Density Estimation on Small Data Sets

Wei-Chia Chen, Ammar Tareen, and Justin B. Kinney*
Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

(Received 12 April 2018; published 19 October 2018)

How might a smooth probability distribution be estimated with accurately quantified uncertainty from a limited amount of sampled data? Here we describe a field-theoretic approach that addresses this problem remarkably well in one dimension, providing an exact nonparametric Bayesian posterior without relying on tunable parameters or large-data approximations. Strong non-Gaussian constraints, which require a nonperturbative treatment, are found to play a major role in reducing distribution uncertainty. A software implementation of this method is provided.

DOI: 10.1103/PhysRevLett.121.160605

The need to estimate smooth probability distributions from a limited number of samples is ubiquitous in data analysis [1]. This "density estimation" problem also presents a fundamental conceptual challenge in statistical learning, important aspects of which remain unresolved. These outstanding problems are especially acute in the context of small data sets, where standard large-data set approximations do not apply. Here we investigate the potential for Bayesian field theory, an area of statistical learning based on field-theoretic methods in physics [2–5], to estimate probability densities in this small-data regime.

Density estimation requires answering two distinct questions. First, what is the *best* estimate for the underlying probability distribution? Second, what do other *plausible* distributions look like? Ideally, one would like to answer these questions by first considering all possible distributions (regardless of mathematical form), then identifying those that fit the data while satisfying a transparent notion of smoothness. Such an approach should not require one to manually identify values for critical parameters, specify boundary conditions, or make invalid mathematical approximations in the small-data regime. However, the most common density estimation approaches, including kernel density estimation (KDE) [1] and Dirichlet process mixture modeling (DPMM) [6,7], do not satisfy these requirements.

Building on Ref. [2], previous work has described a Bayesian field theory approach called density estimation using field theory (DEFT) [8,9] for addressing the density estimation problem in low dimensions. DEFT satisfies all of the above criteria except for the last one: In Refs. [8,9], an appeal to the large-data regime was used to justify a Laplace approximation (i.e., a saddle-point approximation) of the Bayesian posterior. This approximation facilitated the sampling of an ensemble of plausible densities, as well as the identification of an optimal smoothness length scale. Independent but closely related work [10] has also relied heavily on this approximation.

Here we investigate the performance of DEFT in the small-data regime and find that the Laplace approximation advocated in prior work can be catastrophic. This is because non-Gaussian features of the DEFT posterior are critical for suppressing "wisps"—large positive fluctuations that otherwise occur in posterior-sampled densities. We further find that these non-Gaussian effects cannot be addressed perturbatively using Feynman diagrams, as has been suggested in other Bayesian field theory contexts [4,5]. These results are not specific to DEFT but rather reflect the fundamentally nonperturbative nature of the density estimation problem.

Happily, we find that importance resampling [7] can rapidly and effectively correct for the Laplace approximation. The resulting DEFT algorithm, which we have made available in robust and easy-to-use software, thus appears to satisfy all of the above requirements for an ideal density estimation method in one dimension. Tests of DEFT on simulated data show favorable performance relative to KDE and DPMM. We also illustrate the utility of DEFT on real data from the Large Hadron Collider [11] and from the World Health Organization (WHO) [12].

We first recap the DEFT approach to density estimation [8,9]. Consider N data points $\{x_i\}_{i=1}^N$ drawn from a smooth one-dimensional probability distribution $Q_{\text{true}}(x)$ that is confined to an x interval of length L. From these data, we wish to obtain a best estimate Q^* of Q_{true}, as well as an ensemble of plausible distributions with which to quantify the uncertainty in this estimate.

DEFT reparametrizes each candidate distribution Q in terms of a field ϕ via

$$Q(x) = \frac{e^{-\phi(x)}}{\int dx' e^{-\phi(x')}}. \quad (1)$$

After adopting a Bayesian prior that constrains the α-order x derivative of ϕ (denoted by $\partial^\alpha\phi$ in what follows), and accounting for the likelihood of the data given ϕ, one obtains a posterior distribution on ϕ. We represent this posterior as $p(Q|\text{data}, \ell) \propto \exp(-S_\ell[\phi])$ where

$$S_\ell[\phi] = \int \frac{dx}{L}\left(\frac{\ell^{2\alpha}}{2}(\partial^\alpha\phi)^2 + NRL\phi + Ne^{-\phi}\right) \quad (2)$$

is the "posterior action" described in Ref. [9]. In Eq. (2), ℓ is a smoothness length scale that has yet to be determined, and $R(x) = (1/N)\sum_{i=1}^{N}\delta(x - x_i)$ is a histogram (of bin width zero) that summarizes the data. See Supplemental Material Section 1 (SM.1) [13] for details. The behavior of Q under this action $S_\ell[\phi]$ is the primary focus of the present Letter.

$S_\ell[\phi]$ is minimized at the maximum *a posteriori* (MAP) field ϕ_ℓ. The MAP field ϕ_ℓ is unique even in the absence of boundary conditions; see SM.2 [13] for details. Although ϕ_ℓ cannot be solved analytically, it is readily computed as the solution to a convex optimization problem after discretization of the x domain at G equally-spaced grid points. In this discrete representation, R becomes a histogram with bin width $h = L/G$. As long as $h \ll \ell$, the choice of G will not greatly affect ϕ_ℓ. The optimal length scale ℓ^* is identified by maximizing the Bayesian evidence, $p(\text{data}|\ell)$; see SM.3 [13] for details. $Q^* = Q_{\ell^*}$ is then used as our best density estimate. Figures 1(a)–1(c) illustrate this procedure on simulated data.

To characterize the uncertainty in the DEFT estimate Q^*, we sample the Bayesian posterior $p(Q|\text{data}) = \int d\ell\, p(\ell|\text{data})p(Q|\text{data}, \ell)$. Each sample is generated by first drawing ℓ from $p(\ell|\text{data})$, then drawing Q from $p(Q|\text{data}, \ell)$. Previous work [8] has suggested that this sampling task be performed using the Laplace approximation, i.e., approximating $p(Q|\text{data}, \ell)$ with a Gaussian distribution that has the same mean and Hessian. The corresponding action, $S_\ell^{\text{Lap}}[\phi]$, is thus quadratic in $\delta\phi = \phi - \phi_\ell$. This Laplace approximation has the advantage that posterior samples Q can be rapidly and independently generated [8].

Figure 1(d) shows multiple Qs sampled from the Laplace posterior $p_{\text{Lap}}(Q|\text{data}) = \int d\ell\, p(\ell|\text{data})p_{\text{Lap}}(Q|\text{data}, \ell)$. Clearly something is very wrong. Although many of these Qs appear reasonable, others exhibit wisps that have substantial probability mass far removed from the data.

We hypothesized that wisps are an artifact of the Laplace approximation. To correct for potential inaccuracies of this approximation, we adopted an importance resampling approach [7]. For each sampled ϕ, we computed a weight

$$w_\ell[\phi] = \exp\left(S_\ell^{\text{Lap}}[\phi] - S_\ell[\phi]\right). \quad (3)$$

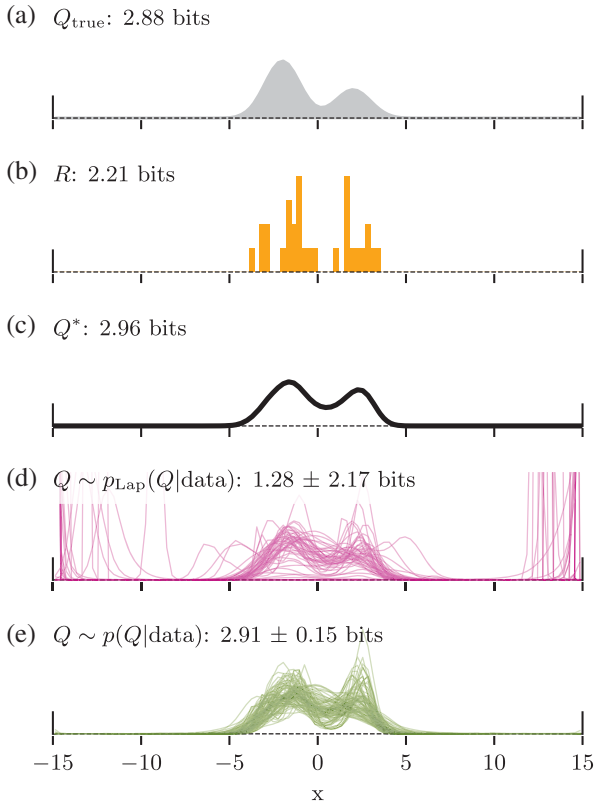

FIG. 1. Density estimation using field theory. (a) A Gaussian mixture distribution $Q_{\text{true}} = \frac{2}{3}\mathcal{N}(-2, 1) + \frac{1}{3}\mathcal{N}(2, 1)$ within the x interval $(-15, 15)$. (b) A histogram R of $N = 30$ data points sampled from Q_{true} and discretized to $G = 100$ grid points. (c) The corresponding estimate Q^* computed by DEFT using $\alpha = 3$ and the same grid as in (b). (d) One hundred distributions sampled from the Laplace-approximated posterior $p_{\text{Lap}}(Q|\text{data})$, which accounts for uncertainty in ℓ as well as in Q. (e) One hundred distributions generated using importance resampling of the Laplace ensemble. The differential entropies of the illustrated distributions are provided.

We then resampled the Laplace ensemble with replacement, selecting each ϕ (and thus Q) with a probability proportional to $w_\ell[\phi]$. A mixture of such resampled ensembles across length scales ℓ was then used to generate an ensemble reflecting $p(Q|\text{data})$; see SM.4 [13] for details. Figure 1(e) shows 100 distributions Q from this resampled posterior. Wisps no longer appear.

Eliminating wisps is especially important when estimating values for summary statistics, such as distribution entropy. In entropy estimation, the goal is to discern a value for the quantity $H_{\text{true}} = H[Q_{\text{true}}]$ where $H[Q] = -\int dx Q(x)\log_2 Q(x)$. Using the DEFT posterior ensemble, we can estimate H_{true} as $\hat{H} \pm \delta\hat{H}$, where $\hat{H} = \langle H\rangle$ and $\delta\hat{H} = \sqrt{\langle H^2\rangle - \langle H\rangle^2}$, with $\langle\cdot\rangle$ denoting a posterior average. Previous work expressed hope that the ensemble provided by the Laplace approximation might serve this

purpose [8]. But in this case, we see that $\hat{H}$ is far less accurate than the point estimates $H[R]$ or $H[Q^*]$, and $\delta\hat{H}$ is enormous [Fig. 1(d)]. Importance resampling fixes both problems: The resulting $\hat{H}$ is closer to H_{true} than either point estimate, and $\delta\hat{H}$ is remarkably small [Fig. 1(e)].

We now turn to the problem of understanding how wisps arise. To this end, we consider the variation in the action upon $\phi_\ell \to \phi_\ell + \delta\phi$. One finds that

$$\delta S_\ell[\phi_\ell + \delta\phi] = \int \frac{dx}{L}\frac{\ell^{2\alpha}}{2}(\partial^\alpha \delta\phi)^2 + \int \frac{dx}{L} V(\delta\phi), \tag{4}$$

where

$$V(\delta\phi) = NLQ_\ell[e^{-\delta\phi} - 1 + \delta\phi]. \tag{5}$$

The first (kinetic) term on the right-hand side of Eq. (4) imposes a smoothness constraint on $\delta\phi$, while the second (potential) term keeps $\delta\phi$ confined to a potential well consistent with the data. See SM.5 [13] for details. Note that V is convex, non-negative, and vanishes when $\delta\phi = 0$. By analogy to equipartition, we define n_{eff}, the effective number of degrees of freedom constrained by the data, as twice the value of the second term in Eq. (4) averaged over the posterior ensemble. Typical fluctuations $\delta\phi$ will therefore exhibit $V(\delta\phi) \sim n_{\text{eff}}/2$.

We now separately consider the "data-rich" regime of the x domain, which we define by $Q_\ell(x) \gg n_{\text{eff}}/2NL$, and the "data-poor" regime corresponding to $Q_\ell(x) \ll n_{\text{eff}}/2NL$. In the data-rich regime, fluctuations are small enough that V adheres well to its Laplace approximation, $V \approx NLQ_\ell\delta\phi^2/2$. Under this nearly symmetric potential, both positive fluctuations $\delta\phi^+$ and negative fluctuations $\delta\phi^-$ are constrained by

$$|\delta\phi^\pm| \sim \delta\phi_{\text{rich}} = \sqrt{\frac{n_{\text{eff}}}{NLQ_\ell}}. \tag{6}$$

By contrast, V is highly asymmetric in the data-poor regime and produces highly asymmetric fluctuations. Positive fluctuations satisfy $\delta\phi^+ \sim n_{\text{eff}}/2NLQ_\ell$, whereas negative fluctuations obey

$$-\delta\phi^- \sim \delta\phi^-_{\text{poor}} = \log\frac{n_{\text{eff}}}{2NLQ_\ell}. \tag{7}$$

See SM.5 [13] for more information.

The key point is that adopting $S_\ell^{\text{Lap}}[\phi]$ in place of $S_\ell[\phi]$ is equivalent to assuming the Laplace approximation for V throughout the entire x domain. Because $\delta\phi_{\text{rich}} \gg \delta\phi^-_{\text{poor}}$ in data-poor regions, the Laplace approximation greatly overestimates the size of downward fluctuations in ϕ. This results in the large upward fluctuations in Q that we identify as wisps. We note that wisps are especially prominent at the x-interval boundaries in Fig. 1 for two reasons: (i) Q_ℓ is especially small here, making these regions very data poor, and (ii) the kinetic term in Eq. (4), which is all that suppresses wisps in data-poor regions, is less effective at constraining $\delta\phi$ because data are present on only one side.

Feynman diagrams provide a general means of correcting for inaccuracies in Laplace approximations [14] and have been advocated in the context of some Bayesian field theory regression problems [4,5]. For density estimation, however, Feynman diagrams are ineffective if any region of the x interval is data poor. This is due to the action $S_\ell[\phi]$ being strongly coupled. For example, in the Bayesian evidence computations used to determine ℓ^*, DEFT estimates the action $Z_\ell = \int \mathcal{D}\phi e^{-S_\ell[\phi]}$ using the Laplace approximation $Z_\ell^{\text{Lap}} = \int \mathcal{D}\phi e^{-S_\ell^{\text{Lap}}[\phi]}$. See SM.3 [13] for details. At first, one might think it possible to correct for potential inaccuracies in this approximation using a series of vacuum diagrams (see SM.6 [13]), i.e.,

$$\log\frac{Z_\ell}{Z_\ell^{\text{Lap}}} = \quad + \quad + \quad + \cdots. \tag{8}$$

However, as described in SM.8 [13], the number of diagrams needed to obtain accurate results is prohibitive when data-poor regions of the x interval are present. Fortunately, one can instead compute nonperturbative corrections to this log ratio using the importance resampling weights in Eq. (3) via

$$\log\frac{Z_\ell}{Z_\ell^{\text{Lap}}} = \log\langle w_\ell\rangle_{\text{Lap}|\ell}. \tag{9}$$

See SM.7 [13] for details.

These results reflect a fundamental yet underappreciated aspect of density estimation: Unless data are observed throughout the x domain, the uncertainties in estimated probability densities require a nonperturbative treatment. Specifically, nonperturbative methods such as the Laplace approximation or Feynman diagrams can only be expected to work if $Q_{\text{true}}(x) \gtrsim 1/NL$ everywhere within the x domain. Very often, however, density estimation is applied to data like that in Fig. 1, which are localized far away from one or both x-interval boundaries. We argue that the analysis of such data will quite generally require a nonperturbative treatment.

To benchmark the performance of DEFT, we quantified its ability to estimate probability densities of known functional form. Specifically, we simulated data sets of varying size N from a variety of Q_{true} distributions, then asked two questions. First, how accurately does Q^* estimate Q_{true}?

Second, how typical is Q_{true} among the distributions in the Bayesian posterior? In both contexts, DEFT was compared to KDE and DPMM. See SM.9 [13] for details on how KDE and DPMM were implemented. Figure 2 shows the results of these performance tests for two different choices of Q_{true}. Figure S3 in Supplemental Material [13] provides analogous results for other Q_{true} distributions.

To answer the first question, we compared the Kullback-Leibler divergence $D_{\text{KL}}(Q_{\text{true}}\|Q^*)$ achieved by each estimator on each data set. Note that smaller values for these divergences indicate better method accuracy. As illustrated in Fig. 2(b), DEFT performed comparably to KDE and DPMM at $N = 10$ and somewhat better at $N = 100$. DEFT appears to have a particular advantage over both KDE and DPMM on Q_{true} distributions that bump up against one or both x-interval boundaries. Also unsurprising is that DEFT performs notably better with $\alpha = 2$, 3, and 4 than with $\alpha = 1$, since $\alpha = 1$ yields nonsmooth Q^* distributions with cusps at each data point [8,15].

To answer the second question, we computed where $D_{\text{KL}}(Q_{\text{true}}\|Q^*)$ falls within the distribution of divergences $D_{\text{KL}}(Q\|Q^*)$ observed for $Q \sim p(Q|\text{data})$. This location is naturally quantified by a p value corresponding to the null hypothesis that $Q_{\text{true}} \sim p(Q|\text{data})$. If Q_{true} is typical of plausible Qs, these p values should be uniformly distributed between 0 and 1. Alternatively, p values clustered close to 0 indicate that the posterior ensemble $p(Q|\text{data})$ overestimates how much Q_{true} diverges from Q^*, whereas p values clustered close to 1 indicate that $p(Q|\text{data})$ underestimates this uncertainty. Figure 2(c) shows our results for the two choices of Q_{true} in Fig. 2(a); the results for other choices of Q_{true} are shown in Fig. S3 [13]. In general, the p values for DEFT (with $\alpha = 2$, 3, and 4) were distributed with remarkable uniformity. DEFT with $\alpha = 1$ tended to overestimate uncertainties, whereas KDE and DPMM tended to underestimate uncertainties.

Finally, we illustrate the capabilities of DEFT using data reported in the initial observation of the Higgs boson [11] (see Fig. S4 [13] for an analysis of data from the WHO). Figure 3(a), which is a reconstruction of Fig. 4 of

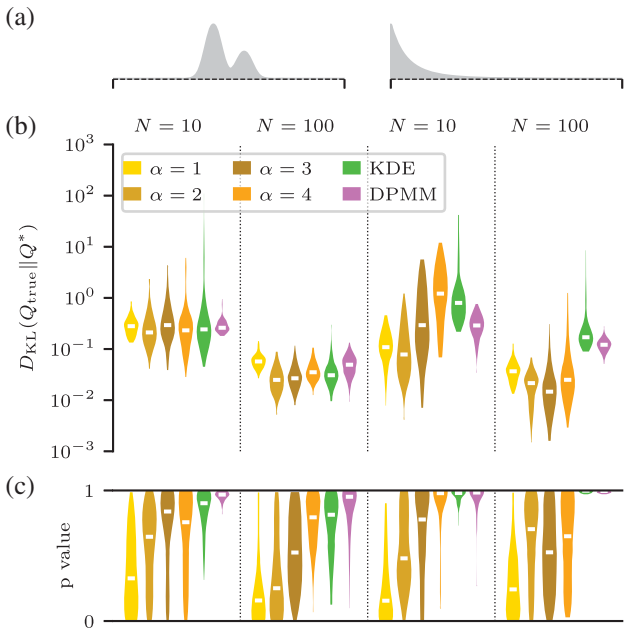

FIG. 2. Performance of DEFT. (a) DEFT, KDE, and DPMM were used to analyze data from two different Q_{true} distributions: the Gaussian mixture from Fig. 1(a) (left) and a Pareto distribution, $Q_{\text{true}}(x) = 3x^{-4}$, confined to the x interval (1,4) (right). (b) One hundred data sets of size $N = 10$ and one hundred data sets of size $N = 100$ were generated for each Q_{true}. For each data set, Q^* was computed by DEFT (using $G = 100$ and $\alpha = 1$, 2, 3, or 4), by KDE, or by DPMM. Violin plots (with median indicated) show the resulting Kullback-Leibler divergences $D_{\text{KL}}(Q_{\text{true}}\|Q^*)$. (c) p values quantifying, for each simulated data set, the location of $D_{\text{KL}}(Q_{\text{true}}\|Q^*)$ within the distribution of $D_{\text{KL}}(Q\|Q^*)$ values observed for $Q \sim p(Q|\text{data})$.

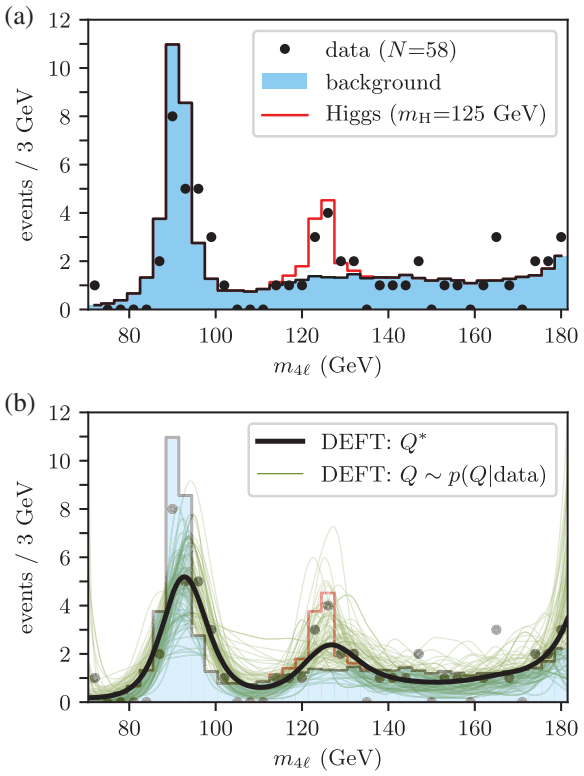

FIG. 3. DEFT applied to Higgs boson data. (a) A reconstruction of Fig. 4 from Ref. [11]. Dots (black) indicate the invariant masses of four-lepton decay events histogrammed across $G = 37$ bins of width 3 GeV each. Also shown are the number of events expected, based on Standard Model simulations, from either background decay processes (blue) or from the decay of a Higgs boson with mass of 125 GeV (red). (b) The optimal density estimate Q^* (black), along with 100 posterior samples $Q \sim p(Q|\text{data})$ (olive) computed by DEFT using the histogram data in panel (a).

Ref. [11], shows a histogram of the invariant masses of $N = 58$ four-lepton events observed by the CMS Collaboration at the Large Hadron Collider. Such events are generated by decays of the Higgs boson via $H \to ZZ \to 4\ell$, but they also arise from a variety of background decay processes. One of the challenges faced by the CMS Collaboration was determining whether these data exhibit a localized excess of events representing a possible Higgs resonance. Figure 3(b) shows DEFT applied to these data using default parameters. Despite Higgs decays representing only ∼10% of the observed events, DEFT detects a prominent local maximum near the Higgs resonance at $m_H = 125$ GeV. The confidence in this maximum can be quantified by sampling $Q \sim p(Q|\text{data})$: 81% of sampled Qs have exactly one local maximum between 110 GeV and 140 GeV (7% have no local maxima and 12% have multiple local maxima), and these maxima occurred at 127.1 ± 3.7 GeV.

Here we have shown that DEFT can effectively address density estimation needs on small data sets in one dimension. DEFT provides point estimates comparable to KDE and DPMM, but it does not suffer from the multiple drawbacks of these other methods. In particular, the only key parameter that the user must specify is a small positive integer α that defines the qualitative meaning of smoothness and which governs how DEFT relates to maximum entropy estimation (see Ref. [9]). In our experience, however, using $\alpha = 3$ seems to work well nearly all of the time. Other parameters, such as the number of grid points G, reflect computational practicalities. These parameters can be chosen automatically and have little effect on the results as long as reasonable values are used.

DEFT thus addresses a major outstanding need, not just in statistical learning theory but also in day-to-day data analysis. To this end, we have developed an open source PYTHON package called SUFTWARE. SUFTWARE allows users to apply DEFT in one dimension to their own data, and in the future it will include additional field-theory-based statistical methods. This implementation is sufficiently fast for routine use: The computations for Fig. 1 take about 0.25 seconds on a standard laptop computer (see SM.10 [13] for a discussion on computational complexity). SUFTWARE has minimal dependencies, is compatible with both PYTHON 2 and PYTHON 3, and is readily installed using the PIP package manager. SUFTWARE homepage [16] for installation and usage instructions.

We thank Kush Coshic for preliminary contributions to this project, as well as Serena Bradde, David McCandlish, and two anonymous referees for helpful feedback. This work was supported by a CSHL/Northwell Health Alliance grant to J. B. K. and by NIH Cancer Center Support Grant No. 5P30CA045508.

*Corresponding author.
jkinney@cshl.edu

[1] B. W. Silverman, *Density Estimation for Statistics and Data Analysis* (Chapman and Hall, London, 1986).
[2] W. Bialek, C. G. Callan, and S. P. Strong, Phys. Rev. Lett. **77**, 4693 (1996).
[3] J. C. Lemm, *Bayesian Field Theory* (Johns Hopkins, Baltimore, 2003).
[4] T. A. Enßlin, M. Frommert, and F. S. Kitaura, Phys. Rev. D **80**, 105005 (2009).
[5] T. Enßlin, arXiv:1301.2556v1.
[6] P. Müller, F. A. Quintana, A. Jara, and T. Hanson, *Bayesian Nonparametric Data Analysis* (Springer, New York, 2015).
[7] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, and A. Vehtari, *Bayesian Data Analysis*, 3rd ed. (CRC Press, New York, 2013), Vol. 109.
[8] J. B. Kinney, Phys. Rev. E **90**, 011301(R) (2014).
[9] J. B. Kinney, Phys. Rev. E **92**, 032107 (2015).
[10] J. Riihimäki and A. Vehtari, Bayesian Anal. **9**, 425 (2014).
[11] CMS Collaboration, Phys. Lett. B **716**, 30 (2012).
[12] World Health Organization Collaboration, *World Health Statistics 2017: Monitoring Health for the SDGs, Sustainable Development Goals* (World Health Organization, Geneva, 2017).
[13] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.121.160605 for a derivation of the DEFT algorithm as well as for other information.
[14] J. Zinn-Justin, *Path Integrals in Quantum Mechanics* (Oxford University, New York, 2010).
[15] I. Nemenman and W. Bialek, Phys. Rev. E **65**, 026137 (2002).
[16] SUFTWARE homepage http://suftware.readthedocs.io