Smoothing, penalized least squares and splines

Douglas Nychka,

`www.image.ucar.edu/~nychka`

- **Penalized least squares smoothers**
- **Properties of smoothers**
- **Splines and Reproducing Kernels**
- **CV and the smoothing parameter**

Supported by the National Science Foundation

Seoul National University Oct 2006

Estimating a curve or surface.

The additive statistical model:

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

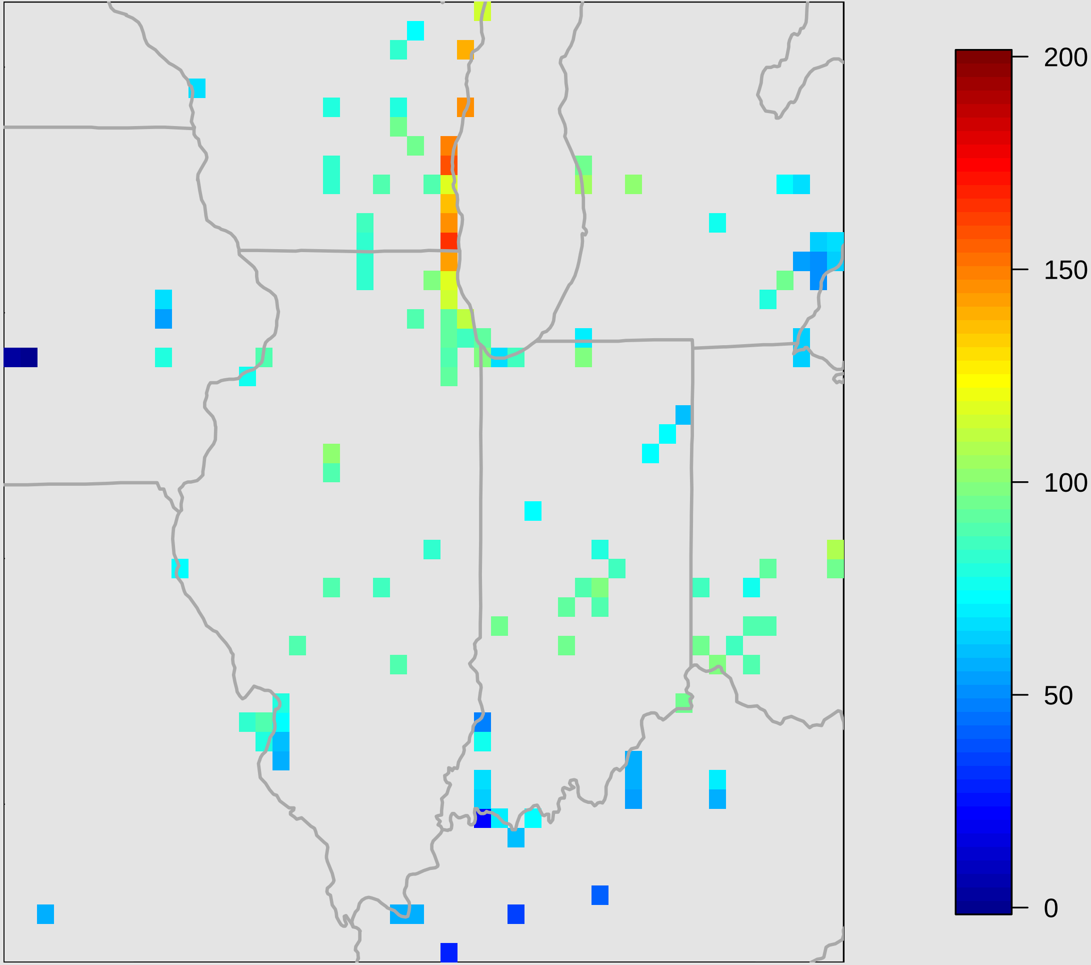
$$y_i = g(x_i) + \epsilon_i$$

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.

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: $||A\boldsymbol{y}|| \leq ||y||$

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

Penalized least squares

Ridge regression

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

$$f(x) = \sum_{l=1}^{n} \beta_k b_k(x)$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)$ are the coefficients.

Let $Xi, k = b_k(x_i)$ so $\boldsymbol{f} = X\boldsymbol{\beta}$

Estimate the coefficients by a penalized least squares

Sum of squares(β) + penalty on β

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (\boldsymbol{y} - [X\boldsymbol{\beta}]_i)^2 + \lambda \boldsymbol{\beta}^T H \boldsymbol{\beta}$$

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

or in general,

- log likelihood + λ penalty on β

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 β,
- set equal to zero,
- solve for β

The monster ...

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

$$\widehat{\boldsymbol{g}} = X\widehat{\boldsymbol{\beta}} = X(X^T X + \lambda H)^{-1} X^T \boldsymbol{y} = A(\lambda)\boldsymbol{y}$$

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 To see why the trace is a good measure we need an alternate form form the ridge regression solution.

A useful decomposition

Find a symmetric positive definite matrix C so that $CX^TXC = I$

U, an orthogonal matrix (i.e. $UU^T = I$) so that $UCHCU^T = D$

Form the magic matrix $G = CU$

$$(XG)^T(XG) = G^T(X^TX)G = I \text{ and } G^THG = D$$

Smoothing matrix

$$A(\lambda) = X(X^TX + \lambda H)^{-1}X^T = XG(I + \lambda D)^{-1}(XG)^T$$

Regression parameters, $\boldsymbol{\beta}$

$$\widehat{\boldsymbol{\beta}}_i = [\boldsymbol{u}]_i/(1 + \lambda D_i)$$

where $\boldsymbol{u} = X^TG\boldsymbol{y}$.

Residual matrix:$I - A(\lambda)$

$$I - A(\lambda) = XG\lambda D(I + \lambda D)^{-1}(XG)^T$$

Effective degrees of freedom

$$tr(A(\lambda)) = tr(XG(I + \lambda D)^{-1}(XG)^T) =$$

$$tr((I + \lambda D)^{-1}(XG)^TXG) = \sum_{i=1}^{n} \frac{1}{1 + \lambda D_k}$$

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:

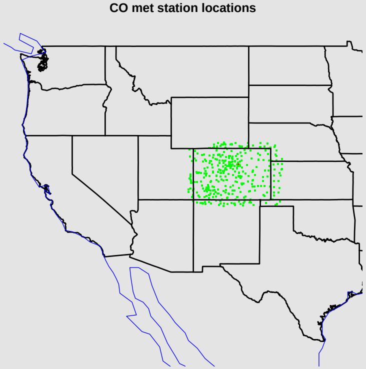
For curve smoothing in one dimension,

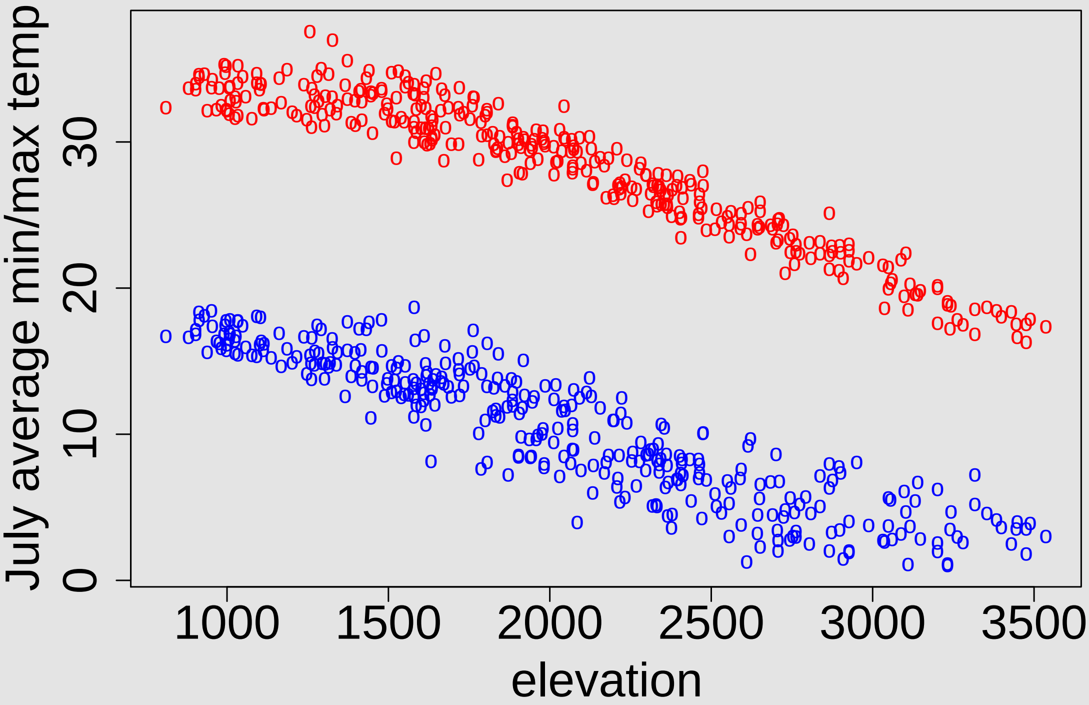
$$\min_{f} \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i))^2 + \lambda \int (f''(x))^2 dx$$

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.

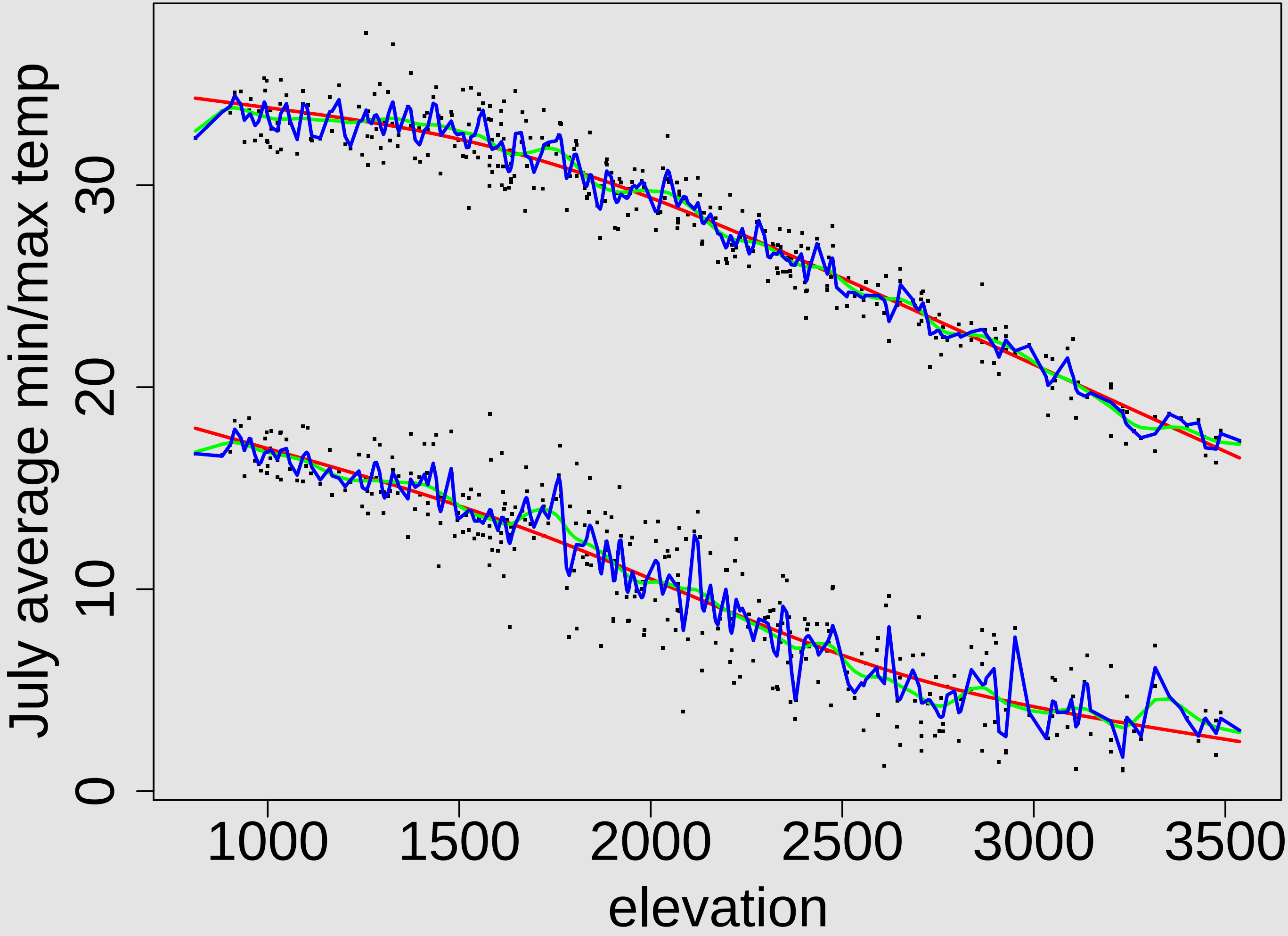
First an example

Climate for Colorado





Cubic splines with different λ s



The fixed and random part of the model

g = low dimensional parametric model + general function

$$y_i = \sum_{j=1}^{n_t} \phi_j(x) d_j + h(x_j) + \epsilon_i$$

$T_{i,J} = \phi_j(x_i)$ **and let** $K_{k,i} = \psi_k(x_i)$

$$g(x) = \sum_{j=1}^{n_t} \phi_j(x) d_j + \sum_{k=1}^{n_p} \psi_k(x) c_k$$

or

$$\hat{g} = T\hat{\boldsymbol{d}} + K\hat{\boldsymbol{c}}$$

Find the parameters by the ridge regression:

$$\min_{\boldsymbol{c},\boldsymbol{d}} (\boldsymbol{y} - T\boldsymbol{d} - K\boldsymbol{c})^T (\boldsymbol{y} - T\boldsymbol{d} - K\boldsymbol{c}) + \lambda \boldsymbol{c}^T \Omega \boldsymbol{c}$$

Can use the same general formula or take advantage of the fact that the penalty in only on $\boldsymbol{c}$.

The cubic smoothing spline

We just need to define the right basis functions and penalty.

A strange covariance:

$$k(u,v) = \begin{cases} u^2v/2 - u^3/6 & \textbf{for } u < v \\ v^2u/2 - v^3/6 & \textbf{for } u \geq v \end{cases}$$

Strange basis functions:

$\phi_1 = 1$, $\phi_2 = x$, $\psi_i(x) = k(x, x_i)$

The penalty:

$\Omega_{i,j} = k(x_i, x_j)$,

Why does this work?

Splines are described by special covariance functions known as reproducing kernels , $k(x, x')$

with $\psi_i(x) = k(x, x_i)$ the choice for cubic splines has the property

$$\int \psi_j^{''}(x)\psi_i^{''}(x)dx = \psi_j(x_i) = k(x_i, x_j)$$

so when $h(x) = \sum_j \psi_j c_j$ and $T^T \boldsymbol{c} = 0$.

$$\int (h^{''}(x))^2 dx = \boldsymbol{c}^T \Omega \boldsymbol{c}$$

So the ridge regression penalty is the same as the integral criterion.

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.

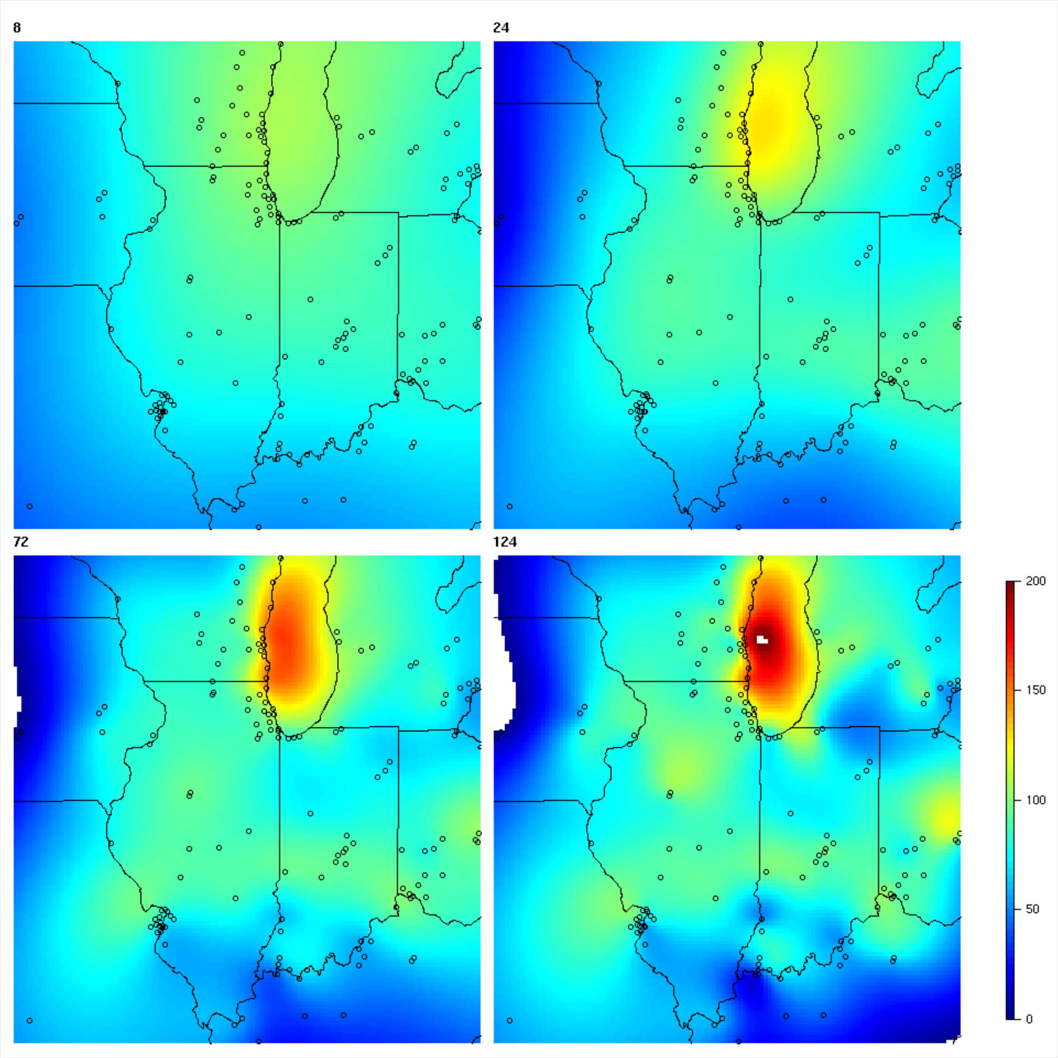
Again, separate off the linear part of f.
$f(x) = \beta_1 + \beta_2 x_1 + \beta_3 x_2 + h(\boldsymbol{x})$

Reproducing Kernel:

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

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

Some thin plate splines for the ozone data



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

$CV(\lambda)$

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

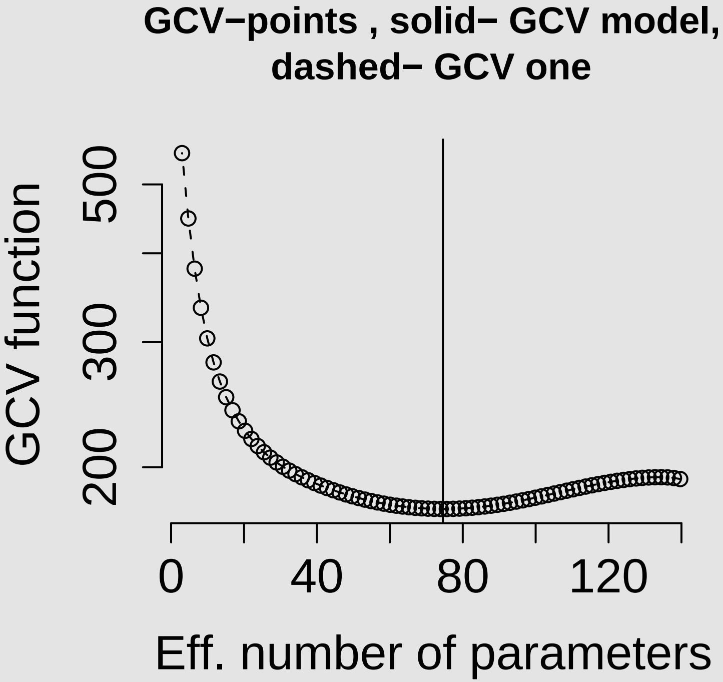
$GCV(\lambda)$

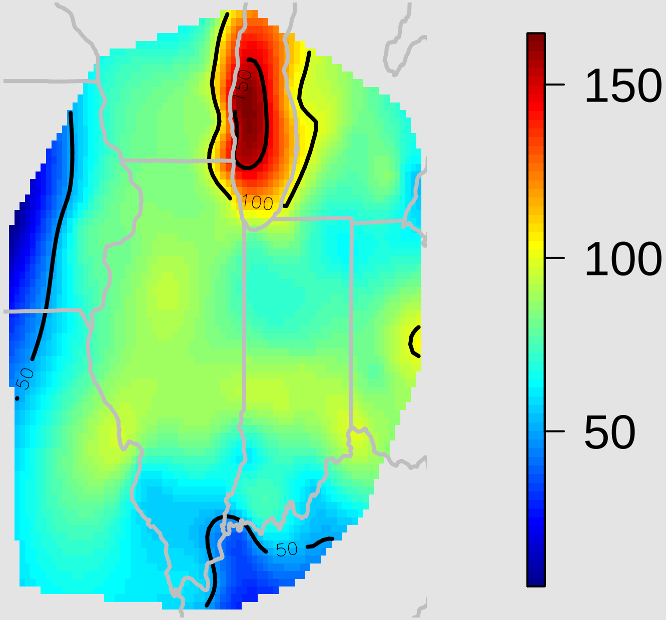
$$(1/n)\frac{\sum_{i=1}^{n}(y_i - \widehat{g}_i)^2}{(1 - \mathbf{tr}A(\lambda)/n)^2}$$

Minimize CV or GCV over λ to determine a good value

GCV for the ozone data

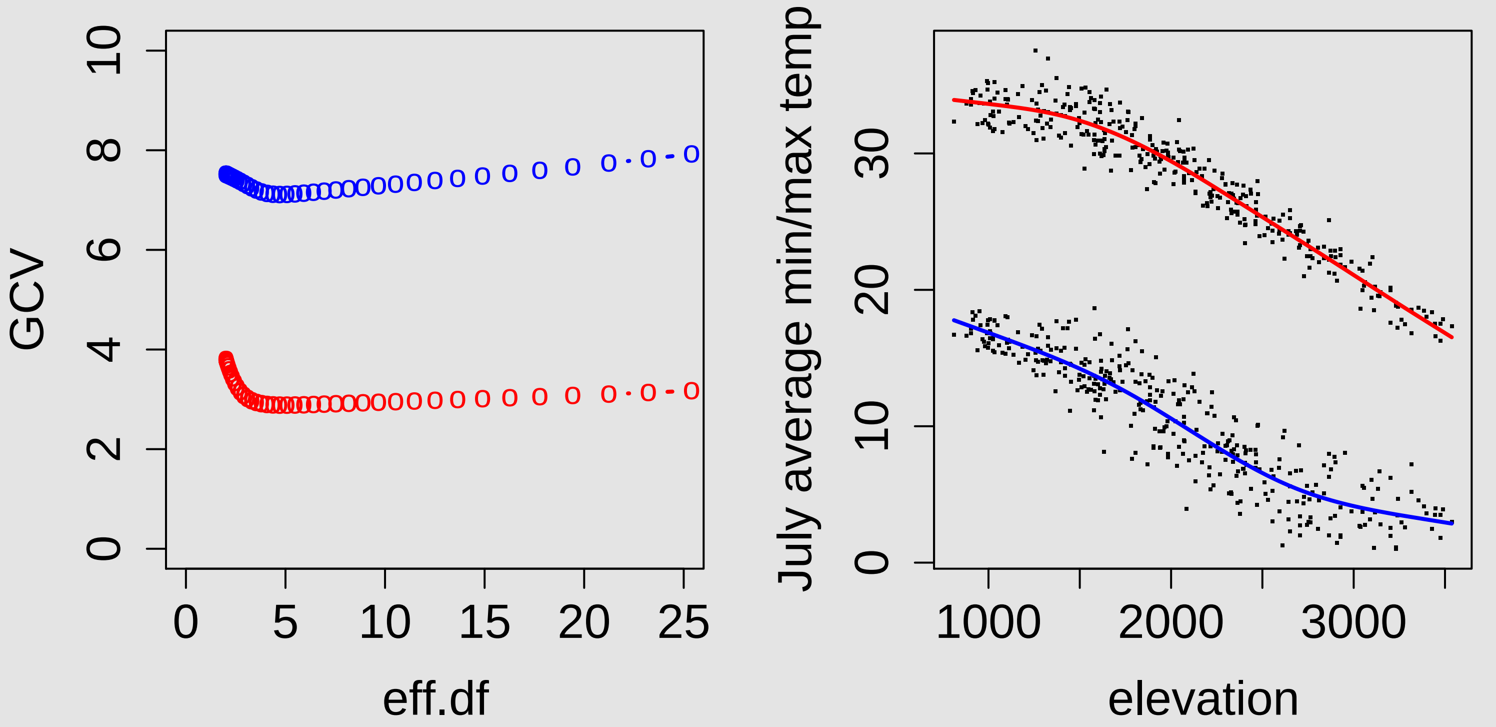
GCV(eff. degrees of freedom), the estimated surface





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!