6th COSPAR Capacity Building Workshop
Sinaia, 4-16 June 2007
Joachim Vogt

Basic Analysis Techniques
& Multi-Spacecraft Data
— Computer Session —

Sheet 2 5 June 2007

2 Density estimation using the kernel method

The probability density function (pdf) is a central concept in statistical data analysis, and the most popular instrument for pdf estimation is the histogram. However, the histogram technique produces pdf estimates which have several drawbacks. Histograms not only depend on the chosen binsize but also on the origin of the discretized data range. Furthermore, they are discontinuous at bin boundaries.

The dependency on origin can be removed if an estimator is constructed directly from the definition of the pdf, the so-called naive estimator. It can be written as a sum of rectangular weight (or window) functions centered around the individual data points. In the kernel estimation technique this weight function is replaced by more general functions. The resulting estimator inherits all its smoothness properties from the weight function and will thus be continuous if the weight function is continuous.

In the following we give a short introduction to pdf (density) estimation and present a worked example. For more information the reader is referred to the book of Silverman [1986], see also the online article[1]. provided by the same author.

2.1 Conceptual introduction and theory

If we consider the measurement of a (continuous) physical quantity X as a realisation of an underlying random process, X is understood as a random variable with a pdf $p(x)$ defined through

$$P(a \leq X \leq b) \;=\; \int_a^b p(x)\,\mathrm{d}x$$

where $P(a \leq X \leq b)$ is the probability to find the observed value of X in the range $[a, b]$. In a statistical sense the system of interest is completely determined by its pdf through the

---

[1] http://nedwww.ipac.caltech.edu/level5/March02/Silverman/

expectation operation $E\{\}$:

$$E\{f(X)\} \;=\; \int_{-\infty}^{\infty} f(x)p(x)\,\mathrm{d}x \ ,$$

for example,

$$\bar{x} \;=\; E\{X\} \;=\; \int_{-\infty}^{\infty} xp(x)\,\mathrm{d}x \quad \text{(mean)} \ ,$$

$$\mathrm{Var}(X) \;=\; E\{(X-E\{X\})^2\} \;=\; \int_{-\infty}^{\infty} (x-\bar{x})^2 p(x)\,\mathrm{d}x \quad \text{(variance)} \ ,$$

which implies that the pdf is of key importance for the statistical characterisation of the underlying physical process. The procedure of estimating the probability density function is usually referred to as *density estimation.*

2.1.1 Histogram technique

The most popular instrument for density estimation is the histogram. The data range is divided into a set of successive and non-overlapping intervals (bins), and the frequencies of occurence in the bins are plotted against the discretised data range. The bin size should be chosen such that a sufficient number of observations falls into each bin. In principle, the bin size may vary across the data range.

The histogram technique is easy to implement computationally. However, the resulting pdf estimate depends on the bin size as well as the origin of the discretised data range, and is discontinuous at the bin boundaries.

2.1.2 The naive estimator

Recall the definition of the pdf as a probability density to write

$$p(x) \;=\; \lim_{h\to 0} P(x-h \le X \le x+h)$$

which allows to directly contruct a pdf estimator $\hat{p}(x)$ that does no longer depend on the origin of the chosen data range discretisation:

$$\hat{p}(x) \;=\; \frac{1}{2hJ}\,\{\text{no. of observations falling into}[x-h, x+h]\} \ .$$

where J denotes the total number of observations. The parameter h controls the effective resolution of the pdf estimate and corresponds to the bin size of the histogram technique.

This definition is equivalent to

$$\hat{p}(x) \;=\; \frac{1}{J}\sum_{j=0}^{J-1} \frac{1}{h} w\left(\frac{x-x_j}{h}\right)$$

where x_j is the j-th datum, and w is the rectangular window (weight) function defined as

$$w(s) \;=\; \begin{cases} 1/2 & , \quad \text{if } |s| \le 1 \;, \\ 0 & , \quad \text{otherwise} \;, \end{cases}$$

This means that in the construction of $\hat{p}(x)$ rectangular boxes of width $2h$ and height $(2JH)^{-1}$ are placed around each datum and then summed up.

2.1.3 The kernel estimator

The naive esimator still yields discontinuous results but the construction can be easily generalised to get continuous pdf estimates. We replace the weight function w by a kernel function $K(x)$ which satisfies

$$\int\limits_{-\infty}^{\infty} K(x)\,\mathrm{d}x \;=\; 1$$

to obtain

$$\hat{p}(x) \;=\; \frac{1}{Jh}\sum_{j=0}^{J-1} K\left(\frac{x-x_j}{h}\right) \;.$$

The kernel function K should be non-negative and is usually continuous. The kernel estimator $\hat{p}$ will then also be smooth because it inherits all smoothness properties from the kernel function.

The parameter h is called window width, bandwidth, or smoothing parameter. It controls the trade-off between the statistical significance of the pdf estimate and its effective resolution.

2.1.4 Variable window width estimators

The kernel technique gives pdf estimates which are smooth and do not depend on the origin but still depend on the smoothing parameter. Variable window width estimators are designed to at least partially remove this dependency. The amount of smoothing is adapted to the ‘local’ density of data. Examples are the nearest neighbor method and the variable kernel method. See Silverman [1986] for details.

2.2 Practical aspects: implementation and a worked example

In order to illustrate the use and the caveats of the kernel method for density estimation, we briefly describe the implementation and then discuss an application to geomagnetic data.

2.2.1 Implementation of the kernel method

Here we choose a straightforward implementation of the kernel method:

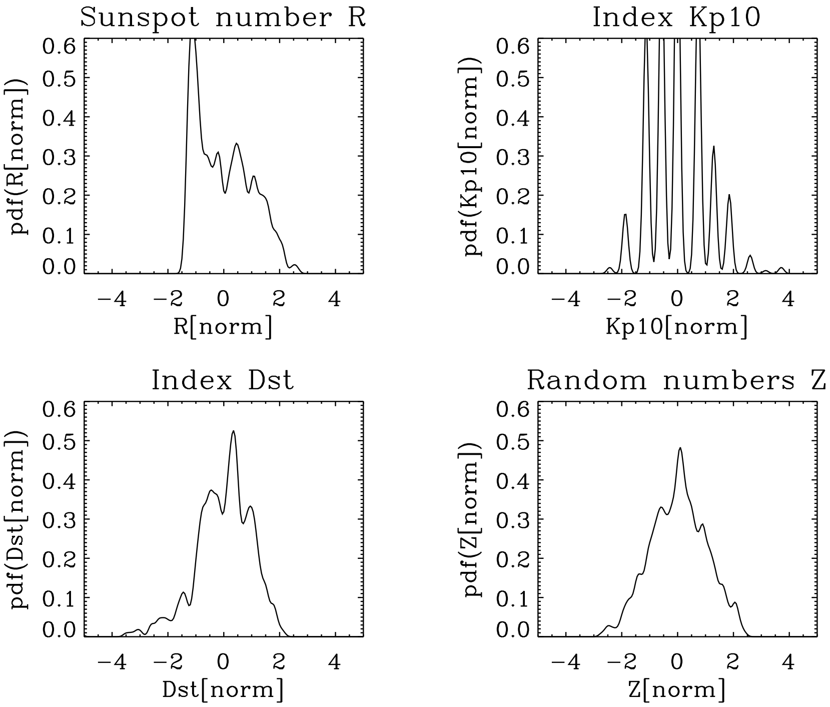

Figure 1: Density estimation using the kernel method. A gaussian kernel function was used with window width $h = 0.1$.

- The kernel function is accessed through a separate procedure `KERNEL`. Three different types can be chosen: a rectangular (which effectively yields the naive estimator), a triangular (gives continuous but non-differentiable estimate), and a gaussian window (yielding the smoothest estimates).
- The procedure `PDFEST` calls `KERNEL` to compute the pdf estimate at individual points u_i of a predefined and equidistant grid (independent of the data values).
- A continuous plot is produced from the resulting pairs $(u_i, \hat{p}(u_i))$.

2.2.2 Application: sunspot numbers and geomagnetic indices

The effect of the smoothing parameter h on the pdf estimate can be easily demonstrated using publicly available data from NASA's OMNIWeb page[2], namely, 27-day averages of sunspots numbers R, rescaled Kp indices Kp10, and ring current indices Dst from 1963 to 2001. Figures 1, 2, and 3 show the result for window widths $h = 0.1, 0.3, 0.5$. The corresponding pdf estimates for a normally distributed random number sequence have been added for comparison. While a too small window width seems to produce spurious spikes and other small-scale features which lack statistical significance (note the effect of the Kp quantisation on the pdf estimate), all essential features are smeared out if the window width chosen is too large.

---

[2] http://nssdc.gsfc.nasa.gov/omniweb/

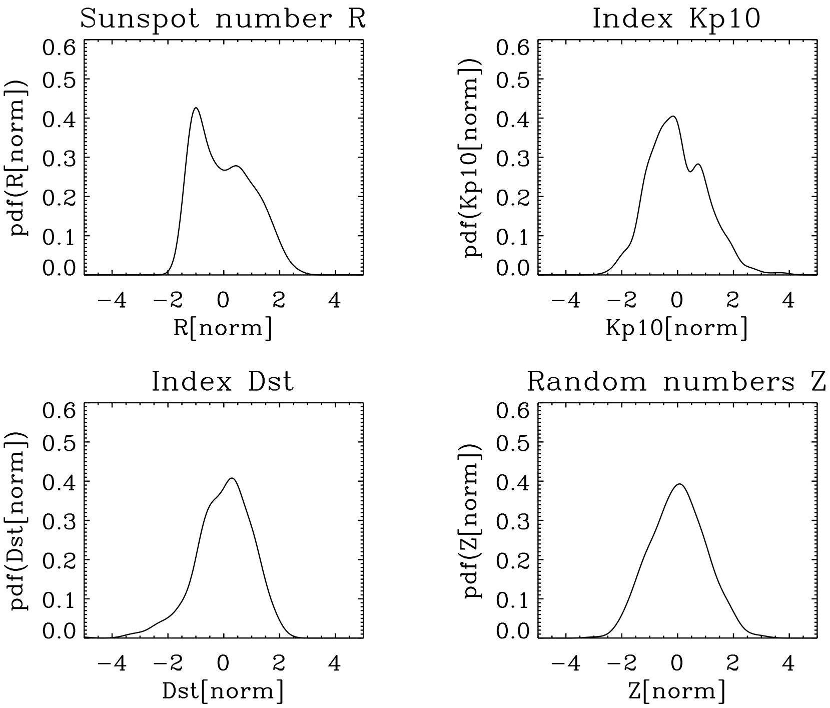

Figure 2: Density estimation using the kernel method. A gaussian kernel function was used with window width $h = 0.3$.

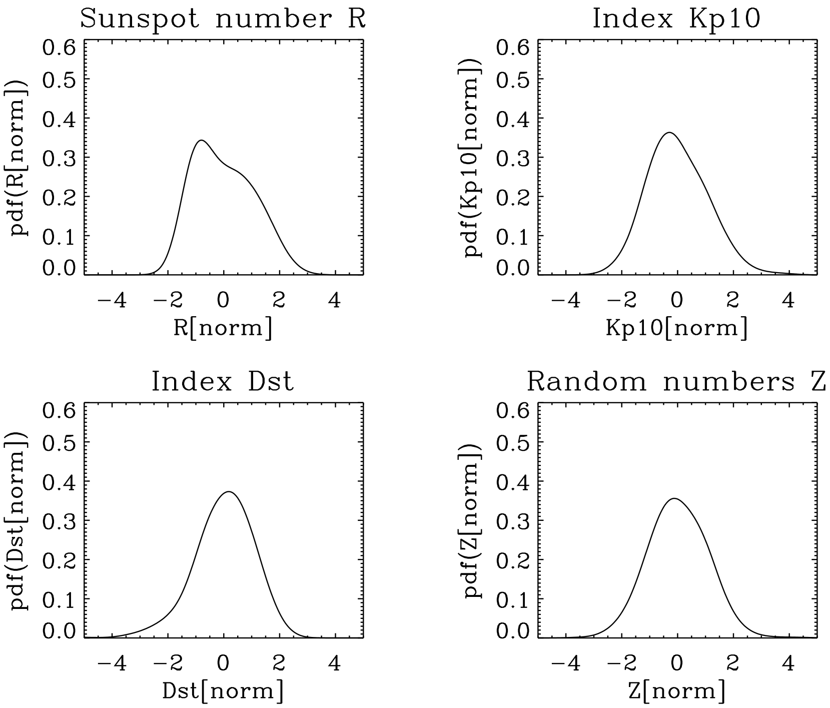

Figure 3: Density estimation using the kernel method. A gaussian kernel function was used with window width $h = 0.5$.

2.2.3 IDL procedures and data files: a worked example

The figures discussed above were produced using IDL. We briefly describe the procedures used in this context and how we arrived at the resulting pdf estimates.

- For convenience all data were put into one file `rkp10dst27.dat` that can be found in the subdirectory `ComputerSessions/BasicAnalysisTechniques_Vogt/ex2/` of the workshop web page directory[3].
- In the same subdirectory, you find several IDL procedures to process the data: the program `read-rkp10dst27.pro` reads the file into an IDL session, `kernel.pro` provides different kernels of the method, `pdfest.pro` carries out the actual pdf estimation, and `plot-pdfest.pro` plots the result.
- The driver routine for pdf estimation is `plot-pdfest.pro` which allows to modify the window width and type easily by changing the corresponding parameters at the beginning of the file. Start the program in IDL in the usual way:
  ```
  IDL> .r plot-pdfest
  ```
  and get the resulting graphics in PS format.

2.3 References

- Silverman, B.W., *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, 1986. Part of the book is also available in html on the internet[4].
- Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., *Numerical Recipes (in C/FORTRAN/...)–The Art of Scientific Computing*, Cambridge University Press, 1992. Available as free pdf from the *Numerical Recipes home page*[5]. Chapter 14 deals with the statistical description of data.
- NIST/SEMATECH *e-Handbook of Statistical Methods*[6].

---

[3] http://www.faculty.iu-bremen.de/jvogt/cospar/cbw6/

[4] http://nedwww.ipac.caltech.edu/level5/March02/Silverman/

[5] http://www.nr.com

[6] http://www.itl.nist.gov/div898/handbook/