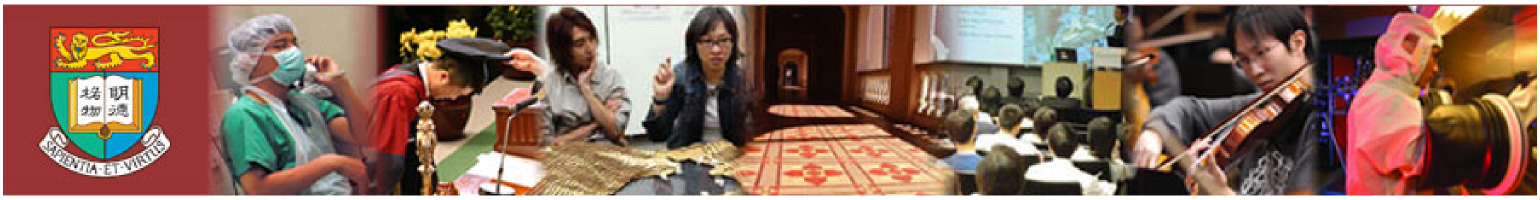

RESEARCH REPORT

Serial No. 15

June 1992

MULTIVARIATE DISCRETE DENSITY ESTIMATION USING KERNEL DENSITIES

by

John Bacon-Shone and John Aitchison

THE UNIVERSITY OF HONG KONG
DEPARTMENT OF STATISTICS

Pokfulam Road, Hong Kong. Tel: (852) 859 2466 Fax: (852) 858 9041
E-Mail: STATIST@HKUCC.HKU.HK or STATIST@HKUCC.BITNET

MULTIVARIATE DISCRETE DENSITY ESTIMATION USING KERNEL DENSITIES

John Bacon-Shone
Department of Statistics
and
Social Sciences Research Centre
University of Hong Kong
Pokfulam Road, Hong Kong

John Aitchison
Department of Mathematics
University of Virginia
Charlottesville, VA
U.S.A.

SUMMARY

Despite intensive recent research into density estimation little attention seems to have been paid to its possible use in describing patterns of variability of univariate or multivariate counts. This paper discusses the relative merits of a number of possible kernels, and illustrates their application to univariate and bivariate count data.

1. Introduction

Although there are now many parametric classes for univariate distributions of counts and some, such as the McKendrick (1926) bivariate Poisson and the Aitchison and Ho (1988) multivariate Poisson-lognormal classes, for multivariate counts, there remain count data sets whose patterns of variability defy satisfactory description by such classes. An excellent example of such a defiant data set is the Hohn and Hellerman (1963) set of bivariate counts of 137 different Potomac River species, one by glass slide and the other by styroform collection. In their analysis of these data Taillie et al (1979) obtain excellent univariate fits using logarithmic series distributions but conclude that their fit to the bivariate counts by a bivariate logarithmic series distribution is very poor.

In view of such parametric difficulties and with the upsurge of research interest in kernel density estimation methodology over the last three decades it seems surprising that no detailed assessment of kernels for count distributions on the set X of non-negative integers, or on its higher-dimensional counterparts X^d, seems to have been undertaken. This is perhaps surprising since there is an embarrassing number of simple kernels from which to choose.

2. Univariate count kernels

To define a kernel on the sample space $Y = \{0, 1, 2, \cdots\}$ of non-negative integers associated with an observed count x we require a probability (density) function $K(y|x, \lambda)$ on Y, centred in some way on x and with the smoothing parameter λ at our disposal. For a data set D of N counts $x_1, \cdots, x_N$ we then use as a kernel density function

$$p(y|D, \lambda) = N^{-1} \sum_{i=1}^{N} K(y|x_i, \lambda).$$

For other sample spaces such as the real line the crucial first step of selecting a suitable kernel K centred on an observation x is easily achieved by choice of a standard mean x providing a suitable location and the scale parameter, the standard deviation λ, acting as smoothing parameter. Such standard selection on the count sample space Y is less apparent since many of the standard distributions, such as the Poisson distribution, have only a single parameter while for two-parameter distributions, such as the negative binomial, it is not obvious how to harness the parameters to provide suitable location and scale. Although as part of our investigation here we shall adapt two standard univariate count distributions to kernel duty we shall see that there are much simpler, though non-standard, means of arriving at flexible classes of kernels.

For a count sample space Y a kernel $(K|x, \lambda)$ corresponding to an observed count x will be determined when we decide on the weights or relative probabilities we wish to place on x and on other possible counts y. Suppose that we use a weighting which is symmetric about x in the sense that the weights for $y = x - j$ and $y = x + j$ are the same and of the form $w_j(\lambda)$, depending only on j and the smoothing parameter λ. The sum of such weights on Y is

$$W(x, \lambda) = W_N(x, \lambda) + W_p(\lambda) + 1$$

where $W_N(x, \lambda) = \sum_{j=1}^{x} w_j(\lambda)$ and $W_p(\lambda) = \sum_{j=1}^{\infty} w_j(\lambda)$. We can then obtain a count kernel by setting

$$K(y|x, \lambda) = w_j(\lambda)/W(x, \lambda) \qquad (|y - x| = j).$$

To obtain a sensible kernel we obviously require $w_j(\lambda)$ to decrease as j increases so that the greatest weight is placed on the observation x with weights decreasing as we move from x. There are many ways in which this can be done and we have selected four, shown in Table 1, corresponding to the geometric, logarithmic and exponential series, and modified logarithmic with the smoothing parameter λ restricted to the interval $0 < \lambda < 1$ to ensure the decreasing nature of the weights. These can clearly be regarded as adaptations by symmetry and truncation of the geometric series, logarithmic series and Poisson distributions to the needs of kernel weighting. For given λ the geometric, logarithmic and exponential kernels are clearly in increasing order in their concentration around the observation x.

While these kernels ensure that the mode of the kernel is at the data point, asymmetry is introduced by the truncation and the mean will always be biased upwards and even the median of the kernel may be biased. One way to avoid this problem is to use a double kernel. This involves placing half of the kernel on the data point and integers above, and the other half on the data point and integers below. This ensures that the median is at the data point.

Another possible extension is to use different smoothing parameters for positive and negative values and this can be implemented for both the single and double kernel.

3. Multivariate count kernels

The obvious extension to multivariate counts is to use a product of the univariate kernels considered above. Any attempt to use a correlated kernel increases computational complexity quite dramatically as the scaling constants no longer have a simple form. We can however extend the class by allowing the smoothing parameters to vary across dimensions.

4. Examples

We start with a relatively tractable data set of counts of accidents sustained by 122 shunters in two consecutive periods from Arbons and Kerrich (1951). In the original paper, the data was fit quite successfully with a bivariate negative-binomial distribution. There is appreciable over-dispension relative to the Poisson distribution and there is sample correlation of 0.26. Table 2 shows the log likelihood for the different kernels when they are fit by cross-validation. Clearly, the results are quite insensitive to the choice of kernel. Allowing the parameter to vary across positive/negative and dimension also has little effect, and the results here for single kernels are closely matched by the double kernel results.

The second data set is from Aitchison and Ho (1988) and is a set of bivariate counts of surface and interior lens faults, with mild negative sample correlation of -0.20. Table 2 shows that again there is very little to choose between the four kernels, or the variants of them.

The third data set is a trivariate set of bacterial counts from three air samplers, also reported by Aitchison and Ho (1988). This data set is interesting in that it has relatively large negative correlation between the counts. It is clear from Table 2 that the exponential model is not competitive, at least while we restrict λ to ensure unimodality of the kernel. The geometric and logarithmic levels both seem more competitive than the modified logarithmic. The picture is similar for the variants on these kernels.

The final data set is the difficult bivariate set from Taillie etal (1979) which has very long tails. There is a slight complication with this data set in that (0,0) combinations are censored, but this is easily adjusted for. It proved impossible to find a finite solution for the exponential kernel as despite careful scaling and double precision arithmetic, the cross-validated likelihood underflowed to zero, regardless of λ. For this data set Table 3 shows that the logarithmic kernel is clearly superior, followed by the modified logarithmic kernel. It is interesting that for this data set, the double kernel is far superior to the single kernel, possibly reflecting that the estimated λ is close to 1 for these kernels making the asymmetry more severe. There is also some benefit in allowing λ to vary, although that advantage is small for the double log kernel, which is the best. Table 4 shows the goodness of fit of the logarithmic model compared with the independent logarithmic series model tried by Taillie et al. (1979). The kernel works well except for the boundary cells at (1,0) and (0,1) where the counts are significantly underestimated.

5. Discussion

The logarithmic kernel works well in all four of our sample data sets with the geometric kernel appearing competitive in the more regular data sets. The exponential kernel appears to be a poor choice, possibly because it dies away so quickly. The limit ratio of successive terms tends to zero for this kernel, while all the other kernels have a ratio with a limit of λ. However, the Potomac data set results are still rather disappointing, indicating that some work still remains to be done in finding kernels that handle data sets with as difficult boundary conditions as this one.

6. Conclusion

Discrete kernels, particularly those based on the logarithmic series provide a useful tool for modelling variability in multivariate count data.

REFERENCES

Aitchison, J. and Aitken, C.G.G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika,* **63**, 413–420.

Aitchison, J. and Ho, C.H. (1988). The multivariate Poisson-lognormal distribution. *Biometrika,* **76**, 643–654.

Arbons, A.G. and Kerrich, J.E. (1951). Accident statistics and the concept of accident proveness. *Biometrics,* **7**, 340–431.

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

Hohn, M.H. and Hellerman, J. (1963). The taxonomy and structure of diatom populations from three eastern North American rivers using three sampling methods. *Transactions of the American Microscopical Society,* **82**, 250–329.

McKendrick, A.G. (1926). Application of mathematics to medical problems. *Proc. Edin. Math. Soc.,* **44**, 98–130.

Taillie, C., Ord, J.K., Mosimann, J.E. and Patil, G.P. (1979). Discrete multivariate distributions. In *Statistical Distribution in Ecological Work,* Ed. J.K. Ord, G.P. Patil and C. Taillie, pp.159–178. Fairland, Maryland: International Cooperative Publishing House.

Table 1 Four simple count kernels $K(y|x,\lambda)$

Kernel	$w_j(\lambda)$	$W(x,\lambda)$	$W_p(\lambda)$
Geometric	λ^j	$\frac{\lambda(1-\lambda^x)}{1-\lambda}$	$\frac{\lambda}{1-\lambda}$
Logarithmic	$1 \quad (j=0)$	$\sum_{j=1}^{x} \frac{\lambda^j}{j}$	$-\log(1-\lambda)$
	$\frac{\lambda^j}{j} \quad (j>0)$		
Exponential	$1 \quad (j=0)$	$\sum_{j=1}^{x} \frac{\lambda^j}{j!}$	$e^\lambda - 1$
	$\frac{\lambda^j}{j!} \quad (j>0)$		
Modified logarithmic	$\frac{\lambda^j}{(j+1)}$	$\sum_{j=1}^{x} \frac{\lambda^j}{(j+1)}$	$\frac{-\log(1-\lambda)}{\lambda} - 1$

Table 2 Cross-validated log likelihoods for simple examples

			Kernel		
Data set	Dimension	Geometric	Logarithmic	Exponential	Mod Log
Shunter	2	−352.4	−351.4	−352.4	−352.7
Lens fault	2	−434.6	−434.9	−434.3	−435.4
Sampler	3	−410.5	−410.8	−434.5*	−414.0

* Solution at boundary $\lambda = 1$.

Table 3 Cross-validated log likelihood for Potomac data set

		Type of kernel			
Scale	Complexity	Geometric	Log	Exponential	Mod Log
Single	Simple	−1285.5	−1068.0	†	−1078.2
Single	Complex	−1191.6	−1050.4		−1055.1
Double	Simple	−1084.0	−1011.1		−1014.2
Double	Complex	−1065.8	−1005.4		−1006.9

Simple means constant λ (1 parameter)

Complex means λ different for positive and negative and for different dimensions. (4 parameters)

†Unable to find a finite solution in double precision arithmetic.

Table 4 Goodness of fit for the Potomac data set for the bivariate logarithmic series and the simple and complex double log kernel

<table>
<tr><th></th><th></th><th colspan="4">X_2</th></tr>
<tr><th></th><th></th><th>0</th><th>1</th><th>2</th><th>3+</th></tr>
<tr><td rowspan="4">X_1</td><td>0</td><td>▨</td><td>13
(11.9,4.2,4.4)</td><td colspan="2">9
(5.7,11.8,18.2)</td></tr>
<tr><td>1</td><td>19
(8.9,5.2,5.6)</td><td colspan="2" rowspan="2">5
(12.0,4.2,4.8)</td><td rowspan="2">5
(8.3,10.0,9.6)</td></tr>
<tr><td>2</td><td rowspan="2">10
(2.7,13.2,19.1)</td></tr>
<tr><td>3+</td><td colspan="2">8
(3.8,10.8,10.3)</td><td>68
(83.6,77.6,65.0)</td></tr>
</table>

Count
(expected)

Overall $\chi^2 = (46.1, 61.1, 60.7)$