A Combined Adaptive-Mixtures/Plug-In Estimator of Multivariate Probability Densities[1]

J. Ćwik and J. Koronacki
Institute of Computer Science, Polish Academy of Sciences
Ordona 21, 01-237 Warsaw, Poland

to appear in
Computational Statistics and Data Analysis
(received August 1995; revised February 1997)

[1]This work was supported by the State Committee for Scientific Research (KBN) under grants 2 P301 001 06 and 8 T11C 019 11.

Abstract: A multivariate extension of the plug-in kernel (and filtered kernel) estimator is proposed which uses asymptotically optimal bandwidth matrix (matrices) for a normal mixture approximation of a density to be estimated (the filtered kernel estimator uses different matrices for different clusters of data). The normal mixture approximation is provided by a recursive version of the EM algorithm whose initial conditions are in turn obtained via an application of the ideas of adaptive mixtures density estimation and AIC-based pruning. Simulations show that the estimator proposed, while it is in fact a rather complex multistage estimation process, provides a very reliable way of estimating arbitrary and highly structured continuous densities on R^2 and, hopefully, R^3.

Key words: nonparametric density estimation; plug-in kernel estimation; recursive EM algorithm; Gaussian clustering algorithm

1 Introduction

Nonparametric probability density estimation in one dimension is now a well developed area with great potential for practical applications (see, e.g., Härdle (1991), Scott (1992), Sheather (1992), Park and Turlach (1992), Cao et. al. (1994), Ciesielski (1991), Wand, Marron and Ruppert (1991), Ruppert and Cline (1994)). The problem, however, becomes much more challenging when data dimensionality is increased to two, let alone to three or four (see Scott (1992), Wand (1992), Wand and Jones (1993), and, for particularly promising estimators proposed so far, Ciesielski (1990), O'Sullivan and Pawitan (1993), Sain, Scott and Baggerly (1994), Wand and Jones (1994)).

In this report, a comparative simulation study of an estimator which combines the ideas of adaptive mixtures and kernel estimation is presented for two-dimensional data. Let $\mathbf{x}_1, \ldots, \mathbf{x}_N$ be a random sample from an unknown distribution with a d-variate probability density f, $f : R^d \to R$, $1 \leq d < \infty$. Let the kernel estimator of f have the form

$$f_N(\mathbf{x}) = N^{-1}|\mathbf{H}|^{-1/2} \sum_{i=1}^{N} K(\mathbf{H}^{-1/2}(\mathbf{x} - \mathbf{x}_i)) \tag{1}$$

where $\mathbf{x}$ is a vector in R^d, $\mathbf{H}$ is a symmetric positive definite $d \times d$ matrix with $\mathbf{H}^{1/2}$ to be referred to as the bandwidth matrix (bandwidth if $d = 1$), and K is a d-variate probability density function. Throughout this paper we assume that K is the standard Gaussian density, i.e.,

$$K(\mathbf{x}) = (2\pi)^{-d/2} \exp(-\frac{1}{2}\mathbf{x}^T\mathbf{x})$$

Loosely speaking, the estimation process consists in: (i) approximating unknown density f by a mixture $\tilde{f}$ of normal densities, with the number of mixture components being finite but otherwise unspecified; (ii) determining bandwidth matrix $\mathbf{H}^{1/2}$ which minimizes the leading two terms in the asymptotic expansion of the mean integrated squared error (MISE) of estimator (1) for the mixture obtained,

$$\text{MISE} = E\int_{R^d} (f_N(\mathbf{x}) - \tilde{f}(\mathbf{x}))^2 d\mathbf{x} \tag{2}$$

(the asymptotic approximation to MISE thus obtained will be referred to as AMISE); (iii) applying estimator (1) with the minimizing $\mathbf{H}^{1/2}$ to the original data. (Preliminary study of this approach was presented by Ćwik and Koronacki (1996a and b).)

The estimator just described can be considered a straightforward extension of the plug-in approach (to kernel density estimation) to dimensions higher than one. According to that approach, the asymptotic approximation of (2) with $\tilde{f}$ replaced by unknown f should be minimized with respect to $\mathbf{H}^{1/2}$, the unknowns in the minimizing bandwidth matrix should be replaced by their consistent estimates, and the resulting matrix should be used in (1) to estimate f from data. In the univariate setting, the plug-in approach is the most recommended one, provided the plug-in estimator of Sheather and Jones (1991) is used; see Sheather (1992), Park and Turlach (1992), Cao et. al. (1994). Recently, the bivariate extension of the plug-in method of Sheather and Jones, based on using estimator (1) with a general bandwidth matrix, has been provided by Wand and Jones (1994). As far as we are aware, however, no efficient plug-in algorithm for choosing the bandwidth matrix for $d = 3$ (let alone more) is known. The problem is that with d increasing the number of unknowns which have to be estimated from data rapidly increases (more precisely, the unknowns in the resulting bandwidth matrix are functionals of the Hessian of f). On the other hand, this problem is completely alleviated by replacing f in the derivation of $\mathbf{H}^{1/2}$ by the approximating mixture $\tilde{f}$.

Moreover, within our approach, it is a straightforward matter to localize the choice of a bandwidth matrix, namely to use different matrices for different data clusters. Indeed, since the density to be estimated is first approximated by a normal mixture, it suffices to apply the plug-in approach to each mixture component separately, and then to use a suitable convex combination of estimators with bandwidth matrices obtained for these component densities. Essentially, the first step of the estimation process, i.e., approximating f by a normal mixture, is based on the idea of adaptive mixtures density estimation (AMDE) developed by Priebe and Marchette (see Priebe (1994)).

Simulations have been performed for bivariate samples from eight normal mixture test densities of Wand and Jones (1991; all but the last test density is included in Wand and Jones (1993)), two gamma mixture test densities and three non-mixture test densities. In addition, the estimators have been applied to the well-known plasma lipid data observed on 320 males suffering from coronary heart disease (see, e.g., Scott (1992)) and to a set of 600 longitude/latitude pairs of the epicentres of the earthquakes in the Mount Saint Helens area (obtained from Professor Tony Qmar from the University of Washington).

The estimator is introduced in Section 2. Simulation results are summarized, and some concluding remarks are given, in Section 3. While admittedly the estimator studied is in fact a multistage, and thus rather complex, estimation process, it can be considered a truly reliable one. A small simulation study which is not reported in the paper has shown that a similar conclusion seems to hold for 3D data as well (see Ćwik and Koronacki (1996b) where estimation of five 3D normal mixtures is studied).

References

Cao R., A. Cuevas and W. Gonzales-Manteiga, A comparative study of several smoothing methods in density estimation, *Comp. Statist. Data Analysis*, 17 (1994) 153-176.

Ciesielski Z., Asymptotic nonparametric spline density estimation in several variables, *International Studies of Numerical Mathematics*, Birkhauser, **94** (1990) 25-53.

Ciesielski Z., Asymptotic nonparametric spline density estimation, *Probab. Math. Statist.*, **12** (1991) 1-24.

Ćwik J. and J. Koronacki, Probability density estimation using a Gaussian clustering algorithm, *Neural Computing & Applications*, **4** (1996a) 149-160.

Ćwik J. and J. Koronacki, Multivariate density estimation: A comparative study, Technical Report, Institute of Computer Science, Polish Acad. Sci., 1996b.

Everitt B.S. and D.J. Hand, *Finite mixture distributions*, Chapman and Hall, 1981.

Friedman J.H., Exploratory projection pursuit, *J. Amer. Statist. Assoc.*, **82** (1987) 249-266.

Härdle W., *Smoothing techniques*, Springer, 1991.

Hartigan J.A. and M.A. Wong, A k-means algorithm, *Applied Statistics*, **28** (1979) 100-108 (also in: P. Griffiths and I.D. Hill (eds.), *Applied statistics algorithms*, Ellis Horwood, 1981).

Johnson, N.L. and S. Kotz, *Distributions in statistics: continuous multivariate distributions*, Wiley, 1972.

Kowalczyk, T. and J. Tyrcha, Multivariate gamma distributions – Properties and shape estimation, *Statistics*, **20** (1989) 465-474.

MacQueen J.B., Some methods for classification and analysis of multivariate observations, *Proc. Fifth Berkeley Symp. Math. Stat. Prob.*, 1967.

Marchette D.J, C.E. Priebe, G.W. Rogers and J.L. Solka, Filtered kernel density estimation, to appear in *Computational Statistics*, 1996.

O'Sullivan F. and Y. Pawitan, Multidimensional density estimation by tomography, *J. R. Statist. Soc. B*, **55** (1993) 509-521.

Park B.U. and B.A. Turlach, Practical performance of several data driven bandwidth selectors, *Comp. Statist.*, **7** (1992) 251-270 and discussion.

Priebe, C.E., Adaptive mixtures, *J. Amer. Statist. Assoc.*, **89** (1994) 796-806.

Ruppert D. and D.B.H. Cline, Bias reduction in kernel density estimation by smoothed empirical transformations, *Ann. Statist.*, **22** (1994) 185-210.

Sain. S, K. Baggerly and D.W. Scott, Cross-validation of multivariate densities, *J. Amer. Statist. Assoc.*, **89** (1994) 807-817.

Scott D.W., *Multivariate density estimation: Theory, practice, and visualization*, Wiley, 1992.

Sheather S.J., The performance of six popular bandwidth selection methods on some real data sets, *Comp. Statist.*, **7** (1992) 225-250 and discussion.

Sheather S.J. and M.C. Jones, A reliable data-based bandwidth selection method for kernel density estimation, *J. Roy. Statist. Soc. ser. B*, **53** (1991) 683-690.

Silverman B.W., *Density estimation for statistics and data analysis*, Chapman and Hall, 1986.

Solka, J.L, W.L. Poston i E.J. Wegman, A visualization technique for studying the iterative estimation of mixture densities, *J. Computational and Graphical Statist.* **4** (1995), 180-197.

Solka, J.L., E.J. Wegman, C.E. Priebe, W.L. Poston i G.W. Rogers, Mixture structure analysis using the Akaike information criterion and the bootstrap, to appear in *Statist. and Computing*, 1966.

Titterington D.M, A.F.M. Smith and U.E. Makov, *Statistical analysis of finite mixture distributions*, Wiley, 1985.

Tråvén H.G.C., A neural network approach to statistical pattern classification by semiparametric estimation of probability density functions, *IEEE Trans. Neural Networks*, **2** (1991) 366-377.

Wand M.P., Error analysis for general multivariate kernel estimators, *J. Nonpar. Statist.*, **2** (1992) 1-15.

Wand M.P. and M.C. Jones, Comparison of smoothing parametrizations in bivariate kernel density estimation, *Technical Report*, Department of Statistics, Rice University, 1991.

Wand M.P. and M.C. Jones, Comparison of smoothing parametrizations in bivariate kernel density estimation, *J. American Statist. Assoc.*, **88** (1993) 520-528.

Wand M.P. and M.C. Jones, Multivariate plug-in bandwidth selection, *Comp. Statist*, **9** (1994) 97-116.

Wand M.P., J.S. Marron and D. Ruppert, Transformations in density estimation, *J. Amer. Statist. Assoc.*, **86** (1991) 343-353 and discussion.