Smoothing, Splines and Mixed Models

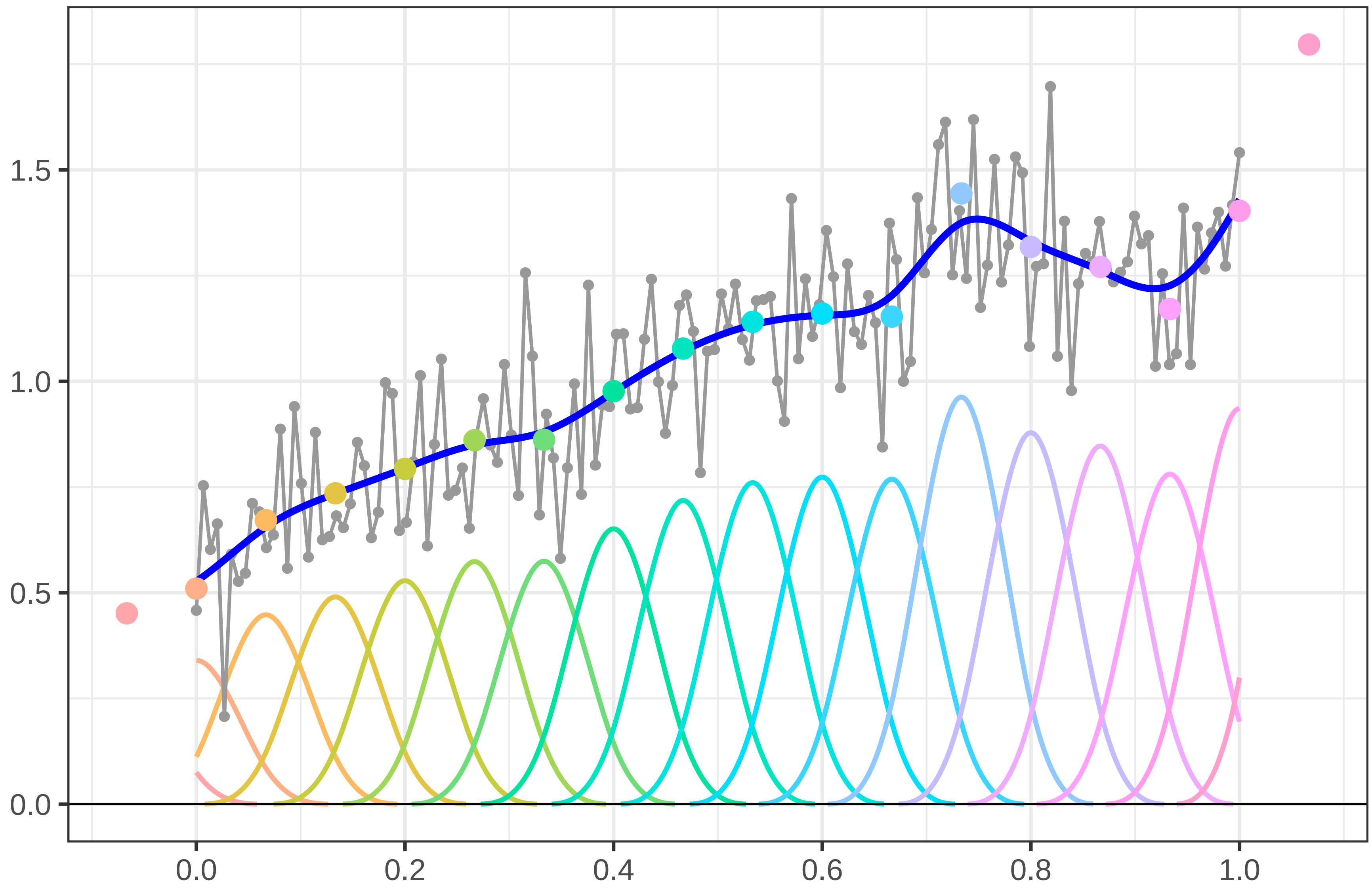
Paul Eilers

Erasmus University Medical Center, Rotterdam

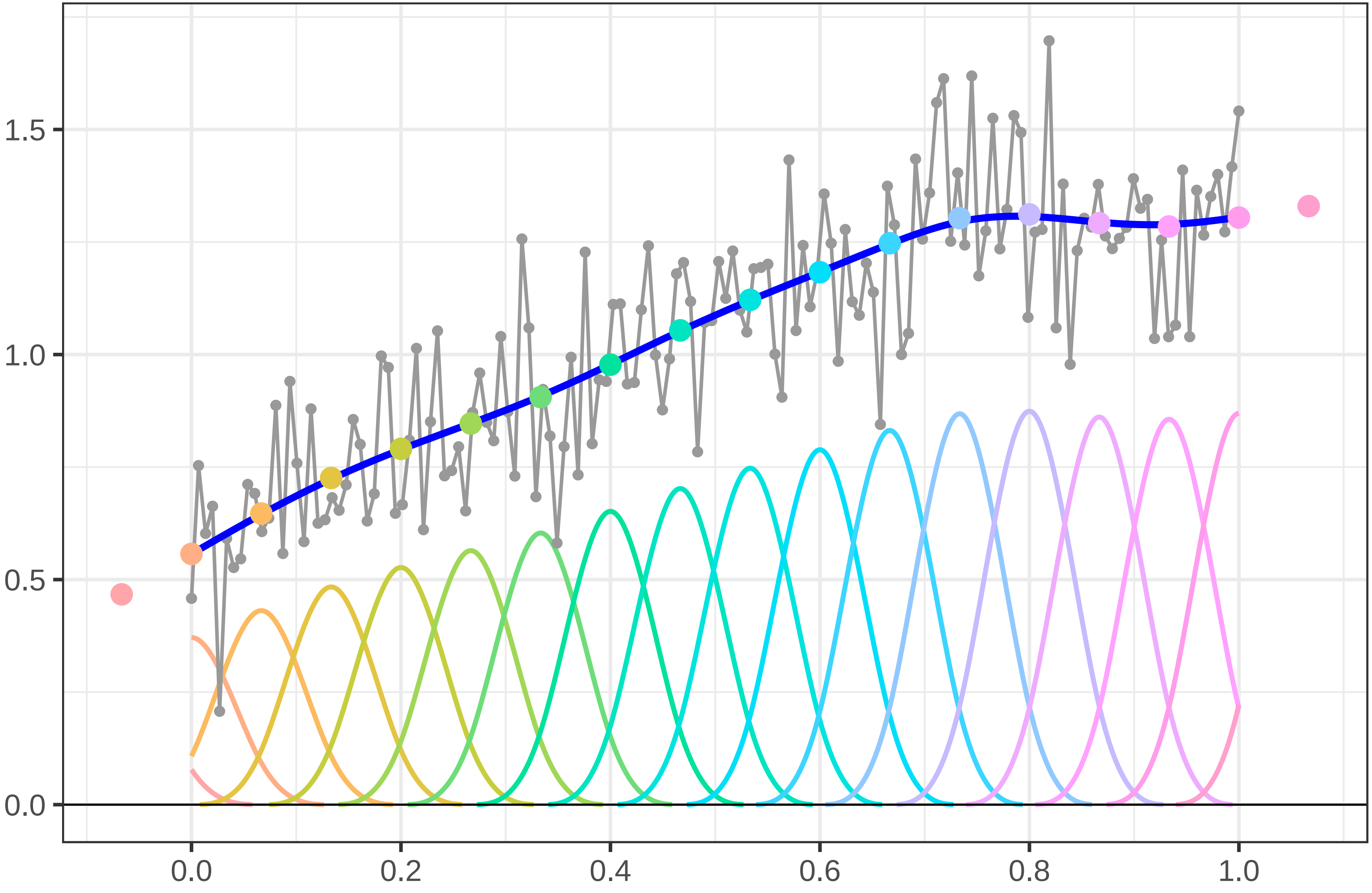
My plan

- Present P-splines as a simple smoothing tool
- First make things complicated: introduce mixed models
- Then simplify the equations
- Making P-splines a simple automatic smoothing tool
- Show examples
- Discuss potential complications

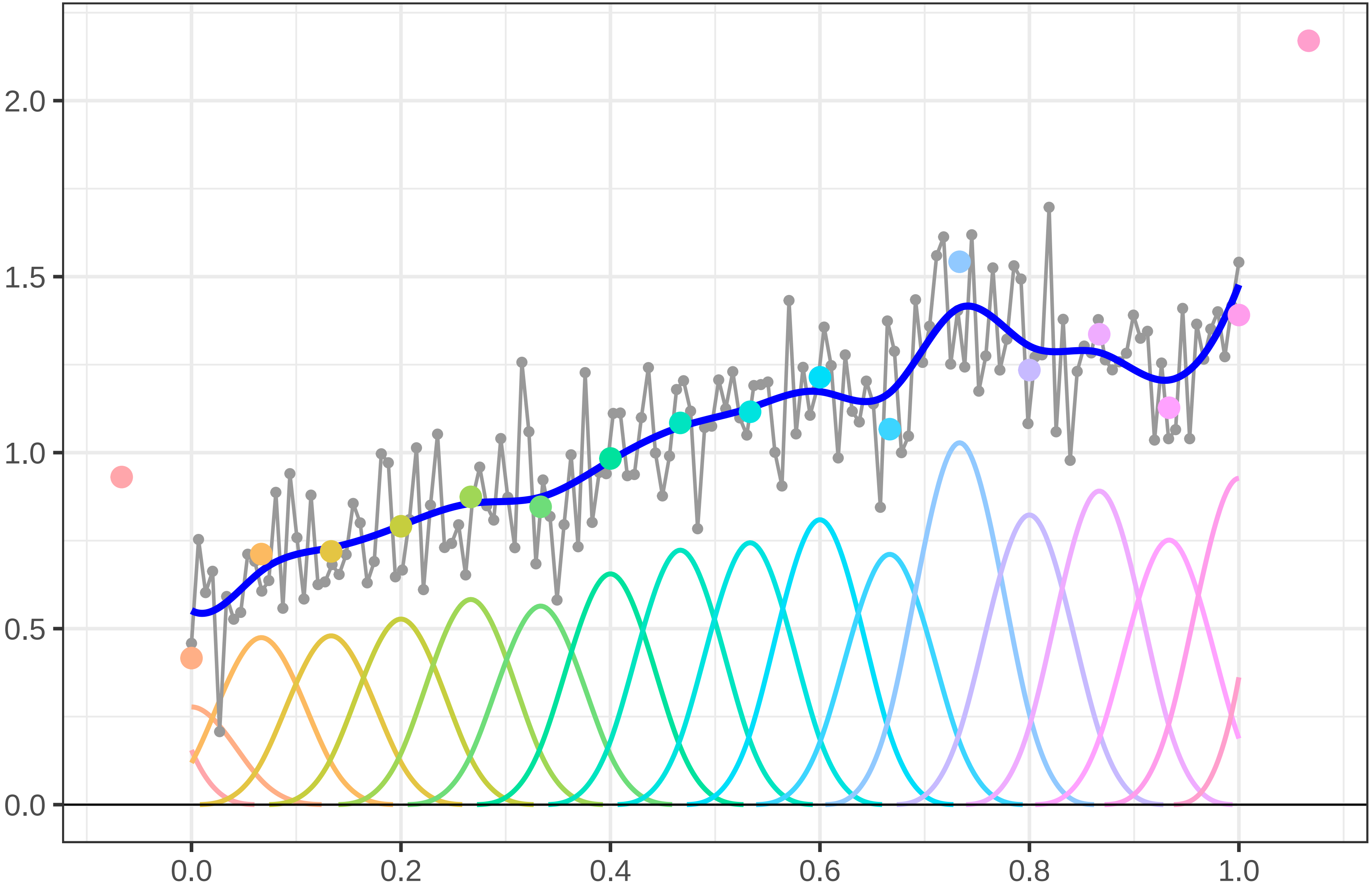
P-splines example: moderate smoothing ($\lambda = 0.1$)



P-splines example: more smoothing ($\lambda = 10$)



P-splines example: less smoothing ($\lambda = 0.01$)

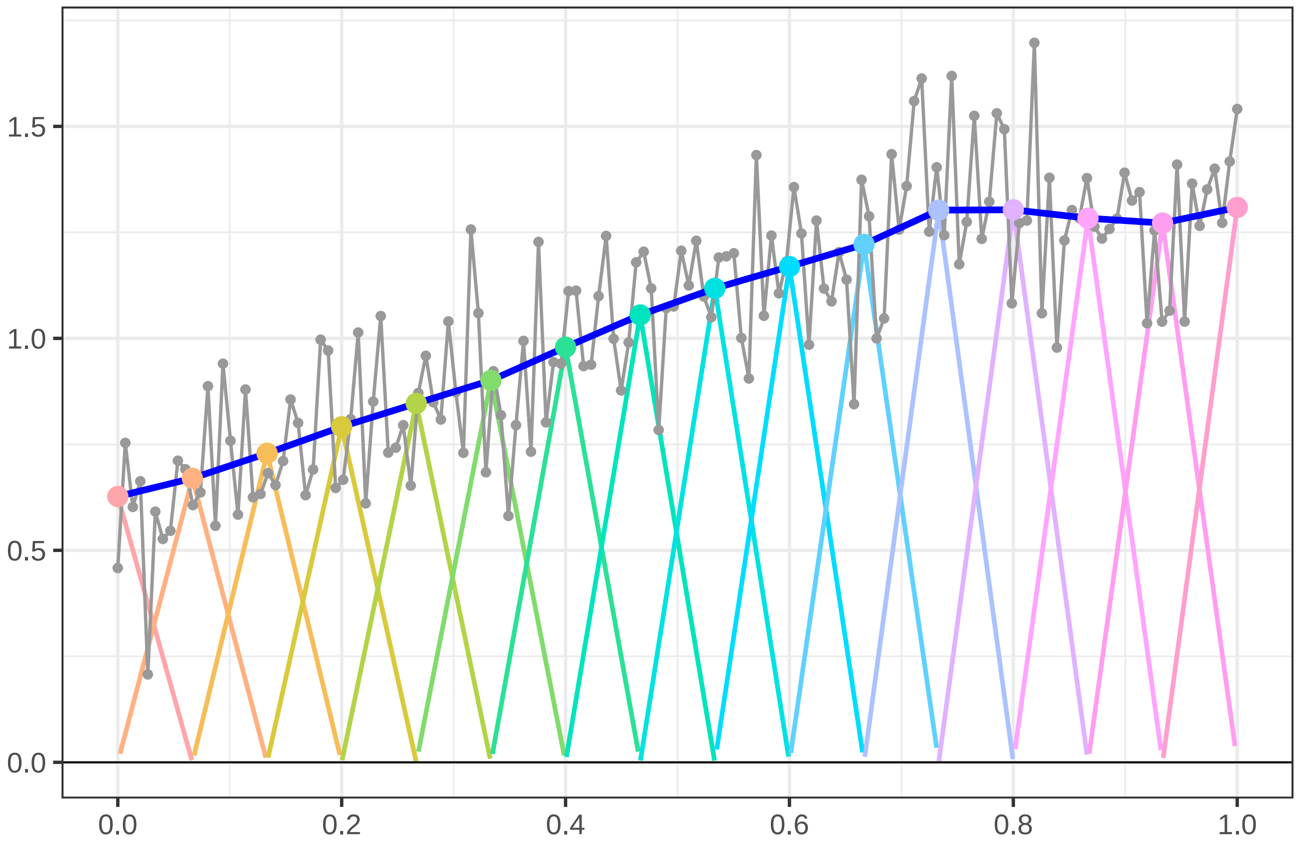


Technical details

- Matrix B with B-splines (the colored humps) in its columns
- Vector of coefficients (the colored dots) a: fitted curve $\mu = Ba$
- Measure of fit: $\|y - \mu\|^2 = \|y - Ba\|^2$
- Add penalty $\lambda\|Da\|^2$, with tuning parameter λ
- Matrix D forms (second order) differences: $Da = \Delta^d a$
- Penalized least squares: $S = \|y - Ba\|^2 + \lambda\|Da\|^2$
- Play with λ to get a pleasing curve (for now)
- We will tune λ automatically with mixed model later

A simplification: linear P-splines

- The segments are linear, the penalty is first order
- Stepping stone to our first mixed model



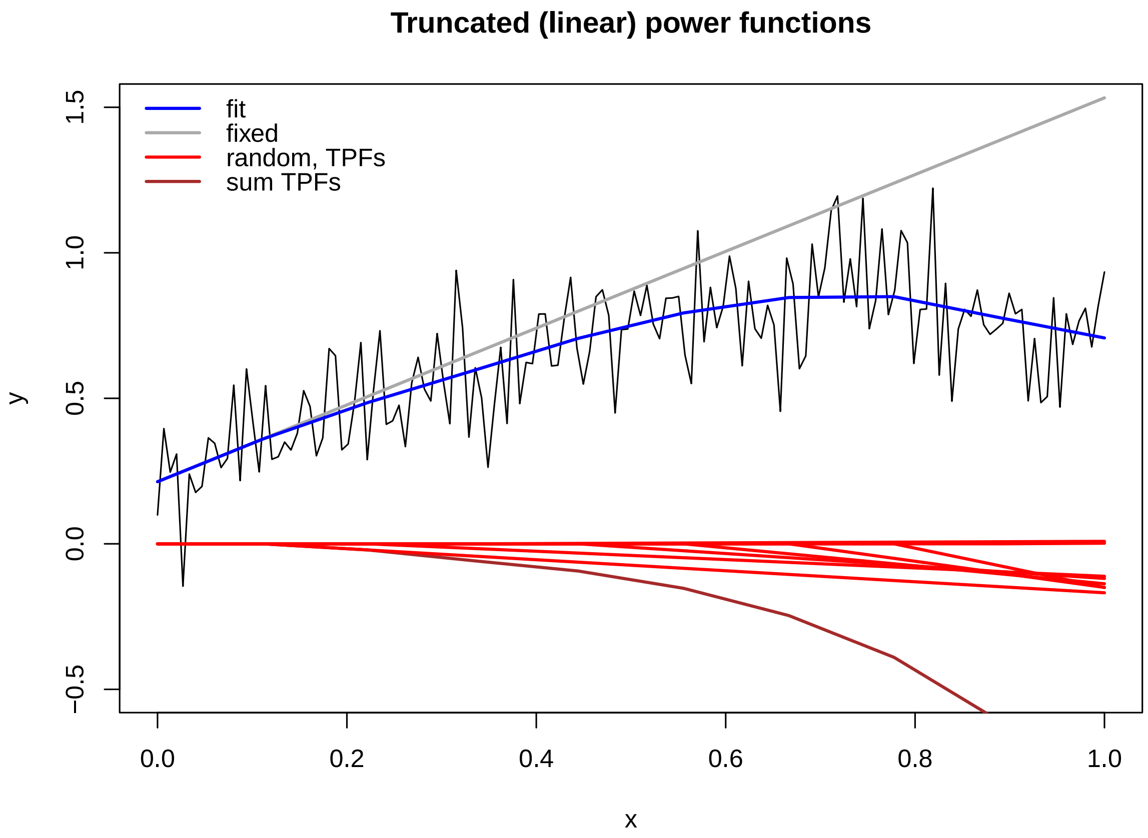
Mixed model with truncated linear function

- Model: $y = \mu + e$; $\mu = X\alpha + Fb = \alpha_0 + \alpha_1 x + Fb + e$
- Centered x, fixed α, errors $e \sim \mathcal{N}(0, \sigma^2)$, random $b \sim \mathcal{N}(0, \tau^2)$
- Truncated linear functions in F: $f_j(x_i) = (x_i - q_j)(x_i > q_j)$
- Objective function: $S = \|y - X\alpha - Fb\|^2 + \kappa\|b\|^2$
- Estimating equations

$$\begin{bmatrix} X'X & X'F \\ F'X & F'F + \kappa I \end{bmatrix} \begin{bmatrix} \hat{\alpha} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} X'y \\ F'y \end{bmatrix}$$

Truncated power functions (TPF) in action

- Individual contributions of TPF give no insight
- In contrast to local B-splines



Mixed model musings

- We estimate and use $\hat{b}$ explicitly: a "conditional" model:
- In orthodox statistics this was suspect
- "You can only estimate fixed parameters"
- "Let's call it BLUP: best linear unbiased predictor"
- Another orthodox idea: see $Fb + e$ as correlated noise
- "Marginal" model, with covariance matrix $C = \tau^2 F'F + \sigma^2 I$
- Estimate $\hat{\alpha} = (X'C^{-1}X)^{-1}X'C^{-1}y$
- Meager result: linear trend $X\hat{\alpha}$ is all you get

Equivalent mixed model for P-splines

- Construct X and Z such that $\mu = X\beta + Zc = Ba$
- Take $X = B\breve{X}$ and $Z = B\breve{Z}$
- With $\breve{X} = [\breve{x}_{jk}]$ $(n \times d)$ with $\breve{x}_{jk} = j^{k-1}$.
- $D\breve{X} = 0$: columns of X lie in the null space of D
- Choose $\breve{Z} = D'(DD')^{-1}$ and $a = \breve{X}\beta + \breve{Z}c$
- $Da = D\breve{X}\beta + DD'(DD')^{-1}c = 0 + c$
- $X = B\breve{X}$ and $Z = B\breve{Z}$ suitable matrices for mixed model

The mixed model equations

- Kernel of log-likelihood

$$L = -\frac{\|y - X\beta - Zc\|^2}{2\sigma^2} - \frac{\|c\|^2}{2\tau^2}$$

- Derivatives with respect to β and c lead to

$$\begin{bmatrix} X'X/\sigma^2 & X'Z/\sigma^2 \\ Z'X/\sigma^2 & Z'Z/\sigma^2 + I/\tau^2 \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{c} \end{bmatrix} = \begin{bmatrix} X'y/\sigma^2 \\ Z'y/\sigma^2 \end{bmatrix}$$

- Multiply by σ^2 and set $\lambda = \sigma^2/\tau^2$

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \lambda I \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{c} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

A useful matrix

- The equations, repeated

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \lambda I \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{c} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

- A useful matrix is

$$K = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} = \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \lambda I \end{bmatrix}^{-1} \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z \end{bmatrix}$$

- We will find that trace(K) is the effective model dimension

Variances

- We call $\rho = \text{trace}(K_{22})$, the effective dimension of c
- Name motivated because we can show $\tau^2 = \|c\|^2/\rho$
- Sum of squares devided by a dimension
- Also $\sigma^2 = \|y - X\hat{\beta} - Z\hat{c}\|^2/(m - d - \rho)$
- With d the order of the differences in the penalty

Henderson

- Generalization: $y = X\beta + Zc + \epsilon$; $c \sim \mathcal{N}(0, G)$ and $\epsilon \sim \mathcal{N}(0, R)$
- G can be block-diagonal for multiple random effects
- The kernel of the deviance (-2 times log-likelihood) is

$$\mathcal{D} = \log |V| + (y - X\beta)' V^{-1} (y - X\beta),$$

- with $V = R + ZGZ'$ (covariance of y, as in marginal model)
- Henderson's equations

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{c} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}$$

Harville's algorithm

- To eliminate fixed effects, Harville (1977) first forms

$$S = R^{-1} - R^{-1}X(X'R^{-1}X)^{-1}X'R^{-1}$$

- Then computes $T = I - (I + Z'SZG)^{-1}$
- And updates τ^2 with $\hat{\tau}^2 = \tilde{c}'\tilde{c}/[q - \mathrm{tr}(\tilde{T})]$
- Where q is length of c
- Updates $\tilde{c}$ and repeat til convergence
- Updates σ^2 with $\sigma^2 = ||y - X\tilde{\beta} - Z\tilde{c}||^2/(m - \mathrm{tr}(\tilde{T}))$

Harville simplified

- Elimination $S = R^{-1} - R^{-1}X(X'R^{-1}X)^{-1}X'R^{-1}$ not attractive
- S is a (large) m-by-m matrix
- Generalization of our matrix K:

$$K = \begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix}^{-1} \begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z \end{bmatrix}$$

- We can prove that $K_{22} = T$
- See appendix E of *Practical Smoothing* (PE and Brian Marx, 2021)

The bottom line for P-splines

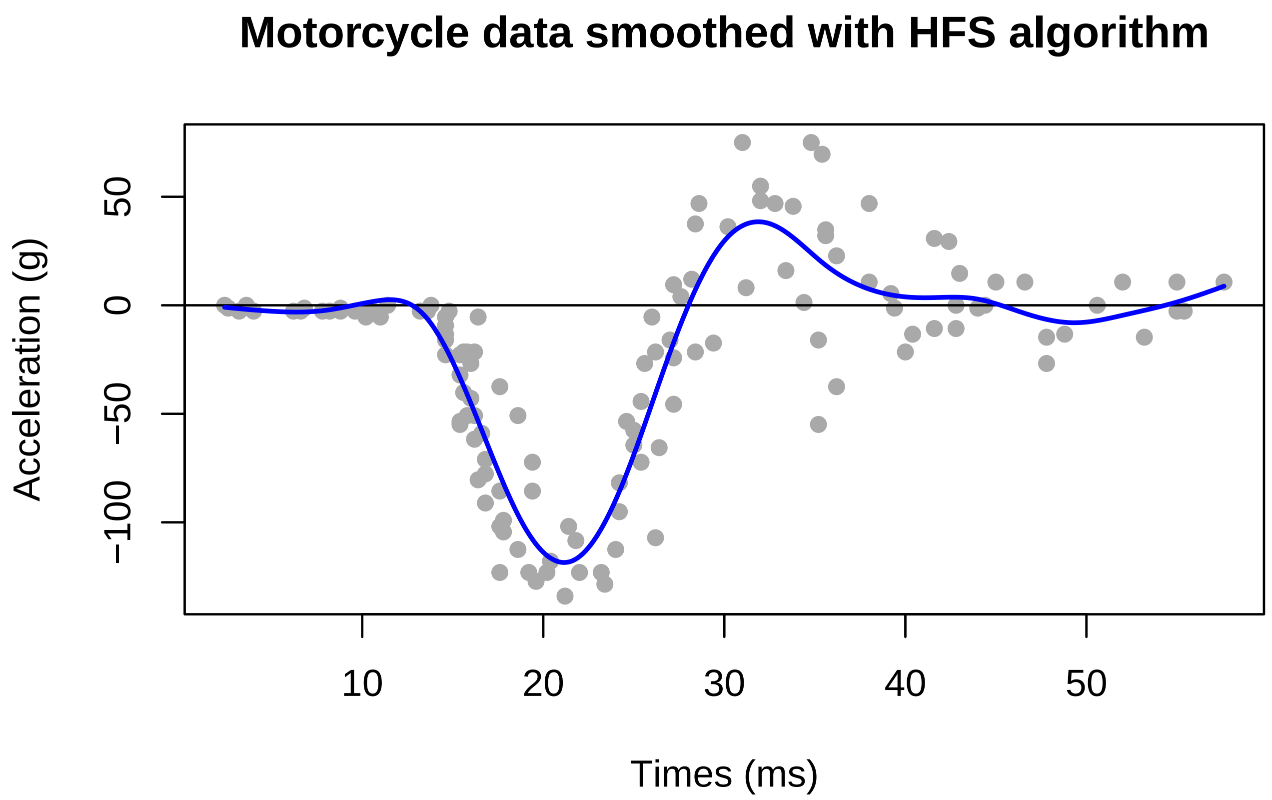
- B-splines in B, penalty matrix in $P = D'D$ (order d)
- Choose a reasonable value for λ, like $\lambda = 1$
- Solve $(B'B + \lambda P)a = B'y$
- Compute $Q = (B'B + \lambda D'D)^{-1}B'B$
- New $ED = \mathrm{trace}(Q)$: effective model dimension
- New $\tau^2 = \|Da\|^2/(ED - d)$: variance of $\|Da\|$
- New $\sigma^2 = \|y - Ba\|^2/(m - ED)$: variance of residuals
- New $\lambda = \sigma^2/\tau^2$ and repeat

On the shoulders of others

- Nickname: HFS algorithm
- H for Harville (*JASA*, 1977)
- F for Fellner (*Technometrics*, 1986)
- S for Schall (*Biometrika*, 1991)
- They made important steps
- Speed of HFS algorithm is quite good

Automatic smoothing in action: motorcycle data

- Motorcycle data: "Fisher Iris" of smoothing
- Convergence in 6 iterations ($\Delta\lambda/\lambda < 10^{-6}$)



The invisible effects

- Classical mixed model have explicit fixed and random effects
- Here they are not directly visible
- One coefficient vector a summarizes P-spline fit
- “Random effects” visible in Da, length $n - d$
- “Fixed effects” hidden, but implicitly present
- Unusual, but no reason to worry
- It stimulates creative thinking about mixed models

The effective dimension according to Ye

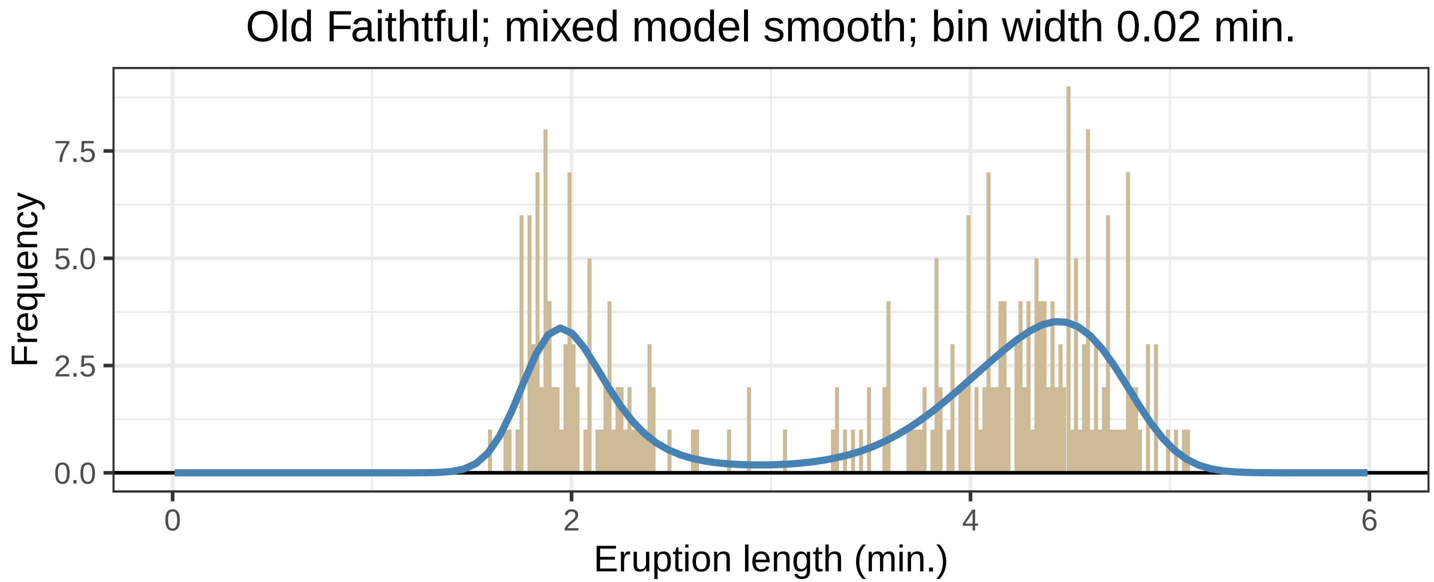
- Remember $\tau^2 = \|Da\|^2/(ED - d)$, the variance of $\|Da\|$
- A sum of squares divided by and effective dimension
- Ye (JASA, 1998) made a principled proposal: $ED = \sum_i \partial \hat{y}_i / \partial y_i$
- Linear model: $\hat{y} = Hy$ with "hat" matrix H and $ED = \text{trace}(H)$
- For P-splines $H = B(B'B + \lambda P)^{-1}B'$
- Cyclic permutation shows equivalence

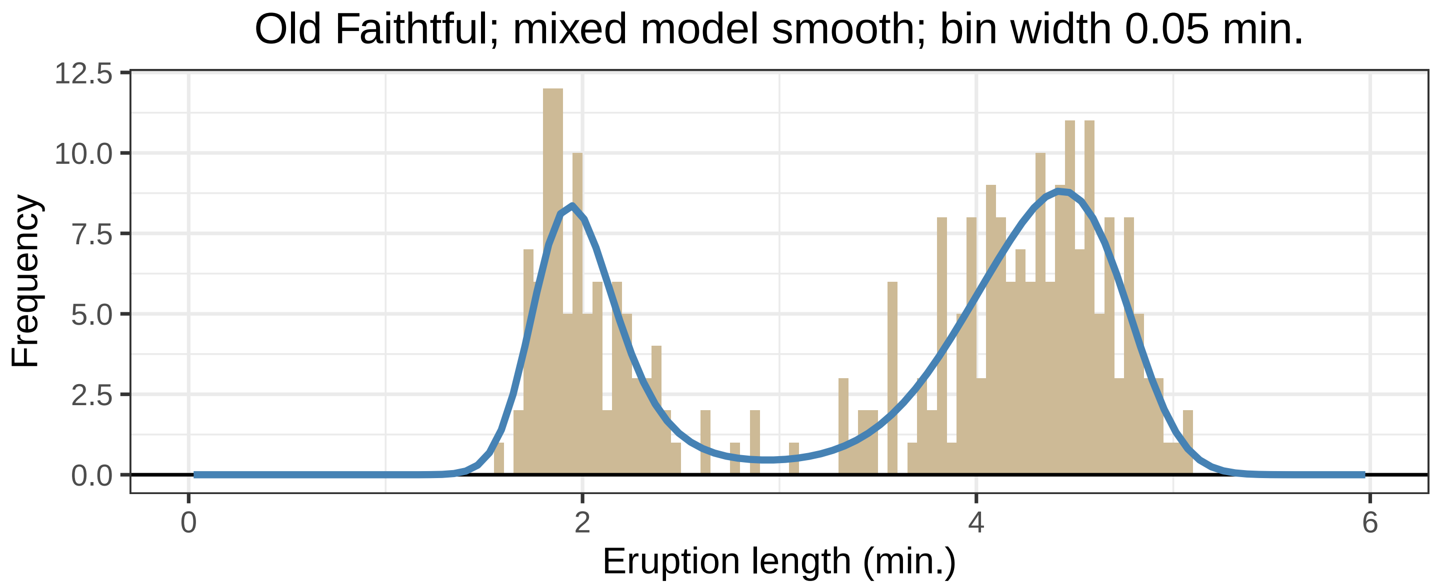
$$\text{trace}[B(B'B + \lambda P)^{-1}B'] = \text{trace}[(B'B + \lambda P)^{-1}B'B]$$

Generalized linear smoothing

- Non-normal data, like counts or fractions
- Example: counts, following Poisson distributions; $y_i \sim \text{Pois}(\mu_i)$
- Model linear predictor $\eta = \log \mu$ with B-splines: $\eta = Ba$
- Combine deviance and penalty: $S = \sum_i D(y_i; \mu_i) + \lambda \|Da\|^2$
- Linearized equations: $(B'MB + \lambda P)\eta = B'(y - \tilde{\mu} + \tilde{M}B\tilde{a})$
- With $M = \text{diag}(\mu)$
- Solved iteratively; usually quickly converging
- Mixed model even easier, because theory says $\sigma^2 \equiv 1$

P-splines and GLM: automatic density smoothing





Generalized additive models (GAM)

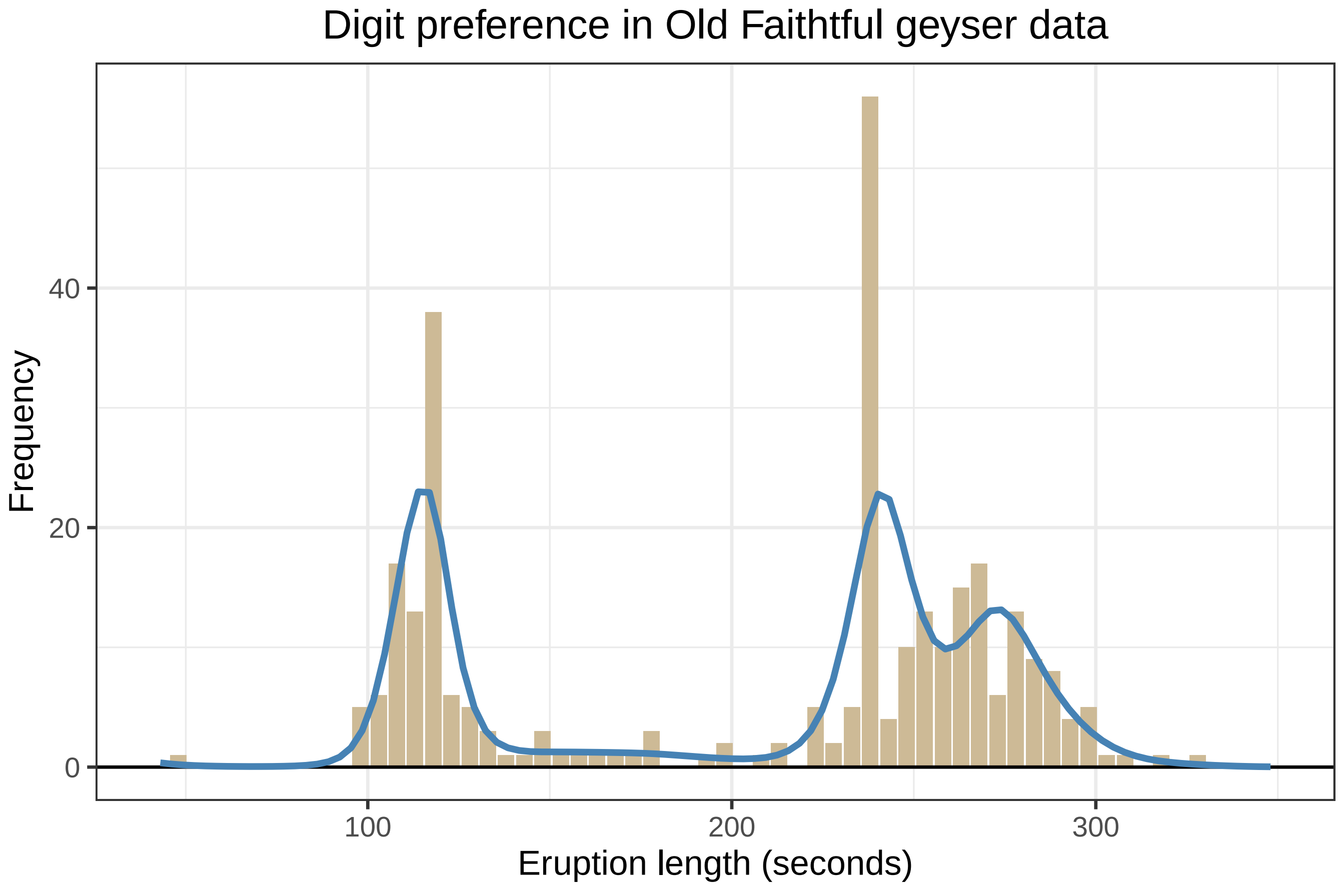
- GAM: sum of several smooth components
- All the (simplified) mixed model theory works
- The equations have a block structure
- With a block per component
- Each block has a partial effective dimension
- Variance (of random effects) easy to compute: pleasant for fitting
- Partial effective dimensions summarize importance of components

A warning

- Automatic smoothing can be dangerous
- Assumptions should hold, e.g. no serial correlation in errors
- This is true for any tool (CV, AIC, BIC, ...)
- Don't trust your results blindly
- Use a generous number of B-splines, rely on the penalty
- Modern computers easily handle hundreds of B-splines
- Small number of B-splines can mask fluctuations
- A very small λ usually is a warning

Example of a problem: digit preference in histogram

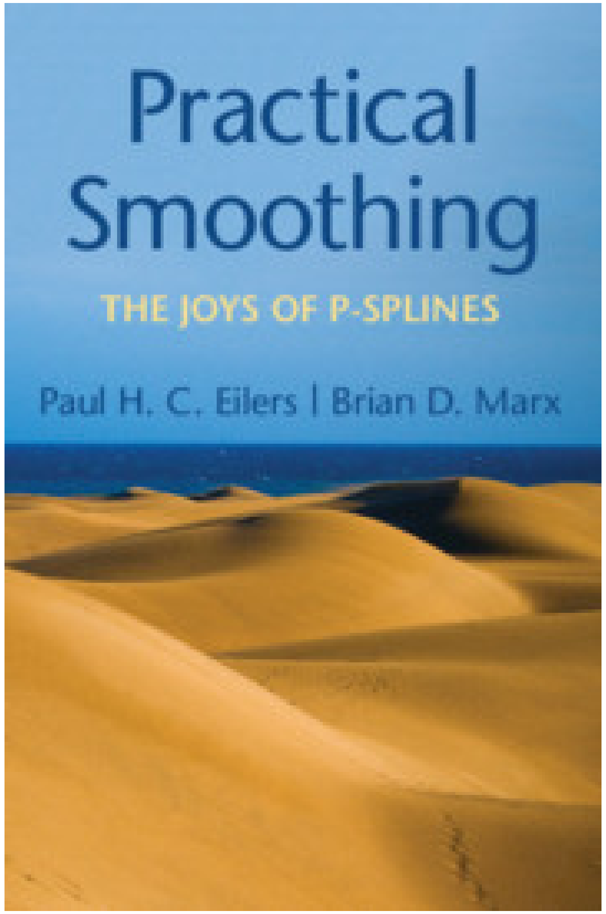
- Part of the Old Faithful data was observed during nights
- Ends of eruptions vague: numbers rounded (2 and 4 minutes)



Wrapping it up

- P-splines can be written as a mixed model
- One option: follow classical pattern (Henderson-Harville)
- Rather complicated and theory is no fun
- Simplifications are possible
- Giving an attractive algorithm for automatic smoothing
- It also works in a generalized linear setting
- Martin Boer will discuss multidimensional P-splines

Advertisement



Paul Eilers & Brian Marx† *Practical Smoothing. The Joys of P-splines*

Cambridge, 2021, GBP 46.99

Software and data in R package JOPS on CRAN