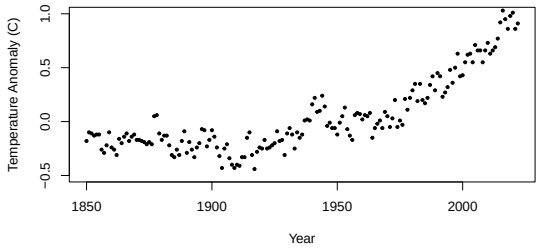
Smooth regression models

Simon Wood
School of Mathematics, University of Edinburgh, U.K.

Global mean temperature anomaly



- Observed mean global annual temperature anomaly relative to 20th century mean.
- Data contain climate component + random weather component. Data are sample from population of all possible mean annual series that could have happened, given underlying climate.
- Questions: which features are annual randomness, which climate trend? Is there any evidence for recent slowing of increase?

Simple temperature model

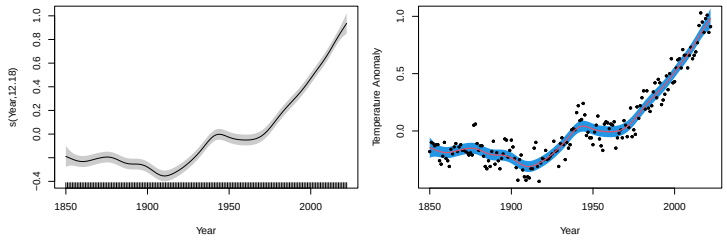
- $\texttt{TA}_i = f(\texttt{Year}_i) + \epsilon_i$, $\epsilon_i \underset{\text{ind.}}{\sim} N(0, \sigma^2)$ and $f(\cdot)$ a smooth function.
- Equivalently $\texttt{TA}_i \underset{\text{ind.}}{\sim} N(f(\texttt{Year}_i), \sigma^2)$.
- R package `mgcv` provides `gam` – `glm` plus smooth terms.

```
require(mgcv)
b <- gam(TA ~ s(Year),data=gt,method="REML")
b

Family: gaussian
Link function: identity

Formula:
TA ~ s(Year)

Estimated degrees of freedom:
8.04  total = 9.04

REML score: -145.3959
```

- Residual checks as GLM OK. Default max DoF of 10 not OK!

More flexible fit

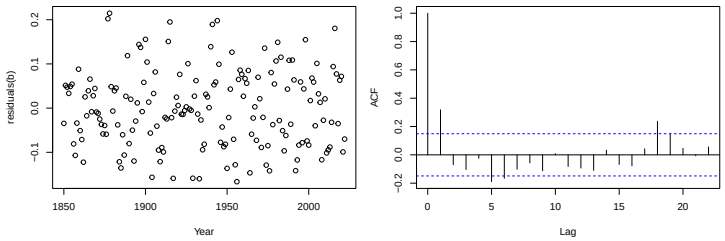
```
b <- gam(TA ~ s(Year,k=30),data=gt,method="REML")
plot(b,scheme=1) ## default plot
## create bespoke plot...
pb <- predict(b,se=TRUE) ## get predictions and s.e.s
ul <- pb$fit + pb$se.fit*2 ## lower conf limit, upper next...
ll <- pb$fit - pb$se.fit*2; n <- length(ul)
with(gt,plot(Year,TA,ylab="Temperature Anomaly",pch=19,cex=.5))
## plot CI...
polygon(c(gt$Year,gt$Year[n:1]),c(ul,ll[n:1]),col=4,border=NA)
with(gt,points(Year,TA,pch=19,cex=.5)) ## make points visible
lines(gt$Year,pb$fit,col=2,lwd=2) ## best fit
```



- No evidence of easing rate of increase.

More checks

```
plot(gt$Year,residuals(b),xlab="Year")
acf(residuals(b))
```

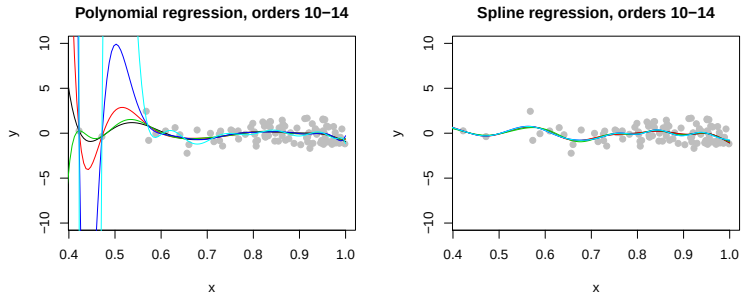


- Residual variance appears constant as assumed by model.
- There is significant residual correlation at lag 1 - see later.
- Also somewhat at lags 5 and 6 – an El Niño effect?

But what model *exactly* is the software using and how? It's time to investigate the main underlying ideas.

Good and bad unknown function models

- We could use polynomials to model unknown functions. i.e. $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \cdots$. Simple, but...
  1. choosing the order of polynomial is clunky, and
  2. they become ever less statistically stable[1] with increasing flexibility (left), unlike equally flexible alternatives (right).

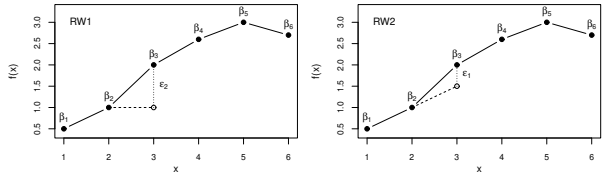


- The instability relates to the polynomial being continuous in all its derivatives. Forgo this and we get more stable behaviour.

[1] numerical stability is not the problem - easy to handle

Models for unknown functions: simple examples

- How should we model an unknown function?
- One way is to treat it as an observation of a stochastic process.
- Consider a piecewise linear function. Evenly spaced 'knots' at $x_1, x_2, \ldots$, parameters $\beta_i = f(x_i)$.



- Can model β_i as random walk. e.g...
  1. Random Gaussian increments, ϵ_i , so $\beta_{i+1} = \beta_i + \epsilon_i$.
  2. Random Gaussian slope changes, ϵ_i, so $\beta_{i+2} = 2\beta_{i+1} - \beta_i + \epsilon_i$.

  ...2 gives visually smoother paths.
- Choose more than enough knots to avoid underfit, and let $\text{var}(\epsilon_i) = \sigma^2$ control function complexity.

RW2 function prior: $\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma^2(\mathbf{D}^\mathsf{T}\mathbf{D})^-)$

- We can write $\boldsymbol{\epsilon} = \mathbf{D}\boldsymbol{\beta}$ where ϵ_i are independent $N(0, \sigma^2)$, i.e.

$$\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdot \end{bmatrix} = \begin{bmatrix} 1 & -2 & 1 & \cdot & \cdot \\ \cdot & 1 & -2 & 1 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdot \end{bmatrix}$$

- What distribution for $\boldsymbol{\beta}$ does this imply? Not unique as $\boldsymbol{\beta}$ has 2 more elements than $\boldsymbol{\epsilon}$.
- Improper Gaussian $\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma^2(\mathbf{D}^\mathsf{T}\mathbf{D})^-)$ works. $\sigma^2(\mathbf{D}^\mathsf{T}\mathbf{D})^-$ is pseudoinverse of precision matrix $\lambda\mathbf{D}^\mathsf{T}\mathbf{D}$ where $\lambda^{-1} = \sigma^2$.
- Why? Form singular value decomposition[2] $\mathbf{D}^\mathsf{T} = \mathbf{U}\mathbf{T}\mathbf{V}^\mathsf{T}$ so that $\mathbf{D}^\mathsf{T}\mathbf{D} = \mathbf{U}\mathbf{T}^2\mathbf{U}^\mathsf{T}$ and $(\mathbf{D}^\mathsf{T}\mathbf{D})^- = \mathbf{U}\mathbf{T}^{-2}\mathbf{U}^\mathsf{T}$.
- $\boldsymbol{\epsilon}$ is a linear transform of $\boldsymbol{\beta}$, so Gaussian, $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{D}\mathbb{E}(\boldsymbol{\beta}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \mathbf{D}(\mathbf{D}^\mathsf{T}\mathbf{D})^-\mathbf{D}^\mathsf{T}\sigma^2 = \mathbf{V}\mathbf{T}\mathbf{U}^\mathsf{T}\mathbf{U}\mathbf{T}^{-2}\mathbf{U}^\mathsf{T}\mathbf{U}\mathbf{T}\mathbf{V}^\mathsf{T}\sigma^2 = \mathbf{I}\sigma^2$

[2] $\mathbf{U}$ column orthogonal $p \times p-2$, $\mathbf{T}$ diagonal, $\mathbf{V}$ orthogonal; $p = \dim(\boldsymbol{\beta})$

Function estimates

- Let's write prior as $\boldsymbol{\beta} \sim N(\mathbf{0}, \mathbf{S}_\lambda^-)$ where $\mathbf{S}_\lambda = \lambda \mathbf{D}^\mathsf{T}\mathbf{D}$.
- Let model be $E(y_i) = \mu_i = f(x_i)$, where y_i is from some tractable distribution, so we can write down a likelihood.
- Posterior modes[3], $\hat{\boldsymbol{\beta}}$ maximize $\pi(\boldsymbol{\beta}|\mathbf{y}) \propto \pi(\mathbf{y}|\boldsymbol{\beta})\pi(\boldsymbol{\beta})$.
- So on the log scale, writing $l(\boldsymbol{\beta}) = \log \pi(\mathbf{y}|\boldsymbol{\beta})$

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\text{argmax}}\; l(\boldsymbol{\beta}) - \boldsymbol{\beta}^\mathsf{T}\mathbf{S}_\lambda\boldsymbol{\beta}/2$$

  – λ controls level of quadratic penalization, hence smoothness.
- We arrived here by considering a *latent Gaussian process* model,
  - but we could have started by constructing the prior precision or covariance for $f(x)$ or $\boldsymbol{\beta}$ directly,
  - or by simply deciding to use a piecewise linear f and penalize the resulting likelihood to control its smoothness.

. . . they are all basically the same in practice, as is treating the Bayesian prior as a frequentist random effect distribution instead.

[3] a.k.a. MAP estimates

Posterior distribution for $\boldsymbol{\beta}$

- Given $\log \pi(\boldsymbol{\beta}|\mathbf{y}) = l(\boldsymbol{\beta}) - \boldsymbol{\beta}^\mathsf{T}\mathbf{S}_\lambda\boldsymbol{\beta}/2 + c$, replace RHS by its 2nd order Taylor approximation about $\hat{\boldsymbol{\beta}}$ and exponentiate.

$$\Rightarrow \pi(\boldsymbol{\beta}|\mathbf{y}) \propto \exp\{-(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\mathsf{T}(\hat{\boldsymbol{\mathcal{I}}} + \mathbf{S}_\lambda)(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})/2\}$$

  i.e. $\boldsymbol{\beta}|\mathbf{y} \sim N\{\hat{\boldsymbol{\beta}}, (\hat{\boldsymbol{\mathcal{I}}} + \mathbf{S}_\lambda)^{-1}\}$ in $n \to \infty$ limit[4].
- Denote this Gaussian approximate posterior $\pi_G(\boldsymbol{\beta}|\mathbf{y})$ for later.
- Can use result directly, or to construct MCMC proposals to simulate from posterior, or as the key ingredient in the INLA approximation (improved tails).
- Here we'll mostly use it directly.

[4] remember $\hat{\boldsymbol{\mathcal{I}}}$ if second derivate matrix (Hessian) of negative log likelihood.

Effective degrees of freedom

- The *degrees of freedom* of a model is usually the number of its parameters that are unknown and free to vary (rather than fixed).
- But smoothing penalties/priors restrict parameters' freedom to vary. How should we then define degrees of freedom?
- Scale by amount penalization has reduced parameter variability.
- $\text{cov}(\boldsymbol{\beta}) \simeq \boldsymbol{\mathcal{I}}^{-1}$ if unpenalized. How to compare to penalized version $(\boldsymbol{\mathcal{I}} + \mathbf{S}_\lambda)^{-1}$? Covariance terms awkward.
- Reparameterize so covariances zero! $\boldsymbol{\beta}' = \mathbf{R}\boldsymbol{\beta}$ where $\mathbf{R}^\mathsf{T}\mathbf{R} = \boldsymbol{\mathcal{I}}$ so $\mathbf{R}^{-1}\mathbf{R}^{-\mathsf{T}} = \boldsymbol{\mathcal{I}}^{-1}$. Unpenalized $\text{cov}(\boldsymbol{\beta}') = \mathbf{R}\mathbf{R}^{-1}\mathbf{R}^{-\mathsf{T}}\mathbf{R}^\mathsf{T} = \mathbf{I}$. Now sum of variances is unpenalized degrees of freedom.
- Penalized $\text{cov}(\boldsymbol{\beta}') = \mathbf{R}(\boldsymbol{\mathcal{I}} + \mathbf{S}_\lambda)^{-1}\mathbf{R}^\mathsf{T}$. So sum of penalized variances is $\text{tr}\{\mathbf{R}(\boldsymbol{\mathcal{I}} + \mathbf{S}_\lambda)^{-1}\mathbf{R}^\mathsf{T}\} = \text{tr}\{(\boldsymbol{\mathcal{I}} + \mathbf{S}_\lambda)^{-1}\mathbf{R}^\mathsf{T}\mathbf{R}\} = \text{tr}\{(\boldsymbol{\mathcal{I}} + \mathbf{S}_\lambda)^{-1}\boldsymbol{\mathcal{I}}\}$ - *effective degrees of freedom*.

Function spaces, bases and basis functions

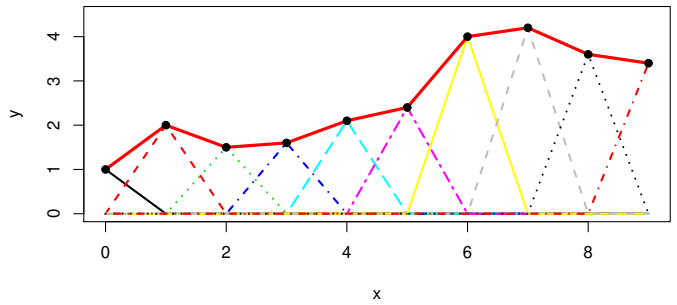
- Working effectively with model functions requires a convenient way of writing them down mathematically (and coding them up).
- Above *'Piecewise linear function with evenly spaced 'knots' at $x_1, x_2, \ldots$, parameters $\beta_i = f(x_i)$'*, describes a space of functions.
- Convenient to represent $f(x)$ using *basis functions* of the space,

$$f(x) = \sum_{j=1}^{p} b_j(x)\beta_j.$$

- Here, the $b_j(x)$ are tent functions, taking value 1 at x_j, descending linearly to 0 at x_{j-1} and x_{j+1}, and being zero outside (x_{j-1}, x_{j+1}).

How the basis works

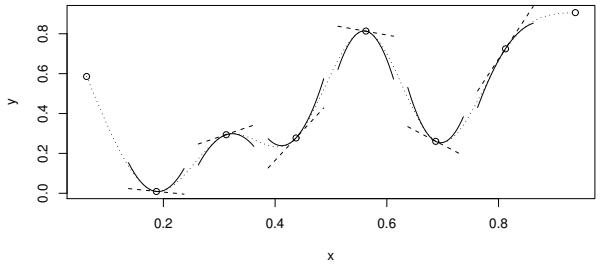
- So the function, $f(x)$, is represented by multiplying each tent function by its coefficient, β_j, and summing the results...



- Given the basis functions and coefficients, we can *predict* the value of f anywhere in the range of the x_j values.

Better bases

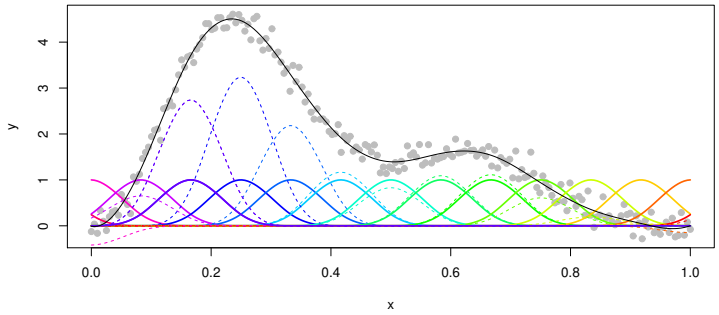
- Piecewise linear functions are statistically stable, but not visually smooth, and even for noiseless data the approximation error is $O(h^2)$ where h is the x spacing of the knots.
- Suppose we seek the function[5], f, that interpolates noise free data observed at $x_1, x_2, \ldots$ while minimizing $\int f''(x)^2 dx$.
  1. The function is a *cubic spline*: it is piecewise cubic, with continuity up to 2nd derivative at the x_j: visually smooth.
  2. It has approximation error $O(h^4)$. A rather high rate.



[5] among all continuous functions with absolutely continuous first derivative

Spline bases

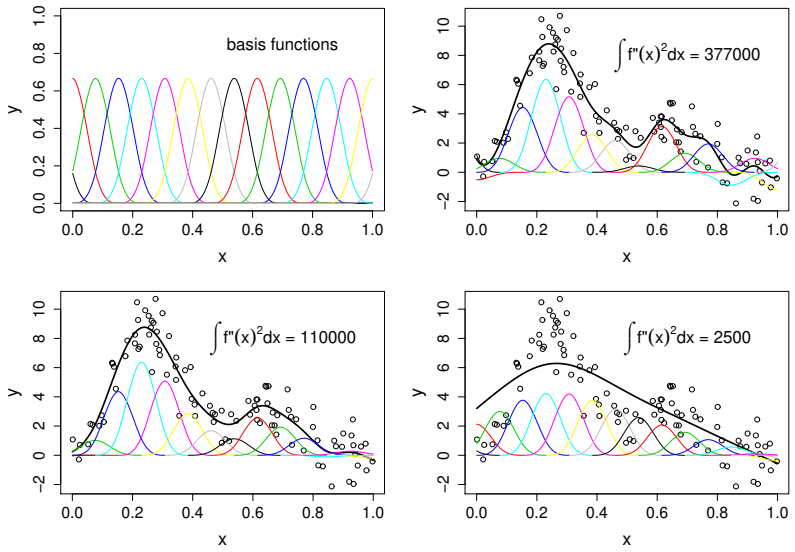
- There are, of course, many possible bases for the space of cubic splines with a given knot sequence, i.e. of different $b_j(x)$ enabling the spline to be written $f(x) = \sum_j \beta_j b_j(x)$.
- A convenient one is the B-spline basis. Here is an example.



Spline penalty

- Given how the spline bases are obtained theoretically, $\int f''(x)^2 dx$ is the natural smoothing penalty.
- Clearly $f''(x) = \sum_j \beta_j b_j''(x) = \boldsymbol{\beta}^\mathsf{T} \mathbf{b}''(x)$, by definition of vector function $\mathbf{b}''(x)$.
- Hence $f''(x)^2 = \boldsymbol{\beta}^\mathsf{T} \mathbf{b}''(x) \mathbf{b}''(x)^\mathsf{T} \boldsymbol{\beta}$.
- So if $\mathbf{S} = \int \mathbf{b}''(x) \mathbf{b}''(x)^\mathsf{T} dx$, then $\int f''(x)^2 dx = \boldsymbol{\beta}^\mathsf{T} \mathbf{S} \boldsymbol{\beta}$:
  - a quadratic penalty;
  - $\boldsymbol{\beta} \sim N(\mathbf{0}, \mathbf{S}^- / \lambda)$ is the equivalent spline smoothing prior;
  - and the spline itself can also be viewed as realization of a latent Gaussian process.
- Clearly the spline model yields exactly the same general structure as the simple piecewise linear random walk processes, so identical methods apply.

spline basis-penalty fit illustrations



How much to penalize: estimating λ[6]

1. *Prediction error optimization*. Which $\boldsymbol{\lambda}$ would be best for predicting data not fitted? Optimize GCV/AIC like criteria, e.g.

$$-2l(\hat{\boldsymbol{\beta}}) + 2\text{EDF}.$$

2. *Marginal likelihood maximisation*. Choose $\boldsymbol{\lambda}$ to maximize the average likelihood of random draws from the prior. i.e. maximize

$$\text{REML} = \int \pi(\mathbf{y}|\boldsymbol{\beta})\pi(\boldsymbol{\beta}|\boldsymbol{\lambda})d\boldsymbol{\beta}$$

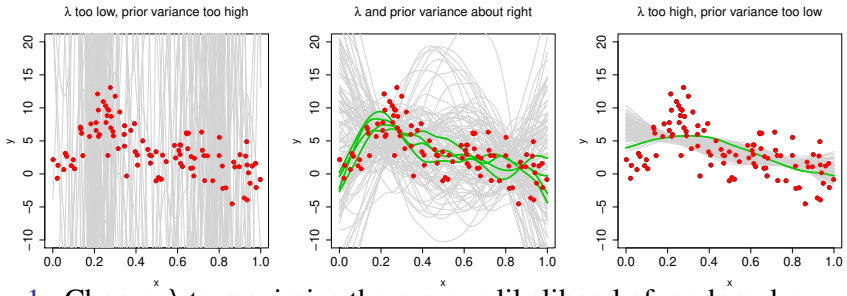
— intractable, but re-using Gaussian approximate posterior, π_G

$$\text{REML} = \pi(\mathbf{y}|\boldsymbol{\lambda}) = \frac{\pi(\mathbf{y}|\hat{\boldsymbol{\beta}})\pi(\hat{\boldsymbol{\beta}}|\boldsymbol{\lambda})}{\pi(\hat{\boldsymbol{\beta}}|\mathbf{y},\boldsymbol{\lambda})} \simeq \frac{\pi(\mathbf{y}|\hat{\boldsymbol{\beta}})\pi(\hat{\boldsymbol{\beta}})|\boldsymbol{\lambda})}{\pi_G(\hat{\boldsymbol{\beta}}|\mathbf{y},\boldsymbol{\lambda})}$$

is tractable: *Laplace Approximation*.

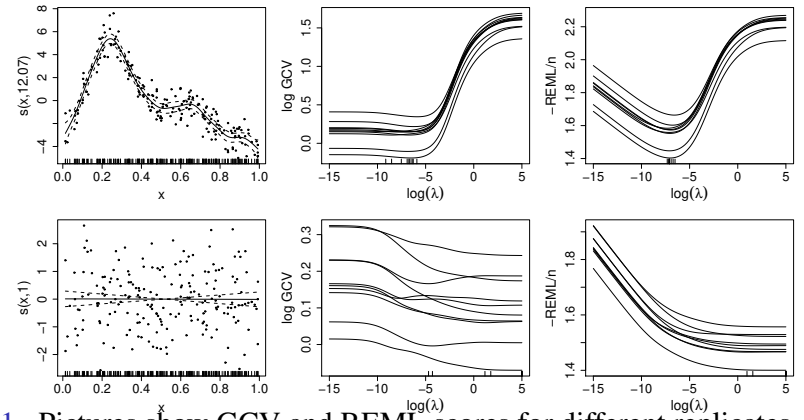
[6] including vector $\boldsymbol{\lambda}$

Marginal likelihood smoothness selection idea



1. Choose λ to maximize the average likelihood of random draws from the prior implied by λ.
2. If λ too low, then almost all draws are too variable to have high likelihood. If λ too high, then draws all underfit and have low likelihood. The right λ maximizes the proportion of draws close enough to data to give high likelihood.

Prediction error vs. likelihood $\boldsymbol{\lambda}$ estimation



1. Pictures show GCV and REML scores for different replicates from same truth.
2. Compared to REML, GCV penalizes overfit only weakly, and so is more likely to occasionally undersmooth.

Credible Intervals

- Given the posterior approximation $\boldsymbol{\beta}|\mathbf{y} \sim N\{\hat{\boldsymbol{\beta}}, (\hat{\boldsymbol{\mathcal{I}}} + \mathbf{S}_\lambda)^{-1}\}$ and the basis expansion for $f(x)$, continuous Bayesian credible intervals are easily constructed.
- $(\hat{\boldsymbol{\mathcal{I}}} + \mathbf{S}_\lambda)^{-1}$ is the covariance matrix implied by sampling variability *plus* the prior expectation of the squared smoothing bias matrix $\mathbf{b}\mathbf{b}^\mathsf{T}$, where $\mathbf{b}$ is the vector of smoothing bias in $\hat{\boldsymbol{\beta}}$.
- By accounting for the uncertainty due to smoothing bias and sampling variability the Bayesian intervals achieve close to nominal coverage, across the function, when treated as frequentist confidence intervals (Nychka, 1988, JASA).

. . . so we have now covered all the ideas and methods used in the initial global temperature example.

Summary

- We can model unknown smooth functions using basis expansions with smoothing priors/penalties to control complexity.
  - also interpretable as latent Gaussian process models.
- Bayesian or penalized likelihood methods used for estimation.
- Smoothing parameters controlling penalization/prior precision estimated by prediction error/cross validation, or by Laplace approximate marginal likelihood.
- Approximate Bayesian posterior gives well calibrated CIs (on average across function).
- `mgcv` implements this approach in R, via a `gam` function that is like `glm` with added smooth terms.