

Density estimation with distribution element trees

Daniel W. Meyer¹

Received: 6 November 2016 / Accepted: 2 May 2017 / Published online: 16 May 2017
© Springer Science+Business Media New York 2017

Abstract The estimation of probability densities based on available data is a central task in many statistical applications. Especially in the case of large ensembles with many samples or high-dimensional sample spaces, computationally efficient methods are needed. We propose a new method that is based on a decomposition of the unknown distribution in terms of so-called distribution elements (DEs). These elements enable an adaptive and hierarchical discretization of the sample space with small or large elements in regions with smoothly or highly variable densities, respectively. The novel refinement strategy that we propose is based on statistical goodness-of-fit and pairwise (as an approximation to mutual) independence tests that evaluate the local approximation of the distribution in terms of DEs. The capabilities of our new method are inspected based on several examples of different dimensionality and successfully compared with other state-of-the-art density estimators.

Keywords Nonparametric density estimation · Adaptive histogram · Kernel density estimation · Adaptive binning · Polynomial histogram · Curse of dimensionality · High dimensional · Big data · Pólya tree · Density estimation tree

Mathematics Subject Classification 62G07 · 62H10 · 62G10

1 Introduction

In this work, we propose a new method for estimating a probability density $p(\mathbf{x})$ of the random variable vector \mathbf{X} at position $\mathbf{x} = (x_1, \dots, x_d)^\top$ of the bounded, d -dimensional probability space Ω based on a given ensemble of n samples $\mathbf{x}_1, \dots, \mathbf{x}_n$. Unlike other methods, our approach is applicable for large and/or high-dimensional datasets. Density estimation methods are essentially categorized into parametric and nonparametric methods (e.g., Härdle et al. 2004; Scott 2015). While we focus in this work on the development of a nonparametric approach, our method can be viewed as a hybrid between the two categories (Yang 2008). Therefore, before we focus on nonparametric methods, we start the following literature review by briefly explaining the concept of parametric density estimation.

In parametric density estimation, a parametric density model $p(\mathbf{x}|\boldsymbol{\theta})$ is given in analytical form. With the parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots)$ estimated from the available data, we arrive at a complete density characterization. This characterization can be evaluated with statistical tests or more precisely composite goodness-of-fit tests such as for example the χ^2 test of Pearson (1900). Here, for a given significance level α , we either do not reject the null hypothesis $p(\mathbf{x}) = p(\mathbf{x}|\boldsymbol{\theta})$ and reject the alternative $p(\mathbf{x}) \neq p(\mathbf{x}|\boldsymbol{\theta})$ or vice versa. For small ensembles, these tests lose their power meaning that they become unable to detect alternatives and instead do not reject the null hypothesis (e.g., Steele and Chaseling 2006). Parametric density estimation is of limited generality, since a predefined density model is needed.

The author is grateful to Marco Weibel for his help during the preparation of this manuscript. Very valuable feedback from an associate editor and two reviewers and helpful input from Oliver Brenner and Florian Müller are gratefully acknowledged. Moreover, the author acknowledges helpful comments from Nina Roth and feedback on the initial version of this manuscript from Patrick Jenny, both from ETH Zürich. The author has been financially supported by ETH Zürich.

✉ Daniel W. Meyer
meyerda@ethz.ch

¹ Institute of Fluid Dynamics, ETH Zürich, Zürich, Switzerland

Nonparametric density estimation is more general as no parametric density model is required. A widely used representative from this category of density estimators is the histogram (Fix and Hodges 1951; Scott 2015, section 3). Here, the probability space is typically discretized into equally spaced bins and the probability density within one bin is set proportional to the number of samples in that bin. To determine the bin width h , the mean integrated square error (MISE), or more precisely its asymptotic approximation referred to as AMISE, is minimized for cases with known $p(\mathbf{x})$; leading for example with a Gaussian $p(\mathbf{x})$ to an h given by the so-called normal reference rule [Härdle et al. 2004, equation (2.26)]. Histograms are conceptually simple, but have certain drawbacks. First of all, they are relatively inaccurate and with the overall number of bins growing exponentially with the number of dimensions d , conventional histograms become prohibitive in cases with large d (Scott 2015, section 3.4.1). More economical and possibly accurate variants with adaptive bin widths are available (e.g., Kogure 1987). These variants attempt to better resolve regions with large density variation with finer bins, while using larger bins in more uniform areas. Adaptive histograms rely on estimates of gradients of $p(\mathbf{x})$ or the use of percentile meshes (Scott 2015, section 3.2.8).

One possible generalization of histograms is the cubic log-spline (Kooperberg and Stone 1991). Here, the one-dimensional probability space is partitioned into bins like in a histogram, but within a bin, a parametric model in the form of a cubic polynomial or spline is applied. The bin bounds or knots are placed subject to a rule that was derived based on experience in fitting log-spline models (Kooperberg and Stone 1991, section 5.1). The coupled polynomial coefficients or spline parameters in turn are determined numerically with a Newton–Raphson method by maximizing the likelihood of the log-spline estimator. The overall number of knots is either calculated by a predefined rule or sequential knot-deletion. A related approach that is applied in multivariate settings is the so-called polynomial histogram. Here, the density in a bin varies according to a linear, quadratic, or higher-order polynomial (Scott and Sagae 1997; Jing et al. 2012). Unlike in the log-spline method, the polynomial coefficients are determined locally based on statistical moments estimated within individual bins (conditional moments). The use of higher-order polynomials enables the use of larger bins globally while maintaining the same MISE compared to conventional histograms (Jing et al. 2012, table 1). This is because to some extent the density variation is accounted for already at the level of an individual bin. Especially in cases where d is large, this reduces computational costs. So far, equidistant and prescribed non-equidistant bin grids were considered (Jing et al. 2012; Zaunders et al. 2016).

Besides histograms, a second important category of nonparametric methods is kernel density estimation (KDE)

(Rosenblatt 1956; Sheather 2004). Here, samples are not grouped into bins, but are equipped with kernel functions, e.g., triangular, Gaussian (Silverman 1998, table 3.1). Similarly like the bin width in histograms, kernels have a certain support or bandwidth h . Optimal global bandwidths were determined with AMISE analysis based on second-order derivatives of $p(\mathbf{x})$ [e.g., Scott 2015, equations (6.18) and (6.50)]. Adaptive more accurate methods that reduce the bandwidth in dense areas and use wider kernels in sparse regions have been documented (e.g., Loftsgaarden and Quesenberry 1965; Achilleos and Delaigle 2012), but are not completely satisfactory yet. For example, Scott (2015, section 6.8) summarized quite recently: ‘Adaptive methods hold much promise, but usually introduce many new parameters that are difficult to estimate, and frequently introduce artifacts of the sample (rather than the underlying density).’ Certain efficient spectral KDE implementations rely on bandwidths that are constant within different directions x_i in probability space (e.g., O’Brien et al. 2016; Silverman 1998, section 3.5). The highly cited KDE method of Botev et al. (2010) on the other hand is adaptive, is available in the form of an efficient spectral implementation (Botev 2007), and reduces boundary bias effects of existing KDE approaches. Recently, a new KDE method for bounded domains that eliminates boundary bias issues was presented in the context of functional data analysis by Petersen and Müller (2016). A high-dimensional implementation of the method of Botev et al. (2010) was recently made available, where the cost per density query scales as an exponential fraction of the number of samples (Botev 2016). Alternative KDE methods were summarized in (Park and Marron 1990; Park and Turlach 1992; Cao et al. 1994; Jones et al. 1996; Silverman 1998; Härdle et al. 2004; Scott 2015).

A density estimator for exponential families—which is a broad class of densities—that has superior convergence properties compared to KDE was presented by Sriperumbudur et al. (2013). This estimator is based on minimizing the Fisher divergence and requires the solution of an $(nd+1) \times (nd+1)$ linear system. Most interestingly, the advantage to KDE grows with increasing dimensionality d as was shown numerically.

Density estimation based on mixture distributions can be viewed as a generalization of KDE, where the unknown distribution is expressed like in KDE as a superposition of probability densities (e.g., Wang and Wang 2015). These densities are referred to as mixture models, and each mixture model is associated with an ensemble subset containing—unlike in KDE—multiple samples. In this context, the so-called Dirichlet process mixture models, that go back to (Ferguson 1973), have received renewed attention after progress was made in the numerical estimation of mixture parameters (e.g., Neal 2000).

To arrive at an efficient adaptive method, tree-based approaches have been proposed more recently (e.g., [Ram and Gray 2011](#); [Wong and Ma 2010](#); [Jiang et al. 2016](#)). These approaches start from the probability space $C = \Omega$, which is typically assumed to be a d -dimensional hypercube. The root cube Ω , or cuboid to be more precise, is recursively subdivided into smaller cuboids, e.g., C_1 and C_2 with $C = C_1 \cup C_2$, based on suitable conditions. More specifically, [Ram and Gray \(2011\)](#) discussed density estimation trees or shorter density trees that are derived from decision trees ([Breiman et al. 1984](#)) and are constructed based on optimal split operations of cuboids at tree nodes. These splits are optimal in the sense that they maximally reduce the integral square error (ISE), i.e., $\text{ISE}(C) > \text{ISE}(C_1) + \text{ISE}(C_2)$. The optimum is found from all possible splits in each dimension. The large number of possibilities renders the method expensive for large datasets [[Ram and Gray 2011](#), equation (9)]. A preset lowest threshold is set to stop the splitting. The splitting process is followed by a tree pruning and cross-validation step. The resulting tree or more precisely its leafs provide a histogram with adaptive bin widths. The tree structure enables a fast density estimation at a cost proportional to the tree depth.

[Wong and Ma \(2010\)](#) and [Jiang et al. \(2016\)](#), on the other hand, have introduced and numerically implemented, respectively, the optional Pólya tree (OPT) method. In this approach, cuboids C are partitioned and uniform cuboid probability densities $p(\mathbf{x}|C_i)$ (or $q(\mathbf{x}|C_i)$ in their work) are assigned according to probabilistic processes involving Bernoulli and Dirichlet random numbers, respectively. The random partitioning process relies on selection and stopping probabilities, $\lambda_i(C|\mathbf{x})$ and $\rho(C|\mathbf{x})$, respectively, that are calculated based on a recursive expression (equation (2.1) in [Jiang et al. 2016](#)) that necessitates partitioning down to subregions with either zero or one sample. This leads to a close to exponential growth in computing time as a function of the number of dimensions ([Jiang et al. 2016](#), figure 1). As a remedy, in naive inexact OPT (NI-OPT) partitioning limits were introduced like smallest number of points in subregions or smallest size of subregions to arrive at manageable trees. Moreover, limited-lookahead OPT (LL-OPT) introduces two additional tuning parameters that control the tree depth for the recursive calculation and thus help to reduce memory requirements and computing times of OPT and NI-OPT. The density estimates $p(\mathbf{x})$ that result from OPT implementations are piecewise uniform within subregions C or bins and are adaptive within individual dimensions.

To relate back to our initial discussion on parametric methods, [Ma and Wong \(2011\)](#) outline an OPT method for goodness-of-fit testing of large datasets against a given base distribution p_0 (or q_0 in their work). Here, instead of uniform cuboid probability densities, p_0 is used, but otherwise the OPT methodology of [Wong and Ma \(2010\)](#) applies. As

a measure of the overall goodness-of-fit, the integral stopping probability $\rho(\Omega)$ was proposed ([Ma and Wong 2011](#), section 3).

In the present contribution, we develop a new nonparametric density estimator that is adaptive and cost efficient. In view of the so-called curse of dimensionality, that is diminishing MISE convergence rates for increasing dimensionality [e.g., [Scott 2015](#), equation (3.67)], methods that enable the treatment of large datasets at small computational costs become important. While our method adopts certain features from the previously introduced polynomial histograms and tree-based techniques, our new method is, however, conceptually and algorithmically simpler and computationally more efficient compared to these approaches.

Our development starts by recognizing that a histogram is essentially a collection of disjoint piecewise uniform distributions. In a histogram with equally sized bins, depending on the true density distribution, a uniform approximation may be accurate in certain bins, while inaccurate in others. Like in the context of parametric methods, we could apply a statistical test to evaluate the goodness-of-fit of the data in individual bins against a uniform null hypothesis. Thus, a natural recursive way of constructing a histogram emerges: Ω is defined as a root hypercuboid that encloses the available data. With a suitable goodness-of-fit test, it is tested whether the uniform null hypothesis is adopted at a given significance level based on the data in the cuboid. If not, the root cuboid is split and the testing and splitting is recursively repeated for all cuboids resulting from this and subsequent splits. Eventually, the test will adopt the uniform hypothesis in all resulting subcuboids. This is because the power of the test diminishes as the number of samples in each subcuboid becomes smaller during consecutive splitting. Finally, an adaptive histogram results, where in each bin the uniform distribution is supported by a positive outcome of a goodness-of-fit test. This approach is not entirely nonparametric as a significance level has to be prescribed. The nested spatial arrangement of cuboids and subcuboids, that is organized in a tree structure, enables the fast query of density estimates locally.

To further reduce the number of bins, a next step is to apply instead of uniform or constant bin-densities linear or higher-order densities similar to polynomial histograms. The corresponding polynomial coefficients are derived from the bin data, and a composite goodness-of-fit test is applied. In this sense, the proposed method is a hybrid between parametric and nonparametric density estimation. With the density inside a bin being a polynomial of a certain order, that approximates the density distribution locally, we refer to it as a distribution element (DE). Since the resulting density estimate is given in the form of a tree with DEs at its leafs, we refer to our approach as DE tree (DET) density estimator. In Sect. 2, we provide details about the formulation of the DET

estimator and in Sect. 3 present comparative applications. Concluding remarks are provided in Sect. 4.

2 Formulation

We aim at accurately estimating the probability density $p(\mathbf{x})$ at different positions $\mathbf{x} = (x_1, \dots, x_d)^\top \in \Omega$ based on a given ensemble including n samples $\mathbf{x}_1, \dots, \mathbf{x}_n$. The probability space Ω is defined by a hypercuboid, i.e.,

$$\Omega = \prod_{i=1}^d [x_{i,l}, x_{i,u}],$$

with lower and upper bounds $x_{i,l}$ and $x_{i,u}$, respectively, of components x_i such that $\mathbf{x}_j \in \Omega \forall j = 1, 2, \dots, n$. This condition is, for example, met if

$$x_{i,l} = \min_{j=1}^n x_{i,j} \text{ and } x_{i,u} = \max_{j=1}^n x_{i,j},$$

where $x_{i,j}$ refers to the i th component of sample \mathbf{x}_j .¹ In our method, Ω is split recursively into m smaller hypercuboids or simply cuboids

$$C_k = \prod_{i=1}^d [x_{i,l}^k, x_{i,u}^k] \subset \Omega$$

with $k = 1, 2, \dots, m$. The cuboids C_k are disjoint and satisfy $\Omega = \bigcup_{k=1}^m C_k$. A cuboid C_k comprises together with a local density $p_k(\mathbf{x})$ the k th DE. Based on all DEs, the DET density estimator is then given by

$$p(\mathbf{x}) = \sum_{k=1}^m p_k(\mathbf{x}) \forall \mathbf{x} \in \Omega. \quad (1)$$

In Sect. 2.1, we define $p_k(\mathbf{x})$, and in Sect. 2.2, we introduce the recursive splitting method that leads to the DET density estimator.

2.1 Distribution elements

Similar to a bin in a histogram, a DE can be viewed as the least complex building block or atom of a density distribution estimate or DET. Therefore, we chose a simple analytical form that is suitable to approximate the density in a small

subregion C_k of Ω . We define the probability density of DE k as

$$p_k(\mathbf{x}) = \begin{cases} \frac{n(C_k)}{n} \prod_{i=1}^d p[x_i | \theta_i(C_k)] & \forall \mathbf{x} \in C_k \\ 0 & \text{otherwise,} \end{cases} \quad (2)$$

where $p[x_i | \theta_i(C_k)]$ are marginal densities of components x_i with local parameter vectors θ_i and $n(C_k)/n$ is the fraction of all samples n that reside in C_k . Therefore, in each DE, the random variables X_i are approximated as statistically independent. This has important implications, since it enables us to break the exponential growth of bins in terms of d as will become clear in the next section.

Insertion of expression (2) in Eq. (1) and integration over the entire probability space Ω reveals that, since $\sum_{k=1}^m n(C_k) = n$, the DET estimator integrates to one and therefore satisfies the normalization condition of a probability density function (PDF). Moreover, if $p[x_i | \theta_i(C_k)] \geq 0 \forall x_i \in [x_{i,l}^k, x_{i,u}^k]$, the DET estimator is nonnegative for $\mathbf{x} \in \Omega$ and therefore is a PDF.

With the uniform marginal density

$$p[x_i | \theta_i(C_k)] = \frac{1}{x_{i,u}^k - x_{i,l}^k}, \quad (3)$$

which defines a so-called constant DE, we obtain from Eq. (2)

$$p_k(\mathbf{x}) = \frac{n(C_k)}{n} \frac{1}{\prod_{i=1}^d (x_{i,u}^k - x_{i,l}^k)},$$

which is the familiar density in a histogram bin (e.g., Scott 2015, section 3.4). Similarly, the density of a linear DE is given as

$$p[x_i | \theta_i(C_k)] = \frac{\left(\frac{x_i - x_{i,l}^k}{x_{i,u}^k - x_{i,l}^k} - \frac{1}{2} \right) \theta_{i,1}(C_k) + 1}{x_{i,u}^k - x_{i,l}^k}, \quad (4)$$

where $\theta_{i,1}(C_k)$ is a slope parameter. We notice that a linear DE with $\theta_{i,1} = 0$ is equivalent to a constant DE. To estimate the slope based on the data points inside C_k , we apply the following minimum mean square error (MMSE) estimator

$$\theta_{i,1}(C_k) = \frac{n(C_k) s_i(C_k)^3}{n(C_k) s_i(C_k)^2 + 144 \langle X_i^2 | C_k \rangle} \quad (5)$$

with $s_i(C_k) = 6(2\langle X_i | C_k \rangle - 1)$. Here, the quantities $\langle X_i | C_k \rangle$ and $\langle X_i^2 | C_k \rangle$ are the mean and variance estimates based on component x_i of the samples contained in cuboid C_k . A derivation of the MMSE slope estimator (5) is included in “Appendix”. To make sure that $p[x_i | \theta_i(C_k)] \geq 0$, we clip $\theta_{i,1}(C_k)$ such that $\theta_{i,1}(C_k) \in [-2, 2]$. In this work, we focus

¹ Fixing the domain bounds based on the data range leads to bounds that are almost certainly too narrow. Accordingly, the resulting density estimates will display a bias toward too high values.

on DEs with uniform and linear marginal PDFs, but higher-order DEs are conceivable as well.

It is pointed out that the polynomials (2) are differently constructed compared to the ones applied by Jing et al. [2012, equations (5) and (6)]. Our DE classification is based on the marginal densities $p[x_i|\theta_i(C_k)]$, while Jing et al. classify the expanded polynomial $p_k(\mathbf{x})$ and allow for statistical dependence among the components x_i .

2.2 Distribution element tree

For the construction of the DET density estimator (1), a constant or linear DE is assigned to the probability space Ω or root cuboid depending on the required order. To verify for the root DE or any subsequent DE k whether the data included in Ω or C_k are compatible with DE density (2), we apply goodness-of-fit tests for the marginal distributions and test independence of the joint distribution. This testing approach is directly implied by the structure of the DE density (2). (a) If one or several of these tests fail, meaning that the null hypothesis given by DE density (2) is rejected, the distribution of the data within C_k is most likely more complex (and there are sufficient data within DE k for rejecting). Consequently, DE k is considered interim and a split along one or two probability space directions is conducted. Next, the testing and splitting process continues for each of the resulting two or four DEs. On the other hand, (b) if DE k passes all tests or contains no data, the splitting process stops, DE k is final, and becomes one of the m DEs that are part of the DET density estimator (1). During the outlined estimator construction process, a tree is emerging with its root given by the DE on Ω , branches to interim DEs resulting from DE splits, and final DEs at the leafs of the tree. More details about the testing and splitting processes follow in the next few paragraphs.

2.2.1 Splitting

If a split of DE k on cuboid C_k along dimension x_i is conducted, cuboid C_k is split either into subcuboids with equal volume (equal-size split) or subcuboids with approximately equal number of samples (equal-score split). If splits in directions x_i and x_j are to be conducted, we first split along direction x_i and then split each of the two resulting DEs along x_j .

2.2.2 Goodness-of-fit testing

To verify whether the data in DE k are compatible with density (2), we apply a two-stage testing sequence. First, we apply goodness-of-fit tests in each of the d directions x_i to verify whether the marginal distributions of the data are compatible with the null hypothesis given by the marginal DE densities (3) or (4), depending on the DE order. If in one

or several directions the null hypothesis is rejected, we split along the rejecting direction with the smallest p -value. In this work, we focus on Pearson's χ^2 goodness-of-fit test (Pearson 1900) with significance level α_g .

This test was originally developed for categorical data, but is often used for continuous variables as well with the continuous data grouped into classes (e.g., Mann and Wald 1942). We define the classes such that the number of samples in

$$n_c = \min \left[\frac{n}{5}, 4\sqrt{\frac{2(n-1)^2}{c^2}} \right] \quad (6)$$

classes, with $c = \sqrt{2} \operatorname{erfcinv}(2\alpha_g)$, is approximately equal. Here, $\operatorname{erfcinv}$ is the inverse complementary error function. In expression (6), the first contribution $n/5$ is the rule of thumb, stating that each class should contain at least five samples (Cochran 1952, section 7), and the second contribution was proposed by Mann and Wald (1942, theorem 1) and maximizes the power of the χ^2 test under certain conditions. With constant and linear DEs, the χ^2 test is conducted with $n_c - 1$ and $n_c - 2$ degrees of freedom, respectively (composite test).

To investigate the influence of the goodness-of-fit test on the performance of the DET method, we apply in Sect. 3.5 in addition to the previously mentioned χ^2 goodness-of-fit test a Kolmogorov–Smirnov (KS) test (Smirnov 1948). Even though the KS test is a nonparametric test in the sense that no discrete classes have to be prescribed, we favor the χ^2 test for reasons of computational efficiency and convenience in cases where large samples and composite tests are involved, respectively.

2.2.3 Independence testing

Second, if $d > 1$ and all tests concerning the marginal distributions are passed without any split, we verify whether the components of the data in DE k are pairwise independent and thus approximately compatible with DE distribution (1). We point out that this approach is approximate, since (mutually) independent random variables, as represented by distribution (1), are pairwise independent, but for $d > 2$ pairwise independence does generally not imply mutual independence (Papoulis 1991, p. 184). If not stated otherwise, we apply Pearson's χ^2 independence test (Pearson 1900) with significance level α_d for pairs of components x_i and x_j with $i = 1, 2, \dots, d-1$ and $j = i+1, i+2, \dots, d$. We use contingency tables with classes such that the samples are equally distributed among $\sqrt{n_c}$ classes in the x_i - and x_j -direction (Bagnato et al. 2012, section 3.1). Moreover, a χ^2 distribution with $(\sqrt{n_c} - 1)^2$ degrees of freedom is applied.

To inspect the role of the independence test statistic on the performance of the DET method, we apply in Sect. 3.5

for comparison an independence test based on Kendall's τ (Kendall 1938). However, for reasons of computational efficiency in high-dimensional cases, we favor the χ^2 independence test.

If for one or several component pairs the independence null hypothesis is rejected, DE splitting is done along the components of the rejecting pair with the smallest p -value. Within component pairs, the component with smallest goodness-of-fit p -value is split first.

2.2.4 Computational cost

The outlined approximate testing sequence avoids the initiation of simultaneous splits in more than two directions, and therefore, the DET estimator is not subject to the exponential growth of bins in terms of d . At each DE in the tree, (a) d goodness-of-fit tests are conducted and if $d > 1$ and no split was induced by these tests, (b) of the order of $d^2/2$ pairwise independence tests follow. This testing sequence is applied at each tree node, that is each interim and final DE. In the case of score-based splitting, an upper bound for the number of interim and final DEs, that determines the DET construction cost, exists. If each final DE contains just one sample, the maximal possible number of final DEs $m = n$. Moreover, if n can be expressed as an integer-valued power of two, the number of all interim and final DEs is equal to

$$1 + 2 + 4 + 8 + \dots + n/2 + n = 2n - 1. \quad (7)$$

Here, the sum starts with 1 representing the root DE and after multiple generations of binary splits ends with n final DEs. If n cannot be written as an integer power of two, the next larger integer that is a power of two is $n' = 2^{\lceil \log_2(n) \rceil}$, which satisfies $n < n' < 2n$. Then based on result (7), an upper bound for the number of all interim and final DEs is given by $2n' - 1 < 4n - 1$. To summarize, when applying equal-score splits, the computational effort for the DET construction scales in the worst case linearly with the number of samples n .

To evaluate the density estimate (1) at a certain location \mathbf{x} , i.e., $p(\mathbf{x})$, the sum (1) involving m final DEs could be evaluated. A more efficient approach, however, exploits the tree structure of the DET estimator: The only nonzero term in sum (1) can be quickly identified by starting at the root DE and by sequentially identifying at the DET forks the interim DE l where $\mathbf{x} \in C_l$. After of the order of n_t decision operations, where n_t is the number of DE splits or tree depth, the leaf or final DE k is reached that contains point \mathbf{x} . Typically $n_t \ll m < n$, which renders the DET estimator computationally efficient for density queries. In the next few sections, we document the capabilities of the DET density estimator for a range of test cases.

3 Density estimation with the DET method

For the following computations, if not mentioned otherwise, significance levels $\alpha_g = \alpha_d = 0.001$ were applied. All computations were carried out on a state-of-the-art laptop computer with a 2.8GHz Intel Core i7 processor. We inspect the accuracy of the DET estimator for a diverse set of one-, two-, four-, and seven-dimensional cases in the following three sections. To this end, we estimate the MISE defined as

$$\left\langle \int_{\Omega} [\hat{p}(\mathbf{x}) - p(\mathbf{x})]^2 d\mathbf{x} \right\rangle, \quad (8)$$

where $\hat{p}(\mathbf{x})$ with $\mathbf{x} \in \Omega$, given for example by the DET method (1), is an estimate of the exact PDF $p(\mathbf{x})$ based on an ensemble of samples and angular brackets represent the expectation with respect to that ensemble. To put our results into perspective with state-of-the-art density estimation, we compare against the adaptive KDE method of Botev et al. (2010), the density estimation tree or in short density-tree method of Ram and Gray (2011), the LL-OPT estimator of Jiang et al. (2016), and finally a conventional histogram, where the bin width in each dimension was determined based on the normal reference rule [Scott 2015, equation (3.66)].

To numerically evaluate the integration over probability space Ω in the MISE (8), we applied for $d = 1$ and 2 trapezoidal rules with equidistant Cartesian grids. As an exception to this, an exponentially stretched grid was applied for the one-dimensional gamma PDF example (following in Sect. 3.1.6) to better resolve the region of high probability near the origin. To assert the accuracy of the resulting MISE, adaptive quadrature methods as outlined by Shampine (2008a, b) were applied as well, leading to virtually the same results as with trapezoidal rules. In connection with the DET estimator, here the integration region Ω was decomposed among all DEs and adaptive quadrature was applied in each element individually. For the four- and seven-dimensional cases, quadrature or trapezoidal rules become too expensive and we resorted to Monte Carlo (MC) integration instead. To this end, we rewrite the MISE (8) as

$$\begin{aligned} & \left\langle \int_{\Omega} \left[\frac{\hat{p}(\mathbf{x})^2}{p(\mathbf{x})} - 2\hat{p}(\mathbf{x}) + p(\mathbf{x}) \right] p(\mathbf{x}) d\mathbf{x} \right\rangle \\ &= \left\langle \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n \left[\frac{\hat{p}(\mathbf{x}_j)^2}{p(\mathbf{x}_j)} - 2\hat{p}(\mathbf{x}_j) + p(\mathbf{x}_j) \right] \right\rangle, \end{aligned} \quad (9)$$

with samples \mathbf{x}_j distributed as prescribed by PDF $p(\mathbf{x})$.

Computationally efficient spectral KDE implementations for $d = 1$ and 2, that operate on equidistant Cartesian grids, are available (Botev 2007). Since these implementations become prohibitive for $d > 2$, an alternative was recently provided by Botev (2016). However, here similar to conventional KDE the computational effort per density evaluation

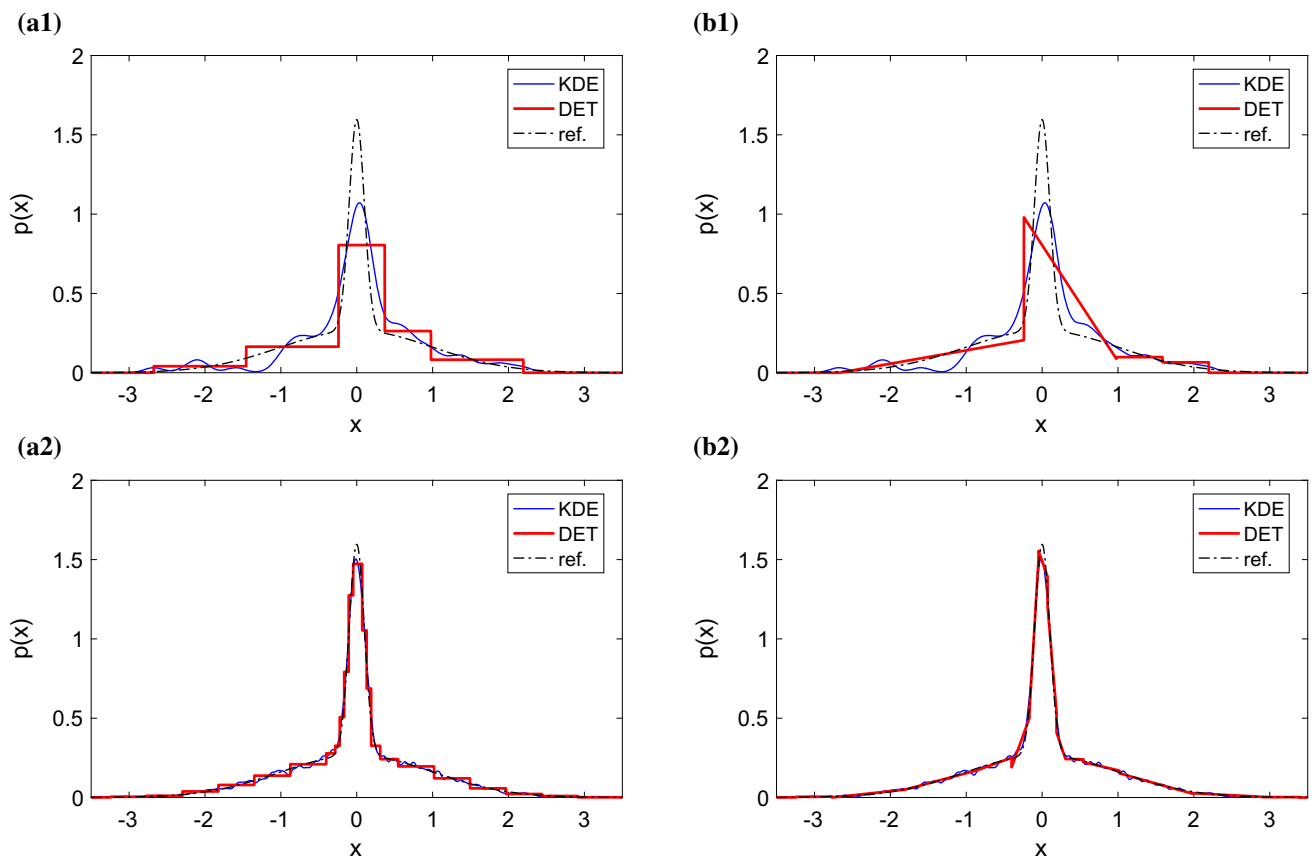


Fig. 1 PDF estimates resulting from (blue thin solid) adaptive KDE and (red thick solid) the size-split DET method with particle ensembles including (1) $n = 100$ and (2) 10^4 samples are compared with (black

dash dot) the reference PDF (10). In panels (a) and (b), DET estimates with constant and linear elements are depicted, respectively

depends on n . A density-tree implementation is available via the MLPACK library (Curtin et al. 2013). For the LL-OPT computations, the implementation cited by Jiang et al. (2016, section 7) was applied. We used the outlined implementations with prescribed standard parameter values.

In the following one- and two-dimensional examples, MISE ensemble averages $\langle \dots \rangle$ were approximated based on 50 samples. Moreover, for the DET method, based on these samples, average number of DEs $\langle m \rangle$ and tree depths $\langle n_t \rangle$ are reported. Here, n_t for one DET sample corresponds to the maximal depth within the tree.

We start our assessment of the DET method, by comparing in Sects. 3.1 and 3.2 the different DET variants to the established KDE method of Botev et al. (2010). In a next set of comparisons contained in Sect. 3.3, we focus on linear DETs with equal-size splits and compare with density trees, LL-OPTs, and histograms.

3.1 One-dimensional examples

If not stated otherwise, for the following one-dimensional adaptive KDE, a grid with $2^{14} = 16,384$ nodes or cosine

modes was deployed (Botev 2007). The following one-dimensional examples involving normal mixture densities were taken from (Marron and Wand 1992).

3.1.1 Kurtotic unimodal PDF

First, we inspect the performance of the DET estimator for the kurtotic unimodal distribution no. 4 in their paper, i.e.,

$$p(x) = \frac{2}{3}N(0, 1) + \frac{1}{3}N(0, 1/10), \quad (10)$$

where $N(\mu, \sigma)$ is the normal distribution with mean μ and standard deviation σ . Exemplary estimates resulting from the DET method and KDE are depicted together with expression (10) in Fig. 1. The DET method adaptively allocates finer bins in the region where part $N(0, 1/10)$ dominates and resorts to larger DEs to represent $N(0, 1)$. The corresponding MISEs are depicted in Fig. 2a as a function of the number of samples n . A power-law decay with exponents between $-1/2$ and -1 is observed for all DET estimators and the adaptive KDE. The linear DET methods and adaptive KDE

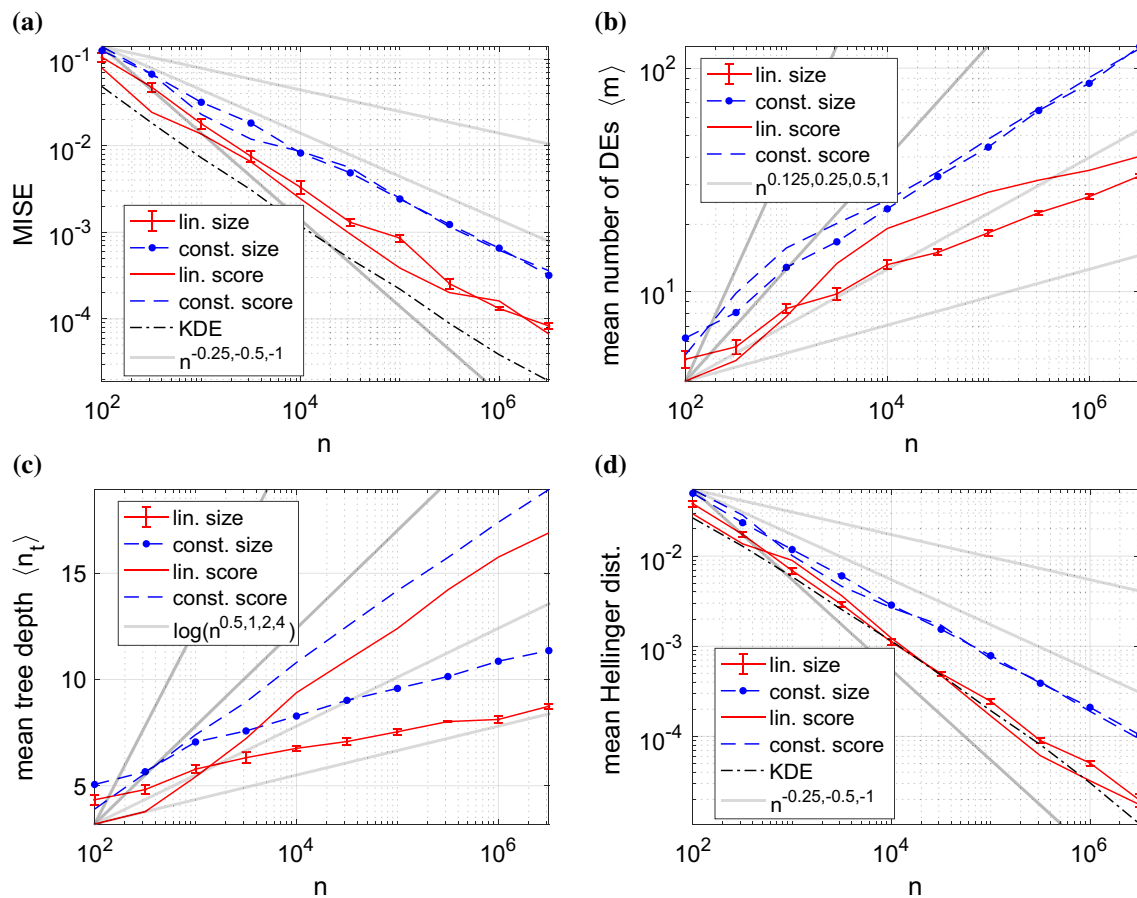


Fig. 2 Estimation of kurtotic unimodal PDF (10). Evolutions of **a** the MISE and **d** the mean Hellinger distance as a function of the number of samples n for DET estimators (red solid and blue dashed) with equal-size (symbols) and -score splits (no symbols) are plotted. The MISE and mean Hellinger distance resulting from the adaptive KDE is included as well (black dash dot). For the DET estimators the resulting mean

numbers of DEs $\langle m \rangle$ and mean tree depths $\langle n_t \rangle$ are given in panels (b) and (c), respectively. Power-law and logarithmic scalings with exponents indicated in the figure legends are depicted (gray thick solid). Exemplary 95% confidence intervals are provided for the linear DET estimator with size splits

share approximately the same decay rate, with KDE having a smaller MISE than the DET variants.

Besides the inferior MISE decay rate, constant DEs require more DEs as is seen in Fig. 2b. Here, the mean number of constant and linear DEs grows approximately with $n^{1/4}$ and $n^{1/8}$ for large n . The number of DEs $\langle m \rangle$ determines the computational cost associated with the DET construction. These growth rates are independent of the splitting method. While the DET splitting method has little effect on $\langle m \rangle$ and the MISE, the tree growth, measured by $\langle n_t \rangle$, is smaller for size- versus score-based splitting (see Fig. 2c). However, in both cases, $\langle n_t \rangle$ increases logarithmically with n to a good approximation.

With m and n_t being discrete variables and the DE splitting being a discrete process, the DET datasets in Fig. 2 are not as smooth as, for example, the MISE decay in the adaptive KDE. Deviations from regular scalings in the present, and more so in the next examples, are especially apparent

for small n , where the DETs are comprised of few splitting levels and DEs. In Fig. 2, exemplary 95% confidence intervals (given a Gaussian likelihood model) illustrate that the ensembles of density estimator realizations used in this work are sufficiently large to keep statistical uncertainties small.

3.1.2 Outlier PDF

In a next step, we study the DET estimator for the outlier distribution no. 5 from (Marron and Wand 1992) given by

$$p(x) = \frac{1}{10}N(0, 1) + \frac{9}{10}N(0, 1/10). \quad (11)$$

The PDF estimates and resulting MISE are plotted in Fig. 3. Except for the slightly worse MISE of the DET variants compared with KDE, all previous observations from the kurtotic unimodal distribution (10) carry over to the present case.

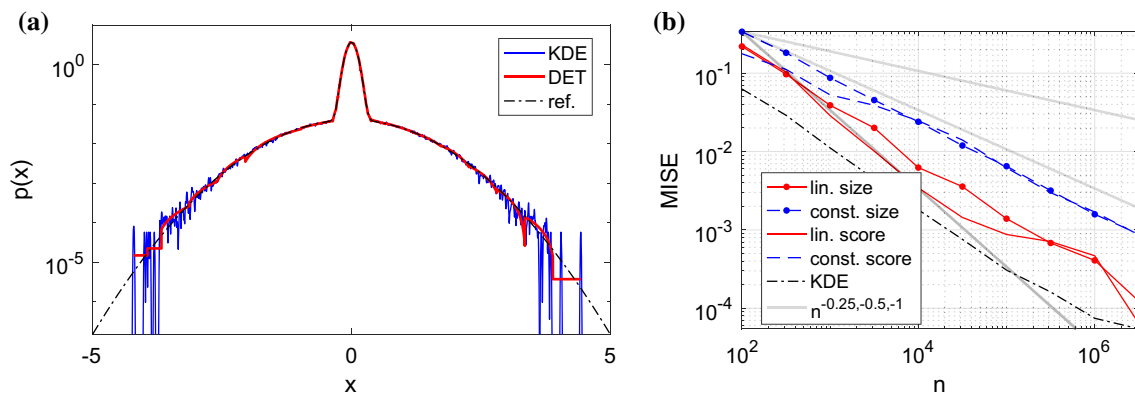


Fig. 3 **a** PDF estimates resulting from (blue thin solid) adaptive KDE and (red thick solid) the size-split linear DET method based on a particle ensembles with $n = 10^6$ samples are compared with (black dash

dot) the reference PDF (11). The corresponding MISE evolution as a function of the number of samples n is shown in panel (b) like in Fig. 2a

3.1.3 Asymmetric claw PDF

The third test distribution is the asymmetric claw distribution no. 12 (Marron and Wand 1992), i.e.,

$$p(x) = \frac{1}{2}N(0, 1) + \sum_{l=-2}^2 \frac{2^{1-l}}{31} N\left(l + \frac{1}{2}, 2^{-l}/10\right). \quad (12)$$

Estimates of this PDF based on the size-based DET estimators and adaptive KDE are included in Fig. 4a and illustrate the adaptive bin width selection of the DET method based on the local sample density. The MISE decay given in panel (b) is slightly closer to the adaptive KDE compared with the previous cases. Again a logarithmic growth of the tree depth as a function of the ensemble size n is recovered with similar depths like in the previous examples as seen in Fig. 2c.

3.1.4 Spiky uniforms PDF

The previous test cases have focused on Gaussian mixtures. The next three examples deal with different PDFs like the spiky uniforms distribution taken from (Wong and Ma 2010, example 6) and given by

$$p(x) = \frac{1}{2}U(0.23, 0.232) + \frac{1}{2}U(0.233, 0.235), \quad (13)$$

where $U(a, b)$ is a uniform distribution defined on the interval $[a, b]$. DET-based density estimates and adaptive KDE are compared in Fig. 5. We can observe that while the DET methods tend to produce oscillations at the bounds of the uniforms, the adaptive KDE is subject to noise apparent in the constant sections of the uniforms. This difficulty is detected as well by the MISE plotted in Fig. 6a, where the DET methods converge faster with approximately $1/n$ to the true density compared to adaptive KDE with $1/\sqrt{n}$. The DET methods become more accurate than KDE for ensembles with

$n > 10^3$. The mean tree depth of the DET method shown in Fig. 6b and the mean number of DEs (not shown) display logarithmic and sublinear growth as observed previously. As an exception to all one-dimensional cases considered, for the spiky uniform distribution, there is no advantage in terms of MISE convergence, number of DEs, and tree depth when using linear vs. constant DEs. Given the piecewise constant form of the PDF, this is expected.

3.1.5 Beta PDF

In our fifth example, we perform density estimation based on data derived from the beta PDF

$$p(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad (14)$$

with $x \in [0, 1]$, parameters $\alpha = 1.05$ and $\beta = 0.8$, and where $\Gamma(x)$ is the gamma function. Exemplary resulting estimates are depicted in Fig. 7. While in the flat region of PDF (14) the constant DET estimator allocates few large elements, increasingly fine bins are placed toward the peak at $x = 1$. The linear DET estimator behaves similarly but requires fewer elements. In comparison with adaptive KDE, one can observe in Fig. 8a that the DET methods converge faster and that the linear DET variants are more accurate. Unlike in the previous examples, it becomes apparent from Fig. 8b that the score-split DET variants lead to smaller trees than their size-split counterparts.

3.1.6 Gamma PDF

The last one-dimensional example considered in this work is the gamma PDF

$$p(x) = \frac{x^{a-1} e^{-x/b}}{b^a \Gamma(a)} \quad (15)$$

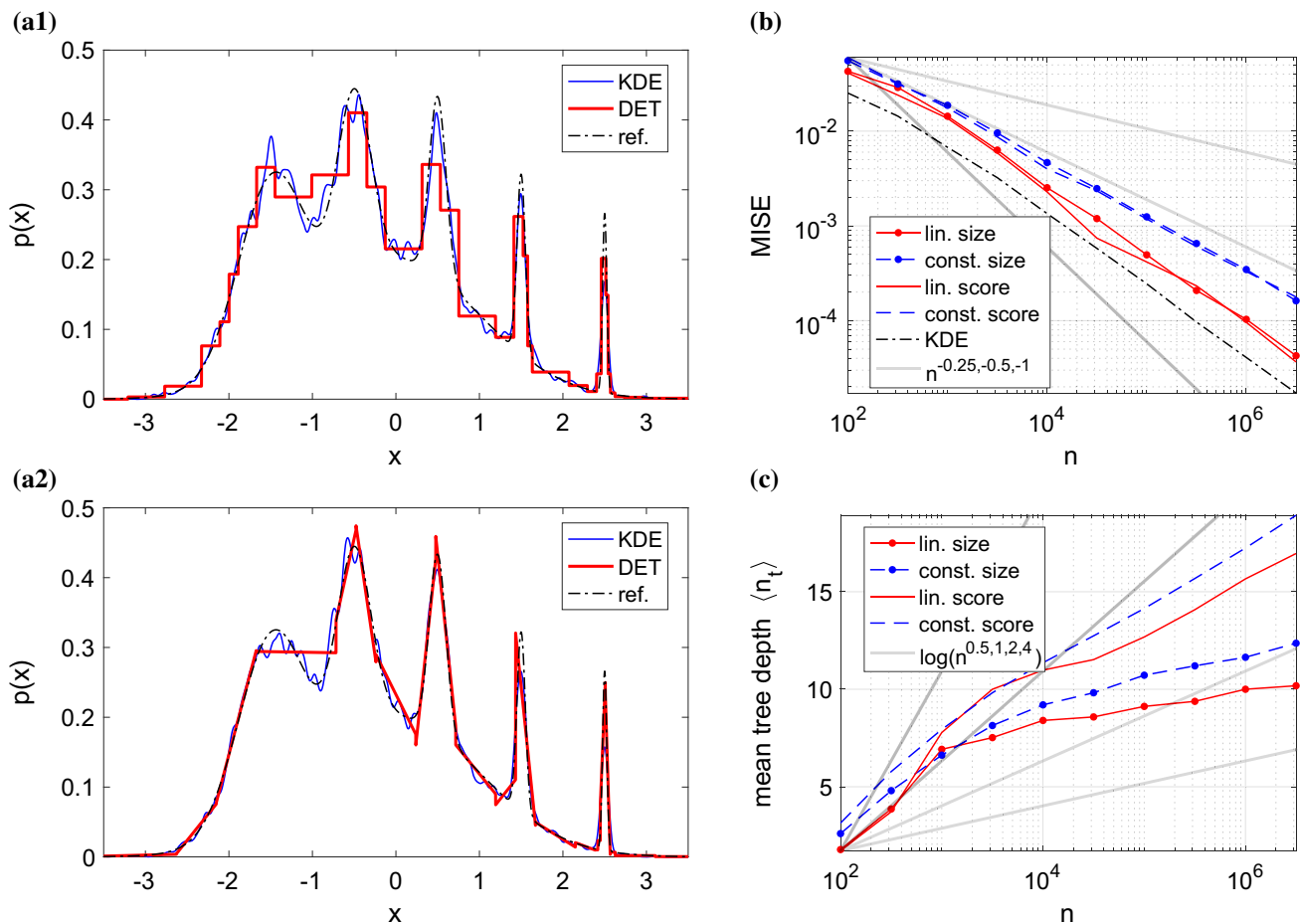


Fig. 4 a PDF estimates resulting from (blue thin solid) adaptive KDE and (red thick solid) size-split (1) constant and (2) linear DETs based on a particle ensembles with $n = 10^4$ samples are compared with (black

dash dot) the reference PDF (12). The corresponding MISE evolution and mean tree depth as a function of the number of samples n are shown in panels (b) and (c) like in Fig. 2

with x going from 0 to ∞ and parameters $a = 2/3$ and $b = 50$. As is visible from Fig. 9, this PDF is skewed and goes to infinity for $x \rightarrow 0$. The MISE results included in the figure document the accuracy of the DET estimators. The ability of the adaptive KDE to resolve the singularity of PDF (15) becomes for $n \approx 10^{4.5}$ limited by the number of grid points or cosine modes applied for its calculation. Increasing this number from 2^{14} to $2^{17} = 131,072$ shifts the resolution limit to larger n as is shown in Fig. 9b. For the gamma PDF as for the beta PDF, score-based splitting is slightly more effective than size-based splitting leading to smaller trees with fewer DEs.

3.2 Two-dimensional examples

For the adaptive KDEs in the following two-dimensional cases, a grid with $2^{10} \times 2^{10} = 1,048,576$ nodes was deployed (Botev 2007).

3.2.1 Bi-variate Gaussian PDF

Before looking at more complex bi-variate distributions, we consider the joint Gaussian PDF

$$p(\mathbf{x}) = \frac{\exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{C}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]}{\sqrt{(2\pi)^2 \det(\mathbf{C})}} \quad (16)$$

with unbound probability space $\mathbf{x} = (x_1, x_2)^\top$, mean vector $\boldsymbol{\mu} = (0, 0)^\top$, and covariance matrix

$$\mathbf{C} = \begin{pmatrix} 4.0 & -2.28 \\ -2.28 & 1.44 \end{pmatrix}.$$

Gaussian PDFs are important in many applications, which emphasizes the present case. Resulting size-split linear DET estimates are compared with KDE and joint PDF (16) in Fig. 10. The adaptivity of the DET method is illustrated in panel (a), where regions of small density variation are represented by large DEs and sections with high varia-

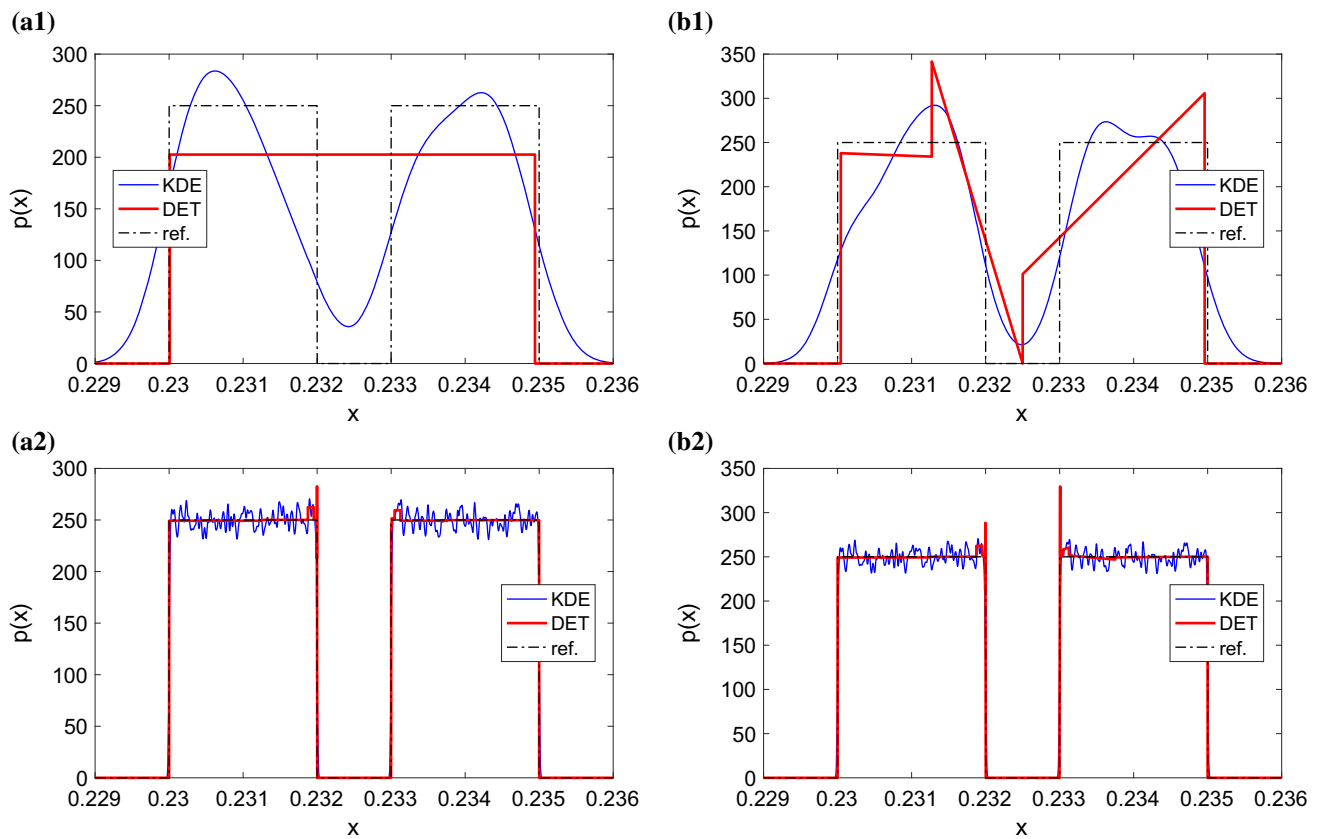


Fig. 5 PDF estimates resulting from (blue thin solid) adaptive KDE and (red thick solid) the size-split DET method with particle ensembles including (1) $n = 100$ and (2) 10^5 samples are compared with (black

dash dot) the reference PDF (13). In panels (a) and (b), DET estimates with constant and linear elements are depicted, respectively

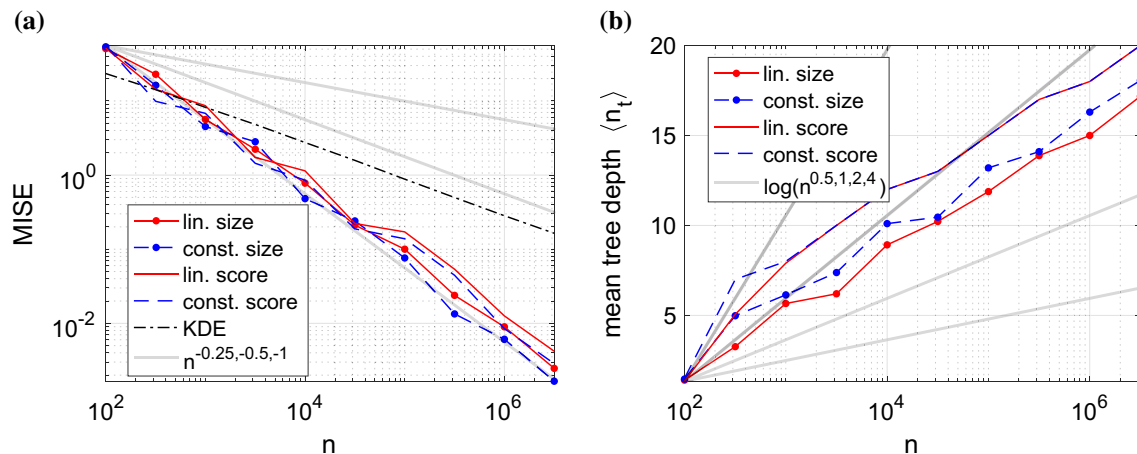


Fig. 6 Estimation of spiky uniform PDF (13). **a** Evolutions of the MISE as a function of the number of samples n for DET estimators (red solid and blue dashed) with equal-size (symbols) and -score splits (no symbols) are plotted. The MISE resulting from the adaptive KDE is

included as well (black dash dot). For the DET estimators, the resulting mean tree depths $\langle n_t \rangle$ are given in panel (b). Power-law and logarithmic scalings with exponents indicated in the figure legends are depicted (gray thick solid)

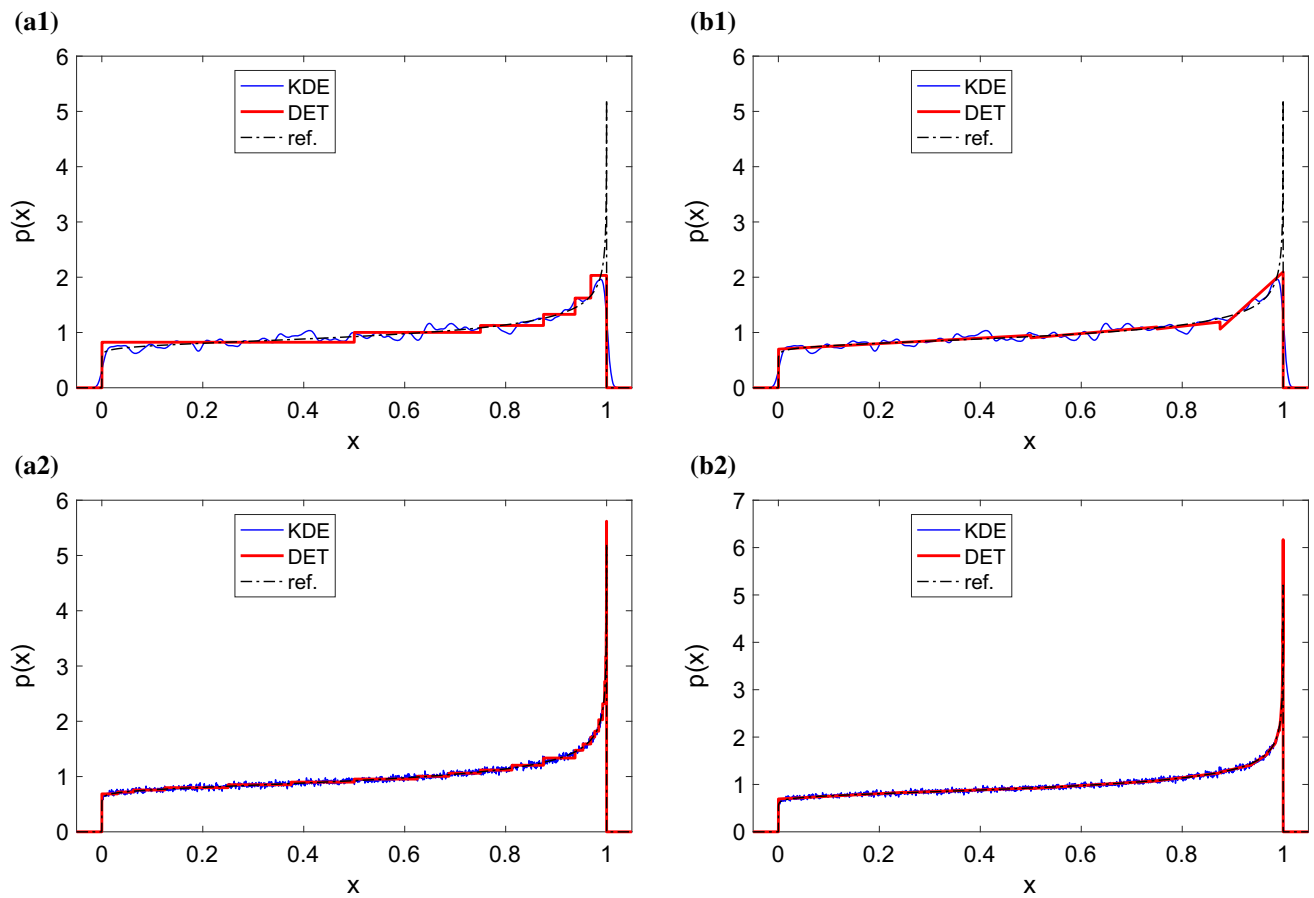


Fig. 7 PDF estimates resulting from (blue thin solid) adaptive KDE and (red thick solid) the size-split DET method with particle ensembles including (1) $n = 10^4$ and (2) 10^6 samples are compared with (black

dash dot) the gamma PDF (14). In panels (a) and (b), DET estimates with constant and linear elements are depicted, respectively

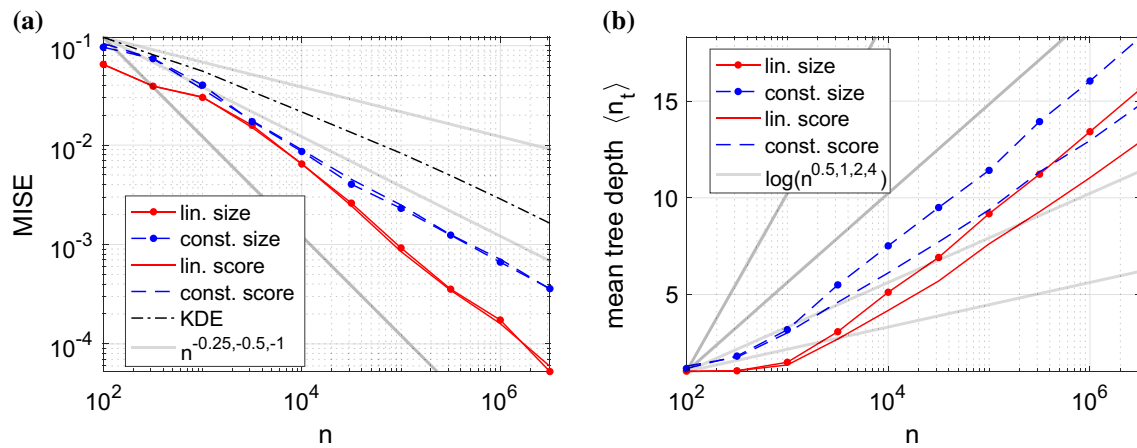


Fig. 8 Estimation of beta PDF (14). See Fig. 6

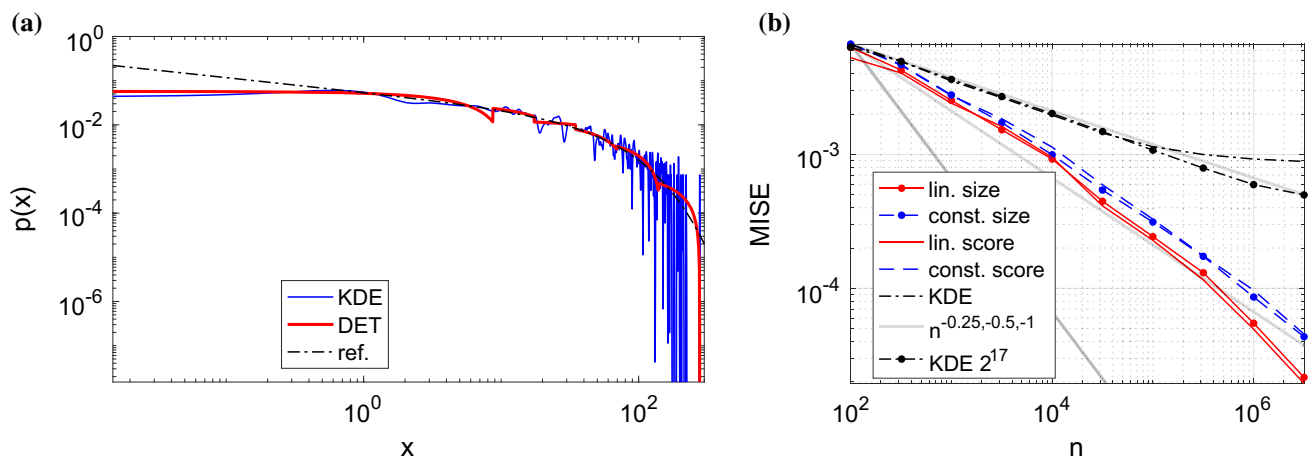
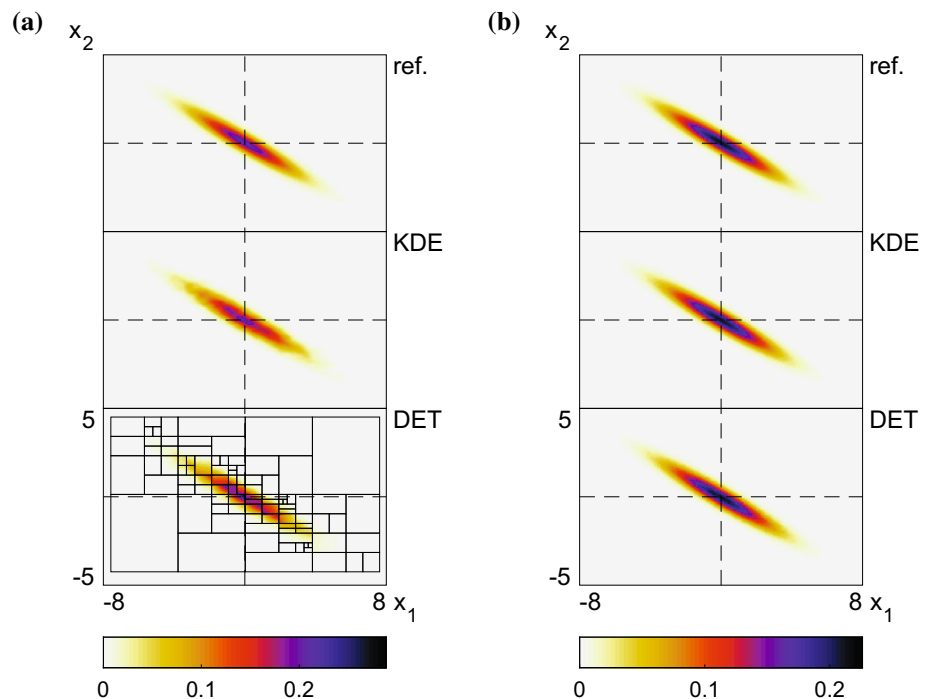


Fig. 9 **a** PDF estimates resulting from (blue thin solid) adaptive KDE and (red thick solid) the size-split linear DET method based on particle ensembles with $n = 10^3$ samples are compared with (black dash dot) the reference PDF (15). The corresponding MISE evolution as a

function of the number of samples n is shown in panel (b) like in Fig. 2a. In addition to the adaptive KDE with 2^{14} grid points (black dash dot), MISE results based on n^{17} points (black dash dot with symbol) are provided

Fig. 10 PDF estimates resulting from adaptive KDE and the size-split linear DET method based on particle ensembles with **a** $n = 10^4$ and **b** $n \approx 10^{6.5}$ samples are compared with the reference PDF (16)



tion were subdivided into several smaller elements.² The decay of the MISE depicted in Fig. 11a is similar to the one-dimensional Gaussian mixture cases, but the convergence rates are slightly smaller (compare, for example, with Fig. 2). Again, adaptive KDE is more accurate than the DET variants and the DE tree growth is logarithmic as seen in Fig. 11c. Moreover for large n , the mean number

of DEs increases sublinearly with approximately $n^{1/4}$ (see Fig. 11b).

Figure 10a reveals a weakness of the present DET method in comparison with KDE, as the DET estimators or DET parametrizations will depend on the orientation of the coordinate system. In case of alignment of the joint Gaussian PDF (16) with the x_1 - x_2 -coordinate system, the DET estimators will require fewer DEs and will be more accurate. This implies the use of a covariance-matrix-based principle axes transform [e.g., Silverman 1998, equation (4.7)], to achieve

² In the singular case of linearly dependent components x_1 and x_2 , many small DEs, resolving the probability peak along the diagonal of the x_1 - x_2 -space, result from a DET estimator.

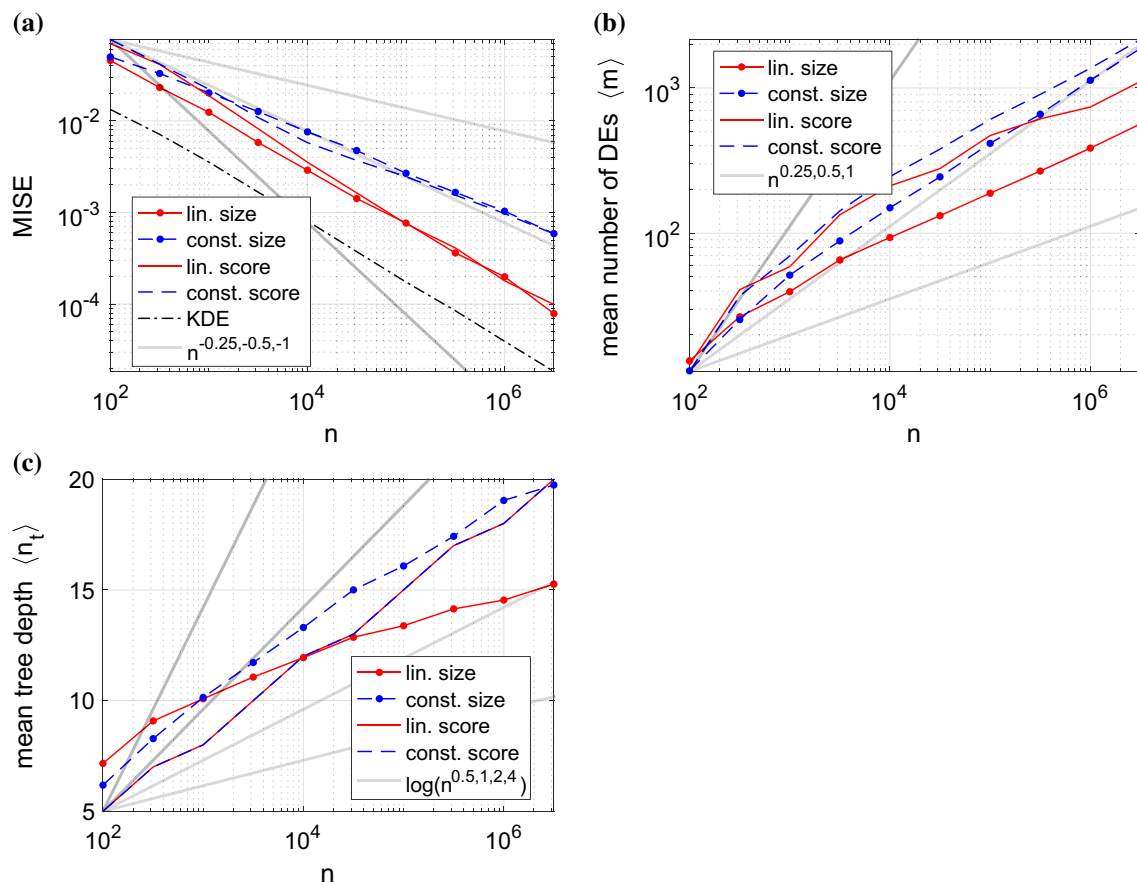


Fig. 11 Estimation of bi-variate Gaussian PDF (16). See Fig. 2. In panel (c), the data series from the score-split DETs coincide

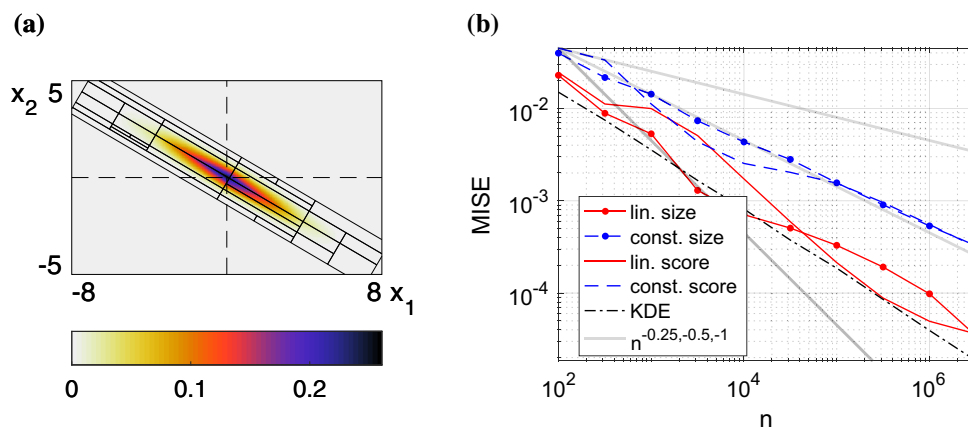


Fig. 12 Estimation of bi-variate Gaussian PDF (16). **a** PDF estimate resulting from the linear size-split DET method combined with a principle axes transform and based on a particle ensemble with $n = 10^4$

samples. **b** Evolutions of the corresponding MISE for the indicated DET estimators and KDE as a function of the number of samples n . All DET variants were combined with a principle axes transform

parametrization invariance and increased accuracy and computational efficiency. An illustration is given in Fig. 12, where results from DET estimators combined with principle axes

transforms are provided. However, in order to demonstrate the versatility of the DET method, this approach is not pursued further in the present work.

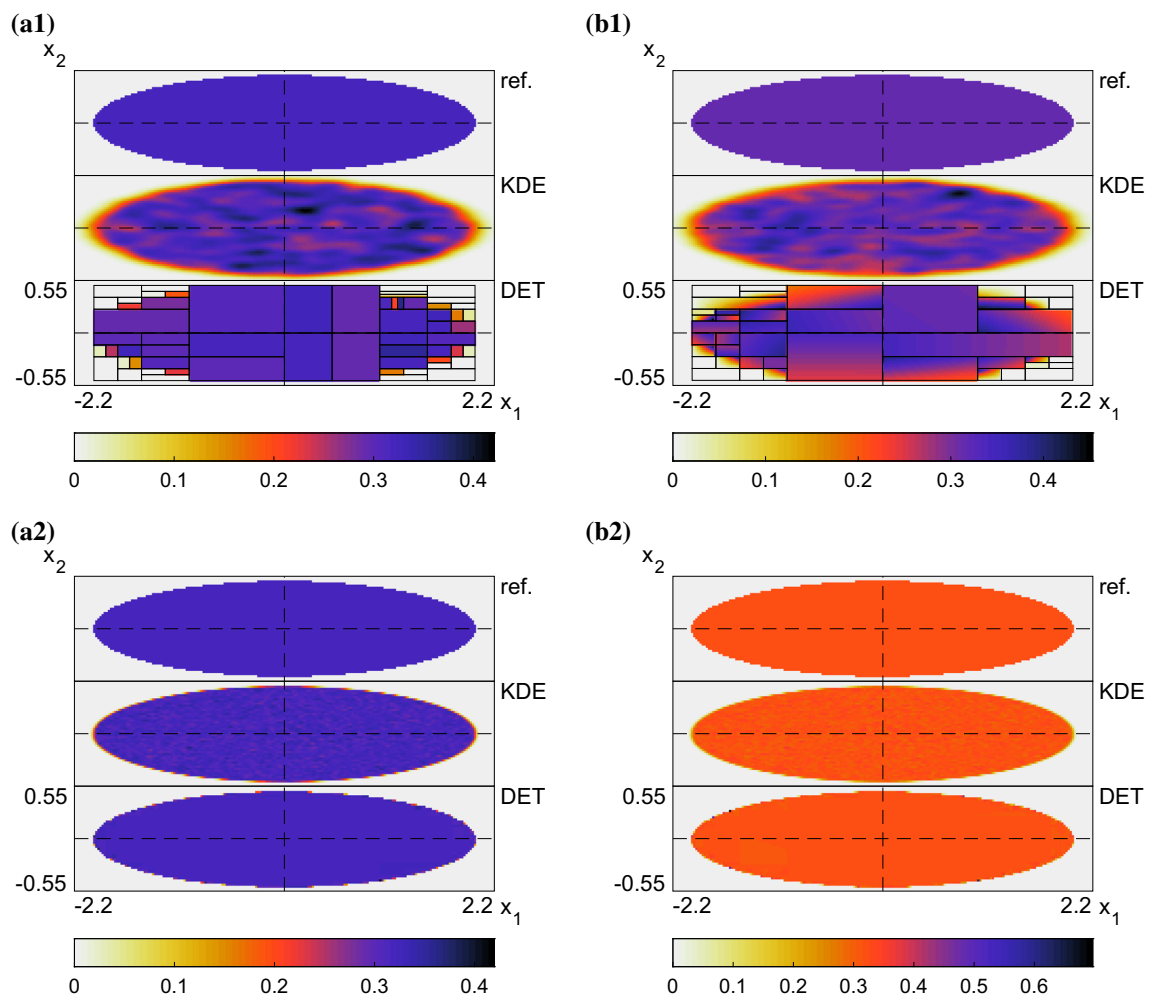


Fig. 13 PDF estimates resulting from adaptive KDE and size-split DET methods with **a** constant and **b** linear DEs based on particle ensembles with (1) $n = 10^4$ and (2) $n \approx 10^{6.5}$ samples are compared with the reference PDF (17)

3.2.2 Uniform on an ellipse

Next, the two-dimensional case with uniform density on an ellipse outlined in (Botev et al. 2010, p. 2942) is inspected. The PDF is given by

$$p(\mathbf{x}) = \begin{cases} 1/\pi & \forall \mathbf{x} \in \{\mathbf{x} | x_1^2 + (4x_2)^2 \leq 4\} \\ 0 & \text{otherwise} \end{cases} \quad (17)$$

and in Fig. 13 KDE and DET-based estimation results based on ensembles stemming from this PDF are provided. Like in the one-dimensional case with uniform PDFs (see Fig. 5a2, b2), constant DEs lead to smaller oscillations at the interface, where PDF (17) switches from 0 to $1/\pi$ (compare ranges of color bars in panels (1) and (2) of Fig. 13). However, while the adaptive KDE displays a noisy density estimate within the ellipse, the DET variants capture the constant density quite accurately. This is reflected in the MISE results shown in Fig. 14a, where for $n > 10^5$

the DET estimators converge faster to the true PDF and become more accurate compared with adaptive KDE. Given the piecewise constant PDF (17), linear DEs are as good as constant elements and like the MISE, the mean number of DEs $\langle m \rangle$ shown in Fig. 14b and the mean tree depth $\langle n_t \rangle$ (not shown) behave similarly for both DE types. For large n , $\langle m \rangle$ increases approximately as \sqrt{n} and $\langle n_t \rangle$ (not shown) grows similarly as in the one-dimensional spiky uniforms case (see Fig. 6b).

One might argue that—in the absence of a principle axes transform—the present setup, with the ellipse aligned with the coordinate system, is in favor of the DET method. In an additional study, the performance of the DET method was inspected for the ellipse rotated by an angle of $\pi/4$. With this modified setup, similar results were found, with the DET variants surpassing KDE at $n \approx 2 \times 10^5$ and with $\langle m \rangle$ starting from 6 elements at $n = 100$ and growing to similar numbers as in the unrotated case.

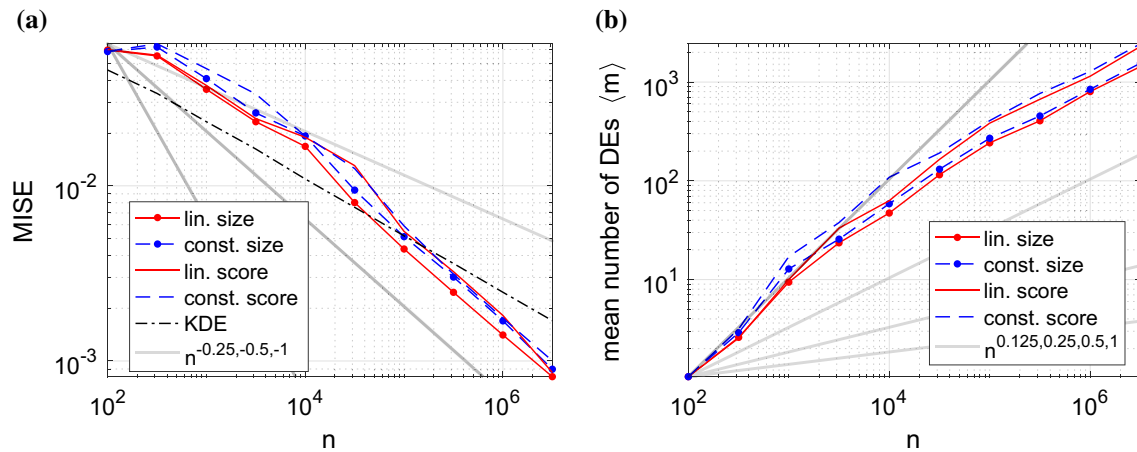


Fig. 14 Estimation of the bi-variate uniform PDF (17). **a** Evolutions of the MISE as a function of the number of samples n for DET estimators (red solid and blue dashed) with equal-size (symbols) and -score splits (no symbols) are plotted. The MISE resulting from the adaptive KDE is

included as well (black dash dot). For the DET estimators, the resulting mean number of DEs $\langle m \rangle$ is given in panel (b). Power-law scalings with exponents indicated in the figure legends are depicted (gray thick solid)

3.2.3 Bi-variate Dirichlet PDF

The third bi-variate case addressed in this work is the Dirichlet PDF given by

$$p(\mathbf{x}) = \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} x_1^{\alpha_1-1} x_2^{\alpha_2-1} (1 - x_1 - x_2)^{\alpha_3-1} \quad (18)$$

$\forall \mathbf{x} \in \{\mathbf{x} | x_1 + x_2 \leq 1 \vee x_1 \geq 0 \vee x_2 \geq 0\}$ and $= 0$ otherwise, with parameters $\alpha_1 = 0.9$, $\alpha_2 = 1.5$, and $\alpha_3 = 3$. The Dirichlet PDF (18) is a bi-variate generalization of the univariate beta PDF (14). PDF (18) is depicted together with KDE results and DET estimates for two differently sized ensembles in Fig. 15. The DET method provides for both ensembles estimates that are in good agreement with the true density. MISE results are compared in Fig. 16a and it is seen that like in the beta PDF example, the linear DET variants are most accurate, followed by their constant counterparts and adaptive KDE. As seen in Fig. 16b, the size-split DET methods produce the smallest trees with the linear DET approach being the most efficient. The mean number of DEs $\langle m \rangle$ (not shown) increases with n like in the bi-variate Gaussian case. The linear DEs and size-based splitting have the smallest $\langle m \rangle$, and all DET variants scale approximately with $n^{1/4}$ for n large (see Fig. 11b).

Rotating the bi-variate Dirichlet PDF by $\pi/4$ has a bigger impact on the DET performance as in the previously discussed ellipse case. While the MISE error decay rates remain similar, the MISEs of the DET variants are larger than KDE for small n and surpass KDE for n between 10^4 and 10^5 .

In addition to the previous MISE-based performance analyses, we have evaluated the accuracy of the DET estimator

with Hellinger (e.g., Wang and Wang 2015) and total variation distances (Shorack 2000, p. 543). These metrics are based on differences of square roots of densities and absolute density differences, respectively, and therefore penalize strong deviations less than the MISE. Since occasional outliers are more likely in the DET estimator and regular small noise is an issue in the adaptive KDE method (see for example Fig. 5b2), the DET estimator was found to perform better than KDE in terms of Hellinger distance metric as opposed to MISE [see exemplary comparison in Fig. 2, panels (a) and (d)]. In terms of total variation distance, the relative performance among the estimators and convergence trends were found to be overall similar to the MISE results.

3.3 Comparisons histogram and tree-based methods

In the previous two sections, we found that the linear DET variant is most accurate with little difference between size- and score-based splitting. Accordingly, we proceed in this and the next sections by focusing on linear DETs with size-based splits. In Fig. 17, a summary of MISE decay curves resulting from the histogram, LL-OPT, density tree, and linear DET estimators for all one- and two-dimensional examples is provided. Due to the rapidly growing computational costs associated with the density-tree estimator, we stop after $n = 31,623 \approx 10^{4.5}$. The computing time for one density-tree estimate of, e.g., the beta PDF example with 10^5 samples took around 10 minutes. Fig. 17 is accompanied by exemplary PDF estimates included in Fig. 18. These estimates resulted from ensembles with $n = 3163$ samples.

One general observation from the MISE curves in Fig. 17 is the good accuracy of the DET estimator. While histograms

Fig. 15 PDF estimates resulting from adaptive KDE and the size-split linear DET method based on particle ensembles with **a** $n = 10^4$ and **b** $n \approx 10^{6.5}$ samples are compared with the reference PDF (18)

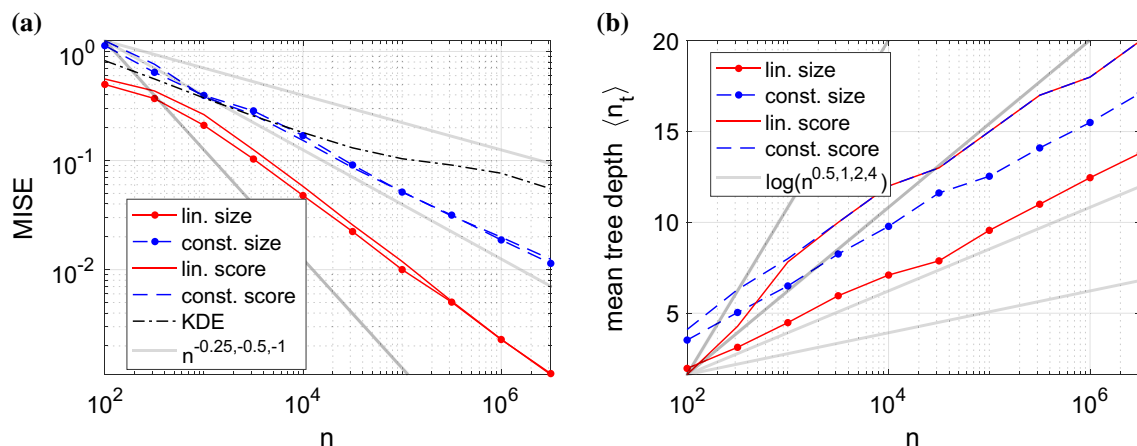
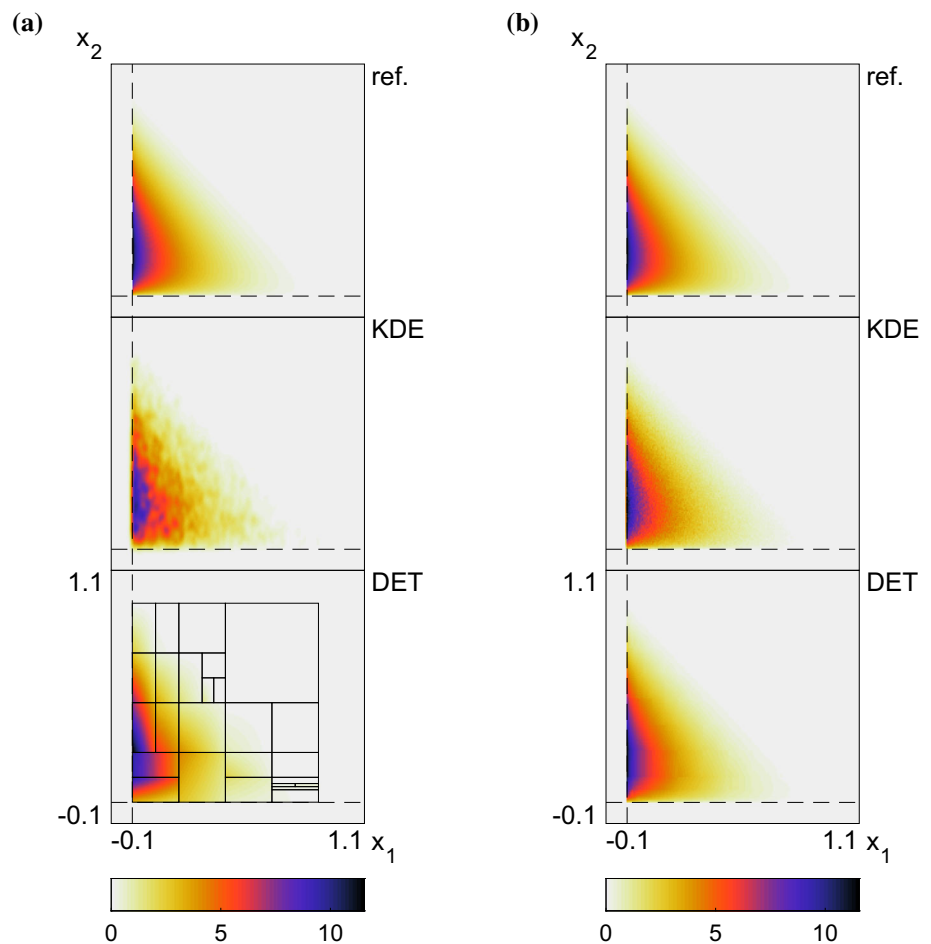


Fig. 16 Estimation of Dirichlet PDF (18). See Fig. 6

with bin widths determined by the normal reference rule are reasonably accurate in the Gaussian mixture cases, they are inaccurate in the other cases. The density-tree estimator leads to mixed results and is most accurate for the examples involving uniform distributions (see Fig. 18b). A MISE

convergence that comes quite close to the DET estimator is resulting from the LL-OPT method. Notable is finally a reduction in MISE decay rates when going from one to two dimensions as seen in Fig. 17. This is an indication of the curse of dimensionality mentioned in the introduction.

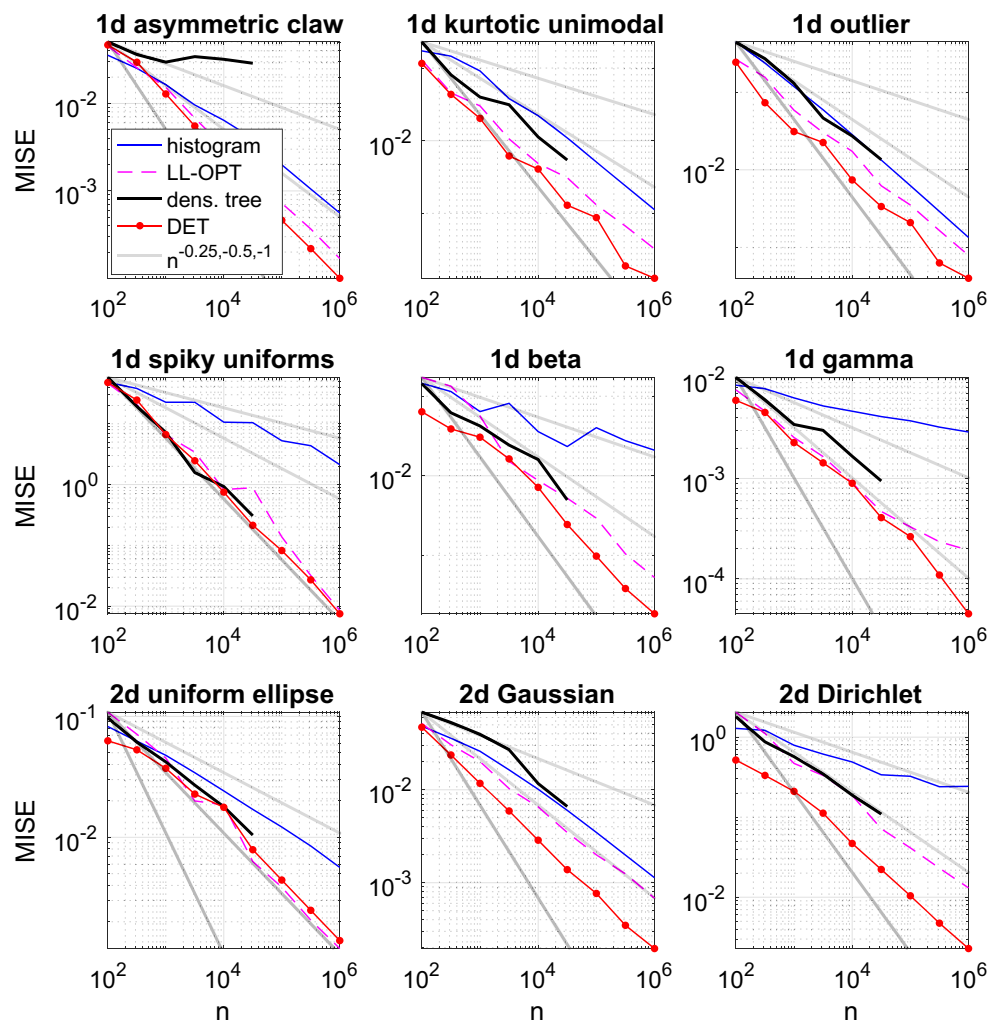


Fig. 17 MISE decay as a function of the number of samples n for the indicated one- and two-dimensional cases. Results from (blue solid) histogram, (magenta dashed) LL-OPT, (black thick solid) density tree,

and (red symbols) size-split linear DET estimators are shown together with (gray thick solid) power-law scalings $n^{-0.25}$, $n^{-0.5}$, and n^{-1}

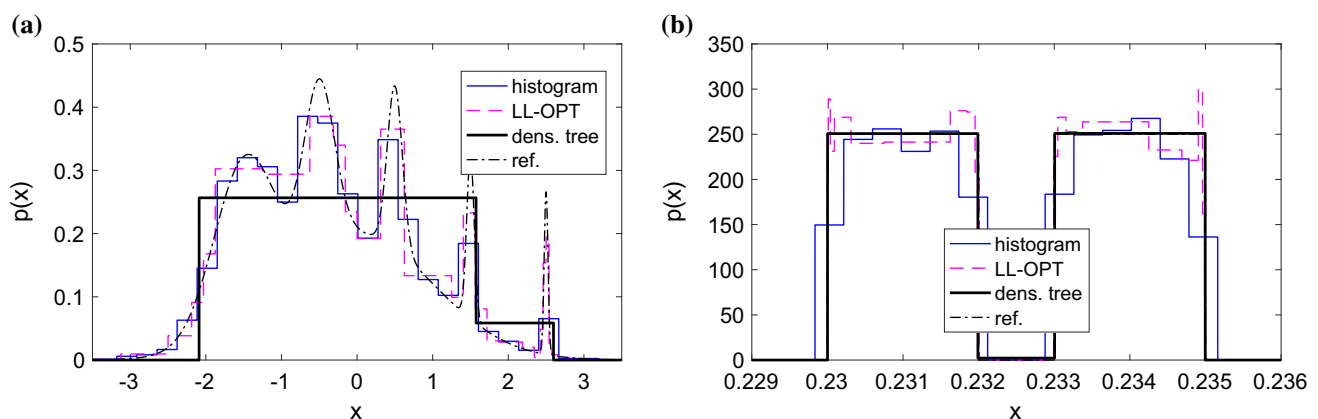


Fig. 18 PDF estimates resulting from (blue thin solid) histogram, (magenta dashed) LL-OPT, and (black thick solid) density tree estimators with $n = 3163$ samples are compared with (black dash dot) the

reference PDFs (12) and (13). In panels (a) and (b), the asymmetric claw (12) and spiky uniforms (13) PDFs, respectively, are depicted

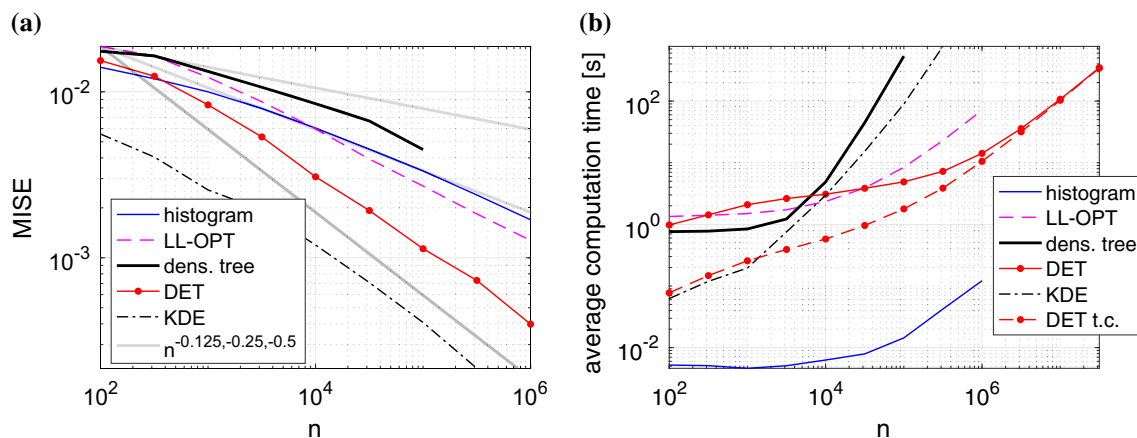


Fig. 19 Estimation of the four-dimensional joint Gaussian PDF. **a** Evolutions of the MISE and **b** computing time as a function of the number of samples n for histograms (blue solid), LL-OPT (dashed magenta), density tree (black thick solid), and DET (red solid symbols) estimators, and adaptive KDE (black dash dot) are plotted. In panel (**b**), the tree construction time for the DET estimator (red dashed symbol) is plotted as well

3.4 Four- and seven-dimensional examples

Next, we assess the performance of the linear DET method with equal-size splits together with the other estimators in examples involving four and seven sample-space dimensions.

3.4.1 High-dimensional Gaussian PDF

In a first multi-dimensional example, we reuse the joint Gaussian PDF (16) with d -dimensional probability space $\mathbf{x} = (x_1, x_2, \dots, x_d)^\top$, mean vector $\boldsymbol{\mu}$ being zero, and $d \times d$ covariance matrix \mathbf{C} . In a first step, we inspect the performance of the different methods for $d = 4$ with the randomly chosen covariance matrix

$$\mathbf{C} = \begin{pmatrix} 1 & -0.344 & 0.141 & -0.486 \\ -0.344 & 1 & 0.586 & 0.244 \\ 0.141 & 0.586 & 1 & -0.544 \\ -0.486 & 0.244 & -0.544 & 1 \end{pmatrix}.$$

In Fig. 19a, MISE decays from the different estimators are compared. KDE performs best, followed by the DET estimator, which shares the same empirically determined convergence rate of approximately $1/\sqrt{n}$. The other estimators perform similarly to histograms. The data series of the density tree and KDE methods stop at $n \approx 10^{4.5}$ and $10^{5.5}$, respectively, due to the rapidly growing computing times of these estimators (see discussion in the introduction 1).

An analysis of computing times per density estimate is provided in Fig. 19b. The reported times are comprised of the construction of the estimator and density queries based on the estimator. The number of query points in the MC

integration is kept constant for different estimators and number of samples n . This explains the plateaus for small n in the LL-OPT, density tree, and histogram estimators, where the computing times for small n are governed by the query effort. For increasing n , the estimator construction becomes noticeable, leading to growing times. In the KDE, the query cost is connected to the number of samples, which leads to a continuous increase in computing time. The relatively small growth in computing time of the DET estimator is based on two factors document in Fig. 20: (panel a) the query time, which is driven by the logarithmically increasing tree depth, grows slowly compared to (b) the tree construction time, which depends on an exponential but sublinear growth in DEs $\langle m \rangle$. The construction time is plotted in Fig. 19b and shown to converge to the total DET computing time, as the query time becomes comparably small for large n .

In a next step, the dimensionality is increased to seven with the covariance matrix

$$\mathbf{C} = \begin{pmatrix} 1 & -0.216 & 0.161 & -0.0496 & 0.0342 & -0.116 & 0.749 \\ -0.216 & 1 & 0.301 & 0.0391 & -0.217 & 0.0189 & -0.381 \\ 0.161 & 0.301 & 1 & 0.574 & -0.312 & 0.109 & 0.386 \\ -0.0496 & 0.0391 & 0.574 & 1 & -0.438 & 0.730 & -0.0572 \\ 0.0342 & -0.217 & -0.312 & -0.438 & 1 & -0.475 & 0.258 \\ -0.116 & 0.0189 & 0.109 & 0.730 & -0.475 & 1 & -0.386 \\ 0.749 & -0.381 & 0.386 & -0.0572 & 0.258 & -0.386 & 1 \end{pmatrix}$$

again chosen randomly. The corresponding MISE and computing time curves for this example are provided in Fig. 21. Compared to the previous four-dimensional case, the MISE convergence rate shown in panel (a) has reduced to $n^{-1/4}$ for the DET estimator (curse of dimensionality). Unlike in the previous case, the DET estimator becomes more accurate than adaptive KDE for $n > 5000$. While the LL-OPT

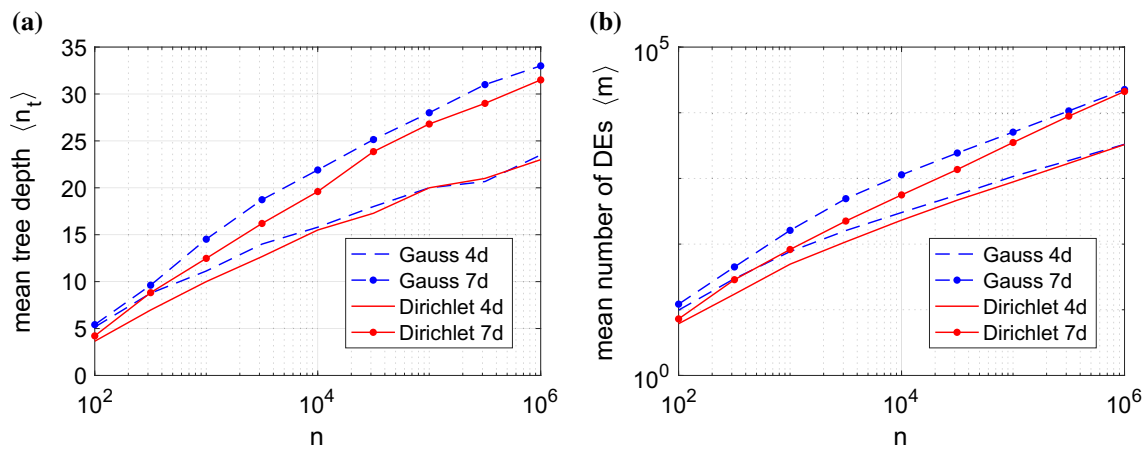


Fig. 20 **a** Mean tree depths $\langle n_t \rangle$ and **b** mean number of DEs $\langle m \rangle$ of the DET estimator in the examples involving the (lines) four- and (symbols) seven-dimensional (blue dashed) Gaussian and (red solid) Dirichlet PDFs

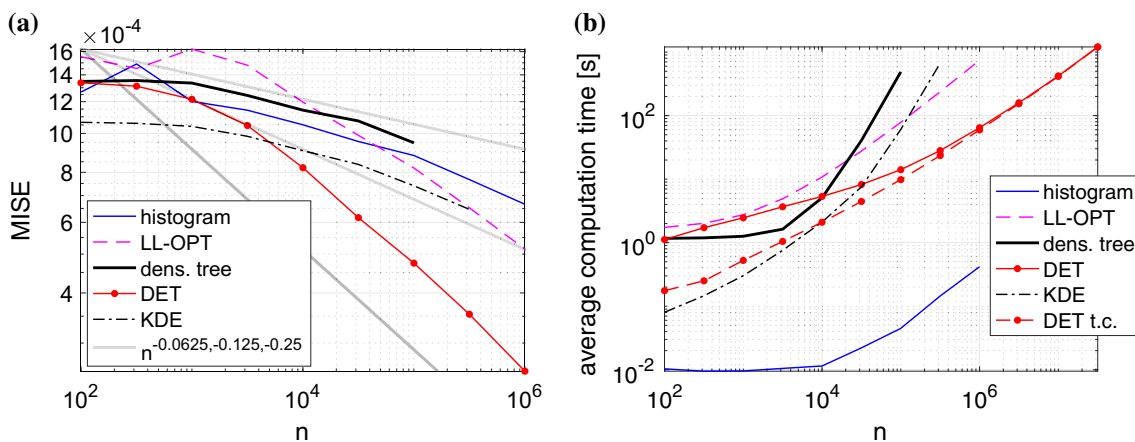


Fig. 21 Estimation of the seven-dimensional joint Gaussian PDF. See Fig. 19

estimator has a similar MISE convergence rate for large n like the DET estimator, the density tree and histogram MISE are similar, like in the previous four-dimensional case. The computing times reported in Fig. 21b show similar trends as well. Unlike in the four-dimensional case, however, we can identify the growth originating from the tree construction process at large n more clearly. Moreover, the computing times have increased noticeably compared to $d = 4$, which is documented as well in Fig. 20, where mean tree depth and number of DEs are higher for larger dimensionality d .

3.4.2 High-dimensional Dirichlet PDF

A d -dimensional generalization of the Dirichlet PDF (18) is given by

$$p(\mathbf{x}) = \frac{\Gamma\left(\sum_{i=1}^{d+1} \alpha_i\right)}{\prod_{i=1}^{d+1} \Gamma(\alpha_i)} \prod_{i=1}^d x_i^{\alpha_i-1} \left(1 - \sum_{i=1}^d x_i\right)^{\alpha_{d+1}-1} \quad (19)$$

$\forall x_i \geq 0$ and $\sum_{i=1}^d x_i \leq 1$ with parameter vector $\alpha = (\alpha_1, \dots, \alpha_{d+1})$ having components $\alpha_i > 0$. Again two examples of different dimensionality are considered, that is a four-dimensional case with

$$\alpha = (6.13, 9.29, 10.6, 8.24, 3.91)^\top$$

and a seven-dimensional case with

$$\alpha = (9, 5.71, 8.96, 4.51, 5.81, 4.06, 10.7, 1.51)^\top.$$

Both parameter vectors were chosen randomly. MISE convergence curves for both cases are plotted in Fig. 22. Except for somewhat smaller asymptotic decay rates, the observations from the previous Gaussian examples carry over to this case. This holds also true for the computing times of the different estimators (not shown) that display similar dependencies on n as seen in the previous section. Finally, by inspecting the mean tree depth and number of DEs in Fig. 20, we can observe that they grow again approximately

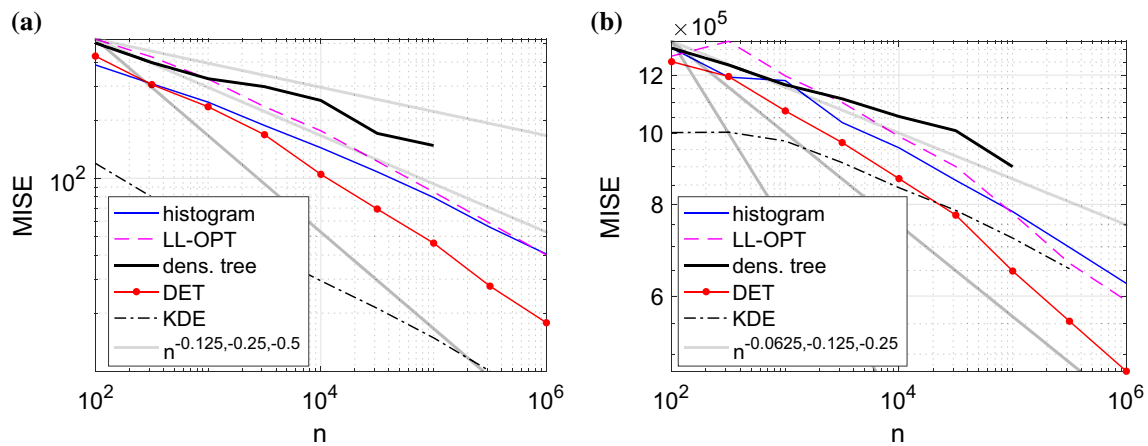


Fig. 22 Estimation of the **a** four- and **b** seven-dimensional Dirichlet PDFs. Evolutions of the MISE as a function of the number of samples n are plotted. See Fig. 19

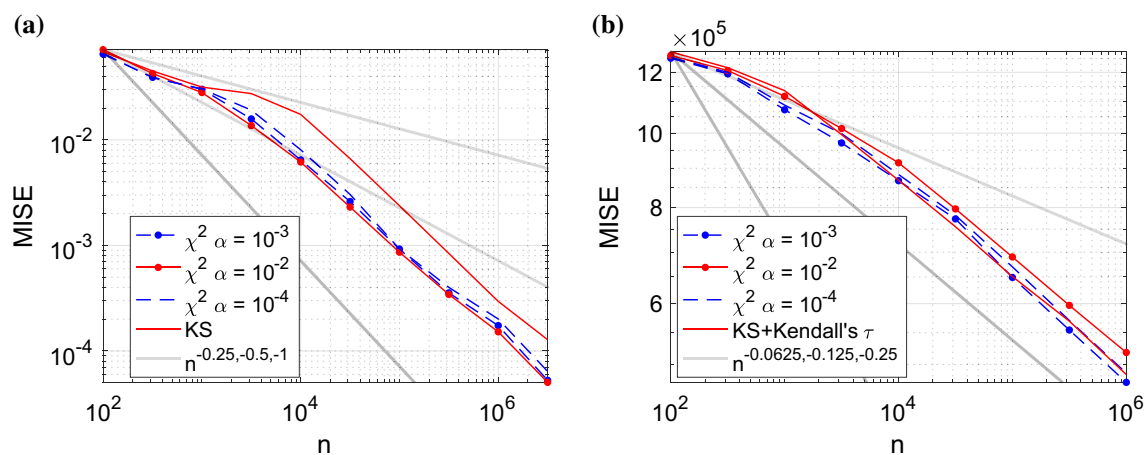


Fig. 23 Estimations of **a** beta PDF (14) and **b** seven-dimensional Dirichlet PDF (19). Evolutions of the MISE as a function of the number of samples n for linear DET estimators with equal-size splits and χ^2 test statistics with significance level (red solid symbol) $\alpha = 0.001$, (blue

dashed symbol) $\alpha = 0.01$, (blue dashed) $\alpha = 0.0001$, and (red solid) KS and Kendall's τ test statistics with $\alpha = 0.001$ are plotted. Power-law and logarithmic scalings with exponents indicated in the figure legends are depicted (gray thick solid)

logarithmically and sublinearly with a dependence on the dimension.

3.5 Dependence on test parameters and test statistic

All previous results were obtained with DET methods that were based on χ^2 test statistics both for goodness-of-fit and independence with identical significance levels $\alpha_g = \alpha_d = 0.001$, respectively. To inspect the influence of the choice of the test statistic on one hand, we performed DET computations with KS and Kendall's τ tests for goodness-of-fit and independence, respectively, while keeping the significance levels unchanged. On the other hand, we varied the significance level $\alpha = \alpha_g = \alpha_d$ by factors of 0.1 and 10, while

maintaining the χ^2 test statistics. No significant changes in the DET estimator performance were found with these variations as is, for example, shown in Fig. 23, where MISE results for the one-dimensional beta and seven-dimensional Dirichlet cases are provided.

By varying, however, the significance level α , a dependence of the mean number of DEs is expected as is seen in Fig. 24. While an increase in α leads to an increase in the number of elements, more elements do not necessarily translate into higher accuracy or reduced MISE. This is due to a bias/variance trade-off, where as α is increased bias/statistical errors de-/increase, respectively, and vice versa. As a result, the total error or MISE does not change significantly.

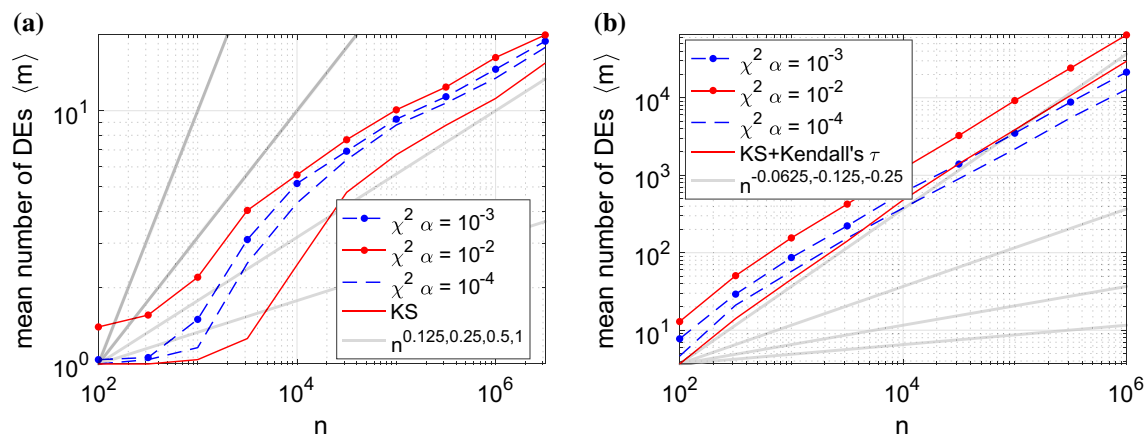


Fig. 24 Estimations of **a** beta PDF (14) and **b** seven-dimensional Dirichlet PDF (19). Evolutions of the mean number of DEs $\langle m \rangle$ are depicted. See Fig. 23

4 Concluding remarks

The DET estimator provides an analytical density representation based on piecewise constant, linear, and possibly higher-order functions. This representation can be efficiently assembled and evaluated at arbitrary probability space positions. Therefore, the adaptivity of the DET method is not limited by a smallest scale that limits, for example, the performance of a grid-based estimation method. Unlike adaptive KDE, DET-based estimates provide limited differentiability, which may be a disadvantage in some applications. Moreover, unlike combined with a principal axes transform common in KDE [e.g., Silverman 1998, equation (4.7)], the DET estimator is not invariant to the orientation of the coordinate system. A corresponding combination was found to enable an increase in the computational efficiency and accuracy of the DET method.

In terms of accuracy, the linear DET method showed, for the different examples considered, similar or better MISE convergence rates compared with adaptive KDE. The latter is, besides the DET method, the most accurate estimator in this study. Constant DEs are advantageous only in cases, where samples stem from piecewise constant PDFs, as they lead to smaller oscillations at PDF discontinuities. The LL-OPT estimator follows in terms of accuracy after the linear DET estimator. Conventional histograms are least accurate, but at the same time computationally inexpensive as long as the dimension is not too high. Density trees were found to provide mixed results in terms of accuracy.

Computing times of the DET method were found to scale in terms of the number of samples favorably compared to the other estimators. In the DET method, the number of DEs, which determines the time for the DET construction,

was found to grow sublinearly. In the case of score-based splitting, there is a theoretical linear upper bound to the DE growth. The splitting method for the DET construction was found to have little effect on the number of DEs, but affects the tree depth. However, irrespective of the method applied, the mean tree depth scales approximately logarithmically with the number of samples. The computational cost of a density evaluation at a specific point is determined by the tree depth. By varying the test parameters that guide the splitting process, i.e., significance levels and test statistics, no significant change in the DET performance was found.

In conclusion, our new DET method is a good candidate for a computationally efficient general-purpose density estimator. The MATLAB implementation of the DET method that was applied in this study is available for download from the MathWorks File Exchange (tag 'distribution element tree').

An important element that allowed us to break the exponential growth of the bin count with respect to the dimension d is the approximation of (mutual) independence by pairwise independence. The relation between pairwise and mutual independence is a subject of ongoing research (e.g., Nelsen and Ubeda-Flores 2012). Besides this important aspect, to further the DET method, it would be interesting to inspect the performance of higher-order DEs and more advanced goodness-of-fit and statistical independence tests.

Appendix: Derivation of MMSE slope estimator

Writing without loss of generality the linear marginal PDF (4) in a simpler form with $x_i \in [0, 1]$ and the subscripts skipped, we obtain

$$p(x|\theta) = \left(x - \frac{1}{2}\right)\theta + 1.$$

By calculating the mean of random variable X based on this PDF, we obtain $\langle X \rangle = \frac{1}{12}(6 + \theta)$, and therefore, can express the slope parameter in terms of this mean as

$$\theta = 6(2\langle X \rangle - 1). \quad (20)$$

In the case of a finite ensemble, we estimate the mean with $\langle X \rangle_n = \frac{1}{n} \sum_{j=1}^n x_j$ and the slope by

$$\hat{\theta} = 6c(2\langle X \rangle_n - 1).$$

Here, c is a correction factor that is determined by minimizing the mean square error (MSE) expressed as

$$\begin{aligned} \langle (\hat{\theta} - \theta)^2 \rangle &= \left\langle \left[6c \left(\frac{2}{n} \sum_{j=1}^n x_j - 1 \right) - \theta \right]^2 \right\rangle \\ &= \left\langle 36c^2 \left(\frac{2}{n} \sum_{j=1}^n x_j - 1 \right)^2 \right. \\ &\quad \left. - 12c \left(\frac{2}{n} \sum_{j=1}^n x_j - 1 \right) \theta + \theta^2 \right\rangle \\ &= 36c^2 \left\langle \frac{4}{n^2} \sum_{j=1}^n \sum_{k=1}^n x_j x_k - \frac{4}{n} \sum_{j=1}^n x_j + 1 \right\rangle \\ &\quad - 12c(2\langle X \rangle - 1)\theta + \theta^2 \\ &= 36c^2 \left(\frac{4}{n^2} \left\langle \sum_{j=1}^n \sum_{k=1}^n x_j x_k \right\rangle - 4\langle X \rangle + 1 \right) \\ &\quad - 12c(2\langle X \rangle - 1)\theta + \theta^2 \\ &= 36c^2 \left(\frac{4}{n^2} \left\langle \sum_{j=1}^n \sum_{k=1, k \neq j}^n x_j x_k + \sum_{j=1}^n x_j^2 \right\rangle - 4\langle X \rangle + 1 \right) \\ &\quad - 12c(2\langle X \rangle - 1)\theta + \theta^2 \\ &= 36c^2 \left[\frac{4(n-1)}{n} \langle X \rangle^2 + \frac{4}{n} \langle X^2 \rangle - 4\langle X \rangle + 1 \right] \\ &\quad - 12c(2\langle X \rangle - 1)\theta + \theta^2. \end{aligned}$$

To determine the minimum MSE, we set

$$\begin{aligned} \frac{d}{dc} \langle (\hat{\theta} - \theta)^2 \rangle &= 72c \left(\frac{4(n-1)}{n} \langle X \rangle^2 + \frac{4}{n} \langle X^2 \rangle \right. \\ &\quad \left. - 4\langle X \rangle + 1 \right) \\ &\quad - 12(2\langle X \rangle - 1)\theta = 0, \end{aligned}$$

which leads for the correction factor to

$$\begin{aligned} c &= \frac{(2\langle X \rangle - 1)n\theta}{6[4(n-1)\langle X \rangle^2 + 4\langle X^2 \rangle - 4n\langle X \rangle + n]} \\ &= \frac{(2\langle X \rangle - 1)n\theta}{6(4n\langle X \rangle^2 - 4\langle X \rangle^2 + 4\langle X^2 \rangle - 4n\langle X \rangle + n)} \end{aligned}$$

$$\begin{aligned} &= \frac{(2\langle X \rangle - 1)n\theta}{6[n(2\langle X \rangle - 1)^2 + 4\langle X^2 \rangle]} \\ &= \frac{6n(2\langle X \rangle - 1)^2}{6[n(2\langle X \rangle - 1)^2 + 4\langle X^2 \rangle]}. \end{aligned} \quad (21)$$

For $n \rightarrow \infty$ the correction factor c goes to one.

References

- Achilleos, A., Delaigle, A.: Local bandwidth selectors for deconvolution kernel density estimation. *Stat. Comput.* **22**(2), 563–577 (2012)
- Bagnato, L., Punzo, A., Nicolis, O.: The autodependogram: a graphical device to investigate serial dependences. *J. Time Ser. Anal.* **33**(2), 233–254 (2012)
- Botev, Z.: Spectral implementation of adaptive kernel density estimator via diffusion. <https://ch.mathworks.com/matlabcentral/fileexchange/14034-kernel-density-estimator> (2007). Accessed 01 Sep 2016
- Botev, Z.: Implementation of adaptive kernel density estimator for high dimensions via diffusion. <https://ch.mathworks.com/matlabcentral/fileexchange/58312-kernel-density-estimator-for-high-dimensions> (2016). Accessed 10 Jan 2017
- Botev, Z.I., Grotowski, J.F., Kroese, D.P.: Kernel density estimation via diffusion. *Ann. Stat.* **38**(5), 2916–2957 (2010)
- Breiman, L., Friedman, J., Stone, C.J., Olshen, R.: *Classification and Regression Trees*. Wadsworth Statistics/Probability. Chapman and Hall/CRC, Boca Raton (1984)
- Cao, R., Cuevas, A., Gonzalez Manteiga, W.: A comparative study of several smoothing methods in density estimation. *Comput. Stat. Data Anal.* **17**(2), 153–176 (1994)
- Cochran, W.G.: The chi square test of goodness of fit. *Ann. Math. Stat.* **23**(3), 315–345 (1952)
- Curtin, R.R., Cline, J.R., Slagle, N.P., March, W.B., Ram, P., Mehta, N.A., Gray, A.G.: Mlpack: a scalable C++ machine learning library. *J. Mach. Learn. Res.* **14**, 801–805 (2013)
- Ferguson, T.S.: A bayesian analysis of some nonparametric problems. *Ann. Stat.* **1**(2), 209–230 (1973)
- Fix, E., Hodges, J.: Discriminatory analysis, nonparametric estimation: consistency properties. Report 4, Project No. 21-49-004, USAF School of Aviation Medicine (1951)
- Härdle, W., Werwatz, A., Müller, M., Sperlich, S.: *Nonparametric and Semiparametric Models*. Springer Series in Statistics, 1st edn. Springer, Berlin (2004)
- Jiang, H., Mu, J.C., Yang, K., Du, C., Lu, L., Wong, W.H.: Computational aspects of optional Pólya tree. *J. Comput. Graph. Stat.* **25**(1), 301–320 (2016)
- Jing, J., Koch, I., Naito, K.: Polynomial histograms for multivariate density and mode estimation. *Scand. J. Stat.* **39**(1), 75–96 (2012)
- Jones, M.C., Marron, J.S., Sheather, S.J.: A brief survey of bandwidth selection for density estimation. *J. Am. Stat. Assoc.* **91**(433), 401–407 (1996)
- Kendall, M.G.: A new measure of rank correlation. *Biometrika* **30**(1–2), 81–93 (1938)
- Kogure, A.: Asymptotically optimal cells for a histogram. *Ann. Stat.* **15**(3), 1023–1030 (1987)
- Kooperberg, C., Stone, C.J.: A study of logspline density estimation. *Comput. Stat. Data Anal.* **12**(3), 327–347 (1991)
- Loftsgaarden, D.O., Quesenberry, C.P.: A nonparametric estimate of a multivariate density function. *Ann. Math. Stat.* **36**(3), 1049–1051 (1965)

- Ma, L., Wong, W.H.: Coupling optional Pólya trees and the two sample problem. *J. Am. Stat. Assoc.* **106**(496), 1553–1565 (2011)
- Mann, H.B., Wald, A.: On the choice of the number of class intervals in the application of the chi square test. *Ann. Math. Stat.* **13**(3), 306–317 (1942)
- Marron, J.S., Wand, M.P.: Exact mean integrated squared error. *Ann. Stat.* **20**(2), 712–736 (1992)
- Neal, R.M.: Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Stat.* **9**(2), 249–265 (2000)
- Nelsen, R.B., Ubeda-Flores, M.: How close are pairwise and mutual independence? *Stat. Probab. Lett.* **82**(10), 1823–1828 (2012)
- O'Brien, T.A., Kashinath, K., Cavanaugh, N.R., Collins, W.D., O'Brien, J.P.: A fast and objective multidimensional kernel density estimation method: fastKDE. *Comput. Stat. Data Anal.* **101**, 148–160 (2016)
- Papoulis, A.: Probability, Random Variables, and Stochastic Processes. McGraw-Hill Series in Electrical Engineering, 3rd edn. McGraw-Hill Inc., New York (1991)
- Park, B.U., Marron, J.S.: Comparison of data-driven bandwidth selectors. *J. Am. Stat. Assoc.* **85**(409), 66–72 (1990)
- Park, B., Turlach, B.: Practical performance of several data driven bandwidth selectors. Report, Université Catholique de Louvain, Center for Operations Research and Econometrics (CORE) (1992)
- Pearson, K.: On the criterion that a given system of deviations from the probable in the case of correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos. Mag.* **50**, 157–175 (1900)
- Petersen, A., Müller, H.G.: Functional data analysis for density functions by transformation to a Hilbert space. *Ann. Stat.* **44**(1), 183–218 (2016)
- Ram, P., Gray, A.G.: Density estimation trees. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, pp. 627–635 (2011)
- Rosenblatt, M.: Remarks on some nonparametric estimates of a density function. *Ann. Math. Stat.* **27**(3), 832–837 (1956)
- Scott, D.W.: Multivariate Density Estimation Theory, Practice, and Visualization. Wiley Series in Probability and Statistics, 2nd edn. Wiley, Hoboken (2015)
- Scott, D.W., Sagae, M.: Adaptive density estimation with massive data sets. In: Proceedings of the Statistical Computing Section, pp. 104–108. ASA, American Statistical Association (1997)
- Shampine, L.F.: Matlab program for quadrature in 2d. *Appl. Math. Comput.* **202**(1), 266–274 (2008a)
- Shampine, L.F.: Vectorized adaptive quadrature in matlab. *J. Comput. Appl. Math.* **211**(2), 131–140 (2008b)
- Sheather, S.J.: Density estimation. *Stat. Sci.* **19**(4), 588–597 (2004)
- Shorack, G.R.: Probability for Statisticians. Springer Texts in Statistics. Springer, Berlin (2000)
- Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Monographs on Statistics and Applied Probability. Chapman and Hall, CRC, Boca Raton (1998)
- Smirnov, N.: Table for estimating the goodness of fit of empirical distributions. *Ann. Math. Stat.* **19**(2), 279–281 (1948)
- Sriperumbudur, B., Fukumizu, K., Kumar, R., Gretton, A., Hyvärinen, A.: Density estimation in infinite dimensional exponential families. [arXiv:1509.04348v2](https://arxiv.org/abs/1509.04348v2) p. 42 (2013)
- Steele, M., Chaseling, J.: Powers of discrete goodness-of-fit test statistics for a uniform null against a selection of alternative distributions. *Commun. Stat. Simul. Comput.* **35**(4), 1067–1075 (2006)
- Wang, X., Wang, Y.: Nonparametric multivariate density estimation using mixtures. *Stat. Comput.* **25**(2), 349–364 (2015)
- Wong, W.H., Ma, L.: Optional Pólya tree and Bayesian inference. *Ann. Stat.* **38**(3), 1433–1459 (2010)
- Yang, Y.: Penalized semiparametric density estimation. *Stat. Comput.* **19**(4), 355 (2008)
- Zaunders, J., Jing, J., Leipold, M., Maecker, H., Kelleher, A.D., Koch, I.: Computationally efficient multidimensional analysis of complex flow cytometry data using second order polynomial histograms. *Cytom. Part A* **89**(1), 44–58 (2016)