Jurnal Teknologi

Full paper

Nonparametric Least Squares Mixture Density Estimation

Chew-Seng Chee[a*]

[a]Department of Mathematics, Faculty of Science and Technology, Universiti Malaysia Terengganu, 21030 Kuala Terengganu, Malaysia

*Corresponding author: chee@umt.edu.my

Article history

Received :21 January 2013
Received in revised form :
7 May 2013
Accepted :25 June 2013

Graphical abstract

$$\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i),$$

Abstract

In this paper, we consider using nonparametric mixtures for density estimation. The mixture density estimation problem simply reduces to the problem of estimating a mixing distribution in the nonparametric mixture model. We focus on the least squares method for mixture density estimation problem. In a simulation experiment, the performance of the least squares mixture density estimator (MDE) and the kernel density estimator (KDE) is assessed by the mean integrated squared error. The performance improvement of MDE over KDE for some common densities is achieved by using cross-validation method for bandwidth selection.

Keywords: Nonparametric mixtures; least squares estimation; kernel density estimation; bandwidth selection; cross-validation

Abstrak

Dalam kertas ini, kami mempertimbangkan penggunaan model bercampur tak berparameter untuk penganggaran fungsi ketumpatan. Masalah penganggaran fungsi ketumpatan secara campuran merupakan masalah penganggaran taburan bercampur dalam model bercampur tak berparameter. Kami menumpukan kepada kaedah kuasa dua terkecil untuk masalah penganggaran fungsi ketumpatan secara campuran. Dalam satu ujikaji simulasi, prestasi penganggar kuasa dua terkecil fungsi ketumpatan secara campuran (MDE) dan penganggar fungsi ketumpatan secara inti (KDE) telah dinilai oleh min ralat kuasa dua terkamir. Peningkatan prestasi MDE terhadap KDE bagi sesetengah fungsi ketumpatan popular telah dicapai dengan menggunakan kaedah silang pengesahan untuk pemilihan lebar jalur.

Kata kunci: Model bercampur tak berparameter; penganggaran kuasa dua terkecil; penganggaran fungsi ketumpatan secara inti; pemilihan lebar jalur; silang pengesahan

1.0 INTRODUCTION

Density estimation is concerned with the problem of estimating a probability density function based on sample data. The books by Silverman [15] and Eggermont and LaRiccia [4] provide an excellent account of methods of density estimation. There are basically three main strands in density estimation, namely the parametric, nonparametric and mixture model approaches. The problem of parametric density estimation simply reduces to estimating the assumed model parameter by some standard method such as maximum likelihood or Bayesian estimation, and finally returns a plug-in density estimate. A parametric approach to density estimation is useful in the presence of a prior knowledge about the true density; otherwise it can be criticized, certainly on the grounds of specification of the model density. Many approaches have been proposed to deal with nonparametric density estimation problems. Specifically, the conventional nonparametric smoothing technique known as kernel density estimation offers a flexible framework for exploring the structure of the data or the shape of the underlying density. Nevertheless, the kernel method which employs a kernel function with bandwidth parameter to smooth out the discrete empirical density function tends to produce a flattened density estimate of the actual density.

Apart from the two standard approaches mentioned above, mixture models, which offer a flexible class of densities and contain the kernel models as special cases, are particularly useful for density estimation because they can approximate many densities reasonably well. Marron and Wand [10] illustrated the flexibility of the class of normal mixtures in covering a variety of different density shapes. Using a mixture model not only can avoid model misspecification, but also provide a sparser estimator compared to the kernel model. Recognizing the advantages of using mixtures for density estimation, Scott and Szewczyk [14] designed a procedure that starts with a kernel density estimate, then sequentially simplifies that overparameterized mixture estimate and finally selects a simplified mixture estimate based on some criterion. Similarly, the so-called variable location kernel density estimator, which automatically indicates a simpler mixture structure after the fitting by maximum likelihood, was studied in Jones and Henderson [5]. Other attempts at using mixtures for

density estimation include, but are not limited to, articles by Roeder and Wasserman [12] and Priebe and Marchette [11].

We, however, prefer the direct specification of mixture models for density estimation and particularly advocate nonparametric mixture models (Lindsay [8]; Böhning [2]). Here, we mention the work of Wang and Chee [17], in which nonparametric mixtures by maximum likelihood estimation were used for density estimation. Basically, for a fixed value of the bandwidth parameter, the mixture density estimation problem simply reduces to the problem of estimating a mixing distribution in the nonparametric mixture model. In the literature, the maximum likelihood method for nonparametric estimation of the mixing distribution has flourished for many years. Recently, nonparametric estimation via the least squares method has been considered; see, e.g., Yuan [18] and Balabdaoui and Wellner [1]. Following the same vein, we consider the least squares functional for nonparametrically estimating the mixing distribution in this paper. We then apply the resulting mixture density to density estimation.

This paper is organized in the following way. Section 2 gives a very brief background description on some density estimators. Section 3 outlines the nonparametric least squares estimation problem, along with some computational aspects of estimating a mixing distribution. Section 4 gives a report on a simulation experiment carried out to investigate the finite sample performance of both the least squares mixture and kernel density estimators. Section 5 contains some concluding remarks.

2.0 SOME DENSITY ESTIMATORS

In this section, we only very briefly describe the fixed bandwidth kernel and mixture density estimators. Kernel density estimation is undoubtedly one of the most popular nonparametric smoothing techniques in the statistical literature. The kernel density estimator (KDE) based on a random sample $x_1, \ldots, x_n \in R$ with common density f is given by

$$\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i), \tag{1}$$

with $K_h(y) = K(y/h)/h$, where K is called a kernel function and $h > 0$ is known as the bandwidth or smoothing parameter. The kernel K is usually taken to be a symmetric and unimodal density function. In practice, it is well-known that the impact of the choice of the kernel function as compared to that of the bandwidth is essentially of little concern. Selecting the bandwidth has been an active and challenging research topic and numerous methods of bandwidth selection have been proposed but none is superior or better than the others in all cases; see, e.g., Jones *et al.* [6] and Loader [9].

As can be seen from (1), an obvious drawback of the kernel method is that all data points are needed in the construction of the KDE. To provide sparse representation for the data, one can use nonparametric mixtures. The density of the nonparametric (location) mixture model is given by

$$\widehat{f}_\beta(x; G) = \int K_\beta(x - \theta)\, \mathrm{d}G(\theta), \tag{2}$$

with $K_\beta(y) = K(y/\beta)/\beta$, where $\beta > 0$ is a bandwidth parameter and G is called the mixing distribution. Here, G is treated as an arbitrary mixing distribution, which can be either discrete, continuous or in any parametric family. The theory of nonparametric estimation of mixtures establishes that under mild conditions the nonparametric maximum likelihood estimate (NPMLE) of G for any fixed β is always discrete (Lindsay [8]). Thus, as far as the method of maximum likelihood is concerned, we may write (2) as

$$\widehat{f}_\beta(x; \boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{j=1}^{J} \pi_j K_\beta(x - \theta_j),$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)^\top$ is a support point vector of distinct elements with corresponding probability mass vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J)^\top$. The vector $\boldsymbol{\pi}$ is elementwise positive and its components sum to unity. With the availability of the NPMLE for a fixed β, it is straightforward to construct the maximum likelihood mixture density estimate. Similar to the kernel density estimation, bandwidth selection remains a challenge for the mixture density estimation (Wang and Chee [17]). In contrast, we shall study the least squares mixture density estimator in this paper.

3.0 NONPARAMETRIC LEAST SQUARES ESTIMATION

The method of least squares has a long tradition in regression estimation. In nonparametric smoothing, the idea of least squares cross-validation is applied to selecting the bandwidth parameter. Further use of this method as a practical estimation method for a variety of parametric models can be found in Scott [13].

The least squares functional based on the observations $x_1, \ldots, x_n$ is defined as

$$\mathcal{L}_\beta(G) = \int \left\{\widehat{f}_\beta(x; G)\right\}^2 \mathrm{d}x - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_\beta(x_i; G). \tag{3}$$

In this paper, we study the problem of minimizing the least squares functional (3) for any fixed β. Also, we consider approximating the estimate of G by a discrete distribution. As a matter of fact, an appropriate discrete distribution can approximate any distribution to any desired level of accuracy. We shall refer to this discrete nonparametric estimate of G as the nonparametric least squares estimate (NPLSE). Due to using a discrete NPLSE, the estimated nonparametric mixture density has a discrete form, similar to that obtained by maximum likelihood estimation. We shall call the resulting density estimate the least squares mixture density estimate (MDE).

The NPLSE is completely characterized by the gradient function given by

$$d(\theta; G) = 2\int \left\{\widehat{f}_\beta(x;\theta)\widehat{f}_\beta(x; G)\right\} \mathrm{d}x - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_\beta(x_i;\theta) - 2\int \left\{\widehat{f}_\beta(x; G)\right\}^2 \mathrm{d}x + \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_\beta(x_i; G).$$

We have that a candidate G^* is the NPLSE if and only if $d(\theta; G^*) \geq 0$ for all θ, and if G^* is the NPLSE, its support points are contained in the set $\{\theta : d(\theta; G^*) = 0\}$. The result given here is an analogous development to that of the nonparametric maximum likelihood estimation of a mixing distribution (Lindsay [8]; Böhning [2]).

By far the most common K used in practice is the standard normal density, and hence we confine our study to the Gaussian kernel. An advantage of using the Gaussian kernel is that we obtain an explicit closed-form expression for the objective functional. Denoting the density of the Gaussian distribution with mean μ and standard deviation σ by $\phi_\sigma(x - \mu)$, we have

$$\int \phi_{\sigma_1}(x - \mu_1)\phi_{\sigma_2}(x - \mu_2)\, \mathrm{d}x = \phi_{\sqrt{\sigma_1^2 + \sigma_2^2}}(\mu_1 - \mu_2).$$

Using the above identity, we have a simple expression for $\mathcal{L}_\beta(G) \equiv \mathcal{L}_\beta(\boldsymbol{\pi}, \boldsymbol{\theta})$:

$$\sum_{k=1}^{J}\sum_{j=1}^{J} \pi_k \pi_j \phi_{\sqrt{2}\beta}(\theta_k - \theta_j) - \frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} \pi_j \phi_\beta(x_i - \theta_j).$$

Also, for this special case, the gradient function can be easily shown to be:

$$d(\theta;G) \equiv d(\theta;\pi,\theta) = 2\sum_{j=1}^{J}\pi_j\phi_{\sqrt{2}\beta}(\theta_j-\theta) - \frac{2}{n}\sum_{i=1}^{n}\phi_\beta(x_i-\theta) - 2\sum_{k=1}^{J}\sum_{j=1}^{J}\pi_k\pi_j\phi_{\sqrt{2}\beta}(\theta_k-\theta_j) + \frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{J}\pi_j\phi_\beta(x_i-\theta_j).$$

Denoting $\mathrm{D}_{k,j} = \phi_{\sqrt{2}\beta}(\theta_k-\theta_j)$ and $\mathrm{b}_j = \frac{1}{n}\sum_{i=1}^{n}\phi_\beta(x_i-\theta_j)$, the objective functional can be written more compactly in matrix form:

$$\mathcal{L}_\beta(\pi,\theta) = \pi^\top \mathrm{D}\pi - 2\pi^\top \mathrm{b}. \tag{4}$$

For the NPLSE computation, we modify the CNM algorithm of Wang [16] which is proposed for computing the NPMLE of a mixing distribution. However, we omit the outline of the algorithm and direct the interested reader to Wang [16]. Basically, both algorithms have the same critical ingredients, i.e., updating the mixing proportions $\boldsymbol{\pi}$, and contracting and expanding the support set $\boldsymbol{\theta}$.

With known $\boldsymbol{\theta}$ and β, minimizing (4) with respect to $\boldsymbol{\pi}$ can be written as follows:

$$\begin{aligned} &\text{minimize} && \|\mathrm{R}\pi - \mathrm{d}\|_2^2 \\ &\text{subject to} && \pi^\top 1 = 1,\ \pi \geq 0, \end{aligned} \tag{5}$$

where $\mathbf{R}$ satisfying $\mathrm{D} = \mathrm{R}^\top\mathrm{R}$ is called the upper triangular Cholesky factor of $\mathbf{D}$, $\mathbf{d}$ is the solution of the lower triangular system $\mathrm{R}^\top\mathrm{d} = \mathrm{b}$, $\|\cdot\|_2$ is the L_2 norm, $0 = (0,\ldots,0)^\top$ and $1 = (1,\ldots,1)^\top$. In principle, $\mathbf{D}$ is a symmetric positive definite matrix as long as all elements of $\boldsymbol{\theta}$ are distinct but in practice, it can be singular. In actual implementation, we add a small positive value to all diagonal elements of $\mathbf{D}$ so that Cholesky factorization can be safely applied to decompose $\mathbf{D}$. Problem (5) can be solved numerically by the NNLS algorithm of Lawson and Hanson [7] after employing the method of Dax [3], which transforms it into the least squares problem with only non-negativity constraints:

$$\begin{aligned} &\text{minimize} && ||\mathrm{P}\tilde{\pi}||_2^2 + |\tilde{\pi}^\top 1 - 1|^2 \\ &\text{subject to} && \tilde{\pi} \geq 0, \end{aligned} \tag{6}$$

where $\mathrm{P} \equiv (\mathrm{r}^{(1)} - \mathrm{d},\ldots,\mathrm{r}^{(J)} - \mathrm{d})$ with $\mathrm{r}^{(j)}$ being the jth column of $\mathbf{R}$. It is established that if $\tilde{\boldsymbol{\pi}}$ solves problem (6), then $\tilde{\pi}/\tilde{\pi}^\top 1$ solves problem (5). We note that there is a little difference between the CNM algorithm for the NPMLE and the algorithm for the NPLSE. For the NPLSE, after updated $\boldsymbol{\pi}$, one does not need to perform a line search.

Those support points with zero masses after $\boldsymbol{\pi}$ being updated are discarded before the next iteration. To obtain new support points, we have to rely on the gradient function. Following the method of Wang [16], we add many good support points, i.e., the localminima of the gradient function, to the support set at each iteration by using the combined Newton-bisection method.

4.0 A SIMULATION EXPERIMENT

We carried out a simulation experiment to investigate the performance of both the KDE and MDE. Four normal mixture densities are considered in this simulation experiment. These densities taken from Marron and Wand [10] are named Gaussian, $\phi_1(x)$, skewed unimodal, $(1/5)\phi_1(x) + (1/5)\phi_{2/3}(x - 1/2) + (3/5)\phi_{5/9}(x - 13/12)$, bimodal, $(1/2)\phi_{2/3}(x-1) + (1/2)\phi_{2/3}(x+1)$ and skewed bimodal, $(3/4)\phi_1(x) + (1/4)\phi_{1/3}(x - 3/2)$.

The bandwidth parameter is crucial to the performance of the KDE and MDE. To select a bandwidth parameter λ, one could use the standard approach of cross-validation (CV). In the V-fold CV, the data set $x_1,\ldots,x_n$ is randomly split into V disjoint partitions, $P_1,\ldots,P_V$, that are roughly equal in size. The λ is chosen so as to minimize the following criterion:

$$\mathrm{CV}(\lambda) = \frac{1}{V}\sum_{v=1}^{V}\int\left\{\hat{f}_\lambda^{(v)}(x)\right\}^2 \mathrm{d}x - \frac{2}{V}\sum_{v=1}^{V}\frac{1}{n_v}\sum_{x_j\in P_v}\hat{f}_\lambda^{(v)}(x_j),$$

where n_v is the number of observations in P_v and $\hat{f}_\lambda^{(v)}$ is the model fitted to all observations but those in P_v. In this study, we use one run of 5-fold CV for the KDE and MDE. For the KDE (or MDE), the candidate density estimates are computed on a grid of values for h (β) ranging from $0.2h_s$ (0.2s) to $2.2h_s$ (1.2s) in steps of $0.05h_s$ (0.05s), where h_s and s denote the Silverman's rule-of-thumb bandwidth (Silverman [15]) and the sample standard deviation respectively.

To evaluate the accuracy of the density estimates, we examine the integrated squared error (ISE) given by

$$\mathrm{ISE}(f,\hat{f}) = \int\left\{f(x) - \hat{f}(x)\right\}^2 \mathrm{d}x,$$

where f is the true density and $\hat{f}$ a density estimate. For each simulated density, we compute the ISE values numerically for each combination of estimator and replication. Subsequently, the mean integrated squared error (MISE) is empirically estimated by the average of these ISE values.

The results of the simulation study based on 100 replications with sample sizes $n = 250$ and $n = 500$ are summarized in Table 1. Shown here are the empirical MISE values with their corresponding standard errors in parentheses. Also, we performed paired sign tests at 5% level of significance and those significant results are marked by an asterisk. On the whole, the MDE achieves better performance when the true densities are the Gaussian, skewed unimodal and bimodal densities. On the other hand, the KDE outperforms the MDE for the skewed bimodal density.

Table 1 Summary of the simulation results in terms of the empirical MISE (X 10^3)

Density Estimator	True Density			
	Gaussian	Skewed Unimodal	Bimodal	Skewed Bimodal
		$n = 250$		
KDE	5.01 (0.41)	6.86 (0.47)	5.61 (0.43)	7.17 (0.40)*
MDE	4.59 (0.52)*	5.94 (0.43)*	5.33 (0.37)	8.02 (0.46)
		$n = 500$		
KDE	2.35 (0.18)	3.40 (0.26)	3.15 (0.20)	3.72 (0.18)*
MDE	1.80 (0.20)*	3.14 (0.36)*	2.57 (0.19)*	4.15 (0.22)

5.0 CONCLUDING REMARKS

In this paper, the practical application of nonparametric mixture models, with a special emphasis on the least squares estimator, in density estimation is described. Specifically, we introduce the least squares mixture density estimator as an alternative to the kernel density estimator. A numerical study was performed to compare the performance of the MDE and KDE. The results show that the mixture method is capable of reducing MISE for some popular densities.

Apart from the improvement in estimation accuracy, the mixture method maintains advantages over the kernel method in terms of model flexibility and sparsity of an estimator (in terms of number of support points) in real applications. In addition, the MDE is easy and fast to implement. Actually, the current work is facilitated by a fast and stable algorithm for computing the mixing distribution estimate in a nonparametric mixture model. These promising benefits make the MDE much more attractive for practical utilization.

References

[1] Balabdaoui, F. and Wellner, J. A. 2010. Estimation of a k-monotone Density: Characterizations, Consistency and Minimax Lower Bounds. *Statistica Neerlandica*. 64(1): 45–70.

[2] Böhning, D. 2000. *Computer-Assisted Analysis of Mixtures and Applications: Meta-Analysis, Disease Mapping and Others.* London: Chapman and Hall.

[3] Dax, A. (1990) The Smallest Point of a Polytope. *Journal of Optimization Theory and Applications*. 64(2): 429–432.

[4] Eggermont, P. P. B. and LaRiccia, V. N. 2001 *Maximum Penalized Likelihood Estimation. Volume I: Density Estimation*. New York: Springer.

[5] Jones, M. C. and Henderson, D. A. 2009 Maximum Likelihood Kernel Density Estimation: On the Potential of Convolution Sieves. *Computational Statistics and Data Analysis*. 53(10): 3726–3733.

[6] Jones, M. C., Marron, J. S. and Sheather, S. J. 1996. A Brief Survey of Bandwidth Selection for Density Estimation. *Journal of the American Statistical Association*. 91(433): 401–407.

[7] Lawson, C. L. and Hanson, R. J. 1974. *Solving Least Squares Problems*. Englewood Cliffs: Prentice-Hall Inc.

[8] Lindsay, B. G. 1995. *Mixture Models: Theory, Geometry and Applications, vol. 5 of NSF-CBMS Regional Conference Series in Probability and Statistics*. Hayward: Institute of Mathematical Statistics.

[9] Loader, C. R. 1999. Bandwidth selection: Classical or plug-in? *The Annals of Statistics*. 27(2): 415–438.

[10] Marron, J. S. and Wand, M. P. 1992. Exact Mean Integrated Squared Error. *The Annals of Statistics*. 20(2): 712–736.

[11] Priebe, C. E. and Marchette, D. J. 2000. Alternating Kernel and Mixture Density Estimates. *Computational Statistics and Data Analysis*. 35(1): 43–65.

[12] Roeder, K. M. and Wasserman, L. A. 1997. Practical Bayesian Density Estimation Using Mixtures of Normals. *Journal of the American Statistical Association*. 92(439): 894–902.

[13] Scott, D. W. (2001) Parametric Statistical Modeling by Minimum Integrated Square Error. *Technometrics*. 43(3): 274–285.

[14] Scott, D. W. and Szewczyk, W. F. 2001. From Kernels to Mixtures. *Technometrics*. 43(3): 323–335.

[15] Silverman, B. W. 1986. *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

[16] Wang, Y. 2007. On fast Computation of the Non-parametric Maximum Likelihood Estimate of a Mixing Distribution. *Journal of the Royal Statistical Society: Series B* (*Statistical Methodology*). 69(2): 185–198.

[17] Wang, Y. and Chee, C.-S. 2012. Density Estimation Using Non-parametric and Semi-parametric Mixtures. *Statistical Modelling*. 12(1): 67–92.

[18] Yuan, M. 2009. State Price Density Estimation via Nonparametric Mixtures. *The Annals of Applied Statistics*. 3(3): 963–984.