Computational Statistics (2008) 23:343–360
DOI 10.1007/s00180-007-0070-z

ORIGINAL PAPER

A filtered polynomial approach to density estimation

Dominik Heinzmann

Accepted: 22 April 2007 / Published online: 3 July 2007

Abstract In this paper, a little known computational approach to density estimation based on filtered polynomial approximation is investigated. It is accompanied by the first online available density estimation computer program based on a filtered polynomial approach. The approximation yields the unknown distribution and density as the product of a monotonic increasing polynomial and a filter. The filter may be considered as a target distribution which gets fixed prior to the estimation. The filtered polynomial approach then provides coefficient estimates for (close) algebraic approximations to (a) the unknown density function and (b) the unknown cumulative distribution function as well as (c) a transformation (e.g., normalization) from the unknown data distribution to the filter. This approach provides a high degree of smoothness in its estimates for univariate as well as for multivariate settings. The nice properties as the high degree of smoothness and the ability to select from different target distributions are suited especially in MCMC simulations. Two applications in Sects. 1 and 7 will show the advantages of the filtered polynomial approach over the commonly used kernel estimation method.

Keywords Density estimation · Empirical transformation · Filtered polynomial · MCMC simulation · Multivariate settings

JEL Classification C63

1 Motivation

Distribution and density estimation is a central concept in statistical data analysis. Given a sample of random variables from a population, one wishes to estimate the

D. Heinzmann (✉)
Institute of Mathematics, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland
e-mail: dominik.heinzmann@math.uzh.ch

underlying unknown cumulative distribution function $F(x)$ and the density function $f(x)$ of the population.

Based on the work of Elphinstone (1983, 1985), we implement an multi-dimensional nonparametric density estimate using a filtered polynomial approach to estimate $F(x)$ and $f(x)$ by $H(t(x))$ and $h(t(x))t'(x)$, where H, called the filter, is a continuous target distribution, h its derivative, $t(x)$ a monotonic increasing transformation and $t'(x)$ its derivative. As we will show, monotonic increasing polynomials can be used to approximate $t(x)$ in the approach and therefore, we refer to this approach as the filtered polynomial density estimation (FPDE). A more detailed discussion of the FPDE is given in Sect. 3. For instance, we focus on applications of the FPDE for showing some of its qualities. We compare our approach to the commonly used nonparametric kernel density estimation (Silverman 1986; Jonathon 1992; Linton and Nielsen 1995). The kernel density estimation was executed by using the function *density* of the software package R (R Development Core Team 2006) with a gaussian kernel and the standard deviation of the smoothing kernel as bandwidth. The algorithm used for the FPDE was written by Heinzmann (2005).

Figure 1 displays on the left side the application of the FPDE approach (solid line) and the kernel density estimation method with a gaussian kernel (dashed line) to a data set of size 50, derived from the density function

$$f(x) = 0.8 f_1(x) + 0.2 f_2(x),$$

where $f_1(x)$ and $f_2(x)$ are the density functions of the two normal distributions $N(-2, 1)$ and $N(2, 1)$. The density estimates are plotted versus the histogram of the data. On the right side of the figure, the estimated transformation $t(y)$ of the FPDE approach (solid line) is plotted against a straight dotted line which indicates an identity transformation of the form $t(y) = x$. Note that $y = (x - \hat{a})/\hat{b}$, where $\hat{a}$ and $\hat{b}$ are the initially estimated transformation parameters which are discussed in Sect. 5. The deviation of the transformation t from the dotted line is a measure on how much the data needs to be modified in order to fit into the filter. For the FPDE, H is arbitrary chosen to be a normal distribution. For the kernel density estimation, the function *density* of the software package R (R Development Core Team 2006) with a gaussian kernel and the standard deviation of the smoothing kernel as bandwidth is used.

Both estimates detects the bimodal nature of the data. But the FPDE shows a higher degree of smoothness since it is a global estimation method where one considers all observations simultaneously for finding an adequate estimate. In contrast, the kernel density estimation is a local estimation method. It builds an estimate of the density from pieces that have been constructed using information provided primarily by observations made in a local neighborhood of each point. Potential applications of the FPDE are discussed in the next section.

2 Potential applications

The FPDE provides three functions of interest. First an estimate, $\hat{F}(x) = H(t(x))$, of the distribution function, then an estimate $\hat{f}(x) = h(t(x))t'(x)$ of the density

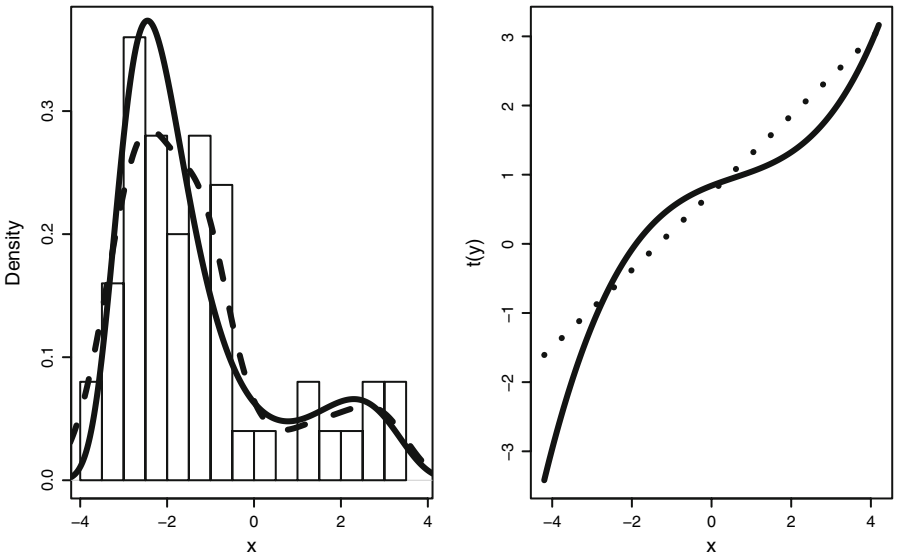

Fig. 1 **a** Application of the FPDE approach (*solid line*) and the kernel density estimation method (*dashed line*) to the data of size 50 derived from a mixture of normal distributions. **b** The transformation $t(y)$ (**solid line**) in the FPDE approach is plotted versus a *straight dotted line* which indicates an identity transformation of the form $t(y) = x$. Note that $y = (x - \hat{a})/\hat{b}$, where $\hat{a} = -1.323$ and $\hat{b} = 1.789$ are the initially estimated mean and SD of the data. (See Sect. 5 for an explanation of the initial data transformation)

function $f(x)$, and finally an estimate of the transformation function $t(x)$. For example, if H is chosen to be a normal distribution, the estimate of $t(x)$ provides an estimated normalizing transformation. The estimate of $t(x)$ of the previous application to a mixture of normal distributions is represented in Fig. 1. The deviation the dashed line corresponds to the identity transformation $t(x) = x$. The deformations of the transformation t (solid line) compared to the straight dotted line indicate the modifications necessary to the normal filter in order to fit the bimodal nature of the data.

Potential applications of filtered polynomial estimation include: (i) Density estimation for data, (ii) Estimation of transformations from empirical data distributions to the normal distribution or to other standard statistical distributions, (iii) Estimation of exceedance probabilities for test statistics or quantiles of estimators in bootstrap applications, and (iv) Discriminant Analysis (Cooley and MacEachern 1998).

3 Transformation model

In this section, the FPDE approach will be validated. We will show that for infinite large samples, the FPDE approach yields an arbitrary close approximation to the true underlying distribution functions $F(x)$ and the corresponding density function $f(x)$. The justification of the approach is divided into three steps.

In the first step, it will be shown that $F(x) = H(t(x))$, where H is the filter and $t(x)$ a monotonic increasing transformation and that $h(t(x))t'(x)$ can be used to estimate

$f(x)$, where h is the derivative of the filter and $t'(x)$ is the derivative of $t(x)$. In a second step, it will be justified that the unknown transformation $t(x)$ can be approximated by a monotonic increasing polynomial $m(x)$ on a bounded and closed interval $[a, b]$, where a and $b \in \mathbb{R}$. Finally, we will verify that the unknown continuous distribution function $F(x)$ can be approximated to any accuracy by $H(m(x))$ on $\mathbb{R}$, where $H(y)$ is a continuous distribution function and $m(x)$ is a monotonic increasing polynomial. The accuracy of the approximation is determined by the order of $m(x)$.

Note that the same transformation approach used in this paper may also be derived by using the Cornish–Fisher expansion (Cornish and Fisher 1937). Based on this connexion, it can be shown that in many practical applications, the accuracy of the FPDE approach is more than sufficient and that it can be computed simpler and more effective than other methods, such as numerical Fourier inversion (Jaschke 2002). The fast and effective performance of the FPDE approach is based on an imbedded structure of the polynomials (see Sects. 5, 6) and favours it compared to alternative methods like the numerical Fourier inversion.

Transformation $t(x)$

Theorem 3.1 *Let x and y be random variables from continuous distribution functions $F(x)$ and $H(y)$, respectively, where $F(x)$ is not known. Suppose that $F(x)$ and $H(y)$ tend to values of 0 and 1 at $-\infty$ and ∞, respectively. If x and y are related by $x = x(y)$, $y = y(x)$, where y is continuous and differentiable in x and x is continuous and differentiable in y, then it is possible to find at least one transformation function $t(x)$ such that $y = t(x) = H^{-1}(F(x))$.*

Proof A proof of the Theorem 3.1 can be found in Kendall and Stuart (1977). □

Corollary 3.2 *The transformation function $t(x)$ in Theorem 3.1 is continuous and monotonic increasing. Hence we can write*

$$F(x) = H(t(x)).$$

Proof Since $F(x)$ and $H(y)$ are continuous and the inverse of $H(y)$ exists, $t(x)$ is continuous and monotonic increasing. Hence $y = t(x) = H^{-1}(F(x))$ can be rewritten as $F(x) = H(t(x))$. □

Theorem 3.3 *Let $H(y)$ and $t(x)$ be given as before and let $h(y)$ and $t'(x)$ be their derivatives. Then $H(t(x))$ is a distribution function and $h(t(x))t'(x)$ is a density function.*

Proof $H(y)$ and $t(x)$ are monotonic increasing. Hence $H(t(x))$ is monotonic increasing. Hence $H(t(x))$ is a distribution function since $H(y)$ is a distribution function itself.

Suppose $H(y)$ has the derivative $h(y)$ and $t(x)$ the derivative $t'(x)$. Then $h(t(x))$ is positive. As the derivative of a monotonic increasing function, $t'(x)$ is positive. And finally $\int_{-\infty}^{\infty} h(t(x))\, t'(x)\, dx = 1$ which indicates that $H(t(x))' = h(t(x))t'(x)$ is a density function. □

Polynomial approximation to $t(x)$

Theorem 3.4 (First fundamental theorem of calculus) *If $g(x)$ is continuous on the closed interval $[a, b]$, where $a, b \in \mathbb{R}$, and $G(x)$ is the indefinite integral of $g(x)$ on $[a, b]$, then*

$$\int_a^b g(x)dx = G(b) - G(a).$$

Proof A proof of the first fundamental theorem of calculus can be found in Krantz (1999). □

Theorem 3.5 (Weierstrass' theorem on approximation of functions) *For any continuous real-valued function $g(x)$ on the interval $[a, b]$, where $a, b \in \mathbb{R}$, there exists a sequence of algebraic polynomials $p_i(x)$ which converges uniformly on $[a, b]$ to the function $g(x)$.*

Proof A proof of Weierstrass' theorem on approximation of functions can be found in Jeffreys and Jeffreys (1988). □

Theorem 3.6 *Let $t(x)$ be a continuous monotonic increasing transformation and $a, b \in \mathbb{R}$. For any $\epsilon > 0$, there is a monotonic increasing polynomial $m(x)$, such that $|t(x) - m(x)| \leq 2\epsilon(b - a)$ for $x \in [a, b]$.*

Proof Suppose $t(x)$ is a continuous monotonic increasing transformation in $[a, b]$, where $a, b \in \mathbb{R}$. Hence its derivative $t'(x)$ is continuous in $[a, b]$. Theorem 3.4 indicates that $t(x) = \xi + \int_0^x t'(u)du$, where $-\infty < x < \infty$ and ξ is an arbitrary constant. Since $t(x)$ is monotonic increasing, $t'(u)$ is positive for all u. Hence by applying Theorem 3.5, one can find a polynomial $p(u)$ for any $\epsilon > 0$ such that

$$|t'(u) - p(u)| \leq \epsilon.$$

We may rewrite this result as

$$p(u) - \epsilon \leq t'(u) \leq p(u) + \epsilon.$$

Since $t'(u)$ is positive, the polynomial $p_\epsilon(u) \doteq p(u) + \epsilon$ is positive for all u. We obtain

$$|t'(u) - p_\epsilon(u)| \leq |t'(u) - p(u)| + \epsilon \leq 2\epsilon.$$

Hence $t'(u)$ can be approximated to any accuracy by a positive polynomial.

Let $m(x) = \xi + \int_0^x p_\epsilon(u)du$ be a monotonic increasing polynomial. Applying Theorem 3.5 yields

$$\big|t(x) - m(x)\big| \leq \left|\int_0^x \big(t'(u) - p_\epsilon(u)\big)du\right| \leq \int_0^x \big|\big(t'(u) - p_\epsilon(u)\big)\big|du,$$

and finally

$$|t(x) - m(x)| \leq 2\epsilon(b-a). \tag{1}$$

Hence the unknown transformation $t(x)$ can be approximated to any accuracy by a monotonic increasing polynomial $m(x)$ on a bounded and closed interval $[a, b]$. □

Filtered polynomial estimation

Theorem 3.7 (Mean value theorem) *Let be $a, b \in \mathbb{R}$ and let $g(x)$ be a continuous function on the closed interval $[a, b]$ and differential on the open interval (a, b). Then there exists some c in (a, b) such that*

$$g'(c) = \frac{g(b) - g(a)}{b - a}.$$

Proof The proof of the mean value theorem can be found in Jeffreys and Jeffreys (1988). □

Theorem 3.8 *Let $F(x)$ and $H(y)$ be continuous distribution functions. Then for any $\epsilon > 0$, there exists a monotonic increasing polynomial $m(x)$ such that*

$$|H(m(x)) - F(x)| < \epsilon.$$

Proof Without loss of generality, we may assume that $m(x) \leq t(x)$. $H(x)$ is continuous and has a continuous density function $h(x)$ on $[a, b]$. Theorem 3.7 states that there is a real value c in $[a, b]$ such that

$$|H(m(x) - H(t(x))| = |h(c)(m(x) - t(x))| = h(c)|(m(x) - t(x))|,$$

where $h(c)$ is finite.

By applying Theorem 3.6, we obtain

$$|H(m(x) - H(t(x))| \leq 2\epsilon(b-a)h(c). \tag{2}$$

Relation (2) indicates that any arbitrarily close approximation $H(m(x))$ to $H(t(x))$ and hence to $F(x)$ can be obtained on the bounded and closed interval $[a, b]$.

To generalize our polynomial approach, we have to investigate the estimation of $F(x)$ in the domain for $x \notin [a, b]$. Note that $H(m(x))$ and $F(x)$ are both monotonic increasing functions.

If $x < a$,

$$H(m(x)) - F(x) \leq H(m(x)) \leq H(m(a)) \leq F(a) + \frac{\epsilon}{2} < \epsilon,$$

and

$$H(m(x)) - F(x) \geq -F(x) \geq -F(a) > -\frac{\epsilon}{2} > -\epsilon.$$

Hence we get

$$|H(m(x)) - F(x)| < \epsilon.$$

If $x > b$,

$$\begin{aligned} F(x) - H(m(x)) &= (1 - H(m(x))) - (1 - F(x)) \leq 1 - H(m(x)) \\ &\leq 1 - H(m(b)) \leq 1 - F(b) + \frac{\epsilon}{2} < \epsilon. \end{aligned}$$

and

$$\begin{aligned} F(x) - H(m(x)) &= (1 - H(m(x))) - (1 - F(x)) \geq -(1 - F(x)) \\ &\geq -(1 - F(b)) > -\frac{\epsilon}{2} > -\epsilon. \end{aligned}$$

So we get

$$|H(m(x)) - F(x)| < \epsilon.$$

□

4 Parametrization of the polynomial

To optimize the computing performance, the FPDE is reparameterized such that we obtain an imbedded structure of the polynomials.

4.1 Preliminaries

Suppose the degree of a monotonic increasing polynomial $m(x)$ is $2k + 1$, where k signifies the stage of our algorithm ($k = 0, 1, 2, \ldots$). Elphinstone (1985) has shown that one can find real-valued functions g_i ($i = 0, \ldots, 2k+1$) of a $(2k+2)$-dimensional vector $\boldsymbol{\Theta} = (\xi, \alpha, \lambda_{11}, \lambda_{12}, \ldots, \lambda_{k1}, \lambda_{k2})$, such that

$$m_{2k+1}(x) = m_{2k+1}(x; \boldsymbol{\Theta}) = \sum_{i=0}^{2k+1} g_i(\boldsymbol{\Theta}) x^i \tag{3}$$

is a monotonic increasing polynomial for all selections of values for $\boldsymbol{\Theta}$.

The construction of $g_i(\boldsymbol{\Theta})$ is based on the functions h_i ($i = 0, \ldots, 2k$) which are real-valued functions of a $(2k + 1)$–dimensional vector $\boldsymbol{\Theta}^*$, such that $\boldsymbol{\Theta}^* = (\alpha, \lambda_{11}, \lambda_{12}, \ldots, \lambda_{k1}, \lambda_{k2})$. Note that $\boldsymbol{\Theta} = (\xi, \boldsymbol{\Theta}^*)$. The $h_i(\boldsymbol{\Theta}^*)$ are chosen such that for all selections of values for $\boldsymbol{\Theta}^*$,

$$p_{2k}(x) = p_{2k}(x; \boldsymbol{\Theta}^*) = \sum_{i=0}^{2k} h_i(\boldsymbol{\Theta}^*) x^i \tag{4}$$

is a positive polynomial. The relation between $h_i(\boldsymbol{\Theta}^*)$ and $g_i(\boldsymbol{\Theta})$ can be written as

$$m_{2k+1}(x;\boldsymbol{\Theta}) = \xi + \int_0^x p_{2k}(t;\boldsymbol{\Theta}^*)dt = \xi + \sum_{i=1}^{2k+1} \frac{h_{i-1}(\boldsymbol{\Theta}^*)}{i} x^i$$
$$= \sum_{i=0}^{2k+1} g_i(\boldsymbol{\Theta}) x^i \,. \tag{5}$$

For every choice of values for $\xi = m(0)$, we get (3).

Let the algorithm be in stage k. Based on the relations (4) and (5), one can replace the search for the optimal coefficients of $m_{2k+1}(x;\boldsymbol{\Theta})$ and $p_{2k}(x;\boldsymbol{\Theta}^*)$ by the search for the optimal $(2k+2)$-dimensional parameter $\boldsymbol{\Theta}$. Let $\boldsymbol{\Theta}_k$ be the optimal value of $\boldsymbol{\Theta}$ in stage k. Then the distribution and density estimates of the FPDE can be computed as $H(m(x;\boldsymbol{\Theta}_k))$ and $h(m(x;\boldsymbol{\Theta}_k))\, p_{2k}(x;\boldsymbol{\Theta}_k^*)$.

4.2 Construction of positive polynomials

Positive polynomials are of even degree, the coefficient multiplying the highest order term is positive, and all roots have even multiplicity.

Suppose that $\boldsymbol{\Gamma} = (\gamma, \gamma_{11}, \gamma_{12}, \ldots, \gamma_{k1}, \gamma_{k2})$. Let $z_j = \gamma_{j1} + i\gamma_{j2}$. A positive polynomial p_{2k} of degree $2k$ may be written as

$$p_{2k}(x;\boldsymbol{\Gamma}) = \gamma \prod_{j=1}^{k} (x - z_j)(x - \bar{z_j}).$$

Let $z_j \bar{z_j} = \gamma_{j1}^2 + \gamma_{j2}^2 > 0$. Therefore

$$p_{2k}(x;\boldsymbol{\Gamma}) = \gamma \prod_{j=1}^{k} z_j \bar{z_j} \prod_{i=1}^{k} \left(\frac{x}{z_j} - 1\right)\left(\frac{x}{\bar{z_j}} - 1\right).$$

Let $w_j = \frac{1}{\bar{z_j}} = \lambda_{j1} + i\lambda_{j2}$ and $\lambda = \gamma \prod_{j=1}^{k} z_j \bar{z_j}$, we have

$$p_{2k}(x;\lambda,\lambda_{11},\lambda_{12},\ldots,\lambda_{k1},\lambda_{k2}) = \lambda \prod_{j=1}^{k} (\bar{w_j} x - 1)(w_j x - 1). \tag{6}$$

In terms of the original parameters, we can write:

$$\lambda = \gamma \prod_{j=1}^{k} (\gamma_{j1}^2 + \tau(\gamma_{j2})), \quad w_j = \lambda_{j1} + i\lambda_{j2} = \frac{1}{\gamma_{j1}^2 + \tau(\gamma_{j2})}(\gamma_{j1} + i\gamma_{j2}).$$

If $|z_j| = \sqrt{\gamma_{j1}^2 + \tau(\gamma_{j2})} \to 0$, than $|w_j| \to \infty$.

To ensure that the coefficient multiplying the highest-order term in the parametrization (6) is positive, we must have $\gamma_{j1}^2 + \tau(\gamma_{j2}) > 0$ for all j and $\lambda > 0$. One possibility to guarantee positivity is to select the function $\tau(.)$ to be $\tau(\gamma_{j2}) = \gamma_{j2}^2$. This choice ensures that $\gamma_{j1}^2 + \tau(\gamma_{j2}) > 0$ without restriction on γ_{j2}. For verifying that the scale parameter λ is greater than 0, we set $\lambda = e^{\alpha}$. Hence we obtain the following parametrization

$$p_{2k}(x; \lambda, \lambda_{11}, \lambda_{12}, \ldots, \lambda_{k1}, \lambda_{k2}) = e^{\alpha} \prod_{j=1}^{k} \left(\left(\lambda_{j1}^2 + \lambda_{j2}^2 \right) x^2 - 2\lambda_{j1} x + 1 \right) \tag{7}$$

which provides the required imbedded structure. Note that these choices of $\tau(.)$ and λ enable us to use an unconstrained search procedure in the FPDE approach.

We can rewrite Eq. (7) in terms of h_i, which are real-valued functions of $\boldsymbol{\Theta}^* = (\alpha, \lambda_{11}, \lambda_{12}, \ldots, \lambda_{k1}, \lambda_{k2})$. We get

$$p_{2k}(x; \boldsymbol{\Theta}^*) = h_0(\boldsymbol{\Theta}^*) + h_1(\boldsymbol{\Theta}^*)x + h_2(\boldsymbol{\Theta}^*)x^2 + \cdots + h_{2k}(\boldsymbol{\Theta}^*)x^{2k}.$$

Further, we note that

$$\begin{aligned} p_{2k}(x; \boldsymbol{\Theta}^*) &= p_{2k-2}(x; \boldsymbol{\Theta}^*)((\lambda_{k1}^2 + \lambda_{k2}^2)x^2 - 2\lambda_{k1}x + 1) \\ &= p_{2k-2}(x; \boldsymbol{\Theta}^*)(d_2^{(k)}(\boldsymbol{\Theta}^*)x^2 + d_1^{(k)}(\boldsymbol{\Theta}^*)x + d_0^{(k)}(\boldsymbol{\Theta}^*)), \end{aligned}$$

where $d_i^{(k)}(\boldsymbol{\Theta}^*)$ is the value of $d_i(\boldsymbol{\Theta}^*)$ in stage k for the polynomial $p_{2k}(x; \boldsymbol{\Theta}^*)$. These values are given by $d_0^{(k)}(\boldsymbol{\Theta}^*) = 1$, $d_1^{(k)}(\boldsymbol{\Theta}^*) = -2\lambda_{k1}x$ and $d_2^{(k)}(\boldsymbol{\Theta}^*) = \lambda_{k1}^2 + \lambda_{k2}^2$.

Finally, the coefficients $h_i(\boldsymbol{\Theta}^*)$ in stage k can be written as

$$h_i^{(k)}(\boldsymbol{\Theta}^*) = h_i^{(k-1)}(\boldsymbol{\Theta}^*)d_0^{(k)}(\boldsymbol{\Theta}^*) + h_{i-1}^{(k-1)}(\boldsymbol{\Theta}^*)d_1^{(k)}(\boldsymbol{\Theta}^*) + h_{i-2}^{(k-1)}(\boldsymbol{\Theta}^*)d_2^{(k)}(\boldsymbol{\Theta}^*),$$

for $i = 0, \ldots, 2k$. If $l < 0$ or $l > 2(k-1)$, we set $h_l^{(k-1)}(\boldsymbol{\Theta}^*) = 0$.

We obtained recursive formulas for the $h_i(\boldsymbol{\Theta}^*)$'s and hence we have an imbedded structure to construct the polynomials.

5 The implemented algorithm

In the previous section, a method to compute the distribution and the density estimates of the FPDE for a given stage k of the algorithm has been discussed. In this section, we will describe the global procedure of the implemented version of the FPDE. First, the algorithm will be presented. Then we will focus on the applied optimization method to find the optimal parameter value for $\boldsymbol{\Theta}$.

5.1 Global procedure

From experience, it has become clear that the sample values should not be too large because when working with high order polynomials, numerical overflow problems may arise (Heinzmann 2005). Hence an initial data transformation is required to ensure that the minimum and the maximum value of the transformed data will be of the same magnitude as that of the filter. Given the data $\boldsymbol{x} = (x_1, \ldots, x_n)$, the initial transformation has the form $(x_i - a)/b$ $(i = 1, \ldots, n)$, where a and b are the mean and the standard deviation of the filter $H(.)$. note that for a Chi-square distribution, $a = 0$ and $b = 1$. The parameters a and b are initially estimated by the maximum likelihood method.

An estimate, $\hat{\boldsymbol{\Theta}}$, of the coefficient vector, $\boldsymbol{\Theta}$, is chosen so as to minimize a measure of appropriateness of the filtered polynomial distribution for the data, $\boldsymbol{x}$. The recommended method in the algorithm is maximum likelihood. Alternative estimates are obtained by minimizing the Anderson–Darling distance between the empirical distribution function for the sample and the estimated distribution function, or the Cramer–von Mises distance between the same two distribution functions (Elfenbein 1978). The minimization of these discrepancy functions is carried out by applying a Newton–Raphson method based on an updating function for non positive Hessian matrices (Sect. 5.2). The minimization leads to estimates for the parameter $\boldsymbol{\Theta}$ and hence for the coefficients of the polynomials.

For selecting the most appropriate degree of the polynomial over all stages of the algorithm, criterions as AIC, BIC, likelihood ratio test and crossvalidation are implemented (Bozdogan 1987; Stone 1977). A definition and application of these criterions to two data sets of the mixture of normal distributions of sizes 10,000 and 40 may be found in Heinzmann (2005). These applications yields:

1. The larger the sample size, the higher the selected degree of the polynomial $m(x)$ by all criterions and hence the closer the approximation to $F(x)$ and $f(x)$.
2. The difference in time to compute the AIC, BIC and likelihood ratio test are marginal. But the computational effort to evaluate the crossvalidation criterion increases heavily by higher order polynomials.
3. Finally, the BIC criterion is the authors's proposed criterion. The BIC tends to select a lower polynomial degree than the others. This is especially favourable for smaller samples where the other criterions select polynomial degrees making the FPDE detecting artificial characteristics in the data (Heinzmann 2005).

5.2 Modified Newton–Raphson method

The minimization of the discrepancy functions is made by a Newton–Raphson procedure (Bailey 1993; Lancaster 1966). Such a procedure requires the Hessian matrix to be positive definite. To guarantee this positivity, we have implemented an update function for the case when we have a non positive definite Hessian matrix. Such update functions are widely used in optimization algorithms (Levenberg 1944). Suppose $\boldsymbol{H}_{k,m}$ is the Hessian matrix and $\boldsymbol{g}_{k,m}$ is the gradient of the discrepancy function and let $\boldsymbol{s}_{k,m}$ be the search direction. If the Hessian matrix at iteration m of stage k of our algorithm

is non positive definite, then we use the following modified Newton–Raphson step:

$$\hat{\boldsymbol{x}}_{k,m+1} = \hat{\boldsymbol{x}}_{k,m} - (\mathbf{H}_{k,m} + \boldsymbol{E}_{k,m})^{-1} \boldsymbol{g}_{k,m} \hat{\boldsymbol{x}}_{k,m}, \tag{8}$$

where $\boldsymbol{E}_{k,m} = 2|\lambda_1|\boldsymbol{I}_k$, $\boldsymbol{I}_k$ is the $(2k+2) \times (2k+2)$ identity matrix and λ_1 is the smallest eigenvalue of $\boldsymbol{H}_{k,m}$. The factor 2 is used to ensure positivity of the Hessian matrix. Based on the singular value decomposition (Golub and Van Loan 1996), we have

$$\boldsymbol{H}_{k,m} + \boldsymbol{E}_{k,m} = \boldsymbol{U}_{k,m}(\boldsymbol{D}_{k,m} + \boldsymbol{E}_{k,m})\boldsymbol{U}_{k,m}^T = \boldsymbol{U}_{k,m}(\boldsymbol{D}_{k,m} + 2|\lambda_1|\boldsymbol{I}_k)\boldsymbol{U}_{k,m}^T,$$

where $\boldsymbol{U}_{k,m}$ is a $(2k+2) \times (2k+2)$ column-orthogonal matrix such that $\boldsymbol{U}_{k,m}\boldsymbol{U}_{k,m}^T = \boldsymbol{I}_k$ and $\boldsymbol{D}_{k,m}$ is a $(2k+2) \times (2k+2)$ diagonal matrix.

Hence the quantity $2|\lambda_1|$ is added only to the eigenvalues of the Hessian matrix $\boldsymbol{H}_{k,m}$. Since λ_1 is the smallest eigenvalue, the matrix $(\boldsymbol{H}_{k,m} + \boldsymbol{E}_{k,m})$ has only positive eigenvalues and is therefore positive definite.

Note that when applying the update function, the selected search direction is no longer optimal. But it can be shown that the performance of the update function approach outperforms simple gradient descent and other conjugate gradient methods in a wide variety of problems (Levenberg 1944; Heinzmann 2005).

6 Polynomial coefficients and its derivatives

For the implementation of the previously introduced Newton–Raphson procedure with the update function for the non positive definite case, an iterative procedure for computing the coefficients of the polynomial and its derivatives is presented in this section. We will adapt the notation introduced in Sect. 4. But for reasons of simplicity, we will suppress the function arguments $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}^*$.

6.1 Coefficients of the positive polynomial

Based on the parametrization of the polynomial evaluated in Sect.4, the procedure of the algorithm in stage k can be described as follows. Let Θ_i denotes the coefficients of the parameter $\boldsymbol{\Theta}$. In particular, we have $\Theta_1 = \xi$ and $\Theta_2 = \alpha$. Suppose that $\tilde{h}_i$ $(i = 1, \ldots 2k-2)$ are the values of the coefficients of the positive polynomial h_i obtained in stage $k-1$ of our algorithm. Then based on the results from Sect. 4, our algorithm calculates in stage k the new values for h_i in three steps.

1. Set $b_0, b_1, \ldots, b_{2k-2}$:

$$b_i = \tilde{h}_i \quad (i = 2, \ldots, 2k-2),$$

2. Set $d_0,\ d_1,\ d_2\ :$

$$\begin{aligned} d_{k,0} &= 1 \\ d_{k,1} &= -2\Theta_{2k+1}, \\ d_{k,2} &= {\Theta_{2k+1}}^2 + {\Theta_{2k+2}}^2. \end{aligned}$$

3. Set $h_0,\ h_1, \ldots,\ h_{2k}\ :$

$$\begin{aligned} h_0 &= b_0 d_{k,0}, \\ h_1 &= b_1 d_{k,0} + b_0 d_{k,1}, \\ h_i &= b_i d_{k,0} + b_{i-1} d_{k,1} + b_{i-2} d_{k,2} \quad (i = 2, \ldots, 2k-2), \\ h_{2k-1} &= b_{2k-2} d_{k,1} + b_{2k-3} d_{k,2}, \\ h_{2k} &= b_{2k-2} d_{k,2}. \end{aligned}$$

We will refer to this procedure which updates the coefficients of the positive polynomial as the *stage procedure*.

6.2 Coefficients of the monotonic increasing polynomial

Given the coefficients h_i $(i = 0, \ldots, 2k)$ of the positive polynomial p_{2k}, we want to calculate the coefficients g_j $(j = 0, \ldots, 2k+1)$ of the monotonic increasing polynomial m_{2k+1}. Any monotonic increasing polynomial m_{2k+1} can be expressed as the integral of a positive polynomial p_{2k} which satisfies $p_{2k}(x) = \frac{d}{dx} m_{2k+1}(x)$. Hence we can find for all m_{2k+1} a $\boldsymbol{\Theta}$, such that

$$m_{2k+1}(x) \;=\; \xi \;+\; \int_0^x \sum_{i=0}^{2k} h_i t^i dt \;=\; \xi \;+\; \sum_{i=1}^{2k+1} \frac{h_{i-1}}{i} x^i \;=\; \sum_{i=0}^{2k+1} g_i x^i,$$

where $g_0 \;=\; \xi \;=\; m(0)$. Therefore, the h_i's and g_j's are related by $g_l \;=\; h_{l-1}/l$ $(l = 1, \ldots, 2k+1)$ and we may solve for the coefficients of the monotonic increasing polynomial.

6.3 Derivatives of the positive polynomial

Since the implemented procedure is based on a Newton–Raphson method, we need to compute the derivatives of the coefficients for the positive polynomial. Based on these derivatives, we then may compute the derivatives of the monotonic increasing polynomial by applying the following relations:

$$\frac{dg_l}{d\Theta_n} = \frac{1}{l}\frac{dh_{l-1}}{d\Theta_n} \quad \text{and} \quad \frac{d^2 g_l}{d\Theta_n^2} = \frac{1}{l}\frac{d^2 h_{l-1}}{d\Theta_n^2} \quad (l = 1, \ldots, 2k+1). \tag{9}$$

Gradient of the positive polynomial

In stage k, the nth element of the gradient of a positive polynomial can be computed in two steps. First, the coefficients $\partial h_i/\partial\Theta_n$ for $i = 0, \ldots, 2k$ are evaluated. Then we compute the nth element of the gradient by using

$$\frac{\partial p_{2k}(x)}{\partial\Theta_n} = \left(\left(\left(\frac{\partial h_{2k}}{\partial\Theta_n}x + \frac{\partial h_{2k-1}}{\partial\Theta_n}\right)x + \frac{\partial h_{2k-2}}{\partial\Theta_n}\right)x \ldots\right)x + \frac{\partial h_0}{\partial\Theta_n}.$$

Note that

$$\frac{\partial h_i}{\partial\Theta_1} = \frac{\partial h_i}{\partial\xi} = 0 \quad \text{and} \quad \frac{\partial h_i}{\partial\Theta_2} = \frac{\partial h_i}{\partial\alpha} = h_i,$$

for $i = 0, \ldots, 2k$.

At stake k, the calculations of $\frac{\partial h_i}{\partial\Theta_n}$ for $n > 3$ is made by applying the following substitutions to the stage procedure if $n = 2k + 1$ or $n = 2k + 2$:

$$d_{k,0} \longleftarrow 0, \quad d_{k,1} \longleftarrow \frac{\partial}{\partial\Theta_n}d_{k,1}, \quad d_{k,2} \longleftarrow \frac{\partial}{\partial\Theta_n}d_{k,2}\,.$$

Hessian matrix of the positive polynomial

For evaluating the (n, m)th component of the Hessian matrix of a positive polynomial in stage k, we proceed in two steps. First we compute

$$\frac{\partial^2 h_l}{\partial\Theta_n\partial\Theta_m} \quad \text{for } l = 0, \ldots, 2k. \tag{10}$$

Then we evaluate the (n, m)th Hessian element as

$$\frac{\partial^2 p_{2k}(x)}{\partial\Theta_n\partial\Theta_m} = \left(\left(\left(\frac{\partial^2 h_{2k}}{\partial\Theta_n\partial\Theta_m}x + \frac{\partial^2 h_{2k-1}}{\partial\Theta_n\partial\Theta_m}\right)x + \frac{\partial^2 h_{2k-2}}{\partial\Theta_n\partial\Theta_m}\right)x \ldots\right)x + \frac{\partial^2 h_0}{\partial\Theta_n\partial\Theta_m}.$$

Note that

$$\frac{\partial^2 h_i}{\partial\Theta_1\partial\Theta_m} = 0 \quad \text{and} \quad \frac{\partial^2 h_i}{\partial\Theta_2^2} = h_i,$$

for $n = 1, \ldots, 2k + 2$ and $i = 0, \ldots, 2k$. To calculate all other derivatives in (10), we group them into three cases. In every case, we modify the stage procedure in a different way.

Case 1 $\frac{\partial^2}{\partial\Theta_2\partial\Theta_m}h_i(x)$ $(m > 2)$. The derivative with respect to Θ_2 are the coefficients themselves. Hence if $m = 2k + 1$ or $m = 2k + 2$, we have the same modifications of the stage procedure as for the calculation of the gradient, i.e.,

$$d_{k,0} \longleftarrow 0, \quad d_{k,1} \longleftarrow \frac{\partial}{\partial\Theta_m}d_{k,1}, \quad d_{k,2} \longleftarrow \frac{\partial}{\partial\Theta_m}d_{k,2}.$$

Case 2 $\frac{\partial^2}{\partial{\Theta_n}^2}h_i(x)$ $(n > 2)$. The following modifications of the stage procedure are made for the case when $n = 2k + 1$ or $n = 2k + 2$:

$$d_{k,0} \longleftarrow 0, \quad d_{k,1} \longleftarrow \frac{\partial^2}{\partial{\Theta_n}^2}d_{k,1}, \quad d_{k,2} \longleftarrow \frac{\partial^2}{\partial{\Theta_n}^2}d_{k,2}.$$

Case 3 $\frac{\partial^2}{\partial\Theta_n\partial\Theta_m}h_i(x)$ $(n, m > 2\,,\ n \neq m)$. Similar to case 2, we modify the stage procedure if $n = 2k + 1$ or $n = 2k + 2$ as

$$d_{k,0} \longleftarrow 0, \quad d_{k,1} \longleftarrow \frac{\partial}{\partial\Theta_n}d_{k,1}, \quad d_{n,2} \longleftarrow \frac{\partial}{\partial\Theta_n}d_{k,2}\,.$$

If $m = 2k + 1$ or $m = 2k + 2$, the modifications are

$$d_{k,0} \longleftarrow 0, \quad d_{k,1} \longleftarrow \frac{\partial}{\partial\Theta_m}d_{k,1}, \quad d_{k,2} \longleftarrow \frac{\partial}{\partial\Theta_m}d_{k,2}\,.$$

It is important to recognize that if $n > 1$ is odd and $m = n + 1$, then

$$\frac{\partial^2}{\partial\Theta_n\partial\Theta_m}h_i = 0\ \forall\ i \quad \text{and hence} \quad \frac{\partial^2}{\partial\Theta_n\partial\Theta_m}p_{2k}(x) = 0.$$

6.4 Derivatives of the monotonic increasing polynomial

Based on the derivatives of the positive polynomial, the derivatives

$$\frac{\partial g_j}{\partial\Theta_n} \quad \text{and} \quad \frac{\partial^2 g_i(x)}{\partial\Theta_n\partial\Theta_m}$$

for $j = 0, \ldots, 2k + 1$ and $n, m = 1, \ldots, 2k + 2$ can be computed by using the relations in (9). Note in this context that $\frac{\partial}{\partial\Theta_1}m_{2k+1}(x) = 1$ and that

$$\frac{\partial^2 m_{2k+1}(x)}{\partial{\Theta_1}^2} = \frac{\partial^2 m_{2k+1}(x)}{\partial\Theta_1\partial\Theta_m} = 0\ (m > 1).$$

Finally the nth element of the gradient can be computed by using

$$\frac{\partial m_{2k+1}(x)}{\partial\Theta_n} = \left(\left(\left(\frac{\partial g_{2k+1}}{\partial\Theta_n}x + \frac{\partial g_{2k}}{\partial\Theta_n}\right)x + \frac{\partial g_{2k-1}}{\partial\Theta_n}\right)x \ldots\right)x + \frac{\partial g_0}{\partial\Theta_n},$$

and the (n, m)th element of the Hessian matrix can be calculated as

$$\frac{\partial^2 m_{2k+1}(x)}{\partial\Theta_n\partial\Theta_m} = \left(\left(\left(\frac{\partial^2 g_{2k+1}}{\partial\Theta_n\partial\Theta_m}x + \frac{\partial^2 g_{2k}}{\partial\Theta_n\partial\Theta_m}\right)x + \frac{\partial^2 g_{2k-1}}{\partial\Theta_n\partial\Theta_m}\right)x \ldots\right)x + \frac{\partial^2 g_0}{\partial\Theta_n\partial\Theta_m}.$$

7 Application of the FPDE

In this section, a second application to a univariate setting is given. The data sets contains 100 test statistics of a MCMC simulation of Multitrait-Multimethod data. The data set is described in detail in Heinzmann (2005). The asymptotic distribution is known to be a Chi-square distribution with undetermined degree of freedom. Hence the filter in the FPDE is specified as a Chi-square distribution. In order to compare the FPDE to the kernel density estimation, the kernel function *density* of the software package R (R Development Core Team 2006) with a gaussian kernel and the standard deviation of the smoothing kernel as bandwidth is selected.

On the left side of Fig. 2, the two density estimates of the FPDE (solid line) and the kernel estimation (dashed line), superposed to the histogram of the data, are represented. The FPDE estimate yields a higher degree of smoothness based on the fact that it is a global estimation method which builds an estimate with all available data on the same time. The kernel estimate as a local estimate detects more of the local features of the data. Since the data is generated by a simulation study where one supposes a high degree of smoothness, the FPDE provides a more appropriate result. The FPDE estimate also provides an algebraic representation of the density estimate which is computationally less expensive to store than all estimated points of the kernel method. The right side of the figure shows the polynomial transformation $m(x)$ of the FPDE approach (solid line) and the straight dotted line which corresponds to an identity transformation of the form $m(x) = x$. Note that in this case, the initial parameters a and b are set to 0 and 1 since the filter is specified as a Chi-square distribution. The plot indicates that only a small transformation of the data is necessary in order to fit into the filter. More precise, only the upper tail of the distribution function needs to be modified in order to get a good estimate.

8 Extension to multivariate settings

Let $\boldsymbol{X}$ be a p-dimensional random vector with expected values $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, which is symmetric and positive definite. For simplicity, set $\boldsymbol{\mu}$ equal to the p-dimensional null vector. The principal component transformation can be written as

$$\boldsymbol{X} \mapsto \boldsymbol{Y} = \boldsymbol{A}^T\boldsymbol{X}, \tag{11}$$

where $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_p]$ is an orthogonal matrix with the eigenvectors $\boldsymbol{a}_i$ as columns. The matrix $\boldsymbol{A}$ satisfies $\boldsymbol{A}^T\boldsymbol{\Sigma}\boldsymbol{A} = \boldsymbol{\Omega}$, where $\boldsymbol{\Omega}$ is a diagonal matrix with the eigenvalues

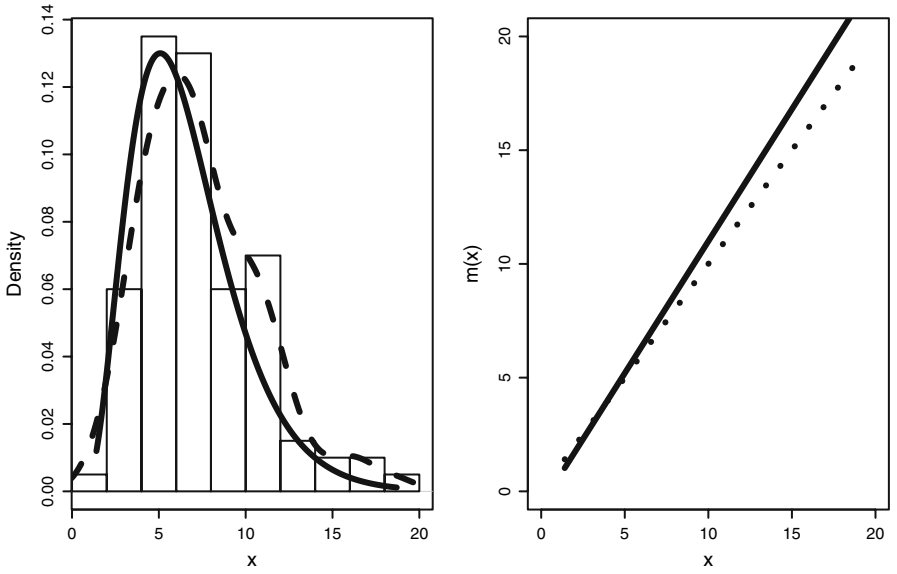

Fig. 2 **a** Application of the FPDE approach (*solid line*) and the kernel density estimation method (*dashed line*) to the Multitrait-Multimethod data with sample size 100. **b** The polynomial transformation $m(x)$ (*solid line*) in the FPDE approach is plotted versus a *straight dotted line* which indicates an identity transformation of the form $m(x) = x$. (Note that in this application, the initial data transformation parameters a and b are 0 and 1, respectively)

of $\boldsymbol{\Sigma}$ as elements such that $\omega_1 \geq \omega_2 \geq \ldots \geq \omega_p \geq 0$. The ith principal component of $\boldsymbol{X}$ can hence be written as $Y_i = \boldsymbol{a}_i^T \boldsymbol{X}$ $(i = 1, \ldots, p)$. The $\boldsymbol{a}_i^T$ are called the vector-loadings. The orthogonality of $\boldsymbol{A}$ implies the uniqueness of the principal components (Mardia et al. 1979).

If $\boldsymbol{X}$ is distributed normally, we may conclude that within the coordinate system defined by $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_p$, the coordinates of $\boldsymbol{X}$ are independent since not correlated. Let $\boldsymbol{x}$ be a realization of $\boldsymbol{X}$. Based on Eq. (11), we can write $\boldsymbol{y} = \boldsymbol{A}^T \boldsymbol{x}$ and we have

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{\boldsymbol{Y}}(\boldsymbol{y}) \cdot det(\boldsymbol{A}^T).$$

Since $det(\boldsymbol{A}) = 1$, we have

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{\boldsymbol{Y}}(\boldsymbol{a}_1^T \boldsymbol{x}, \ldots, \boldsymbol{a}_p^T \boldsymbol{x}) = \prod_{i=1}^{p} f_{Y_i}(\boldsymbol{a}_i^T \boldsymbol{x}).$$

Hence the univariate estimates f_{Y_i} lead to a multivariate estimate $f_{\boldsymbol{X}}$ (Cooley and MacEachern 1998). Applications of the FPDE to multivariate settings provide a high degree of smoothness (Heinzmann 2005) and an algebraic expression of its multivariate estimate.

9 Discussion

The filtered polynomial approach provides coefficient estimates for (close) algebraic approximations to the unknown cumulative distribution and density functions, $F(x)$ and $f(x)$, as well as a transformation (e.g., normalization) from the unknown data distribution to a target distribution (filter).

This approach provides a high degree of smoothness in its estimates for univariate as well as for multivariate settings (Heinzmann 2005).

The main difference between the filtered polynomial method and the kernel method is the procedure to estimate data. The filtered polynomial method can be seen as a global estimation where one considers all observations simultaneously for finding an adequate estimate. Hence we obtain a high degree of smoothness of the estimates as has been shown in the two applications (Sects. 1, 7). The kernel estimation can be seen as a local estimation since it builds its estimates using information provided primarily by observations made in a local neighborhood of each point. This results in a less smoothed representation of the distribution of the data (Sects. 1, 7).

From simulations it becomes clear that for randomly generated data (e.g., in MCMC applications), where one supposes a high level of smoothing, filtered polynomial estimation would be preferred. In situations where one supposes special characteristics in the data (real data), the kernel method would be preferred. But sometimes, a highly smoothed estimation of real data is less complicated to interpret than a less smoothed one. Hence the choice of one of the two methods should be based on the purpose of its application and the origin of the data (Bailey 1993).

Acknowledgments This research project was supervised by Prof. Michael W. Browne (Ohio State University, Columbus, USA) and by Prof. Stephan Morgenthaler (Swiss Federal Institute of Technology, Lausanne, Switzerland).

References

Bailey DH (1993) Multiprecision translation and execution of fortran programs. ACM Trans Math Softw 19(3):288–319

Bozdogan H (1987) Model selection and akaike's information criterion (aic): the general theory and its analytical extensions. Psychometrika 52(3):345–370

Cooley CA, MacEachern SN (1998) Classification via kernel product estimators. Biometrika 85(4):823–833

Cornish E, Fisher R (1937) Moments and cumulates in the specification of distributions. Extrait de la Revue de l'Institute International de Statistique 4:1–14

Elfenbein L (1978) On minimum Von Mises statistic estimators. Ph.D. thesis, George Washington University

Elphinstone CD (1983) A target distribution model for nonparametric density estimation. Commun Stat Theory Methods 12(2):161–198

Elphinstone CD (1985) A method of distribution and density estimation. Ph.D. thesis, University of South Africa, Pretoria

Golub GH, Van Loan CF (1996) Matrix computations. In: The singular value decomposition and unitary matrices, 3rd edn. Johns Hopkins University Press, Baltimore, pp 70–73

Heinzmann D (2005) Computational aspects of filtered polynomial density estimation. Master's thesis, Swiss Federal Institute of Technology at Lausanne, Switzerland. http://www.math.unizh.ch/user/heinzmann/software

Jaschke SR (2002) The cornish fisher expansion in the context of delta gamma normal approximations. J Risk 4(4):33–52

Jeffreys B, Jeffreys H (1988) Method of mathematical physics. In: Mean-value theorems, 3rd edn. Cambridge University Press, Cambridge, pp 446–448
Jonathon DV (1992) Nonlinear vision. In: Nonlinear systems analysis in vision: overview of kernel methods, 1 edn. CRC Press, London, pp 1–37
Kendall M, Stuart A (1977) The advanced theory of statistics: distribution theory, 4 edn, vol 1. Griffin, London. Provided by the Smithsonian/NASA Astrophysics Data System
Krantz SG (1999) Handbook of complex variables.In: The fundamental theorem of calculus along curves. Birkhaeuser, Boston, pp 22–25
Lancaster P (1966) Error analysis for the Newton–Raphson method. Numerische Math 9(1):55–68
Levenberg K (1944) A method for the solution of certain problems in least squares. Quart Appl Math 2:164–168
Linton OB, Nielsen JP (1995) A kernel method of estimating structured nonparametric regression based on marginal integration. biometrika 82(1):93–100
Mardia KV, Kent JT, Bibby JM (1979) Multivariate analysis. In: Principal component analysis. Academic, London, pp 213–229
R Development Core Team (2006) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org
Silverman BW (1986) Density estimation for statistics and data analysis. Chapman and Hall, London
Stone M (1977) An asymptotic equivalence of choice of model by cross validation and akaike's criterion. J R Stat Soc B39:44–47