Regression Methods

Myrto Limnios

Autumn 2025 - Week 9

Regularisation - Basic Notions

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.

Collinearity

- Columns of X **collinear** if there exists a non-zero $v_{p\times 1}$ such that $Xv = 0$, i.e., $\text{rank}(X) < p$, 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 $\text{var}(U^{\mathrm{T}}y) = \sigma^2 I_n$,

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

  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**

$$\hat{\beta} \in \operatorname{argmin}_\beta \|y - X\beta\|^2 + p_\lambda(\beta). \tag{1}$$

- 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 $0 \le \alpha \le 1$ 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.

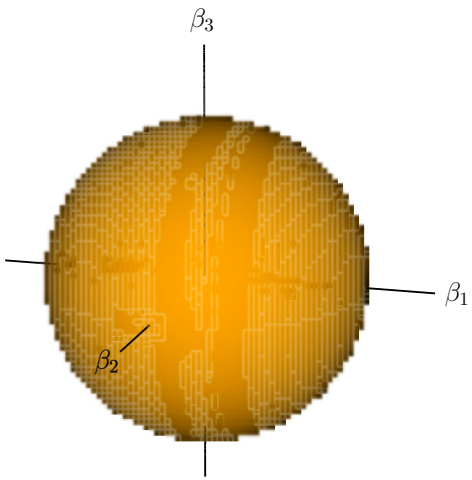
Bound form

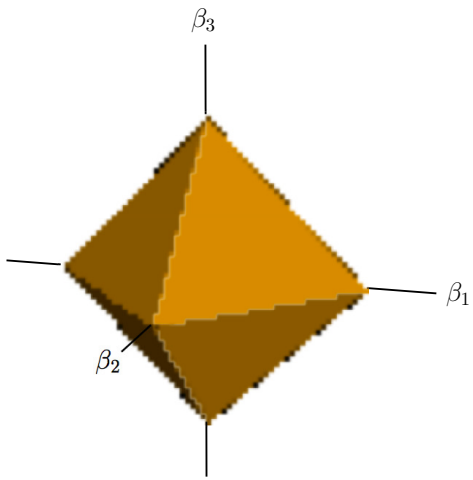
- Equivalently we can take the **bound form** of the minimisation problem, i.e.,

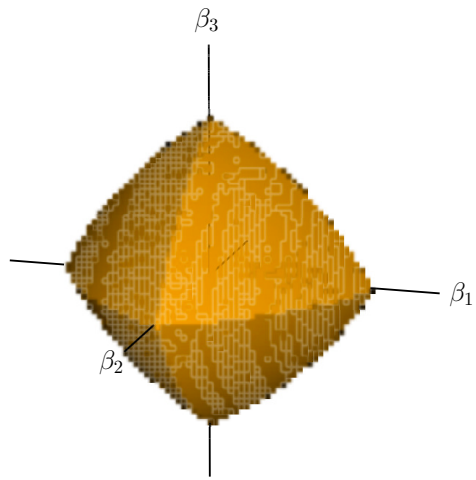
$$\text{minimise}_\beta \quad \|y - X\beta\|_2^2 \quad \text{subject to} \quad p(\beta) \leq t,$$

  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 approach

- 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

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

  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,

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

Bayesian linear model

- The model (2) 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 (1) 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,

$$\begin{aligned}
\log f(\beta \mid y, \sigma^2) &\equiv \frac{1}{\sigma^2}\left\{\beta^{\mathrm{T}} X^{\mathrm{T}} y - \beta^{\mathrm{T}}(X^{\mathrm{T}}X)\beta/2 + \beta^{\mathrm{T}} V_*^{-1}\beta_* - \beta^{\mathrm{T}} V_*^{-1}\beta/2\right\} \\
&= \frac{1}{2\sigma^2}\left\{2\beta^{\mathrm{T}}(X^{\mathrm{T}} y + V_*^{-1}\beta_*) - \beta^{\mathrm{T}}(X^{\mathrm{T}}X + V_*^{-1})\beta\right\}, \qquad (3)
\end{aligned}$$

  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

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

- 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:

$$f(\beta \mid \sigma^2) = \frac{1}{(2\pi)^{p/2}|V_*|_+^{1/2}} \exp\left\{-(\beta-\beta_*)^{\mathrm{T}} V_*^{-}(\beta-\beta_*)/(2\sigma^2)\right\}, \qquad (4)$$

  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.
- (4) 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

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

  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

$$\begin{aligned} \mathrm{E}(\tilde{\beta} \mid \beta) &= (X^{\mathrm{T}}X + V_*^-)^{-1}(X^{\mathrm{T}}X\beta + V_*^-\beta_*) \\ &= \beta + (X^{\mathrm{T}}X + V_*^-)^{-1}V_*^-(\beta_* - \beta), \\ \mathrm{var}(\tilde{\beta} \mid \beta) &= \sigma^2(X^{\mathrm{T}}X + V_*^-)^{-1}X^{\mathrm{T}}X(X^{\mathrm{T}}X + V_*^-)^{-1}. \end{aligned} \tag{5}$$

- 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

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

- 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

$$\mathrm{E}\left(\|A\tilde{\beta} - A\beta\|^2\right) = \sigma^2 \mathrm{tr}\left\{A(X^{\mathrm{T}}X + V_*^-)^{-1} A^{\mathrm{T}}\right\}, \tag{6}$$

  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.

- 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

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

- 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

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

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 9, 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 20

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

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

and the resulting profile log restricted likelihood for λ is

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

Note on Lemma 20

- 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 $\ell_{\mathrm{REML}}(\widehat{\sigma}^2_\lambda, \lambda)$.

Regularisation - Simple Applications

$\hat{\beta} \in \operatorname{argmin} \|y - X\beta\|^2 + p_\lambda(\beta)$ where $p_\lambda(\beta)$ is the penalty function depending on $\lambda > 0$ and $p_\lambda(\beta) = \lambda\|\beta\|_2$ ⇝ ridge regression

$= \lambda\|\beta\|_1$ ⇝ Lasso regression

$= \alpha\|\beta\|_2 + \gamma\|\beta\|_1$ ⇝ Elastic net.

Last session: Motivate the use of penalization with Bayesian calculus:

We showed that: if $\begin{cases} y|\beta,\sigma^2 \sim \mathcal{N}(X\beta, \sigma^2 I) \\ \beta|\sigma^2 \sim \mathcal{N}_p(\beta_a, \sigma^2 V_a) \end{cases}$

then the posterior distribution of $\beta|y,\sigma^2$ is also Gaussian

$$\tilde{y} = X\tilde{\beta} = \underbrace{X^T(X^TX + V_a)^{-1}}_{= H_a}\left(X^TX\hat{\beta} + V_a^{-1}\beta_a\right)$$

if $\beta_a = 0$ then $\tilde{y}$ is just a linear transformation of $\hat{y}$ (OLS: $\hat{y} = X\hat{\beta}$)

Equivalent degrees of freedom $\begin{cases} \tilde{y} = H_a\hat{y} \\ \text{edf} = \operatorname{tr}(H_a) = p - \operatorname{tr}\left((X^TX + V_a)^{-1}V_a\right) \end{cases}$

↳ Define CV, GCV

Penalization is used when # covariate parameters $\gg$ # observations

$\hat{\beta} = \underbrace{(X^TX)^{-1}}_{\text{ill defined}} X^Ty$

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

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

and effective degrees of freedom

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

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, $\mathbf{edf}_\lambda$ decreases from $p = \mathbf{rank}(X)$ to 1. The two are equivalent, but $\mathbf{edf}_\lambda$ 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.

In practice

Multicollinearity problem is that $\det\left[X^\top X\right] \approx 0$
[i.e. $X^\top X$ *almost not invertible*]

A Solution: add a "small amount" of a full rank matrix to $X^\top X$.

For reasons to become clear soon, we *standardise* the design matrix:

- Write $X = (1_n \; X')$, $\beta = (\beta_0 \; \gamma)^\top$
- Recentre/rescale the covariates (columns) defining:
  $Z_j = \frac{\sqrt{n}}{\mathrm{sd}(X'_j)}(X'_j - 1_n \overline{X'}_j)$
  - Coefficients now have common scale
  - Interpretation of β_j slightly different: not "mean impact on response per unit change of explanatory variable", but now "mean impact on response per unit deviation of explanatory variable from its mean, measured in units of standard deviation"
- The Z_j are all orthogonal to 1_n and are of unit norm.

- Since $Z_j \perp 1_n$ for all, j, we can estimate β_0 and γ by two separate regressions (orthogonality).
- Least squares estimators based on the *standardised design matrix* become

$$\widehat{\beta}_0 = \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \widehat{\gamma} = (Z^\top Z)^{-1} Z^\top y.$$

- Ridge regression replaces $Z^\top Z$ by $Z^\top Z + \lambda I_{(p-1)\times(p-1)}$ (i.e. adds a "ridge")

$$\boxed{\widehat{\beta}_0 = \overline{y}, \quad \widehat{\gamma} = (Z^\top Z + \lambda I)^{-1} Z^\top y}.$$

Adding $\lambda I_{(p-1)\times(p-1)}$ to $Z^\top Z$ makes inversion more stable
$\hookrightarrow$ λ called *ridge parameter*.

$\rightarrow$ Motivation of adding the ridge term λI

$\hookrightarrow$ Can see that $(\widehat{\beta}_0 \quad \widehat{\gamma}) = (\overline{y} \quad (Z^\top Z + \lambda I)^{-1} Z^\top y)$ minimizes

$$\|y - \beta_0 1_n - Z\gamma\|_2^2 + \lambda \|\gamma\|_2^2$$

or equivalently

$$\|y - \beta_0 1_n - Z\gamma\|_2^2 \quad \text{subject to} \quad \sum_{j=1}^{p-1} \gamma_j^2 = \|\gamma\|_2^2 \leq r(\lambda)$$

instead of least squares estimator which minimizes

$$\|y - \beta_0 1_n - Z\gamma\|_2^2.$$

Idea: in the presence of collinearity, coefficients are ill-defined: a wildly positive coefficient can be cancelled out by a largely negative coefficient (many coefficient combinations can produce the same effect). By imposing a *size* constraint, we limit the possible coefficient combinations!

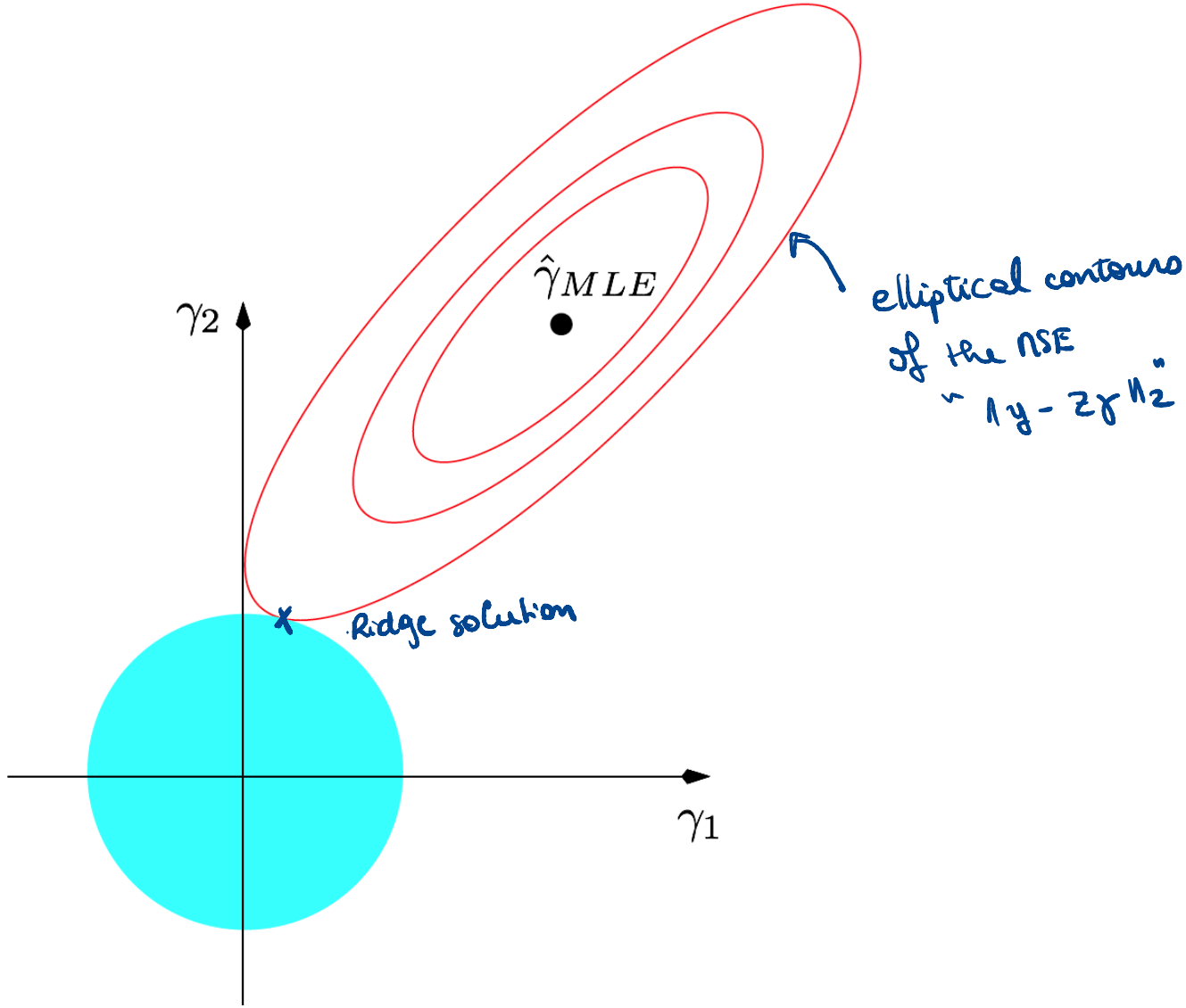


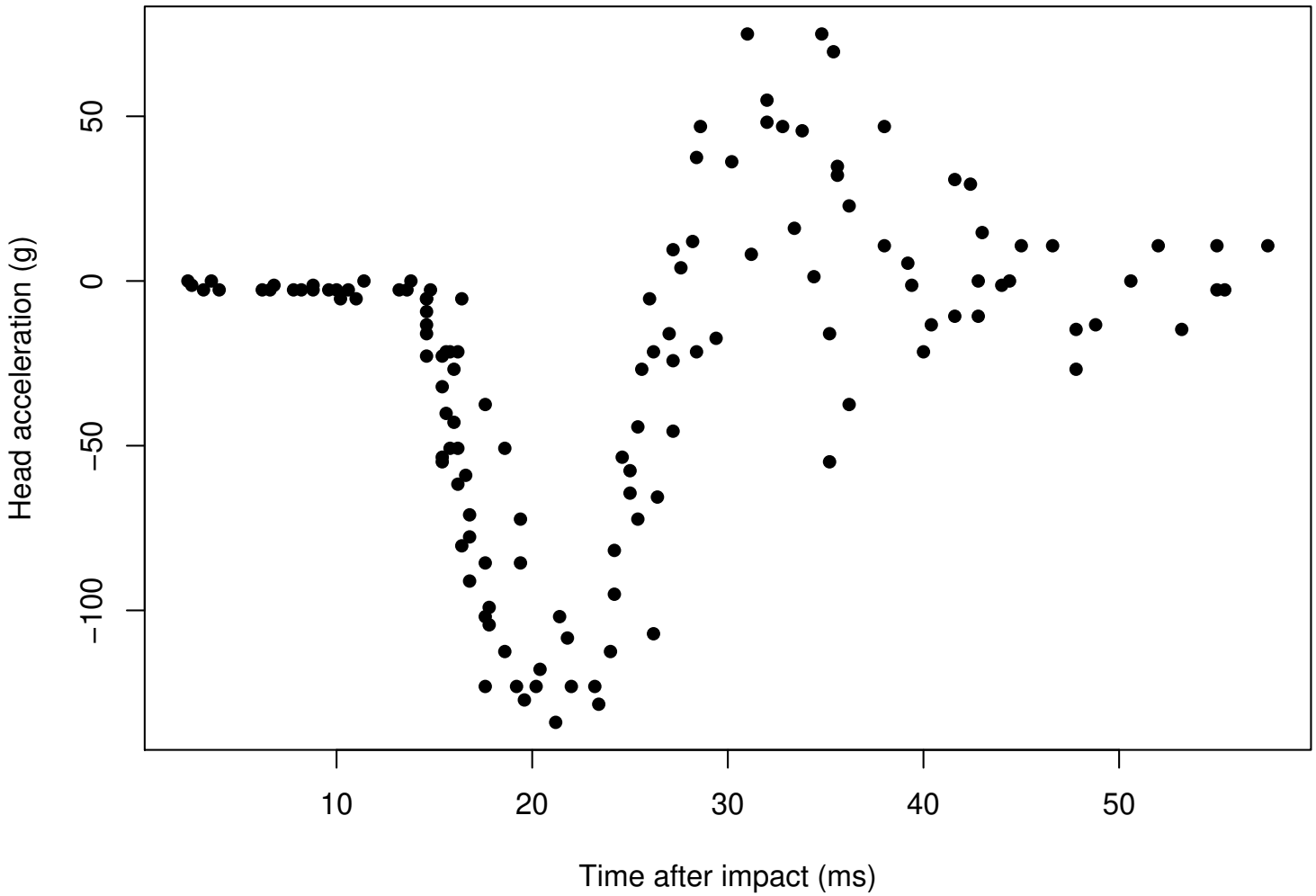
Figure: L^2 Shrinkage [Ridge Regression]

Semiparametric regression

- Normal linear model has two main aspects:
  - **systematic variation**, $\mathrm{E}(y) = \mu$, and $\mu = X\beta$ with parameters β;
  - **stochastic variation**, $y \sim \mathcal{N}_n(\mu, \sigma^2 I_n)$.
- Can relax the stochastic assumption using other distributions or second-order assumptions, but still have parametric model for the systematic part.
- Often want to relax systematic part for more flexible models, for
  - exploratory data analysis — 'will a linear model be adequate?'
  - confirmatory data analysis — 'I've fitted a linear model, is it adequate?'
  - general modelling — 'the data are too complex to expect a simple parametric model to work, so what can I do?'
  - semiparametric modelling — 'I will use a parametric model for the effects of interest, but can I model nuisance effects more flexibly?'
- Most basic tool is the **scatterplot smoother**.

Example: Motorcycle data

Measurements of head acceleration (g) at time after impact (ms) in a simulated motorcycle accident, used to test crash helmets:



Scatterplot smoothing

- Have data $(x_1, y_1), \ldots, (x_n, y_n)$, with $x_- \leq x_1 < \cdots < x_n \leq x_+$ (ahem) and we wish to estimate $\mathrm{E}(y) = \mu(x)$, for $x \in \mathcal{X} = [x_-, x_+]$.
- Suppose that $\mu \in \mathcal{M}$, a function space spanned by n linearly independent basis functions that can be identified by evaluation at $x_1, \ldots, x_n$, and let $\mu_j = \mu(x_j)$.
- Can choose a basis $\{b_1(x), \ldots, b_n(x)\}$ for $\mathcal{M}$ such that $\mu(x) = \sum_{j=1}^{n} \mu_j b_j(x)$ interpolates $(x_1, \mu_1), \ldots, (x_n, \mu_n)$.
- Suppose that $\mathcal{M}$ contains the linear functions on $\mathcal{X}$ and that the second derivatives of the $b_j(x)$ are not all zero, so functions in $\mathcal{M}$ may also be nonlinear in x.
- To estimate μ we minimise a **penalised sum of squares**,

$$\sum_{j=1}^{n} \{y_j - \mu(x_j)\}^2 + \lambda \int_{\mathcal{X}} \{\mu''(x)\}^2 \, \mathrm{d}x, \tag{7}$$

  where the **roughness penalty** imposes smoothness: if $\lambda \to 0$, then $\mu(x_j) \to y_j$ and $\widehat{\mu}$ interpolates, but when $\lambda \to \infty$ even tiny wiggles in μ will give a huge penalty, making $\widehat{\mu}$ linear.
- The penalty does not affect linear functions, so $\mathcal{M} = \mathcal{L} \bigoplus \mathcal{P}$, where $\mathcal{L}$ and $\mathcal{P}$ are the two-dimensional vector space of linear functions on $\mathcal{X}$ and an $(n-2)$-dimensional vector space of nonlinear functions on $\mathcal{X}$, and $\bigoplus$ denotes addition of vector spaces.

- The roughness term is

$$\int_{\mathcal{X}} \{\mu''(x)\}^2 \, \mathrm{d}x = \int_{\mathcal{X}} \left\{ \sum_{j=1}^{n} \mu_j b_j''(x) \right\}^2 \mathrm{d}x = \sum_{i,j=1}^{n} \mu_i \mu_j \int_{\mathcal{X}} b_i''(x) b_j''(x) \, \mathrm{d}x = \mu^{\mathrm{T}} S \mu,$$

  say, where $\mu^{\mathrm{T}} = (\mu_1, \ldots, \mu_n)$.

- $S_{n \times n}$ has (i, j) element $\int_{\mathcal{X}} b_i''(x) b_j''(x) \, \mathrm{d}x$ and is symmetric and positive semi-definite of rank $n - 2$, because linear functions are unpenalised, so $S 1_n = S(x_1, \ldots, x_n)^{\mathrm{T}} = 0$.

- The penalised sum of squares

$$(y - \mu)^{\mathrm{T}}(y - \mu) + \lambda \mu^{\mathrm{T}} S \mu \equiv -2\mu^{\mathrm{T}} y + \mu^{\mathrm{T}}(I_n + \lambda S)\mu,$$

  is minimised by $\widehat{\mu}_\lambda = (I_n + \lambda S)^{-1} y$.

- As λ increases from zero, the fitted value $\widehat{\mu}_\lambda$ shrinks from y towards the straight-line regression fit to y, which is unpenalised.

- The equivalent degrees of freedom are $\mathbf{edf}_\lambda = \mathrm{tr}(H_\lambda) = \sum_{j=1}^{n} (1 + \lambda \delta_j)^{-1}$, where $\delta_1 \geq \cdots \geq \delta_3 > \delta_2 = \delta_1 = 0$ are the eigenvalues of S. As λ increases $\mathbf{edf}_\lambda$ decreases monotonically from $\mathbf{edf}_0 = n$ towards $\mathbf{edf}_\infty = 2$.

- In principle we might take any basis functions, but in practice we usually take local polynomials known as **splines** that have good approximation properties.
- There are many forms of splines, which
  - are often cubic polynomials with finite support between values of x known as **knots**, $x_1^*, \ldots, x_K^*$, and then S is tri-diagonal,
  - sometimes form a **natural cubic spline**, which has $K = n$ and certain optimality properties,
  - are discussed in more detail later.
- If there is no penalisation ($\lambda = 0$) then we have a standard linear model, and spline basis functions are called **regression splines**.
- Under second-order assumptions we choose λ by minimising $\mathrm{CV}(\lambda)$ or $\mathrm{GCV}(\lambda)$.
- Under normal-theory assumptions we can use REML to estimate σ^2 and λ.
- Obvious generalisation allows weight matrix $W = \mathrm{diag}(w_1, \ldots, w_n)$.
- If the $x_1, \ldots, x_n$ are not unique, write $\mathrm{E}(y) = N_{n \times n'} \mu_{n' \times 1}$ in terms of the means μ at the n' unique elements of x, and minimise

$$(y - N\mu)^{\mathrm{T}} W (y - N\mu) + \lambda \mu^{\mathrm{T}} S \mu.$$

  where $S_{n' \times n'}$ arises as before from the roughness penalty on $\mu(x)$.

Linear, quadratic and cubic B-splines