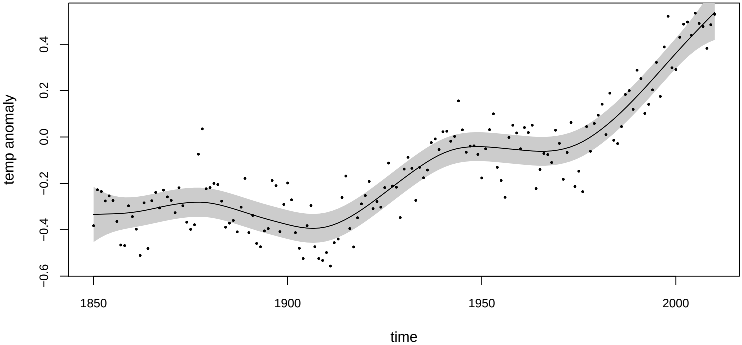
Additive models, load prediction etc.

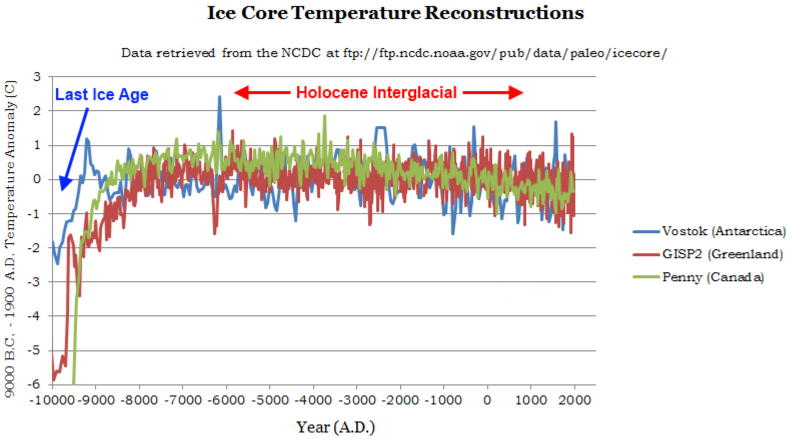
Simon Wood University of Bath, EPSRC funded

What is this?

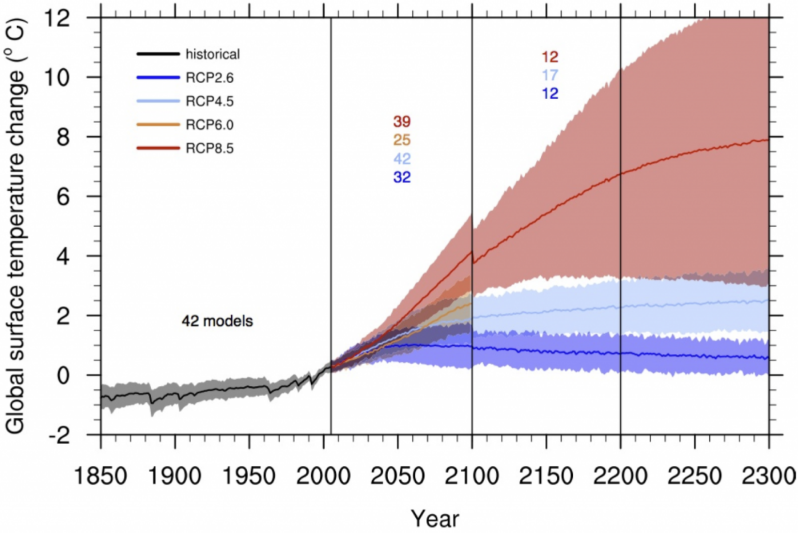


- A smoother applied to some data?
- The solution of a variational problem in a certain reproducing kernel Hilbert space?
- The solution to a variational problem in a Sobolev space incolving a particular semi-norm?
- An intrinsic latent Gaussian random field model with posterior credible region?

Temperature anomaly last 10000y



Projected anomaly IPCC5



Fossil fuel energy produces CO2

. . . but it is relatively easy to match supply to demand.

Renewable energy doesn't produce CO2

. . . but the **big** problem is matching supply and demand.

Demand management

- If we can't control supply so easily, try controlling demand.
- Offer incentives to use power when it is available.
- Only works if:
  1. We can predict supply (weather — quite well sorted out)
  2. We can predict demand, so that we maximize incentives for the behaviour we need. (Hard problem — work needed)
- To re-iterate, incentive based demand management can only work if you can predict what demand would have been without the incentive.
- Better prediction methods is where statistical applied mathematicians can make a real difference.
- Let's look at the maths for one class of predictive models with track record. . .

Smooth prediction models: some background

- Consider a Hilbert space of real valued functions, f, on some domain τ (e.g. $[0, 1]$).
- It is a *reproducing kernel Hilbert space*, $\mathcal{H}$, if evaluation is bounded. i.e. $\exists M$ s.t. $|f(t)| \leq M\|f\|$.
- Then the Riesz representation thm says that there is a function $R_t \in \mathcal{H}$ s.t. $f(t) = \langle R_t, f \rangle$.
- Now consider $R_t(u)$ as a function of t: $R(t, u)$

  $$\langle R_t, R_s \rangle = R(t, s)$$

  — so $R(t, s)$ is known as *reproducing kernel* of $\mathcal{H}$.
- Actually, to every positive definite function $R(t, s)$ corresponds a unique r.k.h.s.

Smoothing

- RKHS are quite useful for constructing smooth models, to see why consider finding $\hat{f}$ to minimize

$$\sum_i \{y_i - f(t_i)\}^2 + \lambda \int f''(t)^2 dt.$$

- Let $\mathcal{H}$ have $\langle f, g \rangle = \int g''(t) f''(t) dt$.
- Let $\mathcal{H}_0$ denote the RKHS of functions for which $\int f''(t)^2 dt = 0$, with finite basis $\phi_1(t), \ldots \phi_M(t)$, say.
- Spline problem seeks $\hat{f} \in \mathcal{H}_0 \oplus \mathcal{H}$ to minimize

$$\sum_i \{y_i - f(t_i)\}^2 + \lambda \|Pf\|^2.$$

Smoothing solution

- $\hat{f}(t) = \sum_{i=1}^{n} c_i R_{t_i}(t) + \sum_{i=1}^{M} d_i \phi_i(t)$. Why?
- Suppose minimizer were $\tilde{f} = \hat{f} + \eta$ where $\eta \in \mathcal{H}$ and $\eta \perp \hat{f}$:
  1. $\eta(t_i) = \langle R_{t_i}, \eta \rangle = 0$.
  2. $\|P\hat{f}\|^2 = \|P\tilde{f}\|^2 + \|\eta\|^2$ which is minimized when $\eta = 0$.
- ...obviously this argument is rather general.
- So if $E_{ij} = \langle R_{t_i}, R_{t_j} \rangle$ and $T_{ij} = \phi_j(t_i)$ then we seek $\hat{c}$ and $\hat{d}$ to minimize

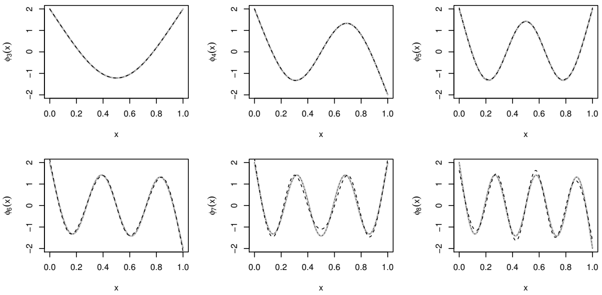
$$\|y - Td - Ec\|_2^2 + \lambda c^{\mathrm{T}} Ec.$$

Computational efficiency: smaller bases

- RKHS approach is elegant and general, but at $O(n^3)$ cost.
- Do we *really* need n coefficients?
- Consider a spline penalty $\int (\nabla^m f)^2 dt = \int f \mathcal{K}^m f dt$, where $\mathcal{K}^m = \nabla^{m*}\nabla^m$ and ∇^{m*} is adjoint of ∇^m w.r.t. $\langle f, g\rangle = \int f(t)g(t)dt$.
- Consider eigenfunctions: $\mathcal{K}^m \phi_i(t) = \Lambda_i \phi_i(t)$, $\Lambda_{i+1} > \Lambda_i \geq 0$.
- Can expand $f(t) = \sum_i \alpha_i \phi_i(t)$ where $\alpha_i = \langle f, \phi_i(t)\rangle$.
- Clearly $\alpha_i \to 0$ (rapidly!) as $i \to \infty$ if $\int (\nabla^m f)^2 dt$ is low.
- Suggestive that we might not need n basis functions.

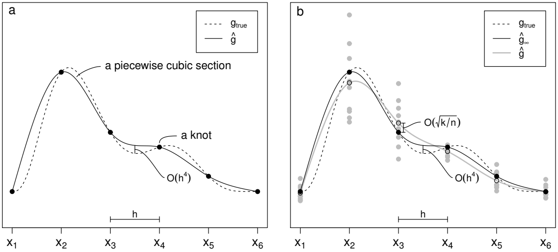
Smaller basis example

- Here are the first few eigenfunctions of $\mathcal{K}^2$...



- So called Demmler-Reinsch basis approximates these... would an L1 penalty on associated coefficients provide a better route to smoothing in the quantile setting?

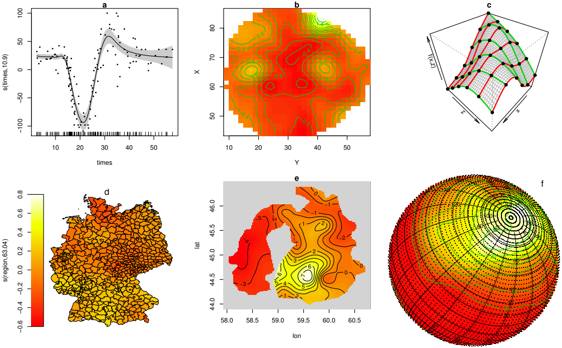
How small a basis: cubic spline example



- A cubic interpolating spline $\hat{g}$ matching a function g_{true} at k evenly spaced (h) knots, has $O(h^4)$ approximation error.
- If we observe g_{true} at each knot n/k times with noise (independently) then $\hat{g}$ has $O(\sqrt{k/n})$ sampling error.
- So $k = O(n^{1/9})$ gives optimal asymptotic error rate.
- With penalization use $k = O(n^{1/9}) - O(n^{1/5})$.

Reduced rank smoothers

- Obtain reduced rank basis by
  1. using spline basis for a representative subset of data, or
  2. using Lanczos methods to find an low order eigenbasis.
- Rich range of smoothers possible. . .



Applicable models

- Models useful in applications, such as load prediction, use multiple smooth terms.
- e.g. $y_i \sim \text{EF}(\mu_i, \phi)$ where $g(\mu_i) = A_i\theta + \sum_j f_j(x_{ji})$.
- Reduced rank spline representation means $g(\mu_i) = X_i\beta$,

$$\hat{\beta} = \underset{\beta}{\text{argmax}} \;\; l(\beta) - \frac{1}{2}\sum \lambda_j \beta^{\mathrm{T}} S_j \beta.$$

— l is log likelihood implied by $\text{EF}(\mu_i, \phi)$.

- Can generalize to models where dependence on f is not additive, and y_i is not EF.
- Have multiple λ_j which need to be estimated.

The Bayesian link

- Suppose we assign a prior density $\beta \sim N(0, \{\sum \lambda_j S_j\}^-)$.
- Then large sample limiting posterior is

$$\beta | y \sim N(\hat{\beta}, \{\mathcal{I} + \sum \lambda_j S_j\}^{-1})$$

  where $\mathcal{I}$ is Fisher information matrix (expected Hessian of -ve log likelihood).
- Estimate λ by marginal likelihood maximization

$$\hat{\lambda} = \underset{\lambda}{\text{argmax}} \int f(y|\beta) f_\lambda(\beta) d\beta$$

  1. Laplace approximate the integral, or
  2. Do integral exactly for linearized 'working model'.

The numerical/computational issues

- At simplest the (-ve) Marginal likelihood has this structure

$$\mathcal{V}(\boldsymbol{\lambda}) = \frac{\|\mathbf{y} - \mathbf{X}\hat{\beta}_\lambda\|^2 + \hat{\beta}_\lambda^{\mathrm{T}}\mathbf{S}_\lambda\hat{\beta}_\lambda}{2\phi} + \frac{\log|\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{S}_\lambda| - \log|\mathbf{S}_\lambda|_+}{2}$$

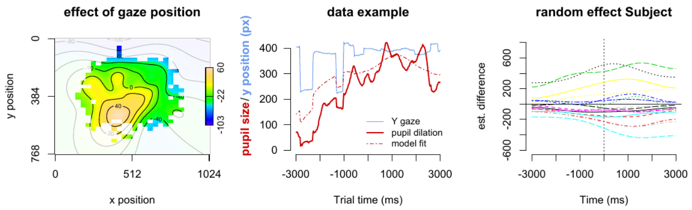
- The determinants require very careful handling.
- Reliable optimization requires at least one, and preferably two, derivatives w.r.t. $\log \lambda$.
- Evaluation and optimization require pivoted QR (very stable) or Cholesky (less so). $O(nk^2)$ cost.
- In big data settings can accumulate $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ or QR decomposition iteratively without forming $\mathbf{X}$.
- Is there a cheaper way?

Parallel methods

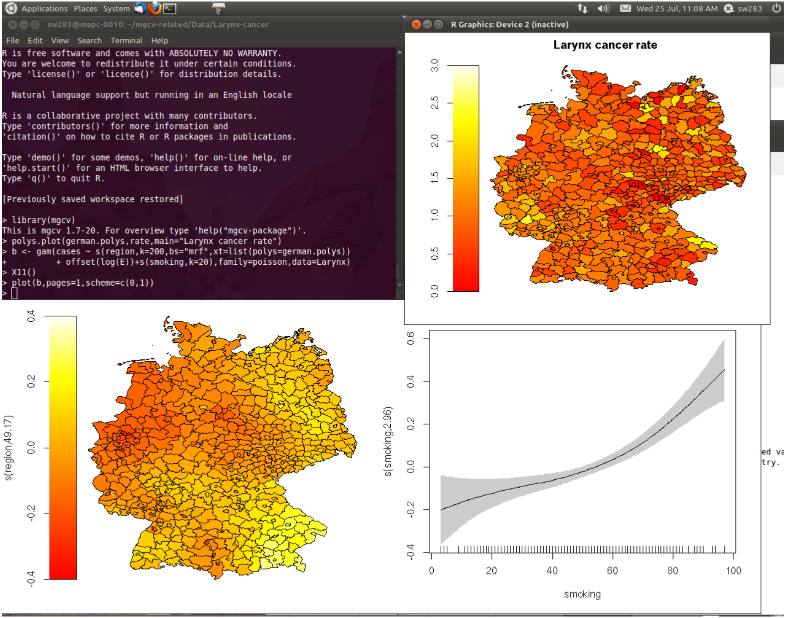
- Can we modify the methods to take advantage of 10-100 core shared memory computers (servers/workstations)?
- Iterative $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ and iterative QR are trivial to split between cores and scale well.
- But we still need a final Cholesky or final QR step. To get that to scale need parallel block pivoted Cholesky or QR.
- These become memory bandwidth limited: very badly in the case of QR.
- Would more modern 'tiling' approaches to QR help to get things to scale?

Software and applications

- The `mgcv` package shipped with R implements a wide variety of smooth model components, automating basis set up and model estimation.
- Future releases should include more scalable methods.
- It is quite widely used ($\sim$ 1500 citations last year), here are some estimates from a pupil dilation experiment about reading and language processing...



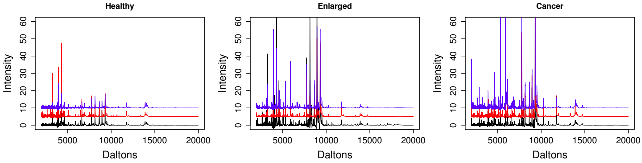
Software and applications



Other applications

- Unemployment and inflation (BoE).
- Mortality rate trends (HSE)
- Fisheries stock assessment (e.g. CSIRO, CEFAS, IFremar).
- Forest health and inventory, remote sensing calibration.
- Air pollution and other epidemiology.
- Medical statistics...

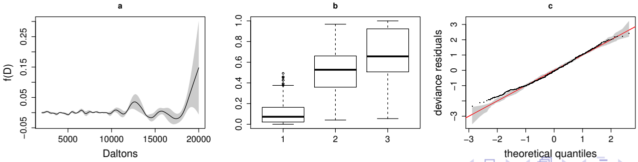
Example: predicting prostate status



- Model category (benign/enlarged/cancer) predicted by latent variable with mean

$$\mu_i = \int f(D)\nu_i(D)dD$$

where $\nu_i(D)$ is i^{th} NI spectrum.

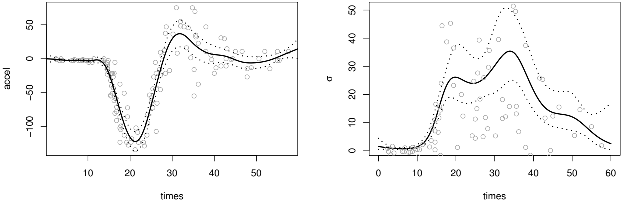


Scale location extensions

- Can extend methods to additively model mean and variance (and skew and. . .)
- Simple example: $y_i \sim N(\mu_i, \sigma_i)$

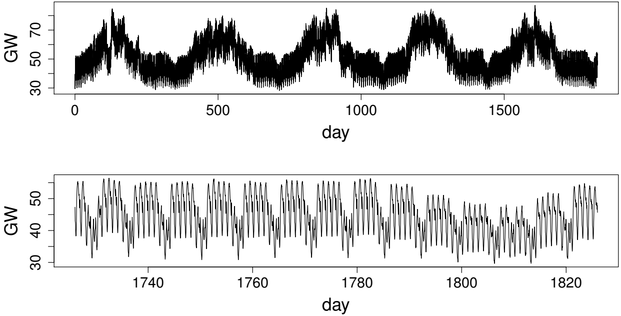
$$\mu_i = \sum_j f_j(x_{ji}), \qquad \log \sigma_i = \sum_j g_j(z_{ji}).$$

- Here is a simple 1-D smoothing example of this. . .



- An alternative to quantile regression?

Back to load prediction



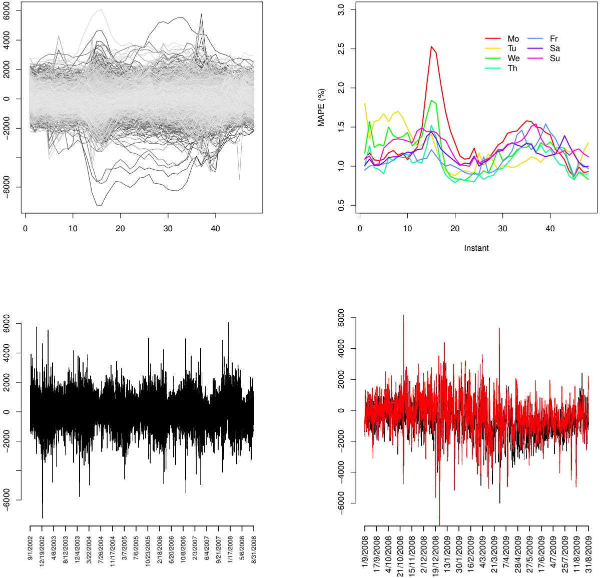
- A predictive smooth additive model...

$$\begin{aligned}\texttt{L}_i = \gamma_j + f_k(\texttt{I}_i, \texttt{L}_{i-48}) + g_1(\texttt{t}_i) + g_2(\texttt{I}_i, \texttt{toy}_i) + g_3(\texttt{T}_i, \texttt{I}_i) \\ + g_4(\texttt{T.24}_i, \texttt{T.48}_i) + g_5(\texttt{cloud}_i) + \texttt{ST}_i h(\texttt{I}_i) + e_i\end{aligned}$$

  if observation i is from day of the week j, and *day class* k.

- $e_i = \rho e_{i-1} + \epsilon_i$ and $\epsilon_i \sim N(0, \sigma^2)$ (AR1).

Residuals



MAPE (%)
Instant
Mo
Tu
We
Th
Fr
Sa
Su
9/1/2002
12/19/2002
4/8/2003
8/12/2003
12/4/2003
3/22/2004
7/26/2004
11/17/2004
3/7/2005
7/6/2005
10/23/2005
2/18/2006
6/20/2006
10/8/2006
2/3/2007
6/4/2007
9/21/2007
1/17/2008
5/6/2008
8/31/2008
1/9/2008
17/9/2008
4/10/2008
21/10/2008
15/11/2008
2/12/2008
19/12/2008
13/1/2009
30/1/2009
16/2/2009
4/3/2009
21/3/2009
7/4/2009
28/4/2009
27/5/2009
17/6/2009
4/7/2009
25/7/2009
11/8/2009
31/8/2009

Open questions. . .

- Full model not quite as good as 48 half hourly models.
- But surely 48 separate models is poor information sharing.
- Is the problem basis size? Can't compute with a large enough basis size to make combined model competitive?
- What is the best way to achieve information sharing and computational efficiency?
- Do we just need better methods for bigger models?
- Or is one big model just wrong?
- Or is the problem that the statistical computing methods are not efficient enough, or are missing something about the structure?
- If we want to model at local level, how should information be shared then?