Stochastic Hydrology and Hydraulics 11: 523-547 © Springer-Verlag 1997

Evaluation of kernel density estimation methods for daily precipitation resampling

Balaji Rajagopalan
Lamont-Doherty Earth Observatory of Columbia University, Palisades, NY, USA

Upmanu Lall and **David G. Tarboton**
Dept. of Civil & Environmental Engineering, Utah Water Res. Lab., Utah State University, Logan, UT 84322, USA

Abstract: Kernel density estimators are useful building blocks for empirical statistical modeling of precipitation and other hydroclimatic variables. Data driven estimates of the marginal probability density function of these variables (which may have discrete or continuous arguments) provide a useful basis for Monte Carlo resampling and are also useful for posing and testing hypotheses (e.g. bimodality) as to the frequency distributions of the variable. In this paper, some issues related to the selection and design of univariate kernel density estimators are reviewed. Some strategies for bandwidth and kernel selection are discussed in an applied context and recommendations for parameter selection are offered. This paper complements the nonparametric wet/dry spell resampling methodology presented in Lall et al. (1996).

1 Introduction

In a recent paper (Lall et al. 1995), a nonparametric approach to a stochastic model for daily precipitation was presented. The salient features of this model were the consideration of alternating wet and dry spells and of a daily rainfall structure within the wet spell. Kernel density estimates (k.d.e.'s) were espoused as effective methods for recovering univariate, multivariate or conditional, discrete and/or continuous probability densities that were needed directly from the historical record. In the process of developing the nonparametric wet/dry spell model in Lall et al. (1996) kernel density estimators of continuous and discrete variables were reviewed and tested with various data sets. Our aim here is to present some of this experience, specifically with the type of data available for modeling dally precipitation as a wet/dry spell model.

The issues relevant to the implementation of the kernel density estimators reviewed here are (a) the specification of the bandwidth of a kernel density estimator for the continuous case, (b) the role of boundary effects in kernel estimation, and (c) the selection of the estimator in the discrete case. The intent is to justify our recommended procedures by example, and to provide a comparison of some of the estimation schemes available in the literature.

Investigations for estimating the probability density function (p.d.f.) of continuous random variables (here, it is the precipitation amount for a day or for a wet spell) are first presented followed by comparisons of methods for the estimation of the probability mass function (p.m.f.) of discrete random variables (here, it is the length of a wet spell or dry spell in days).

2 Kernel density estimation of a continuous random variable

Kernel density estimation for univariate, continuous random variates was reviewed recently by Lall et al. (1993) in the flood frequency estimation context. The presentation here adds a few recent bandwidth estimation methods, and a discussion of the possible utility of boundary kernels with precipitation data. The interested reader is referred to Silverman (1986) for a pragmatic treatment of Kernel density estimation; to Devroye and Györfi (1985) for a rigorous treatment using L1 (absolute value) methods; and to Scott (1992) for a recent monograph with an excellent treatment of multivariate estimation. Chiu (1996) and Jones et al (1996) provide reviews of bandwidth selection methods. Hydrologic applications are reviewed in Lall (1995).

2.1 Basic ideas

Given observations $x_1, x_2, \cdots, x_n$, the kernel density estimator (k.d.e) at any point x is $\hat{f}_n(x)$ is defined as:

$$\hat{f}_n(x) = \sum_{i=1}^{n} \frac{1}{nh_i} K\left(\frac{x - x_i}{h_i}\right) \tag{1}$$

where K(.) is a kernel function centered on the observation x_i, that is usually taken to be a symmetric, positive, probability density function with finite variance; and h_i is a bandwidth or "scale" parameter of the kernel centered at x_i. A fixed kernel density estimator uses a constant bandwidth, h, irrespective of the location of x. The kernel K(.) is a symmetric function centered on the observation x_i, that is positive, integrates to unity, has first moment equal to zero and finite variance. An illustration of how a kernel density estimate is computed is provided in Figure 1. Examples of kernel functions that are often used are provided in Table 1. In this work, we have used the Epanechnikov and the Bisquare kernels.

Other examples (see Devroye and Györfi 1985; Silverman 1986; Scott 1992) of nonparametric density estimators include the k nearest neighbor density estimator, Fourier series estimators, adaptive shifted histograms, frequency polygons, penalized likelihood estimators, and orthogonal series estimators. All these methods can be shown to be equivalent to kernel density estimators with special kernels.

The goal of nonparametric density estimation is to obtain a good pointwise estimate of the underlying p.d.f. Consequently, the performance of the estimator is judged by the pointwise error. The choice of the estimator and the bandwidth is motivated through an analysis of mean squared error (MSE) in estimating the density at a point x, given as

$$\mathrm{MSE}(\hat{f}_n(x)) = E\{[f)(x) - \hat{f}_n(x)]^2\} \tag{2}$$

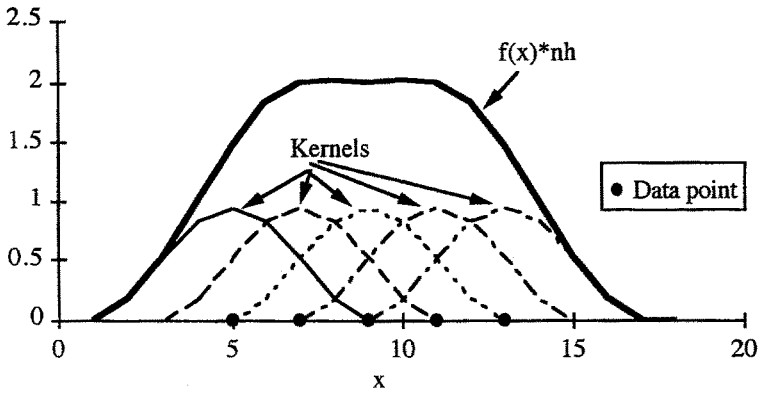

Figure 1. Example of kernel density estimation using 5 equally spaced values (5-13) with Bisquare kernel, h=4.

Table 1. Examples of Continuous Variable Kernel Functions

Kernels	Note $t = (x - x_i)/h$
Normal	$K(t) = (2\pi)^{-1/2} e^{-t^2/2}$
Epanechnikov	$K(t) = 0.74(1 - t^2)$ |t| 1
Bisquare	$K(t) = 0.9375(1 - t^2)^2$ |t| 1

Continuous (Left) Boundary Kernels, Univariate (Müller, 1991)

Note that q=x/h, 0≤q≤1 and x is the point at which the density is estimated, and h is the bandwidth.

for Epanechnikov $K(q,t) = 6(1+t)(q-t)\ \frac{1}{(1+q)^3}\ \left\{1 + 5\left(\frac{1-q}{1+q}\right)^2 + 10\ \frac{1-q}{(1+q)^2}\ t\right\}$

where E[.] denotes the expectation operator. Härdle (1991), p. 59 provides the asymptotic mean square error of the kernel density estimate (equation 1) for differentiable f(x), through a Taylor Series expansion of the MSE as:

$$\mathrm{MSE}(\hat{f}_n(x)) = \left\{0.5h^2 f''(x) \int t^2 K(t)dt\right\}^2 + (nh)^{-1} f(x) \int K^2(t)dt + O(h^2) \qquad (3)$$

The first term in equation (4) is the bias squared and the second is the variance of the estimate at x. Since it is a weighted moving average, the k.d.e. typically underestimates the density at the modes, and overestimates it at the antimodes, corresponding to the bias term that is proportional to $f''(x)$. The Mean Integrated Squared Error (MISE=MSE($\hat{f}_n(x)dx$) and related measures of performance can be developed from equation (3).

Epanechnikov (1969) showed that the MSE optimal kernel (among the class of kernels that are positive everywhere and have first moment and second moment finite), for density estimation is the quadratic kernel bearing his name given in Table 1. He also showed that the asymptotic relative MSE efficiency (MSE($\hat{f}_n(x)$) using kernel/MSE($\hat{f}_n(x)$) using optimal kernel) of any other admissible kernel function (even the rectangular kernel) was always close to one. The reason for this is that different kernels can be made equivalent in this sense through appropriate choices of the bandwidth (Scott, 1992). Consequently it is generally believed that the choice of a kernel function is not very important for density estimation as far as the asymptotic MSE is concerned. However, there are other factors that are important for choosing a kernel function. The differentiability of the kernel function is inherited by the resulting density estimate. The Epanechnikov kernel is not differentiable at the ends of its support. The Bisquare kernel (Table 1) is to be preferred in this regard. Where the random variable is bounded (e.g., precipitation is defined only over [0,]), a kernel with bounded support is to be preferred (e.g. Epanechnikov or Bisquare) over one with infinite support (e.g. Normal) to minimize boundary effects (which will be discussed in section 2.2).

Typically the bandwidth and the kernel are selected by minimizing the estimated average mean integrated square error (AMISE=E[MSE($\hat{f}_n(x)dx$)]). Methods for bandwidth selection are described in section 2.3 and are summarized in Table 2.

Since kernel density estimation is a local averaging process, estimates in the tail (especially for data from long tailed distributions) can be rough (have high variance of estimate) because there will be fewer and fewer data points to average for a fixed bandwidth. A natural way to deal with with such situations is to use a larger h in regions of low density (e.g., tails) and smaller h in regions of high density (e.g., near the modes). Variable bandwidths can reduce the problem of oversmoothing of the modes.

Estimation of a variable bandwidth h_i is more difficult than the estimation of the global bandwidth h. A practical approach is a procedure suggested by Silverman (1986) based on recommendations by Abramson (1982), who showed that choosing h_i proportional to $\hat{f}_n(x_i)^{-1/2}$ could improve the MSE rate of convergence of $\hat{f}_n(x)$ from $O(n^{-4/5})$ to $O(n^{-8/9})$. Here O(.) refers to "terms of the order of", and for comparison the optimal convergence rate for a parametric density estimate is usually $O(n^{-1})$. The strategy is to perturb an appropriate fixed or global bandwidth h into a sequence of bandwidths h_i at each observation x_i as:

Table 2. Choices of Bandwidth Selection for Kernel Estimators of Continuous Variables

Method	Equation	Criteria/Remarks
PR-M	$h_{opt} = 3.03n^{-0.2}$	Based on minimization of MISE, given Epanechnikov kernel and assuming underlying probability density function to be 0.5N(-2,1)+0.5N(2,1).
PR-N	$h_{opt} = 2.13\hat{\sigma}n^{-0.2}$	Based on minimization of MISE, given Epanechnikov kernel and assuming the underlying probability density function to be $N(0,\hat{\sigma}^2)$. $\hat{\sigma}$ is the sample standard deviation.
PR-E	$h_{opt} = 1.97\hat{\sigma}n^{-0.2}$	Based on minimization of MISE, given Epanechnikov kernel and assuming the underlying probability density function to be $Exp(\hat{\sigma})$. $\hat{\sigma}$ is the sample standard deviation.
LSCV	$LSCV(h)=\int \hat{f}^2 - 2n^{-1}\sum_{i=1}^{n} \hat{f}_{-i}(x_i)$	Choose h to minimize LSCV(h) function. $\hat{f}_{-i}$ represents the k.d.e constructed by dropping the i^{th} observation.
MLCV	$MLCV(h)=n^{-1}\sum_{i=1}^{n} \log(\hat{f}_i)$	Choose h to maximize MLCV(h) function.
SJ	refer to equations, 12-15, section 2.3	Based on recursive estimation of MISE.
SJL	Same as SJ, but applied to log transformed data.	

Note: (For details on these methods, refer section 2.3)
PR Parametric reference
LSCV Least squares cross validation
MLCV Maximum likelihood cross validation
SJ Sheather and Jones (1991) procedure
SJL Sheather and Jones (1991) procedure applied to log transformed data

$$h_i = h(\hat{f}_n(x_i))/g)^{-1/2} \quad (4)$$

where g is the geometric mean of $\hat{f}_n(x_i)$. One can iteratively re-estimate $\hat{f}_n(x_i)$ and hence h_i using the latest kernel density estimate. Two to three such iterations were found to be sufficient to achieve pointwise convergence to a fractional tolerance of 0.001 in the resulting density estimate.

2.2 Boundary effects and their treatment

An annoying aspect of kernel estimators of probability densities (both continuous and discrete) is the increased bias within one bandwidth of the boundary (e.g., 0) of the sample space. The bias is a consequence of the increasingly asymmetric distribution of the random variable as one approaches the boundary. Modifications to kernel density estimate are necessitated within a bandwidth of the boundary (e.g. 0 for data from exponential distribution) of the sample space. Two problems are faced for estimation in the boundary region.

The first is that a kernel can extend past the boundary if the bandwidth is larger than the observation at which a kernel function is centered. This leads to a leakage of probability mass, and the resulting $\hat{f}_n(x)$ will not integrate to 1 over the sampling domain. Clearly this problem is aggravated if a kernel with infinite support is used (such as the Gaussian kernel, see Table 1). The boundary problem is illustrated in Figure 2. Consider the continuous univariate random variable $x \in [0,]$, and a fixed bandwidth (h=0.1). For the point of estimation in the Figure 2 (i.e. x=0.01), which is within one bandwidth of the boundary, the interior Epanechnikov kernel is truncated of at the boundary (x=0.0) resulting in the leakage of probability mass. Boundary kernels developed by Müller (1992) alleviate this problem.

The second problem is increased bias that results from the asymmetric distribution of observations around the point of estimate. Let us say that the smallest sample value is x_1, and that x_1 is greater than h. Now if a kernel estimate of $\hat{f}_n(x)$ is needed for x<h, i.e., in the boundary region, all the sample values are to the right of x, leading to an increased bias in the estimate $\hat{f}_n(x)$. Attempts to overcome this bias typically lead to an increased variance due to the relatively few points caught in a bandwidth of the kernel.

A number of methods for dealing with the boundary problems mentioned above have been proposed. We investigated four methods for boundary modification of the kernel estimator.

The first method is "cut and normalize". One computes the area of each kernel that lies within the sample space, and normalizes the truncated kernel to have unit area, by dividing the kernel function by this area. Bias reduction issues are not addressed.

The second method, reflection, augments the data set by reflection of the real data across the boundary. The assumption is that f'(x)=0. There is no basis for this assumption and it is unlikely that it holds for the precipitation data sets.

The third method which is more general, considers the development of special boundary kernels (see Müller, 1988, 1992, and Table 1), that are asymmetric, unbiased, and minimum variance but are not non-negative. These kernels are modified versions of the kernels used in the interior of the sample space, and are derived from

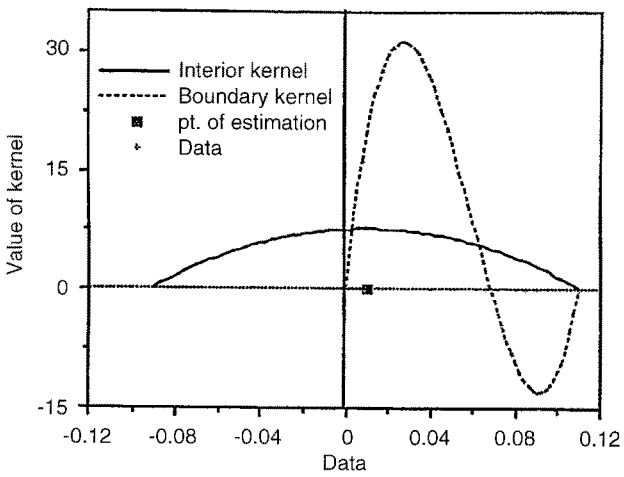

Figure 2. Conceptual figure of the boundary problem in kernel density estimation.

variational conditions (see Müller, 1992 for details). We have investigated such kernels in the univariate case with reasonably good results. Bias of the density estimate is reduced in the boundary region, typically with some increase in the variance of estimate. For the type of data we were dealing with (precipitation or spell length), the density is high near the origin (i.e. 0.01 and 1 respectively), and the possible negative values of the boundary kernel function near the origin do not translate into negative density estimates. For the discrete case, Dong and Simonoff (1994) have developed boundary kernels for the Epanechnikov and Bisquare kernels, (See Table 1 for boundary kernels for Epanechnikov kernel).

A fourth method relevant for data concentrated near the boundary (e.g exponential, log normal) is a logarithmic transform of the data prior to density estimation. Such a transformation can also provide an automatic degree of adaptability of the bandwidth (in real space), thus alleviating the need to choose variable bandwidths with heavily skewed data, and also alleviates problems that the kernel density estimator has with p.d.f. estimates near the boundary (e.g., the origin) of the sample space. The resulting k.d.e. can be written as:

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_x x} K\left(\frac{\log(x) - \log(x_i)}{h_x}\right) \tag{5}$$

where h_x is the bandwidth of the log transformed data. The above estimator worked well for data concentrated near the origin (e.g. exponential type) and hence is recommended.

2.3 Bandwidth selection schemes

In this section we review some choices of bandwidth selection for kernel density estimation for continuous variables. Comparisons of these alternatives with synthetic data are presented next. Rather than reproducing a variety of statistical results, we

shall focus on getting the basic ideas across through a brief review of the univariate, continuous random variable case.

Four methods for selecting the optimal global bandwidth were considered.

(1) Parametric reference (PR) procedure.

The optimal bandwidth h_{opt} and kernel are selected by first minimizing the Mean Integrated Squared Error (MISE), equation (3) integrated with respect to h. The result is the optimal bandwidth h_{opt} and then solving for the optimal kernel (see Silverman, 1986, p.38-42).

The MISE of the fixed, univariate, continuous, k.d.e. and the corresponding optimal global bandwidth h_{opt} are given by Silverman (1986), sec. 3.3 as:

$$\text{MISE}(f_n(p)) \approx (nh)^{-1}R(K) + 0.25h^4\sigma_K^2 R(f'') \tag{6}$$

$$h_{opt} = \{R(K)/(\sigma_K^2 R(f''))\}^{1/5} n^{-1/5} \tag{7}$$

where $R(g)=\int g^2(x)dx$ and $\sigma_g^2 = \int x^2 g(x)dx$. The terms R(K)and σ_K^2 depend only on the known kernel K(.). Consequently, the unknown term in equations 8 and 9 is R(f''), which depends on the unknown density f(x). Now one could fit the "best" parametric model for precipitation, e.g., the exponential, and then "knowing" f(x) compute R(f'') and thereby evaluate h_{opt}. Silverman (1986), p. 47, provides h_{opt} using the normal distribution as a reference. We investigated such schemes, and found that bandwidths selected in this manner can be quite sensitive to the choice of the reference distribution. For example, for a Gaussian kernel, the h_{opt} for a Normal parent p.d.f. is 1.33 times the h_{opt} for an Exponential parent. The need to refer to a parametric model detracts from the utility of this method, but the method is less sensitive to boundary effects while selecting h_{opt}.

From equation (7) observe that knowing the optimal bandwidth h_N for the Normal kernel, the optimal bandwidth h_K for a kernel different from the Normal kernel can be readily evaluated as:

$$h_K = \{(R(K)\sigma_N^2)/(\sigma_K^2 R(N))\}^{1/5} h_N \tag{8}$$

where "N" identifies the Normal kernel, and "K" the kernel of interest. Different kernels can thus be made equivalent.

(2) Least Squares Cross Validation (LSCV, see Silverman (1986), section 3.4). The optimal bandwidth is solved by the minimization of

$$\text{LSCV}(h) = \int f^2 - 2n^{-1}\sum_{i=1}^{n} f_{-1}(x_i) \tag{9}$$

where f_{-i} represents a k.d.e. constructed by dropping the i^{th} observation.

LSCV is prone to undersmoothing where the data exhibits fine structure, and also suffers from a high degree of sampling variability, leading to rather poor MISE convergence rates ($O(n^{-1/10})$ (see, Hall and Marron (1987)). The computational burden and poor convergence rate of this method are discouraging. However, its broad applicability to a wide class of situations renders it popular.

(3) Maximum Likelihood Cross Validation (MLCV, see Silverman (1986), section 3.4).

The optimal bandwidth is solved by the maximization of a pseudo-likelihood criteria given as:

$$\text{MLCV}(h) = n^{-1} \sum_{i=1}^{n} \log(f_{-1}(x_i)) \tag{10}$$

MLCV leads to degenerate solutions if the data is long tailed, and also suffers from the same low convergence rate that characterizes LSCV. The degeneracy can be corrected (Schuster 1985) by excluding a fraction of the right tail data from the MLCV score (not from the density estimate). The subjectivity of the choice of such a cutoff point and the computational burden of the scheme detract from its usage.

(4) Direct minimization of estimated MSE/MISE.

"Plug in" or recursive estimators are methods that use data driven kernel estimates of f(x) and R(f'') (or equivalent measures in the discrete case). Such methods were originally proposed by Woodroofe (1970), and pursued by Scott et al (1977), Scott and Factor (1981) and Sheather (1983, 1986). Improvements by Park and Marron (1990), and Sheather and Jones (1991) (hereafter, SJ) among others have lent stability to these methods and have led to a MISE convergence rate of h_{opt} of the order of $n^{-5/14}$, as well as a reduction in the size of the constants associated with this rate.

A summary of the SJ procedure for the continuous, univariate k.d.e. follows. They developed a kernel estimate $s(\alpha)$ for R(f'') as:

$$S(\alpha) = \{n(n-1)\}^{-1}\alpha^{-5} \sum_{i=1}^{n}\sum_{j=1}^{n} K^{iv}((x_j - x_i)/\alpha) \tag{11}$$

$$\alpha(h) = 1.357\{S(a)/T(b)\}^{1/7}h^{5/7} \tag{12}$$

$$T(b) = -\{n(n-1)\}^{-1}b^{-7} \sum_{i=1}^{n}\sum_{j=1}^{n} K^{iv}((x_i - x_j)/b) \tag{13}$$

$a=0.92\lambda n^{-1/7}$ and $b=0.912\lambda n^{-1/9}$

where α is a bandwidth (not equal to h), and $K^{iv}(.)$ is a special kernel for estimating fourth derivative of the density, $K^{iv}(.)$ is a special kernel for estimating the sixth derivative of the density, and λ is the sample interquartile range $(x_{0.75} - x_{0.25})$, T(b) is an estimate of R(f''') and a, b are bandwidths that are evaluated with reference to a Normal distribution for the derivative kernels considered.

Relatively crude estimates (with reference to a known distribution) of the bandwidths used in estimating R(f'') and R(f''') suffice given that the dependence of the MISE expression equation (6) on these expressions is successively weaker (note the exponents). The optimal bandwidth h_{opt} is now evaluated by computing a and b from the data, evaluating S(a) and T(b), and substituting the equation (13) into equation (12), and equation (12) into equation (11). This leads to a nonlinear expression in terms of h, which is solved using the Newton Raphson method. Sheather and Jones

specify the normal kernel for K(.) and evaluate the derivative kernels as the appropriate derivatives of this kernel. While this is the most attractive data based approach that we tested, it does not consider the boundary behavior of the kernel estimator. In the case where the data is positive and heavily concentrated near the origin, the SJ procedure tends to grossly undersmooth relative to the theoretical optimal bandwidth.

2.4 Comparative results of various bandwidth selection schemes

The most critical aspect of developing the k.d.e. is the specification of the bandwidth. A second factor is the need for specialized treatment near x=0 (i.e., the boundary problem). We compare the different methods outlined in sections 2.2 and 2.3 with two synthetic data sets.

First we sample (C1) from a Gaussian mixture(0.5N(−2,1)+0.5N(2,1)), to demonstrate estimability with location mixtures. The second sample (C2), was generated from an Exponential distribution with mean 0.15, to demonstrate the boundary effect. In each case a sample of size 250 was used. Sample statistics and values of the key parameters in each case are summarized in Table 3. The corresponding p.d.f.'s estimated by selected methods are shown in Figures 3a to 3e.

We consider six estimators for density estimation for the above mentioned data sets. These are: (1) (PR-N) parametric reference assuming the underlying probability density function to be N(0,$\hat{\sigma}^2$), (2) (PR-M), parametric reference assuming the underlying probability density function to be a Gaussian mixture 0.5N(−2,1)+0.5N(2,1), (3) (PR-E) parametric reference assuming the underlying probability density function to be Exp($\hat{\alpha}$), (4) (LSCV) Least squares cross validation, (5) (MLCV) Maximum likelihood cross validation, (5) (S J) Sheather and Jones (1991) procedure, and (6) (SJL) Sheather and Jones (1991) procedure applied to log transformed data. Table 2 summarizes the bandwidth estimation procedures. In the first three methods the term parametric reference means the bandwidth is chosen to be optimal with reference to an assumed underlying parametric distribution. The first five methods, which consider untransformed real space data also use Silverman's method (discussed in section 2.1) to specify a local rather than a fixed global bandwidth. Boundary kernels as defined by Müller (1991) were used to adjust the density estimates near the lower boundary (x≥0), but were not used during bandwidth estimation. The SJL procedure, eliminated the boundary problem and provides some local bandwidth adaption, so no local bandwidth adjustment and no boundary kernels were used.

For data set C1 we used methods PR-N, PR-M,LSCV, MLCV, SJ, while, for data set C2 we used PR-E, LSCV, MLCV, SJ and SJL.

The following observations are apparent from the figures:

1. The parametric reference (PR) procedures work very well as expected when the assumed p.d.f. matches the underlying p.d.f. However, under mis-specification, performance suffers. In case of C 1, the bandwidth from the true reference (PR-M) is 1.0, while from using the normal distribution (i.e. mis-specification) as the reference (PR-N) the bandwidth is 1.76. This results in gross oversmoothing of the two modes present in C1 (see Figure 3a). The parametric reference bandwidth is the best possible estimate of h provided f(x) is known. Of course,

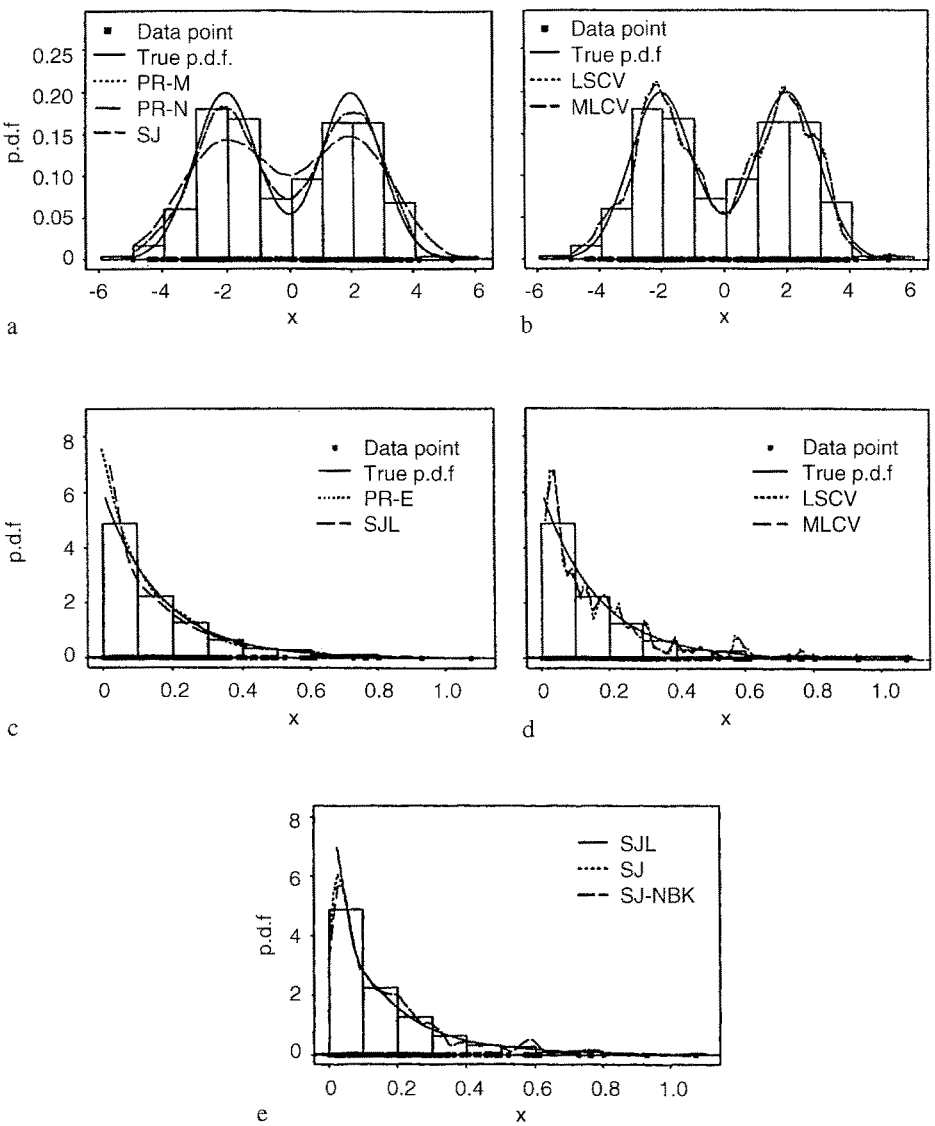

Figure 3. (a) Plot of p.d.f's estimated from PR-M (h=1), PR-N (h=1.76), SJ (h=1.03), the true underlying p.d.f, observed data and histogram of observed data, for the data set C1. (b) Plot of p.d.f's estimated from LSCV (h=0.48), MLCV (h=0.53), the true underlying p.d.f, observed data and histogram of observed data, for the data set C2. (c) Plot of p.d.f's estimated from PR-E (h=0.11), SJ (h=0.04), SIL (h=0.77), the true underlying p.d.f, observed data and histogram of observed data, for the data set C2. (d) Plot of p.d.f's estimated from LSCV (h=0.015), MLCV (h=0.02), the true underlying p.d.f, observed data and histogram of observed data, for the data set C2. (e) Plot of p.d.f's estimated from SJ, SJL, and SJ-NBK (Bandwidth chosen from SJ procedure but boundary kernels are not used). Along with observed data and histogram of observed data, for the data set C2.

Table 3. Statistics (Sample size =250 for each) and Methods for Figure 2

Data	Method (corresponding to Appendix 2)	Global Bandwidth
C1 (Gaussian mixture)	PR-M	1.00
($\bar{x}$=0.00, s=2.26)	PR-N	1.76
	LSCV	0.48
	MLCV	0.53
	SJ	1.03
C2 (Exponential)	PR-E	0.11
($\bar{x}$=0.16, s=0.18)	LSCV	0.015
	MLCV	0.02
	SJ	0.04
	SJL	0.77 (in log space)

Note: $\bar{x}$ is sample mean and s is sample standard deviation.

The SJL estimator is, (Equation 1)

$\hat{f}_n(p) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{hp} K\left(\frac{\ln(p)-(\ln(p_i)}{h}\right)$ with Epanechnikov kernel.

The Parametric reference, LSCV, MLCV and SJ all use, $\hat{f}_n(p) = \sum_{i=1}^{n} \frac{1}{nh_i} K\left(\frac{\ln(p)-\ln(p_i)}{h}\right)$ with Epanechnikov kernel and Müller boundary kernels. Local bandwidths h_i are given by, $h_i = h(f(p_i)/g)^{-1/2}$, where h is global bandwidth, $f(p_i)$ is the kernel density estimate at p_i using the global bandwidth h and g is the geometric mean of $f(p_i)$. These estimators only differ in the procedure used to obtain global bandwidth.

one reason we pursue nonparametric estimates of the p.d.f. is lack of knowledge of the underlying model. In this context, PR estimates with the correct f(x) are useful as a benchmark to compare the performance of fully data driven methods.

2. LSCV and MLCV are prone to undersmoothing especially when the data exhibits fine structure (e.g multiple modes) and is long tailed (see, Hall and Marron (1987)). Also the cross-validation functions (which are minimized for the bandwidth estimation) have spurious local optima (corresponding to clustering of data at different scales) at small bandwidths, (see Hall and Marron (1987)). Thus, we expect small bandwidths from LSCV and MLCV which leads to an undersmoothed density estimate. This can be seen from Figures 3b and 3d, where the estimates from LSCV and MLCV are very rough, suggesting that the variance is high.

3. SJ has been shown to have a better mean integrated square error (MISE) convergence rate than cross validation methods (see Sheather and Jones (1991)) and hence should lead to a better estimate. This is borne out in Figures 3a and 3e, and Table 3. Note that the SJ optimal bandwidth for C1 is close to the optimal bandwidth based on the Gaussian mixture as reference (PR-M). However for C2 the SJ optimal bandwidth is much smaller than the optimal bandwidth for the exponential distribution. This is due to the fact that the boundary effect is not considered while estimating the SJ bandwidth, which is a problem in case C2 but not in C1. In both cases the SJ bandwidth is superior to those chosen by MLCV and LSCV.

   Note that in all these cases, the optimal h is determined without using the boundary kernels, and is perhaps smaller than it would be (to reduce the effect of leakage across the boundary) if boundary kernels were used during bandwidth estimation. This emphasizes the need for proper treatment of the boundary of the domain during all phases of k.d.e. We expect to pursue modifications of the SJ estimator to account for boundaries during bandwidth selection.

4. For C2, in Figures 3c and 3d, we use the Müller boundary kernels (except when using SJL) to reduce the bias at the boundary. Despite this a considerable bias can be observed near the origin in these figures, for each of these estimators. This is a consequence of the high curvature of the target density near the origin, and the "leakage" from the kernels across the boundary at x=0. Figure 3e for the case C2 includes a p.d.f estimated without using boundary kernels (SJ-NBK) along with those from SJ and SJL. The inclusion of boundary kernels in SJ offers only a marginal improvement over SJ in this case, since it still suffers from a bias due to the high curvature of f(x) in this area. SJL, on the other hand does not suffer as much from this problem and hence, performs better.

5. For data sets with a heavy concentration of data near the origin, a log transformation is an attractive choice. We see from Figure 3e that the SJL procedure provides a very competitive k.d.e in this situation. Note that SJL provides local bandwidth adaptation in real space. For the wet day precipitation data, that is usually modeled using an Exponential, or a Gamma distribution, this may be a natural transformation to consider.

Our recommendation of SJL is motivated largely by a desire to deal with the boundary effects and local bandwidth adaptation in a natural way given the nature of the

precipitation data. Where boundary effects are not of concern (e.g., C1) a direct application of SJ would be preferred. Once a modification of SJ to account for boundary effects during bandwidth estimation is successful, SJL need not be the method of choice even in this situation.

3 Kernel density estimation for discrete random variables

Wet spell and dry spell lengths are treated as an integer number of days in our rainfall model (Lall et al., 1995), consequently estimators for discrete data are reviewed here. The presentation of discrete kernel estimators is new to the hydrologic literature, and includes a new estimation method we developed (Rajagoplan and Lall (1995)). For a discussion of the methods for discrete data refer to Hand (1982), Bishop et al. (1975) and Coomans and Broeckaert (1986).

3.1 Basic ideas

The basic concepts of kernel estimation of p.d.f.'s in the continuous case introduced earlier hold for the discrete case as well. In the discrete case one can first estimate the sample relative frequencies. These relative frequencies or multinomial cell proportions can then be "smoothed" using a kernel estimator. The problem of nonparametric smoothing of the multinomial cell proportions has not been studied as extensively as nonparametric density estimation, its counterpart in the continuous case. Here we have a sample $y_1, y_2, \cdots, y_n$ for n multinomial trials with possible outcomes 1,2,..L..,L_{max} with probabilities of occurrence $f_1, f_2, \cdots, f_{L_{max}}$ that are unknown. Estimates $\hat{f}_n(L)$ for any cell L may be obtained as sample relative frequencies ($\tilde{p}_L = n_L/n$), or by smoothing the $\tilde{p}_L$. Hall and Titterington (1989) note that smoothing can be beneficial when there are many cells with small or zero frequencies, i.e. the data are sparse. This is the case with the wet and dry spell length data.

A kernel estimator $\hat{f}_n(L)$ is given as:

$$\hat{f}_n(L) = \sum_{i=1}^{L_{max}} K_d(L, i, h))\tilde{p}_i \tag{14}$$

where h is the bandwidth, L_{max} is the maximum observed spell length and Kd(.) is a discrete kernel (or weight function).

A nonparametric estimator of the discrete probabilities of the wet or dry spell lengths (w or d) would be the maximum likelihood estimator that yields directly the relative frequencies (e.g., (number of w_i)/n_w, for the i^{th} wet spell length w_i in a sample of size n_w). The kernel method is superior to this approach, because (a) it allows extrapolation of probabilities to spell lengths that were unobserved in the sample, and (b) it has higher MSE efficiency (Hall and Titterington, 1987). Three major estimators identified in literature and a fourth one developed by Rajagopalan and Lall (1995) for smoothing probabilities of discrete data, are described. Their performance with synthetic data sets is compared in the following sections.

3.2 Choice of discrete kernel estimators

The estimators considered are (1) The Geometric kernel estimator developed by Wang and Van Ryzin (1981), hereafter WV; (2) Maximum Penalized Likelihood Estimator (MPLE) developed by Simonoff (1983) and the estimator by Hall and Titterington

(1987), hereafter HT; and (4) the Discrete Kernel (hereafter DK) estimator developed by Rajagopalan and Lall (1995). These are summarized in Table 4.

(1) Wang and Van Ryzin (1981) estimator (WV)

The kernel estimator of the probability mass function (p.m.f) of a discrete variable L, (here the length of wet or dry spell with n sample values) given by Wang and Van Ryzin (1981) uses equation (14) with the geometric kernel given as:

$$\begin{aligned} K_d(L,i,h) &= 0.5(1-h)h^{|L-i|} \quad \text{if } |L-i| \geq 1 \qquad h \in [0,1] \\ &= (1-h) \qquad\qquad \text{if } L = i \end{aligned} \tag{15}$$

The bandwidth h can be global or local.

Wang and Van Ryzin (1981) derived optimal global and local bandwidths to minimize the MSE (Mean Square Error = $E[(f(L)\text{-}\hat{f}_n(L))^2]$). They estimate the local bandwidths h(i) by minimizing the approximate MSE of $f_n(i)$, while truncating the geometric kernel at i±2. The resulting expressions are in terms of the unknown true probabilities f(i). They show that substitution of the relative frequencies of i, estimated from the sample as $\tilde{p}_i(\tilde{p}_i = n_i/n)$ in the expressions leads to a strongly consistent procedure. An optimal global bandwidth is obtained by minimizing the average MSE (i.e., $1/n_i$MSE(i)) over the data. Expressions for the optimal global and local bandwidths are given in Table 4.

Note that for small values of h, the estimator is close to the naive maximum likelihood estimator (MLE) (i.e. $\tilde{p}_i$), and for $\tilde{p}_i$ small, h is larger, leading to a higher smoothing, or larger "smearing" of the relative frequencies. An improved extrapolation in the tail of the density can result through the use of the local bandwidths.

(2) Maximum Penalized Likelihood Estimator (MPLE)

The MPLE was first introduced by Good and Gaskins (1971) for continuous variables, and was later extended to the density estimation for discrete variables by Simonoff (1983). Simonoff (1983) proposes a solution for the "category" probabilities $\hat{f}_i$ that maximizes a penalty function given by,

$$\text{LFN} = \text{Log likelihood} - \text{roughness penalty} \tag{16}$$

The idea is to balance the goodness-of-fit of the estimate (i.e., likelihood) with its smoothness (i.e., roughness penalty). The smoothest estimate is obtained if all cell probabilities are equal over the range of Cells considered. With this in mind, the penalized likelihood function is defined as:

$$\text{LFN} = \sum_{i=1}^{L_{max}} n_i \log(\hat{f}_i) - \beta \sum_{i=1}^{L_{max}} \{\log(\hat{f}_i/\hat{f}_{i+1}\}^2 \tag{17}$$

$$\text{where } \sum_{i=1}^{L_{max}} \hat{f}_i = 1 \,, \tag{18}$$

Table 4. Examples of Discrete Kernel Estimators

Wang and VanRvzin (1981) (WV) Geometric Kernel estimator

$$\text{Geometric kernel} \quad K(x) = \begin{cases} 0.5(1-h)h^{|x-x_i|} & \text{if } |x - x_i| \; 1 \qquad h \in [0,1] \\ (1-h) & \text{if } x = x_i \end{cases}$$

$$\text{Global bandwidth} \quad h = \beta_1 \{3/2 + B_1 - B_2 + (n-1)\beta_{10}\}^{-1}$$

$$\text{Local bandwidth} \quad h(i) = d_i \left\{p_i + \tfrac{1}{4} E_i + F_i - G_i(n-1)e_i\right\}^{-1}$$

where,

$$\beta_1 = 1 - \sum_{i=1}^{n} \tilde{p}_i^2 + \tfrac{1}{2} B_1, \; B_1 = \sum_{i=1}^{L_{max}} \tilde{p}_i(\tilde{p}_{i-1} + \tilde{p}_{i+1}), \; B_2 = \sum_{i=1}^{L_{max}} \tilde{p}_i(\tilde{p}_{i-2} + \tilde{p}_{i-2}),$$

$$\beta_{10} = \sum_{i=1}^{L_{max}} \tilde{p}_i^2 - B_1 + \tfrac{1}{4} B_0$$

$$G_i = \tilde{p}_i(\tilde{p}_{i-2} + \tilde{p}_{i+2}), \; F_i = \tilde{p}_i(\tilde{p}_{i-1} + \tilde{p}_{i+1}), \; E_i = (\tilde{p}_{i-1} + \tilde{p}_{i+1}),$$

$$d_i = \tilde{p}_i(1 - \tilde{p}_i) + \tfrac{1}{2} F_i,$$

$$e_i = \left(\tilde{p}_i - \tfrac{1}{2} E_i\right)^2, \; B_0 = \sum_{i=1}^{L_{max}} (\tilde{p}_{i-1} + \tilde{p}_{i+1})^2$$

where, $\tilde{p}_i(\tilde{p}_i = n_i/n)$ are the sample relative frequencies

Maximum Penalized Likelihood Estimator (MPLE) of Simonoff (1983)

$$LFN = \sum_{i=1}^{L_{max}} n_i \log(\hat{f}_i) - \beta \sum_{i=1}^{L_{max}} \left\{\log\left(\hat{f}_i/\hat{f}_{i+1}\right)\right\}^2$$

where $\sum_{i=1}^{L_{max}} \hat{f}_i = 1$, $\beta \geq 0$,is a smootng parameter, and L_{max} is the largest cell (e.g. longest spell length) considered

The smoothing parameter β controls the relative weight assigned to smoothness and consequently has the same role as the bandwidth used in kernel estimators. The LFN function is minimized to solve for each $\hat{f}_i$s (the required cell probability estimates)

Hall and Titterington (1987) HT estimator

$$W(t) = K(t) / \sum_{j=-\infty}^{\infty} K(j/h) \; K(t) \text{ is a continuous r.v. kernel, j is integer}$$

$$1/h \in [0,1]$$

Discrete (Left) Boundary Kernels, Univariate (Dong and Simonoff, 1994)

Note that q=(x-1)/h, $0 \leq q \leq 1$ and x is the point at which the density is estimated

for Epanechnikov $K(q,t) = \frac{-6}{(1+q)^3} t^2 + \frac{3(q^2+1)}{(1+q)^3}$

Table 4 (continued)

DK estimator

Note t=(L-j)/h, and L is point at which density is estimated

Interior region (i.e. L≥h+1)

Quadratic kernel $K(t) = at^2 + b$ for $|t| \leq 1$

$$a = \frac{-3h}{(1-4h^2)} \quad \text{and} \quad b = \frac{3h}{(1-4h^2)}$$

Left Boundary (i.e. 1<L<h+1)

for Quadratic kernel $K(t) = at^2 + b$ for $|t| \leq 1$

$$a = \frac{-D}{2h(h+L)} \times \frac{1}{\left(\frac{E}{4h^3} - \frac{CD}{12h^3(h+L)}\right)} \quad \text{and} \quad b = 1 \left[1 - \frac{aC}{6h^2}\right] \frac{1}{h+L)}$$

where, C=h(h-1)(2h-1)+(L-2)(i-1)(2L-3); D=-h(h-1)+ (L-2)(L-1); E=-(h(h-1))^2+((L-2)(L-1))^2

Left Boundary (i.e. L=1)

for Quadratic kernel $K(t) = at^2 + b$for$|t| \leq 1$

$$a = \frac{-D}{2h^2} \times \frac{1}{\left(\frac{E}{4h^3} - \frac{CD}{12h^4}\right)} \quad \text{and} \quad b = \left[1 - \frac{aC}{6h^2}\right] \frac{1}{h}$$

where, C=h(h-1)(2h-1); D=-h(h-1); E=-(h(h-1))^2

$\beta \geq 0$, is a smoothing parameter, and L_{max} is the largest cell considered (or the longest spell length considered).

The smoothing parameter β controls the relative weight assigned to smoothness and consequently has the same role as the bandwidth used in kernel estimation. Here a data dependent β is used through the following procedure which minimizes asymptotic mean square error.

1. An initial β is chosen as $0.009N(L_{max})^{0.6}(\log(L_{max}))^{0.4})$, where N is the sample size.

2. Given this β, the penalized likelihood (Equation 15) is maximized with respect to $\hat{f}_i, i = 1, \cdots, L_{max}$ using the method of Lagrange multipliers.

3. An optimal β is now estimated by minimizing an asymptotic MSE, defined as an asymptotic approximation to $\sum_{i=1}^{L_{max}}(\hat{f}_i - \pi_i)^2$, where π_i is the unknown probability of cell i. Simonoff (1983) develops this asymptotic MSE expression in terms of the sample relative frequencies $\tilde{p}_i(\tilde{p}_i = n_i/n)$, β and the unknown probability π_i. For π_i he uses the estimates $\hat{f}_i$ from step 2.

4. Steps 2 and 3 are repeated till convergence is achieved.

Simonoff (1983) argues that although a formal proof for the convergence of this procedure is not available, extensive computations have indicated that the scheme does converge. The need to specify L_{max} (in excess of the longest observed spell) detracts from the use of this method. We would prefer a natural extension of the tail of the p.m.f. by the method used, rather than a prior specification of its extent.

(3) Hall and Titterington (1987) estimator (HT)

The HT estimator developed by Hall and Titterington (1987) uses a discrete kernel function formed from a continuous kernel as:

$$K_d(L, j, h) = \frac{K((L-j)/h)}{s(h)} \tag{19}$$

where h>1 and $s(h)=\sum_{i=L-h}^{j=L+h} K(j/h)$. K(.) is any suitable continuous univariate kernel function, with compact support, positive, integrating to one and symmetric. The bandwidth h is selected as a minimizer of a Least Squares Cross Validation (LSCV) function suggested in Hall and Titterington (1987), over a suitable range for h given as

$$LSCV(h) = \sum_{j=1}^{L_{max}} (\hat{f}_n(j))^2 - 2 \sum_{j=1}^{L_{max}} \hat{f}_{n,-j}(j)\tilde{p}_j \tag{20}$$

where, $\hat{f}_{n,-j}(j)$ is the estimate of the p.m.f of spell length j, by dropping all the spells of length j from the data. This method has been shown by Hall and Titterington (1987) to automatically adapt the estimator to an extreme range of sparseness types.

Note that this estimator has the same convolution structure as the kernel density estimator in the continuous case. The HT estimator, uses a standard continuous variate kernel function rescaled by the sum of the weights applied to an integer set of points. This estimator is defined over the set of integers. However wet and dry

spell lengths are counting numbers (integers greater than 1). To avoid the problem of the estimator assigning probability to integers less than 0 (the boundary problem), Dong and Simonoff (1994) developed boundary kernels for Epanechnikov and Bisquare kernels which are given in Table 4. By HT we refer to the HT estimator with the boundary modification of Dong and Simonoff (1994).

For finite samples, some disquieting aspects of the HT estimator become apparent. The non-integer bandwidth leads to an effective kernel that also varies with h in a manner quite different from that prescribed by equation (19). The effective integer support of $K_d(L,j,h)$ in equation (19) is $[(L-h^*),(L+h^*)]$, where h^* is the closest integer greater than or equal to h. HT kernels are defined as quadratics or other polynomials over [L-h,L+h]. Since this is not the effective integer support of the kernel the effective kernel over the space of integers is not the quadratic defined.

Alternatively, it is possible to develop a kernel that recognizes the data to be in integer space, has an integer bandwidth and satisfies all the required conditions in the integer space. This also obviates the need for normalization of the kernel weights as done in HT. We explored this line of thought and, sought a direct, discrete analog of the continuous kernel density estimator, which lead to the development of the discrete kernel (DK) estimator (Rajagopalan and Lall, 1995).

(4) Discrete Kernel Estimator (DK)

Our estimator $\hat{f}_n(L)$ uses equation (14) with discrete Quadratic Kernel (QK) is given as:

$$K_d(L,i,h) = at_i^2 + b \tag{21}$$

where $t_i = \frac{L-i}{h}$. Epanechnikov (1969) showed that the MSE optimal kernel of second order, is the quadratic kernel (QK), also known as the Epanechnikov kernel. Here we need to specify the constants a and b for the interior (i>h+1) and the boundary region ($1 \le i \le h+1$). The constants a and b are solved to satisfy: (a) the kernel function goes to zero for |i-j| $\ge$h, i.e. $K(t_j) = 0$ for $|t_j| \ge 1$, (b) sum of the weights is unity, i.e. $\sum_{j=i-h}^{j=1+h} K\left(\frac{i-j}{h}\right)=1$ and (c) the first moment of the kernel function is zero, i.e. $\sum_{j=i-h}^{j=1+h} K\left(\frac{i-j}{h}\right) t_j=0$. One could choose higher order Beta kernels and derive results similar to these that follow for DQ.

The resulting kernels for the interior and the boundary are given in Table 4. Derivations of these kernels are presented in Rajagopalan and Lall (1995).

Note that the kernel and hence, the estimator $\hat{f}_n(L)$ is expressed strictly in terms of the bandwidth h. An optimal choice of h then completes the definition of the estimator The bandwidth is selected by minimizing the Least Squared Cross Validation function given as,

$$LSCV(h) = \sum_{j=1}^{L_{max}} (\hat{f}_n(j))^2 - 2 \sum_{j=1}^{L_{max}} \hat{f}_{n,-j}(j)\tilde{p}_j \tag{22}$$

where, $\hat{f}_{n,-j}(j)$ is same as defined in earlier Hall and Titterington (1987) also show that cross- validation automatically adapts the estimator to an extreme range of sparseness types. If the multinomial is only slightly sparse, cross-validation, cross-validation will

produce an estimator which is virtually the same as the cell-proportion estimator. As sparseness increases, cross- validation will automatically supply more and more smoothing, to a degree which is asymptotically optimal.

3.3 Comparative results of various discrete kernel estimators

The four methods (WV, MPLE, HT and DK) are compared with two synthetic data sets generated from long tailed distributions (e.g. Geometric distribution). First we use a sample (D 1) from a geometric distribution with π=0.2. The second sample (D2), was generated from a mixture of two geometric distributions defined as (0.3G(π=0.9)+0.7G(π=0.2)). In each case a sample of size 250 was used. We also fitted a geometric distribution (GP) to D1 and D2 using the method of moments. Sample statistics and values of the key parameters in each case are summarized in Table 5. The corresponding probabilities estimated by each method for D1 and D2 are shown in Figure 4.

1. The WV procedure does not smooth the sample proportions ($\tilde{p}_i$) properly. In most cases, there is very little smoothing. In cases where there is some smoothing (e.g., Figure 4a, in the range x=4 to 6), the resulting estimate is rather unsatisfactory, and is inconsistent with the underlying population. We feel that part of this behavior is due to the rapid "drop off" of weight associated with the Geometric kernel, and part due to the method used for selecting the bandwidth h.

2. On the other hand, since the roughness penalty tries to make the p.m.f. uniform, MPLE emphasizes smoothness. Consequently, when the true p.d.f has a high second derivative (e.g., near the origin), MPLE has difficulty distinguishing between "true" curvature and observed variation. The resulting estimate often has a strong downward bias near the origin (Figure 4a). The MPLE is also sensitive to the value specified for L_{max}, the longest spell length considered. As L_{max} is increased, the downward bias at the origin is increased and the entire p.m.f. is "flattened".

3. The GP fit is very good (estimated π=0.1956) for D1 where the true distribution was geometric. As expected, a large bias is incurred near the origin for D2 (see Figures 4b and 4d), where the estimated π was 0.2554.

4. Figures 4b and 4d indicate that HT and DK perform comparably and are the best among the estimators considered. As both these estimators are quite similar in construction this is expected. The estimated p.m.f is smooth, and it also exhibits the least pointwise bias. The HT and DK estimators automatically adapts to a large range of density variation, providing optimal smoothness in finite samples. Unlike parametric fits, the HT and DK estimates are robust to certain kinds of outliers, as shown in Figure 4e. Outliers were added at 45, 50, 75 and 100.

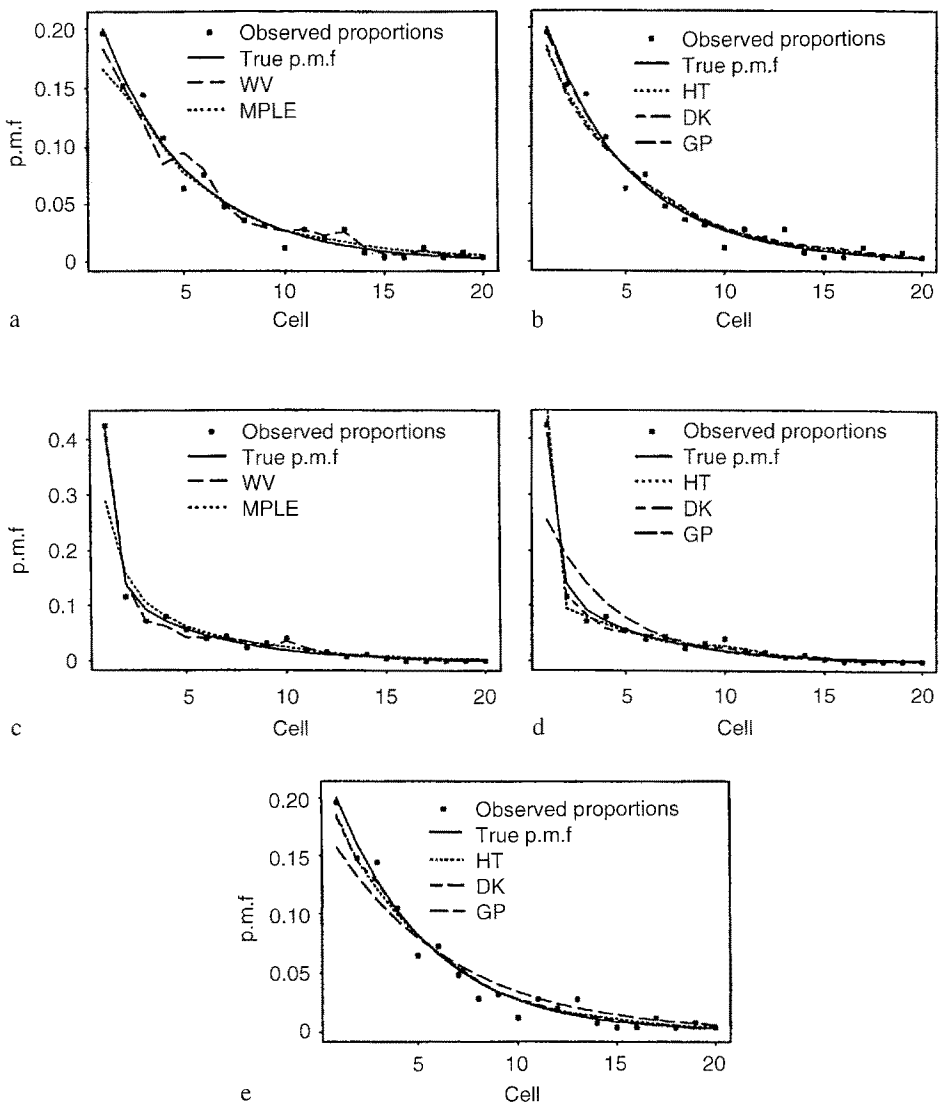

Figure 4. (a) Plot of p.m.f's estimated from WV (h = 0.43), MPLE (β=30.25), the true underlying p.m.f and observed proportions, for the data set D1. (b) Plot of p.m.f's estimated from HT (h=5), DK (h=6), GP (p=0.1956), the true underlying p.m.f and observed proportions, for the data set D1. (c) Plot of p.m.f's estimated from WV (h=0.08), MPLE (β=28.25), the true underlying p.m.f and observed proportions, for the data set D2. (d) Plot of p.m.f's estimated from HT (h=3), DK (h=2), GP (p=0.2554), the true underlying p.m.f and observed proportions, for the data set D2. (e) Plot showing the effect of outliers on fitted Geometric distribution (GP), HT and DK estimate. Outliers at 45,50,75,100 in the data set D1.

Table 5. Statistics (Sample Size = 250 for Each) and Methods for Figure 4

Figure	Data	Estimator	Kernel used	Method of Bandwidth Selection
4a	D1($\bar{x}$=5.11, s=4.19)	WV	Geometric kernel	MSE
		MPLE	- - -	- - -
4b	D1	HT	Epanechnikov kernel	LSCV
		DK	Quadratic kernel	LSCV
4c	D2($\bar{x}$=3.92, s=4.02)	WV	Geometric kernel	MSE
		MPLE	- - -	- - -
4d	D2	HT	Epanechnikov kernel	LSCV
		DK	Quadratic kernel	LSCV

Note: $\bar{x}$ is sample mean and s is sample standard deviation.

Quadratic kernel is the discrete equivalent of the Epanechnikov kernel.

These could be generated if the data were contaminated by a few large values (e.g. from a Geometric distribution with π=0.01). The fitted Geometric distribution, i.e. (GP) is very much affected by the outliers and deviates from the true distribution, especially near the mode (i.e. 1.). The HT and DK estimators still follow the data closely.

It is apparent from the figures that the HT and DK estimators perform the best. Rajagopalan and Lall (1995) found in their Monte Carlo comparisons of HT and DK that they gave comparable results with better approximation of the tail and the modes by DK. DK was also computationally faster, and had a lower variance of optimal bandwidth selection that HT. Consequently it is recommended.

4 Summary and conclusions

Issues in estimating parameters for continuous and discrete kernel density estimators were discussed and recommended procedures were developed through examples.

In summary, we recommend using the SJL procedure for estimating the p.d.f. of wet day precipitation amount. This entails the use of a Epanechnikov (or Quadratic kernel) with log transformed precipitation data with bandwidth chosen in log space using the Sheather Jones (1991) recursive procedure. The resulting density estimate is then transformed to real space. Generally this may be the method of choice for data sets which exhibit a high density near the origin. For discrete data such as spell lengths, we recommend the DK procedure with discrete quadratic kernels in the interior and boundary regions and bandwidth chosen by least squared cross validation.

We found that where the parametric procedure was appropriate, the nonparametric procedure worked nearly as well. Where the parametric model was inappropriate, the nonparametric kernel density estimators were superior. Given that the nonparametric procedures are robust and reproduce different parametric alternatives without prior assumptions, they offer a very general procedure for uniform application across a variety of sites and processes.

Problems with kernel density estimates are high relative bias and variance in the tail of the density if local adaption of the bandwidth is not used. Ability to extrapolate is limited to one bandwidth of the maximum observed value. Where a local bandwidth is used, the local bandwidth at the extreme point of observation is usually quite large and this problem is ameliorated.

The nonparametric modeling framework provides a promising alternative to parametric approach. The assumption free, data adaptiveness and robust nature of the nonparametric estimators makes the model attractive in a broad class of situations.

Acknowledgements

Partial support of this work by the U.S. Forest Service under contract notes, INT-915550-RJVA and INT-92660-RJVA, Amend #1 is acknowledged. We are grateful for discussions with D.S. Bowles, the principal investigator of the project. We thank S.J. Sheather and J.Simonoff for providing codes for implementing the SJ procedure and HT estimator with boundary modification respectively. Finally, we thank H.G. Muller, J. Dong, M.C. Jones and M. Wand for stimulating discussions and provision of relevant manuscripts. The work reported here was also supported in part by the

USGS through their funding of the second author's 1992-93 sabbatical leave, when he worked with BSA,WRD,USGS, National center, Reston, VA.

References

Abramson, I.S. 1982: On bandwidth variation in kernel estimates-a square root law, The Ann. of Statist. 10(4), 1217-1223

Bishop, Y.M.; Fienberg, S.E.; Holland, P.W. 1975: Discrete multivariate analysis: Theory and Practice, MIT Press, Cambridge, Mass

Chiu, S.-T. 1996: A comparative review of bandwidth selection for kernel density estimation, Statistica Sinica 6(1), 129-145

Coomans. D.; Broeckaert, I. 1986: Potential pattern recognition in chemical and medical decision making, John Wiley and sons, New York

Devroye, L.; Györfi, L. 1985: Nonparametric density estimation: The L1 view, John Wiley and Sons, New York

Dong Jianping; Simonoff, J. 1994: The construction and properties of boundary kernels for sparse multinomials, J. Computational and Graphical Statist. 3, 1-10

Epanechnikov, V.A. 1969: Nonparametric estimation of a multidimensional probability density, Theory of Probability and Applications 14, 153-158

Good, I.J.; Gaskins, R.A. 1971: Nonparametric roughness penalties for probability densities, Biometrika 58, 255-277

Hall, P.; Marron, J.S. 1987: Extent to which least-squares cross-validation minimizes integrated square error in nonparametric density estimation, Probability Theory Related Fields 74, 567-581

Hall, P.; Titterington, D.M. 1987: On smoothing sparse multinomial data, Australian J. Statist. 29(1), 19-37

Hand D.J. 1982: Kernel discriminant analysis, Pattern recognition and image processing research studies series, vol. 2, series Ed. J. Kittler, Research Studies press: Chichester

Härdle, W. 1991: Smoothing techniques with implementation in S, Springer Verlag, New York

Jones, M.C.; Marron, J.S.; Sheather, S.J. 1996: Progress in Data-Based Bandwidth Selection for Kernel Density Estimation, Computational Statist. 11 (3), 337-381

Lall, U.; Moon, Y.; Bosworth, K. 1993: Kernel flood frequency estimators: bandwidth selection and kernel choice, Water Resour. Res. 29(4), 1003-1015

Lall, U. 1995: Nonparametric function estimation: Recent hydrologic applications, US National report, 1991-1994, International Union of Geodesy and Geophysics, Reviews of Geophysics Supp., 1093-1102

Lall, U.; Rajagopalan, B.; Tarboton, D.G. 1996: A nonparametric wet/dry spell model for resampling daily precipitation, Water Resour. Res. 32(9), 2803-2823

Müller, H.G. 1988: Nonparametric regression analysis of longitudinal data, Springer Verlag, New York

Müller, H.G. 1992: Smooth optimum kernel estimators near endpoints, Biometrika. 78(3), 521-530

Park, B.U.; Marron, J.S. 1991: Comparison of data-driven bandwidth selectors, Journal of the American Statistical Association 85, 66-72

Rajagopalan, B.; Lall, U. 1995: A kernel estimator for discrete distributions, J. Nonparametric Statist. 4, 409-426

Schuster, E. 1985: Incorporating support constraints into nonparametric estimators of densities, Communication in Statistics A14, 1123-1136

Scott, D.W. 1992: Multivariate density estimation: Theory, Practice and Visualization, Wiley series in probability and mathematical statistics, John Wiley and Sons, New York

Scott, D.W.; Tapia, R.A.; Thompson, J.R. 1977: Kernel density estimation revisited, Nonlinear Analysis 1, 339-372

Scott, D.W.; Factor, L.E. 1981: Monte carlo study of three data-based nonparametric density estimators, J. Amer. Statist. Assoc. 76, 9-15

Sheather, S.J. 1983: A data-based algorithm for choosing the window width when estimating the density at a point, Comput. Statist. Data Anal. 1, 229-238

Sheather, S.J. 1986: An improved data-based algorithm for choosing the window width when estimating the density at a point, Comput. Statist. Data Anal. 4, 61-65

Sheather, S.J.; Jones, M.C. 1991: A reliable data-based bandwidth selection method for kernel density estimation, J. Royal Statist. Soc. B53, 683-690

Silverman, B.W. 1986: Density estimation for statistics and data analysis, Chapman and Hall, New York

Simonoff, J. 1983: A penalty function approach to smoothing large sparse contingency tables, Annals of Statist. 11, 208-218

Wang, M.C.; Van Ryzin, J. 1981: A Class of smooth estimators for discrete distributions, Biometrika 68(1), 301-309

Woodroofe, M. 1970: On choosing a delta-sequence, Annal. Math. Statist. 41, 1665-1671