Statistics 215a – 9/10/03 – D. R. Brillinger

Smoothing scatter plots

Replacing (x,y) by (x,y_x) with y_x smooth and connect the points

datum = smooth + rough

Purposes.

Get clearer view, less detail

See what the data are saying

Reduce impact of isolated points

Reduces irrelevant variation / noise

Preparatory to further processing

Separates rapid changes from less rapid

May suggest simple closed form expression

Variants preserve discontinuities

Statistics 215a – 9/10/03 – D. R. Brillinger

Smoothing – some types

data (x_i,y_i)

I. Parametric regression

e.g. regression line by OLS

nonlocal, infinitely smooth

variance small, 1/n

"bias", (error for specific function), can be large

II. Bin smoother

cut points c_k

cells $c_k \leq x_i < c_{k+1}$

$R_k = \{i|\ c_k \leq x_i < c_{k+1}\}$, $c_0 = -\inf$, $c_K = \inf$
approx equi-sized

$s(x) = \text{ave}\{y_i|i \in R_k, x \in R_k\}$

not smooth, step function

cut(), stepfun(), ksmooth()

III. Running mean

Average over points close to x

$s(x) = \text{ave}\ \{y_j|j \in N(x)\}$

$N(x_i) = \{max(i-k,1),…,i-1,i,i+1,…,min(i+k,n)\}$

Moving/running average

k controls appearance, smooth vs. jagged

span: (2k+1)/n

wiggly, biased, endpoint problem

theory is “easy”

might use r=2k+1 nearest neighbors

IV. Running-line smoother

Replace average above by OLS line

$s(x) = a(x) + b(x)x$

a(x), b(x) OLS for data in N(x)

good at ends

jagged, points equal weight (big change on shift)

loess(), lowess()

V. Kernel smoothers

K(.): kernel function, e.g. pdf

Biweight – $(1-u^2)^2$

$K_b(x)=K(x/b)$, b bandwidth

$s(x) = \Sigma_j\ y_j\ K_b(x-x_j)\ /\ \Sigma_j\ K_b(x-x_j)$

linear in y’s

choice of b is important

surprisingly effective/efficient

endpoints

ksmooth()

VI. Running medians

replace running mean by running median

resistant to outliers

salt-and-pepper noise

repeated running medians

VII. Equivalent kernels

Many studied are linear

$s(x_i) = \Sigma_j S_{ij} y_j$

S is the smoother matrix
may have parameter λ

S_{0j} : the equivalent kernel
plot vs. x_0

Degrees of freedom: tr(S), tr(SS^T), …

VIII. Regression splines

compromise between local and global

piecewise polynomials, separated by knots

smooth joins

e.g. cubic

$s(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \Sigma_j \theta_j (x-\xi_j)_+^3$

$s^{(3)}$ exists, $s^{(2)}$ continuous

Find β, θ by OLS

Knots more difficult

bs() generates a basis

IX. Cubic smoothing splines

solve extremal problem

$$\Sigma_i \{y_i - f(x_i)\}^2 + \lambda \int f''(t)^2 dt$$

closeness to data + smoothness

λ: relative weight

smooth.spline()

X. Locally-weighted running-line

Cleveland's lowess(), loess()

weighted least squares

$\exists$ robust variant

XI. Supersmoother

k-th nearest neighbor LS, k=n/2,n/5,n/20
cross-validation used to choose k for each x interpolating between the three

"fast"

supsmu()

XI. Multiple predictors

spatial data

thin-plate spline

T. Hastie and R. Tibshirani (1990).
Generalized Additive Models. Chapman & Hall

Cross-validation. A method for estimating prediction error and other things.
One tests the procedure on data different from those used to estimate its parameters.
E.g. drop out one observation at a time.

Thin plate spline radial basis functions

d: dimension of space

r: radial distance

m: derivatives in roughness penalty

$r^{2m-d} \log r$,   d even

r^{2m-d} ,   d odd