APPLIED STOCHASTIC MODELS IN BUSINESS AND INDUSTRY
Appl. Stochastic Models Bus. Ind., 2006; **22**:533–546
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asmb.642

Quick multivariate kernel density estimation for massive data sets

K. F. Cheng[1], C. K. Chu[2,*,†] and Dennis K. J. Lin[3]

[1] *Institute of Statistics, National Central University, Chungli, Taiwan*
[2] *Department of Applied Mathematics, National Donghwa University, Hualien, Taiwan*
[3] *Department of Supply Chain and Information Systems, The Pennsylvania State University, University Park, PA 16802, U.S.A.*

SUMMARY

Massive data sets are becoming popular in this information era. Due to the limitation of computer memory space and the computing time, the kernel density estimation for massive data sets, although strongly demanding, is rather challenging. In this paper, we propose a quick algorithm for multivariate density estimation which is suitable for massive data sets. The term quick is referred to indicate the computing ease. Theoretical properties of the proposed algorithm are developed. Its empirical performance is demonstrated through a credit card example and numerous simulation studies. It is shown that in addition to its computational ease, the proposed algorithm is as good as the traditional methods (for the situations where these traditional methods are feasible). Copyright © 2006 John Wiley & Sons, Ltd.

Received 16 May 2005; Revised 5 May 2006; Accepted 21 July 2006

KEY WORDS: conditional density estimation; histogram; marginal density estimation

1. INTRODUCTION

Most statistical inferences heavily depend on distribution theory (the density function). A density can give an intuitive picture of such characteristics as skewness of the distribution or the number of modes. A further advantage of having an estimate of the density is ease of interpretation for non-statisticians. Many statisticians would explain the normal distribution by drawing the familiar bell-shaped curve rather than writing out the explicit formula for the density of the normal distribution. Manku *et al.* [1] note that it is common in the database field to keep summaries of the variables in the form of equi-depth histograms. Density estimation is an important and long-studied problem. One application, for example, is in the area of network routing. Network routing decisions, and hence quality of service for the network users (see Kesidis [2]), could be improved by having more accurate summaries of the distributions of the

*Correspondence to: C. K. Chu, National Donghwa University, Department of Applied Mathematics, Hualien, Taiwan.
† E-mail: chu@mail.ndhu.edu.tw

Copyright © 2006 John Wiley & Sons, Ltd.

historical traffic data, in particular the tails of these distributions. Another application is in the area of Markov chain Monte Carlo (MCMC) analysis where simulations routinely generate massive amounts of data. A further application is in the computation through simulation of critical values and percentile points of new statistics whose distributions are unknown (see Dunn [3]). However, creating and maintaining these histograms can be quite costly. This is especially true for massive data sets.

In the past decade, we have witnessed a revolution in information technology. Routine collection of systematically generated data is now commonplace. As a consequence, massive data sets are becoming more and more common in modern society. Databases with hundreds of fields, billions of records and terabytes of information are not unusual. They arise from sources as diverse as large call centres, internet traffic data, sales transactional records, or satellite feeds. Typical examples, as mentioned in Hand *et al.* [4], include, Barclaycard (U.K.) carries out 350 million transactions a year; Wal-mart makes over 7 billion transactions a year; and AT&T carries over 70 billion long distance calls annually. This phenomenon presents a clear need to be able to process the data accurately and efficiently so that current analyses may be performed before becoming inundated by a continually growing store of data. It becomes very challenging to extract useful features from a large data set because many statistics are difficult to compute by standard algorithms or statistical packages when the data set is too large to be stored in primary memory. As such, the classical methods for density estimation are not feasible for massive data sets. In this paper, we propose a quick, low-storage and efficient approach for multivariate density estimation.

This paper is organized as follows. In Section 2, we first review the ordinary kernel density estimation method and then formally introduce the quick kernel density estimation method. The term *quick* is used here to indicate the computational ease of the proposed method, in addition to its high efficiency (as compared to the ordinary kernel density estimation). Section 3 discusses the proposed methods for marginal and conditional kernel estimation. Asymptotic properties of the proposed method are discussed in Section 4. Section 5 gives a thorough simulation study, followed by a case study of a credit card company. Final conclusions are given in Section 6.

2. QUICK DENSITY ESTIMATION

Consider the data $X_i = (X_{i,1}, \ldots, X_{i,d})$, for $i = 1, \ldots, n$, are independent and identically distributed and have the density function f on R^d. The purpose of this study is to estimate the value of $f(b_J)$, for each $b_J \in [c_0, c_1]^d$. Here, c_0 and c_1 are taken such that all data points fall inside the region $[c_0, c_1]^d$, $b_J = (b_{j_1}, \ldots, b_{j_d})$ are equally spaced partition points of $[c_0, c_1]^d$, $J = (j_1, \ldots, j_d) \in Z^d$, $b_j = j \times b$, and $b\ (= b_n)$ tends to 0 as $n \to \infty$. Of course, the smaller the value of b, the finer the grid of b_J.

2.1. The ordinary kernel density estimators

Given the kernel function K as a probability density function supported on $[-1, 1]$ and the bandwidth $h = h_n$ tending to 0 as $n \to \infty$, the ordinary kernel density estimator $\hat{f}_{\mathrm{ORD}}(b_J)$ [5] for

Copyright © 2006 John Wiley & Sons, Ltd.

DOI: 10.1002/asmb

$f(b_J)$ is defined by

$$\hat{f}_{\mathrm{ORD}}(b_J) = n^{-1}h^{-d}\sum_{i=1}^{n} K_{J,i}$$

for each $b_J \in [c_0, c_1]^d$, where $K_{J,i} = \prod_{u=1}^{d} K\{(b_{j_u} - X_{i,u})/h\}$.

It is well known that, in the sense of having smaller asymptotic mean square error, the optimal kernel function K for $\hat{f}_{\mathrm{ORD}}$ is the Epanechnikov kernel $K(z) = (\frac{3}{4})(1 - z^2)I_{[-1,1]}(z)$ [6]. In practice, for automatically choosing the value of h, the least squares cross-validation criterion [7, 8] is one of the most popular methods. Given the value of b, it is designated to take the selected value $\hat{h}_{\mathrm{ORD}}$ of h as the minimizer of

$$\mathrm{LSCV}_{\mathrm{ORD}}(h) = b^d \sum_{J\in\Psi} \hat{f}_{\mathrm{ORD}}(b_J)^2 - 2n^{-1}\sum_{i=1}^{n} \hat{f}_{\mathrm{ORD},i}(b^*_{J,i})$$

over h. Here $\hat{f}_{\mathrm{ORD},i}$ is $\hat{f}_{\mathrm{ORD}}$ based on the sample with X_i deleted, $b^*_{J,i}$ is the partition point b_J closest to X_i, and Ψ is the collection of the values of the subindex J with $b_J \in [c_0, c_1]^d$.

To obtain the values of $\hat{f}_{\mathrm{ORD}}(b_J)$, for all $b_J \in [c_0, c_1]^d$, the number of values of $K(\cdot)$ needed to be computed is $(\beta^d + n) \times d \times n$. Here and throughout this paper, $\beta = [(c_1 - c_0)/b]$ and the notation $[u]$ denotes the integer part of u. The larger each value of β, d, and n, the more the computation time needed by $\hat{f}_{\mathrm{ORD}}$. For a massive data set, this is typically infeasible.

Take bivariate density estimation, as will be seen in Section 5.1, as a simple illustration. For the specific case, we have $\beta = 1601$, $d = 2$, $n = 10^6$, and 101 values of h. Given the Epanechnikov kernel, each value of $K(\cdot)$ is obtained by performing two subtractions and three multiplications. Using the personal computer for evaluating 1601 values of $\hat{f}_{\mathrm{ORD}}(b_J)$, as will be described in Section 5.1, the computation time for $1601 \times 2 \times 10^6$ values of $K(\cdot)$ is about 832 s. Thus, the computation time for finding $\hat{h}_{\mathrm{ORD}}$ with that data set is estimated to be $832 \times 1601 \times 101/(86\,400 \times 365) + 832 \times 10^6 \times 101/(1601 \times 86\,400 \times 365) = 5.93$ years, where the first quantity is the estimated computation time for $\sum_{J\in\Psi} \hat{f}_{\mathrm{ORD}}(b_J)^2$, and the second one is for $\sum_{i=1}^{n} \hat{f}_{\mathrm{ORD},i}(b^*_{J,i})$. This is clearly unacceptable.

2.2. Quick kernel density estimation

We now give the formulation of our proposed quick kernel density estimator. It is constructed in two stages. In the first stage, the original data X_i are transformed into the equally spaced pseudo-data $\hat{Y}(b_J)$ at $b_J \in R^d$. Given the kernel function W as a probability density function supported on $[-1, 1]$ and the bandwidth $g = rb/2$, where $r \geqslant 1$, $\hat{Y}(b_J)$ are defined by

$$\hat{Y}(b_J) = n^{-1}g^{-d}\sum_{i=1}^{n} W_{J,i}$$

where $W_{J,i} = \prod_{u=1}^{d} W\{(b_{j_u} - X_{i,u})/g\}$. If $r < 1$, then there might be some X_i not used to construct the pseudo-data $\hat{Y}(b_J)$.

In the second stage for producing our proposed estimator, the Nadaraya–Watson estimator [9, 10] is applied to smooth the equally spaced pseudo-data $\{b_J, \hat{Y}(b_J)\}$. Given both the kernel function K and the bandwidth h employed by $\hat{f}_{\mathrm{ORD}}$, our suggested quick estimator $\hat{f}_{\mathrm{QCK}}(b_J)$ for

Copyright © 2006 John Wiley & Sons, Ltd.

DOI: 10.1002/asmb

$f(b_J)$ is defined by

$$\hat{f}_{\mathrm{QCK}}(b_J) = \sum_{I \in \Psi_J} K_{J,I}\, \hat{Y}(b_I) \Big/ \sum_{I \in \Psi_J} K_{J,I}$$

for each $b_J \in [c_0, c_1]^d$. Here $K_{J,I} = \prod_{u=1}^{d} K\{(b_{j_u} - b_{i_u})/h\}$ and Ψ_J denote the collection of the values of the subindex $I = (i_1, \ldots, i_d) \in \Psi$ with $|b_{j_u} - b_{i_u}| \leqslant h$, for each $u = 1, \ldots, d$. By the formulation, the value of $\hat{f}_{\mathrm{QCK}}(b_J)$ converges to that of $\hat{f}_{\mathrm{ORD}}(b_J)$, for each b_J, as the value of b decreases. This is always true, regardless the size of the data set (either massive or non-massive).

As will be shown in Section 4, under some regularity conditions, both estimators $\hat{f}_{\mathrm{ORD}}$ and $\hat{f}_{\mathrm{QCK}}$ for f have the same asymptotic mean square error. Thus, in the sense of having smaller asymptotic mean square error, the optimal kernel function K for producing $\hat{f}_{\mathrm{QCK}}$ is the Epanechnikov kernel, the same as that for producing $\hat{f}_{\mathrm{ORD}}$. There are only two candidates for the optimal kernel W, that is, the uniform kernel $W(z) = (\frac{1}{2})I_{[-1,1]}(z)$ and the triangle kernel $W(z) = (1 - |z|)I_{[-1,1]}(z)$, in the class of two degree polynomials (see Remark 4.1). Also, the value of g has to be taken as $g = rb/2$ with r as a positive integer.

From the computational aspect, for producing $\hat{f}_{\mathrm{QCK}}$, we suggest using only W as the uniform kernel and taking the value of $g = b/2$. In this case, to compute the equally spaced pseudo-data $\{b_J, \hat{Y}(b_J)\}$, we only need to perform the numerical division $1 \times n \times d$ times and the subtraction $1 \times n \times d$ times, and the resulting pseudo-data $\{b_J, \hat{Y}(b_J)\}$ are the d-dimensional histogram at b_J. On the other hand, if W is taken as the triangle kernel, then we need to perform the numerical division $2 \times n \times d$ times and the subtraction $2 \times n \times d$ times to produce the pseudo-data $\{b_J, \hat{Y}(b_J)\}$. Thus, we do not suggest using the triangle kernel to generate $\{b_J, \hat{Y}(b_J)\}$.

After the pseudo-data $\{b_J, \hat{Y}(b_J)\}$ are obtained, given a value of h and a partition point b_J, in order to compute the value of $\hat{f}_{\mathrm{QCK}}(b_J)$, there are $(2[h/b] + 1)^d$ values of $K_{J,I} = \prod_{u=1}^{d} K\{(b_{j_u} - b_{i_u})/h\}$ that are needed to be calculated. Due to the fact that the partition points $\{b_J\}$ are equally spaced, to compute these values of $K_{J,I}$, we only need to compute $2[h/b] + 1$ values of $K(jb/h)$, for $j = 0, \pm 1, \ldots, \pm[h/b]$. Further, the same values of $K_{J,I}$ can be used to evaluate $\hat{f}_{\mathrm{QCK}}$ at all partition points b_J. Thus, to compute $\hat{f}_{\mathrm{QCK}}(b_J)$ at all b_J, there are only $2[h/b] + 1$ values of $K(\cdot)$ needed to be evaluated. On the other hand, to compute the values of $\hat{f}_{\mathrm{ORD}}(b_J)$ at all b_J, the total number of the values of $K(\cdot)$ needed to be computed is $\beta^d \times n \times d$. Comparing the two discussed estimators on computation efficiency, we conclude that, in practice, if $2[h/b] + 1 < \beta^d \times n \times d$, then $\hat{f}_{\mathrm{QCK}}$ has the computation advantage over $\hat{f}_{\mathrm{ORD}}$, whether the given data set is of massive or non-massive size. The larger each value of β, d, and n, or the smaller the value of $[h/b]$, the more significant the computation advantage of our proposed $\hat{f}_{\mathrm{QCK}}$ over $\hat{f}_{\mathrm{ORD}}$.

In practice, for choosing the value of h for constructing $\hat{f}_{\mathrm{QCK}}$, the idea of cross-validation [11] in the non-parametric regression field can be considered. Given the value of b and the equally spaced pseudo-data $\{b_J, \hat{Y}(b_J)\}$, the selected value of h is taken as the minimizer $\hat{h}_{\mathrm{QCK}}$ of

$$CV_{\mathrm{QCK}}(h) = \sum_{J \in \Psi} \{\hat{f}_{\mathrm{QCK},J}(b_J)^2 - \hat{Y}(b_J)\}^2$$

over h. Here $\hat{f}_{\mathrm{QCK},J}$ is $\hat{f}_{\mathrm{QCK}}$ with $\{b_J, \hat{Y}(b_J)\}$ deleted.

In the case of $d = 1$, there are some estimators related to our proposed $\hat{f}_{\mathrm{QCK}}$ in the literature. Considering computation efficiency, Fan and Marron [12] use $g = b$ and the triangle kernel to produce the pseudo-data $\{b_J, \hat{Y}(b_J)\}$, and then apply the local linear estimator [13, 14] to smooth the pseudo-data. Considering asymptotic properties, Wu and Chu [15] and Cheng [16]

Copyright © 2006 John Wiley & Sons, Ltd.

DOI: 10.1002/asmb

employ $g = b/2$ and the uniform kernel to produce the pseudo-data $\{b_J, \hat{Y}(b_J)\}$, and then apply, respectively, the Gasser–Müller estimator [17, 18] and the local linear estimator to smooth the pseudo-data. The Nadaraya–Watson, the Gasser–Müller, and the local linear estimators are popular in the smoothing field. For characteristics of the three estimators, see, for example, Wu and Chu [15] and Chu and Marron [19]. Note that the proposed $\hat{f}_{\mathrm{QCK}}$ has the same asymptotic mean square error as those related estimators. However, the proposed $\hat{f}_{\mathrm{QCK}}$ has the computation advantage over them, because each of the uniform kernel used in the first stage and the Nadaraya–Watson estimator used in the second stage requires only the very least computation burden. In this paper, we are able to further extend $\hat{f}_{\mathrm{QCK}}$ to the multivariate cases.

3. QUICK MARGINAL AND CONDITIONAL KERNEL DENSITY ESTIMATORS

Recall that $J = (j_1, \ldots, j_d) \in Z^d$ and $b_J = (b_{j_1}, \ldots, b_{j_d}) \in R^d$. Given positive integers q and k with $1 \leqslant q, k \leqslant d$ and $q + k \leqslant d$, set $b_{J,q} = (b_{j_1}, \ldots, b_{j_q})$, $b_{J,k,q} = (b_{j_{q+1}}, \ldots, b_{j_{q+k}})$, f_q as the density function of $(X_{i,1}, \ldots, X_{i,q})$, $f_{k|q}$ as the conditional density function of $(X_{i,q+1}, \ldots, X_{i,q+k})$ given a value of $(X_{i,1}, \ldots, X_{i,q})$, and $F_{1|q}$ as the conditional distribution function of $X_{i,q+1}$ given a value of $(X_{i,1}, \ldots, X_{i,q})$. Furthermore, given $p \in (0, 1)$ and $b_{J,q} \in [c_0, c_1]^q$, let $\zeta_{p,q}$ denote the conditional pth quantile of $F_{1|q}(x|b_{J,q})$, that is, the root of $F_{1|q}(x|b_{J,q}) = p$. Assuming that $f_q(b_{J,q}) > 0$ and the root is unique, one can estimate $f_q(b_{J,q})$, $f_{k|q}(b_{J,k,q}|b_{J,q})$, and $\zeta_{p,q}$ by using the values of $\hat{f}_{\mathrm{QCK}}(b_J)$.

The marginal kernel density estimate $\hat{f}_{\mathrm{QCK},q}(b_{J,q})$ for $f_q(b_{J,q})$ is then defined by

$$\hat{f}_{\mathrm{QCK},q}(b_{J,q}) = b^{d-q} \sum_{(j_{q+1},\ldots,j_d):(j_1,\ldots,j_d)\in\Psi} \hat{f}_{\mathrm{QCK}}(b_{j_1}, \ldots, b_{j_d})$$

for each $b_{J,q} \in [c_0, c_1]^q$. Given the values of $\hat{f}_{\mathrm{QCK}}(b_J)$, to obtain such $\hat{f}_{\mathrm{QCK},q}(b_{J,q})$ for all $b_{J,q} \in [c_0, c_1]^q$, the computation effort required is to perform numerical addition $\beta^{(d-q)(d-q+1)/2}$ times.

Set $\hat{f}_{\mathrm{QCK}}(b_{J,q})$ and $\hat{f}_{\mathrm{ORD}}(b_{J,q})$ as the quick and the ordinary kernel density estimators applied directly to the q-dimensional data $(X_{i,1}, \ldots, X_{i,q})$. Through a straightforward calculation, the value of $\hat{f}_{\mathrm{QCK},q}(b_{J,q})$ is the same as that of $\hat{f}_{\mathrm{QCK}}(b_{J,q})$. Combining the result with the discussions for both $\hat{f}_{\mathrm{QCK}}$ and $\hat{f}_{\mathrm{ORD}}$ given in Section 2, the estimator $\hat{f}_{\mathrm{QCK},q}(b_{J,q})$ for $f_q(b_{J,q})$ has the same asymptotic mean square error as $\hat{f}_{\mathrm{ORD}}(b_{J,q})$. But the former has the computation advantage, since it is produced from $\hat{f}_{\mathrm{QCK}}(b_J)$ by only performing numerical addition.

The conditional density estimate $\hat{f}_{\mathrm{QCK},k|q}(b_{J,k,q}|b_{J,q})$ for $f_{k|q}(b_{J,k,q}|b_{J,q})$ is defined by

$$\hat{f}_{\mathrm{QCK},k|q}(b_{J,k,q}|b_{J,q}) = \hat{f}_{\mathrm{QCK},q+k}(b_{J,q+k})/\hat{f}_{\mathrm{QCK},q}(b_{J,q})$$

for each $b_{J,k,q} \in [c_0, c_1]^k$. If $\hat{f}_{\mathrm{QCK},q}(b_{J,q}) = 0$, then take $\hat{f}_{\mathrm{QCK},k|q}(b_{J,k,q}|b_{J,q}) = 0$. To estimate the conditional pth quantile $\zeta_{p,q}$, set

$$\hat{F}_{\mathrm{QCK},1|q}(b_j|b_{J,q}) = b \sum_{i:i\leqslant j} \hat{f}_{\mathrm{QCK},1|q}(b_i|b_{J,q})$$

as an estimate of $F_{1|q}(b_j|b_{J,q})$. Our estimate $\hat{\zeta}_{\mathrm{QCK},p,q}$ for $\zeta_{p,q}$ is taken as the value of b_j such that the corresponding value $\hat{F}_{\mathrm{QCK},1|q}(b_j|b_{J,q})$ is the closest to p over the subindex j. Similar computation procedures can be applied to estimate the values of the marginal and the conditional density functions, and the conditional quantiles for any selected dimension among the d dimensions of the data.

Copyright © 2006 John Wiley & Sons, Ltd.

DOI: 10.1002/asmb

4. ASYMPTOTIC PROPERTIES

The asymptotic behaviours of our proposed quick kernel density estimators will be studied under the following assumptions. These assumptions are commonly made in essentially all classical density estimation literature [5].

(A1) The d-variate density function f is positive on R^d, and each of its second order partial derivatives is Lipschitz continuous on R^d.
(A2) The kernel functions K and W are Lipschitz continuous and symmetric probability density functions with support $[-1, 1]$.
(A3) The values of h and b satisfy $h > b$, and are selected on the interval $H_n = [\mu n^{-1+\delta}, \mu^{-1} n^{-\delta}]$. Here the positive constants μ and δ are arbitrarily small.
(A4) The total number of observations in this density estimation setting is n, with $n \to \infty$. The values of h and b satisfy $h^{-2}b \to 0$ and $n^{-1}h^{-d} \to 0$, as $n \to \infty$.

Theorem 4.1 gives the asymptotic bias and variance of $\hat{f}_{\mathrm{QCK},q}$, and those of $\hat{f}_{\mathrm{QCK},k|q}$ and $\hat{\zeta}_{\mathrm{QCK},p,q}$. The proof is similar to the one given in Wu and Chu [15] and thus omitted here (although it is available through the authors). To state Theorem 4.1, we introduce the following notation. Set $\kappa_S = \int_{-1}^{1} K(u)^2 \, \mathrm{d}u, \kappa_2 = \int_{-1}^{1} u^2 K(u) \, \mathrm{d}u, \beta_q = \sum_{i=1}^{q} f_{q,ii}, w_r = r^{-1} \sum_{i=-\rho}^{\rho} W * W(i/r)$, for $q = 1, \ldots, d$; where, $f_{q,ii}$ is the iith second order partial derivative of the density function f_q, ρ is the largest integer which is strictly less than $2r$, and the notation $*$ denotes convolution.

Theorem 4.1
Given the positive value r, if the assumptions given in Section 2 and (A1)–(A4) hold, then the asymptotic bias and variance of $\hat{f}_{\mathrm{QCK},q}$ and those of $\hat{f}_{\mathrm{QCK},k|q}$ and $\hat{\zeta}_{\mathrm{QCK},p,q}$ can be expressed, respectively, as

$$\mathrm{Bias}\{\hat{f}_{\mathrm{QCK},q}(b_{J,q})\} = \tfrac{1}{2} h^2 \kappa_2 \beta_q(b_{J,q})\{1 + o(1)\} \tag{1}$$

$$\mathrm{Var}\{\hat{f}_{\mathrm{QCK},q}(b_{J,q})\} = n^{-1} h^{-q} f_q(b_{J,q}) \kappa_S^q w_r^q \{1 + o(1)\} \tag{2}$$

$$\begin{aligned}\mathrm{Bias}\{\hat{f}_{\mathrm{QCK},k|q}(b_{J,k,q}|b_{J,q})\} = {} & \tfrac{1}{2} h^2 \kappa_2 \{f_q(b_{J,q})^{-1} \beta_{q+k}(b_{J,q+k}) \\ & - f_q(b_{J,q})^{-2} f_{q+k}(b_{J,q+k}) \beta_q(b_{J,q})\}\{1 + o(1)\}\end{aligned} \tag{3}$$

$$\mathrm{Var}\{\hat{f}_{\mathrm{QCK},k|q}(b_{J,k,q}|b_{J,q})\} = n^{-1} h^{-q-k} \kappa_S^{q+k} w_r^{q+k} f_{q+k}(b_{J,q+k}) f_q(b_{J,q})^{-2} \{1 + o(1)\} \tag{4}$$

$$\begin{aligned}\mathrm{Bias}(\hat{\zeta}_{\mathrm{QCK},p,q}) = {} & \tfrac{1}{2} h^2 \kappa_2 (F_{1|q}/f_{1|q})(\zeta_{p,q}|b_{J,q}) \Big\{ f_q(b_{J,q})^{-1} \beta_q(b_{J,q}) \\ & - \int_{-\infty}^{\zeta_{p,q}} \beta_{q+1}(b_{J,q}, u) \, \mathrm{d}u \Big/ \int_{-\infty}^{\zeta_{p,q}} f_{q+1}(b_{J,q}, u) \, \mathrm{d}u \Big\} \{1 + o(1)\}\end{aligned} \tag{5}$$

Copyright © 2006 John Wiley & Sons, Ltd.

DOI: 10.1002/asmb

$$\mathrm{Var}(\hat{\zeta}_{\mathrm{QCK},p,q}) = n^{-1}h^{-q}\kappa_S^q w_r^q (F_{1|q}/f_{1|q})^2(\zeta_{p,q}|b_{J,q}) f_q(b_{J,q})^{-1}\{1 - w_r + w_r \int_{\zeta_{p,q}}^{\infty} f_{q+1}(b_{J,q},u)\,\mathrm{d}u \Big/ \int_{-\infty}^{\zeta_{p,q}} f_{q+1}(b_{J,q},u)\,\mathrm{d}u\}\{1 + o(1)\} \quad (6)$$

for $0 < p < 1$, and $1 \leqslant q$, $k \leqslant d$ with $q + k \leqslant d$.

Remark 4.1
A comparison between $\hat{f}_{\mathrm{ORD}}$ and $\hat{f}_{\mathrm{QCK}}$ on their asymptotic mean square error can be drawn by comparing Equations (1)–(2) and Equations (9)–(10) of Silverman [5].

- Both estimators have the same asymptotic bias, but their asymptotic variances are not comparable in magnitude since it is not known whether $w_r > 1$ or $w_r < 1$. When $w_r = 1$, both estimators have the same asymptotic variance.
- If W is the uniform kernel for each positive integer r, or if W is the triangle kernel for each even positive integer r, then $w_r = 1$.
- If W is taken as the other kernel in the class of two degree polynomials satisfying the conditions in (A2), then, by numerical computation, we have $w_r > 1$, for each positive integer r.
- Using the Riemann sum approximation and the fact that $W * W$ is a probability density function, if $r \to \infty$, then $w_r \to 1$. Hence, both estimators could have the same asymptotic mean square error.

Remark 4.2
A brief summary is made below for the practical choice of the kernel functions W and K and that of the values of the smoothing parameters r, b, and h for constructing our quick estimators $\hat{f}_{\mathrm{QCK},q}$, $\hat{f}_{\mathrm{QCK},k|q}$, and $\hat{\zeta}_{\mathrm{QCK},p,q}$. By the results in Remark 4.1 and by the consideration of potential computational burden, we suggest only using W as the uniform kernel and the value of r as 1. Furthermore, we suggest using the value of $\hat{h}_{\mathrm{QCK}}$ for $\hat{f}_{\mathrm{QCK}}(b_J)$ to compute all these quick estimators.

5. EMPIRICAL STUDIES

In this section, we provide some empirical studies for the performance of the proposed method. We first show some simulation studies, followed by a case study from commercial banking data. The proposed method works very well in all cases.

5.1. Simulation study

The entire simulation study was performed via a personal computer under the operation system: Microsoft Windows (Second edition), Genuine Intel x86 Family 15 Model 0 Stepping 10 with 1.4 GHz and 100 MB RAM. The software GAUSS was employed for all computations; computer codes are available upon request through the authors. Both univariate and multivariate cases were studied.

Copyright © 2006 John Wiley & Sons, Ltd.

DOI: 10.1002/asmb

5.1.1. Univariate simulation. For the univariate cases, three density functions were considered: the standard normal $N(0,1)$; the mixture of two normals $\frac{9}{10}N(0,1)+\frac{1}{10}N(10,9)$; and the heavy tail Cauchy(0,1). The following steps were taken.

1. For each data set, a random sample of size $n=10^6$ were generated based on the pre-specified density function.
2. Given the value of $b=0.01$, equally spaced pseudo-data were produced by using the uniform kernel and the value $r=1$ on the equally spaced grid of 10 001 values of b_j in $[-50,50]$. The density values were also estimated on the same grid of b_j.
3. The cross-validated bandwidth $\hat{h}_{\mathrm{QCK}}$ was chosen for constructing $\hat{f}_{\mathrm{QCK}}(b_j)$, the values of $CV(h)$ were calculated on the equally spaced logarithmic grid of 101 values of h in [0.01, 0.5].
4. Likewise, the optimal value of bandwidth h_{ISE} was taken as the minimizer of integrated squared error ISE(h), defined by $\mathrm{ISE}(h)=\int_{-50}^{50}\{\hat{f}_{\mathrm{QCK}}(z)-f(z)\}^2\,\mathrm{d}z$. For each given value of h, the value of ISE(h) was approximated by $b^{-1}\sum_{j:b_j\in[-50,50]}\{\hat{f}_{\mathrm{QCK}}(b_j)-f(b_j)\}^2$.
5. After evaluation on the grid, the global minimizers h_{ISE} of ISE(h) and $\hat{h}_{\mathrm{QCK}}$ of $CV(h)$ were taken on the grid.
6. Repeat Steps 1–5 for 100 times. After the values of h_{ISE} for each of the 100 data sets were obtained, the sample average and standard deviation of the corresponding ISE(h_{ISE}) values were calculated. The former measures the best performance of $\hat{f}_{\mathrm{QCK}}$. While the sample average of ISE($\hat{h}_{\mathrm{QCK}}$) over the 100 data sets measures the performance of $\hat{f}_{\mathrm{QCK}}$ which can also be obtained in practice by using the cross-validated bandwidth. The simulation results for the univariate cases are summarized in Table I and Figure 1.

Displayed in Figure 1 are

(1a) Plot of the true $N(0,1)$ density function (bold-faced dashed curve) and five density estimates derived from five sets of the simulated data by $\hat{f}_{\mathrm{QCK}}$ using $\hat{h}_{\mathrm{QCK}}$ (solid curves).
(1b) Plot of the true $N(0,1)$ cumulative distribution function (bold-faced dashed curve) and five cumulative distribution estimates derived from the five data sets employed in (1a) by $\hat{F}_{\mathrm{QCK}}$ using $\hat{h}_{\mathrm{QCK}}$ (solid curves).
(1c) Plot of the true $N(0,1)$ quantile function (bold-faced dashed curve) and five quantile estimates derived from the five data sets employed in (1a) by $\hat{\zeta}_{\mathrm{QCK}}$ using $\hat{h}_{\mathrm{QCK}}$ (solid curves).

The same description given in (1a)–(1c) for the $N(0,1)$ density is then applied to the mixture normal density and displayed in (1d)–(1f), respectively; while the same description given in (1a)–(1c) for the $N(0,1)$ density is applied to Cauchy(0, 1) density and displayed in (1g)–(1i), respectively. Here, for a better visual comparison, the estimates in each (1a)–(1i) have been vertically shifted. It is seen that the proposed method works very well in all cases.

5.1.2. Multivariate simulation. Consider a bivariate normal density with each marginal density being $N(0,1)$ density. The correlation coefficients ρ between two component of the bivariate normal density were taken as $0, 0.3, -0.3, 0.9$, and -0.9. For each given value of ρ, one million random observations were generated. We then applied our procedure for density estimation. Finally, we compare our results with the 'true' underlying densities.

Copyright © 2006 John Wiley & Sons, Ltd.

DOI: 10.1002/asmb

Table I. Mean (standard deviation) of univariate simulation results.

	$N(0,1)$		$0.9N(0,1)+0.1N(10,9)$		Cauchy(0, 1)	
	h_{ISE}	h_{CV}	h_{ISE}	h_{CV}	h_{ISE}	h_{CV}
$f(x)$	0.1607	0.1374	0.1649	0.1465	0.1578	0.1377
	(0.0179)	(0.0284)	(0.0179)	(0.0269)	(0.0151)	(0.0181)
For integrated squared errors (multiplied by 10^{-6})						
$f(x)$	6.032	7.188	5.614	6.424	4.497	5.004
	(1.811)	(3.196)	(1.576)	(2.423)	(1.241)	(1.432)
$F(x)$	8.509	8.261	7.160	6.967	8.626	8.467
	(3.030)	(3.140)	(2.560)	(2.639)	(3.727)	(3.720)
$F^{-1}(x)$	32.50	30.82	50.11	48.67	2231.00	2228.00
	(10.36)	(11.24)	(19.73)	(20.29)	(1656.00)	(1655.00)
Running time (s)	277.6		287.2		286.2	

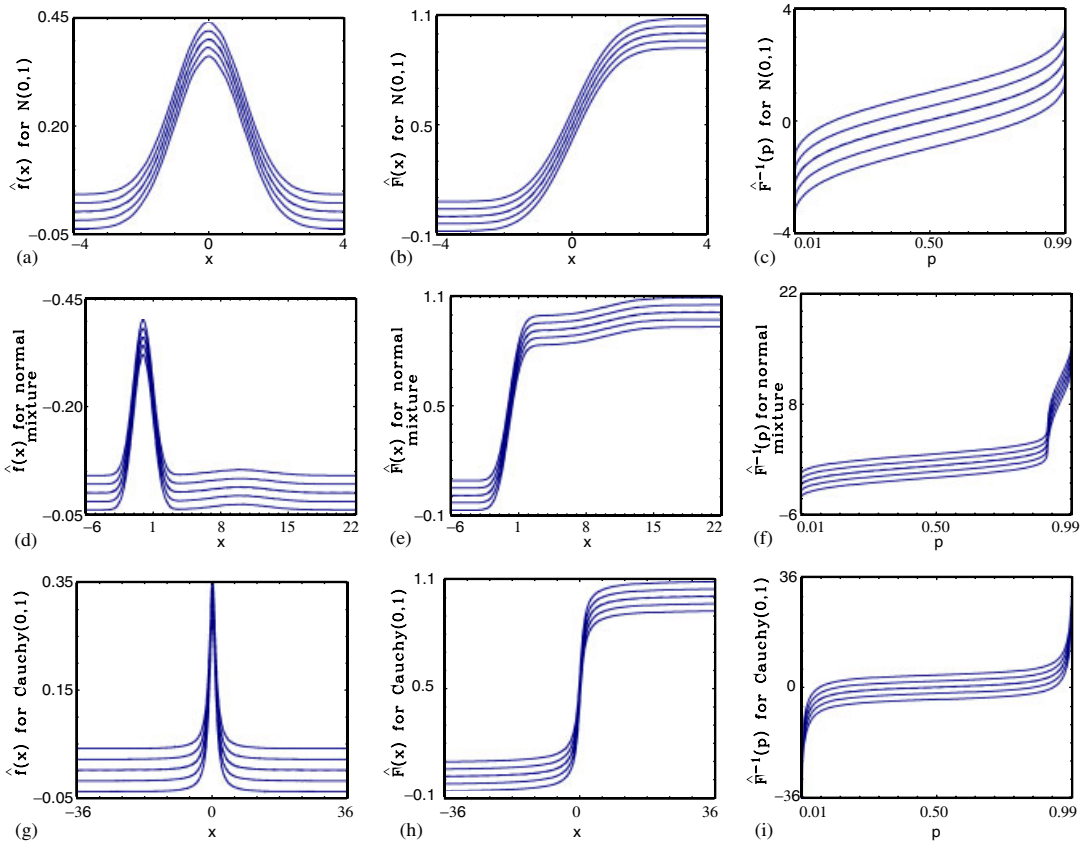

Figure 1. Simulation results: univariate cases.

Copyright © 2006 John Wiley & Sons, Ltd.

DOI: 10.1002/asmb

Specifically, for each bivariate data set, the density function values were computed on the equally spaced grid of 1601×1601 values of (b_i, b_j) in $[-8, 8]^2$. The same grid of the values of h for the univariate data set was also employed by the bivariate data set. The same computation procedures used in the univariate case were also employed in the bivariate case. Note that there is no way to compute $\hat{f}_{\mathrm{ORD}}$, since it costs too much computation time. The simulation results are summarized in Figure 2. Displayed in Figure 2 are

(2a) Plot of conditional density for the bivariate density with $\rho = 0$.
(2b) Conditional distribution for the bivariate density with $\rho = 0$.
(2c) Conditional quantile estimates for the bivariate density with $\rho = 0$.

The same descriptions given in Figures 2(a)–(c) with $\rho = 0$ when applied to $\rho = 0.3$ are displayed in Figures 2(d)–(f); while the same descriptions given in Figures 2(a)–(c) applied to $\rho = 0.9$ are displayed in Figures 2(g)–(i). Note that for having better visual performance, all estimates in Figures 2(a)–(i) were vertically shifted.

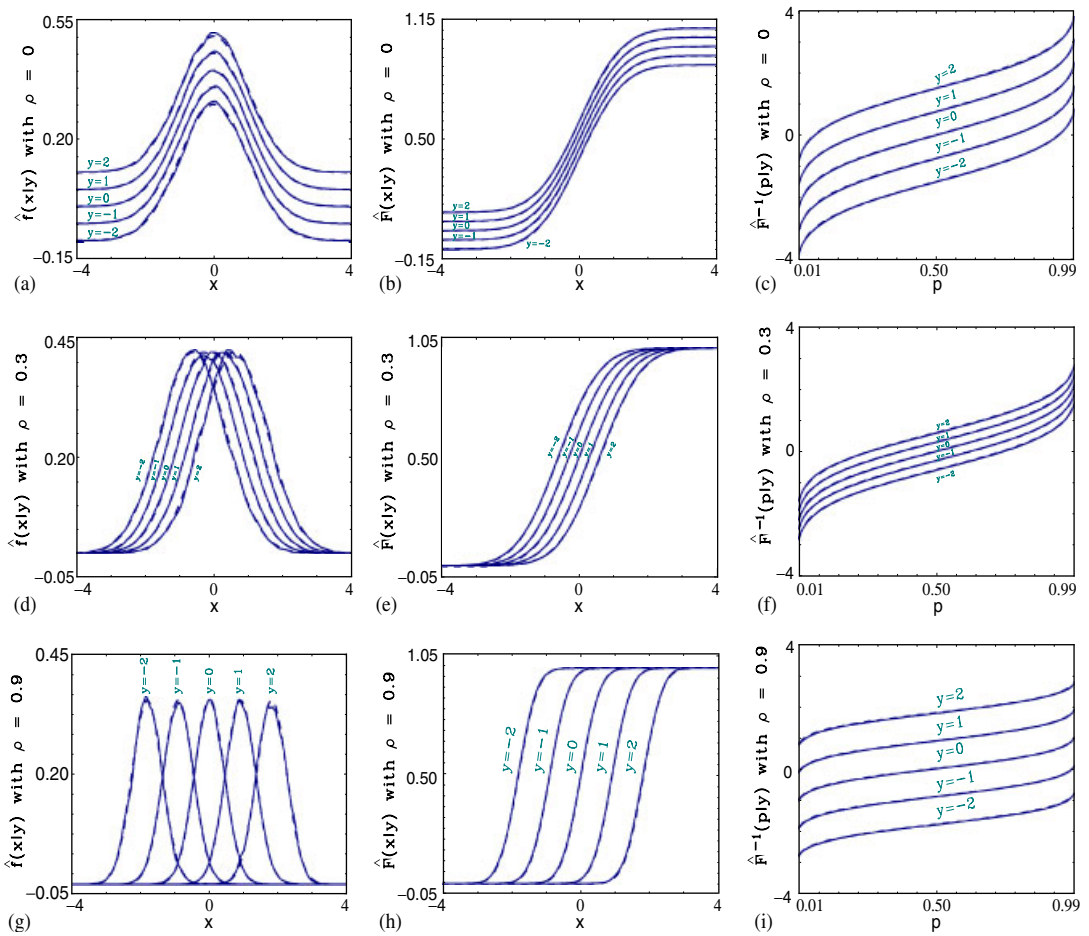

Figure 2. Simulation results: multivariate cases.

Copyright © 2006 John Wiley & Sons, Ltd.

DOI: 10.1002/asmb

5.2. *Credit card example*

A commercial bank in Taiwan kindly offered the data set from one of its branches, which contains $n = 739\,654$ records with more than 30 variables. Among them, four major variables are of particular interest: age (x_1); income (x_2); expenditure (x_3); and credit (x_4). The proposed procedure described in Section 2 is first used to estimate the (four-dimensional) joint density distribution function. While the functional form of the density is rather complicated, some graphical presentations are possible. Figure 3 shows all the pair-wise joint density estimates using the cross-validated bandwidth; while Figure 4 shows all the (four) marginal distributions as well as their corresponding conditional distributions. Figure 3 indicates some potential interaction effects among all variables. This is helpful, as will be discussed below. Note that in Figures 3 and 4, the measurement unit for the income variable is 10^3, for the expenditure variable is 10^3, and for the credit variable is 10^4.

Figure 4(a) is the marginal distribution of age; its three quartiles (25, 50, and 75 percentile) are marked, respectively, by dashed, long-dashed, and solid vertical lines. Figure 4(b) is the conditional distribution of income, given that age is at its 25, 50, and 75 percentile values. Again this is indicted via dashed, long-dashed, and solid vertical lines, respectively. Figure 4(c) is the conditional distribution of expenditure, given that age is at its 25, 50, and 75 percentile values;

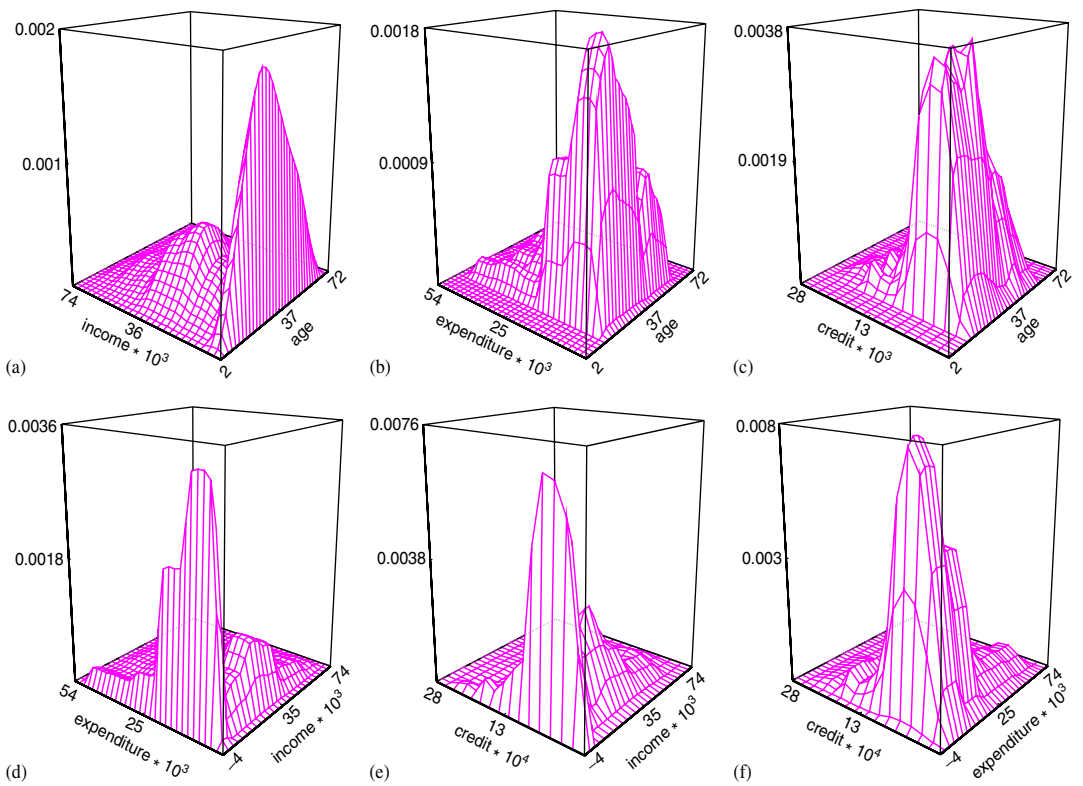

Figure 3. Credit card example: pair-wise joint density estimates.

Copyright © 2006 John Wiley & Sons, Ltd.

DOI: 10.1002/asmb

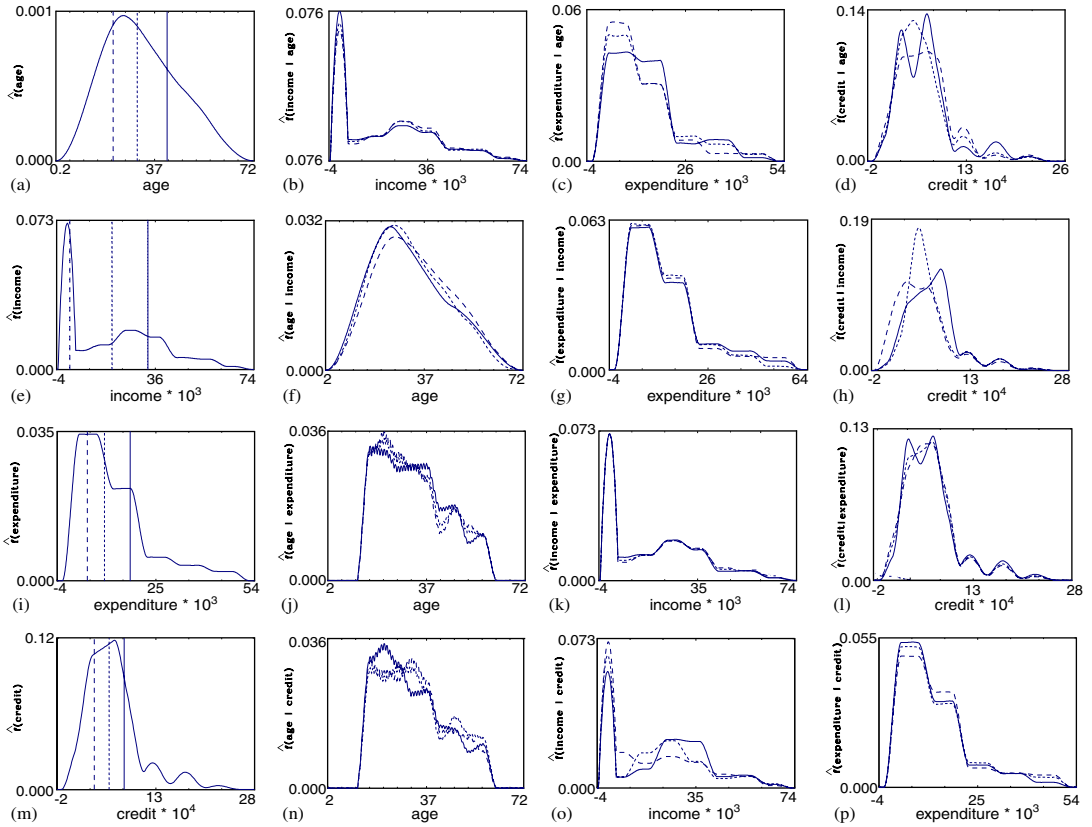

Figure 4. Credit card example: marginal and conditional distributions.

while Figure 4(d) is the conditional distribution of credit, given that age is at its 25, 50, and 75 percentile values. These are, of course, examples for illustration. Once the joint density function is obtained, we are able to (easily) estimate conditional distribution for any percentile. Similar conditional distribution plots were made for income, expenditure and credit as the conditional variables; Figures 4(e)–(h) for income, Figures 4(i)–(l) for expenditure, and Figure 4(m)–(p) for credit, respectively.

The marginal distribution, as in Figures 4(a), (e), (i) and (m), provides a basic understanding of each variable. For example, it is clear that age is near symmetric, while other variables are rather skew, especially income. The high percentage in low income indicates some potential risk issues for the policy of increasing the credit limits, and this is reflected in the current low credit policy, as seen in Figure 4(m).

The conditional plots are also useful. Take Figures 4(b) and (d) as examples: Figure 4(b) indicates that the distribution of income is somewhat irrelevant with age (all three curves are nearly identical); while Figure 4(d) indicates that the distributions of credit for young, middle, and old age group are quite different.

Furthermore, if the variable credit is under concern, we will focus on Figures 4(m), (d), (h) and (l). The marginal distribution in Figure 4(m) shows a high portion of low credit and

Copyright © 2006 John Wiley & Sons, Ltd.

DOI: 10.1002/asmb

relatively small portion for high credit, meaning a new policy on increasing the credit limits should be considered. Conditional on credit, the conditional plot for other variables are displayed in Figures 4(d), (h) and (l). It is shown that expenditure is somewhat irrelevant to credit (all three curves are nearly identical), while age and income have interaction effects with credit, especially for lower credit. Other observations can be made in a similar manner.

A discussion with the banking experts on these plots leads to the following conclusions (some other conclusions are confidential and thus are omitted here). We identify three major targeted groups and impose new policies for each group. These three main targeted groups are: (1) those with young age, high income, middle expenditure and middle credit; (2) those with low middle income, middle expenditure and low credit (age is insignificant here); and (3) those with middle (to old) age, high income, middle expenditure and middle credit. Policy for group (1), for example, is to increase their credit limits, provide more information and opportunities on products, such as coupons, discount or free gifts. Policies for the other two groups were also suggested accordingly.

6. CONCLUSION

While massive data sets are becoming popular, the conventional wisdom for density estimation methods do not seem to be capable, mainly due to the limit of computing memory space and computing times. A quick density estimation, based on the histogram approach, is proposed here. It is shown, theoretically and empirically, that such a density estimation method works well for both univariate and multivariate cases, and is extremely inexpensive, in terms of the computational costs. A successful case study on banking data indicates the power of density estimation. We anticipate more applications of this kind in the near future, especially for multivariate massive data sets type problems.

REFERENCES

1. Manku GS, Rajagopalan S, Lindsay BG. Approximate medians and other quantiles in one pass and with limited memory. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Seattle, Washington, U.S.A. June 1988; 426–435.
2. Kesidis G. Bandwidth adjustments using on-line packet-level adjustments. *SPIE Conference on Performance and Control of Network Systems*, Boston, 19–22 September 1999.
3. Dunn CL. Precise simulated percentiles in a pinch. *The American Statistician* 1991; **45**(3):207–211.
4. Hand DJ, Blunt G, Kelly MG, Adams NM. Data mining for fun and profit. *Statistical Sciences* 2000; **15**:111–131.
5. Silverman BW. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall: New York, 1986.
6. Epanechnikov VA. Nonparametric estimation of a multivariate probability density. *Theory of Probability and its Applications* 1969; **14**:153–158.
7. Rudemo M. Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics* 1982; **9**:65–78.
8. Bowman AW. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* 1984; **71**:353–360.
9. Nadaraya EA. On estimating regression. *Theory of Probability and its Applications* 1964; **9**:141–142.
10. Watson GS. Smooth regression analysis. *Sankhya A* 1964; **26**:359–372.
11. Härdle W. *Applied Nonparametric Regression*. Cambridge University Press: Cambridge, 1990.
12. Fan J, Marron JS. Fast implementations of nonparametric curve estimators. *Journal of Computational and Graphical Statistics* 1994; **3**:35–56.
13. Fan J. Design-adaptive nonparametric regression. *Journal of the American Statistical Association* 1992; **87**:998–1004.
14. Fan J. Local linear regression smoothers and their minimax efficiencies. *Annals of Statistics* 1993; **21**:196–216.
15. Wu JS, Chu CK. Double smoothing for kernel estimators in nonparametric regression. *Journal of Nonparametric Statistics* 1992; **1**:375–386.

Copyright © 2006 John Wiley & Sons, Ltd.

DOI: 10.1002/asmb

16. Cheng MY. A bandwidth selector for local linear density estimators. *Annals of Statistics* 1997; **25**:1001–1013.
17. Gasser T, Müller HG. Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation*, Gasser T, Rosenblatt M (eds), Lecture Notes in Mathematics, vol. 757. Springer: Berlin, 1979; 23–28.
18. Gasser T, Müller HG. Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics* 1984; **11**:171–185.
19. Chu CK, Marron JS. Choosing a kernel regression estimator (with discussion). *Statistical Science* 1991; **6**:404–436.

Copyright © 2006 John Wiley & Sons, Ltd.

DOI: 10.1002/asmb