CHAPTER 1

Density Estimation

The probability density function (PDF), f, is the statistical distribution of a population that, once determined, allows for the maximum extraction of statistical information about the population. Sometimes the probability density function (PDF) - usually just referred to as the density or density function - of the population can be deduced from the nature of the statistical experiment, however this is not always the case. When the population density is unknown it can be estimated parametrically or non-parametrically. In parametric density estimation data is assumed to follow a given distribution. This distribution is estimated by estimating the relevant parameters for the distribution - e.g. the mean and variance for the normal distribution. In non-parametric density estimation data is allowed to speak for itself and initially no assumptions, besides there existing a PDF, is made. In contrast to the parametric case the non-parametric case does not attempt to estimate parameters of the PDF, but rather the PDF itself. The naive approach is to consider what is called the empirical cumulative distribution function (CDF) defined as follows [**Scott2015**]

$$\hat{F}(\vec{x}) \equiv \frac{\#\{\vec{x} \leq \vec{x}\}}{n}, < \quad \forall \vec{x} \in \mathcal{R}^d. \tag{1}$$

From the CDF the density can be estimated as follows

$$\hat{f}(\vec{x}) \equiv \frac{\partial^d \hat{F}(\vec{x})}{\partial x_1 \partial x_2 \dots \partial x_d} = \frac{1}{n} \sum_i \delta^d(\vec{x} - \vec{x}_i). \tag{2}$$

However - as is evident from equation (2) - this density estimate is a distribution of delta functions and as such worth little in terms of representing a true density that is continuous. The empirical density estimate is a direct representation of the data sample. For this reason the variability of the estimate between different samples from the same population is large (i.e. the variance is large). In order to obtain a better estimate - which varies less between different samples of the same population - information external to the sample has to be applied. Such information can be requiring continuity of the density, the support of the density or expectation of the rough shape of the density (e.g. based on reviewing a histogram). Assuming the true density is continuous, some amount of smoothing (with respect to the empirical estimate) has to be applied. The smoothing action can be implemented in a many different ways, many of which are useful for facilitating a discussion and fewer of which are in the end expedient to use in estimating the true density.
This report will focus on univariate (section (1.1)) and multivariate (section 1.2) density estimation using weight function estimators and especially kernel estimators. It will be of particular interest to derive rules of thumb for the parameters of kernel estimation and test the precision of these in numerical experiments. Since there are many relevant and illuminating numerical experiments to considered, the report is structured such that these are located in the text as examples where relevant. The idea behind this is to have the experiment in the vicinity of the relevant text and not have a disjoint pile of experiments in the end.

1.1 UNIVARIATE DENSITY ESTIMATION

A first stab at implementing the smoothing action is to apply the two-sided numerical derivative instead of the derivative operator in equation (2). By doing so the (univariate) density estimate is given by

$$\hat{f}(x) = \frac{\hat{F}(x + \frac{h}{2}) - \hat{F}(x - \frac{h}{2})}{h}, \tag{3}$$

where h is a smoothing parameter.

Example 1.1.
Figure 1 shows a histogram and plots of the CDF and PDF - using equation (1) *and* (3)*, respectively - of the treatment data from* **Silverman86**.

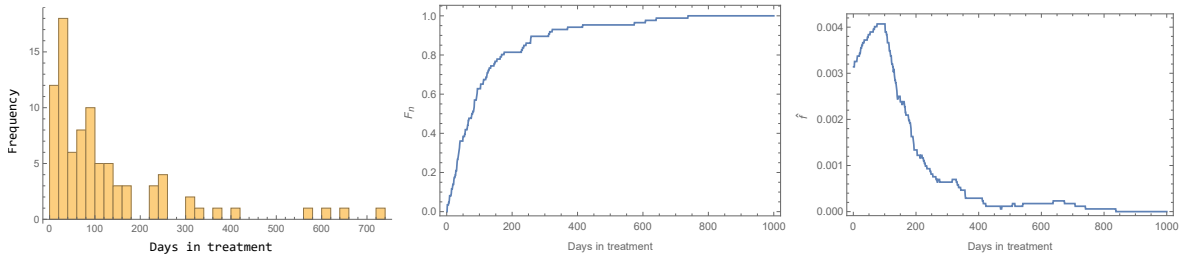

Figure 1: Left: Histogram of the treatment data (50 bins). Middle: Empirical CDF of the treatment data. Right: Density estimate computed via equation (3) with $h = 200$ chosen subjectively.

The above example illustrates the smoothing action of the two-sided numerical derivative relative to the derivative operator (a collection of delta functions). The smoothing by the two-sided derivative is equivalent to sliding a rectangular kernel (of width h) across the data and computing the average within the rectangle as it moves across data. This approach can be generalized to the class of estimators called general weight function estimators; let $w(x, y)$ be a function which satisfy

$$\int_{-\infty}^{\infty} dy w(x, y) = 1 \tag{4}$$

with

$$w(x, y) \geq 0 \quad \forall x, y. \tag{5}$$

The general weight function estimators are then defined, in terms of w, viz **[Silverman86]**

$$\hat{f}(t) = \frac{1}{n} \sum_{i=1}^{n} w(x_i, t). \tag{6}$$

The degree and implementation of the smoothing action is controlled by the function w. Table 1 list some common classes of w.

Estimator	w
Histogram	$w(x,y) = \begin{cases} h^{-1}(x) & \text{if x and y fall in the same bin} \\ 0 & \text{Otherwise} \end{cases}$
Orthogonal series	$w(x,y) = \sum_{\nu=0}^{\Lambda} \phi_\nu(x)\phi_\nu(y)$
Kernel	$w(x,y) = \frac{1}{h} K\left(\frac{y-x}{h}\right)$
Generalized k'th nearest neighbor	$w(x,y) = \frac{1}{d_k(t)} K\left(\frac{y-x}{d_k(t)}\right)$

Table 1

In relation to table 1; Λ is a cutoff that determines the amount of smoothing and $d_k(t)$ is the Euclidean distance to the k'th nearest point from t. Each weight function smooths the data in a different way and has different strengths/weaknesses. The histogram provides an easy, intuitive representation of data, however, this estimate contains discontinuities and has a high dependence on the sampled data. Hence, as far as estimating the true density, the histogram is rather wanting. The orthogonal series method estimates the probability density via developing it from a set of basis functions. The basis functions depend on the support of the density function. In general, when the real line $(-\infty, \infty)$ or half the real line $[0, \infty)$ are the support, Hermite and Laguerre series are recommended **[Efromovich]**. If, on the other hand, the density function has a compact support $[0, 1]$, then a Fourier series can be

used [**Silverman86**]. The drawback of the orthogonal series method is that the density estimate may not be a bona fide probability density, meaning that it may not integrate to unity and it can possibly take on negative values [**Efromovich**]. The kernel estimate and the generalized k'th nearest neighbor estimate are related in that both estimates utilize kernels (see table 2 for some common [**Silverman86**] univariate kernels[1]). The kernels are usually assumed to have continous derivatives to all required orders and to be symmetric functions satisfying [**Silverman86**]

$$\int dtK(t) = 1, \qquad \int dttK(t) = 0 \quad \text{and} \quad \int dtt^2K(t) = k_2 \neq 0. \tag{7}$$

From table 1 it is clear that the difference between the kernel estimate and the k'th nearest neighbor estimate lies in the window width applied. For the former a fixed window width is applied whereas for the latter a position dependent window width is used. The fixed window width is less flexible as compared to the position dependent one. The danger associated with increased flexibility is, as discussed previously, that the variance of the estimate can become too high, meaning that the results depend too much on the particular sample of data and by extension the estimate of the true density suffers. The two methods can be combined into the sample-point[2] adaptive estimate. This estimate consists of a series of steps [**silverman**];

1. Develop a pilot estimate $\tilde{f}$ for which $\tilde{f}(x_i) > 0$.
2. Define a local bandwidth, λ_i viz

$$\lambda_i \equiv \left(\frac{g}{\tilde{f}(x_i)}\right)^{\alpha}, \tag{8}$$

   where

$$g = e^{\frac{1}{n}\sum_i ln(\tilde{f}(x_i))} \tag{9}$$

   is the geometric mean of $\tilde{f}$ and $0 \leq \alpha \leq 1$ is a sensitivity parameter[3].
3. The adaptive estimate is then (for the univariate case) defined by letting $h \to h\lambda_i$ in the kernel estimate. That is

$$\hat{f}(t) = \frac{1}{n}\sum_i \frac{1}{h\lambda_i} K\left(\frac{t - x_i}{h\lambda_i}\right). \tag{10}$$

The adaptive estimate is relatively insensitive of the pilot estimate and so it is expedient to use a pilot estimate that provides both low computation time and analytical complexity.

Kernel	$K(t)$	Efficiency
Epanechnikov	$\begin{cases} \frac{3}{4\sqrt{5}}(1 - \frac{1}{5}t^2) & \text{for } \lvert t\rvert < \sqrt{5} \\ 0 & \text{Otherwise} \end{cases}$	1
Biweight	$\begin{cases} \frac{15}{16}(1 - t^2)^2 & \text{for } \lvert t\rvert < 1 \\ 0 & \text{Otherwise} \end{cases}$	0.9939
Triangular	$\begin{cases} 1 - \lvert t\rvert & \text{for } \lvert t\rvert < 1 \\ 0 & \text{Otherwise} \end{cases}$	0.9859
Gaussian	$\frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}}$	0.9512
Rectangular	$\begin{cases} \frac{1}{2} & \text{for } \lvert t\rvert < 1 \\ 0 & \text{Otherwise} \end{cases}$	0.9295

Table 2

[1] The meaning and derivation of "Efficiency" in Table 2 is covered later.
[2] In the sample point adaptive estimators $h \to h(x_i)$. Alternatively one could let $h \to h(t)$, which would result in another class of adaptive estimators (not considered in this study).
[3] $\alpha = 0$ returns the pilot estimate whereas $\alpha = 1$ is the generalized nearest neighbor estimate.

Example 1.2.

The treatment data from **Silverman86** *denotes the number of days a set of test persons spent in treatment. Hence, the possible values are positive integers. The true density will reflect this parameter space and vanish below* 1*. In order to make an accurate density estimate the kernel should reflect this property as well. Naively applying the kernels of Table 2 will result in a density estimate yielding a finite probability of negative days in treatment. The issue can be alleviated by transforming data such that the distribution more closely resembles the kernel function. After the transformation the data can be transformed back and so a density estimate featuring this central property of the true density can be obtained in this way. The sequence can be captured in one operation by using*

$$f(t)dt = f(a(t))da(t) \Rightarrow f(t) = f(a)d_t a, \tag{11}$$

where a *is any transformation of the data, e.g.* $a(t) = ln(t)$*. Note however that both* $t \to a(t)$ *and* $x_i \to a(x_i)$ *in the transformation. For the treatment data it is expedient to perform a log-transformation of data. This results - for the Gaussian, Epanechnikov and rectangular kernels - in the density estimates*

$$\begin{aligned}
&\textit{Gaussian kernel:} \quad \hat{f}(t) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{ht}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{ln(t)-ln(x_i)}{h}\right)^2}, \\
&\textit{Epanechnikov kernel:} \quad \hat{f}(t) = \frac{1}{n}\sum_{i=1}^{n}\begin{cases}\frac{1}{ht}\frac{3}{4\sqrt{5}}\left(1-\frac{1}{5}\left(\frac{ln(t)-ln(x_i)}{h}\right)^2\right) & \textit{for } \left|\frac{ln(t)-ln(x_i)}{h}\right| < \sqrt{5} \\ 0 & \textit{Otherwise}\end{cases}, \\
&\textit{Rectangular kernel:} \quad \hat{f}(t) = \frac{1}{n}\sum_{i=1}^{n}\begin{cases}\frac{1}{2ht} & \textit{for } \left|\frac{ln(t)-ln(x_i)}{h}\right| < 1 \\ 0 & \textit{Otherwise}\end{cases}.
\end{aligned} \tag{12}$$

The density estimates using the Gaussian, Epanechnikov and rectangular kernels as well as the log-transformed estimates of the three and the adaptive Gaussian kernel are shown in figure 2. From the figure it is clear how the non-log-transformed estimates predict a non-vanishing probability for negative days in treatment and how this is not the case for the log-transformed estimates. From the figure it is also clear that the density estimates using different kernels are quite similar, although the density estimates using the rectangular kernel window is noisy relative to the others.

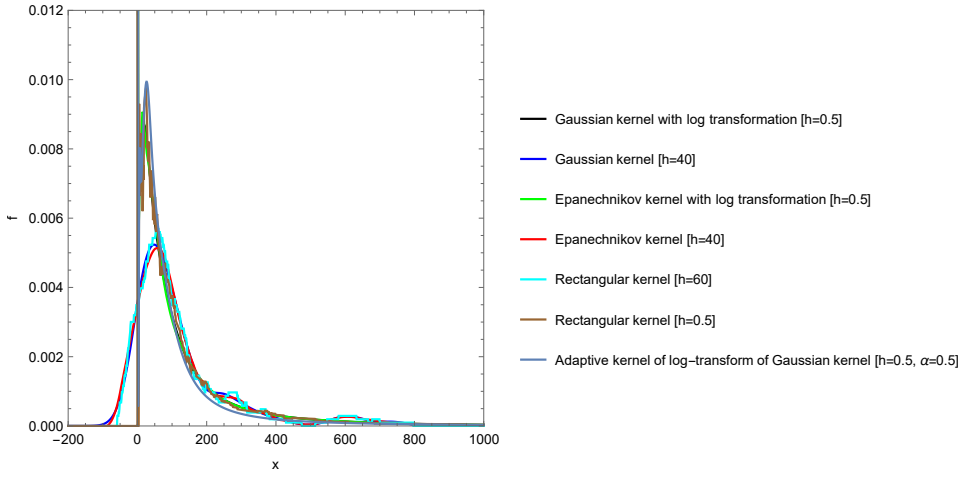

Figure 2

Example 1.2 illustrate two challenges in density estimation, namely determining the smoothing parameter (h) and - in the case of the adaptive estimate - also the sensitivity parameter (α). Intuitively it is expected that the optimal values of h and α should depend on the true density - that is information not contained in the sample. This argument goes in favor of choosing α and h subjectively based on expectations and knowledge external to the sample. That being said it is useful to be aware of the optimal values for h and α given some assumptions and use these as rules of thumb.

1.1.1 *Determining the parameters of density estimation*

The parameters of density estimation (h, α) are conventionally determined such that the difference between the true density f and the density estimate $\hat{f}$ is minimized. There exist many measures of difference between density distributions. In this study the total variation (TV), Kullback-Leibler divergence and mean integrated squared error (MISE) will be considered.

$$\begin{aligned} TV(f,\hat{f}) &= \frac{1}{2}\int |\hat{f}(x) - f(x)|dx, \\ I(f,\hat{f}) &= \int f(x)\log\left(\frac{f(x)}{\hat{f}(x)}\right)dx, \\ MISE[f,\hat{f}] &= \mathbb{E}\left[\int (\hat{f}(x) - f(x))^2\right]dx. \end{aligned} \tag{13}$$

The TV is the largest possible difference between predicted probabilities of two probability distributions. Because of this nice, intuitive interpretation the TV is used to evaluate the numerical experiments conducted in this study. However, there is no simple way of estimating the TV, so a different, somewhat equivalent, measure (of which there are many) is considered for the analytical manipulation associated to parameter determination; the Kullback-Leibler (KL) divergence. This is the procedure used in the maximum likelihood approach to parameter estimation. The two measures are related via the Pinsker inequality

$$TV(f,\hat{f}) \leq \sqrt{\frac{1}{2}I(f,\hat{f})}, \tag{14}$$

from which it is clear that minimizing the Kullback-Leibler divergence will result in a minimization of the TV as well. The last probability measure considered is the MISE. The MISE has the advantage of being (relatively) analytically forgiving however, it does not have the same intuitive probabilistic interpretation as the TV.

Parameter Estimation via the MISE

Parameter estimation via the MISE is (evidently) centered around the MISE, given by

$$\begin{aligned} MISE[f,\hat{f}] &\equiv \mathbb{E}\left[\int_{-\infty}^{\infty}(\hat{f}(x) - f(x))^2 dx\right] \\ &= \int_{-\infty}^{\infty} MSE_x[\hat{f}]dx\Bigg] \\ &= \int_{-\infty}^{\infty}(\mathbb{E}[\hat{f}(x)] - f(x))^2 dx + \int_{-\infty}^{\infty} Var[\hat{f}(x)]dx, \end{aligned} \tag{15}$$

where

$$\begin{aligned} MSE_x[f,\hat{f}] &\equiv \mathbb{E}[(\hat{f}(x) - f(x))^2] \\ &= (\mathbb{E}[\hat{f}(x)] - f(x))^2 + Var[\hat{f}(x)]. \end{aligned} \tag{16}$$

Using the MISE for parameter estimation conventionally proceed via one of two different routes; i) expand the bias and variance in the smoothing parameter (h) and determine the parameters that minimize the asymptotic limit $(h \to 0 \wedge n \to \infty)$ of the MISE ii) estimate the MISE itself from data and iteratively determine the values of h, α that minimize the estimated MISE.

ROUTE 1: THE ASYMPTOTIC MISE In route one the asymptotic limit $(h \to 0 \wedge n \to \infty)$ of the MISE is considered and the parameters are chosen in order to minimize this limit. The regular (non-adaptive) estimate can be considered as the limit of $\alpha \to 0$ of the adaptive estimate, so it will suffice to consider the expansion of the adaptive estimate.
In line with the short description of the asymptotic limit above the straightforward approach would be to perform an expansion of the bias and variance (with $h \to 0 \wedge n \to \infty$) to leading order and set the

sensitivity and smoothing parameters such that the asymptotic MISE (AMISE) is minimized. However, it turns out that the leading order in the bias can be eliminated by a particular choice of the sensitivity parameter, namely $\alpha = \frac{1}{2}$. For this reason $\alpha = \frac{1}{2}$ can be used as a rule of thumb for the sensitivity parameter [**Silverman86**]. However, by eliminating the leading order of the bias, there cannot be an associated rule of thumb for[4] h, and so one is left with choosing h subjectively or using an unrelated rule of thumb. An alternative approach is to pick α to eliminate the next to leading order term in the bias ($\alpha = \frac{1}{4}$) and then pick h to minimize the AMISE. This strategy results in a relevant rule of thumb for h and is further motivated by the fact that the higher order terms in the bias expansion are - in general - significant in magnitude. Motivated by the notion that the higher order terms of the bias are relevant, one can also try to pick the sensitivity parameter to minimize the entire bias to all orders under the assumption that - apart from the factorized numerical coefficient and dependence on α - all orders are of the similar magnitude ($\alpha \simeq 0.55$). However (as is the case for $\alpha = \frac{1}{2}$), this choice of sensitivity parameter leads to unbounded integrals in the leading order of the bias and so there is no associated rule of thumb for h in this case either. Adding this on top of the strong assumption on the magnitude of all orders in the bias expansion makes this last choice not recommended based on the arguments presented here. However, it goes to show that taking $\alpha \sim 0.5$ possibly also reduce the higher order terms in the bias - on top of eliminating the leading order. Hence, based on the arguments presented here, $\alpha \sim 0.5$ would seem the best choice for the sensitivity parameter given a rule of thumb on h is of no interest. However, as it turns out, the difference between taking $\alpha = \frac{1}{4}$ and $\alpha = \frac{1}{2}$ is very small in the numerical examples later considered and so taking $\alpha = \frac{1}{4}$ with the associated rule of thumb for h emerge as the recommended strategy. For the examples considered in one dimension, this strategy proves very effective.

To provide details on the above outline, begin with examining the bias for the adaptive estimate in the univariate case, or - to the same end - the expectation value for the adaptive estimate

$$\begin{aligned}\mathbb{E}[\hat{f}(t)] &= \frac{1}{n}\sum_i \mathbb{E}[w(x_i,t)] \\ &= \frac{1}{h}\int_{-\infty}^{\infty} dx f(x)^{\alpha+1} K\left(\frac{t-x}{h} f(x)^{\alpha}\right) \\ &= \int_{-\infty}^{\infty} du f(t+hu)^{\alpha+1} K\big(u f(t+hu)^{\alpha}\big),\end{aligned} \tag{17}$$

where $u = \frac{x-t}{h}$ and the symmetry of the kernel has been used in going from line one to two. Following **Silverman86** $g = 1$ is assumed in choosing the smoothing and sensitivity parameters. Taylor expanding the integrand reveals (see appendix **??**)

$$\begin{aligned}\mathbb{E}[\hat{f}] =& f + (1-2\alpha)\frac{h^2}{2f^{2\alpha+1}}\int_{-\infty}^{\infty} da a^2 K(a)\big[f\partial_t^2 f - 2\alpha(\partial_t f)^2\big] \\ &- (1-4\alpha)\frac{h^4}{24 f^{4\alpha+3}}\int_{-\infty}^{\infty} da a^4 K(a)\big[8\alpha(2\alpha+1)(4\alpha+1)(\partial_t f)^4 \\ &\qquad + 4\alpha f[4f\partial_t f\partial_t^3 f - 18(4\alpha-1)(\partial_t f)^2\partial_t^2 f \\ &\qquad\quad + 3f(\partial_t^2 f)^2] - f^3\partial_t^4 f\big] + \mathcal{O}(h^5).\end{aligned} \tag{18}$$

As is clear from equation (18); taking $\alpha = \frac{1}{i}$ will eliminate the $\mathcal{O}(h^i)$ term. As mentioned previously; it is often assumed that the leading order is the dominating one and so $\alpha = \frac{1}{2}$ can be suggested as a rule of thumb [**Silverman86**]. However, the $\mathcal{O}(h^4)$ term is significant (it often dominate) in the h-α plane of the bias for even moderate values of α and so it is expedient to instead take, as a rule of thumb, $\alpha = \frac{1}{4}$ such that the higher order term is eliminated. The fact that the higher order term in the bias dominate the choice of α indicate that even higher orders in the bias may be relevant. Considering now only the

[4] Because the next to leading order leads to unbounded integrals for this choice of α.

factorized numerical coefficients and α - let the rest be $\sim \mathcal{O}(1)$ for all orders for simplicity. In this case the integrated squared bias will have the form

$$\begin{aligned}\int_{-\infty}^{\infty}(\mathbb{E}[\hat{f}]-f)^2dx &\sim \left(\sum_{i=1}^{\infty}\frac{(1-2i\alpha)(-1)^i}{(2i)!}\right)^2 \\ &= (cos(1)+\alpha sin(1)-1)^2\end{aligned} \tag{19}$$

Equation (19) is minimized for $\alpha = 0.546 \simeq 0.55$. This value for α relies on the strict requirement that all orders in the expansion of the bias - apart from the factorized numerical coefficients and α - are of the same order.

Example 1.3.

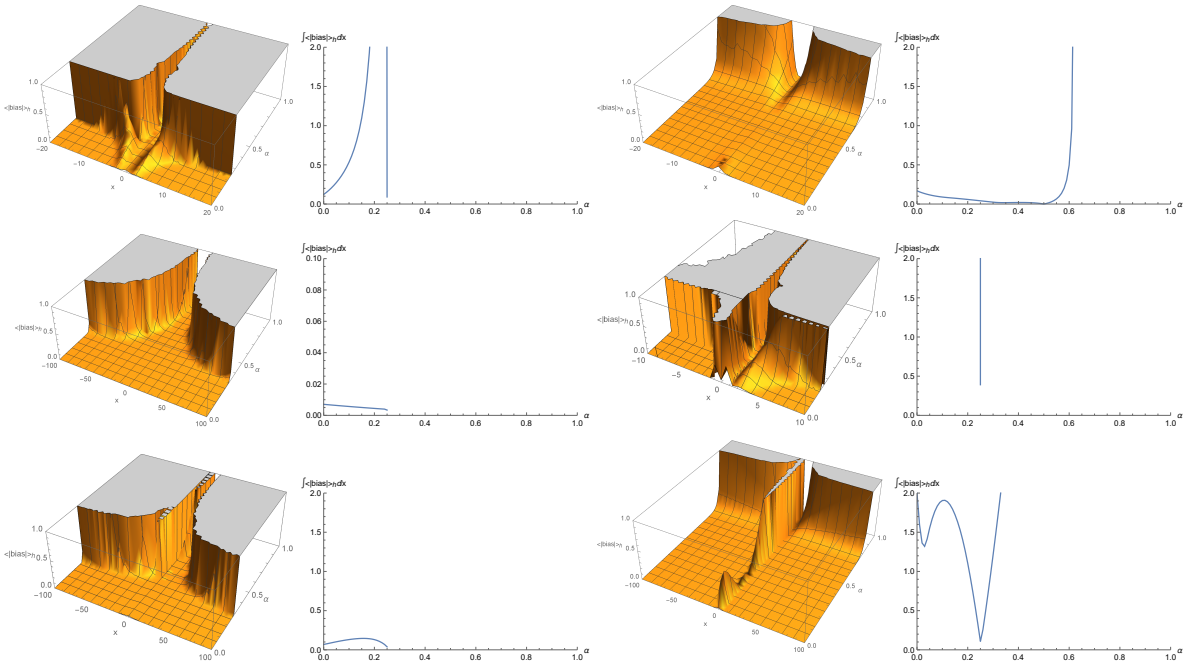

Figure 3: The figure consist of 6 panels each containing two figures. The left figure in each panel show the magnitude of the bias ($< \mathcal{O}(h^5)$) averaged across h as function of f and α. The right figure in each panel show the integral of the left panel over f. The difference between different panels is the true density. Top left panel $f \sim \mathcal{N}(0,1)$ (standard normal distribution). Top right panel $f \sim \mathcal{T}(0,1,1)$ (student-t). Middle left panel $f \sim \mathcal{L}(0,3)$ (logistic distribution). Middle Right panel $f \sim \mathcal{SN}(0,1,5)$ (skewed normal distribution). Bottom left panel $f \sim \mathcal{N}(0,1) + \mathcal{L}(10,2)$. Bottom right panel $f \sim \mathcal{SN}(0,1,5) + \mathcal{T}(5,3,1)$.

A way to analyze the bias (in this example only $< \mathcal{O}(h^5)$ is considered) is to consider the magnitude of the bias averaged over h and integrated over x. Figure 3 illustrates an approximation of this quantity as well as the integrand for different distributions. The approximation lie in the average across h; the average is computed from sampling the integrand for $h \in [0,1]$ with equidistant spacing of 0.01. The figure consist of six panels each containing two figures. The left figure in each panel show the magnitude of the bias ($< \mathcal{O}(h^5)$) averaged across h as function of x and α. The right figure in each panel show the integral of the left panel over x.
From figure 3 it is clear that overall - from the perspective of the analysis in this example - $\alpha = \frac{1}{4}$ is the better choice. For all distributions except the student-T distribution $\int \langle|bias|\rangle_h \, dx$ diverge for $\alpha > \frac{1}{4}$, meaning that the higher order term in the bias dominate over the leading order term in these cases. The student-t distribution is an interesting special case because here the higher order term in the bias almost vanishes at $\alpha = \frac{1}{2}$ meaning that the global minimum is at $\alpha = \frac{1}{2}$. This is, however - as mentioned - a special case. Hence, the assumptions for setting $\alpha = 0.546$ exactly are violated, but knowing that the higher order terms work in this direction provides a lifeline for setting $\alpha \sim \frac{1}{2}$ as a rule of thumb.
Another interesting aspect of figure 3 is that it appears that $\alpha = 0$ is often a reasonable choice. $\alpha = 0$ corresponds to the regular (non-adaptive) estimate, so it is indicated that the regular estimate will often be a reasonable choice.

Moving on to the variance

$$\begin{aligned} Var[\hat{f}(t)] &= \frac{1}{n}\sum_i Var[w(x_i,t)] \\ &= \frac{1}{n}\left[\frac{1}{h^2}\int_{-\infty}^{\infty} dx K\left(\frac{t-x}{h} f(x)^\alpha\right)^2 f(x)^{2\alpha+1} - \left(\frac{1}{h}\int_{-\infty}^{\infty} dx K\left(\frac{t-x}{h} f(x)^\alpha\right) f(x)^{\alpha+1}\right)^2\right] \\ &= \frac{1}{n}\left[\frac{1}{h}\int_{-\infty}^{\infty} du K\big(uf(t+hu)^\alpha\big) f(t+hu)^{2\alpha+1} + \mathcal{O}(h^0)\right] \\ &= \frac{f(t)^{\alpha+1}}{nh}\int_{-\infty}^{\infty} da K(a)^2 + \mathcal{O}(n^{-1}) \end{aligned} \tag{20}$$

Implicitly taking $\alpha = \frac{1}{4}$ (and hereby assuming the $\mathcal{O}(h^2)$ term in the bias dominate) the AMISE can be written

$$AMISE[\hat{f}] = (1-2\alpha)^2 \frac{h^4 k_2^2}{4}\int_{-\infty}^{\infty} dx\left(\frac{f\partial_t^2 f - 2\alpha(\partial_t f)^2}{f^{2\alpha+1}}\right)^2 + \frac{1}{nh}\int_{-\infty}^{\infty} dx f^{\alpha+1}\int_{-\infty}^{\infty} da K^2, \tag{21}$$

where the arguments are suppressed and the regular AMISE is obtained by setting $\alpha = 0$. Equation (21) illustrates what is often referred to as the bias-variance trade off; $\lim_{h\to 0}(bias) = 0$ however $\lim_{h\to 0}(variance) = \infty$ and vice versa for $h \to \infty$. Hence, there is a trade off between the bias and variance. The value of h that minimize the AMISE is referred to as the optimal h. In the case of equation (21), it is given by

$$h_{opt} = \left(\frac{\int_{-\infty}^{\infty} dx f^{\alpha+1}\int_{-\infty}^{\infty} da K^2}{(1-2\alpha)^2 n k_2^2 \int_{-\infty}^{\infty} dx\big(f^{-(2\alpha+1)}[f\partial_t^2 f - 2\alpha(\partial_t f)^2]\big)^2}\right)^{\frac{1}{5}}. \tag{22}$$

Equation (22) illustrates that the optimal value for h depends on the true f. It is also clear that $\lim_{n\to\infty}(h_{opt}) = 0$, however this limit is approached at a slow rate. Because the optimal value of the smoothing parameter depends on the true density there is no standard choice of h that is ideal for all possible true densities (beyond rules of thumb). Conventionally a subjective pick of h is taken based on the expectations of the true distribution. If the true distribution is known a better estimate can be computed. A rule of thumb is to reference the case where $f \sim \mathcal{N}$ and parametrize the dependence on the kernel viz

$$B(K) = \left(\frac{\int_{-\infty}^{\infty} da K(a)^2}{(\int_{-\infty}^{\infty} dx x^2 K(x))}\right)^{\frac{1}{5}}. \tag{23}$$

h_{opt} - derived based on the leading order in the expansion of the bias - for different values of α is shown in table 3 in the second column. In the third column adjusted values of h_{opt} is listed. The adjustment of h_{opt} is heuristically motivated such that the best estimate across true densities is obtained **[Silverman86]**. The adjusted h_{opt} in the first row is taken directly from **Silverman86** whereas the adjusted h_{opt} in the second row is derived from inspiration of the former.

α	$h_{opt}(f, K\sim\mathcal{N})$	Adjusted $h_{opt}(f, K\sim\mathcal{N})$
0	$1.37\hat{\sigma}n^{-\frac{1}{5}}B(K)$	$1.17\min\left(\hat{\sigma}, \frac{\hat{R}}{1.34}\right)n^{-\frac{1}{5}}B(K)$
1/4	$1.30\hat{\sigma}^{\frac{3}{4}}n^{-\frac{1}{5}}B(K)$	$1.11\min\left(\hat{\sigma}^{\frac{3}{4}}, \left(\frac{\hat{R}}{1.34}\right)^{\frac{3}{4}}\right)n^{-\frac{1}{5}}B(K)$
1/2	NA	NA
0.55	NA	NA

Table 3: $\min(x,y)$ picks the smaller of the two quantities, $\hat{\sigma}$ is the estimate of the standard deviation of the true normal distribution and $\hat{R}$ is the estimate of the interquartile range.

Substituting h_{opt} from equation (22) back into the AMISE yields

$$AMISE[\hat{f}] = \frac{5}{4}C(K)\left(\int_{-\infty}^{\infty} dx f^{\alpha+1}\right)^{\frac{4}{5}}\left(\frac{(1-2\alpha)^2}{n^4}\int_{-\infty}^{\infty} dx\left(\frac{f\partial_t^2 f - 2\alpha(\partial_t f)^2}{f^{2\alpha+1}}\right)^2\right)^{\frac{1}{5}}, \tag{24}$$

where

$$C(K) = \left(\sqrt{k_2} \int_{-\infty}^{\infty} dt K(t)^2 \right)^{\frac{4}{5}}. \tag{25}$$

Equation (24) illustrates that K should be chosen such that $C(K)$ is small. **Silverman86** shows that $C(K)$ is minimized by the Epanechnikov kernel. The efficiency of a kernel is therefore defined with respect to the Epanechnikov kernel. Denote the Epanechnikov kernel by K_e, then the efficiency is defined as

$$\begin{aligned} eff(K) &\equiv \left(\frac{C(K_e)}{C(K)} \right)^{\frac{5}{4}} \\ &= \frac{3}{5\sqrt{5}} \left[\frac{1}{\sqrt{\int_{-\infty}^{\infty} dt K(t) t^2 \int_{-\infty}^{\infty} dt' K(t')^2}} \right]. \end{aligned} \tag{26}$$

The efficiency is listed alongside the kernels considered in table 2.

Example 1.4.

Inspired by example 2.3 it is interesting to investigate the AMISE (with up to $\mathcal{O}(h^4)$ in the bias) as a function of h for fixed $\alpha = 0, \frac{1}{4}, \frac{1}{2}, 0.55$. Figure 4 illustrates the AMISE for a random sample of $n = 50$ drawn from six different distributions. The relatively low value of n is chosen such that the effect of the variance is clearly visible (recall the variance is $\propto h^{-1}$).

In the figure the blue line denotes the AMISE for $\alpha = \frac{1}{2}$, the black line for $\alpha = 0$, the brown line for $\alpha = \frac{1}{4}$ and the cyan line for $\alpha = 0.55$. The blue and black dots denote AMISE with $\alpha = 0$ evaluated at $h = 1.06\hat{\sigma} n^{-\frac{1}{5}}$ and $h = 0.9 \min\left(\hat{\sigma}, \frac{\hat{R}}{1.34}\right) n^{-\frac{1}{5}}$, respectively. The red and green dots denote AMISE with $\alpha = \frac{1}{4}$ evaluated at $h = 1.01 \hat{\sigma}^{\frac{3}{4}} n^{-\frac{1}{5}}$ and $h = 0.86 \min\left(\hat{\sigma}^{\frac{3}{4}}, \left(\frac{\hat{R}}{1.34}\right)^{\frac{3}{4}}\right) n^{-\frac{1}{5}}$, respectively. With regards to the dots, note that there is an element of randomness to their location from the randomness of the random sample of $n = 50$ they are based on. However, the placements can be considered representative since the fluctuations between samples are not that great. Note also that there are no instances of complete overlaps between dots, so when a dot is not visible it is outside the plot range at higher h.

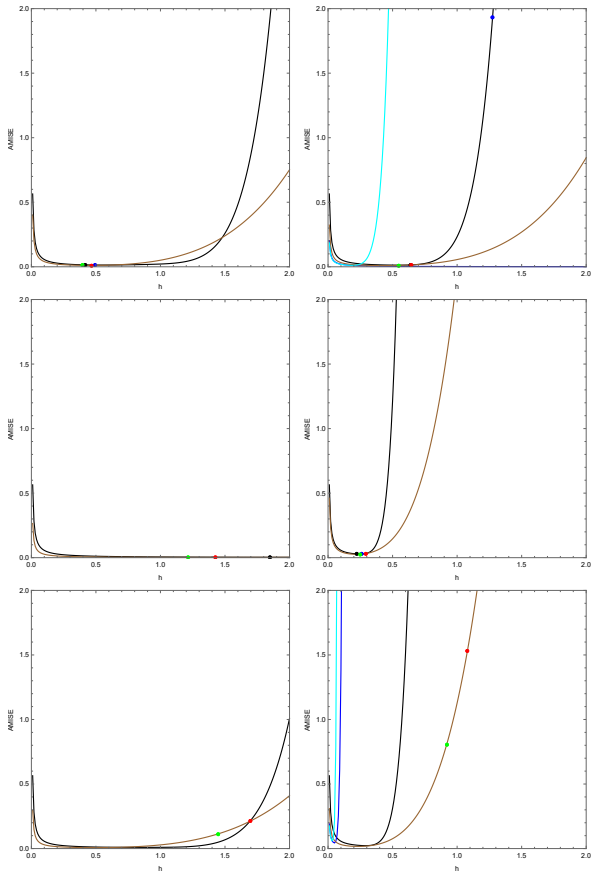

Figure 4: The figure contains six panels which show the AMISE as a function of h for different true densities assuming a Gaussian kernel. The blue line denotes the AMISE for $\alpha = \frac{1}{2}$, the black line for $\alpha = 0$, the brown line for $\alpha = \frac{1}{4}$ and the cyan line for $\alpha = 0.55$. The blue and black dots denote AMISE with $\alpha = 0$ evaluated at $h_{opt}(f \sim \mathcal{N})$ and the adjusted $h_{opt}(f \sim \mathcal{N})$, respectively. The red and green dots denote AMISE with $\alpha = \frac{1}{4}$ evaluated at $h_{opt}(f \sim \mathcal{N})$ and the adjusted $h_{opt}(f \sim \mathcal{N})$, respectively. Top left panel $f \sim \mathcal{N}(0,1)$ (standard normal distribution). Top right panel $f \sim \mathcal{T}(0,1,1)$ (student-t). Middle left panel $f \sim \mathcal{L}(0,3)$ (logistic distribution). Middle Right panel $f \sim \mathcal{SN}(0,1,5)$ (skewed normal distribution). Bottom left panel $f \sim \mathcal{N}(0,1) + \mathcal{L}(10,2)$. Bottom right panel $f \sim \mathcal{SN}(0,1,5) + \mathcal{T}(5,3,1)$.

From the figure several things can be noted; first, the cyan and blue lines only appear in the top right and bottom right panels due to the AMISE being undefined in the other cases. Second, the black line contain both derived orders in the bias with $\alpha = 0$. Hence, the black line does not represent the AMISE that the rule of thumb value for h is obtained from. In this case only the leading order in the bias is included. The brown line, on the other hand, only contain the leading order in the bias as the next to leading order is eliminated by the choice of α. Hence, this line represents the AMISE that the rule of thumb for $\alpha = \frac{1}{4}$ is derived from. For this reason the red dot is located exactly at the minimum of the brown curve in the top left panel. Third, even for this relatively low value of n, the contribution of the variance is negligible for $h \gtrsim 0.3$. Fourth, the black dot, representing the regular (non-adaptive, i.e. $\alpha = 0$) estimate with $h = 0.9 \min\left(\hat{\sigma}, \frac{\hat{R}}{1.34}\right) n^{-\frac{1}{5}}$, provide a good estimate for all uni-modal distributions considered. However, this estimate performs relatively poor in case of the multi-modal distributions considered. The best overall rule of thumb is the green dot, representing the adaptive estimate with $\alpha = \frac{1}{4}$ and $h = 0.86 \min\left(\hat{\sigma}^{\frac{3}{4}}, \left(\frac{\hat{R}}{1.34}\right)^{\frac{3}{4}}\right) n^{-\frac{1}{5}}$.

ROUTE 2: ESTIMATING THE MISE In route two the MISE itself is estimated from the information in the sample and subsequently minimized with respect to h, α. Writing out the MISE

$$\begin{aligned} MISE[\hat{f}] &\equiv \mathbb{E}\left[\int_{-\infty}^{\infty} dx(\hat{f}(x) - f(x))^2\right] \\ &= \int_{-\infty}^{\infty} dx\hat{f}(x)^2 - 2\int_{-\infty}^{\infty} dx f(x)\hat{f}(x) + \int_{-\infty}^{\infty} dx f(x)^2 \\ &\equiv R(\hat{f}, f) + \int_{-\infty}^{\infty} dx f(x)^2, \end{aligned} \tag{27}$$

where the last line define $R(\hat{f}, f)$. f does not depend on h, α, so in this context minimizing R corresponds to minimizing the MISE. The first integral in R is straightforward to write out whereas the second can be estimated by the mean cross-validated estimate of f, that is **[Silverman86]**

$$\hat{R}(h, \alpha) = \int_{-\infty}^{\infty} dx \left[\frac{1}{n}\sum_i \frac{1}{h\lambda_i} K\left(\frac{x - x_i}{h\lambda_i}\right)\right]^2 - \frac{2}{n}\sum_i \hat{f}_{-i}(x_i), \tag{28}$$

where the dependence on α is implicit through λ_i (see equation (8)) and

$$\hat{f}_{-i}(x_i) = \frac{1}{(n-1)}\sum_{j \neq i} \frac{1}{h\lambda_j} K\left(\frac{x_i - x_j}{h\lambda_j}\right). \tag{29}$$

To see that the latter term in equation (28) is a reasonable estimate of the latter integral in R, consider the expectation value of the estimator

$$\begin{aligned} \mathbb{E}\left[\frac{1}{n}\sum_i \hat{f}_{-i}(x_i)\right] &= \mathbb{E}[\hat{f}_{-i}(x_i)] \\ &= \int \hat{f}(x) f(x) dx. \end{aligned} \tag{30}$$

Hence, from equation (28) h, α can be determined by numerically minimizing $\hat{R}$.

Parameter Estimation via the Kullback-Leibler Divergence

As mentioned previously, the Kullback-Leibler divergence is connected via to the TV via the Pinsker inequality (equation (14)) and so minimizing the Kullback-Leibler divergence yields estimates of the parameters that also minimize the TV. The Kullback-Leibler divergence can be estimated viz

$$\hat{I}(h, \alpha) = \frac{1}{n}\sum_i \log\left(\hat{f}_{-i}(x_i)\right). \tag{31}$$

The expectation of $\hat{I}$ is

$$\begin{aligned} \mathbb{E}[\hat{I}] &= \mathbb{E}[\log\left(\hat{f}_{-i}(x_i)\right)] \\ &\simeq \int f(x) \log\left(\hat{f}(x)\right) \\ &= -I(f, \hat{f}) + \int f(x) \log\left(f(x)\right) dx, \end{aligned} \tag{32}$$

where the approximation lies in letting the estimate of f based on $n-1$ observations go to the estimate of f based on n observations (i.e. let $\hat{f}_{-i} \to \hat{f}$). Equation (32) shows that $\hat{I}$ is an unbiased estimator of I up to a constant. Hence, from equation (31) h, α can be determined by numerically minimizing $\hat{I}$.

Example 1.5.

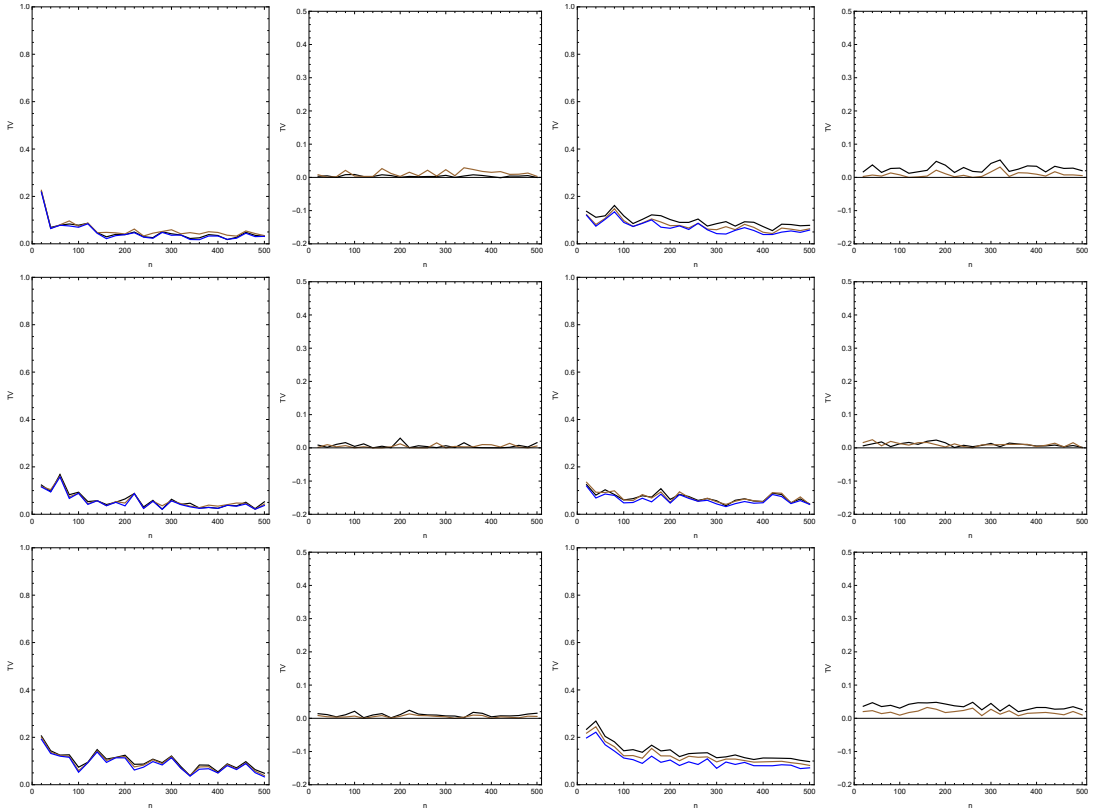

Figure 5: The figure shows $TV(n)$ between different true distributions and different estimates. Blue is an adaptive estimate using an approximation of the optimal h and α (h, α are scanned in a rough way to determine the optimal value), black is an adaptive estimate using an approximation of the optimal h and $\alpha = \frac{1}{4}$ and blue is an adaptive estimate using an approximation of the optimal h and $\alpha = \frac{1}{2}$. The true densities are the same as for example 2.3 and 2.4.

Figure 5 illustrates $TV(n)$ between true densities and i) an adaptive estimate using a approximations of the ideal h and α determine (blue) ii) an adaptive estimate using $\alpha = \frac{1}{4}$ and an approximation of the ideal h (black) and iii) an adaptive estimate using $\alpha = \frac{1}{2}$ and an approximation of the ideal h (brown). All estimates use the Gaussian kernel shown in table 2. The plot consists of six panels each consisting of two figures (being horizontal neighbors). The left figure in each panel is $TV(n)$ whereas the right panel is the difference in TV between the approximately optimal estimate and the remainder. The only thing that differs between panels is the true density. The experiment is conducted by taking 25 random samples of the true distribution of size $n \in [20, 500]$ with equidistant sample sizes ($n = 20, 40, 60, \ldots 500$). For each sample the TV is determined via estimates i), ii) and iii). The approximation of the optimal h, α is determined by performing a grid search for the lowest TV. This is the case in all TV-examples in this study.
From figure 5 it is clear that the TV of estimates i), ii) and iii) is approximately the same for all considered distributions. This means that, as speculated previously, the higher order terms in the bias are significant - otherwise the $\alpha = \frac{1}{2}$ estimate would have been significantly worse. It also means that the choice of α is relatively unimportant when compared to the choice of h. With this in mind, the choice of $\alpha = \frac{1}{4}$ is recommended because the theoretical arguments in favor of this choice can be formed without investigating terms $> \mathcal{O}(h^5)$ and perhaps more importantly, because this choice has a well defined rule of thumb for h associated to it.
From the figure it is also clear that the improvement of the adaptive estimate as a function of n is relatively modest - as speculated in example 2.4. This indicates that the AMISE with $\alpha = \frac{1}{4}$ well approximates the behaviour of the MISE in general.

Example 1.6.

Figure 6 illustrates $TV(n)$ between true densities and i) an adaptive estimate using an approximation of the ideal h and α (blue) ii) an adaptive estimate using $\alpha = \frac{1}{4}$ and the adjusted value for h from table 3 (black), iii) an adaptive estimate with $\alpha = 0$ using an approximation of the ideal h (cyan) and iv) an adaptive estimate with $\alpha = 0$ using the adjusted value for h from table 3 (brown). The experiment and figure is constructed as detailed

in example 2.5.
Several things can be noted from figure 6; first, the adaptive estimate with approximations for ideal h, α improves upon the regular estimate with an approximately ideal h. The improvements are however relatively small (for the considered true densities), meaning that - as indicated in example 2.3 - the regular estimate is not a bad choice in general. Second, for the unimodal distributions all estimates end up with $TV \lesssim 0.1$ as $n \to 500$. The adaptive estimate with approximations of ideal h, α is the best estimate whereas the other three estimates are close behind and roughly equally good. For multimodal distributions however, the regular estimate with the adjusted h from table 3 is relatively poor. Here the adaptive estimate with approximations for ideal h, α is again the best, with the regular estimate with an approximately ideal h and the adaptive estimate with $\alpha = \frac{1}{4}$ and the adjusted h from table 3 relatively close and about equal in TV. Hence, it can be concluded that the regular estimate with approximately ideal h is a good estimate for all (considered) distributions. Interestingly, the adaptive estimate with $\alpha = \frac{1}{4}$ and the adjusted h from table 3 is very similar in terms of TV. The adaptive estimate with approximately ideal h, α improves slightly upon the former two. Lastly, the regular estimate with the adjusted h from table 3 performs well for unimodal distributions but relatively poorly for multimodal distributions.
Lastly, it can be concluded that the results are consistent with the content of example 2.4, in that estimate ii) is almost identical (in terms of TV) to estimate i) for the top left and middle panels. In the remainder there are relatively small differences - as should be the case. To further check the contents of example 2.4, the non-adjusted h can be used instead of the adjusted one (both from table 3) to compute the adaptive estimate with $\alpha = \frac{1}{4}$ for the sample with $n = 500$. Using the adjusted h reveals $TV \simeq 0.08$ whereas using the non-adjusted h reveals $TV \simeq 0.4$. Hence, using the non-adjusted h instead of the adjusted one leads to a significant increase in the TV driven by a large bias - as predicted by example 2.4. Hence, the AMISE with $\alpha = \frac{1}{4}$ from example 2.4 appear consistent with the results in example 2.5 and 2.6. This means that not only does taking $\alpha = \frac{1}{4}$ provide a rule of thumb value for h, but additionally it provides an analytical relation for the AMISE that relatively well approximates the MISE. This is a significant advantage as the AMISE can be used to asses different rules of thumb for h and further more predict how well these perform for different distributions (without the need to perform time consuming numerical experiments).

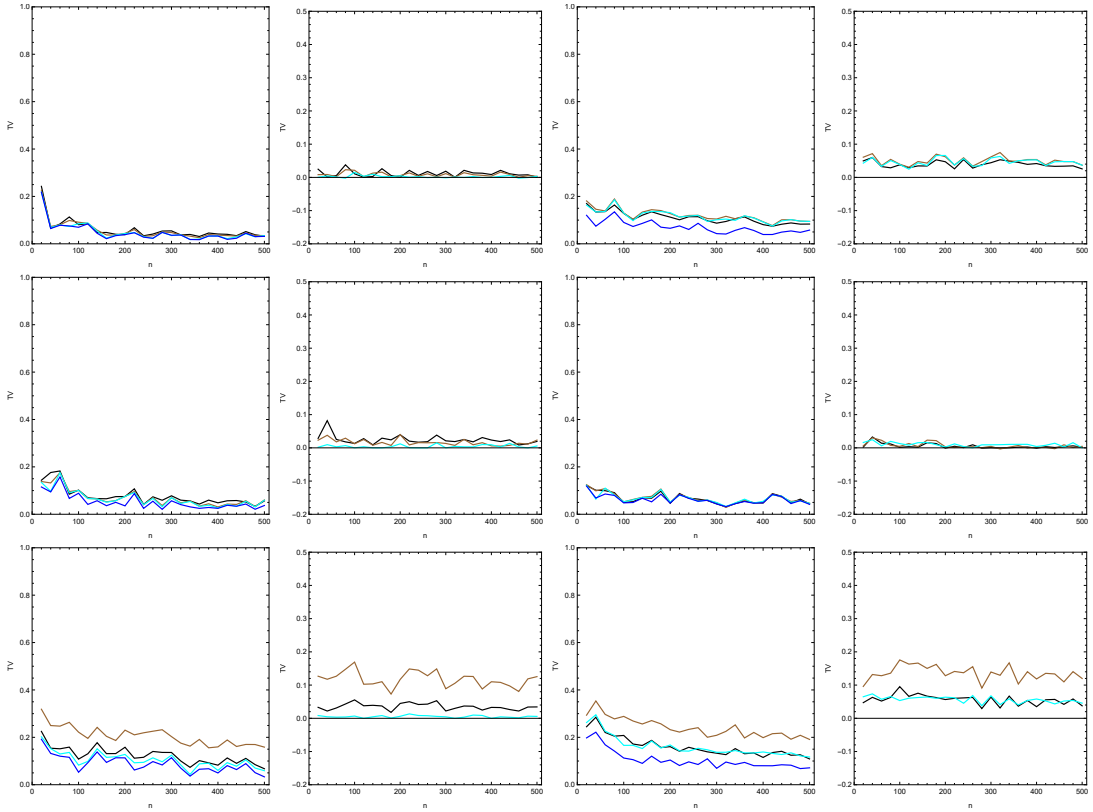

Figure 6: The figure shows $TV(n)$ between different true distributions and different estimates. Blue is an adaptive estimate using an approximation of the ideal h and α, black is an adaptive estimate using $\alpha = \frac{1}{4}$ and the adjusted h from table 3, cyan is an adaptive estimate with $\alpha = 0$ using an approximation of the ideal h and brown is an adaptive estimate with $\alpha = 0$ using the adjusted h from table 3. The true densities are the same as for example 2.5.

Example 1.7.

Figure 7 illustrates $TV(n)$ between true densities and i) an adaptive estimate using an approximation of the ideal h and α (blue) ii) an adaptive estimate using $\alpha = \frac{1}{4}$ and the adjusted value for h from table 3 (cyan), iii) an estimate with h, α chosen via numerically minimizing equation (28) *via a grid search (brown) and iv) an estimate with h, α chosen via numerically minimizing equation* (31) *via a grid search (black). The experiment and figure is constructed as detailed in example 2.5.*

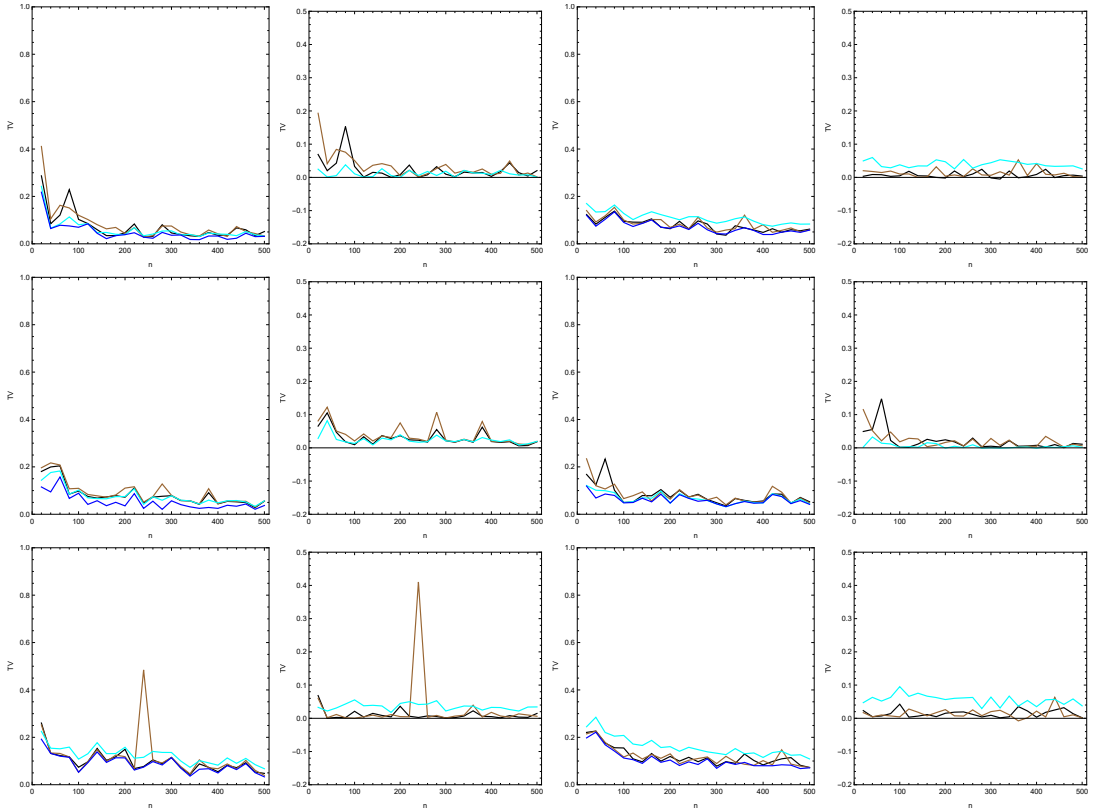

Figure 7: The figure shows $TV(n)$ between different true distributions and different estimates. Blue is an adaptive estimate using an approximation of the ideal h and α, cyan is an adaptive estimate using $\alpha = \frac{1}{4}$ and the adjusted h from table 3, brown is an estimate using h, α as obtained by minimizing equation (28) via a grid search and black is an estimate using h, α as obtained by minimizing equation (31) via a grid search. The true densities are the same as for example 2.5.

As is clear from figure 7, the TVs of estimates iii) and iv) are generally very close to the TV of estimate i). Estimate iii) has a handful of relatively large (relative to the otherwise very small difference in TV between estimate i) and iii)) fluctuations away from the TV of estimate i) (distributed at various values of n) as well as one major fluctuation in case of the bottom left panel. Estimate iv) in general has fluctuations similar to estimate iii) but of smaller magnitude and without the major fluctuation entirely. Hence, estimate iv) emerge the best overall estimate in this example with estimate iii) close behind albeit being prone to fluctuations. Estimate ii) is approximately as good as estimate iii) and iv) (a little worse for the top right panel) for the unimodal distribution and a little worse for the multimodal distributions. That being said, estimate ii) is still a very good estimate and has a few advantages over estimates iii) and iv); first, estimate ii) is less prone to large fluctuations relative to estimate i) as compared to estimate iv) and especially estimate iii). Second, estimate ii) is much more forgiving computationally since it does not rely on a grid search for the optimal values of h, α. Especially estimate iii) is heavy computationally since it involves integrals over all-space for each grid point in the grid search. The computational problems - centered around the integral over all-space - for estimate iii) also increase with dimensionality, for which reason estimate iii) is not considered in higher dimensions.

Example 1.8.

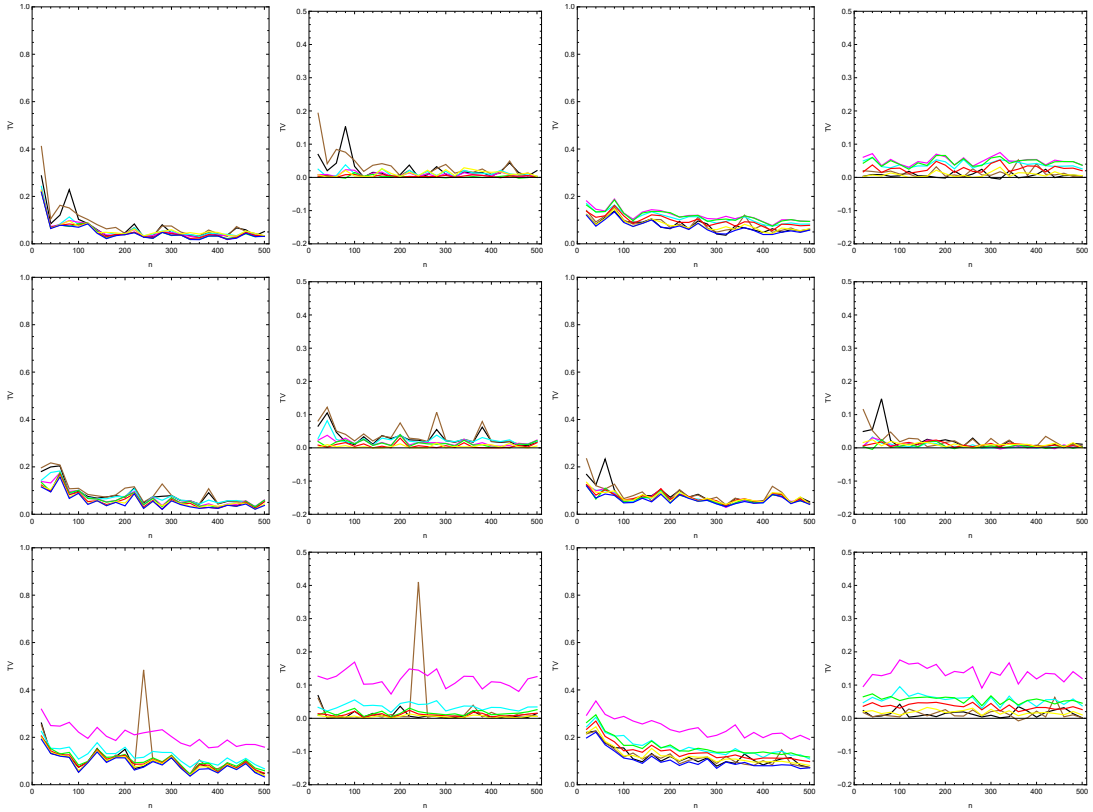

Figure 8: $TV(n)$ between true densities and all estimates reviewed in examples 2.5-2.7. Due to amount of graphs displayed the color designation is described in the text below the figure. The distributions are the same as for example 2.5.

Figure 8 recaps $TV(n)$ between true densities and all estimates reviewed in examples 2.5-2.7. Blue an estimate with approximately optimal h, α determined via a grid search. Red is an estimate with $\alpha = \frac{1}{4}$ and an approximately optimal h determined via a grid search. Yellow is an estimate with $\alpha = \frac{1}{2}$ and an approximately optimal h determined via a grid search. Green is an estimate with $\alpha = 0$ and an approximately optimal h determined via a grid search. Magenta is an estimate with $\alpha = 0$ and the adjusted value for h from table 3. Cyan is an estimate with $\alpha = \frac{1}{4}$ and the adjusted value for h from table 3. Brown is an estimate with h, α obtained by minimizing equation (28) via a grid search and black is an estimate with h, α obtained by minimizing equation (31) via a grid search.

Example 1.9.

The difference between the regular ($\alpha = 0$) and adaptive ($\alpha \neq 0$) estimate is that the former uses a fixed smoothing parameter whereas the latter uses a variable smoothing parameter. Because the regular estimate uses a fixed smoothing parameter, the estimate struggles with the tails of the distribution where the sample data are weighted too heavily. In order to illustrate this effect, consider the regular and adaptive estimates of $f \sim \mathcal{T}(5, 2, 1)$ (student-t distribution) from a random sample of $n = 500$ using a Gaussian kernel illustrated in figure 9. The total variations corresponding to figure 9; $TV_{h_{rot}} = 0.08$ (black), $TV_{h_{opt}} = 0.07$(blue), $TV_{h_{rot},\alpha_{rot}} = 0.06$(green), $TV_{h_{rot},\alpha_{opt}} = 0.05$(red). In terms of TV there is little difference between the regular Gaussian estimate with optimal h and the adaptive estimates. Figure 9 illustrates that the (small) difference comes from the wiggles in the tails of the regular estimates. The challenge of the regular estimates is that if the distribution of data is very peaked the best result is obtained with a low value of the smoothing parameter, however, a low value of the smoothing parameter will give too much weight to data in the tails and so wiggles will occur. This problem is not present (or at least significantly less so) in the adaptive estimates since here the smoothing parameter varies with the position such that a large value of the smoothing parameter can be applied in the tails and a lower towards the peaked center.

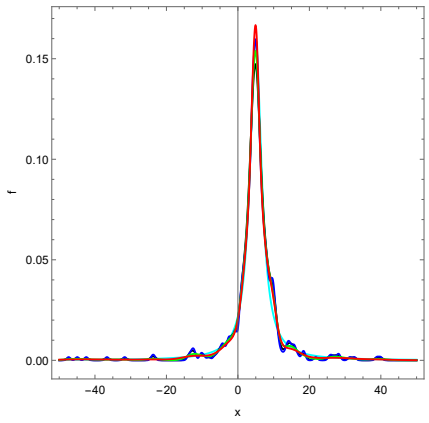

Figure 9: The univariate density estimates of $f \sim \mathcal{T}(5,2,1)$. All estimates use the Gaussian kernel. Cyan is the true density, black is an adaptive estimate with $\alpha = 0$ using the adjusted h from table 3, blue is an adaptive estimate with $\alpha = 0$ using an approximately optimal h, green is an adaptive estimate with $\alpha = \frac{1}{4}$ and an approximately optimal h and red is an adaptive estimate with approximately optimal h, α. The pilot estimate are largely independent of the pilot estimates, however, for completeness sake the relevant rules of thumb values for h are used in the pilot estimates.

1.2 MULTIVARIATE DENSITY ESTIMATION

The generalization of the univariate formalism to the multivariate case proceeds by generalizing the kernel and accounting for the possibility of the different variables having different scales. The latter can be accounted for by linearly transforming data such that

$$\hat{f}(\vec{t}) = \frac{1}{n|H|} \sum_{i=1}^{n} K\big(H^{-1}(\vec{t} - \vec{x}_i)\big), \tag{33}$$

where H - the smoothing matrix - is a $d \times d$ matrix responsible for the linear transformation of data. The multivariate versions of the Epanechnikov and Gaussian kernels are listed in table[5] 4.

Kernel	$K(\vec{w})$
Epanechnikov	$\begin{cases} \frac{1}{2c_d}(d+2)(1-\vec{x}^T\vec{x}) & \text{for } \vec{x}^T\vec{x} < 1 \\ 0 & \text{Otherwise} \end{cases}$
Gaussian	$\frac{e^{-\frac{\vec{x}^T\vec{x}}{2}}}{(2\pi)^{\frac{d}{2}}}$

Table 4

Similarly to the conditions in equation (7) the multivariate kernels (considered here) are required to satisfy **[Scott2015]**

$$\int_{\mathcal{R}^d} K(\vec{w})d\vec{w} = 1, \qquad \int_{\mathcal{R}^d} \vec{w}K(\vec{w})d\vec{w} = \vec{0}, \qquad \int_{\mathcal{R}^d} \vec{w}\vec{w}^T K(\vec{w})d\vec{w} = \tilde{k}_2 \mathbb{1}, \tag{34}$$

where $d\vec{x} = dx_1 dx_2 \dots dx_d$ and $\tilde{k}_2$ is a scalar. The fixed kernel methods generalize to the adaptive kernel method analogously to the univariate case, so again it is a three-step procedure

1. Develop a pilot estimate $\tilde{f}$ for which $\tilde{f}(\vec{x}_i) > 0$.
2. Define a local bandwidth, λ_i viz

$$\lambda_i \equiv \left(\frac{g}{\tilde{f}(\vec{x}_i)}\right)^{\alpha}, \tag{35}$$

[5] c_d is the volume of a d-dimensional unit-sphere. E.g. $c_1 = 2$, $c_2 = \pi, c_3 = \frac{4\pi}{3}, \dots$.

where

$$g = e^{\frac{1}{n}\sum_i ln(\tilde{f}(\vec{x}_i))} \tag{36}$$

is the geometric mean of $\tilde{f}$.

3. The adaptive kernel estimate is then (for the multivariate case)

$$\hat{f}(\vec{t}) = \frac{1}{n|H|}\sum_{i=1}^{n}(\lambda_i)^{-d}K\big(\lambda_i^{-1}H^{-1}(\vec{t}-\vec{x}_i)\big), \tag{37}$$

where again the regular estimate is obtained by taking $\alpha \to 0$.

1.2.1 *Determining the parameters of density estimation*

The parameters of density estimation in the multivariate case are determined analogously to the univariate case. In this study the generalized versions of the total variation (TV), Kullback-Leibler divergence and mean integrated squared error (MISE) will be considered.

$$\begin{aligned} TV(f,\hat{f}) &= \frac{1}{2}\int_{\mathcal{R}^d}|\hat{f}(\vec{x}) - f(\vec{x})|d\vec{x}, \\ I(f,\hat{f}) &= \int_{\mathcal{R}^d} f(\vec{x})\log\left(\frac{f(\vec{x})}{\hat{f}(\vec{x})}\right)d\vec{x}, \\ MISE[f,\hat{f}] &= \mathbb{E}\left[\int_{\mathcal{R}^d}(\hat{f}(\vec{x}) - f(\vec{x}))^2\right]d\vec{x}, \end{aligned} \tag{38}$$

where again the TV and Kullback-Leibler divergence is related via the Pinsker inequality (equation (14)).

Parameter Estimation via the MISE

Using the MISE the parameter estimation proceeds completely analogously to the univariate case via one of two different routes; i) expand the bias and variance in the smoothing parameter (h) and determine the parameters that minimize the asymptotic limit ($h \to 0 \wedge n \to \infty$) of the MISE ii) estimate the MISE itself from data and iteratively determine the values of h, α that minimize the estimated MISE. The only difference is the challenge and changes that come from the multivariate formalism.

ROUTE 1: THE ASYMPTOTIC MISE The first route proceeds analogous to the univariate case. The asymptotic expansion shows that in general (see appendix **??**)

$$\mathbb{E}[\hat{f}] - f = \sum_{i=1}^{\infty}\frac{1-[2i+1-d]\alpha}{(2i)!}\cdot(\dots). \tag{39}$$

Hence - in accordance with the results from the univariate case - the i'th term in the bias can be eliminated by setting $\alpha = \frac{1}{2i+1-d}$. Similarly to the univariate case α can be set to eliminate the leading order or the next to leading order. Contrary to eliminating the leading order, eliminating the next to leading order will result in an associated rule of thumb for h - giving this choice a competitive advantage. Letting $(\dots)$ in equation (39) be of the same order for $d = 1$ results in the bias being minimized for $\alpha \simeq 0.55$ as found in the univariate case. For $d \geq 2$ however, this procedure reveals $\alpha > 1$ and so is not a viable option in this case. Hence, the options remain to eliminate either the leading or next to leading order in the bias. The difference between the two choices of α is little - as seen in the univariate case - and so the choice resulting in an associated rule of thumb value for h may be favoured for just this reason.

For the multivariate adaptive estimate the smoothing matrix and sensitivity parameter are - analogously to the univariate case - determined by investigating the asymptotic expansion of the MISE. The expectation value in this case (see appendix **??**)

$$\mathbb{E}[\hat{f}] = f + \frac{[1-(3-d)\alpha]\tilde{k}_2}{2}\left[\frac{fTr[H^T\tau H] - (3-d)\alpha Tr[H^T\Lambda H]}{f^{2\alpha+1}}\right] + \mathcal{O}(H^3), \tag{40}$$

where $\tau_{jk} = \frac{\partial^2 f}{\partial t_j \partial t_k}$ and $\Lambda_{jk} = \frac{\partial f}{\partial t_j}\frac{\partial f}{\partial t_k}$. The variance term is computed straightforwardly and so the AMISE is

$$\begin{aligned} AMISE[\hat{f}] =& [1-(3-d)\alpha]^2 \frac{\tilde{k}_2^2}{4} \int_{\mathcal{R}^d} \left(\frac{fTr[H^T \tau H] - (3-d)\alpha Tr[H^T \Lambda H]}{f^{2\alpha+1}} \right)^2 d\vec{x} \\ &+ \frac{1}{n|H|} \int f^{\alpha+1} d\vec{x} \int_{\mathcal{R}^d} K^2 d\vec{w}. \end{aligned} \tag{41}$$

Next, parametrize the H-matrix viz

$$H = hA, \tag{42}$$

where $h > 0$ is a scalar. Minimizing the AMISE with respect to h reveals

$$h_{opt} = \left[\frac{d}{[1-(3-d)\alpha]^2 n|A|} \frac{\int f^{\alpha+1} d\vec{x} \int_{\mathcal{R}^d} K^2 d\vec{w}}{\tilde{k}_2^2 \int \left(\frac{fTr[A^T \tau A] - (3-d)\alpha Tr[A^T \Lambda A]}{f^{2\alpha+1}} \right)^2 d\vec{x}} \right]^{\frac{1}{d+4}}. \tag{43}$$

From equation (43) it is clear that - similar to the univariate case - the optimal value for h depends on the true f and again $\lim_{n\to\infty}(h_{opt}) = 0$ however at an increasingly slow rate as d increases. In line with the univariate case a rule of thumb value for h is to take $f \sim \mathcal{N}$, $A = \Sigma^{\frac{1}{2}}$ and defining (see appendix **??**)

$$\tilde{B}(K) \equiv \left(\frac{\int_{\mathcal{R}^d} K^2 d\vec{w}}{\tilde{k}_2^2} \right)^{\frac{1}{d+4}}. \tag{44}$$

h_{opt} - derived based on the leading order in the expansion of the bias (see appendix **??**) - for different values of α is shown in table 5 in the second column. In relation to table 5 and estimating h_{opt} via a rule of thumb, it is interesting to note that for $d = 3$ the procedure of eliminating the next to leading order via α is not possible. For $d = 3$ the factorized α-dependency of the leading order term vanishes and the next to leading order can be cancelled by setting $\alpha = \frac{1}{2}$. However, for $\alpha = \frac{1}{2}$ the integral in the denominator in h_{opt} diverge and so the next to leading order cannot be canceled in this case. In this case the third leading order term in the bias can be eliminated by setting $\alpha = \frac{1}{4}$ and deriving the associated rule of thumb for h using the leading order.

Estimate	α	d	$h_{opt}(f \sim \mathcal{N})$
Regular	0	d	$\left(\frac{4(2\sqrt{\pi})^d}{n(d+2)}\right)^{\frac{1}{d+4}} \tilde{B}(K)$
Cancel LO	$\frac{1}{3-d}$	d	NA
Cancel NTLO	1/3	2	$0.74 n^{-\frac{1}{6}} \lvert S\rvert^{-\frac{5}{36}} \tilde{B}(K)$
Cancel TLO	1/4	3	$0.78 n^{-\frac{1}{7}} \lvert S\rvert^{-\frac{5}{56}} \tilde{B}(K)$

Table 5: $|S|$ denotes the determinat of the sample covariance matrix (see e.g. equation (**??**)). LO, NTLO and TLO abbreviates "leading order", "next to leading order" and "third leading order", repsectively and cancel LO, NTLO and TLO denote choosing α chosen to eliminate these terms.

Substituting h_{opt} from equation (43) back into the AMISE yields

$$AMISE[\hat{f}] = \frac{d+4}{4d^{\frac{d}{d+4}}} \tilde{C}(K) \left(\frac{\int f^{\alpha+1} d\vec{x}}{|A|n} \right)^{\frac{4}{d+4}} \left([1-(3-d)\alpha]^2 I_f \right)^{\frac{d}{d+4}}, \tag{45}$$

where

$$\tilde{C}(K) \equiv \left(\tilde{k}_2^{\frac{d}{2}} \int_{\mathcal{R}^d} K^2 d\vec{w} \right)^{\frac{4}{4+d}}, \quad \text{and} \quad I_f \equiv \int \left(\frac{fTr[A^T \tau A] - (3-d)\alpha Tr[A^T \Lambda A]}{f^{2\alpha+1}} \right)^2 d\vec{x} \tag{46}$$

Similarly to the univariate case the efficiency of kernels can be defined in terms of the Epanechnikov kernel viz

$$eff(K,d) \equiv \left(\frac{\tilde{C}(K_e)}{\tilde{C}(K)} \right)^{\frac{4+d}{d}}. \tag{47}$$

Table 6 shows the efficiency of the Gaussian kernel in table 4 as a function of dimensionality for $d \in [1,5]$. The decline in $\tilde{C}(K)$ of the Gaussian kernel relative to that of the Epanechnikov kernel is somewhat less than the efficiency, with e.g. $\frac{\tilde{C}(K_e,5)}{\tilde{C}K_G,5)} = 0.84$. Since the AMISE is proportional to $\tilde{C}(K)$, the justification for using the Gaussian kernel decrease for increasing d, however, with relatively small impact for $d \lesssim 5$.

d	$eff(K_G, d)$
1	0.95
2	0.89
3	0.82
4	0.75
5	0.68

Table 6: $eff(K,d)$ for Gaussian kernel in table 4.

ROUTE 2: ESTIMATING THE MISE Route two proceeds completely analogously to the univariate case. By taking $H = hS^{\frac{1}{2}}$, $\hat{R}$ (with S given by equation (**??**)) can - by straightforward generalization of equation (28) - be written

$$\hat{R}(h,\alpha) = \int_{\mathcal{R}^d} \left[\frac{1}{nh^d|S|^{\frac{1}{2}}} \sum_{i=1}^{n} (\lambda_i)^{-d} K\big(\lambda_i^{-1} h^{-1} S^{-\frac{1}{2}} (\vec{t} - \vec{x}_i)\big) \right]^2 d\vec{y} - \frac{2}{n} \sum_i \hat{f}_{-i}(\vec{x}_i), \tag{48}$$

now with

$$\hat{f}_{-i}(\vec{x}_i) = \frac{1}{(n-1)h^d|S|^{\frac{1}{2}}} \sum_{j \neq i}^{n} (\lambda_i)^{-d} K\big(\lambda_i^{-1} h^{-1} S^{-\frac{1}{2}} (\vec{x}_j - \vec{x}_i)\big). \tag{49}$$

From equation (48) h, α can be determined by numerically minimizing $\hat{R}$.

Parameter Estimation via the Kullback-Leibler Divergence

Parameter estimation via the Kullback-Leibler divergernce proceeds by straightforward generalizing the univariate case such that

$$\hat{I}(h,\alpha) = \frac{1}{n} \sum_i \log\big(\hat{f}_{-i}(\vec{x}_i)\big). \tag{50}$$

From equation (50) h, α can be determined by numerically minimizing $\hat{I}$.

Example 1.10.

Figure 10 illustrates $TV(n)$ between true densities in two dimensions and i) an adaptive estimate using an approximations of the ideal h and α (blue) ii) an adaptive estimate using $\alpha = \frac{1}{3}$ (eliminating the next to leading order in the bias) and an approximation of the ideal h (black) and iii) an adaptive estimate using $\alpha = 1$ (eliminating the leading order in the bias) and an approximation of the ideal h (brown). All estimates use the Gaussian kernel shown in table 4. Similar to the univariate case the plot consists of six panels each consisting of two figures (being horizontal neighbors). The left figure in each panel is $TV(n)$ whereas the right panel shows the difference in TV between the best estimate in the left panel and the remainder. The only thing that differs between panels is the true density.

The experiment is conducted by taking 25 random samples of the true distribution of size $n \in [20, 500]$ with equidistant sample sizes ($n = 20, 40, 60, \ldots 500$). For each sample the TV is determined via estimates i), ii) and iii). The approximation of the optimal h, α is determined by performing a grid search for the lowest TV.

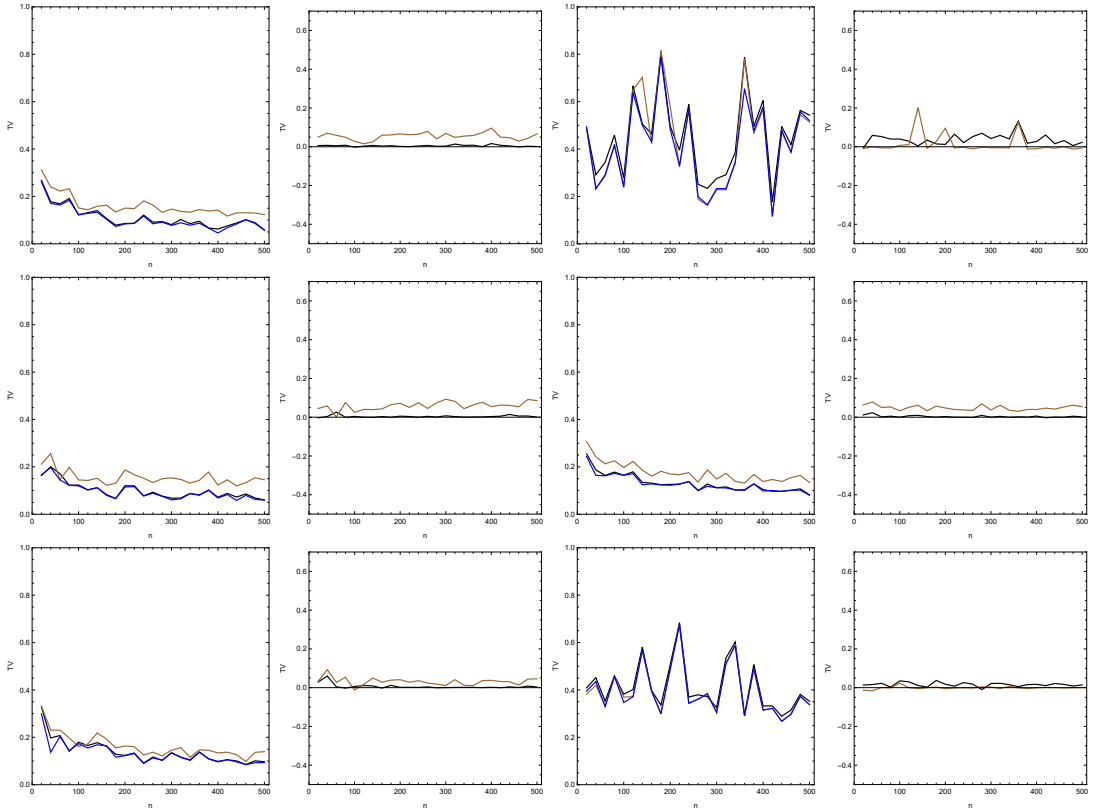

Figure 10: The figure shows $TV(n)$ between different true distributions and different estimates. Blue is an adaptive estimate using an approximation of the optimal h and α (h, α are scanned in a rough way to determine the optimal value), black is an adaptive estimate using an approximation of the optimal h and $\alpha = \frac{1}{3}$ and blue is an adaptive estimate using an approximation of the optimal h and $\alpha = 1$. The true densities are shown in appendix **??**.

From figure 10 several things can be noted; first, in two dimensions choosing α to eliminate the leading order in the bias (i.e. setting $\alpha = 1$ for $d = 2$) results in a worse estimate compared to setting α to eliminate the next to leading order (i.e. setting $\alpha = \frac{1}{3}$ for $d = 2$). This is in contrast to the univariate case where the choice of α was rather unimportant relative to the choice of h. Second, the estimates in the left column and middle right panel all reach $TV \lesssim 0.1$ for $n \to 500$. The top and bottom right panels, however, are significantly worse and contain much more $\sim$random noise of significant magnitude. The $\sim$random noise of significant magnitude - present for all estimates - show that the estimates are highly dependent on the particular sample considered. By extension this means that these distributions are rather extreme in morphology such that consistently drawing representative samples is difficult. The student-T distribution (with the chosen parameters) has an incredibly narrow peak as well as relatively wide tails. In fact the distributions involving the student-T distribution are so extreme that the extra iteration of weighting the data points introduced by the adaptive estimate is not enough to get close to the required morphology. In particular the pilot estimate is far from peaked enough where it should be (the non-adaptive rule of thumb value for h is always used for the pilot estimate). For this reason the h value for the adaptive estimate is required to be very small, meaning that the variance naturally becomes large and so random fluctuations in the estimate from the random samples occur. However, if h is increased to decrease the variance, the peak is not modeled well and so the deviation is large regardless. Hence, it seems the AMISE is simply just high for the distributions in the top right and bottom right panels. This issue could perhaps be alleviated by introducing an extra round of weighting into the adaptive estimate. Investigating this issue further is beyond the scope of this study.

Example 1.11.

Figure 11 illustrates $TV(n)$ between true densities in two dimensions and i) an adaptive estimate using an approximation of the ideal h and α (blue) ii) an adaptive estimate using $\alpha = 0$ and an approximately optimal value for h (cyan) iii) an adaptive estimate with $\alpha = \frac{1}{3}$ and h from table 5 (black) and iv) an adaptive estimate with $\alpha = 0$ and h from table 5 (brown). The experiment and figure is constructed as detailed in example 3.1.

Several things can be noted from figure 11; first, the difference between estimates i) and ii) in terms of TV is very little, meaning that the adaptive ($\alpha \neq 0$) and non-adaptive ($\alpha = 0$) estimates by and large have the same potential in terms of TV. Second, compared to estimate iii), estimate iv) performs significantly worse for panels in the left column and the middle right panel, but significantly better for the top right and bottom right panels. This indicates that estimate iv) uses a relatively low value for h for all distributions. This is an advantage for the more extreme distributions which require so, but a disadvantage otherwise.

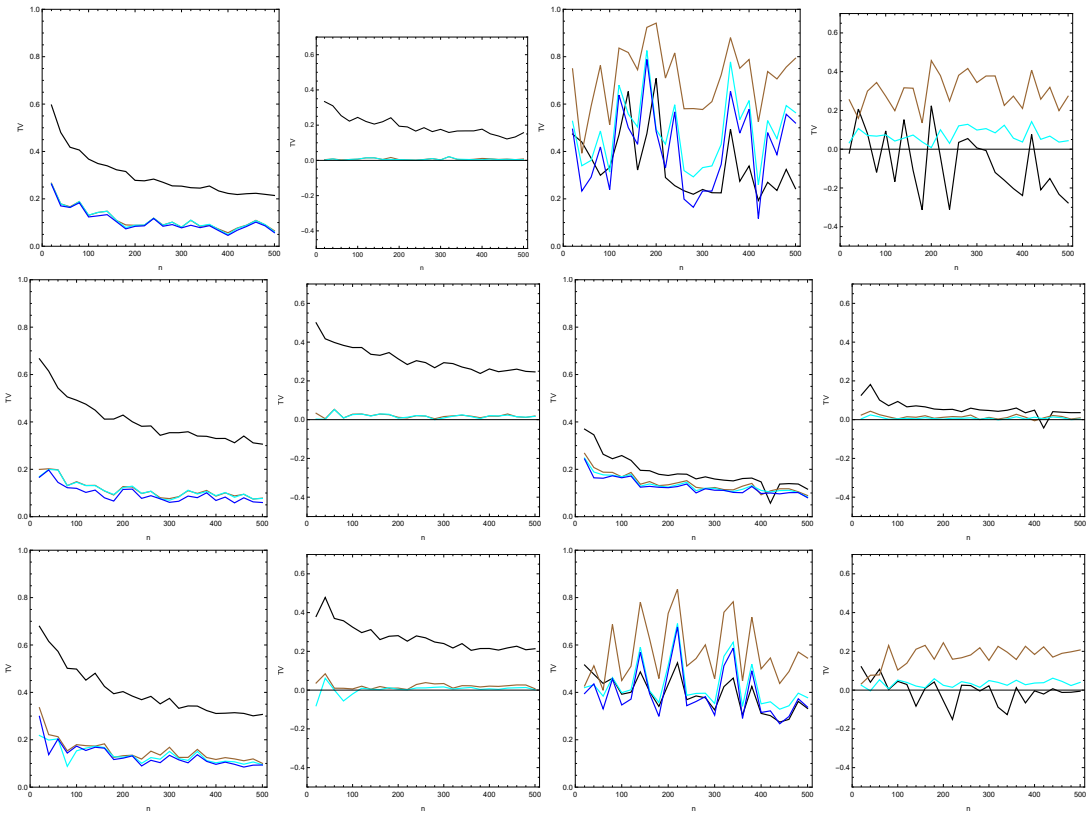

Figure 11: The figure shows $TV(n)$ between different true distributions and different estimates. Blue is an adaptive estimate using an approximation of the ideal h and α, black is an adaptive estimate using $\alpha = \frac{1}{3}$ and h from table 5, cyan is an adaptive estimate with $\alpha = 0$ using an approximation of the ideal h and brown is an adaptive estimate with $\alpha = 0$ using h from table 5. The true densities are the same as for example 3.1.

Example 1.12.

Figure 12 illustrates $TV(n)$ between true densities in two dimensions and i) an adaptive estimate using an approximation of the ideal h and α (blue) ii) an adaptive estimate using $\alpha = \frac{1}{3}$ and h from table 5 (brown), iii) an adaptive estimate using h, α as obtained by minimizing equation (50) *via a grid search (black). The experiment and figure is constructed as detailed in example 3.1.*

As is clear from comparing figure 12 and figure 11; estimate iii) from this example and estimate ii) from example 3.2 is very similar across all considered distributions, however, with estimate ii) from example 3.2 emerging as a slightly less bad estimate in case of the top right and bottom right panels. Because of the similarity of estimate ii) to estimate iii) from example 3.2, this estimate will not be discussed further. The other estimates has been discussed previously.

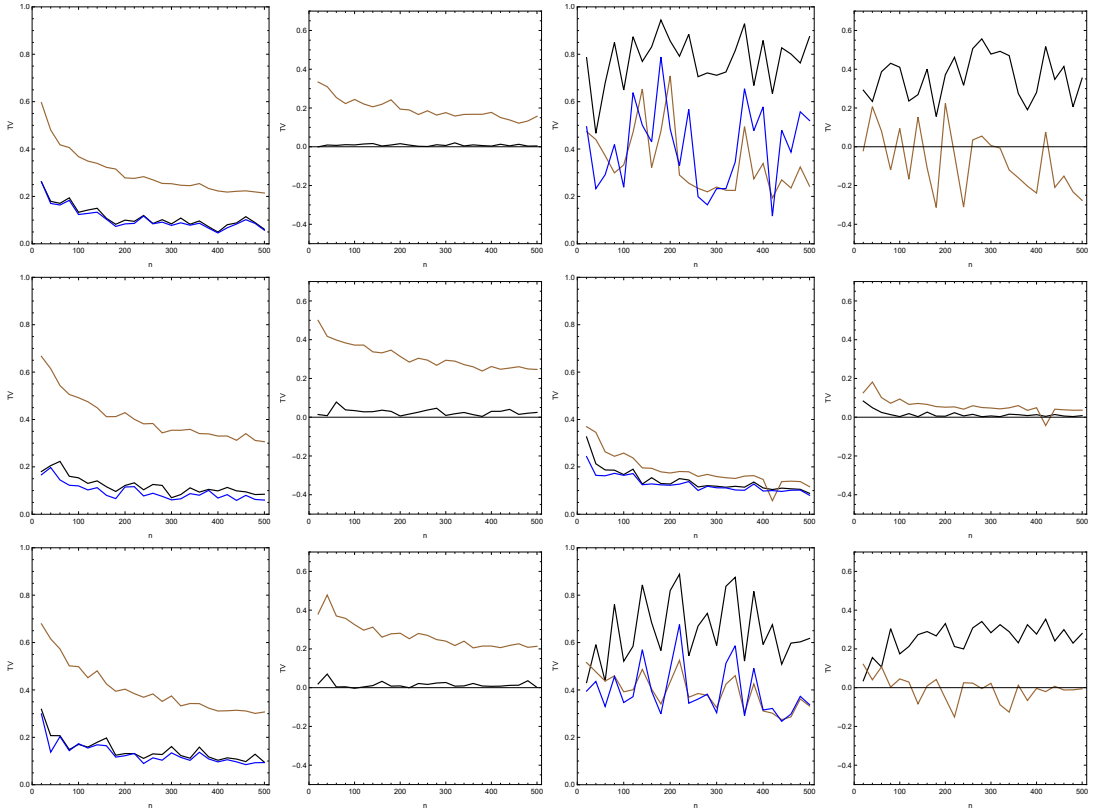

Figure 12: The figure shows $TV(n)$ between different true distributions and different estimates. Blue is an adaptive estimate using an approximation of the ideal h and α, brown is an adaptive estimate using $\alpha = \frac{1}{3}$ and h from table 5 and black is an estimate using h, α as obtained by minimizing equation (50) via a grid search. The true densities are the same as for example 3.1

Example 1.13.

Figure 13 collects all density estimates considered in two dimensions for comparison.

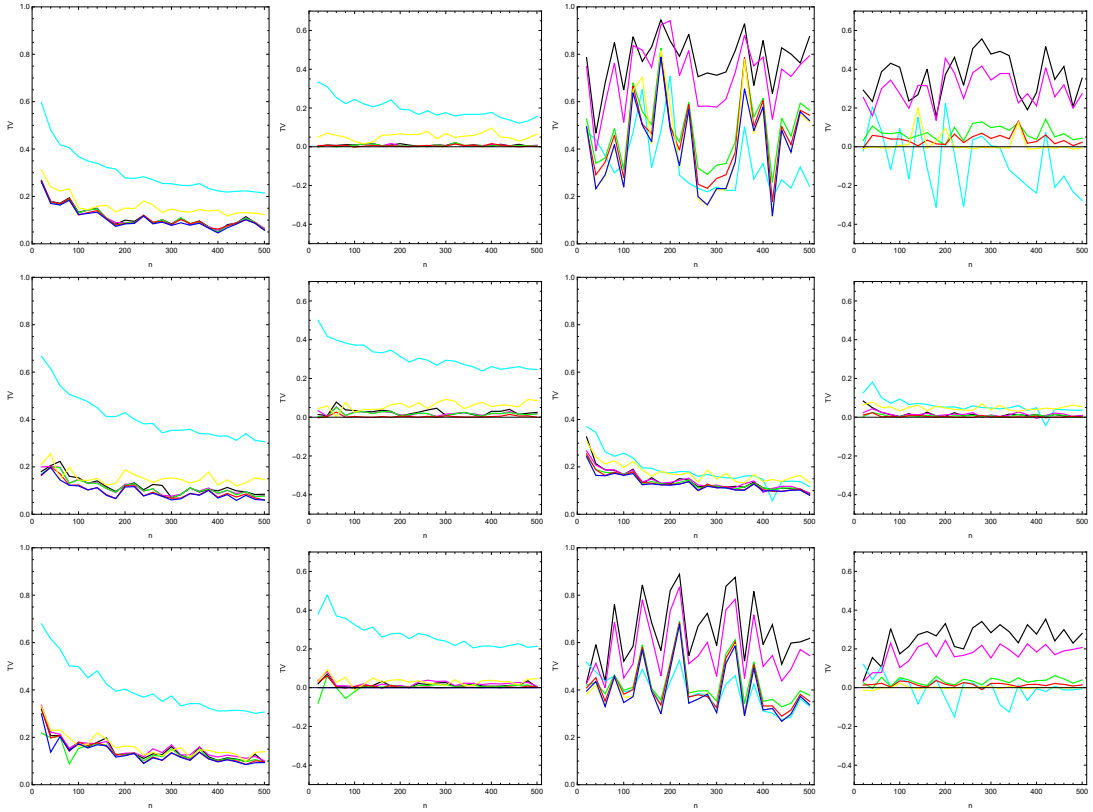

Figure 13: $TV(n)$ between true densities and all estimates reviewed in examples 3.1-3.3. Blue is an adaptive estimate with optimal h, α determined via a grid search. Cyan is an adaptive estimate using $\alpha = \frac{1}{3}$ and h from table 5. Black is an adaptive estimate using h, α as obtained by minimizing equation (50) via a grid search. Magenta is an adaptive estimate with $\alpha = 0$ and h from table 5. Green, red and yellow are adaptive estimates with optimal h determined via grid searches and $\alpha = 0, \frac{1}{3}, 1$, respectively.

Example 1.14.

Figure 14 illustrates $TV(n)$ in three dimensions between true densities and i) an adaptive estimate using an approximation of the ideal h and α (blue) ii) an adaptive estimate using $\alpha = \frac{1}{4}$ and h from table 5 (cyan), iii) an adaptive estimate using $\alpha = 0$ and h from table 5 (brown) iv) an adaptive estimate using h, α as obtained by minimizing equation (50) via a grid search (black) and v) an adaptive estimate using $\alpha = \frac{1}{4}$ and h from table 5 corresponding to $\alpha = 0$. Due to the computational load the experiment is differently than in one and two dimensions; the experiment is conducted by taking 8 random samples of the true distribution of size $n \in [50, 400]$ with equidistant sample sizes ($n = 50, 100, \ldots, 400$). For each sample the TV is determined via estimates i), ii), iii) and iv).

As is clear from figure 14, the three-dimensional results are consistent with the results from two dimensions; the regular estimate with the rule of thumb value for h (estimate iii)) outperforms the adaptive estimate with the rule of thumb for h (estimate ii)) for the distributions not involving the student-T distribution. For the distributions involving the student-T distribution, all estimates perform poorly, but estimate ii) significantly less so. It is curious that the tendencies are the same for two and three dimensions, since in the two-dimensional case the next to leading order in the bias is eliminated whereas in the three-dimensional case the third leading order term in the bias is eliminated (both by choice of α). This suggests that setting α to minimize the higher orders in the bias for $d \geq 2$ may not have a significant impact on the accuracy of the estimated density. On the other hand, from example 3.1 it is clear that with α adjusted to eliminate the next to leading order term, it is possible to obtain approximately ideal estimation. Hence, it is the rule of thumb value for h that is the issue more than the choice of α for estimate ii).

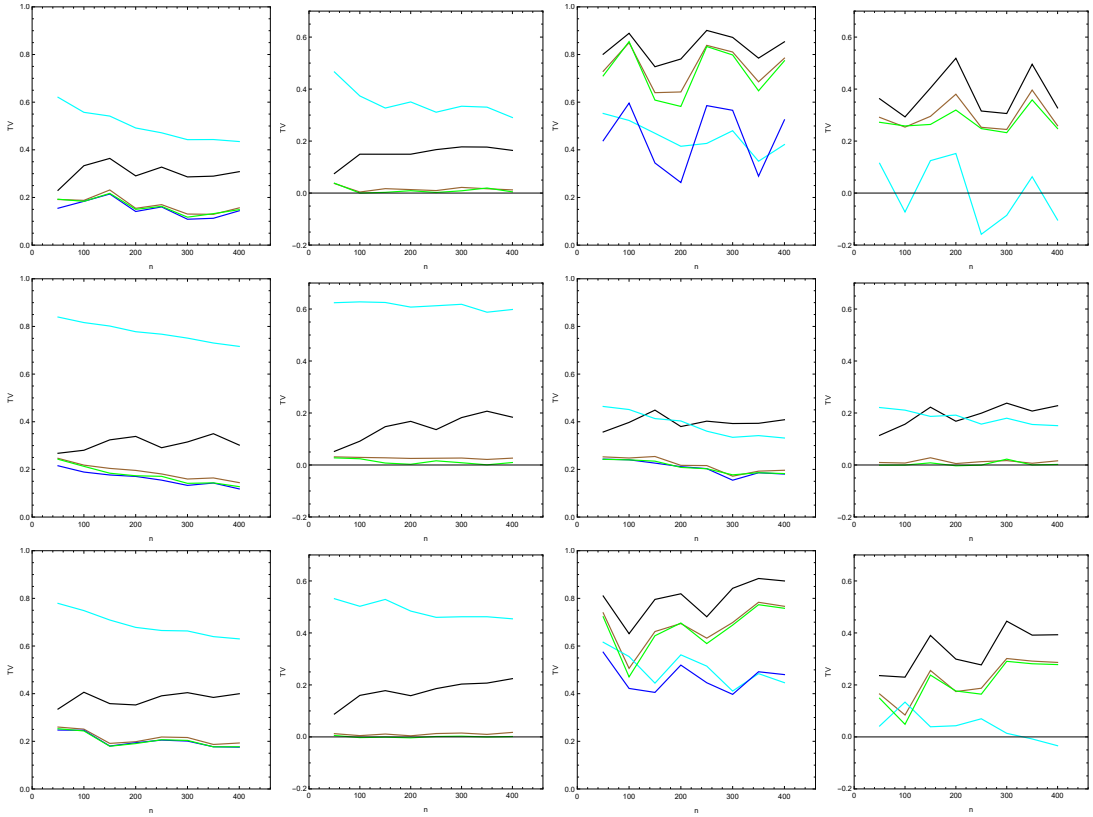

Figure 14: The figure shows $TV(n)$ between different true distributions and different estimates. Blue is an adaptive estimate using an approximation of the ideal h and α. Cyan is an adaptive estimate using $\alpha = \frac{1}{4}$ and h from table 5. Brown is an adaptive estimate using $\alpha = 0$ and h from table 5. Black is an adaptive estimate using h, α as obtained by minimizing equation (50) via a grid search. Green is an adaptive estimate using $\alpha = \frac{1}{4}$ and h from table 5 corresponding to $\alpha = 0$. The difference between different panels is the true density. Top left panel $f \sim \mathcal{N}$ (normal distribution) with $\mu_1 = 1, \mu_2 = 2, \mu_3 = 1, \sigma_{11} = 1, \sigma_{22} = 2, \sigma_{33} = 3, \rho_{12} = 0.3, \rho_{13} = 0.4$ and $\rho_{2,3} = 0.5$. Top right panel $f \sim \mathcal{T}$ (student-T distribution) with one degree of freedom and otherwise the same parameters as the top left panel. Middle left panel $f \sim \mathcal{L}(0,3) \times \mathcal{L}(0,3) \times \mathcal{L}(0,3)$ (logistic distribution). Middle Right panel $f \sim \mathcal{SN}(0,1,5) \times \mathcal{SN}(5,5,10) \times \mathcal{SN}(2,2,10)$ (skewed normal distribution). For the bottom left panel the distribution is a product distribution of the $1:1$ mixtures of different variations of the normal and logistic distributions. For the bottom right panel the distribution is a product distribution of the $1:1$ mixtures of different variations of the sutdent-T and skew normal distribution. For further details, the Reader is referred to the associated Mathematica files.

Naively, one could try to merge the strengths of the adaptive and non-adaptive approaches by taking $\alpha = \frac{1}{3}$ but using the rule of thumb value for h corresponding to $\alpha = 0$ from table 5. This is estimate v) in figure 14. As is clear from the figure, this naive approach actually improves slightly upon estimate iii). In spirit with estimate v) one could also set $\alpha = \frac{1}{2}$ (thereby eliminating the next to leading order in the bias instead of the third leading order) and use the rule of thumb value for h corresponding to $\alpha = 0$ from table 5. This estimate is not shown in figure 14. In terms of TV it is, however, marginally worse than estimate v) for the distributions not involving the student-T distribution and marginally better for the distributions involving the student-T distribution. A similar approach can be applied to the two-dimensional case yielding the same conclusion.
A last note to make from figure 14 is that estimate iv) is worse than estimates iii) and v) for all considered distributions, making this estimate - with the applied coarse grid search for approximately optimal values for h, α - unfavorable. Naturally this conclusion could possibly be changed by increasing the amount of points in the grid search.

1.3 SUMMARY

In this work the formalism of non-parametric density estimation has been introduced and the subject of weight function estimators with associated rules of thumbs for the relevant parameters has been

investigated in depth for both the univariate and multivariate case. Several numerical experiments has been conducted, including some which investigate the total variation between different density estimates and the true density. The total variation as a function of sample size for six different distributions has been used to asses the quality of the different density estimates, including a density estimate with approximations of the ideal parameters and density estimates with varying rules of thumbs and approximations of ideal parameter. Such experiments has been conducted thoroughly for one and two dimensional data and - for computational considerations - less thoroughly for three dimensions. The dimensionality has not been increased beyond three because of the computational cost. From the experiments it has been concluded that tuning the sensitivity parameter (α) of the adaptive estimates is relatively insignificant next to tuning the smoothing parameter (h). For this reason it may be preferential to pick a value of α that come with an associated rule fo thumb value for h. It has been shown that the i'th order term in d dimensions of the bias is proportional to $\alpha = \frac{1}{2i+1-d}$ in general. There is an associated rule of thumb for h if $\alpha = [0, 0.5[$, so as a rule of thumb α can be chosen such to minimize the lowest order in the bias for which this is the case. This strategy (strat1) works well for the univariate case, but less so for the multivariate cases. In the multivariate case it has been shown that choosing α as described above but with the rule of thumb value associated to $\alpha = 0$ yields better results (strat2) which, however, only slightly improve upon the results from taking $\alpha = 0$ with the associated rule of thumb value (strat3) for h (which is a good pick, especially in the multivariate case). Strat2 seems to improve as dimensionality increases. The reason for this reality has not been investigated in an experiment, but it could be related to the pilot estimate which corresponds to an adaptive estimate with $\alpha = 0$. Following **Silverman86** $g = 1$ has been adopted in the AMISE calculations to determine th rule of thumb value for h. Taking $g = 1$ corresponds to ignoring the uncertainty of the pilot estimate. Hence, it is indicated that this assumption breaks down in $d \geq 2$ and by extension that the uncertainty of the pilot estimate should be accounted for in this case. Doing so is beyond the scope of this study.
Another approach to determine h, α is to estimate the loss function itself from the sample - The Kullback-Leibler divergence has been considered for the univariate as well as multivariate case whereas the MISE has been considered for the univariate case - and minimize it via a grid search. This strategy (strat4) yield good results in the univariate case, especially in case of the Kullback-Leibler divergence. However, in the multivariate case strat4 consistently underperform with respect to both strat2 and strat3. It is speculated that the underperformance of strat4 in the multivariate case can be alleviated by increasing the amount of points in the grid search used to determine h, α. This has not been investigated further as it is beyond the scope of this study; a more in depth study of $TV(n)$ - especially for the multivariate case - requires greater computational power than provided by a common laptop.

All in all it is concluded that many different strategies do well in the univariate case, with strat1 and strat4 about equally good with strat 3 slightly worse (strat2 has not been tested for the univariate case). For the multivariate case ($2 \leq d \leq 3$ has been investigated) strat2 emerge as the best option slightly improving upon strat3.