DENSITY ESTIMATION INCLUDING EXAMPLES

Hans-Georg Müller and Alexander Petersen
Department of Statistics
University of California
Davis, CA 95616 USA

In order to gain information about an underlying continuous distribution given a sample of independent data, one has two major options:

- Estimate the distribution and probability density function by assuming a finitely-parameterized model for the data and then estimating the parameters of the model by techniques such as maximum likelihood*(Parametric approach).
- Estimate the probability density function nonparametrically by assuming only that it is "smooth" in some sense or falls into some other, appropriately restricted, infinite dimensional class of functions (Nonparametric approach).

When aiming to assess basic characteristics of a distribution such as skewness*, tail behavior, number, location and shape of modes*, or level sets, obtaining an estimate of the probability density function, i.e., the derivative of the distribution function*, is often a good approach. A histogram* is a simple and ubiquitous form of a density estimate, a basic version of which was used already by the ancient Greeks for purposes of warfare in the 5th century BC, as described by the historian Thucydides in his History of the Peloponnesian War. Density estimates provide visualization of the distribution and convey considerably more information than can be gained from looking at the empirical distribution function, which is another classical nonparametric device to characterize a distribution.

This is because distribution functions are constrained to be 0 and 1 and monotone in each argument, thus making fine-grained features hard to detect. Furthermore, distribution functions are of very limited utility in the multivariate case, while densities remain well defined. However, multivariate densities are much harder to

estimate, due to the curse of dimensionality, see Stone (1994), and there are many additional difficulties when one moves from the one-dimensional to the multivariate case, especially for dimensions larger than 3.

The parametric approach to density estimation is sensible if one has some justification that the data at hand can be modeled by a known parametric family of distributions, such as the Gaussian distribution*. In the Gaussian case it suffices to estimate mean and variance parameters (or the elements of the covariance matrix by empirical sample estimates) in order to specify the corresponding Gaussian density. In more general cases, maximum likelihood estimators* are commonly employed to infer the parameter vector that characterizes the assumed distribution from the data. An advantage of this approach is that one easily obtains confidence regions and statistical tests, where correctly specified models can usually be consistently estimated with associated asymptotically valid inference.

For the nonparametric approach, it is common to only rely on the much weaker assumption that the underlying density is smooth, say twice continuously differentiable. This then facilitates the complex task of estimating the density function, which is an infinite-dimensional functional object. Many nonparametric density estimators are motivated as extensions of the classical histogram. Nonparametric density estimation is an ideal tool for situations where one wants to "let the data speak for themselves" and therefore has a firm place in exploratory data analysis. Some variants such as the "rootogram" (Tukey 1977) or visualizations such as "violin plots" (Hintze and Nelson 1998) have proven particularly useful. In practical settings, one rarely has enough information to safely specify a parametric distribution family, even if it is a flexible class of models like the Pearson family (Johnson and Kotz 1994). If a parametric model is misspecified, subsequent statistical analysis may lead to inconsistent estimators and tests. Misspecification and inconsistent estimation is less likely to occur with the nonparametric density estimation approach.

The increased flexibility of the nonparametric approach, however, has some disadvantages that contribute to make inference more difficult:

(i) asymptotic rates of convergence of the (integrated) mean squared error of density estimates are $n^{-\alpha}$ with $\alpha < 1$, where α depends on the smoothness of the

underlying density but rapidly declines with increasing dimensionality d of the data and therefore is always slower than common rate n^{-1} for parametric approaches;

(ii) each of the various available density estimation techniques requires the choice of one or several smoothing or tuning parameters; and

(iii) the information contained in the density estimate usually cannot be conveniently summarized by a few parameter estimates.

In the following, we equate density estimation with the nonparametric approach. Density estimation in this sense has been standard statistical practice for a long time in the form of constructing histograms. It is a subfield of the area of nonparametric curve or function estimation (smoothing methods) that was very active in the 1970s and 1980s. Many statisticians have moved away from finitely parameterized statistical models in search of increased flexibility as needed for data exploration. This has led to the development of exploratory and model-generating techniques and surging interest in statistical analysis for infinite-dimensional objects such as curves and surfaces. Among the first historical appearances of the idea of smoothing beyond the construction of histogram-type objects are papers by A. Einstein (1914) and Daniell (1946) on the smoothing of periodograms (spectral density function estimation), and Fix and Hodges (1951) on the smoothing of density functions in the context of nonparametric discriminant analysis.

Useful introductions to density estimation and good sources for additional references are previous encyclopedia entries of Wegman (1982) and the now classic book on density estimation by Silverman (1986) which, even 30 years after publication, still provides an excellent introduction to the area. More modern resources are the book by Efromovich (2008) that emphasizes series estimators, the book by Klemelä (2009), with a focus on density estimation as a tool for visualization, and the book by Simonoff (2012) with an overall review of smoothing methods. The new edition of the book by Scott (2015) emphasizes the more difficult multivariate (low-dimensional) case and carefully explores its many complexities. Density estimation also plays a major role in machine learning, classification and clustering. Some clustering methods (again in the low-dimensional case) are based on bump hunting, i.e., locating the modes in the density. Bayes classifiers are based on density ratios that can be imple-

mented via density estimation in the low-dimensional case and, under independence assumptions, also in the higher-dimensional case. Applications of density estimation in classification are discussed in more depth in the books of Izenman (2008) and Hastie, Tibshirani and Friedman (2009), and their relevance for particle physics is one of the themes of the recent book by Narsky and Porter (2014).

Examples and Applications of Density Estimation

When the distribution underlying a given data set possesses a probability density, a good density estimate will often reveal important characteristics of the distribution. Applications of density estimation in statistical inference also include the estimation of Fisher information*, efficiency of nonparametric tests, and the variance of quantile* estimates and medians, for example, as all of these depend on densities or density derivatives. Multivariate density estimation can be used for nonparametric discriminant analysis*, cluster analysis* and for the quantification of dependencies between variables through conditional densities, for instance. A practical limitation is that the dimension of the data must be low or assumptions need to be introduced that render the effective dimension low.

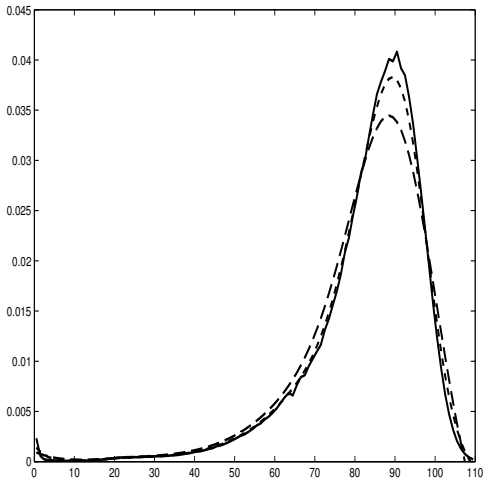
Density estimates are also applied in the construction of smooth distribution function estimates via integration, which then can be used to generate bootstrap* samples from a smooth estimate of the cumulative distribution function rather than from the empirical distribution function (Silverman and Young 1987). Other statistical applications include identifying the nonparametric part in semi-parametric models, finding optimal scores for nonparametric tests, and empirical Bayes methods.

Two examples of density estimation in action are briefly presented in the following.

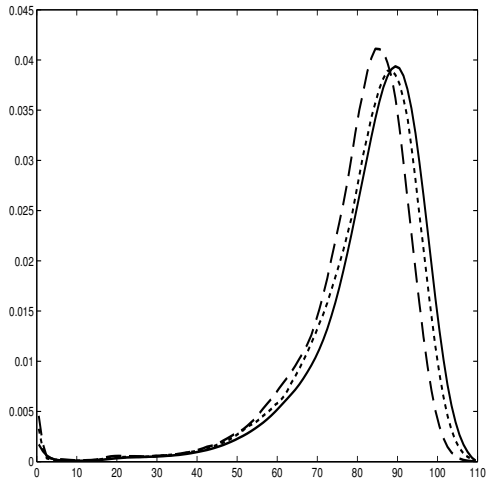
Example 1. Country-specific period lifetables can be used to estimate the distribution of age at death (mortality) for the population. Figure 1(a) shows three estimates of the mortality distribution for the population of Japan in the year 2010. These estimates were obtained by smoothing histograms with local linear smoothers, using three different bandwidths. The smallest bandwidth shows sharp local features in

the estimates, indicating that this density estimate has been undersmoothed. On the other hand, the largest bandwidth considerably decreases the prominence of the mode of the distribution and produces an oversmoothed density estimate.

Figure 1(b) demonstrates density estimates for mortality in Japan for the years 1990, 2000 and 2010. One clearly sees the shift to greater longevity as calendar time increases. From 1990 to 2000, the mode moved from approximately 85 years of age to the low 90s. From 2000 to 2010, it is unclear how much further the mode shifted; however, it appears that the overall mass of the distribution did shift toward more advanced ages. The data are available through the Human Mortality Database, University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany) at www.mortality.org.



(a) The solid, short-dashed and long-dashed lines correspond to smoothed histograms using bandwidths of 0.5, 2 and 3, respectively.



(b) 1990 (long-dash), 2000 (dotted) and 2010 (solid)

Figure 1: Density estimates for the distribution of age at death in Japan, smoothing histograms with local linear fitting. (Left) Estimates for the year 2010 using three different bandwidths. (Right) Estimates for the years 1990, 2000 and 2010 with a smoothing bandwidth of 2.

Example 2. Consider the distribution of the number of eggs laid during the lifetime of a female Mediterranean fruit fly (medfly). In particular, we consider the joint (bi-

variate) distribution of the number of eggs laid during the first 10 days of life and the number laid thereafter. Figure 2 shows the density estimate for $n = 868$ medflies in the form of surface and contour plots. The density estimate suggests the hypothesis that high reproductivity in the first 10 days would be followed by a lower rate afterward and provides many additional details about the antagonistic relationship between early and late reproduction in these flies; see Carey et al. (1998) for further details. The data are available at http://anson.ucdavis.edu/∼mueller/data/data.html.

The data analysis for the first example was performed in `MATLAB`, using the hades package (available at http://www.stat.ucdavis.edu/hades/) for histogram smoothing. For the bivariate data in Example 2, the `R` package `sparr` was used for bivariate density estimation. Various other `R` packages are also available for both univariate and multivariate density estimation. The package `KernSmooth` provides univariate and bivariate density estimation via kernel smoothing, while the `ks` package allows for multivariate density estimation for up to 6 dimensions. In addition, the package `np` includes routines for estimating multivariate conditional densities using kernel methods. Density estimation based on histograms is also implemented in the packages `delt` and `ash`.

Histograms

Let $X_1, ..., X_n$ be a sample of data in $\Re^d$, $d \geq 1$, which possess a probability density function (p.d.f.) f with $\int f(x)dx = 1$ and $f(x) \geq 0$ for all x. A time-honored statistical graphical device for checking distributional properties like symmetry and outliers or for comparing the distribution of various groups is the histogram estimate $\hat{f}_H(x)$ of $f(x)$. We divide the range of the data in $m = m(n)$ subsets called "bins" B_j which may be of equal (commonly) or unequal size (length or volume). Let m_j be the number of data values which fall into B_j and denote the size of B_j by $|B_j|$. If x falls into the bin B_l, the histogram estimate is

$$\hat{f}_H(x) = \frac{m_l}{\sum_{j=1}^{m} m_j |B_j|}. \qquad (1)$$

The number of bins $m = m(n)$ is a tuning parameter and needs to be chosen in dependency on the sample size n.

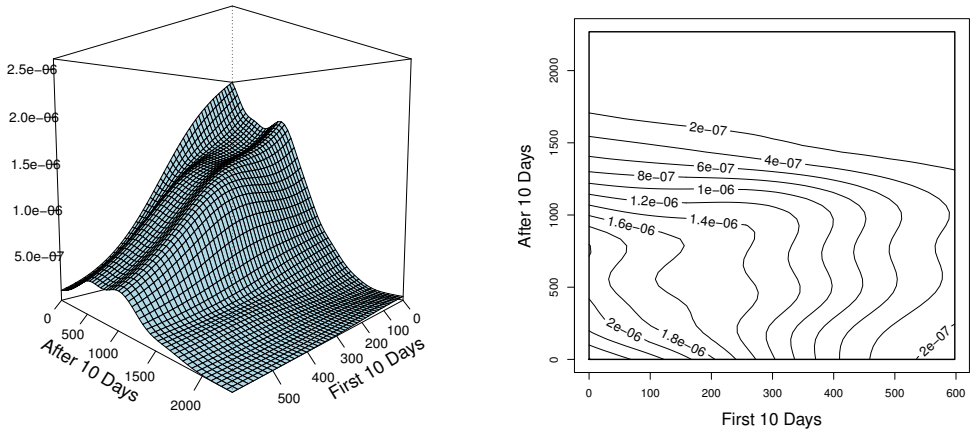


Figure 2: Bivariate density estimate for the number of eggs laid by female medflies within the first 10 days of life and the number laid thereafter, with surface plot on the left and contour plot on the right. A smoothing bandwidth of 75 was used in both directions.

Two disadvantages of this method as compared to other density estimation methods are: (i) Relatively slow asymptotic rate of convergence of the mean squared error. When densities are twice continuously differentiable, the rate is $n^{-2/3}$, as compared to $n^{-4/5}$ for other methods like kernel density estimation, and this is often reflected in worse practical behavior as well; (ii) Discontinuity of these density estimates, when viewed as a function, even though the true densities are assumed to be smooth.

On the plus side, the histogram is easy to understand and to compute in one and higher dimensions, and widely accessible through statistical packages.

Kernel Density Estimation

In the univariate case ($x \in \Re$), smooth function estimates are produced by kernel estimators, which can be motivated by generalizing sliding histograms

$$\frac{F_n(x+h) - F_n(x-h)}{h} = \int_{x-h}^{x+h} \frac{1}{h} dF_n(u),$$

where F_n is the empirical distribution function and dF_n the empirical measure. This

leads to the kernel estimator

$$\hat{f}_K(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) = \int \frac{1}{h} K\left(\frac{x-u}{h}\right) dF_n(u), \tag{2}$$

where $h = h(n)$ is a sequence of bandwidths or smoothing parameters, and K is a kernel function, often with compact support. The kernel method for density estimation was introduced by Rosenblatt (1956) and Parzen (1957) and is often referred to as Rosenblatt-Parzen estimator.

Asymptotics

Writing

$$\hat{f}_K(x) - f(x) = \int \frac{1}{h} K\left(\frac{x-u}{h}\right) (dF_n(u) - dF(u)),$$

asymptotic properties have been obtained by many methods, including strong embedding (Silverman 1978) and empirical processes (Stute 1982, van de Geer 2000). Typical asymptotic results for density estimates are local results like pointwise convergence, asymptotic normality, as well as global results like uniform consistency and distribution of the maximal deviation to obtain uniform confidence bands. For mean and integrated squared errors, as well as other deviation measures, minimax convergence results are available for restricted function classes (Ibragimov and Hasminskii 1980, Tsybakov and Zaiats, 2009) by evaluating the best estimator at the hardest-to-estimate function in the class. Kernel estimators typically attain the minimax rates of convergence.

One set of common assumptions is that the density f is k-times continuously differentiable, the sequence of bandwidths $h = h(n)$ satisfies $h \to 0$ and $nh \to \infty$ as $n \to \infty$, and the kernel satisfies $\int K(x)dx = 1$, $\int K(x)x^j dx = 0$, $j = 0, ..., k-1$, $\int K(x)x^k dx \neq 0$ and some other regularity conditions ("kernel of order k"). By minimizing the leading terms of the asymptotic integrated mean squared error (IMSE) of $\hat{f}_n$, the IMSE-optimal bandwidth for kernels of order two and twice differentiable densities is found to be

$$h^* = c_K \left(n \int f^{(2)}(x)^2 dx\right)^{-1/5},$$

where c_K is a known kernel dependent constant. If this bandwidth, which depends on the unknown derivative $f^{(2)}$ of the density f to be estimated, is used, one obtains

$$\begin{aligned}&\int E\left(\hat{f}_K(x) - f(x)\right)^2 dx \\ &= n^{-4/5}\left[\left(\int f^{(2)}(x)^2 dx\right)\left\{\left(\int K(u)u^2 du\right)\left(\int K(u)^2 du\right)^2\right\}\right]^{1/5} + o\left(n^{-4/5}\right). \qquad (3)\end{aligned}$$

The IMSE-rate of convergence is seen to be $n^{-4/5}$, which falls short of the parametric rate n^{-1}, but is better than the rate $n^{-2/3}$ attained for histograms.

Choice of Bandwidth and Kernel

There exists a large literature on bandwidth selection* (Hall, Marron and Park, 1992; Wand and Jones 1995) and a more limited one on the choice of kernel function K (Granovsky and Müller 1991). In particular the bandwidth choice (more so than the kernel choice) has a strong impact on the quality of the estimated density as it regulates the trade-off between variance and bias.

As for kernel choice, the order of the kernel (number of vanishing moments of the kernel function) determines the rate of convergence, given sufficient smoothness of the density. The density estimate inherits its smoothness from the smoothness of the kernel function employed. According to (3), the kernel shape can be optimized by minimizing functionals like

$$T(K) = \int K^2(x)dx \left(\int K(x)x^2 dx\right)^2 \qquad (4)$$

(see Gasser, Müller and Mammitzsch 1985). The nonnegative kernel which minimizes (4) is the Bartlett-Priestley-Epanechnikov kernel $K(x) = \frac{3}{4}(1-x^2)$ on $[-1,1]$ (Epanechnikov 1969). Other popular nonnegative kernels (and weight functions for (9) below) are the smooth kernels $K(x) = c_\ell(1-x^2)^\ell$ on $[-1,1]$, $\ell > 1$, for suitable constants c_ℓ determined by $\int K = 1$, with the limiting case being the Gaussian kernel. Higher order optimal kernels under sign restrictions for even k are polynomials of order k restricted to $[-1,1]$, and determined by the moment side conditions.

Extensions and Modifications

Many extensions and modifications of the above basic kernel estimation scheme have been considered. Among them:

Multivariate Density Estimation. An application is shown in Fig. 2. The kernel approach can be easily generalized from one to several dimensions. The simplest extension is via product kernels $K(x) = \prod_{i=1}^{d} K_i(x_i)$, where $x \equiv (x_1, ..., x_d)^T \in \Re^d$ and the K_i, $i = 1, ..., d$, are one-dimensional kernel functions (Epanechnikov 1969). Related questions are choice of shape of support of the multidimensional kernel function other than rectangular, and of corresponding multidimensional smoothing parameters, which can be scalars, vectors or matrices. Rates of convergence of IMSE are typically $n^{-4/(d+4)}$ for twice differentiable densities, the "curse of dimensionality" leading to very slow rates for large d (see Scott 1992). Multivariate kernel density estimates can also be used for nonparametric contour estimation. Given a level $\gamma > 0$, the estimated contour is $\left\{x \in \Re^d : \hat{f}_K(x) = \gamma\right\}$; see the right panel of Figure 2 for an example of a contour plot.

Estimation of Derivatives of a Density. A typical approach, which works also for multivariate partial derivatives, is to employ sufficiently smooth kernel functions and to differentiate the density estimate. Such a procedure is found to be equivalent to using less smooth kernel functions with moment properties targeting derivatives (Bhattacharya 1967, Singh 1979, Müller 1984).

Derivatives are of interest for bandwidth choice, construction of confidence regions, mode estimation and estimation of Fisher information*.

Variable Bandwidth Choices and Boundary Kernels. Implementations vary for choosing a different bandwidth for each contribution X_i. Abramson (1982) established that such a scheme can lead to faster rates of convergence.

Another approach is to choose bandwidths which depend on the point x where the density is to be estimated (Müller and Wang 1990) or a mixture of both schemes. Local bandwidth choices lead to smaller IMSEs and, in the multivariate case, to spatial adaptivity. Another possibility is to use varying bandwidths as determined by the k-nearest neighbor* distance (Mack and Rosenblatt 1979). Then $k = k(n)$ corresponds to the smoothing parameter to be determined by the user. Varying bandwidth

choices are particularly useful near boundaries of the domain of the density function.

Since the naive kernel estimator suffers from boundary effects, boundary kernels need to be used (see, e.g., Müller 1993) to avoid boundary bias. A more elegant way to implement such boundary kernels is to use weighted local linear fitting on histograms that utilize small bin widths. Other extensions include recursive and sequential density estimation, estimation of number, location and size of modes, estimation of discontinuities, estimation of density functionals such as $\int f''(t)^2 dt$ and the estimation of conditional densities. Special methods are available and often advantageous for unimodal, isotonic, or otherwise shape-restricted densities.

Other Density Estimation Methods

Besides kernel estimators, various other competing nonparametric density estimation procedures are available. While some of these methods, like orthogonal series estimators which have been around for a long time, have never quite caught on in statistical practice, others like density estimation via nonparametric regression (smoothing of histograms) or wavelet expansions have been used to advantage. Basically, any method of function approximation can be fashioned into a tool for the estimation of densities and more general functions. In terms of understanding and comparing these methods, it is often illuminating to express them in terms of "equivalent" kernels, i.e., kernel estimators, where kernels vary with the location x.

Orthogonal Series Estimators

These estimators are based on the idea of approximating a function by an orthogonal* system of basis functions. The density f allows an expansion

$$f \sim \sum_{i=0}^{\infty} a_i g_i,$$

with

$$a_i = \int f(x) g_i(x) dx$$

in a suitable function (Hilbert) space, where the g_i form an orthonormal system of basis functions. This suggests to estimate f by truncating the orthogonal expansion of f at finitely many, say m, terms and estimating the coefficients a_i, say by

$$\hat{a}_i = \sum_{j=1}^{n} g_i(X_j),$$

with $E\hat{a}_i = a_i$. The orthogonal series density estimator becomes

$$\hat{f}(x) = \sum_{i=0}^{m} \hat{a}_i g_i(x).$$

The basic idea is due to Whittle (1958) and Čencov (1962). Here, $m = m(n)$ assumes the role of the smoothing parameter.

Orthogonal systems $\{g_i\}_{i\geq 0}$ which have been proposed for density estimation include orthonormal polynomials and trigonometric functions $e^{i\pi kx}$. Walter and Blum (1979) introduced the general notion of a delta sequence estimator, which contains kernel and orthogonal series estimators as special cases. Wavelet expansions also have been successfully used for density estimation, thus extending the approach of orthogonal series expansions (Antoniadis, Grégoire and McKeague 1994).

Penalized Maximum Likelihood Estimators

The unrestricted likelihood function* for a density estimate $\hat{f}$ is

$$L(\hat{f}) = \prod_{i=1}^{n} \hat{f}(X_i).$$

By putting atoms at the location of each X_i, L can be made arbitrarily large. Therefore, Good and Gaskins (1980) proposed to introduce a roughness penalty function G and to maximize instead the penalized log likelihood

$$\log L(\hat{f}) = \sum_{i=1}^{n} \log\{\hat{f}(X_i)\} + \alpha G(f),$$

where α is the smoothing parameter.

Good and Gaskins consider roughness penalties $G(f) = \int_{-\infty}^{\infty} \frac{f'(t)^2}{f(t)} dt$ and $G(f) = \int f''(t)^2 dt$. These methods have a Bayesian interpretation, where the roughness penalty corresponds to an improper prior on the function space.

Density Estimation via Nonparametric Regression

Observing that all data occur actually in binned form due to round-offs and computer number representation limitations, these approaches correspond to smoothing a histogram with small equal bin widths with a kernel-based or other nonparametric regression smoother. The problem of estimating a density function is thus transformed to one of appropriately binning the data and then estimating a nonparametric regression function in a setting with fixed equidistant design.

Practically, the X_i in say (5) are replaced by the midpoints of the bins and the Y_i by the number of points falling into the bin with midpoint X_i. In this setting, fixed design regression kernel smoothing methods are asymptotically equivalent, including local least squares (Müller 1987, Fan and Gijbels 1996).

Locally weighted least squares type kernel estimators are an important tool for scatterplot smoothing. Assume one has a scatterplot (X_i, Y_i), $i = 1, \ldots, n$, where (X_i, Y_i) are sampled form the joint distribution (X, Y). When fitting local lines to the scatterplot data, the corresponding estimators $\hat{m}_L$ of $m(x) = E(Y|X = x)$, are obtained as follows:

$$\text{Minimize} \ \sum_{i=1}^{n} \frac{1}{h} K(\frac{x - X_i}{h}) \left[Y_i - (\alpha_0 + \alpha_1 (X_i - x))\right]^2, \tag{5}$$

with respect to α_0, α_1, and set

$$\hat{m}_L(x) = \hat{\alpha}_0.$$

This corresponds to fitting a line locally in a window of size b around x by weighted least squares, where the weights are derived from the kernel function K, and then taking the estimated intercept of the fitted line centered at x as regression function estimate.

If the bin widths of the initial histogram that is smoothed by fitting local linear lines are very small, this method is equivalent to a kernel density estimator that uses

boundary kernels. This method was employed for the density estimates displayed in Figure 1.

Outlook

The basic methodology for density estimation is in place, and future research emphasis will move towards more complex situations and applications. Less explored but practically important situations are those where the dimension of the data for which a density is desired is high, or where one has dependencies in the data, or missing and incomplete data. Other problems arise for densities with inhomogeneous smoothness properties, such as densities with discontinuities, where questions like estimation of support, number and location of discontinuities or break curves are of interest, in particular for the multivariate situation. In such situations, more generalized notions of modes such as modal curves are of interest and have applications in manifold learning. For example, a generalized modal surface may represent the target manifold to be learned from the data.

For the high-dimensional case, the curse of dimensionality forces one to impose sensible constraints. One promising constraint is that the data are on a low-dimensional manifold. In this case, the curse of dimensionality is determined by the dimension of the low-dimensional manifold and not that of the ambient space (Bickel and Lee 2007, Buchman, Lee and Schafer 2011). Therefore density estimation may remain feasible if the manifold is low-dimensional even in the case where the ambient space is high-dimensional.

References

1. Abramson, I.S. (1982). On bandwidth variation in kernel estimates–a square root law. *Ann. Statist.* **10,** 1217-1223. (First demonstration that variable bandwidth choice depending on X_i leads to improved rate of convergence).

2. Antoniadis, A., Grégoire, G. and McKeague, I.W. (1994). Wavelet methods for curve estimation. *J. Amer. Statist. Assoc.* **89**, 1340-1353. (Introduces wavelets for density estimation).

3. Bhattacharya, P.K. (1967). Estimation of a probability density function and its derivatives. *Sankhyā* **A29**, 373-382. (First proposal of density derivative estimators).

4. Bickel, P. and Li, B. (2007). Local polynomial regression on unknown manifolds. *Complex Datasets And Inverse Problems: Tomography, Networks And Beyond. IMS Lecture Notes-Monograph Series.* **54**, 177– 186.

5. Buchman, S.M., Lee, A.B. and Schafer, C.M. (2011). High-dimensional density estimation via SCA: An example in the modelling of hurricane tracks. *Statist. Methodology* **8**, 18–30.

6. Carey, J.R., Liedo, P., Müller, H.G., Wang, J.L., Chiou, J.M. (1998). Relationship of age patterns of fecundity to mortality, longevity, and lifetime reproduction in a large cohort of Mediterranean fruit fly females. *J. of Gerontology – Biological Sciences* **53**, 245–251.

7. Čencov, N.N. (1962). Evaluation of an unknown density from observations. *Soviet Mathematics* **3**, 1559-1562. (Introduces orthogonal series density estimators).

8. Daniell, P.J. (1946). Discussion of paper by M.S. Bartlett. *J. Roy. Statist. Soc. Suppl.* **8**, 88-90. (Introduces spectral smoothing).

9. Einstein, A. (1914). Méthode pour la détermination de valeurs statistiques d'observations concernant des grandeurs soumises à des fluctuations irrégulières. *Arch. Sci. Phys. et Nat. Ser. 4* **37**, 254-256. (Idea of spectral smoothing first mentioned).

10. Epanechnikov, V.K. (1969). Non-parametric estimation of a multivariate probability density. *Theory Probab. Appl.* **14**, 153-158. (Introduces a product kernel method and contains the derivation of the mean squared error optimal nonnegative kernel function, a quadratic polynomial which is often called Epanechnikov kernel—although Bartlett and Priestley independently derived

this kernel much earlier in the context of spectral density estimation. It would therefore more appropriately be referred to as Bartlett-Priestley-Epanechnikov kernel).

11. Fix, E. and Hodges, J.L. (1951). Discriminatory analysis and nonparametric estimation: consistency properties. *Rept. No. 4, Proj. No. 21-49-004,* USAF School of Aviation Medicine, Randolph Field, Texas. (The first appearance of kernel density estimation).

12. Gasser, T., Müller, H.G. and Mammitzsch, V. (1985). Kernels for nonparametric curve estimation. *J. Roy. Statist. Soc.* **B47**, 238-252. (Choice of kernels with IMSE-optimal shape is discussed).

13. Good, I.J. and Gaskins, R.A. (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by the scattering and meteorite data (with Discussion). *J. Amer. Statist. Assoc.* **75**, 42-73. (Introduces the roughness penalty to make maximum likelihood for density estimation work).

14. Granovsky, B. and Müller, H.G. (1991). Optimizing kernel methods: A unifying variational principle. *International Statistical Review* **59**, 373-388. (Reviews various aspects of kernel choice, contains optimality proofs and proposes a class of analytical kernels obtained as limits of polynomial kernels.)

15. Hall, P., Marron, J.S. and Park, B.N. (1992). Smoothed cross-validation. *Probab. Th. Rel. Fields* **92**, 1-20. (An interesting proposal of bandwidth choice by smoothing twice).

16. Hintze, J.L. and Nelson, R.D. (1998). Violin Plots: A Box Plot-Density Trace Synergism. *Am. Statist.* **52** 181-184

17. Johnson, N.L. and Kotz, S. (1994). *Continuous Univariate Distributions,* 2nd Edition. New York: Wiley. (An encyclopedic work on univariate distributions. A sourcebook for designing appropriate parametric models suitable for parametric density estimation).

18. Mack, Y.P. and Rosenblatt, M. (1979). Multivariate k-nearest neighbor density estimates. *J. Multiv. Anal.* **9**, 1-15.

19. Müller, H.G. (1984). Smooth optimum kernel estimators of densities, regression curves and modes. *Ann. Statist.* **12**, 766-774. (Derivative estimation, mode estimation and kernel choice for both are discussed).

20. Müller, H.G. (1987). Weighted local regression and kernel methods for nonparametric curve fitting. *J. Amer. Statist. Assoc.* **82**, 231-238. (Asymptotic equivalence of some kernel regression estimators).

21. Müller, H.G. (1993). On the boundary kernel method for nonparametric curve estimation near endpoints. *Scandinavian J. Statistics* **20**, 313-328. (One of several papers on boundary kernels).

22. Müller, H.G. and Wang, J.L. (1990). Locally adaptive hazard smoothing. *Prob. Th. Rel. Fields* **85**, 523-538. (Efficiency of local bandwidth choice for kernel estimation of hazard functions under censoring).

23. Parzen, E. (1957). On consistent estimates of the spectrum of a stationary time series. *Ann. Math. Statist.* **28**, 329-348. (One of the earliest papers on kernel estimation in the spectral density context).

24. Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27**, 832-837. (A classic. Introduces the kernel estimator for density estimation. It is shown that nonparametric density estimates cannot be finitely unbiased).

25. Silverman, B. W. (1978). Weak and strong uniform consistency of the kernel estimate of a density function and its derivatives. *Ann. Statist.* **6,** 177-184. (One of the early advanced asymptotic results on kernel density estimation).

26. Silverman, B. W. and Young, G. A. (1987). The bootstrap: to smooth or not to smooth? *Biometrika* **74,** 469-479. (The question addressed is whether one

should resample from the empirical distribution function or from a smoothed version of it. There is no universal answer).

27. Singh, R.S. (1979). Mean squared errors of estimates of a density and its derivatives. *Biometrika* **66**, 177-180. (Introduces a class of density derivative estimators).

28. Stone, C.J. (1994). The Use of Polynomial Splines and Their Tensor Products in Multivariate Function Estimation. *Ann. Statist.* **22**, 118-171.+

29. Stute, W. (1982). A law of the iterated logarithm for kernel density estimators. *Ann. Probab.* **10**, 414-422.+ (Using oscillation behavior of empirical processes).

30. Tukey, J.W. (1977). *Exploratory Data Analysis.* Addison-Wesley.

31. Walter, G. and Blum, J.R. (1979). Probability density estimation using delta sequences. *Ann. Statist.* **7,** 328-340. (A unified treatment of kernel and orthogonal series estimators).

32. Whittle, P. (1958). On the smoothing of probability density functions. *J. Roy. Statist. Soc. Series* **B20**, 334-343. (Introduces the idea of orthogonal series density estimators).

Books and Reviews

1. Efromovich, S. (2008). *Nonparametric Curve Estimation: Methods, Theory, and Applications.* New York: Springer.

2. Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications.* CRC Press.

3. Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* New York: Springer.

4. Ibragimov, I.A. and Hasminski, R.Z. (1980). *Statistical Estimation-Asymptotic Theory.* New York: Springer-Verlag. (A good introduction to the minimax approach).

5. Izenman, A.J. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning.* New York: Springer.

6. Klemelä, Jussi (2009). *Smoothing of Multivariate Data: Density Estimation and Visualization.* New Jersey: Wiley.

7. Narsky, I. and Porter, F. (2014). *Statistical Analysis Techniques in Particle Physics.* Weinheim: Wiley-VCH.

8. Scott, D.W. (2015). *Multivariate Density Estimation: Theory, Practice and Visualization.* New Jersey: Wiley. (Impressive data and function visualization, emphasis on multidimensional histograms, averaged shifted histograms and kernel estimators; beginning to intermediate level, easily accessible, contains many problems and makes an excellent textbook).

9. Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall. (Touches briefly on most aspects of density estimation, easily accessible. Includes data illustrations and applications of density estimation to other statistical problems. Level beginning to intermediate).

10. Simonoff, J.S. (2012). *Smoothing Methods in Statistics.* New York: Springer.

11. Tsybakov, A. B. and Zaiats, V. (2009). *Introduction to Nonparametric Estimation.* New York: Springer.

12. Tukey, J.W. (1977). *Exploratory Data Analysis.* Reading, Mass.

13. Van De Geer, S.A. (2000). *Empirical Processes in M-estimation.* Cambridge University Press

14. Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing.* London: Chapman and Hall. (Comprehensive yet concise overview over kernel density estimation, including bandwidth choice and extensions to kernel regression and other kernel estimators, also includes exercises. Level intermediate to advanced).

15. Wegman, E.J. (1982). Density Estimation. *Encyclopedia of Statistical Sciences*, S. Kotz and N.L. Johnson, Ed., **Vol. 2**, 309-315.+