Nonparametric density estimation
or
Smoothing the data

Eric Feigelson

NARIT-EACOA Summer Workshop on Astrostatistics & Astroinformatics
August 2019

Why density estimation?

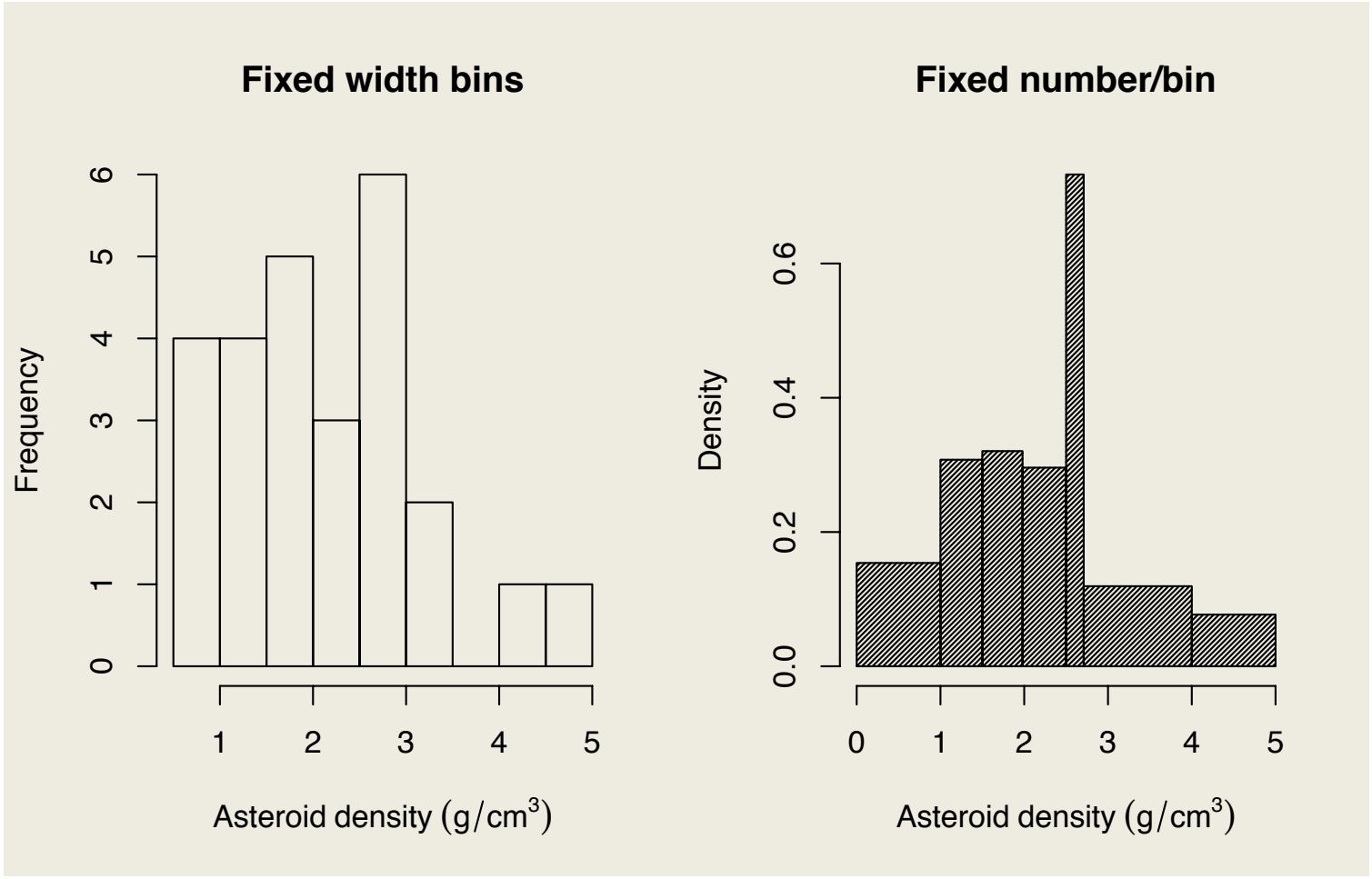
The goal of density estimation is to estimate the unknown probability density function of a random variable from a set of observations. In more familiar language, density estimation smoothes collections of individual measurements into a continuous distribution, replacing points on a line or plane by 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: A first step in density estimation



Problems with the histogram

- Discontinuities between bins are not present in the underlying population
- No mathematical guidance for choosing origin, x_0
- No mathematical guidance for binning (grouping) method: equal spacing, equal # points, etc.
- No mathematical guidance for choosing the `center' of a bin
- Difficult to visualize multivariate histograms

`In terms of various mathematical descriptions of accuracy, the histogram can be quite substantially improved upon'. (Silverman 1986)

`The examination of both undersmoothed and oversmoothed histograms should be routine.' (Scott 1992)

Histograms are useful for exploratory visualization of univariate data, but are not recommended for statistical analysis.

Fit models to the original data points and (cumulative) e.d.f.'s, not the (differential) histogram, unless the data are intrinsically grouped into ordered categories

Kernel density estimation

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

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

Mathematics: 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. See http://en.wikipedia.org/wiki/Kernel_density_estimation

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. and is therefore unknown in most situations.)

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

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

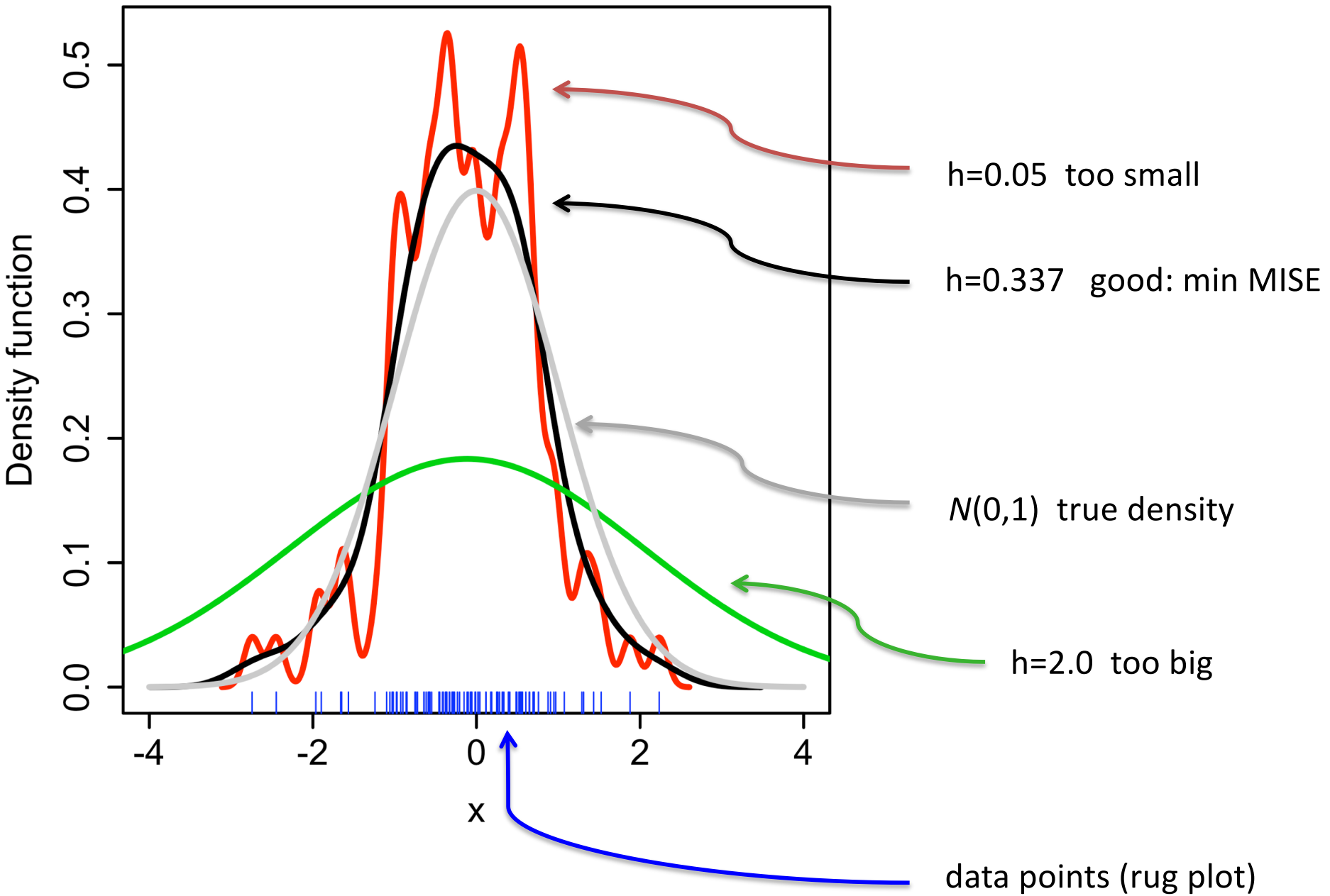
where *A* is the minimum of the standard deviation σ and the interquartile range *IQR/1.34*, and *p* is the number of dimensions

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

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

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

Example of normal kernel smoothing



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. behaviors, confidence intervals for any KDE can be estimated from bootstrap resamples.

Kernel regression

A regression approach to smoothing bivariate or multivariate data …

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

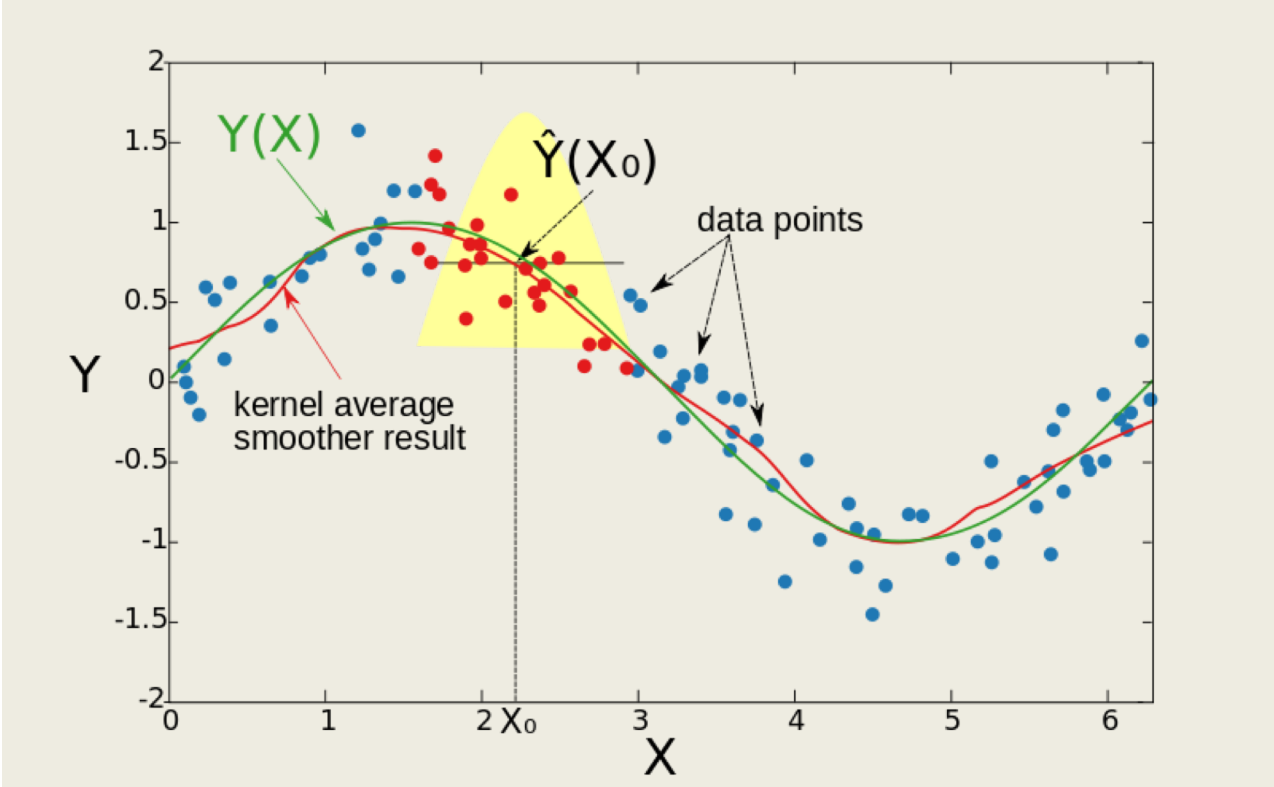
Read “the expected population value of the response variable Y given a chosen value of x is a specified function of x”. A reasonable estimation approach with a limited data set is to find the mean value of Y in a window around x, [x-h]/, x+h/2) with h chosen to balance bias and variance.

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.

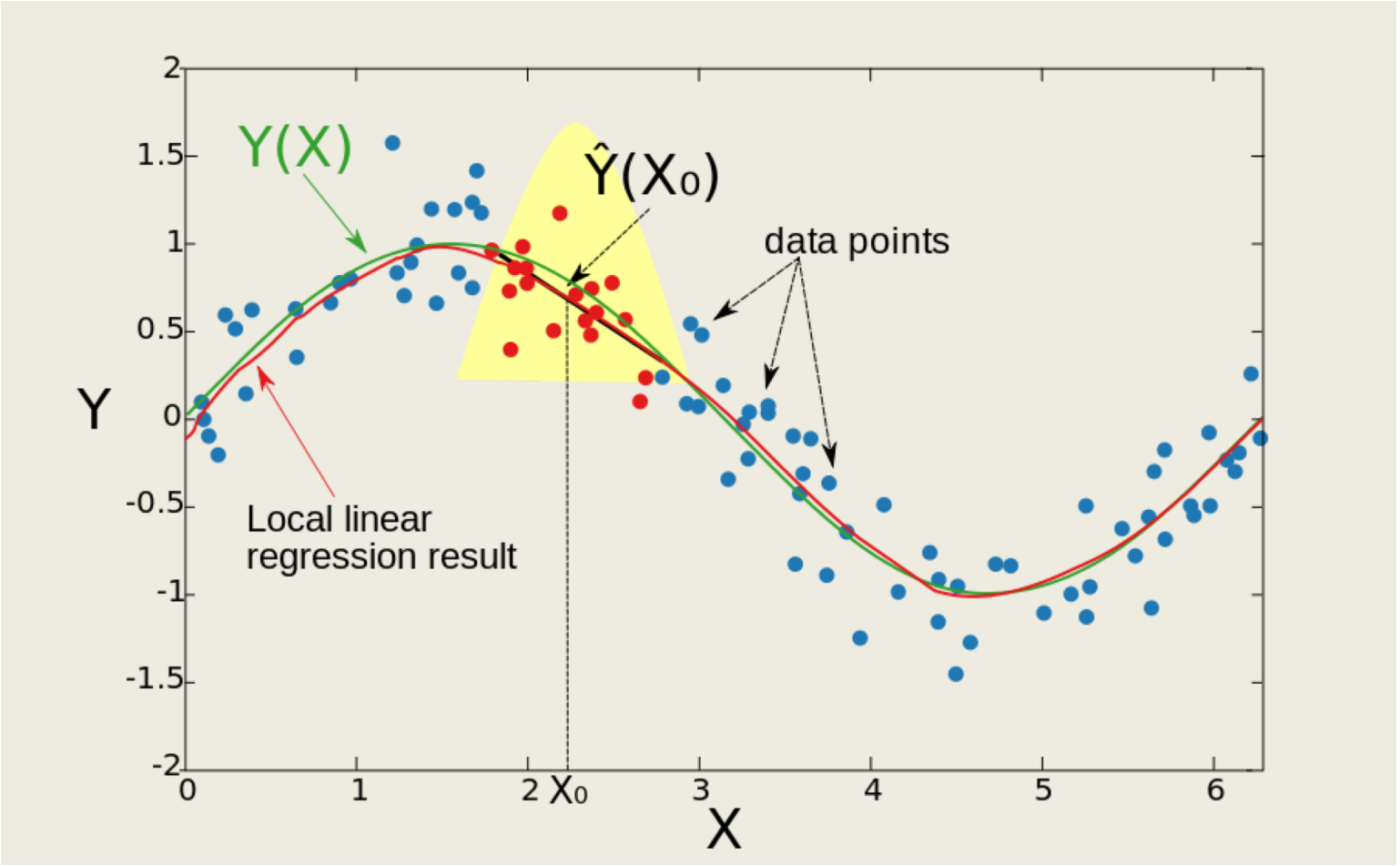
Two common nonparametric regressions

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 (LOESS, kriging ~ Gaussian Process regression)



Spline regression

A spline is a piecewise interpolating function that passes through a series of pre-specified *knots* in a low-dimensional space in a manner that minimizes the curvature under the constraint of continuous first and second derivatives. The function is typically chosen to be a (cubic) polynomial.

Algorithms for spline estimation were developed during the 1960-70s

C. de Boor, On calculating with B-splines, *J. Approx. Theory* **6** (1972), 50–62

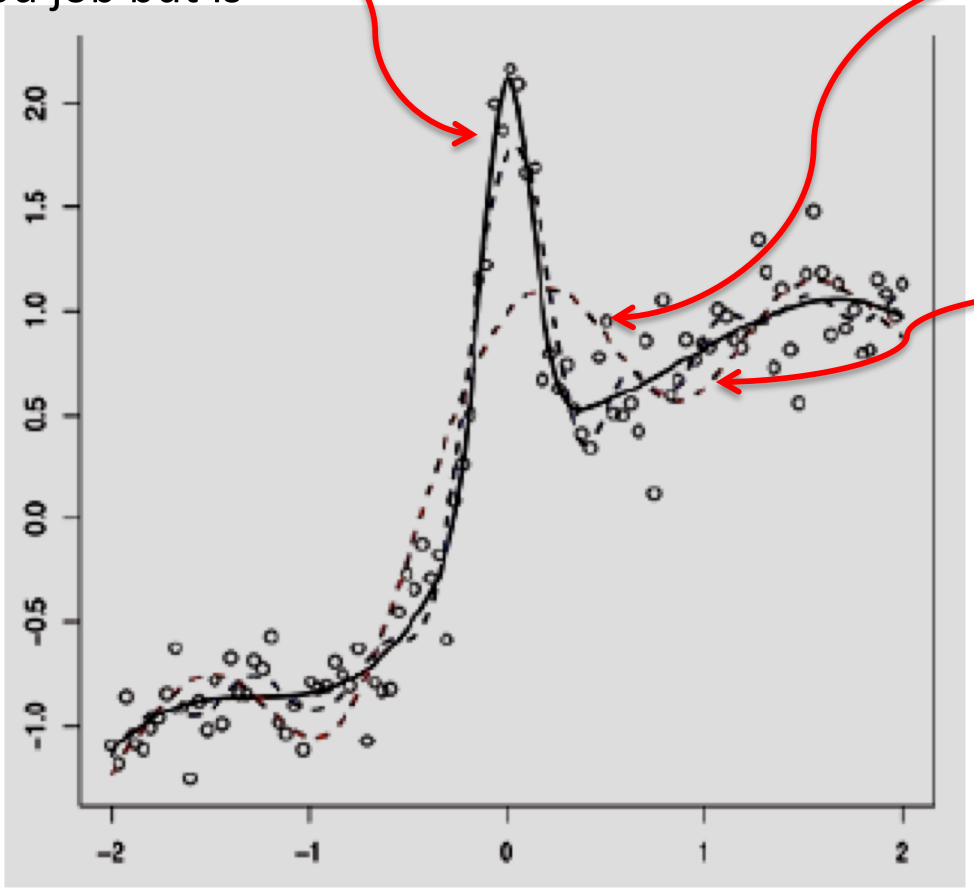
Many variants have been developed: Bezier curves, natural splines, B-splines (basis), NURBS (non-uniform rational B-spline), M-splines (non-negative), I-spline (monotone), T-splines (terminated NURBS), box splines (multivariate B-splines), spline wavelet (wavelet transform based on B-splines), etc.

The challenge of spline knot selection

7 knots chosen by user does a good job but is subjective

5 knots chosen by R misses peak

15 knots chosen by R has too many wiggles in smooth areas



Kass 2008

Modern techniques prune knot selection based on likelihood measures

Large literature on local regression techniques

Extensive software is available in the R/CRAN environment

Some books on local regression:

K. Takezawa, *Introduction to Nonparametric Regression (2005)*

W. Klemela, *Multivariate Nonparametric Regression and Visualization with R and Applications to Finance (2014)*

C. Loader, *Local Regression and Likelihood (1999)*

J.-P. Chiles & P. Delfiner, *Geostatistics: Modeling Spatial Uncertainty (2012)*

D. Ruppert, M. Wand & R. Carroll, *Semiparametric Regression (2003, 2nd ed in press)*

Comment for astronomers

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

Well-established methods like KDE and the NW estimator have asymptotic confidence bands. For many 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.