3-16-2018 4:00 PM

Advances in Semi-Nonparametric Density Estimation and Shrinkage Regression

Hossein Zareamoghaddam, *The University of Western Ontario*

Supervisor: Provost, Serge B., *The University of Western Ontario*
Joint Supervisor: Ahmed, Syed Ejaz, *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Statistics and Actuarial Sciences

ABSTRACT

This thesis advocates the use of shrinkage and penalty techniques for estimating the parameters of a regression model that comprises both parametric and nonparametric components and develops semi-nonparametric density estimation methodologies that are applicable in a regression context.

First, a moment-based approach whereby a univariate or bivariate density function is approximated by means of a suitable initial density function that is adjusted by a linear combination of orthogonal polynomials is introduced. Such adjustments are shown to be mathematically equivalent to making use of standard polynomials in one or two variables. Once extended to apply to density estimation, in which case the sample moments are being utilized, the proposed technique readily lends itself to the modeling of massive univariate or bivariate data sets. As well, the resulting density functions are shown to be expressible as kernel density estimates via the Christoffel-Darboux formula. Additionally, it is established that a set of n observations is entirely specified by its first n moments.

It is also explained 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. An explicit representation of the density function so obtained is derived and jointly sufficient statistics for its parameters are identified. Then, extensions of the proposed methodology to density estimation and multivariate settings are discussed. As a matter of fact, this approach constitutes a generalization of Pearson's system of curves. Several illustrative examples are presented including regression applications.

Finally, an iterative algorithm involving shrinkage and pretest techniques is introduced for estimating the parameters of a certain semi-nonparametric model. It is theoretically established and numerically verified that the proposed estimators are more accurate than the unrestricted ones. This methodology is successfully applied to a mass spectrometry data set.

Key Words: density approximation, joint moments, bivariate Hermite polynomials, bivariate normal distribution, bivariate density estimation, log-density, Pearson curve, semi-nonparametric model, local linear regression, multiple simple regression, shrinkage estimation.

I dedicate this thesis to my wife (Maryam) and my son (Daniel) and to the memory of my father, my mother and my younger brother.

ACKNOWLEDGEMENTS

First of all, I wish to express my sincere gratitude and appreciation to my supervisors, Professors Serge B. Provost and Syed Ejaz Ahmed their inspired guidances and generous support. I am also very grateful to the Department of Statistical and Actuarial Sciences and the Faculty of Graduate Studies for their financial support. I would like to express my thanks to the thesis examiners, Professors Sudhir Paul, Guangyong Zou, Ricardas Zitikis and Jiandong Ren for their valuable suggestions and comments. Thanks are also due to the Faculty, staff and fellow graduate students in the Department of Statistical and Actuarial Sciences.

I wish to express my profound appreciation to Mir Hashem Moosavi, Bangxin (George) Zhao, Rahim Khorasani, Mahdi Zonnourian, Fateme Behjatnia, Fariborz Golara, Freshteh Yasoubi, Saeed Yasoubi and Laleh Rahmani for their encouragement, help and support.

Last but not least, I would like to acknowledge with heartfelt appreciation the love and support of my wife for being with me in all situations.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

A plethora of results pertaining to density approximation or density estimation are available in the literature. In the former case, the exact density function is approximated by means of various statistical or mathematical techniques, while in the latter case, the underlying density function which is unknown is estimated from the available data. There are numerous studies on density approximation/estimation methodologies in univariate case; however, fewer are available for multivariate distributions. Chapter 2 is concerned with density approximation/estimation techniques that make use of polynomial adjustments and are applied to bivariate distributions. A family of distributions known as Pearson's curves is introduced in Chapter 3. In this case, one assumes that the logarithm of the density function of interest has a rational form where the numerator and denominators are polynomials of degrees one and two respectively. An extension of the Pearson frequency curves is proposed and an explicit representation of the resulting density approximant is provided. This approach is extended to apply to bivariate random vectors in Chapter 4. It should be pointed out that the bivariate density approximation/estimation techniques discussed in this thesis can readily be generalized with a view to model multivariate distributions. As well, they can be utilized in the context of regression. Additionally, a parameter estimation technique applying to a certain semi-nonparametric regression model, which was proposed by *Ma et al.* (2015) is improved upon in Chapter 5. In this model, there are n different location and scale parametric components and a common unknown function as the nonparametric component. Stein-type shrinkage and pretest techniques were applied to the

parametric components of the model to increase the level of accuracy of the estimators. The last chapter includes some concluding remarks and points out possible developments.

Note that since this thesis has been submitted in the 'integrated article' format, some redundancies are somewhat inevitable as each chapter is essentially self-contained. The *Mathematica* code utilized in connection with the main numerical examples presented in this dissertation is included in Appendix A.

1.1 Density estimation

1.1.1 Introduction

Over the years, statisticians have devoted much attention to approximating and estimating density functions, and there is a significant body of scientific literature on the subject. At the outset, it should be specified that what is referred to as a *bona fide* continuous density function is a nonnegative continuous function that integrates to one over its support.

Parametric distributions such as the exponential, gamma, beta, and normal, are entirely specified by their parameters. In order to estimate their associated density functions, one only needs to determine the parameters from the available observations. However, specific parametric distributions are often inadequate for modeling purposes. Accordingly, nonparametric density estimation techniques which are more flexible, have been widely studied. One can refer to Parzen (1962), Silverman (1986) and Izenman (1991) for further considerations about various density estimation methodologies. Since the density estimation techniques discussed in this thesis are based on moments, it is now established that the first n sample moments comprise all the information contained in a sample of size n.

Theorem 1.1.1. *A sample of size n is uniquely determined by the first n moments.*

Proof. Let $S = \{x_1, x_2, \ldots, x_n\}$ be a set of points and let $M = \{m_1, m_2, \ldots, m_n\}$ where $m_h = \sum_{i=1}^{n} x_i^h/n$ is the h^{th} sample moment, $h = 1, \ldots, n$. According to the fundamental theorem of algebra, a monic polynomial of degree n, $p(z) = a_0 + a_1 z + \cdots + a_{n-1} z^{n-1} + z^n$, is uniquely defined by its coefficients $\{a_0, a_1, \ldots, a_{n-1}\}$ and it is also uniquely specified by its n roots $\{x_1, x_2, \ldots, x_n\}$. Note that only the case of real roots is of interest in this thesis. Moreover, given S, the coefficients of $p(x)$ can be expressed in terms of the sequence of moments M via the Newton-Girard identity. Accordingly, a given polynomial of degree n, say $p(x)$, can be represented as follows:

$$\prod_{i=1}^{n}(x - x_i) = \sum_{k=0}^{n}(-1)^{n-k} e_{n-k}\, x^k, \tag{1.1.1}$$

where $e_0 = 1$ and

$$e_\ell = \frac{n}{\ell}\sum_{j=1}^{\ell}(-1)^{j-1} e_{\ell-j}\, m_j, \quad \ell = 1, \ldots, n. \tag{1.1.2}$$

Thus, given the first n sample moments associated with S, a sample of size n, one can determines the right hand side of (1.1.1) whose roots are precisely $\{x_1, x_2, \ldots, x_n\}$. This establishes that S is uniquely specified by M. Although S and M contain exactly the same amount information, oftentimes, only a subset of the latter suffices to elicit the distributional characteristics of a given sample. □

1.1.2 Kernel density estimates and histogram

One of the oldest and popular methodologies for estimating an unknown density function is known as kernel density estimation, see for instance (Rosenblatt, 1956), (Epanechnikov,

1969), (Scott, 1979) and references therein. Such estimates which depend on a kernel function $K(\cdot)$ as well as a bandwidth h, have the following functional form:

$$\hat{f}_h(x) = \frac{1}{n\,h}\sum_{i=1}^{n} K\Big(\frac{x - x_i}{h}\Big), \tag{1.1.3}$$

where n denotes the sample size and $\hat{f}_h(x)$ is the estimated density function at the point x. There are several types of kernel functions which, for example, are discussed in (Epanechnikov, 1969), (Scott, 1979), (Silverman, 1986), (Botev *et al.*, 2010), (Xu *et al.*, 2015) and references therein. The Gaussian, Epanechnikov and the tri-cube kernels are the most widely utilized. Additionally, there are various criteria for finding the optimum value of h, known as bandwidth selection techniques whereby on can obtain more accurate density estimate on the basis of a given dataset such as those proposed in (Bowman, 1984), (Park and Marron, 1990), (Sheather and Jones, 1991), (Jones, Marron and Sheather, 1996), (Hall *et al.* 1992), (Park, Turlach, 1992), (Cao *et al.*, 1994) and (Jiang, 2009). Note that every kernel function has to satisfy the following conditions: (1) $K(x) = K(-x)$, (2) $\int K(x)\,\mathrm{d}\,x = 1$ and (3) $0 < \int x^2 K(x)\,\mathrm{d}\,x < \infty$.

Among all the kernel density estimators, the histogram might be the oldest and easiest to implement, see (Rudemo, 1982). Incidentally, it is still widely used. In order to apply this approach to a given dataset, one needs an initial point x_0 and a bandwidth h. Then, the m intervals are usually defined as $D_i = [x_0 + (i-1)h,\, x_0 + i\,h)\,,\ i = 1,\, 2,\, \ldots,\, m$, where each interval is often taken to be closed from the left side and open from the right side. Let $\mathcal{I}_{D_i}(x)$ be the indicator function of the i^{th} interval, and n_i, the number of sample points that fall into D_i with $\sum_{i=1}^{m} n_i = n$. The histogram function is then given by

$$\hat{f}_H(x) = \frac{1}{n\,h}\sum_{i=1}^{m} n_i \mathcal{I}_{D_i}(x). \tag{1.1.4}$$

1.1.3 Density estimation methodologies involving polynomial adjustments

Some density approximation/estimation techniques depend on the moments of a given distribution. Whereas exact moments are used for the purpose of density approximation, sample moments are employed for the purpose of density estimation.

Density approximation/estimation methodologies involving polynomial adjustments are discussed in the next chapter. In this case, the density approximant is the product of an initial (base) density function $f_0(x)$ and a polynomial adjustment of degree m, $p(x) = \sum_{i=0}^{m} a_i x^i$. Thus, the resulting density approximant of degree m is $f_m(x) = f_0(x)\, p(x)$. The coefficients a_i, $i = 0,\ 1,\ \ldots,\ m$, are obtained by solving a linear system of equations that relies on the moments of the target and base density functions.

Now, let

$$\varphi_k(x) = \sum_{\ell=0}^{k} \delta_{k,\ell}\, x^{\ell},\ k = 0, \ldots, m\,, \tag{1.1.5}$$

be polynomials defined on the interval (α, β), which satisfy the orthogonality property,

$$\int_a^b w(x)\, \varphi_i(x)\, \varphi_j(x)\, \mathrm{d}x = \begin{cases} \theta_i & \text{for } \ i = j \\ 0 & \text{for } \ i \neq j, \end{cases} \tag{1.1.6}$$

where $w(x)$ denotes a certain nonnegative weight function whose 'moments' given by $\int_\alpha^\beta x^k\ w(x)\ \mathrm{d}x$, exist for $k = 0, 1, \ldots$, and θ_i will be referred to as the i^{th} degree orthogonality factor. Then, $\{\varphi_0(x), \varphi_1(x), \ldots, \varphi_m(x)\}$ is said to form a set of orthogonal polynomials with respect to $w(x)$.

Provost (2005) developed a density approximation methodology involving linear combinations of orthogonal polynomials as adjustments. Making use of orthogonal polynomials has the advantage that there is no need to solve any linear system of equations to determine the unknown coefficients present in the adjustment. It is explained in the next section that a density approximant computed as the product of a base density and a standard polynomial adjustment is equivalent to an approximant that is obtained as the product of the same initial density and a certain linear combination of orthogonal polynomials. It is also explained that in the context of density estimation, the latter can be expressed as a kernel density estimate.

1.1.3.1 Moment-based density approximants: orthogonal vs standard polynomials

It is shown in this section that the coefficients $a_j, j = 0, 1 \ldots, p$, appearing in the approximant $f_p(x)$, defined as $f_p(x) = c\, w(x) \sum_{i=0}^{p} a_i\, \varphi_i(x)$ where the normalizing constant c is such that $\int_\alpha^\beta c\, w(x)\, \mathrm{d}x = 1$, can be determined by matching the first p moments of $f_p(x)$ to those of $f(x)$, the density function being approximated. First, one can easily establish that the equalities

$$\int_\alpha^\beta x^j f_p(x)\, \mathrm{d}x = \int_\alpha^\beta x^j f(x)\, \mathrm{d}x, \;\; j = 0, 1, \ldots, p, \tag{1.1.7}$$

are mathematically equivalent to

$$\int_\alpha^\beta \varphi_j(x) f_p(x)\, \mathrm{d}x = \int_\alpha^\beta \varphi_j(x) f(x)\, \mathrm{d}x, \;\; j = 0, 1, \ldots, p, \tag{1.1.8}$$

where the $\varphi_j(x)$'s are orthogonal polynomials generated from the base density function by means of the Gram-Schmidt orthogonalization process. Accordingly, if (1.1.8) holds, which amounts to assuming that the first p moments of the approximate distribution are

equal to those associated with the target density function, one has

$$\int_{\alpha}^{\beta} c\, w(x) \sum_{i=0}^{p} a_i\, \varphi_i(x)\, \varphi_j(x)\, \mathrm{d}x = \int_{\alpha}^{\beta} \varphi_j(x) f(x)\, \mathrm{d}x,$$

that is,

$$\sum_{i=0}^{p} c\, a_i \int_{\alpha}^{\beta} w(x)\, \varphi_i(x)\, \varphi_j(x)\, \mathrm{d}x = \int_{\alpha}^{\beta} \varphi_j(x) f(x)\, \mathrm{d}x \tag{1.1.9}$$

or

$$c\, a_j\, \theta_j = \int_{\alpha}^{\beta} \varphi_j(x) f(x)\, \mathrm{d}x\,,$$

so that

$$a_j = \frac{\int_{\alpha}^{\beta} \varphi_j(x) f(x)\, \mathrm{d}x}{c\, \theta_j}\,, \tag{1.1.10}$$

where

$$\begin{aligned} \int_{\alpha}^{\beta} \varphi_j(x) f(x)\, \mathrm{d}x &= \int_{\alpha}^{\beta} \sum_{\ell=0}^{j} \delta_{j,\ell}\, x^{\ell} f(x)\, \mathrm{d}x \\ &= \sum_{\ell=0}^{j} \delta_{j,\ell}\, \mu_X(\ell)\,, \end{aligned} \tag{1.1.11}$$

$\mu_X(\ell)$ denoting the ℓ^{th} moment of the distribution specified by $f(x)$. Thus,

$$a_j = \sum_{\ell=0}^{j} \frac{\delta_{j,\ell}\, \mu_X(\ell)}{c\, \theta_j}\,, \quad j = 0, 1, \ldots, p\,, \tag{1.1.12}$$

and the p^{th} degree density approximant can be expressed as follows:

$$f_p(x) = w(x) \sum_{j=0}^{p} \sum_{\ell=0}^{j} \frac{\delta_{j,\ell}\, \mu_X(\ell)}{\theta_j}\, \varphi_j(x)\,, \tag{1.1.13}$$

where $\theta_j = \int_{\alpha}^{\beta} w(x)\, \varphi_j^2(x)\, \mathrm{d}x$ and $\delta_{j,\ell}$ denotes the coefficient of x^{ℓ} in $\varphi_j(x)$.

1.1.3.2 Kernel representation of the density estimates

Density estimates which are the counterparts of the orthogonal polynomial density approximants discussed in the previous section are shown to admit a certain kernel representation. Let $\{x_1, x_2, \ldots, x_n\}$ be a simple random sample from a population whose distribution is specified by the random variable X. On replacing the exact moments, $\mu_X(\ell)$, in (1.1.13) by the sample moments, $m_X(\ell) = \frac{1}{n}\sum_{i=1}^{n} x_i^{\ell}\,,\ \ell = 0, 1, \ldots, p$, one obtains the p^{th} degree orthogonal polynomial density estimate,

$$\begin{aligned} \hat{f}_p(x) &= w(x) \sum_{j=0}^{p} \frac{\varphi_j(x)}{\theta_j} \sum_{\ell=0}^{j} \delta_{j,\ell}\, m_X(\ell) && (1.1.14) \\ &= \frac{w(x)}{n} \sum_{j=0}^{p} \frac{1}{\theta_j}\, \varphi_j(x) \sum_{\ell=0}^{j} \delta_{j,\ell} \sum_{i=1}^{n} x_i^{\ell} \\ &= \frac{w(x)}{n} \sum_{i=1}^{n} \sum_{j=0}^{p} \frac{1}{\theta_j}\, \varphi_j(x_i)\, \varphi_j(x)\,. && (1.1.15) \end{aligned}$$

On making use of the Christoffel–Darboux formula, that is,

$$\sum_{k=0}^{p} \frac{\varphi_k(x)\, \varphi_k(y)}{\theta_k} = \frac{\delta_{p,p}}{\delta_{p+1,p+1}} \frac{\varphi_{p+1}(x)\, \varphi_p(y) - \varphi_p(x)\, \varphi_{p+1}(y)}{\theta_p\, (x - y)}\,, \tag{1.1.16}$$

cf., e.g., Hildebrand (1956), $\delta_{k,k}$ being the coefficient of x^k in $\varphi_k(x)$, and letting

$$\mathcal{K}_p(x, x_i) = w(x) \sum_{j=0}^{p} \frac{1}{\theta_j}\, \varphi_j(x_i)\, \varphi_j(x) \tag{1.1.17}$$

or equivalently,

$$\mathcal{K}_p(x, x_i) = \frac{w(x)\, \delta_{p,p}}{\delta_{p+1,p+1}} \Big(\frac{\varphi_{p+1}(x)\, \varphi_p(x_i) - \varphi_p(x)\, \varphi_{p+1}(x_i)}{\theta_p\, (x - x_i)} \Big)\,, \tag{1.1.18}$$

one has the following kernel representation of the orthogonal polynomial density estimate:

$$\hat{f}_p(x) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}_p(x, x_i), \tag{1.1.19}$$

which is mathematically equivalent to the representation given in (1.1.14) in terms of the first p sample moments. It can be shown that such kernels integrate to one.

Example 1.1.1. *Legendre Polynomial Kernels* Let $\varphi_k^{\mathrm{L}}(x) = \sum_{\ell=0}^{k} \delta_{k,\ell}^{\mathrm{L}}\, x^\ell$ denote a k^{th} degree Legendre polynomial; then the coefficient of x^ℓ as given explicitly in a suitable form in Provost and Ha (2009) is

$$\delta_{k,\ell}^{\mathrm{L}} = \frac{(-1)^k + (-1)^\ell}{2^{k+1}} \frac{(-1)^{\frac{3k-\ell}{2}}(k+\ell)!}{\Gamma(\frac{k-\ell}{2}+1)\,\Gamma(\frac{k+\ell}{2}+1)\;\ell!}, \quad \ell = 0, 1, \ldots, k,$$

so that $\delta_{n,n}^{\mathrm{L}} = 2n!/\left(2^n (n!)^2\right)$. In this case, the support is $(-1, 1)$, the weight function is $w(x) = 1/2$, and the orthogonality factor is $\theta_p = 2/(2p+1)$. Thus, according to (1.1.18), the p^{th} degree kernel associated with the Legendre polynomials is

$$\mathcal{K}_p(x, x_i) = \frac{(p+1)}{2} \frac{\left(\varphi_{p+1}^{\mathrm{L}}(x)\,\varphi_p^{\mathrm{L}}(x_i) - \varphi_p^{\mathrm{L}}(x)\,\varphi_{p+1}^{\mathrm{L}}(x_i)\right)}{x - x_i}, \tag{1.1.20}$$

where $\varphi_k^{\mathrm{L}}(x) = \sum_{i=0}^{k} \delta_{k,i}^{\mathrm{L}}\, x^i$ denotes a Legendre polynomial of degree k.

In some instances and, in particular, if one wishes to make use of available results on classical orthogonal polynomials, it may be indicated or even necessary to transform the data prior to resorting to the representations the density estimates given in (1.1.14), (1.1.15) or (1.1.19). For example, on letting $y_1, y_2, \ldots, y_n$ be a simple random sample from a distribution specified by the density function, $f(y)$, and making the change of variables $x = g(y)$, where $g(y)$ is a differentiable function of y, the density estimate corresponding to (1.1.19) becomes

$$\hat{f}_p(y) = \frac{|g'(y)|}{n} \sum_{i=1}^{n} \mathcal{K}_p(g(y), g(y_i)), \tag{1.1.21}$$

where $\mathcal{K}_p(\cdot,\cdot)$ is as defined in Equation (1.1.18). Oftentimes, it suffices to apply an affine transformation such as

$$g(y) = \frac{y-\tau}{\nu}\,, \tag{1.1.22}$$

so that support or the first moment (or in some cases the first two moments) of the transformed variable coincide(s) with that (those) of the normalized weight function associated with a given type of orthogonal polynomials. It should be noted that, by making use of the Gram-Schmidt orthogonal generalization process, one can always generate a set of orthogonal polynomials from a suitable weight function, in which case there is no need to apply any transformation to the data.

When one makes use of the linear transformation specified by Equation (1.1.22), the density estimates (1.1.14), (1.1.15) and (1.1.19), respectively become

$$\hat{f}_p(y) = \frac{1}{\nu}\, w\Big(\frac{y-\tau}{\nu}\Big) \sum_{j=0}^{p} \frac{\varphi_j\Big(\frac{y_i-\tau}{\nu}\Big)}{\theta_j} \sum_{\ell=0}^{j} \delta_{j,\ell}\, m_X(\ell) \tag{1.1.23}$$

with

$$m_X(\ell) = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{y_i-\tau}{\nu}\Big)^{\ell} = \frac{1}{\nu^{\ell}}\sum_{k=0}^{\ell}\binom{\ell}{k} m_Y(k)\,(-1)^{\ell-k}\,\tau^{\ell-k}\,,$$

where

$$m_Y(k) = \frac{1}{n}\sum_{i=1}^{n} y_i^k\,,\quad k=0,1,\ldots,\ell\,;$$

$$\hat{f}_p(y) = \frac{1}{n\,\nu}\, w\Big(\frac{y-\tau}{\nu}\Big) \sum_{i=1}^{n}\sum_{j=0}^{p} \frac{1}{\theta_j}\,\varphi_j\Big(\frac{y_i-\tau}{\nu}\Big)\,\varphi_j\Big(\frac{y-\tau}{\nu}\Big); \tag{1.1.24}$$

and

$$\hat{f}_p(y) = \frac{1}{n\,\nu}\sum_{i=1}^{n}\mathcal{K}_p\Big(\frac{y-\tau}{\nu}, \frac{y_i-\tau}{\nu}\Big)\,. \tag{1.1.25}$$

In the case of a density approximant, which is based on $\mu_Y(k) = E(Y^k),\ k = 0.1,\ldots,p,$,

the linear transformation yields the following density approximant corresponding to (1.1.13):

$$f_p(y) = w\Big(\frac{y-\tau}{\nu}\Big) \sum_{j=0}^{p} \sum_{\ell=0}^{j} \frac{\delta_{i,\ell}\, \mu_X(\ell)}{\nu\, \theta_j}\, \varphi_j\Big(\frac{y-\tau}{\nu}\Big)\,, \tag{1.1.26}$$

where

$$\mu_X(\ell) = \frac{1}{\nu^\ell} \sum_{k=0}^{\ell} \binom{\ell}{k} \mu_Y(k)(-1)^{\ell-k}\tau^{\ell-k}\,.$$

As a further refinement, one could define the kernels to be zero subinterval where they are negative and renormalizing the resulting function to obtain a *bona fide* density estimate.

The main results derived in this section, that is, the connection between approximants and estimates and the dual representation of the density estimates, ought to provide valuable insights into the orthogonal polynomial density estimation methodology advocated herein and lead to a heightened appreciation of this approach as a viable alternative to other density estimation techniques.

1.1.4 Pearson's frequency curves

In order to model skewed observations, Pearson (1895) proposed a system of frequency curves known as Pearson's curves. He assumed that the following differential equation holds:

$$\frac{\mathrm{d}\, f_X(x)}{\mathrm{d}\, x} = \frac{x - a_0}{c_2 x^2 + c_1 x + c_0}\, f_X(x), \tag{1.1.27}$$

where $f_X(x)$ is the density function and a_0, c_0, c_1 and c_2 are constant parameters. Then, he identified 13 major types of density curves, depending on the values of the constant parameters, which can be determined as the solutions of the differential equation (1.1.27) by solving an appropriate linear system of equations requiring the first four moments of the

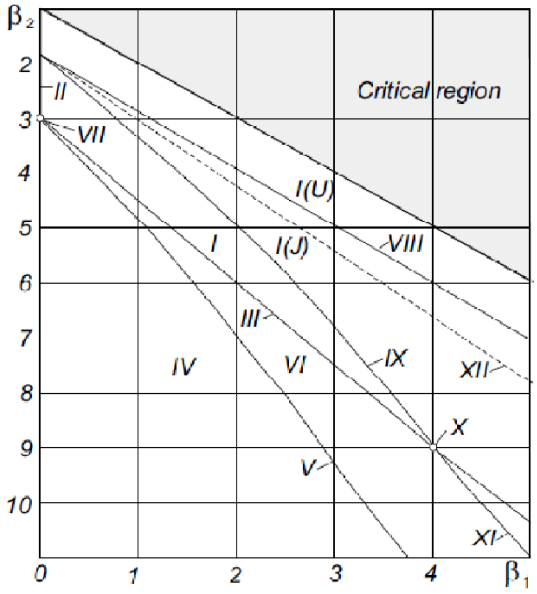


Figure 1.1: The regions of different distribution types of Pearson's frequency curves based on the values of β_1 = skewness squared and β_2 = kurtosis + 3, where the β_2-axis is in reverse order. There are no distributions in the critical region (top right shaded area).

distribution. In Figure 1.1 (Podladchikova, et al., 2003), different Pearson curves can be identified in terms of certain relationships between their skewness and kurtosis. There are many well-known families of distributions, such as the *normal*, *beta*, *gamma*, *chi square*, *Student t* whose density functions could be obtained from this differential equation.

By multiplying both sides of Equation (1.1.27) by x^n, one has

$$x^n\,(c_2x^2 + c_1x + c_0)\frac{\mathrm{d}\,f_X(x)}{\mathrm{d}\,x} = x^n(x - a_0)\,f_X(x). \tag{1.1.28}$$

Then, by integrating by part both sides of (1.1.28) and assuming that $\lim_{|x|\to\infty} x^n f_X(x) = 0$, the following recurrence formula for the moments of $f_X(x)$ is obtained:

$$a_0\mu_n - n\,c_0\mu_{n-1} - (n+1)c_1\mu_n - (n+2)c_2\mu_{n+1} = \mu_{n+1}, \tag{1.1.29}$$

where μ_n is the n^{th} moment. On setting $\mu_{-1} = 0$, with $\mu_0 = 1$, one has the following

linear system of equations:

$$\begin{aligned} -a_0 + c_1 &= 0 \\ c_0 + 3c_2\mu_2 &= -\mu_2 \\ -a_0\mu_2 + 3c_1\mu_2 + 4c_2\mu_3 &= -\mu_3 \\ -a_0\mu_3 + 3c_0\mu_2 + 4c_1\mu_3 + 5c_2\mu_4 &= -\mu_4. \end{aligned} \tag{1.1.30}$$

where μ_k being the k^{th} *central* moments for $k = 0,\ 1,\ 2,\ 3,\ 4$.

1.1.5 Differentiated log-density approximants (DLDA's)

This is a generalization of the Pearson frequency curves discussed in the previous subsection whereby the derivative of logarithm of a density function $f_X(x)$ whose support is $(\alpha,\ \beta)$ is assumed to be a rational function, that is,

$$\frac{\mathrm{d}}{\mathrm{d}x} ln(f_X(x)) = \frac{f'_X(x)}{f_X(x)} = r(x), \tag{1.1.31}$$

where

$$r(x) = \frac{\sum_{i=0}^{\nu} a_i\, x^i}{\sum_{j=0}^{\delta} c_j\, x^j} = \frac{N_\nu(x)}{D_\delta(x)}, \tag{1.1.32}$$

$N_\nu(x)$ 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. The coefficients a_i's and c_j's are determined by solving a linear system that is specified in Chapter 3 where an explicit representation of the approximant,

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

is provided, κ is denoting the normalizing constant. Unlike Padé approximants, DLDA's remain nonnegative.

1.2 Shrinkage, pretest and penalty estimators

Various regression models, both parametric or semi-nonparametric, have been proposed in the literature. For regression models involving a certain number of parameters, it is important to determines the number of significant parameters; otherwise, overfitting may create problems for prediction purposes and other analyses. Additionally, one should know which variables are significant and which ones can be considered as nuisance. In this section, we give an overview of variable selection based on the shrinkage, pretest and penalty strategies, as discussed in Ahmed (2014).

1.2.1 Shrinkage methods: basic concepts

Consider the regression model,

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.2.1}$$

where $\mathbf{y} = (y_1,\, y_2, \ldots,\, y_n)'$ is the response vector, X is an $n \times p$ fixed matrix of coefficients, $\boldsymbol{\beta} = (\beta_1,\, \beta_2,\, \ldots,\, \beta_p)'$ is the unknown vector of parameters and $\boldsymbol{\epsilon} = (\epsilon_1,\, \epsilon_2,\, \ldots,\, \epsilon_n)'$ is the unknown error vector, a prime denoting the transpose of a matrix or a vector. Let

$E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. The following two assumptions, called "regularity conditions", are made for deriving the estimators:

- $\max\limits_{1\leq i\leq n} \mathbf{x_i}'(X'X)^{-1}\mathbf{x_i} \to 0.$ as $n \to \infty$, where $\mathbf{x_i}'$ is the i^{th} row of X.
- $\lim\limits_{n\to\infty} \frac{X'X}{n} = C$, where C is a finite positive definite matrix.

If there is no restriction on $\boldsymbol{\beta}$, the unrestricted estimator (UE) of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}}^{UE} = (X'X)^{-1}X'\mathbf{y}. \tag{1.2.2}$$

Let $\boldsymbol{\beta}$ be divided into two subvectors as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ so that $\boldsymbol{\beta}_1$ is a p_1-vector and $\boldsymbol{\beta}_2$ is a p_2-vector with $p_1 + p_2 = p$ and $0 \leq p_i,\ i = 1, 2$. Suppose that we have the uncertain prior information (UPI) that $\boldsymbol{\beta}_2$ is a vector of nuisance parameters and $\boldsymbol{\beta}_1$ is the set of parameters of interest. Such UPI might be written as a linear restriction $H\boldsymbol{\beta} = \mathbf{h}$ on $\boldsymbol{\beta}$ where H is a known $p_2 \times p$ matrix and $\mathbf{h}$ is a known vector of length p_2. Under the restriction $H\boldsymbol{\beta} = \mathbf{h}$, the restricted estimator (RE) is obtained as

$$\hat{\boldsymbol{\beta}}^{RE} = \hat{\boldsymbol{\beta}}^{UE} - (X'X)^{-1}H'(H(X'X)^{-1}H')^{-1}(H\hat{\boldsymbol{\beta}}^{UE} - \mathbf{h}) \tag{1.2.3}$$

where $\hat{\boldsymbol{\beta}}^{RE}$ is a linear transformation of $\hat{\boldsymbol{\beta}}^{UE}$. If we want to test the validity of the restriction (UPI) as the null hypothesis $H_0 :\ H\boldsymbol{\beta} = \mathbf{h}$, the following test statistic is to be utilized:

$$\phi_n = \frac{(H\hat{\boldsymbol{\beta}}_i - h)'(H(X'X)^{-1}H')^{-1}(H\hat{\boldsymbol{\beta}}_i - h)}{s_e^2},\ \ i = 1, ..., n,$$

where

$$s_e^2 = \frac{(\mathbf{y} - X\hat{\boldsymbol{\beta}}^{UE})'(\mathbf{y} - X\hat{\boldsymbol{\beta}}^{UE})}{n-p}$$

is an estimator of σ^2. Under H_0, ϕ_n has a chi-square distribution with p_2 degrees of freedom.

Shrinkage and pretest estimators are obtained as linear combinations of $\hat{\boldsymbol{\beta}}^{UE}$ and $\hat{\boldsymbol{\beta}}^{RE}$ in terms of the test statistic ϕ_n as described below.

The Stein-type shrinkage estimator (SE) $\hat{\boldsymbol{\beta}}_1^{\,S}$ of $\boldsymbol{\beta}_1$ is calculated as

$$\hat{\boldsymbol{\beta}}_1^{\,S} = \hat{\boldsymbol{\beta}}_1^{\,RE} + (\hat{\boldsymbol{\beta}}_1^{\,UE} - \hat{\boldsymbol{\beta}}_1^{\,RE})\{1 - k\phi_n^{-1}\},$$

where $k = p_2 - 2 \quad (p_2 \geq 3)$. In some cases, the sign of $(1 - k\phi_n^{-1})$ might be negative which may adversely affect the estimators. To overcome this difficulty, Ahmed (2014) used the positive-rule Stein-type estimator (PSE) which is defined as

$$\hat{\boldsymbol{\beta}}_1^{\,S+} = \hat{\boldsymbol{\beta}}_1^{\,RE} + (\hat{\boldsymbol{\beta}}_1^{\,UE} - \hat{\boldsymbol{\beta}}_1^{\,RE})\{1 - k\phi_n^{-1}\}^+,$$

where $a^+ = max\{a,\, 0\}$. An alternative way of evaluating $\hat{\boldsymbol{\beta}}_1^{\,S+}$ is

$$\hat{\boldsymbol{\beta}}_1^{\,S+} = \hat{\boldsymbol{\beta}}_1^{\,RE} + (\hat{\boldsymbol{\beta}}_1^{\,UE} - \hat{\boldsymbol{\beta}}_1^{\,RE})\{1 - k\phi_n^{-1}\}I(\phi_n < k),$$

where $I(\cdot)$ denotes the indicator function.

The pretest technique (PT) of estimating $\boldsymbol{\beta}_1$, the parameter vector of interest, yields the pretest estimator $\hat{\boldsymbol{\beta}}_1^{\,PT}$ which is of the form,

$$\hat{\boldsymbol{\beta}}_1^{\,PT} = \hat{\boldsymbol{\beta}}_1^{\,UE} - (\hat{\boldsymbol{\beta}}_1^{\,UE} - \hat{\boldsymbol{\beta}}_1^{\,RE})I(\phi_n < c_{n,\alpha}),$$

where $c_{n,\alpha}$ is the upper $100\alpha^{\text{th}}$ percentage point of the test statistic ϕ_n. Under the PT, we test the prior information (i.e. UPI) with H_0 before computing the estimator while SE and PSE use the value of the test statistic to obtain the estimator. In this case, PT either rejects or not the presence of $\hat{\boldsymbol{\beta}}_1^{\,RE}$ based on whether $\phi_n > c_{n,\alpha}$ or not.

In addition to the above techniques, there are some popular penalized least squares family of estimators which are for instance available in Ahmed (2014). Under these techniques, a penalized term (involving the parameters) is added to the sum of squared errors to shrink a subset of the parameters to zero. The resulting general form of the penalizing least squares is

$$S(\boldsymbol{\beta}) = (\mathbf{y} - X\boldsymbol{\beta})'(\mathbf{y} - X\boldsymbol{\beta}) + \lambda\pi(\boldsymbol{\beta}), \tag{1.2.4}$$

where $\pi(\boldsymbol{\beta})$, which is called the penalized function, is a function of the parameter vector $\boldsymbol{\beta}$. $\lambda > 0$ is called the tuning parameter and the optimum value is mainly selected by cross validation.

Within this class of estimators, ridge regression (Hoerl, 1958), least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1994), adaptive Lasso (Zou, 2006), the smoothly clipped absolute deviation (SCAD) method (Fan and Li, 2001) and the minimax concave penalty (MCP) (Zhang, 2010) are the most popular techniques. Among them, LASSO and SCAD are more widely used. We shall simply provide a brief review of these techniques, more information about these methods being available in (Ahmed, 2014).

The penalty function given in Equation (1.2.4) for LASSO is the sum of the absolute values of the parameters. Accordingly, the LASSO estimator is obtained as

$$\hat{\boldsymbol{\beta}}^{LASSO} = arg\min_{\beta}\Big\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}x_{i,j}\beta_j)^2 + \lambda\sum_{i=1}^{n}|\beta_j|\Big\}.$$

In this case, the shrinkage and variable selection are done simultaneously.

SCAD is another important penalized estimator which is given by

$$\hat{\boldsymbol{\beta}}^{SCAD} = arg\min_{\beta}\Big\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j)^2 + \lambda\sum_{i=1}^{n} p_{\alpha,\,\lambda}|\beta_j|\Big\},$$

where $p_{\alpha,\,\lambda}(\cdot)$ is the smoothly clipped absolute deviation penalty. The SCAD penalty function is a quadratic spline on $[0,\,\infty)$ with nodes at λ and $\alpha\lambda$, whose first order derivative is

$$p_{\alpha,\,\lambda}(x) = \lambda\{I(|x| \leq \lambda) + \frac{(\alpha\lambda - |x|)^{+}}{(\alpha-1)\lambda} I(|x| > \lambda)\}, \quad x \geq 0,$$

where $\lambda > 0$ and $\alpha > 2$ are the tuning parameters. For the asymptotic analysis of these methods, the reader is referred to Ahmed (2014).

References

Ahmed, S.E. (2014). Penalty, Shrinkage and Pretest Strategies Variable Selection and Estimation, Springer, New York.

Botev, Z.I., Grotowski, J.F. and Kroese, D.P. (2010). Kernel density estimation via diffusion, Annals of Statistics, 38(5), 2916–2957.

Bowman, A.W. (1984). An alternative method of cross-validation for the smoothing of density estimates, Biometrika, 71(2), 353–360.

Cao, R., Cuevas, A. and Manteiga, W.G. (1994). A comparative study of several smoothing methods in density estimation, Computational Statistics and Data Analysis, 17(2), 153–176.

Elderton, W.P. and Johnson, N.L. (1969). Systems of Frequency Curves, Cambridge University Press, Oxford.

Epanechnikov, V.A. (1969). Non-parametric estimation of a multivariate probability density, Theory of Probability and its Applications, 14, 153–158.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, Journal of the American Statistical Association, 96, 1348–1360.

Hajek, J. (1972), Local asymptotic minimax and admissibility in estimation, in: Proc. Sixth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1. University of California Press, Berkeley.

Hall, P., Marron, J.S., Park, B.U. (1992). Smoothed cross-validation, Probability Theory and Related Fields, 92, 1–20.

Hildebrand, F.B. (1956). Introduction to Numerical Analysis, McGraw-Hill, New York.

Hoerl, A.E. (1958). Application of ridge analysis to regression problems, Chemical Engineering Progress, 54–59.

Izenman, A.J., (1991). Recent developments in nonparametric density estimation, Journal of the American Statistical Association, 86, 205–224.

Jiang, M. (2009). Advances in Kernel Density Estimation, PhD Thesis, The University of Western Ontario, London, Canada.

Jones, M.C., Marron, J.S., Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation, Journal of the American Statistical Association, 91(433), 401–407.

Ma, W., Feng, Y., Chen, K. and Ying, Z. (2015). Functional and parametric estimation in a semi- and nonparametric model with application to mass-spectrometry data, The International Journal of Biostatistics, 11(2), 285–303.

Park, B.U., Marron, J.S. (1990). Comparison of data-driven bandwidth selectors, Journal of the American Statistical Association, 85(409), 66–72.

Park, B.U., Turlach, B.A. (1992). Practical performance of several data driven bandwidth selectors (with discussion), Computational Statistics, 7, 251–270.

Parzen, E. (1962). On estimation of a probability density function and mode, The Annals of Mathematical Statistics, 33(3), 1065–1076.

Pearson, K. (1895). Contributions to the mathematical theory of evolution, II: Skew variation in homogeneous material, Philosophical Transactions of the Royal Society, 186, 343–414.

Podladchikova, O., Lefebvre, B., Krasnoselskikh, V., and Podladchikov, V. (2003). Classification of probability densities on the basis of Pearson's curves with application to coronal heating simulations, Nonlinear Processes in Geophysics, European Geosciences Union (EGU), 10(4/5), 323–333.

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

Provost, S.B., and Ha, H.T. (2015). Differentiated logdensity approximants, Statistical Methodology, 26, 61–71.

Provost, S.B., and Ha, H.T. (2009). On the inversion of certain moment matrices, Linear Algebra and its Applications, 430, 2650–2658.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function, The Annals of Mathematical Statistics, 27(3), 832–837.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators, Scandinavian Journal of Statistics, 9(2), 65–78.

Sheather, S.J., Jones, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation, Journal of the Royal Statistical Society, Series B, 53(3), 683–690.

Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis, London: Chapman and Hall/CRC.

Scott, D. (1979). On optimal and data-based histograms, Biometrika, 66(3), 605–610.

Tibshirani, R. (1994). Regression shrinkage and selection via the LASSO, Journal of the Royal Statistical Society, Series B, 58, 267–288.

Xu, X., Yan, Z., Xu, S. (2015). Estimating wind speed probability distribution by diffusion-based kernel density method,. Electric Power Systems Research, 121, 28–37.

Zhang, C.H. (2010). Nearly unbiased variable selection under minimax concave penalty, The Annals of Statistics, 38, 894–942.

Zou, H. (2006). The adaptive LASSO and its oracle properties, Journal of the American Statistical Association, 101, 1418–1429.

Chapter 2

A moment-based bivariate density estimation methodology applicable to big data modeling

2.1 Introduction

When density functions do not have closed form representations or they assume complicated forms that may for instance involve special functions, it may be desirable to replace them by certain relatively simple moment-based approximations. However, unlike the type of approximants being proposed in this chapter, such approximations, which include Pearson curves (Solomon, 1978), Edgeworth expansions (Edgeworth, 1905), Johnson curves (Elderton and Johnson, 1969), Gram-Charlier expansions (Charlier, 1906) and the saddle-point approximations (Daniels, 1954) and (Reid, 1988) can prove inadequate. This is often the case when for example the target distributions are not unimodal.

It may happen that different distributions have the same moments, which is often referred to as the "moment problem". Rao (2001) provided conditions that ensure the uniqueness of a density function with respect to its moment sequence, $\mu(i)$, $i = 0, 1, 2, \ldots$. A sufficient condition is that

$$\sum_{i=1}^{\infty} \frac{\mu(i)t^i}{i!}$$

be absolutely convergent for some $t > 0$.

The density approximation methodology being proposed in this section, that is, adjusting a base density function by means of a polynomial or a linear combination of orthogonal polynomials, is not restricted to univariate density functions. While being conceptually simple and easy to implement, this approach is flexible, and accurate. As well, it is applicable in the context of density estimation.

Incidentally, it should be noted that moment-based density estimation techniques are ideally suited for modeling data sets containing an exceedingly large number of observations, which will be referred to as massive data sets. Indeed, once the moments have been evaluated, which is easily achieved even for extremely large data sets, the determination of the estimated density function does not depend on the sample size. Moreover, when a new set of observations, say, $x_{n_1+1}, \ldots, x_n$, becomes available in addition to an initial data set, $x_1, \ldots, x_{n_1}$, there is no need to make use of each of the n_1 original data points. This is the case since the h^{th} updated moment will then be $\{n_1 m_h + \sum_{i=n_1+1}^{n} x_i^h\}/n$ where m_h denotes the h^{th} sample moment evaluated from the initial data set. In the bivariate case, let $m_{h,\ell} = \sum_{i=1}^{n_1} x_i^h y_i^\ell / n_1$ be the $(h, \ell)^{\text{th}}$ joint sample moments of the initial observations $(x_1, y_1), \ldots, (x_{n_1}, y_{n_1})$. Then, the joint moments can be similarly updated as $\{n_1 m_{h,\ell} + \sum_{i=n_1+1}^{n} x_i^h y_i^\ell\}/n$ given the additional observations $(x_{n_1+1}, y_{n_1+1}), \ldots, (x_n, y_n)$. This property clearly extends to joint moments occurring in higher dimensions. Thus, moment-based methodologies enable one to process large amounts of univariate or multivariate data that often arrive in streams without having to access previously collected observations. Note that, for instance, this is not the case for kernel density estimates which, unlike the proposed estimates, do not have a simple functional representation.

This chapter is organized as follows. A univariate density approximation technique relying on Hermite polynomials and their associated Gaussian type weight function is discussed in this introductory section. As explained in Provost and Ha (2009), approximants

that are based on other initial continuous density functions and their associated orthogonal polynomials can be similarly obtained. The use of standard polynomials as adjustments is treated in Provost (2005). An efficient formula for generating bivariate Hermite polynomials, which can easily be extended to the multivariate case, is provided in Section 2.2. Some properties of such polynomials are also described in that section. In Section 2.3, the density function associated with a bivariate random vector is first approximated in terms of a bivariate standard polynomial. Then, the use of linear combinations of bivariate Hermite polynomials is discussed in connection with the approximation of normal-type bivariate densities, which are often encountered in practice. Algorithms are provided for both approaches, which are shown to produce identical density estimates. This methodology is extended to density estimation in Section 2.4 where several illustrative examples are presented. Some concluding remarks are included in the last section.

The proposed moment-based density approximation methodology is now briefly described for the univariate case. Let Y be a random variable whose density function is to be approximated and $X = (Y - \mu)/\sigma$. First, $f_X(x)$, the density of X, is approximated by means of a certain base density, $\psi_X(x)$, that is adjusted by a polynomial of degree n:

$$f_{X_n}(x) = \psi_X(x) \sum_{k=0}^{n} \xi_k \, x^k. \tag{2.1.1}$$

Clearly, the resulting density approximant for Y is then given by

$$f_{Y_n}(y) = \psi_X((y-u)/s) \sum_{k=0}^{n} \frac{\xi_k}{s} \Big(\frac{y-u}{s}\Big)^k. \tag{2.1.2}$$

As explained in Provost (2005), the coefficients ξ_k can be determined by equating the first n moments of $f_{X_n}(x)$ to those of $f_X(x)$ and solving the resulting linear system of

equations.

Since the adjustment consists of a polynomial, it may happen that some of its roots fall within the support of the target distribution, in which case the approximant will be slightly negative on certain subintervals which, generally, is hardly noticeable graphically. It should be noted that, theoretically, the higher the degree of the adjustment, the closer to zero such negative parts will be. In any case, this issue can easily be addressed by defining the original function to be zero when it is negative and normalizing the resulting function so that it integrates to one, thus yielding a *bona fide* density approximant. Alternatively, one may apply an iterative procedure, namely the P-algorithm proposed by Gajek (1986), in order to obtain legitimate density approximants. Whether in the context of density approximation or density estimation, we shall refer to expressions such as those appearing in Equations (2.1.2) and (2.3.1) as density functions with the understanding that they can readily be made *bona fide*.

When $\psi_X(x)$ is a standard normal density function, Equation (2.1.1) can be written as follows in terms of orthogonal polynomials:

$$f_{X_n}(x) = \psi_X(x) \sum_{k=0}^{n} \eta_k H_k(x) \tag{2.1.3}$$

where $H_k(x),\ k = 0,\, 1, \ldots,$ are univariate modified Hermite polynomials given by

$$\begin{aligned} H_k(x) &= (-1)^k e^{\frac{x^2}{2}} \frac{\mathrm{d}^k}{\mathrm{d}x^k} e^{-\frac{x^2}{2}},\ -\infty < x < \infty, \\ &\equiv \sum_{h=0}^{k} \alpha_{kh}\, x^h. \end{aligned} \tag{2.1.4}$$

These polynomials satisfy the following orthogonality property:

$$\int_{-\infty}^{\infty} e^{-x^2/2} H_i(x)\, H_j(x)\, \mathrm{d}x = \sqrt{2\pi} j!\, \delta_{ij}\,, \tag{2.1.5}$$

where δ_{ij} is equal to 1 if $i = j$ and, 0, otherwise. In Equation (2.1.5), $e^{-x^2/2}$ is the weight function associated with the density function is $\psi_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ whose j^{th} moment is $m_X(j) = \int_{-\infty}^{\infty} x^j \psi_X \mathrm{d}x$. In this instance, the coefficients η_k in Equation (2.1.3) can be determined from equating $\int_{-\infty}^{\infty} H_k(x) f_{X_n}(x)\, \mathrm{d}x$ to $\int_{-\infty}^{\infty} H_k(x) f_X(x)\, \mathrm{d}x$ for $k = 0,\, 1,\, \ldots, n$. On applying the orthogonality property (2.1.5), these coefficients are then obtained as

$$\eta_k = \frac{1}{c_T \theta_k} \Big\{ \sum_{h=0}^{k} \alpha_{kh}\, \mu_X(h) \Big\}, \;\; k = 0,\, 1,\, 2,\, \ldots,\, n, \tag{2.1.6}$$

where $\theta_k = \sqrt{2\pi} k!\,,\; \alpha_{k0},\, \ldots,\, \alpha_{kn}$ are the coefficients obtained from Equation (2.1.4), and $\mu_X(h) = E(X^h)$. Expansions of functions in terms of orthogonal polynomial series such as Hermite, Laguerre, Jacobi and Legendre, essential and approximating properties of orthogonal systems as well as multiply orthogonal series are discussed for instance in Sansone (2004), Alexits (1961) and Szegö (1959); the convergence behavior of such expansions, including the fundamental theorem on the convergence of orthogonal series are also treated therein.

In this case, the determination of the coefficients η_k for $k = 0,\, 1,\, 2,\, \ldots,\, n$, does not require the solving of any linear system of equations. In general, a sequence of orthogonal polynomials can be generated from a given weight function by applying the Gram-Schmidt orthogonalization process. The approximation of a univariate mixture of normal density functions is considered in the following example.

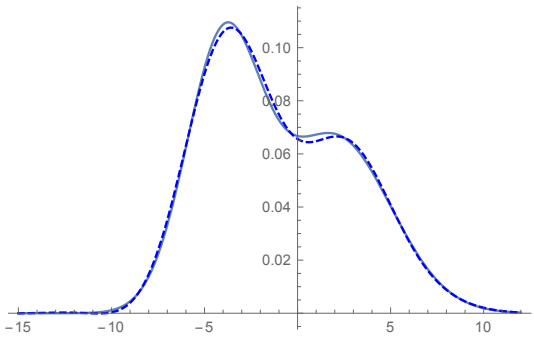


Figure 2.1: Exact (solid line) and approximated (dashed line) density functions

Example 2.1.1. Let $X_1 \sim \mathcal{N}(-4,\, 2)$, $X_2 \sim \mathcal{N}(2,\, 3)$ and $f_X(x) = 1/2\big[f_{X_1}(x)+f_{X_2}(x)\big]$. In Figure 2.1, the dashed line represents the density approximant as determined from the first 15 moments of the mixture by making use of Equation (2.1.3). This approximant is seen to be nearly identical to the exact density function on which it is superimposed. Even more accurate approximations could be obtained by making use of additional moments.

2.2 Bivariate Hermite polynomials

As indicated earlier, univariate Hermite polynomials can be determined from Equation (2.1.4). A few approaches may be utilized for obtaining bivariate Hermite polynomials. Rayner *et al.* (2013) proposed a recurrence formula that cannot be easily extended to higher dimensions. Willink (2005) discussed the use of a simpler formula, which is an extension of Equation (2.1.4), and derived some useful relations for determining multivariate Hermite polynomials of different orders and obtaining the moments of a multivariate normal distribution in terms of Hermite polynomials. Later, Withers and Nadarajah (2010) suggested a related formula. We will hereafter make use of the differentiation formula proposed by Willink (2005). Letting $\mathbf{z} = (x,\, y)'$ and Σ be a positive definite matrix of order 2, the associated bivariate Hermite polynomial of orders r_1 and r_2 can be obtained as follows:

$$H_{r_1,r_2}(\mathbf{z},\,\Sigma) \equiv (-1)^r \exp\big(\mathbf{z}'\Sigma^{-1}\mathbf{z}\big)\frac{\partial^r}{\partial x^{r_1}\partial y^{r_2}}\exp\big(-\mathbf{z}'\Sigma^{-1}\mathbf{z}\big) \tag{2.2.1}$$

where $r = r_1 + r_2$. For simplicity, we will denote $H_{r_1,r_2}(\mathbf{z},\,\Sigma)$ by $H_{r_1,r_2}(x,\,y)$ when there is no ambiguity. This formula can easily be extended to k variables, in which case $\mathbf{z} = (x_1,\,x_2,\,\ldots,\,x_k)'$ and $r = r_1 + r_2 + \cdots + r_k$.

Example 2.2.1. Let $\mathbf{Z} = (X,\,Y)'$ be a centered bivariate Gaussian random vector with covariance matrix $\Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}$. Hermite polynomials of various orders $r\,(= r_1 + r_2)$ were determined by applying Equation (2.2.1). For example, one has

$$H_{1,1}(x,\,y) = \tfrac{1}{5} - \tfrac{3x^2}{25} + \tfrac{7xy}{25} - \tfrac{2y^2}{25},$$
$$H_{1,2}(x,\,y) = -\tfrac{8x}{25} + \tfrac{3x^3}{125} + \tfrac{6y}{25} - \tfrac{13x^2y}{125} + \tfrac{16xy^2}{125} - \tfrac{4y^3}{125},$$
$$H_{2,2}(x,\,y) = \tfrac{8}{25} - \tfrac{33x^2}{125} + \tfrac{9x^4}{625} + \tfrac{52xy}{125} - \tfrac{42x^3y}{625} - \tfrac{22y^2}{125} + \tfrac{61x^2y^2}{625} - \tfrac{28xy^3}{625} + \tfrac{4y^4}{625}.$$

When $\Sigma = I$ and $r_2 = 0$, one has

$$H_{1,0}(x,\,y) = x,$$
$$H_{2,0}(x,\,y) = x^2 - 1,$$
$$H_{3,0}(x,\,y) = x^3 - 3x,$$
$$H_{4,0}(x,\,y) = x^4 - 6x^2 + 3,$$
$$H_{5,0}(x,\,y) = x^5 - 10x^3 + 15x,$$

which are the univariate Hermite polynomials of orders 1 to 5.

In the bivariate case, a concept called the *dual* of a bivariate Hermite polynomial is needed.

Definition 2.2.1. Let $H_{i,j}(x,\, y) = \sum_{\ell=0}^{j} \sum_{k=0}^{i} \delta_{ijk\ell}\, x^k y^\ell$ be a bivariate Hermite polynomial of order $(i,\, j)$ where i and j are nonnegative integers. Then

$$H^*_{u,v}(x,\, y) = \sum_{n=0}^{v} \sum_{m=0}^{u} \delta^*_{uvmn} x^m y^n$$

is called the dual bivariate Hermite polynomial associated with $H_{i,j}(x,\, y)$ with respect to the weight function $g(x,\, y)$. Then, one has

$$\int_{\mathcal{R}^2} \int g(x,\, y)\, H^*_{u,v}(x,\, y)\, H_{i,j}(x,\, y)\, \mathrm{d}x\, \mathrm{d}y = \begin{cases} \theta_{u,v} & \text{if } (i,j) = (u,v), \\ 0 & \text{otherwise.} \end{cases} \tag{2.2.2}$$

Remark 1. The dual of a bivariate Hermite polynomial $H_{u,v}(x,\, y)$ can be obtained as

$$H^*_{u,v}(x,\, y) = E[(x + iX)^u (y + iY)^v]$$

where $i = \sqrt{-1}$ and $(X,\, Y)$ is a centered bivariate normal random vector, see Willink (2005) and Withers and Nadarajah (2010).

The following theorem explains how the joint moments of a bivariate normal vector can be obtained from bivariate Hermite polynomials.

Theorem 2.2.1. *Let* $\mathbf{Z} = (X,\, Y)'$ *be a normal random vector with mean* $\boldsymbol{\mu}$ *and covariance matrix* Σ*, that is,* $\mathbf{Z} \sim \mathcal{N}_2(\boldsymbol{\mu},\, \Sigma)$*; then, according to Willink (2005), the* $(r_1,\, r_2)^{\text{th}}$ *joint moment of* $\mathbf{Z}$ *can be determined as follows:*

$$E(X^{r_1} Y^{r_2}) = H_{r_1,r_2}(-\Sigma^{-1}\boldsymbol{\mu},\, -\Sigma^{-1}). \tag{2.2.3}$$

2.3 Moment-based bivariate density approximation

In this section, the univariate methodology, as described in Section 2.1, is extended to approximate bivariate continuous density functions. First, we discuss the use of standard polynomials as adjustments and then, demonstrate that utilizing linear combination of orthogonal polynomials produces exactly the same approximants. Hereafter, we shall denote the target density function by $f(x,\, y)$. Let $f(x,\, y) = \psi(x,\, y)\, \lambda(x,\, y)$ where $\psi(x,\, y)$ is a suitable initial density approximant whose selection is discussed in Remark 2 and $\lambda(x,\, y) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} c_{i,j} x^i y^j$ is an adjustment. Now, let $f_{p,q}(x,\, y)$ denote the following approximant of orders p and q:

$$f_{p,q}(x,\, y) = \psi(x,\, y)\, \lambda_{p,q}(x,\, y) \tag{2.3.1}$$

where $\lambda_{p,q}(x,\, y) = \sum_{i=0}^{p} \sum_{j=0}^{q} c_{i,j} x^i y^j$ is a truncated polynomial adjustment. Typically, the more fluctuating or jagged a marginal density function is, the greater the highest degree of the corresponding variable appearing in the polynomial adjustment ought to be.

Remark 2. In order to identify a suitable base density function, one may rely for instance on common knowledge about the characteristics of the target distribution or the general features of a histogram of the observations when a data set is available. If only the moments of a distribution being approximated are available, then a uniform base density would be indicated. Only coarse initial estimates or approximants are required since the polynomial adjustments will lead to significant improvements in accuracy. For instance, in order to approximate a bivariate normal-type density function, one would make use of a bivariate Gaussian base density function, as was done in Example 2.3.1. As can be seen from the last three numerical examples included in the next section, other types of base density functions

can also be appropriately selected.

Let the integrated squared difference between $f_{p,q}(x, y)$ and $f(x, y)$ over the support of joint distribution be denoted by

$$ISD(p, q) = \int \int \Big(f_{p,q}(x, y) - f(x, y) \Big)^2 \mathrm{d}x \, \mathrm{d}y. \tag{2.3.2}$$

To quantify the error incurred by approximating the target density function $f(x,\, y)$ with $f_{p,q}(x, y)$ and determine the optimal orders of the adjustment term, we seek values of p and q such that $ISD(p, q)$ reaches a set tolerance level or beyond which $ISD(p, q)$ only decreases marginally.

Let the $(k, \ell)^{\mathrm{th}}$ exact joint moment associated with the density function $f(x,\, y)$ be denoted by $\mu(k, \ell) = \int_{\mathcal{R}^2} \int x^k \, y^\ell \, f(x,\, y) \, \mathrm{d}x \, \mathrm{d}y$ and the $(k, \ell)^{\mathrm{th}}$ joint moment associated with the base density $\psi(x,\, y)$, by $m(k, \ell) = \int_{\mathcal{R}^2} \int x^k \, y^\ell \, \psi(x,\, y) \, \mathrm{d}x \, \mathrm{d}y$. Joint moments could as well be determined by differentiating the moment-generating functions when they are available.

In order to obtain a computable representation of the approximant $f_{p,q}(x,\, y)$, one needs to determine the coefficients $c_{i,j}$ of the truncated polynomial adjustment. To this end, as in univariate case, the joint moments of the exact density $f(x,\, y)$ are equated to those associated with $f_{p,q}(x,\, y)$:

$$\begin{aligned}\mu(k,\ell) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^k y^\ell f_{p,q}(x,\,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^k y^\ell\,\psi(x,\,y)\,\lambda_{p,q}(x,\,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^k y^\ell\,\psi(x,\,y)\sum_{i=0}^{p}\sum_{j=0}^{q} c_{i,j}\,x^i y^j\,\mathrm{d}x\,\mathrm{d}y \\ &= \sum_{i=0}^{p}\sum_{j=0}^{q}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} c_{i,j}\,x^{k+i}y^{\ell+j}\,\psi(x,\,y)\,\mathrm{d}x\,\mathrm{d}y,\end{aligned}$$

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

$$\mu(k,\ell) = \sum_{i=0}^{p}\sum_{j=0}^{q} c_{i,j}\,m(k+i,\,\ell+j),\ k = 0,\,1,\,2,\ldots,\,p,\ \ell = 0,\,1,\,2,\ldots,\,q. \quad (2.3.3)$$

Thus, the $c_{i,j}$'s can be obtained by solving the linear system $M\mathbf{c} = \boldsymbol{\mu}$ where $\mathbf{c}$ and $\boldsymbol{\mu}$ are vectors of dimensions $(p+1)(q+1)$ whose $(i(q+1)+(j+1))^{\text{th}}$ component, $c_{i,j}$ and $\mu(i,j)$, appear in the same order for $i = 0,\,1,\,\ldots,\,p$ and $j = 0,\,1,\,\ldots,\,q$. Increasing p and q should theoretically result in greater accuracy. The generalization to three or more variables is straightforward. The following algorithm describes the process of approximating a continuous density function $f(x,\,y)$ in terms of standard polynomial adjustments.

Algorithm 3.1. Moment-based density approximants expressed in terms of standard polynomial adjustments

Step 1: Let p and q initially equal 3—or a larger integer, which would be indicated in the case of an irregular marginal density function.

Step 2: Evaluate the joint moments of the bivariate vector (X, Y), that is, $\mu(i, j)$ for $i = 0, 1, \ldots, p$ and $j = 0, 1, \ldots, q$.

Step 3: Set the density approximant as $f_{p,q}(x, y) = \psi(x, y)\, \lambda_{p,q}(x, y)$ where $\psi(x, y)$ is the initial base density (selected as per Remark 2) and $\lambda_{p,q}(x, y) = \sum_{i=0}^{p} \sum_{j=0}^{q} c_{i,j} x^i y^j$ is the polynomial adjustment.

Step 4: Obtain the coefficients $c_{i,j}$ by solving the linear system of equations resulting from Equation (2.3.3).

Step 5: Evaluate $ISD(p, q)$ as defined in Equation (2.3.2).

Step 6: Repeat steps 2-5 with larger values of p and/or q until $ISD(p, q)$ is deemed to be sufficiently small.

Note that it can prove useful to standardize or rescale the data before applying Algorithm 3.1 or Algorithm 3.2. This can be achieved as follows:

Let $\boldsymbol{m}' = (m_x, m_y)$ be the mean of a random sample $(x_1, y_1), \ldots, (x_n, y_n)$ whose underlying distribution is specified by the density function $f(x, y)$ and let $S^{-1/2}$ be the inverse of symmetric square root of the sample covariance matrix. Then, letting $\boldsymbol{z}_i = (x_i, y_i)$, $i = 1, \ldots, n$, the standardizing transformation

$$\boldsymbol{z}_i^* = S^{-1/2} \left(\boldsymbol{z}_i - \boldsymbol{m} \right), \tag{2.3.4}$$

is utilized to remove the correlation between the variables with $\boldsymbol{z}_i^* \equiv (x_i^*, y_i^*)'$. When dealing with positive random variables, it suffices to rescale the data, in which case $\mathbf{m}$ is set to $\mathbf{0}$ in the above transformation.

A density estimate is then obtained by applying Algorithm 3.1 to $(x_i^*, y_i^*)'$. The final density estimate $f_{p,q}$ for $(x_1, y_1), \ldots, (x_n, y_n)$ is then obtained by applying the inverse

transformation to the density $f_{p,q}(x^*, y^*)$, which yields

$$\hat{f}_{p,q}(x, y) = \hat{f}_{p,q}(\boldsymbol{z}) = |S^{-1/2}|\, f_{p,q}\left(S^{-1/2}\,(\boldsymbol{z} - \boldsymbol{m})\right) . \tag{2.3.5}$$

Now, let the adjustment be expressed in terms of orthogonal polynomials. When a Gaussian base density is indicated, once the target distribution has been centered, the polynomial adjustment $\lambda_{p,q}(x,\, y)$ consists of a linear combination of bivariate Hermite polynomials as defined in Section 2.2. In this instance, one may utilize the following approximation to $f(x,\, y)$:

$$f_{p,q}(x,\, y) = \psi(x,\, y) \sum_{i=0}^{p} \sum_{j=0}^{q} \eta_{i,j} H_{i,j}(x,\, y) \tag{2.3.6}$$

where $\psi(x,\, y) = 2\phi(\sqrt{2}\,x, \sqrt{2}\,y)$ and $\phi(\cdot, \cdot)$ denotes a standard bivariate normal density function.

The coefficients η_{ij} are obtained by equating $\int\int H^*_{u,v}(x,\, y) f_{p,q}(x,\, y)\,\mathrm{d}x\,\mathrm{d}y$ to $\int\int H^*_{u,v}(x,\, y) f(x,\, y)\,\mathrm{d}x\,\mathrm{d}y$ for $u = 0,\, 1,\, \ldots, p$ and $v = 0,\, 1,\, \ldots, q$, where the dual Hermite polynomial $H^*_{u,v}(x,\, y)$ is as specified in Definition . Denoting the $(k, \ell)^{\mathrm{th}}$ joint moment of the centered target distribution by $\mu(k,\, \ell)$, one has

$$\int\int \psi(x,\, y)\, H^*_{u,v}(x,\, y) \sum_{i=0}^{p} \sum_{j=0}^{q} \eta_{i,j} H_{i,j}(x,\, y)\mathrm{d}x\,\mathrm{d}y = \sum_{\ell=0}^{v} \sum_{k=0}^{u} \delta^*_{uvk\ell}\, \mu(k,\, \ell). \tag{2.3.7}$$

Then, on making use of Equation (2.2.2) with $g(x,\, y)$ replaced by $\psi(x,\, y)$, the left-hand side of Equation (2.3.7) becomes

$$\sum_{i=0}^{p} \sum_{j=0}^{q} \eta_{i,j} \int\int \psi(x,\, y) H^*_{u,v}(x,\, y) H_{i,j}(x,\, y)\,\mathrm{d}x\,\mathrm{d}y = \eta_{u,\,v}\theta_{u,\,v},$$

so that

$$\eta_{u,\,v} = \frac{\sum_{\ell=0}^{v}\sum_{k=0}^{u}\delta^{*}_{uvk\ell}\,\mu(k,\,\ell)}{\theta_{u,\,v}}. \tag{2.3.8}$$

Thus, as in the univariate case, there is no need to solve linear systems of equations when making use of linear combinations of orthogonal polynomials as adjustments. The steps to be utilized for approximating $f(x,\,y)$ in terms of bivariate Hermite orthogonal polynomials are provided in the next algorithm.

Algorithm 3.2. Moment-based density approximants expressed in terms of bivariate Hermite orthogonal polynomials

Step 1: Let p and q initially equal 3 —or a larger integer, which would be indicated in the case of an irregular marginal density function.

Step 2: Evaluate the joint moments of the bivariate vector $(X,\,Y)$ denoted by $\mu(i,\,j)$ for $i = 0,\,1,\,\ldots,\,p$ and $j = 0,\,1,\,\ldots,\,q$.

Step 3: As specified by Equation (2.3.6), set the approximate density as $f_{p,q}(x,\,y) = \psi(x,\,y)\sum_{i=0}^{p}\sum_{j=0}^{q}\eta_{i,j}H_{i,j}(x,\,y)$ where $\psi(x,\,y)$ is a centered Gaussian base density, $H_{i,j}(x,\,y)$ is the associated bivariate Hermite polynomial of order $(i,\,j)$ and $\sum_{i=0}^{p}\sum_{j=0}^{q}\eta_{i,j}H_{i,j}(x,\,y)$ is the adjustment.

Step 4: Evaluate the coefficients $\eta_{i,j}$ by making use of Equation (2.3.8).

Step 5: Evaluate $ISD(p,q)$ as defined in Equation (2.3.2).

Step 6: Increase p and/or q and repeat steps 2-5 until $ISD(p,q)$ is deemed sufficiently small.

The proposed density estimation methodology can readily be extended to the multivariate case as follows. Let $f(\boldsymbol{x})$ denote the density function associated with a $p-$dimensional

random vector $\mathbf{X}' = (\mathrm{X}_1, \ldots, \mathrm{X}_p)$ whose domain is $\mathcal{D} = (\alpha_1, \beta_1) \times \cdots \times (\alpha_p, \beta_p)$ and $\{\varphi_{i_k}(x),\ i_k = 0, 1, \ldots,\ \},\ k = 1, \ldots, p$, be sequences of orthogonal polynomials that can be generated from the normalized nonnegative weight functions $w_k(x)$, $k = 1, \ldots, p$, whose product shall serve as base density. Letting

$$\varphi_{\boldsymbol{i}}(\boldsymbol{x}) = \prod_{k=1}^{p} \varphi_{i_k}(x_k)\,,$$

the $\varphi_{\boldsymbol{i}}(\boldsymbol{x})$'s are seen to form an orthogonal system on $\mathcal{D}$ with respect to the weight function $w(\boldsymbol{x}) = \prod_{k=1}^{p} w_k(x_k)$. Assuming that $f(\boldsymbol{x})$ is square integrable, that is, $\int_{\mathcal{D}} f^2(\boldsymbol{x}) \mathrm{d}\,\boldsymbol{x} < \infty$, defining $\boldsymbol{i}' = (i_1, \ldots, i_p)$ to be a vector of nonnegative integers, and letting $\mathbf{1}' = (1, \ldots, 1)$ be a p-dimensional vector, $f(\boldsymbol{x})$ can be expanded as follows:

$$f(\boldsymbol{x}) = w(\boldsymbol{x}) \sum a_{\boldsymbol{i}}\, \varphi_{\boldsymbol{i}}(\boldsymbol{x})\,,$$

where the summation sign denotes a multiple sum whose indices i_j are going from 0 to n_j, $j = 0, \ldots, p$. Given a simple random sample, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, we define the resulting multivariate orthogonal polynomial density estimate to be

$$\hat{f}_{\boldsymbol{n}}(\boldsymbol{x}) = w(\boldsymbol{x}) \sum \hat{a}_{\boldsymbol{i}}\, \varphi_{\boldsymbol{i}}(\boldsymbol{x}) \tag{2.3.9}$$

where $\boldsymbol{n} = (n_1, \ldots, n_p)'$ and

$$\hat{a}_{\boldsymbol{i}} = \frac{1}{n} \sum_{k=1}^{n} \varphi_{\boldsymbol{i}}(\boldsymbol{x}_k),$$

see Schwartz (1967) or Hall (1983) for selected results. This form assumes that the components of the base density are independently distributed. As previously explained, one can transform the data so that the components be uncorrelated, in which case $w(\boldsymbol{x})$ should prove to be a suitable initial density estimate.

The following theorem formally proves the equivalence between the approximated densities obtained in terms of standard polynomial adjustments and those resulting from linear combinations of orthogonal polynomials as adjustments for the univariate and bivariate cases. Thus, for a given target density function, an identical approximate density function can be obtained *without* having to generate a set of orthogonal polynomials from the selected base density function.

Theorem 2.3.1. *For a given base density function, the approximated density functions obtained by utilizing standard polynomial adjustments are mathematically equivalent to those obtained by making use of linear combinations of orthogonal polynomials as adjustments.*

Proof: The result is first established for the univariate case.

Let $\tilde{f}_n(x) = \psi(x) \sum_{k=0}^{n} c_k \, x^k$, where the c_k's are such that the first n moments associated with the approximated density function $\tilde{f}_n(x)$ agree with those of the target density $f(x)$. Let $\hat{f}_n(x) = \psi(x) \sum_{j=0}^{n} \eta_j H_j(x)$ where the $H_j(x)$'s are the orthogonal polynomials corresponding to the base density function $\psi(x)$ and the coefficients η_i are such that the first n moments of $\hat{f}_n(x)$ also coincide with those of $f(x)$, be also an approximation to $f(x)$.

Then,

$$\begin{aligned}
\hat{f}_n(x) &= \psi(x) \sum_{j=0}^{n} \eta_j \, H_j(x) \\
&= \psi(x) \sum_{j=0}^{n} \eta_j \, \Big(\sum_{k=0}^{j} \alpha_k \, x^k \Big) \\
&= \psi(x) \sum_{j=0}^{n} \sum_{k=0}^{j} \eta_j \, \alpha_k \, x^k \\
&= \psi(x) \sum_{k=0}^{n} \Big(\sum_{j=k}^{n} \eta_j \, \alpha_k \Big) x^k \qquad \text{as } \sum_{j=0}^{n} \sum_{k=0}^{j} \equiv \sum_{k=0}^{n} \sum_{j=k}^{n} \\
&\equiv \psi(x) \sum_{k=0}^{n} c_k \, x^k \qquad \text{with } c_k = \sum_{j=k}^{n} \eta_j \, \alpha_k \\
&\equiv \tilde{f}_n(x) \, . \qquad \square
\end{aligned} \tag{2.3.10}$$

It should be pointed out that the set of orthogonal polynomials generated from a *given* base density is unique and that such polynomials of degree less than or equal to n are linearly independent. Thus they are forming a basis for all standard polynomials of degree at most n. Accordingly, there is a single linear combination of such orthogonal polynomials that will be equal to a specific standard polynomial of degree n.

We now consider the bivariate case. Noting that $\sum_{i=0}^{p} \sum_{k=0}^{i} \equiv \sum_{k=0}^{p} \sum_{i=k}^{p}$ and $\sum_{j=0}^{q} \sum_{\ell=0}^{j} \equiv \sum_{\ell=0}^{q} \sum_{j=\ell}^{q}$, and letting $\tilde{f}_{p,q}(x, \, y) = \psi(x, \, y) \sum_{k=0}^{p} \sum_{\ell=0}^{q} c_{k,\ell} x^k y^\ell$ and $\hat{f}_{p,q}(x, \, y) = \psi(x, \, y) \sum_{i=0}^{p} \sum_{j=0}^{q} \eta_{i,j} H_{i,j}(x, \, y)$, one has

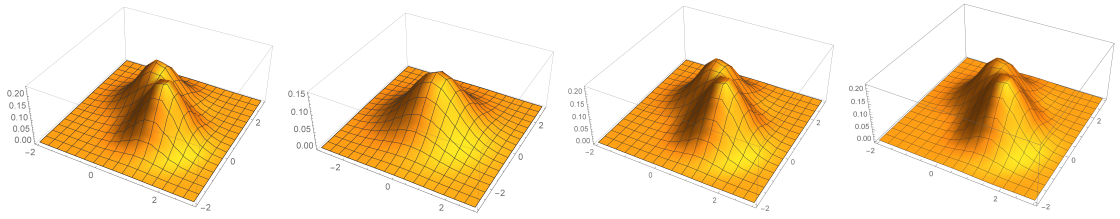
$$\begin{aligned}
\hat{f}_{p,q}(x,\,y) &= \psi(x,\,y)\sum_{i=0}^{p}\sum_{j=0}^{q}\eta_{i,j}H_{i,j}(x,\,y)\\
&= \psi(x,\,y)\sum_{i=0}^{p}\sum_{j=0}^{q}\eta_{i,j}\sum_{k=0}^{i}\sum_{\ell=0}^{j}\alpha_{ijk\ell}\,x^k y^\ell,\\
&= \psi(x,\,y)\sum_{i=0}^{p}\sum_{k=0}^{i}\sum_{j=0}^{q}\sum_{\ell=0}^{j}\eta_{i,j}\,\alpha_{ijk\ell}\,x^k y^\ell,\\
&= \psi(x,\,y)\sum_{k=0}^{p}\sum_{\ell=0}^{q}\Big\{\sum_{i=k}^{p}\sum_{j=\ell}^{q}\eta_{k,\,\ell}\,\alpha_{ijk\ell}\Big\}\,x^k y^\ell\\
&\equiv \psi(x,\,y)\sum_{k=0}^{p}\sum_{\ell=0}^{q}c_{k,\ell}\,x^k y^\ell\\
&= \tilde{f}_{p,q}(x,\,y). \qquad (2.3.11)
\end{aligned}$$

This demonstrates the mathematical equivalence between approximated densities obtained in terms of standard polynomial adjustments and those expressed in terms of linear combinations of orthogonal polynomials when, in each case, the highest degrees of x and y are respectively p and q. This equivalence can be similarly established in higher dimensions. □

The following example illustrates graphically that both approaches produce identical density approximants.

Example 2.3.1. *A mixture of bivariate normal densities.* Let

$$\mathbf{Z}_1 \sim \mathcal{N}_2\left(\begin{pmatrix}1.1\\-0.1\end{pmatrix},\begin{pmatrix}0.33 & 0.03\\0.03 & 0.33\end{pmatrix}\right),\mathbf{Z}_2 \sim \mathcal{N}_2\left(\begin{pmatrix}0.2\\1.2\end{pmatrix},\begin{pmatrix}0.4 & 0.04\\0.04 & 0.4\end{pmatrix}\right) \qquad (2.3.12)$$



(a) Exact density (b) Base density (c) $\tilde{f}_{6,6}(x,\, y)$ (d) $\hat{f}_{6,6}(x,\, y)$

Figure 2.2: Plots in connection with Example 2.3.1

and $f(x,\, y) = \frac{1}{2}(f_{\mathbf{Z}_1}(x,\, y) + f_{\mathbf{Z}_2}(x,\, y))$. The base density $\psi(x,\, y)$ is assumed to have the following distribution:

$$\mathcal{N}_2\left(\begin{pmatrix} 0.65 \\ 0.55 \end{pmatrix},\, \begin{pmatrix} 0.5675 & -0.2575 \\ -0.2575 & 0.7875 \end{pmatrix}\right), \tag{2.3.13}$$

whose mean and covariance matrix coincide with those of the mixture.

Although, increasing p and q improves the accuracy of the approximant, at some point, the improvement as quantified by the integrated squared differences between the exact and approximated density functions as defined by Equation (2.3.2), becomes minimal. Figure 2.2 displays the plots of the exact and base density functions as well as the approximated densities of $f(x,\, y)$ which were adjusted by making use of a bivariate standard polynomial and a linear combination of bivariate Hermite polynomials. For $p = q = 6$, it is seen that the approximants obtained from both types of adjustment are nearly identical to the exact density.

2.4 Density estimation and applications

In this section, we make use of the proposed approximation methodology in the context of density estimation. Let $\{(x_k,\, y_k),\, k = 1,\, \ldots, N\}$ be a dataset whose underlying density

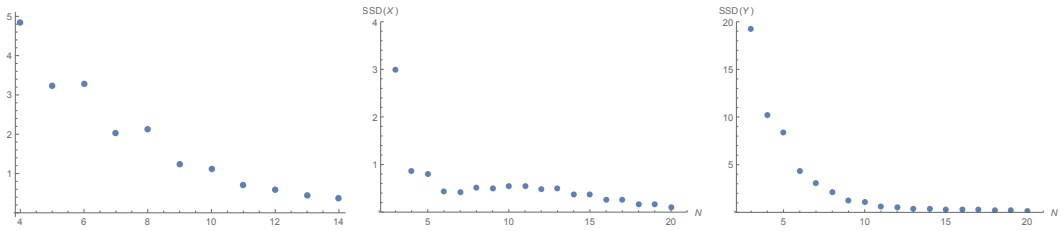
function has to be estimated. In this case, an appropriate initial density function can be chosen by inspecting a bivariate histogram (or scatter plot) of the data or histograms of the observations on each of the variables. For example, if the histograms of the x_k's and the y_k's are more or less normally distributed, then the base density function $\psi_X(x,\, y)$ can be taken to be that of a $\mathcal{N}_2(\hat{\boldsymbol{\mu}}, \hat{\Sigma})$ random vector where $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ are respectively the sample mean and sample covariance matrix. The polynomial adjustment is then determined from the joint sample moments, which are

$$\hat{\mu}_{i,j} = \frac{1}{N}\sum_{k=1}^{N} x_k^i\, y_k^j, \quad i,\, j = 0,\, 1,\, 2, \ldots . \tag{2.4.1}$$

Let ECDF$(x,\, y)$ denote the empirical CDF associated with the dataset and $F_{p,q}(x,\, y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{p,q}(x,y)\mathrm{d}y\mathrm{d}x$ be the CDF obtained from the estimated density function $f_{p,q}(x,\, y)$. The selection of the optimal maximal degree for each variable appearing in the polynomial adjustment, which are denoted by $p^\star$ and $q^\star$, is made in terms of the following sum of squared differences:

$$SSD(p,q) = \sum_{i=1}^{N}(\text{ECDF}(x_i,\, y_i) - F_{p,q}(x_i,\, y_i))^2, \tag{2.4.2}$$

which is used as a goodness-of-fit measure. The optimal degrees of p and q denoted by $p^\star$ and $q^\star$ are chosen to be those that minimize $SSD(p,q)$. Thus, the steps described in Algorithm 3.1 apply to the determination of a density estimate, except that sample moments are utilized in lieu of exact moments and the degree selection criteria is based on $SSD(p,q)$ rather than $ISD(p,q)$. Note that in light of Theorem 2.3.1, there is no need to resort to orthogonal polynomials.



(a) $SSD(n, n)$ vs n (b) $SSD_X(p)$ vs p (c) $SSD_Y(q)$ vs q

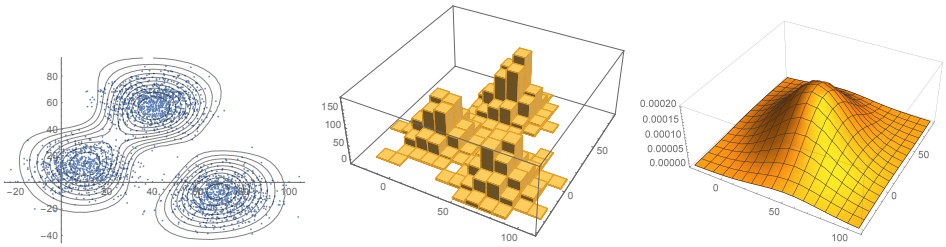
Figure 2.3: Graphs in connection with the selection of the optimal degree Example 2.4.1

Alternatively, it may prove simpler to consider $SSD_X(p)$ and $SSD_Y(q)$, which denote the sums of the squared differences between the ECDF and a CDF estimate obtained from a univariate density estimate that has been adjusted by means of a polynomial of degree p for x and a polynomial of degree q for y, and then select p^* and q^* corresponding to global minima or local minima beyond which further improvements are minimal. It would be appropriate to proceed in this manner whenever one of the variables requires a higher degree adjustment due to its more pronounced variability. This approach is utilized in the following examples.

Example 2.4.1. Consider the observations referred to as the *xclara* data set which is available from the *R* package data base (Struyf *et al.*, 1996).

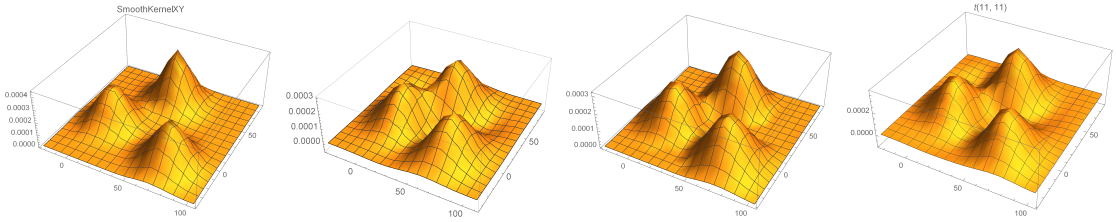
Plots of bivariate and univariate SSD's are shown in Figure 2.3. The graphs of SSD_X and SSD_Y indicate that one could take $p = 7$ and $q = 11$ as suitable degrees for the adjustments in x and y, respectively. In fact, $SSD(7, 11) = 0.73$ and $SSD(11, 11) = 0.74$ which, on taking parsimony into account, suggests that setting $p = 7$ and $q = 11$ is more than adequate, which is confirmed by the plots shown in Figure 2.5. Figure 2.4 contains a scatter plot and a histogram of the data as well as a plot of the base density.

Such density estimates could be utilized in the context of nonparametric regression. Consider for instance the regression model



(a) Scatter plot (b) 3D Histogram (c) Base density

Figure 2.4: Graphs in connection with Example 2.4.1



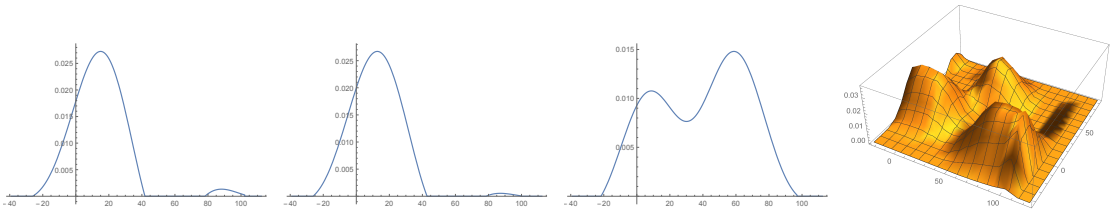
(a) KDE (b) $f_{7,7}(x,\, y)$ (c) $f_{7,11}(x,\, y)$ (d) $f_{11,11}(x,\, y)$

Figure 2.5: Kernel density estimate and $f_{p,q}(x,\, y)$ for $p = 7,\, 11$ and $q = 7,\, 11$

$$y = g(x) + \epsilon, \tag{2.4.3}$$

where $g(x)$ is an unknown continuous function and ϵ denotes a random error term. On the basis of an estimate of the joint density $f_{p,\,q}(x,\, y)$, one can readily determine the conditional density $f_{p,\,q}(y|X = x)$. Although the predicted values of the response variable are often determined by taking the expectation of the conditional density of Y given $X = x$, plots of the conditional density functions such as those appearing in Figure 2.6 prove significantly more informative.

The following example shows that the proposed methodology also applies to non-Gaussian type continuous distributions.



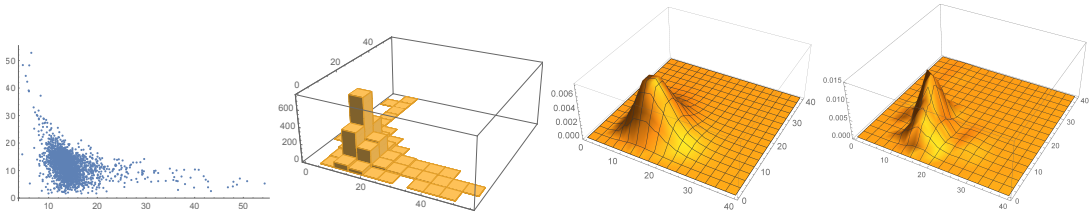
(a) $f_{11,7}(y|x=-15)$ (b) $f_{11,7}(y|x=4)$ (c) $f_{11,7}(y|x=25)$ (d) $f_{11,7}(y|x)$

Figure 2.6: The estimated conditional density function (Example 2.4.1)

Example 2.4.2. 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 "pctUrban": the percentage of people living in areas classified as urban as the X variable and "pctWFarmSelf": the percentage of households with a farm or self employment income in 1989 as the Y variable. Since the same observations are used for the variables X and Y in Example 6 wherein a third variable is also involved, we shall refer to certain figures pertaining to that example.

As can be seen from panels (a) and (b) of Figure 2.8, the marginal distribution of each variable behaves somewhat like a gamma random variable. Accordingly, the initial density estimate (plotted in panel (c) of Figure 2.7) was taken to be the product of two gamma density functions, that is, $\psi(x,\, y) = g_1(\alpha_1,\, \beta_1) \cdot g_2(\alpha_2,\, \beta_2)$ where the parameters were determined as follows: $\alpha_1 = \frac{\bar{x}^2}{\nu - \bar{x}^2} = 10.2251$ and $\beta_1 = \frac{\nu - \bar{x}^2}{\bar{x}} = 1.41277$, where $\bar{x}$ is the sample mean of X and $\nu = \frac{1}{N}\sum_{k=1}^{N} x_k^2$ is the second sample moment of X. The parameters $\alpha_2 = 6.14075$ and $\beta_2 = 1.92751$ were similarly obtained from the observations on Y. Figure 2.7 also includes a scatter plot and a three-dimensional histogram of the bivariate data.

The graphs of SSD_X and SSD_Y shown in Figure 2.8 suggest that $p = 7$ and $q = 4$



(a) Scatter plot (b) 3D Histogram (c) Base density (d) $f_{7,4}(x,\, y)$

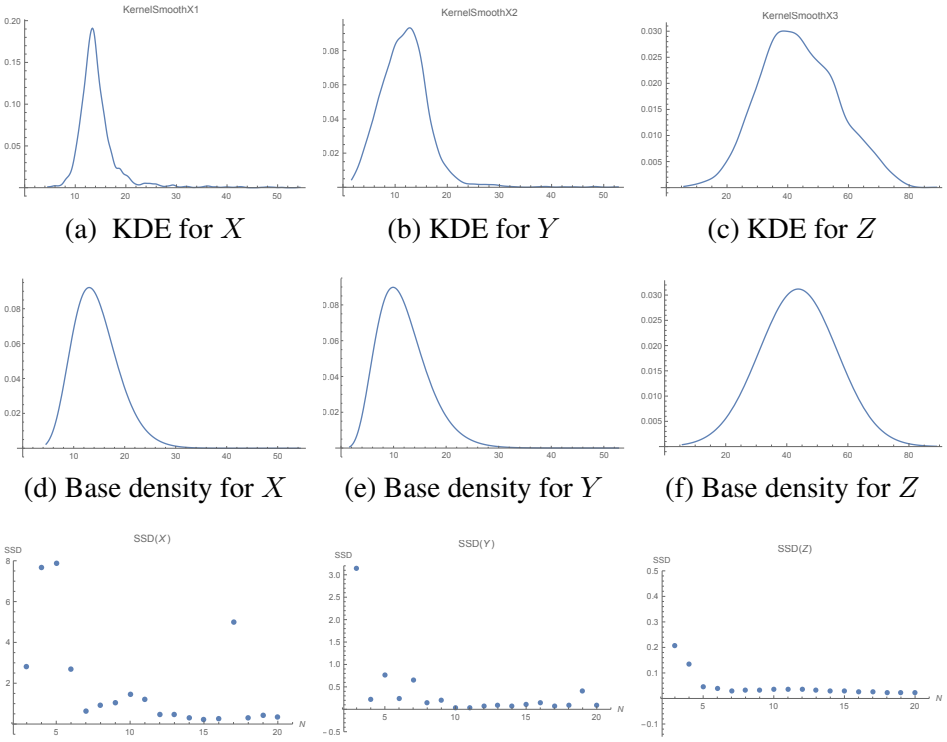
Figure 2.7: Graphs in connection with Example 2.4.2

would be suitable degrees. The resulting density estimate plotted in panel (d) of Figure 2.7 turns out to be similar to the kernel density estimate plotted in panel (a) of Figure 2.9.

Example 2.4.3. In this example, the approach being herein advocated is applied to a trivariate dataset. We modeled the data *CommViolPredUnnormalizedData* with the variables X and Y as previously defined in Example 2.4.2 and a third variable Z, "perCapInc": per capita income, which is modeled by a $\mathcal{N}(43.750, 163.531)$ random variable whose density function is plotted in panel (f) of Figure 2.8.

Figure 2.8 contains plots of the univariate KDE's and base densities for each of the three components. The graphs of the univariate SSD's, also shown in Figure 2.8, suggest that suitable degrees for the adjustments in x, y and z could respectively be 7, 4 and 5. Thus, the base density in x, y and z, that is, the product of the univariate base densities, is adjusted by means of a trivariate polynomial of degrees 7 for x, 4 for y and 5 for z. This yields a density estimate denoted by $f_{7,4,5}(x, y, z)$.

Since this estimated density function cannot be plotted, the bivariate marginal density estimates, that is, $f_{7,4,5}(x, y)$, $f_{7,4,5}(y, z)$ and $f_{7,4,5}(x, z)$ are shown in Figure 2.9, where for example $f_{7,4,5}(x, y) = \int f_{7,4,5}(x, y, z)\,\mathrm{d}z$. They are seen to be quite similar to the corresponding KDE's also plotted in the same figure. It should also be observed that density function of the marginal distribution X and Y, that is, $f_{7,4,5}(x, y)$ as plotted in panel (d) Figure 2.9 agrees with $f_{7,4}(x, y)$, which is the joint density function of X and Y



(a) KDE for X (b) KDE for Y (c) KDE for Z

(d) Base density for X (e) Base density for Y (f) Base density for Z

(g) $SSD_X(p)$ vs p (h) $SSD_Y(q)$ vs q (i) $SSD_Z(r)$ vs r

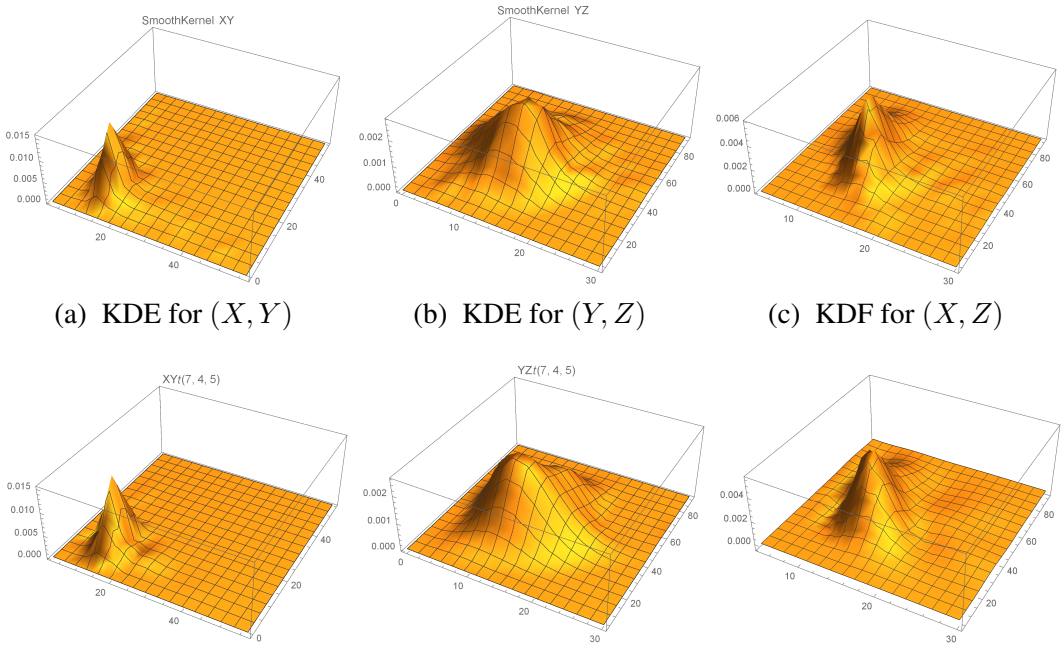
Figure 2.8: Univariate KDE's, base densities and SSD's in connection with Example 2.4.3

plotted in panel (d) of Figure 2.7.

The next example illustrates the usefulness of proposed methodology for modeling large datasets. This approach is most efficient as it only relies on the joint sample moments of a dataset, irrespective of its size.

Example 2.4.4. The dataset being considered, which is called *Covertype*, contains $581,012$ observations on 54 covariates. It was also 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 "Aspect": aspect in azimuth degrees as the X variable and "Slope": slope in degrees as the Y variable. Figure 2.10 displays a three-dimensional histogram of the data and the KDE's for X, Y and $(X,\ Y)$. Uniform and gamma base density functions



(a) KDE for (X, Y) (b) KDE for (Y, Z) (c) KDF for (X, Z)

(d) Marginal PDF $f_{7,4,5}(x, y)$ (e) Marginal PDF $f_{7,4,5}(y, z)$ (f) Marginal PDF $f_{7,4,5}(x, z)$

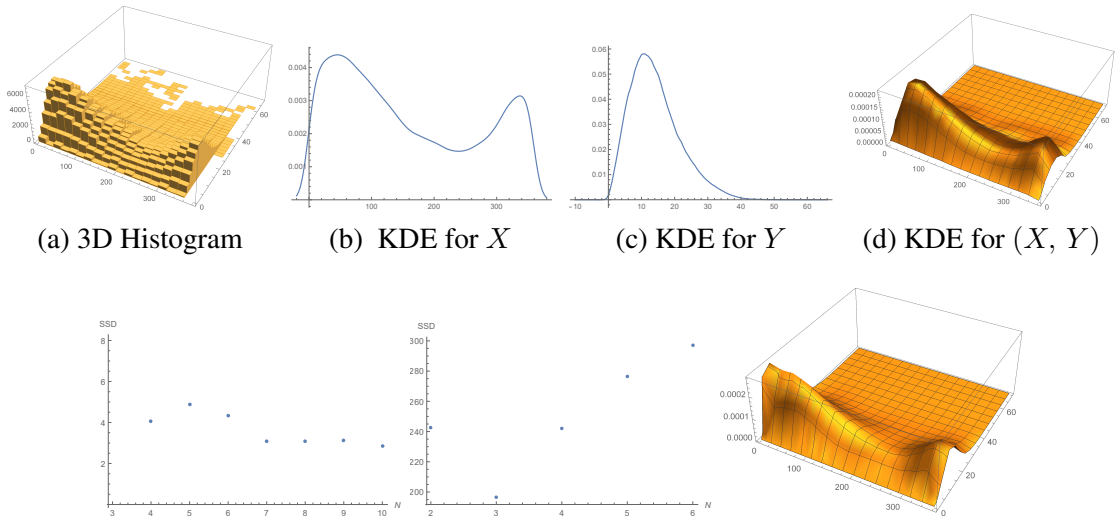
Figure 2.9: Plots of the bivariate KDE's and the bivariate marginals of the proposed density estimates in connection with Example 2.4.3

were used for X and Y respectively. The SSD values associated with X and Y whose minima are respectively 7 and 3, are shown in panels (e) and (f) of Figure 2.10, which also includes a plot of the estimated density function $f_{7,3}(x, y)$ whose features are similar to those exhibited by the kernel density estimate.

The next example shows the application of density estimation in nonparametric regression models.

Example 2.4.5. In this example, we consider the following nonparametric regression model

$$y_i = m(x_i) + \epsilon_i, \ i = 1, \ldots, n \tag{2.4.4}$$



(a) 3D Histogram (b) KDE for X (c) KDE for Y (d) KDE for (X, Y)

(e) $SSD_X(p)$ vs p (f) $SSD_Y(q)$ vs q (g) $f_{7,3}(x, y)$

Figure 2.10: Graphs in connection with Example 2.4.4

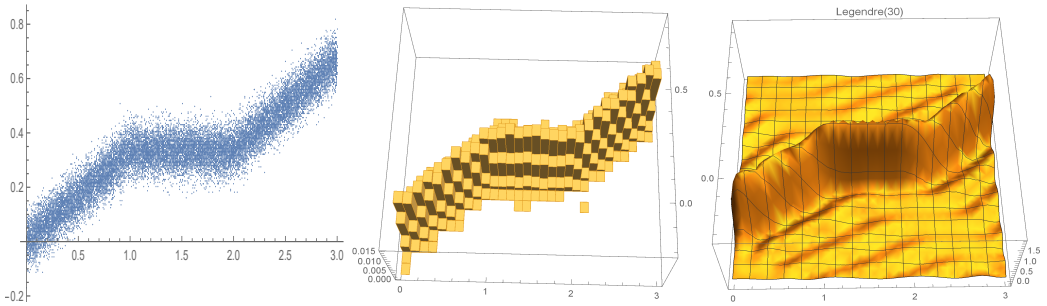
where

$$m(x) = \begin{cases} \frac{x}{3}, & \text{if } 0 \leq x < 1 \\ \frac{1}{3}, & \text{if } 1 \leq x < 2 \\ \frac{x-1}{3}, & \text{if } 2 \leq x \leq 3 \end{cases}$$

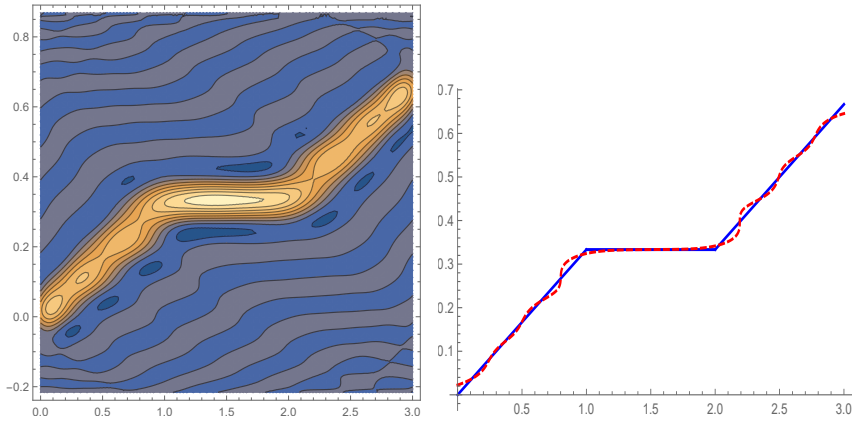
and ϵ_i's follows normal distribution $\mathcal{N}(0, 0.05)$.

The scatter plot of the $n = 2000$ points is shown in Panel (a) of Figure 2.11. The density function was obtained by making use of bivariate Legendre polynomials of orders at most 20 from Equation (2.3.9). The plot of the resulting bivariate density function, the contour plot and the projection of the PDF are shown in panels (b) to (d) in Figure 2.11, respectively.

Example 2.4.6. In this example, the dataset is a given set of 100000 points whose scatter plot shown in Panel (a) of Figure 2.12 represents a square within a triangle. The density function was obtained from Equation (2.3.9) by making use of bivariate Legendre polynomials of orders at most 40 . The plot of the resulting bivariate density function, the contour plot and the projection of the PDF are displayed in panels (b) to (d) in Figure 2.12, respectively.

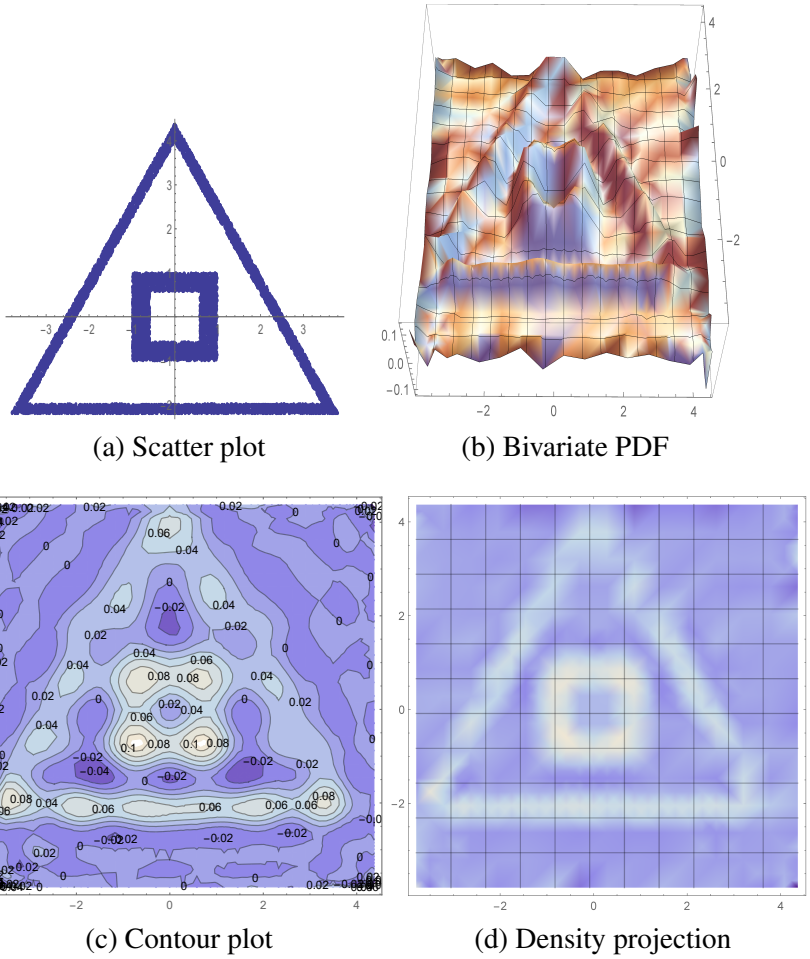


(a) Scatter plot (b) 3D Histogram (c) Bivariate PDF



(d) Contour plot (e) $\hat{m}(x)$ and $m(x)$

Figure 2.11: Scatter plot, Histogram, bivariate PDF, contour plot and estimated regression function in connection with Example 2.4.5



(a) Scatter plot (b) Bivariate PDF

(c) Contour plot (d) Density projection

Figure 2.12: Scatter plot, bivariate PDF, contour plot and density projection in connection with Example 2.4.6

2.5 Concluding remarks

This chapter introduces an efficient methodology for approximating the density functions of continuous random vectors, which makes use of a base density function and a moment-based polynomial adjustment. It was shown that on making use of a sequence of orthogonal polynomials associated with a given base density function, one can directly obtain density approximants or estimates without having to solve any system of linear equations. It was established that a density approximant whose adjustment is expressed in terms of a linear combination of orthogonal polynomials, is identical to that resulting from a standard polynomial adjustment. As well, the proposed approach has been successfully employed for modeling four data sets. The numerical examples illustrate its flexibility as well as its applicability in higher dimensions. All the calculations were carried out with the symbolic computation software *Mathematica*, the code being available from the authors upon request. As it turns out, the resulting density estimates possess a simple functional form that lends itself to algebraic manipulations. Furthermore, as was explained in the Introduction, the proposed methodology is ideally suited to model the massive data sets occurring nowadays in genomics, meteorology and numerous other fields of scientific investigation.

References

Alexits, G. (1961). Convergence Problems of Orthogonal Series, Pergamon Press, New York.

Blackard, J.A. and 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.

Charlier, C.V.L. (1906). Über die Darstellung willkürlicher Funktionen, Arkiv für Mathematik Astronomi och Fysik, 2, 1–35.

Daniels, H.E. (1954). Saddlepoint approximation in statistics, The Annals of Mathematical Statistics, 25(4), 631–659.

Efromovich, S. (1999). Nonparametric Curve Estimation. Springer, New York.

Edgeworth, F.Y. (1905). The law of error, Cambridge Philosophical Society Transactions, 20, 33–66.

Elderton, W.P. and Johnson, N.L. (1969). Systems of Frequency Curves, Cambridge University Press, Oxford.

Gajek, L. (1986). On improving density estimators which are not bona fide functions, The Annals of Statistics, 14(9), 1612–1618.

Hall, P. (1983). Orthogonal series distribution function estimation, with applications, Journal of Royal the Royal Statistical Society, Series B, 45, 81–88.

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

Provost, S.B. and Ha, H.T. (2009). On the inversion of certain moment matrices, Linear Algebra and its Application, 430, 2650–2658.

Rao, C.R. (2001). Linear Statistical Inference and Its Application, Second Edition, Wiley, New York.

Rayner, J.C., Thas, O., Pipelers, P. and Beh, E.J. (2013). Calculating bivariate orthonormal polynomials by recurrence, Australian & New Zealand Journal of Statistics, 55(1), 15–24.

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), 213–238.

Sansone, G. (2004). Orthogonal Functions, Dover Publications, New York.

Schwartz, S.C. (1967). Estimation of probability density by an orthogonal series, The Annals of Mathematical Statistics, 38, 1261–1265.

Solomon, H., and Stephens, M. (1978). Approximations to density functions using Pearson curves, Journal of the American Statistical Association, 73(361), 153–160.

Struyf, A., Hubert, M. and Rousseeuw, P. J. (1996). Clustering in an object-oriented environment. Journal of Statistical Software 1. (http://www.jstatsoft.org/v01/i04)

Szegö, G. (1959). Orthogonal Polynomials. American Mathematical Society, New York.

Willink, R. (2005). Normal moments and Hermite polynomials, Statistics and Probability Letters, 73(3), 271–275.

Withers, C. S. and Nadarajah, S. (2010). Expansions for the multivariate normal, Journal of Multivariate Analysis, 101(5), 1311–1316.

Appendices

A .1. Some useful univariate orthogonal polynomials

Among the widely used orthogonal polynomials, the Legendre, Laguerre, Jacobi and Hermite polynomials are of particular interest in connection with the approximation of density functions. Some of their mathematical properties are included in this section.

- *Legendre Polynomials*:

  This class of orthogonal polynomials is defined on $[-1,\ 1]$ as

  $$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2-1)^n,\ \ n = 0,\ 1, \ldots,$$

  or

  $$P_n(x) = \sum_{i=0}^{n}(-1)^i \binom{n}{i}((1+x)/2)^{n-i}((1-x)/2)^i.$$

  They satisfy the recurrence relation

  $$(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x),$$

  the first five being $P_0(x) = 1,\ P_1(x) = x,\ P_2(x) = \frac{1}{2}(3x^2-1),\ P_3(x) = \frac{1}{2}(5x^3 - 3x),\ P_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3)$, and the orthogonality property

  $$\int_{-1}^{1} P_i(x)P_j(x)\mathrm{d}x = \frac{2}{2n+1}\delta_{ij},$$

  where $\delta_{ij} = 1$ when $i = j$ and 0 otherwise.

- *Laguerre Polynomials*:

  The support of these orthogonal polynomials is $[\,0,\,\infty)$. They are given by

$$L_n(\alpha, x) = \frac{x^{-\alpha}e^x}{n!}\frac{d^n}{dx^n}(e^{-x}x^{n+\alpha}),\;\; n = 0,\,1,\ldots$$

  or

$$L_n(\alpha, x) = \sum_{i=0}^{n}(-1)^i\binom{n+\alpha}{n-i}\frac{x^i}{i!},$$

  and satisfy the recurrence relation

$$(n+1)L_{n+1}(\alpha, x) = (2n+\alpha+1-x)L_{\alpha,x}(x) - (n+\alpha)L_{n-1}(\alpha, x),$$

  the first few generalized Laguerre polynomials being $L_0(\alpha, x) = 1$, $L_1(\alpha, x) = -x+1$, $L_2(\alpha, x) = x^2/2 - (\alpha+2)x + (\alpha+2)(\alpha+1)/2$, as well as the

  orthogonality property

$$\int_0^\infty x^{-\alpha}e^{-x}L_i(\alpha, x)L_j(\alpha, x)\mathrm{d}x = \frac{\Gamma(n+\alpha+1)}{n!}\delta_{ij}.$$

- *Jacobi Polynomials*:

  These orthogonal polynomials which are defined on $[\,-1,\,1]$ can be determined as follows:

$$J_n(\alpha,\,\beta, x) = \sum_{i=0}^{n}\binom{n+\alpha}{i}\binom{n+\beta}{n-i}((x+1)/2)^i((x-1)/2)^{n-i};$$

they satisfy the orthogonality property

$$\int_{-1}^{1}(1-x)^{\alpha}(1+x)^{\beta}J_i(\alpha,\,\beta,x)J_j(\alpha,\,\beta,x)\mathrm{d}x = \frac{2^{\alpha+\beta+1}\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}{(2n+\alpha+\beta+1)\Gamma(n+\alpha+\beta+1)n!}\delta_{ij}.$$

- *Hermite Polynomials*:

  The modified Hermite polynomials are defined on $(-\infty,\,\infty)$; they can be obtained as follows:

$$H_n(x) = (-1)^n e^{\frac{x^2}{2}}\frac{\mathrm{d}^n}{\mathrm{d}x^n}e^{-\frac{x^2}{2}}.$$

  They satisfy the recurrence relation

$$H_{n+1}(x) = xH_n(x) - nH_{n-1}(x)$$

  and orthogonality property

$$\int_{-\infty}^{\infty} e^{-x^2/2}H_i(x)H_j(x)\mathrm{d}x = \sqrt{2\pi}n!\,\delta_{ij}.$$

A .2. Density approximation by means of Legendre and Laguerre polynomials

A .2.1. Legendre polynomials for densities having a compact support

Provost (2005) suggested to use Legendre polynomials to approximate densities with compact supports. For simplicity, at first, we assume that the density function $f_X(x)$ is defined on $[\,-1,\,1]$. Next, we extend the results to an arbitrary closed support $[\,a,\,b]$ by applying a simple linear transformation. We have

$$f_X(x) = \sum_{k=0}^{\infty} \lambda_k p_k(x) \tag{A.2.1}$$

where $p_k(x)$ is a Legendre polynomial of degree k in x, that is,

$$p_k(x) = \sum_{i=0}^{floor[\,k/2]} (-1)^i 2^{-k} \frac{(2k-2i)!}{i!(k-i)!(k-2i)!} x^{k-2i}, \tag{A.2.2}$$

$floor[\,k/2]$ being the largest integer not greater than $k/2$ and

$$\lambda_k = \frac{2k+1}{2} \sum_{i=0}^{floor[\,k/2]} (-1)^i 2^{-k} \frac{(2k-2i)!}{i!(k-i)!(k-2i)!} \mu_X(k-2i) = \frac{2k+1}{2} p_k^* \tag{A.2.3}$$

with $p_k^* = p_k(x)$ wherein X^{k-2i} is replaced by the $(k-2i)^{\text{th}}$ moment of X, which is given by

$$\mu_X(k) = E(X^k) = \int_{-1}^{1} x^k f_X(x)\,\mathrm{d}x,\ k = 0,\ 1,\ 2, \ldots\ .$$

To approximate $f_X(x)$ on the basis of $\mu_X(0) = 1,\ \mu_X(1),\ \ldots, \mu_X(n)$, we make use of the truncated series

$$f_{X_n}(x) = \sum_{k=0}^{n} \lambda_k\, p_k(x).$$

To extend this approximation technique to a density function $f_Y(y)$ having an arbitrary compact support $[\,a,\ b]$ with k^{th} moment,

$$\mu_Y(k) = E(X^k) = \int_a^b y^k\, f_Y(y)\mathrm{d}y,\ k = 0,\ 1,\ 2, \ldots\ ,$$

the following linear transformation is required:

$$X = \frac{2Y - (a+b)}{b-a}$$

the support of X being $[\,-1,\,1]$. After some algebra, one obtains the following approximation:

$$f_{Y_n}(y) = \frac{2}{b-a}\sum_k \lambda_k\, p_k\Big(\frac{2y-(a+b)}{b-a}\Big)$$

where $\lambda_k = \frac{2k+1}{2} p_k((2y-(a+b))/(b-a))$ wherein y^j is replaced by $\mu_Y(j)$.

A .2.2. Laguerre polynomials for densities defined on the positive half-line

Laguerre polynomials are appropriate for approximating the density functions of many statistics defined on the interval $[\,a,\,\infty)$, such as shifted gamma distributions.

Suppose a random variable Y is defined on the interval $[\,a,\,\infty)$ where $0 \le a < \infty$ and denote its j^{th} moment by $\mu_Y(j),\ \ j = 0,\,1,\,2,\,\ldots$. Define

$$c = \frac{\mu_Y(2) - \mu_Y(1)^2}{\mu_Y(1) - a} \quad \text{and} \quad \nu = \frac{\mu_Y(1) - a - c}{c},$$

and set $X = \frac{Y-a}{c}$.

Then, X has the support $[\,0,\,\infty)$ and its density function has the following representation:

$$f_X(x) = x^\nu e^{-x} \sum_{k=0}^{\infty} \delta_j L_j(\nu, x), \tag{A.2.4}$$

where

$$L_j(\nu, x) = \sum_{k=0}^{j} (-1)^k \frac{\Gamma(\nu+j+1)x^{j-k}}{k!(j-k)!\Gamma(\nu+j-k+1)}$$

is a Laguerre polynomial of order j in x with parameter ν and

$$\delta_j = \sum_{k=0}^{j} (-1)^k \frac{j!}{k!(j-k)!\Gamma(\nu+j-k+1)} \mu_X(j-k),$$

which can be expressed as $\delta_j = \frac{j!}{\Gamma(\nu+j+1)} L_j(\nu, x)$ wherein X^k is replaced by $\mu_X[\,k\,]$. Since $Y = cX + a$, we then have

$$\begin{aligned} f_{Y_n}(y) &= (y-a)^{\nu} e^{-\frac{y-a}{c}} \sum_{j=0}^{n} \delta_j L_j(\nu, \frac{y-a}{c}) \\ &= f_{Y_0}(y) \sum_{j=0}^{n} \omega_j L_j(\alpha - 1, \frac{y-a}{\beta}) \end{aligned} \tag{A.2.5}$$

where $\alpha \equiv \nu + 1 = \frac{(\mu_Y(1)-a)^2}{(\mu_Y(2)-\mu_Y(1))^2}$, $\beta \equiv c = \frac{\mu_Y(2)-\mu_Y(1)^2}{\mu_Y(1)-a}$, $f_{Y_0(y)}$ is a gamma distribution with parameters α and β and $\omega_j = \Gamma(\alpha)\delta_j$.

For further results on the approximation of various types of density functions by means of orthogonal and standard polynomials, the reader is referred to Provost (2005). Note that density approximations expressed in terms of modified Hermite polynomials are discussed in Section 2.1.

Chapter 3

An explicit representation of differentiated log-density approximants expressed as rational functions

3.1 Introduction

This chapter focuses on a density approximation (estimation) methodology that yields what is referred to as differentiated log-density approximants. Under this approach, the derivative of the logarithm of the density function $f_X(x)$ is represented as a rational function. This constitutes an extension of Pearson's frequency curve system wherein the density function of a continuous distribution is assumed to satisfy the following equation:

$$\frac{\mathrm{d}}{\mathrm{d}x}\log(f_X(x)) = \frac{-a_0 + x}{c_0 + c_1 x + c_2 x^2}\,, \tag{3.1.1}$$

where the parameters a_0, c_0, c_1 and c_2 are real numbers that can be determined from the first four moments of the distribution. The collection of all the distributions satisfying Equation (3.1.1) is called the family of Pearson distributions, which are described in Elderton (1938) and have been studied by Cramer (1946), Elderton (1953), Bol'shev (1963), Johnson and Kotz (1970), Solomon and Stephens (1978) and Stuart and Ord (1987), among others. Several widely used statistical distributions such as the gamma, beta, normal and uniform belong to this family.

As an extension of (3.1.1), consider the generalized Pearson's system (GPS) wherein approximants satisfy the following differential equation:

$$\frac{\mathrm{d}}{\mathrm{d}x}\log(f_X(x)) = \frac{f_X'(x)}{f_X(x)} = \frac{a_0 + a_1x + \ldots + a_\nu x^\nu}{c_0 + c_1x + \ldots + c_\delta x^\delta}, \tag{3.1.2}$$

where $a_0 + a_1x + \ldots + a_\nu x^\nu \equiv N_\nu(x)$ and $c_0 + c_1x + \ldots + c_\delta x^\delta \equiv D_\delta(x)$, the coefficients a_i's and c_j's being treated as parameters. The probability density function (PDF) $f_X(x)$ generated from (3.1.2) shall be referred as differentiated log-density approximant (DLDA) or a generalized Pearson's density curve. Several authors considered some special cases of (3.1.2) leading to certain PDF's of interest. For instance, an extension of the Pearson family of distributions that was considered by Cobb *et al.* (1983), can generate multimodal univariate distributions when the degree of the numerator is greater than one and the denominator is selected as one of the following polynomials: $D_0(x) = 1$, $-\infty < x < \infty$; $D_1(x) = x$, $0 < x < \infty$; $D_2(x) = x^2$, $0 < x < \infty$; $D_2(x) = x(1-x)$, $0 < x < 1$.

Rossani and Scarfone (2009) solved the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}x}\log(f_X(x)) = \frac{a_0 + a_1x + a_2x^2}{c_0 + c_1x + c_2x^2},$$

in connection with the monitoring of the interactions of the charged particles in an electric or magnetic field. Shakil *et al.* (2010) considered the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}x}\log(f_X(x)) = \frac{a_0 + a_1x + a_2x^2}{c_1x}, \quad c_1 \neq 0, \tag{3.1.3}$$

whose solution is the PDF,

$$f_X(x) = \kappa\, x^{\alpha} e^{-\mu x^2 - \beta x} \tag{3.1.4}$$

with $\mu = -a_2/2c_1$, $\alpha = a_0/b_1$, $\beta = -a_1/b_1$, κ being the normalizing constant. Hamedani (2011) obtained the following PDF:

$$f_X(x) = \kappa\, \nu\, x^{-(\nu+1)}(\beta - \alpha\, x^{2\nu}) e^{-\alpha x^{\nu} - \beta x^{-\nu}}, \quad 0 < x < \big(\frac{\beta}{\alpha}\big)^{\frac{1}{2\nu}}, \tag{3.1.5}$$

where $\alpha,\ \beta,\ \nu > 0$, κ being the normalizing constant, which satisfies the differential equation:

$$\frac{\mathrm{d}}{\mathrm{d}x}\log(f_X(x)) = \frac{\nu\beta^2 - (\nu+1)\beta x^{\nu} - 2\nu\alpha\beta x^{2\nu} - (\nu-1)\alpha x^{3\nu} + \nu\alpha^2 x^{4\nu}}{\beta x^{\nu+1} - \alpha x^{3\nu+1}}. \tag{3.1.6}$$

For other studies in connection with generalizations of the Pearson family of distributions, the interested reader is referred to (Dunning and Hanson, 1977), (Shakil, 2016), (Lefevre *et al.*, 2002), among others.

The traditional Pearson curves for which the degrees of numerator and denominator are respectively one and two while the coefficients of interest in (3.1.2) are not specified in advance are a very particular case of the proposed DLDA's, which results in more accuracy.

A review of different density estimation methodologies is provided in (Silverman, 1986). Reid (1988) discussed the saddlepoint technique to approximate a target density. Certain moment-based techniques were discussed for instance in (Elderton and Johnson,

1969), (Solomon and Stephens, 1978), (Provost, 2005) and (Provost and Ha, 2015). Extensions to bivariate and multivariate density estimates or approximants are discussed in (Scott, 2015), (Zareamoghaddam *et al.*, 2017a) and (Zareamoghaddam *et al.*, 2017b).

This chapter is organized as follows: The DLDA methodology is described in the next section. An explicit expression for the density approximant is derived in Section 3.3 where some special cases of interest are considered as well. In Section 3.4, we carry out a Monte Carlo simulation study involving several well-know distributions to illustrate the efficiency of DLDA methodology. Some concluding remarks are included in the last section.

3.2 Differentiated log-density approximation

This section describes the DLDA technique as applied to the approximation of continuous density functions. 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 the logarithm of $f_X(x)$ can be represented by a rational function, that is,

$$\frac{\mathrm{d}}{\mathrm{d}x}\log(f_X(x)) = \frac{f_X'(x)}{f_X(x)} = r(x), \tag{3.2.1}$$

where

$$r(x) = \frac{\sum_{i=0}^{\nu} a_i x^i}{\sum_{j=0}^{\delta} c_j x^j} = \frac{N_\nu(x)}{D_\delta(x)}, \tag{3.2.2}$$

$N_\nu(x)$ and $D_\delta(x)$ being polynomials in x of orders ν and δ . We shall assume without any loss of generality that c_δ is equal to one. After determining the a_i's and c_j's, by solving the

linear system specified by Equation (3.2.7), which involves a certain number of moments of the target distribution, $f_X(x)$ is approximated as

$$f_{\nu,\delta}(x) = \kappa e^{\int_\alpha^x r(y)\,\mathrm{d}y} \tag{3.2.3}$$

where κ is the normalizing constant.

In light of Equations (3.2.1) and (3.2.2), one has

$$f_{\nu,\delta}(x)\sum_{i=0}^{\nu} a_i\, x^i = f'_{\nu,\delta}(x)\sum_{j=0}^{\delta} c_j\, x^j, \tag{3.2.4}$$

from which the polynomial coefficients can be obtained as follows: Multiplying both sides of Equation (3.2.4) by x^h and integrating over the interval $(\alpha,\, \beta)$ yields

$$\int_\alpha^\beta f_{\nu,\delta}(x)\sum_{i=0}^{\nu} a_i\, x^{i+h}\mathrm{d}x = \int_\alpha^\beta f'_{\nu,\delta}(x)\sum_{j=0}^{\delta} c_j\, x^{j+h}\mathrm{d}x,\ h = 0,\, 1, \ldots,\, \nu+\delta; \tag{3.2.5}$$

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

$$\begin{aligned}\sum_{i=0}^{\nu} a_i \int_\alpha^\beta x^{i+h} f_{\nu,\delta}(x)\mathrm{d}x =& f_{\nu,\delta}(x)\sum_{j=0}^{\delta} c_j x^{j+h}\ |_\alpha^\beta \\ &- \sum_{j=0}^{\delta} c_j(j+h)\int_\alpha^\beta x^{j+h-1} f_{\nu,\delta}(x)\mathrm{d}x,\ h = 0,\, 1, \ldots,\, \nu+\delta.\end{aligned} \tag{3.2.6}$$

Thus, letting μ_h, $h = 0,\, 1, \ldots,\, \nu + \delta$, denote the h^{th} moment of the approximated density function $f_X(x)$, one obtains $\nu + \delta + 1$ linear equations of the following form:

$$\sum_{i=0}^{\nu} a_i\, \mu(i+h) = \sum_{j=0}^{\delta} c_j \left(f_{\nu,\delta}(\beta)\, \beta^{j+h} - f_{\nu,\delta}(\alpha)\, \alpha^{j+h} \right) - \sum_{j=0}^{\delta} c_j (j+h)\, \mu(j+h-1),\ \ h = 0,\, 1, \ldots,\, \nu + \delta, \tag{3.2.7}$$

where $\mu_0 \equiv 1$. In order to determine the unknown coefficients of $r(x)$ as specified by Equation (3.2.2), one needs to solve the linear system resulting from Equation (3.2.7). By replacing the unknown $\mu(h)$ by $\mu_X(h)$, for $h = 0,\, 1, \ldots,\, \nu + \delta$, where $\mu_X(h)$ denotes the h^{th} moment of the distribution being approximated, one obtains the following system of linear equations:

$$\sum_{i=0}^{\nu} a_i\, \mu_X(i+h) = \sum_{j=0}^{\delta} c_j \left(f_{\nu,\delta}(\beta)\, \beta^{j+h} - f_{\nu,\delta}(\alpha)\, \alpha^{j+h} \right) - \sum_{j=0}^{\delta} c_j (j+h)\, \mu_X(j+h-1),\ \ h = 0,\, 1, \ldots,\, \nu + \delta, \tag{3.2.8}$$

where in most cases of interest $f_{\nu,\delta}(\alpha)$ and $f_{\nu,\delta}(\beta)$ can be set equal to zero.

Once the solution of this linear system is obtained, one may solve the differential equation

$$f'_{\nu,\delta}(x) = r(x) f_{\nu,\delta}(x), \tag{3.2.9}$$

where $r(x) = \sum_{i=0}^{\nu} a_i x^i / \sum_{j=0}^{\delta} c_j x^j$, by making use of symbolic computation packages

such as *Mathematica* or *Maple*. Equivalently, one can evaluate the right-hand side of Equation (3.2.3), which as shown in the next section, provides an explicit solution to the differential equation specified by Equation (3.2.9).

Remark 1. The degree of the denominator $D_\delta(x)$ in (3.2.2), that is δ, corresponds to the number of times the density function $f_X(x)$ intersects the abscissa plus the number of points at which this density function is not differentiable. Note that in light of Equation (3.2.3), it is seen that the integrand will become infinite at the roots of $D_\delta(x)$ so that the approximant will be zero at those points. Moreover, since a simple polynomial cannot adequately account for abrupt changes in the slope of a function, one needs to include the number of points of non-differentiability in the set of roots of $D_\delta(x)$. Also, increasing the degree of numerator $N_\nu(x)$ generally leads to more accurate approximations.

3.3 An explicit representation of the density approximant

The following lemmas provide an explicit representation of the density approximation $f_{\nu,\delta}(x)$.

Lemma 3.3.1. *Let* $\lambda_1, \lambda_2, \ldots, \lambda_\delta$ *be the distinct real roots of a polynomial* $D_\delta(y) = \sum_0^\delta c_i\, y^i$, $c_\delta \neq 0$, *and* $M(y) = \sum_{i=0}^{\gamma} a_i\, y^i$, $a_\gamma \neq 0$, *with* $\gamma < \delta$, *be another polynomial. Then, one has*

$$\frac{M(y)}{D_\delta(y)} = \sum_{k=1}^{\delta} \frac{\tau_k}{y - \lambda_k} \tag{3.3.1}$$

where $\tau_k = M(\lambda_k) / \prod_{\ell \neq k} (\lambda_k - \lambda_\ell)$.

Proof: Let

$$\frac{\sum_{i=0}^{\gamma} a_i\, y^i}{\prod_{k=1}^{\delta} (y-\lambda_k)} = \sum_{k=1}^{\delta} \frac{\tau_k}{y-\lambda_k}, \tag{3.3.2}$$

which can be re-expressed as

$$\frac{\sum_{i=0}^{\gamma} a_i\, y^i}{\prod_{k=1}^{\delta} (y-\lambda_k)} = \sum_{k=1}^{\delta} \frac{\tau_k\, A_k}{\prod_{k=1}^{\delta} (y-\lambda_k)}, \tag{3.3.3}$$

where clearly $A_k = \prod_{j\neq k} (y-\lambda_j)$. As the right and left hand sides of (3.3.3) are equal for each y value, one has

$$\sum_{i=0}^{\gamma} a_i\, y^i = \sum_{k=1}^{\delta} \tau_k \prod_{j\neq k} (y-\lambda_j).$$

On substituting λ_k to y in this equation, all the summands the right hand side will vanish except for the k^{th} term, which is $\tau_k \prod_{j\neq k} (\lambda_k-\lambda_j)$. Thus,

$$\tau_k = \Big(\sum_{i=0}^{\gamma} a_i\, \lambda_k^i\Big) \Big/ \prod_{j\neq k} (\lambda_k-\lambda_j).$$

□

Lemma 3.3.2. *Letting $x > \alpha$, $i \geq 2$ and A, B, a and b be constants. One has*

$$\int \frac{A\,y+B}{y^2+a\,y+b}\mathrm{d}y = \int \frac{A(2y+a)/2+(B-aA/2)}{y^2+a\,y+b}\mathrm{d}y$$
$$= \frac{A}{2}\log(y^2+a\,y+b) + (B-\frac{aA}{2})\int \frac{\mathrm{d}y}{y^2+a\,y+b} \qquad (3.3.4)$$

where one of the following three cases is possible depending on the solutions of the equation $y^2+a\,y+b=0$*:*

i. The equation $y^2+a\,y+b$ *has two real roots of* λ_1, λ_2 $(\lambda_1 \neq \lambda_2)$. *By setting* $m^2 = b-\frac{a^2}{4}$*, one has* $\lambda_1 = -a/2+m$ *and* $\lambda_2 = -a/2-m$*, and*

$$\int \frac{A\,y+B}{y^2+a\,y+b}\mathrm{d}y = \int \frac{E_1}{y-\lambda_1}\mathrm{d}y + \int \frac{E_2}{y-\lambda_2}\mathrm{d}y$$
$$= E_1\log|y-\lambda_1| + E_2\log|y-\lambda_2|$$
$$= \log\Big(|y-\lambda_1|^{E_1}|y-\lambda_2|^{E_2}\Big) \qquad (3.3.5)$$

where $E_1 = \frac{A\,\lambda_1+B}{\lambda_1-\lambda_2}$ *and* $E_2 = \frac{A\,\lambda_2+B}{\lambda_2-\lambda_1}$.

ii. The quadratic equation has one repeated real root, that is $\lambda = \lambda_1 = \lambda_2 = -a/2$ *where* $m=0$*. In that case,*

$$\int \frac{A\,y+B}{y^2+a\,y+b}\mathrm{d}y = \frac{A}{2}\log(y-\lambda)^2 + (B-\frac{aA}{2})\int \frac{\mathrm{d}y}{(y-\lambda)^2}$$
$$= A\log(y-\lambda) - (B-\frac{aA}{2})/(y-\lambda). \qquad (3.3.6)$$

iii. The quadratic equation does not have any real root. Then,

$$\begin{aligned}\int \frac{A\,y+B}{y^2+a\,y+b}\mathrm{d}y &= \frac{A}{2}\log(y^2+a\,y+b)+(B-\frac{aA}{2})\int \frac{\mathrm{d}y}{(y^2+a\,y+a^2/4)+b-\frac{a^2}{4}}\\ &= \frac{A}{2}\log(y^2+a\,y+b)+(B-\frac{aA}{2})\int \frac{\mathrm{d}y}{(y+a/2)^2+m^2}\\ &= \frac{A}{2}\log(y^2+a\,y+b)+(B-\frac{aA}{2})\,\frac{1}{m}\arctan\big(\frac{y+a/2}{m}\big).\end{aligned} \tag{3.3.7}$$

Now, letting $\lambda_1,\ \lambda_2,\ \ldots,\ \lambda_\delta$ denote the distinct real roots of the denominator $D_\delta(y)$ of the rational function specified in Equation (3.2.2), one can express $D_\delta(y)$ as a product of monomials, that is, $\prod_{k=1}^{\delta}(y-\lambda_k)$. If $\nu \geq \delta$, one has

$$N_\nu(y)/D_\delta(y) = Q_{\nu-\delta}(y) + R(y)/D_\delta(y), \tag{3.3.8}$$

where $Q_{\nu-\delta}(y) = \sum_{i=0}^{\nu-\delta} q_i\,y^i$ and $R(y) = \sum_{i=0}^{\gamma} d_i\,y^i$, $d_\gamma \neq 0$, $\gamma < \delta$, are the quotient and the remainder of the rational function $r(y)$, respectively. The coefficients q_i and d_i can be determined by making use of symbolic computation packages such as *Maple* or *Mathematica*. Note that $Q(y)$ will be equal to zero whenever $\nu < \delta$. Thus,

$$\begin{aligned} r(y) &= \frac{\sum_{i=0}^{\nu} a_i\,y^i}{\prod_{k=1}^{\delta}(y-\lambda_k)} = \sum_{i=0}^{\nu-\delta} q_i\,y^i + \frac{\sum_{i=0}^{\gamma} d_i\,y^i}{\prod_{k=1}^{\delta}(y-\lambda_k)}\\ &= \sum_{i=0}^{\nu-\delta} q_i\,y^i + \sum_{k=1}^{\delta}\frac{\tau_k}{y-\lambda_k}\end{aligned} \tag{3.3.9}$$

where

$$\tau_k = \Big(\sum_{i=0}^{\gamma} d_i\,\lambda_k^i\Big)\Big/\prod_{j\neq k}(\lambda_k-\lambda_j) \tag{3.3.10}$$

as shown in Lemma 3.3.1, and one has

$$\begin{aligned}\int_\alpha^x r(y)\,\mathrm{d}y &= \int_\alpha^x Q_{\nu-\delta}(y)\,\mathrm{d}y + \sum_{k=1}^{\delta}\int_\alpha^x \frac{\tau_k}{y-\lambda_k}\,\mathrm{d}y \\ &= \sum_{i=0}^{\nu-\delta} q_i\, y^{i+1}/(i+1)\;|_\alpha^x + \sum_{k=1}^{\delta}\tau_k \log|y-\lambda_k|\;|_\alpha^x \\ &= \sum_{i=0}^{\nu-\delta} q_i\left(x^{i+1}-\alpha^{i+1}\right)/(i+1) + \sum_{k=1}^{\delta}\left\{\log|x-\lambda_k|^{\tau_k} - \log|\alpha-\lambda_k|^{\tau_k}\right\} \\ &= \sum_{i=0}^{\nu-\delta} q_i\, x^{i+1}/(i+1) + \log\prod_{k=1}^{\delta}|x-\lambda_k|^{\tau_k} + C. \end{aligned} \tag{3.3.11}$$

where $C = -\sum_{i=0}^{\nu-\delta} q_i\,\alpha^{i+1}/(i+1) - \log\prod_{k=1}^{\delta}|\alpha-\lambda_k|^{\tau_k}$ is a constant.

Thus, using the representation of the approximate density function specified by Equation (3.2.3) that is $f_{\nu,\delta}(x) = \kappa\, e^{\int_\alpha^x \frac{N_\nu(y)}{D_\delta(y)}\,\mathrm{d}y}$, one has

$$\begin{aligned} f_{\nu,\delta}(x) &= \kappa e^{\sum_{i=0}^{\nu-\delta} q_i\, x^{i+1}/(i+1) + \log\prod_{k=1}^{\delta}|x-\lambda_k|^{\tau_k} + C} \\ &= \kappa' e^{\sum_{i=0}^{\nu-\delta} q_i\, x^{i+1}/(i+1)}\prod_{k=1}^{\delta}|x-\lambda_k|^{\tau_k}, \quad x\in(\alpha,\beta), \end{aligned} \tag{3.3.12}$$

where κ' is a normalizing constant (such that $\int_\alpha^\beta f_{\nu,\delta}(x)\,\mathrm{d}x = 1$), q_i is as defined in (3.3.8), q_i being zero whenever $\nu < \delta$, τ_k is defined in (3.3.10), and the λ_k's are the δ distinct roots of $D_\delta(x)$.

We now consider the special cases where $\delta \le 2$. Explicit representations of the approximated density function $f_{\nu,\delta}(x)$ for $\delta = 0,\ 1$ and 2 are specified below.

I. Letting $\delta = 0$, without any loss of generality, one has

$$\frac{f'_X(x)}{f_X(x)} = p_\nu(x) = \sum_{i=0}^{\nu} a_i x^i. \tag{3.3.13}$$

Thus, in light of (3.2.3), the approximated density function has the following representation:

$$f_{\nu,0}(x) = \kappa e^{\int_\alpha^x p_\nu(y)\,\mathrm{d}y} = \kappa' e^{\sum_{i=0}^{\nu} \frac{a_i}{i+1} x^{i+1}}, \tag{3.3.14}$$

where κ and κ' are the appropriate normalizing constants.

II. When $\delta = 1$, so that the denominator is a linear function of the form $D_1(x) = x - \lambda_1$, the rational function $r(x)$ has the following form:

$$r(x) = \frac{\sum_{i=0}^{\nu} a_i x^i}{x - \lambda_1} = Q_{\nu-1}(x) + \frac{\tau_1}{x - \lambda_1}, \tag{3.3.15}$$

where $Q_{\nu-1}(x) = \sum_{i=0}^{\nu-1} q_i x^i$. Accordingly, the approximated density function $f_{\nu,1}(x)$ is obtained as

$$f_{\nu,1}(x) = \kappa\,(x - \lambda_1)^{\tau_1} e^{\sum_{i=0}^{\nu-1} \frac{q_i}{i+1} x^{i+1}}, \tag{3.3.16}$$

where κ is the normalizing constant.

III. When $\delta = 2$, one can let $c_2 = 1$ without any loss of generality. Assuming that there are no repeated roots, the denominator has the form of $D_2(x) = x^2 + c_1\,x + c_0$ and

Equation (3.2.2) has the following representation

$$r(x) = \frac{\sum_{i=0}^{\nu} a_i x^i}{x^2 + c_1\, x + c_0} = Q_{\nu-2}(x) + \frac{A_1\, x + A_0}{x^2 + c_1\, x + c_0}. \tag{3.3.17}$$

$D_2(x) = x^2 + c_1\, x + c_0$ can have either two real roots or two complex roots. First, we assume that $D_2(x)$ has two real roots, λ_1 and λ_2, so that

$$\frac{A_1 x + A_0}{x^2 + c_1\, x + c_0} = \frac{\theta_1}{x - \lambda_1} + \frac{\theta_2}{x - \lambda_2},$$

with $\theta_1 = \frac{A_1\,\lambda_1 + A_0}{\lambda_1 - \lambda_2}$ and $\theta_2 = \frac{A_1\,\lambda_2 + A0}{\lambda_2 - \lambda_1}$ calculated through partial fractions. Then, from Equation (3.2.3), one has

$$\begin{aligned} f_{\nu,2}(x) &= \kappa e^{\int_\alpha^x Q_{\nu-2}(x)\,\mathrm{d}x + \int_\alpha^x \frac{\theta_1}{x-\lambda_1}\,\mathrm{d}x + \int_\alpha^x \frac{\theta_2}{x-\lambda_2}\,\mathrm{d}x}, \\ &= \kappa' e^{\sum_{i=0}^{\nu-2} \frac{q_i}{i+1} x^{i+1}} \prod_{k=1}^{2} |x - \lambda_2|^{\theta_i}, \end{aligned} \tag{3.3.18}$$

where $Q_{\nu-2}(x) = \sum_{i=0}^{\nu-2} q_i x^i$ and, κ and κ' are normalizing constants. The second possibility is that $x^2 + c_1\, x + c_0$ does not have any real root. We know that

$$f_{\nu,2}(x) = \kappa e^{\sum_{i=0}^{\nu-2} \frac{q_i}{i+1} x^{i+1}}\, e^{\int_\alpha^x \frac{A_1 x + A_0}{x^2 + c_1\, x + c_0}\,\mathrm{d}x}. \tag{3.3.19}$$

Thus, the density estimate obtained by evaluating the last integral in Equation (3.3.19) can be expressed as follows by making use of Equation (3.3.7) (Lemma 3.3.2):

$$f_{\nu,2}(x) = \kappa'(x^2 + c_1\, x + c_0)^{\frac{A_1}{2}}\, e^{\sum_{i=0}^{\nu-2} \frac{q_i}{i+1} x^{i+1} + \arctan\left(\frac{x + c_1/2}{\sqrt{c_0 - c_1^2/4}}\right)}, \tag{3.3.20}$$

with κ' being the new normalizing constant. On the other hand, when $D_2(x)$ has the single solution $\lambda = -c_1/2$, by making use of Equation (3.3.6) (Lemma 3.3.2), the

solution of the integral in Equation (3.3.19) is

$$\begin{aligned} f_{\nu,2}(x) &= \kappa e^{\sum_{i=0}^{\nu-2} \frac{q_i}{i+1} x^{i+1}} e^{\int_\alpha^x \frac{A_1 x + A_0}{x^2 + c_1 x + c_0} \, \mathrm{d}x} \\ &= \kappa'(x-\lambda)^{A_1} e^{\sum_{i=0}^{\nu-2} \frac{q_i}{i+1} x^{i+1} - \frac{1}{x-\lambda}}. \end{aligned} \tag{3.3.21}$$

When the DLDA is a differentiable function as is the case for instance of the right-hand side expression in Equation (3.3.18) which, assuming that the end points of the distribution are $\lambda_1 < \lambda_2$ and letting $\theta_j = \frac{q_{j-3}}{j-2}$, $j = 3, \ldots, \nu+2$, can be expressed as

$$\begin{aligned} f_{\nu,2}(x) &= \kappa'(x-\lambda_1)^{\theta_1}(\lambda_2 - x)^{\theta_2} e^{\sum_{i=0}^{\nu-1} \theta_{i+3} x^{i+1}} \\ &= e^{\sum_{j=1}^{\nu+1} p_j(\theta_1, ..., \theta_{j+1}) K_j(x) + q(\theta_1, ..., \theta_{j+1})} \end{aligned} \tag{3.3.22}$$

in the notation of Hogg and Craig (1978, p. 366), where $p_j(\theta_1, \ldots, \theta_{j+1}) = \theta_j$ for $j = 1, \ldots, \nu+1$, $K_1(x) = \log(x - \lambda_1)$, $K_2(x) = \log(\lambda_2 - x)$ and $K_j(x) = x^{j-2}$, $j = 3, \ldots, \nu+2$. Accordingly, on the basis of a sample of size n,

$$\sum_{i=1}^{n} \log(x_i - \lambda_1), \ \sum_{i=1}^{n} \log(\lambda_2 - x_i), \ \sum_{i=1}^{n} x_i, \ldots, \sum_{i=1}^{n} x_i^{\nu}$$

are joint sufficient statistics for $\theta_1, \ldots, \theta_{\nu+2}$, the parameters of the resulting density function. Joint sufficient statistics can be similarly determined for the parameters of the general form of the density approximant given in Equation (3.3.12).

3.4 Numerical results

In the first example, simulation study is carried out to compare the accuracy obtained from the density approximants determined by making use of the DLDA methodology in connection with the sample moments of the simulated distributions and the kernel density

estimation (KDE) technique. The integrated squared error

$$ISE(\hat{f}) = \int \Big(\hat{f}(x) - f(x)\Big)^2 \mathrm{d}x, \tag{3.4.1}$$

is utilized to measure the accuracy of the approximated density $\hat{f}(x)$.

Example 3.4.1. Let $n = 500,\ 1000,\ 2000$ observations be generated randomly from beta(2,10), gamma(2,20), exponential(2), Student $t(5)$ and normal distribution PDF's. The DLDA and KDE methodologies were applied to the generated random values by running Monte Carlo simulations 100 times. The means and standard deviation of the resulting ISE's are included in Table 3.1.

Table 3.1: Average ISE's (SD's in parentheses) for different distributions estimated by applying the DLDA and KDE techniques.

		$n = 500$	$n = 1000$	$n = 2000$
Beta	DLDA	0.0165 (0.0247)	0.0100 (0.0065)	0.0069 (0.0034)
	KDE	0.0270 (0.0142)	0.0158 (0.0079)	0.0101 (0.0041)
Gamma	DLDA	0.0001 (0.0001)	0.0000 (0.0000)	0.0000 (0.0000)
	KDE	0.0002 (0.0001)	0.0001 (0.0001)	0.0001 (0.0000)
Exponential	DLDA	0.0449 (0.0366)	0.0391 (0.0216)	0.0381 (0.0157)
	KDE	0.0641 (0.0148)	0.0544 (0.0077)	0.0457 (0.0062)
Student *t*	DLDA	0.0013 (0.0013)	0.0008 (0.0006)	0.0006 (0.0004)
	KDE	0.0020 (0.0011)	0.0010 (0.0006)	0.0006 (0.0003)
Normal	DLDA	0.001 (0.0004)	0.0003 (0.0002)	0.0002 (0.0001)
	KDE	0.0018 (0.0009)	0.0010 (0.0005)	0.0007 (0.0003)

Example 3.4.2. In this example, the following density function is considered:

$$f(x) = \begin{cases} x/2 & \text{if } 0 \le x \le 1 \\ 1/2 & \text{if } 1 \le x \le 2 \\ (3-x)/2 & \text{if } 2 \le x \le 3 \end{cases} \tag{3.4.2}$$

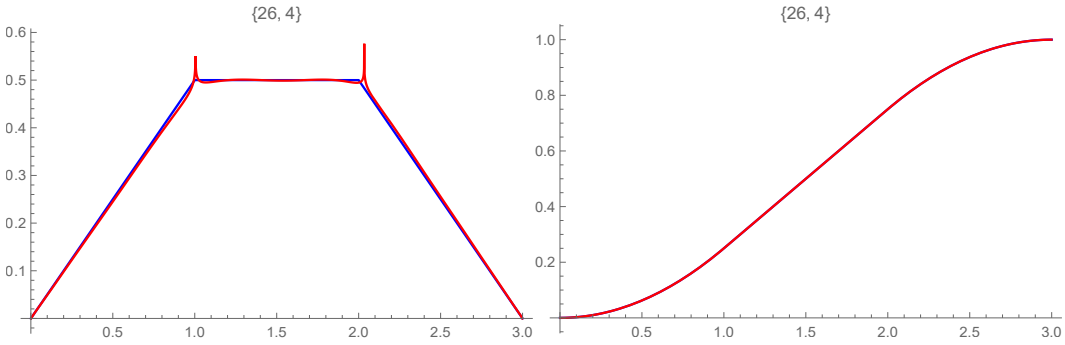


Figure 3.1: Exact and estimated PDF's and CDF's in connection with Example 3.4.2.

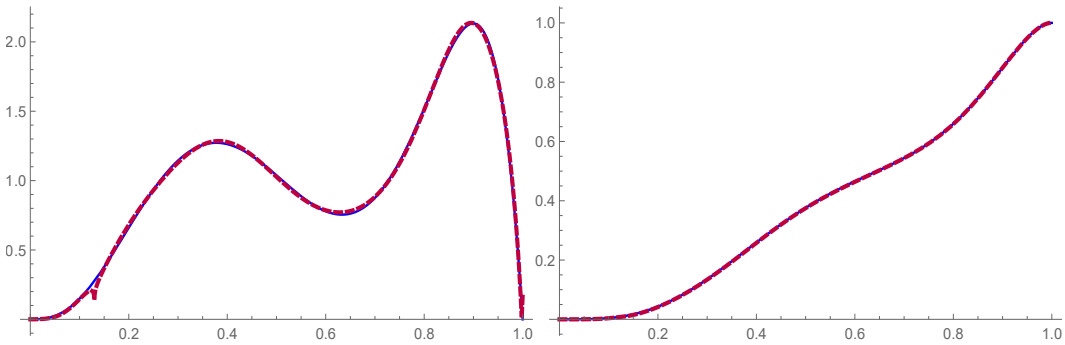


Figure 3.2: Exact and estimated PDF's and CDF's in connection with Example 3.4.3.

In order to approximate this density function, we let $\delta = 4$ be the degree of the denominator in accordance with Remark 1, noting that there are two points where the function is not differentiable and two points of intersection with the abscissa. The exact density and distribution functions as well as the DLDA density approximant $f_{26,4}(x)$ and the corresponding CDF approximation are shown in Figure 3.1.

Example 3.4.3. In this case, the target density function is taken as an equally weighted mixture of $Beta(10,\ 2)$ and $Beta(4,\ 6)$ PDF's. The degree of the denominator is set to two. The DLDA density approximant $f_{8,2}(x)$ and the exact density functions are plotted in Figure 3.2.

References

Bol'shev, L.N. (1963). Asymptotically Pearson transformations, Theory of Probability and its Applications, 8(2), 121–146.

Cobb, L., Koppstein, P., and Chen, N.H. (1983). Estimation and moment recursion relations for multimodal distributions of the exponential family. Journal of the American Statistical Association, 78(381), 124–130.

Cramer, H., (1946). Mathematical Methods of Statistics , Princeton University Press, Princeton.

Dunning, K., and Hanson, J. N. (1977). Generalized Pearson distributions and nonlinear programming. Journal of Statistical Computation and Simulation, 6(2), 115–128.

Elderton, W.P. (1953). Frequency Curves and Correlation, Harren Press, Washington DC.

Elderton, W.P. and Johnson, N.L. (1969). Systems of Frequency Curves, Cambridge University Press, Oxford.

Hamedani, G. G. (2011). Characterizations of the Shakil-Kibria-Singh distribution, Austrian Journal of Statistics, 40(3), 201–207.

Hogg, R.V. and Craig, A.T. (1978). Introduction to Mathematical Statistics, Macmillan Publishing Co., New York.

Johnson, N.L. and Kotz, S. (1970). Distributions in Statistics, Vol. 2: Continuous Univariate Distributions. John Wiley and Sons, New York.

Lefevre, C., Papathanasiou, V., and Utev, S. (2002). Generalized Pearson distributions and related characterization problems. Annals of the Institute of Statistical Mathematics, 54, 731–742.

Pearson, K. (1901). Mathematical contributions to the theory of evolution, X: Supplement to a memoir on skew of variation, Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 197, 343–414.

Pearson, K. (1895). Contributions to the mathematical theory of evolution, II: Skew variation in homogeneous material, Philosophical Transactions of the Royal Society of London, A186, 343–414.

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

Provost, S.B. and Ha, H.T. (2015). Differentiated logdensity approximants, Statistical Methodology, 26, 61–71 .

Reid, N. (1988). Saddlepoint methods and statistical inference, Statistical Science, 3(2), 213–238.

Rossani, A., and Scarfone, A. M. (2009). Generalized Pearson distributions for charged particles interacting with an electric and/or a magnetic field, Physica A, 388, 2354–2366.

Scott, D.W. (2015). Multivariate Density Estimation: Theory, Practice, and Visualization, 2nd Edition, John Wiley and Sons, Hoboken, NJ.

Shakil, M. (2016). Review on generalized Pearson system of probability distributions, Journal of Mathematical Sciences and Mathematics Education, 11(2), 13–33.

Shakil, M., Singh, J. N., and Kibria, B. M. G. (2010). On a family of product distributions based on the Whittaker functions and generalized Pearson differential equation, Pakistan Journal of Statistics, 26(1), 11–125.

Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis Series: Monographs on Statistics and Applied Probability, Chapman and Hall, London.

Solomon, H. and Stephens, M.A. (1978). Approximations to density functions using Pearson curves, Journal of the American statistical association, 73(361), 153–160.

Stuart, A. and Ord, J.K. (1987). Kendall's Advanced Theory of Statistics, Griffin, London.

Zareamoghaddam, H., Provost, S.B. and Ahmed, S.E. (2017a). A log-density estimation methodology applicable to massive bivariate data, International Journal of Statistics and Probability, 6(5), 1–12.

Zareamoghaddam, H., Provost, S.B. and Ahmed, S.E. (2017b). A moment-based bivariate density estimation methodology applicable to big data modeling, Journal of Probability and Statistical Science, 15(2), 135–152.

Chapter 4

A log-density estimation methodology applicable to massive bivariate data

4.1 Introduction

We initially consider the problem of approximating the density function of a univariate 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 proposed in this chapter 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 4.3.1) involving a sample of simulated values.

This chapter is organized as follows. First, the technique being utilized for obtaining univariate DLDA's is developed in Section 4.2, 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 4.3 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 4.4.

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,

$$\frac{\mathrm{d}}{\mathrm{d}x} ln(f_X(x)) = \frac{f_X'(x)}{f_X(x)} = r(x), \tag{4.2.1}$$

where

$$r(x) = \frac{\sum_{i=0}^{\nu} a_i\, x^i}{\sum_{j=0}^{\delta} c_j\, x^j} = \frac{N_\nu(x)}{D_\delta(x)}, \tag{4.2.2}$$

$N_\nu(x)$ 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 (4.2.6).

In light of Equations (4.2.1) to (4.2.2), one has

$$f'_{\nu,\delta}(x)\sum_{j=0}^{\delta} c_j\, x^j = f_{\nu,\delta}(x)\sum_{i=0}^{\nu} a_i\, x^i, \tag{4.2.3}$$

from which the polynomial coefficients can be obtained as follows: Multiplying both sides of Equation (4.2.3) by x^h and integrating over the interval $(\alpha,\ \beta)$ 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 the left-hand side, one has

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

where $h = 0,\ 1, \ldots,\ \nu+\delta$. Note that the first term on the left-hand side, that is, $f_{\nu,\delta}(x)\sum_{j=0}^{\delta} c_j x^{j+h}\ |_\alpha^\beta$, 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 h^{th} moment of the approximated density function $f_{\nu,\delta}(x)$, one obtains $\nu+\delta+1$ linear equations having the following form:

$$-\sum_{j=0}^{\delta} c_j(j+h)\,\mu(j+h-1) = \sum_{i=0}^{\nu} a_i\,\mu(i+h),\ \ h = 0,\ 1, \ldots,\ \nu+\delta, \tag{4.2.4}$$

with $\mu(0) \equiv 1$. In order to determine the unknown coefficients of $r(x)$ as specified by Equation (4.2.2), one needs to solve the linear system resulting from Equation (4.2.4). On replacing the unknown $\mu(h)$ by $\mu_X(h)$, for $h = 0,\, 1, \ldots,\, \nu + \delta$, where $\mu_X(h)$ denotes the h^{th} moment of the distribution being approximated, one obviously obtains the following linear system:

$$-\sum_{j=0}^{\delta} c_j(j+h)\,\mu_X(j+h-1) = \sum_{i=0}^{\nu} a_i\,\mu_X(i+h),\ \ h = 0,\, 1, \ldots,\, \nu + \delta. \qquad (4.2.5)$$

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

$$f'_{\nu,\delta}(x) = r(x) f_{\nu,\delta}(x), \qquad (4.2.6)$$

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

$$ISD(\nu) = \int_{\mathscr{S}} \Big(f_{\nu,\delta}(x) - f_X(x) \Big)^2 \mathrm{d}x. \tag{4.2.7}$$

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 $ISD(\nu_0)$ reaches a set tolerance level or $ISD(\nu)$ 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 (4.2.2) 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 (4.2.4).)

3. Determine the coefficients of the rational function by solving the linear system (4.2.5).

4. Find the solution of the differential equation specified by (4.2.6) 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 $ISD(\nu)$ as defined in Equation (4.2.7).

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

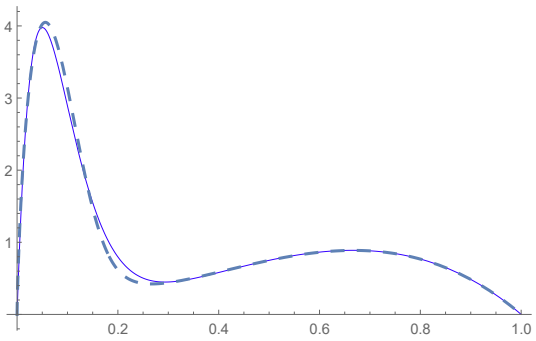


Figure 4.1: Exact (solid line) and approximated (dashed line) density functions in connection with Example 4.2.1.

Example 4.2.1. 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 4.1. 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

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

4.2.1 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 (4.2.2). This gives

rise to the following linear system:

$$\begin{pmatrix} \mu_X(0) & \mu_X(1) & \dots & \mu_X(n) \\ \mu_X(1) & \mu_X(2) & \dots & \mu_X(n+1) \\ \mu_X(2) & \mu_X(3) & \dots & \mu_X(n+2) \\ \vdots & \vdots & \ddots & \vdots \\ \mu_X(n) & \mu_X(n+1) & \dots & \mu_X(2n) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} 0 \\ -\mu_X(0) \\ -2\mu_X(1) \\ \vdots \\ -n\mu_X(n-1) \end{pmatrix}, \tag{4.2.8}$$

which, in matrix form, can be expressed as $M\mathbf{a} = \boldsymbol{\tau}$ 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 $\Re^n$. Then,

$$\begin{aligned} \mathbf{z}^T M \mathbf{z} &= \sum_{i=0}^{n} \sum_{j=0}^{n} z_i z_j \int_\alpha^\beta x^{i+j} f_X(x)\,\mathrm{d}x \\ &= \int_\alpha^\beta \Big(\sum_{i=0}^{n} z_i x^i \Big) \Big(\sum_{j=0}^{n} z_j x^j \Big) f_X(x)\,\mathrm{d}x \\ &= \int_\alpha^\beta \Big(\sum_{i=0}^{n} z_i x^i \Big)^2 f_X(x)\,\mathrm{d}x > 0. \end{aligned}$$

Thus, the linear system specified by Equation (4.2.8) has the unique solution $M^{-1}\boldsymbol{\tau}$. The resulting density approximant, which shall be referred to as a Polynomial Log-density Approximant (PLDA), will then have the following representation:

$$f_{X_n}(x) = \kappa\, e^{\sum_{i=0}^{n} a_i\, x^{i+1}/(i+1)} \tag{4.2.9}$$

where κ is the normalizing constant, which 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, $SSD(\nu) = \sum_{i=1}^{n}(\text{ECDF}(x_i) - F_{\nu,\delta}(x_i))^2$. One could select the degree ν for which $SSD(\nu)$ reaches a minimum value or beyond which $SSD(\cdot)$ 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.

4.3 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 (4.3.1). 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 $v_{11} = \big[\sum_{i=1}^{n}(w_i - \bar{w})^2\big]/n$, $v_{12} = v_{21} = \big[\sum_{i=1}^{n}(w_i - \bar{w})(z_i - \bar{z})\big]/n$ and $v_{22} = \big[\sum_{i=1}^{n}(z_i - \bar{z})^2\big]/n$. 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 w_i's and the z_i's are then uncorrelated (however, in general, they are not independently distributed), and we let

$$f_{\nu 1,\delta 1,\nu 2,\delta 2,p}(x,\, y) = f_{\nu 1,\delta 1}(x) f_{\nu 2,\delta 2}(y)\, \pi_p(x, y), \tag{4.3.1}$$

where $f_{\nu_1,\delta_1}(x)$ and $f_{\nu_2,\delta_2}(y)$ denote the estimated marginal density functions for the standardized vector $(X,\, Y)'$, $\pi_p(x, y)$ 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 (4.3.1) 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 $\mu(k,\ell) = \int_{\mathcal{R}^2} \int x^k\, y^\ell\, f_p(x,\, y)\, \mathrm{d}x\, \mathrm{d}y$ and the $(k,\ell)^{\text{th}}$ joint moment associated with the initial density $\psi(x,\, y)$, by $m(k,\ell) = \int_{\mathcal{R}^2} \int x^k\, y^\ell\, \psi(x,\, y)\, \mathrm{d}x\, \mathrm{d}y$.

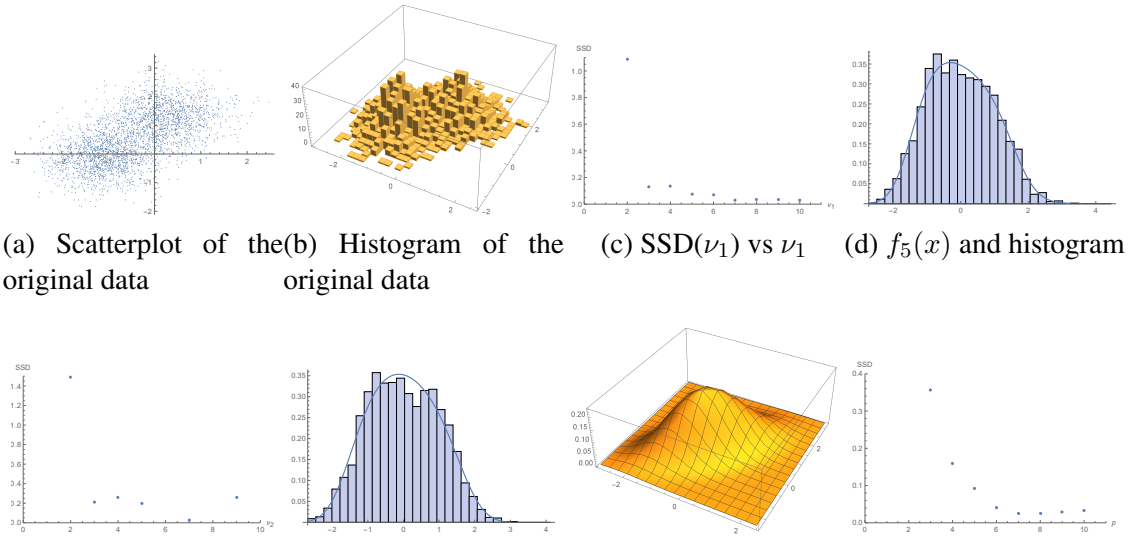
In order to obtain a computable representation of the approximant $f_p(x,\, y)$, one needs to determine the coefficients $d_{i,j}$ of the polynomial adjustment. To this end, the joint moments of the exact density $f(x,\, y)$ are equated to those associated with $f_p(x,\, y)$, that is,

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

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, $c_{i,j}$ and $\mu(i,j)$, appear in the same order for $i = 0,\, 1,\, \ldots,\, p$ and $j = 0,\, 1,\, \ldots,\, p$. Increasing p should



(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 4.2: Dataset, SSD's, estimates of the marginal densities and base density function (Example 4.3.1)

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:

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

where ECDF$(w,\ z)$ denotes the empirical CDF associated with the dataset and $G_p(w_i,\ z_i) = \int_{-\infty}^{w_i} \int_{-\infty}^{z_i} g_p(w,\ z)\, \mathrm{d}z\, \mathrm{d}w$ 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.

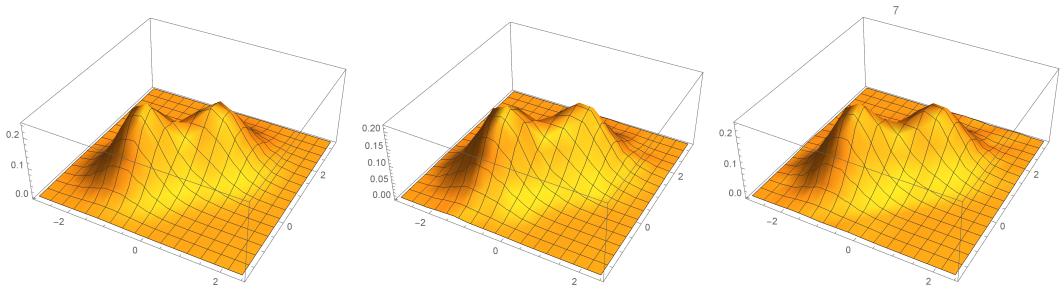
Example 4.3.1. Let $\mathbf{X}_1$ and $\mathbf{X}_2$ be bivariate normal random variables where

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

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, $g_{\mathbf{X}}(w,\, z) = 0.5g_{\mathbf{X}_1}(w,\, z) + 0.5g_{\mathbf{X}_2}(w,\, z)$. 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 4.2. 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 4.2.

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 4.3. 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 esti-



(a) Exact (b) Bivariate KDE (c) Final PDF estimate $g_{5,3,7}(w, z)$

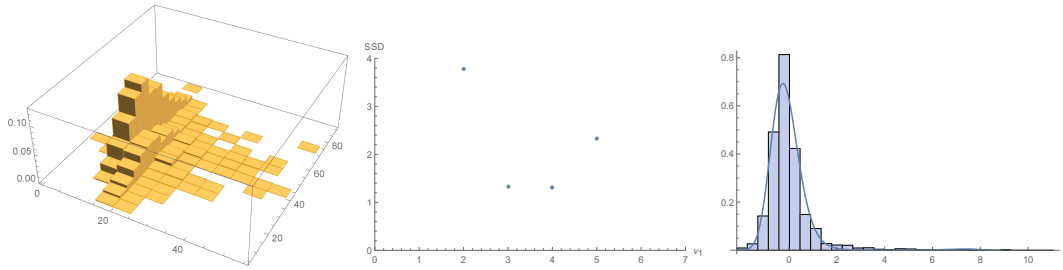
Figure 4.3: Exact and estimated density functions (Example 4.3.1)

mate. 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 (4.2.7) are respectively 0.0006 and 0.0011.

4.4 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.

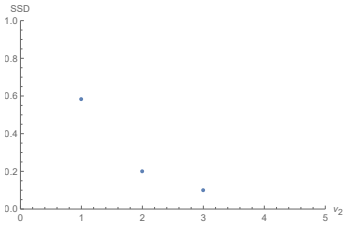
Example 4.4.1. The dataset being modeled in this example was extracted from *CommViolPredUnnormalizedData* which is included in the "UC Irvine Machine Learning Repository dataset". It contains 2315 observation vectors related to communities and crime. It combines socio-economic data from the '90 Census, 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



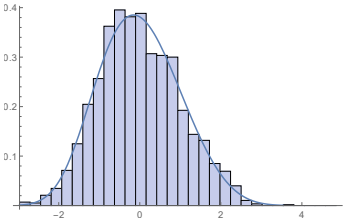
(a) Histogram of the original data

(b) SSD(ν_1) vs ν_1

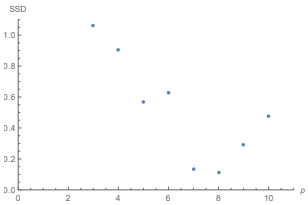
(c) $f_4(x)$ and histogram



(d) SSD(ν_2) vs ν_2



(e) $f_3(y)$ and histogram



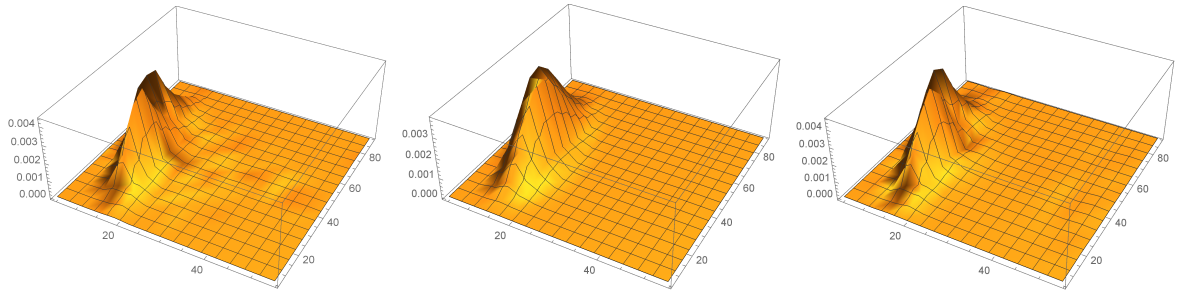
(f) $SSD(p)$ vs p

Figure 4.4: Dataset, SSD's and estimates of the marginal density functions (Example 4.4.1)

income in 1989 as the Z variable.

The histogram of the data is displayed in panel (a) of Figure 4.4. The SSD plots of the estimated marginal density functions of the standardized data shown in panels (b) and (d) of Figure 4.4 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 4.4, are included in Figure 4.5.



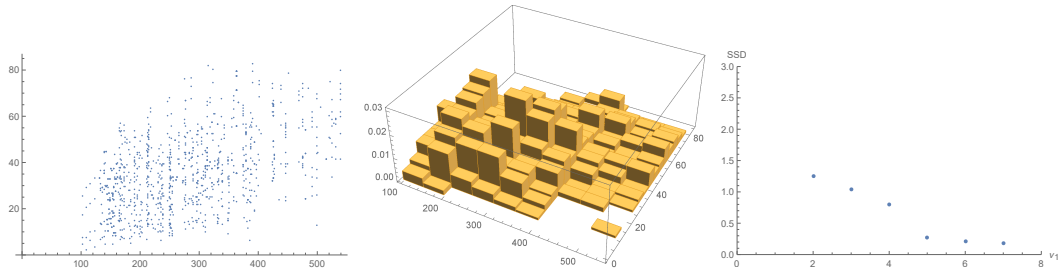
(a) Bivariate KDE (b) Transformed base PDF $\phi(w, z)$ (c) Final PDF estimate $g_{4,3,7}(w, z)$

Figure 4.5: SSD and density estimates (Example 4.4.1)

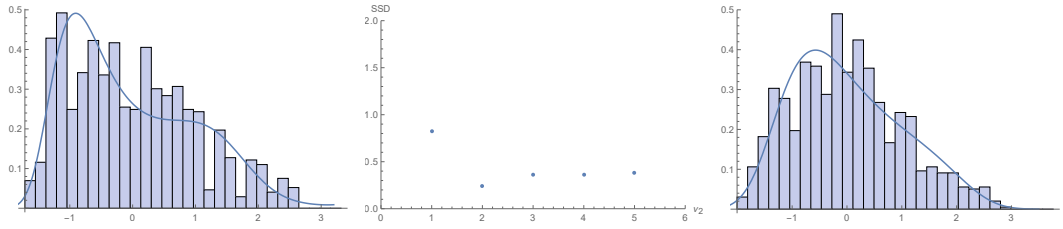
Example 4.4.2. In this example, we consider the dataset *concrete* also included in the "UC Irvine Machine Learning Repository dataset", which contains 1030 observation vectors related to concrete compressive strength in civil engineering. More information about this dataset is available in (Yeh, 1998). We selected "Cement": kg in a m3 mixture as the W variable and concrete compressive strength as the Z variable.

In Figure 4.6, the scatterplot and a histogram of the concrete data are displayed in panels (a) and (b). Likewise, the previous example, the SSD plots of the estimated marginal density functions that are shown in panels (c) and (e) of Figure 4.6 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) are all included in Figure 4.7. 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.

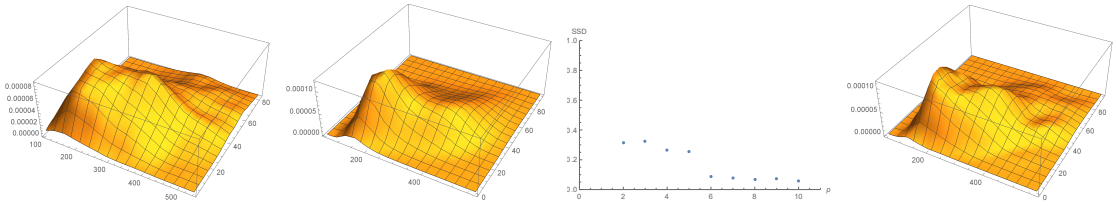


(a) Scatterplot of the original data (b) Histogram of the original data (c) SSD(ν_1) vs ν_1



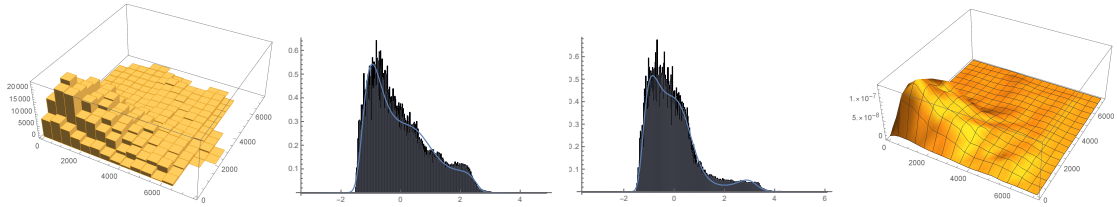
(d) $f_4(x)$ and histogram (e) SSD(ν_2) vs ν_2 (f) $f_3(y)$ and histogram

Figure 4.6: The dataset, SSD(ν_1), SSD(ν_2) and estimates of the marginal density functions (Example 4.4.2)



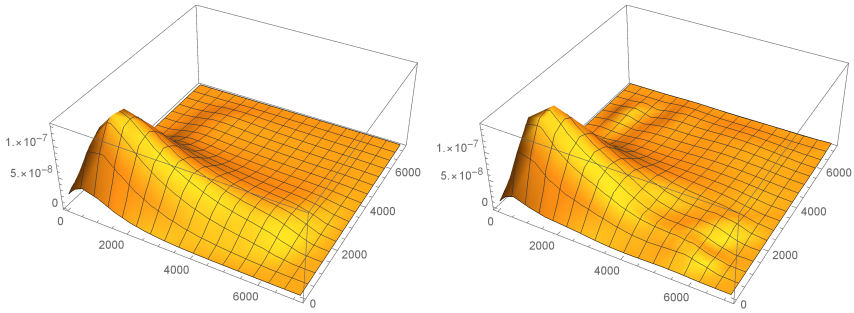
(a) Bivariate KDE (b) Transformed base PDF $\phi(w, z)$ (c) $SSD(p)$ vs p (d) Final PDF estimate $g_{4,3,6}(w, z)$

Figure 4.7: $SSD(p)$ and bivariate density estimates (Example 4.4.2)



(a) Histogram of the original data (b) $f_5(x)$ and histogram (c) $f_6(y)$ and histogram (d) Bivariate KDE

Figure 4.8: Dataset, estimates of the marginal density functions and bivariate KDE (Example 4.4.3)

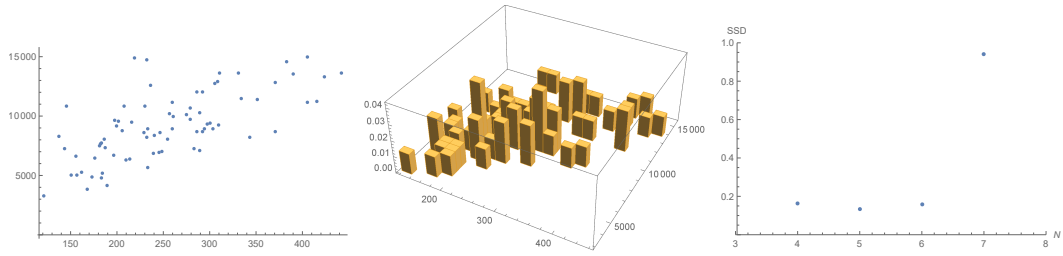


(a) Transformed base PDF $\phi(w, z)$ (b) Final PDF estimate $g_{5,6,7}(w, z)$

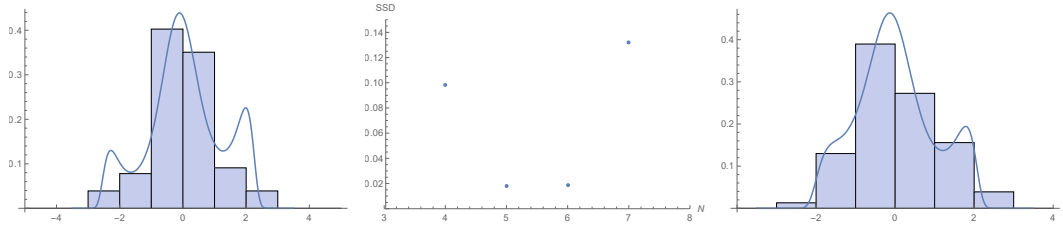
Figure 4.9: Transformed base PDF and final PDF estimate (Example 4.4.3)

Example 4.4.3. 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, 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 4.8 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



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



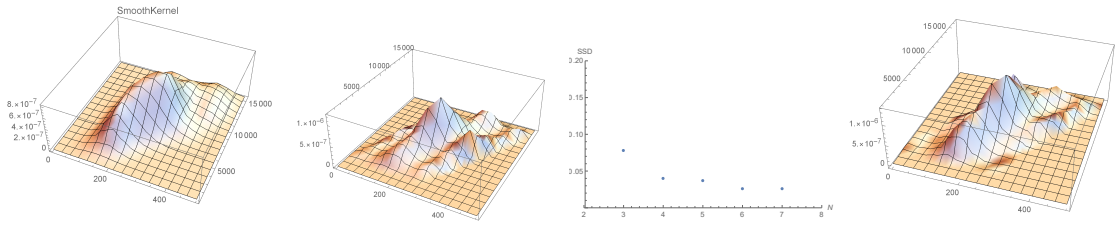
(d) $f_{5,2}(x)$ and histogram (e) SSD(ν_2) vs ν_2 (f) $f_{5,2}(y)$ and histogram

Figure 4.10: Dataset, SSD's and estimates of the marginal density functions (Example 4.4.4)

included in Figure 4.9. We observe that the proposed density estimate is consistent with the histogram of the observations.

Example 4.4.4. 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 4.10. The SSD plots of the estimated marginal density functions that are shown in panels (c) and (e) of Figure 4.10 indicate that $\nu_1 = 5$ and $\nu_2 = 5$ are suitable degrees. The corresponding density functions are plotted in panels (d) and (f).



(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 4.11: SSD and density estimates (Example 4.4.4)

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 4.11. 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.

4.5 Concluding remarks

A technique is developed whereby the derivative of the logarithm of a univariate continuous density function can be 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 product of the marginal density approximants (taken to be a 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 avail-

able from the authors upon request.

References

Blackard, J.A., and 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.

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

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

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

Elderton, W.P. , and Johnson, N.L. (1969). Systems of Frequency Curves, Cambridge University Press, Oxford.

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

Provost, S.B., and Ha, H.T. (2015). Differentiated logdensity approximants, Statistical Methodology, 26, 61–71.

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), 213–238.

Scott, D.W. (2015). Multivariate Density Estimation: Theory, Practice, and Visualization, 2nd Edition, John Wiley and Sons, Hoboken, NJ.

Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis Series: Monographs on Statistics and Applied Probability, Chapman and Hall, London.

Solomon, H., and Stephens, M. (1978). Approximations to density functions using Pearson curves, Journal of the American Statistical Association, 73, 153–160.

Yeh, I.C. (1998). Modeling of strength of high performance concrete using artificial neural networks, Cement and Concrete Research, 28(12), 1797–1808.

Chapter 5

Shrinkage estimation applied to a semi-nonparametric regression model

5.1 Introduction

Consider the n semi-nonparametric regression equations (Ma *et al.*, 2015),

$$y_{it} = \alpha_i + \beta_i m(x_{it}) + \dot{\epsilon}_{it}, \tag{5.1.1}$$

where y_{it} is the t^{th} response value on the i^{th} individual for $i = 1, 2, \ldots, n$ and $t = 1, 2, \ldots, T$, with the corresponding explanatory variable x_{it}. The unknown coefficients α_i and β_i are the location and scale parameters associated with i^{th} model, $m(\cdot)$ is the common nonparametric component which is assumed to be differentiable, and the $\dot{\epsilon}_{it}$'s are the error terms which are assumed to be independently and identically distributed with mean zero and constant variance $\dot{\sigma}^2$. The model (5.1.1) could be represented in matrix form as

$$Y = A + B\, m(X) + \mathcal{E},$$

where $Y = \{y_{it}\}$, $A = (\alpha_1 \mathbf{1}_T, \ldots, \alpha_n \mathbf{1}_T)$, $\mathbf{1}_T = (1, \ldots, 1)'$, $B = \text{Diag}(\beta_1, \ldots, \beta_n)$, $X = \{x_{it}\}$ and $\mathcal{E} = \{\dot{\epsilon}_{it}\}$ for $i = 1, 2, \ldots, n$ and $t = 1, 2, \ldots, T$. The objective consists in estimating the parametric and nonparametric components of this regression model. One may use this model when the observations on several individuals are similarly distributed

but with possibly different response values that could be explained by varying location and scale parameters. One could refer to (Fan and Gijbels, 1996), (Robinson, 1988), (Stone, 1982), (Ruppert *et al.*, 2003) and (Begun *et al.*, 1983) for more information about nonparametric and semi-nonparametric models and related results.

For the purpose of measuring the characteristics (proteins, etc.) of individual molecules, mass spectrometry (MS), which converts those characteristics to ions in some magnetic fields, are utilized to sort and separate the ions and then measure the separated ions relating to such characteristics, see for instance (Roy *et al.*, 2011), (Yasui *et al.*, 2003) and (Guilhaus, 1995). A mass spectrum will usually be presented as a vertical bar graph in which each bar represents an ion having a specific mass-to-charge ratio (m/z), the length of the bar indicating the relative abundance of the ion. The most intense ion is assigned an abundance level of 100, and it is referred to as the base peak. Such measurements are useful to monitor the progression of the disease and to evaluate new treatments. The two more widely used mass spectrometers are SELDI-TOF (surface enhanced laser desorption/ionization time-of-fight) (Baggerly *et al.*, 2004) and MALDI-TOF (matrix assisted laser desorption and ionization time-of-flight) (Baggerly *et al.*, 2003). Model (5.1.1) was proposed for analyzing such mass spectrometry data. In each spectrum, the y-axis is the relative abundance (intensity) and the x-axis is the mass-to-charge ratio (m/z value). In this chapter, we apply this model to a SELDI-TOF mass spectrometry data set collected from a study on liver cancer. Interested readers could refer to (Ma *et al.*, 2015) and references therein for more information about MS data analysis.

When making use of Stein-type shrinkage techniques, one assumes some prior uncertain information (PUI) about the parameters of interest. The Stein-type shrinkage estimators are predicated on the assumed PUI with a view to increase the accuracy of the estimates of the parameters of interest. For some background about various shrinkage

techniques applying to similar problems, one can refer to (Ahmed and Krzanowski, 2004), (Ahmed, 1994), (Chitsaz and Ahmed, 2012a), (Chitsaz and Ahmed, 2012b) and the references therein. We focus on the location and scale parametric components in model (5.1.1). In the remainder, the prior information consists in letting all the scale parameters β_i's be equal, for $i = 1, \ldots, n$, that is, $\beta_1 = \ldots = \beta_n = \beta_0$, β_0 being unknown. The parametric components are then estimated under the full model as well as the restricted model which is specified in Section 5.3 where the Stein-type shrinkage estimators for the location and scale parameters are also discussed. Section 5.4 features a simulation study that confirms the efficiency of the shrinkage-type estimators.

5.2 The original methodology

The multi-step algorithm proposed in (Ma *et al.*, 2015) for the estimation of the parametric components α_i and β_i, $i = 1, 2, \ldots, n$, and the nonparametric function $m(\cdot)$ in (5.1.1) is described in this section. As the authors stated, on the one hand, when the values of the α_i's and β_i's for $i = 1, 2, \ldots, n$, are known, one will end up with nonparametric regression models whereby $m(\cdot)$ is estimated by some local linear regression approach using the kernel function $K(\cdot)$ with $K_h(\cdot) = K(\cdot/h)/h$, h being a suitable bandwidth; on the other hand, if $m(\cdot)$ were known, the model would reduce to n simple linear regression equations in which case location and scale parameters are obtained by applying the least squares approach.

5.2.1 The algorithm

In order to obtain estimates for both the parametric and nonparametric components of (5.1.1), the following iterative algorithm was proposed in (Ma *et al.*, 2015):

Algorithm 1: Iterative algorithm for the semi-nonparametric model

- (a) Set $\alpha_1 = 0$ and $\beta_1 = 1$. The initial kernel estimator of $m(\cdot)$ obtained from (x_{1t}, y_{1t}), $t = 1, ..., T$ via $y_{1t} = m(x_{1t}) + \dot{\epsilon}_{1t}$ is

$$\tilde{m}(x) = \frac{\sum_{t=1}^{T} \omega_{1t}(x) y_{1t}}{\sum_{t=1}^{T} \omega_{1t}(x)} \tag{5.2.1}$$

  where $\omega_{1t}(x) = K_h(x_{1t} - x)(S_{T,2} - (x_{1t} - x)S_{T,1})$ and $S_{T,k} = \sum_{t=1}^{T} K_h(x_{1t} - x)(x_{1t} - x)^k$ for $k = 1, 2$.

- (b) Replacing $m(\cdot)$ by $\tilde{m}(\cdot)$, as obtained in Step (a), the parameters $\alpha_i, \beta_i, i = 1, ..., n$ are estimated by simple linear regression (least squares) as

$$\hat{\beta}_i = \frac{\sum_{t=1}^{T} [\tilde{m}(x_{it}) - \bar{\tilde{m}}(x_{i.})] y_{it}}{\sum_{t=1}^{T} [\tilde{m}(x_{it}) - \bar{\tilde{m}}(x_{i.})]^2}, \;\; \hat{\alpha}_i = \bar{y}_{i.} - \hat{\beta}_i \bar{\tilde{m}}(x_{i.}), \tag{5.2.2}$$

  where $\bar{y}_{i.} = \frac{1}{T} \sum_{t=1}^{T} y_{it}$ and $\bar{\tilde{m}}(x_{i.}) = \frac{1}{T} \sum_{t=1}^{T} \tilde{m}(x_{it})$.

- (c) Given the estimates $\hat{\alpha}_i$ and $\hat{\beta}_i$, the estimate of the function $m(\cdot)$ is updated as

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} \sum_{t=1}^{T} \omega_{it}^*(x) y_{it}^*}{\sum_{i=1}^{n} \sum_{t=1}^{T} \omega_{it}^*(x)}, \tag{5.2.3}$$

  where $\omega_{it}^*(x) = \hat{\beta}_i^2 K_{h^*}(x_{it} - x)[\sum_{i=1}^{n} \hat{\beta}_i^2 S_{T,2}^{*(i)} - (x_{it} - x) \sum_{i=1}^{n} \hat{\beta}_i^2 S_{T,1}^{*(i)}]$, $y_{it}^* = (y_{it} - \hat{\alpha}_i)/\hat{\beta}_i$ and $S_{T,k}^{*(i)} = \sum_{t=1}^{T} K_{h^*}(x_{it} - x)(x_{it} - x)^k$ for $k = 1, 2$.

- (d) Repeat Steps (b) and (c) until convergence is observed for both the parametric and nonparametric components.

Note: The bandwidth h^* might be different from h in order to achieve a better convergence rate, see (Ma *et al.*, 2015).

5.2.2 Bandwidth selection

The proper choices of h and h^* may asymptotically affect the accuracy of the estimators. (Ma *et al.*, 2015) suggested the K-fold cross validation approach for selecting the bandwidth. To this aim, all n individuals are divided randomly into K groups $Z_1, Z_2, \ldots, Z_K$ where Z_k is taken to be the k^{th} test set with $Z_{-k} = (\{1, 2, \ldots, n\} - Z_k)$ as the corresponding training set. Now, Algorithm 1 is applied to Z_{-k} to obtain $\hat{m}_{Z_{-k}}(\cdot, h, h^*)$ as an estimator of $m(\cdot)$ where h and h^* are the bandwidths used for obtaining $\tilde{m}(\cdot)$ and $\hat{m}(\cdot)$, respectively. Then, given $\hat{m}_{Z_{-k}}(\cdot, h, h^*)$, the parametric components of the test set are estimated by the least squares (simple linear regression) approach and the mean square prediction error associated with Z_k is computed as

$$\text{MSPE}(Z_k, h, h^*) = \frac{1}{|Z_k|} \sum_{i \in Z_k} \sum_{t=1}^{T} (y_{it} - \hat{\alpha}_{ki} - \hat{\beta}_{ki}\ \hat{m}_{Z_{-k}}(x_t, h, h^*))^2 \qquad (5.2.4)$$

where $\hat{\alpha}_{ki}$ and $\hat{\beta}_{ki}$ are the estimated regression coefficients in Z_k for $k = 1, \ldots, K$. Then, the optimum values of the bandwidths for h and h^* are chosen to be those that minimize MSPE, that is,

$$(\hat{h}, \hat{h^*}) = \arg\min_{(h,h^*)} \frac{1}{K} \sum_{k=1}^{K} \text{MSPE}(Z_k, h, h^*). \qquad (5.2.5)$$

Asymptotic results

The following asymptotic results are stated in Ma *et al.* (2015). By asymptotic, it is meant that both n and T are considered to be large enough (normally $n \ll T$). The following conditions will be needed to establish the asymptotic theory.

C1: The baseline intensity $m(\cdot)$ is continuous and has a bounded second order derivative.

C2: There exist constants $\alpha > 0$ and $\delta > 0$, such that $f(\cdot)$ the marginal density of x_{it} satisfies $f(x) > \delta$, and $|f(x) - f(y)| \leq c|x - y|^{\alpha}$ for any x and y within the support of $f(\cdot)$.

C3: The conditional variance $\sigma_i^2(x) = Var(y_{it}|x_{it} = x)$ is bounded and continuous in x, where $i = 1, \ldots, n$ and $t = 1, \ldots, T$.

C4: The kernel $K(\cdot)$ is a symmetric probability density function with bounded support. Hence $K(\cdot)$ has the properties: $\int_{-\infty}^{\infty} K(u)\mathrm{d}u = 1$, $\int_{-\infty}^{\infty} uK(u)\mathrm{d}u = 0$ and $\int_{-\infty}^{\infty} u^2K(u)\mathrm{d}u$ is positive and bounded.

Let $X = \{x_{it},\, i = 1, \ldots, n,\; t = 1, \ldots, T\}$, $\mu_l = \int_{-\infty}^{\infty} u^l K(u)\mathrm{d}u$ and $\nu_l = \int_{-\infty}^{\infty} u^l K^2(u)\mathrm{d}u$, for $l \geq 0$.

Lemma 5.2.1. *Assuming that conditions C1-C4 are satisfied,*

$$
\begin{aligned}
&E(\tilde{m}(x) - m(x)|X) = \frac{1}{2}m''(x)\mu_2 h^2 + o(h^2),\\
&Var(\tilde{m}(x)|X) = \frac{\nu_0\,\sigma^2(x)}{T\,h\,f(x)} + o(\frac{1}{T\,h}),\\
&E(\{\tilde{m}(x) - m(x)\}^2|X) = \frac{1}{4}(m''(x)\mu_2)^2h^4 + \frac{\nu_0\,\sigma^2(x)}{T\,h\,f(x)} + o(h^4 + \frac{1}{T\,h}).
\end{aligned}
$$

Lemma 5.2.2. *Suppose that conditions C1-C4 are satisfied and assume that* $E\big(m^2(x_{it})\big)(\sigma^2 + 1) < \infty$ *and* $E\big(m^2(x_{it})\big) > 0$ *for all* $i = 1, \ldots, n$ *and* $t = 1, \ldots, T$. *By restricting the order of* h *to be between* $T^{-1/2}$ *and* $T^{-1/4}$, *the* $\hat{\beta}_i^U$ *values are asymptotically distributed*

as follows:

$$(\hat{\beta}_i^U - \beta) \sim \mathcal{N}(0, {\sigma_1}^2). \tag{5.2.6}$$

For more information about the value of σ_1 and the proofs of Lemmas 5.2.1 and 5.2.2, one can refer to Ma *et al.* (2015).

Letting $\tilde{m}(\cdot)$ be a consistent estimator of the nonparametric component $m(\cdot)$, the nuisance parameter, it follows that

$$(\hat{\alpha}_i^U - \alpha) \sim \mathcal{N}(0, {\sigma_2}^2), \quad i = 1, \ldots, n, \tag{5.2.7}$$

where ${\sigma_2}^2 = {\sigma_1}^2\big(\bar{\tilde{m}}(x_{i.}) + \frac{1}{T})\big)$, $i = 1, \ldots, n$.

5.3 Stein-type location and scale estimators

Once $m(\cdot)$ in (5.1.1) is identified, the problem reduces to solving multiple simple regression models wherein the slopes $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)'$ and the intercepts $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)'$ are estimated via the least squares approach. In fact, $m(\cdot)$ is considered to be the nuisance component whereas $\tilde{m}(\cdot)$ is viewed as its consistent estimator. Thus, the goal is to improve the accuracy of the parametric estimators under the assumption that the slope is the same for all n models, i.e., $\beta_1 = \ldots = \beta_n = \beta_0$ where β_0 is unknown. For simplicity, $m(x_{it})$ is denoted by m_{it} and $M = \{\tilde{m}(x_{it}),\, i = 1, \ldots, n,\ t = 1, \ldots, T\}$. Accordingly, we have the following n simple linear models:

$$\mathbf{y}_i = \boldsymbol{\alpha}_i \mathbf{1}_T + \boldsymbol{\beta}_i\, \mathbf{m}_i + \boldsymbol{\epsilon}_i, \qquad i = 1, 2, \ldots, n, \tag{5.3.1}$$

with $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})'$, $\mathbf{1}_T = (1, \ldots, 1)'$, $\mathbf{m}_i = (m_{i1}, \ldots, m_{iT})'$ and the associated error vector $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iT})'$ with mean vector $\mathbf{0}$ and covariance matrix $\sigma^2 I_T$, I_T being the identity matrix of order T.

In the next section, some Stein-type shrinkage estimators for the slopes and intercepts are computed under the parallelism hypothesis, i.e. $H_o : \boldsymbol{\beta} = \beta_0 \mathbf{1}_n$.

5.3.1 Estimates under the full and restricted models

By making use of the known matrix M, the *unrestricted* (full model) estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ of (5.3.1) obtained by applying the least squares approach are as follows:

$$\hat{\boldsymbol{\beta}}^U = (\hat{\beta}_1^U, \ldots, \hat{\beta}_n^U)' \qquad \text{and} \qquad \hat{\boldsymbol{\alpha}}^U = \bar{\mathbf{y}} - T_n \, \hat{\boldsymbol{\beta}}^U, \tag{5.3.2}$$

where $\bar{\mathbf{y}} = (\bar{y}_{1\cdot}, \ldots, \bar{y}_{n\cdot})'$, $\hat{\beta}_i^U = [\mathbf{m}_i'\mathbf{y}_i - \frac{1}{T}(\mathbf{m}_i'\mathbf{1}_T)(\mathbf{y}_i'\mathbf{1}_T)]/(T\,Q_i)$, $T_n = \mathrm{Diag}(\bar{m}_{1\cdot}, \ldots, \bar{m}_{n\cdot})$, $T\,Q_i = \mathbf{m}_i'\mathbf{m}_i - \frac{1}{T}(\mathbf{m}_i'\mathbf{1}_T)^2$, $\bar{m}_{i\cdot} = \frac{1}{T}(\mathbf{m}_i'\mathbf{1}_T)$ and $\bar{y}_{i\cdot} = \frac{1}{T}(\mathbf{y}_i'\mathbf{1}_T)$.

Furthermore, the unbiased estimator of σ^2 is

$$s^2 = \frac{1}{nT - 2n} \sum_{i=1}^{n} \|\mathbf{y}_i - \hat{\boldsymbol{\alpha}}_i^U \mathbf{1}_T - \hat{\boldsymbol{\beta}}_i^U \, \mathbf{m}_i\|_2^2, \tag{5.3.3}$$

where $\|\cdot\|_2$ denotes the L_2-norm. However, under the null hypothesis, i.e., $H_0 : \boldsymbol{\beta} = \beta_0 \mathbf{1}_n$, the *restricted* estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are

$$\hat{\boldsymbol{\beta}}^R = \frac{\mathbf{1}_n \mathbf{1}_n' D_{22} \hat{\boldsymbol{\beta}}^U}{nTQ} \qquad \text{and} \qquad \hat{\boldsymbol{\alpha}}^R = \hat{\boldsymbol{\alpha}}^U + T_n \, H \, \hat{\boldsymbol{\beta}}^U, \tag{5.3.4}$$

where $H = I_n - [\mathbf{1}_n \mathbf{1}_n' D_{22}]/(nTQ)$ with $nQ = \sum_{i=1}^n Q_i$ and $D_{22}^{-1} = \mathrm{Diag}(TQ_1, \ldots, TQ_n)$.

Theorem 5.3.1. *(Saleh, 2006) Assuming that the conditions of Lemma 5.2.2 are satisfied, then given M, one has*

$$\begin{pmatrix} \hat{\boldsymbol{\alpha}}^U - \boldsymbol{\alpha} \\ \hat{\boldsymbol{\beta}}^U - \boldsymbol{\beta} \end{pmatrix} \sim \mathcal{N}_{2n} \left\{ \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \sigma_1^2 \begin{pmatrix} D_{11} & -T_n D_{22} \\ -T_n D_{22} & D_{22} \end{pmatrix} \right\}, \tag{5.3.5}$$

$$\begin{pmatrix} \hat{\boldsymbol{\alpha}}^R - \boldsymbol{\alpha} \\ \hat{\boldsymbol{\beta}}^R - \boldsymbol{\beta} \end{pmatrix} \sim \mathcal{N}_{2n} \left\{ \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \sigma_1^2 \begin{pmatrix} D_{11}^* & D_{12}^* \\ D_{21}^* & D_{22} \end{pmatrix} \right\}, \tag{5.3.6}$$

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}^U - \boldsymbol{\beta} \\ \hat{\boldsymbol{\beta}}^U - \boldsymbol{\beta}^R \end{pmatrix} \sim \mathcal{N}_{2n} \left\{ \begin{pmatrix} \mathbf{0} \\ H\,\boldsymbol{\beta} \end{pmatrix}, \sigma_1^2 \begin{pmatrix} D_{22} & H D_{22} \\ D_{22} H' & H D_{22} \end{pmatrix} \right\}, \tag{5.3.7}$$

and

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}^R - \beta_0 \mathbf{1}_n \\ \hat{\boldsymbol{\beta}}^U - \boldsymbol{\beta}^R \end{pmatrix} \sim \mathcal{N}_{2n} \left\{ \begin{pmatrix} (\bar{\beta} - \beta_0)\mathbf{1}_n \\ H\,\boldsymbol{\beta} \end{pmatrix}, \sigma_1^2 \begin{pmatrix} \frac{\mathbf{1}_n \mathbf{1}_n'}{nQ} & O \\ O & H\,D_{22} \end{pmatrix} \right\} \tag{5.3.8}$$

with $\bar{\beta}\mathbf{1}_n = \mathbf{1}_n \mathbf{1}_n' D_{22}^{-1} \boldsymbol{\beta}/(nQ)$, $D_{11} = (N)^{-1} + T_n D_{22} T_n$, $N = \mathrm{Diag}(T, \ldots, T)$, $D_{11}^* = (N)^{-1} + T_n \mathbf{1}_n \mathbf{1}_n' T_n/(nQ)$, $D_{12}^* = \frac{1}{nQ} \mathbf{1}_n \mathbf{1}_n' T_n$ *and* $D_{21}^* = D_{12}^{*\prime}$.

5.3.2 Shrinkage estimators

In order to test $H_0 : \boldsymbol{\beta} = \beta_0 \mathbf{1}_n$, the likelihood ratio test statistic,

$$\mathcal{L}_n = \frac{\hat{\boldsymbol{\beta}}^{U\prime} H' D_{22}^{-1} H \hat{\boldsymbol{\beta}}^U}{(n-1)s^2}, \tag{5.3.9}$$

is considered. Saleh (2006) proved that under the null hypothesis, $\mathcal{L}_n$ follows a central F-distribution with $n-1$ and $T-2n$ degrees of freedom, while under the alternative hypothesis, $H_a : \boldsymbol{\beta} \neq \beta_0 \mathbf{1}_n$, it follows the noncentral F-distribution with $n-1$ and $T-2n$

degrees of freedom and noncentrality parameter $\Delta^2/2$ where $\Delta^2 = \hat{\boldsymbol{\beta}}^{U\prime} H' D_{22}^{-1} H \hat{\boldsymbol{\beta}}^U / \sigma^2$.

In terms of $\mathcal{L}_n$, the Stein-type shrinkage estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are given by

$$\hat{\boldsymbol{\beta}}^S = \hat{\boldsymbol{\beta}}^U - c\,\mathcal{L}_n^{-1} H \hat{\boldsymbol{\beta}}^U \qquad \text{and} \qquad \hat{\boldsymbol{\alpha}}^S = \hat{\boldsymbol{\alpha}}^U + c\,\mathcal{L}_n^{-1} T_n H \hat{\boldsymbol{\beta}}^U, \tag{5.3.10}$$

where $c = \frac{(n-3)m}{m+2}$ with $m = n(T-2)$ and $c \to n-3$ as $nT \to \infty$. Note that if $\mathcal{L}_n > c$, then $c\,\mathcal{L}_n^{-1} > 1$, and this may result in over-shrinking, which is undesirable. In order to avoid this problem, one can utilize of the positive-rule shrinkage estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, which are defined as follows:

$$\begin{aligned} \hat{\boldsymbol{\alpha}}^{PS} &= \hat{\boldsymbol{\alpha}}^R\, \mathcal{I}(\mathcal{L}_n < c) + \hat{\boldsymbol{\alpha}}^S\, \mathcal{I}(\mathcal{L}_n > c) \\ &= \hat{\boldsymbol{\alpha}}^U + \{1 - (1 - c\,\mathcal{L}_n^{-1})\, \mathcal{I}(\mathcal{L}_n > c)\} T_n H \hat{\boldsymbol{\beta}}^U, \end{aligned} \tag{5.3.11}$$

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{PS} &= \hat{\boldsymbol{\beta}}^R\, \mathcal{I}(\mathcal{L}_n < c) + \hat{\boldsymbol{\beta}}^S\, \mathcal{I}(\mathcal{L}_n > c), \\ &= \hat{\boldsymbol{\beta}}^U - (1 - c\,\mathcal{L}_n^{-1})\, \mathcal{I}(\mathcal{L}_n > c) H \hat{\boldsymbol{\beta}}^U, \end{aligned} \tag{5.3.12}$$

where $\mathcal{I}(\cdot)$ is the indicator function.

The preliminary test estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are obtained as follows:

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{PT} &= \hat{\boldsymbol{\beta}}^U - H \hat{\boldsymbol{\beta}}^U\, \mathcal{I}(\mathcal{L}_n < c_\alpha), \\ \hat{\boldsymbol{\alpha}}^{PT} &= \hat{\boldsymbol{\alpha}}^U + T_n H \hat{\boldsymbol{\beta}}^U\, \mathcal{I}(\mathcal{L}_n < c_\alpha), \end{aligned} \tag{5.3.13}$$

where c_α (the critical point) is the the upper α-quantile of the central F distribution with $n-1$ and m degrees of freedom.

5.4 Asymptotic properties

In this section, we study the asymptotic properties of the estimates of the unknown parametric components α_i and β_i, $i = 1, \ldots, n$, in (5.1.1), as determined by applying various shrinkage and pre-test techniques. The nonparametric component $m(\cdot)$ shall be targeted as a nuisance parameter. Asymptotic means that, both n and T are assumed to be large enough (normally $n \ll T$).

Theorem 5.4.1. *Assuming that all the required conditions stated in Lemma 5.2.2 are satisfied, the bias associated with intercept and slope estimators are*

$$\mathbf{b}(\hat{\boldsymbol{\beta}}^U) = \mathbf{0}$$

$$\mathbf{b}(\hat{\boldsymbol{\beta}}^R) = -H\boldsymbol{\beta}$$

$$\mathbf{b}(\hat{\boldsymbol{\beta}}^S) = -c(n-1)H\boldsymbol{\beta}E[\chi^{-2}_{n+1}(\Delta^2)]$$

$$\mathbf{b}(\hat{\boldsymbol{\beta}}^{PS}) = -H\boldsymbol{\beta}\{G_{n+1,m}(c_1;\Delta^2) + c_1E[F^{-1}_{n+1,m}(\Delta^2)\mathcal{I}(F_{n+1,m}(\Delta^2) > c_1)]\}$$

$$\mathbf{b}(\hat{\boldsymbol{\beta}}^{PT}) = -H\boldsymbol{\beta}\, G_{n+1,m}(\ell_\alpha;\Delta^2)$$

$$\mathbf{b}(\hat{\boldsymbol{\alpha}}^U) = \mathbf{0}$$

$$\mathbf{b}(\hat{\boldsymbol{\alpha}}^R) = T_nH\boldsymbol{\beta}$$

$$\mathbf{b}(\hat{\boldsymbol{\alpha}}^S) = c(n-1)T_nH\boldsymbol{\beta}E[\chi^{-2}_{n+1}(\Delta^2)]$$

$$\mathbf{b}(\hat{\boldsymbol{\alpha}}^{PS}) = c_1T_nH\boldsymbol{\beta}\,\{E[F^{-1}_{n+1,m}(\Delta^2)] - E[F^{-1}_{n+1,m}(\Delta^2)\mathcal{I}(F_{n+1,m}(\Delta^2) > c_1)]\} + T_nH\boldsymbol{\beta}\, G_{n+1,m}(c_1;\Delta^2)$$

$$\mathbf{b}(\hat{\boldsymbol{\alpha}}^{PT}) = T_nH\boldsymbol{\beta}\, G_{n+1,m}(\ell_\alpha;\Delta^2)$$

where $\ell_\alpha = \frac{n-1}{n+1}F_{n-1,m}(\alpha)$, $\Delta^2 = \hat{\boldsymbol{\beta}}^{U\prime}H'D_{22}^{-1}H\hat{\boldsymbol{\beta}}^U/\sigma^2$, $c_1 = \frac{c(n-1)}{n+1}$, $\mathbf{b}(\hat{\boldsymbol{\beta}}^*) = E(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}\,|\,M)$ *and* $\mathbf{b}(\hat{\boldsymbol{\alpha}}^*) = E(\hat{\boldsymbol{\alpha}}^* - \boldsymbol{\alpha}\,|\,M)$.

Theorem 5.4.2. *Assuming that all the required conditions mentioned in Lemma 5.2.2 are*

satisfied, the MSE associated with intercept and slope estimators are

$$MSE(\hat{\boldsymbol{\beta}}^U) = \sigma^2 D_{22} \qquad MSE(\hat{\boldsymbol{\alpha}}^U) = \sigma^2 D_{11}$$
$$MSE(\hat{\boldsymbol{\beta}}^R) = \frac{\sigma^2 \mathbf{1}_n \mathbf{1}_n'}{TQ} + H\boldsymbol{\beta}\boldsymbol{\beta}'H' \qquad MSE(\hat{\boldsymbol{\alpha}}^R) = \sigma^2 D_{11}^* + T_n H\boldsymbol{\beta}\boldsymbol{\beta}'H'T_n'$$

$$\begin{aligned} MSE(\hat{\boldsymbol{\beta}}^S) &= \sigma^2 D_{22} - c(n-1)\sigma^2 H D_{22}\{2E[\chi_{n+1}^{-2}(\Delta^2)] - (n-3)E[\chi_{n+1}^{-4}(\Delta^2)]\} \\ &\quad + c(n^2-1)(H\boldsymbol{\beta}\boldsymbol{\beta}'H')E[\chi_{n+3}^{-4}(\Delta^2)] \\ MSE(\hat{\boldsymbol{\alpha}}^S) &= \sigma^2 D_{11} - c(n-1)\sigma^2 T_n H D_{22} H' T_n' \{2\Delta^2 E[\chi_{n+1}^{-2}(\Delta^2)] \\ &\quad - (n-3)E[\chi_{n+1}^{-4}(\Delta^2)]\} + c(n^2-1)(T_n H\boldsymbol{\beta}\boldsymbol{\beta}'H'T_n')E[\chi_{n+3}^{-4}(\Delta^2)] \end{aligned}$$

$$\begin{aligned} MSE(\hat{\boldsymbol{\beta}}^{PS}) &= MSE(\hat{\boldsymbol{\beta}}^S) - (\sigma^2 H D_{22} - 2H\boldsymbol{\beta}\boldsymbol{\beta}'H') \\ &\quad \times E[(1 - c_1 F_{n+1,m}^{-1}(\Delta^2))\mathcal{I}(F_{n+1,m}(\Delta^2) < c_1)] \\ &\quad - H\boldsymbol{\beta}\boldsymbol{\beta}'H' E[(1 - c_2 F_{n+3,m}^{-1}(\Delta^2))\mathcal{I}(F_{n+3,m}(\Delta^2) < c_2)] \\ MSE(\hat{\boldsymbol{\alpha}}^{PS}) &= MSE(\hat{\boldsymbol{\alpha}}^S) - (\sigma^2 T_n H D_{22} H' T_n' - 2T_n H\boldsymbol{\beta}\boldsymbol{\beta}'H'T_n') \\ &\quad \times E[(1 - c_1 F_{n+1,m}^{-1}(\Delta^2))\mathcal{I}(F_{n+1,m}(\Delta^2) < c_1)] \\ &\quad - T_n H\boldsymbol{\beta}\boldsymbol{\beta}'H'T_n' E[(1 - c_2 F_{n+3,m}^{-1}(\Delta^2))\mathcal{I}(F_{n+3,m}(\Delta^2) < c_2)] \end{aligned}$$

$$\begin{aligned} MSE(\hat{\boldsymbol{\beta}}^{PT}) &= \sigma^2 D_{22} - \sigma^2 H D_{22} G_{n+1,m}(\ell_\alpha; \Delta^2) + H\boldsymbol{\beta}\boldsymbol{\beta}'H' \\ &\quad \times \{2G_{n+1,m}(\ell_\alpha; \Delta^2) - G_{n+3,m}(\ell_\alpha^*; \Delta^2)\} \\ MSE(\hat{\boldsymbol{\alpha}}^{PT}) &= \sigma^2 D_{11} - \sigma^2 (D_{11} - D_{11}^*) G_{n+1,m}(\ell_\alpha; \Delta^2) + T_n H\boldsymbol{\beta}\boldsymbol{\beta}'H'T_n' \\ &\quad \times \{2G_{n+1,m}(\ell_\alpha; \Delta^2) - G_{n+3,m}(\ell_\alpha^*; \Delta^2)\} \end{aligned}$$

where $\ell_\alpha^* = \frac{n-1}{n+3} F_{n-1,m}(\alpha)$, $c_2 = \frac{c(n-1)}{n+3}$, $D_{11}^* = N^{-1} + \frac{T_n' \mathbf{1}_n \mathbf{1}_n' T_n}{TQ}$, $MSE(\hat{\boldsymbol{\beta}}^*) = E((\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta})' \mid M)$ *and* $MSE(\hat{\boldsymbol{\alpha}}^*) = E((\hat{\boldsymbol{\alpha}}^* - \boldsymbol{\alpha})(\hat{\boldsymbol{\alpha}}^* - \boldsymbol{\alpha})' \mid M)$.

Saleh (2006) showed that by comparing the MSE matrices of different intercept estimators under H_0, the following results hold:

$$
\begin{aligned}
&(i)\, MSE(\hat{\boldsymbol{\alpha}}^U) - MSE(\hat{\boldsymbol{\alpha}}^R) = \sigma^2 T_n H\boldsymbol{\beta}\boldsymbol{\beta}' H' T_n',\\
&(ii)\, MSE(\hat{\boldsymbol{\alpha}}^U) - MSE(\hat{\boldsymbol{\alpha}}^{PT}) = \sigma^2 T_n H\boldsymbol{\beta}\boldsymbol{\beta}' H' T_n' G_{n+1,m}(\ell_\alpha;0),\\
&(iii)\, MSE(\hat{\boldsymbol{\alpha}}^{PT}) - MSE(\hat{\boldsymbol{\alpha}}^R) = \sigma^2 T_n H\boldsymbol{\beta}\boldsymbol{\beta}' H' T_n' (1 - G_{n+1,m}(\ell_\alpha;0)),\\
&(iv)\, MSE(\hat{\boldsymbol{\alpha}}^U) - MSE(\hat{\boldsymbol{\alpha}}^S) = c\,\sigma^2 T_n H\boldsymbol{\beta}\boldsymbol{\beta}' H' T_n', \qquad (5.4.1)\\
&(v)\, MSE(\hat{\boldsymbol{\alpha}}^S) - MSE(\hat{\boldsymbol{\alpha}}^{PS}) = \sigma^2 T_n H\boldsymbol{\beta}\boldsymbol{\beta}' H' T_n'\\
&\qquad\qquad \times (E[(1 - c_1 F_{n+1,p}^{-1}(0))^2 \mathcal{I}(F_{n+1,p}(0) < c_1)]),
\end{aligned}
$$

where the right hand sides of the above expressions are positive semidefinite matrices.

Theorem 5.4.3. *Assuming that all the required conditions mentioned in Lemma 5.2.2 are satisfied, the risk expressions associated with the intercepts and slopes are*

$$
\begin{aligned}
R(\hat{\boldsymbol{\beta}}^U; W) &= \sigma^2 tr(W D_{22}) \qquad & R(\hat{\boldsymbol{\alpha}}^U; W) &= \sigma^2 tr(W D_{11})\\
R(\hat{\boldsymbol{\beta}}^R; W) &= \frac{\sigma^2 \mathbf{1}_n' W \mathbf{1}_n}{TQ} + \boldsymbol{\beta}' H' W H \boldsymbol{\beta} & R(\hat{\boldsymbol{\alpha}}^R; W) &= \sigma^2 tr(W D_{11}^*)\\
& & &\quad + \boldsymbol{\beta}' H' T_n' W T_n H \boldsymbol{\beta}
\end{aligned}
$$

$$
\begin{aligned}
R(\hat{\boldsymbol{\beta}}^S; W) =& \sigma^2 tr(WD_{22}) - c(n-1)\sigma^2 tr(WHD_{22}) \\
& \times \{2E[\chi^{-2}_{n+1}(\Delta^2)] - (n-3)E[\chi^{-4}_{n+1}(\Delta^2)]\} \\
& + c(n^2-1)(\boldsymbol{\beta}' H' WH\boldsymbol{\beta})E[\chi^{-4}_{n+3}(\Delta^2)] \\
R(\hat{\boldsymbol{\alpha}}^S; W) =& \sigma^2 tr(WD_{11}) - c(n-1)\sigma^2 tr(WT_n HD_{22}H'T_n')\{2\Delta^2 E[\chi^{-2}_{n+1}(\Delta^2)] \\
& - (n-3)E[\chi^{-4}_{n+1}(\Delta^2)]\} + c(n^2-1)(\boldsymbol{\beta}' H' T_n' WT_n H\boldsymbol{\beta})E[\chi^{-4}_{n+3}(\Delta^2)]
\end{aligned}
$$

$$
\begin{aligned}
R(\hat{\boldsymbol{\beta}}^{PS}; W) =& R(\hat{\boldsymbol{\beta}}^S; W) - (\sigma^2 tr(WHD_{22}) - 2\boldsymbol{\beta}' H' WH\boldsymbol{\beta})E[(1 - c_1 F^{-1}_{n+1,m}(\Delta^2)) \\
& \times \mathcal{I}(F_{n+1,m}(\Delta^2) < c_1)] - \boldsymbol{\beta}' H' WH\boldsymbol{\beta} E[(1 - c_2 F^{-1}_{n+3,m}(\Delta^2)) \\
& \times \mathcal{I}(F_{n+3,m}(\Delta^2) < c_2)] \\
R(\hat{\boldsymbol{\alpha}}^{PS}; W) =& R(\hat{\boldsymbol{\alpha}}^S; W) - (\sigma^2 tr(WT_n HD_{22}H'T_n') - 2\boldsymbol{\beta}' H' T_n' WT_n H\boldsymbol{\beta}) \\
& \times E[(1 - c_1 F^{-1}_{n+1,m}(\Delta^2))\mathcal{I}(F_{n+1,m}(\Delta^2) < c_1)] \\
& - \boldsymbol{\beta}' H' T_n' WT_n H\boldsymbol{\beta} E[(1 - c_2 F^{-1}_{n+3,m}(\Delta^2))\mathcal{I}(F_{n+3,m}(\Delta^2) < c_2)]
\end{aligned}
$$

$$
\begin{aligned}
R(\hat{\boldsymbol{\beta}}^{PT}; W) =& \sigma^2 tr(WD_{22}) - \sigma^2 tr(WHD_{22})G_{n+1,m}(\ell_\alpha; \Delta^2) + \boldsymbol{\beta}' H' WH\boldsymbol{\beta} \\
& \times \{2G_{n+1,m}(\ell_\alpha; \Delta^2) - G_{n+3,m}(\ell^*_\alpha; \Delta^2)\} \\
R(\hat{\boldsymbol{\alpha}}^{PT}; W) =& \sigma^2 tr(WD_{11}) - \sigma^2 tr[W(D_{11} - D^*_{11})]G_{n+1,m}(\ell_\alpha; \Delta^2) + \boldsymbol{\beta}' H' T_n' WT_n H\boldsymbol{\beta} \\
& \times \{2G_{n+1,m}(\ell_\alpha; \Delta^2) - G_{n+3,m}(\ell^*_\alpha; \Delta^2)\}
\end{aligned}
$$

where W is a positive definite weight matrix, $R(\hat{\boldsymbol{\beta}}^*; W) = E((\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta})' W(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}) \mid M)$ and $R(\hat{\boldsymbol{\alpha}}^*; W) = E((\hat{\boldsymbol{\alpha}}^* - \boldsymbol{\alpha})' W(\hat{\boldsymbol{\alpha}}^* - \boldsymbol{\alpha}) \mid M)$.

The proofs of these results are included in the Appendix.

By making use of Equations (5.4.1), one can conclude that under the null hypothesis $R(\hat{\boldsymbol{\beta}}^R; W) < R(\hat{\boldsymbol{\beta}}^U; W)$, $R(\hat{\boldsymbol{\beta}}^{PT}; W) < R(\hat{\boldsymbol{\beta}}^U; W)$, $R(\hat{\boldsymbol{\beta}}^R; W) < R(\hat{\boldsymbol{\beta}}^{PT}; W)$,

$R(\hat{\boldsymbol{\beta}}^S; W) < R(\hat{\boldsymbol{\beta}}^U; W)$ and $R(\hat{\boldsymbol{\beta}}^{PS}; W) < R(\hat{\boldsymbol{\beta}}^S; W)$.

5.5 Experimental results

In this section, we compare the performance of the four estimators defined in the previous section, first by carrying out Monte Carlo simulation studies and then by applying them to a certain mass spectrometry data set.

5.5.1 Simulation study

Monte Carlo simulation studies are conducted for comparing the shrinkage estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in (5.1.1) with those originally proposed. As a measure of accuracy in connection with such estimators, the "relative error" (RE) is considered. The relative error of $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ with respect to the estimators obtained from the original approach, that is, $(\boldsymbol{\alpha}^U, \boldsymbol{\beta}^U)$, is defined by

$$RE(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) = \frac{Er(\boldsymbol{\alpha}^U, \boldsymbol{\beta}^U)}{Er(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)}, \tag{5.5.1}$$

with

$$Er(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) = \sum_{i=1}^{n} (\hat{\alpha}_i^* - \alpha_i)^2 + (\hat{\beta}_i^* - \beta_i)^2, \tag{5.5.2}$$

where the α_i's and β_i's are the assumed values of the parameters for $i = 1, \ldots, n$. Thus, a larger relative error represents an improvement in accuracy over $(\boldsymbol{\alpha}^U, \boldsymbol{\beta}^U)$, the unrestricted (full model) estimators.

It is indicated to test the performance of the estimators both under H_0 and the alternative hypothesis. We carried out simulation studies for different values of Δ where

$\Delta = \|\boldsymbol{\beta} - \beta_0 \mathbf{1}_n\|_2^2$, Δ being equal to zero under the null hypothesis.

Example 5.5.1. In this example, the exact values of the location and scale parameters of length n, are taken to be $\boldsymbol{\alpha} = (0,\, 0.2,\, 0.5,\, 0.7,\, 1,\, 0,\, 0.2,\, 0.5,\ 0.7,\, 1, \ldots)'$ and $\boldsymbol{\beta} = (1,\, 1,\, \ldots,\, 1)'$ for various values of n and T, and the function $m(x)$ is assumed to be $\sin(x)$. The error terms are assumed to be independently and identically distributed $\mathcal{N}(0, 0.25)$ random variables. In this case, we set $x_{it} = x_t$ for $t = 1,\, \ldots,\, T$, and $i = 1,\, \ldots,\, n$, which is the case in many practical problems such as those arising in mass spectrometry data, the T values of x_t being selected randomly from a uniform distribution on the interval $U(0,\, 20)$. The values of $(h,\, h^*)$ were set to be $(0.1,\, 0.1)$ noting that these values had almost the minimum MSPE for different n and T values based on a 5-fold cross-validation approach.

The scatter plots of the simulated observations (x_{it}, y_{it}), $t = 1,\, \ldots,\, T = 2000$ where $i = 1,\, 2,\, 3,\, 4$, are included in the top panel of Figure 5.1 and the graph of the exact function $m(x) = sin(x)$ is plotted over the interval $(0, 20)$ as a solid black line while the estimated functions $\tilde{m}(\cdot)$ and $\hat{m}(\cdot)$ are shown as blue and red dashed lines in the bottom panel of the same figure. It should be pointed out that, in this example, the estimated nonparametric functions (curves) obtained with the other techniques are very close to those determined with the original (unrestricted) method. Thus, we only plotted the unrestricted estimates of the functional components. It is observed that these estimated functions of $\tilde{m}(\cdot)$ and $\hat{m}(\cdot)$ are in a very close agreement with the original function $m(x) = \sin(x)$.

The relative errors associated with the various estimators are shown in Table 5.1 for several values of n and T when $\Delta = 0$. It is seen that when the null hypothesis is valid, the relative errors of the restricted estimators have the largest values, which is acceptable, while the pretest relative errors are similar or rather close to the restricted values. Then, the positive shrinkage technique provides the next best estimators and, of course, these are

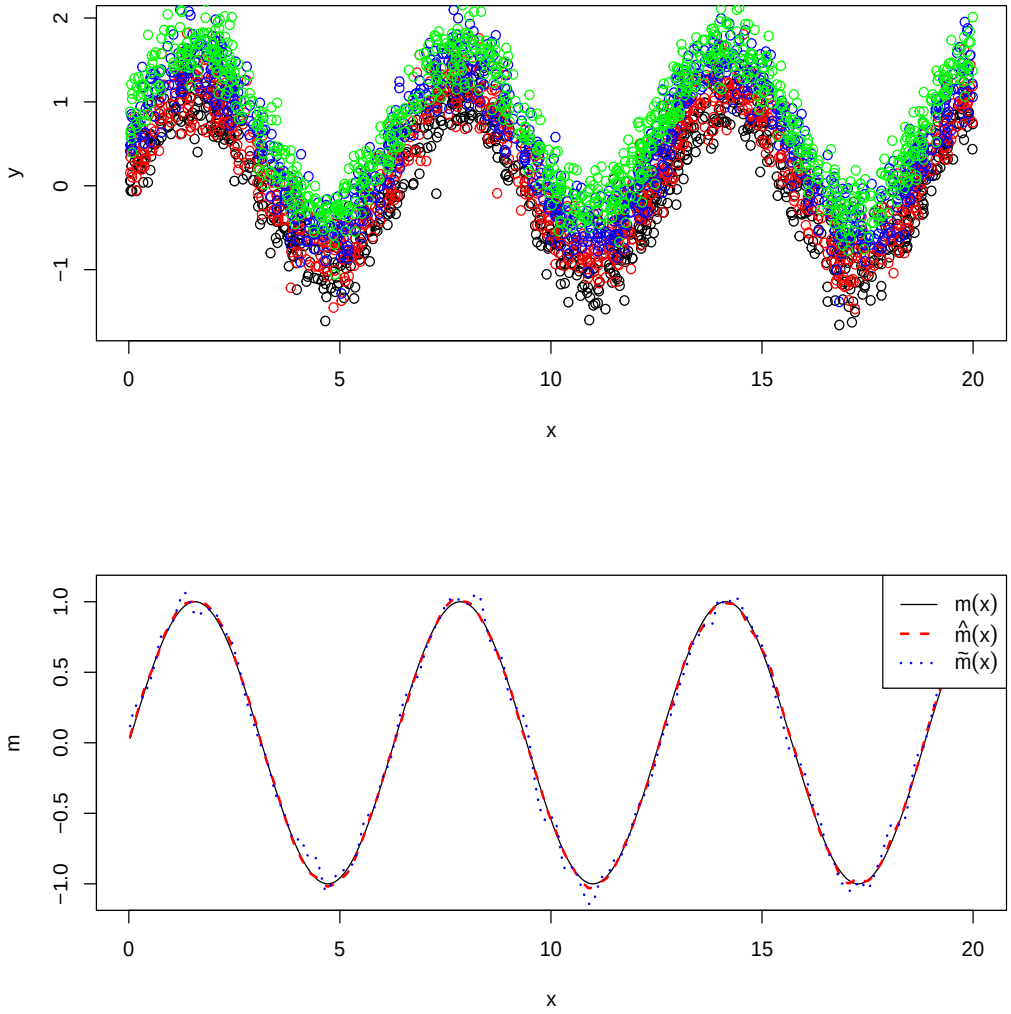


Figure 5.1: Scatter plots of the first four sets of simulated values (top panel); the actual function $m(x) = \sin(x)$ and the estimates $\tilde{m}(x)$ and $\hat{m}(x)$ (bottom panel).

all more accurate than the original estimators as the RE's associated with all of these techniques are greater than 1.4, which confirms that they are more efficient than the standard estimates.

Table 5.1: $RE(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ for certain values of n and T ($\Delta = 0$)

n	T	$RE(\boldsymbol{\alpha}^R, \boldsymbol{\beta}^R)$	$RE(\boldsymbol{\alpha}^{PS}, \boldsymbol{\beta}^{PS})$	$RE(\boldsymbol{\alpha}^{PT}, \boldsymbol{\beta}^{PT})$
10	500	1.7389	1.4152	1.5031
10	1000	1.7440	1.5841	1.7440
10	1500	1.5923	1.5495	1.5923
10	2000	1.9465	1.6802	1.9465
10	2500	1.9358	1.5848	1.9358
10	3000	1.4921	1.3002	1.4921
15	2500	1.6641	1.6045	1.6641
20	2500	1.4713	1.4408	1.4713
25	2500	1.9156	1.7880	1.9156
30	2500	1.6359	1.5931	1.6359
35	2500	1.5651	1.4930	1.5651

In Table 5.2, the relative error of the estimators are shown for various $\Delta \geq 0$ values when $n = 20$ and $T = 1500$. It is observed that the estimators obtained with the restricted method are not as efficient as they are for the case $\Delta = 0$ since the null hypothesis is not valid in this case. However, the relative error associated with the positive shrinkage and pretest estimators are still greater than or equal to one. This corroborates the efficiency of the Stein-type shrinkage techniques regardless of the validity of the null hypothesis.

Example 5.5.2. We now consider the estimation of the location and scale parameters, which are as specified in Example 5.5.1, letting the error terms have a $\mathcal{N}(0,\, 0.25)$ distribution, the functional component being $m(x) = 2/(\exp(x) + \exp(-x))$ on the support $(-7,\, 7)$ in this case. The bandwidths $h = 0.2$ and $h^* = 0.1$ were utilized for various values of n and T.

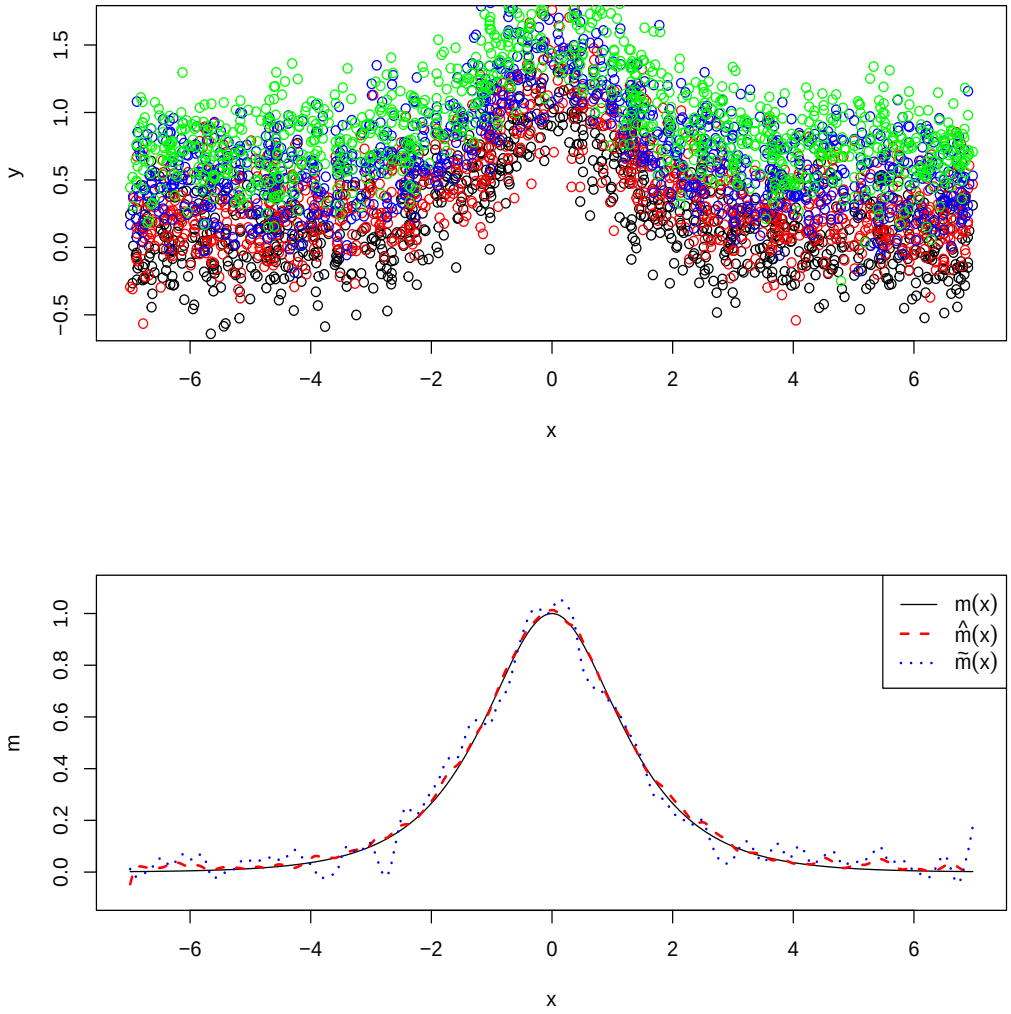


Figure 5.2: Scatter plots of the first four sets of simulated values (top panel); the actual function $m(x) = 2/(exp(x) + exp(-x))$ and the estimates $\tilde{m}(x)$ and $\hat{m}(x)$ (bottom panel).

Table 5.2: $RE(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ for certain $\Delta \geq 0$ with $n = 20$, $T = 1500$, and $m(x) = \sin x$

Δ	$RE(\boldsymbol{\alpha}^R, \boldsymbol{\beta}^R)$	$RE(\boldsymbol{\alpha}^{PS}, \boldsymbol{\beta}^{PS})$	$RE(\boldsymbol{\alpha}^{PT}, \boldsymbol{\beta}^{PT})$
0.00	1.99974	1.90482	1.99974
0.05	0.85712	1.24353	0.97805
0.10	0.36683	1.09256	1.00000
0.15	0.19461	1.05267	1.00000
0.20	0.11837	1.03648	1.00000
0.30	0.05624	1.02329	1.00000
0.40	0.03259	1.01790	1.00000
0.50	0.02124	1.01506	1.00000
0.60	0.01495	1.01332	1.00000
0.70	0.01111	1.01214	1.00000
0.80	0.00860	1.01130	1.00000
0.90	0.00687	1.01065	1.00000
1.00	0.00562	1.01013	1.00000

In Table 5.3, as was the case in the previous example, the relative errors of the estimators are shown for several values of n and T when $\Delta = 0$. In this example, when the null hypothesis is valid, the relative errors of the estimators are even larger than those observed in Example 5.5.1 and thus the efficiency of those estimators is superior in this case.

Table 5.3: $RE(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ for certain values of n and T where $\Delta = 0$ and $m(x) = 2/(\exp(x) + \exp(-x))$

n	T	$RE(\boldsymbol{\alpha}^R, \boldsymbol{\beta}^R)$	$RE(\boldsymbol{\alpha}^{PS}, \boldsymbol{\beta}^{PS})$	$RE(\boldsymbol{\alpha}^{PT}, \boldsymbol{\beta}^{PT})$
10	500	7.1292	1.7358	7.1292
10	1000	3.5141	2.8260	3.5141
10	1500	3.2525	2.5407	3.2525
10	2000	2.9014	2.3964	2.9014
10	2500	2.7548	2.6044	2.7548
10	3000	3.5942	3.4507	3.5942
15	2500	1.7144	1.6971	1.7144
20	2500	3.1391	3.1250	3.1390
25	2500	3.1670	3.0557	3.1670
30	2500	4.8968	4.6132	4.8968
35	2500	2.3400	2.2858	2.3400

In Table 5.4, the relative error of the estimators are shown for different $\Delta \geq 0$ values when $n = 20$ and $T = 1500$. One can again observe that the estimators associated with the restricted method are not as efficient as they are for the case $\Delta = 0$ since the null hypothesis is no longer valid. The relative error associated with the positive shrinkage and pretest estimators are for the most part close to one.

Table 5.4: $RE(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ for certain $\Delta \geq 0$ with $n = 20$, $T = 1500$, and $m(x) = 2/(\exp(x) + \exp(-x))$

Δ	$RE(\boldsymbol{\alpha}^R, \boldsymbol{\beta}^R)$	$RE(\boldsymbol{\alpha}^{PS}, \boldsymbol{\beta}^{PS})$	$RE(\boldsymbol{\alpha}^{PT}, \boldsymbol{\beta}^{PT})$
0.00	3.63055	2.82008	3.18199
0.05	1.87993	1.81232	1.59694
0.10	0.92114	1.26926	0.95903
0.15	0.56465	1.12593	1.00000
0.20	0.37634	1.06943	1.00000
0.30	0.19538	1.02086	1.00000
0.40	0.11837	1.00419	1.00000
0.50	0.07908	0.99689	1.00000
0.60	0.05652	0.99321	1.00000
0.70	0.04243	0.99118	1.00000
0.80	0.03306	0.99001	1.00000
0.90	0.02653	0.98930	1.00000
1.00	0.02180	0.98888	1.00000

5.5.2 Application to a mass spectrometry data set

The shrinkage estimation techniques are now applied to an actual mass spectrometry data set, namely, the SELDI-TOF mass spectrometry data set collected from a study on liver cancer patients conducted at Changzheng Hospital Shanghai. The accuracy of these methods are measured by evaluating their mean squared error (MSE):

$$MSE^* = \frac{1}{T}\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it} - \hat{\alpha}_i^* - \hat{\beta}_i^*\hat{m}(x_t)^*)^2. \tag{5.5.3}$$

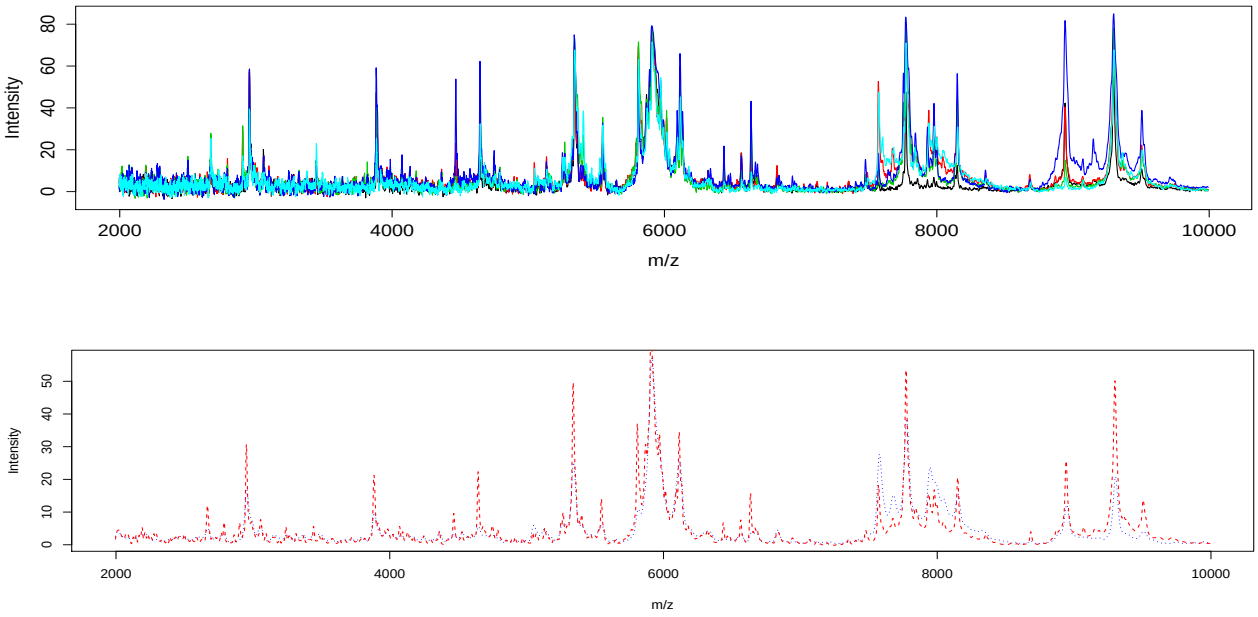


Figure 5.3: The observations on five individuals randomly selected from the mass spectrometry data set (top panel); the estimates $\tilde{m}(x)$ in blue and $\hat{m}(x)$ in red (bottom panel).

Table 5.5 shows the MSE values in connection with different estimating approaches. It is seen that, the MSE values associated with the positive shrinkage technique is the smallest and that the pretest estimator and original method gave the same level of accuracy even though the null hypothesis is not valid. As previously explained, the restricted estimation produces the largest MSE value.

Table 5.5: MSE values obtained from various techniques applied to a mass spectrometry data set.

	Original	R	PS	PT
MSE	13.48971	17.95069	13.38967	13.48971

The plots of the observations of 5 randomly selected individuals from the data sets are presented in the top panel of Figure 5.3, while the plots of the estimates $\tilde{m}(x)$ and $\hat{m}(x)$ of the nonparametric component of the model are shown in the bottom panel of Figure 5.3. In this case, $n = 35$ and $T = 21000$.

References

Ahmed, S.A. (1994). Improved estimation in a multivariate regression model, Computational Statistics and Data Analysis, 17(5), 537–554.

Ahmed, S.E. and Krzanowski, W.J. (2004). Biased estimation in a simple multivariate regression model, Computational Statistics and Data Analysis, 45, 689–696.

Ali, A.M. (1990). Interface of preliminary test approach and empirical Bayes approach to shrinkage estimation, PhD Thesis, Carleton University, Canada.

Baggerly, K.A., Morris, J.S. and Coombes, K.R. (2004). Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments, Bioinformatics, 20, 777–785.

Baggerly, K.A., Morris, J.S., Wang, J., Gold, D., Xiao, L.C., and Coombes, K.R. (2003). A comprehensive approach to the analysis of MALDI-TOF proteomics spectra from serum samples, Proteomics, 3, 1667–1672.

Begun, J., Hall, W.J., Huang, W.M. and Wellner, J.A. (1983). Information and asymptotic efficiency in parametric-nonparametric models, The Annals of Statistics, 11, 432–52.

Chitsaz, S. and Ahmed, S.E. (2012a). Shrinkage estimation for the regression parameter matrix in multivariate regression model, Journal of Statistical Computation and Simulation, 82, 309–323.

Chitsaz, S, and Ahmed, S.E. (2012b). An improved estimation in regression parameter matrix in multivariate regression model, Communications in Statistics - Theory and Methods, 41, 2305–2320.

Fan, J. and Gijbels, I. (1996). Local Polynomial Modelling and Its Applications, London, Chapman and Hall.

Guilhaus, M. (1995). Principles and instrumentation in time-of-flight mass spectrometry, J. Mass Spectrom., 30, 1519–32.

Judge, G.G. and Bock, M.E. (1978). The Statistical Implication of Pretest and Stein-Rule Estimators in Econometrics, Amsterdam, North Holland.

Ma, W., Feng, Y., Chen, K. and Ying, Z. (2015). Functional and parametric estimation in a semi- and nonparametric model with application to mass-spectrometry data, The International Journal of Biostatistics, 11(2), 285-303.

Robinson, P.M. (1988). Root-N consistent semiparametric regression, Econometrica, 55, 931-54.

Roy, P., Truntzer, C., Maucort-Boulch, D., Jouve, T. and Molinari, N. (2011). Protein mass spectra data analysis for clinical biomarker discovery: a global review, Briefings Bioinf, 12, 176–86.

Ruppert, D., Wand, M. and Carroll, R. (2003). Semiparametric Regression, Cambridge University Press, New York.

Saleh, A.K.E. (2006). Theory and Preliminary Test and Stein-type Estimation with Applications, Wiley, New York.

Stone, C.J. (1982). Optimal global rates of convergence for nonparametric regression, The Annals of Statistics, 10, 1040–1053.

Yasui, Y., Pepe, M., Thompson, M.L., Adam, B.L., Wright, G.L. Jr, Qu, Y., Potter, J.D., Winget, M., Thornquist, M. and Feng, Z. (2003). A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection, Biostatistics, 4(3), 449–463.

Appendices

Some asymptotic results are formally proved in Appendix A2.1 and the results of a simulation study that corroborates the usefulness of certain shrinkage and penalty techniques in connection with the multivariate linear regression model are presented in Appendix A2.2.

A2 .1. Proofs of Theorems 5.4.1, 5.4.2 and 5.4.3

The following Lemma will be used to establish the asymptotic results:

Lemma .1.1. *(Judge and Bock, 1978) Let the p-vector $\mathbf{x}$ have a $\mathcal{N}_p(\boldsymbol{\mu}_{\mathbf{x}}, \Sigma_{\mathbf{x}})$ distribution. Then, for a measurable function of ϕ, one has*

$$E[\mathbf{x}\phi(\mathbf{x}'\mathbf{x})] = \boldsymbol{\mu}_{\mathbf{x}} E[\phi(\chi^2_{p+2}(\Delta^2))] \tag{A2.1.1}$$

and

$$E[\mathbf{x}\mathbf{x}'\phi(\mathbf{x}'\mathbf{x})] = \Sigma_{\mathbf{x}} E[\phi(\chi^2_{p+2}(\Delta^2))] + \boldsymbol{\mu}_{\mathbf{x}}\boldsymbol{\mu}'_{\mathbf{x}} E[\phi(\chi^2_{p+4}(\Delta^2))] \tag{A2.1.2}$$

where $\Delta^2 = \boldsymbol{\mu}'_{\mathbf{x}}\Sigma^{-1}_{\mathbf{x}}\boldsymbol{\mu}_{\mathbf{x}}$.

A2 .1.1. Proof of Theorem 5.4.1

The biases associated with $\hat{\boldsymbol{\beta}}^U$, $\hat{\boldsymbol{\alpha}}^U$, $\hat{\boldsymbol{\beta}}^R$ and $\hat{\boldsymbol{\alpha}}^R$ follow from the results stated in Theorem 5.3.1. Additionally, one has

$$\mathbf{b}(\hat{\boldsymbol{\beta}}^S) = E[\hat{\boldsymbol{\beta}}^U - \boldsymbol{\beta} - c\boldsymbol{\mathcal{L}}_n^{-1} H\hat{\boldsymbol{\beta}}^U] = -c(n-1)H\boldsymbol{\beta}E[\phi(\chi^{-2}_{p+2}(\Delta^2))],$$
$$\mathbf{b}(\hat{\boldsymbol{\alpha}}^S) = E[\hat{\boldsymbol{\alpha}}^U - \boldsymbol{\alpha} + c\boldsymbol{\mathcal{L}}_n^{-1} T_n H\hat{\boldsymbol{\beta}}^U] = c(n-1)T_n H\boldsymbol{\beta}E[\phi(\chi^{-2}_{p+2}(\Delta^2))],$$

$$
\begin{aligned}
\mathbf{b}(\hat{\boldsymbol{\beta}}^{PS}) &= E[\hat{\boldsymbol{\beta}}^{R} - \boldsymbol{\beta} + (1 - c\mathcal{L}_n^{-1})H\hat{\boldsymbol{\beta}}^{U}\mathcal{I}(\mathcal{L}_n > c)] \\
&= -H\boldsymbol{\beta}\{G_{n+1,m}(c_1;\Delta^2) + c_1 E[F_{n+1,m}^{-1}(\Delta^2)\mathcal{I}(F_{n+1,m}(\Delta^2) > c_1)]\}, \\
\mathbf{b}(\hat{\boldsymbol{\alpha}}^{PS}) &= E[\hat{\boldsymbol{\alpha}}^{U} - \boldsymbol{\alpha} + \{1 - (1 - c\mathcal{L}_n^{-1})\mathcal{I}(\mathcal{L}_n > c)\}T_n H\hat{\boldsymbol{\beta}}^{U}] \\
&= c_1 T_n H\boldsymbol{\beta}\,\{E[F_{n+1,m}^{-1}(\Delta^2)] - E[F_{n+1,m}^{-1}(\Delta^2)\mathcal{I}(F_{n+1,m}(\Delta^2) > c_1)]\} \\
&\quad + T_n H\boldsymbol{\beta}\, G_{n+1,m}(c_1;\Delta^2),
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{b}(\hat{\boldsymbol{\beta}}^{PT}) &= E[\hat{\boldsymbol{\beta}}^{U} - \boldsymbol{\beta} - H\hat{\boldsymbol{\beta}}^{U}\,\mathcal{I}(\mathcal{L}_n < \mathcal{C}_\alpha)] = -H\boldsymbol{\beta}\, G_{n+1,m}(\ell_\alpha;\Delta^2), \\
\mathbf{b}(\hat{\boldsymbol{\alpha}}^{PT}) &= E[\hat{\boldsymbol{\alpha}}^{U} - \boldsymbol{\alpha} + T_n H\hat{\boldsymbol{\beta}}^{U}\,\mathcal{I}(\mathcal{L}_n < \mathcal{C}_\alpha)] = T_n H\boldsymbol{\beta}\, G_{n+1,m}(\ell_\alpha;\Delta^2).
\end{aligned}
$$

□

A2 .1.2. Proof of Theorem 5.4.2

Likewise, the MSE's associated with $\hat{\boldsymbol{\beta}}^{U}$, $\hat{\boldsymbol{\alpha}}^{U}$, $\hat{\boldsymbol{\beta}}^{R}$ and $\hat{\boldsymbol{\alpha}}^{R}$ can easily be derived from Theorem 5.3.1. The MSE's of other estimators are obtained as follows:

$$
\begin{aligned}
MSE(\hat{\boldsymbol{\beta}}^{S}) &= E[(\hat{\boldsymbol{\beta}}^{U} - c\,\mathcal{L}_n^{-1}H\hat{\boldsymbol{\beta}}^{U} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}^{U} - c\,\mathcal{L}_n^{-1}H\hat{\boldsymbol{\beta}}^{U} - \boldsymbol{\beta})'] \\
&= E[(\hat{\boldsymbol{\beta}}^{U} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}^{U} - \boldsymbol{\beta})'] - 2cHE[\mathcal{L}_n^{-1}\hat{\boldsymbol{\beta}}^{U}(\hat{\boldsymbol{\beta}}^{U} - \boldsymbol{\beta})'] \\
&\quad + c^2 HE[\mathcal{L}_n^{-2}\hat{\boldsymbol{\beta}}^{U}\hat{\boldsymbol{\beta}}'^{U}]H' \\
&= \sigma^2 D_{22} - c(n-1)\sigma^2 H D_{22}\{2E[\chi_{n+1}^{-2}(\Delta^2)] - (n-3)E[\chi_{n+1}^{-4}(\Delta^2)]\} \\
&\quad + c(n^2-1)(H\boldsymbol{\beta}\boldsymbol{\beta}'H')E[\chi_{n+3}^{-4}(\Delta^2)] \\
MSE(\hat{\boldsymbol{\alpha}}^{S}) &= E[(\hat{\boldsymbol{\alpha}}^{U} + c\,\mathcal{L}_n^{-1}T_n H\hat{\boldsymbol{\beta}}^{U} - \boldsymbol{\alpha})(\hat{\boldsymbol{\alpha}}^{U} + c\,\mathcal{L}_n^{-1}T_n H\hat{\boldsymbol{\beta}}^{U} - \boldsymbol{\alpha})'] \\
&= \sigma^2 D_{11} - c(n-1)\sigma^2 T_n H D_{22} H' T_n'\{2\Delta^2 E[\chi_{n+1}^{-2}(\Delta^2)] - (n-3)E[\chi_{n+1}^{-4}(\Delta^2)]\} \\
&\quad + c(n^2-1)(T_n H\boldsymbol{\beta}\boldsymbol{\beta}'H'T_n')E[\chi_{n+3}^{-4}(\Delta^2)]
\end{aligned}
$$

$$
\begin{aligned}
MSE(\hat{\boldsymbol{\beta}}^{PT}) &= E[(\hat{\boldsymbol{\beta}}^{U} - H\hat{\boldsymbol{\beta}}^{U}\,\mathcal{I}(\mathcal{L}_n < \mathcal{C}_\alpha) - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}^{U} - H\hat{\boldsymbol{\beta}}^{U}\,\mathcal{I}(\mathcal{L}_n < \mathcal{C}_\alpha) - \boldsymbol{\beta})'] \\
&= E[(\hat{\boldsymbol{\beta}}^{U} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}^{U} - \boldsymbol{\beta})'] - 2HE[\hat{\boldsymbol{\beta}}^{U}\,\mathcal{I}(\mathcal{L}_n < \mathcal{C}_\alpha)(\hat{\boldsymbol{\beta}}^{U} - \boldsymbol{\beta})'] \\
&\quad + HE[\mathcal{I}(\mathcal{L}_n < \mathcal{C}_\alpha)\hat{\boldsymbol{\beta}}^{U}\hat{\boldsymbol{\beta}}'^{U}]H' \\
&= \sigma^2 D_{22} - \sigma^2 H D_{22} G_{n+1,m}(\ell_\alpha; \Delta^2) + H\boldsymbol{\beta}\boldsymbol{\beta}'H' \\
&\quad \times \{2G_{n+1,m}(\ell_\alpha; \Delta^2) - G_{n+3,m}(\ell^*_\alpha; \Delta^2)\} \\
MSE(\hat{\boldsymbol{\alpha}}^{PT}) &= E[(\hat{\boldsymbol{\alpha}}^{U} + T_n H\hat{\boldsymbol{\beta}}^{U}\,\mathcal{I}(\mathcal{L}_n < \mathcal{C}_\alpha) - \boldsymbol{\alpha}) \times (\hat{\boldsymbol{\alpha}}^{U} + T_n H\hat{\boldsymbol{\beta}}^{U}\,\mathcal{I}(\mathcal{L}_n < \mathcal{C}_\alpha) - \boldsymbol{\alpha})'] \\
&= \sigma^2 D_{11} + 2T_n H\,E[\hat{\boldsymbol{\beta}}^{U}\,\mathcal{I}(\mathcal{L}_n < \mathcal{C}_\alpha)(\hat{\boldsymbol{\alpha}}^{U} - \boldsymbol{\alpha})'] \\
&\quad + T_n HE[\mathcal{I}(\mathcal{L}_n < \mathcal{C}_\alpha)\hat{\boldsymbol{\beta}}^{U}\hat{\boldsymbol{\beta}}'^{U}]H'T_n' \\
&= \sigma^2 D_{11} - \sigma^2 (D_{11} - D^*_{11}) G_{n+1,m}(\ell_\alpha; \Delta^2) + T_n H\boldsymbol{\beta}\boldsymbol{\beta}'H'T_n' \\
&\quad \times \{2G_{n+1,m}(\ell_\alpha; \Delta^2) - G_{n+3,m}(\ell^*_\alpha; \Delta^2)\}
\end{aligned}
$$

$$
\begin{aligned}
MSE(\hat{\boldsymbol{\beta}}^{PS}) &= E(\mathbf{z}\mathbf{z}') \qquad \text{where} \quad \{\mathbf{z} \equiv \hat{\boldsymbol{\beta}}^{U} - (1 - c\,\mathcal{L}_n^{-1})\,\mathcal{I}(\mathcal{L}_n > c)H\hat{\boldsymbol{\beta}}^{U} - \boldsymbol{\beta}\} \\
&= MSE(\hat{\boldsymbol{\beta}}^{S}) - (\sigma^2 H D_{22} - 2H\boldsymbol{\beta}\boldsymbol{\beta}'H') \\
&\quad \times E[(1 - c_1 F^{-1}_{n+1,m}(\Delta^2))\mathcal{I}(F_{n+1,m}(\Delta^2) < c_1)] \\
&\quad - H\boldsymbol{\beta}\boldsymbol{\beta}'H'E[(1 - c_2 F^{-1}_{n+3,m}(\Delta^2))\mathcal{I}(F_{n+3,m}(\Delta^2) < c_2)] \\
MSE(\hat{\boldsymbol{\alpha}}^{PS}) &= E[(\hat{\boldsymbol{\alpha}}^{U} + \{1 - (1 - c\,\mathcal{L}_n^{-1})\,\mathcal{I}(\mathcal{L}_n > c)\}T_n H\hat{\boldsymbol{\beta}}^{U} - \boldsymbol{\alpha}) \\
&\quad \times (\hat{\boldsymbol{\alpha}}^{U} + \{1 - (1 - c\,\mathcal{L}_n^{-1})\,\mathcal{I}(\mathcal{L}_n > c)\}T_n H\hat{\boldsymbol{\beta}}^{U} - \boldsymbol{\alpha})'] \\
&= MSE(\hat{\boldsymbol{\alpha}}^{S}) - (\sigma^2 T_n H D_{22} H'T_n' - 2T_n H\boldsymbol{\beta}\boldsymbol{\beta}'H'T_n')E[(1 - c_1 F^{-1}_{n+1,m}(\Delta^2)) \\
&\quad \times \mathcal{I}(F_{n+1,m}(\Delta^2) < c_1)] - T_n H\boldsymbol{\beta}\boldsymbol{\beta}'H'T_n' \\
&\quad \times E[(1 - c_2 F^{-1}_{n+3,m}(\Delta^2))\mathcal{I}(F_{n+3,m}(\Delta^2) < c_2)].
\end{aligned}
$$

□

For detailed derivations of the MSE's associated with pretest and shrinkage estimators, one can refer to (Ali, 1990).

A2 .1.3. Proof of Theorem 5.4.3

Clearly, in a $p \times p$ matrix $A = [a_{ij}]$, the $tr(A)$ is the sum of the diagonal elements of A. By making use of the properties of the trace and the definition of the risk, that is, $R(\hat{\boldsymbol{\beta}}^*; W) = E((\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta})'W(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}) \,|\, M) = tr(WMSE(\hat{\boldsymbol{\beta}}^*))$, the risk expressions follow directly from the MSE expressions given in Theorem 5.4.2.

□

A2 .2. A simulation study of certain shrinkage and penalty techniques in connection with the multivariate linear regression model

In this section, we compare the relative mean squared error (RMSE) of the restricted, shrinkage, pretest, penalty (LASSO and SCAD) estimators described in Section 1.2 to the unrestricted model estimator for various values of n, p and p_1. Monte Carlo simulation experiments have been carried out to examine the performance of these methods with respect to their MSE's. The following linear model was simulated:

$$y_i = x_{1,i}\beta_1 + x_{2,i}\beta_2 + \cdots + x_{p,i}\beta_p + \epsilon_i,\ i = 1,\ \ldots,\ n,$$

where $x_{1,i}$ and $x_{2,i}$ are selected randomly and independently from a $\mathcal{N}(1,\ 2)$ distribution, $x_{s,i},\ i = 3,\ \ldots,\ p$ are iid $\mathcal{N}(0,\ 1)$ and the error terms ϵ_i are also iid $\mathcal{N}(0.\,1)$. We wish to test the null hypothesis $H_0 :\ \beta_j = 0$ for $j = p_1 + 1,\ \ldots,\ n$.

As discussed earlier, we partition the parameter vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ as $\boldsymbol{\beta}^{(0)} = (\boldsymbol{\beta}_1', \mathbf{0}')'$. Let $\Delta^* = \|\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}\|$ where $\|\cdot\|$ denotes the L_2 norm. To compare the perfor-

mance of these estimators, the MSE of an estimator of $\boldsymbol{\beta}_1$ obtained from each method was determined as well as the value of the MSE for the full model estimator, and the following ratio was considered:

$$\text{RMSE}(\hat{\boldsymbol{\beta}}_1^{UE}, \hat{\boldsymbol{\beta}}_1^{*}) = \frac{\text{MSE}\,(\hat{\boldsymbol{\beta}}_1^{UE})}{\text{MSE}\,(\hat{\boldsymbol{\beta}}_1^{*})},$$

where $\hat{\boldsymbol{\beta}}_1^{*}$ is an estimator of one of the aforementioned types.

In Table A2.1, first, we compare the performance of various estimators first keeping p and p_1 fixed and then keeping n fixed in the second part of the table.

Table A2.1: RMSE's of various estimators for several values of n, p and p_1

n	p	p_1	$\hat{\beta}_1^{RE}$	$\hat{\beta}_1^{S}$	$\hat{\beta}_1^{S+}$	$\hat{\beta}_1^{PT}$	$\hat{\beta}_1^{LASSO}$	$\hat{\beta}_1^{SCAD}$
50	15	5	7.91	2.50	2.51	2.37	1.41	1.23
60	15	5	7.45	2.28	2.26	2.31	1.40	1.95
70	15	5	7.42	2.26	2.29	2.43	1.37	2.06
80	15	5	7.47	2.31	2.34	2.36	1.40	2.03
90	15	5	6.33	2.11	2.12	2.12	1.40	1.90
100	15	5	6.64	2.12	2.12	2.13	1.40	1.87
100	20	7	5.11	1.73	1.73	1.73	1.15	1.62
100	25	10	3.95	1.53	1.55	1.55	1.06	1.49
100	35	15	3.62	1.48	1.49	1.48	1.02	1.47
100	50	20	4.67	1.75	1.79	1.74	1.07	1.77

In Table A2.2 we are interested in observing the effects of changing n and p while keeping $p_1 = 4$ and $\Delta^* = 0$. We see that by increasing p, the accuracy is improving in nearly every case.

It is expected that when $\Delta^* = 0$, the RE's will be the more accurate estimators; however when $\Delta^* > 0$, the scenario changes and the RE's are no longer preferred, as can be seen from the RMSE values presented in Table A2.3.

Table A2.2: RMSE's of various estimators for several values of p and n and $\Delta^* = 0$

n	p	p_1	$\hat{\beta}_1^{RE}$	$\hat{\beta}_1^{S}$	$\hat{\beta}_1^{S+}$	$\hat{\beta}_1^{PT}$	$\hat{\beta}_1^{LASSO}$	$\hat{\beta}_1^{SCAD}$
30	10	4	8.38	2.76	2.76	3.06	1.85	2.31
30	14	4	14.67	3.62	3.73	3.83	2.01	3.15
30	18	4	31.16	5.14	3.64	5.97	2.80	4.61
50	10	4	7.51	2.75	2.75	2.76	1.95	2.32
50	14	4	11.63	3.04	3.05	3.10	1.87	2.58
50	18	4	16.82	3.68	3.72	3.61	1.73	3.11
100	10	4	6.21	2.41	2.40	2.45	1.75	2.02
100	14	4	9.62	2.69	2.69	2.67	1.61	2.50
100	18	4	16.02	3.33	3.50	3.50	1.73	3.06
120	10	4	5.86	2.45	2.42	2.55	1.80	2.19
120	14	4	9.39	2.79	2.76	2.82	1.63	2.39
120	18	4	14.11	2.97	2.99	2.95	1.55	2.68

Table A2.3: RMSE's for several values of Δ^*

Δ^*	n	p	p_1	$\hat{\beta}_1^{RE}$	$\hat{\beta}_1^{S}$	$\hat{\beta}_1^{S+}$	$\hat{\beta}_1^{PT}$	$\hat{\beta}_1^{LASSO}$	$\hat{\beta}_1^{SCAD}$
0	50	15	6	5.10	1.89	1.87	1.94	1.28	1.69
0.06	50	15	6	4.88	2.01	2.02	2.04	1.27	1.75
0.12	50	15	6	3.92	1.90	1.92	1.95	1.25	1.75
0.2	50	15	6	2.97	1.89	1.91	1.89	1.38	1.72
0.35	50	15	6	1.71	1.86	1.86	1.78	1.36	1.61
0.5	50	15	6	1.05	1.76	1.77	1.64	1.30	1.66
0.85	50	15	6	0.44	1.60	1.56	1.54	1.32	1.79
1	50	15	6	0.34	1.66	1.66	1.58	1.34	1.79
2	50	15	6	0.09	1.55	1.55	1.52	1.23	1.68

Chapter 6

Concluding remarks and further research

6.1 Concluding remarks

Certain univariate moment based techniques have been extended to the bivariate case and it was explained that they are applicable in higher dimensions. In Chapter 2, it was assumed that a bivariate density function could be approximated as the product of an initial density function and a bivariate polynomial adjustment. The coefficients of the polynomial adjustment were determined by making use of the joint moments of the exact and initial density functions and a polynomial degree selection criterion was proposed. Then, it was established that adjustments by means of standard and orthogonal polynomials produce the same approximants. Then, this technique was extended to obtain density estimates to estimate an appropriate density function in which case sample moments are utilized in lieu of exact moments of the exact density function. Several illustrative examples were provided. It was also explained that the methodology is applicable to 'big data'. One example involved trivariate data and application of the technique to regression analysis was pointed out.

In the next chapter, a popular class of univariate distributions known as Pearson's family of distributions or Pearson's frequency curves, as well as certain extensions which were previously proposed, were reviewed. Then, a general extension of Pearson's distributions, which is called differentiated log-density approximant (DLDA), was introduced and an explicit solution of the resulting PDF was obtained. It was explained that this methodology also applies in the context of density estimation.

In Chapter 4, an extension of DLDA technique to bivariate random vectors was considered. Using the technique discussed in Chapter 2, a bivariate DLDA density estimation was developed. These chapters conclude with several applications. Unlike kernel density estimates, the density estimates discussed in this thesis have functional representations which makes them amenable to algebraic manipulations.

In Chapter 5, a semi-nonparametric regression model proposed by Ma *et al.* (2015) for analyzing mass spectrometry data sets, was studied. It involved n similar regression models with identical nonparametric component, and n location and scale parameters. The pretest and shrinkage techniques were applied for estimating the parametric components of the model in order to obtain more accurate estimates.

6.2 Further research

Further investigations on the applicability of the density estimation techniques proposed in this dissertation and comparisons to currently used techniques could be carried out. The multivariate density estimation techniques discussed in this thesis could possibly be generalized to estimate the density functions of certain random matrices.

The DLDA technique is an extension of the Pearson curves which makes use of standard polynomials. One could study the possibility of utilizing orthogonal polynomials in both the numerator and the denominator of the rational function. As well, an extension of the methodology that would directly produce bivariate density approximations and the connections to copulas shall be investigated.

The regression methodology discussed in Chapter 5 could be extended to more complex semi-nonparametric models involving several nonparametric components.

Appendix A

The *Mathematica* code utilized in connection with the main numerical examples presented in this dissertation is included in Appendix A.

Appendix A : Mathematica Code

A .1 Modules Used in Chapter 2

Module used for Example 2.1.1. (Univariate mixture normal densities)

```
fY[y_] := fY[y] = 1/2 1/Sqrt[2 π 2] (e^-(y+4)^2/8) + 1/2 1/Sqrt[2 π 3] (e^-(y-2)^2/18);
μY[h_] := μY[h] = 1/2 Integrate[y^h 1/Sqrt[2 π 2] (e^-(y+4)^2/8), {y, -∞, ∞}] + 1/2 Integrate[y^h 1/Sqrt[2 π 3] (e^-(y-2)^2/18), {y, -∞, ∞}]
n = 15;
Table[μY[h], {h, 0, n}];
u = μY[1];
s = Sqrt[μY[2] - μY[1]^2];
μX[j_] := μX[j] = Expand[((Y - u)/s)^j] /. Y^h_. :> μY[h];
w[x_] := e^-x^2/2;
H*k_[x_] := (-1)^k 2^-k/2 HermiteH[k, x/Sqrt[2]];
θk_ := θk = Sqrt[2 π] k!;
fYn[y_] := w[(y - u)/s] Sum[1/(s θi) (H*i[x] /. x^j_. :> μX[j]) H*i[(y - u)/s], {i, 0, n}];
f1Yn[y_] := w[(y - u)/s] Sum[1/(s θi) (H*i[x] /. x^j_. :> Expand[((Y - u)/s)^j] /. Y^h_. :> μY[h]) H*i[(y - u)/s], {i, 0, n}];
S7 = Show[Plot[fY[y], {y, -15, 12}, PlotRange → All], Plot[Evaluate[fYn[x]], {x, -15, 12},
    PlotRange → All, PlotStyle → {Dashing[{0.01, 0.01}], RGBColor[0, 0, 1]}]]
```

Example 2.3.1. Approximation of the density of an equal mixture of bivariate Gaussian densities by means of bivariate orthogonal polynomials

```
Needs["MultivariateStatistics`"]
m1 = {1.1, -.1}; V1 = {{.33, .03}, {0.03, .33}};
m2 = {.2, 1.2}; V2 = {{.4, .04}, {.04, .4}};
A1 = Inverse[V1];
A2 = Inverse[V2];
Plot3D[
 .5 PDF[MultinormalDistribution[m1, V1], {x, y}] + .5 PDF[MultinormalDistribution[m2, V2],
```

135

```
    {x, y}], {x, -2, 3}, {y, -2, 3}, PlotRange → All, PlotLabel → "Exact Density"]
```

$$\text{mgf}[t1\_, t2\_] := \frac{1}{2} e^{m1.\{t1,t2\}+\{t1,t2\}.V1.\{t1,t2\}/2} + \frac{1}{2} e^{m2.\{t1,t2\}+\{t1,t2\}.V2.\{t1,t2\}/2}$$

```
jm[r_, s_] :=
  jm[r, s] = ((Derivative[r, s][mgf] /. {#1 → t1, #2 → t2})[[1]]) /. {t1 → 0, t2 → 0};
jm[0, 0] = 1;
μ = {jm[1, 0], jm[0, 1]}
V = {{jm[2, 0] - jm[1, 0]^2, jm[1, 1] - jm[1, 0] jm[0, 1]},
  {jm[1, 1] - jm[1, 0] jm[0, 1], jm[0, 2] - jm[0, 1]^2}}
A = Inverse[V];
```

$$\text{fbc}[x\_, y\_] := \text{fbc}[x, y] = \frac{e^{-\{x,y\}.\{x,y\}/2}}{2\pi}$$

```
Plot3D[PDF[MultinormalDistribution[{0, 0}, IdentityMatrix[2]], {x, y}],
 {x, -3, 3}, {y, -3, 3},
 PlotRange → All, PlotLabel → "Standardized Base Density"]

Vhi = MatrixPower[V, -1/2];
Vh = MatrixPower[V, 1/2];
Plot3D[.5 PDF[MultinormalDistribution[Vhi.(m1 - μ), Vhi.V1.Vhi], {x, y}] +
  .5 PDF[MultinormalDistribution[Vhi.(m2 - μ), Vhi.V2.Vhi], {x, y}], {x, -3, 3},
 {y, -3, 3}, PlotRange → All, PlotLabel → "Exact Standardized Density"]
```

$$\text{mgfn}[t1\_, t2\_] := \frac{1}{2} e^{(\text{Vhi}.(m1-\mu)).\{t1,t2\}+\{t1,t2\}.\text{Vhi}.V1.\text{Vhi}.\{t1,t2\}/2} + \frac{1}{2} e^{(\text{Vhi}.(m2-\mu)).\{t1,t2\}+\{t1,t2\}.\text{Vhi}.V2.\text{Vhi}.\{t1,t2\}/2}$$

```
mn2[r_, s_] := mn2[r, s] =
   Chop[((Derivative[r, s][mgfn] /. {#1 → t1, #2 → t2})[[1]]) /. {t1 → 0, t2 → 0}];
mn2[0, 0] = 1;
H1[i_, j_] := H1[i, j] = Expand[
    Simplify[(-1)^(i+j) / fbc[x, y] ((Derivative[i, j][fbc] /. {#1 → x, #2 → y}))[[1]]]];
H1[0, 0] = 1;
Unprotect[Power]; e^(0.`x^2+0.`xy+0.`y^2) = 1; Protect[Power];

Hd1[i_, j_] := Hd1[i, j] = Chop[Apart[ ExpectedValue[Expand[(x + i X)^i (y + i Y)^j],
     MultinormalDistribution[{0, 0}, {{1, 0}, {0, 1}}], {X, Y}]]]
```

$$\theta[i\_, j\_, k\_, l\_] := \theta[i, j, k, l] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \text{fbc}[x, y]\, \text{Expand}[H1[i, j]\, Hd1[k, l]]\, dx\, dy;$$

```
h[n_, x1_, y1_] :=
 h[n, x1, y1] = Sum[Sum[(1/(k! l!)) Sum[Sum[(CoefficientList[Hd1[k, l], {x, y}][[i + 1, j + 1]])
          mn2[i, j], {j, 0, l}], {i, 0, k}] Chop[Expand[H1[k, l]]] /. {x → x1, y → y1},
     {l, 0, n}], {k, 0, n}]

ListPlot3D[Flatten[Table[Chop[{w1, w2, Evaluate[fbc[x1, y1] h[9, x1, y1]]} Det[Vhi]] /.
    {x1 → Vhi[[1, 1]] (w1 - μ[[1]]) + Vhi[[1, 2]] (w2 - μ[[2]]),
     y1 → Vhi[[2, 1]] (w1 - μ[[1]]) +
       Vhi[[2, 2]] (w2 - μ[[2]])}, {w1, -2, 3, 1/4}, {w2, -2, 3, 1/4}], 1],
 PlotRange -> All, PlotLabel → 9 ]
```

Example 2.4.1. Bivariate 'xclara' dataset

```
ClearAll[ YY, Y1, Y2, Y, X1, X2];
mydata = Import["xclara.csv", "CSV"];
Unprotect[Power]; 0^0 = 1; Protect[Power];
header = mydata[[1]];
data = mydata[[2 ;;]];
myDataset = Thread[header → #] & /@ data // Map[Association] // Dataset;
data = mydata[[2 ;;]];
Y1 = data[[All, 2]];
Y2 = data[[All, 3]];
n = Length[Y1]
data1 = Transpose[{Y1, Y2}];
Y = Table[{Y1[[j]], Y2[[j]]}, {j, 1, n}];
X1 = Table[Y[[i]][[1]], {i, 1, n}];
X2 = Table[Y[[i]][[2]], {i, 1, n}];

dat1 = Table[{Y1[[j]], Y2[[j]]}, {j, 1, n}];
Plot3D[
 Evaluate[PDF[SmoothKernelDistribution[dat1, "SheatherJones", "Gaussian"], {x, y}]],
 {x, Min[Y1] - 5, Max[Y1] + 5}, {y, Min[Y2] - 5, Max[Y2] + 5}, Filling → Axis,
 Exclusions → None, PlotLabel → "SheatherJones-Gaussian", PlotRange → All]

me1 = {Sum[X1[[j]] / n, {j, n}], Sum[X1[[j]] / n, {j, n}]};
V1 = 1/(n - 1) {{Sum[(X1[[j]] - me1[[1]])^2, {j, n}],
     Sum[(X1[[j]] - me1[[1]]) (X2[[j]] - me1[[2]]), {j, n}]},
    {Sum[(X1[[j]] - me1[[1]]) (X2[[j]] - me1[[2]]), {j, n}],
     Sum[(X2[[j]] - me1[[2]])^2, {j, n}]}};
Vhi = MatrixPower[V1, -1 / 2];
Vh = MatrixPower[V1, 1 / 2];
detVhi = Det[Vhi]
YY = Transpose[Vhi.Transpose[Y - ConstantArray[me1, n]]];
X1 = Table[YY[[i]][[1]], {i, 1, n}];
X2 = Table[YY[[i]][[2]], {i, 1, n}];

me = {0, 0};
V = IdentityMatrix[2];
ddx = (Max[X1] - Min[X1]) / 100;
ddy = (Max[X2] - Min[X2]) / 100;
Needs["MultivariateStatistics`"];
(*   The joint sample moments of orders r and s:   *)
jm[r_, s_] := jm[r, s] = Sum[X1[[j]]^r X2[[j]]^s, {j, n}] / n
(*  The base density   *)
fnt[x_, y_] := fnt[x, y] = PDF[MultinormalDistribution[me, V], {x, y}]
(*Plot3D[fnt[x,y],{x,Min[X1]-2ddx,Max[X1]+2ddx},{y,Min[X2]-2ddy,Max[X2]+2ddy},
```

```
 PlotRange→ All, PlotLabel→"Standardized Base Density"]*)
t1 = 11; t2 = 7;

Off[Inner::"normal"]
f3[L1_List, L2_List] := Inner[Plus, L1, L2, List];
L3 = Flatten[Table[{j, i}, {i, 0, t1}, {j, 0, t2}], 1];
P3 = Table[f3[L3[[i]], L3[[j]]], {i, 1, Length[L3] }, {j, 1, Length[L3]}];
mgf[tt1_, tt2_] := e^(me.{tt1,tt2}+{tt1,tt2}.V.{tt1,tt2}/2
mm2[r_, s_] :=
  mm2[r, s] = ((Derivative[r, s][mgf] /. {#1 → tt1, #2 → tt2})[[1]]) /. {tt1 → 0, tt2 → 0};
mm2[0, 0] = 1;
M4 = Rationalize[Table[mm2[P3[[i, j]][[1]], P3[[i, j]][[2]]],
    {i, Length[L3]}, {j, Length[L3]}], 10^-25];
Zv1[x_, y_] := Zv1[x, y] = Flatten[Table[x^j y^i, {i, 0, t1}, {j, 0, t2}], 1];
jm[r_, s_] := jm[r, s] = Sum[X1[[j]]^r X2[[j]]^s, {j, n}] / n
Gms[i_] := Gms[i] = jm [L3[[i, 1]], L3[[i, 2]]] ;
μ = Table[Gms[i], {i, Dimensions[L3][[1]]}] // N;
c4 = LinearSolve[M4, μ];
t5[x_, y_] := t5[x, y] = fnt[x, y] c4.Zv1[x, y];
cst = NIntegrate[t5[x, y],
  {x, Min[X1] - 5 ddx, Max[X1] + 5 ddx}, {y, Min[X2] - 5 ddy, Max[X2] + 5 ddy}]
ft[x_, y_] := detVhi t5[Vhi[[1, 1]] (x - me1[[1]]) + Vhi[[1, 2]] (y - me1[[2]]),
    Vhi[[2, 1]] (x - me1[[1]]) + Vhi[[2, 2]] (y - me1[[2]])] / cst
Plot3D[ft[x, y], {x, Min[Y1] - 5 ddx, Max[Y1] + 5 ddx}, {y, Min[Y2] - 5 ddy, Max[Y2] + 5 ddy},
 PlotRange → All, PlotLabel → t[t1, t2]]
F1[x_, y_] := F1[x, y] = NIntegrate[ft[x1, y1], {x1, Min[Y1], x}, {y1, Min[Y2], y}]
(*Timing[F1[X1[[2500]],X2[[2500]]]]*)
Off[NIntegrate::izero]
Off[NIntegrate::"ncvb"]
edis = EmpiricalDistribution[Y];
EmpCDF = Table[CDF[edis, Y[[j]]], {j, n}];
EstCDF = Table[F1[Y1[[j]], Y2[[j]]], {j, n}];
Sum[(EmpCDF[[j]] - Max[0, EstCDF[[j]] ])^2, {j, n}]
```

Example 2.4.3. Trivariate ‘CommViolPredUnnormalizedData’ dataset

```
ClearAll[ Y0, Y1, Y2, Y3, Y, X1, X2, X3];
mydata = Import["CommViolPredUnnormalizedData.csv", "CSV"];
header = mydata[[1]];
data = mydata[[2 ;;]];
Unprotect[Power]; 0^0 = 1; Protect[Power];
myDataset = Thread[header → #] & /@ data // Map[Association] // Dataset;
data = mydata[[2 ;;]];
Y1 = Table[data[[i, 16]], {i, Length[data]}];
Y2 = Table[data[[i, 13]], {i, Length[data]}];
Y3 = Table[data[[i, 22]], {i, Length[data]}];
n = Length[Y1]
```

```
Y = Table[{Y1[[j]], Y2[[j]]}, {j, 1, n}];
data1 = Table[{Y1[[j]], Y2[[j]], Y3[[j]]}, {j, 1, n}];
mea1 = Mean[Y1]; mea2 = Mean[Y2]; mea3 = Mean[Y3];
Var = Covariance[data1];
Vhi = MatrixPower[Var, -1/2]; detVhi = Det[Vhi];
Vh = MatrixPower[Var, 1/2]; detVh = Det[Vh];
X1 = Vhi[[1, 1]] (Y1 - ConstantArray[mea1, n]) + Vhi[[1, 2]] (Y2 - ConstantArray[mea2, n]) +
   Vhi[[1, 3]] (Y3 - ConstantArray[mea3, n]) + 2.5;
X2 = Vhi[[2, 1]] (Y1 - ConstantArray[mea1, n]) + Vhi[[2, 2]] (Y2 - ConstantArray[mea2, n]) +
   Vhi[[2, 3]] (Y3 - ConstantArray[mea3, n]) + 2.5;
X3 = Vhi[[3, 1]] (Y1 - ConstantArray[mea1, n]) + Vhi[[3, 2]] (Y2 - ConstantArray[mea2, n]) +
   Vhi[[3, 3]] (Y3 - ConstantArray[mea3, n]);

me1 = Sum[X1[[j]]/n, {j, n}];
me2 = Sum[X2[[j]]/n, {j, n}];
me3 = Sum[X3[[j]]/n, {j, n}];
x2bar1 = Sum[(X1[[j]]^2)/n, {j, n}];
x2bar2 = Sum[(X1[[j]]^2)/n, {j, n}];
V3 = 1/(n - 1) Sum[(X3[[j]] - me3)^2, {j, n}];
alpha1 = me1^2/(x2bar1 - me1^2);
beta1 = (x2bar1 - me1^2)/me1;
alpha2 = me2^2/(x2bar2 - me2^2);
beta2 = (x2bar2 - me2^2)/me2;
jm[n1_, n2_, n3_] := Sum[X1[[j]]^n1 X2[[j]]^n2 X3[[j]]^n3, {j, n}]/n
jm[0, 0, 0];
Off[MLE::shdw];
Needs["MultivariateStatistics`"];

f0X1[x_] := PDF[GammaDistribution[alpha1, beta1], x]
f0X2[y_] := PDF[GammaDistribution[alpha2, beta2], y]
f0X3[z_] := PDF[NormalDistribution[me3, Sqrt[V3]], z]
fnt1[x_, y_, z_] := f0X1[x] f0X2[y] f0X3[z]
moments1[h1_] := beta1^h1 Gamma[alpha1 + h1]/Gamma[alpha1];  moments1[0] = 1;
moments2[h2_] := beta2^h2 Gamma[alpha2 + h2]/Gamma[alpha2];  moments2[0] = 1;
mgf[tt1_] := e^(me3 (tt1) + tt1^2 V3/2)
moments3[r_] := ((Derivative[r][mgf] /. {#1 → tt1})[[1]]) /. tt1 → 0;  moments3[0] = 1;
mm3[p_, q_, r_] := moments1[p] moments2[q] moments3[r];

t1 = 7;  t2 = 5;  t3 = 4;
Off[Inner::"normal", Inner::intpm]
f3[L1_List, L2_List] := Inner[Plus, L1, L2, List]
L3 = Flatten[Table[{k, j, i}, {i, 0, t3}, {j, 0, t2}, {k, 0, t1}], 2];
L = Length[L3];
P3 = Table[f3[L3[[i]], L3[[j]]], {i, 1, L}, {j, 1, L}];
M4 = Rationalize[
   Table[mm3[P3[[i, j]][[1]], P3[[i, j]][[2]], P3[[i, j]][[3]]], {i, L}, {j, L}], 10^-25];
Zv1[x_, y_, z_] := Zv1[x, y, z] = Flatten[
```

```
    Table [x^k y^j z^i, {i, 0, t3}, {j, 0, t2}, {k, 0, t1}], 2];
Gms[i_] := Gms[i] = jm [L3[[i, 1]], L3[[i, 2]], L3[[i, 3]]] ;
mu = Table[Gms[i], {i, Dimensions[L3][[1]]}] // N;
c4 = LinearSolve[M4, mu];
t33[x_, y_, z_] := fnt1[x, y, z] c4.Zv1[x, y, z]
cst = NIntegrate[t33[x, y, z], {x, 0, 20}, {y, -6, 6}, {z, 0, 20}]
t4[x_, y_, z_] :=
 detVhi t33[Vhi[[1, 1]] (x - mea1) + Vhi[[1, 2]] (y - mea2) + Vhi[[1, 3]] (z - mea3) + 2.3,
    Vhi[[2, 1]] (x - mea1) + Vhi[[2, 2]] (y - mea2) + Vhi[[2, 3]] (z - mea3),
    Vhi[[3, 1]] (x - mea1) + Vhi[[3, 2]] (y - mea2) + Vhi[[3, 3]] (z - mea3) + 2.5] / cst

(*Marginal joint density of (X, Y) :*)
tm1[x_, y_] := NIntegrate[t4[x, y, z], {z, Min[Y3] - 2, Max[Y3] + 2}];
tmp1[x_, y_] :=
  If[Min[Y1] < x < Max[Y1] && Min[Y2] < y < Max[Y2] && tm1[x, y] ≥ 0, tm1[x, y], 0];
Plot3D[tmp1[x, y], {x, Min[Y1] - 2, Max[Y1] + 2}, {y, Min[Y2] - 2, Max[Y2] + 2},
 PlotRange → All, PlotLabel → XY t[t1, t2, t3]]

dat1 = Table[{Y1[[j]], Y2[[j]]}, {j, 1, n}];
D = SmoothKernelDistribution[dat1];
Plot3D[PDF[D, {x, y}], {x, Min[Y1] - 2, Max[Y1] + 2},
 {y, Min[Y2] - 2, Max[Y2] + 2}, PlotLabel → SmoothKernel XY, PlotRange → All]
ListPlot[dat1, PlotLabel → ListPlot XY]


(*Marginal joint density of (Y,Z):*)
tm2[y_, z_] := NIntegrate[t4[x, y, z], {x, Min[Y1] - 2, Max[Y1] + 2}];
tmp2[y_, z_] :=
  If[Min[Y2] < y < Max[Y2] && Min[Y3] < z < Max[Y3] && tm2[y, z] ≥ 0, tm2[y, z], 0];
Plot3D[tmp2[y, z], {y, Min[Y2] - 2, 30}, {z, Min[Y3] - 2, Max[Y3] + 2},
 PlotRange → All, PlotLabel → YZ t[t1, t2, t3]]

dat2 = Table[{Y2[[j]], Y3[[j]]}, {j, 1, n}];
D2 = SmoothKernelDistribution[dat2];
Plot3D[PDF[D2, {y, z}], {y, Min[Y2] - 2, 30},
 {z, Min[Y3] - 2, Max[Y3] + 2}, PlotLabel → SmoothKernel YZ, PlotRange → All]
ListPlot[dat2, PlotLabel → ListPlot YZ]

(*Marginal joint density of (X,Z):*)
tm3[x_, z_] := NIntegrate[t4[x, y, z], {y, 0, 55}];
tmp3[x_, z_] := Max[tm3[x, z], 0];
Plot3D[tmp3[x, z], {x, 0, 30}, {z, Min[Y3] - 2, Max[Y3] + 2},
 PlotRange → All, PlotLabel → XZ t[t1, t2, t3]]

dat3 = Table[{Y2[[j]], Y3[[j]]}, {j, 1, n}];
D3 = SmoothKernelDistribution[dat3];
Plot3D[PDF[D3, {x, z}], {x, 0, 30}, {z, Min[Y3] - 2, Max[Y3] + 2},
 PlotLabel → SmoothKernel XZ, PlotRange → All]
ListPlot[dat3, PlotLabel → ListPlot XZ]
```

Example 2.4.4. Bivariate ‘covtype’ dataset

```
ClearAll[ YY, Y1, Y2, Y, X1, X2];
mydata = Import["covtype.csv", "CSV"];
Unprotect[Power]; 0^0 = 1; Protect[Power];
header = mydata[[1]];
data = mydata[[2 ;;]];
myDataset = Thread[header → #] & /@ data // Map[Association] // Dataset;
data = mydata[[2 ;;]];
Y1 = data[[All, 2]];
Y2 = data[[All, 3]];
n = Length[Y1]
data1 = Transpose[{Y1, Y2}];
Unprotect[Power]; 0^0 = 1; Protect[Power];
Y = Table[{Y1[[j]], Y2[[j]]}, {j, 1, n}];
X1 = Table[Y[[i]][[1]], {i, 1, n}];
X2 = Table[Y[[i]][[2]], {i, 1, n}];
me1 = {∑_{j=1}^{n} X1[[j]] / n, ∑_{j=1}^{n} X2[[j]] / n};
V1 = 1/(n - 1) {{Sum[(X1[[j]] - me1[[1]])^2, {j, n}],
    Sum[(X1[[j]] - me1[[1]]) (X2[[j]] - me1[[2]]), {j, n}]},
   {Sum[(X1[[j]] - me1[[1]]) (X2[[j]] - me1[[2]]), {j, n}],
    Sum[(X2[[j]] - me1[[2]])^2, {j, n}]}};
Vhi = MatrixPower[V1, -1/2];
Vh = MatrixPower[V1, 1/2];
detVhi = Det[Vhi];
YY = Transpose[Vhi.Transpose[Y - ConstantArray[me1, n]]];
X1 = Table[YY[[i]][[1]], {i, 1, n}];
X2 = Table[YY[[i]][[2]], {i, 1, n}];
me = {0, 0};
V = IdentityMatrix[2];
ddx = 2 (Max[X1] - Min[X1]) / (Max[X1] + Min[X1]);
ddy = 2 (Max[X2] - Min[X2]) / (Max[X2] + Min[X2]);
Needs["MultivariateStatistics`"];
jm[r_, s_] := jm[r, s] = Sum[X1[[j]]^r X2[[j]]^s, {j, n}] / n
fnt[x_, y_] := fnt[x, y] = PDF[MultinormalDistribution[me, V], {x, y}]

t1 = 7; t2 = 4;
Off[Inner::"normal"]
f3[L1_List, L2_List] := Inner[Plus, L1, L2, List];
L3 = Flatten[Table[{j, i}, {i, 0, t1}, {j, 0, t2}], 1];
P3 = Table[f3[L3[[i]], L3[[j]]], {i, 1, Length[L3] }, {j, 1, Length[L3]}];
mgf[tt1_, tt2_] := e^(me.{tt1,tt2}+{tt1,tt2}.V.{tt1,tt2}/2
mm2[r_, s_] :=
```

```
  mm2[r, s] = ((Derivative[r, s][mgf] /. {#1 → tt1, #2 → tt2})[[1]]) /. {tt1 → 0, tt2 → 0};
mm2[0, 0] = 1;
M4 = Rationalize[Table[mm2[P3[[i, j]][[1]], P3[[i, j]][[2]]],
    {i, Length[L3]}, {j, Length[L3]}], 10^-25];
Zv1[x_, y_] := Zv[x, y] = Flatten[Table[x^j y^i, {i, 0, t1}, {j, 0, t2}], 1];
jm[r_, s_] := jm[r, s] = Sum[X1[[j]]^r X2[[j]]^s, {j, n}] / n
Gms[i_] := Gms[i] = jm [L3[[i, 1]], L3[[i, 2]]] ;
μ = Table[Gms[i], {i, Dimensions[L3][[1]]}] // N;
c4 = LinearSolve[M4, μ];
t5[x_, y_] := t5[x, y] = fnt[x, y] c4.Zv1[x, y];
Plot3D[t5[x, y], {x, Min[X1] - 5 ddx, Max[X1] + 5 ddx},
 {y, Min[X2] - 5 ddy, Max[X2] + 5 ddy}, PlotRange → All, PlotLabel → T5[t1, t2]]
tpositive[x_, y_] := tpositive[x, y] = Max[t5[x, y], 0];
Off[NIntegrate::"slwcon"]
cst = NIntegrate[tpositive[x, y],
  {x, Min[X1] - 5 ddx, Max[X1] + 5 ddx}, {y, Min[X2] - 5 ddy, Max[X2] + 5 ddy}]
ft[x_, y_] := detVhi t5[Vhi[[1, 1]] (x - me1[[1]]) + Vhi[[1, 2]] (y - me1[[2]]),
    Vhi[[2, 1]] (x - me1[[1]]) + Vhi[[2, 2]] (y - me1[[2]])] / cst
Plot3D[ft[x, y], {x, Min[Y1] - 10, Max[Y1] + 15}, {y, Min[Y2] - 5, 50}, PlotRange → All]
Histogram3D[data1]
𝒟 = SmoothKernelDistribution[data1];
Plot3D[PDF[𝒟, {x, y}], {x, Min[Y1] - 10, Max[Y1] + 15},
 {y, Min[Y2] - 5, Max[Y2]}, PlotRange → All]
```

Example 2.4.5. Applying this technique to nonparametric regression y=m(x)+ε

```
Off[NIntegrate::"slwcon"]
ClearAll[G, K, a, c, Z1, PE, PA1, rts, Sol]
I_u_,v_[z_] := If[u ≤ z < v, 1, 0]
(*The exact function*)
f_E[x_] := f_E[x] = (I_0,1[x] x + I_1,2[x] + I_2,3[x] (x - 1)) / 3
n = 20000;
α = 0; β = 3; ν = 5; δ = 2 ;
PE = Plot[f_E[y], {y, 0, 3}, PlotRange → All,
   PlotLabel → Nonparametric, PlotStyle → RGBColor[0, 0, 1]];
Y1 = RandomVariate[UniformDistribution[{0, 3}], n];
eps = RandomVariate[NormalDistribution[0, 0.05], n];
Y2 = Table[f_E[Y1[[i]]] + eps[[i]], {i, n}];
Y = Table[{Y1[[i]], Y2[[i]]}, {i, n}];
data1 = Table[{Y1[[j]], Y2[[j]]}, {j, 1, n}];
SCAT = ListPlot[Y, PlotRange → All, AxesOrigin → {0, 0}]
X1 = Y1;
X2 = Y2;

fhatLegendre[x_, y_, data_, n_] := Module[{μ, std, newdata, w, cw, nn, d,
     dd, a, b, cov, invcov, xx, yy, X, ax0, bx0, xx1, yy1, by0, ay0, fx, aa, bb},
    nn = Length[data];
    xx1 = Table[data[[i]][[1]], {i, 1, nn}];
```

```
    yy1 = Table[data[[i]][[2]], {i, 1, nn}];
    ax0 = Min[xx1] - (Quantile[xx1, 0.1] - Min[xx1]);
    bx0 = Max[xx1] + (Max[xx1] - Quantile[xx1, 0.9]);
    ay0 = Min[yy1] - (Quantile[yy1, 0.1] - Min[yy1]);
    by0 = Max[yy1] + (Max[yy1] - Quantile[yy1, 0.9]);
    (*ax0=-15;
    bx0=15;
    ay0=-15;
    by0=15;*)
    xx = (2 xx1 - (ax0 + bx0))/(bx0 - ax0);
yy = (2 yy1 - (ay0 + by0))/(by0 - ay0);
w[z1_, z2_] := 1;
    cw = 1/4;
a[i_, j_] := If[i == 0 && j == 0, 0, Sqrt[(2 i + 1)/2] Sqrt[(2 j + 1)/2]
      Mean[Expand[LegendreP[i, x]] Expand[LegendreP[j, y]] /. {x → xx, y → yy}]];
aa = Table[a[i, j], {i, 0, n}, {j, 0, n - i}];
    d[i_, j_] := If[i == 0 && j == 0, 0, (2 i + 1)/2 (2 j + 1)/2
      Mean[Expand[LegendreP[i, x]^2] Expand[LegendreP[j, y]^2] /. {x → xx, y → yy}]];
    dd = Table[d[i, j], {i, 0, n}, {j, 0, n - i}];
    b[i_, j_] := If[i == 0 && j == 0, 0,
      If[0 < (nn aa[[i + 1, j + 1]]^2 - dd[[i + 1, j + 1]])/((nn - 1) aa[[i + 1, j + 1]]^2) < 1,
       (nn aa[[i + 1, j + 1]]^2 - dd[[i + 1, j + 1]])/((nn - 1) aa[[i + 1, j + 1]]^2), 0]];
    bb = Table[b[i, j], {i, 0, n}, {j, 0, n - i}];
fx = (4/((bx0 - ax0) (by0 - ay0)))
      (cw + Sum[Sum[aa[[i + 1, j + 1]] bb[[i + 1, j + 1]] Sqrt[(2 i + 1)/2] Sqrt[(2 j + 1)/2] LegendreP[i, z1]
          LegendreP[j, z2], {j, 0, n - i}], {i, 0, n}] /. {z1 → (2 x - (ax0 + bx0))/(bx0 - ax0), z2 → (2 y - (ay0 + by0))/(by0 - ay0)};
    fx
   ]
  (*dt1=apple[[1]];*)
dt1 = data1;
Clear[x]
m = 30
  fapple1 = fhatLegendre[x, y, data1, m];
nn = Length[dt1]
  xx1 = Sort[Table[dt1[[i]][[1]], {i, 1, nn}]];
yy1 = Sort[Table[dt1[[i]][[2]], {i, 1, nn}]];
```

143

```
  sa1 = Min[dt1] - (Quantile[xx1, 0.05] - xx1[[1]])
  sb1 = Max[dt1] + (yy1[[nn]] - Quantile[xx1, 0.95])
  fhat[a_, b_] := fhat[a, b] = fapple1 /. {x → a, y → b}
  ft0[x_, y_] := ft0[x, y] = Max[fhat[x, y], 0]
  cst = NIntegrate[ fhat[x, y], {x, 0, 3}, {y, sa1, sb1}]
  cstPLUS = NIntegrate[ ft0[x, y], {x, 0, 3}, {y, sa1, sb1}]
  ft[x_, y_] := ft[x, y] = ft0[x, y]
  ft[1, 1] // N
PTPLUS = Plot3D[fhat[x, y] / cstPLUS,
    {x, 0, 3}, {y, sa1, sb1}, PlotRange → All, PlotLabel → Legendre[m]]
ContourPlot[fhat[x, y] / cstPLUS, {x, 0, 3}, {y, sa1, sb1},
   PlotRange → All, ContourLabels → Automatic]
DensityPlot[fhat[x, y] / cstPLUS, {x, 0, 3}, {y, sa1, sb1}, Mesh → 10]
Histogram3D[data1, Automatic, "Probability"]
fx[x_] := fx[x] = FindMaximum[ft[x, y] / 3, {y}][[2, 1, 2]]
REGRE = Plot[fx[x], {x, 0, 3}, PlotStyle → {Red, Dashed, Thickness[0.02]}]
AA = Plot[f_E[x], {x, 0, 3}, PlotStyle → {Blue, Thick}]
Show[AA, REGRE]
```

Example 2.4.6. Square within a triangle

```
<< Histograms`
(*<<Graphics`Graphics`*)
SetDirectory["H:\\research\\my Package"];
Get["Density.m"]
Off[FindRoot::"cvmit"]
Sim[n_] := Module[{r, θ, x2, y2, xy2, xy3, xyc, xys, xyt},
  r = Sqrt[RandomReal[{0, 6.3 π}, n]];
  θ = RandomReal[{0, 2 π}, n];
x2 = r Cos[θ];
  y2 = r Sin[θ];
xy2 = Table[{x2[[i]], y2[[i]]}, {i, 1, n}];
Triangle[list_] :=
   Max[-2 list[[2]], list[[2]] - list[[1]] √3, list[[2]] + list[[1]] √3];
xyt = Select[xy2, Triangle[#] < 4π/3 + 0.2 && Triangle[#] > 4π/3 - 0.2 &];
  circle[list_] := √(list[[1]]^2 + list[[2]]^2);
(*xyc=Select[xy2, circle[#] <1.7+0.2&& circle[#]>1.7-0.2&];*)
  square[list_] := Max[Abs[list[[1]]], Abs[list[[2]]]];
xys = Select[xy2, square[#] < π/4 + 0.2 && square[#] > π/4 - 0.2 &];
  (*xy3=Flatten[{xyc,xys,xyt},1];*)
  xy3 = Flatten[{xyt, xys}, 1];
  xy3
 ]
n0 = 100 000;
SeedRandom[22];
```

```
dt = Sim[n0];
ListPlot[dt]
nn = Length[dt]
fhatLegendre1[x_, y_, data_, n_] := Module[
  {μ, std, newdata, w, cw, nn, d, dd, a, aa, b, cov, invcov, xx, yy, X, a0, b0, xx1, yy1},
  nn = Length[data];
  xx1 = Sort[Table[data[[i]][[1]], {i, 1, nn}]];
yy1 = Sort[Table[data[[i]][[2]], {i, 1, nn}]];
  a0 = Min[data] - (xx1[[3]] - xx1[[1]]);
  b0 = Max[data] + (yy1[[nn]] - yy1[[nn - 2]]);
  newdata = (2 data - (a0 + b0))/(b0 - a0);
xx = Table[newdata[[i]][[1]], {i, 1, nn}];
yy = Table[newdata[[i]][[2]], {i, 1, nn}];
w[z1_, z2_] := 1;
  cw = 1/4;
X = (2 {x, y} - (a0 + b0))/(b0 - a0);
a[i_, j_] := If[i == 0 && j == 0, 0, Sqrt[(2 i + 1)/2] Sqrt[(2 j + 1)/2]
     Mean[Expand[LegendreP[i, x]] Expand[LegendreP[j, y]] /. {x → xx, y → yy}]];
aa = Table[a[i, j], {i, 0, n}, {j, 0, n - i}];
  d[i_, j_] := If[i == 0 && j == 0, 0, (2 i + 1)/2 (2 j + 1)/2
     Mean[Expand[LegendreP[i, x]^2] Expand[LegendreP[j, y]^2] /. {x → xx, y → yy}]];
  dd = Table[d[i, j], {i, 0, n}, {j, 0, n - i}];
  b[i_, j_] := If[i == 0 && j == 0, 0,
    If[0 < (nn aa[[i + 1, j + 1]]^2 - dd[[i + 1, j + 1]])/((nn - 1) aa[[i + 1, j + 1]]^2) < 1,
     (nn aa[[i + 1, j + 1]]^2 - dd[[i + 1, j + 1]])/((nn - 1) aa[[i + 1, j + 1]]^2), 0]];
  bb = Table[b[i, j], {i, 0, n}, {j, 0, n - i}];
(2/(b0 - a0))^2 (cw + Sum[Sum[aa[[i + 1, j + 1]] bb[[i + 1, j + 1]] Sqrt[(2 i + 1)/2] Sqrt[(2 j + 1)/2]
          LegendreP[i, z1] LegendreP[j, z2], {j, 0, n - i}], {i, 0, n}]) /. {z1 → X[[1]], z2 → X[[2]]}
 ]
ODS1[data_, end_] := Module[{M, nn, cov, invcov, μ, std,
   newdata, xx, yy, a, d, aa, dd, J, JM, de, temp, b, a0, b0, xx1, yy1},
  M = end;
  nn = Length[data];
  xx1 = Sort[Table[data[[i]][[1]], {i, 1, nn}]];
yy1 = Sort[Table[data[[i]][[2]], {i, 1, nn}]];
  a0 = Min[data] - (xx1[[3]] - xx1[[1]]);
  b0 = Max[data] + (yy1[[nn]] - yy1[[nn - 2]]);
```

```
   newdata = (2 data - (a0 + b0))/(b0 - a0);
xx = Table[newdata[[i]][[1]], {i, 1, nn}];
yy = Table[newdata[[i]][[2]], {i, 1, nn}];
a[i_, j_] := If[i == 0 && j == 0, 0, Sqrt[(2 i + 1)/2] Sqrt[(2 j + 1)/2]
      Mean[Expand[LegendreP[i, x]] Expand[LegendreP[j, y]] /. {x → xx, y → yy}]];
   d[i_, j_] := If[i == 0 && j == 0, 0, ((2 i + 1)/2) ((2 j + 1)/2)
      Mean[Expand[LegendreP[i, x]^2] Expand[LegendreP[j, y]^2] /. {x → xx, y → yy}]];
aa = Table[a[i, j], {i, 0, M}, {j, 0, M - i}];
   dd = Table[d[i, j], {i, 0, M}, {j, 0, M - i}];
   J[m_] := Sum_{i=0}^{m} Sum_{j=0}^{m-i} (2 dd[[i + 1, j + 1]] - (nn + 1) aa[[i + 1, j + 1]]^2);
   JM = Table[J[m], {m, 0, M}];
   de = Range[0, M, 1];
test1 = Table[{de[[i]], JM[[i]]}, {i, 1, Length[JM]}];
temp = Table[If[test1[[i, 2]] == Min[JM], test1[[i]][[1]], 0], {i, 1, Length[JM]}];
Total[temp]
 ]
fcom1 = fhatLegendre1[x, y, dt, 50];
nn = Length[dt];
xx1 = Sort[Table[dt[[i]][[1]], {i, 1, nn}]];
yy1 = Sort[Table[dt[[i]][[2]], {i, 1, nn}]];
coma0 = Min[dt] - (xx1[[3]] - xx1[[1]]);
comb0 = Max[dt] + (yy1[[nn]] - yy1[[nn - 2]]);
Plot3D[fcom1 /. {x → x, y → y}, {x, coma0, comb0}, {y, coma0, comb0}]
ContourPlot[fcom1, {x, coma0, comb0}, {y, coma0, comb0}, ContourLabels → Automatic]
DensityPlot[fcom1, {x, coma0, comb0}, {y, coma0, comb0}, Mesh → 10]
```

A .2 Modules Used in Chapter 3

Example 3.4.1. Monte Carlo simulation study

```
ClearAll[G, K, a, c, Z1, PE, PA1, rts, Sol]
Unprotect[Power]; 0^0 = 1; Protect[Power];
𝒟1 = BetaDistribution[2, 10]; (*Beta Distribution*)
(* 𝒟1=ExponentialDistribution[2]; (*Exponential Distribution*)*)
(* 𝒟1=GammaDistribution[2,20]; (*Gamma Distribution*)*)
(* 𝒟1=StudentTDistribution[5]; (*Exponential Distribution*) *)
f1[x_] := f1[x] = PDF[𝒟1, {x}]
F1[x_] := F1[x] = CDF[𝒟1, {x}]
```

```
ISD[m_, 𝒟1_] :=
 Module[{data, Y1, X1, edis1, EmCDF1, me1, V1, LB, UB, Sol, g1, h1, ISD1, ISD2},
  ClearAll[G, K, a, c, Z1, rts, data, Y1, X1, edis1, EmCDF1,
   me1, V1, LB, UB, Sol, g1, h1, ISD1, , ISD2, KSDf, KSDh];
  f1[x_] := f1[x] = PDF[𝒟1, {x}];
  F1[x_] := F1[x] = CDF[𝒟1, {x}];
  data = RandomVariate[𝒟1, m];
  Y1 = data;
  n = Length[Y1];
  edis1 = EmpiricalDistribution[Y1];
  EmCDF1 = Table[CDF[edis1, Y1[[j]]], {j, 1, n}];
  me1 = ∑_{j=1}^{n} Y1[[j]] / n;
  V1 = Sqrt[1/(n - 1) Sum[(Y1[[j]] - me1)^2, {j, n}]];
  X1 = Table[(Y1[[i]] - me1) / V1, {i, 1, n}];
  μX[h_] := μX[h] = Mean[X1^h];
  α = Min[X1]; β = Max[X1]; ν = 3; δ = 2;
  LS[r_, ν_, δ_] :=
   LS[r, ν, δ] = (∑_{i=0}^{ν} K[i] μX[r + i] == (r + δ) μX[r + δ - 1] + ∑_{j=0}^{δ-1} (r + j) G[j] μX[r + j - 1]);
  Z1 = Solve[Table[LS[r, ν, δ], {r, 0, ν + δ}],
    Flatten[{Table[K[i], {i, 0, ν}], Table[G[j], {j, 0, δ - 1}]}]];
  a = Table[Z1[[1, i + 1, 2]], {i, 0, ν}];
  c = Table[Z1[[1, i + 1, 2]], {i, ν + 1, ν + δ}];
  c = Append[c, 1];
  rt = Solve[∑_{i=0}^{δ} c[[i + 1]] y^i == 0, y];
  rts = Table[rt[[i, 1, 2]], {i, 1, δ}];
  Sol =
   DSolve[w'[x] == -(∑_{i=0}^{ν} a[[i + 1]] x^i) / (∏_{i=1}^{δ} (x - rt[[i, 1, 2]])) w[x], w[x], x] // Chop //
    Simplify;
  S_DC = Sol[[1, 1, 2]] / C[1];
  S_D[y_] := S_D[y] = S_DC /. {x → y};
  LB = Max[Min[rts] + 0.02, Min[X1]];
  UB = Min[Max[rts] - 0.02, Max[X1]];
  Off[NIntegrate::slwcon, NIntegrate::ncvb];
  Nc1 = NIntegrate[S_D[x], {x, LB, UB}];
  S1[x_] := S1[x] = 1/Nc1 S_D[x];
  g1[z_] := g1[z] = S1[(z - me1) / V1] / V1;
  h1[x_] := h1[x] = (Re[g1[x]] - f1[x])^2;
  ISD1 = NIntegrate[h1[x], {x, 0, 1}];

  KSDf[x_] := PDF[SmoothKernelDistribution[data], x];
```

147

```
  KSDh[x_] := KSDh[x] = (KSDf[x] - f1[x])^2;
  ISD2 = NIntegrate[KSDh[x], {x, 0, 1}];

  Return[{ν, δ, n, LB, UB, ISD1, ISD2}]]

ISDs = Table[ISD[1000, 𝒟1], {i, 100}];
MatrixForm[ISDs]
Mean[Table[ISDs[[i, 6]], {i, 100}]]
StandardDeviation[Table[ISDs[[i, 6]], {i, 100}]]
Mean[Table[ISDs[[i, 7]], {i, 100}]]
StandardDeviation[Table[ISDs[[i, 7]], {i, 100}]]

(*Normal distribution*)

ClearAll[G, K, a, c, Z1, PE, PA1, rts, Sol]
Unprotect[Power]; 0^0 = 1; Protect[Power];
𝒟1 = NormalDistribution[0, 1];
(*𝒟2=MultinormalDistribution[{.2,1.2},{{.4,.04},{.04,.4}}];
𝒟=MixtureDistribution[{1/2,1/2},{𝒟1,𝒟2}];*)
f1[x_] := f1[x] = PDF[𝒟1, {x}]
F1[x_] := F1[x] = CDF[𝒟1, {x}]
(*data=RandomVariate[𝒟1,3000];*)
Plot[PDF[𝒟1, x], {x, -4, 4}]


ISD[m_, 𝒟1_] :=
 Module[{data, Y1, X1, edis1, EmCDF1, me1, V1, LB, UB, Sol, g1, h1, ISD1, ISD2, ISD3},
  ClearAll[K0, K1, a1, a0, Z0, Z1, data, cz0, cz1, fA, LS, LS0, Y1, X1, edis1, EmCDF1,
   me1, V1, LB0, UB0, S0, f0, g1, h1, ISD1, ISD2, KSDf, KSDh, ISD3, KSDf3, KSDh3];
  f1[x_] := f1[x] = PDF[𝒟1, {x}];
  F1[x_] := F1[x] = CDF[𝒟1, {x}];
  data = Sort[RandomVariate[𝒟1, m]];
  Y1 = data;
  n = Length[Y1];
  edis1 = EmpiricalDistribution[Y1];
  EmCDF1 = Table[CDF[edis1, Y1[[j]]], {j, 1, n}];
  me1 = Sum[Y1[[j]], {j, 1, n}] / n;
  V1 = Sqrt[1/(n - 1) Sum[(Y1[[j]] - me1)^2, {j, n}]];

  X1 = Table[(Y1[[i]] - me1) / V1, {i, 1, n}];
  μX[h_] := μX[h] = Mean[X1^h];
  α = Min[X1]; β = Max[X1]; ν = 5; δ = 2;

  (*Begin initial estimation*)
  LB0 = -3.5; UB0 = 3.5;
  LS0[r_, ν_] := LS0[r, ν] = (Sum[K0[i] μX[r + i], {i, 0, ν}] == -r μX[r - 1]);
  Z0 = Solve[Table[LS0[r, ν], {r, 0, ν}], Table[K0[i], {i, 0, ν}]];
  a0[i_] := Z0[[1, i + 1, 2]];
```

```
  g0[x_] := g0[x] = Exp[∑_{i=0}^{ν} (a0[i] x^(i+1) / (i + 1))];
  cz0 = NIntegrate[g0[x], {x, LB0, UB0}];
  S0[x_] := S0[x] = g0[x] / cz0;
  f0[z_] := f0[z] = S0[(z - me1) / V1] / V1;
(*End initial estimation*)
  LS[r_, ν_] := LS[r, ν] = ( ∑_{i=0}^{ν} K1[i] μX[r + i] == -S0[α] α^r + S0[β] β^r - r μX[r - 1] );
  Z1 = Solve[Table[LS[r, ν], {r, 0, ν}], Table[K1[i], {i, 0, ν}]];
  a1 = Table[Z1[[1, i + 1, 2]], {i, 0, ν}];
  a1 = Table[Z1[[1, i + 1, 2]], {i, 0, ν}];
  g1[x_] := g1[x] = Exp[∑_{i=0}^{ν} (a1[[i + 1]] x^(i+1) / (i + 1))];
  cz1 = NIntegrate[g1[x], {x, LB0, UB0}];
  S1[x_] := S1[x] = g1[x] / cz1;
  fA[z_] := fA[z] = S1[(z - me1) / V1] / V1;
  h1[x_] := h1[x] = (Re[fA[x]] - f1[x])^2;

  ISD1 = NIntegrate[h1[x], {x, -4, 4}];
  KSDf[x_] := PDF[SmoothKernelDistribution[data], x];
  KSDh[x_] := KSDh[x] = (KSDf[x] - f1[x])^2;
  ISD2 = NIntegrate[KSDh[x], {x, -4, 4}];

  Return[{ν, δ, n, LB0, UB0, ISD1, ISD2}]]



ISDs = Table[ISD[1000, 𝒟1], {i, 100}];
MatrixForm[ISDs]
Mean[Table[ISDs[[i, 6]], {i, 100}]]
StandardDeviation[Table[ISDs[[i, 6]], {i, 100}]]
Mean[Table[ISDs[[i, 7]], {i, 100}]]
StandardDeviation[Table[ISDs[[i, 7]], {i, 100}]]
```

149

Example 3.4.2. Trapezoidal density approximant

```
ClearAll[G, K, a, c, Z1, PE, PA1, rts, Sol]
I_{u_,v_}[z_] := If[u ≤ z ≤ v, 1, 0]
(*The exact density*)
f_E[x_] := f_E[x] = I_{0,1}[x] x / 2 + 0.5 I_{1,2}[x] + I_{2,3}[x] (-x + 3) / 2
F_E[z_] := F_E[z] = Rationalize[∫_0^z f_E[x] dx, 10^-300];
μ_X[h_] := μ_X[h] = Rationalize[∫_0^3 x^h f_E[x] dx, 10^-300];
α = 0; β = 3; ν = 26; δ = 4;
PE =
  Plot[f_E[y], {y, 0, 3}, PlotRange → All, PlotLabel → {ν, δ}, PlotStyle → RGBColor[0, 0, 1]];
LS[r_, ν_, δ_] := LS[r, ν, δ] =
  (∑_{i=0}^{ν} K[i] μ_X[r + i] == (r + δ) μ_X[r + δ - 1] + ∑_{j=0}^{δ-1} (r + j) G[j] μ_X[r + j - 1])
Z1 = Solve[Table[LS[r, ν, δ], {r, 0, ν + δ}],
    Flatten[{Table[K[i], {i, 0, ν}], Table[G[j], {j, 0, δ - 1}]}]] // N
a = Table[Z1[[1, i + 1, 2]], {i, 0, ν}];
c = Table[Z1[[1, i + 1, 2]], {i, ν + 1, ν + δ}];
c = Append[c, 1];
rt = Solve[∑_{i=0}^{δ} c[[i + 1]] y^i == 0, y] // Chop;
rts = Table[rt[[i, 1, 2]], {i, 1, δ}];
Sol = DSolve[w'[x] == -(∑_{i=0}^{ν} a[[i + 1]] x^i) / (∏_{i=1}^{δ} (x - rt[[i, 1, 2]])) w[x], w[x], x] // Chop //
    Simplify;
S_DC = Sol[[1, 1, 2]] / C[1];
S_D[y_] := S_D[y] = S_DC /. {x → y}
LB = 0; UB = 3;
Off[NIntegrate::slwcon, NIntegrate::ncvb]
Nc1 = NIntegrate[S_D[x], {x, Min[rts], Max[rts]}]
f_A1[x_] := f_A1[x] = 1 / Nc1 S_D[x]
PA1 = Plot[Re[f_A1[y]], {y, Min[rts], Max[rts]},
    PlotRange → All, PlotLabel → {ν, δ}, PlotStyle → RGBColor[1, 0, 0]];
Show[
 PE,
 PA1]
```

Example 3.4.3. Mixture of Beta densities

```
ClearAll[G, K, a, c, Z1, PE, PA1, rts, Sol]
Unprotect[Power]; 0^0 = 1; Protect[Power];
𝒟1 = BetaDistribution[10, 2];
f1[x_] := f1[x] = PDF[𝒟1, {x}]
𝒟2 = BetaDistribution[4, 6];
f2[x_] := f2[x] = PDF[𝒟2, {x}]
𝒟 = MixtureDistribution[{1/2, 1/2}, {𝒟1, 𝒟2}];
PP1 = Plot[f1[y], {y, 0, 1}, PlotRange → All,
  PlotStyle → RGBColor[0, 0, 1], AxesOrigin → {0, 0}]
PP2 = Plot[f2[y], {y, 0, 1}, PlotRange → All,
  PlotStyle → RGBColor[0, 0, 1], AxesOrigin → {0, 0}]
I_u_,v_[z_] := If[u ≤ z ≤ v, 1, 0]
(*The exact density*)
f_E[x_] := f_E[x] = PDF[𝒟, {x}]
F_E[z_] := F_E[z] = Rationalize[∫_0^z f_E[x] dx, 10^-300];
μ_X[h_] := μ_X[h] = Rationalize[∫_0^1 x^h f_E[x] dx, 10^-300];
α = 0; β = 1; ν = 8; δ = 2 ;
PE = Plot[f_E[y], {y, 0, 1}, PlotRange → All,
  PlotStyle → RGBColor[0, 0, 1], AxesOrigin → {0, 0}]
PE2 = Plot[F_E[y], {y, 0, 1}, PlotRange → All,
  PlotStyle → RGBColor[0, 0, 1], AxesOrigin → {0, 0}]

LS[r_, ν_, δ_] :=
 LS[r, ν, δ] = (∑_{i=0}^{ν} K[i] μ_X[r + i] == (r + δ) μ_X[r + δ - 1] + ∑_{j=0}^{δ-1} (r + j) G[j] μ_X[r + j - 1])
Z1 = Solve[Table[LS[r, ν, δ], {r, 0, ν + δ}],
   Flatten[{Table[K[i], {i, 0, ν}], Table[G[j], {j, 0, δ - 1}]}]] // N
a = Table[Z1[[1, i + 1, 2]], {i, 0, ν}] // N
c = Table[Z1[[1, i + 1, 2]], {i, ν + 1, ν + δ}] // N
c = Append[c, 1]
rt = Solve[∑_{i=0}^{δ} c[[i + 1]] y^i == 0, y] // N // Chop

rts = Table[rt[[i, 1, 2]], {i, 1, δ}]
```

```
p1[y_] := p1[y] = Sum[a[[i + 1]] y^i, {i, 0, ν}]
p2[j_] := p2[j] = Product[If[j ≠ i, (rts[[j]] - rts[[i]]), 1], {i, 1, δ}]
theta[j_] := theta[j] = - p1[rts[[j]]] / p2[j]
{quot, rem} = PolynomialQuotientRemainder[Sum[a[[i + 1]] y^i, {i, 0, ν}], Sum[c[[i + 1]] y^i, {i, 0, δ}], y];
CoQuot = CoefficientList[quot, y];
CoRem = CoefficientList[rem, y];
{quot, rem} // N
n2 = Length[CoQuot];
f_paper[x_] := f_paper[x] = Exp[- Sum[CoQuot[[i]] x^i / i, {i, 1, n2}]] Product[Abs[x - rts[[j]]]^theta[j], {j, 1, δ}]
Nc2 = NIntegrate[f_paper[x], {x, 0, 1}];
LB = 0; UB = 1;
f_pap[x_] := f_pap[x] = 1/Nc2 f_paper[x]
PA2 = Plot[f_pap[y], {y, LB, UB}, PlotRange → All,
  PlotStyle → {Thickness[0.008], Dashed, RGBColor[0.75, 0, 0.25]}]
Show[PE, PA2]
F_pap[x_] := F_pap[x] = NIntegrate[f_pap[y], {y, 0, x}]
PA3 = Plot[F_pap[y], {y, LB, UB}, PlotRange → All,
  PlotStyle → {Thickness[0.008], Dashed, RGBColor[0.75, 0, 0.25]}]
Show[
 PE2,
 PA3]
```

A .3 Modules Used in Chapter 4

Example 4.2.1. Univariate mixture of Beta densities

```
α = 0; β = 1; ν = 4 ; δ = 4 ;
f_E[x_] := f_E[x] = (PDF[BetaDistribution[2, 20], x] + PDF[BetaDistribution[3, 2], x]) / 2
F_E[z_] := F_E[z] = Rationalize[Integrate[f_E[x] dx, {x, 0, z}], 10^-300];
μ_X[h_] := μ_X[h] = Rationalize[Integrate[x^h f_E[x] dx, {x, 0, 1}], 10^-300];
PE = Plot[f_E[y], {y, α, β}, PlotRange → All,
  PlotLabel → {ν, δ}, PlotStyle → RGBColor[0, 0, 1]]
```

```
G[δ] = 1;
G[δ - 1] = G1 - (α + β);
G[δ - 2] = G2 + α β - G1 (α + β);
G[δ - 3] = G1 α β - G2 (α + β);
G[δ - 4] = G2 α β;
LS[r_, ν_, δ_] :=
```

$$LS[r, \nu, \delta] = \left(\sum_{i=0}^{\nu} K[i]\, \mu_X[r+i] == (r+\delta)\, \mu_X[r+\delta-1] + \sum_{j=0}^{\delta-1} (r+j)\, G[j]\, \mu_X[r+j-1]\right)$$

```
Z1 = Solve[Table[LS[r, ν, δ], {r, 0, ν + δ - 2}],
    Flatten[{Table[K[i], {i, 0, ν}], Table[{G1, G2}]}]];
a_i_ := a_i = Z1[[1, i + 1, 2]]
G1 = Z1[[1, ν + 1 + 1, 2]];
G2 = Z1[[1, ν + 1 + 2, 2]];
```

$$rt = \text{Solve}\left[\sum_{i=0}^{\delta} c_i\, y^i == 0,\ y\right] // N // \text{Chop};$$

```
rts = Flatten[Table[If[Im[rt[[i, 1, 2]]] < 10^-12, Re[rt[[i, 1, 2]]], {}], {i, 1, δ}]] // N;

tn =
  Sort[Flatten[Table[If[α - (β - α) / 15 < rts[[j]] < α + (β - α) / 15 || β - (β - α) / 15 < rts[[
          j]] < β + (β - α) / 15, rts[[j]], { }], {j, Length[rts]}]]] // N;
```

$$Sol = \text{DSolve}\left[w'[x] == -\left(\sum_{i=0}^{\nu} a_i\, x^i\right) \Big/ \left(\prod_{i=1}^{\delta} (x - rt[[i, 1, 2]])\right) w[x],\ w[x],\ x\right] // \text{Chop} //$$

```
    Simplify ;
S_DC = Sol[[1, 1, 2]] / C[1];
S_D[y_] := S_D[y] = S_DC /. {x → y}
LB = 0; UB = 1;
Nc1 = NIntegrate[S_D[x], {x, LB, UB}];
```

$$f_{A1}[x\_] := f_{A1}[x] = \frac{1}{Nc1}\, S_D[x]$$

```
PA1 = Plot[f1_A[y], {y, LB, UB}, PlotRange → All,
  PlotLabel → {ν, δ}, PlotStyle → RGBColor[1, 0, 0]]
Show[
 PE,
 PA1]
```

Example 4.3.1. Mixture of bivariate normal densities

```
Unprotect[Power]; 0^0 = 1; Protect[Power];
Needs["MultivariateStatistics`"]
𝒟1 = MultinormalDistribution[{-1.1, -.1}, {{.33, .03}, {0.03, .33}}];
𝒟2 = MultinormalDistribution[{.2, 1.2}, {{.4, .04}, {.04, .4}}];
𝒟3 = MixtureDistribution[{1/2, 1/2}, {𝒟1, 𝒟2}];
fe[x_, y_] := fe[x, y] = PDF[𝒟3, {x, y}]
Fe[x_, y_] := Fe[x, y] = CDF[𝒟3, {x, y}]

me1 = ∑_{j=1}^{n} Y1[[j]] / n;

me2 = ∑_{j=1}^{n} Y2[[j]] / n;

Var = Covariance[data];
Vhi = MatrixPower[Var, -1/2];
Vh = MatrixPower[Var, 1/2];
detVhi = Det[Vhi];

X1 = Vhi[[1, 1]] (Y1 - ConstantArray[me1, n]) + Vhi[[1, 2]] (Y2 - ConstantArray[me2, n]);
X2 = Vhi[[2, 1]] (Y1 - ConstantArray[me1, n]) + Vhi[[2, 2]] (Y2 - ConstantArray[me2, n]);

H1 = Histogram[X1, 25];
H2 = Histogram[X2, 25];
μX[h_] := Mean[X1^h]
μY[h_] := Mean[X2^h]
```

Marginal density for X1 :

```
ν1 = 5; LB = Min[X1]; UB = Max[X1];
LS[r_, ν1_] := LS[r, ν1] = (∑_{i=0}^{ν1} K[i] μX[r + i] == -r μX[r - 1])
Z1 = Solve[Table[LS[r, ν1], {r, 0, ν1}], Table[K[i], {i, 0, ν1}]];
a1[i_] := Z1[[1, i + 1, 2]];
g1[x_] := g1[x] = Exp[∑_{i=0}^{ν1} (a1[i] x^(i+1) / (i + 1))]
cz = NIntegrate[g1[x], {x, LB, UB}];
S1[x_] := S1[x] = g1[x] / cz
P1 = Plot[S1[y], {y, LB, UB}, PlotRange → All, PlotLabel → ν1]
```

Marginal density for X2 :

```
ν2 = 3; LB2 = Min[X2]; UB2 = Max[X2];
LS2[r_, ν2_] := LS2[r, ν2] = (Sum[K2[i] μY[r + i], {i, 0, ν2}] == -r μX[r - 1])
Z2 = Solve[Table[LS2[r, ν2], {r, 0, ν2}], Table[K2[i], {i, 0, ν2}]];

a2[i_] := Z2[[1, i + 1, 2]];
g2[y_] := g2[y] = Exp[Sum[(a2[i] y^(i+1) / (i + 1)), {i, 0, ν2}]]
cz2 = NIntegrate[g2[y], {y, LB2, UB2}];
S2[y_] := S2[y] = g2[y] / cz2
P2 = Plot[S2[y], {y, LB2, UB2}, PlotRange → All]

ListPlot[data, AxesOrigin → {0, 0}]
Histogram3D[data]
```

Initial joint density function of (X1, X2) :

```
g[x_, y_] := g[x, y] = S1[x] S2[y]
Plot3D[g[x, y], {x, LB, UB}, {y, LB2, UB2}, PlotRange → All, PlotLabel → G[ν1, ν2]]

ft[x_, y_] := detVhi
  g[Vhi[[1, 1]] (x - me1) + Vhi[[1, 2]] (y - me2), Vhi[[2, 1]] (x - me1) + Vhi[[2, 2]] (y - me2)]

Plot3D[ft[x, y], {x, Min[Y1], Max[Y1]}, {y, Min[Y2], Max[Y2]}, PlotRange → All]
```

Kernel Smooth density function of Y1 and Y2 :

```
𝒟 = SmoothKernelDistribution[data];
Plot3D[PDF[𝒟, {x, y}], {x, Min[Y1], Max[Y1]},
 {y, Min[Y2], Max[Y2]}, PlotLabel → SmoothKernel XY, PlotRange → All]
Histogram3D[data]
```

Polynomial adjustment with degree “MSTerms”:

```
Off[NIntegrate::izero, NIntegrate::ncvb, NIntegrate::slwcon]
m1[i_, j_] := m1[i, j] = NIntegrate[x^i y^j g[x, y], {x, LB, UB}, {y, LB2, UB2}]
MSTerms = 7;
f3[L1_List, L2_List] := Inner[Plus, L1, L2, List];
L3 = Flatten[Table[{j, i}, {i, 0, MSTerms}, {j, 0, MSTerms}], 1];
P3 = Table[f3[L3[[i]], L3[[j]]], {i, 1, Length[L3]}, {j, 1, Length[L3]}];
M3 = Rationalize[
   Table[m1[P3[[i, j]][[1]], P3[[i, j]][[2]]], {i, Length[L3]}, {j, Length[L3]}], 10^-25];
Zv[x_, y_] := Zv[x, y] = Flatten[Table[x^j y^i, {i, 0, MSTerms}, {j, 0, MSTerms}], 1];
Gms[i_] := Gms[i] = Sum[X1[[j]]^L3[[i,1]] X2[[j]]^L3[[i,2]] / n, {j, 1, n}];
μ = Table[Gms[i], {i, Dimensions[L3][[1]]}] // N;
c3 = LinearSolve[M3, μ];
```

155

```
t4[x_, y_] := t4[x, y] = g[x, y] c3.Zv[x, y];
t5[x_, y_] := t5[x, y] = If[t4[x, y] < 0 || x > UB || x < LB || y > UB2 || y < LB2, 0, t4[x, y]]
cii = NIntegrate[t4[x, y], {x, LB, UB}, {y, LB2, UB2}] // N
t6[t1_, t2_] := t6[t1, t2] = t4[t1, t2] / cii
f[u_, w_] := f[u, w]  = detVhi t4[Vhi[[1, 1]] (u - me1) + Vhi[[1, 2]] (w - me2),
      Vhi[[2, 1]] (u - me1) + Vhi[[2, 2]] (w - me2)] / cii


ISDest = NIntegrate[(f[u, w] - fe[u, w]) ^2 , {u, Min[Y1], Max[Y1]}, {w, Min[Y2], Max[Y2]}]

F[u_, w_] := F[u, w]  = NIntegrate[f[x1, y1], {x1, Min[Y1], u}, {y1, Min[Y2], w}]
EstCDF = Table[F[Y1[[j]], Y2[[j]]], {j, 1, n}];

MSTerms
error = Sum[(EmpCDF[[j]] -  Max[0, EstCDF[[j]] ])^2, {j, 1, n}]
Plot3D[f[u, w] , {u, Min[Y1], Max[Y1]},
 {w, Min[Y2], Max[Y2]}, PlotRange → All, PlotLabel → MSTerms]
```

Example 4.4.1. CommViolPredUnnormalizedData

```
ClearAll[X1, X2, Y1, Y2, data, n]
Needs["Histograms`"];
Unprotect[Power]; 0^0 = 1; Protect[Power];
mydata = Import["CommViolPredUnnormalizedData.csv", "CSV"];

header = mydata[[1]];
data = mydata[[2 ;;]];
myDataset = Thread[header → #] & /@ data // Map[Association] // Dataset;
data = mydata[[2 ;;]];
Y1 = Table[data[[i, 13]], {i, Length[data]}];
Y2 = Table[data[[i, 22]], {i, Length[data]}];

data = Table[{Y1[[i]], Y2[[i]]}, {i, Length[data]}];
n = Length[Y1]

edis = EmpiricalDistribution[data];
EmpCDF = Table[CDF[edis, data[[j]]], {j, 1, n}];
me1 = ∑_{j=1}^{n} Y1[[j]] / n;
me2 = ∑_{j=1}^{n} Y2[[j]] / n;
Var = Covariance[data];
Vhi = MatrixPower[Var, -1 / 2];
Vh = MatrixPower[Var, 1 / 2];
detVhi = Det[Vhi];

X1 = Vhi[[1, 1]] (Y1 - ConstantArray[me1, n]) + Vhi[[1, 2]] (Y2 - ConstantArray[me2, n]);
X2 = Vhi[[2, 1]] (Y1 - ConstantArray[me1, n]) + Vhi[[2, 2]] (Y2 - ConstantArray[me2, n]);

Hist[data_, n_] := Module[{ran},
  ran = Max[data] - Min[data];
Histogram[data, HistogramScale → 1,
   HistogramRange → {Min[data] - 0.4 * ran, Max[data] + 0.4 * ran},
   HistogramCategories → Range[Min[data], Max[data] + 0.2 ran, ran / n]]
 ]
H1 = Hist[X1, 25];
H2 = Hist[X2, 25];
μX[h_] := Mean[X1^h]
μY[h_] := Mean[X2^h]
```

Marginal density for X1 :

```
ν1 = 4; LB = Min[X1]; UB = Max[X1];
LS[r_, ν1_] := LS[r, ν1] = (Sum[K[i] μX[r + i], {i, 0, ν1}] == -r μX[r - 1])
Z1 = Solve[Table[LS[r, ν1], {r, 0, ν1}], Table[K[i], {i, 0, ν1}]];
a1[i_] := Z1[[1, i + 1, 2]];
g1[x_] := g1[x] = Exp[Sum[a1[i] x^(i+1) / (i + 1), {i, 0, ν1}]]
cz = NIntegrate[g1[x], {x, LB, UB}];
S1[x_] := S1[x] = g1[x] / cz
f1[z_] := f1[z] = S1[(z - me1) / V1] / V1
P1 = Plot[S1[y], {y, LB, UB}, PlotRange → All, PlotLabel → ν1];
ν1
Show[H1, P1]
```

Marginal density for X2 :

```
ν2 = 3; LB2 = Min[X2]; UB2 = Max[X2];
LS2[r_, ν2_] := LS2[r, ν2] = (Sum[K2[i] μY[r + i], {i, 0, ν2}] == -r μX[r - 1])
Z2 = Solve[Table[LS2[r, ν2], {r, 0, ν2}], Table[K2[i], {i, 0, ν2}]];

a2[i_] := Z2[[1, i + 1, 2]];
g2[y_] := g2[y] = Exp[Sum[a2[i] y^(i+1) / (i + 1), {i, 0, ν2}]]
cz2 = NIntegrate[g2[y], {y, LB2, UB2}];
S2[y_] := S2[y] = g2[y] / cz2
f2[z_] := f2[z] = S2[(z - me2) / V2] / V2
P2 = Plot[S2[y], {y, LB2, UB2}, PlotRange → All];
ν2
Show[H2, P2]

ListPlot[data, AxesOrigin → {0, 0}]
Histogram3D[data]
```

Initial joint density function of (X1, X2) :

```
g[x_, y_] := g[x, y] = S1[x] S2[y]

Plot3D[g[x, y], {x, LB, UB}, {y, LB2, UB2}, PlotRange → All, PlotLabel → G[ν1, ν2]]
```

Kernel Smooth density function of Y1 and Y2 :

```
𝒟 = SmoothKernelDistribution[data];
Plot3D[PDF[𝒟, {x, y}], {x, Min[Y1], Max[Y1]},
 {y, Min[Y2], Max[Y2]}, PlotLabel → SmoothKernel XY, PlotRange → All]
Histogram3D[data]
ListPlot[data, PlotLabel → ListPlot XY]
```

Polynomial adjustment with degree “MSTerms”:

```
m1[i_, j_] := m1[i, j] = NIntegrate[x^i y^j g[x, y], {x, LB, UB}, {y, LB2, UB2}]
MSTerms = 7;
f3[L1_List, L2_List] := Inner[Plus, L1, L2, List];
L3 = Flatten[Table[{j, i}, {i, 0, MSTerms}, {j, 0, MSTerms}], 1];
P3 = Table[f3[L3[[i]], L3[[j]]], {i, 1, Length[L3]}, {j, 1, Length[L3]}];
M3 = Rationalize[
   Table[m1[P3[[i, j]][[1]], P3[[i, j]][[2]]], {i, Length[L3]}, {j, Length[L3]}], 10^-25];
Zv[x_, y_] := Zv[x, y] = Flatten[Table[x^j y^i, {i, 0, MSTerms}, {j, 0, MSTerms}], 1];
Gms[i_] := Gms[i] = Sum[X1[[j]]^L3[[i,1]] X2[[j]]^L3[[i,2]] / n, {j, 1, n}];
μ = Table[Gms[i], {i, Dimensions[L3][[1]]}] // N;
c3 = LinearSolve[M3, μ];

t4[x_, y_] := t4[x, y] = g[x, y] c3.Zv[x, y];
t5[x_, y_] := t5[x, y] = If[t4[x, y] < 0 || x > UB || x < LB || y > UB2 || y < LB2, 0, t4[x, y]]
cii = NIntegrate[t4[x, y], {x, LB, UB}, {y, LB2, UB2}] // N
t6[t1_, t2_] := t6[t1, t2] = t4[t1, t2] / cii
f[u_, w_] := f[u, w]  = detVhi t4[Vhi[[1, 1]] (u - me1) + Vhi[[1, 2]] (w - me2),
      Vhi[[2, 1]] (u - me1) + Vhi[[2, 2]] (w - me2)] / cii
```

Estimated density for p=7:

```
Plot3D[f[u, w] , {u, Min[Y1], Max[Y1]}, {w, Min[Y2], Max[Y2]}, PlotRange → All]

Off[NIntegrate::izero]
Off[NIntegrate::"ncvb"]
Off[NIntegrate::"slwcon"]
F[u_, w_] := F[u, w]  = NIntegrate[f[x1, y1], {x1, Min[Y1], u}, {y1, Min[Y2], w}]
EstCDF = Table[F[Y1[[j]], Y2[[j]]], {j, 1, n}];

MSTerms
error = Sum_{j=1}^{n} (EmpCDF[[j]] - Max[0, EstCDF[[j]] ])^2
Plot3D[f[u, w] , {u, Min[Y1], Max[Y1]},
 {w, Min[Y2], Max[Y2]}, PlotRange → All, PlotLabel → MSTerms]
```

159

Example 4.4.2. Concrete dataset

```
ClearAll[X1, X2, Y1, Y2, data, n]
Needs["Histograms`"];
Needs["MultivariateStatistics`"]
mydata = Import["Concrete_Data.csv", "CSV"];
header = mydata[[1]];
data = mydata[[2 ;;]];
Unprotect[Power]; 0^0 = 1; Protect[Power];
myDataset = Thread[header → #] & /@ data // Map[Association] // Dataset;
data = mydata[[2 ;;]];
Y1 = Table[data[[i, 1]], {i, Length[data]}];
Y2 = Table[data[[i, 9]], {i, Length[data]}];
n = Length[Y1]
data = Table[{Y1[[j]], Y2[[j]]}, {j, 1, n}];
edis = EmpiricalDistribution[data];
EmpCDF = Table[CDF[edis, data[[j]]], {j, 1, n}];
me1 = ∑_{j=1}^{n} Y1[[j]] / n;
me2 = ∑_{j=1}^{n} Y2[[j]] / n;
Var = Covariance[data];
Vhi = MatrixPower[Var, -1 / 2];
Vh = MatrixPower[Var, 1 / 2];
detVhi = Det[Vhi];

X1 = Vhi[[1, 1]] (Y1 - ConstantArray[me1, n]) + Vhi[[1, 2]] (Y2 - ConstantArray[me2, n]);
X2 = Vhi[[2, 1]] (Y1 - ConstantArray[me1, n]) + Vhi[[2, 2]] (Y2 - ConstantArray[me2, n]);

Hist[data_, n_] := Module[{ran},
  ran = Max[data] - Min[data];
Histogram[data, HistogramScale → 1,
   HistogramRange → {Min[data] - 0.4 * ran, Max[data] + 0.4 * ran},
   HistogramCategories → Range[Min[data], Max[data] + 0.2 ran, ran / n]]
 ]
H1 = Hist[X1, 25];
H2 = Hist[X2, 25];
μX[h_] := Mean[X1^h]
μY[h_] := Mean[X2^h]
```

Marginal density for X1 :

```
ν1 = 4; LB = -3; UB = 3.1;
LS[r_, ν1_] := LS[r, ν1] = (Sum[K[i] μX[r + i] == -r μX[r - 1], {i, 0, ν1}])
Z1 = Solve[Table[LS[r, ν1], {r, 0, ν1}], Table[K[i], {i, 0, ν1}]];
a1[i_] := Z1[[1, i + 1, 2]];
g1[x_] := g1[x] = Exp[Sum[(a1[i] x^(i+1) / (i + 1)), {i, 0, ν1}]]
cz = NIntegrate[g1[x], {x, LB, UB}];
S1[x_] := S1[x] = g1[x] / cz
f1[z_] := f1[z] = S1[(z - me1) / V1] / V1
P1 = Plot[S1[y], {y, LB, UB}, PlotRange → All, PlotLabel → ν1];
ν1
Show[H1, P1]
```

Marginal density for X2 :

```
ν2 = 3; LB2 = -3; UB2 = 3.3;
LS2[r_, ν2_] := LS2[r, ν2] = (Sum[K2[i] μY[r + i] == -r μX[r - 1], {i, 0, ν2}])
Z2 = Solve[Table[LS2[r, ν2], {r, 0, ν2}], Table[K2[i], {i, 0, ν2}]];

a2[i_] := Z2[[1, i + 1, 2]];
g2[y_] := g2[y] = Exp[Sum[(a2[i] y^(i+1) / (i + 1)), {i, 0, ν2}]]
cz2 = NIntegrate[g2[y], {y, LB2, UB2}];
S2[y_] := S2[y] = g2[y] / cz2
f2[z_] := f2[z] = S2[(z - me2) / V2] / V2
P2 = Plot[S2[y], {y, LB2, UB2}, PlotRange → All];
ν2
Show[H2, P2]
```

Initial joint density function of (X1, X2) :

```
g[x_, y_] := g[x, y] = S1[x] S2[y]
```

Polynomial adjustment with degree “MSTerms”:

```
m1[i_, j_] := m1[i, j] = NIntegrate[x^i y^j g[x, y], {x, LB, UB}, {y, LB2, UB2}]
MSTerms = 6;
f3[L1_List, L2_List] := Inner[Plus, L1, L2, List];
L3 = Flatten[Table[{j, i}, {i, 0, MSTerms}, {j, 0, MSTerms}], 1];
P3 = Table[f3[L3[[i]], L3[[j]]], {i, 1, Length[L3]}, {j, 1, Length[L3]}];
M3 = Rationalize[
   Table[m1[P3[[i, j]][[1]], P3[[i, j]][[2]]], {i, Length[L3]}, {j, Length[L3]}], 10^-25];
Zv[x_, y_] := Zv[x, y] = Flatten[Table[x^j y^i, {i, 0, MSTerms}, {j, 0, MSTerms}], 1];
Gms[i_] := Gms[i] = Sum[X1[[j]]^L3[[i,1]] X2[[j]]^L3[[i,2]] / n, {j, 1, n}];
μ = Rationalize[Table[Gms[i], {i, Dimensions[L3][[1]]}], 10^-25] // N;
c3 = LinearSolve[M3, μ];
t4[x_, y_] := t4[x, y] = g[x, y] c3.Zv[x, y];
t5[x_, y_] := t5[x, y] = If[t4[x, y] < 0 || x > UB || x < LB || y > UB2 || y < LB2, 0, t4[x, y]]
cii = NIntegrate[t4[x, y], {x, LB, UB}, {y, LB2, UB2}] // N
t6[t1_, t2_] := t6[t1, t2] = t4[t1, t2] / cii
f[u_, w_] := f[u, w]  = detVhi t4[Vhi[[1, 1]] (u - me1) + Vhi[[1, 2]] (w - me2),
     Vhi[[2, 1]] (u - me1) + Vhi[[2, 2]] (w - me2)] / cii

Off[NIntegrate::izero]
Off[NIntegrate::"ncvb"]
F[u_, w_] := F[u, w]  = NIntegrate[f[x1, y1], {x1, 50, u}, {y1, 0, w}]
EstCDF = Table[F[Y1[[j]], Y2[[j]]], {j, 1, n}];
MSTerms
error = Sum[(EmpCDF[[j]] - Max[0, EstCDF[[j]] ])^2, {j, 1, n}]
Plot3D[f[u, w] , {u, 50, 570}, {w, 0, 95}, PlotRange → All]
```

Example 4.4.3. Covertype dataset

```
ClearAll[X1, X2, Y1, Y2, data, n]
mydata = Import["covtype.csv", "CSV"];
Unprotect[Power]; 0^0 = 1; Protect[Power];
header = mydata[[1]];
data = mydata[[2 ;;]];
myDataset = Thread[header → #] & /@ data // Map[Association] // Dataset;
data = mydata[[2 ;;]];
Y1 = data[[All, 6]];
Y2 = data[[All, 10]];
n = Length[Y1]
data  = Table[{Y1[[j]], Y2[[j]]}, {j, 1, n}];
```

```
me1 = ∑_{j=1}^{n} Y1[[j]] / n;

me2 = ∑_{j=1}^{n} Y2[[j]] / n;

Var = Covariance[data];
Vhi = MatrixPower[Var, -1/2];
Vh = MatrixPower[Var, 1/2];
detVhi = Det[Vhi];
X1 = Vhi[[1, 1]] (Y1 - ConstantArray[me1, n]) + Vhi[[1, 2]] (Y2 - ConstantArray[me2, n]);
X2 = Vhi[[2, 1]] (Y1 - ConstantArray[me1, n]) + Vhi[[2, 2]] (Y2 - ConstantArray[me2, n]);

H1 = Histogram[X1, Automatic, "Probability"]
H2 = Histogram[X2, Automatic, "Probability"]
μX[h_] := Sum[X1[[j]]^h, {j, n}] / n
μY[h_] := Sum[X2[[j]]^h, {j, n}] / n
μXY[r_, h_] := Sum[X1[[j]]^r X2[[j]]^h, {j, n}] / n

ListPlot[data, PlotRange → All, AxesOrigin → {0, 0}]
Histogram3D[data, {40, 30}]
𝒟 = SmoothKernelDistribution[data];
Plot3D[PDF[𝒟, {x, y}], {x, Min[Y1], Max[Y1]}, {y, Min[Y2], Max[Y2]}, PlotRange → All]

mxy = Table[μXY[r, h], {r, 0, 14}, {h, 0, 14}]

ν1 = 5; LB = -2.5; UB = 3.5;
LS[r_] := LS[r] = ( ∑_{i=0}^{ν1} K[i] mxy[[r + i + 1, 1]] == -r mxy[[r, 1]] )
Z1 = Solve[Table[LS[r], {r, 0, ν1}], Table[K[i], {i, 0, ν1}]];
a1[i_] := Z1[[1, i + 1, 2]];
g1[x_] := g1[x] = Exp[ ∑_{i=0}^{ν1} (a1[i] x^{i+1} / (i + 1)) ]
cz = NIntegrate[g1[x], {x, LB, UB}];
S1[x_] := S1[x] = g1[x] / cz

ν2 = 6; LB2 = -2; UB2 = 4.1;
LS2[r_, ν2_] := LS2[r, ν2] = ( ∑_{i=0}^{ν2} K2[i] mxy[[1, r + i + 1]] == -r mxy[[1, r]] )
Z2 = Solve[Table[LS2[r, ν2], {r, 0, ν2}], Table[K2[i], {i, 0, ν2}]];

a2[i_] := Z2[[1, i + 1, 2]];
g2[y_] := g2[y] = Exp[ ∑_{i=0}^{ν2} (a2[i] y^{i+1} / (i + 1)) ]
cz2 = NIntegrate[g2[y], {y, LB2, UB2}];
S2[y_] := S2[y] = g2[y] / cz2
g[x_, y_] := g[x, y] = S1[x] S2[y]

ft[x_, y_] := detVhi
  g[Vhi[[1, 1]] (x - me1) + Vhi[[1, 2]] (y - me2), Vhi[[2, 1]] (x - me1) + Vhi[[2, 2]] (y - me2)]
```

163

```
Plot3D[ft[x, y], {x, 0, Max[Y1]}, {y, 0, Max[Y2]}, PlotRange → All]
```

Polynomial adjustment with degree “MSTerms”:

```
m1[i_, j_] := m1[i, j] = NIntegrate[x^i y^j g[x, y], {x, LB, UB}, {y, LB2, UB2}]
MSTerms = 7;
f3[L1_List, L2_List] := Inner[Plus, L1, L2, List];
L3 = Flatten[Table[{j, i}, {i, 0, MSTerms}, {j, 0, MSTerms}], 1];
P3 = Table[f3[L3[[i]], L3[[j]]], {i, 1, Length[L3]}, {j, 1, Length[L3]}];
M3 = Rationalize[
   Table[m1[P3[[i, j]][[1]], P3[[i, j]][[2]]], {i, Length[L3]}, {j, Length[L3]}], 10^-25];
Zv[x_, y_] := Zv[x, y] = Flatten[Table[x^j y^i, {i, 0, MSTerms}, {j, 0, MSTerms}], 1];
(*Gms[i_]:=Gms[i]=Sum[X1[[j]]^L3[[i,1]]   X2[[j]]^L3[[i,2]] /n,{j,1,n}];*)
Gms[i_] := Gms[i] = mxy[[L3[[i, 1]] + 1, L3[[i, 2]] + 1]]
μ = Rationalize[Table[Gms[i], {i, Dimensions[L3][[1]]}], 10^-25] // N;
c3 = LinearSolve[M3, μ];
t4[x_, y_] := t4[x, y] = g[x, y] c3.Zv[x, y];
t5[x_, y_] := t5[x, y] = If[t4[x, y] < 0 || x > UB || x < LB || y > UB2 || y < LB2, 0, t4[x, y]]
cii = NIntegrate[t4[x, y], {x, LB, UB}, {y, LB2, UB2}] // N;
t6[t1_, t2_] := t6[t1, t2] = t4[t1, t2] / cii
f[u_, w_] := f[u, w]  = detVhi t4[Vhi[[1, 1]] (u - me1) + Vhi[[1, 2]] (w - me2),
     Vhi[[2, 1]] (u - me1) + Vhi[[2, 2]] (w - me2)] / cii
MSTerms
Plot3D[f[x, y] , {x, 0, Max[Y1]}, {y, 0, Max[Y2]}, PlotRange → All]
```

Example 4.4.4. Flood dataset

```
ClearAll[X1, X2, Y1, Y2, data, n]
(*Needs["Histograms`"];*)
Unprotect[Power]; 0^0 = 1; Protect[Power];
data = {{121, 3306}, {137, 8327}, {143, 7235}, {146, 10818}, {151, 5057}, {156, 6620},
    {157, 5002}, {162, 5236}, {168, 3826}, {173, 4892}, {176, 6460}, {181, 7502},
    {182, 7684}, {183, 4780}, {183, 7748}, {184, 5167}, {186, 8026}, {187, 7350},
    {189, 4189}, {196, 6728}, {197, 9645}, {200, 9177}, {202, 9581}, {206, 8788},
    {208, 10853}, {210, 6334}, {214, 6414}, {216, 9506}, {219, 14890}, {229, 8637},
    {230, 10828}, {232, 8177}, {232, 14769}, {233, 5650}, {233, 8923}, {236, 12577},
    {239, 6865}, {240, 8409}, {245, 6907}, {246, 8640}, {248, 6989}, {255, 8041},
    {257, 10174}, {260, 8949}, {260, 11127}, {261, 9957}, {275, 10128}, {279, 9763},
    {279, 10659}, {283, 7241}, {286, 8711}, {286, 12035}, {289, 7133}, {289, 10299},
    {292, 8692}, {292, 12057}, {294, 8918}, {297, 9352}, {300, 9406}, {303, 8900},
    {306, 12740}, {309, 12882}, {310, 9266}, {311, 13593}, {331, 13602}, {334, 11437},
    {343, 8192}, {351, 11401}, {371, 8704}, {371, 12825}, {383, 14559}, {390, 13543},
    {405, 11174}, {405, 15003}, {416, 11272}, {424, 13315}, {442, 13608}};
Y1 = Table[data[[i, 1]], {i, Length[data]}];
Y2 = Table[data[[i, 2]], {i, Length[data]}];
n = Length[Y1]
Y = Table[{Y1[[i]], Y2[[i]]}, {i, n}];
me1 = Sum[Y1[[j]], {j, 1, n}] / n;
me2 = Sum[Y2[[j]], {j, 1, n}] / n;
Var =
  1/(n - 1) {{Sum[(Y1[[j]] - me1)^2, {j, n}], Sum[(Y1[[j]] - me1) (Y2[[j]] - me2), {j, n}]},
    {Sum[(Y1[[j]] - me1) (Y2[[j]] - me2), {j, n}], Sum[(Y2[[j]] - me2)^2, {j, n}]}};
Var // N;
Vhi = MatrixPower[Var, -1/2];
Vh = MatrixPower[Var, 1/2];
detVhi = Det[Vhi] // N;
V1 = Var[[1, 1]];
V2 = Var[[2, 2]];
X1 = Vhi[[1, 1]] (Y1 - ConstantArray[me1, n]) + Vhi[[1, 2]] (Y2 - ConstantArray[me2, n]);
X2 = Vhi[[2, 1]] (Y1 - ConstantArray[me1, n]) + Vhi[[2, 2]] (Y2 - ConstantArray[me2, n]);

edis = EmpiricalDistribution[data];
EmpCDF = Table[CDF[edis, data[[j]]], {j, 1, n}];

μX[h_] := Mean[X1^h]
μY[h_] := Mean[X2^h]

H1 = Histogram[X1, Automatic, hspec = "Probability"]
H2 = Histogram[X2, Automatic, hspec = "Probability"]
```

Marginal density for X1 :

```
ν1 = 5;
δ = 2;
LB = -3.5;
UB = 3.5;

Dn[x_] := Expand[(x - LB) (x - UB)]
c = CoefficientList[Dn[x], x]
LS[r_, ν1_, δ_] := LS[r, ν1, δ] = (Sum[K[i] μX[r + i], {i, 0, ν1}] == Sum[(r + j) c[[j + 1]] μX[r + j - 1], {j, 0, δ}])
Z1 = Solve[Table[LS[r, ν1, δ], {r, 0, ν1}], Table[K[i], {i, 0, ν1}]];

K_i_ := K_i = Z1[[1, i + 1, 2]]
C_i_ := C_i = c[[i + 1]]
Z1 // N
Table[K_i, {i, 0, ν1}] // N
Table[C_i, {i, 0, δ}] // N
Sum[C_i y^i, {i, 0, δ}] // N

mi1 = FindMinimum[{Sum[K_i x^(i+1) / (i + 1), {i, 0, ν1}], LB ≤ x ≤ -2}, {x, -2.5}];

a = mi1[[2, 1, 2]];

ma1 = FindMinimum[{Sum[K_i x^(i+1) / (i + 1), {i, 0, ν1}], 2 ≤ x ≤ UB}, {x, 2.5}];

b = ma1[[2, 1, 2]];

Sol = DSolve[p'[x] == -((Sum[K_i x^i, {i, 0, ν1}]) / ((x - a) (x - b))) p[x], p[x], x] // Chop // Simplify ;

S_DC = Sol[[1, 1, 2]] / C[1];
Sol1[z_] := S_DC /. {x → z}
cn1 = NIntegrate[Sol1[x], {x, a, b}];
g1a[x_] := Sol1[x] / cn1
I_u_,v_[z_] := If[u < z < v, 1, 0]
S1[x_] := S1[x] = I_a, b[x] g1a[x]
P1 = Plot[S1[y], {y, a, b}, PlotRange → All, PlotLabel → ν1];
Show[H1, P1]
```

Marginal density for X2 :

```
ν2 = 5; δ2 = 2; LB = -3.5; UB = 3.5;
Dn2[x_] := Expand[(x - LB) (x - UB)]
c2 = CoefficientList[Dn2[x], x]
LS2[r_, ν2_, δ_] := LS2[r, ν2, δ] = (Sum[K2[i] μY[r + i], {i, 0, ν2}] == Sum[(r + j) c2[[j + 1]] μY[r + j - 1], {j, 0, δ}])
Z2 = Solve[Table[LS2[r, ν2, δ2], {r, 0, ν2}], Table[K2[i], {i, 0, ν2}]];

K2_i_ := K2_i = Z2[[1, i + 1, 2]]
C2_i_ := C2_i = c2[[i + 1]]
Z2 // N
Table[K2_i, {i, 0, ν2}] // N
Table[C2_i, {i, 0, δ2}] // N
Sum[C2_i y^i, {i, 0, δ2}] // N

mi11 = FindMinimum[{Sum[K2_i x^(i+1) / (i + 1), {i, 0, ν2}], LB ≤ x ≤ -2}, {x, -2.5}];
aa = mi11[[2, 1, 2]];
ma11 = FindMinimum[{Sum[K2_i x^(i+1) / (i + 1), {i, 0, ν2}], 2 ≤ x ≤ UB}, {x, 2.5}];
bb = ma11[[2, 1, 2]];
Sol22 =
  DSolve[p'[x] == -((Sum[K2_i x^i, {i, 0, ν2}]) / ((x - aa) (x - bb))) p[x], p[x], x] // Chop // Simplify ;
S2_DC = Sol22[[1, 1, 2]] / C[1];
Sol2[z_] := S2_DC /. {x → z}
cn11 = NIntegrate[Sol2[x], {x, aa, bb}];
g2b[x_] := Sol2[x] / cn11
I_u_,v_[z_] := If[u < z < v, 1, 0]
S2[y_] := S2[y] = I_aa, bb[y] g2b[y]
P2 = Plot[Re[S2[y]], {y, aa, bb}, PlotRange → All, PlotLabel → ν2];
Show[H2, P2]

ListPlot[data, AxesOrigin → {0, 0}]
Histogram3D[data, 20, "Probability"]
g[x_, y_] := g[x, y] = S1[x] S2[y]
Plot3D[S1[u] S2[w], {u, a, b}, {w, aa, bb}, PlotRange → All]
ft[x_, y_] := detVhi
  g[Vhi[[1, 1]] (x - me1) + Vhi[[1, 2]] (y - me2), Vhi[[2, 1]] (x - me1) + Vhi[[2, 2]] (y - me2)]
Plot3D[ft[u, w], {u, 0, 500}, {w, 1000, 15500}, PlotRange → All]
```

Kernel Smooth density function of Y1 and Y2 :

```
𝒟 = SmoothKernelDistribution[data];
Plot3D[PDF[𝒟, {x, y}], {x, 0, 500}, {y, 1000, 15500}, PlotRange → Full]
ListPlot[data]
```

Polynomial adjustment with degree “MSTerms”:

```
m1[i_, j_] := m1[i, j] = NIntegrate[x^i y^j g[x, y], {x, a, b}, {y, aa, bb}]
```

167

```
MSTerms = 7;
f3[L1_List, L2_List] := Inner[Plus, L1, L2, List];
L3 = Flatten[Table[{j, i}, {i, 0, MSTerms}, {j, 0, MSTerms}], 1];
P3 = Table[f3[L3[[i]], L3[[j]]], {i, 1, Length[L3]}, {j, 1, Length[L3]}];
M3 = Rationalize[
   Table[m1[P3[[i, j]][[1]], P3[[i, j]][[2]]], {i, Length[L3]}, {j, Length[L3]}], 10^-25];
Zv[x_, y_] := Zv[x, y] = Flatten[Table[x^j y^i, {i, 0, MSTerms}, {j, 0, MSTerms}], 1];
Gms[i_] := Gms[i] = Sum[X1[[j]]^L3[[i,1]] X2[[j]]^L3[[i,2]] / n, {j, 1, n}];
μ = Rationalize[Table[Gms[i], {i, Dimensions[L3][[1]]}], 10^-25] // N;
c3 = LinearSolve[M3, μ] // N
t4[x_, y_] := t4[x, y] = g[x, y] c3.Zv[x, y];
cii = NIntegrate[t4[x, y], {x, a, b}, {y, aa, bb}] // N
t6[t1_, t2_] := t6[t1, t2] = t4[t1, t2] / cii
ff[u_, w_] := ff[u, w] = detVhi t4[Vhi[[1, 1]] (u - me1) + Vhi[[1, 2]] (w - me2),
     Vhi[[2, 1]] (u - me1) + Vhi[[2, 2]] (w - me2)] / cii

Off[NIntegrate::izero]
Off[NIntegrate::"ncvb"]
F[u_, w_] := F[u, w] = NIntegrate[ff[x1, y1], {x1, 0, u}, {y1, 1000, w}]
EstCDF = Table[F[Y1[[j]], Y2[[j]]], {j, 1, n}];

MSTerms
error = Sum[(EmpCDF[[j]] - Max[0, EstCDF[[j]]])^2, {j, 1, n}]
Plot3D[ff[u, w] , {u, 0, 500}, {w, 1000, 15500}, PlotRange → All]
```

SSE of KDE:

```
𝒟 = SmoothKernelDistribution[data];
Plot3D[PDF[𝒟, {x, y}], {x, 0, 500},
 {y, 1000, 15500}, PlotLabel → SmoothKernel , PlotRange → Full]
F_kde[u_, w_] := CDF[𝒟, {u, w}]
KdeCDF = Table[F_kde[Y1[[j]], Y2[[j]]], {j, 1, n}];
error = Sum[(EmpCDF[[j]] - Max[0, KdeCDF[[j]]])^2, {j, 1, n}]
```

Vita

Hossein Zareamoghaddam

Education

- B.Sc., Pure Mathematics, Zabol University, Iran, 2000-2004.

- M.Sc., Applied Mathematics, Guilan University, Iran, 2005-2007.

- Ph.D., Statistics, The University of Western Ontario, Canada, 2014-2018,

Supervisors: S.B. Provost and S.E. Ahmed

Conference Presentations

- Sixth International Workshop on the Perspectives on High-Dimensional Data Analysis (HDDA-VI), Toronto ON, 2016 (presentation).

- Statistical Society of Canada (SSC) Annual Meeting, St. Catherines, ON, 2016 (presentation).

Published or Working Papers

- Zareamoghaddam, H., Provost, S.B. and Ahmed, S.E., (2017). A log-density estimation methodology applicable to massive bivariate data, *International Journal of Statistics and Probability*, 6(5), 1–12.

- Zareamoghaddam, H., Provost, S.B. and Ahmed, S.E., (2017). A moment-based bivariate density estimation methodology applicable to big data modeling, *Journal of Probability and Statistical Science*, 15(2), 135–152.

- Diedrichsen, J., Provost, S.B. and Zareamoghaddam, H., On the distribution of cross-validated Mahalanobis distances, https://arxiv.org/abs/1607.01371.

- Shrinkage estimation applied to a semi-nonparametric regression model. Working paper.

- An explicit representation of differentiated log-density approximants. Working paper.