Nonparametric density estimation

or

Smoothing the data

Eric Feigelson

Brief lecture and R tutorial

Harvard-Smithsonian Center for Astrophysics

January 2014

Why density estimation?

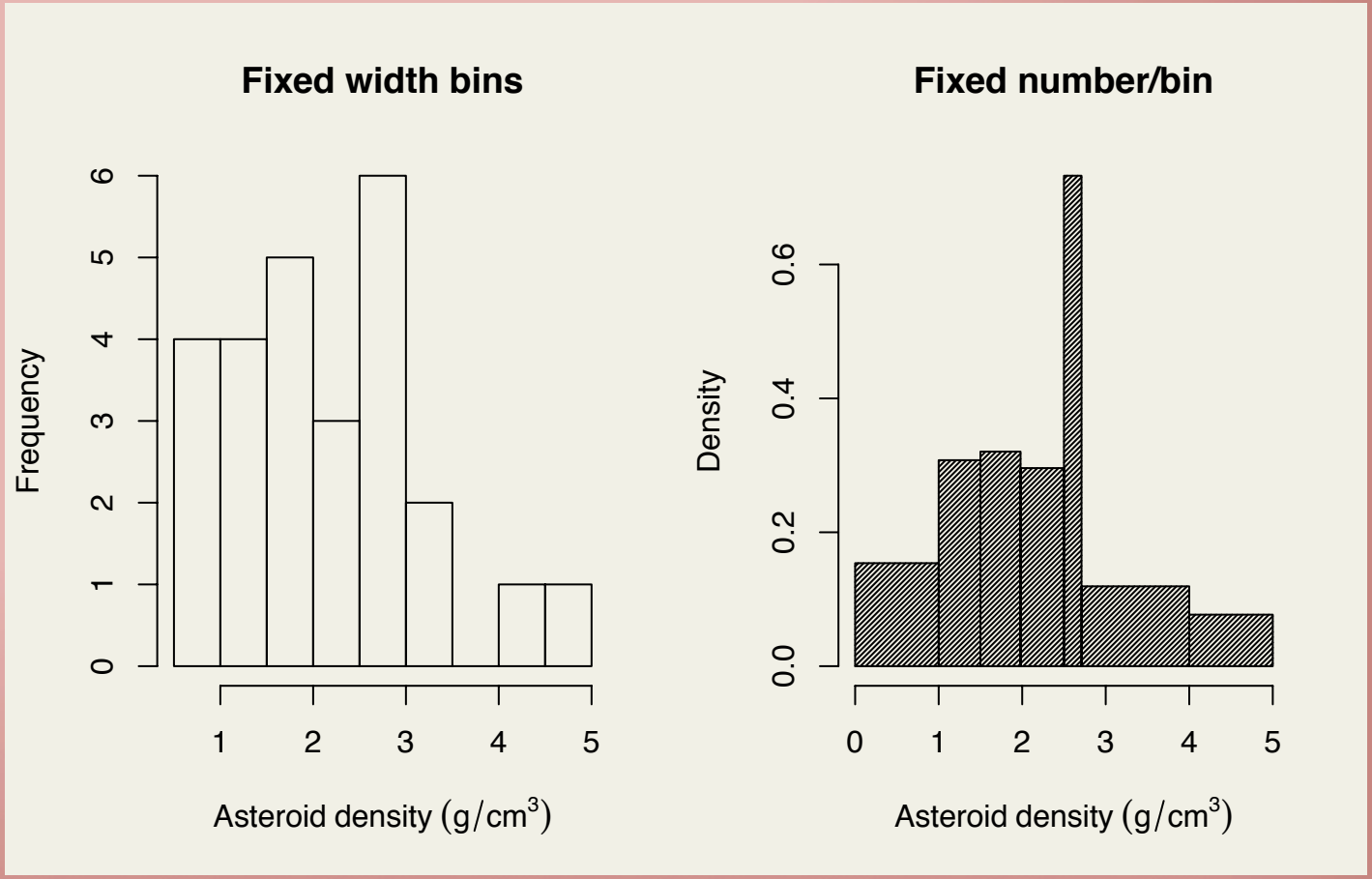
The goal of density estimation is to estimate the unknown probability density function of a random variable from a set of observations. In more familiar language, density estimation smoothes collections of individual measurements into a continuous distribution, replacing dots on a scatterplot by a smooth estimator curve or surface.

When the parametric form of the distribution is known (e.g., from astrophysical theory) or assumed (e.g., a heuristic power law model), then the estimation of model parameters is the subject of regression (MSMA Chpt. 7). Here we make no assumption of the parametric form and are thus involved in ***nonparametric density estimation***.

Astronomical applications

- Galaxies in a rich cluster → underlying distribution of baryons
- Lensing of background galaxies → underlying distribution of Dark Matter
- Photons in a Chandra X-ray image → underlying X-ray sky
- Cluster stars in a Hertzsprung-Russell diagram → stellar evolution isochrone
- X-ray light curve of a gamma ray burst afterglow → temporal behavior of a relativistic afterglow
- Galaxy halo star streams → cannibalism of satellite dwarf galaxy

Histograms are commonly used



but histograms have serious problems for inference

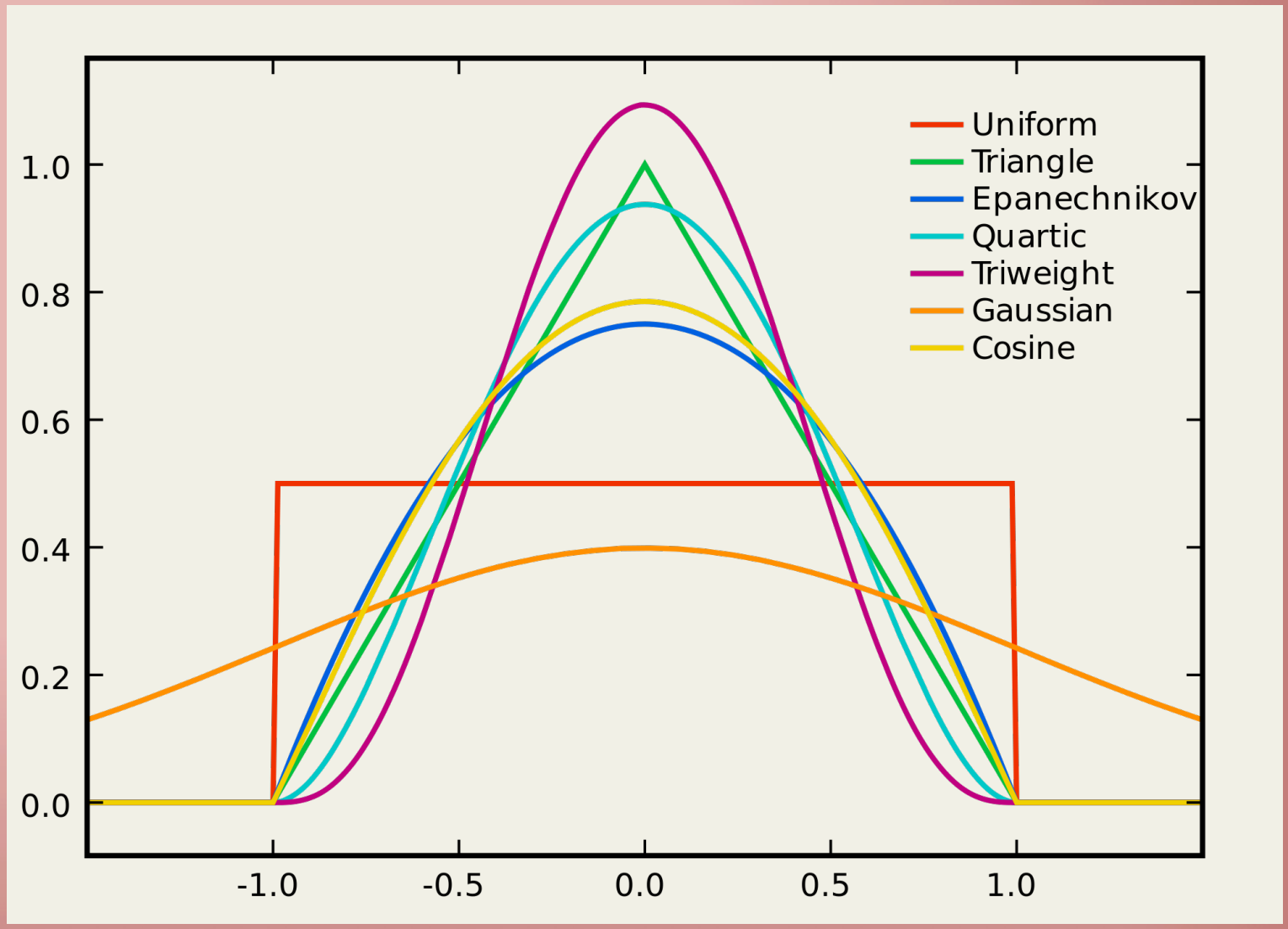
Kernel density estimation

The most common nonparametric density estimation technique convolves discrete data with a normalized kernel function to obtain a continuous estimator:

$$\hat{f}_{kern}(x,h) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-x_i}{h})$$

A kernel must integrate to unity over $-\infty < x < \infty$, and must be symmetric, K(u) = K(-u) for all u. If K(u) is a kernel, then a scaled $K^*(u) = \lambda K(\lambda u)$ is also a kernel.

A normal (Gaussian) kernel a good choice, although theorems show that the minimum variance is given by the Epanechnikov kernel (inverted parabola). The uniform kernel (`boxcar', `Heaviside function') give substantially higher variance.



http://en.wikipedia.org/wiki/Kernel_density_estimation
http://en.wikipedia.org/wiki/Kernel_(statistics)

The choice of bandwidth is tricky!

A narrow bandwidth follows the data closely (small bias) but has high noise (large variance). A wide bandwidth misses detailed structure (high bias) but has low noise (small variance).

Statisticians often choose to minimize the L_2 risk function, the **mean integrated square error (MISE)**,

$$MISE(\hat{f}_{kern}) = E\int [\hat{f}_{kern}(x) - f(x)]^2 dx$$

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

$$\sim c_1 h^4 + c_2 h^{-1}$$

(The constant c_1 depends on the integral of the second derivative of the true p.d.f.)

The `asymptotic mean integrated square error' (AMISE) based on a 2nd-order expansion of the MISE is often used.

KDE: Choice of bandwidth

The choice of bandwidth *h* is more important than the choice of kernel function. Silverman's `rule of thumb' that minimizes the MISE for simple distributions is

$$h_{r.o.t.} = 0.9An^{-1/5} \qquad h_{opt,j} = \sigma_j n^{-1/(p+4)}$$

where A is the minimum of the standard deviation σ and the interquartile range IQR/1.34.

More generally, statisticians choose kernel bandwidths using **cross-validation**. Important theorems written in the 1980s show that maximum likelihood bandwidths can be estimated from resamples of the dataset. One method is leave-on-out samples with (n-1) points, and the likelihood is

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n} log\hat{f}_{-1,kern}(x_i)$$

Alternatives include the `least squares cross-validation' estimator, and the `generalized cross-validation' (GCV) related to the Akaike/Bayesian Information Criteria.

Rarely recognized by astronomers …

Confidence bands around the kernel density estimator can be obtained:

- For large samples and simple p.d.f. behaviors, confidence intervals for normal KDEs can be obtained from asymptotic normality (i.e. the Central Limit Theorem)

- For small or large samples and nearly-any p.d.f. behviors, confidence intervals for any KDE can be estimated from bootstrap resamples.

Kernel regression

A regression approach to smoothing bivariate or multivariate data …

$$E(Y \mid x) = f(x)$$

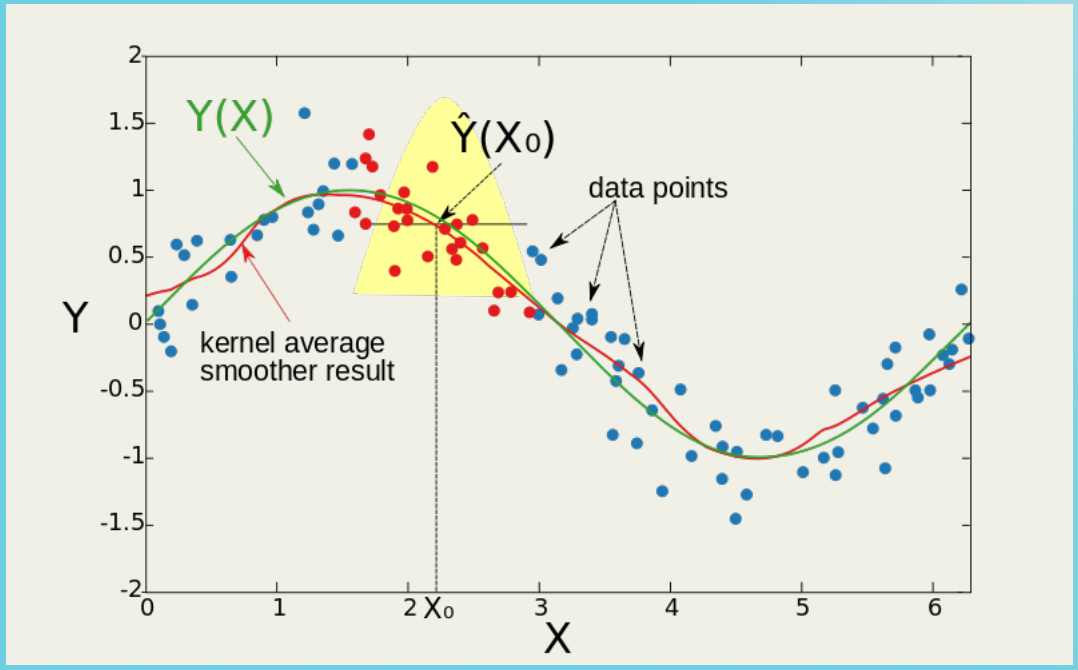
Read "the expected population value of the response variable Y given a chosen value of x is a specified function of x". A reasonable estimation approach with a limited data set is to find the mean value of Y in a window around x, [x-h]/, x+h/2) with h chosen to balance bias and variance.

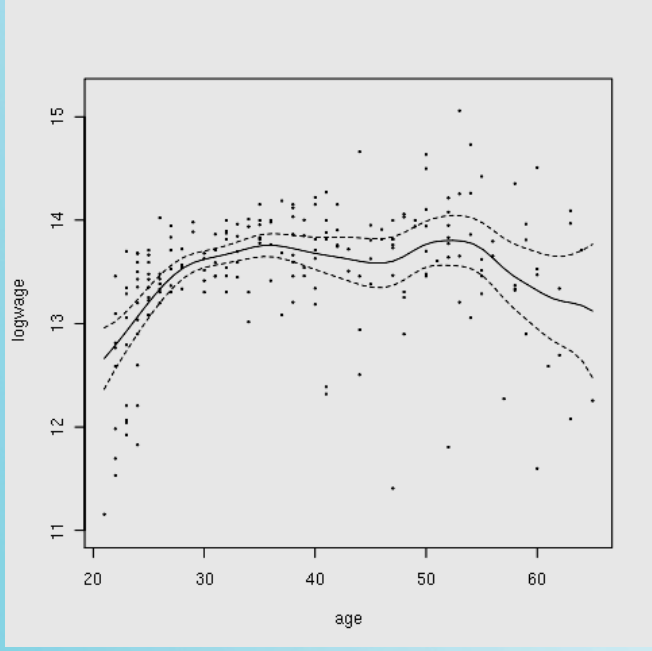
A more effective way might include more distant values of x downweighted by some kernel p.d.f. function such as $N(0,h^2)$. This called *kernel regression,* a type of *local regression.* The `best fit' might be obtained by locally weighted least squares or maximum likelihood.

Two common nonparametric regressions

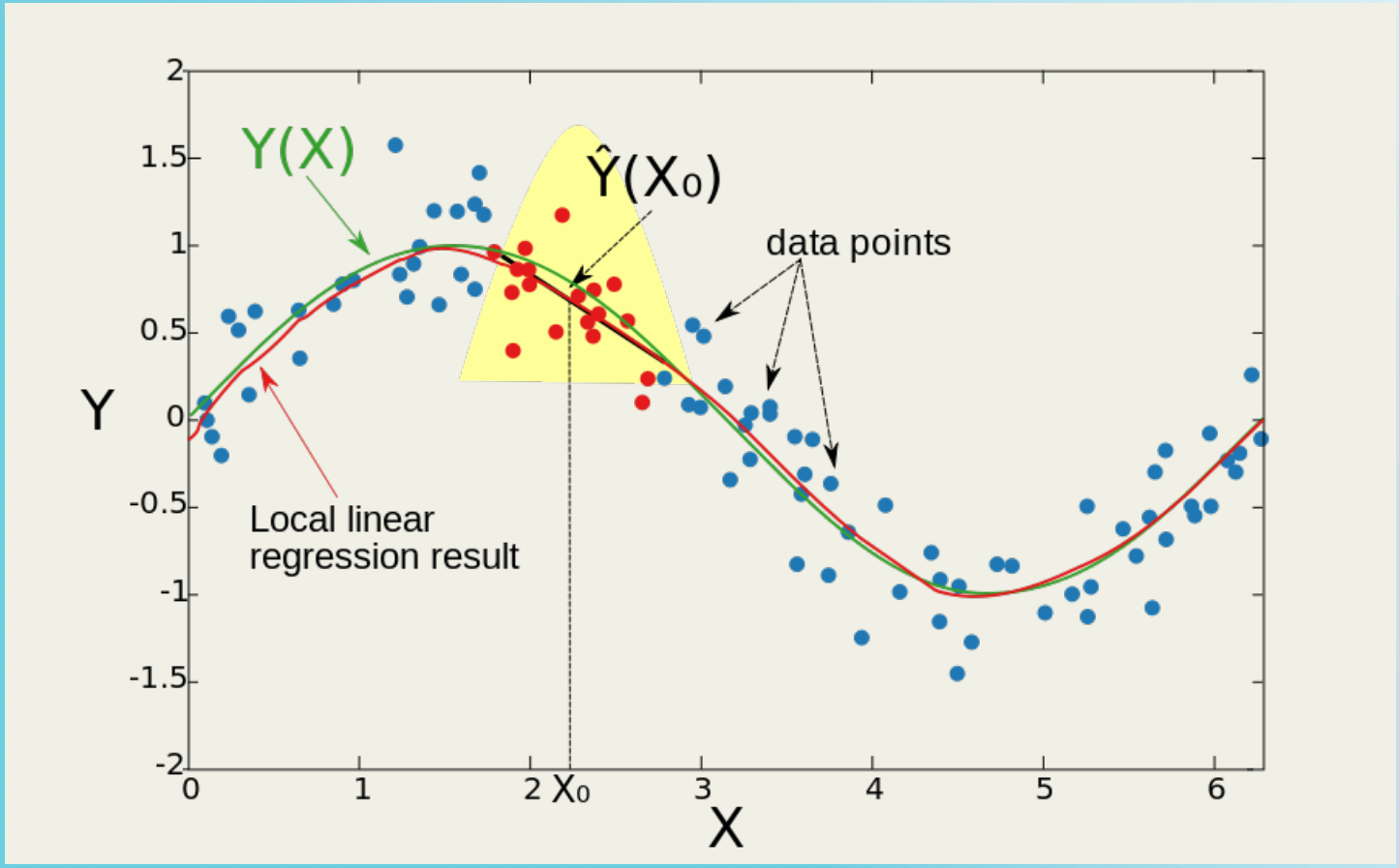
Nadaraya-Watson estimator

$$\hat{r}_{NW}(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h_x}\right)}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_x}\right)}.$$





Local polynomial smoother (LOWESS, LOESS, Friedman's supersmoother)



Spline regression

The function f(x) can be any (non)linear function but is often chosen to be a polynomial. If the polynomials are connected together at a series of *knots* to give a smooth curve, the result is called a *spline.*

It is easy to define a cubic spline having knots at ζ_1, $\zeta_2, \ldots, \zeta_p$. Let $(x - \zeta_j)_+$ be equal to $x - \zeta_j$ for $x \geq \zeta_j$ and 0 otherwise. Then the function

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x-\zeta_1)^3_+ + \beta_5 (x-\zeta_2)^3_+ + \ldots + \beta_{p+3} (x-\zeta_p)^3_+ \quad (6)$$

is twice continuously differentiable, and is a cubic polynomial on each segment $[\zeta_j, \zeta_{j+1}]$. Furthermore, the nonparametric regression model (1) becomes an ordinary linear regression model so that standard least-squares software may be used to obtain spline-based curve fitting. For example, suppose we

Kass 2008

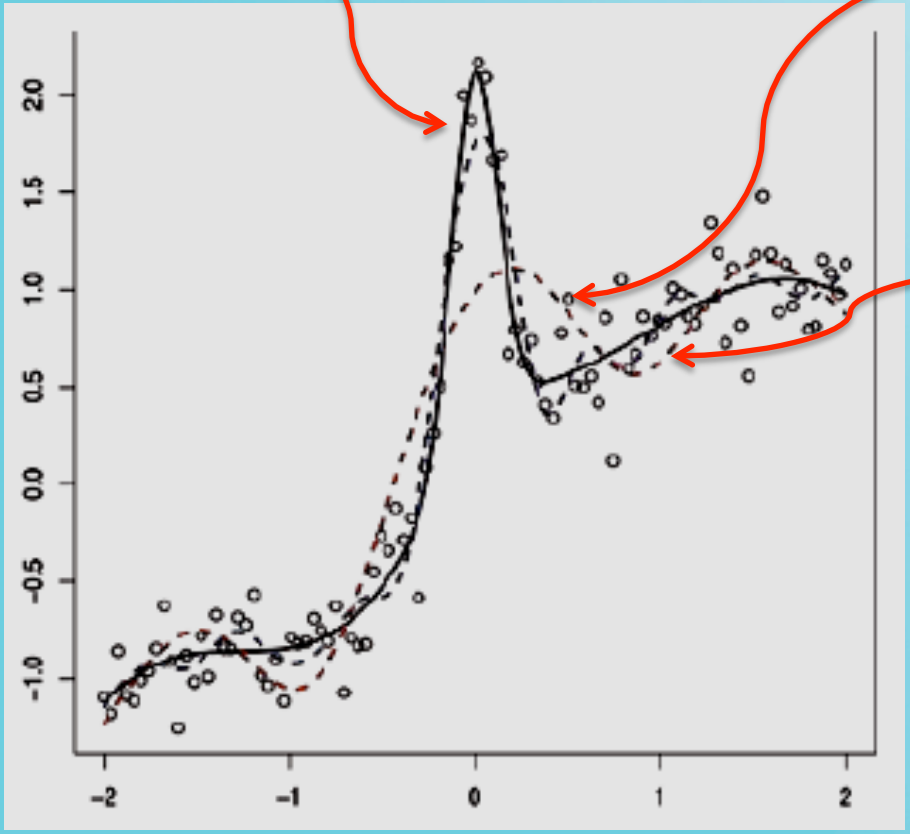
Variants that avoid high correlation among the fitted parameters are *B-splines* and *natural splines.*

The challenge of spline knot selection

7 knots chosen by user

5 knots chosen by R

15 knots chosen by R



Kass 2008

Adaptive splines

When the rapidity of changes in the data is not constant, then a constant bandwidth h, or evenly-spaced knots will perform poorly.

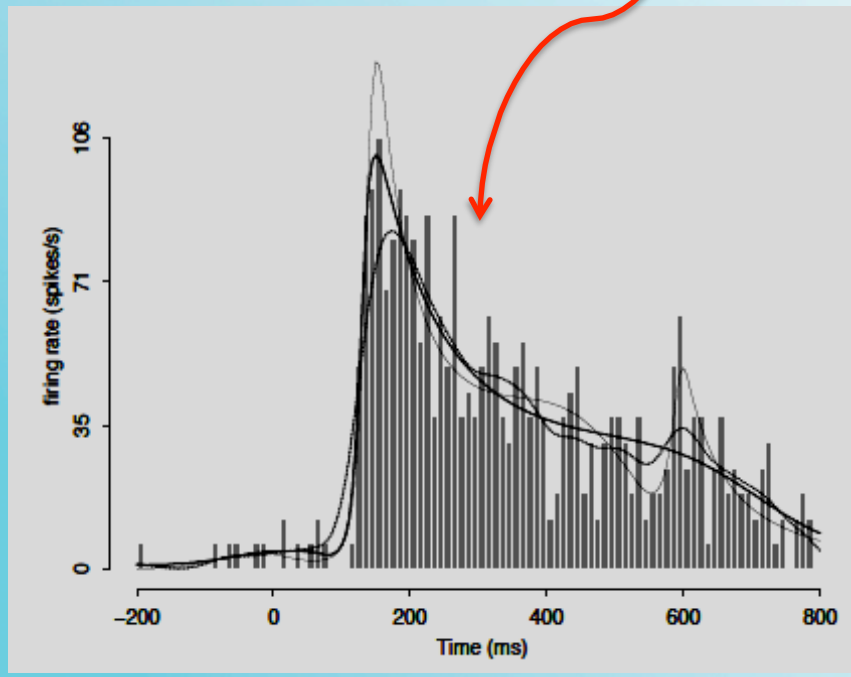
Smoothing splines involve a 'penalty' so that the spline coefficients are small when f'' is small (the function changes slowly), and high when f'' is large.

Bayesian Adaptive Regression Splines (BARS) find knot locations as posterior distributions of Bayes' Theorem using MCMC computations. The chosen knot locations are selected by minimizing the BIC. The performance for complicated distributions (even with discontinuities) can be excellent. Unlike competing *wavelet* methods, confidence intervals can be estimated.

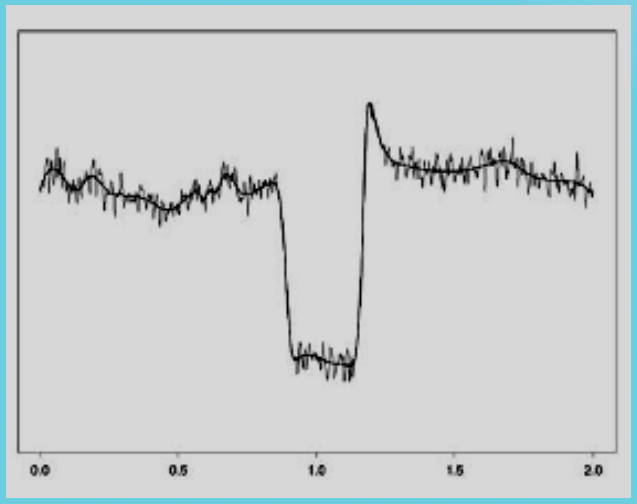
Two applications of BARS

Neuronal spikes or gamma ray burst?

Poisson regression



Gaussian regression



http://www.stat.cmu.edu/~kass/bars/
with C code and R wrappers
Wallstrom et al. (J. Stat. Software 2008)

Dotted: kernel density estimator
Gray: logspline
Thick: BARS

Comment for astronomers

Due to unfamiliarity with kernel density estimation and nonparametric regressions, astronomers too often fit data with heuristic simple functions (e.g. linear, linear with threshold, power law, ….). Unless scientific reasons are present for such functions, it is often wiser to let the data speak for themselves, estimating a smooth distribution from data points nonparametrically. A variety of often-effective techniques are available for this.

Well-established methods like KDE and NW estimator have asymptotic confidence bands. For all methods, confidence bands can be estimated by bootstrap methods within the `window' determined by the (local/global) bandwidth. Astronomers thus do not have to sacrifice `error analysis' using nonparametric regression techniques.