T E C H N I C A L

R E P O R T

0420

DATA-DRIVEN BOUNDARY ESTIMATION IN DECONVOLUTION PROBLEMS

A. DELAIGLE and I. GIJBELS

I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Data-driven boundary estimation in deconvolution problems*

A. Delaigle
Department of Statistics
University of California at Davis
360 Kerr Hall
One Shields Avenue
Davis, CA 95616, USA.

and

I. Gijbels
Institut de Statistique
Université catholique de Louvain
20 Voie du Roman Pays
B-1348 Louvain-la-Neuve
Belgium.

July 2004

Abstract

This paper deals with estimation of the support of a density function when only a contaminated sample from the density is available. Delaigle and Gijbels (2003) introduced a kernel-based method and studied theoretical bias and variance of the estimator. The present paper deals with the practical implementation issues of this method, and is a necessary supplement to the theoretical results to get to a data-driven method that is widely applicable. We propose two such practical data-driven procedures, and show via simulations that they perform well for a wide variety of densities (including quite difficult cases). The methods can also be applied for error-free data and as such also present data-driven procedures for estimation of boundaries in the case of non-contaminated data. Moreover they can be applied for estimating discontinuities of a density, as is shown. We illustrate the use of the proposed data-driven boundary estimation procedures in frontier estimation.

Key words and phrases: Bandwidth, Boundary points, Data-driven procedure, Deconvolution problem, Density estimation, Diagnostic function, Discontinuity points, Endpoints.

1 Introduction

In this paper, we consider data-driven kernel estimation of boundary points or discontinuity points of a density, based on a sample that contains random measurement error. The

* This research was supported by the IAP research network nr P5/24 of the Belgian State (Belgian Science Policy). The first author gratefully acknowledges support from the BAEF (Belgian American and Educational Foundation).

contamination problem, often referred to as a deconvolution problem, arises in many data applications, since in practice it is frequent to have non negligible measurement error on the data. See for example Mendelsohn and Rice (1982) for an application in microfluorimetry, or Carroll, Ruppert and Stefanski (1995) for examples in studies on public health (e.g. dietary studies). Estimation of the boundary of the support of a density function is a first necessary step when estimating a density with support different from the whole real line since, then, modifications of the usual kernel-type estimators are needed for consistent estimation of the density at the endpoints of its support. Boundary estimation is also of interest on its own, since it is the basic problem in, for example, frontier estimation in econometric efficiency analyses.

In density deconvolution problems, the interest is in estimating the unknown density f_X of a random variable X, but observations from X are contaminated by an error Z, such that one actually only observes

$$Y = X + Z \; . \tag{1.1}$$

The error term Z is assumed to be independent of X, and to have a known density f_Z, with error variance $\mathrm{Var}(Z) = \sigma_Z^2$. An i.i.d. sample $Y_1, \cdots, Y_n$ from (1.1) is available. Nonparametric kernel-type estimation of f_X based on contaminated data was studied by Carroll and Hall (1988), Stefanski and Carroll (1990), Fan (1991a,b,c, 1992) and Delaigle and Gijbels (2002, 2004a,b) among others. When the support of the density f_X is not the whole real line, one needs to introduce modifications to the usual estimation methods, as is done in, for example, Zhang and Karunamuni (2000). Such modifications however require knowledge of the support of the unknown density. In this paper we focus on data-driven procedures for estimating this boundary based on contaminated data.

In the error-free case, when i.i.d. data $X_1, \cdots, X_n$ from X are available, simple, asymptotically unbiased and consistent estimators for the endpoints τ_ℓ and τ_r of the support $[\tau_\ell, \tau_r]$ of a density f_X are respectively $\min(X_1, \cdots, X_n)$ and $\max(X_1, \cdots, X_n)$. In case of a contaminated sample $Y_1, \cdots, Y_n$ one can consider $\min(Y_1, \cdots, Y_n)$ and $\max(Y_1, \cdots, Y_n)$, but these quantities are consistent estimators of $\tau_\ell + c$ and $\tau_r + d$ respectively, where c and d are the endpoints of the support of Z. For some illustrations on this see Delaigle and Gijbels (2003). See also Figure 2.6 in Section 2.

Delaigle and Gijbels (2003) introduced a consistent method for estimating the support of a density based on a contaminated sample. We now briefly describe this method. For a given bandwidth parameter $h = h_n > 0$ and a given kernel function K, the deconvolution estimator of f_X is defined as

$$\widehat{f}_X(x; h) = \frac{1}{nh} \sum_{j=1}^{n} K_Z\Big(\frac{x - Y_j}{h}\Big), \tag{1.2}$$

where $K_Z(u) = (2\pi)^{-1} \int e^{-itu} \varphi_K(t)/\varphi_Z(t/h)\, dt$, with φ_L the Fourier transform of a function or a random variable L. See Carroll and Hall (1988) and Stefanski and Carroll (1990) for an introduction to this estimator. Throughout this paper, we consider real and symmetric kernel functions K. This guarantees that K_Z is real and symmetric.

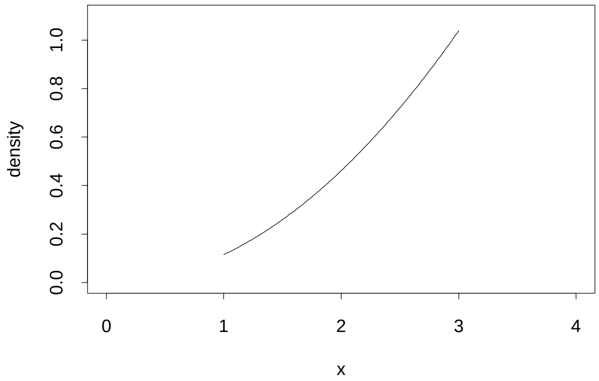


Figure 1.1: *Example of a density with a compact support.*

Figure 1.1 presents an example of a density f_X of support $[1, 3]$. For estimating the endpoints 1 and 3, Delaigle and Gijbels (2003) propose to look at the diagnostic function

$$\widehat{J}(x) = \frac{1}{nh} \sum_{i=1}^{n} K'_Z\Big(\frac{x - Y_i}{h}\Big) \,, \tag{1.3}$$

which is in fact h times the derivative of the deconvolution kernel density estimator of f_X, as given in (1.2). The multiplicative factor h is only introduced for (theoretical) convenience. The kernel estimate of f_X is a smooth function and is a consitent estimate of f_X in continuity points of the unknown density. In the context of a density with support different from the whole real line, it is to be expected that this estimate will be continuous but with large (in absolute value) derivatives when approaching the endpoints (large positive derivative for a left endpoint and large negative derivative for a right endpoint). Note that we exploit here the (inconsistent) behaviour of the kernel estimate at the boundary. A left boundary point τ_ℓ can thus be estimated via $\text{argmax}_x \widehat{J}(x)$, whereas a right boundary point τ_r can be estimated by $\text{argmin}_x \widehat{J}(x)$.

These estimators of the endpoints τ_ℓ and τ_r of the support clearly depend on the smoothing parameter h. Delaigle and Gijbels (2003) studied the theoretical properties of these consistent estimators, including bias, variance and mean squared error. Their theoretical results also allowed to discuss, on theoretical grounds, the choice of an optimal bandwidth (in the mean squared error sense). One of the interesting findings is that the behaviour of a theoretical optimal bandwidth, as well as the existence of an analytical expression for it, depends on the tail behaviour of the characteristic function of Z. For certain classes of error densities, such a theoretical optimal bandwidth cannot be found via explicit expressions or does not even exist. For more details we refer to Delaigle and Gijbels (2003). In that paper, it was also illustrated that the performance of the estimation method heavily depends on the choice of h. With a well-chosen bandwidth the simulations showed a quite good finite sample performance. The aim of the present paper is exactly to propose practical choices of parameters, and as such to derive a fully data-driven estimation procedure.

As an illustration of the estimation method, we plot, in Figure 1.2, the diagnostic function $\widehat{J}(\cdot)$ for $n = 100$ data from model (1.1) with density f_X as in Figure 1.1, and error Z a Laplace density. We plot the diagnostic function $\widehat{J}(\cdot)$ for different values of the bandwidth h. The actual values of the boundary points $\tau_\ell = 1$ and $\tau_r = 3$ are indicated by vertical lines.

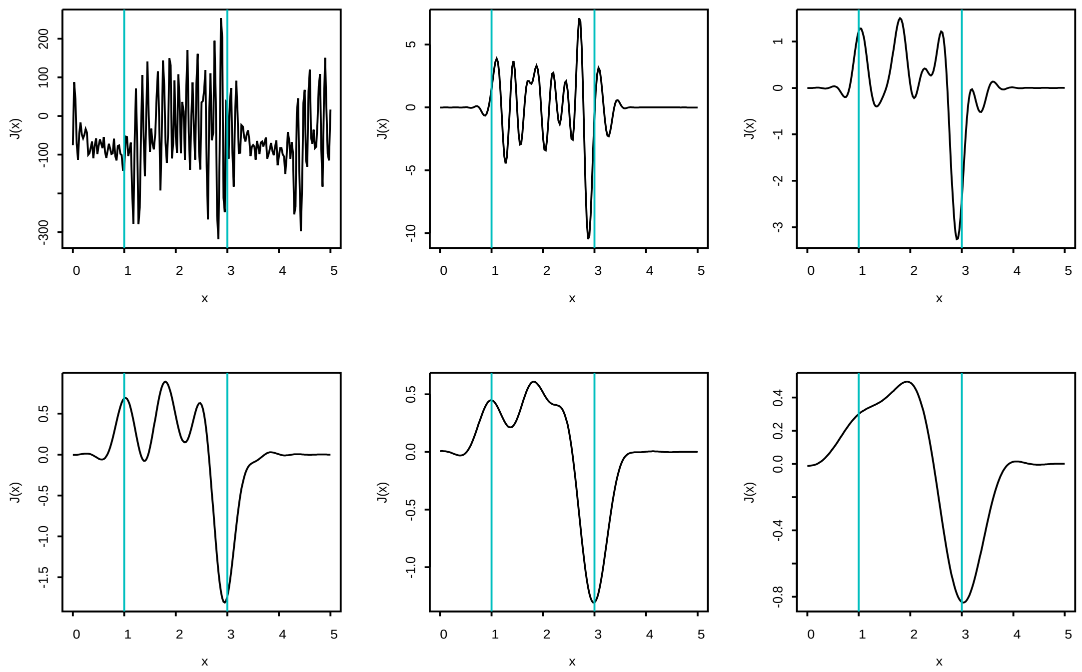


Figure 1.2: *Typical shape of $\widehat{J}(x)$ for increasing bandwidths (from left to right and top to bottom) for a Laplace contaminated sample from the density supported on* $[1, 3]$ *depicted in Figure 1.1.*

From these representative graphs, we see that the values of the global maximizer and minimizer of $\widehat{J}(\cdot)$ depend on the bandwidth chosen. In Figure 1.2, the first bandwidth is too small and introduces a lot of artificial peaks, and the last bandwidth is too large, making almost disappear the peak close to the left boundary point. This is related to the fact that $\widehat{J}(x)$ is the sum of small bumps built around each observation, and as for density estimation, the width of these bumps is dictated by the value of the bandwidth: a too small bandwidth results in a very rough diagnostic function with a lot of minima and maxima, and a too large bandwidth results in an oversmoothed diagnostic function, with too few minima or maxima. Hence a good choice of the bandwidth is a necessity.

Procedures, similar in spirit as the one above, with appropriately chosen diagnostic functions have been proposed by Wu and Chu (1993), Chu and Cheng (1996), Couallier (1999), Gijbels, Hall and Kneip (1999) and Gijbels and Goderniaux (2004), among others, to estimate change-points in a density or regression function in the error-free case. The current context is rather specific in the sense that we deal with estimation of endpoints of the support of a density and that we only have a contaminated sample.

The paper is organized as follows. In Section 2 we discuss the implementation issues of the estimation method, including the bandwidth issue. We introduce an estimation procedure for the endpoints based on a threshold parameter, as well as a stepwise estimation procedure. In Section 3 we present an extensive simulation study and demonstrate the finite sample performance of the discussed procedures by considering a wide range of densities. In Section 4 we show how to adapt the method to estimation of discontinuities of a density f_X. Finally, in Section 5 we apply the method in a data analysis from an econometric application, where the interest is to estimate the frontier function.

2 Implementation of the method

In this section, we discuss practical implementation issues for the estimation method based on the diagnostic function $\widehat{J}(\cdot)$.

To illustrate further the complexity of the estimation problem we consider the density in Figure 2.1 (a), for which $\tau_\ell = 0$ and $\tau_r = 6$. This density has a large peak in the interior of the support (at the point 3) and, as a result, the derivative f'_X, depicted in Figure 2.1 (b), shows a minimum and a maximum at interior points. As a consequence, for finite samples, $\widehat{f'_X}$ (and thus $\widehat{J}$) will also show large (absolute) values in the neighbourhood of these interior points. This can be seen as a kind of identification problem. Figure 2.2 presents the diagnostic function $\widehat{J}(\cdot)$ for increasing values of the bandwidth, based on a sample of size $n = 100$ from that density contaminated with Laplace error. From that figure and Figure 1.2, we see that it seems preferable to estimate τ_ℓ and τ_r by local extrema of $\widehat{J}$. More precisely, we consider

$$\begin{aligned}&\text{estimator of } \tau_\ell = \text{the left-most local maximizer of } \widehat{J}(x)\\&\text{estimator of } \tau_r = \text{the right-most local minimizer of } \widehat{J}(x).\end{aligned}$$

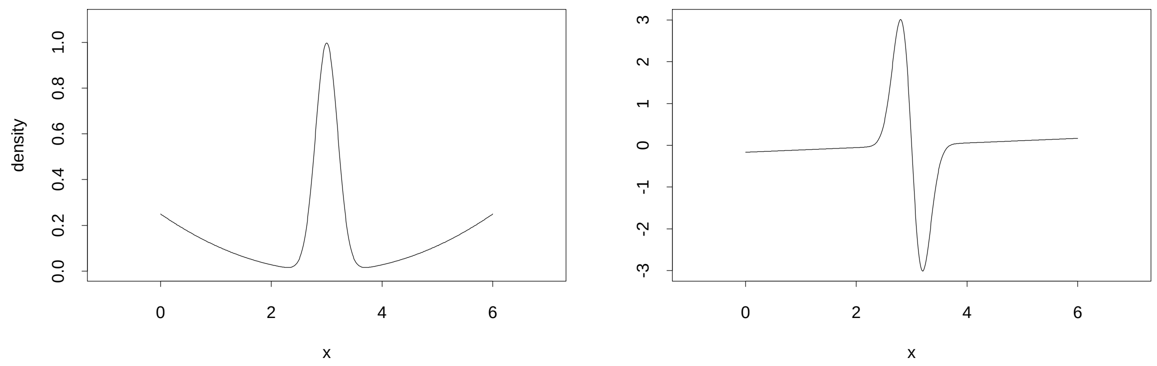


Figure 2.1: *A density (left panel) and its derivative (right panel).*

In the sequel we provide two methods that will further deal with the choice of the bandwidth. A first method combines the choice of a threshold parameter and a simple data-driven choice for h. The second method relies on a quite different approach to deal with the above mentioned identification problem and the choice of h inherent to this.

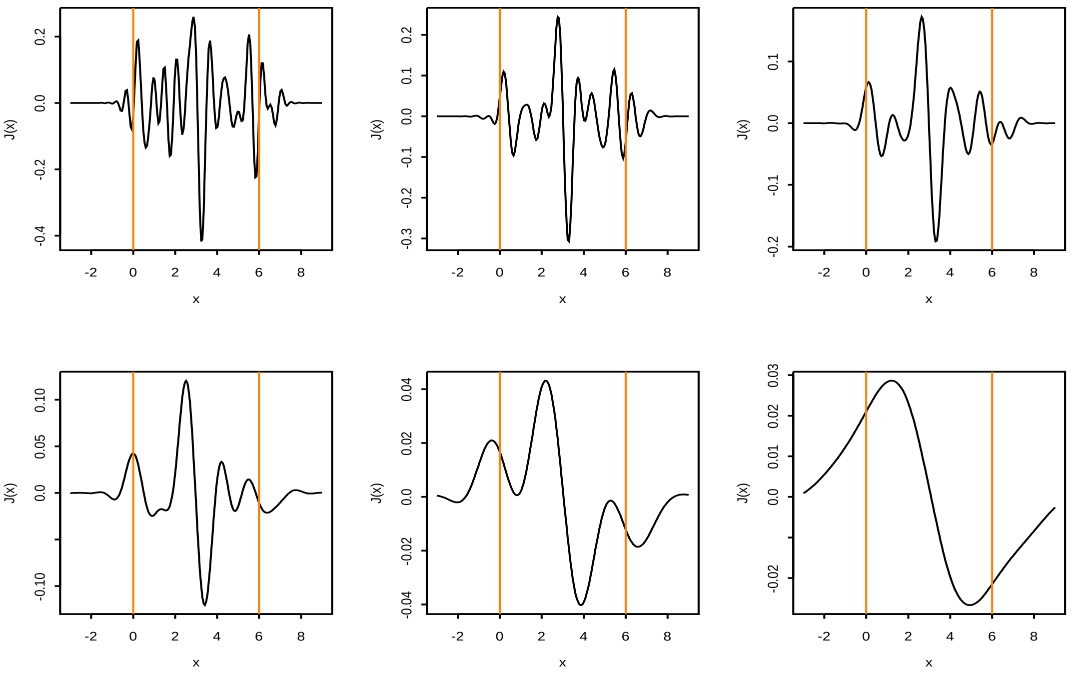


Figure 2.2: *Typical shape of* $\widehat{J}(x)$ *for increasing bandwidths (from left to right and top to bottom) from a Laplace contaminated sample from a density supported on* $[0, 6]$.

2.1 The threshold method

Looking at Figures 1.2 and 2.2, we notice that the diagnostic function can show some small spurious bumps in the extremities (away from the boundary points), especially in the graphs with smaller bandwidths. When searching for the left-most maximizer or the right-most minimizer of the diagnostic function, we do not want to take these small bumps into account. In many cases, these bumps are artifacts caused by the particular shape of K'_Z, as illustrated in Figure 2.3 for the kernel $K = \frac{48\cos x}{\pi x^4}\left(1 - \frac{15}{x^2}\right) - \frac{144\sin x}{\pi x^5}\left(2 - \frac{5}{x^2}\right)$, a commonly-used kernel in the kernel deconvolution context, and a Laplace error. The disregard of the small bumps in Figures 1.2 and 2.2 can be accomplished by introducing a certain *threshold value*.

Another reason why we need a threshold is that if f_Z has a non-compact support, $\min(Y_1, \ldots, Y_n)$ and $\max(Y_1, \ldots, Y_n)$ get further and further away from the boundary of f_X as n increases. For example, for a Gaussian error (where $c = -\infty$ and $d = +\infty$) $\min(Y_1, \cdots, Y_n)$ and $\max(Y_1, \cdots, Y_n)$ tend to $-\infty$ and $+\infty$ respectively. In this case, as the sample size increases, the probability of observing very small or very large values of Y_i increases. Since the diagnostic function builds a bump around each observation Y_i, this could result in local extrema far away from a boundary point.

The threshold should be large enough to ignore spurious bumps, but small enough so as not to exclude the bumps of interest. Since, in general, the latter have a size rather close to the global maximum or minimum of $\widehat{J}$ (see for example Figure 1.2), it seems

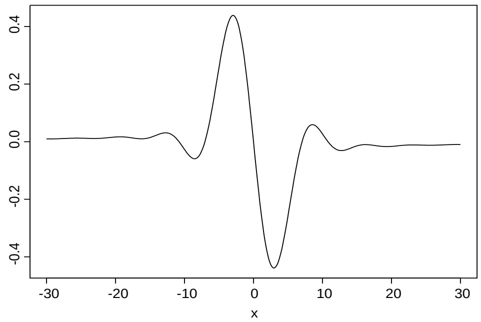


Figure 2.3: *Derivative of K_Z for a commonly-used kernel K and a Laplace(σ) error with $\sigma^2/h^2 = 5$.*

appropriate to define

$$\begin{aligned} \widehat{\tau}_\ell &= \text{the left-most local maximizer } M \text{ of } \widehat{J} \text{ that satisfies } \widehat{J}(M) \geq p \cdot \max_x \widehat{J}(x) \\ \widehat{\tau}_r &= \text{the right-most local minimizer } m \text{ of } \widehat{J} \text{ that satisfies } \widehat{J}(m) \leq p \cdot \min_x \widehat{J}(x), \end{aligned} \tag{2.1}$$

with $0 \leq p \leq 1$. Considering such a left-most local maximizer and right-most local minimizer, instead of the global maximizer and minimizer, only makes a difference for finite samples. For n tending to infinity, this local maximizer (minimizer) coincides with the global maximizer (minimizer).

In Section 3.1, we compare the results obtained using the threshold as in (2.1), with different values of p. From our extensive simulation study, we can see that there is a whole range of good p-values and that generally, the choice $p = 0.3$ seems to be an appropriate choice, meaning that we disregard bumps of size smaller than 30% of the largest (in absolute value) global bump.

It remains to deal with the choice of the bandwidth. One, theoretical, option would be to take the bandwidth h for which the asymptotic mean squared error of the estimator $\widehat{\tau}$ (where $\widehat{\tau}$ denotes $\widehat{\tau}_\ell$ or $\widehat{\tau}_r$) is minimal. Such a theoretical bandwidth exists and has been derived, in Delaigle and Gijbels (2003), for ordinary smooth error densities f_Z, i.e. error densities for which the characteristic function behaves as a polynomial of degree $-\beta$ (with $\beta > 0.5$) in the tails. Examples of ordinary smooth densities are Laplace densities and gamma densities. For supersmooth error densities (i.e. densities for which the characteristic function behaves as a negative exponential in the tails), expressions for theoretical optimal bandwidths, in the same asymptotic mean squared error sense, are not available. Examples of supersmooth densities are Gaussian and Cauchy densities.

From the asymptotic expressions for the mean squared error of $\widehat{\tau}$ established in Delaigle and Gijbels (2003), we can deduce, for ordinary smooth error densities with $\beta > 0.5$, an asymptotic theoretical optimal bandwidth, defined as the value h that minimizes

$$h^4 D_\tau^2 + hR(K_Z'')B_\tau n^{-1} \tag{2.2}$$

with respect to h, where

$$B_\tau = \frac{f_Y(\tau_+) + f_Y(\tau_-)}{2} , \quad D_\tau = K(0)\,[f'_X(\tau_+) - f'_X(\tau_-)] , \quad R(K''_Z) = \int (K''_Z(u))^2 \, du ,$$

with f_Y the density of Y. Further, $g(x_+)$ (respectively $g(x_-)$) denotes the right-hand limit (respectively the left-hand limit) of a function g. Note that the definition of K_Z depends on h, and hence that the explicit expression of the function to be minimized in (2.2) depends on the particular function K_Z.

In this problem of boundary estimation in the deconvolution context, the theoretical optimal bandwidth defined via (2.2) depends on unknown quantities, namely $f_Y(\tau_+) + f_Y(\tau_-)$ and $f'_X(\tau_+) - f'_X(\tau_-)$. In a simulation study we know all these quantities and hence we can calculate and consider this theoretical optimal bandwidth. In real data application, we need to have practical choices of the bandwidth and these quantities have to be estimated. Unfortunately, this is not an easy task. A plug-in type of bandwidth estimator, for example, would require a pilot estimator, say $\widehat{\tau}_0$, of τ. Then, the unknown quantities could be estimated by their kernel estimators $\widehat{f}_Y(\widehat{\tau}_{0+}) + \widehat{f}_Y(\widehat{\tau}_{0-})$ and $\widehat{f}'_X(\widehat{\tau}_{0+}) - \widehat{f}'_X(\widehat{\tau}_{0-})$ and then plugged into (2.2). Such a procedure however may prove to be very inaccurate and is not guaranteed to work well, since it necessitates the estimation of pointwise quantities.

As an alternative, we propose to use an approximation of the bandwidth optimal, in the mean integrated squared error sense, to estimate the derivative of f_X pretending that f_X is continuous and differentiable. More precisely, we define $h = h_{\mathrm{der}}$ the value of h that minimizes the asymptotic mean integrated squared error for estimation of f'_X, i.e.

$$\frac{1}{2\pi n h^3} \int t^2 \, |\varphi_K(t)|^2 \, |\varphi_Z(t)|^{-2} \, dt + \frac{h^4}{4} R(f_X^{(3)}) \mu_{K,2} , \tag{2.3}$$

with respect to h, where $R(f_X^{(3)}) = \int \left(f_X^{(3)}(x)\right)^2 dx$, with $f_X^{(3)}$ the third order derivative of f_X, and $\mu_{K,2} = \int x^2 K(x)\, dx$. Similarly to the normal reference rule (see for example Wand and Jones (1995)), we then estimate the unknown quantity $R(f_X^{(3)})$ by referring to a family of normal densities, i.e. supposing that $f_X = \mathrm{N}(\mu_X; \sigma_X^2)$. We then find that

$$R(f_X^{(3)}) = \frac{6!}{(2\sigma_X)^7 \; 3! \sqrt{\pi}} ,$$

where, from (1.1), σ_X^2 can be consistently estimated by

$$\widehat{\sigma}_X^2 = \widehat{\sigma}_Y^2 - \mathrm{Var}(Z) , \tag{2.4}$$

where $\widehat{\sigma}_Y^2$ is the empirical variance of $Y_1, \cdots, Y_n$, and $\mathrm{Var}(Z)$ is the known error variance. The practical bandwidth is then the value of h, denoted by $\widehat{h}_{\mathrm{der}}$, that minimizes expression (2.3) with $R(f_X^{(3)})$ replaced by its estimated value

$$\widehat{R(f_X^{(3)})} = \frac{6!}{(4\widehat{\sigma}_X^2)^{7/2} \; 3! \sqrt{\pi}} , \tag{2.5}$$

using (2.4).

We implemented the threshold-based method with this practical bandwidth and compared it with the method when using the theoretical optimal bandwidth that minimizes (2.2). From the simulation results we see that this derivative-based bandwidth seems to be a reasonable practical choice. It can be shown that this practical bandwidth satisfies the conditions imposed on the bandwidth for establishing the theoretical results (such as consistency) for the estimator $\widehat{\tau}$ of τ as exposed in Delaigle and Gijbels (2003).

2.2 The stepwise method

In this section, we propose an alternative data-driven procedure for estimating the endpoints of the support. It is inspired by ideas from Gijbels, Hall and Kneip (1999), who, in the error-free regression context, use a stepwise scheme to estimate a discontinuity point. See Gijbels, Hall and Kneip (1999) for a motivation and detailed explanation for the stepwise scheme. The method is based on the fact that, using a very large bandwidth, the diagnostic function will be oversmoothed and have one or two very broad bumps corresponding to the global maximum and/or minimum. See Figures 1.2 and 2.2. In a first step we estimate the boundary points by the global minimizer and maximizer of $\widehat{J}$ for this bandwidth. The idea is then to follow the maximizer or minimizer of $\widehat{J}$ when h decreases: on a decreasing grid of bandwidths, search for the local maximizer and/or minimizer of $\widehat{J}$ closest to the previous one(s). This s-stage procedure can be written as follows:

Step 0. Search for the global minimizer m_0 (respectively maximizer M_0) of the diagnostic function, with a very large bandwidth h_0;

Step 1. Using a bandwidth $h_1 < h_0$, search for the local minimizer m_1 (respectively maximizer M_1) of $\widehat{J}$ nearest to m_0 (respectively M_0);

$\vdots$

Step i. Using a bandwidth $h_i < h_{i-1}$, search for the local minimizer m_i (respectively maximizer M_i) of $\widehat{J}$ nearest to m_{i-1} (respectively M_{i-1});

$\vdots$

Step s. At stage s, estimate the right (respectively the left) boundary point by the local minimizer m_s (respectively maximizer M_s) of $\widehat{J}$ nearest to m_{s-1} (respectively M_{s-1}), using the bandwidth $h_s < h_{s-1}$.

We take h_0 equal to 4 times $\widehat{h}_{\text{der}}$. In general, this bandwidth is far too large for our boundary estimation problem, since it smooths the derivative too much. Hence it surely is a good starting point for the algorithm. Similarly to Gijbels, Hall and Kneip (1999), we took a geometric decreasing grid for h, which means that, as h gets smaller, we let it decrease with smaller steps (and hence are more careful). In our simulations, we took

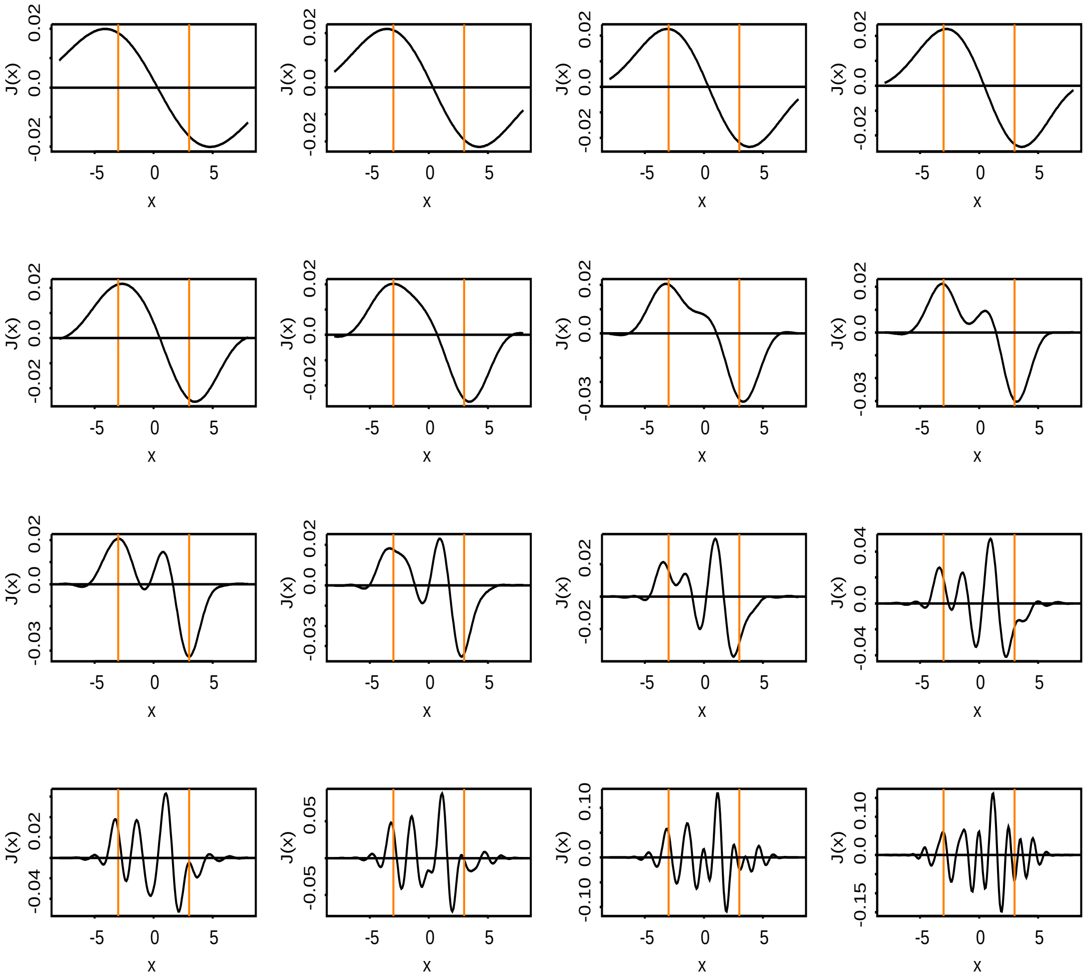


Figure 2.4: *Diagnostic function for the sinus density (Density #4) for decreasing bandwidths* $h_0, \ldots, h_{15}$ *(from left to right and from top to bottom), from a sample of size* $n = 100$ *contaminated by a Laplace error with* $\operatorname{Var} Z / \operatorname{Var} X = 10\%$*. The endpoints are indicated by two vertical lines.*

$h_{i+1} = 0.85\, h_i$. Note that the number of iteration steps need not be the same for the left and right boundary points.

Figure 2.4 shows on an example (Density #4 in Figure 3.1) that if the stopping point is well chosen (in this case, stop after 7 iterations for the left endpoint and after 8 iterations for the right endpoint), this procedure indeed gives good results.

It is clear from the above description that we do not really use the fact that we are estimating boundary points rather than ordinary discontinuity points. The procedure can be improved exploiting this knowledge so as to deal with more involved cases, such as in Figure 2.5 (simulation for Density #3 in Figure 3.1). Consider the estimation of the left boundary point 0. We see that the procedure starts well, but at step 8 (ninth picture), the left bump divides in two secondary bumps: one that moves to the left in the next steps, and the other one that stays more to the right. In step 8, it is the right

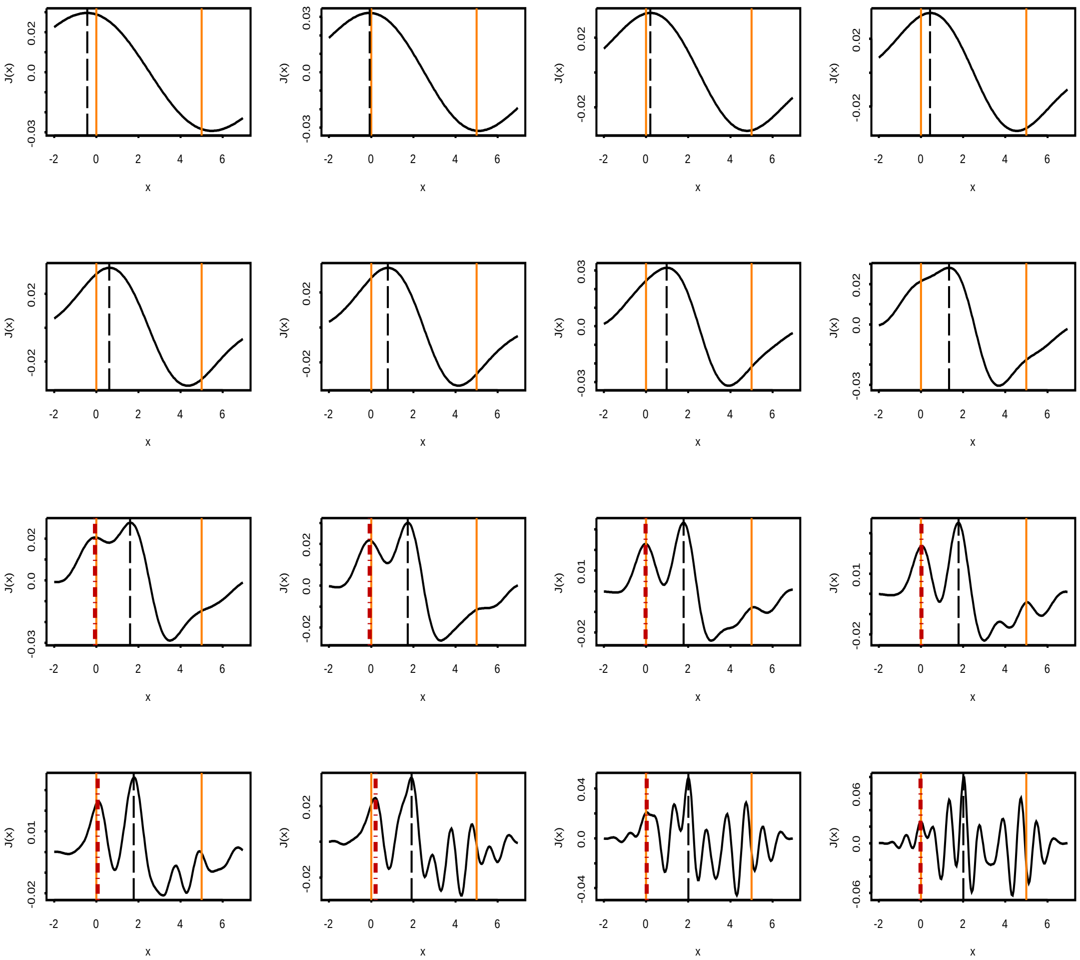


Figure 2.5: *Diagnostic function for the concave density (Density #3) for decreasing bandwidths* $h_0, \ldots, h_{15}$ *(from left to right and from top to bottom), from a sample of size* $n = 100$ *contaminated by a Laplace error with* $\operatorname{Var} Z / \operatorname{Var} X = 10\%$. *The endpoints are indicated by two vertical lines. The steps for estimating the left boundary point are indicated by the dashed-dotted vertical line.*

secondary bump that is the closest to the bump of step 7. However, we see that following this right bump in the next steps will take us further and further away from the left endpoint $\tau_\ell = 0$. By following the left secondary bump of step 8, we would have got closer and closer to 0. We observe a similar phenomenon for the estimation of the right endpoint. Of course, what we should do is use the information that we are estimating boundary points i.e. points at the extreme left or right of the support of f_X. This can be done as follows. The first time (say at step k) a secondary maximum (respectively minimum) bump appears, i.e. the bump of step $k-1$ divides in two bumps at step k, we take M_k corresponding to the left-most secondary maximum bump (respectively we take m_k corresponding to the right-most secondary minimum bump). In the next steps, we follow the regular rule, i.e. search for the local maximizer M_i (respectively minimizer

m_i) of $\widehat{J}$ nearest to M_{i-1} (respectively m_{i-1}). In Figure 2.5 we indicated with a dashed vertical line what happens when we always follow the bump next to the previous bump. The dashed-dotted vertical line indicates the steps for the proposed procedure. We stop the iterations when the difference between two successive approximations falls under a certain value, which, for simplicity, can be fixed to a small constant value c. For example, stop at step i^* if $|\widehat{\tau}_{i^*} - \widehat{\tau}_{i^*-1}| < c\,|\widehat{\tau}_1 - \widehat{\tau}_0|$, where $\widehat{\tau}_i$ denotes the obtained estimator of τ at step i. Such a procedure is simple to use and works reasonably well in practice. It is also possible to choose a stopping criterion that depends on the sample size. See Section 3.

The most difficult densities are those that have large derivatives (in absolute value) at interior points, as explained in the beginning of Section 2. For these densities, we are facing an identification problem: the procedure should distinguish between a real boundary point and a point with a large (absolute) derivative. Our procedure deals with these cases by following the left-most or right-most bump in the secondary bumps. In general, for easier densities (with no identification problem), no secondary bump appears before we stop. Note that in this second approach, the issue of the small artificial bumps in the extremeties discussed in Section 2.1 is dealt with inherently by tracing back the local maximizers (minimizers) on an increasingly-finer grid of h-values, and by the stopping criterion.

An alternative way to deal with the identification problem would be to adapt, to our context, the practical identification algorithm as proposed by Gijbels, Hall and Kneip (1999) to identify a real discontinuity point in the regression error-free case. In Section 4 we will provide a small comparison, via simulations, between our proposed data-driven procedure and this alternative procedure.

In Figure 2.6, we show the results of the proposed stepwise procedure when estimating the single (left) boundary point of the density $f_X(x) = 5e^{-5\,x}.1_{\{x>0\}}$. We generated 100 samples of size $n = 100$ and $n = 500$ contaminated by a Laplace error with $\mathrm{Var}(Z)/\,\mathrm{Var}(X) = 10\%$. Each point in the bottom panels of Figure 2.6 represents one of the 100 replications of the estimation of the left endpoint 0. The top panels show the boxplots based on these 100 estimated values. We compare the results with the naive estimator $\min(Y_1, \ldots, Y_n)$. From Figure 2.6 it is clear that the proposed data-driven stepwise method performs quite well, and much better than the naive (and inconsistent) method. The results improve as the sample size increases.

In Section 3.2 we present extensive simulation results for the stepwise method, also for the densities considered in Figures 2.4 and 2.5. From these results we can conclude that the method performs quite well on a variety of examples, and thus can be recommended in general.

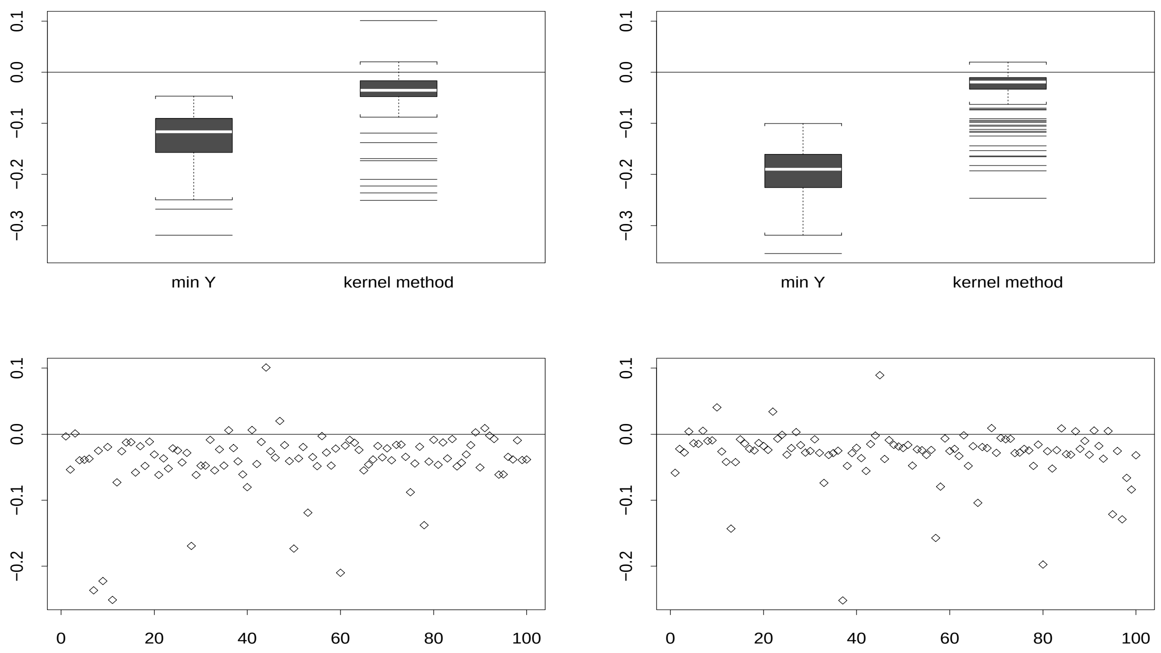


Figure 2.6: *Estimation of the left endpoint of an exponential density, from samples of size* $n = 100$ *(left panels) or* $n = 500$ *(right panels) for Laplace contaminated samples, with* $\operatorname{Var} Z/\operatorname{Var} X = 10\%$. *Top panels compare the naive method (smallest observation) with our kernel method. Bottom panels show scatterplots of the results for the kernel method.*

3 Simulation study

In our simulation study, we consider a variety of different types of densities: some 'easy' densities and some more involved densities. These densities are

#1. A uniform density: $f_X(x) = 1/3 \cdot 1_{[0,3]}(x)$;

#2. A convex density: $f_X(x) = 3/200\ (x^2 + 3) \cdot 1_{[-4,4]}(x)$;

#3. A concave density: $f_X(x) = 3/175(-x^2 + 6x + 5) \cdot 1_{[0,5]}(x)$;

#4. A sinus type density: $f_X(x) = (\sin^2(x/2) + 2)/(15 - \sin 3) \cdot 1_{[-3,3]}(x)$;

#5. A polynomial density : $f_X(x) = 3x^2/26 \cdot 1_{[1,3]}(x)$;

#6. A mixed concave-peak density: $f_X(x) = [0.8 \times\ 3/152(-x^2 + 6x + 5) + 0.2\varphi_{3,0.01}] \cdot 1_{[1,5]}(x)$;

#7. A mixed convex-peak density: $f_X(x) = 0.5[(x-3)^2/18 + \varphi_{3,0.04}] \cdot 1_{[0,6]}(x)$;

#8. A multimodal density: $f_X(x) = (1.1 - \sin(20\,x))/1.29\,.\,1_{[0,1.2]}(x)$;

#9. An exponential density: $f_X(x) = 5\exp(-5x)1_{\{x>0\}}$,

where φ_{μ,σ^2} denotes the density of a $\mathrm{N}(\mu;\sigma^2)$. These 9 densities are presented in Figure 3.1.

Densities 1, 2 and 4 are easy densities: they do not present any particularity that could perturb the estimation of the boundary points. Note that for Density 3, the jumps have a different size, and it should be easier to estimate the right endpoint where the density shows the largest jump. Density 5 is a bit more difficult because the left endpoint has a much smaller jump size than the right endpoint, what will cause some difficulties in recovering the left boundary point. The difficulties of Densities 6 and 7 come from the large peak at the interior of the support: in the neighbourhood of the peak, the derivative of the density has a large (absolute) value. Finally, the difficulty of Density 8 comes from the various oscillations at the interior of the support. Here again, this could cause large (absolute) values of the derivatives of the density estimator. Density 9 is the only density with one boundary point and an unbounded support.

In all simulations, we generated 100 samples from model (1.1) with f_X one of the above densities and a Laplace error Z such that the noise-to-signal ratio equals 0.10, i.e. $\operatorname{Var} Z/\operatorname{Var}(X) = 10\%$. For the kernel K in (1.2) and (1.3) we took $K = \frac{48\cos x}{\pi x^4}\left(1 - \frac{15}{x^2}\right) - \frac{144\sin x}{\pi x^5}\left(2 - \frac{5}{x^2}\right)$, for which a typical K_Z' is depicted in Figure 2.3.

We also conducted simulations for a Gaussion error Z, and these led to similar conclusions. Hence we do not report on these simulations.

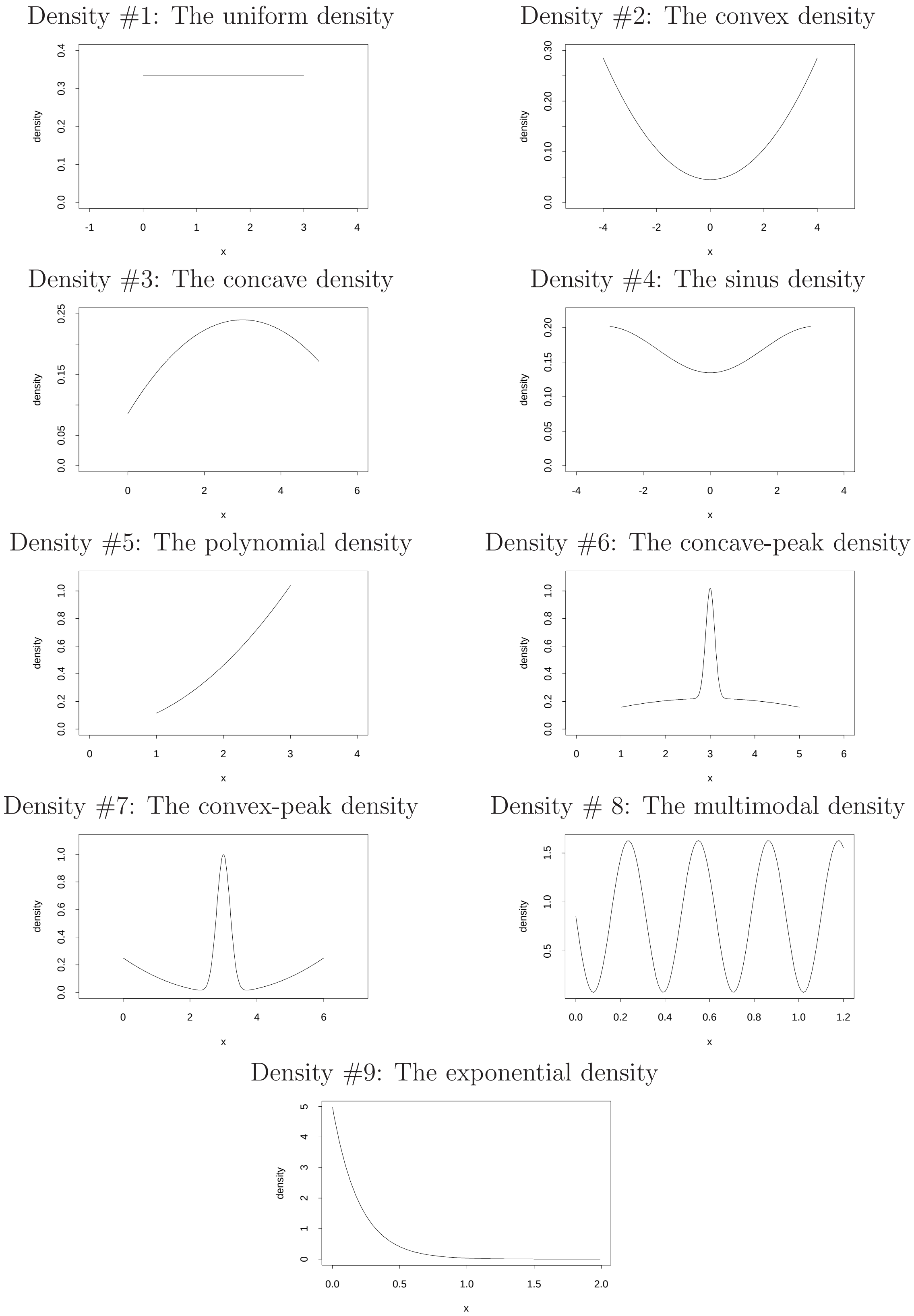


Figure 3.1: *Densities with a support different from the whole real line.*

3.1 The threshold method

We now illustrate the performance of the threshold method (2.1) presented in Section 2.1. For the bandwidth, we took either the theoretical optimal, defined via (2.2), or the practical derivative-based bandwidth $\widehat{h}_{\text{der}}$ defined via (2.3) and (2.5). We consider various values for the threshold parameter p: 0.2, 0.25, 0.3, 0.35 and 0.4.

Figures 3.2, 3.3 and 3.4 show, for the nine densities in Figure 3.1, boxplots of 100 replicated estimators using the theoretical asymptotic optimal bandwidth and the 5 values of p (the first group of five boxplots), or using the derivative-based bandwidth and the same values of p (the second group of five boxplots). We also present, in a first boxplot, the results for the naive (inconsistent) estimators $\min(Y_1, \ldots, Y_n)$ and $\max(Y_1, \ldots, Y_n)$. It should be mentioned that the theoretical asymptotic optimal bandwidth (see (2.2)) cannot be calculated for the uniform density (Density #1) since, for this density, the derivatives $f'_X(\tau\pm)$ are zero, and hence the asymptotic bias part is zero, leading to a degenerate minimization problem. For this density we therefore fixed the bandwidth to be the average of the best bandwidth (in Squared Error sense) over all simulations for producing the first group of five boxplots. This resulted in $h = 0.21$. From Figures 3.2, 3.3 and 3.4 we see that, depending on the case, different values of p are preferable, but from these results (and other non reported results) it seems that $p = 0.3$ works well 'on the average'. It is also clear that our method improves the results of the naive method.

When comparing the results for the two different choices of the bandwidth, we see that the derivative-based bandwidth performed reasonably well, and in some cases even better than the theoretical optimal bandwidth. In some cases, it leads to a biased estimator.

3.2 The stepwise estimation method

We now illustrate the data-driven stepwise procedure of Section 2.2 on the densities of Figure 3.1. We used the following stopping rule: stop at step i^* if $|\widehat{\tau}_{i^*} - \widehat{\tau}_{i^*-1}| < n^{-2/5}|\widehat{\tau}_1 - \widehat{\tau}_0|$. In Figures 3.5—3.7 we present, for all 100 simulations, the estimated values for the boundary point(s), and this for sample sizes $n = 100$ and $n = 250$. The results for left and right boundary points are shown in the same picture, where the horizontal axes represent the true values of the boundary points. These plots allow to have a more detailed view of the simulation results than the (summarizing) boxplots. As we can see, the stepwise method works quite nicely for all densities, including the difficult Densities #6 and #7. We also see the improvement when the sample size increases.

3.3 Comparison of implemented methods

In this section we compare the results obtained for the threshold method taking $p = 0.3$ and the derivative-based bandwidth $\widehat{h}_{\text{der}}$ (see Section 2.1) with these of the stepwise method (see Section 2.2). For each of the nine densities in Figure 3.1, we present the results of 100 simulations, for sample sizes $n = 50$, $n = 100$ and $n = 250$, in a boxplot. The first boxplot in each group of two boxplots summarizes the results for the threshold method (indicated by "m1"), and the second boxplot these for the stepwise method (indicated by "m2"). From Figures 3.8 and 3.9, we see that both methods perform quite well, and that none of the two methods can be declared as uniformly outperforming the other one. So, both methods can be recommended.

4 Application to estimation of discontinuities

4.1 The estimation method

The proposed method can also be adapted to estimation of a discontinuity point of a density function. The major difference between estimation of a boundary point and a discontinuity point is that in case of a boundary point we have some additional information, namely that we are at the endpoints of the range of the X-measurements. This information has been exploited in the threshold and the stepwise procedures as discussed in Section 2. Clearly, such information cannot be used when estimating a discontinuity point. Therefore, for discontinuity points, we follow the procedure of Gijbels, Hall and Kneip (1999) that they proposed in case of identification problems. For a single discontinuity point, the procedure reads as follows:

Step 0. Start with a large bandwidth h_0. Search for all local maxima $M_{1,0}, \cdots, M_{k,0}$ of the absolute value of the diagnostic function $\widehat{J}$;

Step 1. Using a bandwidth $h_1 < h_0$, search for the local maximizer $M_{j,1}$ of $|\widehat{J}|$ nearest to $M_{j,0}$, for $j = 1, \cdots, k$;

$\vdots$

Step i**.** Using a bandwidth $h_i < h_{i-1}$, search for the local maximizer $M_{j,i}$ of $|\widehat{J}|$ nearest to $M_{j,i-1}$, for $j = 1, \cdots, k$;

$\vdots$

Step s**.** At stage s, using the bandwidth $h_s < h_{s-1}$, find the local maximizer $M_{j,s}$ of $|\widehat{J}|$ nearest to $M_{j,s-1}$, for $j = 1, \cdots, k$, . Estimate the discontinuity point by $\widehat{\tau}_s = M_{\widehat{j},s}$, where $\widehat{j} = \text{argmax}_{j=1,2\cdots,k} |M_{j,s} - M_{j,0}|$.

The stopping rule is as in Section 2.2. Further we take $h_0 = \widehat{h}_{\text{der}}$, since, in this case, we need to start with a not too big bandwidth in order to have a certain number of local maxima at the start.

In case we need to estimate m (with $m > 1$) number of discontinuity points, we would consider in Step s the m largest values among the k values $|M_{j,s} - M_{j,0}|$.

4.2 Some simulation results

We now illustrate the finite sample performance of this data-driven method for estimating a discontinuity point. We consider two simulation models: a mixture of two exponential densities and a mixture of two normal densities. More precisely, let

$$\begin{aligned} &\text{Density \#10}: \quad f_X(x) = q\exp(x)\cdot 1_{\{x\leq 0\}} + (1-q)5\exp(-5x)\cdot 1_{\{x>0\}} \\ &\text{Density \#11}: \quad f_X(x) = \varphi_{0,\sigma_1^2}\cdot 1_{\{x\leq 0\}} + \varphi_{0,0.25}\cdot 1_{\{x>0\}}\ , \end{aligned} \tag{4.1}$$

where, for q, we consider the values $0, 0.25, 0.50$ and 0.75. For the normal density φ_{0,σ_1^2} we consider the values 1 and 2 for σ_1. These densities have also been considered by Chu and Cheng (1996) in their simulation study. Density #10 is a mixture of two exponential densities, with mixing proportion q. Note that for the extreme case $q = 0$ we again have a boundary estimation problem. Density #10 with $q = 0$ is exactly the same as Density #9. This means that the simulation results for this case will allow a comparison between the adapted procedure from Gijbels, Hall and Kneip (1999) and our data-driven procedure proposed for boundary points in Section 2.2. Figure 4.1 presents these mixtures of exponential densities for the considered values of the mixing proportion q. Clearly, the jump size of the discontinuity decreases with increasing values of q. Density #11 is a mixture of normal densities, and the two different mixtures are presented in Figure 4.2. The size of the jump increases with increasing values of σ_1.

To get a feeling of the difficulty of the estimation tasks for the various densities, we provide, in Figures 4.3 and 4.4, histograms of a sample of size $n = 2000$ drawn from each mixture of densities. We consider two situations: the error-free case, and the case where the samples are contaminated with a Laplace error with $\operatorname{Var} Z/\operatorname{Var} X = 10\%$. The distinction between these two cases allows us to really sense the extra difficulty of estimating a discontinuity point when there are measurement errors. In the latter case, the sample *seems* not to contain information about the discontinuity.

For each of the above densities, we simulated samples of sizes $n = 50, 100$ and $n = 250$ and applied the data-driven method for estimating the discontinuity point. Figure 4.5 summarizes simulation results for the mixtures of exponential densities. For each of the three non-zero values of q, we present boxplots for the three sample sizes. As we can see, the method adapted to this discontinuity detection context works quite nicely. The results are better for smaller values of q. For the special case $q = 0$, we have in fact a problem of boundary estimation. For comparison purpose, we present, in Figure 4.6, boxplots for results from the boundary estimation method of Section 2.2 as well as for the discontinuity estimation method of this section, and this for 100 simulated samples of sizes $n = 50, 100$ and 250. As can be seen from Figure 4.6, the boundary estimation method performs better, as expected, since it exploits the fact that we know that we are estimating a boundary point.

Figure 4.7 finally summarizes the simulation results for Density #11 for the two values of σ_1. For these mixtures of normal densities, the data-driven discontinuity estimation method works fine as well.

5 Application to frontier estimation

In this section, we apply our methodology of boundary estimation to the estimation of a frontier. This problem is generally multivariate, and we restrict to the bivariate case, i.e. to the case where the data of interest are observations of a couple of random variables (W, X). The variable W may for example represent the input or cost of a production, and X the output that is produced. The interest is to try to determine the maximal level of output that can be produced for a given level of input, i.e., for each value w of W, try to determine the right boundary point of the variable $X|W = w$. The aim is thus to estimate the boundary of the support of (W, X).

In most cases, the output X cannot be measured exactly and is observed with an error Z. In this problem, usually referred to as a stochastic frontier estimation problem, the n observations $(W_1, Y_1), \ldots, (W_n, Y_n)$ come from the couple of random variables (W, Y), where

$$Y = X + Z,$$

or this conditionally on $W = w$. If we further assume that X and Z are independent, we see that this problem may be viewed as a deconvolution problem.

Hall and Simar (2002) propose to decompose this bivariate problem of curve estimation into several univariate problems of estimation of boundary points. For m values $w_1, \ldots, w_m$ of the variable W, they create a subsample $\mathcal{Y}$ of the variable $Y|W = w$ in the following way: for $k = 1, \ldots, m$, take $\mathcal{Y}_k$ to be the sample of observations Y_i for which the W coordinate belongs to the interval $[w_k - h, w_k + h]$, with h a bandwidth to be chosen, i.e.

$$\mathcal{Y}_k = \{Y_i : W_i \in [w_k - h, w_k + h]\}.$$

Then, for $k = 1, \ldots, m$, search for the boundary point of the variable $Y|W = w_k$ using the univariate sample $\mathcal{Y}_k$. Finally, estimate the frontier curve by running a Nadaraya-Watson kernel smoother through the m estimated boundary points.

To illustrate the method, we use the data of Christensen and Greene (1976), concerning 123 electric utility industries. These data are also used by Hall and Simar (2002) for estimating the maximal level of output Q given the cost of production C. More precisely, they consider the bivariate data $(W, Y) = (\log C, \log Q)$ presented in Figure 5.1. In order to avoid edge effects, they restrict their study to the values of $\log C$ in the interval [1.5,5]. They estimate the $m = 50$ boundary points of the variables $Y|W = w_1, \ldots, Y|W = w_{50}$ where $w_1, \ldots, w_{50}$ are 50 equidistant values in the interval [1.5,5]. Finally the frontier curve is estimated by running a Nadaraya-Watson kernel smoother (with quartic kernel and bandwidth $h = 0.51$) through those 50 estimated boundary points.

This method used by Hall and Simar (2002) to estimate the 50 boundary points assumes that the error present in the data is small. In this type of problems, such an assumption seems realistic, since the noise-to-signal ratio is usually small. Below, we estimate the same 50 boundary points using our methodology, which does not require the error variance to be small, but requires the knowledge of the error density.

To simplify the problem, we assume that the error Z is independent of W. We will compare the obtained results assuming a Laplace error, with different values of the noise-to-signal ratio $\mathrm{NSR} = \mathrm{Var}(Z)/\mathrm{Var}(X|W = w)$. The variance of $X|W = w_k$ may be estimated through the empirical variance of $Y|W = w_k$. However, the sizes of the subsamples $\mathcal{Y}_k$ are very small (from 3 to 26), hence we rather estimate the variance of $Y|W = w_k$ by the empirical variance of Y. This estimator is of course inconsistent but is just used to give a rough idea of the variance of the error present in the data.

Figure 5.2 compares the results obtained assuming no error in the data (dashed curve) or a Laplace error, with NSR equal to 0.05, 0.15, or 0.3 (continuous line). Since the 50 subsamples $\mathcal{Y}_1, \dots, \mathcal{Y}_{50}$ have a small size (from 3 to 26), we use a stopping rule similar to Hall and Simar (2002): stop at step i^* if $|\widehat{\tau}_{i^*+1} - \widehat{\tau}_{i^*}| > |\widehat{\tau}_{i^*} - \widehat{\tau}_{i^*-1}|$ or if $|\widehat{\tau}_{i^*+1} - \widehat{\tau}_{i^*}| < 0.01$. The curves drawn were obtained using the Nadaraya-Watson estimator as described above. The $+$'s represent the observations. Note that the dashed curve is not exactly the same as that found by Hall and Simar (2002), because we did not apply the second-order correction they propose.

We see that assuming error in the data results in an estimator of the frontier which is almost always below the estimator obtained when assuming no error in the data. The difference is small but increases as the noise-to-signal ratio increases. Assuming Gaussian errors gave quite similar results (not reported here). In this example, since we have no information on the actual measurement error, it is hard to say which estimate is the best one, and if the differences between the curves are important or not. In this bivariate context, a simulation study could be helpful to see the improvement that can be expected by using this deconvolution kernel estimator compared to the methods classically used to estimate a stochastic frontier.

References

Carroll, R.J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, **83**, 1184–1186.

Carroll, R.J., Ruppert, D. and Stefanski, L. (1995). *Measurement error in nonlinear models.* Chapman and Hall, London.

Christensen, L. and Greene, R. (1976). Economies of scale in U.S. electric power generation. *Journal of Political Economy*, **84**, 655-676.

Chu, C.K. and Cheng, P.E. (1996). Estimation of jumps points and jump values of a density function. *Statistica Sinica*, **6**, 79–95.

Couallier, V. (1999). Estimation non paramétrique d'une discontinuité dans une densité. *Comptes Rendus de l'Académie des Sciences, I*, **t.329**, 633–636.

Delaigle, A. and Gijbels, I. (2002). Estimation of integrated squared density derivatives from a contaminated sample. *Journal of the Royal Statistical Society*, Series B, **64**, 869–886.

Delaigle, A. and Gijbels, I. (2003). Boundary estimation and estimation of discontinuity points in deconvolution problems. Institut de Statistique, Université catholique de Louvain, *Discussion Paper* # 0320.

Delaigle, A. and Gijbels, I. (2004a). Bootstrap bandwidth selection in kernel density estimation from a contaminated sample. *The Annals of the Institute of Statistical Mathematics*, **56**, 19–47.

Delaigle, A. and Gijbels, I. (2004b). Practical bandwidth selection in deconvolution kernel density estimation. *Computational Statistics and Data Analysis*, **45**, 249–267.

Fan, J. (1991a). Asymptotic normality for deconvolution kernel density estimators. *Sankhya A*, **53**, 97–110.

Fan, J. (1991b). Global behaviour of deconvolution kernel estimates. *Statistica Sinica*, **1**, 541–551.

Fan, J. (1991c). On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, **19**, 1257–1272.

Fan, J. (1992). Deconvolution with supersmooth distributions. *The Canadian Journal of Statistics*, **20**, 155–169.

Gijbels, I. and Goderniaux, A.-C. (2004). Bandwidth selection for change point estimation in nonparametric regression. *Technometrics*, **46**, 76–86.

Gijbels, I., Hall, P. and Kneip, A. (1999). On the estimation of jump points in smooth curves. *Annals of the Institute of Statistical Mathematics*, **51**, 231–251.

Hall, P. and Simar, L. (2002). Estimating a changepoint, boundary or frontier in the presence of observation error. *Journal of the American Statistical Association*, **97**, 523–534.

Mendelsohn, J. and Rice, J. (1982). Deconvolution of microfluorometric histograms with B splines. *Journal of the American Statistical Association*, **77**, 748–753.

Stefanski, L. and Carroll, R.J. (1990). Deconvoluting kernel density estimators. *Statistics*, **2**, 169–184.

Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing*. Chapman and Hall, London.

Wu, J.S. and Chu, C.K. (1993). Kernel-type estimators of jump points and jump values of a regression function. *The Annals of Statistics*, **21**, 1545–1566.

Zhang, S. and Karunamuni, R. (2000). Boundary bias correction for nonparametric deconvolution. *Annals of the Institute of Statistical Mathematics*, **52**, 612–629.

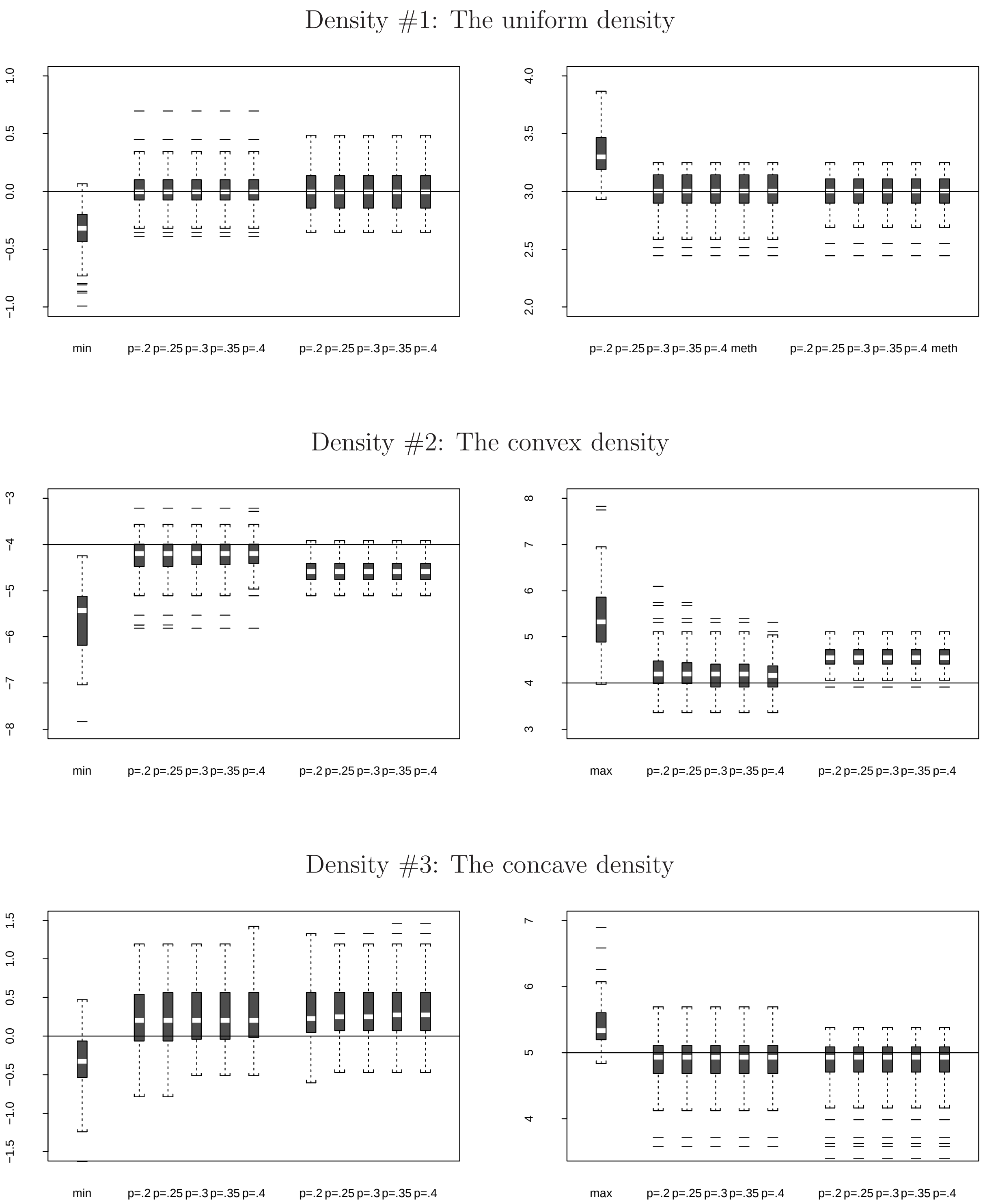


Figure 3.2: *Comparison of different thresholds with the optimal bandwidth (first group of five boxplots) or the derivative-based bandwidth (second group of five boxplots) for Laplace contaminated samples of size* $n = 100$*, with* $\operatorname{Var} Z / \operatorname{Var} X = 10\%$*, for the left endpoint (left picture) or the right endpoint (right picture) for Densities #1,2 and 3 (top panels, center panels and bottom panels). The first boxplot in each picture shows results for the naive (inconsistent) method.*

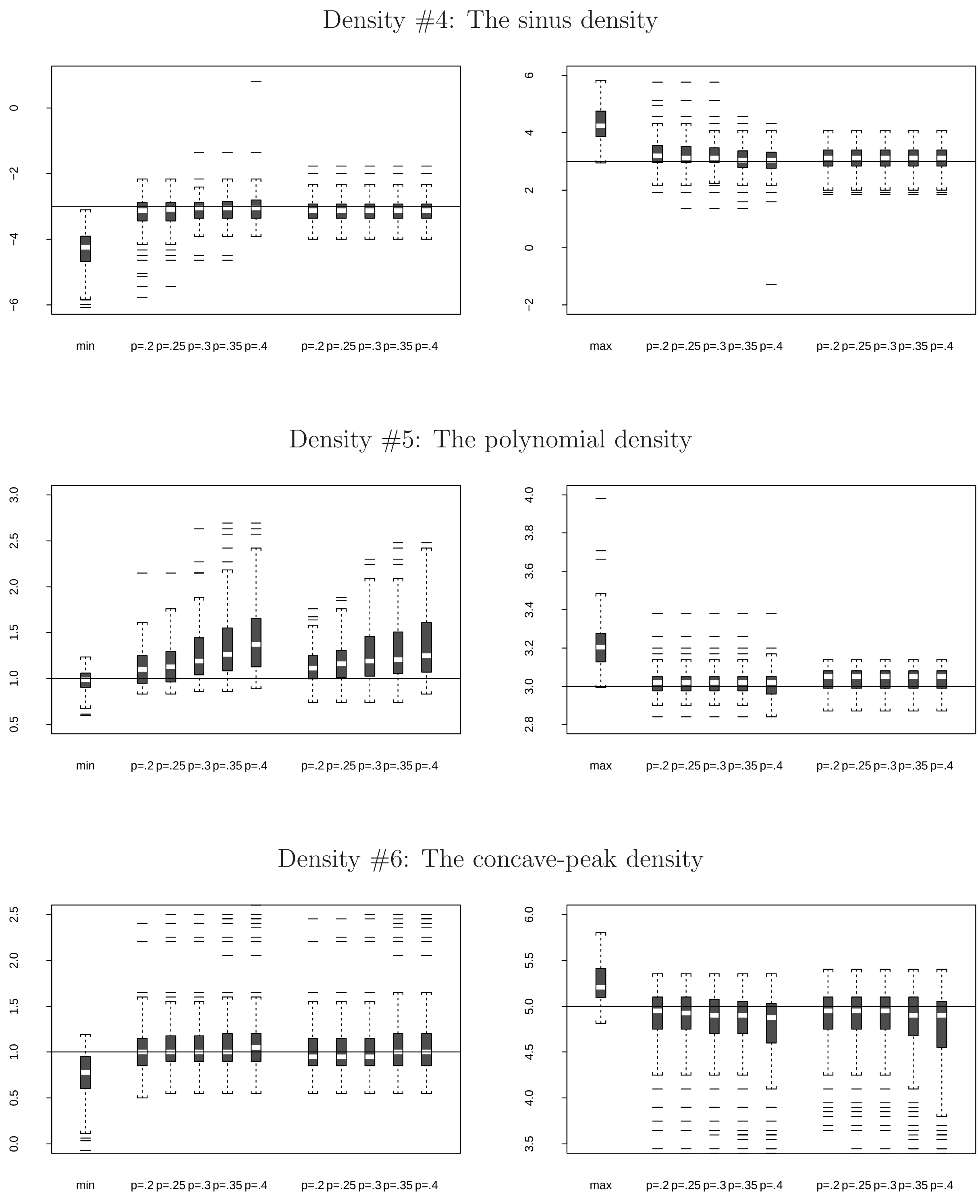


Figure 3.3: *Comparison of different thresholds with the optimal bandwidth (first group of five boxplots) or the derivative-based bandwidth (second group of five boxplots) for Laplace contaminated samples of size* $n = 100$*, with* $\operatorname{Var} Z/\operatorname{Var} X = 10\%$*, for the left endpoint (left picture) or the right endpoint (right picture) for Densities #4,5 and 6 (top panels, center panels and bottom panels). The first boxplot in each picture shows results for the naive (inconsistent) method.*

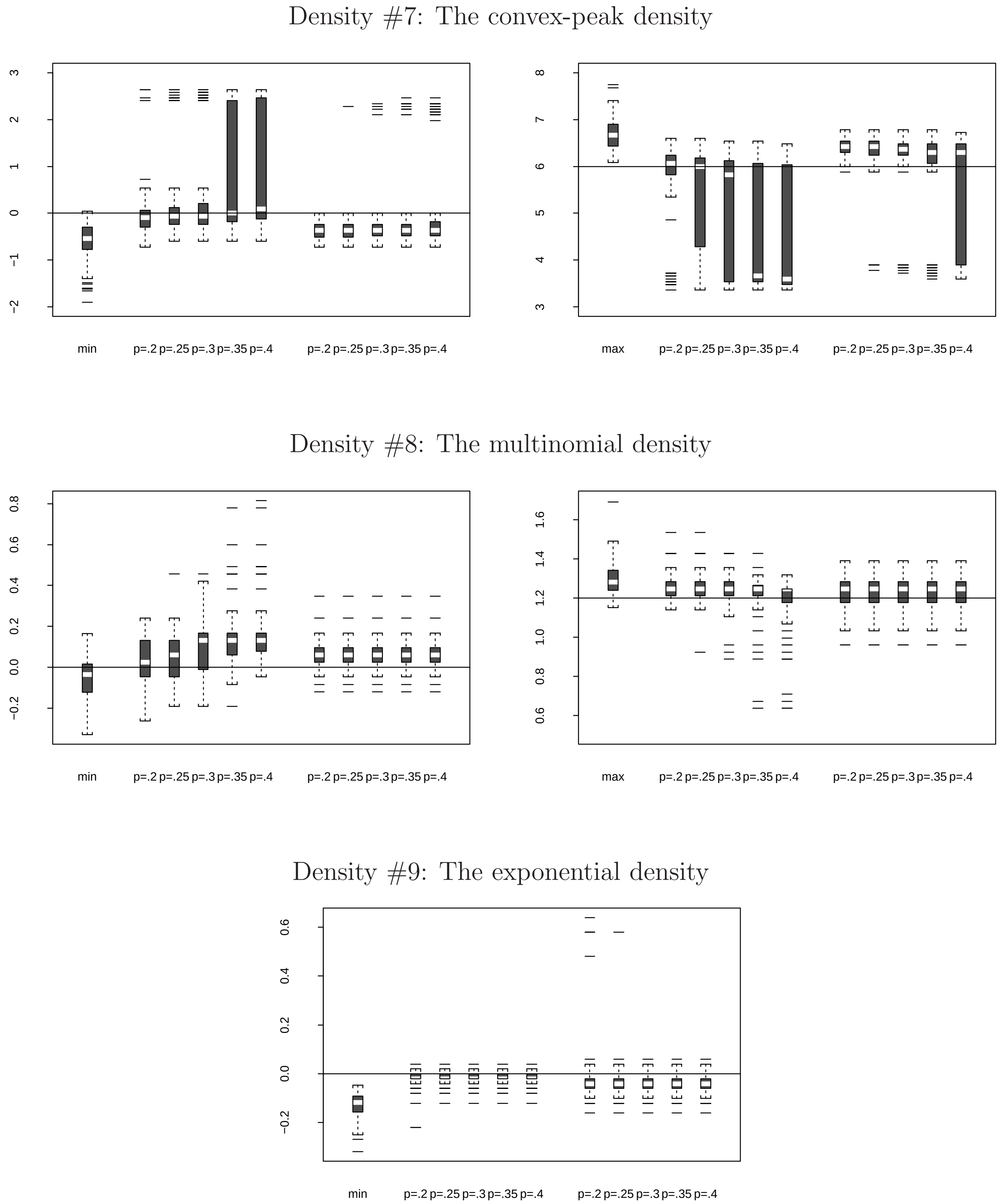


Figure 3.4: *Comparison of different thresholds with the optimal bandwidth (first group of five boxplots) or the derivative-based bandwidth (second group of five boxplots) for Laplace contaminated samples of size* $n = 100$*, with* $\operatorname{Var} Z/\operatorname{Var} X = 10\%$*, for the left endpoint (left picture) or the right endpoint (right picture) for Densities #7,8 and 9 (top panels, center panels and bottom picture). The first boxplot in each picture shows results for the naive (inconsistent) method.*

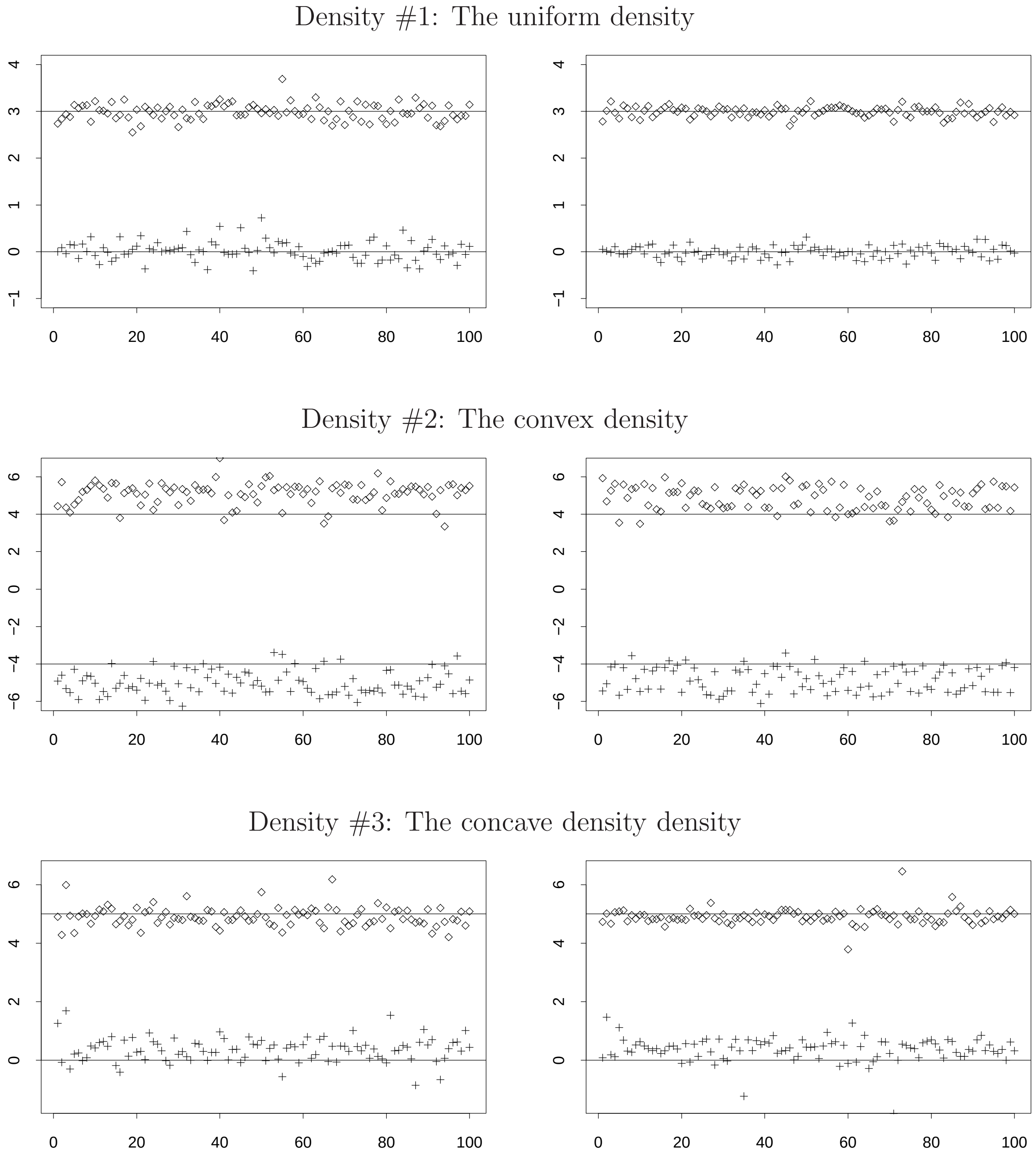
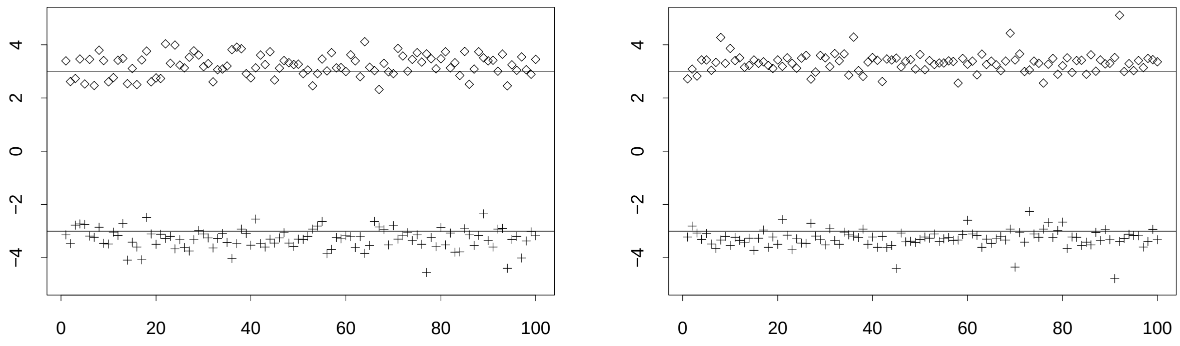
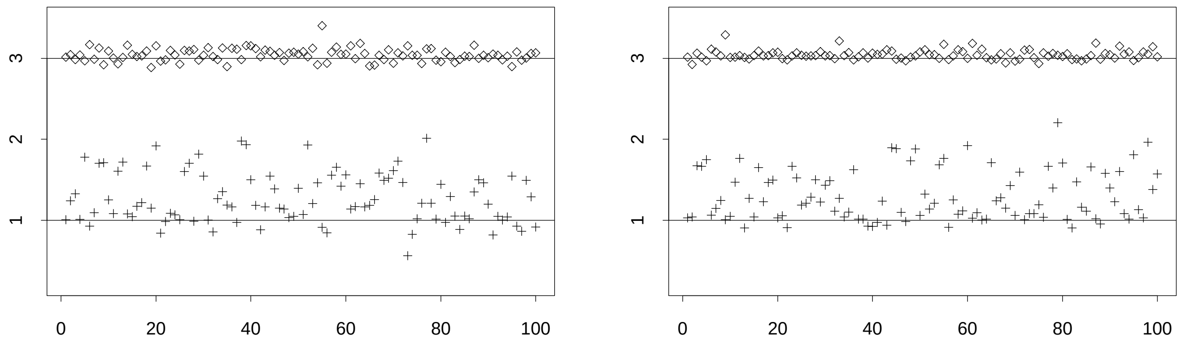


Figure 3.5: *Data-driven stepwise estimation of the endpoints of a density from samples of size* $n = 100$ *(left panels) or* $n = 250$ *(right panels) for Laplace contaminated samples, with* $\operatorname{Var} Z / \operatorname{Var} X = 10\%$*. The points indicated with the character "◇" present the results for the right boundary point and these indicated with "+" for the left boundary point. Results for Densities #1—#3.*

Density #4: The sinus density



Density #5: The polynomial density



Density #6: The concave-peak density

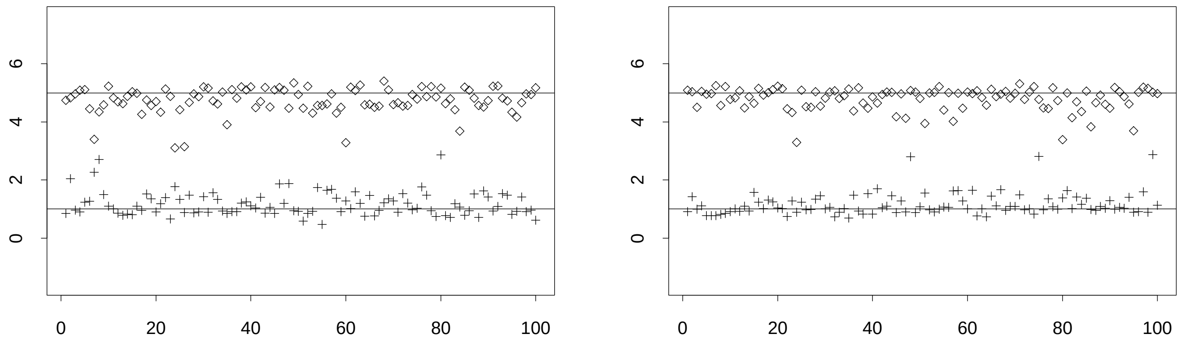


Figure 3.6: *Data-driven stepwise estimation of the endpoints of a density from samples of size* $n = 100$ *(left panels) or* $n = 250$ *(right panels) for Laplace contaminated samples, with* $\operatorname{Var} Z / \operatorname{Var} X = 10\%$*. The points indicated with the character "◇" present the results for the right boundary point and these indicated with "+" for the left boundary point. Results for Densities #4—#6.*

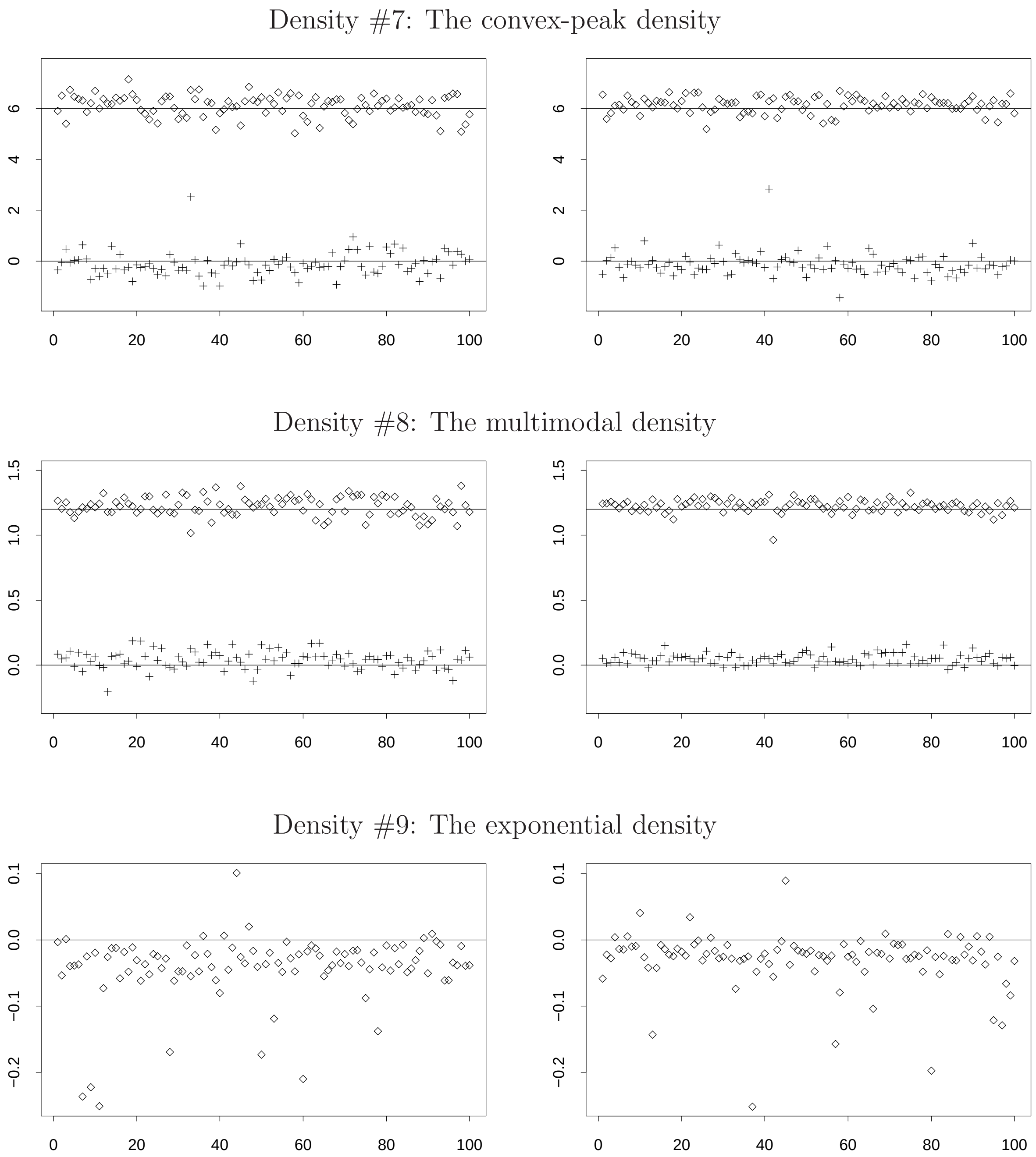


Figure 3.7: *Data-driven stepwise estimation of the endpoints of a density from samples of size* $n = 100$ *(left panels) or* $n = 250$ *(right panels) for Laplace contaminated samples, with* $\operatorname{Var} Z / \operatorname{Var} X = 10\%$*. The points indicated with the character "◇" present the results for the right boundary point and these indicated with "+" for the left boundary point. Results for Densities #7—#9.*

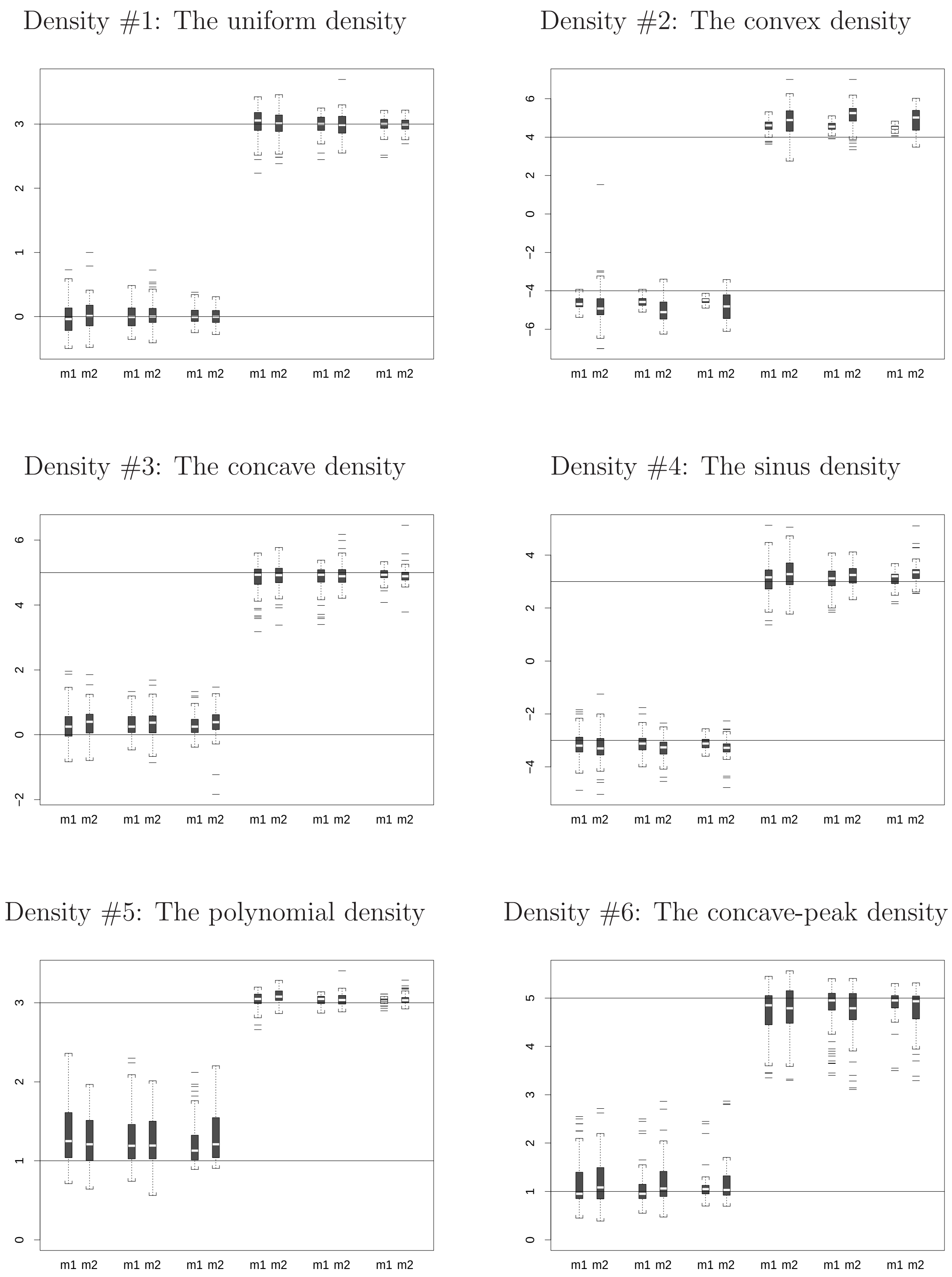


Figure 3.8: *Comparison of the two procedures: the threshold method, using* $p = 0.3$*, with bandwidth* $\widehat{h}_{der}$ *(method m1) and the data-driven procedure (method m2), for samples of size* $n = 50, 100$ *and* 250 *for Laplace contaminated samples, with* $\operatorname{Var} Z / \operatorname{Var} X = 10\%$*. The first (respectively second) boxplot in a group of two summarizes the results for method m1 (respectively method m2). The first, second, third group of two boxplots are for respectively sample sizes 50, 100 and 250. Results for Densities #1—#6.*

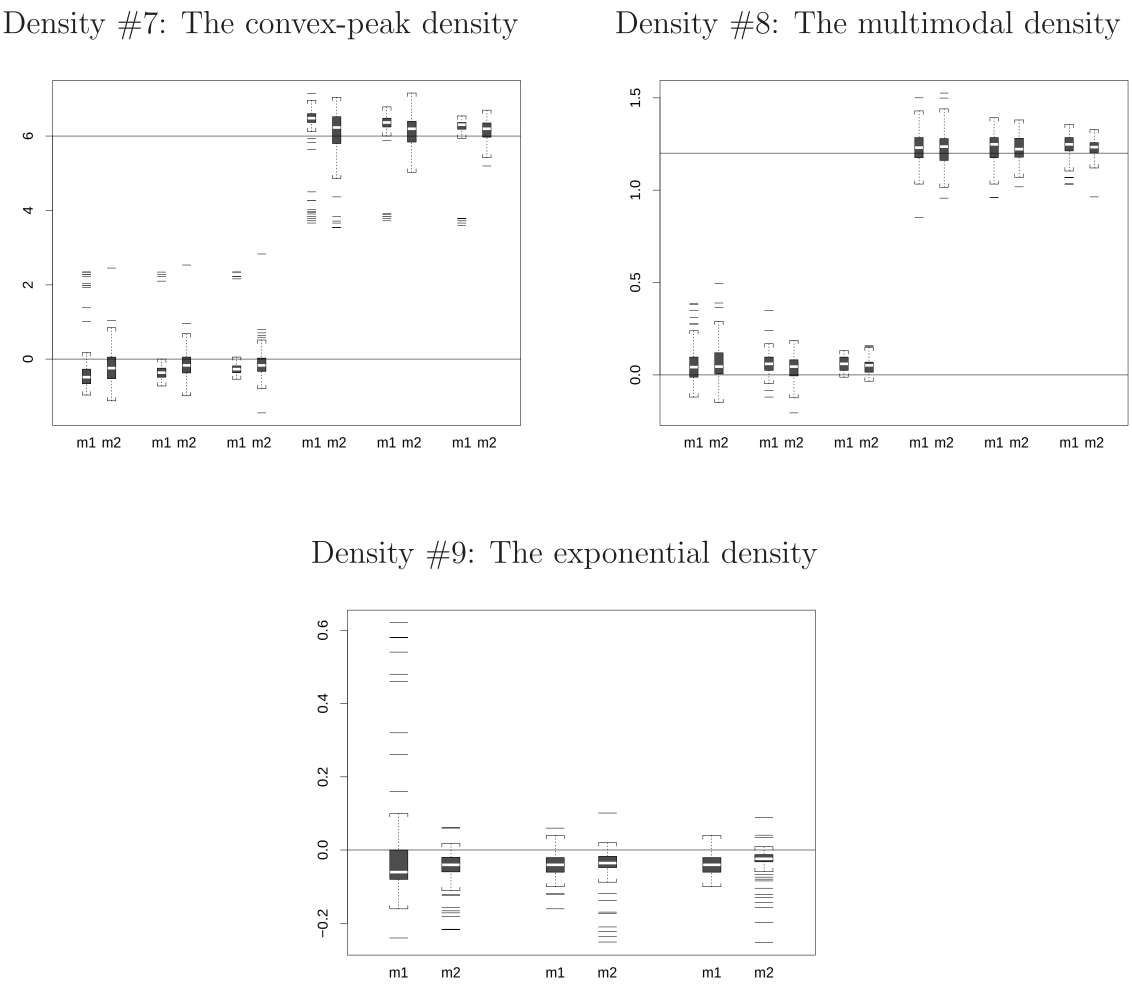


Figure 3.9: *Comparison of the two procedures: the threshold method, using* $p = 0.3$, *with bandwidth* $\widehat{h}_{der}$ *(method m1) and the data-driven procedure (method m2), for samples of size* $n = 50, 100$ *and* 250 *for Laplace contaminated samples, with* $\operatorname{Var} Z / \operatorname{Var} X = 10\%$. *The first (respectively second) boxplot in a group of two summarizes the results for method m1 (respectively method m2). The first, second, third group of two boxplots are for respectively sample sizes 50, 100 and 250. Results for Densities #7—#9.*

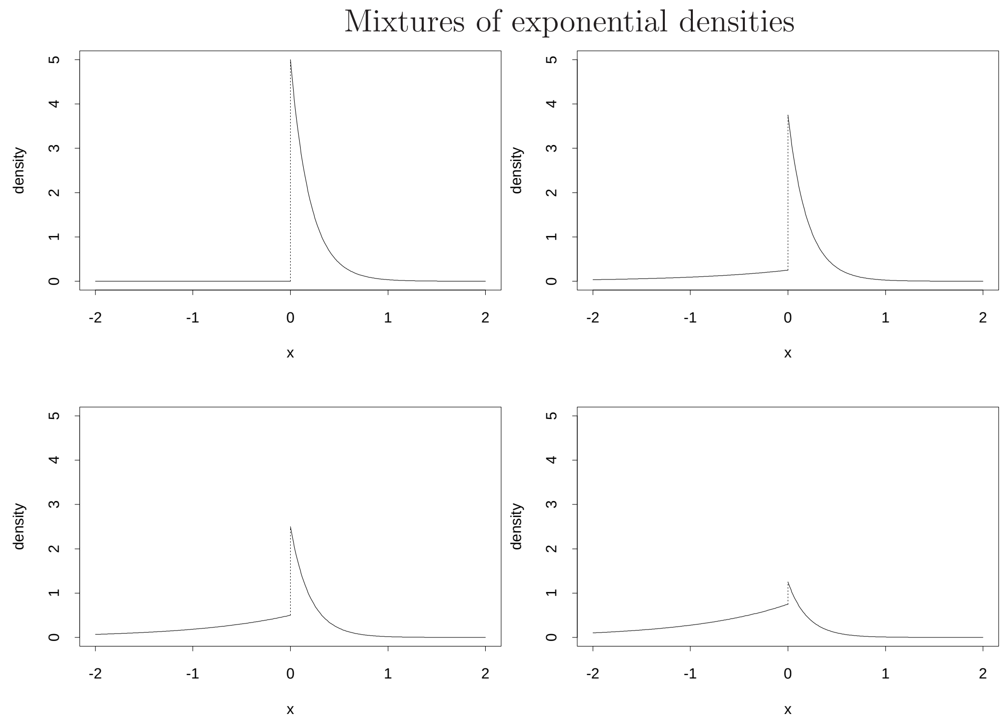


Figure 4.1: *Mixtures of exponential densities as defined in (4.1). From left to right and from top to bottom: mixing proportion* $q = 0, 0.25, 0.5$ *and* 0.75.

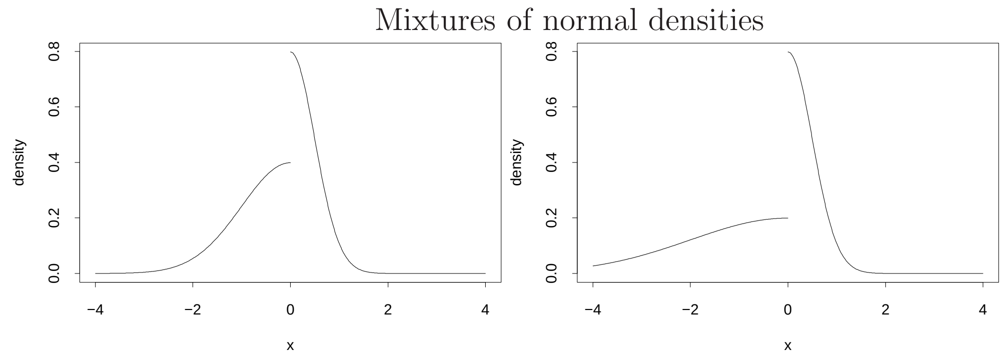


Figure 4.2: *Mixtures of normal densities as defined in (4.1). Left panel:* $\sigma_1 = 1$*, Right panel:* $\sigma_1 = 2$.

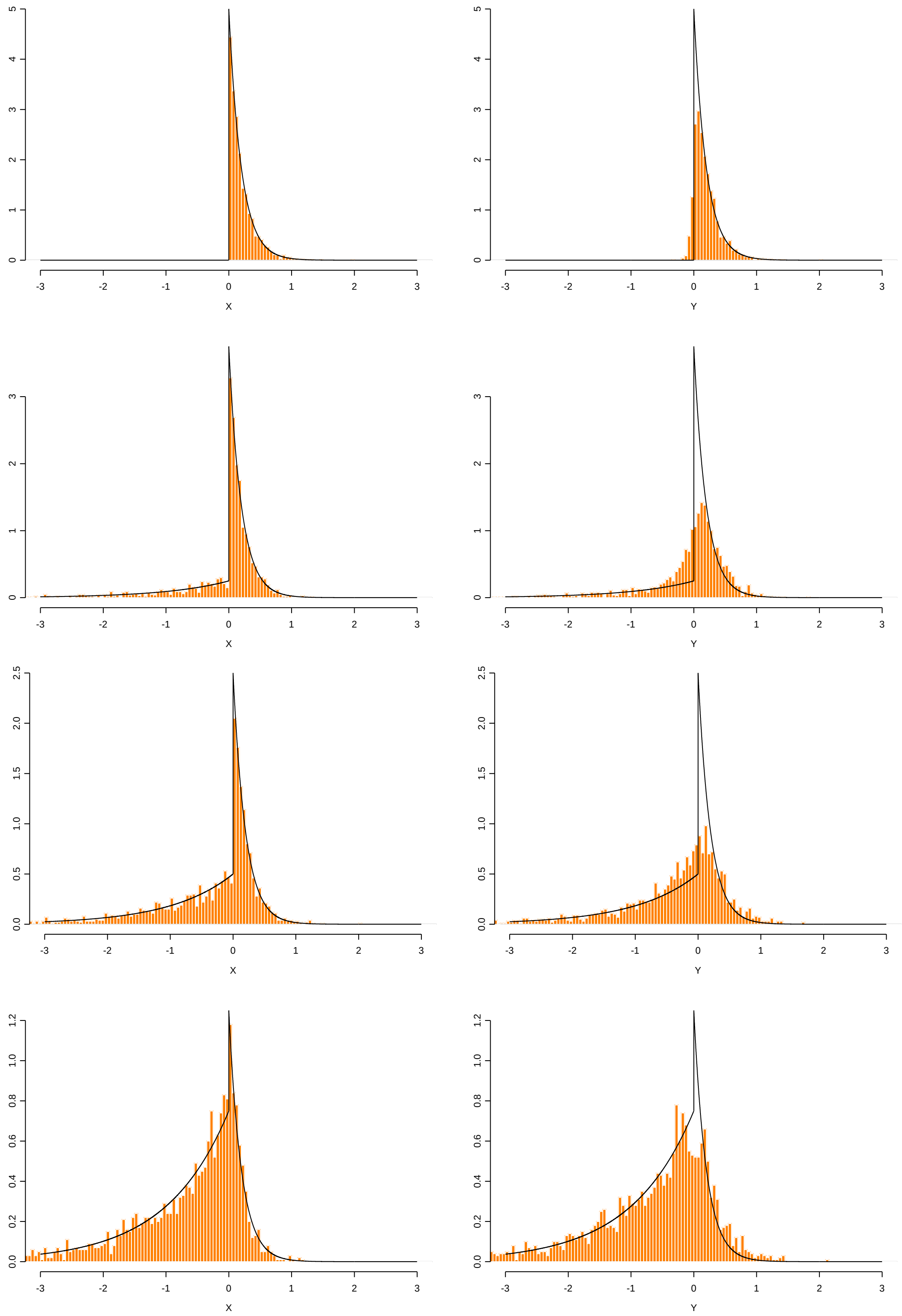


Figure 4.3: *Histograms of samples of size* $n = 2000$ *drawn from the mixtures of exponential densities with* $q = 0, 0.25, 0.50$ *and* 0.75 *(from top to bottom panels), either without measurement error (left side) or contaminated by Laplace error with* $\operatorname{Var} Z / \operatorname{Var} X = 10\%$ *(right side). The true densities are superimposed for comparison.*

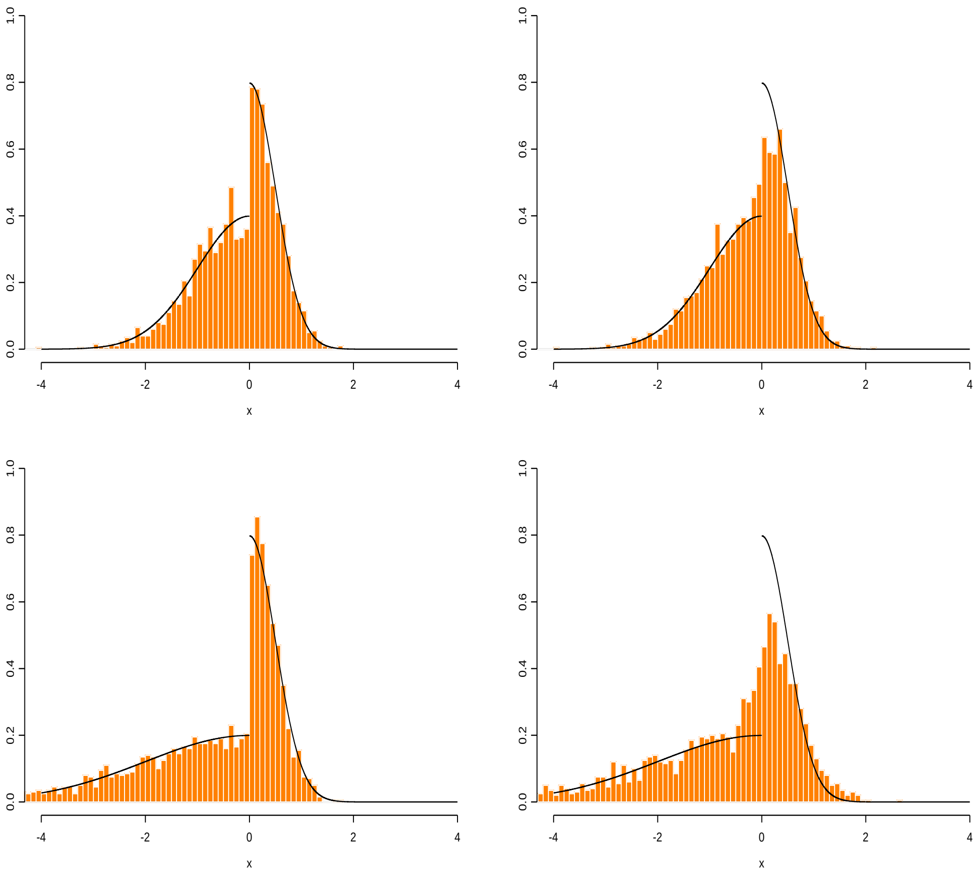


Figure 4.4: *Histograms of samples of size* $n = 2000$ *drawn from the mixtures of normal densities with* $\sigma_1 = 1$ *(top panels) or* $\sigma_1 = 2$ *(bottom panels), either without measurement error (left side) or contaminated by Laplace error with* $\operatorname{Var} Z / \operatorname{Var} X = 10\%$ *(right side). The true densities are superimposed for comparison.*

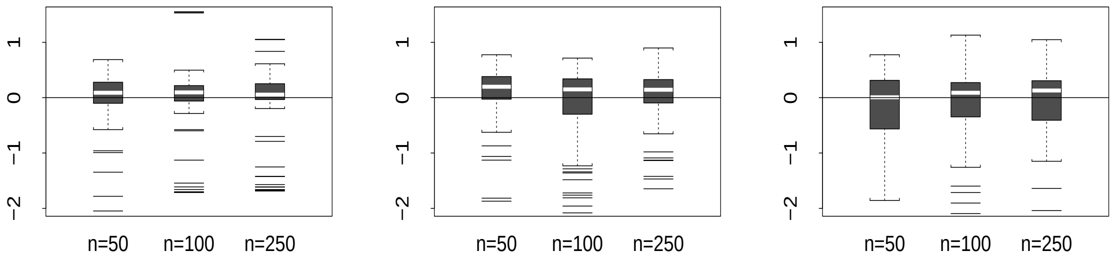


Figure 4.5: *Simulation results for the mixtures of exponential densities with* $q = 0.25, 0.50$ *and* $q = 0.75$ *(left, middle and right) for sample sizes* $n = 50, 100$ *and* 200 *(first, second and third boxplot) contaminated by Laplace error with* $\operatorname{Var} Z / \operatorname{Var} X = 10\%$.

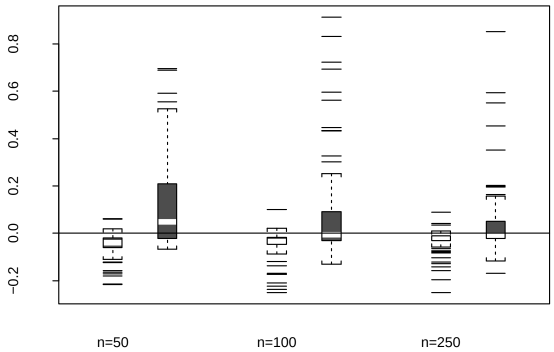


Figure 4.6: *Simulation results for the exponential density ($q = 0$), for samples of sizes $n = 50, 100$ and 250 (first, second and third group of boxplots) contaminated by Laplace error with $\operatorname{Var} Z/\operatorname{Var} X = 10\%$. Results for the boundary estimation method of Section 2.2 (first boxplot) and the data-driven discontinuity estimation method of Section 4 (second boxplot).*

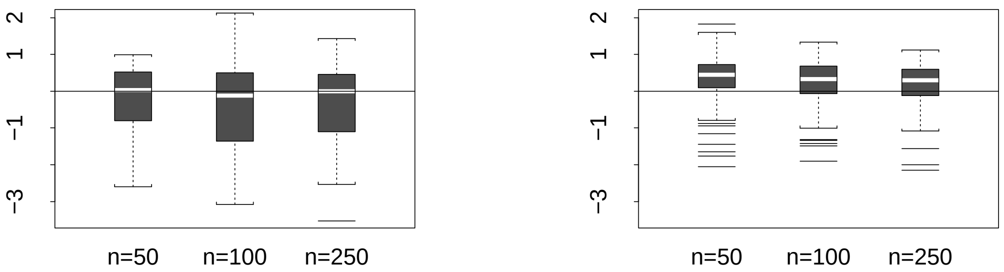


Figure 4.7: *Simulation results for the mixtures of normal densities for sample sizes $n = 50, 100$ and 250 (first, second and third boxplot) contaminated by Laplace error with $\operatorname{Var} Z/\operatorname{Var} X = 10\%$. Left panel: $\sigma_1 = 1$. Right panel: $\sigma_1 = 2$.*

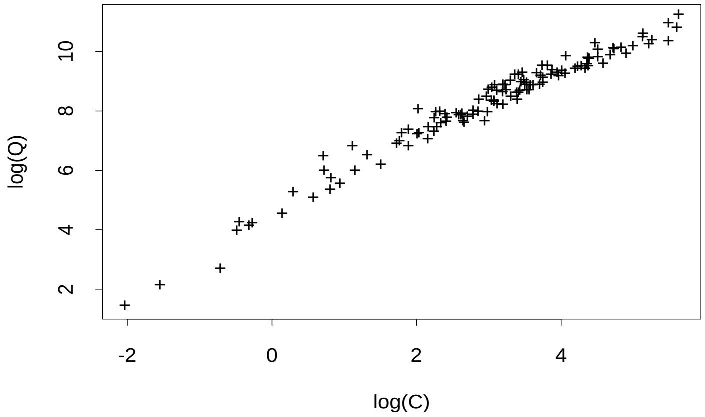


Figure 5.1: *Electric utility data: cost (C) and output (Q).*

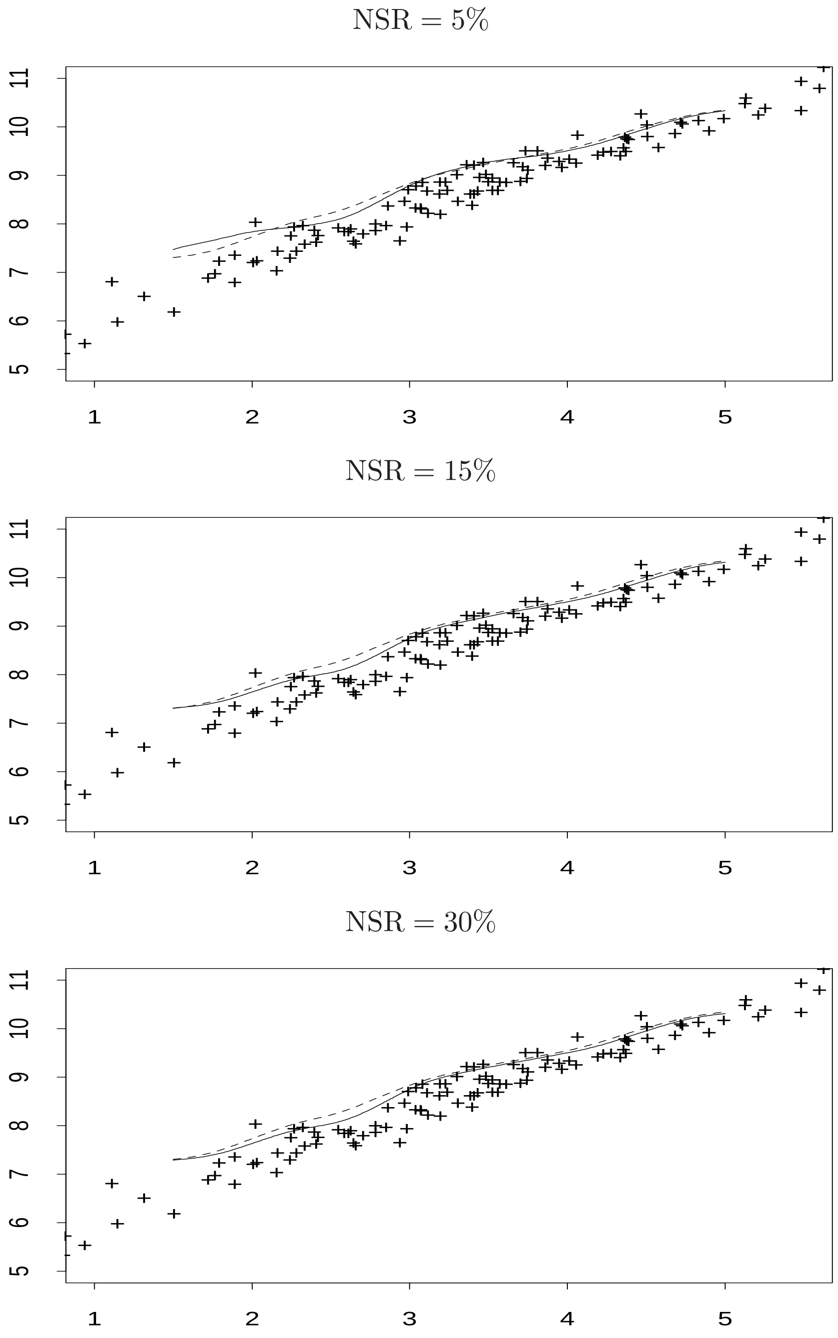


Figure 5.2: *Estimation of the frontier of the electric utility data, if we assume that the error is Laplace with different values of the noise-to-signal ratio NSR. The dashed curve represents the frontier estimation assuming no error in the data. The +'s represent the observations.*