Multivariate Density Estimation: A Comparative Study[1]

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

submitted to
Neural Computing and Applications

[1]This work was supported in part by the State Committee for Scientific Research (KBN) under grant 2 P301 001 06.

Abstract – This paper is a continuation of the authors' earlier work [1], where a version of the Tråvén's [2] Gaussian clustering neural network (being a recursive counterpart of the EM algorithm) has been investigated. A comparative simulation study of the Gaussian clustering algorithm studied in [1], two versions of plug-in kernel estimators and a version of Friedman's projection pursuit algorithm is presented for two- and three-dimensional data. Simulations show that the projection pursuit algorithm is a good or a very good estimator provided, however, that the number of projections is suitably chosen. Although practically confined to estimating normal mixtures, the simulations confirm general reliability of plug-in estimators and show the same property of the Gaussian clustering algorithm. Indeed, the simulations confirm the earlier conjecture that this last estimator provides a way of effectively estimating arbitrary and highly structured continuous densities on R^d, at least for small d, either by using this estimator itself or, rather, by using it as a pilot estimator for a newly proposed plug-in estimator.

Keywords: Gaussian clustering neural network; Recursive EM algorithm; Plug-in kernel estimator; Projection pursuit; Non-parametric density estimation.

1 Introduction

Non-parametric estimation of multivariate probability densities is an important ingredient of many pattern recognition problems. In 1990, Specht [3] noticed that the most widely used non-parametric estimator of an unknown density, namely the kernel estimator, has the form of a neural network. The problem, however, was that no estimator can work reliably until its smoothing parameters (bandwidth matrix) are chosen correctly. No wonder that, for years, research has been focused on finding reliable methods of adaptive, i.e., data-based, choice of these parameters. In fact, it took more than a decade to bring the area of non-parametric density estimation to the state of practical applicability, and that being the case only if the data are one-dimensional (see, e.g., Härdle [4], Scott [5], Sheather [6], Park and Turlach [7], Cao et. al. [8], Ciesielski [9], Wand et. al. [10] and Ruppert and Cline [11]). In the univariate setting, the most recommended approach to choosing the estimator's smoothing parameter is now the so-called plug-in approach. At the same time, more studies are still needed when data dimensionality is increased to two, let alone to three or four (see Scott [5], Wand [12], Wand and Jones [13], and, for particularly promising methods proposed so far, Ciesielski [14], O'Sullivan and Pawitan [15], Sain, Scott and Baggerly [16], Wand and Jones [17]).

In this report, a comparative simulation study of several density estimators for two- and three-dimensional data is presented. (Moreover, a small simulation study which is not reported in the paper has shown that similar conlusions to those obtained for 2D and 3D data seem to hold for 4D data as well.) The following estimators are compared: two versions of plug-in kernel estimators, a version of Friedman's projection pursuit algorithm, and a "Gaussian clustering" (GC) neural network earlier studied in [1]. The last of these algorithms, which is a version of the Tråvén's [2] network, is in fact a recursive counterpart of the EM algorithm and, thus, is also a version of the recursive EM algorithm of Titterington [18]. One of the plug-in kernel estimators considered is an "ideal" or "purely theoretical" version of the estimator with the bandwidth matrix proposed by Wand and Jones [17]. Namely, in a simulation study, the densities to be estimated are in fact known and this knowledge can be used to compute bandwidth matrices which minimize the leading terms in the asymptotic expansion of the estimator's mean integrated squared error. In the other plug-in estimator, the Sheather-Jones [19] bandwidths are used for each co-ordinate separately.

In the 2D case, simulations have been performed for samples from twelve (out of fifteen) example normal mixture densities used by Ćwik and Koronacki [1] and borrowed from Wand and Jones [20]; all but one test density are included in Wand and Jones [13]. In the 3D case, five example normal mixture densities have been used, four of them used earlier by Ćwik and Koronacki [1] and the fifth included in, and borrowed from, Wand's program for minimizing asymptotic mean integrated squared error of a general kernel estimator (the program has been kindly provided by the author). Using normal mixtures as test densities is motivated by the fact that, actually, any density can be approximated by such a mixture to within a preassigned accuracy (say, in L_1 norm). 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 [5]; it is worthwhile to note that the data can hardly be believed to come from a normal mixture density).

The estimators studied are briefly recalled in Section 2. Simulation results are summarized, and some concluding remarks are given, in Section 3. In particular, the GC algorithm, as was conjectured in [1], has indeed been found to be very promising. Interestingly, while the GC algorithm may sometimes prove unreliable, e.g., for strongly kurtotic normal mixtures or for some non-mixture densities, it can still be used to our advantage as it provides a ready means for an extension of the (strict-sense non-parametric) plug-in approach to higher dimensions. This last possibility is briefly discussed in Section 3 too, and it is dealt with in detail in a separate report (see Ćwik and Koronacki [21]; in that report, simulation results for several non-mixture densities are also given, thus confirming general validity of recommendations based on the current study).

References

1. Ćwik J, Koronacki J. Probability density estimation using a Gaussian clustering algorithm. Neural Computing & Applications 1996; 4: 149-160.

2. Tråvén HGC. A neural network approach to statistical pattern classification by semiparametric estimation of probability density functions. IEEE Trans. Neural Networks 1991; 2: 366-377.

3. Specht DF. Probabilistic Neural Networks. Neural Networks 1990; 3: 109-118.

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

5. Scott DW. Multivariate density estimation: Theory, practice, and visualization. Wiley, 1992.

6. Sheather SJ. The performance of six popular bandwidth selection methods on some real data sets. Comp. Statist. 1992; 7: 225-250 and discussion.

7. Park BU, Turlach BA. Practical performance of several data driven bandwidth selectors. Comp. Statist. 1992; 7: 251-270 and discussion.

8. Cao R, Cuevas A, Gonzales-Manteiga W. A comparative study of several smoothing methods in density estimation. Comp. Statist. Data Analysis 1994; 17: 153-176.

9. Ciesielski Z. Asymptotic nonparametric spline density estimation Probab. Math. Statist. 1991; 12: 1-24.

10. Wand MP, Marron JS, Ruppert D. Transformations in density estimation. J. Amer. Statist. Assoc. 1991; 86: 343-353 and discussion.

11. Ruppert D, Cline DBH. Bias reduction in kernel density estimation by smoothed empirical transformations. Ann. Statist. 1994; 22: 185-210.

12. Wand MP. Error analysis for general multivariate kernel estimators. J. Nonpar. Statist. 1992; 2: 1-15.

13. Wand MP, Jones MC. Comparison of smoothing parametrizations in bivariate kernel density estimation. J. American Statist. Assoc. 1993; 88: 520-528.

14. Ciesielski Z. Asymptotic nonparametric spline density estimation in several variables. International Studies of Numerical Mathematics, Birkhauser, 1990; 94: 25-53.

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

16. Sain S, Baggerly K, Scott DW. Cross-validation of multivariate densities. J. Amer. Statist. Assoc. 1994; 89: 807-817.

17. Wand MP, Jones MC. Multivariate plug-in bandwidth selection. Comp. Statist. 1994; 9: 97-116.

18. Titterington DM. Recursive parameter estimation using incomplete data. J. R. Statist. Soc. B (1984); 46: 257-267.

19. Sheather SJ, Jones MC. A reliable data-based bandwidth selection method for kernel density estimation. J. Roy. Statist. Soc. B 1991; 53: 683-690.

20. Wand MP, Jones MC. Comparison of smoothing parametrizations in bivariate kernel density estimation. Technical Report, Department of Statistics, Rice University, 1991.

21. Ćwik J. and J. Koronacki, A combined adaptive-mixtures/plug-in estimator of multivariate probability densities. To be submitted, 1996.

22. Friedman JH, Stuetzle W, Schroeder A. Projection pursuit density estimation. Amer. Statist. Assoc. 1984; 79: 599-608.

23. Friedman JH. Exploratory projection pursuit. J. Amer. Statist. Assoc. 1987; 82: 249-266.

24. Huber PJ. Projection pursuit. Ann. Statist. 1985; 13: 435-475 and discussion.

25. Priebe CE. Adaptive mixtures J. Amer. Statist. Assoc. 1994; 89: 796-806.

26. Silverman BW. Density estimation for statistics and data analysis. Chapman and Hall, 1986.