted, because it is not related to the scale of X.
- The inverse exists even if $X^{\mathrm{T}}X$ is singular, but if it is invertible then

$$\widehat{\beta}_\lambda = (X^{\mathrm{T}}X + \lambda S)^{-1}(X^{\mathrm{T}}X + \lambda S - \lambda S)(X^{\mathrm{T}}X)^{-1} X^{\mathrm{T}} y = \widehat{\beta} - \lambda (X^{\mathrm{T}}X + \lambda S)^{-1} S \widehat{\beta},$$

  so as $\lambda \to \infty$ all the elements of $\widehat{\beta}_\lambda$ tend to zero, other than the first. This corresponds to reducing the prior variance to zero, thereby giving the data themselves less and less influence on the elements of $\widehat{\beta}_\lambda$ other than the first, and thus stabilises the estimator.

Example: Cement data

```
> cement
   x1 x2 x3 x4     y
1   7 26  6 60  78.5
2   1 29 15 52  74.3
3  11 56  8 20 104.3
4  11 31  8 47  87.6
5   7 52  6 33  95.9
6  11 55  9 22 109.2
7   3 71 17  6 102.7
8   1 31 22 44  72.5
9   2 54 18 22  93.1
10 21 47  4 26 115.9
11  1 40 23 34  83.8
12 11 66  9 12 113.3
13 10 68  8 12 109.4
```

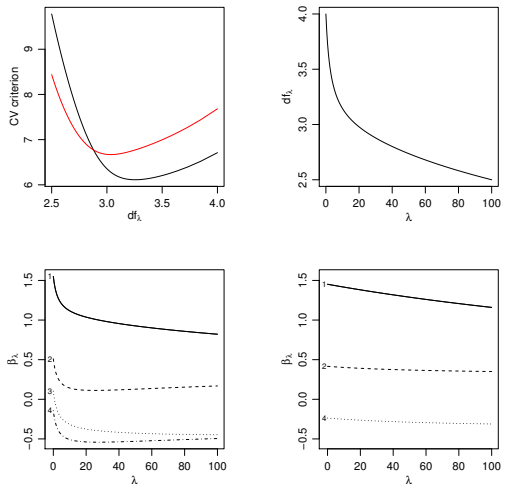
Example: Cement data

	Full model		Reduced model	
Parameter	Estimate	Standard error	Estimate	Standard error
β_0	62.41	70.07	71.64	14.14
β_1	1.55	0.74	1.45	0.12
β_2	0.51	0.72	0.42	0.19
β_3	0.10	0.75		
β_4	–0.14	0.71	–0.24	0.17

- ☐ The next slide shows results for ridge fits for these models.
- ☐ Looks like 3 df is optimal for prediction.
- ☐ Software often preprocesses X and y by either
  - centering both, by subtracting column means, or
  - centering y and centering and scaling X, so the column means are zero and the column variances are unity.
- ☐ The singular values for the centred X matrix are 78.8, 28.5, 12.2, 1.7, and those for the centred and scaled X matrix are 5.18, 4.35, 1.50, 0.14, so it matters which is used.
- ☐ The singular values for the (centred) reduced matrix are 78.8, 19.8 and 9.15.
- ☐ The shrinkage due to increasing λ occurs more slowly for the reduced model.

Example: Cement data/Ridge analysis

Top left: CV (black) and GCV (red) as functions of degrees of freedom df_λ. Top right: dependence of df_λ on λ. Bottom left: $\widehat{\beta}_\lambda$ as a function of λ, with all four covariates. Bottom right: $\widehat{\beta}_\lambda$ as a function of λ, with x_1, x_2, and x_4 only.



Comments

- ☐ The literature on ridge regression is very large and very dispersed, with many variants and many connections to ML techniques.
- ☐ Be careful with software: any pre-processing of X is not always described.