Regression Methods

Myrto Limnios

Autumn 2025 - Week 12

Regularisation - General Framework

Generalisations

- We've discussed estimation of a single function $\mu(x)$, but in applications we may have
  - covariates to be treated parametrically,
  - several smooth functions,
  - non-normal response variable,
  - unmeasured effects possibly referred to as random effects.
- To include ordinary covariates and allow for weights, we write

$$y \mid b \sim (B\theta, \sigma^2 W), \quad B\theta = X\beta + Zb,$$

where $B = (X, Z)$ is $n \times d$, $\theta = (\beta^{\mathrm{T}}, b^{\mathrm{T}})^{\mathrm{T}}$ is $d \times 1$, $d = p + q$ and

  - the $n \times p$ matrix X represents the ordinary covariates, plus any unpenalised columns for smooth components,
  - the $p \times 1$ parameter vector β is unpenalized,
  - the $n \times q$ matrix Z represents the bases for any smooth functions,
  - the $q \times 1$ vector b is penalized,
  - the $n \times n$ diagonal matrix $W = \mathrm{diag}(w_1, \ldots, w_n)$ contains positive weights,

and everything 'goes through as before'.

Additivity and identifiability

- Consider the **additive model**

$$\mathrm{E}(y) = \mu_1(x) + \mu_2(z),$$

where μ_1, μ_2 belong to suitable classes of smooth functions; if

$$x \equiv \text{time}, \quad z \equiv \text{space},$$

then μ_1 is defined on $\mathcal{X}_1 \subset \mathbb{R}$ and μ_2 is defined on $\mathcal{X}_2 \subset \mathbb{R}^2$.

- There is an identifiability problem, since we could map

$$\mu_1(x) \mapsto \mu_1(x) + a, \quad \mu_2(z) \mapsto \mu_2(z) - a, \quad a \in \mathbb{R},$$

and the fitted values would not change, so we must constrain μ_1 and μ_2.

- As before, we use bases for μ_1 and μ_2, writing

$$\mathrm{E}(y) = Zb = \begin{pmatrix} Z_1(x) & Z_2(z) \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix},$$

where we penalise the q_1 elements of b_1 and the q_2 elements of b_2.

Ensuring identifiability

- The identifiability problem is solved by **centering** the fitted smooth, i.e., enforcing
$$1_n^{\mathrm{T}} Z_{n\times q} b_{q\times 1} = 0$$
for each smooth term.
- In general we can use a QR decomposition. If $C_{a\times q} b_{q\times 1} = 0_{a\times 1}$, with $a < q$, write
$$C_{q\times a}^{\mathrm{T}} = Q_{q\times q} R_{q\times a} = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix} \begin{pmatrix} R_1 \\ 0 \end{pmatrix},$$
where Q is orthogonal,
  - Q_1 has dimension $q \times a$,
  - Q_2 has dimension $q \times (q-a)$, and
  - R_1 has dimension $a \times a$ and is upper triangular.

  Then if we set $b_{q\times 1} = Q_2 b'_{(q-a)\times 1}$, we have
$$Cb = R^{\mathrm{T}} Q^{\mathrm{T}} b = \begin{pmatrix} R_1^{\mathrm{T}} & 0 \end{pmatrix} \begin{pmatrix} Q_1^{\mathrm{T}} \\ Q_2^{\mathrm{T}} \end{pmatrix} Q_2 b' = \begin{pmatrix} R_1^{\mathrm{T}} & 0 \end{pmatrix} \begin{pmatrix} 0 \\ I_{q-m} \end{pmatrix} b' = 0.$$
- Thus the constraint is satisfied if we replace $Z_{n\times q}$ by $(ZQ_2)_{n\times(q-1)}$; this reduces b to dimension $(q-1)\times 1$.

Penalty formulation

- Minimise

$$(y - B\theta)^{\mathrm{T}} W (y - B\theta) + \theta^{\mathrm{T}} S_\lambda \theta = (y - X\beta - Zb)^{\mathrm{T}} W (y - X\beta - Zb) + \theta^{\mathrm{T}} S_\lambda \theta$$

  where S_λ is a sum of symmetric positive semi-definite $d \times d$ matrices S_m, such that

$$\theta^{\mathrm{T}} S_\lambda \theta = \theta^{\mathrm{T}} \left(\sum_{m=1}^{M} \lambda_m S_m \right) \theta = \sum_{m=1}^{M} \lambda_m b_m^{\mathrm{T}} S_m^* b_m, \quad \lambda_m \geq 0,$$

  where S_m^* is the non-zero diagonal block of S_m and b has sub-vectors $b_1, \ldots, b_M$.

- With $M = 2$, β, b_1 and b_2 are vectors of respective lengths p, q_1 and q_2, and S_1^* and S_2^* are square matrices of sides q_1 and q_2, so

$$\theta = \begin{pmatrix} \beta \\ b_1 \\ b_2 \end{pmatrix}, \quad S_\lambda = \lambda_1 S_1 + \lambda_2 S_2 = \lambda_1 \begin{pmatrix} 0 & 0 & 0 \\ 0 & S_1^* & 0 \\ 0 & 0 & 0 \end{pmatrix} + \lambda_2 \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & S_2^* \end{pmatrix},$$

  with S_1 and S_2 partitioned conformably with θ.

- Let S_λ^* denote the $q \times q$ corner of S_λ corresponding to b; here $S_\lambda^* = \mathrm{diag}(\lambda_1 S_1^*, \lambda_2 S_2^*)$.

- Note that $|S_\lambda|_+ = |S_\lambda^*|_+$.

Estimation

- For fixed λ, the minimiser and fitted values for

$$(y - B\theta)^{\mathrm{T}} W (y - B\theta) + \theta^{\mathrm{T}} S_\lambda \theta$$

are

$$\widehat{\theta}_\lambda = (B^{\mathrm{T}} W B + S_\lambda)^{-1} B^{\mathrm{T}} W y, \quad \widehat{y}_\lambda = B\widehat{\theta}_\lambda = B(B^{\mathrm{T}} W B + S_\lambda)^{-1} B^{\mathrm{T}} W y = H_\lambda y.$$

- If the unpenalized least squares estimator $\widehat{\theta} = (B^{\mathrm{T}} W B)^{-1} B^{\mathrm{T}} W y$ exists, then

$$\widehat{\theta}_\lambda = (B^{\mathrm{T}} W B + S_\lambda)^{-1} B^{\mathrm{T}} W B \widehat{\theta} = \widehat{\theta} - (B^{\mathrm{T}} W B + S_\lambda)^{-1} S_\lambda \widehat{\theta} = P_\lambda \widehat{\theta},$$

and if $\widehat{y}$ is the unpenalised fitted value, then

$$\widehat{y}_\lambda = \widehat{y} - B(B^{\mathrm{T}} W B + S_\lambda)^{-1} S_\lambda \widehat{\theta}.$$

- Now we must decide
  - how many degrees of freedom for each smooth?
  - how to select the smoothing parameters?

Amount of smoothing

- We write
$$\widehat{\theta}_\lambda = P_\lambda \widehat{\theta},$$
say, where P_λ shows how penalisation shrinks $\widehat{\theta}$ towards $\widehat{\theta}_\infty = (\widehat{\beta}^{\mathrm{T}}, 0^{\mathrm{T}})^{\mathrm{T}}$.
- If $\lambda \approx 0$, then $P_\lambda \approx I_{p+q}$ and the degrees of freedom of the two fits are both $\approx p+q$, but as $\lambda \to \infty$, P_λ tends to the projection matrix onto the column space of $X_{n\times p}$.
- On slide 13 week 11 with just one smooth term we defined
$$\mathbf{edf}_\lambda = \mathrm{tr}(H_\lambda) = \mathrm{tr}(P_\lambda) = \sum_{r=1}^{p+q} P_{\lambda,rr} \in (p, p+q),$$
which gives the usual definition for a linear model.
- If $\theta^{\mathrm{T}} = (\beta^{\mathrm{T}}, b_1^{\mathrm{T}}, \ldots, b_M^{\mathrm{T}})$, we define the **effective degrees of freedom** $\mathbf{edf}_{\lambda_m}$ associated to the mth smooth as being the sum of those $P_{\lambda,rr}$ that correspond to the elements of b_m in θ.
- To choose the vector λ we use either
  - CV(λ) or GCV(λ) (second-order assumptions),
  - REML (normal-theory assumptions).
- Must optimise over (log) λ, e.g., by grid search (CV/GCV) or other methods (REML).

Inference

- So far we have discussed only 'point estimation' of a smooth function $\mu(x)$, but in applications we also want
  - pointwise confidence intervals for smooth functions,
  - overall confidence bands for (say) $\{\mu(x) : x \in \mathcal{S}\}$, where $\mathcal{S}$ is some subset of $\mathcal{X}$, and
  - tests of hypotheses such as 'is the spline part needed?' and 'is the curve monotonic?'
- Under the normal model we have the Bayesian interpretation from week 9

$$\theta \mid y, \sigma^2, \lambda \sim \mathcal{N}_d\left(\widehat{\theta}_\lambda, V_\lambda\right), \quad V_\lambda = \sigma^2 (B^{\mathrm{T}} W B + S_\lambda)^{-1},$$

  from which we can simulate to find bounds for any function $A(\theta)$.
- If $A(\theta) = A_{m\times d}\theta$, then

$$A\theta \mid y, \sigma^2, \lambda \sim \mathcal{N}_m(A\widehat{\theta}_\lambda, AV_\lambda A^{\mathrm{T}}),$$

  and generalisation Eq. (6) s11.w9 gives that its mean square error is

$$\mathrm{MSE} = \mathrm{E}\left(\|A\widehat{\theta}_\lambda - A\theta\|^2\right) = \mathrm{tr}(AV_\lambda A^{\mathrm{T}}),$$

  which takes into account both estimation error and prior uncertainty about θ.

Average coverage probabilities

- Bayesian credible intervals have good frequentist properties, averaged over the domain of x.
- Let the random index variable J choose the m rows a_j^{T} of A with equal probabilities, and aim to choose constants d and c_j such that the **average coverage probability**

$$\mathrm{ACP} = \Pr\left\{|a_J^{\mathrm{T}}\widehat{\theta}_\lambda - a_J^{\mathrm{T}}\theta| \le dc_J\right\} = 1-\alpha;$$

  i.e., ACP has a desired value averaged over y, θ and J.
- The random variable

$$a_J^{\mathrm{T}}(\widehat{\theta}_\lambda - \theta)/c_J = a_J^{\mathrm{T}}\{\widehat{\theta}_\lambda - \mathrm{E}(\widehat{\theta}_\lambda)\}/c_J + a_J^{\mathrm{T}}\{\mathrm{E}(\widehat{\theta}_\lambda) - \theta\}/c_J = S + T,$$

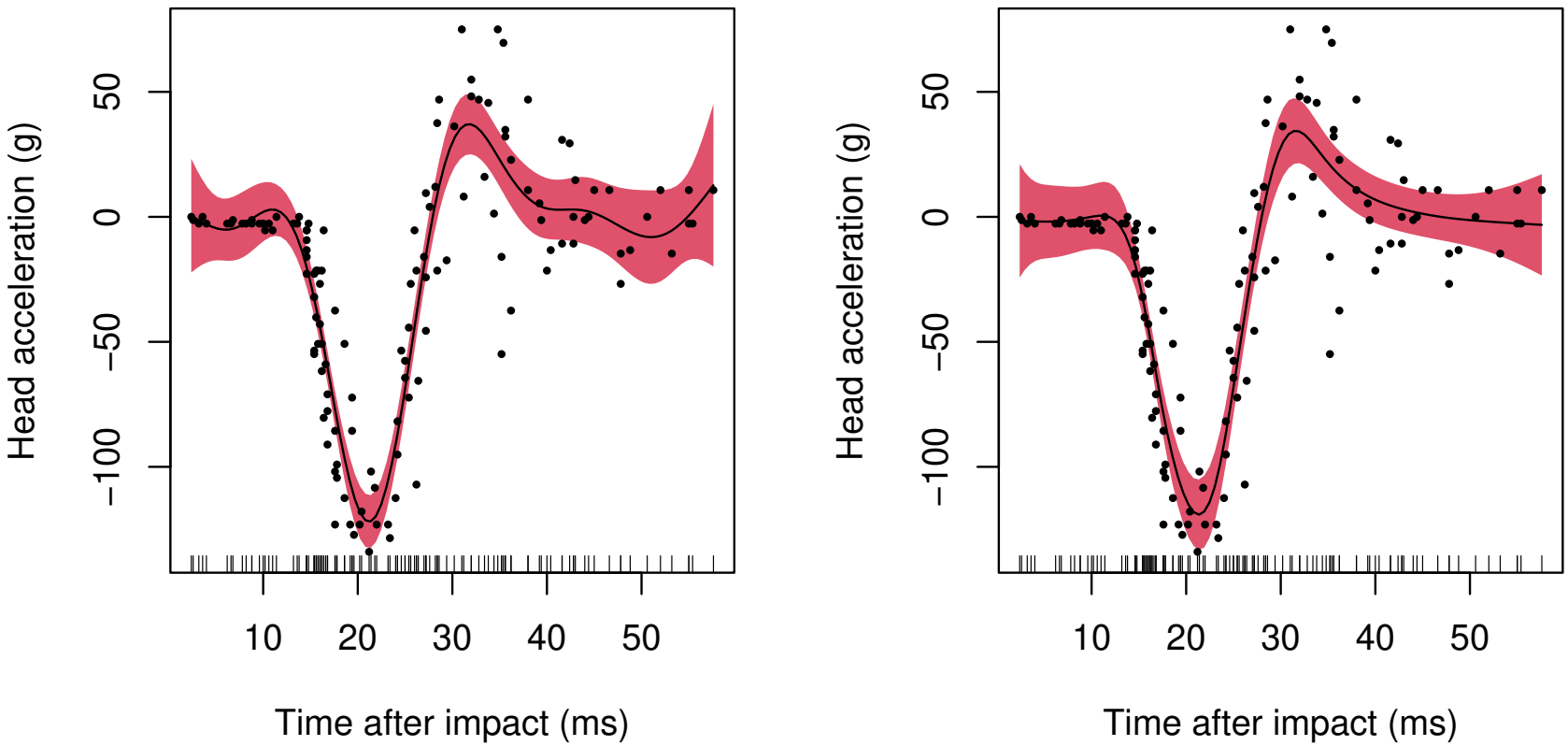
  say, has a mixture of normal distributions, where
  - S is approximately normal and $\mathrm{E}(S) = 0$,
  - T is random (because of J) with $\mathrm{E}(T) \approx 0$, but $\mathrm{var}(T) \ll \mathrm{var}(S)$.
- We now choose $C = \mathrm{diag}(c_1, \ldots, c_m) = \mathrm{diag}(AV_\lambda A^{\mathrm{T}})^{1/2}$, so that

$$\mathrm{var}(S+T) \approx m^{-1}\mathrm{E}\left\{\|C^{-1}A(\widehat{\theta}_\lambda - \theta)\|^2\right\} = m^{-1}\mathrm{tr}\left(C^{-1}AV_\lambda A^{\mathrm{T}}C^{-1}\right) = 1,$$

  and then setting $d = z_{1-\alpha/2}$ gives the required value for ACP.
- This ignores estimation error for σ^2 and λ.

Example: Motorcycle data

Standard (left) and adaptive (right) spline fits, the latter with $K = 40$ and $L = 5$, and 95% pointwise confidence intervals:

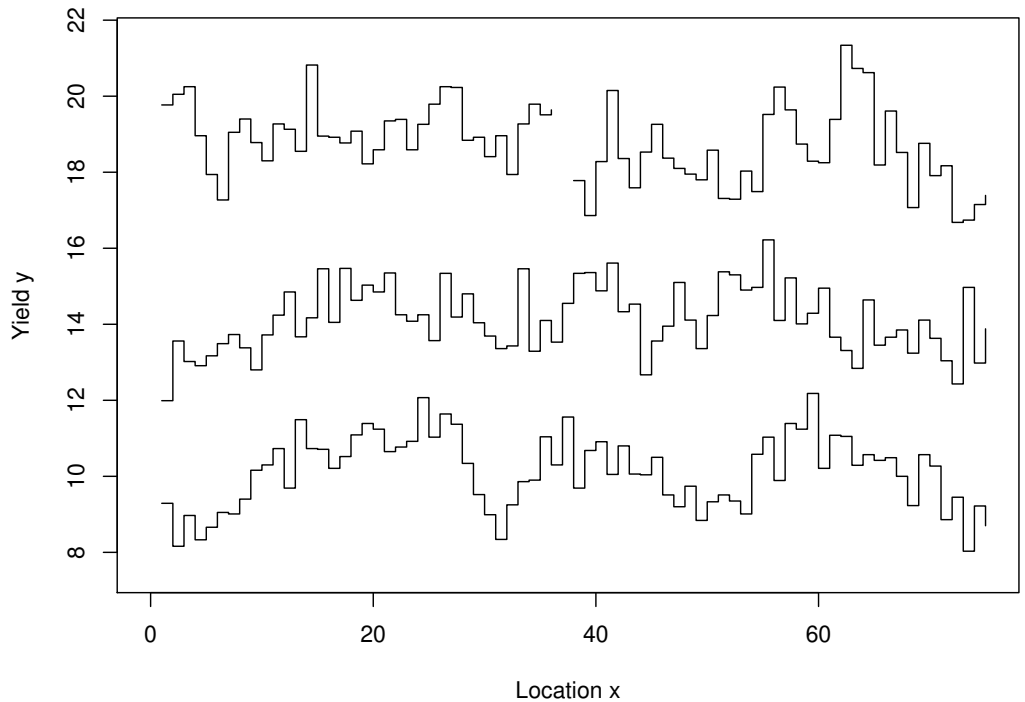


Example: Spring barley data

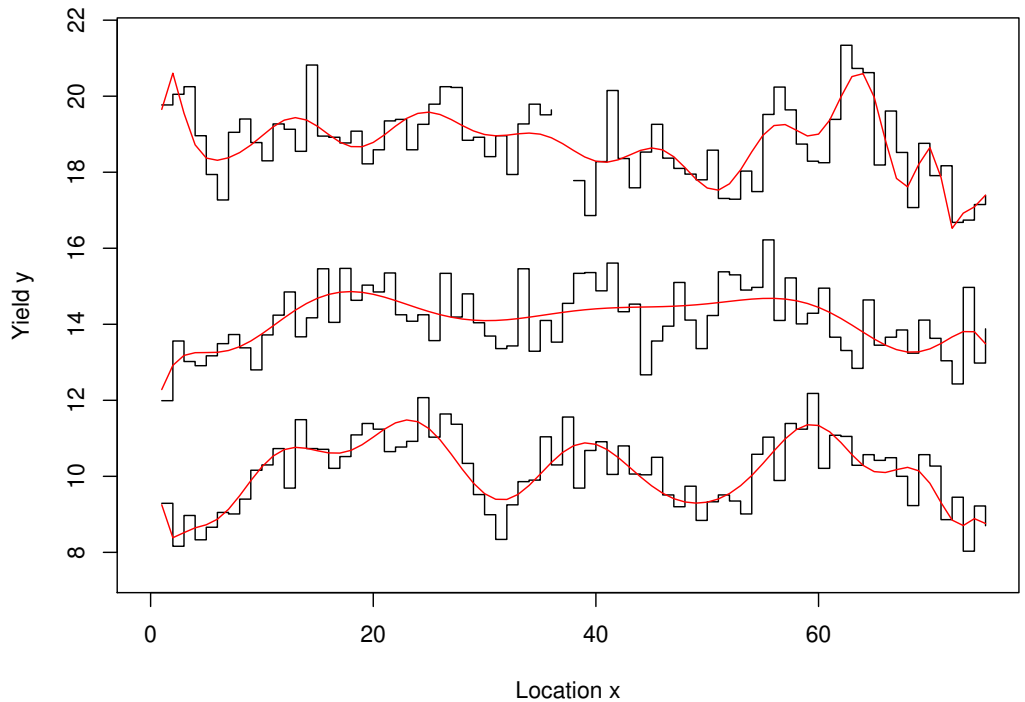
Plot yield at harvest for 75 varieties of spring barley sown in 3 blocks each of 75 plots:

Location x	Block 1		Block 2		Block 3	
	Variety	Yield y	Variety	Yield y	Variety	Yield y
1	57	9.29	49	7.99	63	11.77
2	39	8.16	18	9.56	38	12.05
3	3	8.97	8	9.02	14	12.25
4	48	8.33	69	8.91	71	10.96
5	75	8.66	29	9.17	22	9.94
6	21	9.05	59	9.49	46	9.27
7	66	9.01	19	9.73	6	11.05
8	12	9.40	39	9.38	30	11.40
9	30	10.16	67	8.80	16	10.78
10	32	10.30	57	9.72	24	10.30
11	59	10.73	37	10.24	40	11.27
12	50	9.69	26	10.85	64	11.13
13	5	11.49	16	9.67	8	10.55
14	23	10.73	6	10.17	56	12.82
15	14	10.71	47	11.46	32	10.95
16	68	10.21	36	10.05	48	10.92
17	41	10.52	64	11.47	54	10.77
18	1	11.09	63	10.63	37	11.08
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

Example: Spring barley data



Spring barley data and polynomial fits



Example: Spring barley data

- We fit a model with parametric variety effects and smooth effects for the fertility patterns in the blocks,

$$y_{n\times 1} \sim (X_{n\times 75}\beta_{75\times 1} + Z_1 b_1 + Z_2 b_2 + Z_3 b_3, \sigma^2 I_n),$$

  where

  - $n = 224$, as one of the responses is missing,
  - X is a matrix of indicators (0/1) of which variety is in which plot in each block,
  - β are the variety effects, with the model parametrized without an overall mean,
  - Z_m of dimension $n \times (p_m + q_m)$ corresponds to the basis functions for the smooth in block m, and
  - b_m are of dimensions $(p_m + q_m) \times 1$, for $m = 1, 2, 3$, corresponding to the smooth effects, and
  - $p_m + q_m = 9$ by default (after centering) when using gam in R package mgcv.

- Taking $p_m = 2$ would correspond to null smooth $\beta_0 + \beta_1 x$ for each block (i.e., linear fertility pattern), but the identifiability constraints impose $\beta_0 = 0$. Hence in fact $p_m = 1$ for a linear baseline smooth and the degrees of freedom for the smooth terms lie in $[1, 9]$ (see slide 17).

Example: Spring barley data

```
library(SMPracticals)
data(barley)

library(mgcv)

# ML fit of variety as fixed effect, with GCV estimation of lambdas,
# with splines for fertility gradients within each block

fit.gcv <- gam(y~Variety-1+s(Location,by=Block),data=barley)

# fit of variety as fixed effect, with REML estimation of lambdas,
# with splines for fertility gradients within each block

fit <- gam(y~Variety-1+s(Location,by=Block),method="REML",data=barley)

# REML fit with variety as a random effect and splines for fertilities

fit.re <- gam(y~s(Variety,bs="re")+s(Location,by=Block),method="REML",
                    data=barley)
```

Example: Spring barley data

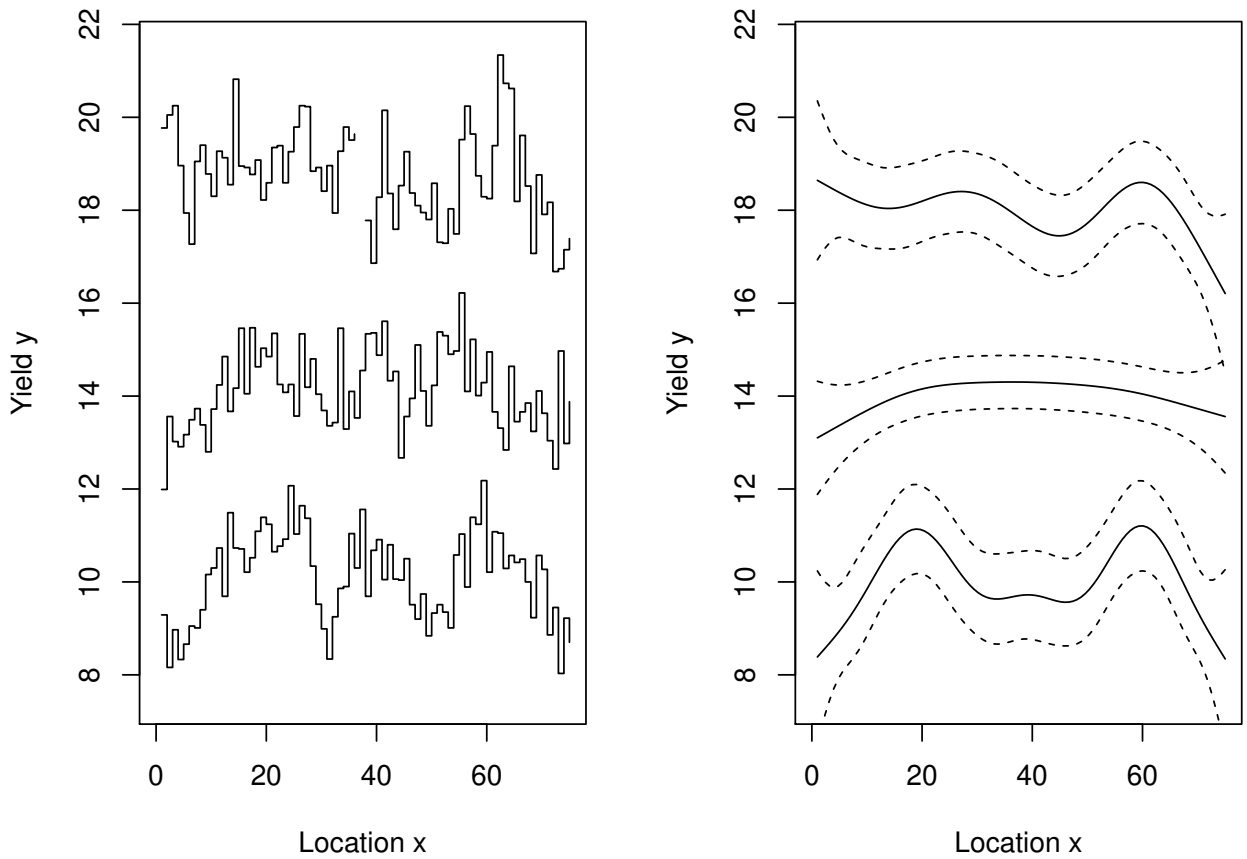
- Using GCV the smooths have $\mathrm{df}_\lambda = 8.3, 6.8, 6.3$, with $\widehat{\sigma} = 0.65$ and AIC $= 513.1$, the residual degrees of freedom is $224 - 75 - 8.3 - 6.8 - 6.3 \approx 130.6$, with SEs around 0.4 for the estimated variety effects (0.54 for variety 27).
- Using REML the smooths have $\mathrm{df}_\lambda = 7.2, 3, 6.1$, with $\widehat{\sigma} = 0.66$ and AIC $= 518.3$, the residual degrees of freedom is 132.7, with SEs around 0.4 for the estimated variety effects (0.53 for variety 27).
- The estimated smoothing parameters are $\widehat{\lambda}_1 = 0.0029$, $\widehat{\lambda}_2 = 0.18$ and $\widehat{\lambda}_3 = 0.0078$.
- The effective degrees of freedom for the smooth terms, with the totals:

Block	$P_{\lambda,rr}$									Total
1	1.00	1.07	0.90	0.7	0.65	0.17	0.38	1.31	1	7.18
2	0.61	0.21	0.12	−0.2	0.03	−0.26	0.01	1.49	1	3.00
3	0.99	1.04	0.76	0.4	0.41	−0.18	0.18	1.47	1	6.07

- The $P_{\lambda,rr}$ need not be positive, though their total for each smooth is positive.
- In applications it would be wise to check whether increasing q_m would lead to very different fits.

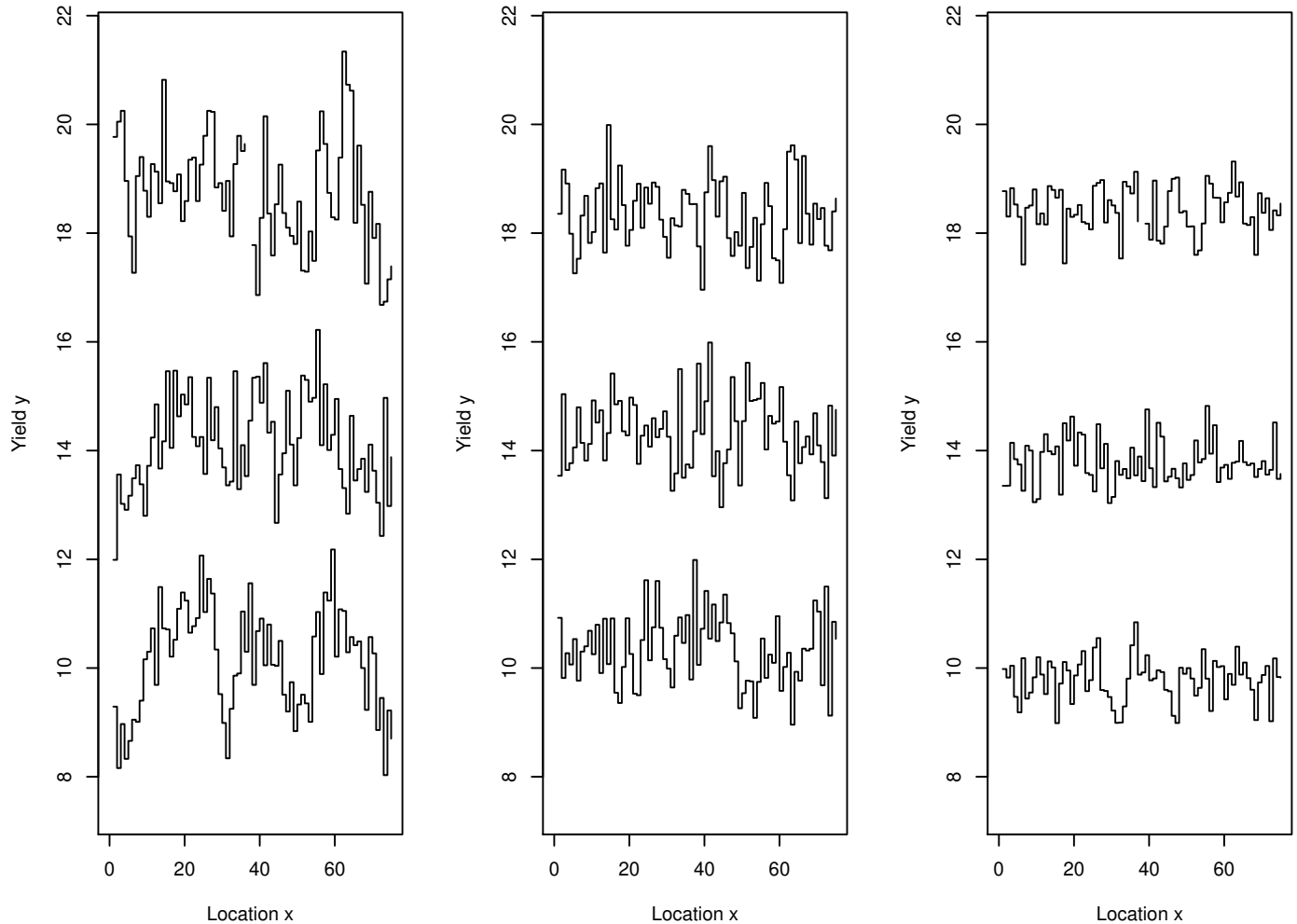
Example: Spring barley data

Left: data (offset by adding 4 and 8 to blocks 2 and 3). Right: estimated fertility patterns (with estimated df 7.2, 3, 6.1) and 95% unconditional pointwise confidence intervals, fitted using REML. The intervals are wider for blocks 1 and 3.



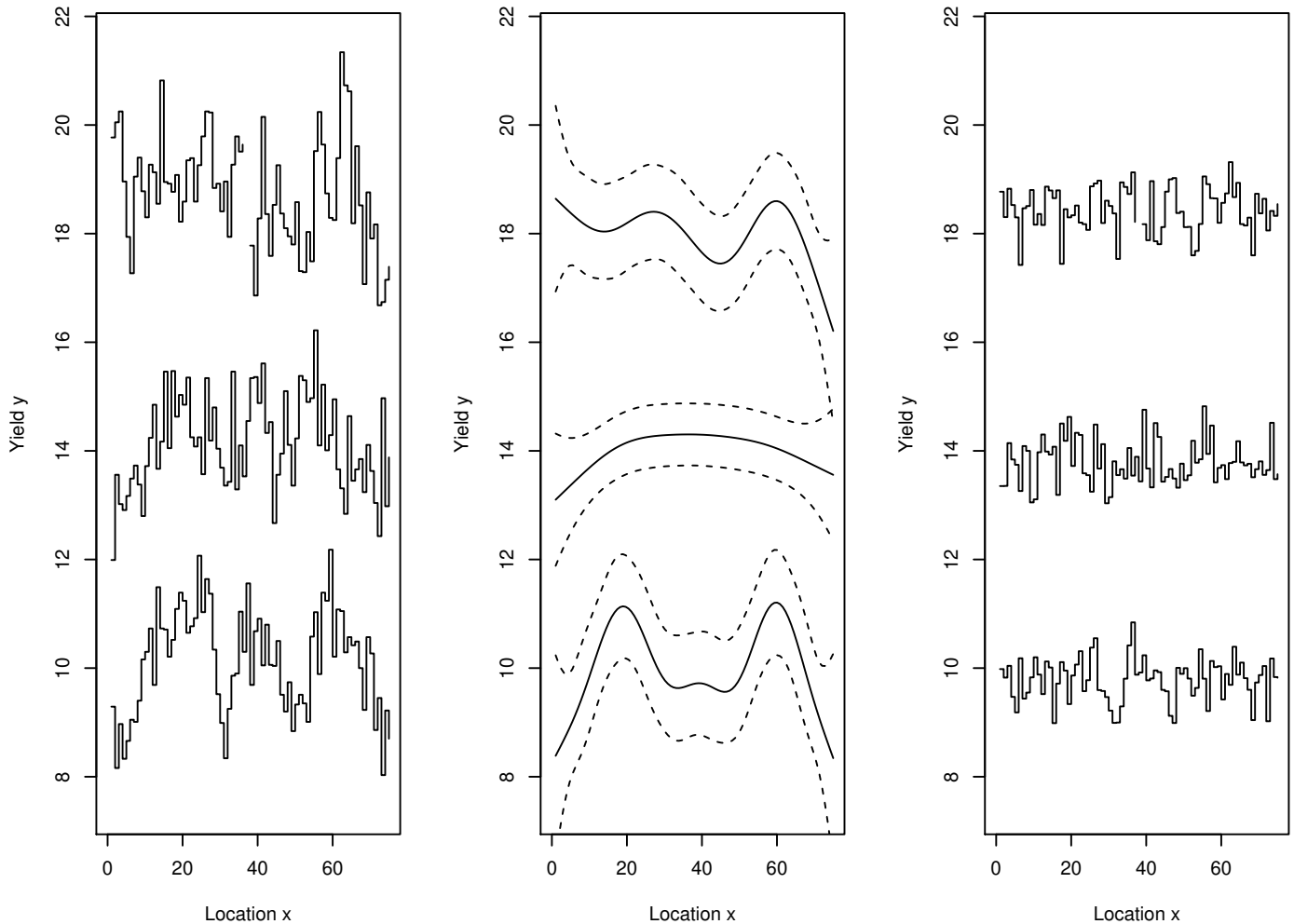
Example: Spring barley data

Left: data (offset by adding 4 and 8 to blocks 2 and 3). Center: Estimated variety effects (also offset). Right: residuals (also offset, and showing serial autocorrelation?)



Example: Spring barley data

Left: data (offset by adding 4 and 8 to blocks 2 and 3). Center: estimated fertility patterns (REML), also offset. Right: residuals.



Example: Spring barley data

- Should the varieties be treated as randomly selected from a population of varieties?
- If so, we use the same basis matrix X as in the previous model, but add a penalty matrix $\lambda_\beta S_\beta$ and minimise the penalised sum of squares

$$(y - B\theta)^{\mathrm{T}}(y - B\theta) + \theta^{\mathrm{T}} S_\lambda \theta,$$

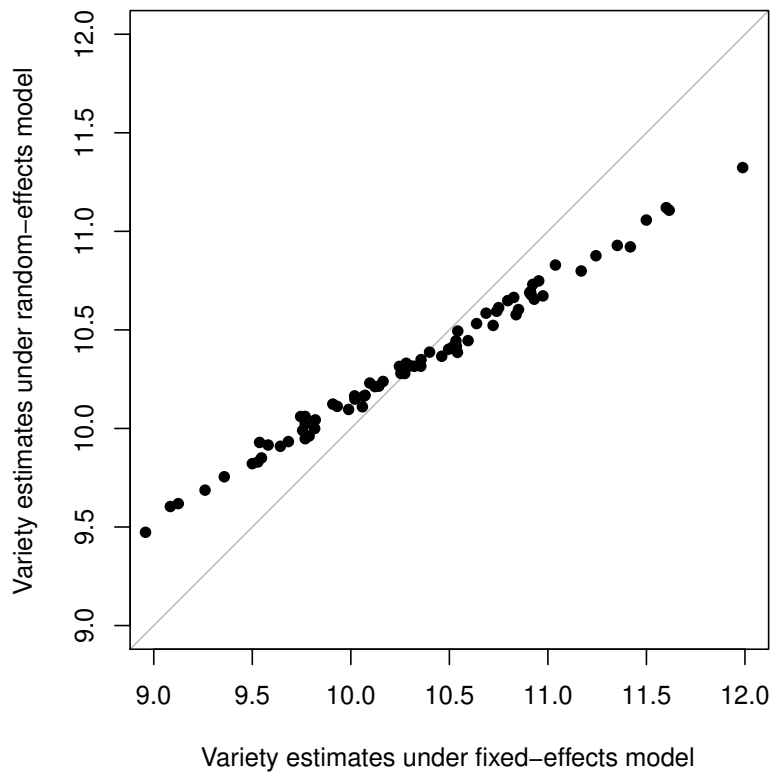
  where

$$S_\lambda = \lambda_\beta S_\beta + \lambda_1 S_1 + \lambda_2 S_2 + \lambda_3 S_3,$$

  where $S_\beta = \mathrm{diag}(I_{75}, 0)$.
- The effective degrees of freedom are then 44.8 for β and 7.5, 3.9 and 6.4 for the splines.
- The optimal smoothing parameters are $\widehat{\lambda}_\beta = 1.76$, $\widehat{\lambda}_1 = 0.0027$, $\widehat{\lambda}_2 = 0.073$ and $\widehat{\lambda}_3 = 0.0070$.
- The fixed-effects model has 75 degrees of freedom for β, so this is substantial shrinkage; the estimated standard deviation drops from 0.65 to 0.39.
- The estimates under the random-effects model have standard errors around 0.31 (0.36 for variety 27), compared to 0.41 (0.54 for variety 27) for the fixed-effects model.
- The next slide compares the estimates.

Example: Spring barley data

Comparison of estimated variety effects under fixed-effects and random-effects models:



Comments

- Penalised estimation extends the basic smoothers to include
  - parametric terms in models,
  - several smooth terms,
  - spatial and more complex smoothing,
  - 'random effect' parameters,

  and extends to **generalized additive models** in a natural way.
- The baseline variance σ^2 and smoothing parameter(s) λ are estimated using cross-validation under second-order assumptions or REML under normality.
- The empirical Bayes formulation allows inference on parameters and smooth functions in a unified way — usually ignoring the uncertainty for σ^2 and λ is not too critical.
- In practice n and d may be very big, so direct matrix inversion is computationally painful, and then indirect methods (e.g., based on the Woodbury formula) are needed to compute $\widehat{\theta}_\lambda$ and V_λ.

Regularisation - Generalized Additive Models

Generalized additive model

- Now we write
$$\mathrm{E}(y) = \mu, \quad g(\mu) = \eta = B\theta = X\beta + Zb,$$
where
  - y follows a GLM (or more general) distribution,
  - $g(\cdot)$ is a link function,
  - the rest is as before ...

  giving a **generalized additive model (GAM)**.
- For a general treatment, suppose we have a penalized log likelihood,
$$\ell_\lambda(\theta) = \ell(\theta) - \tfrac{1}{2}\theta^{\mathrm{T}} S_\lambda \theta = \sum_{j=1}^{n} \ell_j\{\eta_j(\theta)\} - \tfrac{1}{2}\theta^{\mathrm{T}} S_\lambda \theta,$$
where $\theta_{d\times 1}$ (with $d = p + q$) contains $\beta_{p\times 1}$ and $b_{q\times 1}$, the latter penalized using a symmetric positive semidefinite $d \times d$ matrix S_λ, and the underlying observations $y_1, \ldots, y_n$ giving likelihood contributions $\ell_1, \ldots, \ell_n$ are assumed to be independent.
- Now we apply the argument leading to the IWLS algorithm to ℓ_λ, leading to the **penalized iterative weighted least squares (PIWLS)** algorithm.

PIWLS

- For fixed λ, we apply (ridge regression) iterative weighted least squares with update step
$$\widehat{\theta}_\lambda = (B^{\mathrm{T}} W B + S_\lambda)^{-1} B^{\mathrm{T}} W z,$$
where S_λ is the penalty matrix, and
$$\begin{aligned} B_{n\times d} &= \partial\eta/\partial\theta^{\mathrm{T}}, \quad \text{(design matrix)} \\ W_{n\times n} &= \mathrm{diag}(w_1,\ldots,w_n), \quad w_j = \{\mathrm{E}(-\partial^2\ell_j/\partial\eta_j^2)\}, \quad \text{(weights)} \\ u_{n\times 1} &= \partial\ell/\partial\eta, \quad \text{(score vector)}, \\ z_{n\times 1} &= B\theta + W^{-1}u, \quad \text{(adjusted dependent variable)}. \end{aligned}$$
It is easier (but less stable) to use the (random) $-\partial^2\ell_j/\partial\eta_j^2$ in place of $\mathrm{E}(-\partial^2\ell_j/\partial\eta_j^2)$.
- Thus to obtain (penalized) MLEs $\widehat{\theta}_\lambda$ we use the **PIWLS algorithm**:
- fix λ and take an initial $\widehat{\theta}_\lambda$. Repeat
  - compute η, B, W, u, z;
  - compute new $\widehat{\theta}_\lambda = (B^{\mathrm{T}} W B + S_\lambda)^{-1} B^{\mathrm{T}} W z$;

  until changes in $\ell_\lambda(\widehat{\theta}_\lambda)$ (or $\widehat{\theta}_\lambda$, or both) are lower than some tolerance.
- We may add a line search: if $\ell_\lambda(\widehat{\theta}_{\lambda,\mathrm{new}}) < \ell_\lambda(\widehat{\theta}_{\lambda,\mathrm{old}})$, halve the step length and try again.

Derivation of PIWLS algorithm

- To find the estimate $\widehat{\theta}_\lambda$ starting from a trial value θ, we make a Taylor series expansion in the score equation

$$0 = \frac{\partial \ell_\lambda(\widehat{\theta}_\lambda)}{\partial \theta} \doteq \frac{\partial \ell_\lambda(\theta)}{\partial \theta} + \frac{\partial^2 \ell_\lambda(\theta)}{\partial \theta \partial \theta^{\mathrm{T}}}(\widehat{\theta}_\lambda - \theta),$$

where

$$\frac{\partial \ell_\lambda(\theta)}{\partial \theta} = B^{\mathrm{T}} u(\theta) - S_\lambda \theta$$

$$\frac{\partial^2 \ell_\lambda(\theta)}{\partial \theta_r \partial \theta_s} = \sum_{j=1}^{n} \frac{\partial \eta_j(\theta)}{\partial \theta_r} \frac{\partial^2 \ell_j(\theta)}{\partial \eta_j^2} \frac{\partial \eta_j(\theta)}{\partial \theta_s} + \sum_{j=1}^{n} \frac{\partial^2 \eta_j(\theta)}{\partial \theta_r \partial \theta_s} u_j(\theta) + S_{\lambda,r,s},$$

where $B \equiv B(\theta) = \partial \eta / \partial \theta^{\mathrm{T}}$. If we use the approximation

$$-\frac{\partial^2 \ell_\lambda(\theta)}{\partial \theta \partial \theta^{\mathrm{T}}} \doteq B^{\mathrm{T}} W B + S_\lambda, \quad W = \operatorname{diag}\left\{-\mathrm{E}\left(\partial^2 \ell_j / \partial \eta_j^2\right)\right\},$$

where the diagonal matrix of second derivatives is replaced by its expectation, then

$$\begin{aligned} 0 &\doteq B^{\mathrm{T}} u(\theta) - S_\lambda \theta - (B^{\mathrm{T}} W B + S_\lambda)(\widehat{\theta}_\lambda - \theta) \\ &= B^{\mathrm{T}} u(\theta) + B^{\mathrm{T}} W B \theta - (B^{\mathrm{T}} W B + S_\lambda)\widehat{\theta}_\lambda. \end{aligned}$$

If $B^{\mathrm{T}} W B + S_\lambda$ is invertible, this gives

$$\widehat{\theta}_\lambda \doteq (B^{\mathrm{T}} W B + S_\lambda)^{-1} B^{\mathrm{T}}(u + W B \theta) = (B^{\mathrm{T}} W B + S_\lambda)^{-1} B^{\mathrm{T}} W z,$$

where $z = B\theta + W^{-1} u$, as required.

Relation with least squares

- With fixed λ, the penalized MLE

$$\widehat{\theta}_\lambda = (B^{\mathrm{T}} W B + S_\lambda)^{-1} B^{\mathrm{T}} W z$$

results from fixing θ, and then iteratively solving the minimization problem

$$\min_\theta \left\| \begin{pmatrix} W^{1/2} z \\ 0 \end{pmatrix}_{(n+d)\times 1} - \begin{pmatrix} W^{1/2} B \\ Q_\lambda \end{pmatrix}_{(n+d)\times d} \theta_{d\times 1} \right\|^2,$$

where Q_λ is a matrix square root of S_λ, i.e., $Q_\lambda^{\mathrm{T}} Q_\lambda = S_\lambda$.

- The corresponding smoothing matrix is taken to be

$$H_\lambda = B(B^{\mathrm{T}} W B + S_\lambda)^{-1} B^{\mathrm{T}} W,$$

and the effective degrees of freedom for a smooth component are defined as the sum of the corresponding diagonal elements of

$$P_\lambda = (B^{\mathrm{T}} W B + S_\lambda)^{-1} B^{\mathrm{T}} W B,$$

with both H_λ and P_λ evaluated at the final step of the iteration.

Approaches to iteration

- Having chosen how to choose λ for fixed θ, there are two main algorithms:
  - **performance iteration** — repeat { fix λ, update θ with one step of PIWLS, update λ } to convergence;
  - **outer iteration** — repeat { fix λ, iterate PIWLS to convergence, update λ } to convergence.
- Performance iteration
  - can be faster,
  - but since the objective function for θ changes at each step, it may not converge—especially in the context of **concurvity** (collinearity for curves . . .), when two or more smooth functions are (almost) confounded.
- Outer iteration
  - is computationally more burdensome,
  - but will converge to a (local) optimum.

Choice of λ

- The choice of λ can be based on the marginal density of y,

$$f(y;\beta,\lambda) = \int f(y \mid b;\beta) f(b;\lambda)\,\mathrm{d}b,$$

  which has no closed form in general (but is Gaussian if both fs are Gaussian).

- Various ways to approximate the integral:
  - quadrature (doesn't work well when $\dim(b)$ is high);
  - simulation (e.g., importance sampling, same problems as quadrature);
  - Laplace approximation;
  - use the EM algorithm to avoid approximating the integral.
- We focus on Laplace approximation that provides well-founded inference if the number of random effect increases at most at rate $n^{1/3}$.

Laplace approximation

Lemma 21

Let $h(u)$ be a smooth convex function defined for $u \in \mathbb{R}^d$, with a minimum at $u = \tilde{u}$, where $\partial h(\tilde{u})/\partial u = 0$ and the matrix of partial derivatives $h_2 \equiv \partial^2 h(\tilde{u})/\partial u \partial u^{\mathrm{T}}$ is positive definite, and let

$$I_n = \int_{\mathbb{R}^d} e^{-nh(u)} \, \mathrm{d}u.$$

Then $I_n = \tilde{I}_n \left\{1 + O(n^{-1})\right\}$, and its **Laplace approximation** *is*

$$\tilde{I}_n = \frac{(2\pi)^{d/2}}{|nh_2|^{1/2}} e^{-nh(\tilde{u})}.$$

- For marginal density approximation we let $\theta = (\beta^{\mathrm{T}}_{p\times 1}, b^{\mathrm{T}}_{q\times 1})^{\mathrm{T}} \sim \mathcal{N}_d(0, S_\lambda^-)$, and write

$$f(y;\beta,\lambda) = \int f(y;\theta) f(\theta;\lambda) \, \mathrm{d}\theta = \frac{|S_\lambda|_+^{1/2}}{(2\pi)^{d/2}} \int \exp\left\{\ell_\lambda(\theta)\right\} \, \mathrm{d}\theta,$$

where β is unpenalised, $|S_\lambda|_+$ is the product of the non-negative eigenvalues of S_λ, and

$$\ell_\lambda(\theta) = \ell(\theta) - \tfrac{1}{2}\theta^{\mathrm{T}} S_\lambda \theta = O(n);$$

the assumptions of Lemma 21 should be satisfied by $h(u) \equiv -n^{-1}\ell_\lambda(\theta)$.

Proof of Lemma 21 (exercise)

- Close to $\tilde{u}$ a Taylor series expansion gives

  $$h(u) \doteq h(\tilde{u}) + h'(\tilde{u})^{\mathrm{T}}(u-\tilde{u}) + \tfrac{1}{2}(u-\tilde{u})^{\mathrm{T}} h''(\tilde{u})(u-\tilde{u}) = h(\tilde{u}) + \tfrac{1}{2}(u-\tilde{u})^{\mathrm{T}} h_2 (u-\tilde{u})$$

  so if we set $z = (nh_2)^{1/2}(u-\tilde{u})$ then $u = \tilde{u} + (nh_2)^{1/2} z$, $\mathrm{d}u/\mathrm{d}z = (nh_2)^{-1/2}$, and arguing heuristically (ignoring the third and higher terms),

  $$\begin{aligned} I_n &\doteq e^{-nh(\tilde{u})} \int e^{-n(u-\tilde{u})^{\mathrm{T}} h_2 (u-\tilde{u})/2} \, \mathrm{d}u \\ &= e^{-nh(\tilde{u})} \int_{-\infty}^{\infty} e^{-z^2/2} \frac{\mathrm{d}u}{\mathrm{d}z} \, \mathrm{d}z \\ &= \left(\frac{(2\pi)^d}{|nh_2|} \right)^{1/2} e^{-nh(\tilde{u})}, \end{aligned}$$

  because the d-dimensional normal density has unit integral.

- A more detailed accounting is needed to get the error term. Take the scalar case ($d = 1$) for simplicity. We start by writing

$$\begin{aligned} nh(u) &\doteq nh(\tilde{u}) + \tfrac{1}{2}nh_2(u-\tilde{u})^2 + \tfrac{1}{6}nh_3(u-\tilde{u})^3 + \tfrac{1}{24}nh_4(u-\tilde{u})^4 + \cdots \\ &= nh(\tilde{u}) + \tfrac{1}{2}z^2 + \tfrac{1}{6}\frac{h_3/h_2^{3/2}}{n^{1/2}}z^3 + \tfrac{1}{24}\frac{h_4/h_2^2}{n}z^4 + O(n^{-3/2}) \\ &= nh(\tilde{u}) + \tfrac{1}{2}z^2 + \frac{A}{n^{1/2}}z^3 + \frac{B}{n}z^4 + O(n^{-3/2}) \end{aligned}$$

say. Hence

$$\begin{aligned} e^{-nh(u)} &= e^{-nh(\tilde{u})-\frac{1}{2}z^2}\left\{1 - \frac{A}{n^{1/2}}z^3 - \frac{B}{n}z^4 + \tfrac{1}{2}\left(-\frac{A}{n^{1/2}}z^3 - \frac{B}{n}z^4\right)^2 + O(n^{-3/2})\right\} \\ &= e^{-nh(\tilde{u})-\frac{1}{2}z^2}\left\{1 - \frac{A}{n^{1/2}}z^3 - \frac{B}{n}z^4 + \tfrac{1}{2}\frac{A^2}{n}z^6 + O(n^{-3/2})\right\}. \end{aligned}$$

- As the odd moments of the normal density are zero, integration with respect to z leaves only the n^{-1} term and the next remaining term is $O(n^{-2})$. The fourth and sixth moments of the standard normal distribution are respectively 3 and 15, and

$$15A^2/2 - 3B = 15(h_3/h_2^{3/2}/6)^2/2 - 3\{h_4/(24h_2)\} = \frac{15h_3^2}{72h_2^3} - \frac{h_4}{8h_2^2} = \frac{5h_3^2}{24h_2^3} - \frac{h_4}{8h_2^2},$$

as required. The same argument works for $m > 1$.

Comments on Laplace approximations

- The $O(1/n)$ error is relative, so the approximation is often surprisingly accurate;
- since the odd moments of the normal density are all zero, the expansion has only terms whose orders are even powers of $n^{-1/2}$, i.e., $n^{-1}, n^{-2}, \ldots$;
- $\tilde{I}_n$ involves only h and the hessian matrix h_2 at $\tilde{u}$, so is easily found, numerically if necessary;
- the series is asymptotic, so the partial sums may not converge, and including additional terms may not be useful;
- as most of the normal probability lies within ± 3 standard deviations of the mean, the limits of the integral are almost irrelevant provided they are far enough away from $\tilde{u}$;
- if

$$I_n = \int_{-\infty}^{\infty} e^{-nh(u)}\,\mathrm{d}u, \quad J_n = \int_{-\infty}^{\infty} e^{-nh^*(u)}\,\mathrm{d}u,$$

where $h^*(u) = h(u) + O(n^{-1})$, then

$$(I_n/J_n) \div (\tilde{I}_n/\tilde{J}_n) = 1 + O(n^{-2}),$$

so two Laplace approximations can be better than one.

Approximate REML

- Laplace approximation gives the approximate restricted log likelihood

$$\ell_{\rm p}(\lambda) \equiv \tfrac{1}{2}\log|S_\lambda|_+ - \tfrac{1}{2}\log|B^{\rm T}WB^{\rm T} + S_\lambda| + \ell(\widehat{\theta}_\lambda) - \tfrac{1}{2}\widehat{\theta}_\lambda^{\rm T} S_\lambda \widehat{\theta}_\lambda + O_p(n^{-1}),$$

  where $O_p(n^{-1})$ is a (random) term of order n^{-1} and

$$\widehat{\theta}_\lambda = (B^{\rm T}WB + S_\lambda)^{-1}B^{\rm T}Wz$$

  results from iterating PIWLS to convergence for fixed λ and satisfies $\partial\ell_\lambda(\widehat{\theta}_\lambda)/\partial\theta = 0$.

- The expression for $\widehat{\theta}_\lambda$ contains

$$B \equiv B(\widehat{\theta}_\lambda), \quad W \equiv W(\widehat{\theta}_\lambda), \quad z = B(\widehat{\theta}_\lambda)\widehat{\theta}_\lambda + W^{-1}(\widehat{\theta}_\lambda)u(\widehat{\theta}_\lambda),$$

  which involve the first two derivatives of the log likelihood contributions ℓ_j.

- Newton–Raphson maximization of $\ell_{\rm p}(\lambda)$ requires its first two derivatives, so we need

$$\frac{\partial\widehat{\theta}_\lambda}{\partial\lambda}, \quad \frac{\partial^2\widehat{\theta}_\lambda}{\partial\lambda\partial\lambda^{\rm T}},$$

  which will involve the third and fourth derivatives of the ℓ_j . . . could be painful.

- A version of this is implemented in mgcv.

UK monthly AIDS reports 1983–1992

Diagnosis period		Reporting-delay interval (quarters):									Total reports to end
Year	Quarter	0†	1	2	3	4	5	6	···	≥14	of 1992
		.	.	.	.	.	.	.	.	.	.
		.	.	.	.	.	.	.	.	.	.
		.	.	.	.	.	.	.	.	.	.
1988	1	31	80	16	9	3	2	8	···	6	174
	2	26	99	27	9	8	11	3	···	3	211
	3	31	95	35	13	18	4	6	···	3	224
	4	36	77	20	26	11	3	8	···	2	205
1989	1	32	92	32	10	12	19	12	···	2	224
	2	15	92	14	27	22	21	12	···	1	219
	3	34	104	29	31	18	8	6	···		253
	4	38	101	34	18	9	15	6	···		233
1990	1	31	124	47	24	11	15	8	···		281
	2	32	132	36	10	9	7	6	···		245
	3	49	107	51	17	15	8	9	···		260
	4	44	153	41	16	11	6	5	···		285
1991	1	41	137	29	33	7	11	6	···		271
	2	56	124	39	14	12	7	10	···		263
	3	53	175	35	17	13	11	2			306
	4	63	135	24	23	12	1				258
1992	1	71	161	48	25	5					310
	2	95	178	39	6						318
	3	76	181	16							273
	4	67	66								133

AIDS data

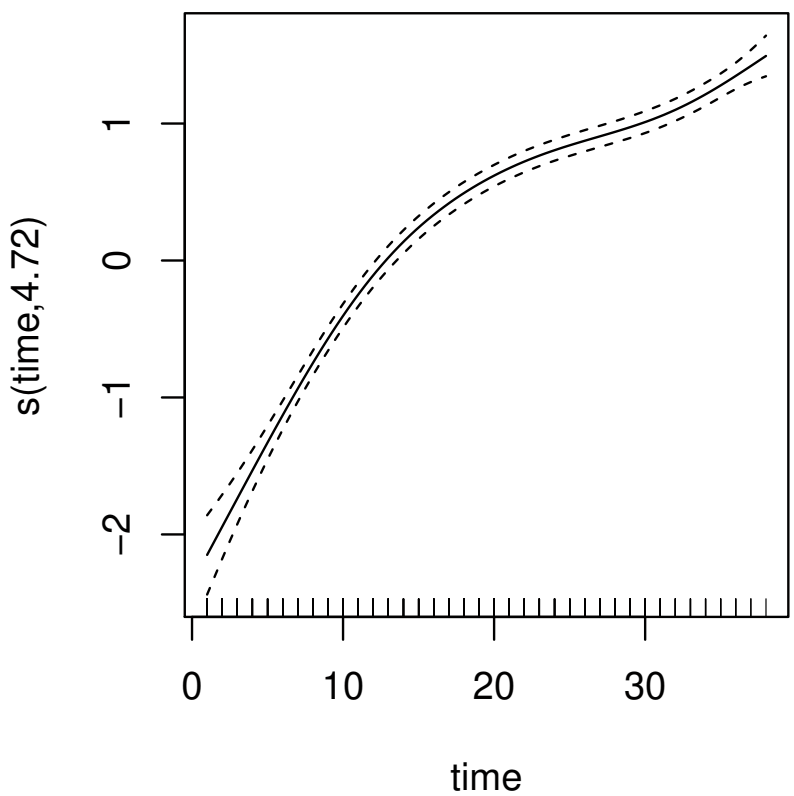
- Chain-ladder model: number of reports in row j and column k is Poisson, with mean
$$\mu_{jk} = \exp(\alpha_j + \beta_k),$$
but
  - why should there be different parameters α_j and β_k for every row and column?
  - Wouldn't smooth variation be more plausible?
- Better models (maybe?):
$$\mu_{jk} = \exp\{s(j) + \beta_k\}, \quad \mu_{jk} = \exp\{s(j) + s(k)\},$$
where the time effect $s(j)$ and the delay effect $s(k)$ vary smoothly.
- Should also account for the overdispersion ...

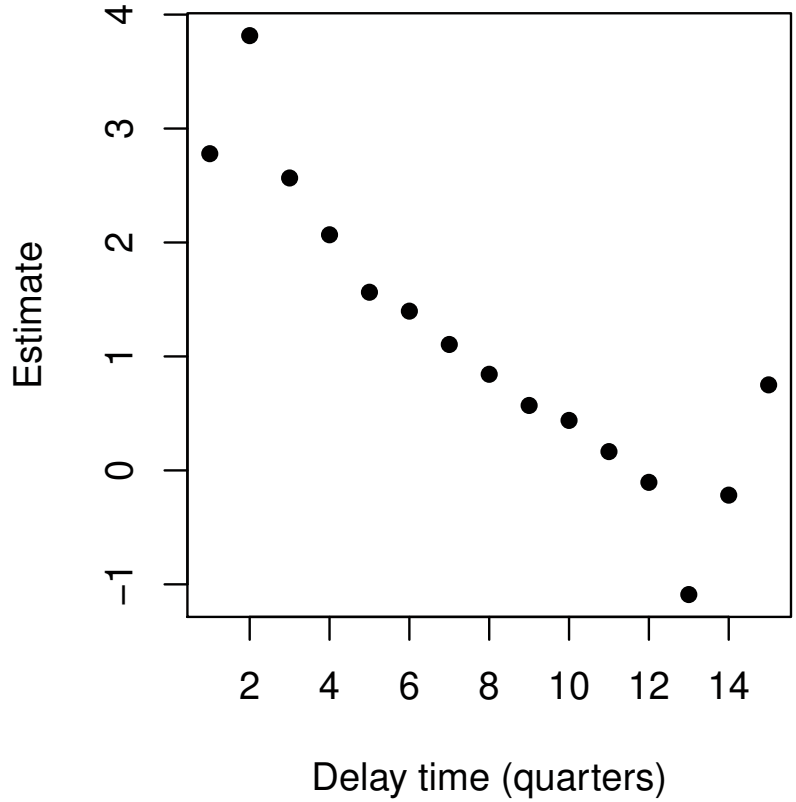
Example: AIDS data

```
library(mgcv); library(boot)
data(aids)
aids.in <- aids[c(1:570)[as.logical(1-aids$dud)],] # these are elements in the two-way table with data, dud=1 has no data
aids.glm <- glm(y~factor(time)+factor(delay),family=quasipoisson,data=aids.in)
aids.gam1 <- mgcv::gam(y~s(time,k=20)+factor(delay)-1,family=quasipoisson,data=aids.in)
plot(aids.gam1,page=1)
> anova(aids.gam1)

Formula:
y ~ s(time, k = 20) + factor(delay)

Parametric Terms:
              df     F p-value
factor(delay) 14 261.6  <2e-16

Approximate significance of smooth terms:   # Ref.df can be ignored
          edf Ref.df     F p-value
s(time) 4.891  6.129 189.1  <2e-16

aids.gam2 <- mgcv::gam(y~s(time,k=20)+s(delay,k=15),family=quasipoisson,data=aids.in)
> anova(aids.gam2)

Formula:
y ~ s(time, k = 20) + s(delay, k = 15)

Approximate significance of smooth terms:
            edf Ref.df     F p-value
s(time)   4.896  6.134 189.0  <2e-16
s(delay) 11.453 12.754 285.5  <2e-16
```

The fits are very similar, but aids.gam2 has slightly lower AIC of 792.0 compared to 792.1 — these are so similar that the choice should be based on interpretability rather than on AIC.

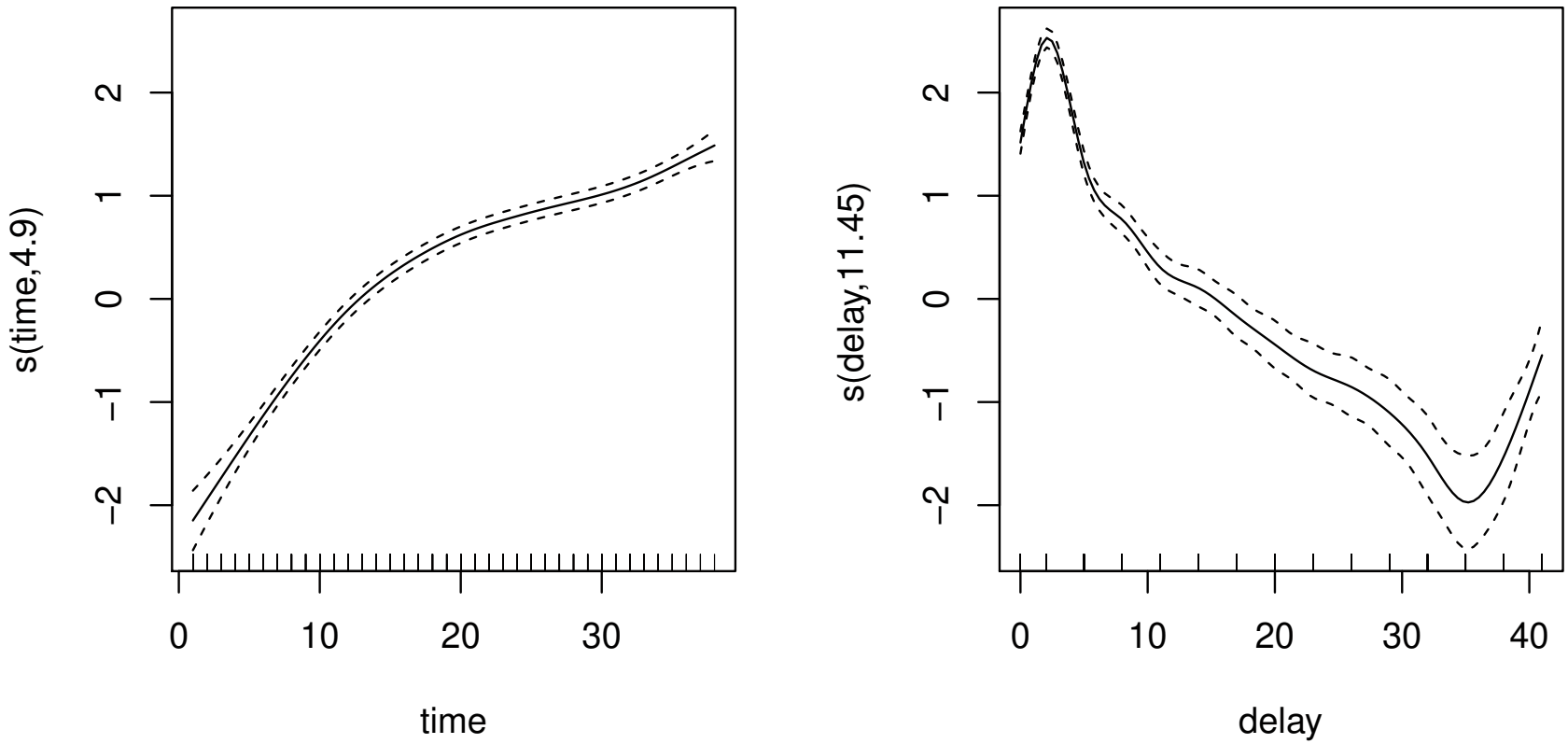
Example: AIDS data





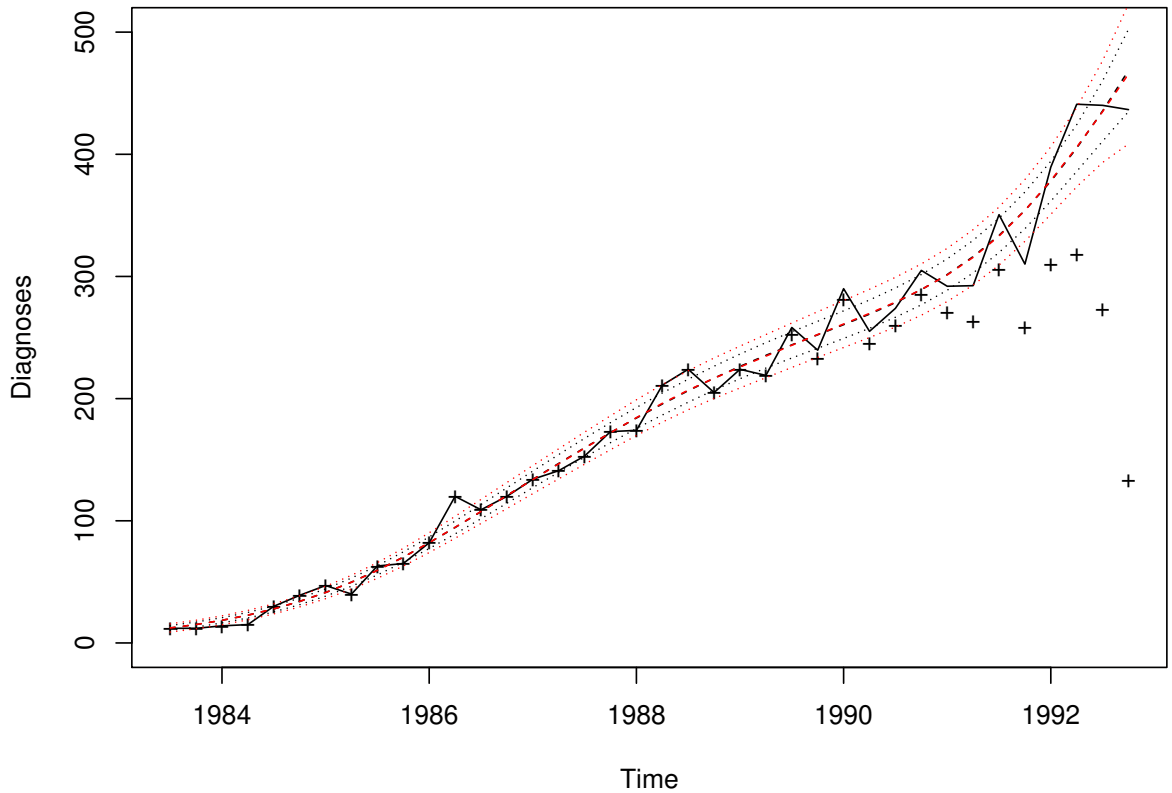
Estimates of (centered) smooth functions $s(j)$ and parameters based on plot(aids.gam1) and coef(aids.gam1).

Example: AIDS data



Estimates of (centered) smooth functions $s(j)$ and $s(k)$ based on `plot(aids.gam2)`.

Example: AIDS data



Numbers of recorded deaths (+), with estimated mean deaths per quarter based on chain-ladder model (solid) and on Poisson (black dashes) and quasi-likelihood GAMs with Poisson variance function $V(\mu) = \mu$ (red dashes). The last two estimates have 95% pointwise confidence intervals (dots) based on the fit (treating the smoothing parameters as fixed). To make these I had to compute the fitted means for the missing lower right triangle of the data table.

Closing

- The basic ideas of regression, dependence of a response on explanatory variables, extend far beyond the linear model, to
  - non-linear dependence on explanatory variables;
  - general response distributions (Poisson, binomial, ...);
  - random effects models—some parameters treated as random, and others as fixed;
  - smooth curve fitting by basis function methods in (generalized) additive models.
- Unifying themes are:
  - (semi-)parametric modelling using basis functions;
  - maximum likelihood inference;
  - estimation using iterative weighted least squares algorithms;
  - penalized fitting to allow for random effects/basis functions;
  - analysis of deviance;
  - residuals and other diagnostics.