Modified Least-squares Density Estimates

Serge B. Provost
The University of Western Ontario
Department of Statistical & Actuarial Sciences
London, Canada
provost@stats.uwo.ca

Abstract

It is often the case that the moments of a continuous distribution whose support is confined to a closed interval can be readily determined, while its exact density function either does not lend itself to numerical evaluation or proves to be mathematically intractable. In such instances, the moments of the distribution can be used in conjunction with a Hilbert matrix to provide an accurate polynomial approximation of the density function. The density approximants so obtained are in fact least-squares approximating polynomials, which can also be expressed as linear combinations of Legendre orthogonal polynomials or as sums involving an explicit representation of the elements of the inverse of a Hilbert matrix. Interestingly, when sample moments are being used, such polynomial estimates can be expressed as kernel density estimates. It will be demonstrated via a kernel analysis that, on extending the range of a sample, one can obtain estimates that are smoother and exhibit less fluctuation in the tails. This will be illustrated with the `Buffalo snowfall' data set. Another example involves approximating the distribution of the distance between two points that are randomly distributed inside a unit cube. An extension of the proposed technique to the bivariate case will be illustrated with successive observations from the geyser eruption data set.

REFERENCES

Alexits, G. (1961). Convergence Problems of Orthogonal Series. Pergamon Press, New York.

Burden, R. L. and Faires, J. D. (1997). Numerical Analysis, 6th Ed. Brooks/Cole, New York.

Devroye, L. (1989). On random variate generation when only moments or Fourier coefficients are known. Mathematics and Computers in Simulation 31, 71-89.

Elderton, W. P. and Johnson, N. L. (1969). Systems of Frequency Curves. Cambridge University Press.

Freedman, D. and Diaconis, P. (1981). On the histogram as a density estimator: L_2 theory. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebeite 57, 453-476.

Hall, P. (1982). Comparison of two orthogonal series methods of estimating a density and its derivatives on an interval. Journal of Multivariate Analysis 12, 432-449.

Izenman, A. J. (1991). Recent development in nonparametric density estimation. Journal of the American Statistical Association 86, 205-224.

Mathai, A. M. (1999). An Introduction to Geometric Probability: Distributional Aspects with Applications. Gordon and Breach, The Netherlands.

Pusey, M. (1970). Representations of one orthonormal set in terms of another. Biometrika 57, 445-450.

Sansone, G. (1959). Orthogonal Functions. Interscience Publishers Inc. New York.

Scott, D. W. (1979). On optimal and data-based histograms. Biometrika 66, 605-610.

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

RÉSUMÉ

Il est souvent plus facile de déterminer les moments d'une loi de probabilité que d'en obtenir la fonction de densité exacte. Nous montrons qu'il est toujours possible d'approximer une fonction de densité continue possédant un support compact en résolvant un système d'équations linéaires qui est fonction d'une matrice de Hilbert et ne requiert que les moments de cette loi. L'approximation qui en résulte est en fait un polynôme de moindres carrés lequel peut également être representé par une combinaison linéaire de polynômes orthogonaux de Legendre ou encore par une somme dont les coefficients sont fonctions d'une représentation explicite des éléments de l'inverse d'une matrice de Hilbert. Il est intéressant de constater que lorsqu'on applique cette technique à des moments échantillonnaux, le polynôme qui estime la fonction de densité peut être representé par un estimateur à noyaux. Or, une analyse de ces noyaux démontre qu'en accroissant l'étendue de l'échantillon, on peut obtenir un estimateur plus lisse dont le comportement s'avère plus régulier, particulièrement au niveau des ailes. En guise d'exemple, nous nous servirons du jeu de données bien connu sur les précipitations annuelles de neige à Buffalo. Dans un second exemple, nous approximons la fonction de densité de la distance entre deux points aléatoirement distribués à l'intérieur d'un cube. Nous généralisons la technique proposée au cas bidimensionnel et utilisons à titre illustratif les données successives sur les durées d'éruption d'un geyser.