Journal of Official Statistics
Vol. 4, No. 1, 1988, pp. 47–51

A Note on Choice of Bivariate Histogram Bin Shape

David W. Scott[1]

Abstract: Regular tiling of the plane may be accomplished with square, triangular, or hexagonal tiles. The statistical properties of bivariate histograms with square bin shapes are well-known. Here other choices of regular bin shapes are evaluated. Hexagonal bins are shown to be best asymptotically but only marginally better than square bins, which are 98% efficient. Equilateral triangular bins are only 91% efficient and usually should be avoided.

Key words: Bivariate histogram; bin shape.

1. Introduction

The square tiling pattern shown in Fig. 1a is the simplest bivariate histogram construction. It is natural to compare the statistical efficiency of square bins and other bin patterns with respect to a global error measure such as integrated mean squared error

$$\text{IMSE} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E[\hat{f}(x, y) - f(x, y)]^2 \, dx \, dy.$$

This problem corresponds to comparing kernel shapes in density estimation considered by Epanechnikov (1969). Improved histograms may be realized in several ways. For example, square bins may be stretched into rectangular bins, or the bins may be rotated away from the co-ordinate axes, or a completely nonregular adaptive mesh may be constructed. Here the focus is only on the shape of the histogram bin in a regular pattern. For real applications the pattern of bins should be generalized to include stretching and possibly rotation; see Hüsemann (1986).

2. Global Error for Several Bin Shapes

In this section, the integrated mean squared error is provided for the tilings shown in part (a) of Fig. 1–5. For comparison purposes each tile has area h^2 as in part (b) of Fig. 1–5. The total IMSE is found by aggregating integrated variance and squared bias estimates for individual bins, in which an arbitrary point is chosen as an origin for a Taylor series; see Scott (1979). Somewhat arbitrarily, these origin points were chosen to form a rectangular lattice.

[1] Department of Statistics, Rice University, Houston, TX 77251–1892, U.S.A.

Acknowledgments: This research was supported by the Office of Naval Research under grant N00014–85–K–0100. The author would like to thank the referees for their kind suggestions.

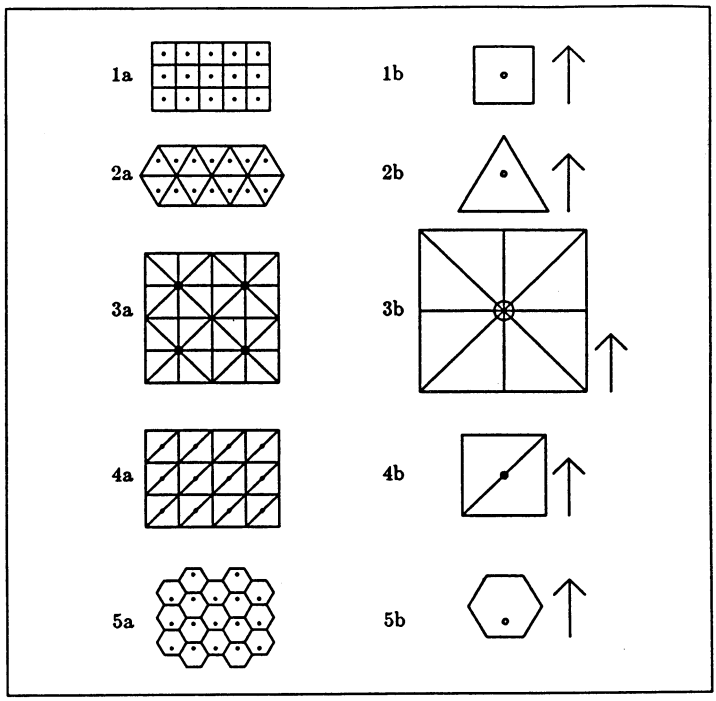

Fig. 1–5

(a) The five bivariate histogram bin or tile patterns considered in the paper are illustrated. The area of each tile is h^2.

(b) The smallest collection of repeatable tiles is shown at expanded scale, together with its common origin indicated by a small circle. Each arrow is of length h. Notice the collection of origins in part (a) of the Figures forms a rectangular lattice.

2.1. *Square bins*

The general approach to computing the bias and variance terms for the integrated mean squared error is illustrated in the case of square bins. To avoid technical issues, assume the density f is such that the necessary derivatives exist and approximation errors vanish. Let the sample count for the bin containing (x, y) be denoted by $v(x, y)$. The bivariate histogram is defined by

$$\hat{f}(x, y) = \frac{v(x, y)}{nh^2} . \tag{2.1}$$

Clearly $Ev(x, y) = np(x, y)$ where $p(x, y)$ is the corresponding bin probability mass. Consider a typical bin centered at the origin as shown in Fig. 1b:

$$p(x, y) = p(0,0) = \int_{-h/2}^{h/2} \int_{-h/2}^{h/2} f(x, y)\, dx\, dy .$$

Using a Taylor series and observing that $x^2 + y^2 \leq h^2/2$,

$$f(x, y) = f(0,0) + xf_x(0,0) + yf_y(0,0) + O(h^2), \tag{2.2}$$

where $\partial f / \partial x$ is denoted by f_x, it follows that

$$p(x, y) = h^2 f(0,0) + O(h^4) \tag{2.3}$$

and

$$E\hat{f}(x, y) = f(0,0) + O(h^2).$$

Using (2.2),

$$\text{Bias}(x, y) = -xf_x(0,0) - yf_y(0,0) + O(h^2).$$

Hence the integrated squared bias for the bin is

$$\int_{-h/2}^{h/2}\int_{-h/2}^{h/2} \text{Bias}(x, y)^2\, dx\, dy =$$

$$\frac{h^4}{12}[f_x(0,0)^2 + f_y(0,0)^2] + O(h^6). \tag{2.4}$$

Equation (2.4) generalizes to other bins if the point (0,0) is replaced with the respective bin center. Summing over all bins and using standard numerical approximations, the total integrated squared bias is

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \text{Bias}(x, y)^2\, dx\, dy = \frac{h^2}{12}(I_x^2 + I_y^2) + O(h^4), \tag{2.5}$$

where

$$I_x^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_x(x, y)^2\, dx\, dy$$

and similarly for I_y^2. Following (2.1) and (2.3), $\text{Var}\, v(x, y) = np(x, y)[1-p(x, y)]$ and

$$\text{Var}\,\hat{f}(x, y) = \frac{f(0,0)}{nh^2} + O\left(\frac{1}{n}\right). \tag{2.6}$$

Next integrate (2.6) over the bin and sum over all bins. Since f integrates to 1,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \text{Var}\,\hat{f}(x, y)\, dx\, dy = \frac{1}{nh^2} + O\left(\frac{1}{n}\right). \tag{2.7}$$

Adding (2.5) and (2.7), we obtain the well-known expression (Lecoutre (1985))

$$\text{IMSE} = \frac{1}{nh^2} + \frac{h^2}{12}(I_x^2 + I_y^2) + O\left(\frac{1}{n}\right). \tag{2.8}$$

2.2. *Triangular bin choices*

For other bin patterns, compute the bias and variance estimates for the basic bin shape as in Section 2.1 and then extend to the plane. For the variance estimate, a close examination of the derivation of equations (2.6) and (2.7) shows that the integrated variance for any choice of bin shape is simply the inverse of the sample size times the area of an individual bin. Hence the integrated variance term in the IMSE will always be $1/nh^2$ for any bin pattern if each bin is scaled so that the area equals to h^2; see part (b) of Fig. 1–5.

2.2.1. Equilateral triangles

Equilateral triangles form an appealing pattern shown in Fig. 2a. Analysis of this pattern or the pattern rotated 180 degrees leads to an expression similar to (2.4) except the divisor is $6\sqrt{3}$; hence

$$\text{IMSE} = \frac{1}{nh^2} + \frac{h^2}{6\sqrt{3}}(I_x^2 + I_y^2) + O\left(\frac{1}{n}\right), \tag{2.9}$$

which is inferior to (2.8) since $6\sqrt{3} = 10.39 < 12$.

2.2.2. Alternating diagonal cuts

A simple but technically nonregular pattern based upon right triangles with two equal sides is shown in Fig. 3a. The bias expression for individual right triangles corresponding to (2.4) is

$$\frac{h^4}{9}\,[f_x(0,0)^2 + f_y(0,0)^2 \pm f_x(0,0)\,f_y(0,0)],$$

with sign on the third term the same as the slope of the hypotenuse of the right triangle. The pattern in Fig. 3a may be built up from the basic eight-bin tile shown in Fig. 3b. The eight $\pm f_x(0,0)\,f_y(0,0)$ terms cancel leading to

$$\text{IMSE} = \frac{1}{nh^2} + \frac{h^2}{9}\,(I_x^2 + I_y^2) + O(\tfrac{1}{n}). \quad (2.10)$$

This pattern is clearly inferior to (2.9).

2.2.3. Diagonal cuts

Consider a pattern similar to Fig. 3a without alternating diagonal cuts, shown in Fig. 4a. This pattern may be built from the basic two-bin tile shown in Fig. 4b. Since all the hypotenuses have slope of + 1, the $f_x(0,0)\,f_y(0,0)$ terms do not cancel and

$$\text{IMSE} = \frac{1}{nh^2} + \frac{h^2}{9}\,(I_x^2 + I_y^2 + I_{xy}) + O(\tfrac{1}{n}), \quad (2.11)$$

where

$$I_{xy} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_x(x, y)\,f_y(x, y)\,dx\,dy.$$

Now I_{xy} may be negative if f is positively correlated. In fact, $I_{xy} = -(I_x^2 + I_y^2)/2$ is attainable, in which case (2.11) looks like equation (2.8) with the divisor 12 replaced by 18, a substantial improvement. This hints at the importance of considering rotation of bin patterns. However, it should be noted that rotating square bins will provide better results than diagonal cut triangles.

2.3. Hexagonal bins

Finally consider a hexagonal tiling shown in Fig. 5a and 5b. A similar but tedious computation reveals that

$$\text{IMSE} = \frac{1}{nh^2} + \frac{h^2}{36\sqrt{3}/5}\,(I_x^2 + I_y^2) + O(\tfrac{1}{n}), \quad (2.12)$$

which finally represents an improvement over (2.8), since $36\sqrt{3}/5 = 12.47$!

3. Relative Efficiency

It is straightforward to show that when comparing the IMSE's of the form (2.8), (2.9), (2.10), or (2.12) but not (2.11), the *ratio* of asymptotically *optimal* IMSE's is simply the square root of the ratio of the respective bias coefficients. Therefore, relative to regular hexagon bins, square bins are $\sqrt{5/3\ \sqrt{3}} = 98.1\%$ efficient while equilateral triangle bins are only $\sqrt{5/6} = 91.3\%$ efficient. The alternating diagonal cuts are only $\sqrt{\sqrt{3}/2} = 93.1\%$ efficient compared to equilateral triangles and only $\sqrt{5/4\ \sqrt{3}} = 85.0\%$ efficient relative to the hexagon bins.

4. Conclusions and Discussion

Among regular isomorph partitions of the plane, it has been shown that equilateral triangle bins asymptotically are a poor choice, although for small samples triangular bins may be superior (Matérn (1986)). This is somewhat unexpected in light of correspond-

ing results for bivariate frequency polygons (Scott (1985)). From the practical point of view, while hexagon bins are optimal, square bins are only 2% inefficient and are much easier to implement in computer software. However, Carr, Littlefield, Nicholson, and Littlefield (1987) have found other grounds to prefer hexagonal bins, based upon graphical perception of data displaying some correlation. The authors also present an algorithm for hexagonal binning.

In R^3, the cube is the only one of the five regular polyhedra which allows an equipartition of space. Many other nonregular tilings are commonly used and it would be interesting to evaluate their asymptotic performance (Conway and Sloane (1982)) and extend to adaptive tilings (Dodge (1986)).

5. References

Carr, D.B., Littlefield, R.J., Nicholson, W.L., and Littlefield, J.S. (1987): Scatterplot Matrix Techniques for Large N. Journal of the American Statistical Association, 82, pp. 424–436.

Conway, J.H. and Sloane, N.J.A. (1982): Varonoi Regions of Lattices, Second Moments of Polytopes and Quantization. IEEE Transactions Information, Theory, Vol. IT-28, pp. 211–226.

Dodge, Y. (1986): Some Difficulties Involving Nonparametric Estimation of a Density Function. Journal of Official Statistics, 2(2), pp. 193–202.

Epanechnikov, V.A. (1969): Nonparametric Estimation of a Multidimensional Probability Density. Theory of Probability and its Applications, 14, pp. 153–158.

Hüsemann, J.A.S. (1986): Histogram Estimators of Bivariate Densities. Technical Report 86–5, Department of Mathematical Sciences, Rice University, Houston, Texas.

Lecoutre, J.-P. (1985): The L_2-Optimal Cell Width for the Histogram. Statistics and Probability Letters, 3, pp. 303–306.

Matérn, B. (1986): Spatial Variation. Springer-Verlag, Berlin.

Scott, D.W. (1979): On Optimal and Data-Based Histograms. Biometrika, 66, pp. 605–610.

Scott, D.W. (1985): Frequency Polygons: Theory and Application. Journal of the American Statistical Association, 80, pp. 348–354.

Received September 1987
Revised March 1988