Nonparametric and robust statistics

Eric Feigelson

3rd INPE Advanced School in Astrophysics:
Astrostatistics 2009

Outline

Motivation for nonparametric/robust methods I

Most standard statistical procedures treat situations where:

- the underlying dataset or population is homogeneous without outliers or multiple populations
- the data are selected from the population with bias (i.i.d)
- the physical phenomenon is described by a parametric model
- the accuracy of measurement is invariant across the sample (homoscedasticity)

But there is little reason to believe these assumptions all apply in real studies! In many cases, our observations have limitations: e.g. inhomogeneous with background, instrumental effects, or interlopers of different populations. In other cases, our understanding is too primitive to establish them: e.g. astrophysical theories often do not predict whether the relationship is linear, or the scatter is Gaussian, in linear or logarithmic variables.

Motivation for nonparametric/robust methods II

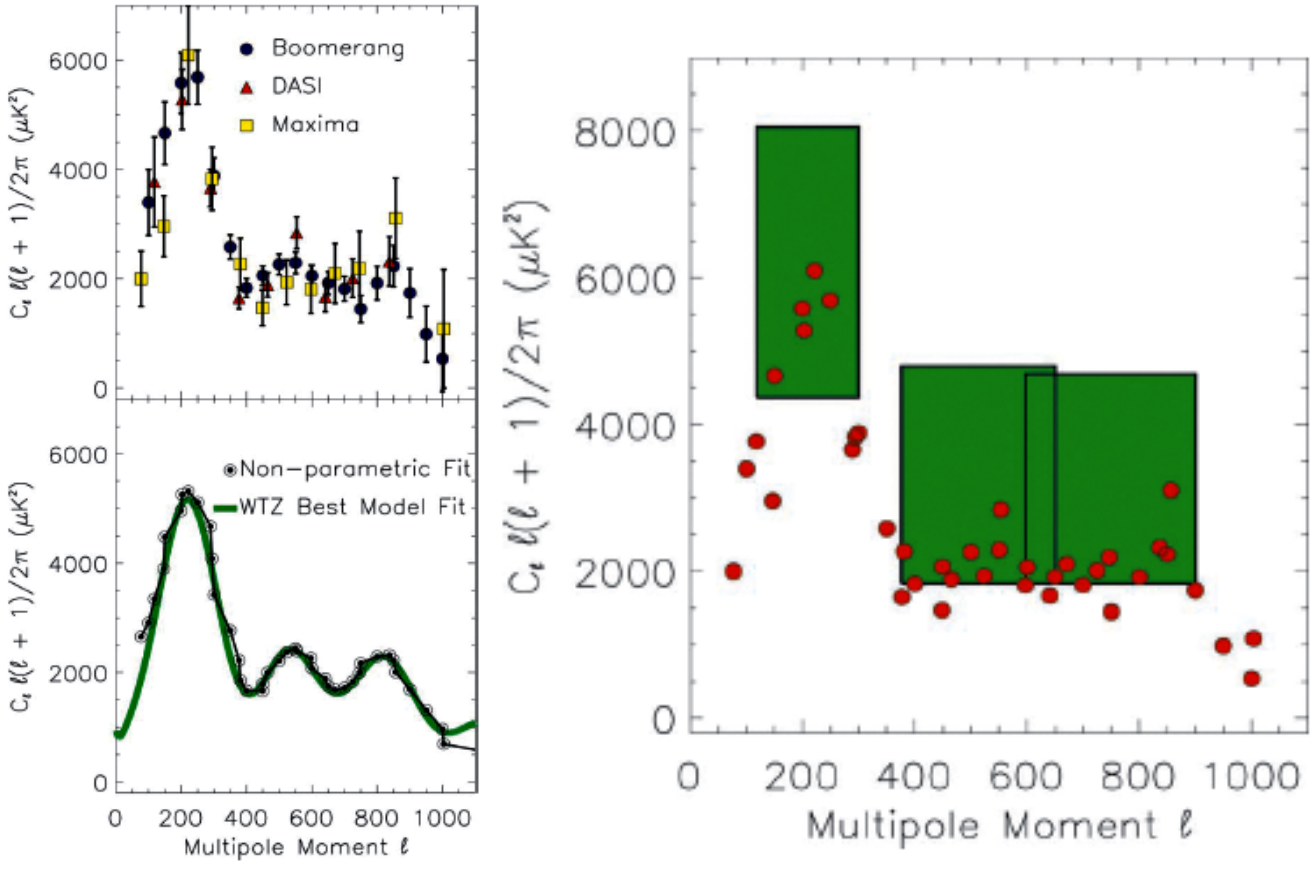
A simplistic parametric model may obfuscate real phenomena. Consider the radial profile of starlight in elliptical galaxies. Thousands of studies have assumed and argued among several parametric models:

1. $I \propto (1 + r/r_c)^{-2}$ (Hubble, 1930)
2. $\log I = -3.33[(r/r_c)^{1/4} - 1]$ (de Vaucouleurs, 1948)
3. $I \propto [(1 + r^2/r_c^2)^{-1/2} - (1 + r_t^2/r_c^2)^{-1/2}]^2$ (King, 1962)
4. $\rho \propto (r/r_c)^{-1}(1 + r/r_c)^{-2}$ (Navarro et al., 1997)

But we now know that luminous elliptical galaxies are complicated triaxial structures dominated by Dark Matter formed by sequences of galaxy mergers. None of these models have a realistic physical basis. Woodroofe/Mateo and Merritt et al. apply modern nonparametric modeling to infer the Dark Matter distribution in elliptical galaxies directly from the data without unreliable assumptions (ApJ, 2005–).

An astrostatistical application

In cosmology, Miller/Nichol/Wasserman/Genovese (ApJ 2002) establish the three-peak structure of the cosmic microwave background fluctuation spectrum using nonparametric methods, but without the high precision of parametric modeling.



Examples of nonparametric methods

Nonparametric methods treat both deficiencies with models and problems with the data.

- Distribution-free (e.g., independent of normality or powerlaw assumptions)
- Tests of hypotheses without parameters
- Density estimation (shape of distributions without parameterization)
- k-sample tests, correlation coefficients
- Robust against outliers (often using rank statistics, M-estimators)
- Classificatory variables

Estimates of the underlying population distribution

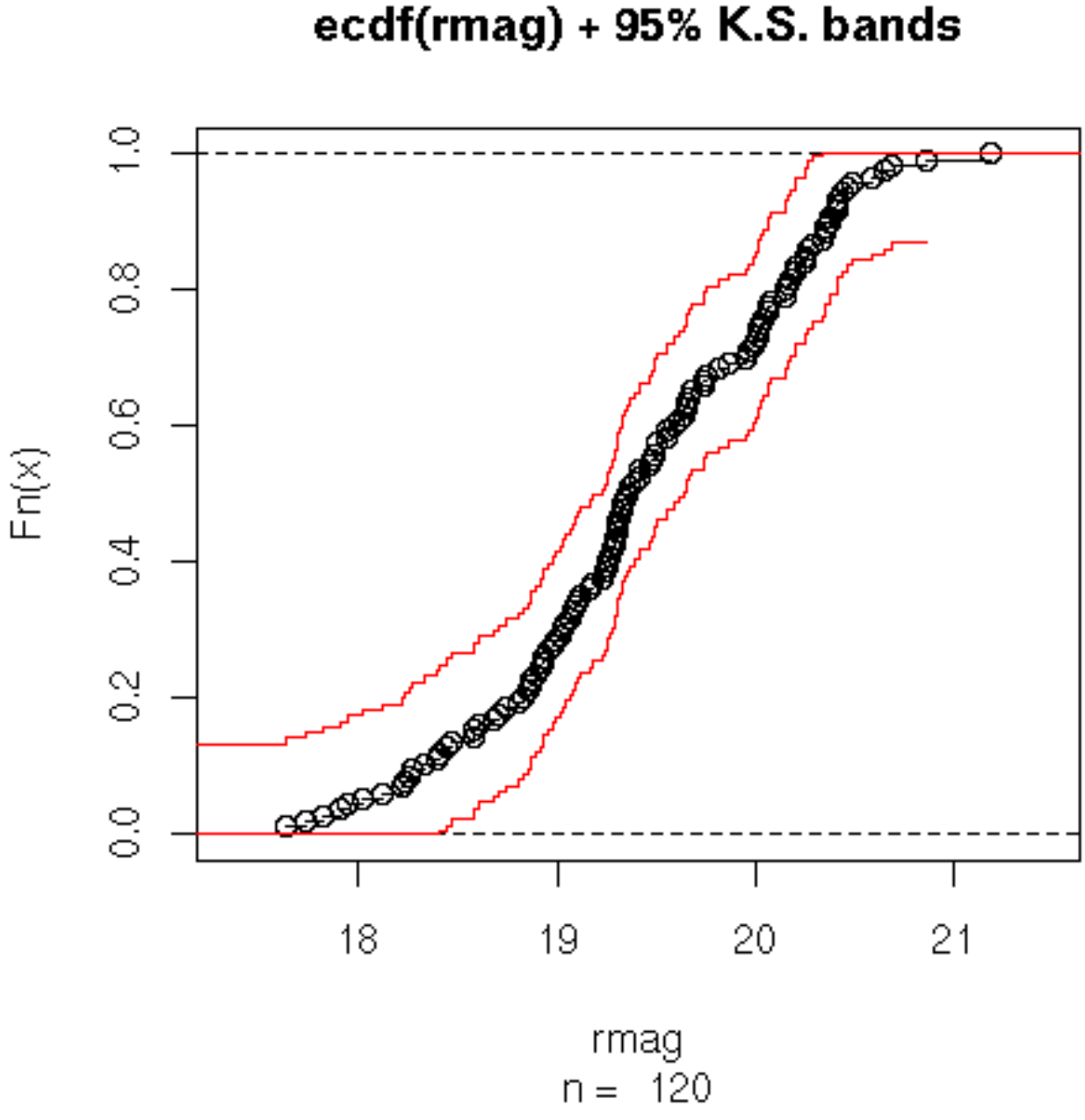
The *empirical distribution function (e.d.f.)* of an i.i.d. sample from an unknown continuous population is the unbiased and asymptotically consistent estimator of the underlying p.d.f. The e.d.f. consists of the data values presented as a cumulative, normalized, discontinuous step-function,

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I[X_i \leq x]$$

with mean and variance

$$E[\widehat{F}_n] = F(x)$$

$$var[\widehat{F}_n] = \frac{F(x)[1 - F(x)]}{n}.$$


ecdf(rmag) + 95% K.S. bands
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
18
19
20
21
rmag
n = 120

Hypothesis tests using the e.d.f. I

We can test the null hypothesis that the data are drawn from a specified p.d.f.($H_o : F(x) = F_0(x)$, one-sample test) or that different datasets are drawn from the same p.d.f. ($H_o : F_1(x) = F_2(x)$, two- or k-sample test). Three statistics and tests are available:

Kolmogorov-Smirnov $M_n = \max_x |\widehat{F}_n(x) - F_0(x)|$

Cramer-von Mises $W_n^2 = n \sum_{i-1}^{n} (\widehat{F}_n(x_i) - F_0(x_i))^2$

Anderson-Darling $A_n^2 = n \sum_{i=1}^{n} \frac{(\widehat{F}_n(x_i) - F_0(x_i))^2}{F_0(x_i)(1-F_0(x_i)}$

Hypothesis tests using the e.d.f. II

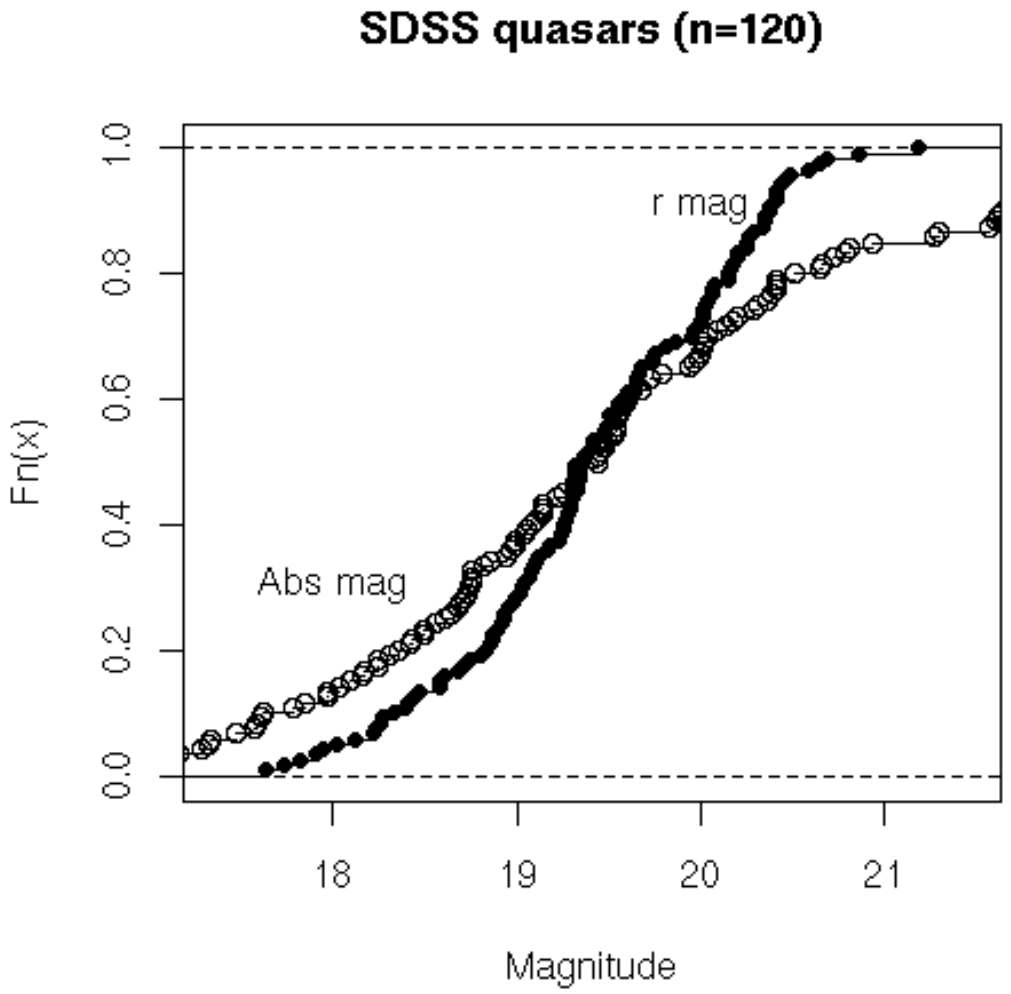
The K-S test is sensitive only to global differences in the distributions. The CvM test is generally better, and the A-D test is best for maintaining sensitivity towards the edges. Astronomers overuse the K-S test!

Critical values for these tests are tabulated for small-N and have formulae for large-N. These are valid for any distribution.

WARNING 1: Tabulated significance levels for K-S one-sample test are not valid if the model parameters were estimated from the dataset under study. The model must be known in advance (e.g. from another dataset). Otherwise, estimate the significance levels using bootstrap resamples.

WARNING 2: None of these e.d.f. tests are distribution-free in 2 or more dimensions.

Comparison of K-S and CvM 2-sample tests



Absolute magnitude values are offset to same median as r mags
Results of 2-sample tests: KS gives P=0.05, CvM gives P=0.004

Quantile function

The quantile (= percentile) function is the inverse of the e.d.f.: What value of x corresponds to a specified value of $F(x)$. Convenient for reducing very large samples. Q-Q plots useful to compare distributions.

The *median* (50% quantile) is the most robust and stable statistic of location of a distribution. Estimated as middle value if n is odd, and mean of two central values if n is even. Better than mean unless normality is known. Robust measure of dispersion is the *Median Absolute Deviation (MAD)* normalized to the Gaussian standard deviation:

$$MADN(X) = Med|X_i - Med(X)|/0.6745.$$

The quantile function is not accurate for small-N samples (e.g., 25% and 75% quartiles easy for n=8 but difficult for n=9.

Bayesian nonparametrics

"Nonparametric Bayesian inference is an oxymoron and a misnomer. Bayesian inference by definition always requires a well-defined probability model ... Nonparametric Bayesian inference traditionally refers to ... inference comparable to classical nonparametric inference, such as kernel density estimation, scatterplot smoothers, etc."

Müller & Quintana, Stat. Sci. 2004

When nonparametrics doesn't work well

Parametric functions known Use frequentist or Bayesian likelihood-based methods.

Multivariate problems OK if variables are independent (e.g. Kendall's τ rank correlation coefficient). But generally, there is no unique ranking (e.g. 2-dimensional K-S test probabilities are not distribution-free).

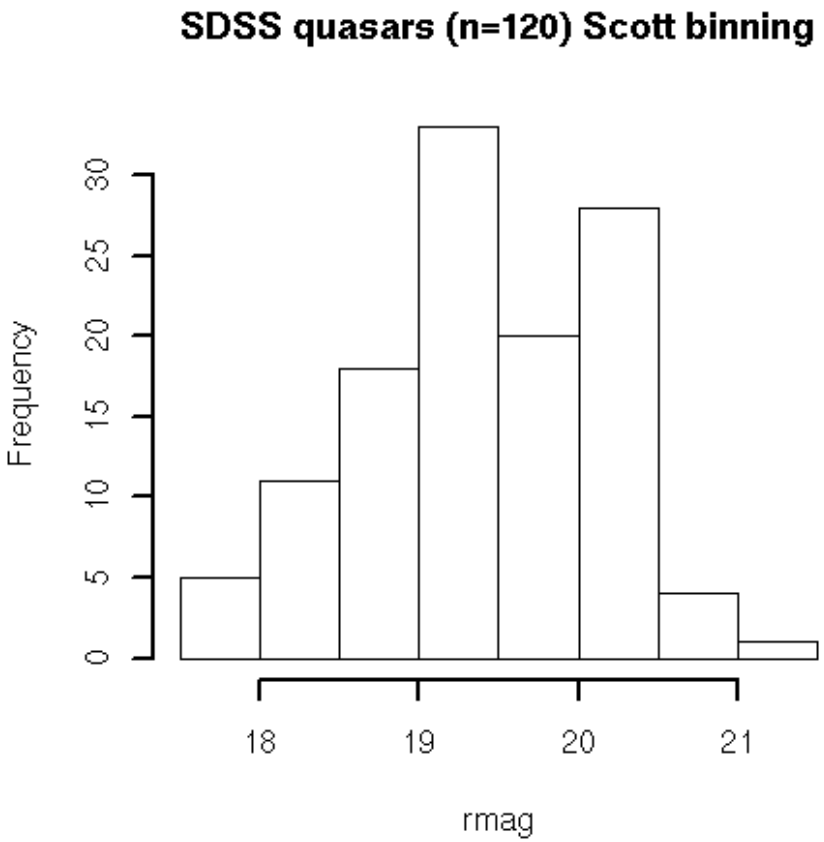
Nonparametric density estimation: Histograms

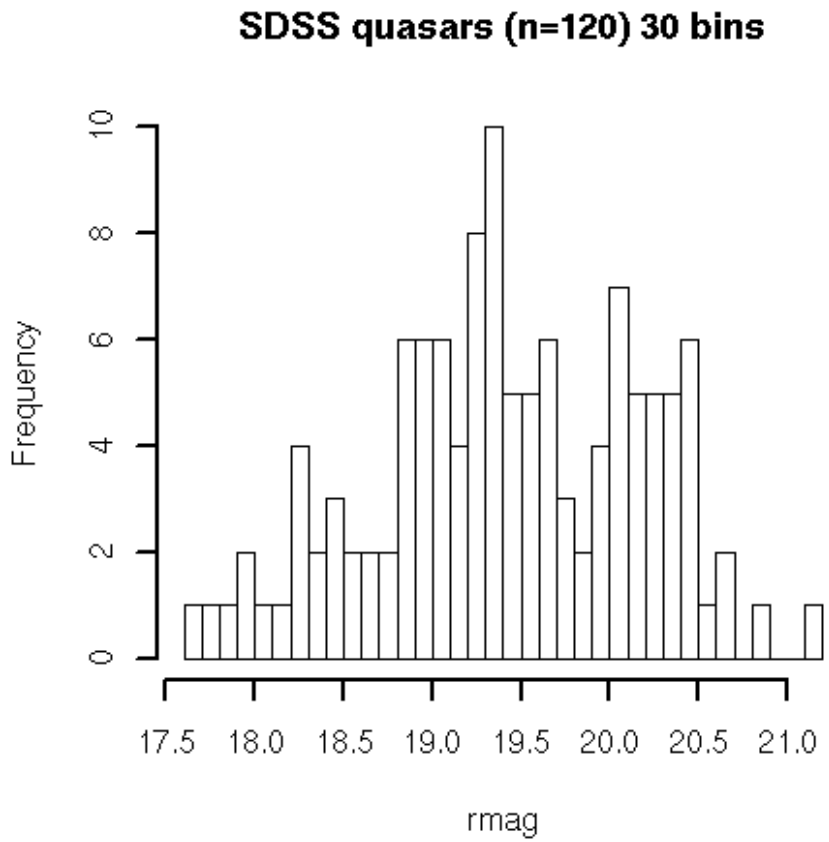
Astronomers are often worried about the interdependency in the e.d.f. prefer estimating the differential distribution using a histogram. Normality of the bin counts is useful ($\sigma \sim \sqrt{N_{bin}}$). However, histograms have a number of important difficulties:

- estimating a smooth distribution by a discrete function
- loss of information within the bin. Particularly bad for skewed distributions like powerlaw (= Pareto) relationships.
- arbitrary choice of bin width. Heuristic choices (e.g. $\Delta = 3.5 s.d./N^{1/3}$, Scott 1992) or optimize in bias-variance plane by cross-validation using parts of the dataset.
- arbitrary choice of zero-point
- no treatment of sample heterogeneity (e.g. outliers)

Statisticians do not recommend use of histograms for statistical evaluation of empirical distributions

Comparison of histograms





Kernel density estimation I

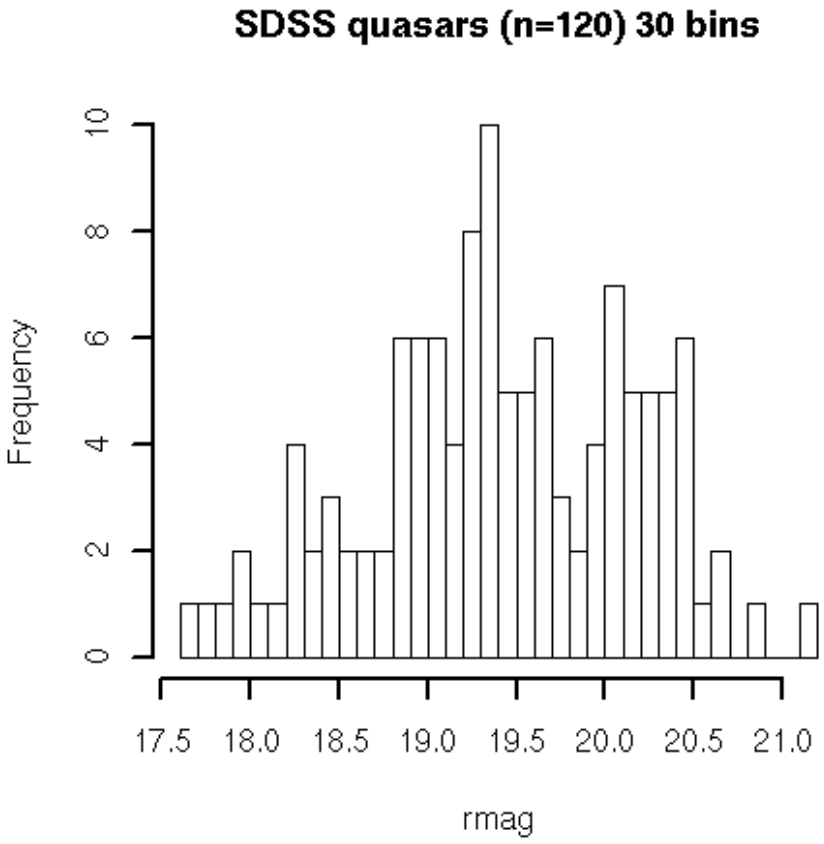
Define the *kernel density estimator* by

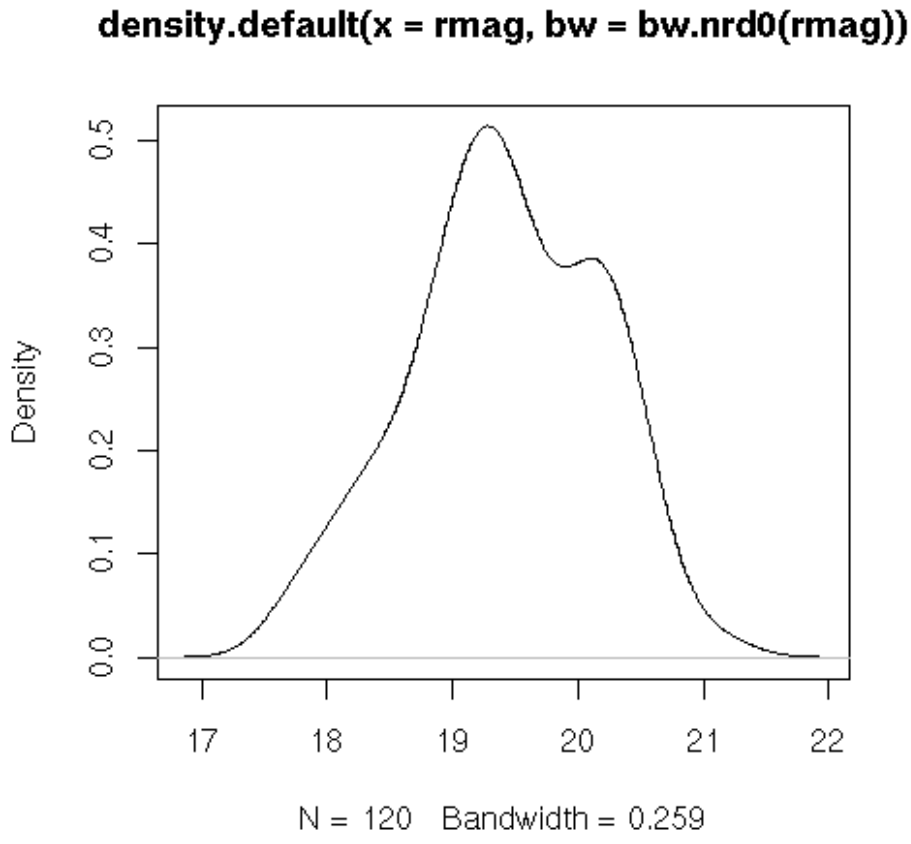
$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right).$$

where K is a normalized, symmetric function so that $\int_{-\infty}^{\infty} \hat{f}_n(x)dx = 1$. Common choices:

- Uniform kernel: $K(u) = \frac{1}{2} I[|u| \leq 1]$
- Triangle kernel: $K(u) = (1 - |u|) I[|u| \leq 1]$
- Epanechnikov kernel: $K(u) = \frac{3}{4}(1 - u^2) I[|u| \leq 1]$
- Gaussian kernel: $K(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}u^2)$

Comparison of histogram and k.d.e.



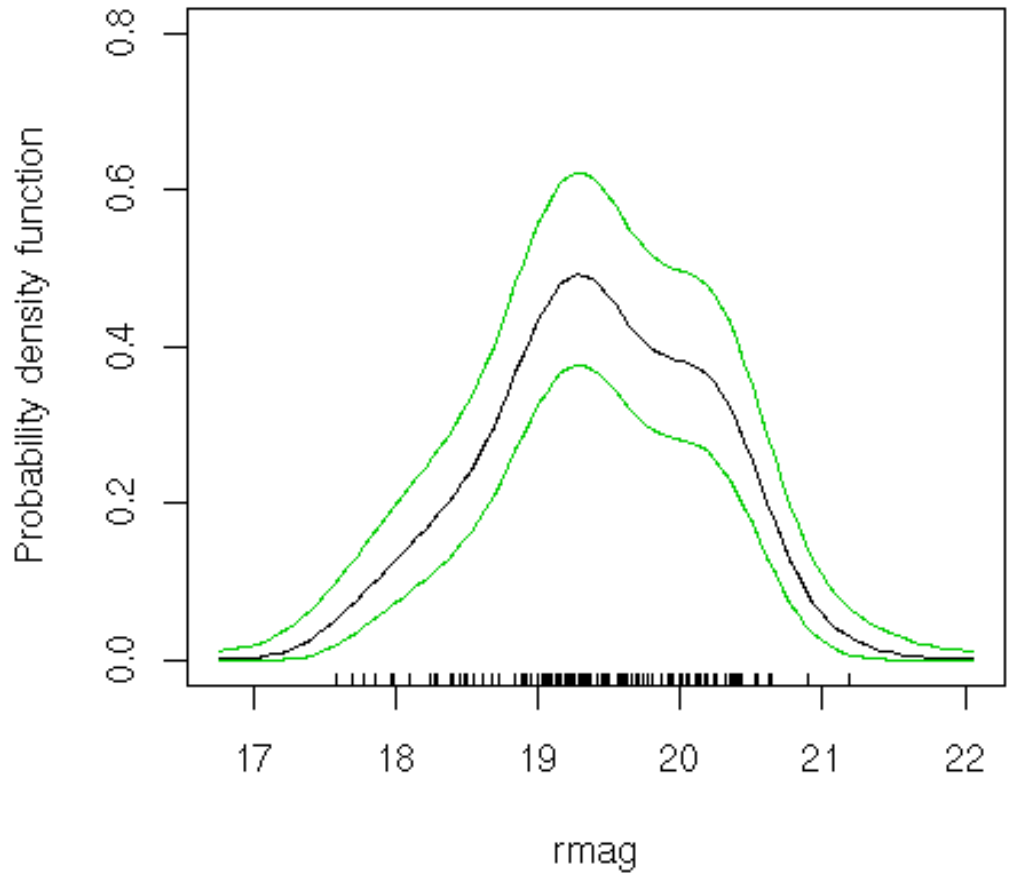


Silverman, cross-validation and Sheather-Jones bandwidths range $0.22 < \Delta < 0.30$.

Kernel density estimation II

Compared to histograms, kernel estimators

- no arbitrary choice of origin
- smoother than the discontinuous histogram
- flexible choice of bandwidth balancing bias vs. variance. For Gaussian kernel, the optimal (minimum MISE = mean integrated square error) bandwidth is $h_{MISE} = 1.06\hat{\sigma}n^{-\frac{1}{5}}$ where $\hat{\sigma}$ is sample standard deviation. Generally, an optimal bandwidth h_x can be found through cross-validation.
- KDE theory easily generalized to multivariate data
- Confidence bands obtained by bootstrap or cross-validation


Probability density function
0.0
0.2
0.4
0.6
0.8
17
18
19
20
21
22
rmag

Kernel density estimation III

Consider a bivariate dataset (x, y) where we seek a density estimate

$$m(x) = E[Y|X = x] = \int y \frac{f(x, y)}{f(x)} dy$$

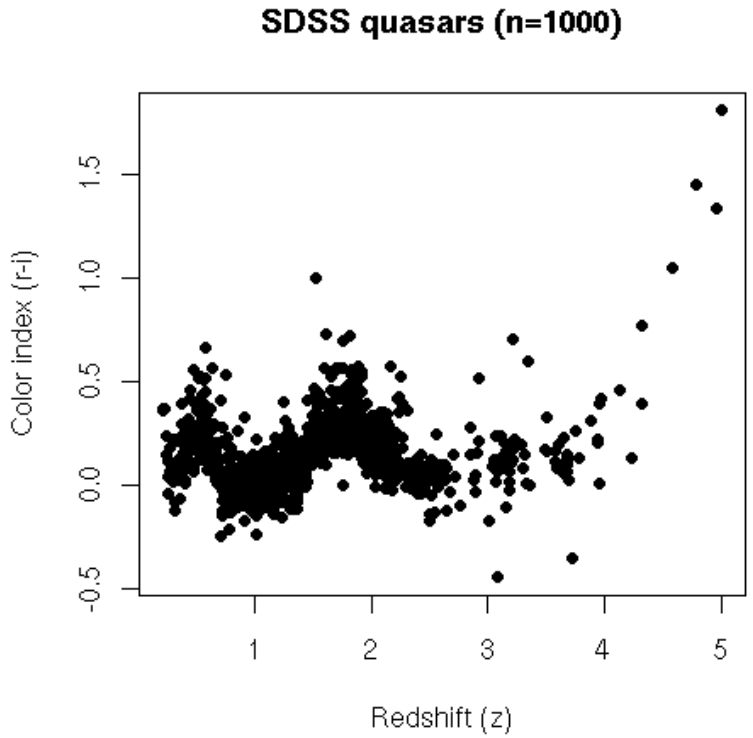
where $f(x)$ and $f(x, y)$ are the marginal density of X and the joint density of X and Y respectively. Substitute the empirical version,

$$\hat{m}_{NW}(x) = \frac{\sum_{i=1}^{n} K((x - X_i)/h_x) Y_i}{\sum_{i=1}^{n} K((x - X_i)/h_x)} \equiv \sum_{i=1}^{n} W_{hi}(x) Y_i,$$

where h_x now is a function of x. This is the *Nadaraya-Watson kernel estimator* which is a weighted average of the Y_i in a neighborhood around x. The NW KDE is a weighted least squares estimator with many advantageous properties, adapting the kernel width to the local density of points.

The nonparametric regression problem

The NW estimator treats complex problems involving multivariate data which cannot be fit by any reasonable parametric function:



A huge recent effort develops methodology to address such problems, extensions of nonparametric density estimation (i.e., smoothing) but with variance estimates. It is variously called local, nonparametric or semi-parametric regression.

The LOESS estimator

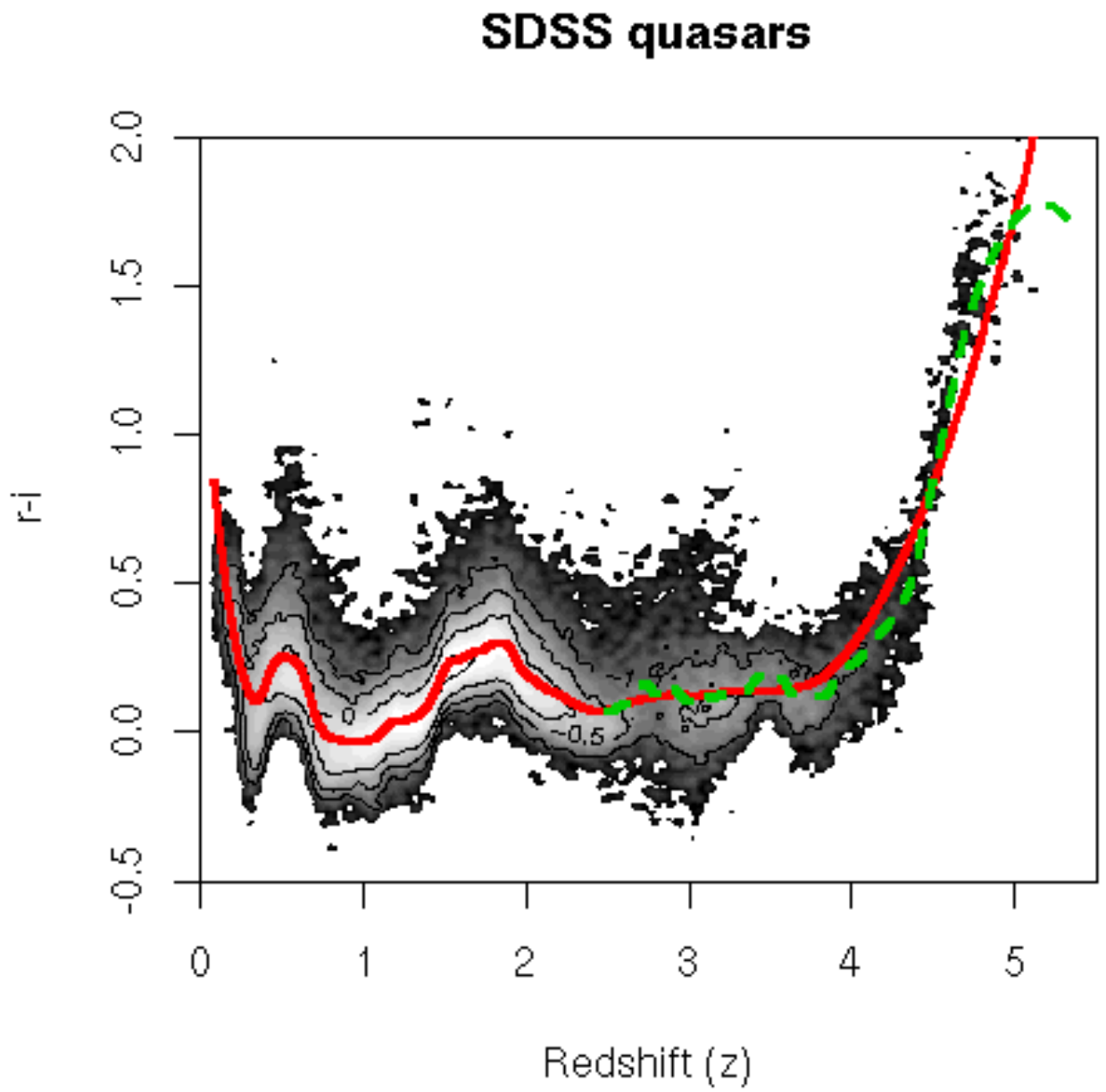
The Nadaraya-Watson estimator essentially gives locally weighted averages of unevenly-spaced data. The LOESS estimator is a generalization giving locally weighted polynomial regressions of unevenly-spaced data. Here the weights $W_i(x)$ become the 'hat matrix' **H** where the bandwidths h_x now varies inversely with the local data density. h_x is related to k-nearest neighbor distance estimators. The 'span' regulating h_x values is again estimated through cross-validation.

The LOESS estimator at x can also be viewed as the polynomial function $\hat{p}(x)$ that minimizes

$$\sum_{i=1}^{n}[Y_i - p(X_i)]^2 Wt(|x - X_i|/\Delta_x) \quad \text{where } \Delta_x = max|X_i - x|$$

with an appropriate weighting function Wt.

The LOESS estimator; SDSS quasars (N=77,429)



Conclusions

- Nonparametric statistics are extremely valuable to astronomers when the parametric functions are not astrophysically well-established
- Nonparametric methods are good when common assumptions are violated (normality, linearity, homogeneity)
- The e.d.f. is a basic estimator, and e.d.f.-based tests are valuable. Don't misuse the Kolmogorov-Smirnov test!
- Medians and MADs are excellent measures of location and spread. Quantiles are often useful, especially with outliers.
- Histograms are helpful exploratory tools, but don't use them for statistical/scientific inference!
- Kernel density estimation and its extensions with adaptive kernels (N-W, loess) are important for astronomy.

References

An Introduction to Modern Nonparametric Statistics by J. J. Higgins (2004, Thomson/Brooks-Cole) is a readable undergraduate-level text. It covers traditional topics (univariate and multivariate methods, one-sample and k-sample tests, tests for trends and correlation) as well as brief treatments of censoring, bootstrap methods, smoothing and robust estimation.

Practical Nonparametric Statistics by W. J. Conover (1999, 3rd ed., Wiley) is a thorough and authoritative presentation of classical nonparametrics. The text covers binomial distribution tests, contingency tables, rank tests, and Kolmogorov-Smirnov-type tests.

Introduction to Nonparametric Regression by K. Takezawa. One of several new advanced texts on density estimation. It treats kernel density estimation, local regression, multivariate regression, histogram smoothing, and pattern recognition. **R** code is included.