PHYSICAL REVIEW E **90**, 011301(R) (2014)

Estimation of probability densities using scale-free field theories

Justin B. Kinney*

Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

(Received 14 January 2014; published 11 July 2014)

The question of how best to estimate a continuous probability density from finite data is an intriguing open problem at the interface of statistics and physics. Previous work has argued that this problem can be addressed in a natural way using methods from statistical field theory. Here I describe results that allow this field-theoretic approach to be rapidly and deterministically computed in low dimensions, making it practical for use in day-to-day data analysis. Importantly, this approach does not impose a privileged length scale for smoothness of the inferred probability density, but rather learns a natural length scale from the data due to the tradeoff between goodness of fit and an Occam factor. Open source software implementing this method in one and two dimensions is provided.

DOI: 10.1103/PhysRevE.90.011301 PACS number(s): 02.50.−r, 11.10.−z, 02.60.−x

Suppose we are given N data points, $x_1, x_2, \ldots, x_N$, each of which is a D-dimensional vector drawn from a smooth probability density $Q_{\text{true}}(x)$. How might we estimate Q_{true} from these data? This classic statistics problem is known as "density estimation" [1] and is routinely encountered in nearly all fields of science. Ideally, one would first specify a Bayesian prior $p(Q)$ that weights each density $Q(x)$ according to some sensible measure of smoothness. One would then compute a Bayesian posterior $p(Q|\text{data})$ identifying which densities are most consistent with both the data and the prior. However, a practical implementation of this straightforward approach has yet to be developed, even in low dimensions.

This Rapid Communication discusses one such strategy, the main theoretical aspects of which were worked out by Bialek *et al.* [2]. One first assumes a specific smoothness length scale ℓ. A prior $p(Q|\ell)$ that strongly penalizes fluctuations in Q below this length scale is then formulated in terms of a scalar field theory. The maximum *a posteriori* (MAP) density Q_ℓ, which maximizes $p(Q|\ell,\text{data})$ and serves as an estimate of Q_{true}, is then computed as the solution to a nonlinear differential equation. This approach has been implemented and further elaborated by others [3–10]; a connection to prior statistics literature on "penalized likelihood" should also be noted [1].

The question of how to choose ℓ remains. Bialek *et al.* argued on theoretical grounds that the data themselves will typically select a natural value for this smoothness length scale due to the competing influences of goodness of fit and an Occam factor [11]. Specifically, if one adopts a "scale-free" prior $p(Q)$, defined as a linear combination of scale-dependent priors $p(Q|\ell)$, then the posterior distribution over length scales $p(\ell|\text{data})$ will become sharply peaked in the large data limit. This important insight was confirmed computationally by Nemenman and Bialek [3] and provides a compelling alternative to cross validation, the standard method of selecting length scales in statistical smoothing problems [1].

*Corresponding author: jkinney@cshl.edu

However, computing $p(\ell|\text{data})$ requires first computing Q_ℓ at every relevant length scale, i.e., solving an infinite compendium of nonlinear differential equations. Nemenman and Bialek [3] approached this problem by independently computing Q_ℓ at a series of length scales chosen by a standard optimization routine. Although this method yielded important results, it also has significant limitations. First, there is no guarantee that this strategy will find the globally optimum length scale. Second, this approach was observed to be very computationally demanding and no implementation has since been developed for general use. Indeed, even simple performance comparisons to more standard density estimation methods have yet to be reported.

Here I describe a rapid and deterministic homotopy method for computing Q_ℓ to a specified accuracy at all relevant length scales. This makes low-dimensional density estimation using scale-free field-theoretic priors practical for use in day-to-day data analysis. The open source "density estimation using field theory" (DEFT) software package [12] provides a Python implementation of this algorithm for one-dimensional (1D) and two-dimensional (2D) problems. Simulation tests show favorable performance relative to standard Gaussian mixture model (GMM) and kernel density estimation (KDE) approaches [1].

Following Refs. [2,3] we begin by defining $p(Q)$ as a linear combination of scale-dependent priors $p(Q|\ell)$,

$$p(Q) = \int_0^\infty d\ell\; p(Q|\ell)\; p(\ell). \tag{1}$$

Adopting the Jeffreys prior $p(\ell) \sim \ell^{-1}$ renders $p(Q)$ covariant under a rescaling of x [11]. Our ultimate goal will be to compute the resulting posterior,

$$p(Q|\text{data}) = \int_0^\infty d\ell\; p(Q|\ell,\text{data}) p(\ell|\text{data}). \tag{2}$$

As in Ref. [3], we limit our attention to a D-dimensional cube having volume $V = L^D$. We further assume periodic boundary conditions on Q, and impose G^D grid points (G in each dimension) at which Q will be computed.

To guarantee that each density is positive and normalized, we define Q in terms of a real scalar field ϕ via

$$Q(x) = \frac{e^{-\phi(x)}}{\int d^D x'\, e^{-\phi(x')}}. \tag{3}$$

Each Q corresponds to multiple different ϕ, but there is a one-to-one correspondence with fields ϕ_{nc} that have no constant Fourier component. Using this fact, we adopt the standard path integral measure $\mathcal{D}\phi_{nc}$ as the measure on Q space. We also define the prior $p(Q|\ell)$ in terms of a field theory on ϕ_{nc},

$$p(Q|\ell) = \frac{1}{Z_\ell^0} \exp\left[-\int d^D x \frac{\ell^{2\alpha-D}}{2} \phi_{nc} \Delta \phi_{nc} \right], \tag{4}$$

where α is a positive integer and $\Delta = (-\nabla^2)^\alpha$ is a differential operator that formalizes our notion of "smoothness." Z_ℓ^0 is the corresponding normalization factor. This prior effectively constrains the α-order derivatives of ϕ_{nc}, strongly dampening Fourier modes that have wavelength much less than ℓ.

Applying Bayes's rule to this prior yields the following exact expression for the posterior [13],

$$p(Q|\ell,\text{data}) = \frac{1}{Z_\ell^N} \int_{-\infty}^{\infty} d\phi_c \, e^{-S_\ell[\phi]}, \tag{5}$$

where

$$S_\ell[\phi] = \int d^D x \left[\frac{\ell^{2\alpha-D}}{2} \phi \Delta \phi + NR\phi + \frac{N}{V} e^{-\phi} \right] \tag{6}$$

is an "action" that constrains the field ϕ, $R(x) = N^{-1} \sum_{n=1}^N \delta(x - x_n)$ is the raw data density, $\phi(x) = \phi_{nc}(x) + \phi_c$, and $Z_\ell^N = Z_\ell^0 \, \Gamma(N)(V/N)^N \, p(\text{data}|\ell)$.

The action S_ℓ was described by Ref. [2] and explored in later work [3,8–10]. It also corresponds to one form of log penalized likelihood discussed in earlier statistics literature [1]. Note, however, that Ref. [2] used $Q(x) = \text{const} \times e^{-\phi(x)}$ in place of Eq. (3) and explicitly enforced normalization in their prior $p(Q|\ell)$. Equation (6) was then derived using a large N saddle point approximation. By contrast, Eqs. (5) and (6) are exact under the formulation presented here.

The MAP density Q_ℓ corresponds to the field ϕ_ℓ that minimizes the action in Eq. (6). To find this minimum we set $\delta S_\ell / \delta \phi = 0$, which gives a nonlinear differential equation for ϕ_ℓ,

$$\ell^{2\alpha-D} \Delta \phi_\ell + N \left[R - \frac{e^{-\phi_\ell}}{V} \right] = 0. \tag{7}$$

The central finding of this Rapid Communication is that, instead of independently solving Eq. (7) at various length scales ℓ, we can compute ϕ_ℓ to a specified accuracy at all length scales of interest by using a homotopy method [14]. First we define $t = \ln(N/\ell^{2\alpha-D})$, a convenient reparametrization of ℓ. Next we differentiate Eq. (7) with respect to t, obtaining

$$[e^{-t}\Delta + Q_\ell] \frac{d\phi_\ell}{dt} = Q_\ell - R, \tag{8}$$

where $Q_\ell(x) = e^{-\phi_\ell(x)}/V$ is the probability density corresponding to ϕ_ℓ. If ϕ_ℓ is known at any specific length scale ℓ_i, we can compute it at any other length scale ℓ_f—and at all length scales in between—by integrating Eq. (8) from ℓ_i to ℓ_f. Because $S_\ell[\phi]$ is a strictly convex function of ϕ, each ϕ_ℓ so identified will be the unique minimum. Moreover, because Eq. (6) is exact, each corresponding Q_ℓ will fit the data optimally even when N is small. And since the matrix representation of $e^{-t}\Delta + Q_\ell$ is sparse, $d\phi_\ell/dt$ can be rapidly computed at each successive value of t using standard sparse matrix methods.

To identify a length scale ℓ_i from which to initiate the integration of Eq. (8), we look to the large length scale limit where a weak field approximation can be used to compute ϕ_{ℓ_i}. Linearizing Eq. (7) and solving for ϕ_ℓ gives, for $|k| > 0$, $\hat{\phi}_\ell(k) = -\frac{V\hat{R}(k)}{1+\exp[\tau_k - t]}$, where hats denote Fourier transforms, $k \in \mathbb{Z}^D$ indexes the Fourier modes of the volume V, and each $\tau_k = \ln[(2\pi|k|)^{2\alpha} L^{D-2\alpha}]$ is a log eigenvalue of $V\Delta$. To guarantee that none of the Fourier modes of ϕ_{ℓ_i} are saturated, ℓ_i should correspond to a value t_i that is sufficiently less than $\min_{|k|>0} \tau_k$. This implies $\ell_i \gg N^{\frac{1}{2\alpha-D}} L$. Similarly, we terminate the integration of Eq. (8) at a length scale ℓ_f above which Nyquist modes saturate. This yields the criterion $\ell_f \ll n^{\frac{1}{2\alpha-D}} h$, where $h = L/G$ is the grid spacing and $n = N/G^D$ is the number of data points per voxel.

Having computed ϕ_ℓ at every relevant length scale, a semiclassical approximation yields

$$p(\ell|\text{data}) = \text{const} \times p(\ell) \frac{e^{-S_\ell[\phi_\ell]}}{\sqrt{\ell^{2\alpha-D} \det[\Delta + e^t Q_\ell]}}. \tag{9}$$

The ratio in Eq. (9) is equal to the MAP density likelihood times an Occam factor. This likelihood quantifies goodness of fit and steadily increases as ℓ gets smaller. The Occam factor, by contrast, decreases as ℓ gets smaller due to the decreasing fraction of model space consistent with the data [11]. As discussed by Refs. [2,3], this tradeoff causes $p(\ell|\text{data})$ to peak at a nontrivial data-determined length scale ℓ^*.

The length scale prior $p(\ell)$ must decay faster than ℓ^{-1} in the infrared in order for $p(\ell|\text{data})$ to be normalizable. The need for such regularization reflects a redundancy among the priors $p(Q|\ell)$ at large ℓ that results from the volume V supporting a limited number of long wavelength Fourier modes. Similar concerns hold in the ultraviolet due to our use of a grid. We therefore set $p(\ell) = 0$ for $\ell > \ell_i$ and for $\ell < \ell_f$.

Our best estimate of Q_{true} is given by the MAP density Q^* corresponding to the length scale ℓ^* that maximizes $p(\ell|\text{data})$. Uncertainties in this estimate can be computed by sampling $Q \sim p(Q|\text{data})$. We do this by first sampling $\ell \sim p(\ell|\text{data})$, then selecting Q according to a semiclassical approximation of $p(Q|\ell,\text{data})$ by choosing

$$\phi(x) = \phi_\ell(x) + \sum_{j=1}^{G^D} \frac{\eta_j}{\sqrt{\ell^{2\alpha-D} \lambda_j^\ell}} \psi_j^\ell(x), \tag{10}$$

where each ψ_j^ℓ is a normalized eigenfunction of the operator $\Delta + e^t Q_\ell$, λ_j^ℓ is the corresponding eigenvalue, and each η_j is an independent normally distributed random variable. If $p(\ell|\text{data})$ is strongly peaked, all ψ_j^ℓ and λ_j^ℓ can be evaluated at $\ell = \ell^*$ to reduce the computational burden.

Figure 1 illustrates the key steps of the DEFT algorithm. First, the user specifies a data set $\{x_n\}_{n=1}^N$, a bounding box for the data, and the number of grid points to be used. A histogram of the data is then computed using bins that are centered on each grid point [Fig. 1(a)]. Next, length scales ℓ_i and ℓ_f are chosen. Equation (8) is then integrated to yield ϕ_ℓ at a set of length scales between ℓ_i and ℓ_f chosen automatically

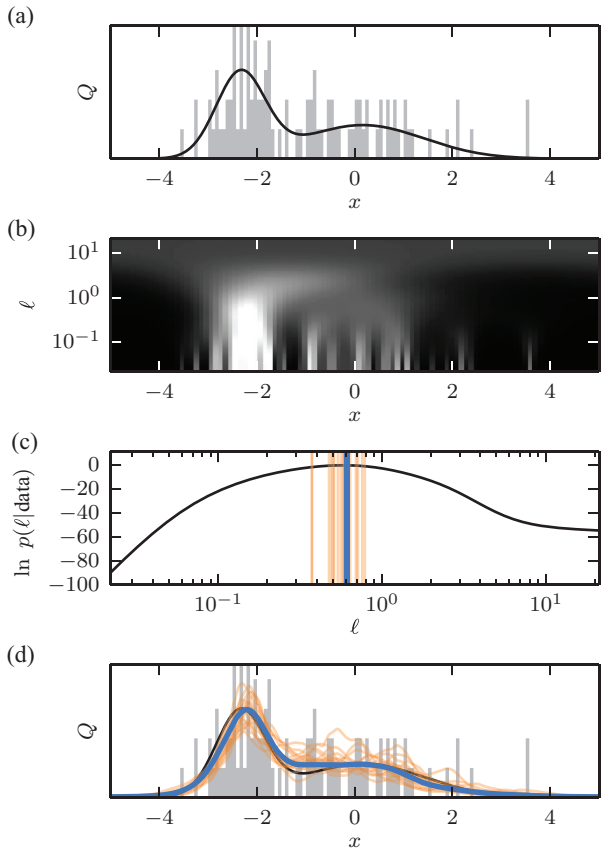

FIG. 1. (Color) Illustration of DEFT in 1D. (a) An example density $Q_{\text{true}}(x)$ (black) along with a normalized histogram (gray, 100 bins) of $N = 100$ sampled data points. (b) Heat map showing all of the MAP densities $Q_\ell(x)$ computed at $G = 100$ grid points using the data from (a) and the parameter $\alpha = 2$; lighter shading corresponds to higher probability. (c) Posterior probability for each length scale shown in (b); the y axis is shifted so that $\ln p(\ell^*|\text{data}) = 0$. (d) The estimated density $Q^*(x)$ (blue) along with 20 densities (orange) sampled from $p(Q|\text{data})$ using Eq. (10); corresponding length scales are shown in (c).

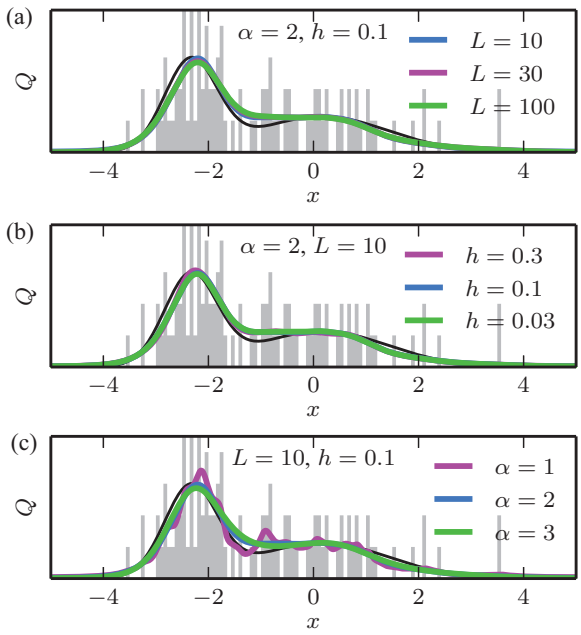

FIG. 2. (Color) Robustness of DEFT to changes in runtime parameters. $Q^*(x)$ was computed using the data from Fig. 1 and various choices for (a) the length L of the bounding box, (b) the grid spacing h, and (c) the order α of the derivative constrained by the field theory prior. $L = 10$ corresponds to the bounding box shown, and $h = 0.1$ is the grid spacing used for the histogram (gray). $Q_{\text{true}}(x)$ is shown in black.

by a standard ordinary differential equation (ODE) solver to achieve the desired accuracy. Equation (9) is then used to compute $p(\ell|\text{data})$ at each of these length scales, after which ℓ^* is identified. Finally, a specified number of densities are sampled from $p(Q|\text{data})$ using Eq. (10).

DEFT is not completely scale free because both the box size L and grid spacing h are prespecified by the user. In practice, however, Q^* appears to be very insensitive to the specific values of L and h as long as the data lie well within the bounding box and the grid spacing is much smaller than the inherent features of Q_{true}; see Figs. 2(a) and 2(b).

It is interesting to consider how the choice of α affects Q^*. As Bialek *et al.* have discussed [2], this field-theoretic approach produces ultraviolet divergences in ϕ_ℓ when $\alpha < D/2$. Above this threshold, increasing α typically increases the smoothness of Q^*, although not necessarily by much [see Fig. 2(c)]. However, larger values of α may necessitate more data before the principal Fourier modes of Q_{true} appear in Q^*. Increasing α also reduces the sparseness of the Δ matrix, thereby increasing the computational cost of the homotopy method.

To assess how well DEFT performs in comparison to more standard density estimation methods, a large number of data sets were simulated, after which the accuracy of Q^* produced by various estimators was computed. Specifically, the "closeness" of Q_{true} to each estimate Q^* was quantified using the natural geodesic distance [15],

$$D_{\text{geo}}(Q_{\text{true}}, Q^*) = 2\cos^{-1}\left[\int d^D x \sqrt{Q_{\text{true}} Q^*}\right]. \quad (11)$$

As shown in Fig. 3, DEFT performed substantially better when $\alpha = 2$ or 3 than when $\alpha = 1$. This likely reflects the smoothness of the simulated Q_{true} densities. DEFT outperformed the KDE method tested here and, for $\alpha = 2$ and 3, performed as well or better than GMM. This latter observation suggests nearly optimal performance by DEFT, since each simulated Q_{true} was indeed a mixture of Gaussians.

In two dimensions, DEFT shows a remarkable ability to discern structure from a limited amount of data (Fig. 4). As in 1D, larger values of α give a smoother estimate Q^*. However, DEFT requires substantially more computational power in 2D than in 1D due to the increase in the number of grid points and the decreased sparsity of the Δ matrix. For instance, the computation shown in Fig. 1 took about 0.3 s, while the DEFT computations in Fig. 4 took about 1–3 s each [16].

Field-theoretic density estimation faces two significant challenges in higher dimensions. First, the computational approach described here is impractical for $D \gtrsim 3$ due to the enormous number of grid points that would be needed. It should be noted, however, that the 1D field theory discussed

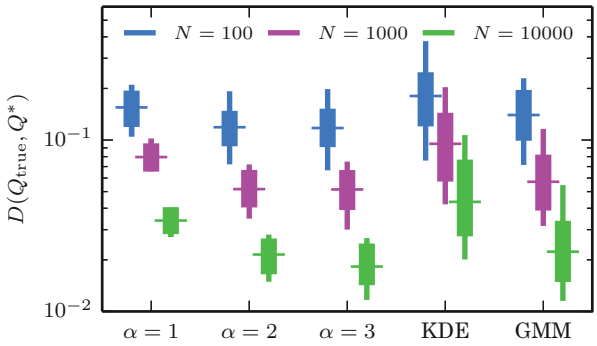

FIG. 3. (Color) Comparison of 1D density estimation methods. One hundred different densities $Q_{\text{true}}(x)$ were generated, each as the sum of five random Gaussians. Data sets of various size N were then drawn from each Q_{true}, after which estimates Q^* were computed using DEFT ($G = 100$, various α), KDE (using Scott's rule to set kernel bandwidth), and GMM (using the Bayesian information criterion to choose the number of components). Accuracy was quantified using the geodesic distance $D_{\text{geo}}(Q_{\text{true}}, Q^*)$ shown in Eq. (11). Box plots indicate median, interquartile range, and 5%–95% quantiles.

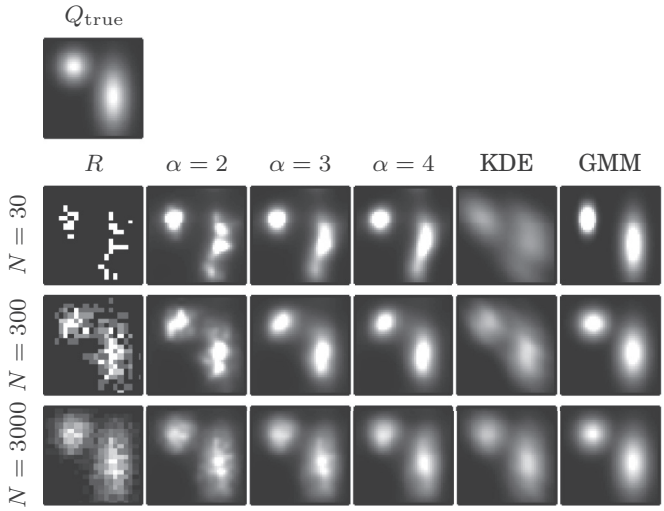

FIG. 4. Density estimation in 2D. Shown is a simulated density Q_{true} composed of two Gaussians, normalized histograms R for sampled data sets of various size N, and resulting density estimates Q^* computed using DEFT ($G = 20$), KDE, and GMM. The grayscale in all plots is calibrated to Q_{true}.

by Holy [4] allows Q_ℓ to be computed without using a grid. It may be possible to extend this approach to higher dimensions.

The "curse of dimensionality" presents a more fundamental problem. As discussed by Bialek *et al.* [2], this manifests in the fact that increasing D requires a proportional increase in α, i.e., in one's basic notion of "smoothness." This likely indicates a fundamental problem with high dimensional priors of the form shown in Eq. (4). Using a different operator Δ, e.g., one with reduced rotational symmetry, might provide a way forward.

I thank Gurinder Atwal, Anne-Florence Bitbol, Daniel Ferrante, Daniel Jones, Bud Mishra, Swagatam Mukhopadhyay, and Bruce Stillman for helpful conversations. I also thank the anonymous referees for providing valuable feedback. Support for this work was provided by the Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory.

[1] B. W. Silverman, *Density Estimation for Statistics and Data Analysis* (Chapman and Hall/CRC, London, 1986).
[2] W. Bialek, C. G. Callan, and S. P. Strong, Phys. Rev. Lett. **77**, 4693 (1996).
[3] I. Nemenman and W. Bialek, Phys. Rev. E **65**, 026137 (2002).
[4] T. E. Holy, Phys. Rev. Lett. **79**, 3545 (1997).
[5] V. Periwal, Phys. Rev. Lett. **78**, 4671 (1997).
[6] T. Aida, Phys. Rev. Lett. **83**, 3554 (1999).
[7] D. M. Schmidt, Phys. Rev. E **61**, 1052 (2000).
[8] J. C. Lemm, *Bayesian Field Theory* (The Johns Hopkins University Press, Baltimore, MD, 2003).
[9] I. Nemenman, Neural Comput. **17**, 2006 (2005).
[10] T. A. Enßlin, M. Frommert, and F. S. Kitaura, Phys. Rev. D **80**, 105005 (2009).
[11] V. Balasubramanian, Neural Comput. **9**, 349 (1997).
[12] Available at http://github.com/jbkinney/13_deft.
[13] The identity $a^{-N} = \frac{1}{\Gamma(N)} \int_{-\infty}^{\infty} du \exp[-Nu - ae^{-u}]$ is used here with $a = (N/V) \int d^D x\, e^{-\phi_{nc}(x)}$ and $u = \phi_c$.
[14] E. L. Allgower and K. Georg, *Numerical Continuation Methods: An Introduction* (Springer, Berlin, 1990).
[15] J. Skilling, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 27th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, edited by K. H. Knuth, A. Caticha, J. L. Center, A. Giffin, and C. C. Rodríguez, AIP Conf. Proc. No. 954 (AIP, Melville, NY, 2007), p. 39.
[16] Computation times were assessed on a computer having a 2.8 GHz dual core processor, 16 GB of RAM, and running the Canopy Python distribution (Enthought).