Fundamental Kernel Density Estimation for Computation

Michael Liou [*]
Department of Statistics, University of Wisconsin, Madison

May 11, 2018

Abstract

We cover fundamental principles of the Kernel Density Estimate (KDE) in a computational setting. Influence of kernel choice and bandwidth tuning on the density estimates are illustrated. Asymptotic bounds and decision theoretic frameworks are presented for the KDE. Computational complexity and efficiency of the Fast Fourier Transform (FFT) procedure in KDE calculation are shown, and bandwidth choices are also shown to nearly attain the asymptotic lower bound computationally. Multivariate extensions of the univariate first principles are briefly mentioned.

Keywords: nonparametric, statistics, fast fourier transform, curse of dimensionality

[*] All code is available at www.github.com/lioumens/stat771_final

1 Introduction

Density estimation is a useful, nonparametric way to estimate the underlying probability density function. This is common in any scenario in which we are trying to determine a pattern of the underlying random process from an observed dataset. Density estimation can provide insight to the pattern of the underlying distribution, especially when the underlying distribution is complicated. Here, we explore some of the fundamental computational results that have shaped the univariate density estimation. We then extend some of these principles into a multidimensional setting and look at how current research areas have addressed sparsity issues and parameter tuning in higher dimensions. We illustrate these results with a mix of mature code bases and self-coded crude approximations.

2 Theoretical Background

2.1 Estimators

The kernel density estimator with bandwidth $h > 0$ and kernel K is defined as

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

There are two parameters to specify for this estimator, the kernel K and the bandwidth h. It is well known from theory and practice that the kernel does not have a significant on the final density estimate. On the other hand, the bandwidth will have a large effect on the overall fit of the estimate. The bandwidth determines the scaling parameter of the kernel. Visually, the bandwidth determines how "squiggly" the estimate looks. The KDE estimate at one particular point is the sum of the Gaussian kernel at every point of evaluation.

The difficulty with kernel density estimation is the varying regions of high density and low density sample regions, in which low bandwidth more appropriate for low density regions and high bandwidth is more appropriate for high density regions. Thus, many other estimators that build off of this basic idea have been proposed (Silverman 2018, Mack & Rosenblatt 1979*a*, Silverman 1982*b*, Mack & Rosenblatt 1979*b*).

Figure 1: Comparison of different bandwidths

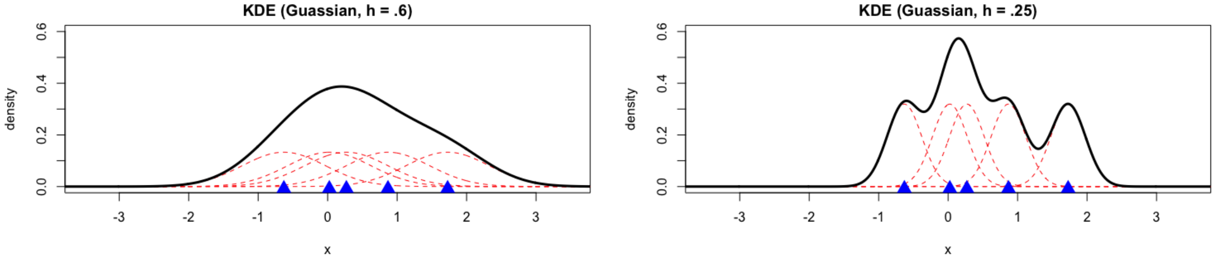


Figure 2: Comparison of different kernels

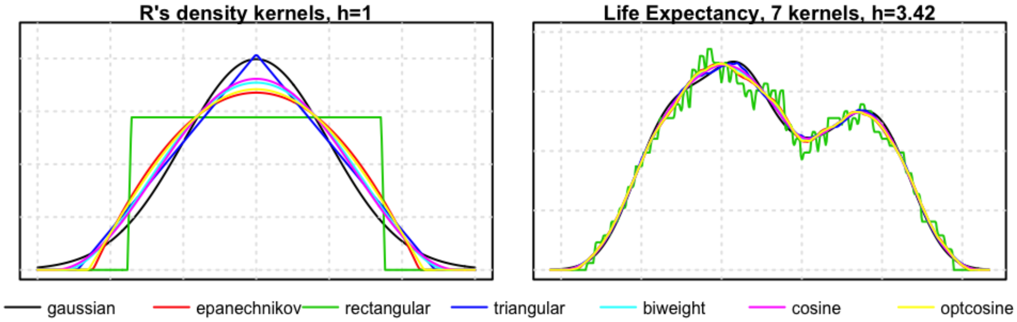


Table 1: Non-Exhaustive List of Density Estimators

Estimator	Formula	Defined Terms
Adaptive Smooth (Location)	$\frac{1}{nh(x)} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h(x)}\right)$	$h(x)$ is a function of the location x
Adaptive Smooth (Sample)	$\frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_i} K\left(\frac{x-X_i}{h_i}\right)$	h_i is a function of the ith sample point
Nearest Neighbor	$\frac{1}{nd_k(x)} \sum_{i=1}^{n} K\left(\frac{x-X_i}{d_k(x)}\right)$	$d_k(x)$ is the Euclidean distance between x and the increasing kth nearest neighbor sequence
Orthogonal Series	$\sum_{i=1}^{n} \theta_i \phi_i(x)$	ϕ_i is an orthogonal basis, $\theta = \mathbf{E}\left[\frac{1}{n}\sum_i \phi_i(x)\right]$
Max Penalized Likelihood	$\arg\max_f \frac{1}{n} \sum_{i=1}^{n} \log f(X_i) - \frac{1}{2}\lambda \left[\log f, \log f\right]$	$[f, f] = \int D(f)D(f)$, $D(f)$ is linear differential operator

2.2 Decision theory

In order to evaluate an estimator in a decision theoretic framework, a commonly defined loss function is the Mean Integrated Squared Error (MISE), which averages the estimator's performance over all samples that could be observed. We can show that, under some regularity conditions, there is a bias-variance trade-off with the MISE.

$$\begin{aligned}
\mathrm{MISE}(\hat{f}) &= \mathbf{E}\left[\int_{-\infty}^{\infty} (\hat{f}(x) - f(x))^2 \, dx\right] \\
&= \int_{-\infty}^{\infty} \mathbf{E}\left[(\hat{f}(x) - f(x))^2\right] \, dx \qquad (1) \\
&= \int_{-\infty}^{\infty} \mathrm{Var}\left(\hat{f}(x)\right) + \left(\mathrm{bias}(\hat{f}(x))\right)^2 \, dx
\end{aligned}$$

To calculate the bias term for the kernel density estimator, we substitute a Taylor expansion and integrate.

$$\begin{aligned}
\mathbf{E}\left[\hat{f}(x)\right] &= f(x) + \frac{1}{2}h^2\sigma_K^2 f''(x) + o(h^2) \\
\left[\mathbf{E}\left[\hat{f}(x)\right] - f(x)\right]^2 &= \frac{1}{4}h^4\sigma_K^4 \left(f''(x)\right)^2 + o(h^4) \\
\int \left[\mathbf{E}\left[\hat{f}(x)\right] - f(x)\right]^2 dx &= \frac{1}{4}h^4\sigma_K^4 \left(\int f''(x)^2 \, dx\right) + o(h^4)
\end{aligned}$$

To calculate the variance term, we do a Taylor expansion about x, and change of variable $t = \frac{x - X_i}{h}$.

$$\begin{aligned}
\mathrm{Var}\left(\hat{f}(x)\right) &= \frac{1}{nh}\int K(t)^2 f(x - ht) \, dt - \frac{1}{n}\left[\mathbf{E}\left[\frac{1}{h}K\left(\frac{x - X_i}{h}\right)\right]\right]^2 \\
&= \frac{1}{nh}\int K(t)^2[f(x) + o(1)] \, dt - \frac{1}{n}\left[f(x) + o(1)\right]^2 \\
&= \frac{1}{nh}f(x)\left(\int K(t)^2 \, dt\right) + o\left(\frac{1}{nh}\right) \\
\int \mathrm{Var}\left(\hat{f}(x)\right) dx &= \frac{1}{nh}\left(\int K(t)^2 \, dt\right) + o\left(\frac{1}{nh}\right)
\end{aligned}$$

Hence, if we put these together back into MISE, we can evaluate our density estimate as a function of the bandwidth parameter and the data. The terms without the order approximation is know as the Asymptotic MISE (AMISE) (Wasserman 2006).

$$MISE(h) = \frac{1}{nh}\left(\int K(t)^2\,dt\right) + \frac{1}{4}h^4\sigma_K^4\left(\int f''(x)^2\,dx\right) + o\left(\frac{1}{nh} + h^4\right)$$

We often estimate this term with cross-validation, with an estimator off by a true constant term.

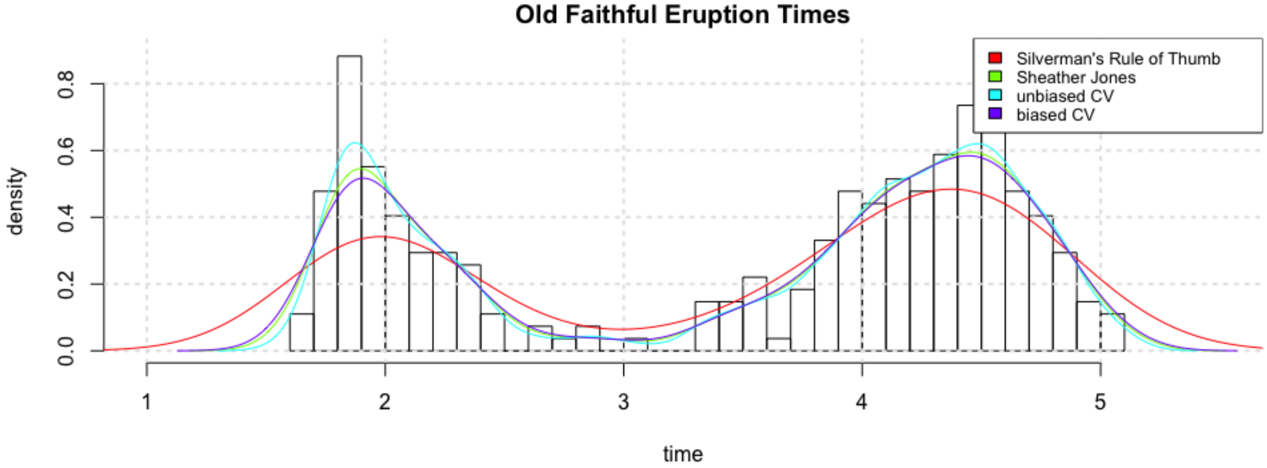
$$\hat{J}(h) = \int \left(\hat{f}_n(x)\right)^2\,dx - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{(-i)}(X_i) \tag{2}$$

2.3 Bandwidth Estimation

Since good bandwidth choice has shown to be highly influential on the quality of the density estimate, many rules have been developed to choose a good bandwidth. Some of the more common methods are shown briefly below

Name of Rule	Estimation Procedure
Silverman's Rule of Thumb	$h_n = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{1/5}$ where $\hat{\sigma} = \min\left(s, \frac{IQR}{1.34}\right)$
Sheather-Jones	See (Sheather & Jones 1991)
Cross-Validation	See (Scott & Terrell 1987)

Figure 3: Comparison of Bandwidth Selection Rules



2.4 Curse of Dimensionality

Density estimation in practice is generally used with lower dimensional data. Consider a sample $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$ from a d-dimensional density function f. As the sample space increases in dimension, the pairwise distance between each of the points increases, and our space will be increasingly sparse, making it more difficult to estimate the underlying density function. We would thus require an exponentially increasing amount of data corresponding to the dimension to achieve a similar MSE lower bound. This is known as the *curse of dimensionality* (Scott 2015). We focus on the kernel density estimator. In higher dimensions, our general kernel density estimator is defined by

$$\hat{f}_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} |\mathbf{H}|^{-1/2} K \left(\mathbf{H}^{-1/2} (\mathbf{x} - \mathbf{X_i}) \right)$$

For higher dimensions, it has been shown that the optimal MISE for an estimator runs on the order of $\mathcal{O}\left(n^{-4/(d+4)}\right)$. Based on this lower bound for MSE, we see that as $n \to \infty$ our estimator $\hat{f}(x)$ converges in probability to $f(x)$. The bandwidth selection also becomes more difficult because instead of a single parameter, we must now estimate a matrix of bandwidths.

3 Computation

3.1 Using the FFT

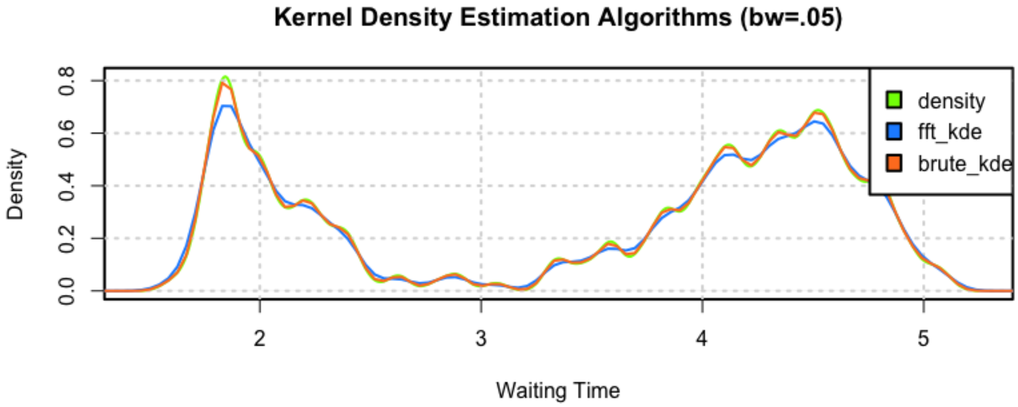
In situations the density estimate needs to be calculated on a grid of points, such as plotting, calculating the kernel density estimator directly is inefficient, requiring $\mathcal{O}\left(n^2 2^g\right)$ time to run on a dataset with n data points and g grid points. If we note that the kernel density estimate is simply a convolution of the empirical density function $g(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - x_i)$ (where δ is the generalized Dirac delta function) and the symmetric kernel $K_h(x)$ with bandwidth parameter h, we can utilize the Fast Fourier Transform to reduce the computational complexity to $\mathcal{O}\left(ng \log n\right)$ (Wand 1994, Lohne 2017). This calculation, due to rounding and underflow, the estimate of the density may lead to small negative values in some cases. We can simply smooth these estimates by setting them to zeros (Silverman

1982*a*). This is the method (slightly simplified) used by the function `density()` in `R`. The number of grid points we discretize the space will also have a significant effect on the algorithm speed. In particular, the direct calculation will grow exponentially with the grid we discretize over.

Algorithm 1 Univariate KDE with Fast Fourier Transform

```
function KDE_FFT($X$, Bandwidth)
    mesh $\leftarrow [\min X - \mathrm{sd}(X), \max X + \mathrm{sd}(X)]$
    meshcount $\leftarrow \mathrm{hist}(X, \mathrm{mesh})$
    meshkernel $\leftarrow$ dKern(mesh)                    ▷ density values of kernel over mesh
    density $\leftarrow \mathrm{FFT}^{-1}(\mathrm{FFT}(\mathrm{meshcount}) * \mathrm{FFT}(\mathrm{meshkernel}))$
    density $\leftarrow$ SmoothZeros(density)
    return density
end function
```

Figure 4: Old Faithful Data (N=272)

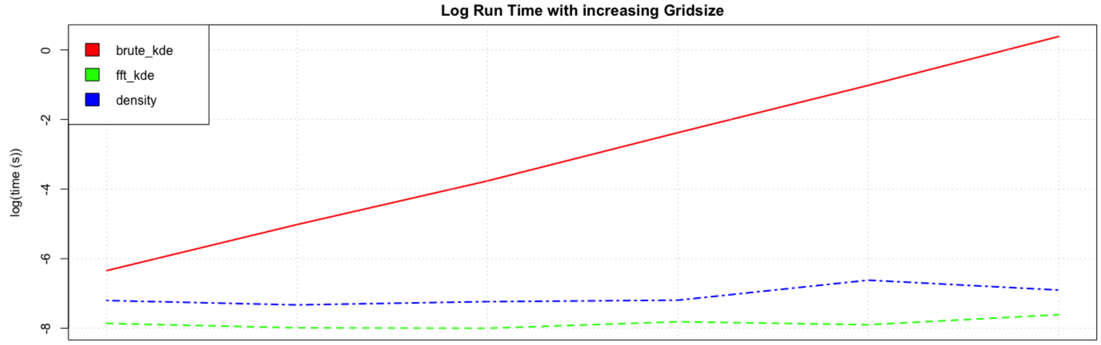


The routines `brute_kde()` and `fft_kde()` were coded myself from scratch, and we compare computational speeds to `density()` provided by `R` in the univariate case on the eruption times of the Old Faithful Geyser Data in Yellowstone National Park (Azzalini & Bowman 1990). For comparability, we calculate the discretization over a grid of 512 points which is default for the `density()` function. From Figure 4 we can see that all the algorithms seem to match up fairly well on our Old Faithful Eruption data. We can attribute the slight deviations to rounding errors and having used a different discrete grid. From Table 3 we see that there is over 1000x speed-up from using a brute force approach to using the FFT. Our hand coded algorithm is also slightly faster than the standard `density()` function. This is likely because we do not do error checking or argument processing in our algorithms, such as weights or kernel changing.

Table 3: Computational Time comparison (avg of 10 runs)

Algorithm	`brute_kde`	`fft_kde`	`density` from R
Time(s)	0.3667	0.000464	0.00073

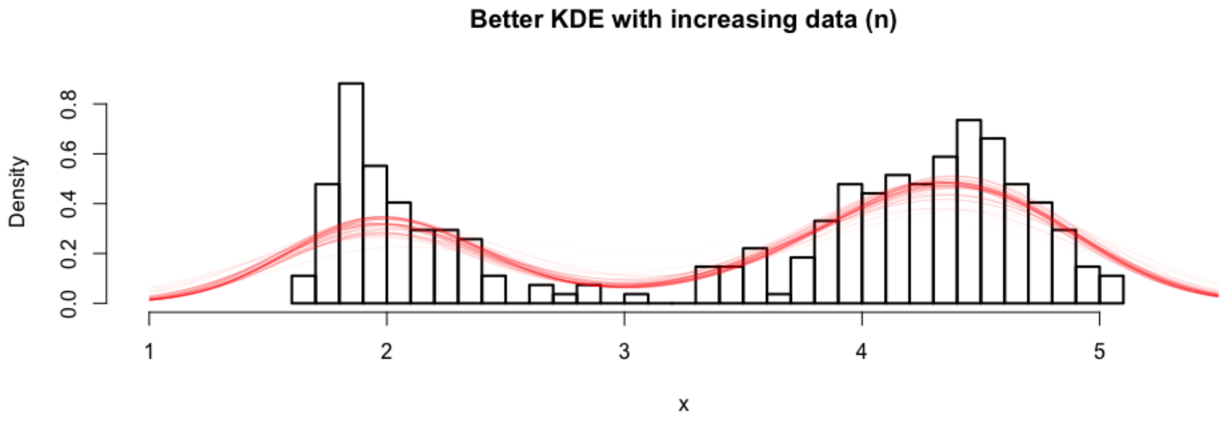
Figure 5: Comparison of Algorithms dependency on Gridsize



3.2 Accuracy with Data

In the above theoretical work, we saw that in the univariate case, the best estimator with optimal bandwidth selection would have a convergence rate of $\mathcal{O}\left(n^{-4/5}\right)$ as a function of the number of data points n. We illustrate the result by calculating the cross-validated risk estimator as shown in Eq. 2 with a near-optimal bandwidth selector, Silverman's Rule of Thumb.

Figure 6: The KDE estimator with Silverman's Rule of Thumb converges closer and closer to the "true" distribution with more data. Faded densities are estimates with less data.



3.3 Higher Dimensions

Again, we would like to use the FFT in higher dimensions because of the massive computational speedup for calculating the KDE. Here we present a two dimensional example based on a UNICEF dataset on life expectancy and child mortality. The problem of using a definitive FFT-based algorithm for unconstrained bandwidth matrices is an open question, though some significant progress was made recently (Gramacki & Gramacki 2017). The technical details are omitted, but the main result is an observation that the generalization presented in (Wand & Jones 1994), has two orderings, the *zero-padding* and *wrap-arround ordering*. The wrap-around ordering will only support kernels that are oriented with the axes, and thus, reshaping the FFT input matrix to use only zero padded procedures will remove some symmetry aliasing. This order was causing some irregularity in the density estimate as shown in Figure 8. They also suggest a scaling parameter of the mesh size based on the largest eigenvalue of the bandwidth matrix **H**. This is an interesting result, but doesn't guarantee that weaknesses will not be found with the zero-padded ordered matrix. This result has already been integrated into the package `ks` in `R`.

Figure 7: Estimated risk of estimator with increasing data nears the theoretical lower bound of $n^{-4/5}$

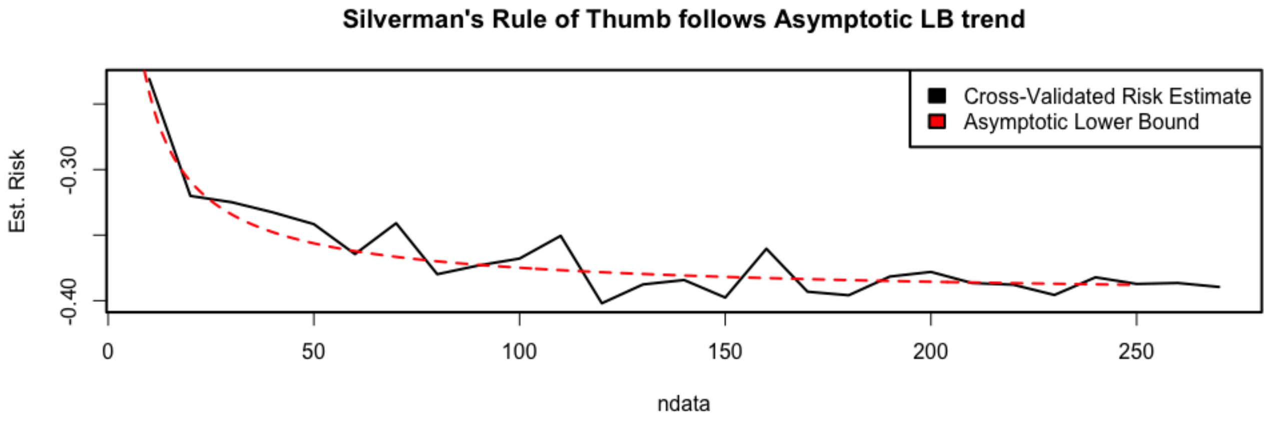
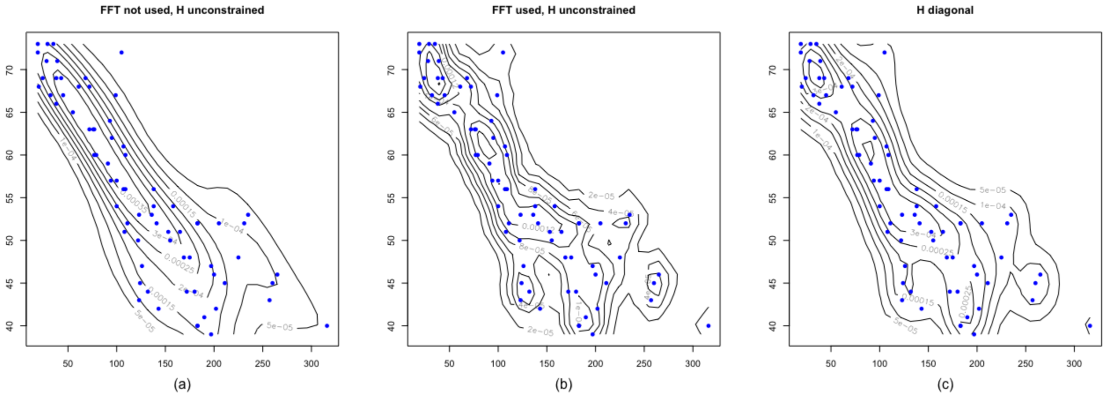


Figure 8: Reproduction of Figure 1 of (Gramacki & Gramacki 2017) showing weakness of KernSmooth FFT implementation. Note (a) is different from (b) and (c)



4 Conclusions

Density estimation has seen many useful applications for understanding of underlying data structure or as a first step in downstream analysis. We've shown some asymptotic results that have led to a wide array of other estimators, in particular the kernel density estimate. The quality of the KDE is highly subject to bandwidth changes, but less so to the kernel choice. The fundamental computational results in one dimension of using the FFT was explored, and we can see at least a 1000-fold increase in computational speed depending on the chosen gridsize. We note the work of extending these results to the multivariate case has matured significantly since the 1990s, which has reduced the computational complexity and increased the accuracy to the point they have been integrated into common software packages (Chen 2017). In future years, we are excited to see maturity in other density estimation techniques such as orthogonal series estimators, splines, or tree-based density estimation.

References

Azzalini, A. & Bowman, A. W. (1990), 'A look at some data on the old faithful geyser', *Applied Statistics* pp. 357–365.

Chen, Y.-C. (2017), 'A tutorial on kernel density estimation and recent advances', *Biostatistics & Epidemiology* **1**(1), 161–187.

Gramacki, A. & Gramacki, J. (2017), 'Fft-based fast computation of multivariate kernel density estimators with unconstrained bandwidth matrices', *Journal of Computational and Graphical Statistics* **26**(2), 459–462.
URL: *https://doi.org/10.1080/10618600.2016.1182918*

Lohne, M. (2017), 'The computational complexity of the fast fourier transform'.

Mack, Y. & Rosenblatt, M. (1979*a*), 'Multivariate k-nearest neighbor density estimates', *Journal of Multivariate Analysis* **9**(1), 1 – 15.
URL: *http://www.sciencedirect.com/science/article/pii/0047259X79900654*

Mack, Y. & Rosenblatt, M. (1979*b*), 'Multivariate k-nearest neighbor density estimates', *Journal of Multivariate Analysis* **9**(1), 1–15.

Scott, D. W. (2015), *Multivariate density estimation: theory, practice, and visualization*, John Wiley & Sons.

Scott, D. W. & Terrell, G. R. (1987), 'Biased and unbiased cross-validation in density estimation', *Journal of the American Statistical Association* **82**(400), 1131–1146.
URL: *http://www.jstor.org/stable/2289391*

Sheather, S. J. & Jones, M. C. (1991), 'A reliable data-based bandwidth selection method for kernel density estimation', *Journal of the Royal Statistical Society. Series B (Methodological)* **53**(3), 683–690.
URL: *http://www.jstor.org/stable/2345597*

Silverman, B. (1982*a*), 'Algorithm as 176: Kernel density estimation using the fast fourier transform', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **31**(1), 93–99.

Silverman, B. W. (1982*b*), 'On the estimation of a probability density function by the maximum penalized likelihood method', *The Annals of Statistics* pp. 795–810.

Silverman, B. W. (2018), *Density estimation for statistics and data analysis*, Routledge.

Wand, M. (1994), 'Fast computation of multivariate kernel estimators', *Journal of Computational and Graphical Statistics* **3**(4), 433–445.

Wand, M. P. & Jones, M. C. (1994), *Kernel smoothing*, Crc Press.

Wasserman, L. A. (2006), *All of nonparametric statistics: with 52 illustrations*, Springer.