Smoothing data and splines

SAMSI Summer School August, 2009

Douglas Nychka,

National Center for Atmospheric Research

Outline

- Penalized least squares smoothers

- Properties of smoothers

- Cubic and thin-plate splines

- Cross-validation for finding the smoothing parameter

National Science Foundation *Short Course MAR 2009*

Estimating a curve or surface.

An additive statistical model:

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

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

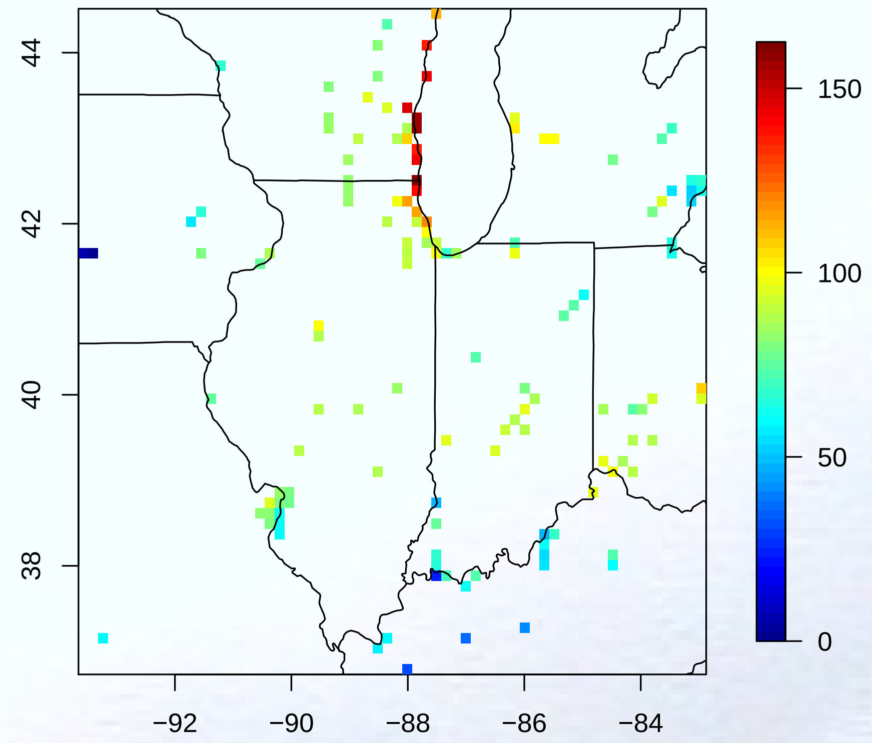
ϵ_i's are random errors
and g is an unknown, smooth function.

The goal is to estimate g based on the observations

A two dimensional example

Predict surface ozone where it is not monitored.

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



Penalized least squares

Ridge regression

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

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

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

Let $X_{i,k} = b_k(x_i)$ so $\hat{\boldsymbol{g}} = X\hat{\boldsymbol{\beta}}$

Penalized least squares.

minimize over β:

Sum of squares($\boldsymbol{\beta}$) + penalty on $\boldsymbol{\beta}$

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

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

In general
- log likelihood (y,β) + penalty (β)

minimizing this makes sense as an estimate.

Spatial statistics estimates:
the basis ($\{b_k\}$) and the penalty (H)
based on a spatial covariance.

Bayesian posterior mode:
The penalty can also be a log prior density for β

Once we have the parameter estimates these can be used to evaluate $\hat{g}$ at any point.

Solution to the Ridge Regression

Just calculus ...

- Take derivatives of the penalized likelihood w/r to $\boldsymbol{\beta}$,
- set equal to zero,
- solve for $\boldsymbol{\beta}$

The monster ...

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

The hat matrix for prediction

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

There is a transformation , G *so that*

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

(D is diagonal and XG orthogonal)

Linear smoothers

The vector of predictions:

$$\hat{\boldsymbol{g}} = \begin{pmatrix} \hat{\boldsymbol{g}}(\boldsymbol{x}_1) \\ \hat{\boldsymbol{g}}(\boldsymbol{x}_2) \\ \vdots \\ \hat{\boldsymbol{g}}(\boldsymbol{x}_n) \end{pmatrix} \tag{1}$$

The smoother matrix: $\hat{\boldsymbol{g}} = A\boldsymbol{y}$

- A is an $n \times n$ matrix
- eigenvalues of A are in the range [0,1].
- $\hat{g}(x)$ in between the data found by interpolating the predictions at the observations.
- $||A\boldsymbol{y}|| \leq ||y||$

For ridge regression $(I + \lambda D)^{-1}$ is the smoothing function.

Effective degrees of freedom

For linear regression trace of $X^T(X^TX)^{-1}X^T$ gives the number of parameters. (Because it is a projection matrix)

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

- $\text{tr}A(\lambda)$ monotonically increases as λ decreases
- $\text{tr}A(0) =$ number of basis functions
- $\text{tr}A(\infty) =$ number of basis functions *not* penalized.

- effective degrees of freedom is a better parametrization than the smoothing parameter.

The classic cubic smoothing spline

Splines are the solutions to variational problems.

For curve smoothing in one dimension,

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

The second derivative measures the roughness of the fitted curve.

Form of the solution

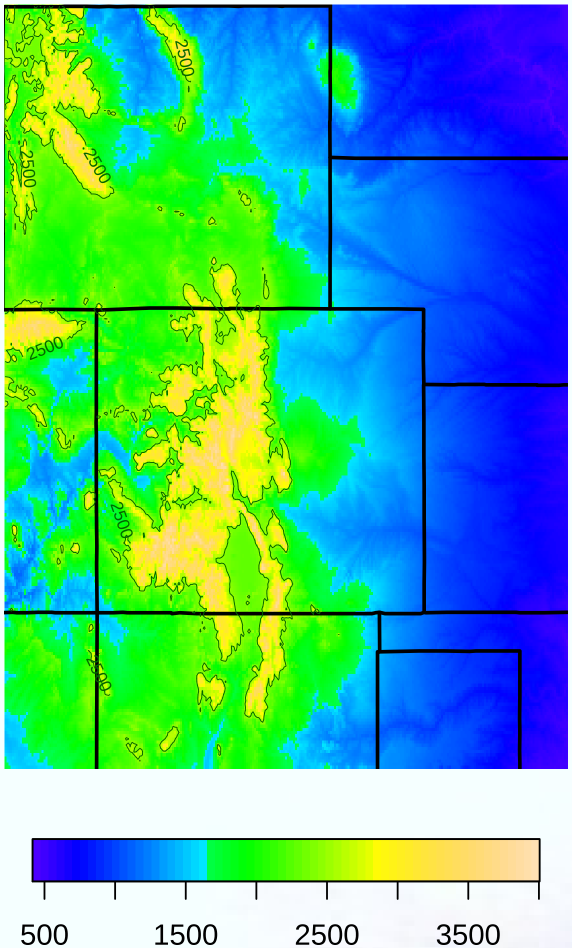
$\widehat{g}$ is continuous and with continuous first and second derivatives

It is a piecewise, cubic polynomial in between the observation points.

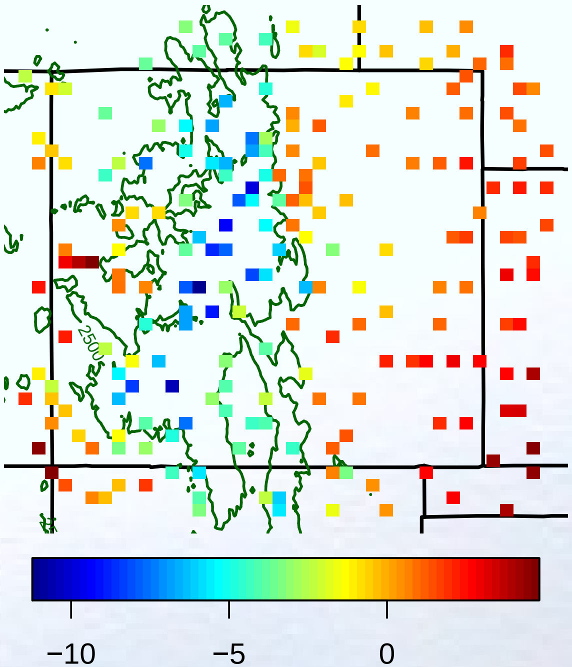
What does this have to do with ridge regression?

Climate for Colorado

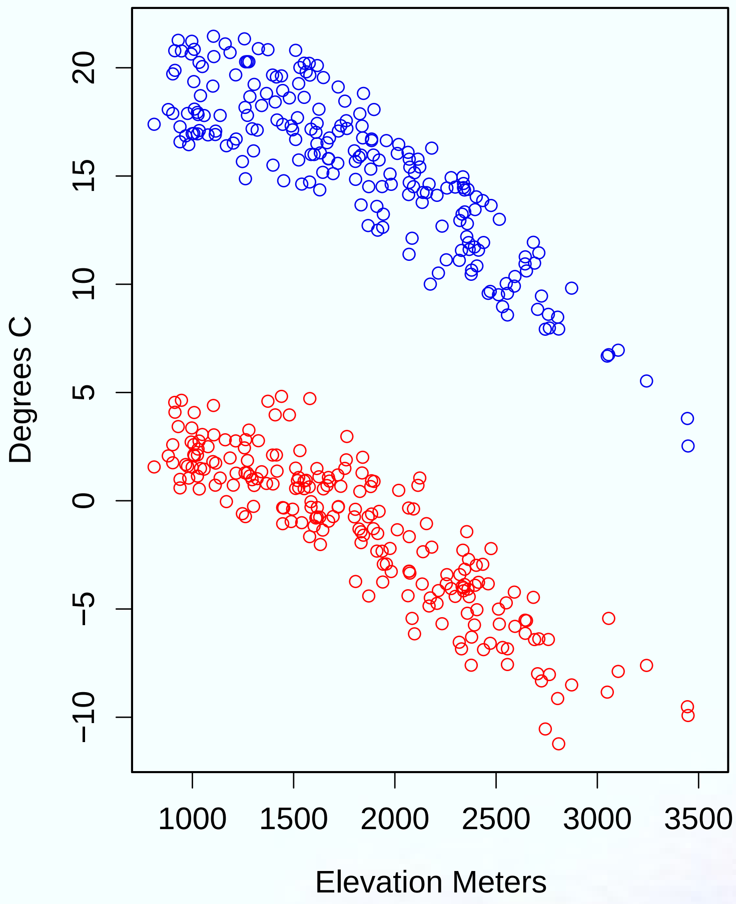
Elevations



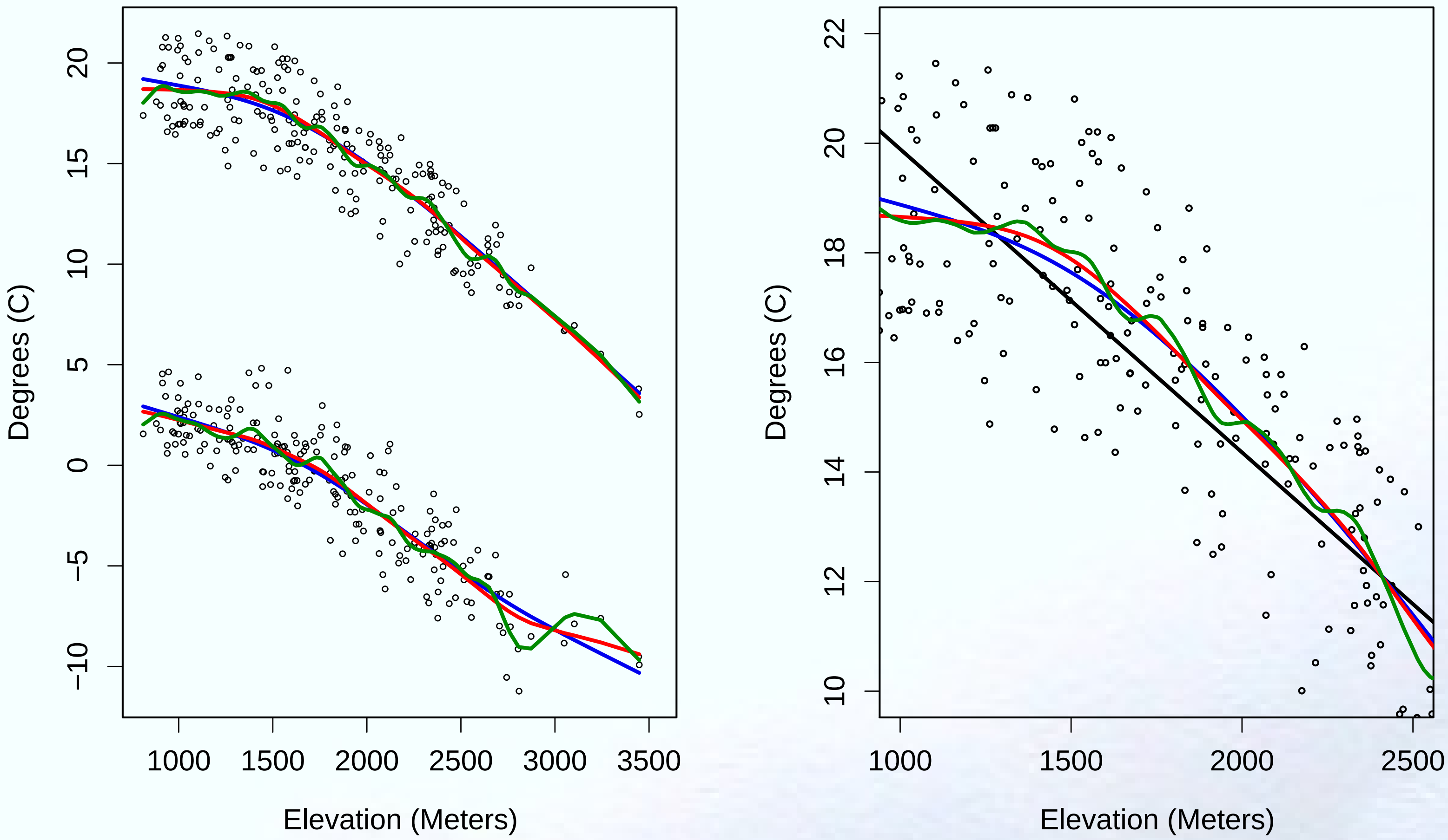
Spring average daily max temperatures



Max/Min spring temperatures



Cubic splines with different λ s



Form of the spline estimate

Estimate =

low dimensional parametric model + general function

Penalty matrix "hard-wired" to basis functions.

Divide the basis functions into two parts $\{\phi_j\}$ and $\{\psi_k\}$ and only penalize the second set.

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

Form (continued)

$$\widehat{g}(x) = \sum_{j=1}^{n_t} \phi_j(x)\widehat{d}_j + \sum_{k=1}^{n_p} \psi_k(x)\widehat{c}_k$$

Ω derived from $\{\psi_k\}$

In matrix format:

$T_{i,J} = \phi_j(x_i)$, $K_{k,i} = \psi_k(x_i)$ and ... $\Omega = K$

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

Find the parameters by the ridge regression:

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

Solution:

$\hat{d} = (T^T M^{-1} T)^{-1} T^T M^{-1} \boldsymbol{y}$ (GLS)

$M = K + \lambda I$

$\hat{c} = (KK^T + \lambda K)^{-1}(\boldsymbol{y} - T\hat{\boldsymbol{d}}) = (K + \lambda I)^{-1}(\boldsymbol{y} - T\hat{\boldsymbol{d}})$

The cubic smoothing spline

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

A strange covariance:

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

Friends and strangers

Friends: $\phi_1(x) = 1$, $\phi_2(x) = x$,

Strangers: $\psi_i(x) = k(x, x_i)$

The penalty matrix: $\Omega_{i,j} = k(x_i, x_j)$,

Why does this work?

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

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 $T^T \boldsymbol{c} = 0$.

$$\int (h''(x))^2 dx = \int (\sum_j \psi_j''(x) c_j)^2 dx = \boldsymbol{c}^T K \boldsymbol{c}$$

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.

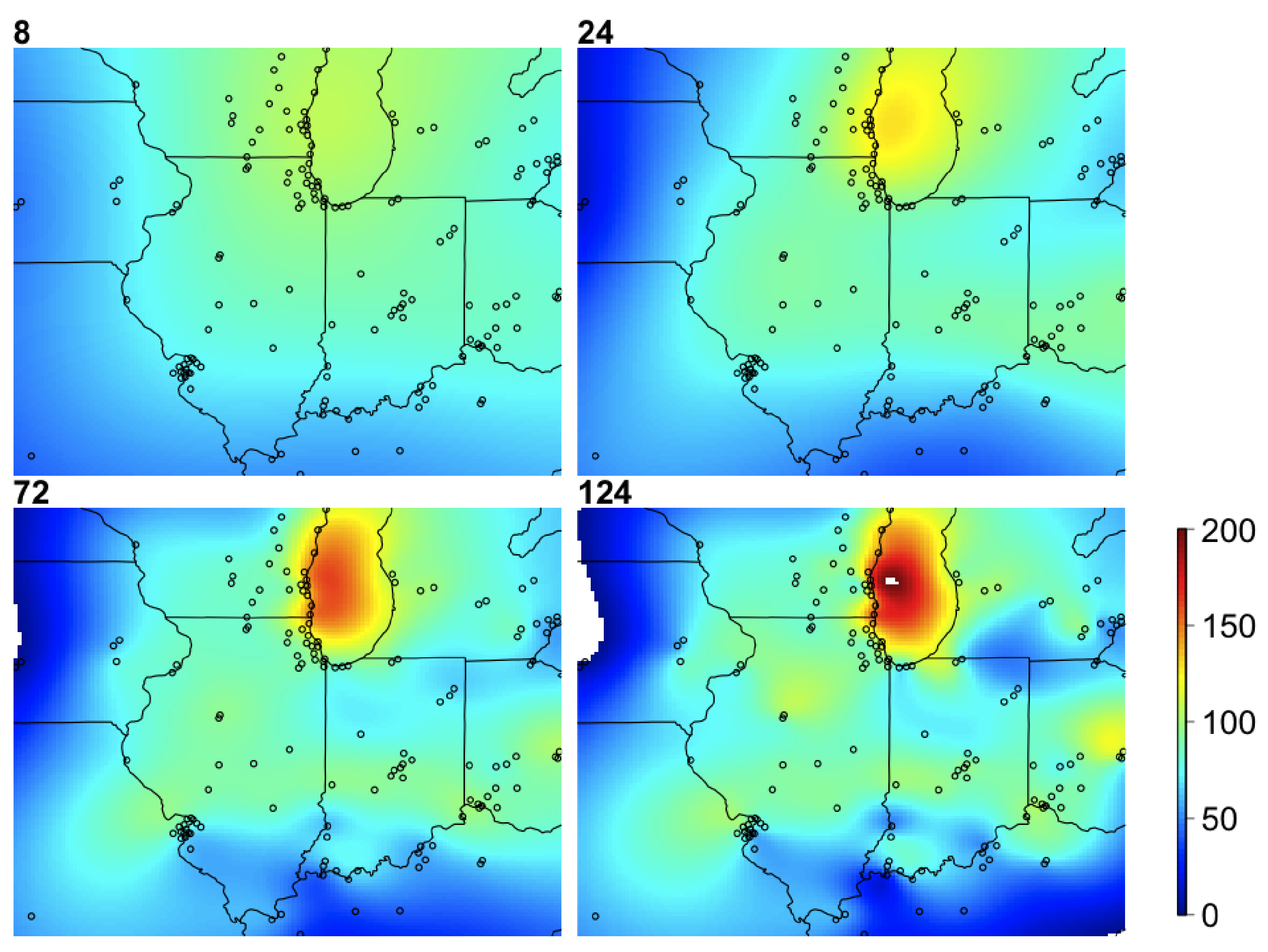
Again, separate off the linear part of f.

$f(x) = \beta_1 + \beta_2 x_1 + \beta_3 x_2 + h(\boldsymbol{x})$

Thin plate spline kernel:

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

Estimates 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 $(\boldsymbol{x}_i, \widehat{g}_{-1})$ to the data does not change the estimate.

CV and Generalized CV criterion

$CV(\lambda)$

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

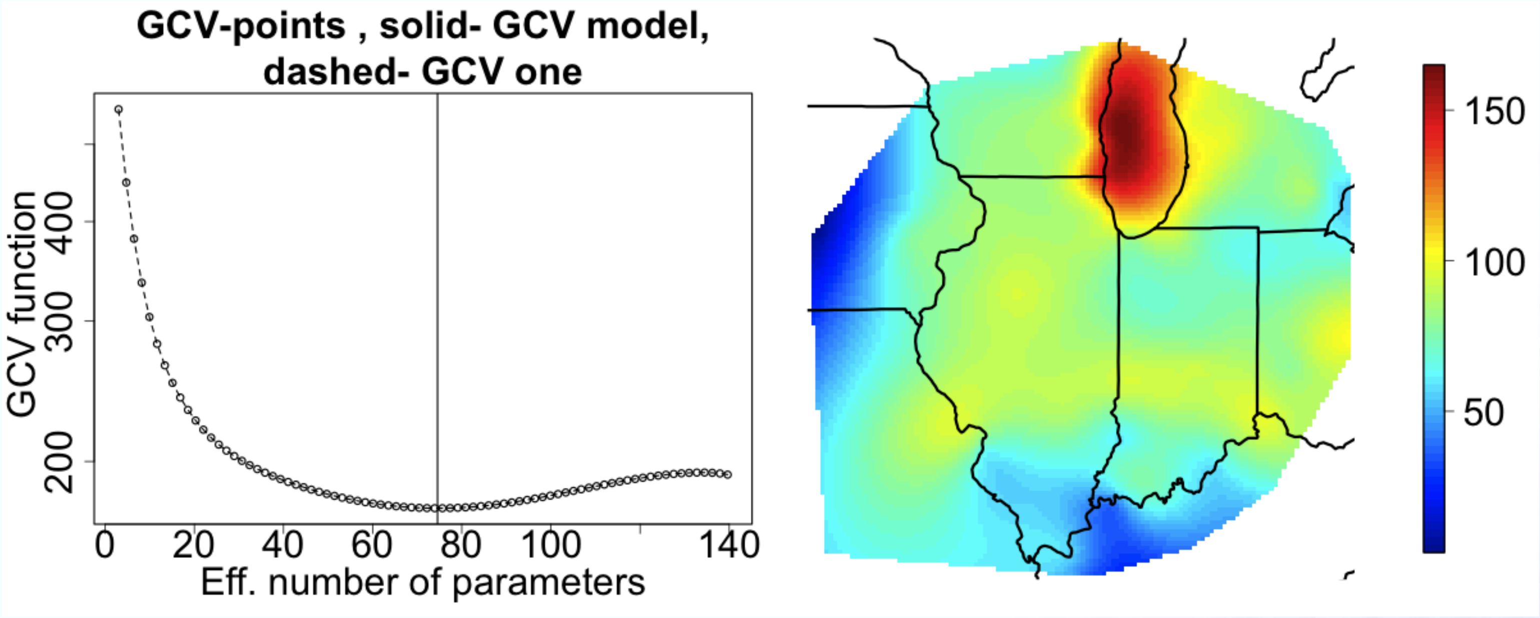
$GCV(\lambda)$

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

Minimize CV or GCV over λ to determine a good value

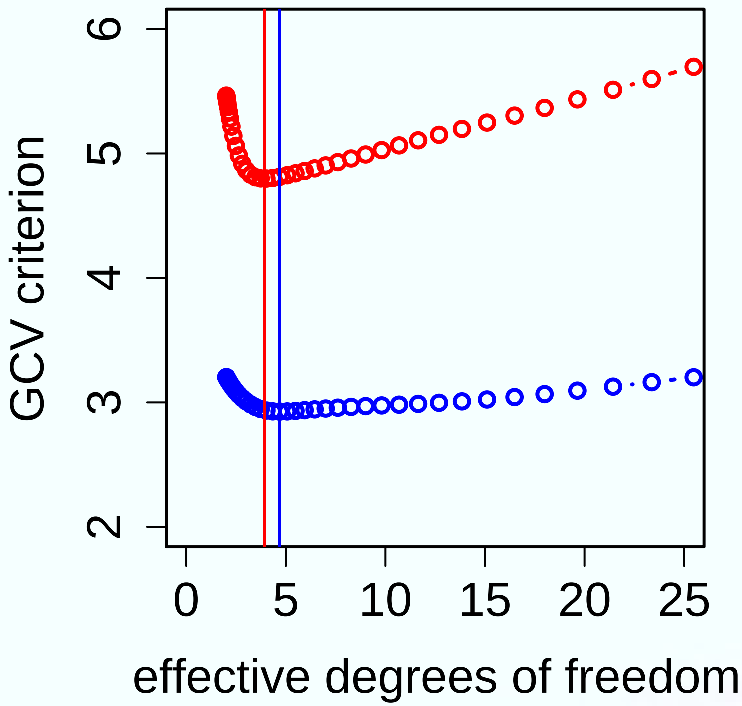
GCV for the ozone data

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



GCV for the climate data

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



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!