3.5 General Framework

slide 239

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,
  - random effects (later).
- 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

$$y = B\theta + \varepsilon = \begin{pmatrix} X_1 & X_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} Z_1(x) & Z_2(z) \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + \varepsilon,$$

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 193 with just one smooth term we defined

$$\mathrm{edf}_\lambda = \mathrm{tr}(H_\lambda) = \mathrm{tr}(P_\lambda) = \sum_{r=1}^{p+q} P_{\lambda,rr} \in (p, p+q),$$

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** edf_{λ_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
  - $\mathrm{CV}(\lambda)$ or $\mathrm{GCV}(\lambda)$ (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 slide 191,

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

  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 of (21) 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**

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

  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: Average coverage probability

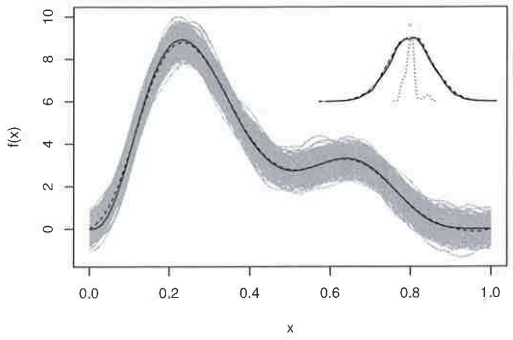
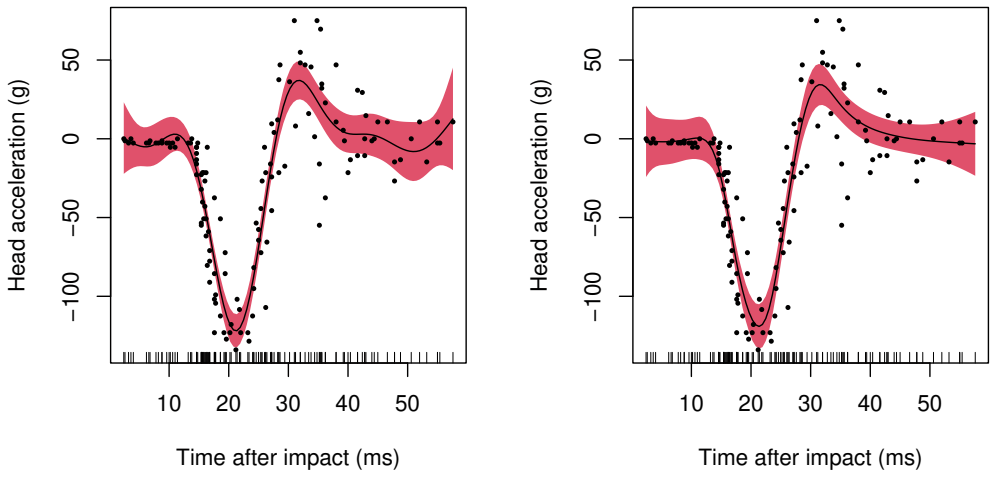


Figure 6.7 *The Nychka (1988) idea. The main black curve shows a true function $f(x)$, while the grey curves show 500 replicate spline estimates $\hat{f}(x)$. The dashed curve is $\mathbb{E}\hat{f}(x)$. Inset at top* ***right are*** *scaled* ***kernel smooth*** *estimates of the distributions of the sampling error, $\hat{f} - \mathbb{E}\hat{f}$ (****continuous*** *black); the* ***bias,*** *$\mathbb{E}\hat{f} - f$, evaluated at a random x (dotted) and $\hat{f} - f$ evaluated at a random x (dashed). In grey is the normal approximation to the dashed curve. Evaluation at a random x turns the bias into a random variable, which has substantially lower variance than the approximately normal $\hat{f} - \mathbb{E}\hat{f}$. Hence the sum of the randomized bias and sampling error is approximately normally distributed. The variance of this sum turns out to be well approximated by the Bayesian posterior covariance for f.*

(Wood, 2017)

Example: Motorcycle data

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

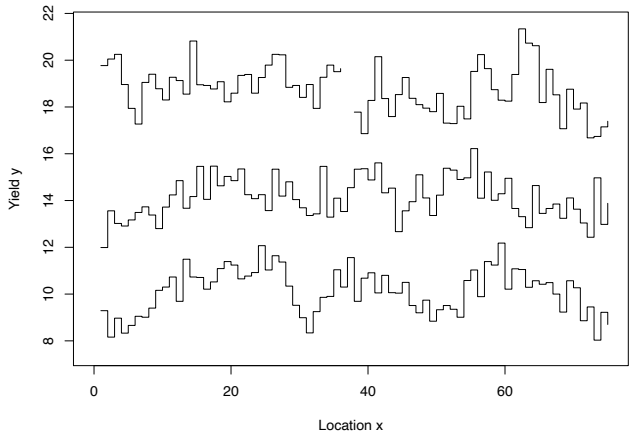


Example: Spring barley data

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

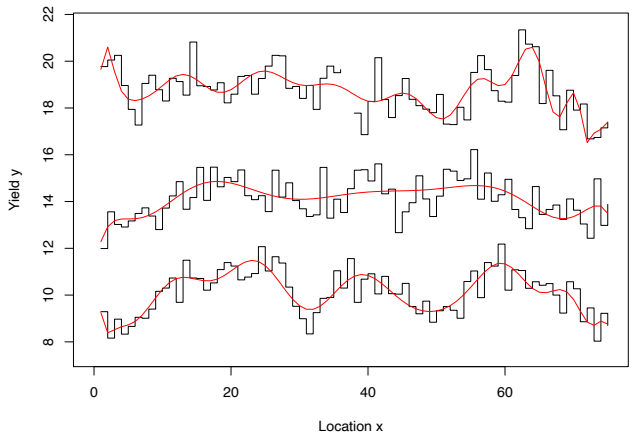
Location x	Block 1		Block 2		Block 3	
	Variety	Yield y	Variety	Yield y	Variety	Yield y
1	57	9.29	49	7.99	63	11.77
2	39	8.16	18	9.56	38	12.05
3	3	8.97	8	9.02	14	12.25
4	48	8.33	69	8.91	71	10.96
5	75	8.66	29	9.17	22	9.94
6	21	9.05	59	9.49	46	9.27
7	66	9.01	19	9.73	6	11.05
8	12	9.40	39	9.38	30	11.40
9	30	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



Yield as a function of location for the three blocks, with yields for blocks 2 and 3 offset by the addition of 4 and of 7 respectively. Value 37 in block 3 is missing.

Spring barley data and polynomial fits



Yield as a function of location for the three blocks, with yields for blocks 2 and 3 offset by the addition of 4 and of 7 respectively, with fitted polynomials of degrees 20, 10 and 50.

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 255).

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)
```

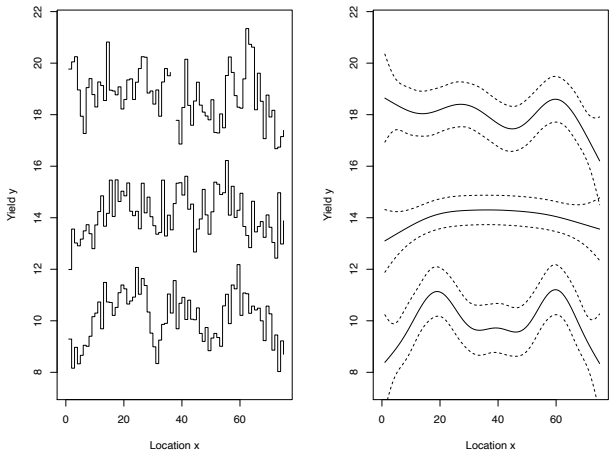

Example: Spring barley data

- Using GCV the smooths have $\mathrm{df}_\lambda = 8.3$, 6.8, 6.3, with $\widehat{\sigma} = 0.65$ and $\mathrm{AIC} = 513.1$, 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 $\mathrm{AIC} = 518.3$, 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

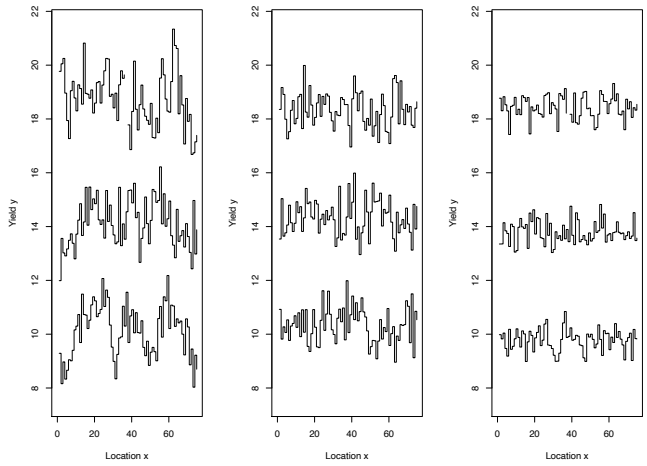
- The $P_{\lambda,rr}$ need not be positive, though their total for each smooth is positive.
- In applications it would be wise to check whether increasing q_m would lead to very different fits.

Example: Spring barley data



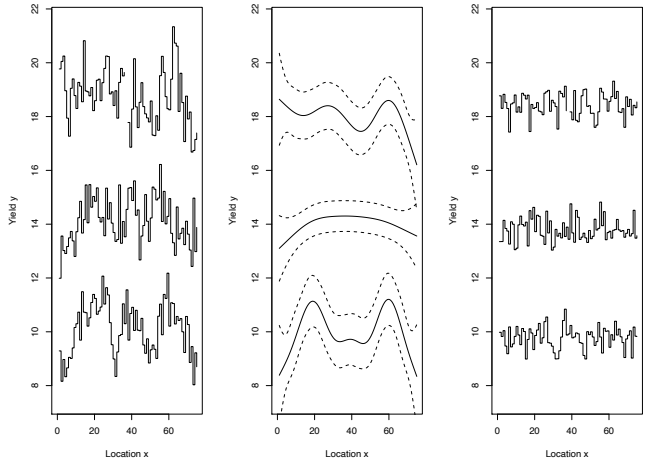
Left: data (offset by adding 4 and 8 to blocks 2 and 3).
Right: estimated fertility patterns (with estimated df 7.2, 3, 6.1) and 95% unconditional pointwise confidence intervals, fitted using REML. The intervals are wider for blocks 1 and 3.

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

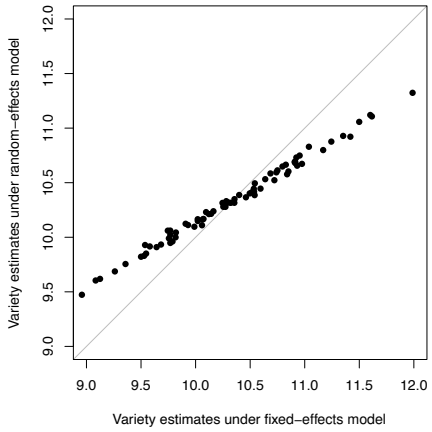
$$S_\lambda = \lambda_\beta S_\beta + \lambda_1 S_1 + \lambda_2 S_2 + \lambda_3 S_3,$$

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.

3.6 Components of Variance

Background and motivation

- ☐ All the models so far have involved just one level of randomness, corresponding to 'measurement error' on individual responses.
- ☐ Complex layering of randomness can arise in applications, and then conclusions may depend on how it is dealt with.
- ☐ Two conceptually different set-ups (which may give the same models):
  - observational/experimental setup generates several layers of randomness;
  - we find it useful to treat the parameters of some model as drawn from a distribution.

  The first concerns logical properties of the data, whereas the second is a modelling assumption.

Example: Blood pressure

- ☐ Blood pressure data: $P = 25$ patients each made $V = 16$ visits to a clinic, and on each occasion their systolic and diastolic blood pressures were measured twice.
- ☐ Consider just the diastolic pressure. We expect there to be variation
  - between patients,
  - between visits within patients, and
  - between measurements within visits,

  which we could model as

$$y_{pvm} = \mu + b_p + e_{pv} + \varepsilon_{pvm}, \quad p = 1, \ldots, P, v = 1, \ldots, V, m = 1, \ldots, M,$$

  where
  - μ is the population mean diastolic blood pressure (DBP),
  - b_p is the difference between the patient and population mean DBP,
  - e_{pv} is the difference between this and the mean DBP on the vth visit, and
  - ε_{pvm} is the difference between the mean DBP for the pth patient at the vth visit and the mth measurement on that visit,

  and

$$b_p \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_b^2) \perp\!\!\!\perp e_{pv} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_e^2) \perp\!\!\!\perp \varepsilon_{pvm} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2).$$

Example: Blood pressure

patno	patient	visno	dbp1	dbp2	sbp1	sbp2
1307	1	7	95	85	150	130
1307	1	8	85	85	140	140
1307	1	9	90	90	150	150
1307	1	10	80	80	135	135
1307	1	11	80	80	130	125
1307	1	12	85	85	150	155
.						
.						
1307	1	19	80	80	130	130
1307	1	20	80	80	140	140
1307	1	21	90	85	145	140
1307	1	22	75	75	130	130
1418	2	7	104	106	160	148
1418	2	8	98	104	158	162
.						
.						
9202	25	21	91	90	142	139
9202	25	22	80	78	162	160

Fixed and random effects

Chimpanzee	Word									
	1	2	3	4	5	6	7	8	9	10
1	178	60	177	36	225	345	40	2	287	14
2	78	14	80	15	10	115	10	12	129	80
3	99	18	20	25	15	54	25	10	476	55
4	297	20	195	18	24	420	40	15	372	190

- Times (min) for four chimpanzees to learn each of ten words.
- A possible model for log time is

$$y_{cw} \mid \alpha_c, \beta_w \overset{\text{ind}}{\sim} \mathcal{N}(\mu + \alpha_c + \beta_w, \sigma^2), \quad c = 1, \ldots, C = 4, w = 1, \ldots, W = 10.$$

- The α_c and/or the β_w would be considered as constant **fixed effects** if we were interested in the relative linguistic abilities of these particular chimps and/or if we planned further tests with these particular words.
- Either (or both) of the α_c and β_w might be considered to be **random effects** if they were thought to be sampled from a larger population whose variation is of interest.

Two distinctions

- ☐ We distinguish **fixed** and **random** effects (above).
- ☐ We distinguish **nested** and **crossed** effects:
  - in the blood pressure data, replicate measurements at each visit are **nested** within visit, because there is no logical connection between $y_{p,v_1,1}$ and $y_{p,v_2,1}$ (we could permute the final index m within each patient/visit combination without changing the data structure). Likewise if we ignore any possible time effects between visits, we could consider that visits are nested within patients;
  - in the chimp data, the effects are **crossed**, because permuting chimps or words would entail permuting entire rows or columns of the data table: there is a logical connection between y_{c_1w} and y_{c_2w}, and between y_{cw_1} and y_{cw_2};
- ☐ In R syntax, with patient and visit number declared as factors, for nested effects we write

```
y ~ patient/visno
```

read as 'separate effects for visit number within the levels of patient' and for crossed effects with chimp and word declared as factors we write

```
y ~ chimp + word
```


Nested model ANOVA

- ☐ For the nested model

$$y_{pvm} = \mu + b_p + e_{pv} + \varepsilon_{pvm}, \quad p = 1, \ldots, P, v = 1, \ldots, V, m = 1, \ldots, M,$$

and with a dot and bar denoting averaging over that index, we write

$$y_{pvm} - \overline{y}_{...} = y_{pvm} - \overline{y}_{pv\cdot} + \overline{y}_{pv\cdot} - \overline{y}_{p\cdot\cdot} + \overline{y}_{p\cdot\cdot} - \overline{y}_{...},$$

and note that

$$\begin{aligned} y_{pvm} - \overline{y}_{pv\cdot} &= \varepsilon_{pvm} - \overline{\varepsilon}_{pv\cdot}, \\ \overline{y}_{pv\cdot} - \overline{y}_{p\cdot\cdot} &= e_{pv} + \overline{\varepsilon}_{pv\cdot} - (\overline{e}_{p\cdot} + \overline{\varepsilon}_{p\cdot\cdot}), \\ \overline{y}_{p\cdot\cdot} - \overline{y}_{...} &= b_p + \overline{e}_{p\cdot} + \overline{\varepsilon}_{p\cdot\cdot} - (\overline{b}_{\cdot} + \overline{e}_{\cdot\cdot} + \overline{\varepsilon}_{...}), \end{aligned}$$

so the overall sum of squares is

$$\begin{aligned} \sum_{p,v,m} (y_{pvm} - \overline{y}_{...})^2 &= \sum_{p,v,m} (y_{pvm} - \overline{y}_{pv\cdot})^2 + \sum_{p,v,m} (\overline{y}_{pv\cdot} - \overline{y}_{p\cdot\cdot})^2 + \sum_{p,v,m} (\overline{y}_{p\cdot\cdot} - \overline{y}_{...})^2 \\ &= \sum_{p,v,m} (y_{pvm} - \overline{y}_{pv\cdot})^2 + M \sum_{p,v} (\overline{y}_{pv\cdot} - \overline{y}_{p\cdot\cdot})^2 + VM \sum_{p} (\overline{y}_{p\cdot\cdot} - \overline{y}_{...})^2, \end{aligned}$$

where these terms are independent sums of squares for variables that are

$$\mathcal{N}(0, \sigma^2), \quad \mathcal{N}(0, \sigma_e^2 + \sigma^2/M), \quad \mathcal{N}\{0, \sigma_b^2 + \sigma_e^2/V + \sigma^2/(VM)\}.$$

Nested model ANOVA II

- Hence

$$\begin{aligned}\sum_{p,v,m}(y_{pvm}-\overline{y}_{pv\cdot})^2 &\sim \sigma^2\chi^2_{PV(M-1)},\\ \sum_{p,v,m}(\overline{y}_{pv\cdot}-\overline{y}_{p\cdot\cdot})^2 &\sim M(\sigma_e^2+\sigma^2/M)\chi^2_{P(V-1)} \stackrel{\mathrm{D}}{=} (M\sigma_e^2+\sigma^2)\chi^2_{P(V-1)},\\ \sum_{p,v,m}(\overline{y}_{p\cdot\cdot}-\overline{y}_{\cdot\cdot\cdot})^2 &\sim VM\left(\sigma_b^2+\tfrac{\sigma_e^2}{V}+\tfrac{\sigma^2}{VM}\right)\chi^2_{P-1} \stackrel{\mathrm{D}}{=} (VM\sigma_b^2+M\sigma_e^2+\sigma^2)\chi^2_{P-1},\end{aligned}$$

  and we can estimate the components of variance σ^2, σ_e^2 and σ_b^2 from the ANOVA table.
- The interpretation of the ANOVA depends on whether we regard $\delta_b^2=\sum_p(b_p-\overline{b}_\cdot)^2$ and $\delta_e^2=\sum_{p,v}(e_{pv}-\overline{e}_{p\cdot})^2$ as random or fixed:

Term	df	Sum of squares	E(Mean square) when terms below random		
			ε	ε, e	ε, e, b
Between patients	$P-1$	$\sum(\overline{y}_{p\cdot\cdot}-\overline{y}_{\cdot\cdot\cdot})^2$	$VM\delta_b^2+M\delta_e^2+\sigma^2$	$VM\delta_b^2+M\sigma_e^2+\sigma^2$	$VM\sigma_b^2+M\sigma_e^2+\sigma^2$
Between visits within patients	$P(V-1)$	$\sum(\overline{y}_{pv\cdot}-\overline{y}_{p\cdot\cdot})^2$	$M\delta_e^2+\sigma^2$	$M\sigma_e^2+\sigma^2$	$M\sigma_e^2+\sigma^2$
Between measures within visits	$PV(M-1)$	$\sum(y_{pvm}-\overline{y}_{pv\cdot})^2$	σ^2	σ^2	σ^2

Nested and crossed ANOVA

- Nested analysis of the blood pressure data:

```
summary( aov(dbp ~ patient/visno, data=blood.dia) )
               Df Sum Sq Mean Sq F value Pr(>F)
patient        24  23059   960.8  124.29 <2e-16 ***
patient:visno 375  39082   104.2   13.48 <2e-16 ***
Residuals     400   3092     7.7
```

- Likewise, crossed analysis of the chimpanzee data:

```
 summary( aov(log(y)~chimp+word,data=chimps) )
            Df Sum Sq Mean Sq F value  Pr(>F)
chimp        3   5.33   1.778   2.719  0.0642 .
word         9  45.69   5.077   7.765 1.5e-05 ***
Residuals   27  17.65   0.654
```

  There are $C-1$ degrees of freedom for chimps, $W-1$ for words, and $(C-1)(W-1)$ for the residual.
- In both cases, we can use the ANOVA table to estimate the variance components and then perform **synthesis of variance**: e.g., how large would W need to be to distinguish the learning abilities of two chimps with probability 0.95?

Example: Blood pressure

- Solving the equations

$$\sigma^2 = 7.7, \quad M\sigma_e^2 + \sigma^2 = 104.2, \quad VM\sigma_b^2 + M\sigma_e^2 + \sigma^2 = 960.8,$$

gives (in units of millimeters of mercury, mmHg)

$$\widehat{\sigma} = 2.8, \quad \widehat{\sigma}_e = 6.9, \quad \widehat{\sigma}_b = 5.2,$$

so the largest variation is between different visits within patients, while that between measurements on a single visit is smallest.

- Different comparisons require appropriate baseline variances:
  - if we are interested in how patient p's response varies from visit to visit, we use

$$\overline{y}_{pv_1\cdot} - \overline{y}_{pv_2\cdot} = \mu + b_p + e_{pv_1} + \overline{\varepsilon}_{pv_1\cdot} - (\mu + b_p + e_{pv_2} + \overline{\varepsilon}_{pv_2\cdot}) \sim \mathcal{N}(0, 2\sigma_e^2 + 2\sigma^2/M),$$

as a basis for a test of a significant difference, whereas to compare average blood pressures for two different patients we use

$$\overline{y}_{p_1\cdot\cdot} - \overline{y}_{p_2\cdot\cdot} = b_{p_1} + \overline{e}_{p_1\cdot} + \overline{\varepsilon}_{p_1\cdot\cdot} - (b_{p_2} + \overline{e}_{p_2\cdot} + \overline{\varepsilon}_{p_2\cdot\cdot}) \sim \mathcal{N}\{0, 2\sigma_b^2 + 2\sigma_e^2/V + 2\sigma^2/(VM)\}.$$

- **Split-unit designs** are set up to make the most important comparisons within units (here patients) and less important ones between units, and the ANOVA reflects this.

General form

- We could have written the nested model above as

$$y = 1_n\mu + X_b b + X_e e + \varepsilon,$$

with design matrices X_b and X_e for the patient and visit-within-patient effects.

- Then if
  - b and e are treated as fixed (ordinary parameters),

$$y \sim \mathcal{N}_n(1_n\mu + X_b b + X_e e, \sigma^2 I_n),$$

  - b is treated as fixed but $e \sim \mathcal{N}_{PV}(0, \sigma_e^2 I_{PV})$, then

$$y \sim \mathcal{N}_n(1_n\mu + X_b b, \sigma_e^2 X_e X_e^{\mathrm{T}} + \sigma^2 I_n),$$

  - and if $b \sim \mathcal{N}_P(0, \sigma_b^2 I_P)$ independent of $e \sim \mathcal{N}_{PV}(0, \sigma_e^2 I_{PV})$, then

$$y \sim \mathcal{N}_n(1_n\mu, \sigma_b^2 X_b X_b^{\mathrm{T}} + \sigma_e^2 X_e X_e^{\mathrm{T}} + \sigma^2 I_n).$$

- Hence random e or b give patterned covariance matrices depending on their variances.

Summary

- ☐ Components of variance ANOVA is easily performed directly for balanced data.
- ☐ Standard ANOVA tables have different interpretations, depending on which components of variance are taken to be random or fixed.
- ☐ Extensions are needed to deal with more complex settings, with unbalanced data, or with non-linear or non-normal errors — hence **mixed models**, i.e., models with both random and fixed parts, arising in many different settings (and with different names):
  - components of variance (as above),
  - classical experimental design (split-plot designs, ...),
  - repeated measures,
  - longitudinal models,
  - multi-level models,
  - hierarchical models.
- ☐ Can subsume linear versions into the **linear mixed model**, which can be extended to nonlinear models, GLMs, ...

3.7 Linear Mixed Model

Linear mixed model

- ☐ The **linear mixed model** may be written as

$$y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + Z_{n\times q}b_{q\times 1} + \varepsilon_{n\times 1}, \quad b \sim N_q(0,\Omega_b), \quad \varepsilon \sim \mathcal{N}_n(0,\Omega),$$

  where
  - β represents the **fixed effects**,
  - b represents the **random effects**, and
  - usually $\Omega = \sigma^2 I_n$.
- ☐ This has the same structure as when smoothing, with the columns of Z giving the structure of the random effects.
- ☐ Equivalently,

$$y \mid b \;\sim\; \mathcal{N}_n(X\beta + Zb, \Omega), \quad b \;\sim\; \mathcal{N}_q(0,\Omega_b),$$

  which gives marginal response distribution

$$y \sim \mathcal{N}_n(X\beta, Z\Omega_b Z^{\mathrm{T}} + \Omega), \quad Z\Omega_b Z^{\mathrm{T}} + \Omega = \sigma^2\Delta^{-1}(\psi),$$

  say, with ψ the vector of distinct variance ratios appearing in Δ^{-1} (e.g., $\sigma_b^2/\sigma^2, \ldots$).
- ☐ Although Ω is often diagonal, $Z\Omega_b Z^{\mathrm{T}}$ is not, so inverting $Z\Omega_b Z^{\mathrm{T}} + \Omega$ involves $O(n^3)$ flops in general, and we should avoid working with Δ.

Maximum likelihood estimation

☐ Let $\tilde{b}$ denote the MLE of b for fixed β (and ψ). Then

$$\begin{aligned} f(y;\beta,\sigma^2,\psi) &= \int f(y \mid b;\beta,\sigma^2,\psi) f(b;\sigma^2,\psi)\,\mathrm{d}b \\ &= f(y,\tilde{b};\beta,\sigma^2,\psi) \times \frac{(2\pi)^{q/2}}{|Z^{\mathrm{T}}\Omega^{-1}Z+\Omega_b^{-1}|^{1/2}} \\ &\propto \frac{f(y \mid \tilde{b};\beta,\sigma^2,\psi) f(\tilde{b} \mid \sigma^2,\psi)}{|Z^{\mathrm{T}}\Omega^{-1}Z+\Omega_b^{-1}|^{1/2}}, \end{aligned}$$

so (apart from additive constants) $-2\log f(y;\beta,\sigma^2,\psi)$ equals

$$(y-X\beta-Z\tilde{b})^{\mathrm{T}}\Omega^{-1}(y-X\beta-Z\tilde{b}) + \tilde{b}^{\mathrm{T}}\Omega_b^{-1}\tilde{b} + \log\{|\Omega||\Omega_b||Z^{\mathrm{T}}\Omega^{-1}Z+\Omega_b^{-1}|\}.$$

☐ The first two (quadratic) terms here depend on β and b, so given ψ and σ^2 we can find $\widehat{\beta}_\psi$ and $\tilde{b}(\widehat{\beta},\psi)$ explicitly, and thus obtain $\ell_{\mathrm{p}}(\psi)$.

☐ By noting that

$$f(b \mid y;\beta,\sigma^2,\psi) = f(y \mid b;\beta,\sigma^2,\psi) f(b;\sigma^2,\psi) / f(y;\beta,\sigma^2,\psi)$$

and taking logs, we obtain

$$b \mid y \sim \mathcal{N}_q\left\{\tilde{b}, (Z^{\mathrm{T}}\Omega^{-1}Z+\Omega_b^{-1})^{-1}\right\}, \quad \tilde{b} = \left(Z^{\mathrm{T}}\Omega^{-1}Z+\Omega_b^{-1}\right)^{-1} Z^{\mathrm{T}}\Omega^{-1}(y-X\beta).$$

Note on maximum likelihood estimation

☐ Suppressing the parameters β, σ^2 and ψ for now, we write the log integrand in

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

in the form

$$\log f(y,b) = \log f(y,\tilde{b}) - \tfrac{1}{2}(b-\tilde{b})^{\mathrm{T}} H(\tilde{b})(b-\tilde{b}),$$

where the linear term of the Taylor series equals zero, because it is evaluated at the maximising value $\tilde{b}$, and the given Taylor series is exact because the log likelihood is quadratic.

☐ On ignoring terms not involving b we have

$$-2\log f(y,b) = -2\log f(y \mid b) - 2\log f(b) \equiv (y-X\beta-Zb)^{\mathrm{T}}\Omega^{-1}(y-X\beta-Zb) + b^{\mathrm{T}}\Omega_b^{-1}b,$$

so

$$H(b) \equiv H = Z^{\mathrm{T}}\Omega^{-1}Z + \Omega_b^{-1}$$

does not depend on b, and thus

$$\begin{aligned} f(y) &= f(y,\tilde{b}) \int \exp\left\{-\tfrac{1}{2}(b-\tilde{b})^{\mathrm{T}} H (b-\tilde{b})\right\}\,\mathrm{d}b \\ &= f(y,\tilde{b}) \times (2\pi)^{q/2}|H|^{-1/2} = f(y,\tilde{b}) \times \frac{(2\pi)^{q/2}}{|Z^{\mathrm{T}}\Omega^{-1}Z+\Omega_b^{-1}|^{1/2}}, \end{aligned}$$

as announced; the integral equals the normalising constant for a $\mathcal{N}_q(\tilde{b}, H^{-1})$ density.

Inference on β

- Since
$$y \sim \mathcal{N}_n(X\beta, Z\Omega_b Z^{\mathrm{T}} + \Omega),$$
weighted least squares gives
$$\widehat{\beta} = \{X^{\mathrm{T}}(Z\Omega_b Z^{\mathrm{T}} + \Omega)^{-1}X\}^{-1}X^{\mathrm{T}}(Z\Omega_b Z^{\mathrm{T}} + \Omega)^{-1}y,$$
with
$$\widehat{\beta} \sim \mathcal{N}_p\left[\beta, \{X^{\mathrm{T}}(Z\Omega_b Z^{\mathrm{T}} + \Omega)^{-1}X\}^{-1}\right],$$
where in general we need $O(n^3)$ flops to invert the $n \times n$ matrix $Z\Omega_b Z^{\mathrm{T}} + \Omega$.
- For cheaper calculation of $\mathrm{var}(\widehat{\beta})$, we use the inversion formulae and obtain
$$\begin{pmatrix} \mathrm{var}(\widehat{\beta})_{p\times p} & \cdot \\ \cdot & \cdot \end{pmatrix} = \begin{pmatrix} X^{\mathrm{T}}\Omega^{-1}X & X^{\mathrm{T}}\Omega^{-1}Z \\ Z^{\mathrm{T}}\Omega^{-1}X & Z^{\mathrm{T}}\Omega^{-1}Z + \Omega_b^{-1} \end{pmatrix}^{-1}_{d\times d},$$
where $d = p + q$, which involves only $O\{nd^2\}$ flops, as Ω is usually diagonal.
- Note that $\mathrm{var}(b \mid y) = (Z^{\mathrm{T}}\Omega^{-1}Z + \Omega_b^{-1})^{-1}$ can be obtained as a by-product.
- In practice these formulae are evaluated at the MLEs $\widehat{\sigma}^2$ and $\widehat{\psi}$ and used to compute confidence intervals etc. for elements of β.

Inference on random effects

- **Conventional terminology**: we **estimate** parameters β and **predict** random variables b.
- To find the best predictor $\tilde{b}(y)$ of b we minimise
$$\mathrm{E}_{b,y}\left[\left\{\tilde{b}(y) - b\right\}^{\mathrm{T}}\left\{\tilde{b}(y) - b\right\}\right],$$
which gives $\tilde{b}(y) = \mathrm{E}(b \mid y)$, with (Woodbury formula):
$$\begin{aligned} \mathrm{E}(b \mid y) &= \left(Z^{\mathrm{T}}\Omega^{-1}Z + \Omega_b^{-1}\right)^{-1} Z^{\mathrm{T}}\Omega^{-1}\left(y - X\beta\right), \\ \mathrm{var}(b \mid y) &= \left(Z^{\mathrm{T}}\Omega^{-1}Z + \Omega_b^{-1}\right)^{-1}. \end{aligned}$$
- Replace parameters β, σ^2, ψ by estimates to get **best linear unbiased predictor (BLUP)** $\tilde{b}$ and its estimated variance.
- Residuals
$$\begin{aligned} y - X\widehat{\beta} &= Z\tilde{b} + y - X\widehat{\beta} - Z\tilde{b} \\ &= Z\tilde{b} + \left\{I_n - Z\left(Z^{\mathrm{T}}\widehat{\Omega}^{-1}Z + \widehat{\Omega}_b^{-1}\right)^{-1} Z^{\mathrm{T}}\widehat{\Omega}^{-1}\right\}\left(y - X\widehat{\beta}\right), \end{aligned}$$
split into two parts, with $Z\tilde{b}$ attributable to random effects, and the second the usual residual $y - X\widehat{\beta}$ shrunk towards zero; this estimates ε.

Note on conditional mean and variance

☐ First we write

$$\tilde{b}(y) - b = \tilde{b}(y) - \mathrm{E}(b \mid y) + \mathrm{E}(b \mid y) - b,$$

expand $\{\tilde{b}(y) - b\}^{\mathrm{T}}\{\tilde{b}(y) - b\}$ and take expectation over b conditional on y to get

$$\mathrm{E}\left[\{\tilde{b}(y) - b\}^{\mathrm{T}}\{\tilde{b}(y) - b\} \mid y\right] = \left\{\tilde{b}(y) - \mathrm{E}(b \mid y)\right\}^{\mathrm{T}} \left\{\tilde{b}(y) - \mathrm{E}(b \mid y)\right\} + \mathrm{var}(b \mid y),$$

which is minimised when $\tilde{b}(y) = \mathrm{E}(b \mid y)$. Any other choice will give a larger expectation when we take E_y, so this is optimal.

☐ To obtain $\mathrm{E}(b \mid y)$, we note that

$$\begin{pmatrix} y \\ b \end{pmatrix} \sim \mathcal{N}_{n+q}\left\{\begin{pmatrix} X\beta \\ 0 \end{pmatrix}, \begin{pmatrix} \Omega + Z\Omega_b Z^{\mathrm{T}} & Z\Omega_b \\ \Omega_b Z^{\mathrm{T}} & \Omega_b \end{pmatrix}\right\},$$

so using standard formulae for conditional normal distributions, we have

$$\begin{aligned} \mathrm{E}(b \mid y) &= \Omega_b Z^{\mathrm{T}} \left(\Omega + Z\Omega_b Z^{\mathrm{T}}\right)^{-1} (y - X\beta), \\ \mathrm{var}(b \mid y) &= \Omega_b - \Omega_b Z^{\mathrm{T}} \left(\Omega + Z\Omega_b Z^{\mathrm{T}}\right)^{-1} Z\Omega_b. \end{aligned}$$

☐ The Woodbury formula applied to the conditional variance gives

$$\mathrm{var}(b \mid y) = \left(Z^{\mathrm{T}}\Omega^{-1}Z + \Omega_b^{-1}\right)^{-1}$$

as required.

☐ For the conditional mean we apply the Woodbury formula to $(\Omega + Z\Omega_b Z^{\mathrm{T}})^{-1}$ and get

$$\begin{aligned} \mathrm{E}(b \mid y) &= \Omega_b Z^{\mathrm{T}} \left\{\Omega^{-1} - \Omega^{-1}Z\left(\Omega_b^{-1} + Z^{\mathrm{T}}\Omega^{-1}Z\right)^{-1} Z^{\mathrm{T}}\Omega^{-1}\right\} (y - X\beta) \\ &= \Omega_b \left\{I_q - Z^{\mathrm{T}}\Omega^{-1}Z\left(\Omega_b^{-1} + Z^{\mathrm{T}}\Omega^{-1}Z\right)^{-1}\right\} Z^{\mathrm{T}}\Omega^{-1}(y - X\beta) \\ &= \Omega_b \left\{\Omega_b^{-1}\left(\Omega_b^{-1} + Z^{\mathrm{T}}\Omega^{-1}Z\right)^{-1}\right\} Z^{\mathrm{T}}\Omega^{-1}(y - X\beta), \end{aligned}$$

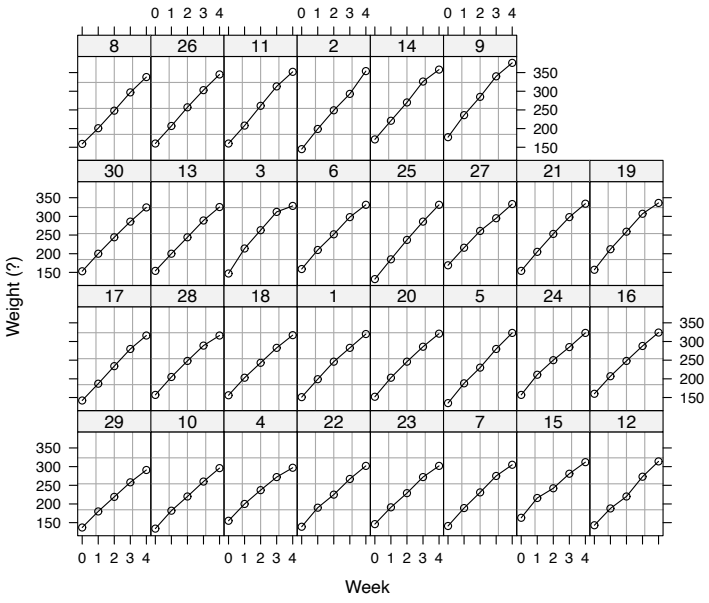
as required, where we wrote the term in braces in the second line as $I - B(A + B)^{-1} = A(A + B)^{-1}$, with $A = \Omega_b^{-1}$ and $B = Z^{\mathrm{T}}\Omega^{-1}Z$.

Example: Rat growth

Weights (units unknown) of 30 young rats over a five-week period

	Week						Week				
	1	2	3	4	5		1	2	3	4	5
1	151	199	246	283	320	16	160	207	248	288	324
2	145	199	249	293	354	17	142	187	234	280	316
3	147	214	263	312	328	18	156	203	243	283	317
4	155	200	237	272	297	19	157	212	259	307	336
5	135	188	230	280	323	20	152	203	246	286	321
6	159	210	252	298	331	21	154	205	253	298	334
7	141	189	231	275	305	22	139	190	225	267	302
8	159	201	248	297	338	23	146	191	229	272	302
9	177	236	285	340	376	24	157	211	250	285	323
10	134	182	220	260	296	25	132	185	237	286	331
11	160	208	261	313	352	26	160	207	257	303	345
12	143	188	220	273	314	27	169	216	261	295	333
13	154	200	244	289	325	28	157	205	248	289	316
14	171	221	270	326	358	29	137	180	219	258	291
15	163	216	242	281	312	30	153	200	244	286	324

Example: Rat growth



Example: Rat growth

Example 33 (Rat growth data)

☐ *Write*

$$y_{jt} = \beta_0 + b_{j0} + (\beta_1 + b_{j1})x_{jt} + \varepsilon_{jt}, \quad t = 1, \ldots 5, j = 1, \ldots, 30,$$

where the random variables (b_{j0}, b_{j1}) *have a joint normal distribution with mean vector zero and unknown variance matrix and the* $\varepsilon_{jt} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$*. In matrix terms,*

$$\begin{pmatrix} y_{j1} \\ \vdots \\ y_{j5} \end{pmatrix} = \begin{pmatrix} 1 & x_{j1} \\ \vdots & \vdots \\ 1 & x_{j5} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} 1 & x_{j1} \\ \vdots & \vdots \\ 1 & x_{j5} \end{pmatrix} \begin{pmatrix} b_{j0} \\ b_{j1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{j1} \\ \vdots \\ \varepsilon_{j5} \end{pmatrix}, \quad j = 1, \ldots, 30;$$

the overall model with $n = 150$ *is obtained by stacking these expressions.*

☐ *We set* $(x_{j1}, \ldots, x_{j5}) = (0, \ldots, 4)$*, so that* β_0 *is the mean weight in week 1.*

☐ $p = 2$ *parameters;* $q = 60$ *since two random variables per rat.*

Example: Rat growth

```
> rat.growth
    rat week   y
1     1    0 151
2     1    1 199
3     1    2 246
4     1    3 283
5     1    4 320
6     2    0 145
...
> fit.reml <- lme(fixed= y~week, random=~week|rat, data=rat.growth)
> summary(fit.reml)
Linear mixed-effects model fit by REML
 Data: rat.growth
      AIC      BIC    logLik
  1096.58 1114.563 -542.2899

Random effects:
 Formula: ~week | rat
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 10.932986 (Intr)
week         3.534747 0.184
Residual     5.817426

Fixed effects: y ~ week
                Value Std.Error  DF  t-value p-value
(Intercept) 156.05333 2.1589786 119 72.28109       0
week         43.26667 0.7275228 119 59.47122       0
 Correlation:
     (Intr)
week 0.007
```

Example: Rat growth

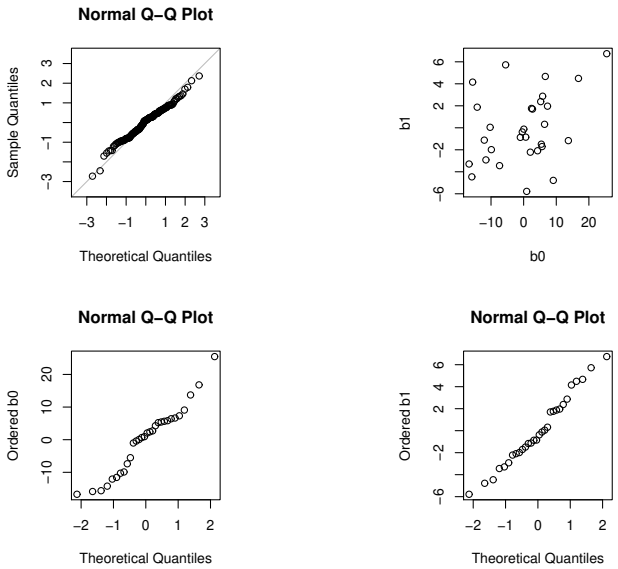
Results from fit of mixed model to rat growth data, using REML. Values in parentheses are for ML fit. In each case $\widehat{\sigma}^2 = 5.82^2$.

Parameter	Fixed		Random	
	Estimate	Standard error	Variance	Correlation
Intercept	156.05	2.16 (2.13)	10.93^2 (10.71^2)	
Slope	43.27	0.73 (0.72)	3.53^2 (3.46^2)	0.18 (0.19)

- ☐ REML estimates of Ω_b slightly larger than ML estimates, but effect is small since $p = 2$.
- ☐ Estimated mean weight in week 1 is 156, but SD of individual rats around this is 11.
- ☐ Correlation between slope and intercept is small but positive: initially heavier rats tend to gain weight faster.
- ☐ Variation around individual slopes is given by $\widehat{\sigma}$, smaller than for the intercept variance.
- ☐ Shrinkage of intercept estimates, shown on next page, is small in this case.
- ☐ Residuals look acceptably normal.

Example: Rat growth

Residuals and random effects



Comments

- Testing for non-zero variance components involves tests on the boundary of the parameter space, which have nasty asymptotic properties: if $\psi = 0$, then a likelihood ratio statistic for testing $\psi = 0$ satisfies $W \dot{\sim} \frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$ as $n \to \infty$, meaning that

$$\mathrm{P}_0(W = 0) = \tfrac{1}{2}, \quad \mathrm{P}_0(W > w) = \tfrac{1}{2}\mathrm{P}(\chi^2_1 > w), \quad w > 0.$$

  Unfortunately,
  - $\mathrm{P}_0(W = 0)$ can be very different from $\frac{1}{2}$ even in large samples, and
  - in more complex problems, the limiting distribution can be much more complex.
- Sometimes clearer to write a mixed model in **multi-level model** form

$$y = X\beta + Z_L b_L + \cdots + Z_0 b_0,$$

  where the $q_l \times 1$ vectors b_l are all mutually independent with means zero and variance matrices Ω_l, so $Y \sim \mathcal{N}_n(X\beta, \sum_{l=0}^{L} Z_l \Omega_l Z_l^{\mathrm{T}})$, where $Z_0 = I_n$, $b_0 = \varepsilon$ and $\Omega_0 = \sigma^2 I_n$.
- The same basic approaches apply in **nonlinear mixed models** and **generalized linear mixed models (GLMMs)**, but integrals appear everywhere and have to be approximated numerically, leading to nastier computations.

3.8 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

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

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

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.

Note: 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

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

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

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

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

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

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

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

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

Relation with least squares

- With fixed λ, the penalized MLE

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

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

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

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

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

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

Numerical example from Wood (2011, JRSSB)

The usual methods (AIC, GCV, ...) for choosing λ are available, but we focus on likelihood methods; see below.

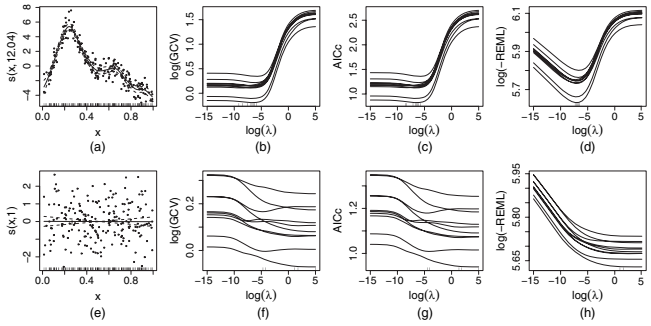


Fig. 1. Example comparison of GCV, AICc and REML criteria: (a) some (x,y)-data modelled as $y_i = f(x_i) + \varepsilon_i$, ε_i independent and identically distributed $N(0,\sigma^2)$ where smooth function f was represented by using a rank 20 thin plate regression spline (Wood, 2003); (b)–(d) various smoothness selection criteria plotted against logarithmic smoothing parameters, for 10 replicates of the data (each generated from the same 'truth') (note how shallow the GCV and AICc minima are relative to the sampling variability, resulting in rather variable optimal λ-values (which are shown as a rug plot), and a propensity to undersmooth; in contrast the REML optima are much better defined, relative to the sampling variability, resulting in a smaller range of λ-estimates); (e)–(h) are equivalent to (a)–(d), but for data with no signal, so that the appropriate smoothing parameter should tend to ∞ (note GCV's and AICc's occasional multiple minima and undersmoothing in this case, compared with the excellent behaviour of REML; ML (which is not shown) has a similar shape to REML)

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.

Laplace approximation

Lemma 34 *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 $\theta = (\beta_{p\times 1}^{\mathrm{T}}, b_{q\times 1}^{\mathrm{T}})^{\mathrm{T}} \sim \mathcal{N}_d(0, S_\lambda^-)$, 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 34 should be satisfied by $h(u) \equiv -n^{-1}\ell_\lambda(\theta)$.

Note on Lemma 34

☐ 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 $m > 1$, but it is more of a bloodbath.

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

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

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

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

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 $\ell_{\mathrm{p}}(\lambda)$ requires its first two derivatives, so we need

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

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	$0^{\dagger}$	1	2	3	4	5	6	$\cdots$	$\geq$14	of 1992
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1988	1	31	80	16	9	3	2	8	$\cdots$	6	174
	2	26	99	27	9	8	11	3	$\cdots$	3	211
	3	31	95	35	13	18	4	6	$\cdots$	3	224
	4	36	77	20	26	11	3	8	$\cdots$	2	205
1989	1	32	92	32	10	12	19	12	$\cdots$	2	224
	2	15	92	14	27	22	21	12	$\cdots$	1	219
	3	34	104	29	31	18	8	6	$\cdots$		253
	4	38	101	34	18	9	15	6	$\cdots$		233
1990	1	31	124	47	24	11	15	8	$\cdots$		281
	2	32	132	36	10	9	7	6	$\cdots$		245
	3	49	107	51	17	15	8	9	$\cdots$		260
	4	44	153	41	16	11	6	5	$\cdots$		285
1991	1	41	137	29	33	7	11	6	$\cdots$		271
	2	56	124	39	14	12	7	10	$\cdots$		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?):

$$\mu_{jk} = \exp\{s(j) + \beta_k\}, \quad \mu_{jk} = \exp\{s(j) + s(k)\},$$

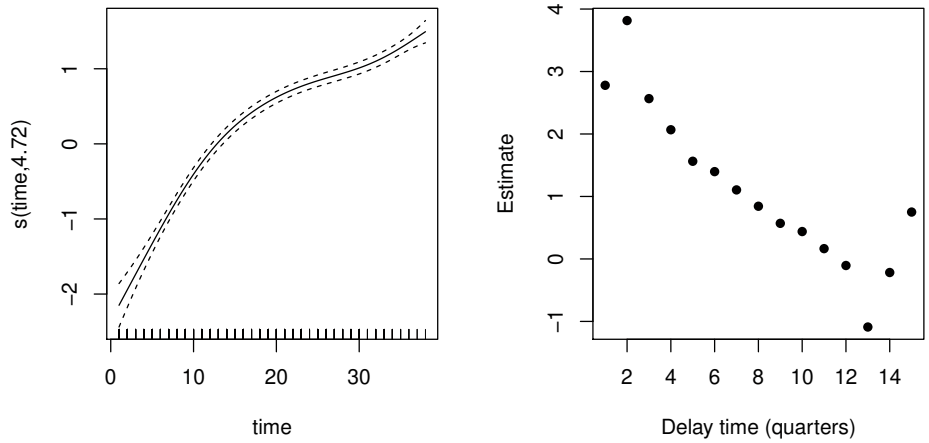
  where the time effect $s(j)$ and the delay effect $s(k)$ vary smoothly.
- Should also account for the overdispersion ...

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
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
```

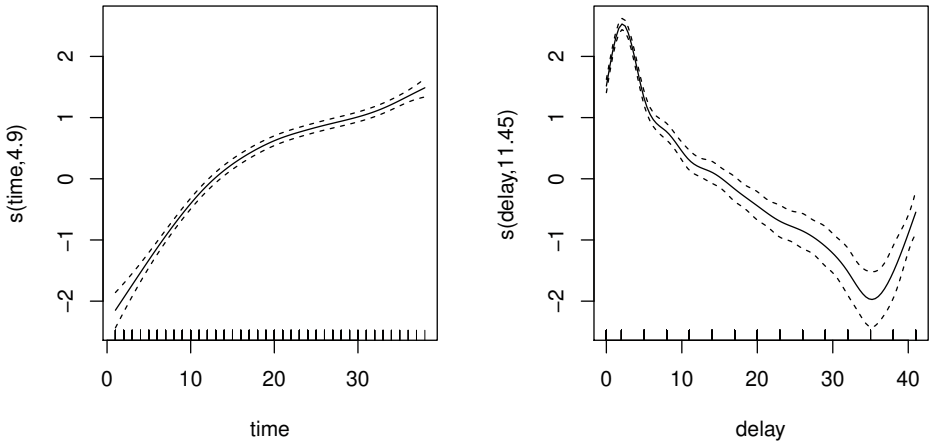
The fits are very similar, but aids.gam2 has slightly lower AIC of 792.0 compared to 792.1 — these are so similar that the choice should be based on interpretability rather than on AIC.

Example: AIDS data



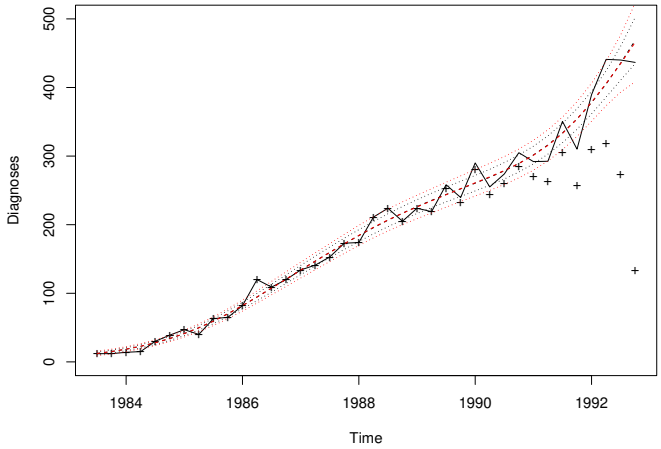
Estimates of (centered) smooth functions $s(j)$ and parameters based on plot(aids.gam1) and coef(aids.gam1).

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.