NONPARAMETRIC DENSITY ESTIMATION

C. R. LONGBOTHAM, SHELL DEVELOPMENT CO.

Abstract

Nonparametric density estimation is a recently applied statistical tool useful in exploratory data analysis and in graphical presentation of data. The underlying density of the data is estimated without assumptions regarding the form of the true distribution except that it is continuous. We describe a macro that may be called to give a nonparametric estimate of a univariate density. An introduction to nonparametric density estimation is included.

Introduction to Nonparametric Density Estimation

The goal of nonparametric density estimation is to estimate the density function, f(x), of a continuous distribution function, F(x). A simple estimate of the density is a histogram for given interval widths and boundaries. A density estimate is useful for presentation of data as well as exploratory data analysis - to visually inspect properties of the density such as skewness, number and position of modes, unusual features, etc. A nonparametric density estimate may be thought of as a smoothed histogram. Several methods are available for estimation. An estimate of f(x) at any point $x \in (-\infty,\infty)$, given a random sample, X1, ..., XN, is a weighted average of a function of the distances $|x - Xi|$, i=1,...,N. The weights are determined by the kernel, K(·), and the bandwidth, h. The kernel specifies the shape of the distribution of the weights and the bandwidth specifies the sphere of influence. The estimate of f(x) for an arbitrary point x is

$$\hat{f}(x) = \frac{1}{Nh} \sum_{i=i}^{N} K\left(\frac{x-Xi}{h}\right).$$

Four kernels are given as options in the macro described in the next section. Two of these are the triangular kernel,

$$K(y) = \begin{cases} 1 - |y| , & \text{if } |y| < 1 \\ 0, & \text{otherwise} \end{cases}$$

and the quartic kernel,

$$K(y) = \begin{cases} \frac{15}{16} \cdot (1-y^2)^2, & \text{if } |y| < 1 \\ 0, & \text{otherwise} \end{cases}$$

The other two kernels are the density functions of the Normal distribution and the Cauchy distribution.

The bandwidth may be chosen automatically by the computer or the user may choose a bandwidth or list of bandwidths for calculation. The choice of bandwidth is much more critical for the shape of the density estimate than the choice of kernel. An optimal bandwidth depends mainly on N and the shape of the true density. Since the true density is not known, an approximation to the optimal bandwidth is made. A bandwidth that is too large will cause the density estimate to be overly smooth and too narrow a bandwidth will give a bumpy appearance to the density. An investigator may choose a density estimate smoother or bumpier than that chosen by the approximately optimal bandwidth because of expectations about density smoothness, modes, tendency for data to clump around certain values, etc.

Program Description

The program is written as a macro that may be called within any program after the observations have been placed in a data step. User-selected options, with defaults accompanied by an *, are

1) type of kernel: quartic*, triangular, normal, cauchy
2) bandwidth selection:
   a) single value selected by user
   b) list of values selected by user
   c) single value selected by computer
   d) list of three values chosen by computer*
3) span of density estimate (and plot): entire range of observations*, portion of range
4) cumulative probability calculations: calculate F(x) for x values where f(x) calculated, do not calculate any F(x) values*
5) x values for density: automatically chosen*, user specified.

The form of the macro with variable names and default values is

```
DENSITY(DATA=A, VAR=X, KERNEL=1, HFIRST=-99999,
HLAST=-99999, HINC=-999, XFIRST=-99999,
XLAST=-99999, XINC=-99999, CUMPROB=-1).
```

If a value for a variable is not specified in the call of macro, the default value is used. For example, following is a simple program to call the macro using all the defaults except the data set and variable name. The macro is in a SAS.SORC file called DENSPROG (a TSO implementation).

```
% INCLUDE SASSORC (DENSPROG);
% DENSITY (DATA = SASDATA.EDUC, VAR=AGE)
```

A table of the macro variables and meaningful values is given below

Variable	Values
DATA	Any valid data set name
VAR	Any numeric variable name
KERNEL	0=triangular, 1=quartic, 2=normal, 3=cauchy
HFIRST	-99999 or positive; smallest bandwidth in list of bandwidths, if default, computer chooses bandwidth by formula of Silverman's
HLAST	-99999 or positive; largest bandwidth in list of bandwidths; if default, computer chooses same bandwidth as for HFIRST

HINC -99999, negative, positive; if positive, HINC is step size in list of bandwidths; if -99999, only one bandwidth, HFIRST used; if other negative, three bandwidths in list, HFIRST, .7*HFIRST and 1.3*HFIRST.

XFIRST Any real value; first x value in domain of f(x); if default, XFIRST is slightly less than smallest x value.

XLAST Any real value; last x value in domain of F(x); if default, XLAST is larger than largest x value.

XINC -99999 or positive; distance between x values in list; if default, XINC = (XLAST-XFIRST)/60.

CUMPROB Negative or positive; if positive, compute and print cumulative probabilities to out file, else do not compute and print probabilities. (NOTE: If kernel = 0 or 1, set XFIRST = -99999, if kernel = 2, set XFIRST < (smallest data value - 3* bandwidth). Cumulative probabilities are not accurate if kernel = 3)

Example

Metallic tensile test samples (called dogbones) were taken from development units. The development units were all processed identically. The density estimate (see Figure 1) clearly is bimodal. Once we discovered this feature of the data we looked for reasons to explain the bimodality.

We found the two modes to correspond closely to two extreme locations where dogbones were taken on the units (Figures 2 and 3). The density of the dogbones taken between the two extremes was again bimodal (Figure 4).

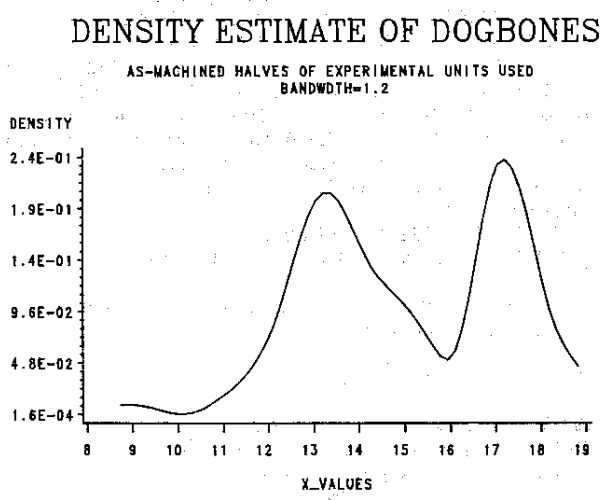

Figure 1.

ALL POSITIONS, 85 OBSERVATIONS

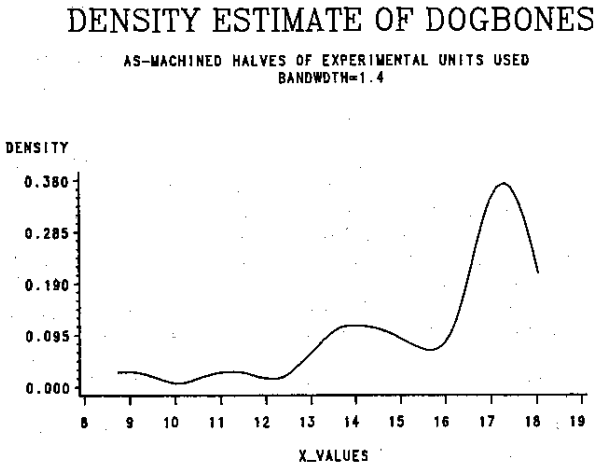

Figure 2.

POLE POSITION, 24 OBSERVATIONS

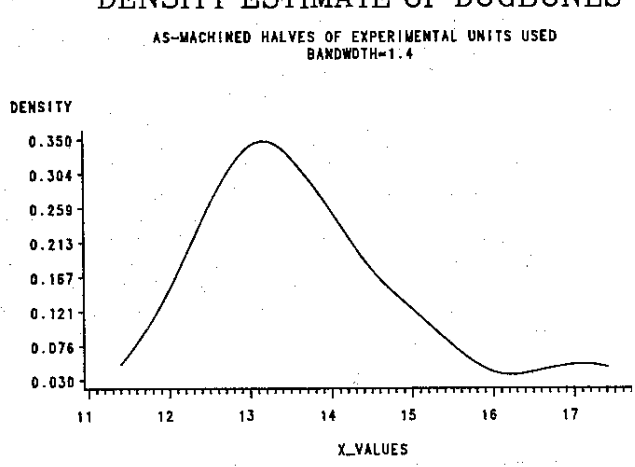

Figure 3.

WAIST POSITION, 25 OBSERVATIONS

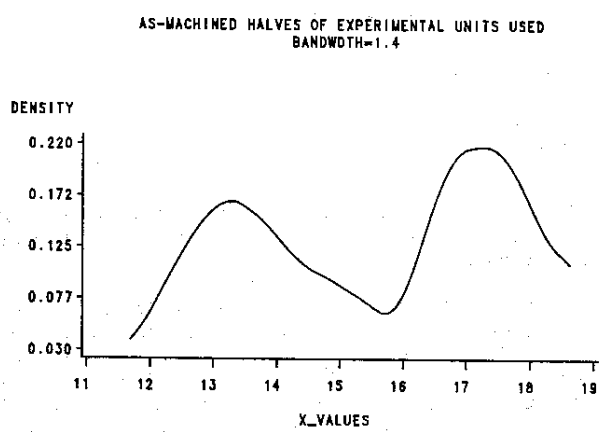

Figure 4.

MID POSITION, 36 OBSERVATIONS

Appendix - Macro Listing

```
*************************************************************************
**   A MACRO TO GIVE NONPARAMETRIC DENSITY ESTIMATES OF A DENSITY      **
**   GIVEN A RANDOM SAMPLE FROM A POPULATION. USER CHOOSES PARAMETERS **
**   TO BALANCE SMOOTHNESS AND BIAS (BANDWIDTH, KERNEL), AND TO        **
**   SPECIFY THE RANGE OF THE DATA OVER WHICH THE DENSITY IS TO BE FIT**
**        PROGRAM WRITTEN BY: C. ROGER LONGBOTHAM                      **
**        WHILE AT ROCKWELL INTERNATIONAL, ROCKY FLATS PLANT           **
*************************************************************************;
OPTIONS LS=76;
GOPTIONS DEV=IBM3279;
%MACRO DENSITY(DATA=A,VAR=X,KERNEL=1,HFIRST=-99999,HLAST=-99999,
    HINC=-999,XFIRST=-99999,XLAST=-99999,XINC=-99999,CUMPROB=-1);
DATA A;
   SET &DATA;
   X = &VAR;
   KEEP X;
PROC SORT DATA=A;
   BY X;
PROC IML;
  START DEFAULTS;
      **************        DEFINE PARAMETER VALUES       **************;
      ***  DEFAULTS FOR PARAMETERS ARE ACHIEVED BY SETTING VALUES EQUAL
      ***  TO -99999 FOR NEXT 6 PARAMETERS, KERNL=1, AND CUMP<0. OTHER-
      ***  WISE SET HF=SMALLEST BANDWIDTH >0,HL=LARGEST BANDWIDTH >0,
      ***  HI=INCREMENT FROM HF TO HL TO COMPUTE BANDWIDTH, XF=SMALLEST
      ***  X VALUE FOR REGION OF INTEREST, XL=LARGEST X VALUE FOR REGION
      ***  (DEFAULT REGION IS RANGE OF DATA), DX=DISTANCE BETWEEN X
      ***  VALUES, KERNL=0 FOR TRIANGULAR, 1 FOR QUARTIC, 2 FOR NORMAL,
      ***  AND 3 FOR CAUCHY. CUMP>0 FOR CALCULATION OF CUMULATIVE D.F.
      ***  (IF CUMP>0, XF NEEDS TO BE <= SMALLEST X VALUE IF KERNL=0 OR
      ***  1, <= (X VALUE)-3*BANDWIDTH IF KERNL=2, AND CUMPROBS NOT
      ***  ACCURATE IF KERNL=3);
      HF = &HFIRST;
      HL = &HLAST;
      HI = &HINC;
      XF = &XFIRST;
      XL = &XLAST;
      DX = &XINC;
      KERNL = &KERNEL;
      CUMP = &CUMPROB;
      FINISH;
   RUN DEFAULTS;
   USE A;
   READ ALL INTO XA;
   N = NROW(XA);
   CONST = 15/16;                         **********************************;
   SIGMASQR = .32653;                     *   SCALE CONSTANTS FOR KERNELS   *;
   GCONST = SQRT(2*3.14159*SIGMASQR);* DEFINED HERE. CONST FOR QUARTIC,*;
   CCONST = .5/3.14159;                   * GCONST,CCONST FOR NORMAL, CAUCHY*;
   START VARIANCE;                        **********************************;
      *** CALCULATE STD. DEV. AND QUARTILE ESTIMATE OF VARIABILITY***
      *** FOR USE IN CALCULATING INITIAL, DEFAULT BANDWIDTH USING ***
      *** SILVERMAN'S FORMULA.                                    ***;
      SUM = XA(|+,|);
      CSS = (XA#XA)(|+,|) - SUM#SUM/N;
      STDDEV = SQRT(CSS/(N-1));
      Q1 = FLOOR(((N+3)/4 || ((N+6)/4));
      Q1 = (XA(|Q1,|))(|+,|)/2;
      Q3 = CEIL(((3*N+1)/4) || ((3*N-2)/4));
      Q3 = (XA(|Q3,|))(|+,|)/2;
      QUARTSIG = (Q3 - Q1)/1.34;
      FINISH;
   RUN VARIANCE;
   START INITIAL;   *** MODULE TO TRANSLATE PARAMETER OPTIONS;
      IF XF=-99999 THEN XF=XA(|1,1|);
      IF XL=-99999 THEN XL=XA(|N,1|);
      IF XL <= XF THEN DO;
         PRINT 'EITHER LARGEST X VALUE CHOSEN IS TOO SMALL';
         PRINT 'OR ALL DATA VALUES ARE THE SAME';
         STOP;END;
      IF DX <= 0 THEN DO;
         INC = (XL-XF)/60;
         RINC = 10##(FLOOR(LOG10(INC))-1);
         DX = ROUND(INC,RINC);
         END;
      IF XF=XA(|1,1|) THEN XF=XF-DX;
      NX = INT((XL-XF)/DX) + 3;
      HDEFAULT = .9*MIN(STDDEV,QUARTSIG)*N##(-.2); ** SILVERMAN'S FORM.;
      IF HF = -99999 THEN HF = HDEFAULT;
      IF HF <= 0 THEN DO;
         PRINT 'HFIRST MUST BE POSITIVE IF DEFAULT NOT SPECIFIED';
         STOP; END;
      IF HL = -99999 THEN HL = HF;
      IF HI = -99999 THEN HI = 100000;
      MINH = (HL-HF)/10;
      IF HI < 0 THEN DO;
         HI = .3*HF;
         HL = HF + HI;
         HF = HF - HI;
         END;
      IF HL ¬= HF & HI < MINH THEN HI = MINH;
      FINISH;
   RUN INITIAL;
   *** NEXT MODULE TO COMPUTE CUMULATIVE PROBABILITIES, IF CHOSEN;
   START CUMPROBS;
      NK = NX + 1;
      FCN = FNX//0 0 0;
      FCH = FCN(|1,3|);
      FCN(|NK,2|) = 0;
      FCUM = J(NK+1,2,0);
      FCUM(|1,1|) = XF - .5*DX;
      FCUM(|1,2|) = (FCN(|1,1|)*DX)/2;
      DO K = 2 TO NK;
         FCUM(|K,1|) = FCUM(|K-1,1|) + DX;
         FCUM(|K,2|) = FCUM(|K-1,2|) + ((FCN(|K-1,1|)+FCN(|K,1|))*DX)/2;
         END;
      FCUM(|NK+1,1|) = FCUM(|NK,1|) + DX;
      FCUM(|NK+1,2|) = (FCUM(|NK,2|) + 1)/2;
      VARLABEL = {X_VALUE CUMPROB};
      VARH = {BANDWDTH};
      PRINT 'CUMULATIVE PROBABILITIES CORRESPONDING TO THE',
            'NONPARAMETRIC DENSITY ESTIMATE'
            FCH (|COLNAME=VARH|),,
            FCUM (|COLNAME=VARLABEL|);
      FINISH;
   FNX = J(NX,3,0);
   *** NEXT MODULE CALCULATES DENSITY ESTIMATE(S) AT SPECIFIED X VALUES;
   START DENS;
      VARS = {DENSITY X_VALUES BANDWDTH};
      CREATE PLOT FROM FNX (|COLNAMES=VARS|);
      DO H = HF TO HL BY HI;
         NUH = N*H;
         X = XF - DX;
         DO I = 1 TO NX;
            X = X + DX;
            Y = (J(1,N,X) - XA)/H;
            Z = 1 >< ABS(Y);
            IF KERNL = 1 THEN KY = CONST*((1-Z#Z)##2);
               ELSE IF KERNL=2 THEN KY=EXP(-.5*Y#Y/SIGMASQR)/GCONST;
               ELSE IF KERNL=3 THEN KY=CCONST/(.25+Y#Y);
               ELSE KY = 1-Z;
            FNX(|I,1|) = SUM(KY)/(NUH);
            FNX(|I,2|) = X;
            END;
         FNX(|,3|) = H;
         IF CUMP > 0 THEN RUN CUMPROBS;
         APPEND FROM FNX;
      END;
      FINISH;
   RUN DENS;
   CLOSE PLOT;
   QUIT;
***  NEXT PLOT THE DENSITIES, CHANGE GPLOT AND OPTIONS DEPENDING ON ***
***  DEVICE AVAILABLE.                                              ***;
PROC SORT DATA=PLOT;
   BY BANDWDTH;
PROC GPLOT DATA=PLOT;
   BY BANDWDTH;
   TITLE 'NONPARAMETRIC DENSITY ESTIMATE';
   SYMBOL1 I=JOIN;
   PLOT DENSITY*X_VALUES;
%MEND DENSITY;
```

References

Silverman, B. W. 1986. **Density Estimation for Statistics and Data Analysis.** Chapman and Hall, London.

Tapia, Richard A. and Thompson, James R. 1978. **Nonparametric Density Estimation.** John Hopkins University Press, Baltimore.

Comments or questions to the author are welcome.

C. Roger Longbotham
Shell Development Co.
P. O. Box 1380
Houston, TX 77251-1380
(713) 493-8298