A Log-Density Estimation Methodology Applicable to Massive Bivariate Data

Hossein Zareamoghaddam[1], Serge B. Provost[1] & S. Ejaz Ahmed[2]

[1] Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Canada

[2] Department of Mathematics and Statistics, Brock University, St. Catharines, Canada

Correspondence: Hossein Zareamoghaddam, Department of Statistical and Actuarial Sciences, University of Western Ontario, London, ON., N6A 5B7, Canada. E-mail: hzareamo@uwo.ca

Abstract

First, it is shown that a univariate *bona fide* density approximation can be obtained by assuming that the derivative of the logarithm of the density function under consideration is expressible as a rational function or a polynomial. Then, the density function of a bivariate continuous random vector is approximated by standardizing it and applying a polynomial adjustment to the product of the density approximants of the marginal distributions. As well, it is explained that this approach can easily be extended to the estimation of density functions. For illustrative purposes, the proposed methodology is applied to several datasets. Since this technique is solely based on sample moments, it readily lends itself to the modeling of large datasets.

Keywords: log-density, density approximation, joint moments, bivariate density estimation

1. Introduction

We initially consider the problem of approximating the density function of a continuous random variable. Obtaining an accurate density approximation can prove useful when the exact density function of a statistical quantity such as an estimator or a test statistic may not be tractable or have a simple closed form. The flexible methodology that is proposed relies on the moments of the target distribution and can even be utilized to approximate irregular or multimodal density functions.

There exist several types of density estimates and approximants. However, many of these techniques will fail to provide adequate approximations, especially when the target density is not a smooth unimodal function. Silverman (1986) provides a survey of the various available methodologies while (Reid, 1988) focuses on the saddlepoint approximation. Moment-based techniques are described for instance in (Elderton & Johnson, 1969), (Solomon & Stephens, 1978) and (Provost, 2005). Efromovich (1999) presents a unified account of nonparametric approaches to density estimation. Other types of nonparametric density estimates that are based on the L_1 norm are presented in (Devroye, 1985) while both parametric and nonparametric approaches are discussed in (Eggermont, 2001). The multivariate case is extensively treated in (Scott, 2015).

The bivariate density estimation methodology that is introduced in this paper relies on a univariate density approximation technique that produces differentiated log-density approximants (DLDA's) whereby the derivative of the logarithm of a density function is assumed to be expressible as a rational function. This approach only necessitates the moments of a distribution up to some particular order; accordingly, when used in conjunction with sample moments, it enables one to process large amounts of data that often arrive in streams without having to access previously collected observations. Upon solving a system of linear equations, the coefficients of the rational function can easily be determined, the density approximant being then obtained by solving a differential equation. This density estimation technique is then applied to each of the marginal distributions of a standardized bivariate sample; the product of the resulting density estimates serves as a base density that is adjusted by means of a bivariate polynomial whose coefficients are determined from the joint sample moments of the standardized dataset being modeled as well as those associated with the base density function. The resulting expressions assume relatively simple functional representations that can lend themselves to algebraic manipulations; this is not the case for kernel density estimates, which incidentally may not be as accurate, as is suggested by a numerical example (Example 3.1) involving a sample of simulated values.

This paper is organized as follows. First, the technique being utilized for obtaining univariate DLDA's is developed in Section 2, including the special case where the derivative of the logarithm of the target density function is assumed to be a polynomial. The bivariate case is then considered in Section 3 where DLDA's are utilized to obtain approximants

or estimates of the marginal density functions whose product is adjusted by a bivariate polynomial. To illustrate the applicability of the new methodology, several numerical examples are presented in Section 4.

2. Differentiated Log-Density Approximation

This section summarizes the results obtained in (Provost & Ha, 2015) wherein a novel technique for approximating continuous univariate density functions is introduced. This approach will be utilized in the next section to approximate the density functions associated with each of the marginal distributions of a standardized bivariate random vector.

We now explain how differentiated log-density approximants are determined. Let $f_X(x)$ be a continuous density function defined on the interval $(\alpha, \beta) \equiv \mathscr{S}$. It is assumed that the derivative of logarithm of $f_X(x)$ can be represented by a rational function, that is,

$$\frac{\mathrm{d}}{\mathrm{d}x} ln(f_X(x)) = \frac{f'_X(x)}{f_X(x)} = r(x), \tag{1}$$

where

$$r(x) = \frac{\sum_{i=0}^{\gamma} a_i x^i}{\sum_{j=0}^{\delta} c_j x^j} = \frac{N_\nu(x)}{D_\delta(x)}, \tag{2}$$

$N_\nu(x)$ and $D_\delta(x)$ being polynomials in x of orders ν and δ. Without any loss of generality, c_δ, the coefficient of x^δ in the denominator of $r(x)$, is set equal to one. After determining the a_i's and c_j's, by solving a linear system involving a certain number of moments of the target distribution, $f_X(x)$ is approximated as

$$f_{\nu,\delta}(x) = \kappa\, e^{\int_\alpha^x r(y)\,\mathrm{d}y},$$

where κ is the normalizing constant, which is such that the integral of $f_{\nu,\delta}(x)$ from α to β numerically integrates to one, and $e^{\int_\alpha^x r(y)\,\mathrm{d}y}$ is the solution of the differential equation specified by (6).

In light of Equations (1) to (2), one has

$$f'_{\nu,\delta}(x) \sum_{j=0}^{\delta} c_j x^j = f_{\nu,\delta}(x) \sum_{i=0}^{\nu} a_i x^i, \tag{3}$$

from which the polynomial coefficients can be obtained as follows: Multiplying both sides of Equation (3) by x^h and integrating over the interval (α, β) yields

$$\int_\alpha^\beta f'_{\nu,\delta}(x) \sum_{j=0}^{\delta} c_j x^{j+h}\mathrm{d}x = \int_\alpha^\beta f_{\nu,\delta}(x) \sum_{i=0}^{\nu} a_i x^{i+h}\mathrm{d}x, \ h = 0, 1, \ldots, \nu + \delta.$$

On interchanging the sum and the integral on each side of this equation and proceeding by parts for integrating the left-hand side, one has

$$f_{\nu,\delta}(x) \sum_{j=0}^{\delta} c_j x^{j+h} \Big|_\alpha^\beta - \sum_{j=0}^{\delta} c_j(j+h) \int_\alpha^\beta x^{j+h-1} f'_{\nu,\delta}(x)\mathrm{d}x = \sum_{i=0}^{\nu} a_i \int_\alpha^\beta x^{i+h} f_{\nu,\delta}(x)\,\mathrm{d}x, \ h = 0, 1, \ldots, \nu + \delta.$$

Note that the first term on the left-hand side, that is, $f_{\nu,\delta}(x) \sum_{j=0}^{\delta} c_j x^{j+h} \big|_\alpha^\beta$, will be zero whenever $f_{\nu,\delta}(\alpha) = f_{\nu,\delta}(\beta) = 0$, which is the case for most densities of interest. Thus, omitting this term and letting μ_h, $h = 0, 1, \ldots, \nu + \delta$, denote the hth moment of the approximated density function $f_{\nu,\delta}(x)$, one obtains $\nu + \delta + 1$ linear equations having the following form:

$$-\sum_{j=0}^{\delta} c_j(j+h)\,\mu(j+h-1) = \sum_{i=0}^{\nu} a_i\,\mu(i+h), \ h = 0, 1, \ldots, \nu + \delta, \tag{4}$$

with $\mu(0) \equiv 1$. In order to determine the unknown coefficients of $r(x)$ as specified by Equation (2), one needs to solve the linear system resulting from Equation (4). On replacing the unknown $\mu(h)$ by $\mu_X(h)$, for $h = 0, 1, \ldots, \nu + \delta$, where $\mu_X(h)$ denotes the hth moment of the distribution being approximated, one obviously obtains the following linear system:

$$-\sum_{j=0}^{\delta} c_j(j+h)\,\mu_X(j+h-1) = \sum_{i=0}^{\nu} a_i\,\mu_X(i+h), \ h = 0, 1, \ldots, \nu + \delta. \tag{5}$$

Once the solution of this linear system is obtained, one still has to solve the differential equation

$$f'_{\nu,\delta}(x) = r(x)f_{\nu,\delta}(x), \tag{6}$$

where $r(x) = \sum_{i=0}^{\nu} a_i x^i / \sum_{j=0}^{\delta} c_j x^j$, which can easily be achieved by making use of symbolic computation packages such as *Mathematica* or *Maple*.

Remark 1 The degree δ is set to be the number of times the density function (or its components in the case of mixtures) intersects the abscissa plus the number of points at which the density function is not differentiable, the roots of $D_\delta(x)$ corresponding to the intersection points and the points of non-differentiability as the case may be. For instance, in the case of a triangular distribution, one would let $\delta = 3$, and for the mixture of density functions described in Example 2.1, δ was set equal to 4.

For a given δ, let the integrated squared difference (or error) between the approximate density function $f_{\nu,\delta}(x)$ and the exact density function $f_X(x)$ over the support of the distribution be denoted by

$$\mathrm{ISD}(\nu) = \int_{\mathscr{S}} \left(f_{\nu,\delta}(x) - f_X(x) \right)^2 \mathrm{d}x. \tag{7}$$

Remark 2 In order to quantify the discrepancy between the approximate density $f_{\nu,\delta}(x)$ and the target density $f_X(x)$ and to determine the optimal order of the numerator of $r(x)$, we seek the value ν_0 such that $\mathrm{ISD}(\nu_0)$ reaches a set tolerance level or $\mathrm{ISD}(\nu)$ only decreases marginally beyond ν_0.

The following algorithm summarizes the DLDA procedure for approximating a univariate continuous density function $f_X(x)$.

Algorithm Differentiated log-density approximation methodology

1. Let $\nu = 0$ be the initial order of $N_\nu(x)$ as specified in Equation (2) and δ, the order of $D_\delta(x)$, be selected as per Remark 1. (It should be noted that, in most cases of interest, ν is greater than or equal to two.)

2. Evaluate the moments of the random variable X, that is, $\mu_X(i)$ for $i = 0, 1, \ldots, r$, where $r = 2\nu + \delta$ if $\delta \leq \nu$ and $r = 2\delta + \nu - 1$ if $\delta > \nu$. (These moments replace those associated with the approximated distribution appearing in Equation (4).)

3. Determine the coefficients of the rational function by solving the linear system (5).

4. Find the solution of the differential equation specified by (6) by making use of a symbolic computation package and normalize the resulting function to obtain a *bona fide* density function $f_{\nu,\delta}(x)$.

5. Evaluate $\mathrm{ISD}(\nu)$ as defined in Equation (7).

6. Repeat Steps 2-5 with larger values of ν until $\mathrm{ISD}(\nu)$ is deemed to be sufficiently small as per Remark 2.

Example 2.1 Suppose that $f_X(x)$ is the univariate density function of a mixture of two equally weighted beta distributions with parameters (2, 20) and (3, 2). In this example, we set $\nu = 4$ and $\delta = 4$. The plots of the exact and approximate density functions are shown in Figure 1. In this case, after rounding to three decimals, the coefficients are $a_0 = 0.031$, $a_1 = -0.709$, $a_2 = 3.070$, $a_3 = -3.210$, $a_4 = 0.147$, $c_0 = 0$, $c_1 = -0.040$, $c_2 = 0.407$, $c_3 = -1.367$, $c_4 = 1$ and the density approximant is

$$f_{4,4}(x) = x^{0.770}(1 - x)^{0.997}(0.040 - 0.367x + x^2)^{0.621} e^{-0.147+1.551\, arctan(2.337-12.731\, x)}/0.036.$$

2.1 Polynomial Log-density Approximants

As a particular case, one may assume that the differentiated log-density function is a polynomial of order n, that is,

$$\frac{\mathrm{d}}{\mathrm{d}x} ln(f_X(x)) = \sum_{i=0}^{n} a_i x^i,$$

in which case $c_0 = 1$ and the other c_i's are equal to zero in Equation (2). This gives rise to the following linear system:

$$\begin{pmatrix} \mu_X(0) & \ldots & \mu_X(n) \\ \mu_X(1) & \ldots & \mu_X(n+1) \\ \vdots & & \vdots \\ \mu_X(n) & \ldots & \mu_X(2n) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} 0 \\ -\mu_X(1) \\ \vdots \\ -n\mu_X(n-1) \end{pmatrix}, \tag{8}$$
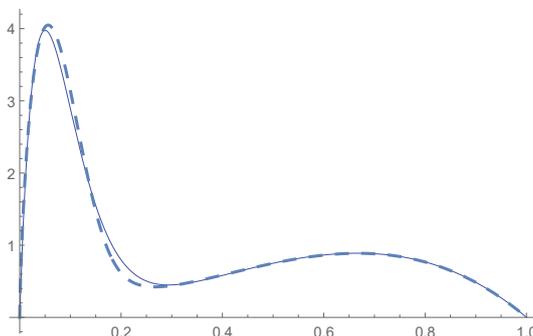
Figure 1. Exact (solid line) and approximated (dashed line) density functions in connection with Example 2.1.

which, in matrix form, can be expressed as $M\mathbf{a} = \tau$ where $\mathbf{a}$ is the vector of unknown coefficients. We now show that M is a positive definite matrix. Suppose that $\mathbf{z}$ is an arbitrary non-null vector of $\mathfrak{R}^n$. Then,

$$\mathbf{z}^T M \mathbf{z} = \sum_{i=0}^{n} \sum_{j=0}^{n} z_i z_j \int_{\alpha}^{\beta} x^{i+j} f_X(x)\, \mathrm{d}x$$

$$= \int_{\alpha}^{\beta} \Big(\sum_{i=0}^{n} z_i x^i \Big) \Big(\sum_{j=0}^{n} z_j x^j \Big) f_X(x)\, \mathrm{d}x$$

$$= \int_{\alpha}^{\beta} \Big(\sum_{i=0}^{n} z_i x^i \Big)^2 f_X(x)\, \mathrm{d}x > 0.$$

Thus, the linear system specified by Equation (8) has the unique solution $M^{-1}\tau$. The resulting density approximant, which shall be referred to as a Polynomial Log-density Approximant (PLDA), will then have the following representation:

$$f_{X_n}(x) = \kappa\, e^{\sum_{i=0}^{n} a_i x^{i+1}/(i+1)} \tag{9}$$

where κ the normalizing constant, is determined by numerical integration.

Remark 3 The DLDA (PLDA) methodology can be applied in the context of density estimation by replacing the exact moments of the target distribution by the sample moments associated with a given dataset. In this case, the degree of $N_\nu(x)$ is determined in terms of the sum of the squared differences between empirical distribution function and the estimated CDF obtained from $f_{\nu,\delta}(x)$, that is, $\mathrm{SSD}(\nu) = \sum_{i=1}^{n} (\mathrm{ECDF}(x_i) - F_{\nu,\delta}(x_i))^2$. One could select the degree ν for which $\mathrm{SSD}(\nu)$ reaches a minimum value or beyond which $\mathrm{SSD}(\cdot)$ does not decrease significantly. A suitable degree for $D_\delta(x)$ can be determined by following the guidelines provided in Remark 1 on the basis of a preliminary density estimate such as a histogram.

3. Bivariate Density Estimation

In this section, the DLDA methodology, as described in the previous section, is initially utilized to approximate each of the marginal density functions of a standardized bivariate random vector $(X, Y)'$. A bivariate polynomial adjustment is then applied to the product of the marginal density approximants to produce a bivariate density approximation. As well, it is explained that the proposed bivariate density approximation methodology can be utilized in the context of density estimation by substituting joint sample moments of given orders to the corresponding exact joint moments of a target distribution.

When X and Y are independent random variables, their joint density function can be expressed as the product of the marginal density functions, that is, $f_{X,Y}(x, y) = f_X(x)f_Y(y)$. However, in general the variables forming a random vector are not independently distributed even after standardizing it, and some adjustment to the product of the *approximate* or *estimated* marginal density functions is needed. We are proposing to apply a bivariate polynomial adjustment to the standardized vectors, which yields a density of the form specified in Equation (10). The density approximant/estimate corresponding to the original bivariate distribution/data is then obtained by applying the inverse transformation.

Now, letting (w_i, z_i), $i = 1, \ldots, n$, constitute a dataset with sample mean $(\bar{w}, \bar{z})$, an estimate of the covariance matrix V is required in order to standardize these n observation vectors. Let this estimate be the m.l.e. of V, that is, $\hat{V} = \{v_{ij}\}$, where $v_{11} = 1/n[\sum_{i=1}^{n} (w_i - \bar{w})^2]$, $v_{12} = v_{21} = 1/n[\sum_{i=1}^{n} (w_i - \bar{w})(z_i - \bar{z})]$ and $v_{22} = 1/n[\sum_{i=1}^{n} (z_i - \bar{z})^2]$. The standardized data is

then obtained as

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} = \hat{V}^{-1/2} \begin{pmatrix} w_i - \bar{w} \\ z_i - \bar{z} \end{pmatrix},$$

$\hat{V}^{-1/2}$ denoting the inverse of the symmetric square root of $\hat{V}$. The x_i's and the y_i's are then uncorrelated (however, in general, they are not independently distributed), and we let

$$f_{\nu 1, \delta 1, \nu 2, \delta 2, p}(x, y) = f_{\nu 1, \delta_1}(x) f_{\nu 2, \delta_2}(y) \pi_p(x, y), \tag{10}$$

where $f_{\nu_1, \delta_1}(x)$ and $f_{\nu_2, \delta_2}(y)$ denote the estimated marginal density functions for the standardized vector $(X, Y)'$ and $\pi_p(x, y)$ is a bivariate polynomial adjustment of order p in each variable. Note that whenever $\delta_i = 0$, $i = 1, 2$, the subscript δ_i is omitted on both sides of Equation (10) and that the subscript p is omitted when there is no polynomial adjustment.

The degrees ν_1, δ_1 and ν_2, δ_2 associated with the density estimates of X and Y are obtained in accordance with the guidelines provided in Remark 3. Due to the presence of a polynomial adjustment, smoother estimates (of lesser degrees) of the marginal density functions could be utilized.

Obtaining the Coefficients of the Polynomial Adjustment

The coefficients of the polynomial adjustment $\pi_p(x, y) = \sum_{i=0}^{p} \sum_{j=0}^{p} d_{i,j} x^i y^j$ can be determined as follows. For simplicity, we denote the estimated density function $f_{\nu 1, \delta 1, \nu 2, \delta 2, p}(x, y)$ by $f_p(x, y)$, and $f_{\nu 1, \delta 1}(x) f_{\nu 2, \delta 2}(y)$ by $\psi(x, y)$ so that

$$f_p(x, y) = \psi(x, y) \pi_p(x, y).$$

Let the $(k, \ell)^{\text{th}}$ joint moment associated with the exact density function $f(x, y)$ be denoted by $\mu(k, \ell) = \int_{\mathcal{R}^2} \int x^k y^\ell f(x, y) \, dx \, dy$ and the $(k, \ell)^{\text{th}}$ joint moment associated with the initial density $\psi(x, y)$, by $m(k, \ell) = \int_{\mathcal{R}^2} \int x^k y^\ell \psi(x, y) \, dx \, dy$.

In order to obtain a computable representation of the approximant $f_p(x, y)$, one needs to determine the coefficients $d_{i,j}$ of the polynomial adjustment. To this end, the joint moments of the exact density $f(x, y)$ are equated to those associated with $f_p(x, y)$, which yields

$$\begin{aligned} \mu(k, \ell) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^k y^\ell f_p(x, y) \, dx \, dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^k y^\ell \psi(x, y) \pi_p(x, y) \, dx \, dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^k y^\ell \psi(x, y) \sum_{i=0}^{p} \sum_{j=0}^{p} d_{i,j} x^i y^j \, dx \, dy \\ &= \sum_{i=0}^{p} \sum_{j=0}^{p} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} d_{i,j} x^{k+i} y^{\ell+j} \psi(x, y) \, dx \, dy, \end{aligned}$$

for $k = 0, \ldots, p$ and $\ell = 0, \ldots, p$, which yields the following $(p + 1)^2$ linear equations:

$$\mu(k, \ell) = \sum_{i=0}^{p} \sum_{j=0}^{p} d_{i,j} m(k + i, \ell + j), \quad k = 0, 1, 2, \ldots, p \text{ and } \ell = 0, 1, 2, \ldots, p.$$

Thus, the $d_{i,j}$'s can be obtained by solving the linear system $M\mathbf{d} = \boldsymbol{\mu}$ where $\mathbf{d}$ and $\boldsymbol{\mu}$ are vectors of dimensions $(p + 1)^2$ whose $(i(p + 1) + (j + 1))^{\text{th}}$ components, $d_{i,j}$ and $\mu(i, j)$, appear in the same order for $i = 0, 1, \ldots, p$ and $j = 0, 1, \ldots, p$. Note that increasing p should theoretically result in greater accuracy. The generalization to three or more variables is straightforward.

The selection of the optimal degree p associated with $\pi_p(x, y)$ is then made in terms of the following sum of squared differences:

$$\text{SSD}(p) = \sum_{i=1}^{n} (\text{ECDF}(w_i, z_i) - G_p(w_i, z_i))^2,$$

(a) Scatterplot of the original data

(b) Histogram of the original data

(c) SSD(ν_1) vs ν_1

(d) $f_5(x)$ and histogram

(e) SSD(ν_2) vs ν_2

(f) $f_3(y)$ and histogram

(g) Transformed base PDF $\phi(w, z)$

(h) SSD(p) vs p
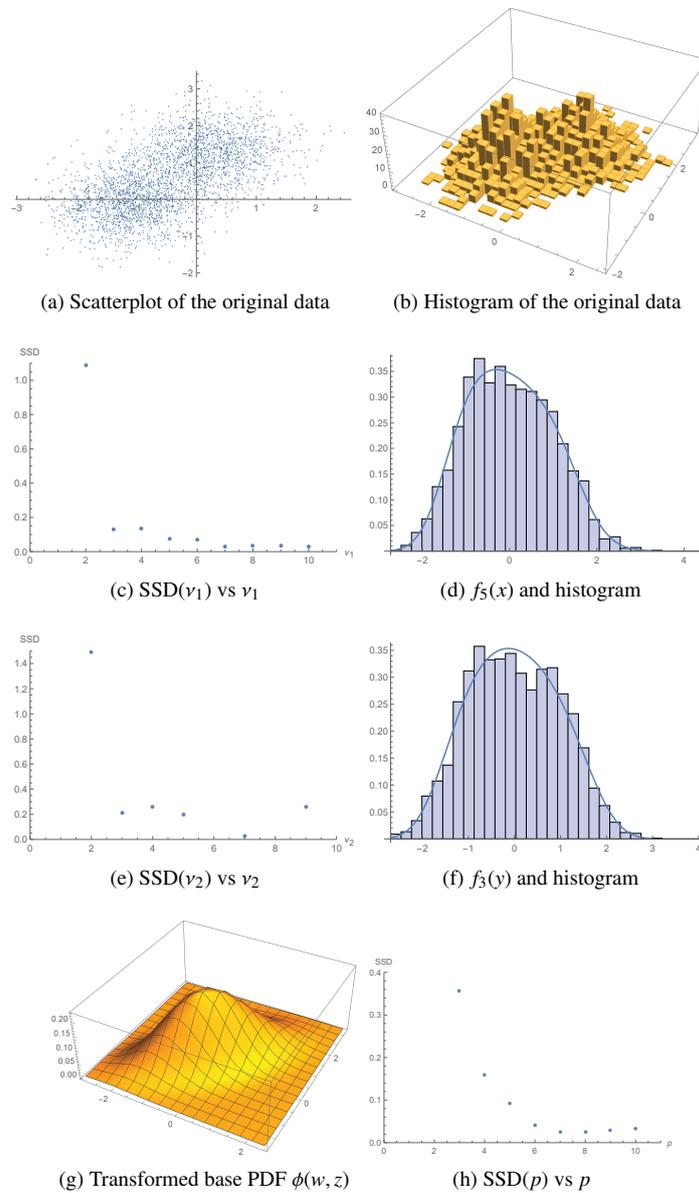
Figure 2. Dataset, SSD's, estimates of the marginal densities and base density function (Example 3.1)

(a) Exact   (b) Bivariate KDE   (c) Final PDF estimate $g_{5,3,7}(w, z)$
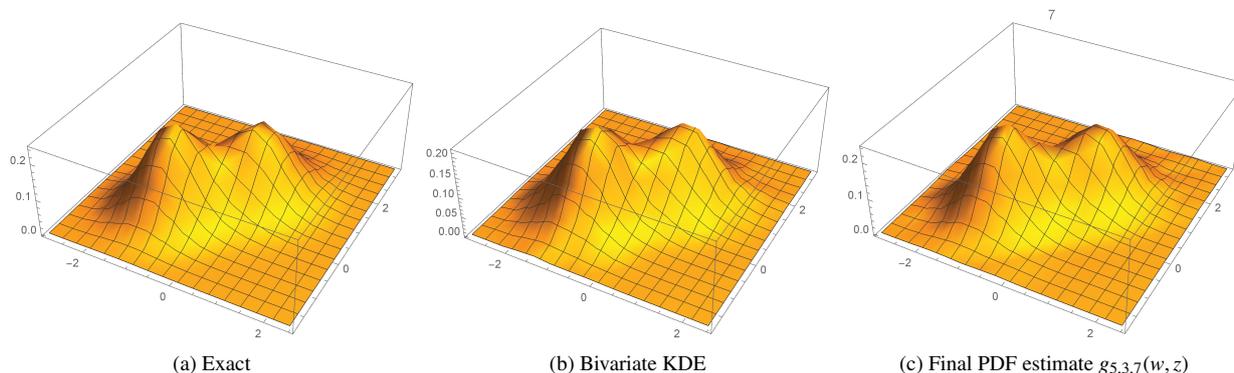
Figure 3. Exact and estimated density functions (Example 3.1)

where ECDF(w, z) denotes the empirical CDF associated with the dataset and $G_p(w_i, z_i) = \int_{-\infty}^{w_i} \int_{-\infty}^{z_i} g_p(w, z) \, dz \, dw$ is the CDF determined from the final density estimate $g_{\nu_1, \delta_1, \nu_2, \delta_2, p}(w, z) \equiv g_p(w, z)$, which is obtained by applying the inverse of the standardizing transformation to $f_p(x, y)$. Note that the base density function, that is, $\psi(x, y)$ will be denoted by $\phi(w, z)$ after applying this inverse transformation and that the degree of each estimated marginal density function is determined from the SSD function applied to the corresponding component of the original dataset.

The following example which makes use of a simulated dataset, enables one to gauge the accuracy of the density estimate obtained by making use of the proposed technique.

Example 3.1 Let $\mathbf{X}_1$ and $\mathbf{X}_2$ be bivariate normal random variables where

$$\mathbf{X}_1 \sim \mathcal{N}_2 \left(\begin{pmatrix} -1.1 \\ -0.1 \end{pmatrix}, \begin{pmatrix} 0.33 & 0.03 \\ 0.03 & 0.33 \end{pmatrix} \right) \quad \text{and} \quad \mathbf{X}_2 \sim \mathcal{N}_2 \left(\begin{pmatrix} 0.2 \\ 1.2 \end{pmatrix}, \begin{pmatrix} 0.4 & 0.04 \\ 0.04 & 0.4 \end{pmatrix} \right),$$

whose density functions are denoted by $g_{\mathbf{x}_1}(w, z)$ and $g_{\mathbf{x}_2}(w, z)$, respectively and let $\mathbf{X}$ denote the random vector resulting from an equally weighted mixture of their respective density functions, that is, $g_{\mathbf{x}}(w, z) = 0.5 g_{\mathbf{x}_1}(w, z) + 0.5 g_{\mathbf{x}_2}(w, z)$. Three thousand bivariate data points were generated from this mixture.

The scatterplot and a 3D histogram of the data are displayed in panels (a) and (b) of Figure 2. As per Remark 2, the SSD plots of the estimated marginal density functions shown in panels (c) and (e) indicate that $\nu_1 = 5$ and $\nu_2 = 3$ are suitable polynomial degrees. The corresponding univariate density functions are plotted in panels (d) and (f). The transformed base density estimate $g_{5,3}(x, y)$ obtained from the product of the estimated marginal densities which were determined by applying the PLDA methodology and the SSD(p) values are respectively plotted in panels (g) and (h) of Figure 2.

The bivariate exact density $g_{\mathbf{X}}(w, z)$, the kernel density estimate and the final joint density estimate $g_{5,3,7}(w, z)$ wherein $p = 7$ is selected as the optimal degree for the polynomial adjustment based on the SSD(p) values are included in Figure 3. It is observed that the plots of the estimated density based on the proposed methodology and the kernel density estimate obtained by applying *Silverman's rule of thumb* are in very close agreement with that of the exact density function, the SSD(7) value obtained from the final density estimate $g_{5,3,7}(w, z)$ being 0.0249 while it is 0.1147 for the kernel density estimate. The ISD values of the estimated density function and the kernel density estimate which were determined by making use of the bivariate counterpart of Equation (7) are respectively 0.0006 and 0.0011.

4. Illustrative Numerical Examples

Four applications of the proposed bivariate density estimation methodology, which involve actual datasets are presented in this section. In the first three instances, the derivative of the log-density estimate is assumed to be a polynomial whereas it is taken to be a rational function in the fourth one.

Example 4.1 The dataset being modeled in this example was extracted from '*CommViolPredUnnormalizedData*' which is included in the UC Irvine Machine Learning Repository dataset. It contains 2315 observation vectors related to communities and crime. It combines socio-economic data from the '90 Census, law enforcement data from the 1990 Law Enforcement Management and Admin Stats survey, and crime data from the 1995 FBI UCR, see (Redmond, 1990). We selected 'pctWFarmSelf': the percentage of households with a farm or self employment income as the W variable and 'perCapInc': per capita income in 1989 as the Z variable.

(a) Histogram of the original data (b) SSD(ν_1) vs ν_1 (c) $f_4(x)$ and histogram

(d) SSD(ν_2) vs ν_2 (e) $f_3(y)$ and histogram (f) SSD(p) vs p
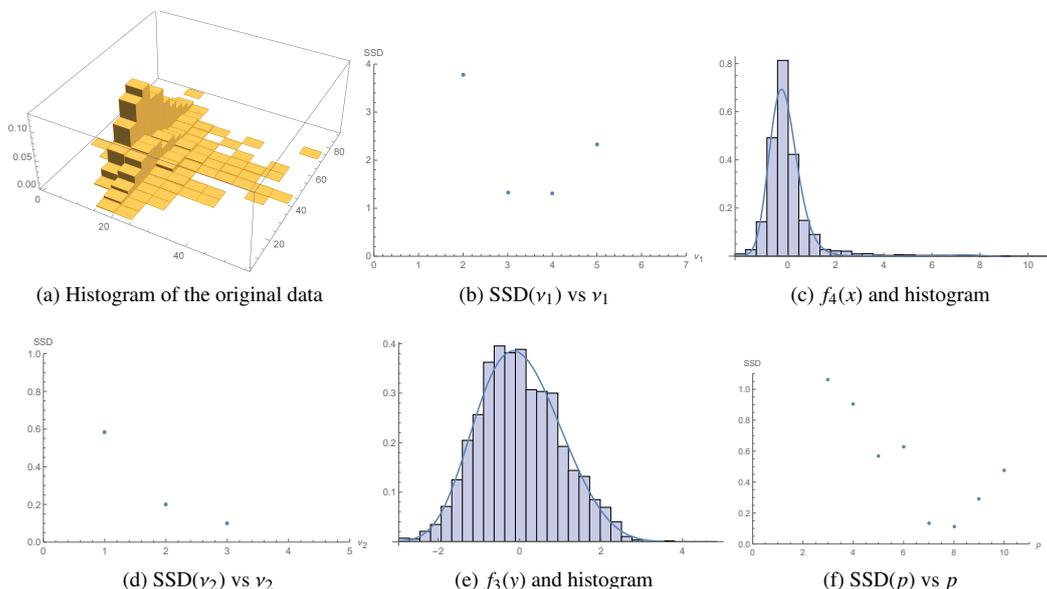
Figure 4. Dataset, SSD's and estimates of the marginal density functions (Example 4.1)

(a) Bivariate KDE (b) Transformed base PDF $\phi(w, z)$ (c) Final PDF estimate $g_{4,3,7}(w, z)$
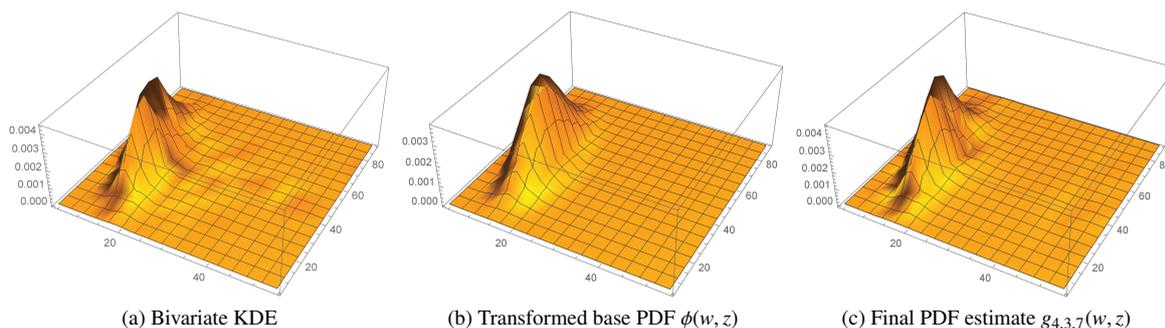
Figure 5. Density estimates (Example 4.1)

The histogram of the data is displayed in panel (a) of Figure 4. The SSD plots of the estimated marginal density functions of the standardized data shown in panels (b) and (d) of Figure 4 indicate that $\nu_1 = 4$ and $\nu_2 = 3$ are suitable degrees. The corresponding density functions are plotted in panels (c) and (e). (The SSD's for $\nu_1 = 6$ and $\nu_2 = 4$ are not shown as they were comparatively too large.)

The bivariate kernel density estimate, the transformed (by means of the inverse of the standardizing transformation) base density estimate $\phi(w, z)$ obtained from the product of the estimated marginal densities based on the PLDA methodology and the final joint density estimate $g_{4,3,7}(w, z)$ wherein $p = 7$ is selected as the optimum degree for the polynomial adjustment, as indicated by the SSD(p) values plotted in panel (f) of Figure 4, are included in Figure 5.

Example 4.2 In this example, we consider the dataset '*Concrete*' also included in the UC Irvine Machine Learning Repository dataset, which contains 1030 observation vectors related to concrete compressive strength in civil engineering. More information about this dataset is available in (Yeh, 1998). We selected 'Cement': kg in a m3 mixture as the *W* variable and concrete compressive strength as the *Z* variable.

In Figure 6, the scatterplot and a histogram of the '*Concrete*' data are displayed in panels (a) and (b). The SSD plots of the estimated marginal density functions that are shown in panels (c) and (e) of Figure 6 indicate that $\nu_1 = 4$ and $\nu_2 = 3$ are suitable degrees while the corresponding univariate estimated density functions are plotted in panels (d) and (f). Large SSD values were omitted.

The bivariate kernel density estimate, the transformed base density estimate $\phi(w, z)$ obtained from the product of the estimated marginal densities based on the PLDA methodology and the final joint density estimate $g_{4,3,6}(w, z)$ wherein $p = 6$ is selected as the optimum degree for the polynomial adjustment based on the SSD(p) values plotted in panel (c)

(a) Scatterplot of the original data

(b) Histogram of the original data

(c) SSD(ν_1) vs ν_1

(d) $f_4(x)$ and histogram

(e) SSD(ν_2) vs ν_2

(f) $f_3(y)$ and histogram
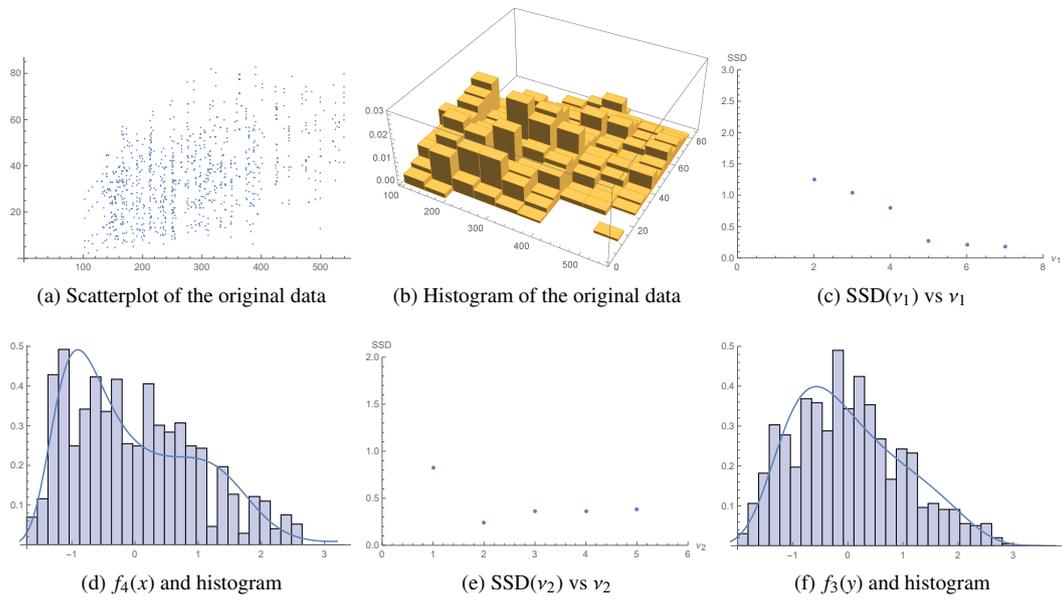
Figure 6. The dataset, SSD(ν_1), SSD(ν_2) and estimates of the marginal density functions (Example 4.2)

(a) Bivariate KDE

(b) Transformed base PDF $\phi(w, z)$

(c) SSD(p) vs p

(d) Final PDF estimate $g_{4,3,6}(w, z)$
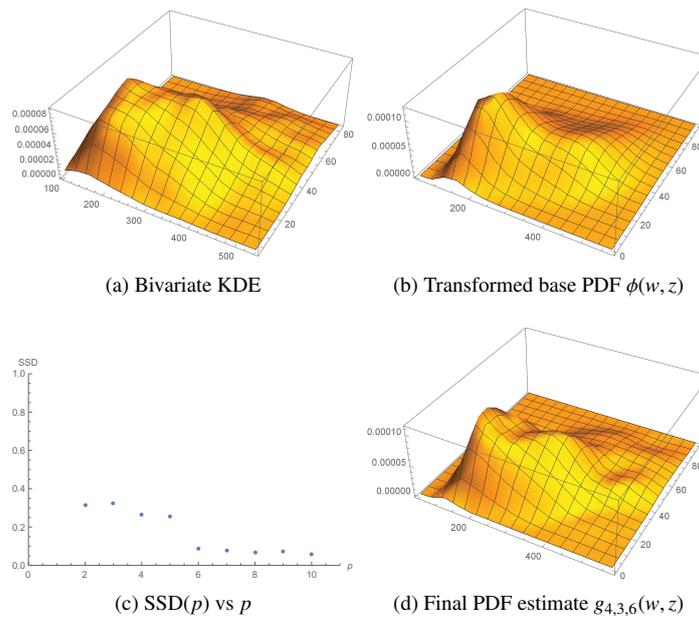
Figure 7. SSD(p) and bivariate density estimates (Example 4.2)

(a) Histogram of the original data   (b) $f_5(x)$ and histogram   (c) $f_6(y)$ and histogram   (d) Bivariate KDE
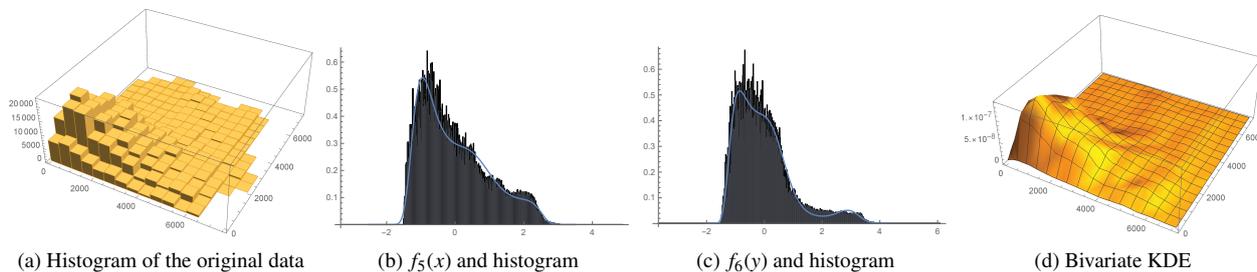
Figure 8. Dataset, estimates of the marginal density functions and bivariate KDE (Example 4.3)

(a) Transformed base PDF $\phi(w, z)$ (b) Final PDF estimate $g_{5,6,7}(w, z)$
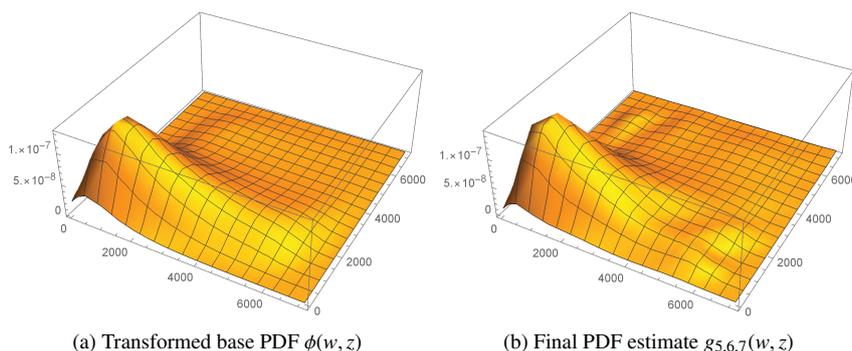
Figure 9. Transformed base PDF and final PDF estimate (Example 4.3)

are all included in Figure 7. We observe that the proposed density estimate which has an SSD of 0.0886 reflects the most salient features of the histogram more accurately than the kernel density estimate for which the SSD is 0.1369.

Example 4.3 The dataset being considered, which is called '*Covertype*', contains $581,012$ observations. It was extracted from the UC Irvine Machine Learning Repository dataset. This data was analyzed in (Blackard and Denis, 2000) in connection with forest cover studies. We selected horizontal distance in meters to nearest roadway as the W variable and horizontal distance in meters to nearest wildfire ignition points as the Z variable. Figure 8 displays a three-dimensional histogram of the data as well as $f_5(x)$ and $f_6(y)$, the estimated marginal density functions, and the corresponding histogram plots. The bivariate kernel density estimate is shown in panel (d).

The transformed base density estimate $\phi(w, z)$ and the final joint density estimate $g_{5,6,7}(w, z)$ wherein $p = 7$ is the selected degree of the polynomial adjustment are both included in Figure 9. We observe that the proposed density estimate is consistent with the histogram of the observations.

Example 4.4 Finally, the bivariate dataset being considered and referred to as the '*Flood*' data was collected in the Madawaska Basin, Quebec, from 1990-1995. It includes 77 observations. The first component of the data is the peak value and the second one is the volume. In this case, the DLDA methodology is applied and it is appropriate to let $\delta_1 = \delta_2 = 2$.

The scatterplot and a histogram of the data are displayed in panels (a) and (b) of Figure 10. The SSD plots of the estimated marginal density functions that are shown in panels (c) and (e) of Figure 10 indicate that $\nu_1 = 5$ and $\nu_2 = 5$ are suitable degrees. The corresponding density functions are plotted in panels (d) and (f).

The bivariate kernel density estimate, the transformed density estimate $\phi(w, z)$, that is, the transformed product of the estimated marginal densities, and the final joint density estimate $g_{5,2,5,2,6}(w, z)$ wherein $p = 6$ is selected as the optimal degree for the polynomial adjustment based on the SSD(p) values plotted in panel (c) are all included in Figure 11. The SSD associated with the proposed density estimate, that is, 0.0265 is about a third of that corresponding to the kernel density estimate, which is 0.0782.

5. Concluding Remarks

A technique is developed whereby the derivative of the logarithm of a univariate continuous density function is approximated by a rational function, which enables one to obtain *bona fide* density approximants for each of the marginals of a standardized continuous bivariate density function. Then, a bivariate density approximant is determined by adjusting the

(a) Scatterplot of the original data   (b) Histogram of the original data   (c) SSD(ν_1) vs ν_1

(d) $f_{5,2}(x)$ and histogram   (e) SSD(ν_2) vs ν_2 (f) $f_{5,2}(y)$ and histogram
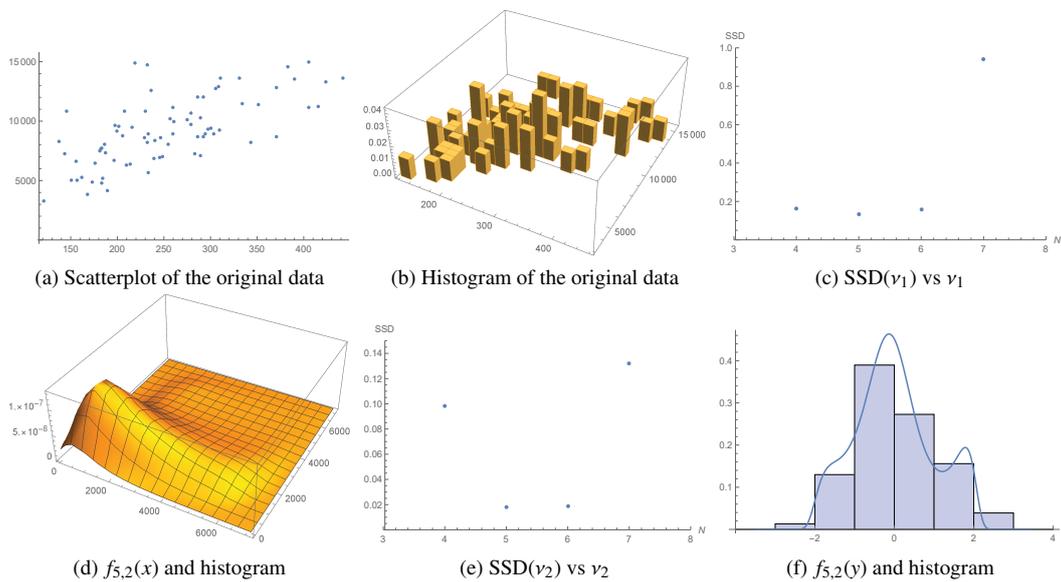
Figure 10. Dataset, SSD's and estimates of the marginal density functions (Example 4.4)

(a) Bivariate KDE   (b) Transformed base PDF $\phi(w, z)$

(c) SSD(p) vs p (d) Final PDF estimate $g_{5,2,5,2,6}(w, z)$
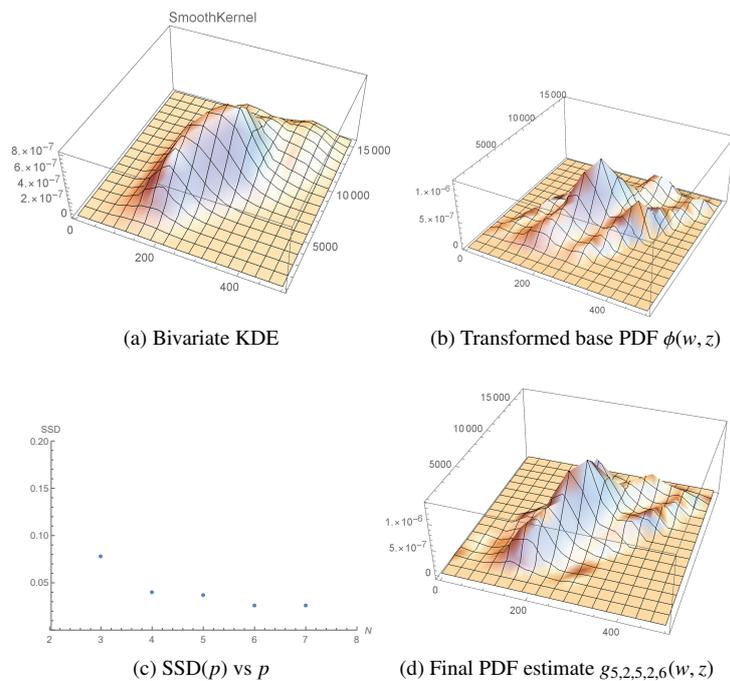
Figure 11. SSD and density estimates (Example 4.4)

product of the marginal density approximants (taken as base density function) by means of a bivariate polynomial whose coefficients are determined from the joint moments associated with the standardized target and base density functions. The methodology is then extended to be applicable in the context of density estimation on the basis of a set of observations by making use of their joint sample moments. This approach, which is well suited for modeling massive datasets, can readily be applied in multivariate settings. The *Mathematica* code utilized to carry out the calculations and to produce the graphs is available from the authors upon request.

Acknowledgements

References

Blackard, J. A., & Denis, J. D. (2000). Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables, *Computers and Electronics in Agriculture, 24*(3), 131–151. https://doi.org/10.1016/S0168-1699(99)00046-0

Devroye, L., & Gyrfi, L. (1985). Nonparametric Density Estimation: The L_1 View. *John Wiley & Sons*, New York.

Efromovich, S. (1999). Nonparametric Curve Estimation: Methods, Theory and Applications. *Springer Science & Business Media*, New York.

Eggermont, P. P. B., & LaRiccia, V. N. (2001). Maximum Penalized Likelihood Estimation: Volume I: Density Estimation. *Springer Science & Business Media*, New York.

Elderton, W. P., & Johnson, N. L. (1969). Systems of Frequency Curves. *Cambridge University Press*, Oxford. https://doi.org/10.1016/S0168-1699(99)00046-0

Provost, S. B. (2005). Moment-based density approximants. *The Mathematica Journal, 9*, 727–756.

Provost, S. B., & Ha, H. T. (2015). Differentiated logdensity approximants. *Statistical Methodology, 26*, 61–71. https://doi.org/10.1016/j.stamet.2015.02.005

Redmond, M. (1990). Computer Science, *La Salle University*, Philadelphia. (http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized)

Reid, N. (1988). Saddlepoint methods and statistical inference, *Statistical Science, 3*(2), 213-C238. https://doi.org/10.1214/ss/1177012906

Scott, D. W. (2015). Multivariate Density Estimation: Theory, Practice, and Visualization, 2nd Edition, *John Wiley & Sons*, Hoboken, NJ. https://doi.org/10.1002/9781118575574

Silverman, B. W. (1986). Density Estimation for Statistics and Data Analysis Series: Monographs on Statistics and Applied Probability. *Chapman and Hall*, London. https://doi.org/10.1007/978-1-4899-3324-9

Solomon, H., & Stephens, M. (1978). Approximations to density functions using Pearson curves. *Journal of the American Statistical Association, 73*, 153-C160. https://doi.org/10.1080/01621459.1978.10480019

Yeh, I. C. (1998). Modeling of strength of high performance concrete using artificial neural networks. *Cement and Concrete Research, 28*(12), 1797–1808. https://doi.org/10.1016/S0008-8846(98)00165-3

Copyrights