Astronomical time series (and more on R)

Eric Feigelson

3rd INPE Advanced School in Astrophysics: Astrostatistics 2009

Outline

1. Time series in astronomy
2. Frequency domain methods
3. More about R
4. R in action: Nonparametric/robust statistics

Time series in astronomy

- Periodic phenomena: binary orbits (stars, extrasolar planets); stellar rotation (radio pulsars); pulsation (helioseismology, Cepheids)
- Stochastic phenomena: accretion (CVs, X-ray binaries, Seyfert gals, quasars); scintillation (interplanetary & interstellar media); jet variations (blazars)
- Explosive phenomena: thermonuclear (novae, X-ray bursts), magnetic reconnection (solar/stellar flares), star death (supernovae, gamma-ray bursts)

Difficulties in astronomical time series

Gapped data streams:

Diurnal & monthly cycles; satellite orbital cycles; telescope allocations

Heteroscedastic measurement errors:

Signal-to-noise ratio differs from point to point

Poisson processes:

Individual photon/particle events in high-energy astronomy

Important Fourier Functions

Discrete Fourier Transform

$$d(\omega_j) = n^{-1/2} \sum_{t=1}^{n} x_t exp(-2\pi i t \omega_j)$$

$$d(\omega_j) = n^{-1/2} \sum_{t=1}^{n} x_t cos(2\pi i \omega_j t) - i n^{-1/2} \sum_{t=1}^{n} x_t sin(2\pi i \omega_j t)$$

Classical (Schuster) Periodogram

$$I(\omega_j) = |d(\omega_j)|^2$$

Spectral Density

$$f(\omega) = \sum_{h=-\infty}^{h=\infty} exp(-2\pi i \omega h)\gamma(h)$$

Fourier analysis reveals nothing of the evolution in time, but rather reveals the variance of the signal at different frequencies.

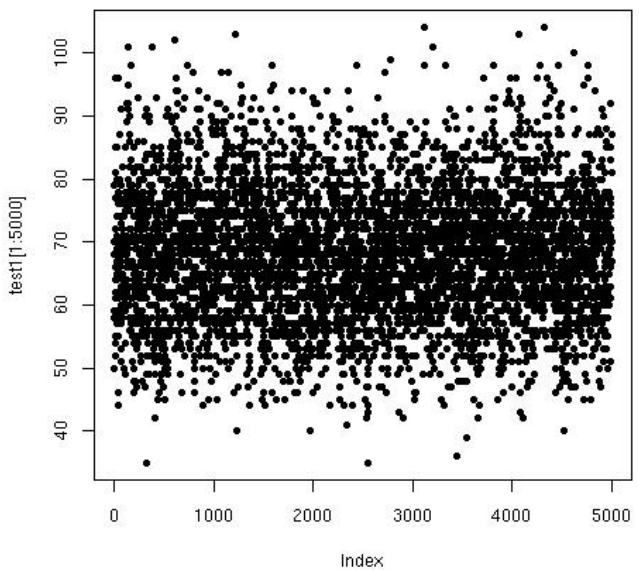
It can be proved that the classical periodogram is an estimator of the spectral density, the Fourier transform of the autocovariance function.

Fourier analysis has restrictive assumptions: an infinitely long dataset of equally-spaced observations; homoscedastic Gaussian noise with purely periodic signals; sinusoidal shape

Formally, the probability of a periodic signal in Gaussian noise is $P \propto e^{d(\omega_j)/\sigma^2}$. But this formula is often not applicable, and probabilities are difficult to infer.

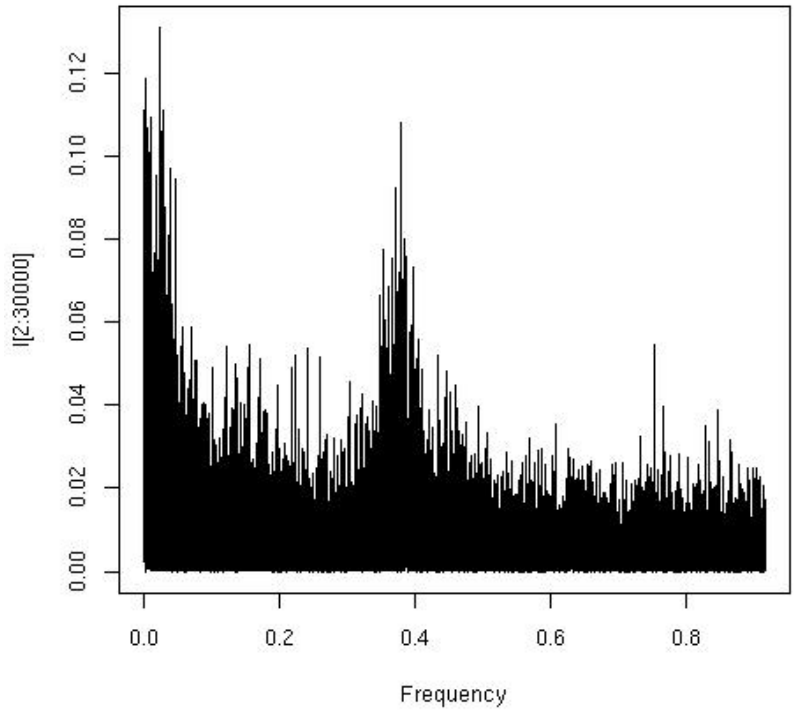
Ginga observations of X-ray binary GX 5-1

GX 5-1 is a binary star system with gas from a normal companion accreting onto a neutron star. Highly variable X-rays are produced in the inner accretion disk. XRB time series often show 'red noise' and 'quasi-periodic oscillations', probably from inhomogeneities in the disk. We plot below the first 5000 of 65,536 count rates from Ginga satellite observations during the 1980s.




```
gx=scan("GX.dat")
t=1:5000
plot(t,gx[1:5000],pch=20)
```

Fast Fourier Transform of the GX 5-1 time series reveals the ‘red noise’ (high spectral amplitude at small frequencies), the QPO (broadened spectral peak around 0.35), and white noise.




```
f = 0:32768/65536
I = (4/65536)*abs(fft(gx)/sqrt(65536))^ 2
plot(f[2:60000],I[2:60000],type="l",xlab="Frequency")
```

Limitations of the spectral density

But the classical periodogram is not a good estimator! E.g. it is formally 'inconsistent' because the number of parameters grows with the number of datapoints. The discrete Fourier transform and its probabilities also depends on several strong assumptions which are rarely achieved in real astronomical data: evenly spaced data of infinite duration with a high sampling rate (Nyquist frequency), Gaussian noise, single frequency periodicity with sinusoidal shape and stationary behavior. Formal statement of strict stationarity: $P\{x_{t_1} \leq c_1, ...s_{x_K} \leq c_k\} = P\{x_{t_{1+h}} \leq c+1, ..., x_{t_{k+h}} \leq c_k\}$.

Each of these constraints is violated in various astronomical problems. Data spacing may be affected by daily/monthly/orbital cycles. Period may be comparable to the sampling time. Noise may be Poissonian or quasi-Gaussian with heavy tails. Several periods may be present (e.g. helioseismology). Shape may be non-sinusoidal (e.g. elliptical orbits, eclipses). Periods may not be constant (e.g. QPOs in an accretion disk).

Improving the spectral density I

The estimator can be improved with **smoothing**,

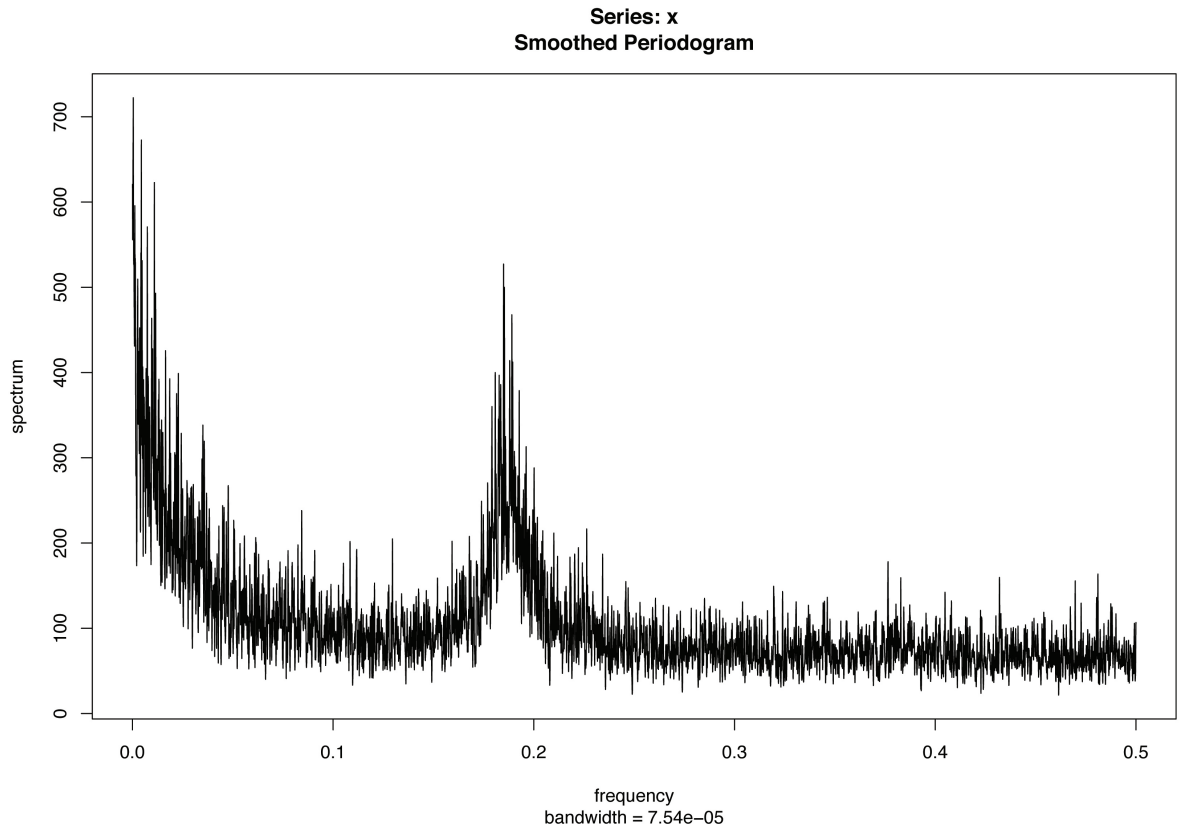
$$\hat{f}(\omega_j) = \frac{1}{2m_1} \sum_{k=-m}^{m} I(\omega_{j-k}).$$

This reduces variance but introduces bias. It is not obvious how to choose the smoothing bandwidth m or the smoothing function (e.g. Daniell or boxcar kernel).

Tapering reduces the signal amplitude at the ends of the dataset to alleviate the bias due to leakage between frequencies in the spectral density produced by the finite length of the dataset. Consider for example the cosine taper

$$h_t = 0.5[1 + cos(2\pi(t - \bar{t})/n)]$$

applied as a weight to the initial and terminal n datapoints. The Fourier transform of the taper function is known as the spectral window. Other widely used options include the Fejer and Parzen windows and multitapering. Tapering decreases bias but increases




```
png(file="GX_sm_tap_fft.png")
k = kernel("modified.daniell", c(7,7))
spec = spectrum(gx, k, method="pgram", taper=0.3, fast=TRUE, detrend=TRUE, log="no")
dev.off()
```

Improving the spectral density II

Pre-whitening is another bias reduction technique based on removing (filtering) strong signals from the dataset. It is widely used in radio astronomy imaging where it is known as the CLEAN algorithm, and has been adapted to astronomical time series (Roberts et al. 1987).

A variety of **linear filters** can be applied to the time domain data prior to spectral analysis. When aperiodic long-term trends are present, they can be removed by spline fitting (high-pass filter). A kernel smoother, such as the moving average, will reduce the high-frequency noise (low-pass filter). Use of a parametric autoregressive model instead of a nonparametric smoother allows likelihood-based model selection (e.g. BIC).

Improving the spectral density III

Harmonic analysis of unevenly spaced data is problematic due to the loss of information and increase in aliasing.

The **Lomb-Scargle periodogram** is widely used in astronomy to alleviate aliasing from unevenly spaced:

$$d_{LS}(\omega) = \frac{1}{2}\left(\frac{[\sum_{t=1}^{n} x_t cos\omega(x_t-\tau)]^2}{\sum_{i=1}^{n} cos^2\omega(x_t-\tau)} + \frac{[\sum_{t=1}^{n} x_t sin\omega(x_t-\tau)]^2}{\sum_{i=1}^{n} sin^2\omega(x_t-\tau)}\right)$$

where $\quad tan(2\omega\tau) = (\sum_{i=1}^{n} sin2\omega x_t)(\sum_{i=1}^{n} cos2\omega x_t)^{-1}$

d_{LS} reduces to the classical periodogram d for evenly-spaced data. Bretthorst (2003) demonstrates that the Lomb-Scargle periodogram is the unique sufficient statistic for a single stationary sinusoidal signal in Gaussian noise based on Bayes theorem assuming simple priors.

Some other methods for periodicity searching

Phase dispersion measure (Stellingwerf 1972) Data are folded modulo many periods, grouped into phase bins, and intra-bin variance is compared to inter-bin variance using χ^2. Non-parametric procedure well-adapted to unevenly spaced data and non-sinusoidal shapes (e.g. eclipses). Very widely used in variable star research, although there is difficulty in deciding which periods to search (Collura et al. 1987).

Minimum string length (Dworetsky 1983) Similar to PDM but simpler: plots length of string connecting datapoints for each period. Related to the Durbin-Watson roughness statistic in econometrics.

Rayleigh and Z_n^2 tests (Leahy et al. 1983) for periodicity search Poisson distributed photon arrival events. Equivalent to Fourier spectrum at high count rates.

Bayesian periodicity search (Gregory & Loredo 1992) Designed for non-sinusoidal periodic shapes observed with Poisson events. Calculates odds ratio for periodic over constant model and most probable shape.

Conclusions on spectral analysis

For challenging problems, smoothing, multitapering, linear filtering, (repeated) pre-whitening and Lomb-Scargle can be used together. Beware that aperiodic but autoregressive processes produce peaks in the spectral densities. Harmonic analysis is a complicated 'art' rather than a straightforward 'procedure'.

It is extremely difficult to derive the significance of a weak periodicity from harmonic analysis. Do not believe analytical estimates (e.g. exponential probability), as they rarely apply to real data. It is essential to make simulations, typically permuting or bootstrapping the data keeping the observing times fixed. Simulations of the final model with the observation times is also advised.

Nonstationary time series

Non-stationary periodic behaviors can be studied using **time-frequency Fourier analysis**. Here the spectral density is calculated in time bins and displayed in a 3-dimensional plot.

Wavelets are now well-developed for non-stationary time series, either periodic or aperiodic. Here the data are transformed using a family of non-sinusoidal orthogonal basis functions with flexibility both in amplitude and temporal scale. The resulting wavelet decomposition is a 3-dimensional plot showing the amplitude of the signal at each scale at each time. Wavelet analysis is often very useful for noise threshholding and low-pass filtering.

Time domain methods

These are covered in a previous lecture by Dr. Salazar. Useful methods include:

- Autocorrelation function
- Partial autocorrelation function
- Autoregressive moving average model (ARMA)
- Extended ARMA models: VAR (vector autoregressive), ARFIMA (ARIMA with long-memory component), GARCH (generalized autoregressive conditional heteroscedastic for stochastic volatility)
- State space models
- Extended state space models: non-stationarity, hidden Markov chains, etc. MCMC evaluation of nonlinear and non-normal (e.g. Poisson) models

Statistical texts and monographs

D. Brillinger, Time Series: Data Analysis and Theory, 2001

C. Chatfield, The Analysis of Time Series: An Introduction, 6th ed., 2003

G. Kitagawa & W. Gersch, Smoothness Priors Analysis of Time Series, 1996

M. B. Priestley, Spectral Analysis and Time Series, 2 vol, 1981

R. H. Shumway and D. S. Stoffer, Time Series Analysis and Its Applications (with R examples), 2nd Ed., 2006

Astronomical references

Bretthorst 2003, "Frequency estimation and generalized Lomb-Scargle periodograms", in Statistical Challenges in Modern Astronomy

Dworetsky 1983, "A period-finding method for sparse randomly spaced observations ....", MNRAS 203, 917

Gregory & Loredo 1992, "A new method for the detection of a periodic signal of unknown shape and period", ApJ 398, 146

Leahy et al. 1983, "On searches for periodic pulsed emission: The Rayleigh test compared to epoch folding",. ApJ 272, 256

Roberts et al. 1987, "Time series analysis with CLEAN ...", AJ 93, 968

Scargle 1982, "Studies in astronomical time series, II. Statistical aspects of spectral analysis of unevenly spaced data", ApJ 263, 835

Stellingwerf 1972, "Period determination using phase dispersion measure", ApJ 224, 953

Vio et al. 2005, "Time series analysis in astronomy: Limits and potentialities, A&A 435, 773

R: A Language and Environment for Statistical Computing

R is the public-domain version of the commercial S-Plus statistical computing package. Integrates data manipulation, graphics and statistical analysis. Uniform documentation and coding standards.

Fully programmable C-like language, similar to IDL. Specializes in vector or matrix inputs; not designed for maps, images or movies.

Easily downloaded from http://www.r-project.org with Windows, Mac or UNIX binaries. Tutorials available in dozens of books (most since 2005) and on-line. 1500 add-on packages collected in Comprehensive R Archive Network http://www.cran.r-project.org.

Some methods covered in Base R

arithmetic & linear algebra, bootstrap resampling, empirical distribution tests, exploratory data analysis, generalized linear modeling, graphics, robust statistics, linear programming, local and ridge regression, maximum likelihood estimation, multivariate analysis, multivariate clustering, neural networks, smoothing, spatial point processes, statistical distributions & random deviates, statistical tests, survival analysis, time series analysis

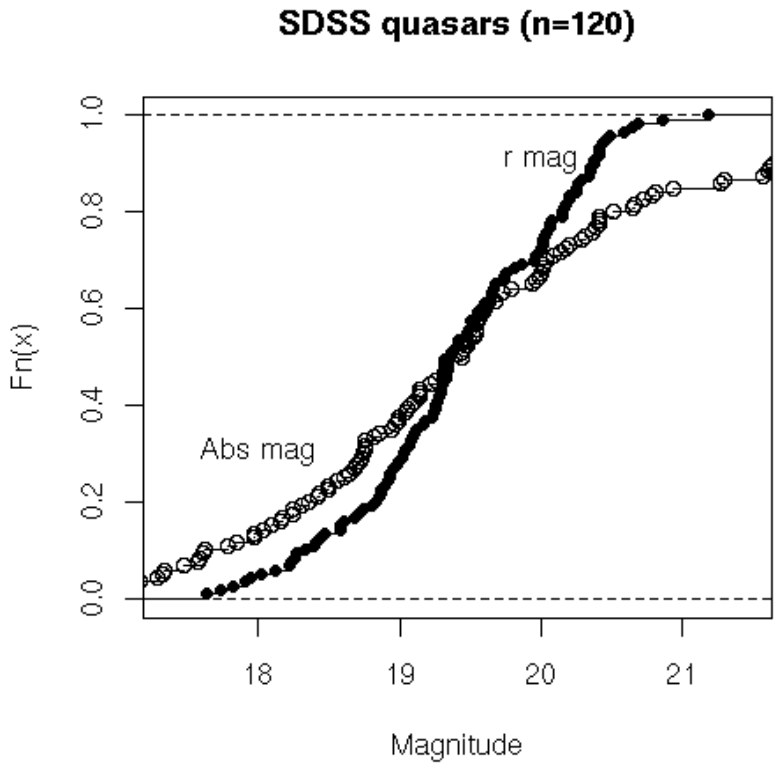
Some methods covered in CRAN

Bayesian computation & MCMC, classification & regression trees, geostatistical modeling, hidden Markov models, irregular time series, kernel-based machine learning, least-angle & lasso regression, likelihood ratios, map projections, mixture models & model-based clustering, nonlinear least squares, multidimensional analysis, multimodality test, multivariate time series, multivariate outlier detection, neural networks, non-linear time series analysis, nonparametric multiple comparisons, omnibus tests for normality, orientation data, parallel coordinates plots, partial least squares, principal curve fits, projection pursuit, quantile regression, random fields, random forest classification, ridge regression, robust regression, self-organizing maps, shape analysis, space-time ecological analysis, spatial analyisis & kriging, spline regressions (MARS, BRUTO), tesselations, three-dimensional visualization, wavelet toolbox

R links to other systems

Interfaces	BUGS, C, C++, Fortran, Java, Perl, Python, Xlisp, XML
I/O	ASCII, binary, bitmap, cgi, FITS, ftp, gzip, HTML, SOAP, URL
Graphics	Grace, GRASS, Gtk, Matlab, OpenGL, Tcl/Tk, Xgobi
Applied math	GSL, lsoda, LAPACK, PVM
Text processor	LaTeX

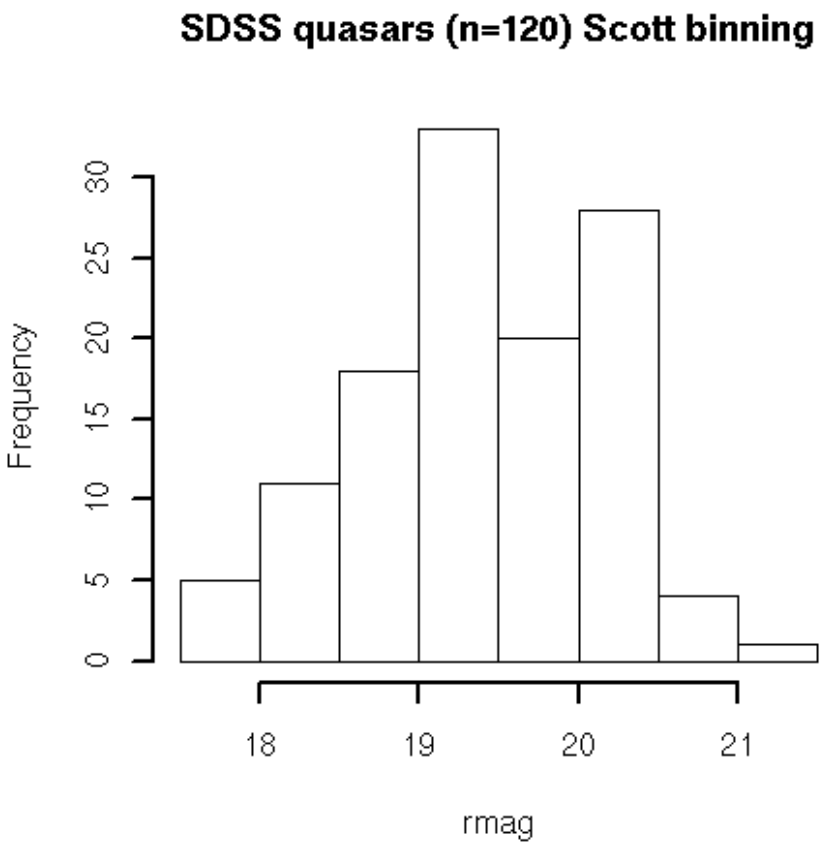
R in action: Nonparametric/robust statistics

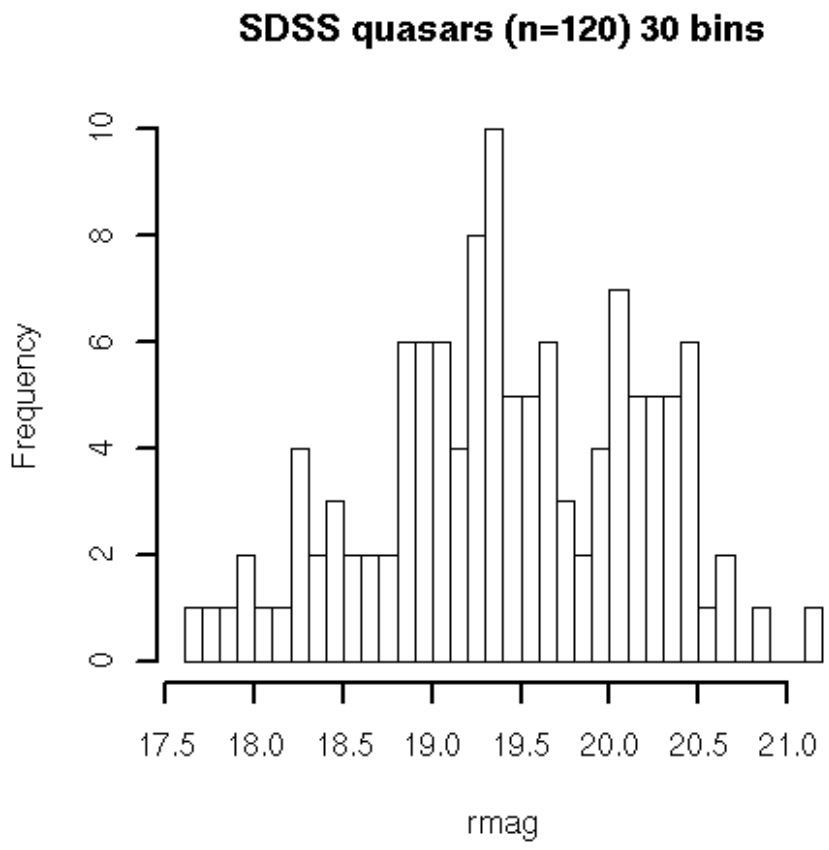


Absolute magnitude values are offset to same median as r mags

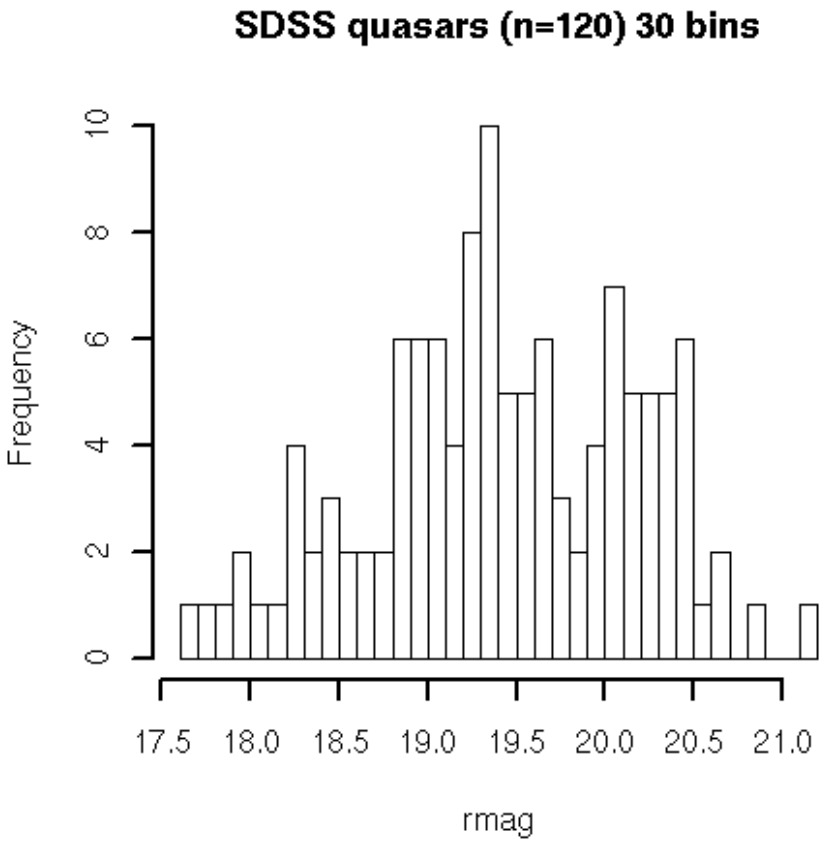
Results of 2-sample tests: KS gives P=0.05, CvM gives P=0.004

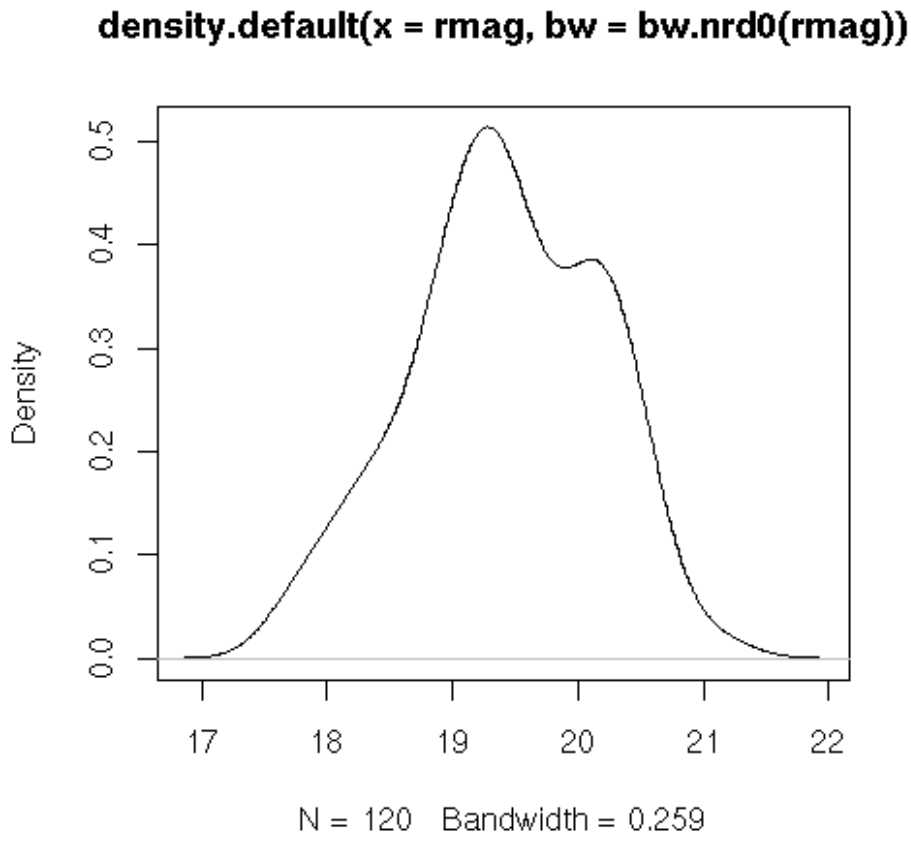
Comparison of histograms





Comparison of histogram and k.d.e.





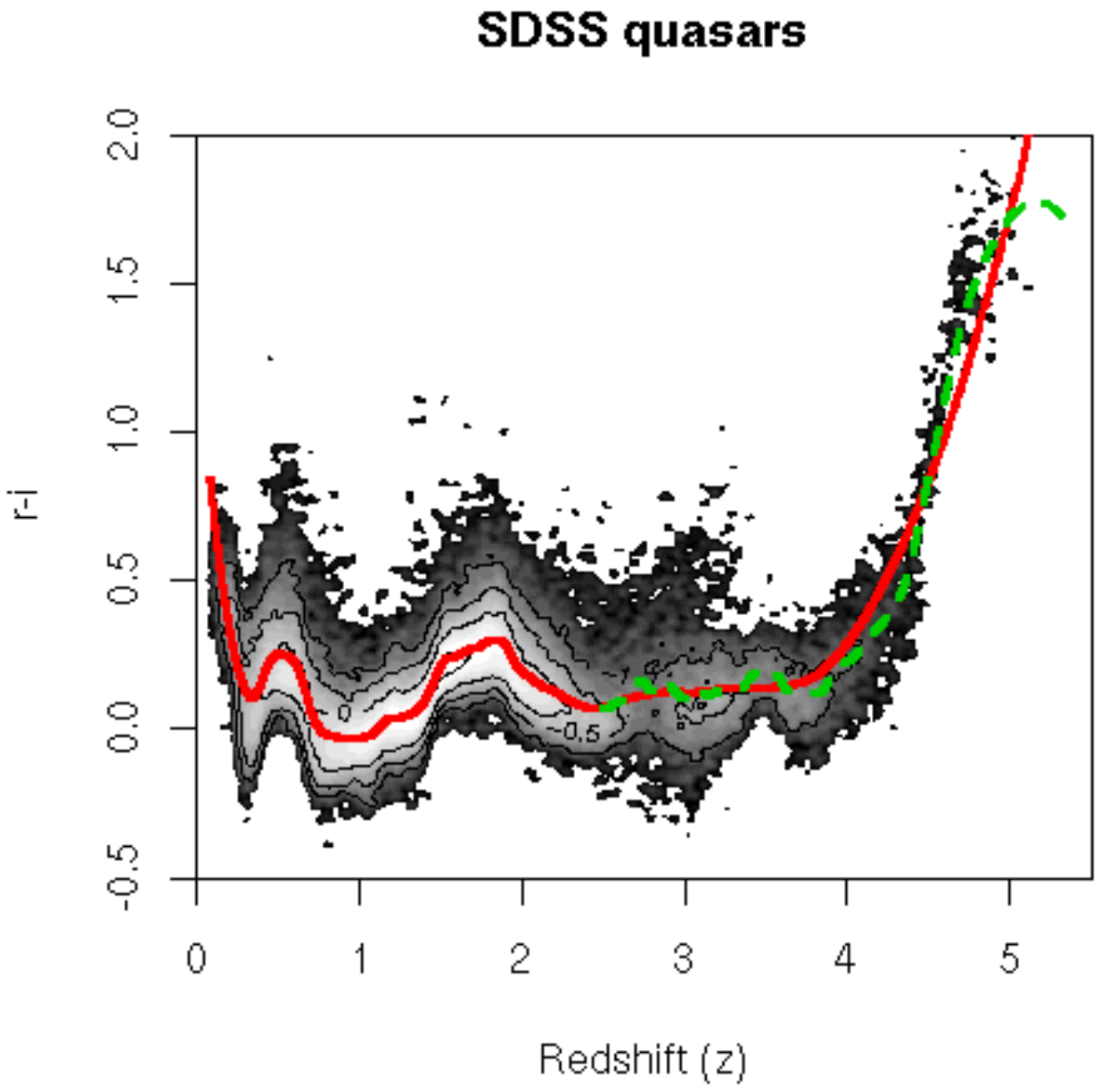
Silverman, cross-validation and Sheather-Jones bandwidths range $0.22 < \Delta < 0.30$.

R scripts for the SDSS quasar r magnitudes

```
# Read dataset of 120 SDSS quasar r magnitudes
qso=
read.table("http://astrostatistics.psu.edu/datasets/SDSS_QSO.dat",
dim(qso) ; names(qso) ; summary(qso)
rmag=qso[1:120,9]
amag=qso[1:120,17]

# Plot e.d.f. with confidence bands install.packages('sfsmisc')
; library('sfsmisc')
ecdf.ksCI(rmag)

# Plot e.d.f.'s
wilcox.test(rmag,amag,conf.int=T)
Absmag=amag+44.7 # sets equal medians
plot(ecdf(rmag),pch=20,xlab="Magnitude")
plot (ecdf(Absmag),add=T) ; text(21,0.7,lab='r mag')
```

```
# Run e.d.f. 2-sample tests
ks.test(rmag,Absmag)
install.packages('cramer') ; library(cramer)
cramer.test(rmag,Absmag)

# Plot histograms and kernel density estimators
hist(rmag,breaks='scott') ; hist(rmag,breaks=30)
plot(density(rmag, bw=bw.nrd0(rmag)))

# Plot k.d.e. with confidence bands
install.packages('sm') ; library(sm)
help('sm.density')
sm.density(rmag) ; tt=sm.density(rmag)
lines(tt$eval.points,tt$upper,col=3) ;
lines(tt$eval.points,tt$lower,col=3)
```

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



R script for the SDSS quasar LOESS plot

```
# Read SDSS quasar sample, N=77,429. Clean bad
photometry
qso=
read.table("http://astrostatistics.psu.edu/datasets/SDSS_QSO.dat",
q1=qso[qso[,10] < 0.3,] ; q1=q1[q1[,12]<0.3,]
dim(q1) ; names(q1) ; summary(q1)
r_i=q1[,9]-q1[,11] ; z=q1[,4]

# Plot two-dimensional smoothed distribution
install.packages('ash') ; library(ash)
nbin=c(500,500) ; ab= matrix(c(0.0,-0.5,5.5,2.),2,2)
bins=bin2(cbind(z1,r_i1),ab,nbin)
f=ash2(bins,c(5,5)) ; f$z=log10(f$z)
image(f$x,f$y,f$z,zl=c(-
2,0.5),col=gray(seq(0,1,by=0.05)),xl=c(0,5.5))
contour(f$x,f$y,f$z,zlim=c(-1,0.5),nlevels=4,add=T)
```

```
# Construct loess local regression lines
z1=q1[,4][order(z)] ; r_i1=r_i[order(z)]
locfit1=loess(r_i1 z1,span=0.1,data.frame(x=z1,y=r_i1))
lines(z1,predict(locfit1),lwd=2,col=2)

z2=z1[z1>2.5]; r_i2=r_i1[z1>2.5]
locfit2=loess(r_i2 z2,span=0.1,data.frame(x=z2,y=r_i2))
lines(z2,predict(locfit2),lwd=2,lty=2,col=3)

# Save evenly-spaced loess fit to file x1=seq(0.0,2.5,by=0.02)
; x2=seq(2.52,5.0,by=0.02)
locfitdat1=predict(locfit1,data.frame(x=x1))
locfitdat2=predict(locfit2,data.frame(x=x2))
write(rbind(x1,locfitdat1) sep=' ',ncol=2,file='qso.txt')
write(rbind(x2,locfitdat2),sep='
',ncol=2,file='qso.txt',append=T)
```