Smoothing and basis expansions

Simon Wood

Penalizing a different sort of complexity

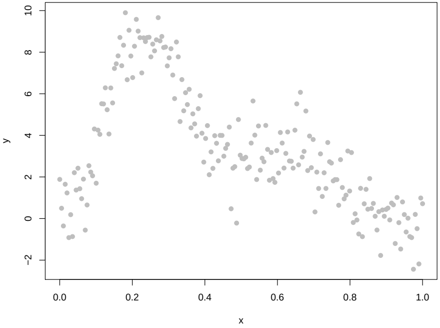
- So far we have considered the case of (generalized) linear models where we need to penalize the complexity of having too many predictors of unknown importance.
- For the most part we approached this task prioritizing predictive performance, therefore selecting the penalty parameter for optimal predictive performance in (cross) validation.
- A different sort of model complexity arises when we are unsure of the form of the relationship between a predictor and a response. e.g. for the model

  $$y_i = f(x_i) + \epsilon_i \quad \epsilon_i \underset{\text{iid}}{\sim} N(0, \sigma^2)$$

  should the unknown function, f, be smooth or wiggly?
- And is prediction error the only way to decide?

A simple example

- Here are some $x - y$ data with a noisy non-linear relationship



- A model along the lines of 'y is some smooth function of x observed with noise' seems appropriate, but how smooth or complex a function is not clear.

Bases and smoothness

- Let's look further at the model

$$y_i = f(x_i) + \epsilon_i \quad \epsilon_i \underset{\text{iid}}{\sim} N(0, \sigma^2)$$

  where f is an unknown 'smooth' function.

- A practical way forward is to introduce a *basis expansion*

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

  where the *basis functions*, $b_j(x)$ are chosen to have convenient properties and the β_j will have to be estimated.

- We also need to define 'smooth': e.g. a small value of

$$\int f''(x)^2 dx$$

Basis penalty smoothing

- To avoid bias from an overly restrictive model, we choose p to be moderately large.
- But large p risks high uncertainty in our inference about f.
- As in the penalized linear model case, there is a bias-variance trade-off.
- To control the trade-off we can use penalized estimation:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \int f''(x)^2 dx$$

where $X_{ij} = b_j(x_i)$ and $\lambda \geq 0$ is a smoothing (regularization) parameter.

The penalty is quadratic in $\boldsymbol{\beta}$

- $f(x) = \sum_{j=1}^{p} \beta_j b_j(x)$, so it follows that $f''(x) = \sum_{j=1}^{p} \beta_j b_j''(x)$.
- Defining vector $\mathbf{d}(x)$ where $d_j(x) = b_j''(x)$ then $f''(x) = \boldsymbol{\beta}^\mathsf{T}\mathbf{d}(x)$.
- In consequence

$$\int f''(x)^2 dx = \int \boldsymbol{\beta}^\mathsf{T}\mathbf{d}(x)\mathbf{d}(x)^\mathsf{T}\boldsymbol{\beta} dx = \boldsymbol{\beta}^\mathsf{T}\mathbf{S}\boldsymbol{\beta}$$

  where $S_{ij} = \int d_i(x)d_j(x)dx.$*
- For some bases, S_{ij} can be computed exactly. e.g. *B-splines*.
- So our fitting problem is now the L_2 penalized

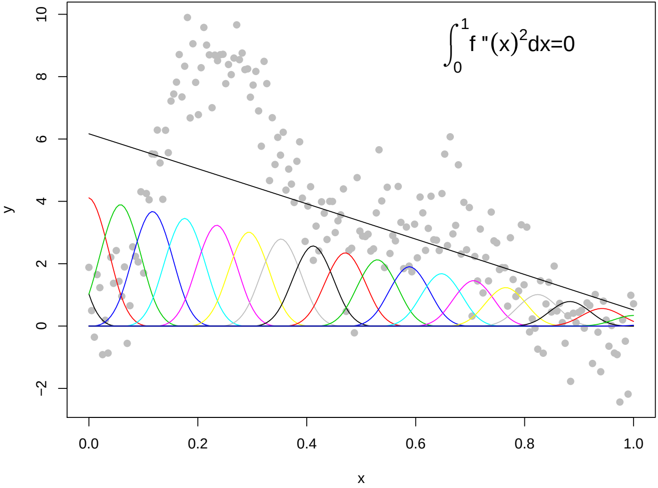
$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \, \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\boldsymbol{\beta}^\mathsf{T}\mathbf{S}\boldsymbol{\beta}.$$

- Let's see the basis-penalty smoother in action ...

---

*this works for other orders of derivative in the penalty too.

Penalized B-spline basis smoothing as λ reduced



$\hat{\boldsymbol{\beta}}$, $\hat{\lambda}$ etc.

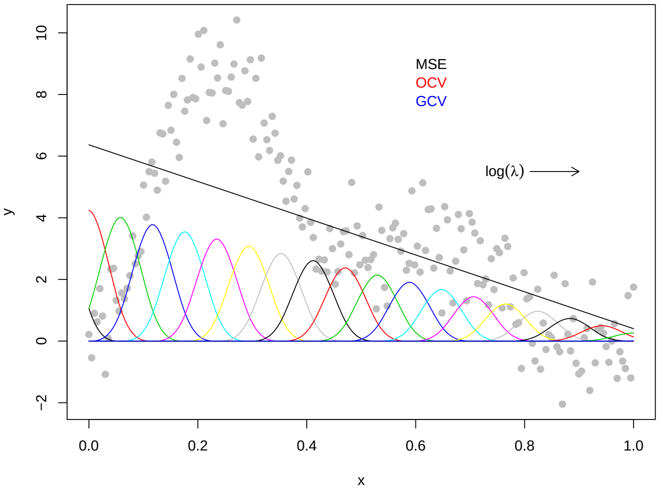
- $\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \boldsymbol{\beta}^\mathsf{T}\mathbf{S}\boldsymbol{\beta}$ has exactly the same form as the ridge regression problem covered earlier, except that $\mathbf{S}$ replaces $\mathbf{I}$ in the penalty.
- It follows that
  1. $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$.
  2. The fitted values are $\hat{\boldsymbol{\mu}} = \mathbf{A}\mathbf{y}$ where $\mathbf{A} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{S})^{-\mathbf{1}}\mathbf{X}^\mathsf{T}$.
  3. As before, the ordinary cross validation criterion is

$$\text{OCV} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_i^{[-i]})^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{(1 - A_{ii})^2}$$

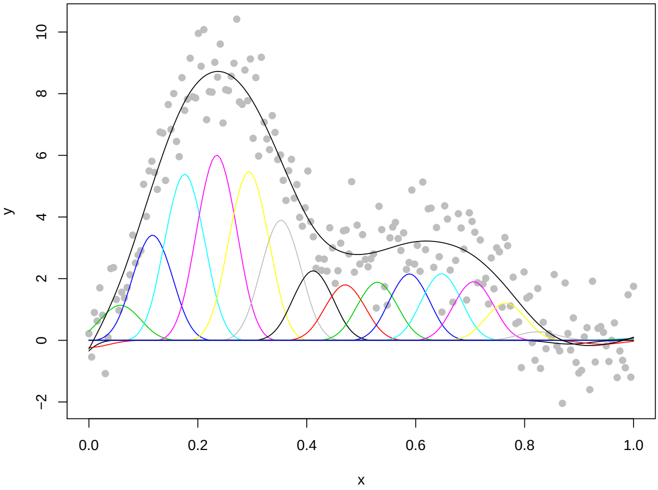
- So we can estimate λ by OCV or the weight averaged version

$$\text{GCV} = \frac{n\|\mathbf{y} - \hat{\boldsymbol{\mu}}\|^2}{\{n - \text{trace}(\mathbf{A})\}^2}$$

Cross validating for λ



The cross validated fit



The Bayesian perspective

- As with ridge regression, we can view the smoothing penalty as induced by a prior $\boldsymbol{\beta} \sim N(\mathbf{0}, \mathbf{S}^{-}\sigma^2/\lambda)$
- The prior here is an *improper* Gaussian, as the prior precision matrix, $\mathbf{S}\lambda/\sigma^{\mathbf{2}}$, is not full rank[†]
- Notice also that $\pi(\boldsymbol{\beta}) \propto \exp\{-\lambda\boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}\boldsymbol{\beta}/(2\sigma^2)\}$ – an exponential prior on wiggliness of f.
- The posterior follows as before, but with **S** in place of **I**

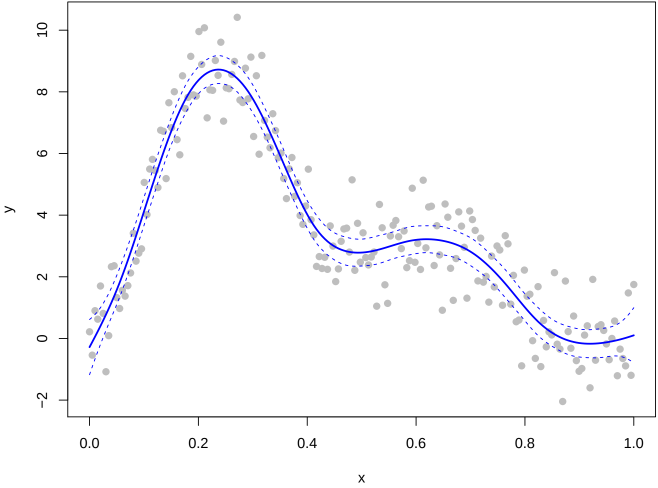
$$\boldsymbol{\beta}|\mathbf{y} \sim N(\hat{\boldsymbol{\beta}}, (\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{S})^{-1}\sigma^2)$$

- Using this with the cross validated $\hat{\lambda}$ is a sort of *Empirical Bayes* method. e.g. we can immediately obtain credible intervals for f.

[†] **S** is rank deficient by the dimension of the space of functions it does not penalize. e.g. 2 for the cubic spline penalty.

95% Bayesian Credible Interval

- If $\hat{f}(x_i) = \tilde{\mathbf{x}}^\mathsf{T}\hat{\boldsymbol{\beta}}$ then $\text{var}\{\hat{f}(x_i)\} = \tilde{\mathbf{x}}^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{S})^{-1}\tilde{\mathbf{x}}\sigma^2$, so ...



Estimating λ from the marginal likelihood

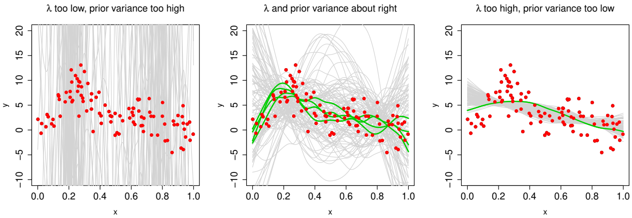
- Formulation in terms of Bayesian smoothing priors raises the possibility of taking a fully Bayesian approach to inference about λ, or of estimating λ to maximise the marginal likelihood.
- Here we will concentrate on maximising the marginal likelihood

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

- At first sight this is not as intuitive as the cross validation approaches to λ choice, but actually it does something quite intuitive…

ML λ estimation is intuitive

- Look at the marginal likelihood expression again $\pi(\mathbf{y}|\lambda) = \int \pi(\mathbf{y}|\boldsymbol{\beta})\pi(\boldsymbol{\beta}|\lambda)d\boldsymbol{\beta}$ — it is the average likelihood of random draws from the prior.
- So by maximizing it we choose λ to maximise the average likelihood of draws from the prior.



- In each panel the curves are randomly drawn from $\pi(\boldsymbol{\beta}|\lambda)$ (but centred) and the green ones have likelihood above a threshold.

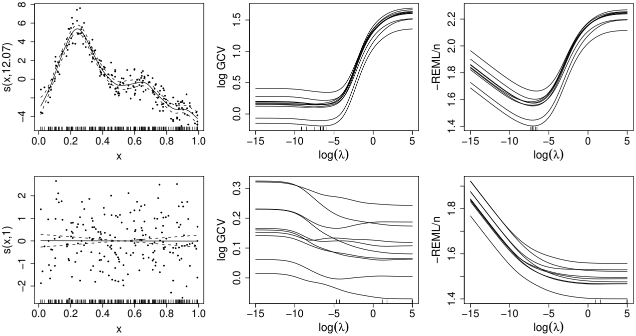
ML computation

- Rather than integrating to find $\pi(\mathbf{y}|\lambda)$ we can use the identity
$$\pi(\mathbf{y}|\lambda) = \pi(\mathbf{y}|\hat{\boldsymbol{\beta}})\pi(\hat{\boldsymbol{\beta}}|\lambda)/\pi(\hat{\boldsymbol{\beta}}|\mathbf{y},\lambda),$$
i.e. $\log\pi(\mathbf{y}|\lambda) = \log\pi(\mathbf{y}|\hat{\boldsymbol{\beta}}) + \log\pi(\hat{\boldsymbol{\beta}}|\lambda) - \log\pi(\hat{\boldsymbol{\beta}}|\mathbf{y},\lambda)$.
- All the $\pi(\cdot)$ are Gaussian, and plugging them in, in turn, yields[‡]
$$2\log\pi(\mathbf{y}|\lambda) = -\frac{\|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \lambda\hat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{S}\hat{\boldsymbol{\beta}}}{\sigma^2} + \log|\lambda\mathbf{S}/\sigma^2|_+ - \log|\mathbf{X}^{\mathsf{T}}\mathbf{X}/\sigma^2 + \lambda\mathbf{S}/\sigma^2| - n\log(2\pi\sigma^2)$$
— note the additional indirect dependence on λ via $\hat{\boldsymbol{\beta}}$.
- $\log\pi(\mathbf{y}|\lambda)$ can be (numerically) optimized w.r.t. λ and σ^2 to estimate these. It is also sometimes referred to as *REML*.

[‡] $|\mathbf{B}|_+$ is the product of the positive eigenvalues of $\mathbf{B}$.

ML versus Cross Validation

- The marginal likelihood typically has a more pronounced optimum than cross validation criteria, and less chance of developing multiple optima, as these simulations show…



- In consequence it is less prone to occasional severe undersmoothing.

Effective degrees of Freedom

- To optimize λ, differentiate $2\log \pi(\mathbf{y}|\lambda)$ w.r.t. λ and set to zero[§]

$$-\hat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{S}\hat{\boldsymbol{\beta}}/\sigma^2 + \text{tr}(\mathbf{S}^{-}\mathbf{S}/\lambda) - \text{tr}\{(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{S}/\sigma^2\} = 0$$

- To optimize σ^2, differentiate $2\log \pi(\mathbf{y}|\lambda)$ w.r.t. σ^2 and set to zero. Noting the preceding equality this yields

$$\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/\sigma^2 + \text{tr}\{(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{X}\} - n = 0$$

- So $\hat{\sigma}^2 = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/[n - \text{tr}\{(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{X}\}]$ suggesting treating $\text{tr}\{(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{X}\}$ as the *Effective Degrees of Freedom* of the smooth model.
- The EDF varies smoothly from p at $\lambda = 0$ to the rank deficiency of $\mathbf{S}$ as $\lambda \to \infty$. This corresponds to the previous example smooth varying from something very wiggly to a straight line fit.

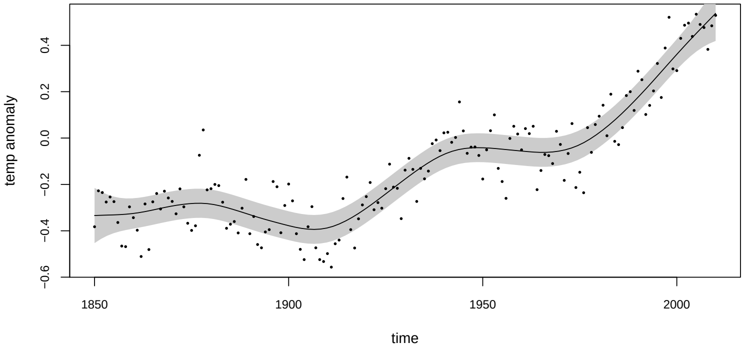
[§] note: the derivatives of $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}\boldsymbol{\beta}$ w.r.t. $\boldsymbol{\beta}$ are zero at $\hat{\boldsymbol{\beta}}$, by definition.

Effective Degrees of Freedom and shrinkage

- Without penalization the coefficient estimates would be $\tilde{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$.
- With penalization they are $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$.
- So $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}\tilde{\boldsymbol{\beta}}$.
- Hence the leading diagonal elements of $(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}$ are $\partial\hat{\beta}_i/\partial\tilde{\beta}_i$ and can be thought of as shrinkage factors.
- So when we sum them up to get the EDF, the result is $p\times$ the average shrinkage factor.
- Note that $\text{tr}\{(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}\} = \text{tr}\{\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{X}^\mathsf{T}\}$, from general properties of the trace.
- For the last example smooth plotted the EDF was almost exactly 11 (but generally there is no reason for it to be integer).

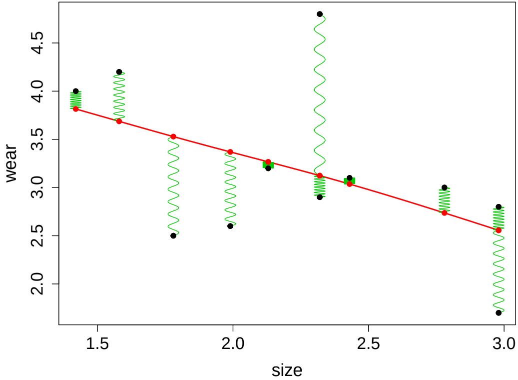
Example

- If this is all a bit abstract, here is a penalized spline smoother with marginal likelihood λ estimation and 95% Bayesian credible interval applied to separating weather from climate in the global temperature series (from the last IPCC report) ...



Why spline bases?

- In introducing penalized basis expansions, B-splines were chosen for their ‘convenient properties’. Why exactly?
- To answer this imagine physically representing f by a flexible strip (e.g. of wood) attached to the data with vertical springs.
- Now consider what happens if the stiffness of the strip is varied:



Splines

- The strip (known as a spline) adopts the position minimising the sum of its bending energy and the energy stored in the springs.
- Mathematically[¶] that is

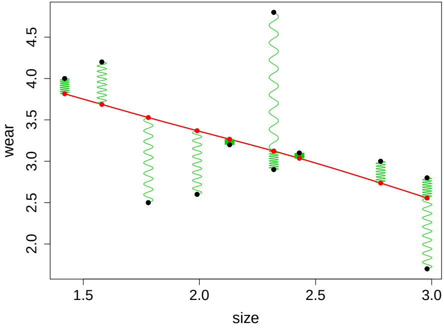
$$\hat{f} = \underset{f}{\text{argmin}} \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int f''(x)^2 dx \tag{1}$$

- Notice that the optimization is over all smooth functions — no basis is being assumed up front.
- In other words: we decide what we mean by 'fitting the data' and what we mean by 'smooth' and seek the *function* optimizing a weighted sum of lack of fit and lack of smoothness.
- It turns out that the solution to (1) can be represented with an n dimensional basis of known functions (independent of λ).

[¶]there is some idealisation here: the spline deformation is assumed small, and we use special vertical extension mathematical springs with zero energy at zero length.

Large deformations

- Obviously once we have defined the spline mathematically we don't need to restrict ourselves to the small deformation regime used in formulating the spline objective...



- The basis of piecewise cubic polynomials between adjacent x_is, continuous to 2^{nd} derivative, is correct for (1) by an integration by parts argument. But consider a more general construction.

Spline objective to basis: 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. $|f(t)| \leq M\|f\|_{\mathcal{H}}$.
- 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 and RKHS

- 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 $\phi_1(t), \phi_2(t)$, say.
- Spline problem seeks $\hat{f} \in \mathcal{H}_0 \oplus \mathcal{H}$ to minimize

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

where P is the projection into $\mathcal{H}$.

Smoothing basis and reproducing kernels

- $\hat{f}(t) = \sum_{i=1}^{n} c_i R_{t_i}(t) + \sum_{i=1}^{2} d_i \phi_i(t)$. 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. $\|P\tilde{f}\|_{\mathcal{H}}^2 = \|P\hat{f}\|_{\mathcal{H}}^2 + \|\eta\|_{\mathcal{H}}^2$ 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

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

- RKHS approach is elegant and general, but at $O(n^3)$ cost.

Other spline basis properties

- Obviously any invertible linear combination of spline basis functions defines a valid basis, we are free to choose.
- The B-splines used earlier are one such choice: they have good numerical stability and *compact support*, meaning that they are zero, apart from over some finite portion of the real line. This leads to sparse $\mathbf{X}$ matrices, for example.
- Another important property of splines is good approximation theoretic properties.
- Suppose we use a cubic spline basis to *interpolate* observations of a smooth function $g(x)$ spaced at most h apart on the x axis. Then $|g(x) - \hat{f}(x)| = O(h^4)$.
- Typically $h \propto n^{-1}$ where n is number of observations. $O(n^{-4})$ is a rather high rate!

Reduced rank smoothing bases

- The full spline bases have dimension n. In many applications this leads to $O(n^3)$ computational cost. Is it really necessary?
- We could use a spline basis constructed for a size $p < n$ set of nicely spaced data ('knots') to model the whole size n dataset[‖].
- In the unpenalized cubic spline basis case this entails an approximation error/bias of $O(p^{-4})$.
- The standard deviation of such a fit is the $O(\sqrt{p/n})$ of regression.
- So to minimize MSE asymptotically we need $p \propto n^{1/9}$.
- In the penalized case $p \propto n^{1/5}$ is about right. Clearly $p = n$ is indeed statistically wasteful.
- In practice we either choose p points to use for basis construction, or use rank p eigen-approximations.

[‖] which is what was done in the preceding examples!

Sum to zero constraints

- Often it is useful to include a smooth function $f(x)$ in a larger model that already includes an intercept, α.
- Identifiability problem! We can not estimate α *and* $f(x)$ without a constraint.
- $\alpha = 0$ doesn't help if we want to add in another smooth function.
- A better option is to constrain $f(x)$ with a sum-to-zero constraint

$$\sum_{i=1}^{n} f(x_i) = 0 \Rightarrow \mathbf{1}^{\mathsf{T}}\mathbf{X}\boldsymbol{\beta} = 0$$

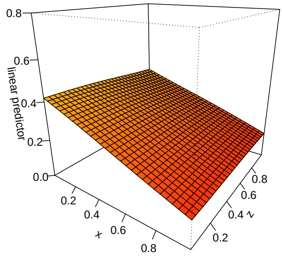
- An obvious way to meet the right hand version is to subtract its mean from each column of $\mathbf{X}$ (there are alternatives of course).
- No change in f's shape: we just shift basis functions up or down.
- But it leaves the centred $\mathbf{X}$ rank deficient by one, as its intercept component has been eliminated. To restore full rank, drop the least variable column** of the centred $\mathbf{X}$ (+ associated parameter).

**the 'least variable' part enhances numerical stability and ensures we never leave in a 0 column.

Multi-dimensional smooths

- The obvious way to generalize from one dimensional smoothing to multidimensional is to base splines on a multidimensional analogue of 1D spline penalties.
- Thin plate splines do that with an isotropic penalty:

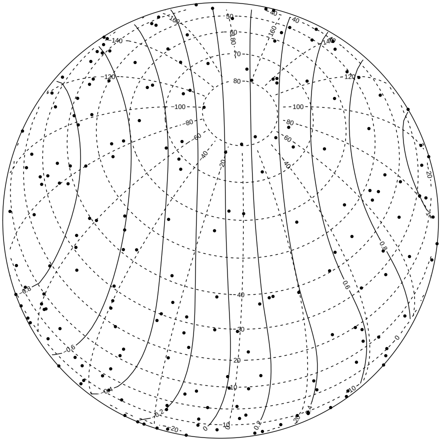
$$\lambda \int f_{xx}^2 + 2f_{xz}^2 + f_{zz}^2 dxdz \quad \text{(2D second order example)}$$



- Different dimensions and orders of derivative are also possible.

Other geometries

- …are possible. A thin plate spline on the sphere for example.

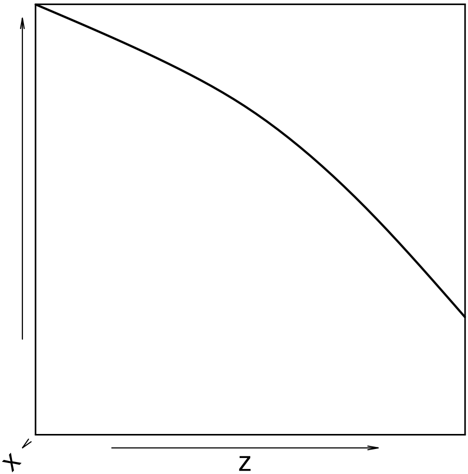


Smooth interactions

- If the arguments of a smooth measure different types of quantities (e.g. distance and time) then it makes no sense to treat them isotropically as a thin plate spline does.
- We don't know what their relative scaling should be[††].
- But scale invariant smooth interactions can be constructed by combining 1D splines.
- The trick is to apply the usual statistical notion of an interaction between variables, x and z, say. In particular
    1. The effect of z is itself dependent on x.
    2. i.e. the parameters for the z effect vary with x.
- Given basis expansions for the smooth effects $f_z(z)$ and $f_x(x)$ this idea is easily applied to smooths.
- Simply let the coefficients of f_z be smooth functions of x...

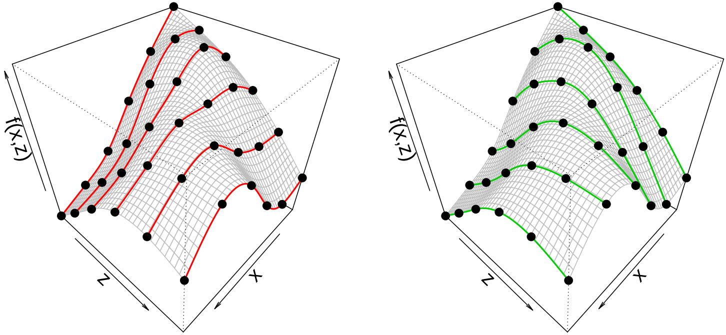
[††] doing something arbitrary like scaling to the unit square assumes we do know.

Tensor product basis construction



Tensor product penalties

- To avoid relative scaling assumptions, we need a separate penalty with its own smoothing parameter for each covariate direction.
- For example, sum up the spline penalties for the red curves and the green curves separately.



Mathematical formulation of tensor product smooths

- Let $b_{zj}(z)$ and $b_{xi}(x)$ be the basis functions for f_z and f_x with penalty matrices $\mathbf{S}_x$ and $\mathbf{S}_z$. The *marginal* smoothers.
- The tensor product basis construction shown above gives:

$$f(x,z) = \sum_i \sum_j \beta_{ij} b_{zj}(z) b_{xi}(x)$$

- With double penalties

$$\boldsymbol{\beta}^{\mathsf{T}} \mathbf{I} \otimes \mathbf{S}_z \boldsymbol{\beta} \text{ and } \boldsymbol{\beta}^{\mathsf{T}} \mathbf{S}_x \otimes \mathbf{I} \boldsymbol{\beta}$$

- The construction generalizes to any number of marginals and multi-dimensional marginals.
- Can start from any marginal bases & penalties (including mixtures of types).

Smooth ANOVA

- Sometimes people like to separate a multi-dimensional smooth into main effects and interactions. e.g.

$$f_x(x) + f_z(z) + f_{xz}(x, z)$$

- For identifiability we must exclude the basis for functions $f_x(x) + f_z(z)$ from the basis for $f_{xz}(x, z)$.
- Easily done using exactly the mechanism used in parametric statistical models: apply sum-to-zero identifiability constraints to the marginal bases used to construct $f_{xz}(x, z)$.
- The constraint removes the constant function from the basis for f_x, so that its product with the basis for f_z does not include a copy of the f_z basis (and vice versa).

Isotropy versus scale invariance

- Smooth fits to data. In the bottom row the x variable has been divided by 5 before fitting. TPS is drastically affected by the scaling and the tensor product smooth not at all.