A number–of–modes reference rule for density estimation under multimodality

Jochen Einbeck*, James Taylor
Department of Mathematical Sciences, Durham University,
Science Laboratories, South Road, DH1 3LE Durham City, UK
{jochen.einbeck, james.taylor1}@durham.ac.uk

Abstract

We consider kernel density estimation for univariate distributions. The question of interest is as follows: Given that the data analyst has some background knowledge on the modality of the data (for instance, "data of this type are usually bimodal"), what is the adequate bandwidth to choose? We answer this question by extending Silverman's idea of "normal–reference" to that of "reference to a Gaussian mixture". The concept is illustrated in the light of real data examples.

Keywords and Phrases: Bandwidth selection, kernels, multiple modes, asymptotic mean integrated squared error, Gaussian mixture models.

1 Background

Given i.i.d. replicates $X_1, \ldots, X_n$ of a univariate random variable X with density f and standard deviation σ_X, we consider the kernel density estimator

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right), \tag{1}$$

where K is a kernel function and h is the bandwidth. The estimator (1) was originally proposed in Rosenblatt (1956) and its properties investigated in

*corresponding author

Parzen (1962) and Silverman (1978). The task of selecting h is extremely important in determining the smoothness of the estimate and has been extensively investigated, with many publications covering the subject over the last three decades. A large class of bandwidth selection tools makes, in one form or another, use of the *mean integrated squared error*,

$$\mathrm{MISE}(f, \hat{f}_h) = E \int \{\hat{f}_h(x) - f(x)\}^2 \, dx, \tag{2}$$

though approaches based on other loss functions such as the Kullback–Leibler divergence have also been considered (Bowman, 1984). A well-known technique, which selects h by minimizing an empirically estimated quantity whose expectation is identical to (2), was suggested independently by Rudemo (1982) and Bowman (1984), and is known as least-squares cross-validation (LSCV). An alternative concept, tracing back to Parzen (1962), is to base the bandwidth selection problem on an asymptotic version of (2). For small bandwidths ($h \longrightarrow 0$) and large sample sizes ($nh \longrightarrow \infty$), the MISE approximates

$$D(f)\frac{h^4}{4}\left[\int u^2 K(u)\, du\right]^2 + \frac{1}{nh}\int K^2(u)\, du, \tag{3}$$

where

$$D(f) = \int [f'']^2(x)\, dx \tag{4}$$

is a functional of the density f. Minimizing (3) w.r.t. h yields

$$h_{\mathrm{opt}} = \kappa_0 D^{-1/5}(f) n^{-1/5}\,, \tag{5}$$

where $\kappa_0 = [\int u^2 K(u)\, du]^{-2/5}[\int K^2(u)\, du]^{1/5}$ is a (known) constant only depending on the kernel. Silverman (1986) proposed to approximate the unknown quantity $D(f)$ by the value $D(\phi_{\sigma_X})$ which would be obtained if f was normally distributed with standard deviation σ_X ("normal reference"), i.e.

$$D(\phi_{\sigma_X}) = \int [\phi''_{\sigma_X}]^2(x)\, dx = \sigma_X^{-5}\int [\phi'']^2(x)\, dx = \frac{3}{8\sqrt{\pi}}\,\sigma_X^{-5} \approx 0.212\,\sigma_X^{-5}. \tag{6}$$

[The density ϕ denotes the Gaussian density function, $\phi(x) = (2\pi)^{-1/2}e^{-x^2/2}$, and $\phi_{\sigma_X}(x) = \sigma_X^{-1}\phi(x/\sigma_X)$.] In the special case of a Gaussian kernel K, one has $\kappa_0 = 0.776$, yielding the bandwidth selector

$$h^*_{\mathrm{opt}} = 1.06\,\sigma_X\, n^{-1/5}. \tag{7}$$

An important issue is the estimation of σ_X, a natural candidate for which is the sample standard deviation $s = [n-1]^{-1/2}[\sum(X_i - \bar{X})^2]^{1/2}$, where $\bar{X}$ is the sample mean. An alternative choice is the robust "hybrid" measure of spread, $A = \min(s, IQR/1.34)$, which will usually take its first argument for multimodal, and its second argument for skew data, respectively, in this manner avoiding gross oversmoothing in either case (Silverman, 1986). Refinements of this technique using improved measures of spread were provided by Janssen et al. (1995) and Zhang & Wang (2009).

However, the issue of oversmoothing has not only to do with the spread, but also with the constant 1.06, which stems from the normal reference assumption. To address this problem, Silverman suggested substituting the figure 1.06 generally by the smaller value 0.9, yielding $h_S \equiv 0.9An^{-1/5}$, without justifying this specific choice of constant further. Intuitively, the more modes the data are expected to have, the smaller the bandwidth has to be relative to the standard deviation in order to enable an adequate degree of resolution. This paper addresses the question of *how much smaller* the bandwidth should be chosen, given some prior anticipation on the modality, for instance based on expert knowledge. Eventually, we seek a function, say $c(m)$, so that, given a prior notion on the number m of modes, a suitable bandwidth is found by

$$h_m^* = c(m)sn^{-1/5}. \tag{8}$$

Of course, s could again be replaced by A herein. Since this paper focuses on the problem of density estimation under multimodality, in which case one will generally have $A = s$, we refrain from this modification for the sake of simplicity.

The remainder of this article is organized as follows. In Section 2 we attempt to quantify the necessary reduction of the bandwidth under multimodality by replacing the concept of "normal reference" with that of "reference to a Gaussian mixture". The technique is worked into a simple rule of thumb in Section 3. A small simulation study is provided in Section 4, before we finish with a Discussion in Section 5.

2 Reference to a Gaussian mixture

We work in this section with a general, not necessarily Gaussian, kernel function K, and consider (5) as the starting point of our analysis. Obviously, the crucial quantity in this expression is $D(f)$. If the data are multimodal,

the normal reference rule will underestimate $D(f)$, hence overestimate h. Hence, we attempt to approximate $D(f)$ more accurately, and this can be achieved by making reference to a mixture

$$\varphi_m(x) \equiv \varphi(x|\theta_m) \equiv \sum_{k=1}^{m} \pi_k \phi_{\sigma_k}(x - \mu_k)$$

of m normal densities $\phi_{\sigma_k}(x - \mu_k)$ centered at locations μ_k, with standard deviations σ_k, and associated mixture probabilities π_k, $k = 1, \ldots, m$, bundled into a parameter vector $\theta_m = \{\pi_k, \mu_k, \sigma_k\}_{1 \le k \le m} \in \mathbb{R}^{3m}$. [These are effectively only $3m - 1$ parameters since $\sum_k \pi_k = 1$.] All these quantities can be estimated straightforwardly through the EM algorithm (Laird, 1978) using standard software (we used Einbeck et al., 2007), yielding estimates $\hat{\theta}_m = \{\hat{\pi}_k, \hat{\mu}_k, \hat{\sigma}_k\}_{1 \le k \le m}$. The estimated mixture density is then given by

$$\hat{\varphi}_m(x) = \varphi(x|\hat{\theta}_m) = \sum_{k=1}^{m} \hat{\pi}_k \phi_{\hat{\sigma}_k}(x - \hat{\mu}_k)$$

and the corresponding integral $D(\hat{\varphi}_m)$ can be calculated exactly using Theorem 4.1 of Marron & Wand (1992),

$$D(\hat{\varphi}_m) = \sum_{k=1}^{m} \sum_{\ell=1}^{m} \hat{\pi}_k \hat{\pi}_\ell \phi^{(iv)}_{\sqrt{\hat{\sigma}_k^2 + \hat{\sigma}_\ell^2}}(\hat{\mu}_k - \hat{\mu}_\ell), \tag{9}$$

or computed numerically using software such as Mathematica. [The notation (iv) signifies a fourth derivative.] Hence, one may approximate (5) by

$$h_m = \kappa_0 D^{-1/5}(\hat{\varphi}_m) n^{-1/5}\,. \tag{10}$$

We illustrate this criterion by means of two simple real data examples. Firstly, we consider data featuring the log-energy consumption, in kg oil equivalent per capita in the year 2007, for a sample of $n = 135$ countries (retrieved from the World bank data base[1]). A rug plot of the data is provided in Figure 1, with several particular countries highlighted for the sake of interest. One observes that the world is essentially divided into two major clusters in this respect; corresponding to the so-called developing and developed countries, respectively. In fact, this bimodal structure has been

[1] http://data.worldbank.org/indicator/EG.USE.PCAP.KG.OE

prevalent for many years already. Though the gap has started to become closer in recent years, it would still be appropriate for an expert to assume, as a working assumption, that data of this type possess two distinct modes. The resulting density estimate, using a Gaussian kernel with h_2 according to (10), is provided in Figure 1 (left), along with the densities obtained using h_1 (being identical to (7) with σ_X estimated by s), and h_S. One observes that, as expected, h_2 resolves the bimodal structure the most, providing the deepest dip between the two modes. The estimated density would remain bimodal, but with a yet more pronounced dip, by using h_3 or h_4 according to Table 1, but as there is not much of a justification for the use of such bandwidths for these data, we abstain from providing the corresponding estimated density curves.

Secondly, $n = 876$ measurements of traffic flow (in vehicles/5 min.) were taken from 10–12/07/07 on a Californian freeway (retrieved from PeMS[2]). Traffic engineers will have some notion that such data tend to have *at least* two distinct modes, one corresponding to freeflow, and another one to busy, possibly congested, traffic. Figure 1 (right) shows the estimated density curves using $h_1, \ldots, h_4$. One observes that, using h_1, the estimated density is in fact bimodal, but still appears oversmoothed. Anticipating $m = 2$ modes resolves the structure better, and unveils a *third* mode for small flow values, which can be traced back to a period of unusual activity on 12/07 between 2 and 3am. Anticipating $m = 3$ modes gives an indication of a potential fourth and fifth mode for flow values of around 70 and 125 veh/5 min, respectively, but going beyond $m > 3$ leads to a clearly overfitted result. A complete breakdown of the values of $D(\hat{\varphi}_m)$, for $m = 1, \ldots, 4$, as well as the resulting bandwidths h_m, is provided in Table 1, for both datasets.

3 Rule of thumb

In practice, it is impractical to fit a Gaussian mixture just for the sake of kernel density bandwidth selection. Firstly, the fitted mixture constitutes a density estimate in its own right already. Secondly, the task of estimating the mixture and computing $D(\hat{\varphi}_m)$ is quite laborious. Thirdly, as pointed out by Jones (2000), for the estimation of $D(f)$, a certain degree of oversmoothing may even be beneficial. As seen for the traffic flow data, the integral $D(\hat{\varphi}_m)$ can depend sensitively on the value of m, especially if m is misspecified.

[2] http://pems.dot.ca.gov/

Therefore, it would be desirable to produce a simple rule of thumb based on the ideas from Section 2, which does not require the actual fitting of the mixture, and is robust (to some degree) to misspecification of m.

We approach this objective by making some simplifying assumptions. We restrict the shape of the mixture density to an equal mixture of m normal densities with standard deviation σ, which are placed at equidistant locations $\mu_k, k = 1, \ldots, m$. Given these assumptions, and noting that the integral over the squared second derivatives is a location invariant functional, we can write the position of the locations w.l.o.g. as $\mu_k = kd\sigma$, with a distance parameter d. It remains a simplified parameter vector $\theta_m^* = \{\sigma, d, m\}$, and the mixture density takes the form

$$\varphi_m^*(x) = \varphi(x|\theta_m^*) = \frac{1}{m}\sum_{k=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{x - kd\sigma}{\sigma}\right)^2\right\}. \tag{11}$$

Using lengthy but otherwise straightforward algebra (see appendix), one derives

$$D(\varphi_m^*) = \frac{3}{8\sqrt{\pi}m\sigma^5}\left[1 + F(m,d)\right], \tag{12}$$

where

$$F(m,d) = \frac{2}{m}\sum_{s=1}^{m-1}(m-s)e^{-\frac{d^2s^2}{4}}\left[1 - s^2d^2 + \frac{s^4d^4}{12}\right].$$

[In the special case $m = 2$, an equivalent formulation of this result was provided by Zhang and Wang (2009).] Substituting (12) into the expression for h_{opt}, (5), one obtains

$$h_{\text{opt}} = \kappa_0\left(\frac{8\sqrt{\pi}}{3}\right)^{1/5} m^{1/5}n^{-1/5}\sigma\left[1 + F(m,d)\right]^{-1/5}. \tag{13}$$

It is important to recall here that σ is the *component* standard deviation, which is different from the *overall* standard deviation, previously denoted by σ_X. However, simple algebra shows that for a random variable X with mixture density (11), one has

$$\sigma_X^2 = \text{Var}(X) = \sigma^2\left(1 + (m^2 - 1)\frac{d^2}{12}\right). \tag{14}$$

So, σ^2 can be estimated by $s^2/(1 + (m^2 - 1)\frac{d^2}{12})$, where s is the overall sample standard deviation. Substituting this into (13), and using now $\kappa_0 = 0.776$

for a Gaussian kernel, yields

$$h_{\text{opt}} = 1.06 m^{-\frac{4}{5}} n^{-\frac{1}{5}} s \frac{2\sqrt{3}}{d\sqrt{1 + (\frac{12}{d^2} - 1)/m^2}\,[1 + F(m,d)]^{\frac{1}{5}}}. \tag{15}$$

This is still a bulky expression, which involves the unknown quantity d, which the practising data analyst will not want to estimate. Hence, a practicable default choice is needed. If one takes $d = 2\sqrt{3}$, which is a reasonable assumption as it means a slight overlap of distributions (see Figure 2 left), with clearly distinguishable modes, then (14) boils down to the simpler form $\sigma_X = m\sigma$. Furthermore, it is worth looking at the surface $F(m,d)$, which is provided in Figure 2 (right). One observes that, for $d \geq 1/2$ one has strictly $|F(m,d)| < 1$, and that in fact $F(m,d) \approx 0$ for a wide range of values of m and d. For the special choice $d = 2\sqrt{3}$, one has $F(2, 2\sqrt{3}) = 0.050$, $F(3, 2\sqrt{3}) = 0.067$, and $F(4, 2\sqrt{3}) = 0.076$, all of which are fairly close to zero. [For $d < 1/2$, values of $|F(m,d)| > 1$ can be observed, but these are irrelevant for our purposes since, realistically, we are only interested in $d > 2$, with $d = 2$ being the largest value of d for which the two normals just don't separate.] In addition, it should be noted that $F(m,d)$ enters into the equation only in terms of a fifth root, so that we can effectively assume $[1 + F(m,d)]^{\frac{1}{5}} \approx 1$. Performing all these simplifications in (15), the expression for the optimal bandwidth simplifies significantly, and becomes

$$h_m^* = 1.06 m^{-\frac{4}{5}} s n^{-\frac{1}{5}}. \tag{16}$$

This gives a simple rule of thumb, $c(m) = 1.06 \times m^{-4/5}$, which, just as the normal reference rule (7), only makes use of the spread of the data, and differs from this one merely by the factor $m^{-4/5}$.

Table 2 gives the resulting factors at a glance. Looking at the row for $c(m)$ one observes that, except for $m = 1$, all values are significantly smaller than Silverman's constant $c(m) \equiv 0.9$. Silverman's objective was to provide *one* constant which serves reasonably well for any modality, accepting that it will "slightly oversmooth" for $m \geq 2$. If one's prior belief distribution on the expected modality has a strong weight on $m = 1$, then a factor of 0.9 still seems to be in line with the results from Table 2.

Before investigating its performance more thoroughly in Section 4, we apply this rule–of–thumb tentatively on the two real data sets introduced in Section 2. Figure 3 (left) compares the normal reference bandwidth $h_1 = h_1^*$

with the bandwidths h_2 and h_2^*. We see that h_2 and h_2^* do not differ strongly and yield similar densities, with the latter one yielding a slightly more pronounced dip. Figure 3 (right) is the analogous image to Figure 1 (right) but using now the rule–of–thumb. We see that both bandwidth selectors yield very similar results, but with the rule–of–thumb method behaving less temperamentally for higher values of m. The numeric values of all used bandwidths are provided in Table 1.

4 Simulation study

We have carried out a simulation study in order to investigate the efficiency of the proposed rule of thumb. Before explaining the setup of the study, it is important to clarify what the technique is supposed to achieve. Crucially, the objective is *not* to reproduce the anticipated number of modes. For instance, when setting $m = 1$, the objective is clearly not to obtain a unimodal density estimate, but to obtain the *best* density estimate based on the reference to a unimodal distribution. The quality of a density estimate can be measured by an empirical version of (2),

$$\mathrm{MSE}(f, \hat{f}_h) = \frac{1}{N} \sum_{i=1}^{N} \{\hat{f}_h(z_i) - f(z_i)\}^2, \tag{17}$$

where $z_1, \ldots, z_N$ is an appropriate set of grid points. The question that we investigate in this study is, hence:

Given that the data are generated from a distribution of known modality, does one achieve the best MSE when exactly this number of modes is used for m in (16)?

We will see below that the answer to this question turns out to be 'yes' throughout. In what follows we will work with a grid of size $N = 200$, ranging from $z_1 = \min_k\{\mu_k - 3\sigma_k\}$ to $z_N = \max_k\{\mu_k + 3\sigma_k\}$. We begin with data simulated from an "ideal" scenario, i.e. data from an equal and equidistant mixture of m Gaussians with equal standard deviation $\sigma = 1$ and distance $d = 2\sqrt{3}$. That is, the rule–of–thumb is in this case exact and produces precisely the asymptotically optimal bandwidth. Figure 4 (left) shows the mixture densities (a)–(d) for $m = 1, 2, 3$ and 4 components. 200 data sets

of size $n = 500$ are generated from each of (a) to (d), and the densities are estimated using rule (16), each for different values for m.

Figure 4 (right) gives the resulting MSEs, where the value of m used in the rule–of–thumb is provided in the horizontal axis label. For comparison, Silverman's rule h_S is also included and symbolized by an S. We observe that, for all of (a) to (d), the MSEs tend to be minimal when the modality was correctly anticipated. Table 3 provides additionally the percentages of times that each value of m led to the winning MSE (the bandwidth h_S is excluded from this analysis). Clearly, using the correct choice of m leads to the best MSE, and deviating from this in either direction deteriorates the fit.

We proceed with investigating more complex scenarios in which rule (16) is *indeed* only a rule–of–thumb. Graphs of the densities (e)–(h) used for this simulation are provided in Figure 5 (left). The precise specifications from which these densities are generated are provided in Table 4. One observes from Figure 5 (right) that, even under this harder scenario, the rule of thumb does a good job in selecting the bandwidth, and at all occasions we achieve the best MSEs when the correct modality is anticipated. This is confirmed by considering the lower part of Table 3.

Some comments concerning density (e) are in order. Firstly, this density highlights that the number of mixture components is generally just an upper bound for the number of modes. Secondly, this example demonstrates that for use in the rule of thumb (16), it is really *the number of modes* rather than *the number of mixture components* which matters. One further observes from the two top right panels in Figure 5 that, for densities (e) and (f), Silverman's bandwidth h_S, using the hybrid measure of spread A, works quite well. Indeed, if h_S had been included in the comparison for these densities in Table 3, then this bandwidth would have won in 40% and 36%, respectively, of the cases. This is actually not surprising since these densities are quite skew, and the IQR component of the hybrid measure A was introduced precisely to serve this case. The proportions of wins for h_S drop to 30% and 10%, respectively, for densities (a) and (b), and to 0% when the underlying density was at least trimodal (not shown).

5 Discussion

Extending Silverman's idea of normal reference towards the "reference to a Gaussian mixture", we have provided a simple rule of thumb for density

estimation under multimodality. The application of this rule requires the specification of an "anticipated" modality. As pointed out by M. Aitkin at occasion of the Conference of Applied Statistics in Ireland 2011, this aspect entails the danger of circularity: If a density estimate (such as a histogram or kernel density estimate) is used to become informed about the modality, this modality will depend on the initial smoothing parameter used. In fact, if one iterates this procedure, one is likely to end up with ever decreasing bandwidths, and an ever increasing number of modes, which is obviously unacceptable. To avoid such circularity, it is important that the "anticipated modality" stems from an external source such as prior knowledge, expert opinion, etc. We have provided two real data examples in which it was realistic to assume that such knowledge is available, and we believe that it is realistic to have such prior information in a wider range of applications.

We have found that there is no need to estimate the actual mixture parameters, since a simple approximation based on an "idealized" mixture performs equally well, and tends to behave in a more stable manner for higher numbers of modes. We have shown in the simulation study that the concept of modality-dependent bandwidths is sensible: Using the "true" modality in the rule–of–thumb has led to minimal MSEs under all investigated scenarios.

We have seen that, for use in the rule–of–thumb, it is the number of modes rather than the number of mixture components which matters. Expert opinion on the modality will often be motivated by the presence of several groups, subpopulations or "components" which drive the data-generating process. Though this provides a reasonable starting point for the choice of m in (16), one should be aware that the actual number of modes could be smaller than the number of mixture components. It should also be pointed out that, even though the concept of a mixture may be a reasonable surrogate, the data–generating mechanism may have worked very differently. For instance, the traffic flow data originally form a time series of clearly non–independent character. In fact, it is the dependence (cases closely together in time are likely to belong to the same cluster) which induces the multimodality in this example. As a working assumption, it still seems acceptable to think of these data as i.i.d. realizations from a two–component mixture structure, in conformity with the setup outlined at the beginning of the Introduction.

Summarizing, we believe that we have formulated a very simple tool for bandwidth selection for multimodal distributions, which operates by multiplying the normal reference rule by $m^{-4/5}$, where m is the anticipated number of modes.

Appendix

Derivation of (12)

Using equation (9) for the special case of a mixture density of type (11), one has

$$D(\varphi_m^*) = \frac{1}{m^2}\left[\sum_{k=1}^{m}\phi_{\sigma\sqrt{2}}^{(iv)}(0) + 2\sum_{s=1}^{m-1}(m-s)\phi_{\sigma\sqrt{2}}^{(iv)}(sd\sigma)\right] \tag{18}$$

Simple calculus shows that

$$\phi_\sigma^{(iv)}(x) = \frac{1}{\sqrt{2\pi}\sigma^5}\left[3 - 6\frac{x^2}{\sigma^2} + \frac{x^4}{\sigma^4}\right]e^{-\frac{x^2}{2\sigma^2}}$$

so that $\phi_{\sigma\sqrt{2}}^{(iv)}(0) = 3/(8\sqrt{\pi}\sigma^5)$ and

$$\phi_{\sigma\sqrt{2}}^{(iv)}(sd\sigma) = \frac{3}{8\sqrt{\pi}\sigma^5}\left[1 - s^2d^2 + \frac{1}{12}s^4d^4\right]e^{-\frac{s^2d^2}{4}}.$$

Plugging these into (18) gives

$$D(\varphi_m^*) = \frac{3}{8\sqrt{\pi}m\sigma^5}\left[1 + \frac{2}{m}\sum_{s=1}^{m-1}(m-s)\left(1 - s^2d^2 + \frac{s^4d^4}{12}\right)e^{-\frac{d^2s^2}{4}}\right],$$

which is (12).

Acknowledgements

The second author was supported by an EPSRC DTA scholarship. We wish to thank one of the anonymous referees for a useful suggestion which enhanced the validity and the results of our simulation.

References

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

Einbeck, J., J. Hinde and R. Darnell (2007). A new package for fitting random effect models – The **npmlreg** package. *R News* **7**, 26–30.

Janssen, P., J.S. Marron, N. Veraverbeke and W. Sarle (1995). Scale measures for bandwidth selection. *Journal of Nonparametric Statistics* **5**, 359–380.

Jones, M.C. (2000). Rough-and ready assessment of the degree and importance of smoothing in functional estimation. *Statistica Neerlandica* **54**, 37–46.

Laird, N.M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, **73**, 805–811.

Marron, J.S. and M.P. Wand (1992). Exact mean integrated squared error. *Annals of Statistics*, **20**, 712–736.

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

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

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

Silverman, B.W. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *Annals of Statistics* **6**, 177–184.

Silverman, B.W. (1986). *Density Estimation*. Chapman and Hall, London.

Zhang, J. and X. Wang (2009). Robust normal reference bandwidth for kernel density estimation. *Statistica Neerlandica* **13**, 13–23.

Table 1: Overview of results for the energy use and the traffic flow data.

Data set	m	1	2	3	4
Energy use	$D(\hat{\varphi}_m)$	0.151	0.961	2.93	2.98
	h_m	0.425	0.291	0.235	0.234
	modes observed	2	2	2	2
	h_m^*	0.425	0.244	0.177	0.141
	modes observed	2	2	3	4
Traffic flow	$D(\hat{\varphi}_m)$	6.24e-10	1.75e-08	8.44e-08	1.74e-06
	h_m	13.89	7.13	5.20	2.84
	modes observed	2	3	5	8
	h_m^*	13.89	7.97	5.77	4.57
	modes observed	2	3	3	5

Table 2: Multimodal correction factor $m^{-4/5}$ for $m = 1, ..., 8$ modes.

m	1	2	3	4	5	6	7	8
$m^{-4/5}$	1.000	0.574	0.415	0.330	0.276	0.238	0.211	0.189
$c(m)$	1.060	0.609	0.440	0.350	0.293	0.253	0.223	0.201

Table 3: Out of 200 simulations, percentage of times that the minimal MSE is achieved when anticipating m modes.

	m					
density	1	2	3	4	5	6
(a)	**91**	9	0	0	0	0
(b)	2	**87**	11	0	0	0
(c)	0	11	**77**	12	0	0
(d)	0	0	24	**64**	12	0
(e)	**86**	14	0	0	0	0
(f)	12	**86**	2	0	0	0
(g)	0	3	**52**	42	3	0
(h)	0	0	15	**58**	27	0

Table 4: Specification of the mixture parameters used to generate densities (e)–(h).

density	m	μ_k	σ_k	π_k
(e)	2	0, 1	1, 0.5	0.8, 0.2
(f)	2	0, 0.7	0.2, 0.4	0.4, 0.6
(g)	3	0, 2, 3	0.8, 0.3, 0.3	0.1, 0.4, 0.5
(h)	4	0, 1, 2, 3	0.3, 0.3, 0.3, 0.3	0.2, 0.3, 0.1, 0.4

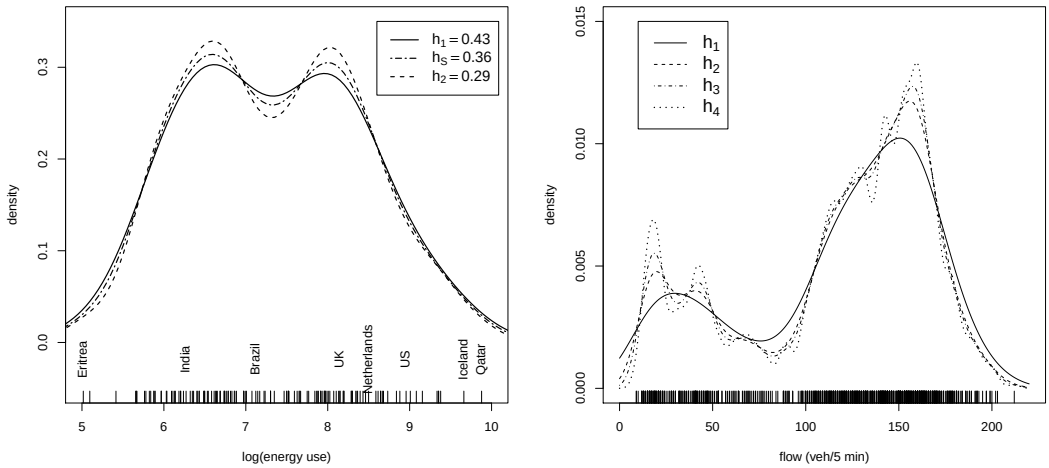

Figure 1: Left: energy consumption per capita in 2007, and density estimates using h_1, h_2, and h_S; right: estimated densities for traffic flow data using h_j, $j = 1, \ldots, 4$; each with rug plots providing the raw data.

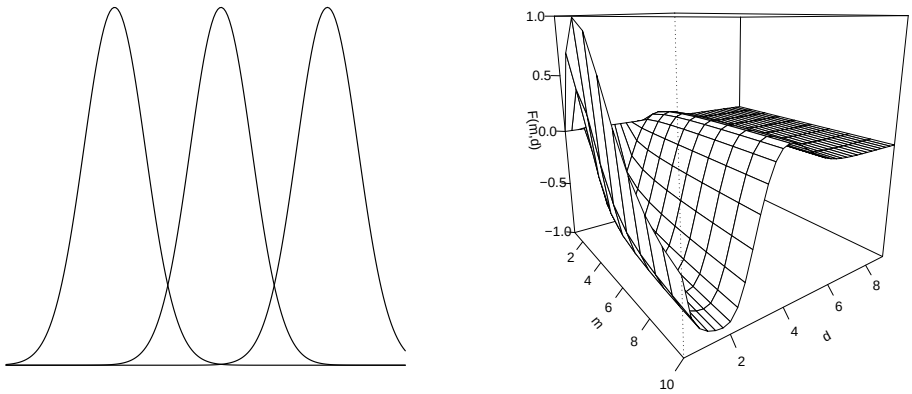

Figure 2: Left: $m = 3$ normals, each separated by $d = 2\sqrt{3}$ standard deviations; right: the surface $F(m, d)$ for $1 \leq m \leq 10$, $0.5 \leq d \leq 9$.

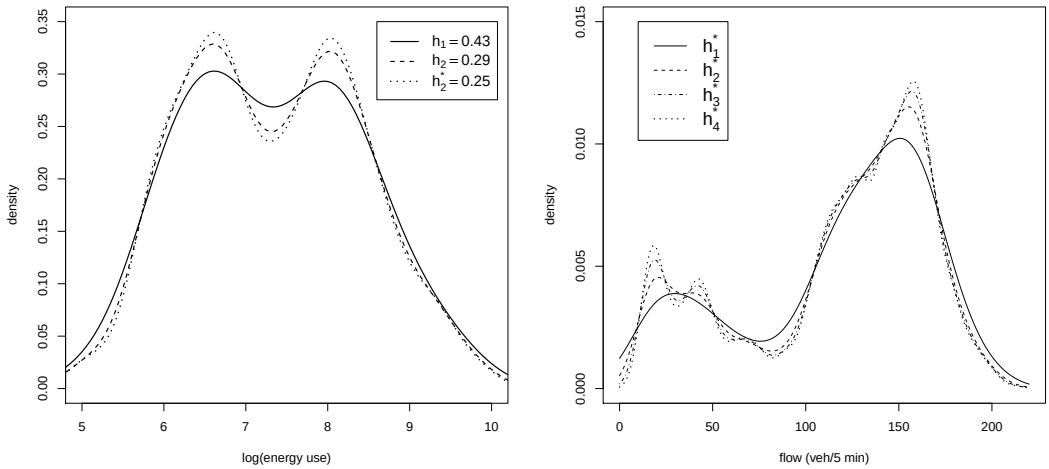

Figure 3: Rule–of–thumb applied to real data sets. Left: estimated density of energy data using bandwidth h_2^*, in comparison with "exact" mixture-based bandwidth h_2, and the normal reference bandwidth $h_1 = h_1^*$; right: estimated densities of traffic flow data using bandwidths $h_j^*, j = 1, \ldots, 4$.

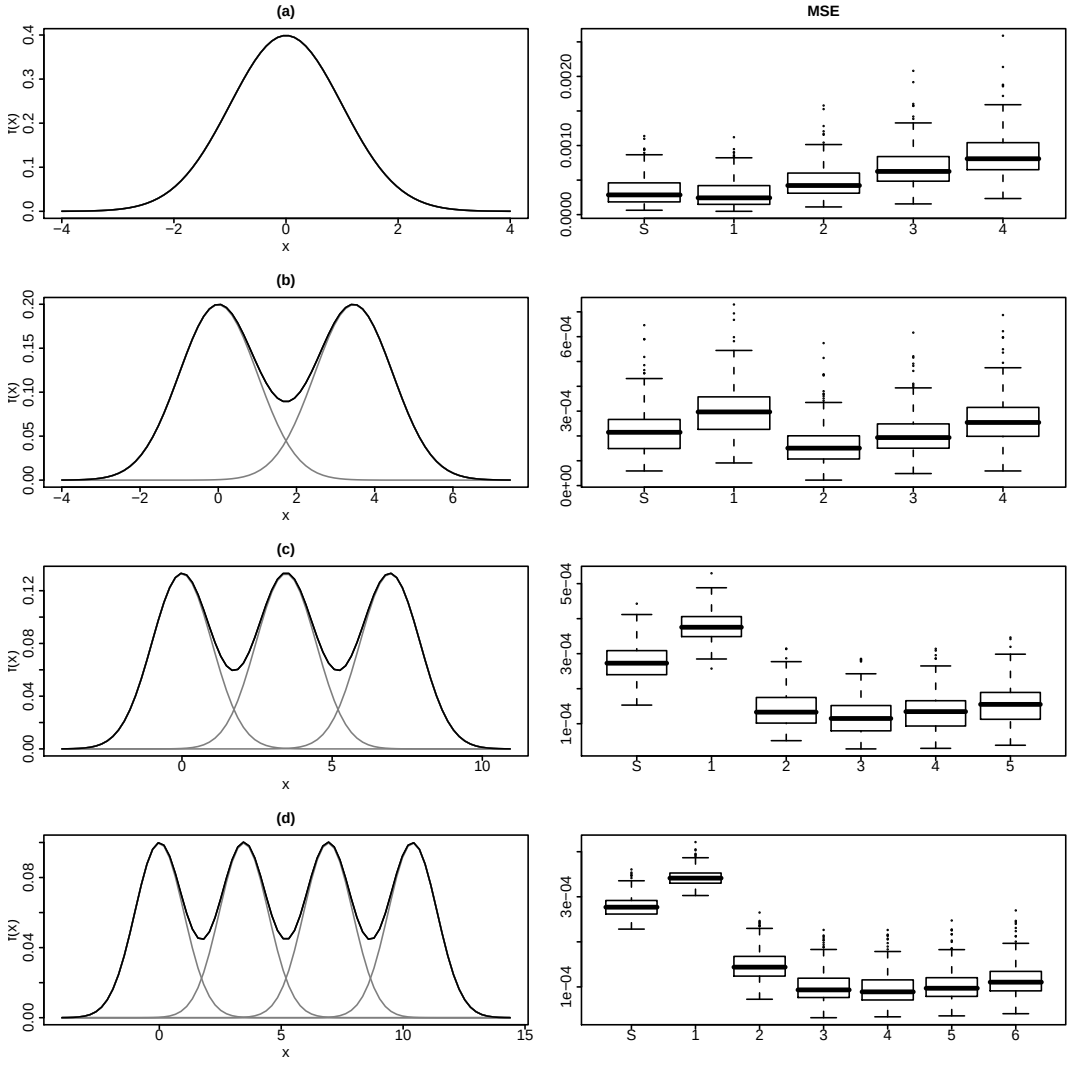

Figure 4: Left: generating densities (black) with probability-weighted component densities (grey); right: boxplots of MSEs for different bandwidths, for scenarios (a) to (d).

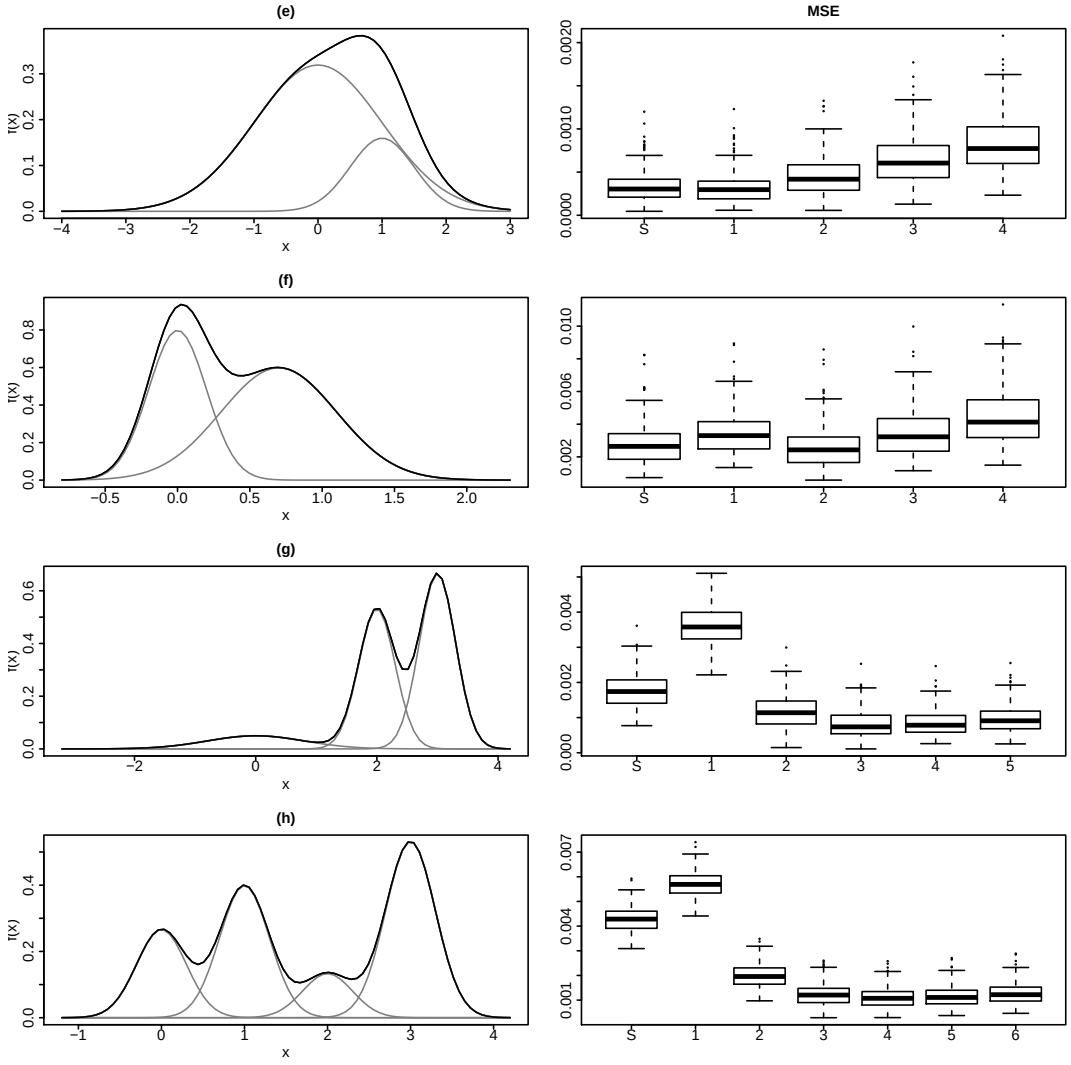

Figure 5: Left: generating densities (black) with probability-weighted component densities (grey); right: boxplots of MSEs for different bandwidths, for scenarios (e) to (h).