A non-parametric k-nearest neighbor entropy estimator

Damiano Lombardi, Sanjay Pant

A non-parametric k-nearest neighbor entropy estimator

Damiano Lombardi and Sanjay Pant*
Inria Paris-Rocquencourt, BP 105, 78153 Le Chesnay Cedex, France
Sorbonne Universités, UPMC Univ Paris 06, 4 Place Jussieu, 75252 Paris cedex 05, France
CNRS, UMR 7598 Laboratoire Jacques-Louis Lions, Paris, France
(Dated: February 11, 2016)

A non-parametric k-nearest neighbor based entropy estimator is proposed [a]. It improves on the classical Kozachenko-Leonenko estimator by considering non-uniform probability densities in the region of k-nearest neighbors around each sample point. It aims at improving the classical estimators in three situations: first, when the dimensionality of the random variable is large; second, when near-functional relationships leading to high correlation between components of the random variable are present; and third, when the marginal variances of random variable components vary significantly with respect to each other. Heuristics on the error of the proposed and classical estimators are presented. Finally, the proposed estimator is tested for a variety of distributions in successively increasing dimensions and in the presence of a near-functional relationship. Its performance is compared with a classical estimator and shown to be a significant improvement.

I. INTRODUCTION

Entropy is a fundamental quantity in information theory that finds applications in various areas such as coding theory and data compression [1]. It is also a building block for other important measures, such as *mutual information* and *interaction information*, that are widely employed in the areas of computer science, machine learning, and data analysis. In most realistic applications, the underlying true probability density function (pdf) is rarely known, but samples from it can be obtained via data-acquisition, experiments, or numerical simulations. An interesting problem then, is to estimate the entropy of the underlying distribution only from a finite number of samples. The approaches to perform such a task can broadly be classified into two categories: *parametric* and *non-parametric*. In the parametric approach the form of the pdf is assumed to be known and its parameters are identified from the samples. This, however, is a strong assumption and in most realistic cases an *a priori* assumption on the form of the pdf is not justified. Consequently, non-parametric approaches where no such assumption is made have been proposed [2]. One such approach is to first estimate the pdf through histograms or kernel density estimators (KDE) [3–5], and then to compute the entropy by either numerical or Monte-Carlo (MC) integration. Other alternatives include methods based on sample spacings for one-dimensional distributions [6, 7] and k-nearest neighbors (kNN) [8–11].

While KDE based entropy estimation is generally accurate and efficient in low dimensions, the method suffers from the *curse of dimensionality* [12]. On the other hand the kNN based estimators are computationally efficient in high dimensions, but not necessarily accurate, especially in the presence of large correlations or functional dependencies [13]. The latter problem has recently been addressed by estimating the local non-uniformity through principal component analysis (PCA) in [13]. In the current work, a different approach to overcome the aforementioned limitations associated with kNN based entropy estimators is presented. The central idea is to estimate the probability mass around each sample point by a local Gaussian approximation. The local approximation is obtained by looking at p-neighbors around the sample point. This procedure has two distinct advantages: first, that the tails of the true probability distribution are better captured; and second, that if the probability mass in one or more directions is small due to large correlations (near-functional dependencies), or due to significant variation in the marginal variances of the random variable components, the non-uniformity is inherently taken into account. These two features allow the entropy to be estimated in high dimensions with a significantly lower error when compared to classical estimators.

The structure of the work is as follows: first, the classical and the new kNN estimators are presented in section II; then, the heuristics on the errors of the two estimators are presented in section III; and finally, numerical test cases are presented in section IV for a variety of distributions in successively increasing dimensions.

II. FORMULATION OF THE ENTROPY ESTIMATOR

Let the random variable under consideration be $\mathbf{X} \in \mathbb{R}^d$ and its probability density be denoted by $p_{\mathbf{X}}(\mathbf{x})$. Its entropy is defined as

$$H(\mathbf{X}) = \int_{\mathcal{X}} p_{\mathbf{X}}(\mathbf{x}) \log\left(\frac{1}{p_{\mathbf{X}}(\mathbf{x})}\right) \, d\mathbf{x} \qquad (1)$$

where $\mathcal{X}$ is the support of $p_{\mathbf{X}}(\mathbf{x})$. The goal is to estimate $H(\mathbf{X})$ from N finite samples, $\mathbf{x}_i \quad i = 1 \dots N$, from

* {Damiano.Lombardi, Sanjay.Pant}@inria.fr
[a] The final version of the paper is available on Phyical Review E, Vol.93, 2016.

the distribution $p_{\mathbf{X}}(\mathbf{x})$. A Monte-Carlo estimate of the entropy can be written as

$$\hat{H}(\mathbf{X}) = \frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{1}{p_{\mathbf{X}}(\mathbf{x}_i)}\right). \tag{2}$$

However, since $p_{\mathbf{X}}(\mathbf{x}_i)$ is unknown, an estimate $\hat{p}_{\mathbf{X}}(\mathbf{x}_i)$ must be substituted in equation (2) to obtain $\hat{H}(\mathbf{X})$.

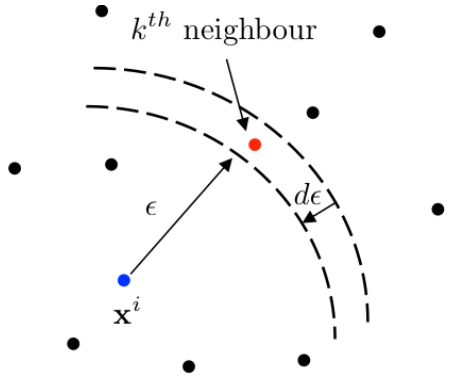


FIG. 1. (Color online) A depiction of k-nearest neighbor and ε-ball.

The key idea is to estimate $\hat{p}_{\mathbf{X}}(\mathbf{x}_i)$ through k-nearest neighbors (kNN) of $\mathbf{x}_i$. Consider the probability density $p_k(\varepsilon)$ of ε, the distance from $\mathbf{x}_i$ to its kNN (see Figure 1). The probability $p_k(\varepsilon)d\varepsilon$ is the probability that exactly one point is in $[\varepsilon, \varepsilon + d\varepsilon]$, exactly $k-1$ points are at distances less than the kNN, and the remaining points are farther than the kNN. Then it follows that

$$p_k(\varepsilon)d\varepsilon = \binom{N-1}{1}\frac{dP_i(\varepsilon)}{d\varepsilon}d\varepsilon\binom{N-2}{k-1}(P_i(\varepsilon))^{k-1}\ (1-P_i(\varepsilon))^{N-k-1} \tag{3}$$

where, $P_i(\varepsilon)$ is the probability mass of an ε-ball centered at a sample point $\mathbf{x}_i$. The region inside the ε-ball is $||\mathbf{x}-\mathbf{x}_i|| < \varepsilon$ and is denoted by denoted by $\mathcal{B}(\varepsilon, \mathbf{x}_i)$. The probability mass in $\mathcal{B}(\varepsilon, \mathbf{x}_i)$ is

$$P_i(\varepsilon) = \int_{\mathcal{B}(\varepsilon,\mathbf{x}_i)} p_{\mathbf{X}}(\mathbf{x})\ d\mathbf{x}. \tag{4}$$

The expected value of $\log(P_i)$ can be obtained from equations (3) and (4)

$$\mathbb{E}(\log P_i) = \int_0^\infty \log P_i(\varepsilon)\ p_k(\varepsilon)\ d\varepsilon = \psi(k) - \psi(N) \tag{5}$$

where ψ is the digamma function.

If the probability mass in $\mathcal{B}(\varepsilon, \mathbf{x}_i)$ can be written in the following form

$$P_i \approx \eta_i\ p_{\mathbf{X}}(\mathbf{x}_i) \tag{6}$$

then, by considering the logarithm and taking expectations on both sides of equation (6), and using equations (5) and (2), the entropy estimate can be written as

$$\hat{H}(\mathbf{X}) = \psi(N) - \psi(k) + \frac{1}{N}\sum \log \eta_i. \tag{7}$$

In what follows the classical manner to obtain equation (6) and the new estimator are presented.

A. Classical estimators

The classical estimates by Kozachenko and Leonenko [8, 11], and similarly by Singh et. al. [10], assume that the probability density $p_{\mathbf{X}}(\mathbf{x})$ is constant inside $\mathcal{B}(\varepsilon, \mathbf{x}_i)$. For example Kozachenko and Leonenko [8, 11] assume that

$$P_i \approx c_d\ \varepsilon^d\ p_{\mathbf{X}}(\mathbf{x}_i) \tag{8}$$

where c_d is the volume of the d-dimensional unit-ball ($\mathcal{B}(\varepsilon, \mathbf{x}_i)$ with $\varepsilon = 1$). The expression for c_d depends on the type of norm used to calculate the distances; for example, for maximum (L_∞) norm $c_d = 2^d$ and for euclidean (L_2) norm $c_d = \pi^{d/2}/\Gamma(1+d/2)$, where Γ is the Gamma function. Using equation (8) in equations (6) and (7), the entropy estimate can be written as

$$\hat{H}(\mathbf{X}) = \psi(N) - \psi(k) + \log(c_d) + \frac{d}{N}\sum_{i=1}^{N}\log\left(\varepsilon(i)\right) \tag{9}$$

where $\varepsilon(i)$ is the distance of the i^{th} sample to its k^{th} nearest neighbor. This estimator is referred as KL estimator in the remainder of this article.

B. The kpN estimator

Although the classical estimator works well in low-dimensions, it presents with large errors when the dimensionality of the random variable is high or the pdf in $\mathcal{B}(\varepsilon, \mathbf{x}_i)$ shows high non-uniformity. The latter may result from: i) presence of a near-functional relationship (leading to high correlation) between two or more components of the random variable $\mathbf{X}$ [13]; and ii) high variability in the marginal variances of $\mathbf{X}$ in $\mathcal{B}(\varepsilon, \mathbf{x}_i)$. In the remainder of the manuscript the term non-uniformity is used to imply the aforementioned features. The primary cause of high error in the KL estimator is the assumption of constant density in each $\mathcal{B}(\varepsilon, \mathbf{x}_i)$. This may be unjustified when the true probability mass is likely to be high only on a small sub-region of $\mathcal{B}(\varepsilon, \mathbf{x}_i)$. In such cases, a constant density assumption in $\mathcal{B}(\varepsilon, \mathbf{x}_i)$ leads to and overestimation of the probability mass and hence the entropy estimate [13]. To remedy this, an alternate formulation for η_i in equation (6) is sought. Contrary to a constant density assumption, the probability density in $\mathcal{B}(\varepsilon, \mathbf{x}_i)$ is represented as

$$p_{\mathbf{X}}(\mathbf{x}) \approx \rho \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{S}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{10}$$

where $\boldsymbol{\mu}$ and $\mathbf{S}$ represent the empirical mean and covariance matrix of the p neighbors of the point $\mathbf{x}_i$. Essentially, the probability density is assumed to be proportional to a Gaussian function approximated by using p-nearest neighbors of $\mathbf{x}_i$. The idea is that the p-neighbors

would capture the local non-uniformity of the true probability density inside $\mathcal{B}(\varepsilon, \mathbf{x}_i)$. From a physical point of view, p is reflective of the characteristic length of changes in the true probability distribution.

Following equation (10), to obtain the form of equation (6), the proportionality constant ρ is obtained by requiring that the value of the local Gaussian approximation be equal to the true pdf at $\mathbf{x}_i$

$$p_{\mathbf{X}}(\mathbf{x}) \approx p_{\mathbf{X}}(\mathbf{x}_i)\ \frac{g(\mathbf{x})}{g(\mathbf{x}_i)}, \tag{11}$$

where

$$g(\mathbf{x}) = \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{S}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right), \tag{12}$$

$$g(\mathbf{x}_i) = \exp\left(-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu})^T\mathbf{S}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\right). \tag{13}$$

Consequently, the probability mass in $\mathcal{B}(\varepsilon, \mathbf{x}_i)$ can be written as

$$P_i = p_{\mathbf{X}}(\mathbf{x}_i)\ \frac{1}{g(\mathbf{x}_i)}\ G_i \tag{14}$$

where

$$G_i = \int\limits_{\mathcal{B}(\varepsilon,\mathbf{x}_i)} g(\mathbf{x})\ d\mathbf{x} \tag{15}$$

Using equation (14) in equations (6) and (7), the entropy estimate can be written as

$$\hat{H}(\mathbf{X}) = \psi(N) - \psi(k) - \frac{1}{N}\sum_{i=1}^{N}\log\left(g(\mathbf{x}_i)\right) + \frac{1}{N}\sum_{i=1}^{N}\log G_i. \tag{16}$$

The above estimator for entropy is referred as the kpN estimator. In this estimator, while the evaulation of $g(\mathbf{x}_i)$ is straightforward, the evaluation of G_i in equation (15) for each sample point is not trivial, especially in high dimensions. Before describing a computationally efficient method to evaluate this integral in the next section, a graphical demonstration of the difference in the integrals of probability density considered by the KL and kpN estimators is shown in Figure 2. Two different points – one near the tails and one near the mode – of a Gaussian distribution are shown. While near the mode of the distribution the approximations to the integral of the probability density are similar for the two estimators, in the tails the integral is better captured by the kpN estimator as a local Gaussian is constructed. This difference, while insignificant in low dimensions can have a significant impact in higher dimensions (demonstrated in section IV).

The drawbacks of the KL estimator were also addressed recently in [13] where a local PCA is performed on the empirical covariance matrix of the k-neighbors. The eigenvalues and eigenvectors of the empirical covariance matrix are then used to rotate and rescale $\mathcal{B}(\varepsilon, \mathbf{x}_i)$.

Algorithm 1: Algorithm to estimate kpN entropy

Input:
- $\mathbf{x}_i \in \mathbb{R}^d$, $i = 1 \dots N$: the samples
- k: the number of nearest neighbors for calculating $\mathcal{B}(\varepsilon, \mathbf{x}_i)$
- p: the number of nearest neighbors for calculating the local Gaussian approximation ($p \geq k$)

Output: $\hat{H}(\mathbf{X})$: the kpN entropy estimate

for $i \leftarrow 1$ **to** N **do**
  $\{\mathbf{x}_i\}^p \leftarrow$ set of p-nearest neighbors of $\mathbf{x}_i$ (L_∞ norm)
end

$\hat{H}(\mathbf{X}) = \psi(N) - \psi(k)$
for $i \leftarrow 1$ **to** N **do**
  $\varepsilon_i \leftarrow L_\infty$ distance to the k-th nearest neighbor of $\mathbf{x}_i$
  $\mathcal{B}(\varepsilon, \mathbf{x}_i) \leftarrow \mathbf{x}_i \pm \varepsilon_i\ \mathbf{e}$; $\mathbf{e}$ being the canonical basis
  $\mu_i \leftarrow$ mean of $\{\mathbf{x}_i\}^p$
  $\mathbf{S}_i \leftarrow$ covariance of $\{\mathbf{x}_i\}^p$
  $G_i \leftarrow$ integral in equation (15) through EMPGP of μ_i and $\mathbf{S}_i$ (section II C)
  $g(\mathbf{x}_i) \leftarrow$ equation (13)
  $\hat{H}(\mathbf{X}) \leftarrow \hat{H}(\mathbf{X}) + N^{-1}\left[\log(G_i) - \log(g(\mathbf{x}_i))\right]$
end

The intuition is to transform $\mathcal{B}(\varepsilon, \mathbf{x}_i)$ such that the hypothesis of constant density in the transformed $\mathcal{B}(\varepsilon, \mathbf{x}_i)$ is better justified. On the contrary, in the present approach the constant density hypothesis is changed to a Gaussian density (estimated empirically via p neighbors) without transforming the ball $\mathcal{B}(\varepsilon, \mathbf{x}_i)$

C. Gaussian integral in boxes

In order to compute the function G_i a multivariate Gaussian definite integral inside $\mathcal{B}(\varepsilon, \mathbf{x}_i)$ has to be computed. Since we adopt the L_∞ distance, this operation amounts to computing the integral of a multivariate Gaussian inside a box. Among the methods proposed in the literature (see for instance [14]), the Expectation Propagation Multivariate Gaussian Probability (EPMGP) method, proposed in [15], is chosen. The method is based on the introduction of a fictitious probability distribution, whose Kullback-Leibler distance with respect to the original distribution is minimised, inside the box. Since the minimisation of the Kullback-Leibler distance is equivalent, for the present setting, to the moment matching, the zero-th, first and second moments of the fictitious distribution match the ones of the original distribution. The zero-th order moment, in particular, is the sought integral value. This method, as shown in [15], is precise in computing the definite Gaussian integral when the domain is a box.

Algorithm 1 shows the steps to obtain the kpN estimate.

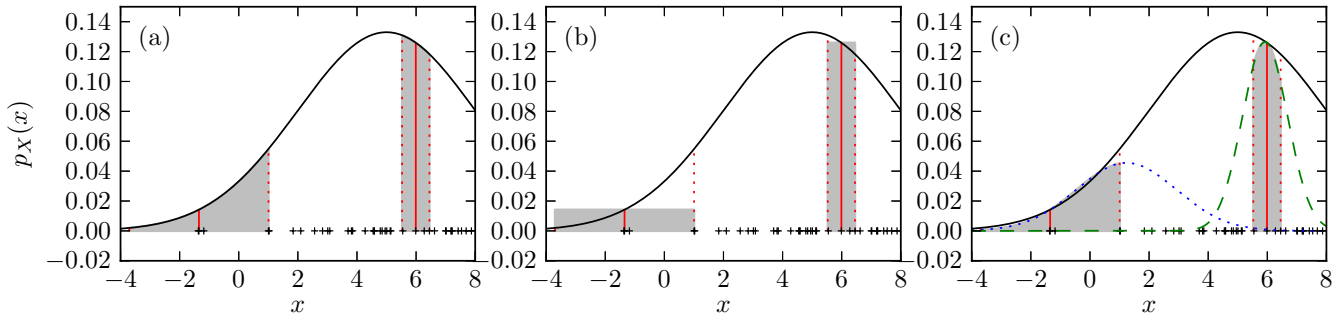


FIG. 2. (Color online) Demonstration of the differences between KL and kpN estimators. In each plot, the true distribution (Gaussian) is shown in solid (black) line and the 50 samples are shown with '+' markers. For the two points (shown in solid vertical lines), the integration region $\mathcal{B}(\varepsilon, \mathbf{x}_i)$ with $k = 3$ is shown with dashed vertical lines, and the integrals are shown in shaded grey. In the left panel, the true area of integration is shown. The centre panel shows the KL approximation to this area, and the right panel shows the area approximations by the kpN estimator with $p = 10$. The local Gaussian approximations for the kpN estimator are shown in dotted (blue) and dashed (green) curves.

III. HEURISTICS ON THE ERROR

In this section analytical heuristics on the error are presented to motivate the approach proposed in this work. First, the error of the KL estimator is derived. The result shows that the estimate is sensitive to both the space dimension and non-uniformity of the pdf in $\mathcal{B}(\varepsilon, \mathbf{x}_i)$.

In what follows, $\mathcal{B}(\varepsilon, \mathbf{x}_i) = [\mathbf{x}_i - \varepsilon_i, \mathbf{x}_i + \varepsilon_i]^d$. Let $p_i(\boldsymbol{\xi})$ be the probability density in $\mathcal{B}(\varepsilon, \mathbf{x}_i)$. In each ball, it is supposed to be $p_i(\boldsymbol{\xi}) \in C^2(\mathcal{B}(\varepsilon, \mathbf{x}_i))$. Albeit quite strong, this regularity is introduced for sake of simplicity of the heuristics. The probability mass is $P_i = \int_{\mathcal{B}(\varepsilon,\mathbf{x}_i)} p_i \, d\boldsymbol{\xi}$.

A. KL estimator error analysis

The error of the KL estimator is analysed. It comprises of two contributions: a statistical error related to the MC integration and an analytical error, resulting from the hypothesis of constant density in $\mathcal{B}(\varepsilon, \mathbf{x}_i)$.

1. *Error in the approximation of probability mass*

The analytical contribution to the error is analysed in this section (see details in Appendix). By considering a second order Taylor expansion of the pdf in $\mathcal{B}(\varepsilon, \mathbf{x}_i)$, the probability mass can be approximated by:

$$P_i \approx P_i^{(\mathrm{KL})} + \frac{1}{2} \int_{\mathcal{B}(\varepsilon,\mathbf{x}_i)} (\boldsymbol{\xi} - \mathbf{x}_i)^T H_{\mathbf{x}_i} (\boldsymbol{\xi} - \mathbf{x}_i) \, d\boldsymbol{\xi}, \quad (17)$$

where $P_i^{(\mathrm{KL})}$ is the probability mass resulting from constant density assumption in the KL estimator, and $H_{\mathbf{x}_i}$ is the Hessian of the pdf computed at $\mathbf{x}_i$.

Let the error in the approximation of P_i be $e_{P_i}^{(\mathrm{KL})} := |P_i - P_i^{(\mathrm{KL})}|$. Then:

$$\frac{|\lambda_i^{min}|}{3} d 2^{d-1} \varepsilon_i^{d+2} \leq e_{P_i}^{(\mathrm{KL})} \leq \frac{|\lambda_i^{max}|}{3} d 2^{d-1} \varepsilon_i^{d+2}, \quad (18)$$

where $\lambda^{min,max}$ denote respectively the minimum and maximum eigenvalues of the Hessian. The lower bound can thus vanish. Concerning the upper bound, note the dependence on the dimension d as well as on the maximum eigenvalue, which can be very large in the presence of non-uniformity of the pdf.

2. *Error in entropy estimation*

Let $H^{(\mathrm{KL})}$ denote the KL entropy estimate. After some derivation and by introducing the approximation of the KL estimator in the ball, it holds:

$$\psi(k) - \psi(N) = \frac{1}{N} \sum_i^N \log(p_i) + \frac{d}{N} \sum_i^N \log(2\varepsilon_i) + \frac{1}{N} \sum_i^N \log\left(1 + \frac{h_i}{P_i^{(\mathrm{KL})}}\right), \quad (19)$$

where $h_i = \frac{1}{2} \int_{\mathcal{B}(\varepsilon,\mathbf{x}_i)} (\boldsymbol{\xi} - \mathbf{x}_i)^T H_{\mathbf{x}_i} (\boldsymbol{\xi} - \mathbf{x}_i) \, d\boldsymbol{\xi}$. After some algebra, the following expression for the entropy estimation is obtained:

$$H - H^{(\mathrm{KL})} = e_S + \frac{1}{N} \sum_i^N \log\left(1 + \frac{h_i}{P_i^{(\mathrm{KL})}}\right), \quad (20)$$

where e_S is the statistical error due to the MC approximation, and the last term on the right hand side is the analytical error.

Eq.(38) and the standard log-inequality (see Appendix) allows to state the upper and lower bounds for

the error:

$$|H - H^{(\mathrm{KL})}| \leq e_S + \frac{d\ 2^{d-1}}{3N} \sum_i^N \frac{|\lambda_i^{max}|}{P_i^{(\mathrm{KL})}} \varepsilon_i^{d+2}. \quad (21)$$

$$|H - H^{(\mathrm{KL})}| \geq \left| e_S + \frac{d\ 2^{d-1}}{3N} \sum_i^N \frac{\lambda_i^{min}\ \varepsilon_i^{d+2}}{P_i^{(\mathrm{KL})} + \frac{|\lambda_i^{max}| d 2^{d-1}}{3} \varepsilon_i^{d+2}} \right| \quad (22)$$

The error is thus bounded by the statistical error and an analytical contribution. If the distribution is piecewise linear, then the analytical contribution vanishes ($\lambda_i^{max} = 0, \forall i$ in Eq.(48)) since the Hessian vanishes. This corresponds to a particular case that hardly represents realistic probability distributions. The lower bound Eq.(46) can vanish for particular distributions. The analysis of the expressions reveals that, given a target distribution, the error in the entropy estimate can be significant in the presence of non-uniformity (high λ^{max}), and when the dimension (d) is high.

B. The kpN estimator error analysis

The analysis presented for the KL estimator is repeated in this section for the kpN estimator. The error analysis shows that the choice made allows to keep the structure of the KL estimator while mitigating the analytical contribution to the error. The main difference is in the approximation of the probability mass.

1. Error in the approximation of the probability mass

The details of the computation are presented in the Appendix. The main difference with respect to the KL estimator consists in the fact that, by constructing a Gaussian osculatory interpolant (empirically identified by using $p-$neighbors), an approximation of the Hessian of the distribution is obtained. This estimate can be rough, but is beneficial in two cases: when the probability distributions are in a high dimensional space, or the pdf in $\mathcal{B}(\varepsilon, \mathbf{x}_i)$ exhibits non-uniformity.

The probability mass approximation in the kpN estimator is denoted by $P_i^{(G)}$ and it is defined as:

$$P_i^{(G)} = P_i^{(\mathrm{KL})} + \frac{p(\mathbf{x}_i)}{2g(\mathbf{x}_i)} \int_{\mathcal{B}(\varepsilon,\mathbf{x}_i)} (\boldsymbol{\xi} - \mathbf{x}_i)^T \left[\nabla\nabla g|_{\mathbf{x}_i}\right] (\boldsymbol{\xi} - \mathbf{x}_i)\, d\boldsymbol{\xi}, \quad (23)$$

so that is it the sum of the probability mass of the KL estimator and a term that approximate the Hessian of the distribution. The error estimate is:

$$e_{P_i}^{(G)} = \frac{1}{2} \int_{\mathcal{B}(\varepsilon,\mathbf{x}_i)} (\boldsymbol{\xi} - \mathbf{x}_i)^T \left[\nabla\nabla R|_{\mathbf{x}_i}\right] (\boldsymbol{\xi} - \mathbf{x}_i)\ d\boldsymbol{\xi}, \quad (24)$$

where R is the difference between the target distribution and its gaussian approximation inside the box.

2. Error in the approximation of the entropy

By repeating the same analysis as for the KL estimator, the following upper and lower bounds are obtained:

$$|H - H^{(G)}| \geq \left| e_S + \frac{d\ 2^{d-1}}{3N} \sum_i^N \frac{\zeta_i^{min}\ \varepsilon_i^{d+2}}{P_i^{(\mathrm{KL})} + \frac{|\zeta_i^{max}| d 2^{d-1}}{3} \varepsilon_i^{d+2}} \right|, \quad (25)$$

$$|H - H^{(G)}| \leq e_S + \frac{d\ 2^{d-1}}{3N} \sum_i^N \frac{|\zeta_i^{max}|}{P_i^{(\mathrm{KL})}} \varepsilon_i^{d+2}, \quad (26)$$

where $\zeta^{min,max}$ are the maximum and minimum eigenvalues of the Hessian of the residual R.

Let us remark that the behaviour is the same for the KL estimator and the kpN estimator in terms of functional dependence with respect to the space dimension. However, by approximating the Hessian (avoiding a bad choice of k and p is important to this end), ζ_i^{max} can be significantly lower than λ_i^{max}. This has two potential advantages: first, that in the presence of non-uniformity, the upper bound on the kpN error is smaller; and second that, even if correlations are not significant, a lower ζ_i^{max} results in a lower rate of increase of error with increasing dimensions.

IV. NUMERICAL TESTCASES

In this section, the numerical experiments are presented. The first test case aims at validating the proposed approach against analytical results, in simple settings.

Then, several relevant properties of the methods are investigated in more complicated settings, that frequently occur when realistic datasets are considered. First, the robustness in dimension increase is investigated. Then, the entropy estimation in presence of functional dependency leading to high correlation is shown.

A. Analysis of estimator: effect of k, p, and N

To assess the effect of the parameters k, p, and N, in the kpN estimator, three probability distributions in two, three, and four dimensions are considered. A summary of the these distributions is presented in Table I. For all the three distributions, the number of samples N are varied from 1000 to 32000, k is varied from 1 to 10, and p/N is varied from 0.01 to 0.10. For each set of these parameters an $N_{\mathrm{ens}} = 1000$ independent kpN entropy estimates are calculated and the corresponding mean and variance of the error with respect to the analytically known true entropy is calculated. These results for the 2-D Gaussian, 3-D Gamma, and 4-D Beta distributions are shown in Figures 3, 4, and 5, respectively.

TABLE I. Summary of the distributions for the analysis of k, p, and N

Distribution	**Parameters**							
2-D Multivariate Normal (correlation coefficient r)			mean $[0.0, 0.0]$		variance $[1.0, 1.0]$		r 0.5	
3-D Gamma distribution (Independent along each dimension)		k_1 1.5	θ_1 2.0	k_2 3.0	θ_2 2.5	k_3 20.0	θ_3 1.0	
4-D Beta distribution (Independent along each dimension)	α_1 2.0	β_1 2.0	α_2 2.0	β_2 5.0	α_3 0.5	β_3 0.5	α_4 5.0	β_4 1.0

From these plots it is observed that the variance of the error decreases with increasing N as expected. Furthermore, the variance appears to be high for $k = 1, 2$ and then lower and approximately invariant with increasing k. This is consistent with the behaviour of the KL estimator [11]. Recall that the parameter p is reflective of the length-scale of changes in the probability density. For a Gaussian distribution, it is clear that a higher p will result in lower error as the local Gaussian approximations will better approximate the true distribution. This is observed in Figure 3a. A similar behaviour is observed for the Gamma distribution (Figure 4a), but for the Beta distribution (Figure 5a) a clear optimal range of p/N varying from 0.01 to 0.05 can be identified. The length-scale of the density variation will in general not be known *a priori* (especially in higher dimensions where the samples are hard to visualise) and consequently a large p/N should be avoided. From Figures 3a, 4a, and 5a, it is observed that unless a particularly bad combination of the N, k, and p, parameters – specifically low N, high k, and small p/N – is chosen, the errors across the entire spectrum of parameter variations are less than 10%. Overall, based on the performance of the estimator across the three significantly different distributions considered, k is recommended to be chosen between 3 and 5, and p/N between 0.02 and 0.05.

B. Dimension increase

The properties of the kpN estimator regarding robustness to dimension increase are investigated. For all these tests $N = 10000$, $k = 4$, $p/N = 0.02$ are fixed. The method is compared to the standard KL estimator in multi-dimensional uncorrelated Gaussian, Gamma and Beta distributions. The dimension ranges from 4 to 80. For all the cases, the quantity of interest is the relative error, defined as $e = \frac{|H^* - H|}{H^*}$, where H^* is the analytical value. The distributions used to test the method are quite regular and smooth. Moreover, no correlation is considered, the only focus being the behaviour with respect to the dimension increase. For all the tests, the computations were repeated $N_{\text{ens}} = 1000$ times, and corresponding mean-values and variances are reported.

1. Multi-dimensional Gaussian

The first test case is the entropy computation of a multi-dimensional Gaussian:

$$p^* = \prod_i^d g_i(\mu_i, \sigma_i^2), \tag{27}$$

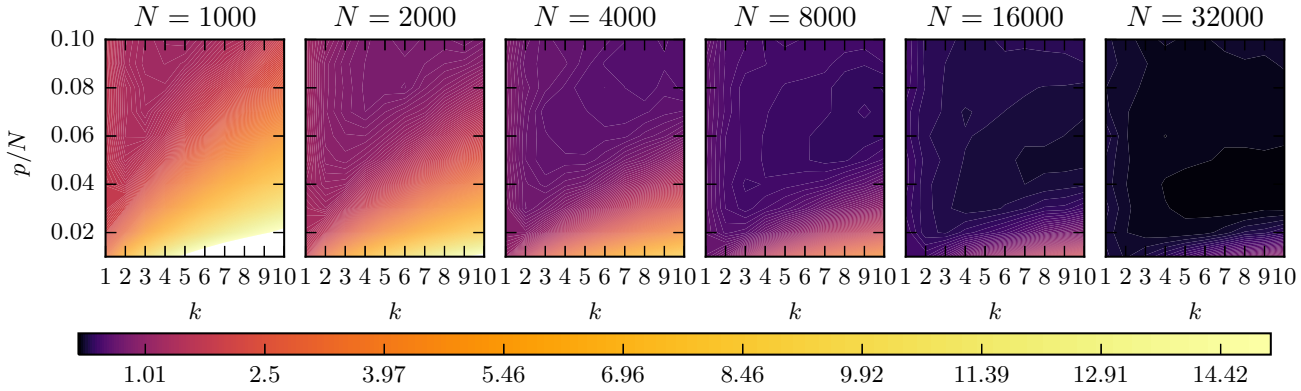
where d is the space dimension, $g_i(\mu_i, \sigma_i)$ is a univariate Gaussian distribution with mean μ_i and variance σ_i^2, and $\mu_i = 0$, $\forall i$. The variance σ_i^2 ranges uniformly in $[0.2, 2]$, *i.e.* $\sigma_i^2 = 1.8(i-1)/(d-1) + 0.2$. The results are summarised in Fig.6. The relative error at low dimension is higher than that of the KL estimator. This is due to the fact that the parameters adopted are not optimal for this distribution, given the number of sample (a higher value of p/N would provide a better result). The kpN error is significantly smaller when the dimension increases: namely, at dimension $d = 80$, it has an error which is less than 10%, while the KL estimator has an error which is about three times larger, despite the fact that the probability distribution is quite regular.

2. Multi-dimensional Gamma

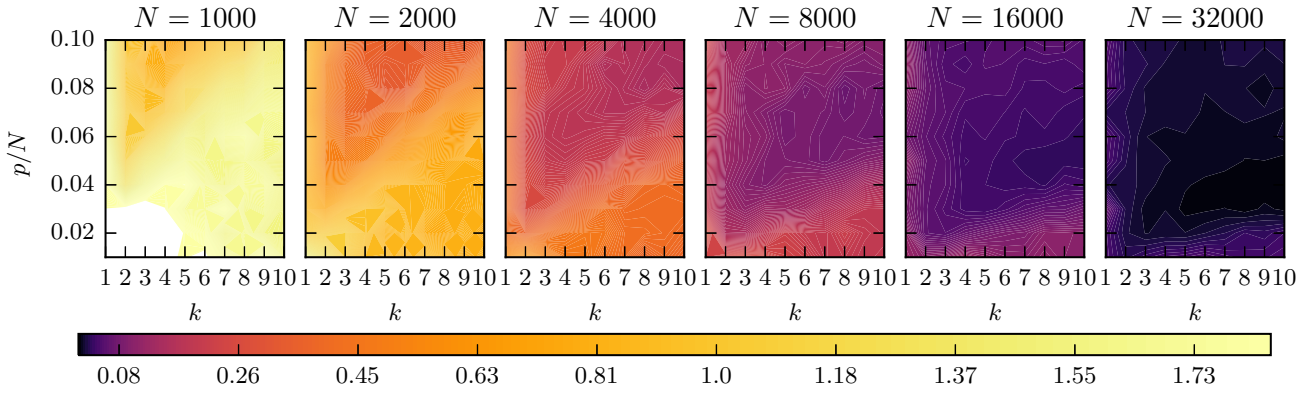
The case of a multivariate Gamma distribution is commented. Similarly to earlier case, the distribution is defined as a product of univariate distributions:

$$p^* = \prod_i^d \gamma_i(k_i, \theta_i), \tag{28}$$

where k_i and θ_i are the shape and scale parameters of the Gamma distribution $\gamma_i(k_i, \theta_i)$. The shape parameter k_i varies uniformly in $[0.5, 5.0]$ while the scale parameter θ_i varies in $[1.0, 2.0]$. The results are shown in Fig.7. For this case, the kpN estimator always outperforms the KL estimator. Note that the error is not necessarily monotonic with respect to the dimension of the space. This depends on the particular nature of the distribution as well as on the parameters k and p adopted. Nonetheless, the kpN error is less than 5% across the entire range of dimensions considered, while the KL error grows up to 35%.



(a) % relative error in the entropy estimate



(b) % variance of relative error in the entropy estimate

FIG. 3. (Color online) kpN entropy estimate for 2D-Gaussian distribution with correlation $r = 0.5$ (see Table I)

3. *Multi-dimensional Beta*

The last test case shown concerns the entropy estimation for a multivariate Beta distribution of the form:

$$p^* = \prod_i^d B_i(\alpha_i, \beta_i), \tag{29}$$

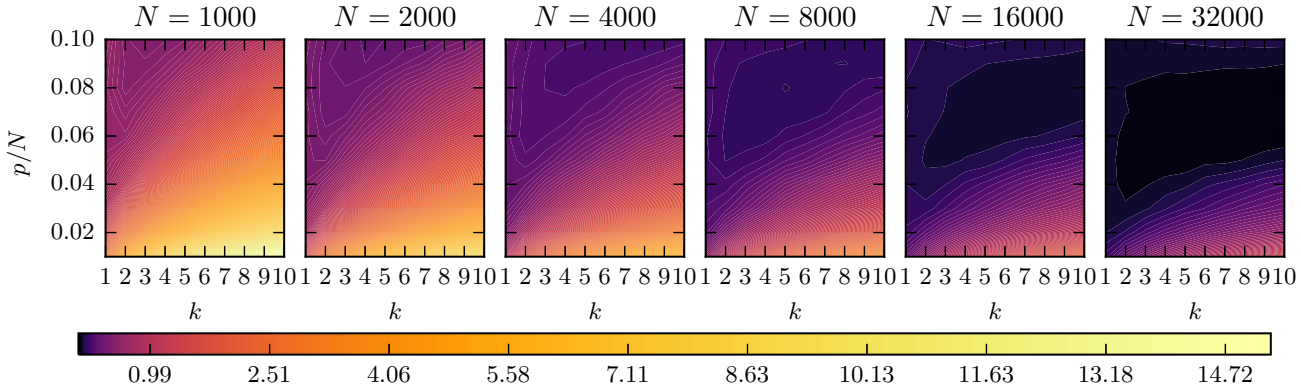
where $B_i(\alpha_i, \beta_i)$ is a Beta distribution with shape parameters α_i and β_i, α_i varies in $[0.5, 5.0]$, and β_i varies in $[0.5, 5.0]$. The results of the kpN and KL entropy estimates are shown in Figure 8. This test appears to be most critical as, on average, the errors on both KL and kpN estimates are higher when compared to the previous Gaussian and Gamma distributions. This may partly be due to pathological nature of the Beta distribution for particular choices of the α and β parameters (for example $\alpha = \beta = 0.5$), or (although unclear why) due to the fact that the Beta distribution has only a finite support over [0.0,1.0] in all dimensions. From Figure 8, the error of the kpN estimator is always less than 20% whereas the KL estimate has a relative error of about 150%, which is almost one order of magnitude higher.

4. *Discussion*

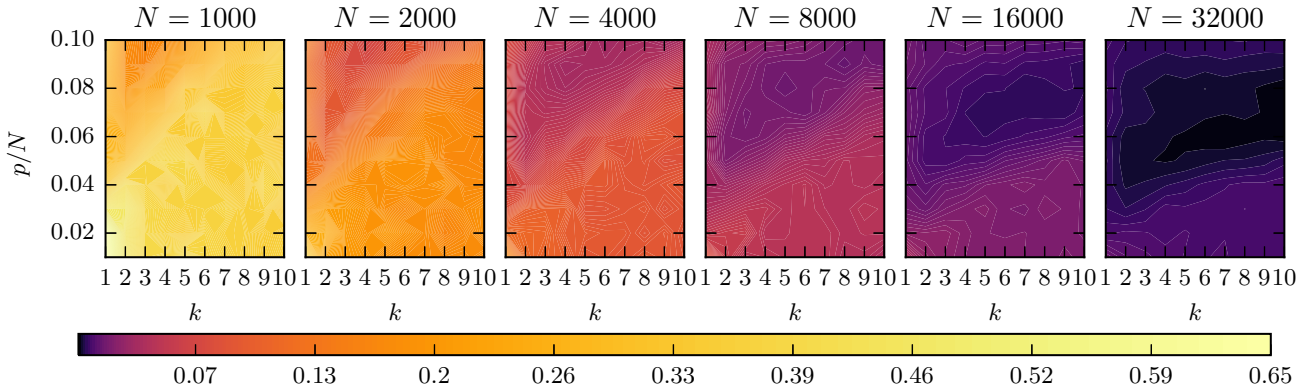
The three tests presented aim at investigating the behaviour of the estimator with respect to the dimension of the space. The error of the KL estimator is monotonic and grows quite fast, because the analytical contribution to the error grows significantly with the dimension, when the number of sample is kept fixed. On the contrary, the kpN estimator proposed manages to mitigate this error by providing a rough estimate of the Hessian of the distribution in each box. The proposed kpN estimator is more robust to the dimension increase, or, conversely, given a certain dimension of the space, it allows to estimate the entropy by using a smaller number of samples. This feature is particularly appealing when dealing with the analysis of realistic datasets.

C. Functional dependency and correlation

Another interesting aspect that occurs frequently when realistic applications are considered is the possible presence of correlation. In this section, the robustness of the



(a) % relative error in the entropy estimate



(b) % variance of relative error in the entropy estimate

FIG. 4. (Color online) kpN entropy estimate for 3D-Gamma distribution with shape parameters shown in Table I

entropy estimators is investigated with respect to correlations and functional dependencies between the components of a random vector.

1. *Varying correlation*

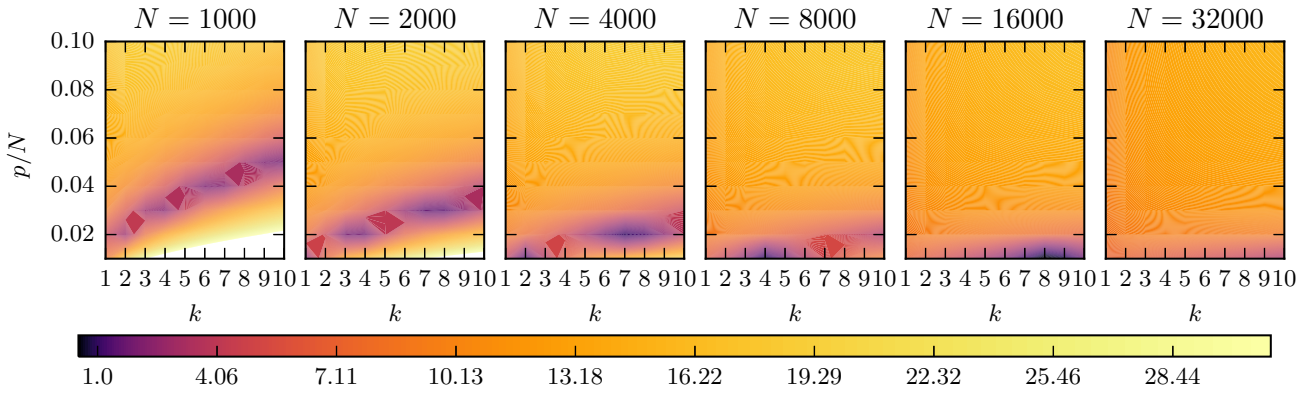
To study the effect of varying correlations the entropy of a 20 dimensional zero-mean multivariate Gaussian distribution with the following covariance matrix is considered:

$$C_{ij} = \begin{cases} 1 & \text{if } i = j \\ r & \text{if } i \neq j \end{cases} \qquad i, j = 1, 2, \ldots, 20 \tag{30}$$

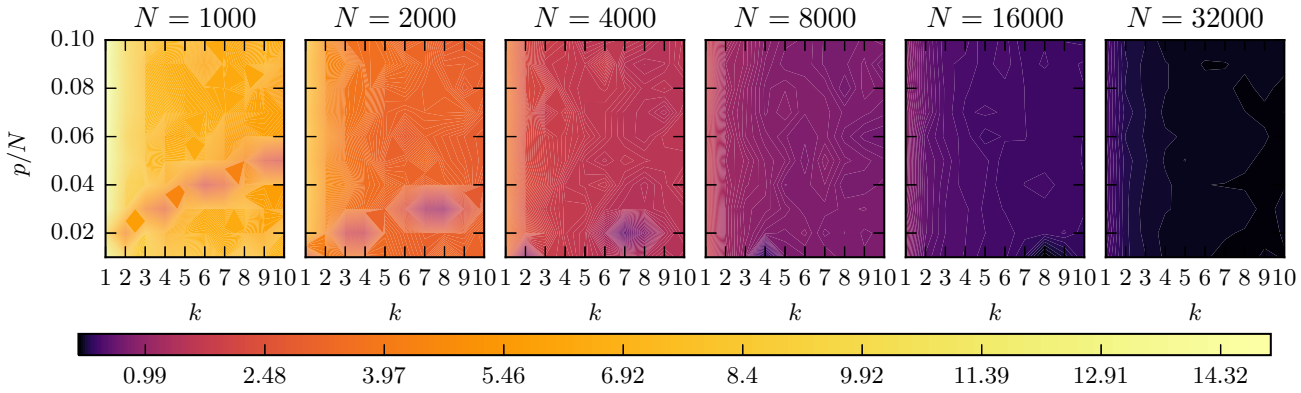
where $0 \leq r < 1$ is the identical coefficient of correlation between any two components of the random vector under consideration. Figure 9 shows the results of the KL and kpN estimators for r varying between 0 and 0.99. For both the estimators $k = 4$ and $N = 10,000$. For the kpN estimator four different values of p/N are chosen in $\{0.02, 0.03, 0.04, 0.05\}$. For the kpN estimator it is observed that the error decreases as p/N increases. Nonetheless, for the lowest value of chosen $p/N = 0.02$, the kpN estimator performs significantly better than the KL estimator. The behaviour of the two estimators is particularly different when the correlations are high, *i.e.* when $r > 0.9$. It is noticeable that in this region the KL estimator shows a rapid increase in the error. On the contrary, the kpN estimator shows a more stable behaviour. It should be noted that in lower dimensions (< 10), the differences between the KL and kpN estimators are less significant even in the presence of large correlations. The case of extreme correlations, for example $r > 0.99$, can be modelled as near functional dependencies that are explored in the next section.

2. *Functional dependencies*

In this section, the kpN method is compared to the KL method for fixed parameters: $N = 5000$, $k = 4$, $p/N = 0.02$. A simple test case is proposed to study near functional dependencies between random variables: the entropy of a Gaussian distribution on a linear manifold



(a) % relative error in the entropy estimate



(b) % variance of relative error in the entropy estimate

FIG. 5. (Color online) kpN entropy estimate for 4D-Beta distribution with shape parameters shown in Table I

is computed, with different levels of noise. The system is

$$y = tx + \nu, \tag{31}$$

where x is a normal random variable with zero mean and unit variance, $t \in \mathbb{R}^+$ is a positive scalar, and ν is a normal random variable with zero mean and variance σ_n^2.

The system output y is observed at discrete times $t_i = \{1, \ldots, 9\}$, providing $y_i = y(t_i)$. The objective is to estimate the entropy of the joint probability distribution of $[x, y_1, \ldots, y_i]$ for increasing i. For this test case, two different levels of noise are considered, namely $\sigma_n^2 = \{10^{-1}, 10^{-3}\}$. The joint dimension increases up to $d = 10$. The results (in terms of absolute error and variance) are shown in Fig.10 for $\sigma_n^2 = 10^{-1}$ and $\sigma_n^2 = 10^{-3}$. When the dimension is low, the performances of the KL estimator and that of the proposed kpN estimator are comparable, *i.e.* no significant difference in error is observed in terms of both the means and the variances. When the dimension increases, depending on the level of noise, the KL estimator starts deviating from the true estimate, whereas the proposed kpN estimator provides a significantly better result. The higher the noise level, the better is the behaviour of the classical KL estimator. This apparently paradoxical result can be explained by considering the analytical heuristics proposed. When the level of noise is higher, the samples are less correlated and, thus, the maximum eigenvalue of the Hessian is, on average, smaller. The joint distribution being more regular, a better entropy estimate is obtained by the classical KL estimator. The kpN estimator, on the other hand, is more robust to variations in noise-levels as based on the p-neighbors the covariance of the local Gaussian approximation adjusts accordingly.

V. CONCLUSIONS AND PERSPECTIVES

A new k-nearest neighbor based entropy estimator, that is efficient in high dimensions and in the presence of large non-uniformity, is proposed. The proposed idea relies on the introduction of a Gaussian osculatory interpolation, which in-turn is based on an empirical evaluation of p-nearest neighbors. By this introduction, the local non-uniformity of the underlying probability distribution is captured, while retaining all the appealing com-

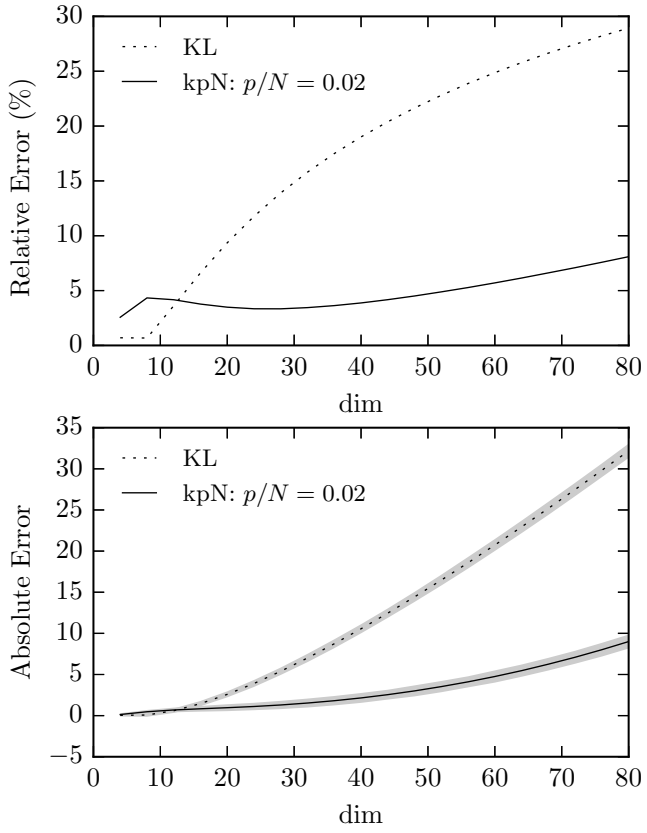


FIG. 6. Error analysis of a multivariate gaussian: shaded regions indicate ±10 standard deviation in the error estimate

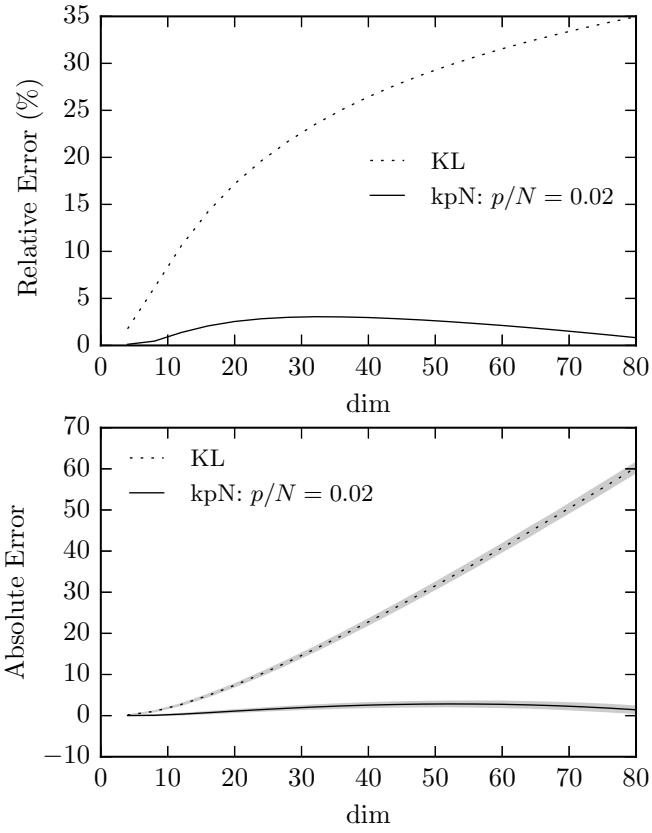


FIG. 7. Error analysis of a multivariate gamma: shaded regions indicate ±10 standard deviation in the error estimate

putational advantages of classical kNN estimators. The robustness of the new estimator is tested for a variety of distributions – ranging from infinite support Gamma distributions to finite support Beta distributions – in successively increasing dimensions (up to 80). Furthermore, a case of direct functional relationship leading to high correlations between the components of a random variable is considered. Across all the tests, the new estimator is shown to consistently outperform the classical kNN estimator.

The main perspective of the current work is that the proposed estimator can be used as a building block to construct estimators for other quantities of interest such as mutual information, particularly in high dimensions. Another perspective is the development of strategies to automatically adapt p based on properties of the cloud of local samples.

VI. APPENDIX

In this Appendix, the details of the error heuristics are presented. Let us recall the notation and the main hypotheses. The ε-ball is denoted by $\mathcal{B}(\varepsilon, \mathbf{x}_i) = [\mathbf{x}_i - \varepsilon_i, \mathbf{x}_i + \varepsilon_i]^d$. Let $p_i(\boldsymbol{\xi}) \in C^2(\mathcal{B}(\varepsilon, \mathbf{x}_i))$ be the probability density in $\mathcal{B}(\varepsilon, \mathbf{x}_i)$. The probability mass in $\mathcal{B}(\varepsilon, \mathbf{x}_i)$ is $P_i = \int_{\mathcal{B}(\varepsilon, \mathbf{x}_i)} p_i \, d\boldsymbol{\xi}$.

A. KL estimator error analysis

The error of the KL estimator is analysed. First, the error on the probability mass in a generic $\mathcal{B}(\varepsilon, \mathbf{x}_i)$ is computed, and the result is used to compute the error on the entropy.

1. *Error in the approximation of the probability mass*

The analytical contribution to the error is due to the approximation of the probability mass P_i. Consider a Taylor expansion of p_i centered around x_i:

$$P_i = \int_{\mathcal{B}(\varepsilon, \mathbf{x}_i)} p(\mathbf{x}_i) + (\boldsymbol{\xi} - \mathbf{x}_i) \cdot \nabla p|_{\mathbf{x}_i} + \frac{1}{2}(\boldsymbol{\xi} - \mathbf{x}_i)^T H_{\mathbf{x}_i} (\boldsymbol{\xi} - \mathbf{x}_i) + o(|\boldsymbol{\xi} - \mathbf{x}_i|^2) \, d\boldsymbol{\xi}, \quad (32)$$

where $H_{\mathbf{x}_i}$ is the Hessian computed in x_i. The first term of the series yields the KL approximation $P_i^{(\mathrm{KL})}$, the sec-

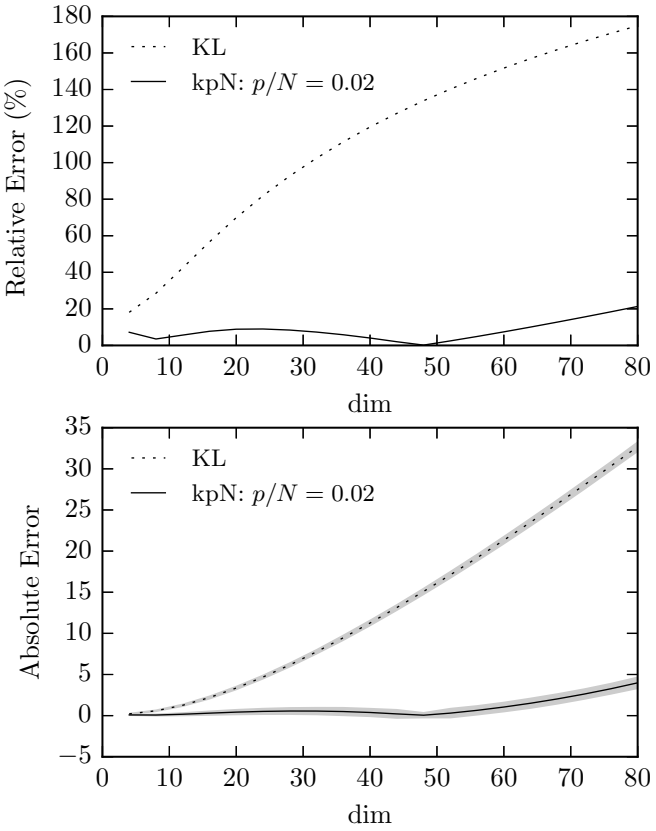


FIG. 8. Error analysis of a multivariate beta: shaded regions indicate ±10 standard deviation in the error estimate

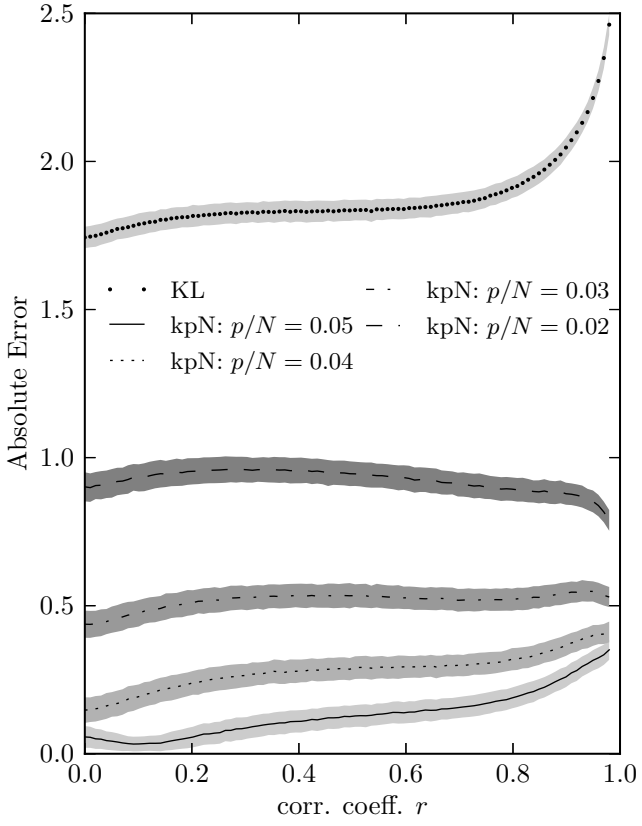


FIG. 9. 20-dimensional Normal distribution with varying correlation coefficient r. Covariance matrix elements are $C_{ij} = 1$ for $i = j$ and $C_{ij} = r$ for $i \neq j$ where $i, j = 1, 2, \ldots, 20$. Parameter values are $k = 4$ for both kpN and KL estimators. Shaded regions indicate ±1 standard deviation in the error estimate

ond term vanishes since it is the integral of an even function over a symetric interval, the third term represent the error of the approximation:

$$P_i \approx P_i^{(\mathrm{KL})} + \frac{1}{2}\int_{\mathcal{B}(\varepsilon,\mathbf{x}_i)} (\boldsymbol{\xi}-\mathbf{x}_i)^T H_{\mathbf{x}_i}(\boldsymbol{\xi}-\mathbf{x}_i)\; d\boldsymbol{\xi}, \quad (33)$$

obtained by discarding the higher order terms. Since $P_i \geq 0$, let us make the hypothesis that this hold even for the truncated approximation, *i.e.*:

$$\frac{\left|\frac{1}{2}\int_{\mathcal{B}(\varepsilon,\mathbf{x}_i)}(\boldsymbol{\xi}-\mathbf{x}_i)^T H_{\mathbf{x}_i}(\boldsymbol{\xi}-\mathbf{x}_i)\; d\boldsymbol{\xi}\right|}{P_i^K} \leq 1. \quad (34)$$

The integral $h_i = \frac{1}{2}\int_{\mathcal{B}(\varepsilon,\mathbf{x}_i)}(\boldsymbol{\xi}-\mathbf{x}_i)^T H_{\mathbf{x}_i}(\boldsymbol{\xi}-\mathbf{x}_i)\; d\boldsymbol{\xi}$ is estimated. A standard result on the quadratic forms is used:

$$\lambda_i^{min}|\boldsymbol{\xi}-\mathbf{x}_i|^2 \leq (\boldsymbol{\xi}-\mathbf{x}_i)^T H_{\mathbf{x}_i}(\boldsymbol{\xi}-\mathbf{x}_i) \leq \lambda_i^{max}|\boldsymbol{\xi}-\mathbf{x}_i|^2, \quad (35)$$

where $\lambda_i^{min,max}$ are the minimum and maximum eigenvalues of $H_{\mathbf{x}_i}$. Then, the bounds on h_i are simply obtained by computing the integral over $\mathcal{B}(\varepsilon, \mathbf{x}_i)$:

$$\int_{\mathcal{B}(\varepsilon,\mathbf{x}_i)} |\boldsymbol{\xi}-\mathbf{x}_i|^2\; d\boldsymbol{\xi} = \sum_j^d \int_{\mathcal{B}(\varepsilon,\mathbf{x}_i)} (\xi_j - x_{i,j})^2\; d\boldsymbol{\xi}. \quad (36)$$

By virtue of the symetry of the ball, this integral can be computed for just one j and then multiplied by d. Let $\mathcal{B}(\varepsilon, \mathbf{x}_i) = [x_{i,j}-\varepsilon_i, x_{i,j}+\varepsilon_i] \times [x_{i,k}-\varepsilon_i, x_{i,k}+\varepsilon_i]^{d-1}$, $k \neq j$. It holds:

$$\int_{\mathcal{B}(\varepsilon,\mathbf{x}_i)} (\xi_j - x_{i,j})^2\; d\boldsymbol{\xi} = (2\varepsilon)^{d-1}\int_{-\varepsilon}^{\varepsilon} \eta^2\; d\eta = (2\varepsilon)^{d-1}\frac{2}{3}\varepsilon^3. \quad (37)$$

By putting together the bounds in Eq.(35) and the result in Eq.(37), the error approximation is obtained. Let $e_{P_i}^{(\mathrm{KL})} := |P_i - P_i^{(\mathrm{KL})}|$. Then:

$$\frac{|\lambda_i^{min}|}{3} d 2^{d-1}\varepsilon_i^{d+2} \leq e_{P_i}^{(\mathrm{KL})} \leq \frac{|\lambda_i^{max}|}{3} d 2^{d-1}\varepsilon_i^{d+2}. \quad (38)$$

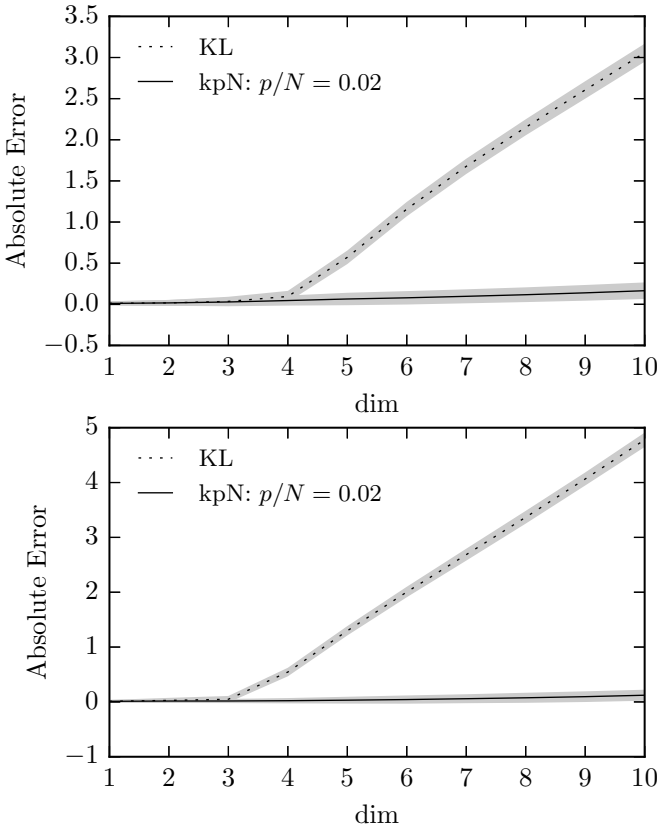


FIG. 10. Error analysis for a gaussian on a linear manifold. Absolute error with respect to the dimension, level of noise: $\sigma_n^2 = 10^{-1}$ (left) and $\sigma_n^2 = 10^{-3}$ (right). Shaded regions indicate ± 3 standard deviation in the error estimate

2. Error in the approximation of the entropy

The error on the entropy estimate is obtained by a derivation of the KL estimator. The KL estimator is obtained by equating $\mathbb{E}\{\log(P)\} = \psi(k) - \psi(N)$. Let $H^{(\mathrm{KL})}$ denote the entropy estimated by using the KL estimator. We write:

$$\psi(k) - \psi(N) = \frac{1}{N}\sum_i^N \log\left(P_i^{(\mathrm{KL})} + h_i\right). \tag{39}$$

The properties of the logarithm are used, leading to:

$$\psi(k) - \psi(N) = \frac{1}{N}\sum_i^N \log\left(P_i^{(\mathrm{KL})}\right) + \frac{1}{N}\log\left(1 + \frac{h_i}{P_i^{(\mathrm{KL})}}\right). \tag{40}$$

The KL approximation of P_i^K is introduced:

$$\psi(k) - \psi(N) = \frac{1}{N}\sum_i^N \log(p_i) + \frac{d}{N}\sum_i^N \log(2\varepsilon_i) + \frac{1}{N}\sum_i^N \log\left(1 + \frac{h_i}{P_i^{(\mathrm{KL})}}\right). \tag{41}$$

After some algebra, it holds:

$$H - H^K = e_S + \frac{1}{N}\sum_i^N \log\left(1 + \frac{h_i}{P_i^{(\mathrm{KL})}}\right), \tag{42}$$

where e_S is the statistical error due to the MC approximation, and the last term on the right hand side is the analytical error.

The use of the result presented in Eq.(38) and of a standard log-inequality allows to state upper and lower bounds for the error.

Indeed, the hypothesis in Eq.(34) allows to make use of the following:

$$\frac{x}{1+x} \leq \log(1+x) \leq x. \tag{43}$$

After having set $x = h_i/P_i^K$, we have:

$$\frac{h_i}{h_i + P_i^{(\mathrm{KL})}} \leq \log\left(1 + \frac{h_i}{P_i^{(\mathrm{KL})}}\right) \leq \frac{h_i}{P_i^{(\mathrm{KL})}}. \tag{44}$$

In order to get a lower bound, the left hand side is studied. It holds:

$$\frac{h_i}{h_i + P_i^{(\mathrm{KL})}} \geq \frac{\min(h_i)}{\max(h_i) + P_i^{(\mathrm{KL})}} = \frac{\lambda_i^{min} d\ 2^{d-1}\varepsilon_i^{d+2}}{\lambda_i^{max} d\ 2^{d-1}\varepsilon_i^{d+2} + 3P_i^K}. \tag{45}$$

The use of this results allows to state the lower bound for the error:

$$|H - H^{(\mathrm{KL})}| \geq \left| e_S + \frac{d\ 2^{d-1}}{3N}\sum_i^N \frac{\lambda_i^{min}\varepsilon_i^{d+2}}{P_i^{(\mathrm{KL})} + \frac{|\lambda_i^{max}|d2^{d-1}}{3}\varepsilon_i^{d+2}} \right|. \tag{46}$$

In order to derive the upper bound, the right hand side of the logarithmic inequality is studied:

$$\frac{h_i}{P_i^{(\mathrm{KL})}} \leq \frac{\max(h_i)}{P_i^{(\mathrm{KL})}} = \frac{\lambda_i^{max} d 2^{d-1}\varepsilon_i^{d+2}}{3P_i^{(\mathrm{KL})}}. \tag{47}$$

By using this, the upper bound reads:

$$|H - H^{(\mathrm{KL})}| \leq e_S + \frac{d\ 2^{d-1}}{3N}\sum_i^N \frac{|\lambda_i^{max}|}{P_i^{(\mathrm{KL})}}\varepsilon_i^{d+2}. \tag{48}$$

B. Analysis of the kpN estimator

As commented above, the main difference is in the approximation of the probability mass in $\mathcal{B}(\varepsilon, \mathbf{x}_i)$. In particular, an osculatory interpolation with an empirically estimated multivariate Gaussian is constructed.

1. Error in the approximation of the probability mass

The probability density distribution inside the ball is approximated by:

$$p(\boldsymbol{\xi}) = p(\mathbf{x}_i)\frac{g(\boldsymbol{\xi})}{g(\mathbf{x}_i)} + R(\boldsymbol{\xi}), \tag{49}$$

where $g := \exp\left(-\frac{1}{2}(\boldsymbol{\xi}-\boldsymbol{\mu})^T\mathbf{S}^{-1}(\boldsymbol{\xi}-\boldsymbol{\mu})\right)$, where $\boldsymbol{\mu}, \mathbf{S}$ are the empirically evaluated mean and covariance, R is the residual of the approximation. Since $p(\boldsymbol{\xi} = \mathbf{x}_i) = p(\mathbf{x}_i)$ by construction of the approximation, it follows $R(\mathbf{x}_i) = 0$. The Taylor expansion of the probability density distribution centred around $\mathbf{x}_i$ is computed for the gaussian approximation:

$$p(\boldsymbol{\xi}) \approx p(\mathbf{x}_i) + (\boldsymbol{\xi}-\mathbf{x}_i)\cdot\left(\frac{p(\mathbf{x}_i)}{g(\mathbf{x}_i)}\nabla g|_{\mathbf{x}_i} + \nabla R|_{\mathbf{x}_i}\right) + \frac{1}{2}(\boldsymbol{\xi}-\mathbf{x}_i)^T K_{\mathbf{x}_i}(\boldsymbol{\xi}-\mathbf{x}_i), \tag{50}$$

where $K_{\mathbf{x}_i} = \frac{p(\mathbf{x}_i)}{g(\mathbf{x}_i)}\nabla\nabla g|_{\mathbf{x}_i} + \nabla\nabla R|_{\mathbf{x}_i}$ is the Hessian computed for the gaussian approximation.

The expression is used to compute the probability mass. Remark that, as before, the linear contribution vanishes identically due to the symetry of the ball. It holds, at second order:

$$P_i = \int_{\mathcal{B}(\varepsilon,\mathbf{x}_i)} p(\boldsymbol{\xi})\, d\boldsymbol{\xi} \approx P_i^{(\mathrm{KL})} + \frac{1}{2}\int_{\mathcal{B}(\varepsilon,\mathbf{x}_i)} (\boldsymbol{\xi}-\mathbf{x}_i)^T H_{\mathbf{x}_i}(\boldsymbol{\xi}-\mathbf{x}_i)\, d\boldsymbol{\xi}, \tag{51}$$

where H denotes the Hessian of the target distribution. On the other hand:

$$P_i = \int_{\mathcal{B}(\varepsilon,\mathbf{x}_i)} p(\boldsymbol{\xi})\, d\boldsymbol{\xi} \approx P_i^{(\mathrm{KL})} + \frac{1}{2}\int_{\mathcal{B}(\varepsilon,\mathbf{x}_i)} (\boldsymbol{\xi}-\mathbf{x}_i)^T K_{\mathbf{x}_i}(\boldsymbol{\xi}-\mathbf{x}_i)\, d\boldsymbol{\xi}. \tag{52}$$

The expression for $K_{\mathbf{x}_i}$ is introduced, allowing to understand what the gaussian approximation does in terms of approximating the mass:

$$P_i = P_i^{(\mathrm{KL})} + \frac{1}{2}\int_{\mathcal{B}(\varepsilon,\mathbf{x}_i)} (\boldsymbol{\xi}-\mathbf{x}_i)^T\left[\frac{p(x_i)}{g(x_i)}\nabla\nabla g|_{\mathbf{x}_i}\right](\boldsymbol{\xi}-\mathbf{x}_i)\, d\boldsymbol{\xi} + \frac{1}{2}\int_{\mathcal{B}(\varepsilon,\mathbf{x}_i)} (\boldsymbol{\xi}-\mathbf{x}_i)^T\left[\nabla\nabla R|_{\mathbf{x}_i}\right](\boldsymbol{\xi}-\mathbf{x}_i)\, d\boldsymbol{\xi}. \tag{53}$$

What is retained in the present approximation is the first term, the error thus reducing to the last term of the expansion (equate the Taylor expansion Eq.(51) with Eq.(53)). The mass approximation is denoted by $P_i^{(G)}$ and it can be defined as:

$$P_i^{(G)} = P_i^{(\mathrm{KL})} + \frac{p(\mathbf{x}_i)}{2g(\mathbf{x}_i)}\int_{\mathcal{B}(\varepsilon,\mathbf{x}_i)} (\boldsymbol{\xi}-\mathbf{x}_i)^T\left[\nabla\nabla g|_{\mathbf{x}_i}\right](\boldsymbol{\xi}-\mathbf{x}_i)\, d\boldsymbol{\xi}. \tag{54}$$

Roughly speaking, the mass is the sum of the mass obtained by the KL hypothesis plus an additional term that results from the approximation of Hessian of the target distribution by means of the Hessian of the empirically estimated gaussian.

The error is denoted by $e_{P_i}^{(G)} := |P_i - P_i^{(G)}|$:

$$e_{P_i}^{(G)} = \frac{1}{2}\int_{\mathcal{B}(\varepsilon,\mathbf{x}_i)} (\boldsymbol{\xi}-\mathbf{x}_i)^T\left[\nabla\nabla R|_{\mathbf{x}_i}\right](\boldsymbol{\xi}-\mathbf{x}_i)\, d\boldsymbol{\xi}. \tag{55}$$

If the distribution is Gaussian and it is perfectly estimated through the samples, this term vanishes. Remark that, the behaviour of the error as function of the dimension is exactly the same as for the KL estimator, but if the Hessian of the Gaussian estimates the Hessian of the target distribution, the upper bound on the error will be smaller.

2. Error in the approximation of the entropy

The error on the entropy estimate is computed by following exactly the same strategy as for the KL estimator. The upper and lower bounds have the same expression, except that the eigenvalues appearing (namely $\zeta_i^{min,max}$) in the expressions are those of the Hessian of the residual R.

The lower bound reads:

$$|H - H^{(G)}| \geq \left| e_S + \frac{d\ 2^{d-1}}{3N}\sum_i^N \frac{\zeta_i^{min}\varepsilon_i^{d+2}}{P_i^{(\mathrm{KL})} + \frac{|\zeta_i^{max}|d2^{d-1}}{3}\varepsilon_i^{d+2}} \right|. \tag{56}$$

And the upper bound is:

$$|H - H^{(G)}| \leq e_S + \frac{d\ 2^{d-1}}{3N}\sum_i^N \frac{|\zeta_i^{max}|}{P_i^{(\mathrm{KL})}}\varepsilon_i^{d+2}. \tag{57}$$

[1] T. M. Cover and J. A. Thomas, *Elements of information theory* (John Wiley & Sons, 2012).

[2] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. C. Van der Meulen, Int. J. Math. Stat. Sci. **6**, 17 (1997).

[3] B. W. Silverman, *Density estimation for statistics and data analysis*, Vol. 26 (CRC press, 1986).

[4] L. Devroye and L. Györfi, *Nonparametric density estimation: the L1 view*, Wiley series in probability and mathematical statistics (Wiley, 1985).

[5] D. W. Scott, *Multivariate density estimation: theory, practice, and visualization* (John Wiley & Sons, 2015).

[6] P. Hall, in *Mathematical Proceedings of the Cambridge*

Philosophical Society, Vol. 96 (Cambridge Univ Press, 1984) pp. 517–532.
[7] E. Dudewicz and E. Van der Meulen, in *New Perspectives in Theoretical and Applied Statistics*, edited by W. W. E. M.L. Puri, J. Vilaplana (Wiley, New York, 1987).
[8] L. F. Kozachenko and N. N. Leonenko, Probl. Inf. Transm. **23**, 9 (1987).
[9] A. B. Tsybakov and E. Van der Meulen, Scandinavian Journal of Statistics , 75 (1996).
[10] H. Singh, N. Misra, V. Hnizdo, A. Fedorowicz, and E. Demchuk, Amer. J. Math. Management Sci. **23**, 301 (2003).
[11] A. Kraskov, H. Stögbauer, and P. Grassberger, Phys. Rev. E **69**, 066138 (2004).
[12] A. G. Gray and A. W. Moore, in *SDM* (SIAM, 2003) pp. 203–211.
[13] S. Gao, G. V. Steeg, and A. Galstyan, Arxiv (http://arxiv.org/abs/1411.2003) (2015).
[14] A. Genz, J. Comp. Graph. Stat. **1**, 141 (1992).
[15] J. Cunningham, P. Hennig, and S. Lacoste-Julien, Arxiv (http://arxiv.org/abs/1111.6832) (2012).