Smoothing, penalized least squares and splines

Douglas Nychka,
www.image.ucar.edu/~nychka

- Locally weighted averages
- Penalized least squares smoothers
- Properties of smoothers
- Splines and Reproducing Kernels
- The interpolation proof
- CV and the smoothing parameter

Supported by the National Science Foundation *L' Ecole D'Ete, Ovronnaz Sep 2005*

Estimating a curve or surface.

The additive statistical model:

Given n pairs of observations (x_i, y_i), $i = 1, \ldots, n$

$$y_i = g(x_i) + \epsilon_i$$

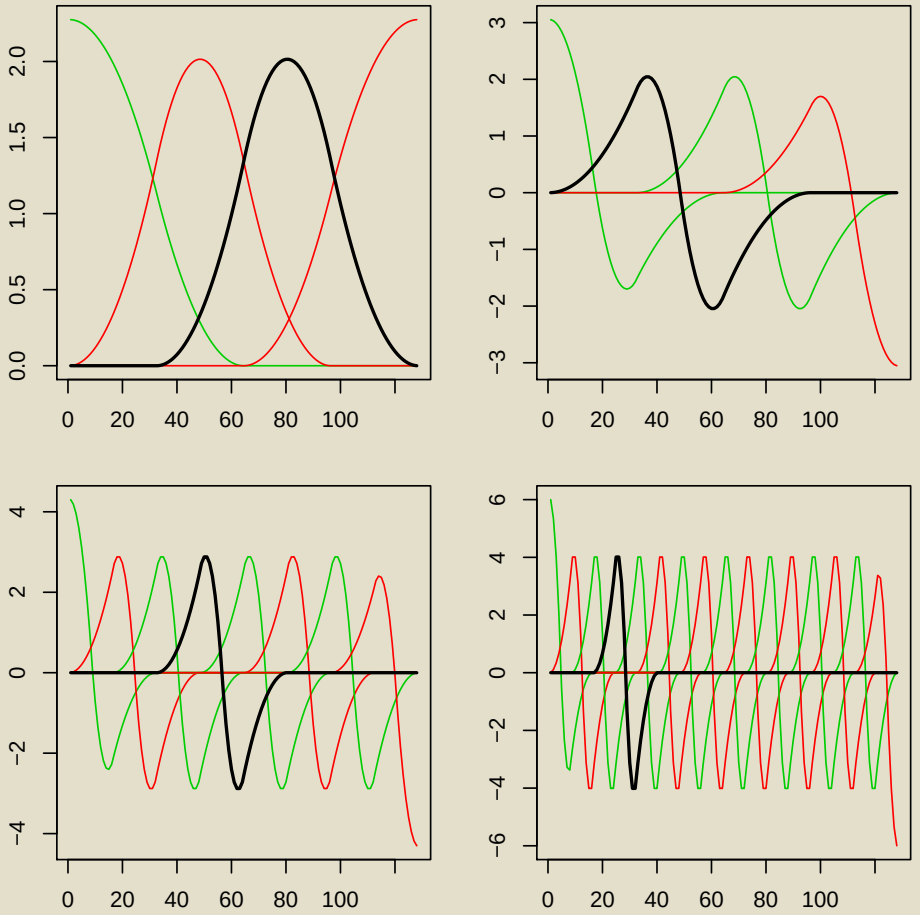
ϵ_i's are random errors
and g is an unknown smooth function.

The goal is to estimate a function g based on the observations

A 2-d example

Predict surface ozone where it is not monitored.

Ambient daily ozone in PPB June 16, 1987, US Midwestern Region.



Local linear surface estimates

Use local information to predict unobserved values

Use a linear regression based on close by observations.

$$y_i = \beta_1 + \textbf{lon}_i\beta_2 + \textbf{lat}_i\beta_3 + \epsilon_i$$

find $\widehat{\beta}$ by least squares.

The prediction at location $(-88, 41)$ is just a weighted average of the observations.

$$\widehat{g}(-88, 41) = \widehat{\beta}_1 \ + \textbf{-88}\widehat{\beta}_2 + \textbf{41}\widehat{\beta}_3 = \sum_{i=1,n} w_i y_i$$

A kernel estimator

Determine the weights based on the distance to the prediction points

$$w_i \sim (1/h)K((\boldsymbol{x} - \boldsymbol{x}_i)/h)$$

(and normalize so that the weights sum to one.)

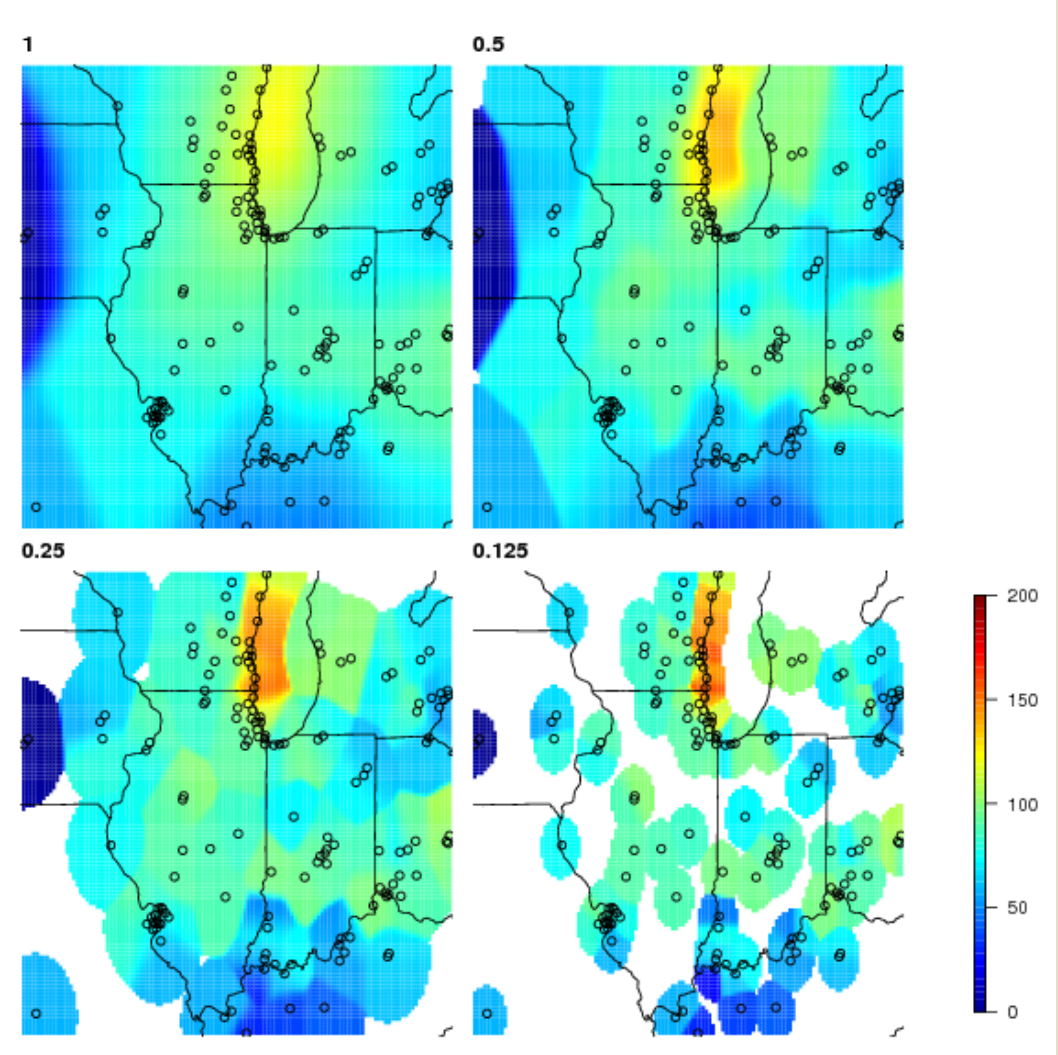
Kernel: K

K is bump shaped e.g. a normal

Bandwidth: h

h controls the spread of K as h gets large the estimate is just the average.

Some kernel estimates for ozone



Linear smoothers

Let $\hat{g} = g(x_1), ..., g(x_n)$ be the prediction vector at the observed points.

A smoother matrix satisfies

$\hat{g} = Ay$ where

- A is an $n \times n$ matrix
- eigenvalues of A are in the range [0,1].

Note: $||Ay|| \leq ||y||$

Usually values in between the data are filled in by interpolating the predictions at the observations.

Problems with local regression and kernels

- How large should the neighborhood/bandwidth be?
- What is the uncertainty of the prediction?
- Predicting in between observations is *ad hoc* and can get weird when the error is small.

Problems with local regression and kernels

- How large should the neighborhood/bandwidth be?
- What is the uncertainty of the prediction?
- Predicting in between observations is *ad hoc* and can get weird when the error is small.

But the theoretical properties of kernel estimators are well understood ...

$$E\left[g(\boldsymbol{x}) - \widehat{g}(\boldsymbol{x})\right]^2 = h^4 K_2/4 + \frac{\sigma^2}{nh} K_0$$

$$\text{MSE} = \text{Bias}^2 + \text{Variance}$$

Penalized least squares

Ridge regression

Start with your favorite n basis functions $\{\psi_k\}_{k=1}^n$ The estimate has the form

$$f(x) = \sum_{l=1}^{n} \theta_k \psi_k(x)$$

where $\theta = (\theta_1, \ldots, \theta_n)$ are the coefficients.

Let $W_{i,k} = \psi_k(x_i)$ so $\boldsymbol{f} = W\boldsymbol{\theta}$

Estimate the coefficients by a penalized least squares

Sum of squares(θ) + penalty on θ

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (\boldsymbol{y} - [W\boldsymbol{\theta}]_i)^2 + \lambda \boldsymbol{\theta}^T B \boldsymbol{\theta}$$

with $\lambda > 0$ a hyperparameter and B a nonnegative definite matrix.

Estimate the coefficients by a penalized least squares

Sum of squares(θ) + penalty on θ

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (\boldsymbol{y} - [W\boldsymbol{\theta}]_i)^2 + \lambda \boldsymbol{\theta}^T B \boldsymbol{\theta}$$

with $\lambda > 0$ a hyperparameter and B a nonnegative definite matrix.

or in general,

- log likelihood + λ penalty on θ

In any case once we have the parameter estimates these can be used to evaluate $\widehat{g}$ at any point.

The form of the smoother matrix

Just calculus ...

- Take derivatives of the penalized likelihood w/r to θ,
- set equal to zero,
- solve for θ

The form of the smoother matrix

Just calculus ...

- Take derivatives of the penalized likelihood w/r to θ,
- set equal to zero,
- solve for θ

The monster ...

$$\hat{\boldsymbol{\theta}} = (W^T W + \lambda B)^{-1} W^T \boldsymbol{y}$$

$$\hat{\boldsymbol{g}} = W\hat{\boldsymbol{\theta}} = W(W^T W + \lambda B)^{-1} W^T \boldsymbol{y} = A(\lambda)\boldsymbol{y}$$

Why is this a smoothing matrix?

If W is symmetric

$$A(\lambda) = W(W^T W + \lambda B)^{-1} W^T = (I + \lambda W^{-1} B W^{-1})^{-1}$$

Effective degrees of freedom in the smoother

For linear regression trace $A(\lambda)$ gives us the number of parameters. (Because it is a projection matrix)

By analogy, $\mathrm{tr}A(\lambda)$ is measure of the effective number of degrees of freedom attributed to the smooth surface

Effective degrees of freedom in the smoother

For linear regression trace $A(\lambda)$ gives us the number of parameters. (Because it is a projection matrix)

By analogy, tr$A(\lambda)$ is measure of the effective number of degrees of freedom attributed to the smooth surface

A useful decomposition

Recall: $A(\lambda) = (I + \lambda W^{-1} B W^{-1})^{-1}$

One can always find an orthogonal matrix, U so that

$U^T U = I$ and $(W^{-1} B W^{-1}) = U \Gamma U^T$

where Γ is diagonal.

$$A(\lambda) = (I + \lambda U \Gamma U^T)^{-1} = U(I + \lambda \Gamma)^{-1} U^T$$

A simple formula for the trace

So

$$\mathbf{tr}A(\lambda) = \mathbf{tr}U(I + \lambda\Gamma)^{-1}U^T = \mathbf{tr}(I + \lambda\Gamma)^{-1}U^TU$$

$$= \mathbf{tr}(I + \lambda\Gamma)^{-1} = \sum_{i=1}^{n} \frac{1}{1 + \lambda\Gamma_{ii}}$$

The Gamma's are all nonnegative so this must be increasing as λ decreases.

The relationship is one-to-one with λ and independent of the data so we can always talk about λ in terms of the effective degrees of freedom.

Splines

One obtains a spline estimate using a specific basis and a specific penalty matrix. Splines are confusing because the basis is a bit mysterious.

The classic cubic smoothing spline:

For curve smoothing in one dimension,

$$\min_f \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i))^2 + \lambda \int (f''(x))^2 dx$$

The second derivative measures the roughness of the fitted curve.

Splines

One obtains a spline estimate using a specific basis and a specific penalty matrix. Splines are confusing because the basis is a bit mysterious.

The classic cubic smoothing spline:

For curve smoothing in one dimension,

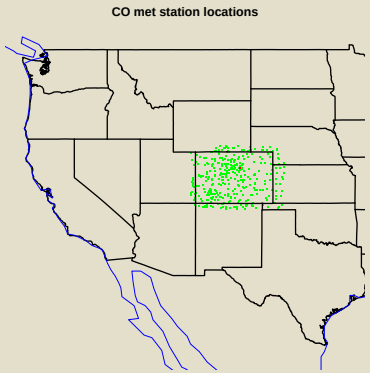
$$\min_f \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i))^2 + \lambda \int (f''(x))^2 dx$$

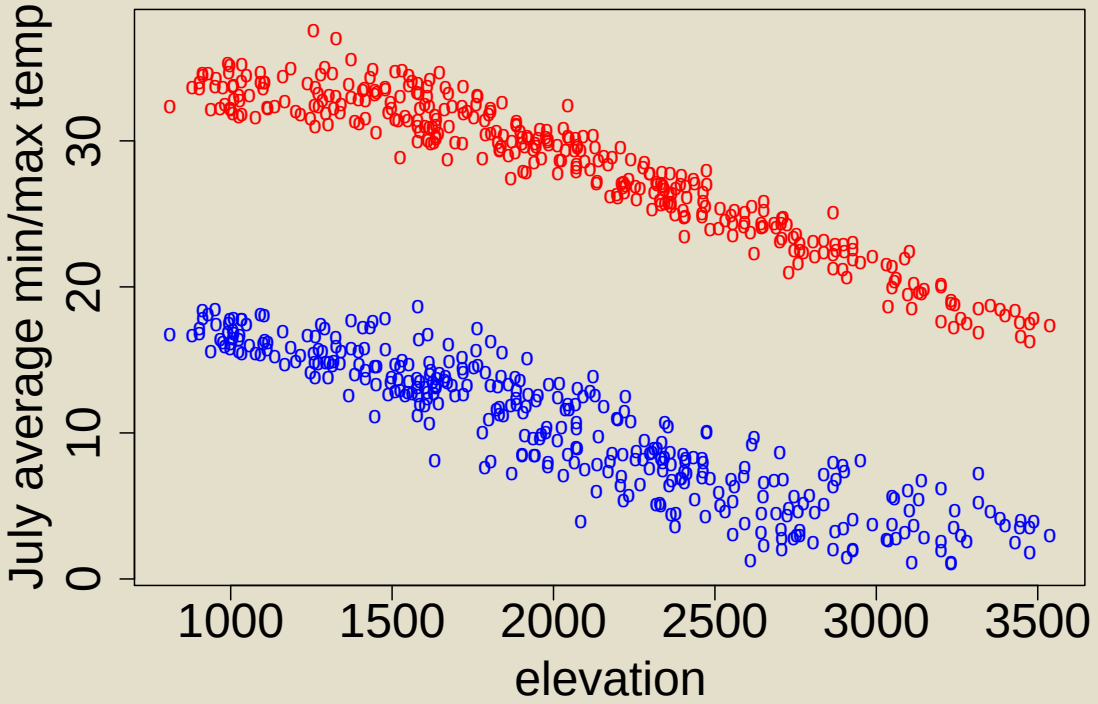
The second derivative measures the roughness of the fitted curve.

The solution, is continuous up to its second derivative and is a piecewise cubic polynomial in between the observation points.

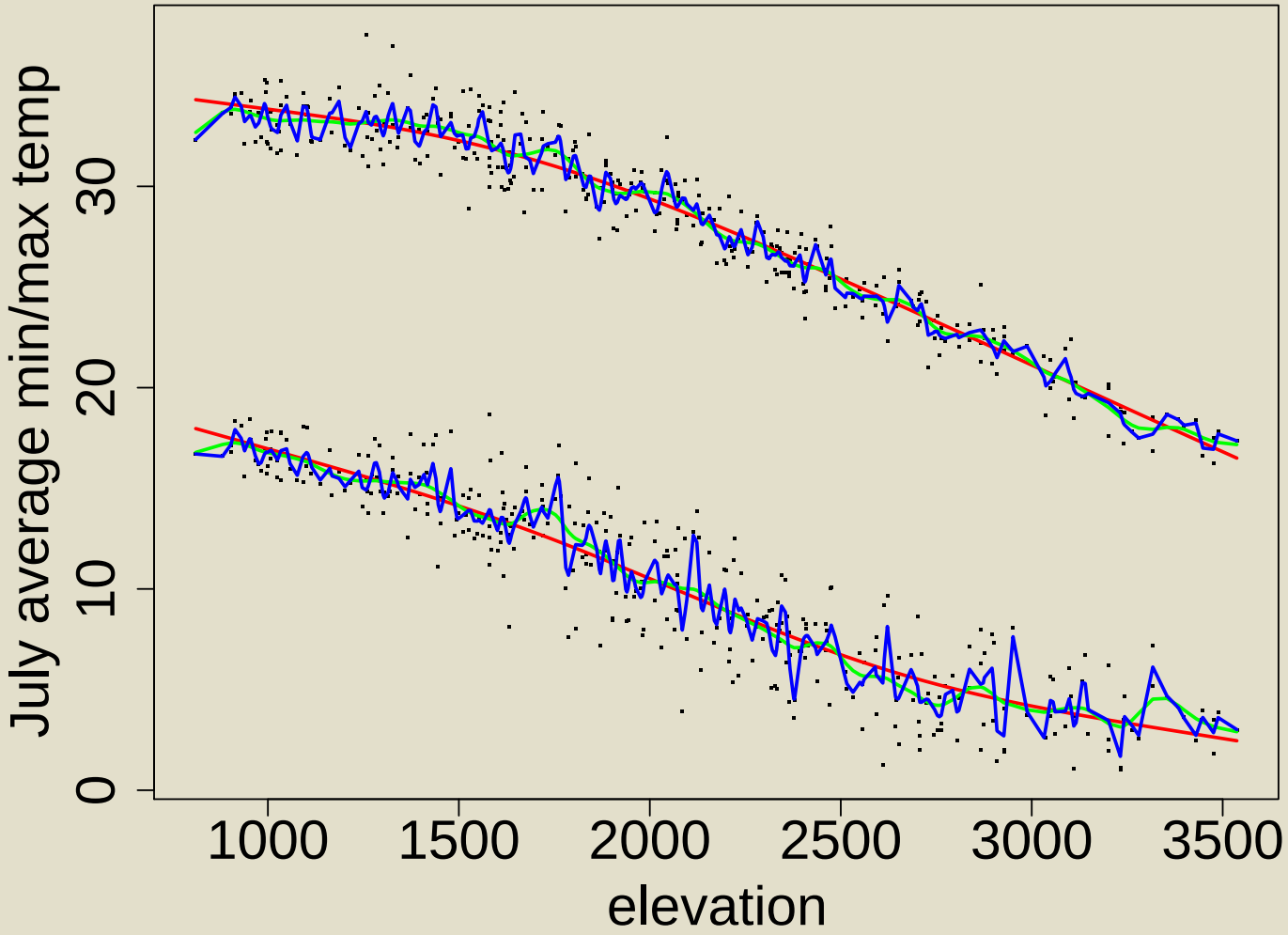
Where does this come from?

Climate for Colorado





Cubic splines with different λ s



Some abstraction

Reproducing kernels

Think of these as covariance functions ... $k(\boldsymbol{x}, \boldsymbol{x}')$

To get basis functions we will hold one argument fixed at some value of $\boldsymbol{x}$ *and let the other vary.*

A Space of functions

Let $\mathcal{H}$ **be the smallest Hilbert space with inner product** $\langle \cdot, \cdot \rangle$ **such that** $k(x, \cdot) \in \mathcal{H}$ **for all** x **and**

$$\langle k(x, \cdot), k(x', \cdot) \rangle = k(x, x')$$

k reproduces itself!

Some abstraction

Reproducing kernels

Think of these as covariance functions ... $k(\boldsymbol{x}, \boldsymbol{x}')$

To get basis functions we will hold one argument fixed at some value of $\boldsymbol{x}$ *and let the other vary.*

A Space of functions

Let $\mathcal{H}$ **be the smallest Hilbert space with inner product** $\langle \cdot, \cdot \rangle$ **such that** $k(x, \cdot) \in \mathcal{H}$ **for all** x **and**

$$\langle k(x, \cdot), k(x', \cdot) \rangle = k(x, x')$$

k reproduces itself!

For f **in** $\mathcal{H}$**,** $\langle k(x, \cdot), f \rangle = f(x)$ **!**

A variational problem

With $\mathcal{H}$ and its inner product,

$$\min_{f \in \mathcal{H}} \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i))^2 + \lambda \langle f, f \rangle$$

A variational problem

With $\mathcal{H}$ and its inner product,

$$\min_{f \in \mathcal{H}} \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i))^2 + \lambda \langle f, f \rangle$$

The solution

Choose basis functions $k(., \boldsymbol{x}_i)$ for $1 \leq i \leq n$

$$\widehat{f}(x) = \sum_{i=1}^{n} \widehat{\boldsymbol{\theta}}_i k(x, x_i)$$

Choose roughness penalty matrix $B = W$

A variational problem

With $\mathcal{H}$ and its inner product,

$$\min_{f \in \mathcal{H}} \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i))^2 + \lambda \langle f, f \rangle$$

The solution

Choose basis functions $k(., \boldsymbol{x}_i)$ for $1 \leq i \leq n$

$$\widehat{f}(x) = \sum_{i=1}^{n} \widehat{\boldsymbol{\theta}}_i k(x, x_i)$$

Choose roughness penalty matrix $B = W$

this gives

$$\widehat{\boldsymbol{\theta}} = (WW^T + \lambda W)^{-1} W^T \boldsymbol{y} = (W + \lambda I)^{-1} \boldsymbol{y}$$

$$\widehat{g} = W\widehat{\boldsymbol{\theta}} = W(W + \lambda I)^{-1} \boldsymbol{y} = (I + \lambda W^{-1})^{-1} \boldsymbol{y}$$

Covariances

Here $W_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$
this looks like a covariance matrix doesn't it.

The Proof

The main part is to use the minimum interpolation properties of splines. It is easiest to prove this in more generality for any reproducing kernel Hilbert space.

A simple proof is to guess at the minimizing solution and use the optimal interpolation results to show that there can not be another solution. – *more on this later.*

A 1-d cubic smoothing spline

The key step is a decomposition: $f(x) = \beta_1 + \beta_2 x + h(x)$

$$\min_f \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i))^2 + \lambda \int (f''(x))^2 dx$$

becomes

$$\min_{\beta,h} \sum_{i=1}^{n} (y_i - \beta_1 + \beta_2 x_i + h(x_i))^2 + \lambda \int (h''(x))^2 dx$$

$\mathcal{H}$:

All functions with 2 derivatives

Inner product:

related to the integral

Usual (Wahba style) Reproducing Kernel:

$k(\boldsymbol{x}, \boldsymbol{x}') = |\boldsymbol{x} - \boldsymbol{x}'|^3$ **+ linear terms**

A less scary reproducing kernel for the same cubic spline problem

On the interval $[0,1]$

Let $k(u,v) = u^2v/2 - u^3/6$ **for** $u < v$
and
$k(u,v) = v^2u/2 - v^3/6$ **for** $u \geq v$

$k(.,v)$ **for fixed** v**, is a function that is a cubic polynomial from** 0 **to** v **and is then a linear function for** $u > v$**. It is twice differentiable.**

Let $\mathcal{H}$ **be a Hilbert space of functions on** $[0,1]$ **where** $f(0) = 0$ **and** $f'(0) = 0$ **and with inner product**

$$\langle f, g \rangle = \int_0^1 f''(u)g''(u)du$$

k is a reproducing kernel for this space!

Show this just using integration by parts.

More on cubic splines

To be explicit the cubic smoothing spline solution has the form:

$$\hat{f}(x) = \hat{\beta}_1 + \hat{\beta}_2 x + \sum_{i=1}^{n} \hat{\theta}_i k(x, x_i)$$

Besides verifying that the reproducing on the previous page works with the integrated second derivative inner product one can also check that it is the covariance function for integrated Brownian motion,$B(t)$:

$$X(u) = \int_0^u B(t)dt,$$

with B(0)=0. i.e.

$$E(X(u)X(v)) = k(u, v)$$

So as we will see later, the Bayesian prior associated with the cubic smoothing spline is an integrated Brownian motion.

The last details ...

Given the reproducing kernel, the problem is

$$\min_{\boldsymbol{\beta},\boldsymbol{\theta}} ||\boldsymbol{y} - X\boldsymbol{\beta} - W\boldsymbol{\theta}||^2 + \lambda\boldsymbol{\theta}^T W\boldsymbol{\theta}$$

X is the regression matrix with columns 1 and $\{x_i\}$

First minimize over θ with β fixed. Plug in solution and then minimize over β.

$$\widehat{\boldsymbol{\beta}} = (X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1}\boldsymbol{y}$$

$\Omega = (W + \lambda I)$

$$\widehat{\boldsymbol{\theta}} = (W + \lambda I)^{-1}(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}})$$

A 2-d thin plate smoothing spline

$$\min_f \sum_{i=1}^{n} (y_i - f_i)^2 + \lambda \int_{\Re^2} \left(\frac{\partial^2 f}{\partial^2 u}\right)^2 + 2\left(\frac{\partial^2 f}{\partial u \partial v}\right)^2 + \left(\frac{\partial^2 f}{\partial^2 v}\right)^2 du dv$$

Collection of second partials is invariant to a rotation.

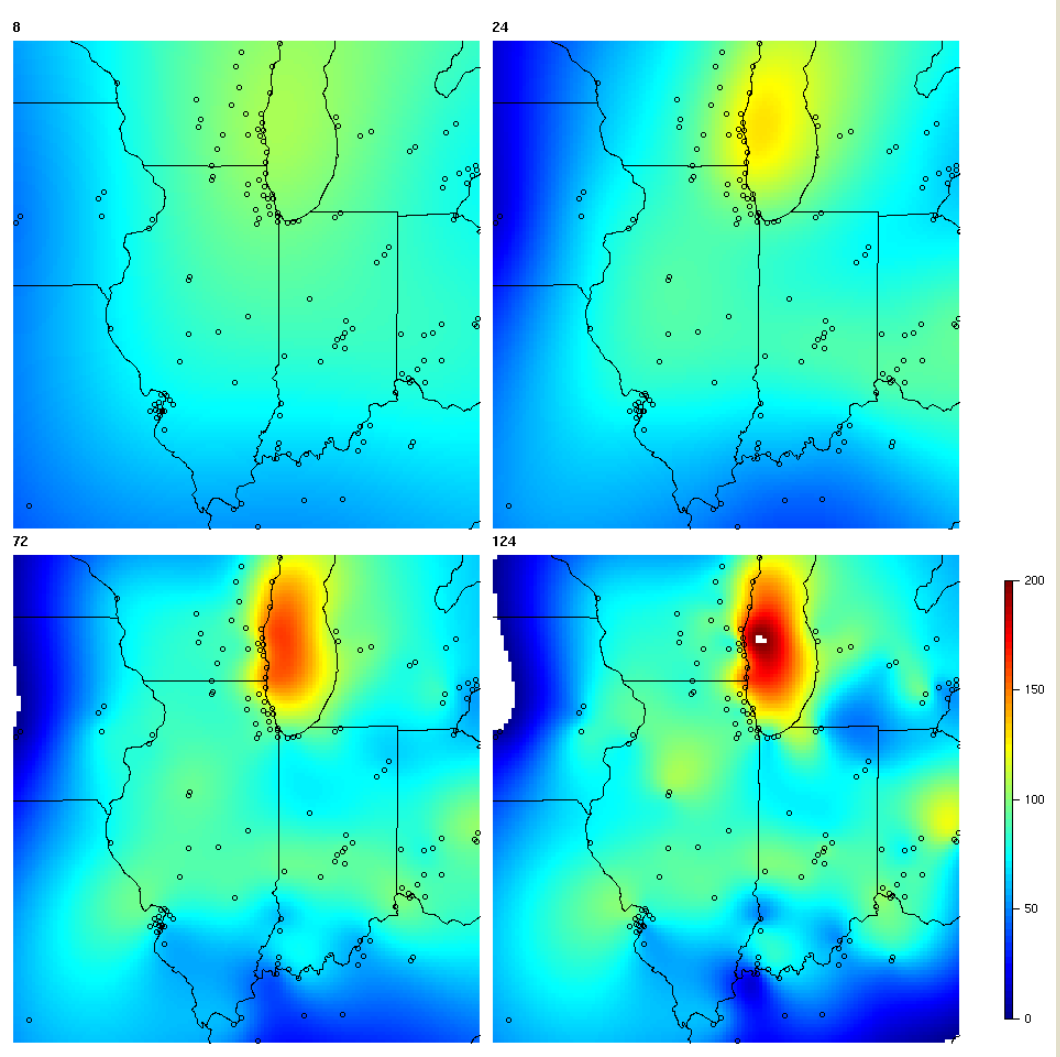
Again, separate off the linear part of f.
$f(x) = \beta_1 + \beta_2 x_1 + \beta_3 x_2 + h(\boldsymbol{x})$

Reproducing Kernel:

$$k(\boldsymbol{x}, \boldsymbol{x}') = ||\boldsymbol{x} - \boldsymbol{x}'||^2 log(||\boldsymbol{x} - \boldsymbol{x}'||) + \textbf{linear terms}$$

leading to basis functions that are bumps at the observation locations.

Some thin plate splines for the ozone data



The interpolation problem: a editorial

Splines arose in numerical analysis with the interpolation problem:

Find a curve g *that minimizes* $\int (g''(x))^2 dx$,
subject to $g(x_i) = y_i$

The interpolation problem: a editorial

Splines arose in numerical analysis with the interpolation problem:

Find a curve g *that minimizes* $\int (g''(x))^2 dx$,
subject to $g(x_i) = y_i$

We already know how to do this by letting $\lambda \to 0$

- **Statisticians are, of course, suspicious of fitting data exactly**
- **1-d splines have some very fast computational properties** $-$ *that do not extend to higher dimensions!*
- **Assuming a reproducing kernel is equivalent to a model for the unknown function. What is that model?**

The general interpolation problem

Given $\{x_i, y_i\}$ and a reproducing kernel Hilbert space.

Find a function g in the space so that minimizes $\langle g, g \rangle$.

subject to the constraints $g(\boldsymbol{x}_i) = \boldsymbol{y}_i$, $1 \leq i \leq n$

The solution

$$g = \sum_{i=1}^{n} \theta_i k(., x_i)$$

Where θ is determined by solving a system of n linear equations to guarentee the interpolation constraints.

Interpolation Proof

Suppose

$$g = \sum_{i=1}^{n} \theta_i k(., x_i)$$

interpolates $\{x_i, y_i\}$ **and so does some other function** h**.**

We will show that $\langle h, h \rangle > \langle g, g \rangle$**.**

Let $\delta = h - g$ **and so** $\langle \delta, \delta \rangle > 0$

$$\langle h, h \rangle = \langle g + \delta, g + \delta \rangle = \langle g, g \rangle + 2\langle g, \delta \rangle + \langle \delta, \delta \rangle$$

The cool part, using the reproducing property

$$\langle g, \delta \rangle = \langle \sum_{i=1}^{n} \theta_i k(., x_i), \delta \rangle = \sum_{i=1}^{n} \theta_i \langle k(., x_i), \delta \rangle = \sum_{i=1}^{n} \theta_i \delta(x_i) = 0$$

because $\delta(x_i) = h(x_i) - g(x_i) = y_i - y_i = 0$**.**
So $\langle h, h \rangle = \langle g, g \rangle + \langle \delta, \delta \rangle > \langle g, g \rangle$ **Done!**

A proof of the smoothing spline solution

The proof is by contradiction and uses the interpolation result.

Let $\hat{g}$ be the smoothing spline obtained as a linear combination of the kernel basis functions and possibly a linear or low order polynomial. This is found as a penalized smoother by plugging this form into the penalized least squares criterion and minimizing by ordinary calculus.

Suppose there is an h that has a smaller penalized least squares than $\hat{g}$.

Construct a g that interpolates h at the $\boldsymbol{x}_i$ and uses the same basis functions as $\hat{g}$. By the interpolation result we know that $\langle h, h\rangle > \langle g, g\rangle$ and since both h and g have the same residuals sums of squares the penalized least squares criterion will be smaller for g. Thus , h can not be a minimizer.

Because $\hat{g}$ however, has the same form as the interpolant and minimizes the criterion. Thus it must in fact be the minimizer over all functions in the Hilbert space.

Choosing λ by Cross-validation

Sequentially leave each observation out and predict it using the rest of the data. Find the λ that gives the best out of sample predictions.

Refitting the spline when each data point is omitted, and for a grid of λ values is computationally demanding.

Fortunately there is a shortcut.

The magic formula

residual for $g(x_i)$ having omitted y_i

$$(y_i - \widehat{g}_{-i}) = (y_i - \widehat{g}_i)/(1 - A(\lambda))_{i,i}$$

This has a simple form because adding a data pair $(x_i, \widehat{g}_{-1})$ to the data does not change the estimate.

CV and Generalized CV criterion

$CV(\lambda)$

$$(1/n)\sum_{i=1}^{n}(y_i - \widehat{g}_{-i})^2 = (1/n)\sum_{i=1}^{n}\frac{(y_i - \widehat{g}_i)^2}{(1 - A(\lambda))_{i,i})^2}$$

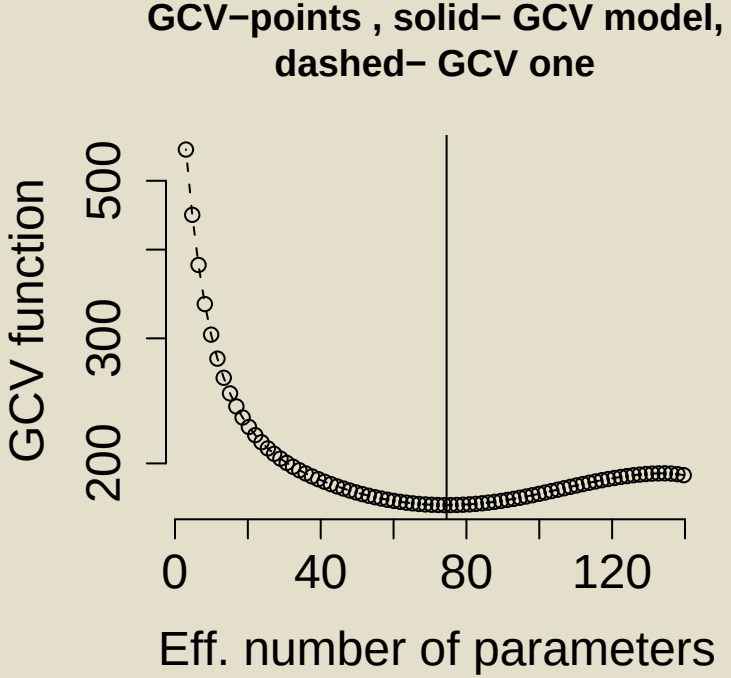
$GCV(\lambda)$

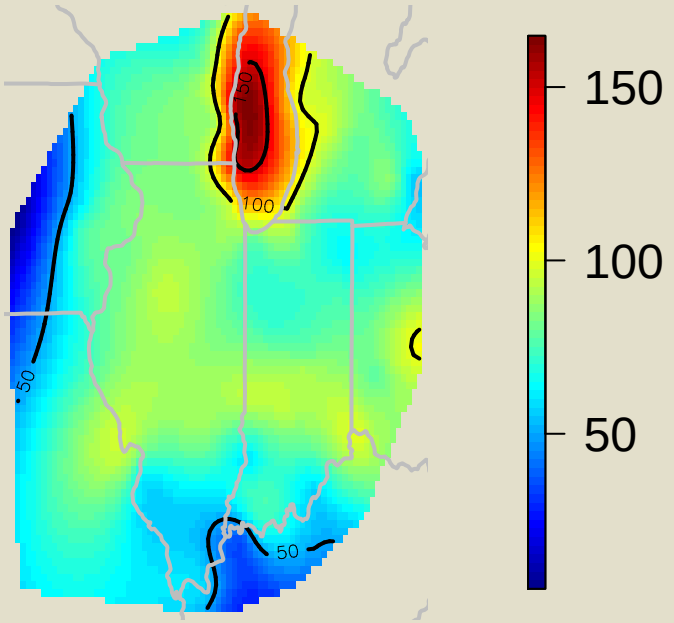
$$(1/n)\frac{\sum_{i=1}^{n}(y_i - \widehat{g}_i)^2}{(1 - \mathbf{tr}A(\lambda)/n)^2}$$

Minimize CV or GCV over λ to determine a good value

GCV for the ozone data

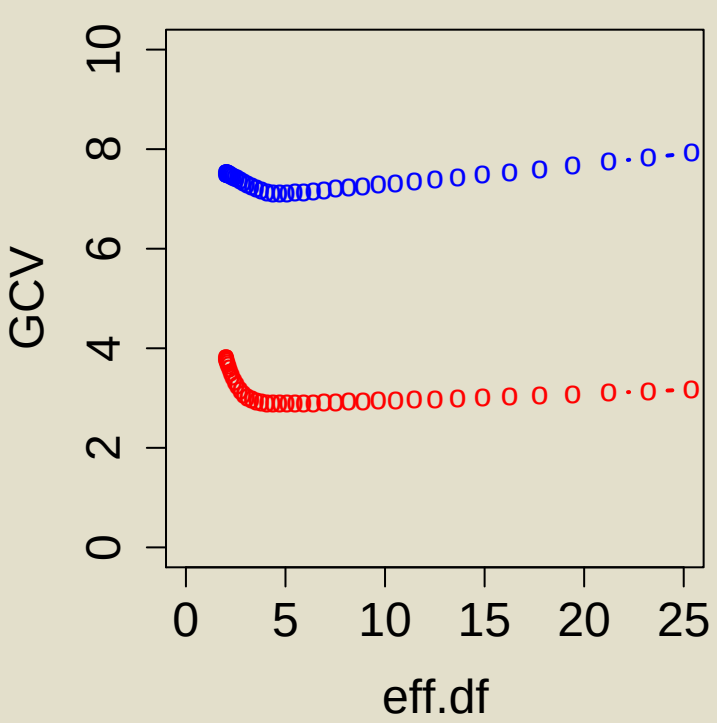
GCV(eff. degrees of freedom), the estimated surface

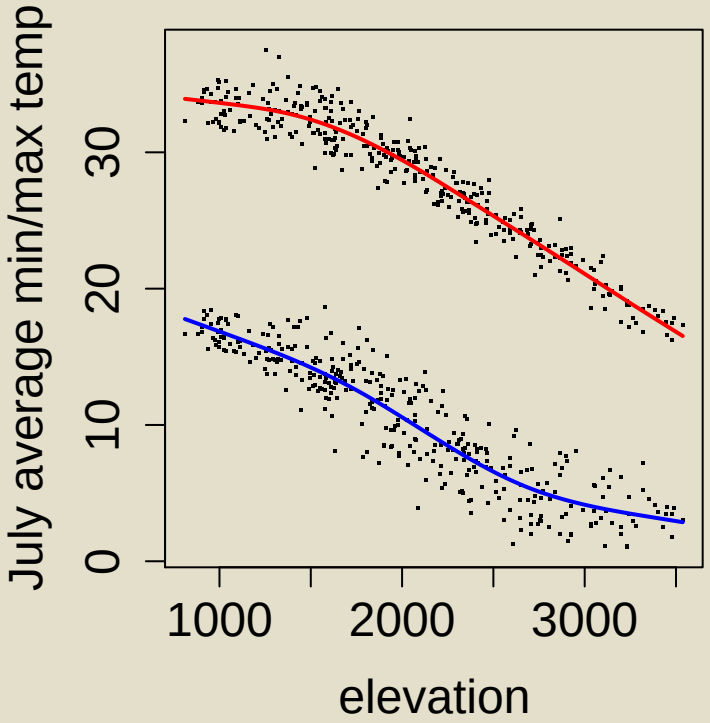




GCV for the climate data

GCV(eff. degrees of freedom), the estimated surface





Summary

We have formulated the curve/surface fitting problem as penalized least squares.

Splines treat estimating the entire curve but also have a finite basis related to a covariance function (reproducing kernel).

One can use CV or GCV to find the smoothing parameter.

Thank you!