Self-consistent method for density estimation

Alberto Bernacchia* and Simone Pigolotti†

June 11, 2013

Abstract

The estimation of a density profile from experimental data points is a challenging problem, usually tackled by plotting a histogram. Prior assumptions on the nature of the density, from its smoothness to the specification of its form, allow the design of more accurate estimation procedures, such as Maximum Likelihood. Our aim is to construct a procedure that makes no explicit assumptions, but still providing an accurate estimate of the density. We introduce the self-consistent estimate: the power spectrum of a candidate density is given, and an estimation procedure is constructed on the assumption, to be released *a posteriori*, that the candidate is correct. The self-consistent estimate is defined as a prior candidate density that precisely reproduces itself. Our main result is to derive the exact expression of the self-consistent estimate for any given dataset, and to study its properties. Applications of the method require neither priors on the form of the density nor the subjective choice of parameters. A cutoff frequency, akin to a bin size or a kernel bandwidth, emerges naturally from the derivation. We apply the self-consistent estimate to artificial data generated from various distributions and show that it reaches the theoretical limit for the scaling of the square error with the dataset size.

1 Introduction

Every scientist has encountered the problem of estimating a continuous density from a discrete set of data points. This may happen, for example, when determining a probability distribution from a finite Monte Carlo sample (Binder, 1986), rounding off the shape of a galaxy from a collection of stars (Ripley and Sutherland, 1990), or assessing the instantaneous firing rate of a neuron from a discrete set of action potentials (Kass et al., 2005). In all those cases, one can adopt two different approaches: either assuming a given functional form for the density *a priori*, specified by a certain number of parameters, or renouncing any prior knowledge (beyond that a density exists and, in some cases, that is smooth). These two approaches lead, respectively, to parametric and non-parametric estimates. We will focus on the latter approach, although we will assume the knowledge of the density as a reasoning tool, to be released *a posteriori*.

*Department of Neurobiology, Yale University, 333 Cedar Street, SHM-C400D, New Haven, Connecticut, alberto.bernacchia@yale.edu

†The Niels Bohr International Academy, The Niels Bohr Institute, Blegdamsvej 17, DK-2100 Copenhagen, Denmark.

The most popular non-parametric method is simply plotting a histogram, but more sophisticated procedures have been developed. Kernel Density Estimation (KDE) has been widely studied (Silverman, 1986; Parzen, 1961; Wand and Jones, 1995): instead of counting the number of points in separate bins, KDE constructs a smoothed picture of the data as a superposition of kernel functions centered at the coordinates of data points. More formally, given a sample of N data points (real numbers), denoted by $\{X_j\}$ $(j = 1, ..., N)$, the KDE estimate $\hat{f}(x)$ is written as

$$\hat{f}_{KDE}(x) = \frac{1}{hN} \sum_{j=1}^{N} K\left(\frac{x - X_j}{h}\right) \qquad (1)$$

where $K(x)$ is the smoothing kernel and h is the bandwidth. Usually, the choice of $K(x)$ is not crucial (Silverman, 1986), while h, which controls the degree of smoothing, has to be carefully adjusted: the more concentrated data points are, the less smoothing is necessary in order to obtain a good estimate of their density. An alternative non-parametric method is the Maximum Penalized Likelihood (MPL, Good and Gaskins (1971)), also known in the physics literature as a regularization of Field Theory (Bialek et al., 1996; Holy, 1987; Schmidt, 2000): it consists in performing a functional average of densities weighted by their likelihood and by a measure of their smoothness.

In general, each non-parametric method depends on the arbitrary choice of an adjustable parameter, such as the bin size in histograms, the bandwidth in KDE, or the cutoff frequency in MPL and Field Theory. Each of them regularizes the estimate and avoids overfitting of the data points. In most cases this corresponds to low-pass filtering, i.e. cutting the high frequencies inherent to the discrete dataset, and preventing the estimate from merely reproducing a narrow peak at each data point. However, it would be desirable to devise methods involving the least possible number of parameters, since their determination usually involves some specific assumptions on the distribution to be estimated (e.g. varying the cutoff parameter in MPL and Field Theory precisely corresponds to different choices of the Bayesian prior (Bialek et al., 1996)). Cross-validation techniques have been previously applied for this purpose (Bowman, 1984), but they are computationally expensive and have been seldom applied in the literature.

In this study we show that a self-consistent approach leads to the emergence of a natural cutoff frequency, and an estimate of the density whose performance approaches the theoretical limit for the scaling of the square error with the dataset size. We start from the observation, made in (Watson and Leadbetter, 1963), that a unique "optimal" convolution kernel can be derived as a function of the power spectrum of the (unknown) density to be estimated. This result alone is of little use, since the power spectrum of the true density is not known *a priori*. However, in Section 2 we exploit the result by defining the "self-consistent" estimate as the one whose associated optimal kernel, applied to the sample dataset, returns the estimate itself. Our main result is to derive the exact expression of the self-consistent estimate for any given dataset, and to study its properties. In Section 3 we test the method on three different problems: the estimates of Gaussian, Cauchy and Comb distribution. In all cases we show that the self-consistent estimate outperforms existing methods and the mean integrated square error reaches the optimal theoretical scaling $\sim N^{-1}$. Technical material is presented in the appendices: in Appendix 1 we replicate and extend the re-

sults on the existence of the optimal kernel. In Appendix 2 we provide details of the derivation of the self-consistent estimate. In Appendix 3 we prove that the self-consistent estimate converge almost surely to the true distribution for large N.

2   The self-consistent estimate

In this section, we define the self-consistent estimate, we derive its exact expression, and we study its properties. We start from a result derived by Watson and Leadbetter (1963), that we replicate and extend in Appendix 1. The basic result is that a unique, optimal convolution kernel can be derived as a function of the power spectrum of the density to be estimated, where "optimal" is intended as minimizing the mean integrated square difference between the true density and its estimate.

Given a sample of N data points (real numbers), denoted by $\{X_j\}$ ($j = 1 \ldots N$), each independently drawn from a probability density distribution $f(x)$, we write the estimate as

$$\hat{f}(x) = \frac{1}{N} \sum_{j=1}^{N} K(x - X_j) \tag{2}$$

where we assume $f, \hat{f} \in L^2$. Note that (2) does not depend on any bandwidth h, contrary to the KDE estimate (1). Instead of choosing an arbitrary shape for the kernel K, and looking for an optimal bandwidth (see e.g. Silverman (1986)), we rather look for an optimal shape of the kernel. It turns out that the Fourier transform $\kappa_{opt}(t)$ of the optimal kernel $K_{opt}(x)$ is equal to (see Appendix 1 and Watson and Leadbetter (1963))

$$\kappa_{opt}(t) = \frac{N}{N - 1 + |\phi(t)|^{-2}} \tag{3}$$

where $\phi(t)$ is the Fourier transform of the true density $f(x)$ (characteristic function). The optimal kernel $K_{opt}(x)$ is symmetric with respect to $x = 0$, where it takes its maximum value. Note that $|\phi(t)|$ in Eq.(3) requires the knowledge of the true density, which is not available, hence Eq.(3) cannot be used to compute the estimate in Eq.(2) from the sample observations $\{X_j\}$ alone. We show in the following how to circumvent this problem by a self-consistent approach. Eq.(3) has been previously derived by Watson and Leadbetter (1963), and has been used for assessing the performance of specific kernels (Davis, 1977), as well as for constructing blockwise estimators (Efromovich, 2008).

Although Eq.(3) cannot be used to compute the density estimate, we make a step further and we write the Fourier transform $\hat{\phi}(t)$ of the density estimate $\hat{f}(x)$ in Eq.(2), using the transformed kernel (3), as

$$\hat{\phi}(t) = \Delta(t)\kappa_{opt}(t) = \frac{N\Delta(t)}{N - 1 + |\phi(t)|^{-2}}. \tag{4}$$

where $\Delta(t)$ is the empirical characteristic function, i.e.

$$\Delta(t) = \frac{1}{N} \sum_{j=1}^{N} e^{itX_j} \tag{5}$$

Our approach is to construct an iterative procedure based on Eq.(4), and to determine its exact fixed point. We replace the unknown term ϕ in Eq.(4) with an initial guess $\hat{\phi}_0$, and we denote the resulting estimate as $\hat{\phi}_1$. Then, we try to obtain an improved estimate $\hat{\phi}_2$ by using a kernel which is optimal for $\hat{\phi}_1$. By iterating this procedure, we construct the following sequence of estimates

$$\hat{\phi}_{n+1} = \frac{N\Delta}{N - 1 + |\hat{\phi}_n|^{-2}}. \tag{6}$$

We search for a fixed point of the iteration, namely an estimate $\hat{\phi}_{sc}$ for which

$$\hat{\phi}_{sc} = \frac{N\Delta}{N - 1 + |\hat{\phi}_{sc}|^{-2}} \tag{7}$$

This coincides with the density whose corresponding optimal kernel applied to the data sample gives back the density itself. We call the resulting estimate a "self-consistent estimate". We derive in Appendix 2 the stable solution of Eq.(7), which is equal to

$$\hat{\phi}_{sc}(t) = \frac{N\Delta(t)}{2(N-1)} \left[1 + \sqrt{1 - \frac{4(N-1)}{N^2|\Delta(t)|^2}} \right] I_A(t) \tag{8}$$

where I is the indicator function ($I_A(t) = 1$ if $t \in A$, $I_A(t) = 0$ if $t \notin A$) and A is the set of "accepted" frequencies, i.e. the frequencies giving a nonzero contribution to the estimate. In order for (8) to be a stable solution of (7), the set A must be contained in B ($A \subseteq B$), where $t \in B$ if and only if

$$|\Delta(t)|^2 \geq \frac{4(N-1)}{N^2} \tag{9}$$

This condition sets a threshold for the amplitudes of frequencies t below which $\hat{\phi}_{sc}(t) = 0$. Hence, the contribution of small amplitude waves is neglected, and this automatically determines the range of frequencies to be considered for the estimate. In most practical situations, the filter will cut the high frequency bands, but the filter is not constrained to be low-pass, and it can rather select different frequency bands.

The condition $A \subseteq B$ leaves the arbitrary choice of a subset of frequencies among those above the threshold set by condition (9). As demonstrated in Appendix 3, the self-consistent estimate converges almost surely to the true density, provided that A is bounded, where the bound grows with N (in addition, the characteristic function is required to be integrable). In practical applications, a bounded interval must be necessarily implemented, and we selected

$$A = B \cap [-t^*, t^*] \tag{10}$$

where, as defined above, $B = \{t : |\Delta(t)|^2 \geq \frac{4(N-1)}{N^2}\}$, and t^* is set in such a way that the inequality (9) holds in one half of the interval $[-t^*, t^*]$. The estimate is not sensitive to the choice of this interval (in applications, a 50% change in the interval resulted, on average, in 1% change in the estimate, see Section 3).

Since $\hat{\phi}_{sc}$ and A are bounded, the self-consistent estimate in Fourier space, Eq.(8), can be antitransformed back to the estimate in real space, i.e.
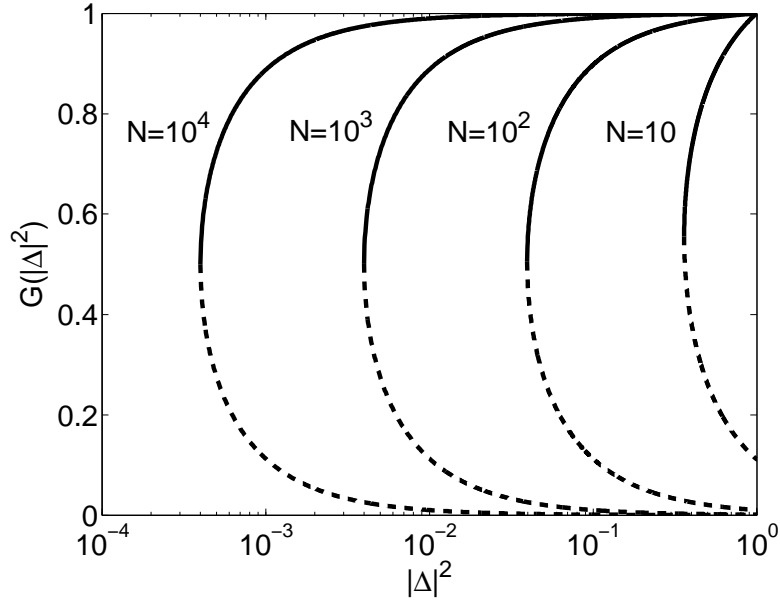
Figure 1: Amplitude gain of the self-consistent estimate (full line), $G = |\hat{\phi}_{sc}|/|\Delta|$, as a function of the squared input amplitude $|\Delta|^2$ and for different values of the sample size N. Dashed line shows the unstable solution (see Appendix 2). The amplitude gain exists only if the inequality (9) is satisfied, and is always smaller than 1, implying that the self-consistent estimate attenuates the amplitudes in the input. The exception is for $|\Delta|^2 = 1$, for which $G = 1$, corresponding to the normalization condition. G tends to one for large values of N (while the unstable solution vanishes), implying that the estimate tends to reproduce all frequencies of the input in that case.

$$\hat{f}_{sc}(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-itx} \hat{\phi}_{sc}(t) dt \tag{11}$$

Applications of the method are considered in the next section, here we describe its properties. A graphical illustration of the filtering properties of the estimate is given in Fig.1, where the amplitude gain $G = |\hat{\phi}_{sc}|/|\Delta|$ is plotted.

The self-consistent estimate is normalized, i.e. $\int_{-\infty}^{+\infty} \hat{f}_{sc}(x) dx = 1$ or, equivalently, $\hat{\phi}_{sc}(0) = 1$, as a consequence of the normalization of the empirical density ($\Delta(0) = 1$). Beside the zero frequency that is kept intact (corresponding to the normalization condition), the self-consistent estimate attenuates all other frequencies ($|\Delta| \leq 1$ implies $|\hat{\phi}_{sc}| \leq |\Delta|$, see Fig.1). Because $\hat{\phi}_{sc}(t)$ is continuous and infinitely differentiable at $t = 0$, all the moments of the self-consistent estimate $\hat{f}_{sc}(x)$ exist. The mean and variance of $\hat{f}_{sc}(x)$ are equal to (see Appendix 2)

$$E(x) = \frac{1}{N} \sum_{j=1}^{N} X_j \tag{12}$$

$$\mathrm{Var}(x) = \frac{1}{N-2} \sum_{j=1}^{N} (X_j - E(x))^2 \tag{13}$$

While the mean is equal to the sample mean, the variance is larger than the sample variance as well as than its unbiased estimator, which is normalized by $\frac{1}{N-1}$ instead of $\frac{1}{N-2}$.

A drawback of the self-consistent estimate is that it is not guaranteed to be non-negative, while the true density is non-negative (note that $|\hat{\phi}_{sc}(t)|^2 \leq 1$ holds from Eq.(8), because $|\Delta|^2 \leq 1$, but that is a necessary and not sufficient condition for $\hat{f}_{sc}(x)$ to be non-negative). However, we seldom observed negative values in simulations. Those can be corrected without any error cost by translating the estimate downward until the positive part is normalized to one, and setting to zero the negative part (Glad et al., 2003; Ushakov, 1999). In general, the restriction to a strictly non-negative estimate has a cost, in terms of the mean integrated square error $E(I)$, quantified by the decay exponent α of the error as a function of the sample size, $E(I) \propto N^{-\alpha}$. Among the estimation procedures that properly give non-negative results, histograms, MPL and Field Theories have $\alpha = 2/3$ (Bialek et al., 1996; Holy, 1987), which is also the limit of KDE when the density is discontinuous (Ushakov, 1999). For a continuous density, KDE improves to $\alpha = 4/5$ (Silverman, 1986). Applications of estimates that are allowed to be negative reach better performance, like "m-th order" kernels (Wand and Jones, 1995; Hall and Marron, 1987; Berlinet, 1993), that have $\alpha = \frac{2m}{2m+1}$ (provided that the density is $m-1$-th differentiable), while infinite order kernels (Devroye, 1992), and the Sinc kernel (Davis, 1977; Schmidt, 2000; Glad et al., 2007), yield $\alpha = 1$ (for infinitely differentiable densities, besides logarithmic terms). Hence, releasing the requirement of a non-negative density estimate allows to improve the performance. The optimal scaling $\alpha = 1$ is also reached by parametric estimators, such as Maximum Likelihood which are, however, strictly non-negative. We present below numerical results suggesting that the self-consistent estimate (8) also reaches $\alpha = 1$ for infinitely differentiable densities.
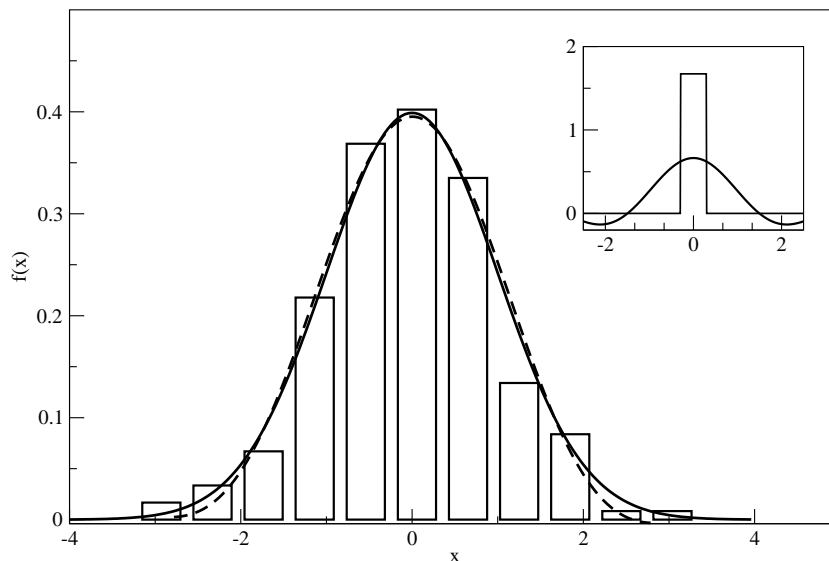
Figure 2: Illustrative example of an estimate of a Gaussian density from $N = 200$ sample points. The true density is given by the full line. The dashed line is the self-consistent estimate, compared with a histogram having optimal binwidth. The inset shows the self-consistent kernel and the bin width of the histogram. (Scott, 1979).

3 Applications to artificial data

In this section we apply the self-consistent estimate to artificial data. The estimate $\hat{f}_{sc}(x)$ is constructed from a sample of N data points $\{X_j\}$ $(j = 1 \ldots N)$, by using Eqs.(5),(8-11). As an illustrative example, we present in Fig. 2 the application of the self-consistent estimate to a Gaussian sample of $N = 200$ points with respect to a histogram (where we chose the optimal binwidth, see Scott (1979)).

We test more systematically the performance of the self-consistent estimate on artificial datasets generated from three distributions: Gaussian, Cauchy and Comb (Marron and Wand, 1992). We compare the performance of the self-consistent estimate (SC) with two kernel estimators, the Gaussian kernel (KG, see Silverman (1986)) and the Trapezoidal kernel (KT, see Politis (2003)), the latter being representative of non strictly positive estimators. For the Cauchy distribution we also show the results for a kernel with a locally adaptive bandwidth (APT, see Silverman (1986); Hossjer (1996)). For each density $f(x)$ and each estimator, we generate 100 samples, each composed of N points randomly and independently drawn from $f(x)$. The performance is evaluated in terms of the mean integrated square error, Eq(22), where the mean is calculated over the 100 sample realizations. We repeat the above procedure for different values of N, ranging from 10^2 to 10^6. In addition to the simulation results, we also present the theoretical bound given by the optimal kernel (3), i.e. the best an estimate of the type of Eq.(2) can achieve (OPT, see also Eq.(28)). In the case of the Gaussian density we also show the theoretical bound given by Maximum Likelihood (ML). The bandwidths for the Gaussian and Trape-
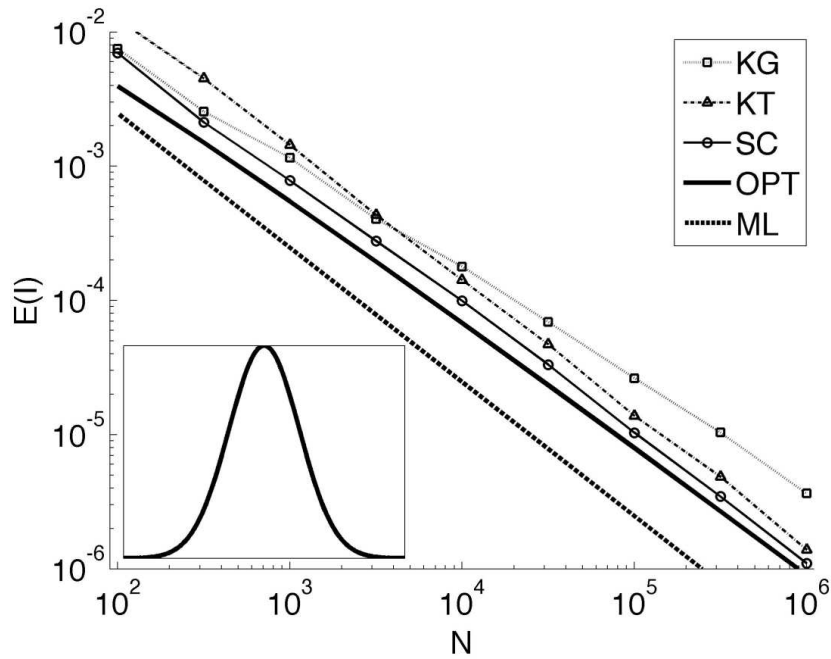
7

Figure 3: Mean integrated square error $E(I)$ for the estimate of a Gaussian distribution (inset), as a function of the sample size N. Standard errors are about 5% and smaller than the size of symbols. Density estimates: Gaussian kernel (KG), Trapezoidal kernel (KT), self-consistent estimate (SC), each point is an average over 100 realizations of the sample. Theoretical bounds: optimal kernel (OPT), maximum likelihood (ML). SC is applied without any prior knowledge, while ML assumes that the density is Gaussian and OPT requires its power spectrum in advance.

zoidal kernels are chosen from the sample data following established empirical rules: $h = 0.79 * iq * N^{-1/5}$ for the Gaussian kernel (iq is the interquartile range, see Eq.(3.29) in Silverman (1986)), while for the Trapezoidal kernel $h^{-1} = 2\min\{m : |\Delta(m+s)|^2 < c^2\log(N)/N, \forall s \in (0, K_N)\}$, where we chose $K_N = \log(N)$ and we averaged results over $2^{-2} \leq c^2 \leq 2^2$ (see Politis (2003)).

First, we study the Gaussian distribution, $f(x) = e^{-x^2/2}/\sqrt{2\pi}$. Fig. 3 shows the mean integrated square error $E(I)$ as a function of the sample size N, for the self-consistent estimate (SC), the Gaussian and Trapezoidal kernels (KG, KT), the optimal bound (OPT) and the Maximum Likelihood (ML). For the Gaussian kernel with a fixed bandwidth h, the exact expression for the error is

$$E(I_{KG}) = \frac{1}{2\sqrt{\pi}} \left(\frac{1}{Nh} - \frac{1}{N\sqrt{1+h^2}} + 1 - \frac{2}{\sqrt{1+h^2/2}} + \frac{1}{\sqrt{1+h^2}} \right) \quad (14)$$

The average value of the bandwidth is $h = 1.06 * N^{-1/5}$, which gives an approximate value of the error $E(I_{KG}) \simeq 0.33 * N^{-4/5}$. The error of the optimal estimate is given by Eq.(28), and for a Gaussian density is equal to

$$E(I_{OPT}) = \frac{\frac{N}{2\sqrt{\pi}(N-1)}\mathrm{Li}_{\frac{1}{2}}(1-N) - 1}{N-1} \simeq \frac{\sqrt{\log(N)}}{\pi N} \quad (15)$$

where Li is the Polylogarithm function, defined by $\mathrm{Li}_s(z) = \sum_{k=1}^{\infty} \frac{z^k}{k^s}$. The error for ML is equal to

$$E(I_{ML}) = \frac{7}{16\sqrt{\pi}}N^{-1} \quad (16)$$

Fig.3 shows that the error of the SC estimate is consistently smaller than both kernel estimates, KG and KT. Both SC and KT approach the theoretical OPT scaling $\sim N^{-1}\sqrt{\log(N)}$ (see Davis (1977)), while ML scales $\sim N^{-1}$. We stress that both ML and OPT require prior knowledge of the density to be estimated. The former needs to know that the density is Gaussian, the latter needs its spectrum in advance, while the self-consistent method achieves the same scaling without any prior assumption.

The second application is the estimate of a Cauchy distribution, $f(x) = [\pi(1+x^2)]^{-1}$. The interest of this case comes from the difficulties of binning long-tailed distributions, especially when the variance diverges. For the Gaussian kernel with a fixed bandwidth, the error is

$$E(I_{KG}) = \frac{1}{2\sqrt{\pi}Nh} + \frac{1}{2\pi} + \frac{N-1}{2\sqrt{\pi}Nh}e^{h^{-2}}\mathrm{erfc}\left(h^{-1}\right) - \frac{\sqrt{2}}{\sqrt{\pi}h}e^{2h^{-2}}\mathrm{erfc}\left(\sqrt{2}h^{-1}\right) \quad (17)$$

The average bandwidth is $h = 1.58 * N^{-1/5}$, for which $E(I_{KG}) \simeq 0.55 * N^{-4/5}$. The error of the optimal estimate, Eq.(28), in the case of the Cauchy distribution, yields

$$E(I_{OPT}) = \frac{\frac{N}{2\pi(N-1)}\log(N) - \frac{1}{2\pi}}{N-1} \simeq \frac{\log(N)}{2\pi N} \quad (18)$$

In order to cope with long-tailed distributions, kernel methods have been generalized to the "adaptive" kernel, which allows the bandwidth to vary locally
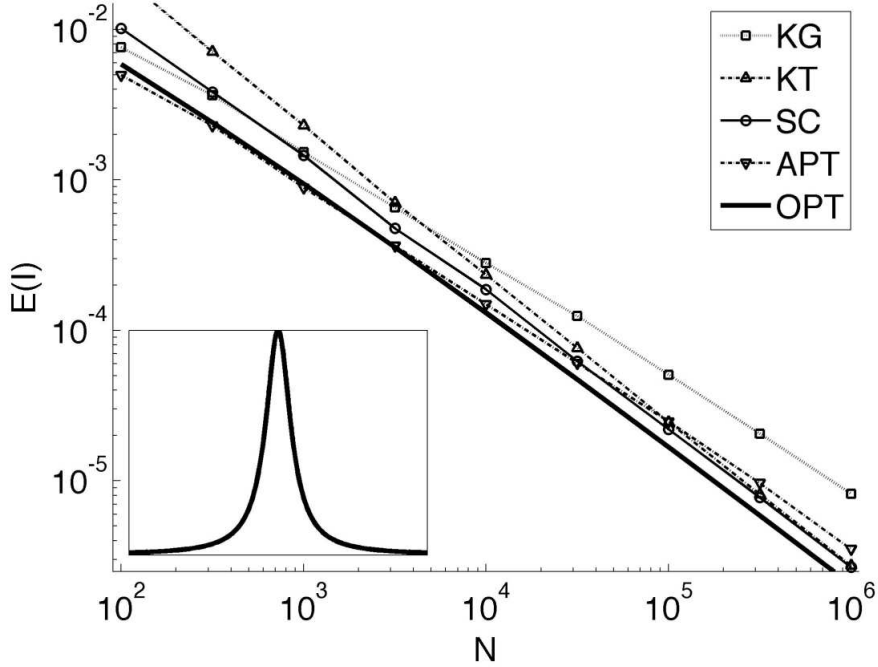
9

Figure 4: Mean square error $E(I)$ for the estimate of a Cauchy distribution (inset) as a function of the sample size N. Standard errors are about 5% and smaller than the size of symbols. Density estimates: Gaussian kernel (KG), Trapezoidal kernel (KT), adaptive kernel (APT), self-consistent estimate (SC), each point is an average over 100 realizations of the sample. Theoretical bound: optimal kernel (OPT). SC is applied without any prior knowledge, APT applies to long-tailed distributions, OPT requires the power spectrum of the true density in advance.

according to a first estimate of the density (Silverman, 1986; Hossjer, 1996). In Fig. 4 we show the mean square error as a function of N for SC, KG, KT estimates, the adaptive kernel estimate (APT), and the OPT bound. For small datasets, the adaptive kernel method performs best. However, the self-consistent method still shows better scaling with N, and its error is lower for large sample sizes, with a crossover occurring for N between 10^4 and 10^5. Again, the SC estimate performs better than both kernel estimates KG and KT. Both SC and KT estimates approach the OPT scaling. We stress that the adaptive method requires some prior knowledge: it is used when one knows that the distribution is long-tailed and, again, the optimal estimate requires prior knowledge of its power spectrum. Conversely, we applied the self-consistent method blindly, in the same way as we did in the Gaussian case.

The last application is the Comb distribution (Marron and Wand (1992)), which is multimodal, where different modes have different widths, and its transform is affected by a large interval of frequencies. Fig.5 shows the results for SC, KG, KT estimates and the OPT bound (which is computed numerically

using Eq.(28)). Note that KG performs poorly in this case, because the empirical bandwidth is unable to capture the large interval of frequencies of the Comb distribution. Instead, both SC and KT embrace an appropriate interval of frequencies, estimated from the empirical characteristic function, they perform much better than KG and approach the OPT scaling.

Finally, we remark that the self-consistent estimate may perform poorly when applied to particular families of density functions. As explained in Appendix 3, if the density is not square integrable or its characteristic function is not integrable, then the self-consistent estimate is not guaranteed to converge to the true density for large N. Simulations (not shown) suggest that the self-consistent estimate does not perform well when applied to the box function (whose characteristic function is not integrable) and to the χ^2 distribution with one degree of freedom (which is not square integrable).
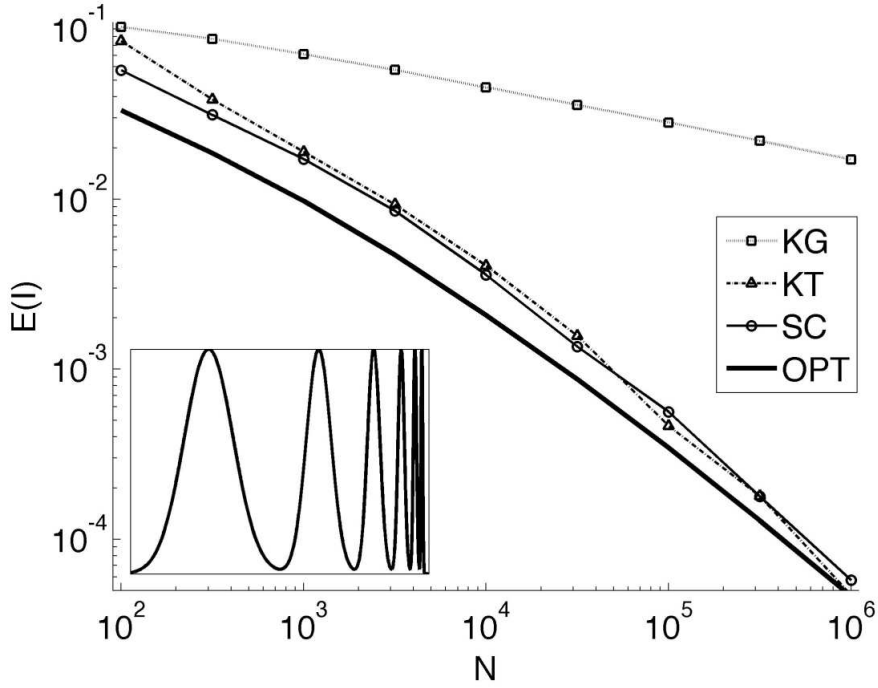

Figure 5: Mean square error $E(I)$ for the estimate of a Comb distribution (inset) as a function of the sample size N. Standard errors are about 1% and smaller than the size of symbols. Density estimates: Gaussian kernel (KG), Trapezoidal kernel (KT), self-consistent estimate (SC), each point is an average over 100 realizations of the sample. Theoretical bound: optimal kernel (OPT). SC is applied without any prior knowledge, OPT requires the power spectrum of the true density in advance.

4 Discussion

We presented a method that estimates a density in a self-consistent way from a finite sample. This approach produces a unique estimate where no parameters have to be adjusted, and the only prior is the belief in the self-consistent procedure. On the other hand, the cases in which there exists a widely accepted theoretical framework for the system at hand would rather point to a Bayesian, parametric approach in which a specific form for the density is postulated.

The self-consistent estimate converges to the true density for large N and it has a better performance and scaling than existing methods for all the examples we studied. Together with the simplicity of its implementation, these features make it preferable for applications, especially when large datasets are available. However, if the true density is not square integrable or its characteristic function is not integrable, the self-consistent estimate is not guaranteed to converge and it may perform poorly.

A frequency filter emerges naturally in the derivation of the self-consistent estimate, as a function of the data sample, and prevents overfitting of the data. When long tails or power-law behavior is suspected, one is tempted to use logarithmic binning, but that has been shown to be highly inaccurate (Clauset et al., 2009; Goldstein et al., 2004), and sometimes leading to the conclusion that the density has power-law tails when it has not. The self-consistent method constitutes a good alternative when there are no solid theoretical grounds to assume (or discard) a power law.

For small values of N, adaptive bandwidth kernels may perform better than the self-consistent kernel (see Fig.4). Note that if the bandwidth is allowed to vary locally, the performance of the estimate is not bounded by the optimal kernel performance, since it does not belong to estimates of the type of Eq.(2), and can in principle perform better. Future studies will be devoted to extend the self-consistent approach to estimates in which the kernel may vary locally. For example, a generic non-adaptive estimate can be converted into an adaptive one by a transformation of x leading to a space with uniform measure (Ruppert and Kline, 1994; Periwal, 1997).

The self-consistent method can be applied concretely to any problem in which a density is sampled and there is no prior knowledge of its functional shape. For instance, the mode of the instantaneous firing rate of a cortical neuron may indicate whether the choice of a behaving monkey is triggered by a memorized object or a spatial cue (Olson et al., 2000). Another application could be the density estimate of the spacings between zeros of the Riemann Zeta function, based on very large numerical dataset, which could help investigate the related mathematical conjectures (Odlyzko, 1987). Finally, the method can be used to analyze samples obtained from Monte Carlo simulations (Binder, 1986).

We remark that the method could not be applied to the case of integer numbers. In general, when data points are uniformly spaced by multiples of any constant length, the Fourier transform of their density is periodic. An amplitude threshold and a cutoff frequency would be meaningless in that case, and this would preclude any practical application. However, when data points are integers there is usually no need for filtering, and it is sufficient to use a histogram, counting the occurrences of each number.

Finally, a straightforward generalization is to consider a finite interval $[a, b]$ instead of the entire line $(-\infty, +\infty)$, by using Fourier series instead of Fourier

12

transforms. Another possibility is to apply the method to high-dimensional distributions. In our derivation, the relevant variables are scalar quantities, and the $d-$dimensional analogy is obtained by using the $d-$dimensional Fourier transform, i.e. by performing the integral $\frac{1}{(2\pi)^d} \int_{\Re^d} dt^d$ instead of $\frac{1}{2\pi} \int_{\Re} dt$. Among many possible implementations, the method could be then applied to the analysis of multielectrode neuronal recordings (Brown et al., 2004), multivariate financial data (Breymann et al., 2003), and reconstruction of Ramachandran angle distributions (Kleywegt and Jones, 1996).

Appendix 1: Derivation of the optimal kernel

In this section, we show that a unique "optimal" convolution kernel can be derived as a function of the power spectrum of the density to be estimated. A similar result has been presented in Watson and Leadbetter (1963). Given a sample of N data points (real numbers), denoted by $\{X_j\}$ $(j = 1 \ldots N)$, each independently drawn from a probability density distribution $f(x)$, we write the estimate as

$$\hat{f}(x) = \frac{1}{N} \sum_{j=1}^{N} K(x - X_j) \tag{19}$$

The true density f is assumed to be normalized, i.e. $\int_{-\infty}^{+\infty} f(x)dx = 1$.

We look for a kernel $K(x)$ such that the estimate (19) minimizes the mean integrated square error

$$E(I) = E \int_{-\infty}^{+\infty} \left[\hat{f}(x) - f(x) \right]^2 dx \tag{20}$$

where $E(\cdot)$ denotes an average over all the possible realizations of the data sample $\{X_j\}$. In order to minimize (20), we follow a procedure for signal deconvolution (the Wiener filter (Wiener, 1949)). We introduce the Fourier transform of the unknown density (characteristic function)

$$\phi(t) = \int_{-\infty}^{+\infty} e^{itx} f(x)dx \tag{21}$$

The normalization condition now reads $\phi(0) = 1$. We also call $\hat{\phi}(t)$ and $\kappa(t)$ the Fourier transforms of the estimate $\hat{f}(x)$ and of the kernel $K(x)$ respectively. The mean integrated square error (20) corresponds to the mean square distance between the true density f and the estimate $\hat{f}$, in terms of the Euclidean metric in the Hilbert space L^2 (we assume $f, \hat{f}, \phi, \hat{\phi} \in L^2$). By means of Parseval's theorem, we rewrite (20) in Fourier space as

$$E(I) = \frac{1}{2\pi} E \int_{-\infty}^{+\infty} \left| \hat{\phi}(t) - \phi(t) \right|^2 dt \tag{22}$$

It is straightforward to perform the average in Fourier space. Applying the convolution theorem to Eq.(19), the transformed estimate is equal to $\hat{\phi}(t) = \kappa(t)\Delta(t)$, where

$$\Delta(t) = \frac{1}{N} \sum_{j=1}^{N} e^{itX_j} \tag{23}$$

is the empirical characteristic function. Note that $\Delta \notin L^2$ for any finite value of N, but $\hat{\phi} \in L^2$ by assumption. Using $E(\Delta) = \phi$, and $E(|\Delta|^2) = |\phi|^2 + N^{-1}\left(1 - |\phi|^2\right)$, we can rewrite the error as

$$E(I) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \left\{ N^{-1}|\kappa|^2 \left(1 - |\phi|^2\right) + |\phi|^2|1 - \kappa|^2 \right\} dt. \qquad (24)$$

Since $f(x)$ is a density, it is normalized and non-negative, which implies $|\phi|^2 \leq 1$. Then, the first term in the integral, proportional to $1/N$, is non-negative: it corresponds to the error due to the finite size of the sample, while the second term does not depend on N. These two sources of error are known as error variance and error bias respectively (Silverman, 1986). Among the possible choices of the kernel, we search for the one minimizing the mean integrated square error. Since Eq.(24) is quadratic in κ, it is straightforward to find its global minimum, by setting to zero the functional derivative of $E(I)$ with respect to κ, i.e.

$$2\pi \frac{\delta E(I)}{\delta \kappa^*} = N^{-1}\kappa \left(1 - |\phi|^2\right) - |\phi|^2(1 - \kappa) = 0 \qquad (25)$$

where the asterisk denotes complex conjugate. This yields a unique, optimal kernel, which in Fourier space reads

$$\kappa_{opt}(t) = \frac{N}{N - 1 + |\phi(t)|^{-2}} \qquad (26)$$

The optimal kernel satisfies the normalization condition $\kappa_{opt}(0) = 1$, because $\phi(0) = 1$, and is a real function. Since the density f is real, then $|\phi(t)| = |\phi(-t)|$, which implies that $\kappa_{opt}(t)$ is an even function. Then its antitransform $K_{opt}(x)$, the optimal kernel in the real space, is also real and even and, because expression (26) is non-negative, $K_{opt}(x)$ takes the maximum value at $x = 0$, i.e. at the coordinate of each data point in Eq.(19).

Using the expression for the transformed optimal kernel, Eq.(26), we rewrite the estimate, Eq.(19), in Fourier space, as

$$\hat{\phi}(t) = \Delta(t) \frac{N}{N - 1 + |\phi(t)|^{-2}}. \qquad (27)$$

which we call the "optimal estimate". The optimal estimate satisfies the normalization condition, $\hat{\phi}(0) = 1$, because $\Delta(0) = 1$ and $\phi(0) = 1$. For infinite sample size ($N \to \infty$), the optimal estimate reduces to the true density with probability one, i.e. $\hat{\phi}(t) \to \phi(t)$, because $\Delta(t) \to \phi(t)$ (see Csorgo and Totik. (1983), Ushakov (1999) for the detailed sufficient conditions), while the fractional term, the kernel, tends to one. This is because an infinite sample would reproduce the true density itself, without the need of any transformation of the data. For finite N, the optimal estimate cuts the frequencies that have less power in the true density, and hence are more subject to noise, i.e. frequencies t whose power is of the order $|\phi(t)|^2 \simeq 1/N$ or less.

The above procedure is analogous to the derivation of the Wiener filter for signal deconvolution (Wiener, 1949): the prior knowledge of the power spectrum of both the signal, the unknown density, and the noise establishes a unique criterion for optimal signal to noise separation (note that, once the signal spectrum is given, $|E(\Delta)|^2 = |\phi|^2$, the assumption of independence of data

14

points allows the noise spectrum to be written as a function of the signal, i.e. $E(|\Delta|^2) - |E(\Delta)|^2 = N^{-1}(1 - |\phi|^2))$.

We conclude this section by deriving the minimum square error obtained by the application of the optimal kernel, that will be useful for assessing the performance of practical applications of the method. By substituting the expression (26) of the optimal kernel in (24), the associated minimum square error can be written, after some algebra, as

$$\min_K E(I) = \frac{K_{opt}^{(N)}(0) - K_{opt}^{(1)}(0)}{N - 1} \tag{28}$$

where we made explicit the dependence of the optimal kernel on the sample size N, by writing $K_{opt} = K_{opt}^{(N)}(x)$.

Finally, note that $|\phi(t)|^2 \leq 1$ and $|\Delta(t)|^2 \leq 1$, which implies from Eq.(27) that also the optimal estimate satisfies $|\hat{\phi}(t)|^2 \leq 1$. While this is a necessary condition for the antitransform $\hat{f}(x)$ to be non-negative, it is not sufficient, and $\hat{f}(x)$ is not guaranteed to be a non-negative density.

Appendix 2: Derivation of the self-consistent estimate

In this section we derive the expression for the self-consistent estimate $\hat{\phi}_{sc}$, Eq.(8), and we study its stability. We start from Eq.(6), the iterative map that we rewrite here

$$\hat{\phi}_{n+1} = \frac{N\Delta}{N - 1 + |\hat{\phi}_n|^{-2}}. \tag{29}$$

We search for a fixed point of the iteration, namely $\hat{\phi}_{sc}$ such that

$$\hat{\phi}_{sc} = \frac{N\Delta}{N - 1 + |\hat{\phi}_{sc}|^{-2}} \tag{30}$$

We derive in the following the two solutions (beyond the null solution) of Eq.(30) and we show that only one solution is stable with respect to the iteration, Eq.(29). We start by taking the absolute value of Eq.(30), in order to obtain an equation for the single unknown variable $|\hat{\phi}_{sc}|$ (note that $\hat{\phi}_{sc}$ is complex valued). Then, we multiply the expression by the denominator and by $|\hat{\phi}_{sc}|$ (leaving the null solution $\hat{\phi}_{sc} = 0$), obtaining a simple quadratic equation

$$(N - 1)\left|\hat{\phi}_{sc}\right|^2 + 1 = N|\Delta|\left|\hat{\phi}_{sc}\right| \tag{31}$$

Provided that $|\Delta|^2 \geq \frac{4(N-1)}{N^2}$, this equation has the following two solutions, denoted by the superscript $\pm$

$$\left|\hat{\phi}^{\pm}\right| = \frac{N|\Delta|}{2(N - 1)}\left[1 \pm \sqrt{1 - \frac{4(N - 1)}{N^2|\Delta|^2}}\right] \tag{32}$$

This solution gives the absolute value of $\hat{\phi}^{\pm}$. By replacing this expression back into the right hand side of Eq.(30), we obtain the solution for $\hat{\phi}^{\pm}$, given by

15

$$\hat{\phi}^{\pm} = \frac{N\Delta}{2(N-1)} \left[1 \pm \sqrt{1 - \frac{4(N-1)}{N^2|\Delta|^2}} \right] \tag{33}$$

While $\hat{\phi}^+$ is normalized, $\hat{\phi}^-$ is not, i.e. when $t = 0$, $\Delta(0) = 1$ implies $\hat{\phi}^+(0) = 1$, while $\hat{\phi}^-(0) = \frac{1}{N-1}$. The two solutions are of very different magnitudes: for large N the solution $\hat{\phi}^-$ vanishes ($|\hat{\phi}^-| \simeq \frac{1}{N|\Delta|}$), while $\hat{\phi}^+$ stays finite ($|\hat{\phi}^+| \simeq |\Delta|$). For all values of $|\Delta|$ for which the solutions $\hat{\phi}^{\pm}$ exist, the following relation holds: $|\hat{\phi}^+||\hat{\phi}^-| = \frac{1}{N-1}$.

We show that $\hat{\phi}^+$ is a stable solution of the iteration, while $\hat{\phi}^-$ is unstable. This can be seen by taking the absolute value of Eq.(29) and computing the derivative

$$\frac{d|\hat{\phi}_{n+1}|}{d|\hat{\phi}_n|} \bigg|_{|\hat{\phi}_n|=|\hat{\phi}^{\pm}|} = 1 \mp \sqrt{1 - \frac{4(N-1)}{N^2|\Delta|^2}} \tag{34}$$

This implies that, provided that the two solutions exist, i.e. provided that $|\Delta|^2 > \frac{4(N-1)}{N^2}$, then $\hat{\phi}^+$ is stable (derivative smaller than one) and $\hat{\phi}^-$ is unstable (derivative larger than one). When $|\Delta|^2 = \frac{4(N-1)}{N^2}$, the two solutions annihilates in a saddle node bifurcation. For $|\Delta|^2 < \frac{4(N-1)}{N^2}$ only the null solution, $\hat{\phi}_{sc}=0$, is available. That is always stable, as can be checked by computing

$$\lim_{|\hat{\phi}_n| \to 0} \frac{d|\hat{\phi}_{n+1}|}{d|\hat{\phi}_n|} = 0 \tag{35}$$

In summary, when $|\Delta|^2 \geq \frac{4(N-1)}{N^2}$, the iteration (29) reaches $\hat{\phi}_{sc} = 0$ for $|\hat{\phi}_0| < |\hat{\phi}^-|$ and $\hat{\phi}_{sc} = \hat{\phi}^+$ for $|\hat{\phi}_0| \geq |\hat{\phi}^-|$, while for $|\Delta|^2 < \frac{4(N-1)}{N^2}$, the unique, globally stable solution is $\hat{\phi}_{sc} = 0$. As described in the main text, we define the set B of t values as

$$B = \left\{ t : |\Delta(t)|^2 \geq \frac{4(N-1)}{N^2} \right\} \tag{36}$$

Hence, $\hat{\phi}_{sc}(t) = 0$ when $t \notin B$. However, $t \in B$ does not guarantee that $\hat{\phi}_{sc}(t) = \hat{\phi}^+(t)$, since a small initial guess $|\hat{\phi}_0(t)| < |\hat{\phi}^-(t)|$ would determine $\hat{\phi}_{sc}(t) = 0$. Then, we define A as the set of t values for which $|\hat{\phi}_0(t)| \geq |\hat{\phi}^-(t)|$, and hence $\hat{\phi}_{sc}(t) = \hat{\phi}^+(t)$. The arbitrary choice of the initial guess translates into the arbitrary choice of the set A, provided that $A \subseteq B$, as required by the existence of the nonzero solutions. This is summarized in Eq.(8), and concludes the derivation.

We conclude this section by calculating the mean and variance of the self-consistent estimate $\hat{f}_{sc}$, Eq.(11). In a finite neighborhood of $t = 0$, $\hat{\phi}_{sc}(t) = \hat{\phi}^+(t)$, which is continuous and infinitely differentiable at $t = 0$. Then, the mean and variance can be computed by the derivatives of $\hat{\phi}^+$ and $|\hat{\phi}^+|^2$ at $t = 0$, namely

$$E(x) = -i \frac{d\hat{\phi}^+}{dt} \bigg|_{t=0} = \frac{d\hat{\phi}^+}{d\Delta} \bigg|_{\Delta=1} \left(-i \frac{d\Delta}{dt} \bigg|_{t=0} \right) \tag{37}$$

$$\text{Var}(x) = -\frac{1}{2}\frac{d^2|\hat{\phi}^+|^2}{dt^2}\bigg|_{t=0} = \frac{d|\hat{\phi}^+|^2}{d|\Delta|^2}\bigg|_{|\Delta|^2=1}\left(-\frac{1}{2}\frac{d^2|\Delta|^2}{dt^2}\bigg|_{t=0}\right) \qquad (38)$$

where we used the chain rule and the fact that $|\Delta|^2$ is an even function, hence $\frac{d|\Delta|^2}{dt}\big|_{t=0} = 0$. It is straightforward to show that the terms in round brackets in Eqs.(37,38) are, respectively, equal to the sample mean and sample variance. Because $|\Delta|$ is even, then $\frac{d\hat{\phi}^+}{d\Delta}\big|_{\Delta=1} = 1$ and, by differentiating Eq.(31), we obtain $\frac{d|\hat{\phi}^+|^2}{d|\Delta|^2}\big|_{|\Delta|^2=1} = N/(N-2)$. Finally, Eqs.(37,38) can be rewritten as

$$E(x) = \frac{1}{N}\sum_{j=1}^{N} X_j \qquad (39)$$

$$\text{Var}(x) = \frac{1}{N-2}\sum_{j=1}^{N}(X_j - E(x))^2 \qquad (40)$$

Appendix 3: Asymptotic convergence of $\hat{f}_{sc}$

In this section we investigate the asymptotic (large N) behavior of the self-consistent estimate. We study the sufficient conditions for the estimate $\hat{f}_{sc}$ to converge to the true density f for $N \to \infty$. In particular, we prove the following

Theorem

If the true density $f(x)$ is square integrable and its transform is integrable, then the self-consistent density estimate $\hat{f}_{sc}(x)$, defined by Eqs.(5),(8-11), converges almost surely to the true density for large N, under the additional assumptions

$$\lim_{N\to\infty} t^* = \infty \qquad (41)$$

$$\lim_{N\to\infty} \frac{t^*}{\sqrt{N}} = 0 \qquad (42)$$

Proof

Because the true density $f(x)$ and the self-consistent estimate $\hat{f}_{sc}(x)$ are both square integrable, we can express them as Fourier transform of, respectively, $\phi(t)$ and $\hat{\phi}_{sc}(t)$. By assumption, the characteristic function is integrable, i.e.

$$\int |\phi(t)|dt < \infty \qquad (43)$$

In the following sequence of inequalities, we find an upper bound for the difference between the true density and its estimate, and we use the fact that $\hat{\phi}_{sc}(t) = 0$ for $|t| > t^*$. In order to prove the theorem, we show that the upper bound tends to zero for large N.

$$\left| \hat{f}_{sc}(x) - f(x) \right| = \left| \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-itx} \left[\hat{\phi}_{sc}(t) - \phi(t) \right] dt \right| \le$$

$$\le \frac{1}{2\pi} \int_{-\infty}^{+\infty} \left| e^{-itx} \right| \left| \hat{\phi}_{sc}(t) - \phi(t) \right| dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \left| \hat{\phi}_{sc}(t) - \phi(t) \right| dt =$$

$$= \frac{1}{2\pi} \int_{-t^*}^{t^*} \left| \hat{\phi}_{sc}(t) - \phi(t) \right| dt + \frac{1}{2\pi} \int_{|t|>t^*} |\phi(t)| \, dt =$$

$$= \frac{1}{2\pi} \int_{-t^*}^{t^*} \left| \hat{\phi}_{sc}(t) - \Delta(t) + \Delta(t) - \phi(t) \right| dt + \frac{1}{2\pi} \int_{|t|>t^*} |\phi(t)| \, dt \le$$

$$\le \frac{1}{2\pi} \int_{-t^*}^{t^*} \left| \hat{\phi}_{sc}(t) - \Delta(t) \right| dt + \frac{1}{2\pi} \int_{-t^*}^{t^*} |\Delta(t) - \phi(t)| \, dt + \frac{1}{2\pi} \int_{|t|>t^*} |\phi(t)| \, dt$$

where $\Delta(t)$ is the empirical characteristic function, see Eq.(5). For increasing N, t^* increases following the limiting bounds given by Eqs.(41,42). Then, the second integral in the last expression tends to zero because of theorem 1 in ref. (Csorgo and Totik., 1983), while the third integral tends to zero because of the integrability of the characteristic function $\phi(t)$, Eq.(43). In order to prove the theorem, we have to demonstrate that also the first integral tends to zero for large N. We use the expression of $\hat{\phi}_{sc}$, Eq.(8), and we denote by Δ^+ and Δ^- the set of values of t for which $|\Delta(t)|^2$ is, respectively, above or below the threshold set by Eq.(9), i.e. $|\Delta(t)|^2 \ge 4(N-1)/N^2$ or $|\Delta(t)|^2 < 4(N-1)/N^2$. Then, the first integral is equal to

$$\frac{1}{2\pi} \int_{(-t^*,t^*) \cap \Delta^+} |\Delta(t)| \left\{ 1 - \frac{N}{2(N-1)} \left[1 + \sqrt{1 - \frac{4(N-1)}{N^2|\Delta(t)|^2}} \right] \right\} dt +$$

$$+ \frac{1}{2\pi} \int_{(-t^*,t^*) \cap \Delta^-} |\Delta(t)| \, dt$$

where the term in curly brackets is non-negative. The first term can be expanded by means of the inequality $\sqrt{1-x} \ge 1 - \sqrt{x}$ for $x \in (0,1)$, while in the second term we substitute the integrand with its maximum value in the interval, and we use the fact that the length of the interval is smaller than $2t^*$. Then, the above expression is smaller than

$$\le \frac{1}{2\pi} \int_{(-t^*,t^*) \cap \Delta^+} \left[\frac{1}{\sqrt{N-1}} - \frac{|\Delta(t)|}{N-1} \right] dt + \frac{2t^* \sqrt{N-1}}{\pi N} \le$$

$$\le \frac{1}{2\pi} \int_{(-t^*,t^*) \cap \Delta^+} \left[\frac{1}{\sqrt{N-1}} + \frac{|\Delta(t)|}{N-1} \right] dt + \frac{2t^* \sqrt{N-1}}{\pi N} \le$$

$$\le \frac{t^*}{\pi} \left[\frac{1}{\sqrt{N-1}} + \frac{1}{N-1} + \frac{2\sqrt{N-1}}{N} \right]$$

where we used $|\Delta(t)| \le 1$ in the last inequality. Because of Eq.(42), the last expression tends to zero for large N, thus proving the theorem.

Comment

The above theorem assumes that the true density is square integrable and its transform integrable. We expect the self-consistent estimate to perform poorly when applied to densities that do not meet those criteria, and its asymptotic consistency is not guaranteed. Examples of such distributions include the box function and the χ^2 distribution with one degree of freedom.

The condition (42) gives the limit on the frequency bound t^*, setting the maximum increase of t^* with N. In simulations, the magnitude of the bound t^* depends on the threshold of the empirical characteristic function, Eq.(9), in such a way that its amplitude is larger than the threshold in half of the interval $(-t^*, t^*)$. Now we argue that this choice satisfies the assumptions of the theorem, without giving a formal derivation.

To simplify the argument, we assume that the distribution is sufficiently smooth so that the following limit exists

$$\lim_{|t| \to \infty} t|\phi(t)| = 0, \tag{44}$$

notice that, if the above limit exist, then it must be equal to zero as a consequence of the integrability of the characteristic function, Eq.(43). We also approximate the empirical characteristic function $\Delta(t)$ with the true characteristic function $\phi(t)$, which is reasonable provided that N is large and t does not increases exponentially with N (see Csorgo and Totik. (1983)). Then, the threshold condition (9) becomes $|\phi(t)|^2 \geq 4(N-1)/N^2$. Hence, t^* increases in such a way that, for large N, the order of magnitude of the characteristic function is $|\phi(t^*)| \sim 1/\sqrt{N}$. Substituting this expression in Eq.(44) we get

$$\lim_{t^* \to \infty} t^*|\phi(t^*)| = \lim_{N \to \infty} \frac{t^*}{\sqrt{N}} = 0 \tag{45}$$

which correspond to condition (42). On the other hand, since the true cumulative distribution is continuous, then $|\phi(t)| > 0$ for any finite value of t, and this implies that t^* tends to infinity for large N, as required by the condition (41).

In conclusion, the above theorem gives the sufficient conditions for the asymptotic consistency of the estimate, especially concerning the finite interval of frequencies set by the bound t^*. From the above arguments, we expect the recipe for t^* used in simulations to guarantee the asymptotic convergence of the estimate.

Acknowledgements

References

BERLINET, A. (1993). Hierarchies of higher order kernels. *Probab. Theory Relat. Fields*, **94**, 489-504.

BIALEK, W., CALLAN, C.G., and STRONG, S.P. (1996). Field theories for learning probability distributions. *Phys. Rev. Lett.*, **77** 4693-4697.

BINDER, K. (1986). *Monte Carlo methods in statistical physics.* Springer, Verlag NY.

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

BREYMANN W., DIAS A., and EMBRECHT P. (2003). Dependence Structures for Multivariate HighFrequency Data in Finance. *Quantitative Finance* **3**(1) 1-14.

BROWN, E.N., KASS, R.E. and MITRA, P.P. (2004). Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature Neuroscience* **7**(5) 456-461.

CLAUSET, A., SHALIZI, C.,R., NEWMAN, M.,E.,J. (2009). Power-law distributions in empirical data. *SIAM Review* **51**(4) 661-703.

CSORGO, S. and TOTIK, V. (1983). On how long interval is the empirical characteristic function uniformly consistent?. *Acta Sci. Math.*, **45** 141-149.

DAVIS, K.B. (1977). Mean Integrated Square Error Properties of Density Estimates. *Ann. Stat.* **5** 530-535.

DEVROYE, L. (1992). A note on the usefulness of superkernels in density estimation. *Ann. Stat.*, **20**, 2037-2056.

EFROMOVICH, S. (2008). Adaptive estimation of and oracle inequalities for probability densities and characteristic functions. *Ann. Stat.*, **36**, 1127-1155.

GLAD, I.K., HJORT, N.L., and USHAKOV N.G. (2003) Correction of Density Estimators that are not Densities. *Scand. J. Stat.*, **30**, 415-427.

GLAD, I.K., HJORT, N.L. and USHAKOV, N. G. (2007) Density estimation using the sinc kernel. *Preprint Statistics*, **2**:1-20.

GOLDSTEIN, M.L, MORRIS S. A., and YEN G.G. (2004). Problems with fitting to the power-law distribution. *Europ. Phys. J. B* **41** 255-258.

GOOD, I.J. and GASKINS, R.A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika*, **58** 255.

HALL, P. and MARRON, J.S. (1987) Choice of kernel order in density estimation. *Ann. Stat.*, **16** 161-173.

HOLY, T.E. (1997). Analysis of Data from Continuous Probability Distributions. *Phys. Rev. Lett.* **79** 3545-3548.

HOSSJER O. (1996) Asymptotic bias and variance for a general class of varying bandwidth density estimators. *Probab. Theory Relat. Fields* **105**, 159-192.

KASS, R. E., VENTURA, V., and BROWN, E. N. (2005). Statistical issues in the analysis of neuronal data. *J Neurophysiol*, **94**(1) 8-25.

KLEYWEGT, G. J. and JONES, T. A. (1996). Phi/Psi-chology: Ramachandran revisited. *Structure* **4**(12) 1395-1400.

MARRON, J. S. and WAND, M. P. (1992) Exact mean integrated square error. *Ann. Stat.*, **20**(2)712-736.

MANN, H. B. and WALD, A. (1948). On Stochastic Limit and Order Relationships. *Ann. Math. Stat.* **14**(3) 217-226.

ODLYZKO, A.M. (1987). On the distribution of spacings between zeros of the Zeta function. *Math. Comp.* **48** 273-308.

OLSON, C.,R., GETTNER, S.N., VENTURA, V., CARTA, R., and KASS, R.E. (2000). Neuronal Activity in Macaque Supplementary Eye Field During Planning of Saccades in Response to Pattern and Spatial Cues. *J. Neurophys.* **84** 1369-1384.

PARZEN, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Stat.*, **33** 1065-1076

PERIWAL, V. (1997). Reparametrization invariant statistical inference and gravity. *Phys. Rev. Lett.*, **78** 4671-4674.

POLITIS, D.N. (2003) Adaptive bandwidth choice. *Nonparam. Stat.*, **15**(4-5)517-533.

RIPLEY B. D. and SUTHERLAND I. (1990). Finding spiral structures in images of galaxies. *Phyl. Trans. Roy. Soc.* **332**(1672) pp.477-485.

RUPPERT, D., and KLINE, B.H. (1994). Bias reduction in kernel density estimation by smoothed empirical transformations. *Ann. Stat.* **22** 185-210.

SCHMIDT, D.M. (2000). Continuous probability distributions from finite data. *Phys. Rev. E* **61** 1052-1055.

SCOTT, D. (1979). On optimal and data-based histograms. *Biometrika* **66** 605-610.

SILVERMAN, B.W. (1986). *Density estimation for statistics and data analysis*. Chapman & Hall.

USHAKOV, B.W. (1999). *Selected topics in characteristic functions*. Utrecht: VSP.

WAND, M.P. and JONES. M.C. (1995). *Kernel Smoothing*. Chapman & Hall.

WATSON, G.S., LEADBETTER, M.R. (1963). On the Estimation of the Probability Density. *Ann. Math. Stat.* **34** 480-491.

WIENER, N. (1949). *Extrapolation, Interpolation, and smoothing of stationary time series*. New York, Wiley.