Rapid and deterministic estimation of probability densities using scale-free field theories

Justin B. Kinney
Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 11724, USA
(Dated: 18 April 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 new 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.

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 \(Q_{\text{true}}\) 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 paper discusses one such strategy, the main theoretical aspects of which were worked out by Bialek et al. in 1996 [2]. One first assumes a specific smoothness length scale \(\ell\). 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 \(Q_{\text{ell}}\), which maximizes \(p(Q|\ell, \text{data})\) and serves as an estimate of \(Q_{\text{true}}\), is then computed as the solution to a nonlinear differential equation. This approach has been implemented and further elaborated by others [3][11]: a connection to previous literature on “maximum penalized likelihood” [1] should also be noted.

Left lingering is the question of how to choose the length scale \(\ell\). Bialek et al. argued, however, that the data themselves will typically select a natural length scale due to the balancing effects of goodness-of-fit (i.e. the posterior probability of \(Q_{\ell}\)) and an Occam factor (reflecting the entropy of model space [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].

However computing \(p(\ell|\text{data})\) requires first computing \(Q_{\ell}\) at every relevant length scale, i.e. solving an infinite compendium of nonlinear differential equations. Nemenman and Bialek [3] approached this problem by computing \(Q_{\ell}\) at a finite, pre-selected set of length scales. Although this strategy yielded important results, it also has significant limitations. First, it is unclear how to choose the set of length scales needed to estimate \(Q_{\text{true}}\) to a specified accuracy. Second, as was noted [3], this strategy is very computationally demanding. Indeed, no implementation of this approach has since been developed for general use, and performance comparisons to more standard density estimation methods have yet to be reported.

Here I describe a rapid and deterministic homotopy method for computing \(Q_{\ell}\) 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, available at github.com/jbkinney/13_deft provides a Python implementation of this algorithm for 1D and 2D problems. Simulation tests show favorable performance relative to standard Gaussian mixture model (GMM) and kernel density estimation (KDE) approaches [4].

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

\[
p(Q) = \int_0^\infty d\ell \ p(Q|\ell) \ p(\ell).
\]  (1)

Adopting the Jeffrey’s 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,

\[
p(Q|\text{data}) = \int_0^\infty d\ell \ p(Q|\ell, \text{data}) \ p(\ell|\text{data}).
\]  (2)

As in [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 $\phi$ as

$$Q(x) = \frac{e^{-\phi(x)}}{\int d^Dx' e^{-\phi(x')}}.$$  

(3)

Each $Q$ corresponds to a different $\phi$, but there is a one-to-one correspondence with fields $\phi_{nc}$ that have a constant Fourier component. Using this fact, we adopt the standard path integral measure $D\phi_{nc}$ as the measure on $Q$-space, and define the prior $p(\phi_{nc}|\ell)$ in terms of a field theory on $\phi_{nc}$. In this paper we consider the specific class of priors discussed by [2], i.e.

$$p(\phi_{nc}|\ell) = \frac{1}{Z_{\ell}^N} \exp \left[ - \int d^Dx \frac{\ell^{2\alpha-D}}{2} \phi_{nc} \Delta \phi_{nc} \right].$$  

(4)

Here we take $\alpha$ to be a positive integer and define the differential operator $\Delta = (-\nabla^2)^\alpha$. $Z_{\ell}^N$ is the corresponding normalization factor. This prior effectively constrains the $\alpha$-order derivatives of $\phi_{nc}$, strongly dampening Fourier modes that have wavelength much less than $\ell$.

Applying Bayes’s rule to this prior yields the following exact expression for the posterior [12]:

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

(5)

where

$$S_{\ell}^N[\phi] = \int d^Dx \left[ \frac{\phi \Delta \phi}{2} + e^t R \phi + e^t e^{-\phi} \right]$$

(6)

is the “action” described by [2] and explored in later work [3, 8–10]. Here, $R(x) = N^{-1} \sum_{n=1}^N \delta(x - x_n)$ is the raw data density, $\phi(x) = \phi_{nc}(x) + \phi_c$, $t = \ln(N/\ell^{2\alpha-D})$, $\beta = \ell^{2\alpha-D}$, and $Z_{\ell}^N = Z_0^N \Gamma(N/(V N)^{ \alpha})$.

In this paper we consider the specific class of priors discussed by [2], i.e. the field that minimizes $S_{\ell}^N$. Setting $\delta S_{\ell}^N/d\phi = 0$ gives the nonlinear differential equation,

$$\Delta \phi + e^t \left[ R - Q_{\ell} \right] = 0,$$  

(7)

where $Q_{\ell}(x) = e^{-\phi_{nc}(x)}/V$ is the probability density corresponding to $\phi_{\ell}$.

The central finding of this paper is that, instead of computationally solving Eq. 7 at select length scales $\ell$, we can compute $\phi_{\ell}$ at all length scales of interest using a convex homotopy method [13]. First we differentiate Eq. 7 with respect to $t$, yielding

$$\exp(-\Delta + Q_{\ell}) \frac{d\phi_{\ell}}{dt} = Q_{\ell} - R.$$  

(8)

If we know $\phi_{\ell}$ at any specific length scale $\ell_i$, we can determine $\phi_{\ell}$ at any other length scale $\ell_j$ - and at all length scales in between - by integrating Eq. 8 from $\ell_i$ to $\ell_j$. Because $S_{\ell}^N[\phi]$ is a strictly convex function of $\phi$, each $\phi_{\ell}$ so identified will uniquely minimize this action. Moreover, because Eq. 8 is exact, each corresponding $Q_{\ell}$ will fit the data optimally even when $N$ is small. And since the matrix representation of $e^{-\Delta + Q_{\ell}}$ 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 $\ell_i$ from which to initiate integration of Eq. 8, we look to the large length scale limit, where a weak-field approximation can be used to compute $\phi_{\ell_i}$. Linearizing Eq. 7 and solving for $\phi_{\ell_i}$ gives, for $|k| > 0$, $\hat{\phi}_{\ell_i}(k) = -\frac{V R(k)}{\exp(\tau_{\ell_i}^k)}$, where hats denote Fourier transforms, $k \in \mathbb{Z}^D$ indexes the Fourier modes of the volume $V$, and each $\tau_k = \ln(\frac{2\pi |k|}{2\alpha L^{D-2\alpha}})$ is a log eigenvalue of $V\Delta$. To guarantee that none of the Fourier modes of
The length scale prior $\ell_i$ is identified. The MAP probability, $p(\phi_t|\ell, \text{data})$, is represented by the exponential term. This is the “goodness-of-fit”, which steadily increases as $\ell$ gets smaller. This is multiplied by an Occam factor (the inverse rooted quantity), which steadily decreases as $\ell$ get smaller due to the increasing entropy of model space. As discussed by \cite{2,3}, this tradeoff causes $p(\ell|\text{data})$ to peak at a nontrivial, data-determined length scale $\ell^*$. The length scale prior $p(\ell)$ must decay faster than $\ell^{-1}$ in the infrared in order for $p(\ell|\text{data})$ to be normalizable. The need for such regularization reflects redundancy among the priors $p(\phi_\ell|\ell)$ for $\ell > \ell_i$ that results from the volume $V$ supporting only 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 $\ell < \ell_f$.

The posterior density $p(Q|\text{data})$ can be sampled by first choosing $\ell \sim p(\ell|\text{data})$, then selecting $\phi \sim p(\phi|\ell, \text{data})$. This latter step simplifies when $p(\ell|\text{data})$ is strongly peaked about a specific $\ell^*$ because, in this case, the eigenvalues and eigenfunctions of $\Delta + e^\ell \mathcal{Q}_i$ do not depend strongly on which $\ell$ is selected and therefore need to be computed only once. $p(\phi|\ell, \text{data})$ can then be sampled by choosing

$$\phi(x) = \phi_t(x) + \sum_{j=1}^{G^D} \frac{\eta_j}{\beta \lambda_j} \psi_j(x), \quad (10)$$

where each $\psi_j$ is an eigenfunction of the operator $\Delta + e^\ell \mathcal{Q}_i$ satisfying $\int d^Dx \psi_j(x)^2 = 1$, $\lambda_j$ is the corresponding eigenvalue, and each $\eta_j$ is a normally distributed random variable.

Fig. 1 illustrates 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. 1a). Next, length scales $\ell_i$ and $\ell_f$ are chosen. Eq. (8) is then integrated to yield $\phi_t$ at a set of length scales between $\ell_i$ and $\ell_f$ chosen automatically by the ODE solver to achieve the desired accuracy. Eq. (9) is then used to compute $p(\ell|\text{data})$ at each of these length scales, after which $\ell^*$ 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 pre-specified 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...
As in 1D, larger values of $\alpha$ give a smoother $Q^*$. 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 $\Delta$ matrix. For instance, the computation shown in Fig. 1 took about 0.3 sec, while the DEFT computations shown in Fig. 4 took about 1-3 sec each [15].

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 by Holy [4] allows $Q_\ell$ 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 increasing in $\alpha$, i.e. in one’s notion of “smoothness.” This likely indicates a fundamental problem with using $\Delta = (-\nabla^2)^{\alpha}$ to define high dimensional priors. Using a different operator for $\Delta$, 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. Support for this work was provided by the Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory.

* Please direct correspondence to jkinney@cshl.edu

[1] P. P. B. Eggermont and V. N. LaRiccia, Maximum Penalized Likelihood Estimation: Volume 1: Density Estimation (Springer, 2001).
[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, Physical review letters 83, 3554 (1999).
[7] D. M. Schmidt, Phys. Rev. E 61, 1052 (2000).
[8] J. C. Lemm, Bayesian Field Theory (Johns Hopkins, 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] The identity $a^{-N} = \frac{1}{N!} \int_{-\infty}^{\infty} du \exp \left[ -Nu - ae^{-u} \right]$ is used here with $a = (N/V) \int d^D x \ e^{-\phi_{\text{true}}(x)}$ and $u = \phi_c$.
[13] E. L. Allgower and K. Georg, Numerical Continuation Methods: An Introduction (Springer, 1990).
[14] J. Skilling, AIP Conf. Proc. 954, 39 (2007).
[15] 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).