Computational Statistics and Data Analysis 101 (2016) 148–160

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A fast and objective multidimensional kernel density estimation method: fastKDE

Travis A. O'Brien [a,b,*], Karthik Kashinath [a], Nicholas R. Cavanaugh [a], William D. Collins [a,c], John P. O'Brien [d,a]

[a] Lawrence Berkeley National Lab, Berkeley, CA, USA
[b] University of Davis, Davis, CA, USA
[c] University of California, Berkeley, CA, USA
[d] University of California, Santa Cruz, CA, USA

HIGHLIGHTS

- A multidimensional, fast, and robust kernel density estimation is proposed: fastKDE.
- fastKDE has statistical performance comparable to state-of-the-science kernel density estimate packages in R.
- fastKDE is demonstrably orders of magnitude faster than comparable, state-of-the-science density estimate packages in R.
- A Python-based implementation of fastKDE is available at https://bitbucket.org/lbl-cascade/fastkde.

ARTICLE INFO

Article history:
Received 22 July 2015
Received in revised form 11 January 2016
Accepted 28 February 2016
Available online 7 March 2016

Keywords:
Empirical characteristic function
ECF
Kernel density estimation
Histogram
Nonuniform FFT
NuFFT
Multidimensional
KDE

ABSTRACT

Numerous facets of scientific research implicitly or explicitly call for the estimation of probability densities. Histograms and kernel density estimates (KDEs) are two commonly used techniques for estimating such information, with the KDE generally providing a higher fidelity representation of the probability density function (PDF). Both methods require specification of either a bin width or a kernel bandwidth. While techniques exist for choosing the kernel bandwidth optimally and objectively, they are computationally intensive, since they require repeated calculation of the KDE. A solution for objectively and optimally choosing both the kernel shape and width has recently been developed by Bernacchia and Pigolotti (2011). While this solution theoretically applies to multidimensional KDEs, it has not been clear how to practically do so.

A method for practically extending the Bernacchia–Pigolotti KDE to multidimensions is introduced. This multidimensional extension is combined with a recently-developed computational improvement to their method that makes it computationally efficient: a 2D KDE on 10^5 samples only takes 1 s on a modern workstation. This fast and objective KDE method, called the fastKDE method, retains the excellent statistical convergence properties that have been demonstrated for univariate samples. The fastKDE method exhibits statistical accuracy that is comparable to state-of-the-science KDE methods publicly available in R, and it produces kernel density estimates several orders of magnitude faster. The fastKDE method does an excellent job of encoding covariance information for bivariate samples. This property allows for direct calculation of conditional PDFs with fastKDE. It is demonstrated how this capability might be leveraged for detecting non-trivial relationships between quantities in physical systems, such as transitional behavior.

© 2016 The Authors and Lawrence Berkeley National Laboratory. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

* Correspondence to: 1 Cyclotron Rd, MS74R-316C, Berkeley, CA, 94720, USA. Tel.: +1 510 495 8047.
E-mail address: TAOBrien@lbl.gov (T.A. O'Brien).

http://dx.doi.org/10.1016/j.csda.2016.02.014
0167-9473/© 2016 The Authors and Lawrence Berkeley National Laboratory. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Numerous facets of scientific research implicitly or explicitly call for the estimation of probability densities. For example, the ubiquitous histogram is one of the most basic and commonly used tools for displaying information about relative occurrences of one or more quantities. Univariate histograms are by far the most common, and they are often used to ascertain or display probabilistic information that is otherwise difficult to encapsulate using simple measures of location or spread: e.g., Schumacher and Houze (2003) and Bony et al. (2004). Multivariate probability density functions (PDFs), which encode information about the joint occurrence of two or more related variables, are less commonly analyzed. However, multivariate PDF analyses expose or encode information that can provide unique insights and open new research pathways: e.g., Larson et al. (2002), Marchand et al. (2010), Lee et al. (2014) and AghaKouchak et al. (2014).

As with univariate histograms, multivariate histograms are the workhorse for multivariate PDF analysis. More sophisticated methods than the histogram for estimating PDFs, including both parametric and non-parametric methods, are used less frequently in contemporary scientific research. However, these more sophisticated methods can provide unique and valuable benefits above what histograms can provide. Parametric methods, such as maximum likelihood parameter estimation, can yield compactly-encoded and high-fidelity quantitative information about the relative distribution of quantities. For example, if one can assume that two quantities are jointly and normally distributed, then a maximum likelihood estimation of the normal distribution parameters will yield 5 parameters corresponding to the means and covariances that uniquely and completely describe the relationship between the two quantities. Non-parametric methods, such as kernel density estimation (KDE), can yield an empirical estimate of the true PDF, without assuming any form for the distribution.

These existing PDF analytic methods also have some drawbacks that can be prohibitive in certain situations. Parametric analysis requires an assumption about the form of the underlying distribution. If the distribution can be isolated to one of a finite number of forms, then a model selection method (e.g., Wit et al., 2012) can be used to choose the best PDF model. However, if such an assumption cannot be made, for example because the PDF may result from a mixture of some large number of unknown PDFs, then parametric methods may be unsuitable. Non-parametric methods (histograms and KDEs) are less restrictive, but both methods require selection of either a bin width or a kernel and kernel bandwidth. This choice is not always straightforward, and it may require some user-intervention for common use cases (Silverman, 1986). The difficulty of making this choice compounds with the addition of variables; the kernel bandwidth becomes a kernel bandwidth matrix for multivariate KDE, and there is a similarly complex choice to be made about bin boundaries for histograms. There are a variety of automatic bandwidth selection methods available, (e.g., smoothed cross validation Duong and Hazelton, 2005), but these methods have a number of drawbacks, such as poor performance for large sample sizes and/or complex distributions and high computational expense (Heidenreich et al., 2013). In a review of automatic selection methods, Heidenreich et al. (2013) recommend a variety of different methods, depending on dataset characteristics (including sample size, distribution smoothness, and skewness). These drawbacks appear to have inhibited their widespread adoption as a fundamental data analysis tool.

To get around this difficulty of a potentially subjective choice of kernel shape and kernel bandwidth, Bernacchia and Pigolotti (2011) derive a method for objectively determining both the kernel shape and the kernel bandwidth (Luedicke and Bernacchia, 2014 implemented it in Stata). Effectively they define a Fourier-based (and typically low-pass) filter on the empirical characteristic function (ECF) of a given dataset that yields an empirical kernel that is optimal in the sense that the integrated, squared difference between the resulting KDE and the true PDF is minimized. They coin this the 'self-consistent' density estimate. O'Brien et al. (2014) show that a non-uniform fast Fourier transform (nuFFT) can be used to calculate the ECF in a computationally efficient and accurate way; they use this new method to automate the estimation of thousands of univariate PDFs from the output of a climate model. Theoretically, the Bernacchia and Pigolotti (2011) optimal KDE can be applied to datasets of arbitrary dimensionality, and O'Brien et al. (2014) note that the nuFFT can also be applied to efficiently calculate multidimensional ECFs. However, the Fourier filters implemented in both of these studies are inherently one dimensional (Bernacchia and Pigolotti, 2011; O'Brien et al., 2014).

In this manuscript we augment the Bernacchia and Pigolotti (2011) KDE method such that it can be applied to datasets of arbitrary dimensionality. We show that when combined with the nuFFT-based ECF calculation of O'Brien et al. (2014), the resulting multivariate PDF estimation method is optimal, fast, and unencumbered by the need for user-selected parameters (Section 2). We further demonstrate that this method can be used to directly, rapidly, and accurately infer conditional probability distributions (Section 3). We apply this method to examine the distribution of precipitation and temperature in California, USA conditioned on global mean temperature (Section 4), which shows a number of intriguing features that prompt further investigation.

2. The self-consistent KDE in arbitrary dimensions

In order to describe the way in which we have augmented the KDE method of Bernacchia and Pigolotti (2011), it is necessary to review some relevant details of the method. Bernacchia and Pigolotti (2011) originally presented their derivation for a univariate KDE, but a trivial substitution of vector notation into their derivation shows that it applies also to multivariate KDE. Consider a generic, multivariate kernel density estimate $P_{KDE}(\vec{x})$ on D-dimensional multivariate data $\vec{x}_j$ (for $j = 1 \ldots N$ observations of variables $\vec{x} = (x^1, \ldots, x^D)$) using the arbitrarily-shaped kernel $K(\vec{x})$, which is equivalent to

the convolution of the kernel function and the set of delta functions centered on the data:

$$P_{KDE}(\vec{x}) \equiv \frac{1}{N}\sum_{j=1}^{N} K(\vec{x}-\vec{x}_j)$$
$$= \frac{1}{N}\sum_{j=1}^{N}\int_{\mathscr{R}^D} K(\vec{s})\cdot\delta(\vec{x}-\vec{x}_j-\vec{s})\,d\vec{s}, \tag{1}$$

where $\delta(\vec{x})$ is the Dirac delta function (which can be viewed as the limit of a normal distribution as all elements of the covariance matrix approach 0).

The KDE can be represented equivalently by its inverse Fourier transform pair ϕ_{KDE}:

$$\phi_{KDE}(\vec{t}) \equiv \mathscr{F}^{-1}\left(P_{KDE}(\vec{x})\right)$$
$$= \kappa(\vec{t})\cdot\mathscr{C}(\vec{t}),$$

where $\mathscr{F}^{-1}$ represents the multidimensional inverse Fourier transform from data coordinates $\vec{x}$ to frequency-space coordinates $\vec{t}$, $\kappa \equiv \mathscr{F}^{-1}(K)$ is the inverse Fourier transform of the kernel, and $\mathscr{C}$ is the ECF of the data defined as:

$$\mathscr{C}(\vec{t}) \equiv \frac{1}{N}\sum_{j=1}^{N} e^{i\vec{x}_j\cdot\vec{t}}. \tag{2}$$

Bernacchia and Pigolotti (2011) derive a transform kernel $\hat{\kappa}$ that statistically minimizes the mean squared difference between the true PDF P and the resulting optimal KDE $\hat{P}_{KDE}$. This optimal transform kernel is defined as:

$$\hat{\kappa}(\vec{t}) \equiv \frac{N}{2(N-1)}\left[1+\sqrt{1-\frac{4(N-1)}{N^2|\mathscr{C}(\vec{t})|^2}}I_{\vec{A}}(\vec{t})\right], \tag{3}$$

where $I_{\vec{A}}(\vec{t})$ represents a frequency filter that is 1 for the set of frequencies $\vec{A}$ used in the KDE and 0 otherwise. We use the nomenclature of Bernacchia and Pigolotti (2011), where the set $\vec{A}$ is referred to as the set of 'accepted' frequencies.

In extending the Bernacchia and Pigolotti (2011) KDE method to arbitrary dimensions, there are a couple of considerations. The ECF can be calculated in a computationally efficient manner using a nuFFT as shown by O'Brien et al. (2014). The nuFFT method naturally extends to arbitrary dimensions (Greengard and Lee, 2004), so a multidimensional ECF can efficiently be calculated using nuFFT methods. However, the selection of the filter $I_{\vec{A}}$ requires special consideration for multidimensional KDE.

The choice of the filter $I_{\vec{A}}$ is critical for the success of this KDE method. Primarily, the filter must be specified such that

$$|\mathscr{C}(\vec{t})|^2 \geq \mathscr{C}_{\min}^2 = 4(N-1)N^{-2} \quad \text{for } \vec{t}\in\vec{A}. \tag{4}$$

This primary filter threshold $\mathscr{C}_{\min}$ is necessary for stability of the estimation method (i.e., positivity of the argument of the square root in Eq. (3) for $\hat{\kappa}$). Secondarily, the set $\vec{A}$ may exclude an additional subset of otherwise acceptable frequencies. The choice of this secondary set of excluded frequencies is somewhat arbitrary, though it must be bounded and it must grow with increasing N for $\hat{P}_{KDE}$ to converge to the true distribution as N increases (Bernacchia and Pigolotti, 2011). Bernacchia and Pigolotti (2011) define the original univariate filter by setting a cutoff frequency t^* such that half of the ECF values within $[-t^*, t^*]$ are above $|\mathscr{C}_{\min}|$. O'Brien et al. (2014) choose a cutoff frequency that is defined by the first occurrence of three consecutive frequencies below $|\mathscr{C}_{\min}|$; ECF values at frequencies above this cutoff are set to 0.

Neither of the filters described by Bernacchia and Pigolotti (2011) or O'Brien et al. (2014) extend simply to multiple dimensions. The natural multidimensional extension of the Bernacchia and Pigolotti (2011) filter would correspond to choosing a cutoff threshold frequency hypersurface $\mathscr{T}$ (which corresponds to a curve in $\mathscr{R}^2$ and a surface in $\mathscr{R}^3$) such that approximately half of the ECF values contained within are above $|\mathscr{C}_{\min}|$. It is not immediately clear how to objectively and automatically choose such a surface for arbitrary dimensionality. The natural multidimensional extension of the O'Brien et al. (2014) filter would correspond to finding a contiguous and closed hypervolume of below-threshold values ($|\mathscr{C}|^2 < |\mathscr{C}_{\min}|^2$) and setting values of the ECF outside that hypervolume to 0. For $\mathscr{R}^2$ this would correspond to a hypothetical 2D *moat* of below-threshold values surrounding the origin in frequency space. If such a moat exists, finding its contiguous hypervolume objectively and automatically is relatively straightforward using a standard flood-fill search algorithm. A contiguous set of below-threshold frequencies is guaranteed to separate high and low frequencies in $\mathscr{R}^1$; this property allowed its simple use in determining a cutoff threshold in the filter of O'Brien et al. (2014). Unfortunately however, this property is not guaranteed to hold in higher dimensions. It is entirely possible that, for example in $\mathscr{R}^2$, a contiguous hypervolume (area) of below-threshold values surrounding the origin does not exist. For a filter to be generically useful for multidimensional KDE, it should be guaranteed to work in all sensible cases.

Instead of defining the set of accepted frequencies $\vec{A}$ such that above-threshold values ($|\mathscr{C}|^2 \geq |\mathscr{C}_{\min}|^2$) within some hypersurface are accepted and those without the hypersurface are rejected, we have defined a filter based on selecting a

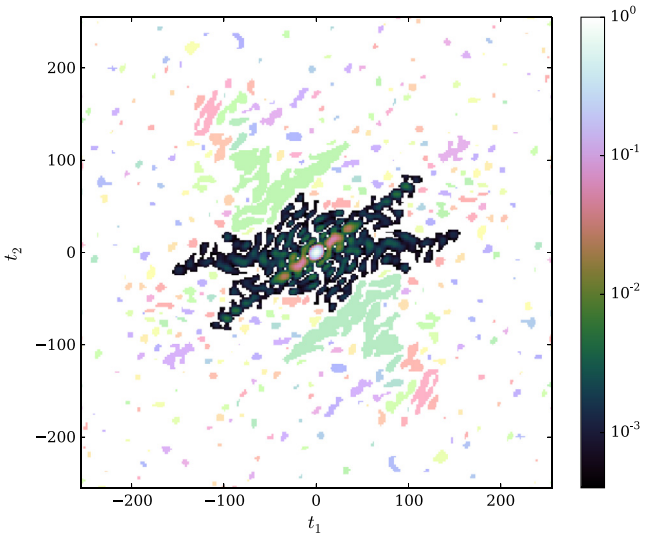

Fig. 1. The square of the empirical distribution function $|\mathscr{C}|^2$ from 10,000 samples of the 'mixture distribution' discussed in Section 3.2 (and shown in Fig. 4), for $|\mathscr{C}|^2 \geq |\mathscr{C}_{\min}|^2$. There are two color schemes present in this figure. The predominantly dark, multicolored colored 'X-shaped' region in the center corresponds to values of $|\mathscr{C}|^2$ for the first contiguous hypervolume (the area containing $\vec{t} = \vec{0}$); the colorbar at right applies to colors in this region. The lightly-colored, monotone areas away from the first contiguous hypervolume correspond to additional contiguous hypervolumes (areas) with $|\mathscr{C}|^2 \geq |\mathscr{C}_{\min}|^2$. The colors of these areas are arbitrary and only serve to visually differentiate nearby contiguous areas from one another. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

subset of contiguous hypervolumes of above-threshold values. We assume that the multidimensional ECF consists of a finite set of contiguous hypervolumes of above-threshold values. At least one such contiguous hypervolume containing the $\vec{t} = \vec{0}$ frequency is guaranteed to exist, since $\mathscr{C}(\vec{0}) = 1$ due to normalization, and $|\mathscr{C}_{\min}|^2 < 1$. In our Python/Cython-based implementation of this KDE method, we sort these contiguous hypervolumes based on the distance from their centers-of-mass to the origin. The user of the method has the option to specify how many of these hypervolumes (as an integer value or a fraction of the total number), in order of their distance to the origin, are retained in the set $\vec{A}$. By default (and throughout this manuscript), only the single hypervolume centered at $\vec{t} = \vec{0}$ is used. Fig. 1 depicts this filter on one of the distributions discussed in Section 3.2. The colored contours in the center of Fig. 1 show the single hypervolume that is included in the construction of the KDE; the additional contiguous hypervolumes are 'discarded' (set to 0).

This filter satisfies the convergence conditions described by Bernacchia and Pigolotti (2011). The set of frequencies included in this *lowest contiguous hypervolume filter* are bounded since they will always be contained within a finite-sized hypervolume around the origin. Further, for a given number of contiguous hypervolumes included in the set $\vec{A}$, this bound grows as the number of data points increases, since $|\mathscr{C}_{\min}|^2 \propto N^{-1}$ (meaning the volumes of the contiguous hypervolumes grow with increasing N). With these two conditions met, this filter should allow the KDE to converge toward the true PDF as the number of data points increases.

3. Convergence properties in multiple dimensions

3.1. Convergence on multidimensional normal distributions

To demonstrate the efficacy of the lowest hypervolume filter, we examine the convergence properties of our implementation of the Bernacchia and Pigolotti (2011) KDE method with the nuFFT-based speed improvement of O'Brien et al. (2014). For simplicity, hereon we refer to this implementation as the *fastKDE* method. Fig. 2(a) shows the integrated, squared error (ISE) of the fastKDE as a function of the number of data samples denoted by N. We approximate this integral using the midpoint rule:

$$\text{ISE} = \sum_{i_1=1}^{i_1=M_1} \cdots \sum_{i_D=1}^{i_D=M_D} \left(\hat{P}_{KDE}(\vec{x}_i) - P(\vec{x}_i)\right)^2 \prod_{j=1}^{D} \Delta x_j, \tag{5}$$

where P is the true PDF; M_1 and M_D are the number of points in the 1st and Dth dimensions (and there is a corresponding summation symbol for each dimension); $\vec{x}_i = (x^1_{i_1}, \ldots, x^D_{i_D})$ is a regular multidimensional lattice of data samples; and Δx_j is the spacing of the grid in the jth dimension (which is constant to due the use of regular grids).

For direct comparison with univariate results from Bernacchia and Pigolotti (2011) and O'Brien et al. (2014), we test the convergence of a univariate standard normal distribution. Fig. 2(a) shows that the ISE decays close to, but slightly slower

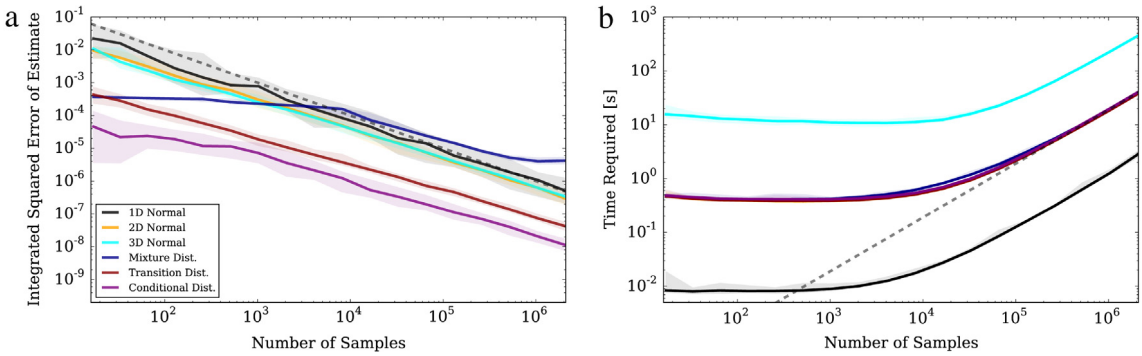

Fig. 2. (a) The integrated, squared error of KDE estimates, as a function of the number of data samples, of samples from several distributions: 1, 2, and 3 dimensional normal distributions, a non-trivial mixture of normal distributions, and a distribution representing transitional behavior (the mixture and transition PDFs are introduced and discussed in Section 3.2). The solid lines depict the median of 30 simulations, and the shaded swaths depict the 5–95 percentile range. The gray dashed line shows the theoretical maximum convergence rate N^{-1} for reference. Note that the ISE technically has different units when calculated from KDE estimates of different dimensionality, so the absolute values of these curves cannot be compared, though it is valid to compare the curve slopes on a log–log plot. (b) As in (a), but for the time required to perform the KDE. The gray dashed line shows N^1 for reference.

than, N^{-1}. This is consistent with theoretical arguments about the convergence rate of the method, and it is consistent with empirical results (Bernacchia and Pigolotti, 2011; O'Brien et al., 2014). We perform the same tests for bivariate and trivariate standard normal distributions, and Fig. 2 shows that the method exhibits similarly good performance even for multivariate data. The convergence rate appears to systematically decrease as dimensionality increases, which is presumably due to the curse of dimensionality (Scott, 2008). However, the decrease is relatively minor, with the exponent of N in the convergence rates for the 1, 2, and 3 dimensional KDEs being -0.92 ± 0.03, -0.91 ± 0.01, and -0.89 ± 0.01 respectively.

Fig. 2(b) shows that the time required to compute the KDE asymptotically scales as N^1 for all three dimensionalities. This result is consistent with the complexity analysis of O'Brien et al. (2014), which suggests that the speed of the nuFFT-based KDE should scale as $\mathcal{O}(N \cdot q^D + M^D \cdot \log(M^D))$, where q is the size of the convolution kernel used in the nuFFT, N is the number of samples, and M^D is the total number of grid points on which the KDE is estimated. For the convergence tests shown in Fig. 2, we use $q = 28$ (which yields approximately double precision accuracy for the nuFFT approximation) and $M = 257$ for each dimension. Accordingly, for $N \cdot q^D \gg M^D \cdot \log(M^D)$, the number of operations required for the estimate (and therefore the time T required) should scale as $T \propto q^D N^1$.

The timing curves in Fig. 2(b) exhibit two intriguing features indicative of additional factors playing important roles in the computational aspects of the method: the curvature of the 3D timing curve, and the relative spacing of the timing curves. The timing curve for the 3D standard normal distribution has a minimum at approximately 10^4 samples; the 3D KDE for a sample size of 16 takes more time than for a sample size of 8192. We hypothesize that this is associated with the lowest contiguous hypervolume filter. The filter relies on an implementation of a classical flood-fill search algorithm, which first searches the ECF for a new above-threshold value and then recursively searches all neighbors for other above-threshold values.[1] Once the recursive search is exhausted, the algorithm repeats until the entire ECF space has been searched. While this flood-fill search requires at least M^D floating point comparisons, there are additional overhead costs associated with the management of the linked-list structure that we use to track the list of unsearched neighbors. If the ECF space consists of a large number of relatively small contiguous hypervolumes, then the search will incur a relatively large overhead, which could make the entire KDE method take more time than would otherwise be expected. The second intriguing feature of Fig. 2 is that the vertical spacing of the timing curves is smaller than predicted by the complexity analysis. The computational time should scale like q^D, meaning that the ratio of timings for the same number of points in the 3D vs 2D cases (or 2D vs 1D) should be $q^D/q^{D-1} = q = 28$. This ratio, as expressed by the vertical spacing of these curves, is roughly the same for the 3D vs 2D and 2D vs 1D cases, in accord with this prediction. But the ratio is approximately 13 instead of 28. This means that the timing of this KDE method scales slower with increasing dimensionality that we would expect based on a simple complexity analysis of the algorithm. It is not immediately clear why the dimensional scaling of this method is better than we anticipated.

3.2. Convergence on complex distributions

Overall, Fig. 2 demonstrates that the fastKDE method exhibits both the excellent convergence properties of the Bernacchia and Pigolotti (2011) method and the excellent computational efficiency of the O'Brien et al. (2014) method for multivariate KDE. By design the normal distribution tests have no covariance structure, and so they do not illustrate whether the optimal kernel used in the fastKDE method can handle input data with non-trivial covariance. In developing this method, we

[1] While the algorithm is recursive, we specifically avoid the use of recursive functions in our implementation, since we have found that such an implementation can result in stack violations for large contiguous volumes.

anticipated that covariance should not be an issue, since the kernel effectively inherits the covariance properties of the underlying data. This can be shown theoretically by considering a 2nd-order Taylor expansion of the optimal kernel (Eq. (3)) about $\vec{t} = \vec{0}$:

$$\hat{\kappa}\left(\vec{t} \approx \vec{0}\right) \approx 1 + \frac{1}{2!}\sum_{k=1}^{D}\sum_{j=1}^{D}\frac{2}{N-2}\left(\frac{\partial^2 \mathscr{C}}{\partial t_j \partial t_k}(\vec{t} = \vec{0})\right) t_j\, t_k \tag{6}$$

$$= 1 - \frac{1}{N-2}\sum_{k=1}^{D}\sum_{j=1}^{D}\sigma_{jk} t_j t_k. \tag{7}$$

Eqs. (6) and (7) are relatively simple compared to Eq. (3) because at $\mathscr{C}(\vec{t} = \vec{0}) = 1$ (normalization), and because derivatives of the ECF at $(\vec{t} = \vec{0})$ are proportional to sample moments: the 1st order derivatives of the ECF are identically 0 at the origin assuming that the sample mean is removed prior to transformation,[2] and the 2nd-order derivatives of the ECF yield the 2nd-order moments (σ_{jk}) of the input samples. In comparison, the 2nd-order Taylor expansion of the ECF is

$$\mathscr{C}\left(\vec{t} \approx \vec{0}\right) = 1 - \frac{1}{2}\sum_{k=1}^{D}\sum_{j=1}^{D}\sigma_{jk} t_j t_k, \tag{8}$$

which differs from Eq. (7) only by a factor of $2(N-2)^{-1}$ in the second term.

This implies that in the vicinity of the origin, the optimal kernel has the same shape and orientation as the ECF, and therefore it has the same shape and orientation in real space. The main difference in the two Taylor approximations is that the optimal kernel has the factor of $2(N-2)^{-1}$ in the second term, which originates from the 2nd-order derivatives of the kernel. This implies that these derivatives, and therefore the variances and covariances of the optimal kernel, asymptotically approach 0 as N approaches infinity; the kernel asymptotically approaches a multidimensional delta function as N increases. For finite N, however, the kernel effectively inherits its main shape and orientation from the underlying data.

We demonstrate that these properties result in convergence even for PDFs with non-trivial covariance structure by examining the convergence properties of the fastKDE method on two additional bivariate PDFs: a PDF with a relatively subtle transition in the relationship between two variables, and a complex mixture of normal distributions. The first distribution, which we label the *transition PDF* (P_T), is derived as follows: given an abscissa variable, the ordinate variable is sampled by first transforming the abscissa values by a sigmoid function and then adding samples from a Normal distribution $\mathscr{N}$. The abscissa variable is sampled from a Gamma distribution Γ, with the abscissa values centered over the transition point of the sigmoid function. We construct a PDF in this manner to simulate a noisy transition in a bivariate physical system:

$$P_T(x, y) \equiv \Gamma\left(x_0 - x + k\theta | k, \theta\right) \cdot \mathscr{N}\left(y | \mu(x), \sigma_y\right), \quad \text{where} \tag{9}$$

$$\mu(x) \equiv \Delta y \cdot \tanh(x - x_0) + y_0, \quad \text{and} \tag{10}$$

$$P_T(y|x) = \mathscr{N}\left(y | \mu(x), \sigma_y\right). \tag{11}$$

Fig. 3(a) shows $P_T(x, y)$ along with a fastKDE estimate based on 10,000 samples from $P_T(x, y)$ using only the lowest contiguous hypervolume in the ECF filter. The contours of the fastKDE estimate and P_T, which use the same contour interval, overlap strongly. This is true even in the region near point at $x = 305$ where the mean of y transitions from a value of 198 to a value of 202, which induces a localized covariance in x and y. To illustrate the degree to which this non-trivial covariance information is correctly encoded by the fastKDE method, we estimate the conditional distribution $P_T(y|x)$ by dividing the bivariate fastKDE estimate $\hat{P}(x, y)$ by the marginal fastKDE estimate $\hat{P}(x)$. Fig. 3(b) shows that the contours of constant $P_T(y|x)$ from the fastKDE conditional estimate follow those of the true conditional distribution; in particular, the contours in the fastKDE estimate exhibit the sigmoid transition at $x = 305$. However, the fastKDE estimate of the conditional is somewhat noisier: especially in data sparse regions, such as above $x = 310$. While it may appear that this transition PDF has hardly any covariance structure, such subtle transitions in real physical systems can result in profound changes in system behavior. Therefore, the ability to reliably discern such transitions from noisy data can be quite important for explaining the behavior of physical systems.

Similar to the convergence of the fastKDE estimates of the standard normal distributions in Section 3.1, Fig. 2(a) shows that the fastKDE estimate of the transition PDF falls as a power law of N. However, in contrast to the standard normal examples, the error improves at a slower rate: -0.803 ± 0.004. Despite the slower error improvement with increasing N, this power law behavior shows that the fastKDE estimate asymptotically approaches the true transition PDF as samples are added. This suggests that as theoretically predicted above, (a) the kernel is shaped and oriented in such a way that it provides a reasonably good PDF estimate for low N, and (b) the kernel approaches a delta function as N approaches infinity, giving data points only a very local influence on the PDF. The time required to compute the fastKDE estimate (Fig. 2(b)) is nearly identical to that required for the bivariate normal distribution: approximately 10 s for 2,000,000 samples.

[2] In practice, we remove the sample mean prior to applying the nuFFT, and we add it back to the resulting real-space PDF estimate.

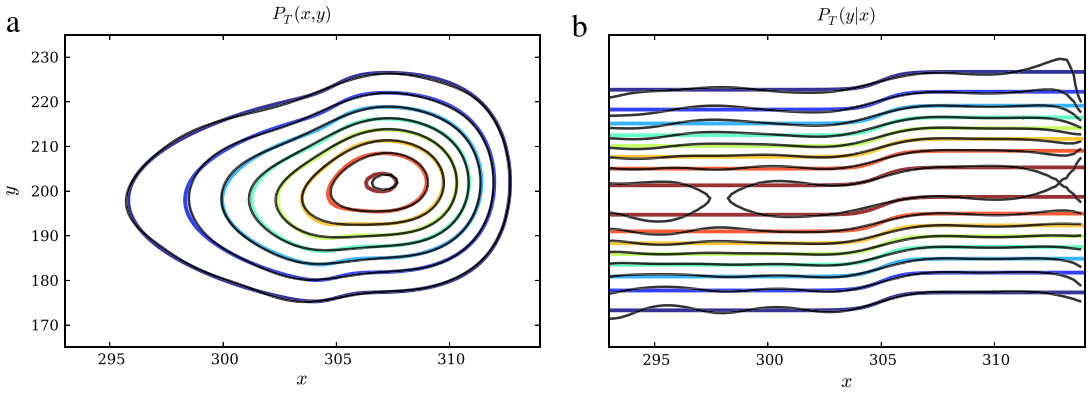

Fig. 3. The true transition PDF described by Eqs. (9)–(11) (color contours) and the fastKDE estimate (black contours) of 10,000 samples from P_T, with $x_0 = 305$, $k = 5$, $\theta = 2$, $y_0 = 200$, $\Delta y = 4$, and $\sigma_y = 12$. (a) The full joint distribution $P_T(x, y)$, and (b) the conditional distribution $P_T(y|x)$. The contour intervals for the true PDF and the fastKDE estimate are the same within (a) and within (b). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Interestingly, the mean ISE of the conditional estimate (calculated as the mean of the ISE of the individual $\hat{P}(y|x)$ curves for each x value) exhibits similar asymptotic convergence for relatively large N. For N lower than approximately 10^3, the mean ISE decreases relatively slowly with increasing N. For N greater than approximately 10^3, the mean ISE of the conditional estimate decays at a rate almost identical to that of the joint distribution. The additional time required to estimate the marginal of x is essentially negligible, and so the time required to compute the conditional fastKDE is nearly identical to that of the joint fastKDE.

From the perspective of data analysis, these are excellent properties for a data analysis tool. Many data analyses performed in the physical sciences assume a simple (linear) covariance structure: e.g., regression and/or correlation analysis. While the Spearman rank correlation coefficient deviates from the assumption of linearity inherent in the more common Pearson correlation coefficient, it still only provides an indication of the degree to which two quantities are related. In the synthetic example provided in Fig. 3(b), the conditional PDF shows what such analyses cannot: that there is a localized and relatively subtle transition in the two quantities. The ability to accurately, objectively, and rapidly estimate conditional PDFs could provide opportunities for physical insight that would be difficult or impossible to obtain using other methods.

The second non-trivial bivariate distribution against which we evaluate the fastKDE method is a mixture of normal distributions. This mixture-distribution is specifically designed to challenge KDE methods: it has a complex structure that includes both long- and shortwave components, it has major structures oriented along several axes, and none of these structures are aligned with axes corresponding to the abscissa or ordinate values. We define this mixture distribution as a random sample from any one of 350 normal distributions. The first 349 distributions are normal distributions, with variances of 0.3 and no covariance, sampled with equal probability (1/678); and the last distribution is a broader normal distribution, with its covariance matrix defined by the covariance matrix of the means of the 349 normal distributions, sampled with probability 349/678.

The locations of the first 349 normal distributions are defined as follows. The logo of the lead author's home institution was resized to 54 by 72 pixels, and color intensity values were discretized to either 1 or 0. These values were rendered as an ASCII text value and hand modified to remove artifacts of the resizing and digitizing processes. The image-space indices of the non-zero values in the ASCII text, of which there are 349, are chosen as the initial coordinates of the first 349 normal distributions. These initial coordinates are then transformed such that the coordinates are centered at (0, 0) and they are rotated by an arbitrary 37° degrees counterclockwise. The location of the last normal distribution is chosen to be centered on (0, 0) with a width and height that spans the first 349 normal distributions and with principal axes parallel to the abscissa and ordinate axes. This produces a mixture-distribution with a long-wave feature (the last normal distribution) combined with a complex structure of relatively shortwave features (the 349 normal distributions) with major structures aligned with two different axes: along the coordinate axes and along 37° and 127° counterclockwise from the abscissa axis. Fig. 4(a) shows the resulting PDF.

Fig. 4(b) and (c) show a fastKDE estimate of 10^6 and 10^4 samples from the mixture-distribution shown in Fig. 4(a). The estimate from 10^4 points shown in Fig. 4(c) represents the general features of the true PDF, with proper orientation, including the wording at the bottom right of the PDF and the orientation and general form of the building to the upper left of the wording. However, this estimate lacks many of the fine details, and one can barely discern that the feature in the lower right indeed corresponds to wording. It appears that the optimal kernel used in the fastKDE estimate of 10^4 samples contains a bandwidth that provides a compromise between the high variance (high wavelength) of the overall PDF and the small variance (low wavelength) features. In contrast, the estimate based on 10^6 points resolves well all of the high-frequency features. However, it appears that the use of a relatively narrow bandwidth kernel that allows resolution of these high-frequency features comes at the detriment of the low-frequency (high variance) background normal distribution. The imprint of individual kernels in the background PDF is evident in Fig. 4(b). Fig. 2 shows that the integrated, squared

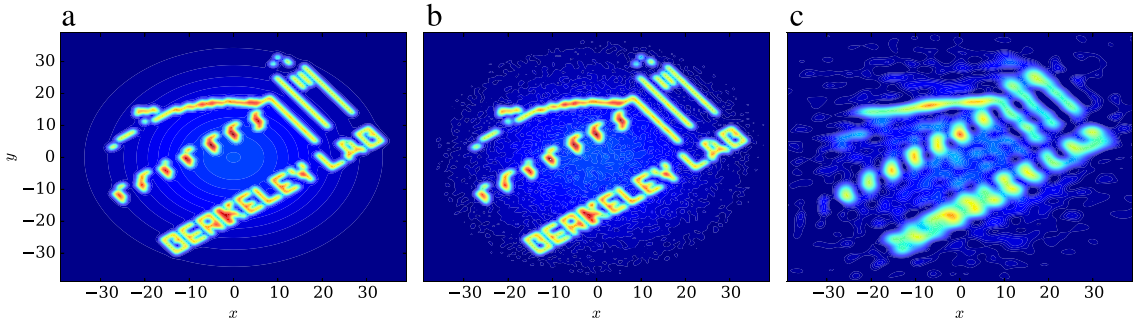

Fig. 4. A non-trivial mixture of normal distributions: (a) the underlying PDF, (b) a fastKDE estimate on 10^6 samples, and (c) a fastKDE estimate on 10^4 samples.

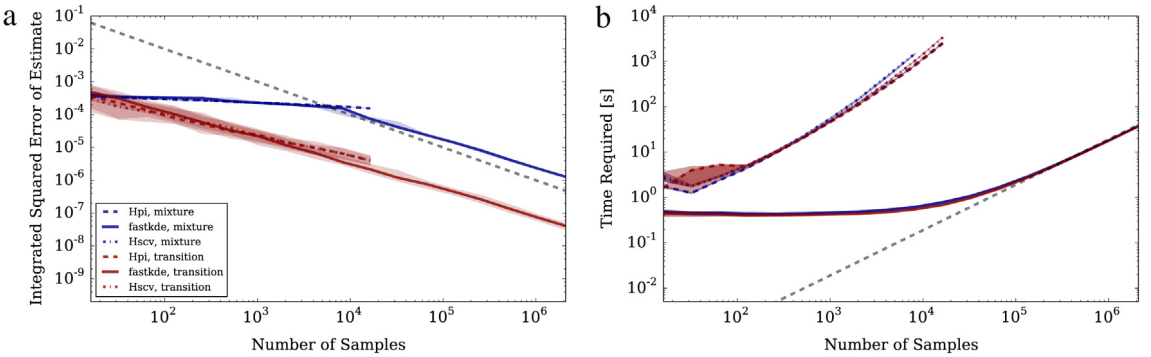

Fig. 5. (a) The integrated, squared error of KDE estimates, as a function of the number of data samples, on samples from the mixture and transition distributions described in Section 3.2. The blue and red lines depict the median of 30 simulations, and the shaded swaths depict the 5–95 percentile ranges. Solid curves represent the fastKDE estimate, and the dashed and dash-dotted curves represent KDE estimates from the 'ks' package of 'R' using the plug-in (Hpi) and smoothed cross-validation (Hscv) bandwidth selections respectively. The gray dashed line shows the theoretical maximum convergence rate N^{-1} for reference. (b) As in (a), but for the time required to perform the KDE. The gray dashed line shows N^1 for reference. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

error of the fastKDE estimate follows a complicated decay as samples are added. Regardless, the error decay suggests that the fastKDE method asymptotes to the underlying mixture-distribution as the number of samples increases.

3.3. Comparison with existing bandwidth selection methods

To demonstrate the performance of fastKDE relative to a publicly-available KDE package with automatic bandwidth selection methods, we repeated the convergence and timing tests shown in Fig. 2 using the kernel smoothing, `ks`, package from R. We focus on the two complex distributions described in Section 3.2: the transition and mixture distributions. We use two bandwidth selection methods available in the `ks` package that represent two mainstream, contemporary approaches to multivariate bandwidth selection: the plug-in selection method (Wand and Jones, 1995), and the smoothed cross validation selection method (Duong and Hazelton, 2005). Though our implementation of the fastKDE algorithm is written in Python, we use R for this analysis for two reasons: (1) Python does not presently have such automatic bandwidth selection methods as part of its core numerical/mathematical libraries, and (2) R is a widely used statistical language that arguably has the most cutting-edge bandwidth selection methods available.

Fig. 5(a) shows the integrated, squared error (similar to Fig. 2) for the fastKDE and R estimates. For direct comparison, all KDEs are performed on the same samples, and identical 513×513 grids are used for both fastKDE and R (though different grids are of course used for the transition and mixture distribution samples). The R-based estimates were run only up to sample-sizes of 16,384 and 8192 (transition and mixture distribution respectively) due to excessive computational time. The transition distribution exhibits almost identical error convergence characteristics as shown in Fig. 2. In contrast, the mixture distribution appears to exhibit better convergence for large sample sizes; note the steepening of the fastKDE ISE curve that occurs near $N = 10^4$, which indicates faster convergence. The 513×513 grid used in Fig. 5 has finer resolution than the grid used for Fig. 2, but the fastKDEs are otherwise identical. As a result, the estimates shown in Fig. 5 permit higher frequencies. It appears that for a PDF with so much structure, permitting high frequency components allows the fastKDE to continue to improve as points are added. The converse is true, as shown in Fig. 2(a); having too low a resolution effectively saturates the error in the fastKDE, such that additional points have no impact on the quality of the resulting KDE.

Both the plug-in and smoothed cross-validation bandwidth selection methods offer similar performance in R. For the transition distribution, both converge at a rate similar to the fastKDE method, although it appears that the error for the

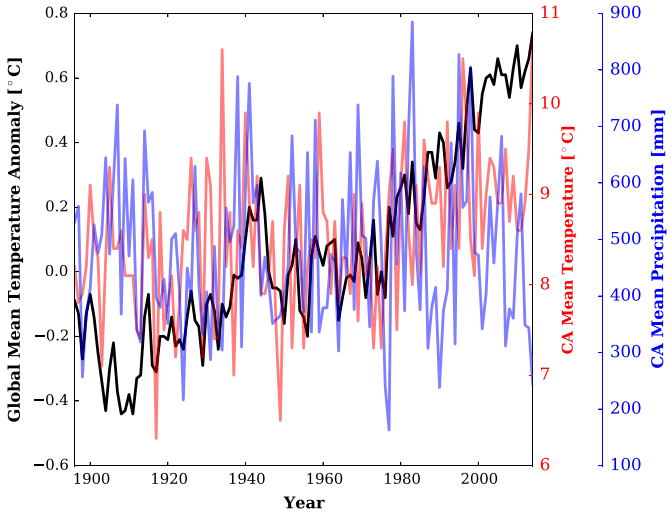

Fig. 6. A timeseries of global mean temperature (black line, left vertical axis), California (CA) mean temperature (red line, first right vertical axis), and CA mean precipitation (blue line, second right vertical axis). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

fastKDE is lower for sample sizes larger than approximately 10^4. For the mixture distribution, the convergence rate is similar to the fastKDE for small sample sizes, but neither method appears to exhibit the sharp jump in convergence rate that occurs at a sample size of approximately 10^4. It is not clear what is the origin of this convergence rate discontinuity for the fastKDE method, nor why neither the plug-in nor smoothed cross validation methods exhibit this jump in convergence. Examination of Fig. 1 shows that there are several large, contiguous hypervolumes that are barely detached from the first contiguous hypervolume. It is possible that the jump in convergence rate is associated with these large hypervolumes abruptly joining the first contiguous hypervolume as sample size increases and $|\mathscr{C}_{\text{min}}|^2$ lowers. We recreated a version of Fig. 1 with 20,000 samples instead of 10,000 (not shown), and indeed this results in the first hypervolume joining with the large, previously disconnected regions.

The main difference between fastKDE and the R-based KDE methods is in both the absolute time required for each KDE, and in the rate at which that time increases as samples are added; Fig. 5(b) shows the timing for all methods. Both the fastKDE and R-based KDE estimates were run on Cray XE6 nodes at the National Energy Research Scientific Computing Center (NERSC; on the machine called 'hopper'). Each node contains 2 twelve-core AMD 'MagnyCours' 2.1-GHz processors and 32 GB of 1333 MHz RAM per node. Despite being run on the same computational architecture, the absolute differences in timings between fastKDE and R should be interpreted cautiously due to large possible differences related to programming language overhead and to code structure. Regardless, it is notable that fastKDE consistently performs faster than either R-based method, with fastKDE being approximately 4 orders of magnitude faster for sample sizes of approximately 10^4. We note that we ceased testing the R-based methods when the time-per-KDE exceeded approximately 2 h because the total wall time required for our (admittedly linear) test code to run would exceed the total allowable 96 h wall time for standard computational jobs on the 'hopper' system at NERSC.

The most notable difference between the two methods is that fastKDE asymptotically approaches N^1 scaling for large sample sizes, whereas the R-based methods have a dependence on sample size that is clearly steeper than linear. Therefore, even if the performance of the R-based KDE methods could be tuned to be absolutely faster than fastKDE for the range of sample sizes shown, the super-linear dependence of the R methods on sample size would mean that fastKDE would inevitably be faster at some larger sample size.

4. Joint occurrence of temperature and precipitation

To demonstrate the utility of this method on a real dataset, we apply the fastKDE method to a dataset of 119 values of global mean temperature, and temperature and precipitation from California, USA. The California data are generated by spatially averaging station observations within the state and by temporally averaging values within the wet season, taken here to be the 6-month period beginning in November and ending in April the following calendar year (Karl and Koss, 1984). The data begin in November, 1895 and extend through April, 2014, which result in 119, semi-annual wet-season data points, extending from 1896 to 2014. The global mean temperature anomalies (relative to the 20th-century mean) are constructed as spatiotemporal annual averages of temperature data from the Global Historical Climatology Network-Monthly (GHCN-M) data set and the International Comprehensive Ocean-Atmosphere Data Set (ICOADS) (Smith et al., 2008). These data are publicly available from NOAA's National Centers for Environmental Information (NCEI) (formerly known as the National Climatic Data Center (NCDC)). Fig. 6 shows these data.

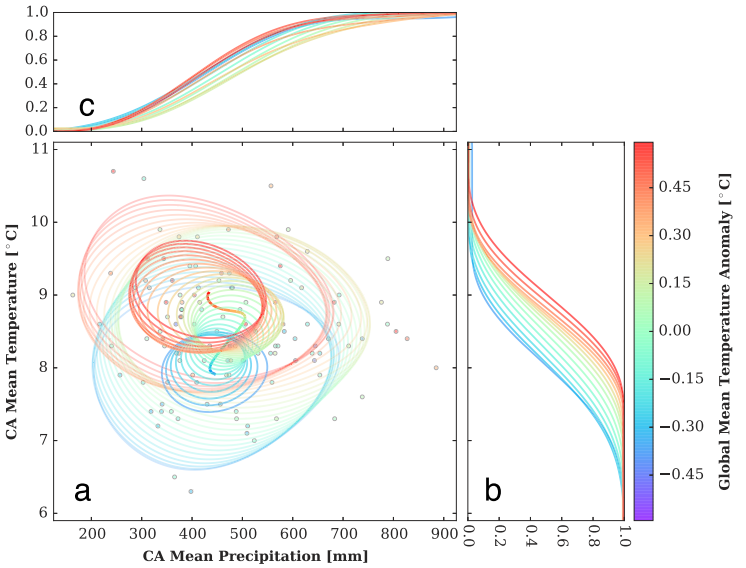

Fig. 7. (a) A fastKDE estimate of temperature versus precipitation, calculated as annual spatiotemporal averages over stations within California, USA (filled contours), and conditioned on the global mean temperature anomaly. The dots show the data points that went into the fastKDE. The colored, closed contours depict two levels of constant probability (inner and outer contours) from the conditional PDF. The thick colored line depicts the mean CA temperature and precipitation conditioned on global temperature anomaly. For the points, the closed contours, and the thick line, colors correspond to global mean temperature anomaly. (b) The survival function of temperature conditioned on global mean temperature anomaly; colors correspond to global temperature anomaly (c) as in (b), but for the cumulative distribution of precipitation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

It is well established that the statistics of temperature have not been stationary during the industrial era due to the effects of greenhouse gas on global temperatures (Pachauri et al., 2014). We hypothesize that the joint PDF of California temperature and precipitation is correspondingly nonstationary. If this hypothesis is correct, then statistics that treat this dataset as stationary, such as the return intervals estimated on this dataset by AghaKouchak et al. (2014), may need a more sophisticated treatment that explicitly accounts for this nonstationarity.

In order to test this hypothesis, we apply the fastKDE method to the dataset described above and shown in Fig. 6. We use the trivariate fastKDE to directly estimate the joint PDF of CA temperature and precipitation conditioned on global mean temperature. Fig. 7(a) depicts the 3D conditional PDF as a projection onto the CA precipitation–temperature plane, with contour colors representing the global mean temperature. The blue closed contours, which correspond to the relatively cool global temperatures from late in the 19th century, are asymmetric with their primary axis oriented from lower-right to upper-left in the figure. The orientation of these contours indicates a positive correlation between CA temperature and precipitation, with warm years that tend to be associated with wet years. In contrast, the red closed contours, which correspond to the relatively warm global temperatures of the present, exhibit a negative correlation; warm years tend to be associated with dry years. Following the contours from the bottom to the top, the transition from positive to negative correlation occurs in the vicinity of global mean temperature anomalies of 0.15 °C (yellow contours).

The conditional relationship between CA temperature and precipitation exhibits a similar transition. The thick colored line in the center of the projected PDF in Fig. 7(a) shows the mean CA temperature and precipitation conditioned on global mean temperature anomaly (same color coding as the contours). It is clear from this curve that as global mean temperature increases, so does CA mean temperature. The same is not true for precipitation. In the cooler part of the century (the blue portion of the curve), mean CA precipitation increases linearly as CA mean temperature increases. This relationship has the same sign as the positive correlation indicated by the conditional probability contours. However, when global mean temperature anomalies reach approximately 0.15 °C, the direction of the relationship reverses, such that mean precipitation decreases as mean temperature increases. Again, this relationship has the same sign as the negative correlation indicated by the contours. Fig. 6 shows that global mean temperature anomalies go permanently above 0.15 °C just prior to 1980. This timeframe is consistent with a well-documented shift in the climate of the north Pacific Ocean that occurred in 1977 (e.g., Deser and Phillips, 2006). We hypothesize that this change in both the correlation and in the trend in the mean indicates a transition in the nature of weather systems that transport moisture to California stemming from the 1977 climate shift. We will evaluate this hypothesis in a future manuscript. It is notable that the precipitation–temperature relationship appears to curve toward the left at the end, which may suggest that California is approaching another transition.

The nonstationarity of this dataset is also clearly evident in the conditional distribution of CA mean temperature shown in Fig. 7(b). As global mean temperatures increase, the distribution of temperatures monotonically shifts toward warmer temperatures. This shift results in some dramatic changes in the probabilities of extremely warm and extremely cold years. For global mean temperature anomalies of −0.3 °C, the probability of CA temperatures greater than 10 °C is $P \approx 0.01$.

However, for global mean anomalies of approximately 0.6 °C, such warm years are far more common ($P \approx 0.08$). Likewise, years colder than 7 °C occur with probability $P \approx 0.1$ for global mean temperature anomalies of −0.3 °C, whereas they almost never occur ($P < 0.01$) for global mean anomalies of 0.6 °C. This nonstationarity is particularly dramatic when framed in terms of conditional recurrence intervals. What used to be a 1-in-100 year event (CA temperature greater than 10 °C) is now an approximately 1-in-13 year event.

The conditional distribution of CA precipitation, shown in Fig. 7(c), is also quite non-stationary, though it does not exhibit a monotonic shift like the temperature distribution in Fig. 7(b). Rather than simply shifting, the precipitation distribution appears to broaden as global mean temperature anomalies increase toward 0.15 °C, and the distribution narrows again beyond. Similarly, and in accord with Fig. 7(a), the precipitation distribution also appears to shift slightly toward wetter values as global anomalies approach 0.15 °C, and they rebound toward drier values beyond. It is clear from Fig. 7(a) and (c) that the relationship between global mean temperature and precipitation is not simple, though it is not immediately clear why. This non-monotonicity should be investigated in future work.

5. Discussion

We have taken the optimal KDE methodology of Bernacchia and Pigolotti (2011), derived and implemented a multidimensional filter, and extended the non-uniform FFT ECF-calculation technique of O'Brien et al. (2014) to create the fastKDE method: a fast, objective, and multidimensional KDE technique. The fastKDE method lacks a need to specify parameters that control the quality of the PDF estimation method (i.e., histogram bin width and kernel bandwidth), which is a limitation inherent in existing non-parametric multidimensional PDF estimation methods. It also has the excellent convergence properties inherent in the original Bernacchia and Pigolotti (2011) method. Section 3 shows that this convergence property applies to PDFs with a range of covariance characteristics.

This method does have some subjectivity in terms of the choice of filter used on the ECF. In all of the tests shown in this manuscript, we use the lowest contiguous hypervolume filter, described in Section 2, with only the hypervolume closest to $\vec{t} = \vec{0}$. Since this single lowest contiguous hypervolume contains only ECF values centered around the $\vec{0}$ frequency, it effectively acts as a low-pass filter that smooths the resulting KDE. We implemented this filter to be flexible in terms of the number of contiguous hypervolumes used. Experimentation with the number of included hypervolumes in a PDF estimate (not shown) demonstrates that if too many hypervolumes are included in the estimate, the resulting PDF is modulated with high-frequency oscillations that are clearly artificial. Admittedly, we have focused our attention on PDFs that are essentially smooth and which are therefore well-suited to such a low-pass filter. A low-pass filter may not be a good choice for distributions that contain a wide range of frequencies, such as distributions with Comb marginals. Bernacchia and Pigolotti (2011) show that their method is general enough to even work well on a Comb distribution, but the filter must be constructed to allow a suitably wide range of frequencies. Using only a single lowest contiguous hypervolume would not likely satisfy this need, but it may suffice to simply use a higher number of hypervolumes. Further development and testing of this methodology on such broadband distributions may be beneficial.

Fat tailed distributions present another challenge to this method. For example, applying this method directly to samples from a Pareto distribution yields a very poor representation of the true PDF. This difficulty arises because the fastKDE bandwidth is fixed for all areas of the data space, but the data samples are spread over a huge range (sometimes multiple orders of magnitude, depending on power law slope and sample size). This inhibits multiple kernels from overlapping, which is the essential requirement for a successful KDE. Fortunately, this is a known issue that generally effects KDE methods, and a good solution is simply to transform the data to a more compressed space prior to the KDE, and then to subsequently transform the PDF back to the original space (Wand et al., 1991). For example, given samples χ_i from a Pareto distribution, the PDF estimate is much better if the fastKDE is computed on $\log(\chi_i)$. Transforming the fastKDE back to data space simply involves exponentiating the values of the fastKDE grid and dividing the fastKDE estimate by the values of the new grid:

$$\begin{aligned} \eta_i &\equiv \log(\chi_i) \\ \hat{P}'(\chi) &= \frac{1}{\chi}\hat{P}_{KDE}\left(\chi = e^{\eta}\right). \end{aligned} \tag{12}$$

This results in a transformed fastKDE $\hat{P}'$ calculated on a grid with log-spacing.

This method is designed to apply to arbitrarily high dimensionality, and our Python-based implementation of this method allows for that generality. However, there are two practical factors that may inhibit its use for high dimensional systems, and both are related to a curse of dimensionality. The first curse of dimensionality is that the effective information density of each sample is reduced as dimensionality increases. This means, for example, that 10,000 trivariate samples yield less information for a KDE than 10,000 bivariate samples. This is a general issue for KDE methods. Recently Nagler and Czado (in preparation) have proposed an intriguing solution that may solve this issue. Instead of directly performing a full D-dimensional KDE, they break the problem down using a copula approach. They use a KDE to estimate the marginals separately, and they assume that the copula can be decomposed into a product of bivariate copulas (a vine copula approach). Estimation of this copula then only requires estimation of two-dimensional PDFs for an arbitrarily-dimensioned problem, and Nagler and Czado (in preparation) argue that this allows one to avoid this curse of dimensionality. Such an approach could conceivably be applied to the fastKDE method to allow higher dimensional KDE estimates.

Computational storage requirements are the second curse of dimensionality that may inhibit application of this method to higher dimensional samples. The fastKDE method, and indeed any KDE method, requires that the KDE (and any intermediate values, such as the ECF) be stored on a D-dimensional grid. Assuming an equal number of points M in each dimension, this grid then requires that M^D floating point values be stored. If we assume $M = 33$ (a relatively coarse grid), a dimensionality of 20, and floating point values, the resulting storage requirement is $S = 32 \cdot 9^{20} \approx 10^{20}$ bytes. For reference, this is approximately the same order of magnitude as all data stored on Earth in 2007 (Hilbert and López, 2011). However, the vine copula approach of Nagler and Czado (in preparation) might provide a solution to this curse of dimensionality as well. If the KDE can be constructed as the product of the marginal KDEs and bivariate copula KDEs, then this implies that all of the information in the full KDE is encoded in the marginal and bivariate KDEs. A D-dimensional KDE using the Nagler and Czado (in preparation) method requires storage of D marginal densities and $D(D-1)/2$ bivariate densities, leading to an exponentially less stringent storage requirement of $S' = D \cdot M + D(D-1)/2 \cdot M^2$. Storing the KDE in this manner would require some special algorithmic considerations, since the full PDF values would have to be reconstructed on an as-needed basis by multiplying the marginals and the bivariate copulas.

While the fastKDE method may benefit from some further methodological development in terms of extending to problems of high-dimensionality, we have clearly demonstrated its functionality for lower dimensional problems. The ability of the method to encode high-fidelity covariance information is a promising capability from a data analysis point of view. As Fig. 3(b) shows, the fastKDE method can be used to directly estimate conditional distributions. This is a boon to data analysts looking for non-trivial relationships among variables. The venerable tradition of examining scatter plots, correlations, and regressions would fail to provide useful information about the transition inherent in the x and y variables of Fig. 3(a). Even simply looking at the bivariate PDF in Fig. 3(a) shows only the slightest hint that there is a non-trivial relationship between x and y. In contrast, the conditional distribution immediately shows that the y value exhibits a transition near a specific value of x. It should be noted that the fastKDE method reliably reproduces this subtle transition (a jump of $\Delta y = 4$) despite the original samples having a spread in y of $\sigma_y = 12$.

6. Summary

We designed the fastKDE method with the intent that it be used as a general-purpose KDE tool. The simplicity of not having to deal with bandwidth matrix specification, combined with the speed of the method, makes it suitable as such. We have demonstrated that the fastKDE method is applicable to KDE estimates of at least 3 variables (Sections 3.1 and 4), and we have shown a number of examples of its successful use on bivariate samples (Sections 3.1, 3.2 and 4). It is particularly encouraging that the fastKDE yields a high-enough fidelity representation of the coupling of two variables that a good representation of a non-trivial conditional distribution can be directly estimated from the fastKDE (Section 3.2).

It is worth noting that the fastKDE method takes approximately 6 s to operate on the trivariate temperature–precipitation dataset used in Section 4 (in contrast to R, which takes approximately 60 s using either bandwidth selection method on the same dataset). The bivariate distributions shown in Fig. 4(b) and (c) took only a fraction of a second compute. This is consistent with the timing values shown in Fig. 2(b). Even for a dataset of 10^6 points, a bivariate fastKDE calculation only takes approximately 10 s to calculate, and a trivariate distribution takes less than two minutes. This, along with the simple interface to our implementation of the fastKDE method, makes its speed and ease of use comparable to that of typical histogram functions available in Python, R, and other programming languages. The statistical performance of fastKDE is comparable to the performance of state-of-the-science KDE methods available in R, with the benefit that the computational cost is several orders of magnitude smaller. This method has the added benefit that there is no need for user-selected parameters, such as bin width or kernel bandwidth. These properties make the fastKDE suitable as a general-purpose PDF estimation tool.

We have implemented the fastKDE method in Python and Cython; Cython is used to develop the loop-intensive portions of the code (e.g., the ECF calculation and the nuFFT). The package includes both an object-oriented interface and a simple functional interface. The object-oriented interface allows for more complex operations, such as a 'summation' operation that leverages the linearity of the Fourier transform. This allows the fastKDE method to be utilized in a map-reduce framework that enables parallelization of the KDE calculation. In such a framework, the samples are mapped to participating computational cores, and the ECF is calculated on each core; the summation operation then reduces the ECF information from each core into a single ECF, after which the filter operation and the back Fourier transform are applied to yield the full fastKDE. The simple functional interface, which allows fastKDE estimation of PDFs and conditional PDFs, operates in a manner similar to the histogram and KDE routines available in the `numpy` and `SciPy` Python libraries.

Our Python/Cython-based implementation of the fastKDE method is available under a Berkeley System Distribution license, and it is free for academic use. The code is hosted at https://bitbucket.org/lbl-cascade/fastkde.

Acknowledgments

The authors would like to thank two anonymous reviewers whose comments greatly helped improve the quality of the manuscript. The authors would also like to thank Dr. Chris Paciorek of UCB for helpful comments in the framing of the manuscript. This research was supported by the Director, Office of Science, Office of Biological and Environmental Research

of the US Department of Energy Regional and Global Climate Modeling Program (RGCM) (ESD13052) and used resources of the National Energy Research Scientific Computing Center (NERSC) (m1949 and m1517), also supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231.

References

AghaKouchak, A., Cheng, L., Mazdiyasni, O., Farahmand, A., 2014. Global warming and changes in risk of concurrent climate extremes: Insights from the 2014 California drought. Geophys. Res. Lett. (ISSN: 1944-8007) 41 (24), 8847–8852. http://dx.doi.org/10.1002/2014GL062308.

Bernacchia, A., Pigolotti, S., 2011. Self-consistent method for density estimation. J. R. Stat. Soc. Ser. B Stat. Methodol. (ISSN: 1467-9868) 73 (3), 407–422. http://dx.doi.org/10.1111/j.1467-9868.2011.00772.x.

Bony, S., Dufresne, J.-L., Le Treut, H., Morcrette, J.-J., Senior, C., 2004. On dynamic and thermodynamic components of cloud changes. Clim. Dynam. (ISSN: 1432-0894) 22 (2–3), 71–86. http://dx.doi.org/10.1007/s00382-003-0369-6.

Deser, C., Phillips, A.S., 2006. Simulation of the 1976/77 climate transition over the north pacific: Sensitivity to tropical forcing. J. Clim. (ISSN: 1520-0442) 19 (23), 6170–6180. http://dx.doi.org/10.1175/jcli3963.1.

Duong, T., Hazelton, M.L., 2005. Cross-validation bandwidth matrices for multivariate kernel density estimation. Scand. J. Statist. (ISSN: 1467-9469) 32 (3), 485–506. URL: http://dx.doi.org/10.1111/j.1467-9469.2005.00445.x.

Greengard, L., Lee, J.-Y., 2004. Accelerating the nonuniform fast Fourier transform. SIAM Rev. 46 (3), 443–454 http://dx.doi.org/10.1137/S003614450343200X. URL: http://epubs.siam.org/doi/abs/10.1137/S003614450343200X.

Heidenreich, N.-B., Schindler, A., Sperlich, S., 2013. Bandwidth selection for kernel density estimation: a review of fully automatic selectors. AStA Adv. Stat. Anal. (ISSN: 1863-818X) 97 (4), 403–433. http://dx.doi.org/10.1007/s10182-013-0216-y.

Hilbert, M., López, P., 2011. The world's technological capacity to store, communicate, and compute information. Science 332 (6025), 60–65. http://dx.doi.org/10.1126/science.1200970. URL: http://www.sciencemag.org/content/332/6025/60.abstract.

Karl, T., Koss, W., N. C. D. C. (US), 1984. Regional and National Monthly, Seasonal, and Annual Temperature Weighted by Area, 1895–1983. In: Historical Climatology Series, National Climatic Data Center, URL: https://books.google.com/books?id=3OjSHAAACAAJ.

Larson, V.E., Golaz, J.-C., Cotton, W.R., 2002. Small-scale and mesoscale variability in cloudy boundary layers: Joint probability density functions. J. Atmos. Sci. (ISSN: 1520-0469) 59 (24), 3519–3539. http://dx.doi.org/10.1175/1520-0469(2002)059<3519:SSAMVI>2.0.CO;2.

Lee, H., Kim, J., Waliser, D.E., Loikith, P.C., Mattmann, C.A., McGinnis, S., 2014. Using joint probability distribution functions to evaluate simulations of precipitation, cloud fraction and insolation in the North America Regional Climate Change Assessment Program (NARCCAP). Clim. Dynam. (ISSN: 1432-0894) 45 (1–2), 309–323. http://dx.doi.org/10.1007/s00382-014-2253-y.

Luedicke, J., Bernacchia, A., 2014. Self-consistent density estimation. Stata J. 14 (2), 237–258 (22). URL: http://www.stata-journal.com/article.html?article=st0334.

Marchand, R., Ackerman, T., Smyth, M., Rossow, W.B., 2010. A review of cloud top height and optical depth histograms from MISR, ISCCP, and MODIS. J. Geophys. Res. (ISSN: 0148-0227) 115 (D16), 1–25. http://dx.doi.org/10.1029/2009jd013422.

Nagler, T., Czado, C., (in preparation). Evading the curse of dimensionality in multivariate kernel density estimation with simplified vines, arXiv Preprint arXiv:1503.03305.

O'Brien, T.A., Collins, W.D., Rauscher, S.A., Ringler, T.D., 2014. Reducing the computational cost of the ECF using a nuFFT: A fast and objective probability density estimation method. Comput. Statist. Data Anal. (ISSN: 0167-9473) 79, 222–234 URL: http://www.sciencedirect.com/science/article/pii/S016794731400173X.

Pachauri, R.K., Allen, M., Barros, V., Broome, J., Cramer, W., Christ, R., Church, J., Clarke, L., Dahe, Q., Dasgupta, P., et al. 2014. Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.

Schumacher, C., Houze, R.A., 2003. Stratiform rain in the tropics as seen by the TRMM precipitation Radar*. J. Clim. (ISSN: 1520-0442) 16 (11), 1739–1756. http://dx.doi.org/10.1175/1520-0442(2003)016<1739:SRITTA>2.0.CO;2.

Scott, D.W., 2008. The curse of dimensionality and dimension reduction. In: Multivariate Density Estimation. John Wiley & Sons, Inc., pp. 195–217. URL: http://dx.doi.org/10.1002/9780470316849.ch7.

Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis, Vol. 26. CRC Press.

Smith, T.M., Reynolds, R.W., Peterson, T.C., Lawrimore, J., 2008. Improvements to NOAA's historical merged land-ocean surface temperature analysis (1880–2006). J. Clim. (ISSN: 1520-0442) 21 (10), 2283–2296. http://dx.doi.org/10.1175/2007jcli2100.1.

Wand, M., Jones, M., 1995. Kernel Smoothing. In: Monographs on Statistics and Applied Probability, vol. 60. Chapman & Hall, London.

Wand, M.P., Marron, J.S., Ruppert, D., 1991. Transformations in density estimation. J. Amer. Statist. Assoc. (ISSN: 1537-274X) 86 (414), 343–353. http://dx.doi.org/10.1080/01621459.1991.10475041.

Wit, E., Heuvel, E.v.d., Romeijn, J.-W., 2012. All models are wrong...: an introduction to model uncertainty. Stat. Neerl. (ISSN: 0039-0402) 66 (3), 217–236. http://dx.doi.org/10.1111/j.1467-9574.2012.00530.x.