High-Dimensional Entropy Estimation for Finite Accuracy Data: *R*-NN entropy estimator

Jan Kybic[1]

Center for Machine Perception, Czech Technical University, Prague, Czech Republic
kybic@fel.cvut.cz,
http://cmp.felk.cvut.cz/∼kybic

Abstract. We address the problem of entropy estimation for high-dimensional finite-accuracy data. Our main application is evaluating high-order mutual information image similarity criteria for multimodal image registration. The basis of our method is an estimator based on k-th nearest neighbor (NN) distances, modified so that only distances greater than some constant R are evaluated. This modification requires a correction which is found numerically in a preprocessing step using quadratic programming. We compare experimentally our new method with k-NN and histogram estimators on synthetic data as well as for evaluation of mutual information for image similarity.

1 Introduction

Nonparametric entropy and mutual information estimation from finite number of samples is an important tool in diverse domains such as statistics [1], computational chemistry [2], or measuring information contents of signals such as neural spike trains [3]. For multimodal image registration, mutual information is the image similarity measure of choice [4–6]. Instead of measuring mutual information of scalar image intensities, in some cases it is advantageous to use more complex multidimensional features, such as color, output of spatial filters, texture descriptors, or intensities of neighborhood pixels [7–10]. However, due to the lack of good estimators, most approaches are limited to low dimensions or have to use strong assumptions such as normality. Histogram and kernel estimators do not work well in high dimensions ($d \gtrapprox 5$) [3, 11–13] when the bins are simultaneously too large and almost empty. Nearest neighbor (NN) distance estimators [3, 14–16] look promising, if their two principal problems can be circumvented — the computational complexity of the nearest neighbor search [17] and the artifacts and singularities when applied to finite accuracy (quantized) data. Here we attempt to solve the second problem by using a new estimator called R-NN, combining k-NN and kernel estimator approaches.

1.1 Entropy Estimation

Let us have N samples $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, from an unknown probability density $f(\mathbf{x})$ in a d-dimensional space, $\mathbb{R}^d$. The task is to estimate the Shannon

information entropy $H_{\text{true}}(f) = -\int f(\mathbf{x}) \log \mathbf{x} \, \mathrm{d}\mathbf{x}$. We consider estimators of the form

$$H(X) = \frac{1}{N} \sum_{i=1}^{N} h(\mathbf{x}_i; X) \qquad \text{with} \qquad h(\mathbf{x}_i; X) \approx -\log f(\mathbf{x}_i) \tag{1}$$

If $h(\mathbf{x}_i; X)$ is an unbiased estimator of $-\log f(\mathbf{x}_i)$, then $H(X)$ is an unbiased estimator of $H_{\text{true}}(f)$ [15].

1.2 Kernel estimator

For $H(X)$ to be a non-parametric estimator, $h(\mathbf{x}_i; X)$ must depend only on the neighborhood of $\mathbf{x}_i$ [13]. We describe the neighborhood compactly by the dependency of the number of points $q(r)$ on the neighborhood size[1] r.

$$q_i(r) = \big|\{\|\mathbf{x} - \mathbf{x}_i\|_\infty \le r; \mathbf{x} \in X\}\big| \tag{2}$$

Fixing the neighborhood radius R, we obtain an estimator for $-\log f(\mathbf{x}_i)$

$$h_{\text{ker}}(\mathbf{x}_i; X) = -\log \frac{q_i(R)}{V_d(R) N} \tag{3}$$

where $V_d(R) = (2R)^d$ is the volume of the neighborhood. This is a plug-in kernel estimator for a constant kernel, equivalent to an averaged shifted histogram (ASH) [13].

1.3 Nearest-neighbor estimator

The distance to the k-th nearest neighbor is

$$\varrho_i(k) = \min_{r \ge 0} \big\{ r; q_i(r) > k \big\} \tag{4}$$

An estimator based on the distance $\varrho_i(1)$ to the nearest-neighbor (NN) of $\mathbf{x}_i$ is due to [14] and was later extended to k-th nearest-neighbor (k-NN) [15, 16]. Its formulation for the ℓ_∞ norm [17] is

$$h_{\text{NN}}(\mathbf{x}_i; X) = h_0(r, k) = -\psi(k) + \psi(N) + d \log 2r \qquad \text{for} \qquad r = \varrho_i(k) \tag{5}$$

where ψ is the digamma function[2]. The estimator (5) is asymptotically unbiased [3, 14]. Its variance can be reduced by choosing higher k [16]. The k-NN estimator works reasonably well even in high dimensions (we have tested it for $d = 25 \sim 50$) and for small sample sizes. The computational bottleneck is the nearest neighbor search (all-NN search) but acceleration techniques exist, based

[1] We are using the ℓ_∞ (maximum) norm for better compatibility with rectangular bins. The ℓ_2 (Euclidean) norm can also be used with minimal changes.

[2] $\psi(k) = -\gamma + \sum_{i=1}^{k-1} 1/i$, where where $\gamma \approx 0.577$ is the Euler constant.

on space partitioning and approximative search [17–22]. Graph-based estimators for Rényi entropy [23, 24] behave similarly to the k-NN estimators.

Real data is often quantized or known only with limited accuracy. Due to the presence of the $\log r$ factor in (5), the results will fluctuate highly if small values of $\varrho_i(k)$ are inaccurate; if some $\varrho_i(k)$ are zero, the estimator diverges. A possible solution is to add low-amplitude perturbation to the data [15] or to switch to a histogram-like estimator for $r < R$, for some fixed R [17]. Singh [16] advocates the use of $k > 1$. However, the first approach increases variance, the second performs poorly in the transition region, and the third does not guarantee to eliminate the problem.

1.4 Proposed approach

Consider the case of finite accuracy data where no distances smaller than a given R can be reliably measured. The k-NN estimator h_{NN} (5) works well for low densities f, when the distance between neighboring points is much larger than the measurement accuracy. Conversely, the kernel estimator h_{ker} (3) works best for high densities f, when the distance between points is smaller than the kernel size. Hence, we propose to construct a new estimator, called R-NN, combining the advantages of the two approaches with a smooth transition between [25]. A numerically calculated correction is used to preserve unbiasedness.

2 Method

For a fixed R, we take the k-NN estimator h_{NN} (5), varying the k for each $\mathbf{x}_i$ so that $\varrho_i(k) > R$. This gives us a naive R-NN estimator $h_{\mathrm{naive}}(\mathbf{x}_i; X) = h_0(r, k)$ where (from now on we will drop the subscript i for brevity)

$$h_0(r, k) = -\psi(k) + \psi(N) + d \log 2r \quad \text{with} \quad k = q(R), \quad r = \varrho(k) \tag{6}$$

The expected value of $h_0(r, k) = h_{\mathrm{naive}}(\mathbf{x}_i; X)$ for a fixed $\mathbf{x}_i$ is

$$\mathrm{E}\big[h_0\big]_{X \setminus \{\mathbf{x}_i\}} = \sum_{k=1}^{N-1} \int_{r>R} h_0(k, r)\, p(k, r)\, \mathrm{d}r \tag{7}$$

where $p(k, r)$ is the probability density of observing k points (including $\mathbf{x}_i$) in the neighborhood R and the $(k+1)$-th point (the k-th NN) at distance $r > R$. It can be obtained using the trinomial formula [15]

$$p(k, r) = \frac{(N-1)!}{(k-1)!(N-k-1)!} (2^d R^d f)^{k-1} (1 - 2^d R^d f)^{N-k-1} 2^d d\, r^{d-1} f \tag{8}$$

where we assume that $f = f(\mathbf{x}_i)$ is constant in a sufficiently large neighborhood of $\mathbf{x}_i$. The restriction $r > R$ makes the naive estimator (6) biased. The

expected value $\mathrm{E}\big[h_0\big]$ can be calculated numerically (see Appendix A). A typical dependency of the bias B_0 on the density f

$$B_0(f) = \log f + \mathrm{E}\big[h_0\big] \tag{9}$$

is shown in Figure 1. The bias goes to zero for $f \to 0$ as the h_0 estimator approaches h_{NN}, and also as $f \to \infty$ as h_0 approaches[3] h_{ker}. The numerical calculation of $B_0(f)$ becomes inaccurate for high f. A deeper problem is that the assumption of a locally constant f in a neighborhood r is contradictory for high f, since the probability $V_d(r)f$ must not exceed 1. We therefore constrain $f < f_{\max}$ for a suitable $f_{\max}$.

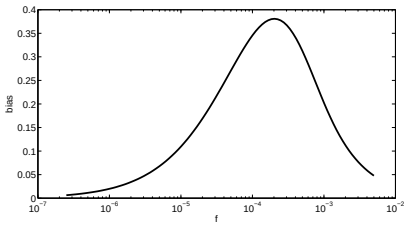

Fig. 1. Bias $B_0(f)$ of the naive estimator h_0 for $N = 1000$, $d = 2$ and $R = 1$.

2.1 Corrected *R*-NN Estimator

Let us add a correction $\tilde{h}$ to the naive estimator h_0

$$h(r,k) = h_0(r,k) + \tilde{h}(r,k) \quad \text{with} \quad k = q(R), \quad r = \varrho(k) \tag{10}$$

so that the corrected estimator h is unbiased

$$\mathrm{E}\big[h\big] = -\log f \qquad \text{for} \quad f < f_{\max} \tag{11}$$

Note that we can assume without loss of generality that $R = 1$. The correction for general R' is then obtained as $\tilde{h}'(r,k) = \tilde{h}(r/R', k)$ for $f' < (R')^{-d} f_{\max}$. To see it, consider estimating entropy $H(R'X)$ and $f' = (R')^{-d} f$.

As finding $\tilde{h}$ analytically seems to be difficult, we attempt a numerical solution. We require (11) to hold for $f_1 = f_{\min}, f_2, \ldots, f_F = f_{\max}$ for sufficiently small $f_{\min}$ and distributing $\log f_i$ uniformly. The correction $\tilde{h}$ is represented as a linear combination

$$\tilde{h}(r,k) = [k \le K] \sum_{i=1}^{M} a_{ik} \varphi_i(r) \tag{12}$$

[3] using the fact that $\psi(N) \approx \log N$ for large N

where the basis functions φ_i are piecewise linear on each interval $[r_i, r_{i+1}]$ and satisfying $\varphi_i(r_j) = \delta_{ij}$. We choose $r_1 = R$, sufficiently large r_M and K, and uniformly distributed $\log r_i$.

2.2 Quadratic Programming Formulation

The expected value of the estimator h (10) with correction $\tilde{h}$ (12) is

$$\mathrm{E}\big[h\big] = \mathrm{E}\big[h_0\big] + \mathrm{E}\big[\tilde{h}\big] = \mathrm{E}\big[h_0\big] + \sum_{k=1}^{K}\sum_{i=1}^{M} a_{ik} P_{ik} \tag{13}$$

$$\text{where} \quad P_{ik}(f) = \int\limits_{r>R} \varphi_i(r)\, p(k,r)\, \mathrm{d}r \tag{14}$$

See Appendix B for details on calculating P_{ik}. For numerical reasons, we shall require the bias of h to be bounded by some small constant γ for all $f \in \{f_1, \ldots, f_F\}$. Using (9,10,13) leads to a system of $2F$ linear inequalities

$$-\gamma \le B_0(f) + \sum_{k=1}^{K}\sum_{i=1}^{M} a_{ik} P_{ik}(f) \le \gamma \qquad \text{for} \quad f \in f_1, \ldots, f_F \tag{15}$$

In addition we shall require $a_{Mk} = 0$ for all $k \le K$ to prevent the discontinuity of $h(r,k)$ at $r = r_M$. Then (15) can be written in a matrix form as

$$\mathsf{A}\mathbf{a} \le \mathbf{c} \tag{16}$$

where $\mathbf{a} = (a_{11}, \ldots, a_{M-1,K})$ is a linearized vector of unknowns.

To prevent indeterminacy of (16), we use a quadratic programming formulation: minimize a quadratic criterion $Q(\mathbf{a})$ under the conditions (16) with

$$Q(\mathbf{a}) = \frac{1}{2}\mathbf{a}^T \mathsf{H}\mathbf{a} + \mathbf{u}^T\mathbf{a} \tag{17}$$

A natural choice for Q would be the variance $\mathrm{Var}[h](f)$. Unfortunately, this choice requires numerical integration and is both time-consuming and inaccurate, leading to an ill-posed or infeasible minimization problem. We have therefore decided to minimize the following simple finite-difference-based criterion instead, taking advantage of the fact that Q serves primarily as a regularization, the final solution is determined mainly by the constraints (16).

$$Q(\mathbf{a}) = \sum_{k=1}^{K}\sum_{i=1}^{M-1} (a_{i+1,k} - a_{i,k})^2 + \sum_{k=1}^{K-1}\sum_{i=1}^{M-1} (a_{i,k+1} - a_{i,k})^2 \tag{18}$$

The criterion (18) expresses our preference for 'smooth' $\tilde{h}$, motivated by the well-known formula $\mathrm{Var}[g(x)] \approx g'\big(\mathrm{E}[x]\big)^2 \mathrm{Var}[x]$. The corresponding Hessian H is very sparse, positive definite, and easy to calculate. The quadratic programming problem (17),(16) is solved by the MINQ algorithm [26].

It remains to determine a good value of γ in (15). We proceed iteratively, starting with $\gamma = \max_f |B_0(f)|$ and halving γ in each step. We stop if the solution cannot be found, or if the criterion increase is suspiciously large compared to the previous one, which signals overfitting.

3 Experiments

A typical shape of the correction $\tilde{h}(k, r)$ is shown in Figure 2 together with the shape of the uncorrected and corrected estimators h_0 and h. The width and position of the peak in $\tilde{h}$ depends on N and d. The parameter M influences the smoothness and F the accuracy. We found that $M = 100$ and $F = 1000$ give good results, with calculation of the estimator parameters taking several minutes. Higher values of M and F require more time and the calculation is often numerically unstable. The estimation itself is as fast as the k-NN estimator, with typical image similarity criterion taking between several seconds and one minute to evaluate, if acceleration techniques for the neighborhood search are used [17].

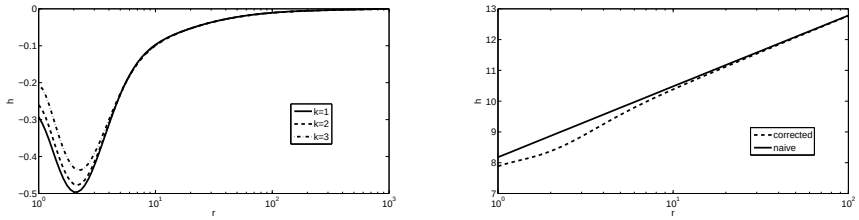

Fig. 2. The correction $\tilde{h}(r, k)$ for $N = 1000$, $d = 1$, shown as a function of r for $k = 1, 2, 3$ (left). The naive and corrected estimators (h_0 resp. h) for $k = 1$ (right).

3.1 Entropy estimation of normal data

Table 1 shows the bias, variance and mean squared error (MSE) for the k-NN, R-NN and histogram estimators (with optimal bin-width [13]) for estimation of entropy of normal data with unit covariance matrix in two dimensions ($d = 2$) from $N = 1000$ sample points. The experiment was repeated 100 times. For very small R the R-NN estimator is equivalent to the k-NN estimator; for higher R the variance decreases while bias remains essentially constant until R becomes comparable to the standard deviation of the data. Histogram estimator has a slightly lower variance but much higher bias.

3.2 Entropy estimation for quantized normal data

In Figure 3 we compare the k-NN, histogram (bin size 1), and R-NN estimators on 2D isotropic normal data quantized with step 1 as a function of the standard

estimator	bias	variance	MSE
k-NN, $k = 1$	0.0099	0.0660	0.0045
k-NN, $k = 2$	0.0101	0.0519	0.0028
k-NN, $k = 5$	0.0197	0.0423	0.0022
k-NN, $k = 10$	0.0300	0.0405	0.0025
histogram	0.0349	0.0386	0.0027
R-NN, $R = 10^{-4}$	0.0099	0.0660	0.0045
R-NN, $R = 10^{-3}$	0.0100	0.0661	0.0045
R-NN, $R = 10^{-2}$	0.0102	0.0628	0.0040
R-NN, $R = 10^{-1}$	**0.0080**	0.0433	**0.0019**
R-NN, $R = 1$	0.0437	**0.0358**	0.0032

Table 1. Estimating entropy of unit covariance normal data using k-NN, R-NN and histogram estimators, $d = 2$, $N = 1000$. Best values in each column are set in bold.

deviation σ. For σ comparable with the quantization step, histogram has the lowest bias and the lowest MSE. For σ much larger than the quantization step, the k-NN estimators perform best in terms of bias and MSE. It appears indeed that $k = 4$ is a good choice [16]. The R-NN estimators are a good compromise — they are almost as good as k-NN estimators for large σ (the difference is negligible for $\sigma = 10^3 \sim 10^4$) while offering significant improvement for small σ. In the high dimensional case ($d = 10$, MSE shown in Figure 3, bottom right), R-NN estimators outperform k-NN for small σ. Note that high R values may not be used in this case, since often no points fall outside the R neighborhood. Conversely, the k-NN estimator performs badly due to high bias for $k \geq 3$.

3.3 Mutual information as an image similarity criterion

We have evaluated mutual information $I(X, Y) = H(X) + H(Y) - H(X, Y)$ between two scalar images (as it is done in image registration) as a function of their horizontal shift. The difficulty of this particular case lies in suppressing quantization artifacts (peaks) for integer shifts and obtaining a smooth dependence on the shift for easy optimization. We use 100×100 pixel centered regions from approximately registered T1 and T2 magnetic resonance images (Figure 4) of the same brain slice [27]. For the histogram estimator (Figure 5, top left), bin size is critical; bins too small lead to quantization artifacts, while bins too large distort the curve shape. k-NN estimator performs poorly in this case, for $k \leq 8$ there are strong peaks on integers, for $k > 8$ the dependency is lost. The result of the R-NN estimator, for $R \geq 10$, is as good as the histogram.

Finally, we show also the color MI criterion as a function of shift for color colposcopy images [28]. While the k-NN estimator has artifacts for all k tested, the R-NN estimation is quite usable for $R = 20$, albeit noisy.

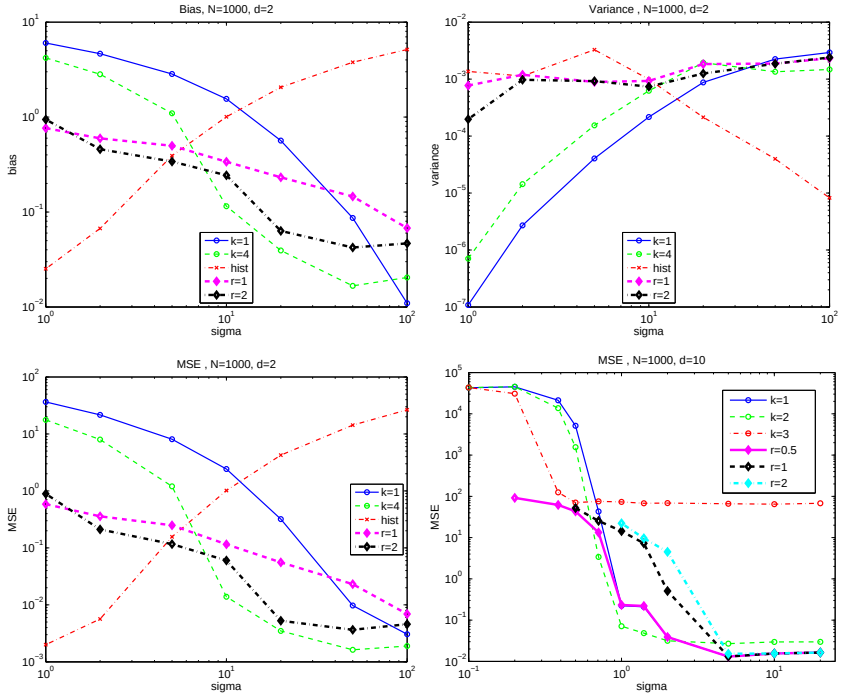

Fig. 3. The bias (top left), variance (top right), and MSE (bottom left) for several entropy estimators — k-NN with $k = 1$, $k = 4$, histogram estimator, and R-NN with $r = 1, r = 2$ — as a function of the standard deviation. The input are $N = 1000$ samples of a 2D ($d = 2$) Gaussian random variable with covariance matrix $\sigma^2 I$, rounded to the nearest integer. Bottom right image is the MSE for $d = 10$, $N = 1000$ for k-NN estimator with $k = 1, 2, 3$ and R-NN with $r = 0.5, r = 1.0, r = 2.0$.

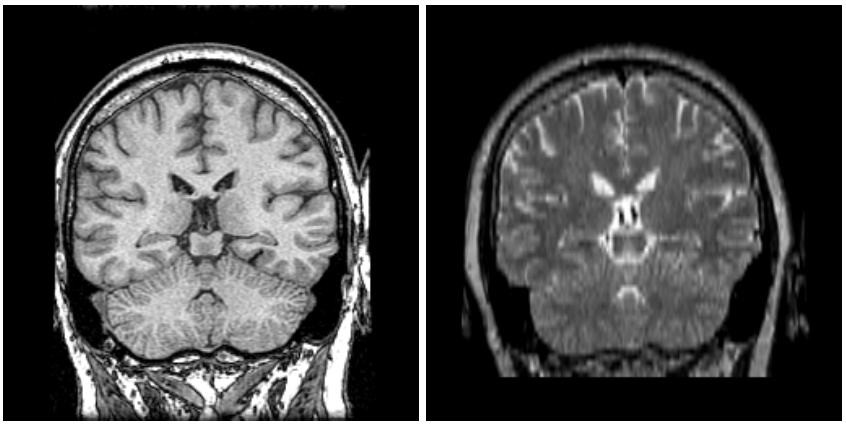

Fig. 4. 2D MRI images used for testing, T1 (left) and T2 (right).

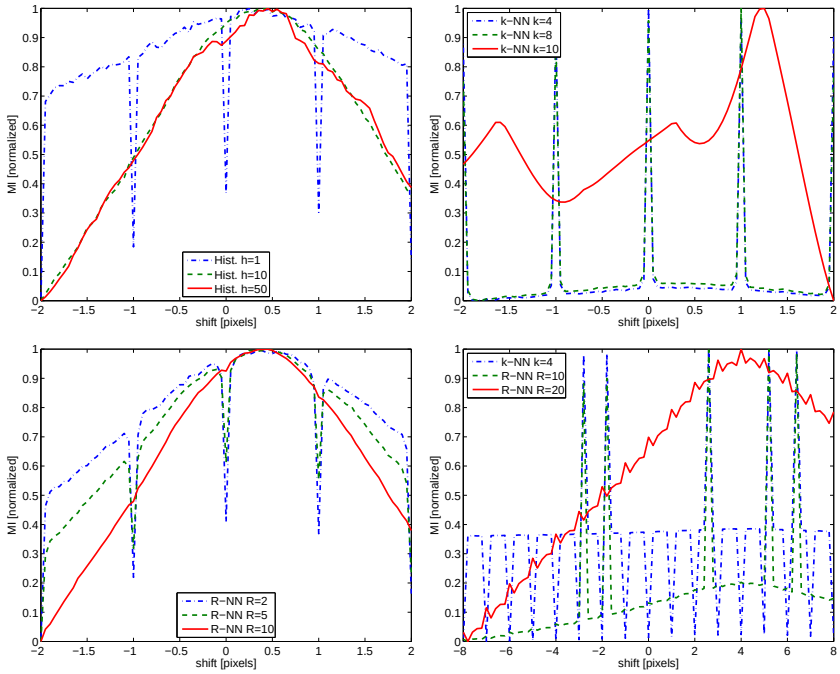

Fig. 5. Scalar mutual information criterion as a function of the horizontal shift, evaluated by histogram estimator (top left), k-NN estimator (top right), and R-NN estimator (bottom left). All values were normalized to $[0, 1]$ for easy visualization. The optimum shift is around 0.5 pixels. The bottom right graph shows the color MI criterion as a function of shift for two color colposcopy images for k-NN and R-NN estimators.

4 Conclusions

We have presented a new entropy estimator based on two quantities, the number of points k in a neighborhood of size R and the distance to the closest point r farther than R. The estimator behaves like an averaged histogram for high densities and like an NN estimator for low densities, smoothly varying between the two, combining their particular strengths. Finding the estimator is formulated as a constrained optimization problem (quadratic programming). The limited accuracy of this numerical procedure especially for high values of N, d, and f is currently the biggest setback of the new method. Nevertheless, our experiments show that the new method outperforms its ancestors in a number of situations, it deals successfully with the limited accuracy effects and is usable in a practical setting.

Acknowledgements

This work was sponsored by the Czech Ministry of Education, Project MSM6840770012.

A Expected Value of the Naive Estimator

We rewrite expression (7) by taking terms independent of r out of the integral and by substituting $\xi = 2^d r^d f$:

$$\mathrm{E}\big[h_0\big] = \sum_{k=1}^{N-1} \frac{(N-1)!}{(k-1)!(N-k-1)!} (2^d R^d f)^{k-1} I_1(N,k,f,R)$$

$$\text{with} \quad I_1(N,k,f,R) = \int_\alpha^1 \big(\log\xi - \log f + \psi(N) - \psi(k)\big)(1-\xi)^{N-k-1}\,\mathrm{d}\xi$$

where $\alpha = 2^d R^d f$ and the upper integration limit comes from the fact that ξ is a probability. The integral I_1 is broken into two:

$$I_1 = \big(-\log f + \psi(N) - \psi(k)\big)\int_\alpha^1 (1-\xi)^{N-k-1}\,\mathrm{d}\xi + \underbrace{\int_\alpha^1 \log\xi (1-\xi)^{N-k-1}\,\mathrm{d}\xi}_{I_2}$$

The first integral is standard, the second one can be integrated by parts

$$I_2 = \frac{1}{N-k}\log\alpha\,(1-\alpha)^{N-k} + \frac{1}{N-k}\underbrace{\int_\alpha^1 \frac{(1-\xi)^{N-k}}{\xi}\,\mathrm{d}\xi}_{I_3}$$

After substitution $z = 1-\xi$, we obtain:

$$I_3 = \int_0^{1-\alpha} \frac{z^{N-k}}{1-z}\,\mathrm{d}z = -\int_0^{1-\alpha} \sum_{i=0}^{N-k-1} z^i\,\mathrm{d}z + \int_0^{1-\alpha}\frac{1}{1-z}\,\mathrm{d}z = -\sum_{i=1}^{N-k}\frac{(1-\alpha)^i}{i} - \log\alpha$$

The last expression is delicate to calculate, since $\sum_{i=1}^{\infty}\frac{(1-\alpha)^i}{i} = -\log\alpha$.

B Calculating the Coefficients $\boldsymbol{P}$

The integral (14) can be written as follows

$$P_{ik} = \frac{(N-1)!}{(k-1)!(N-k-1)!}(2^d R^d f)^{k-1}\big(I^- + I^+\big)$$

$$\text{with} \quad I^- = \int_{r_{i-1}}^{r_i} (1-2^d r^d f)^{N-k-1}\,\frac{r-r_{i-1}}{r_i - r_{i-1}}\,2^d r^{d-1} d\,\mathrm{d}r$$

$$I^+ = \int_{r_i}^{r_{i+1}} (1-2^d r^d f)^{N-k-1}\,\frac{r_{i+1}-r}{r_{i+1}-r_i}\,2^d r^{d-1} d\,\mathrm{d}r$$

Substituting $\xi = 2^d r^d f$ and $\xi_i = 2^d r^d f_i$ we get

$$I^+ = \frac{1}{r_{i+1}-r_i}\left(r_{i+1}\int_{\xi_i}^{\xi_{i+1}} (1-\xi)^{N-k-1}\mathrm{d}\xi - \frac{1}{2f^{1/d}}\int_{\xi_i}^{\xi_{i+1}} \xi^{1/d}(1-\xi)^{N-k-1}\,\mathrm{d}\xi\right)$$

and similarly for I^-. The first integral is standard, the second one is related to the non-regularized incomplete beta function:

$$\int_0^{\xi_i} \xi^{1/d}(1-\xi)^{N-k-1}\,\mathrm{d}\xi = B(\xi_i; \frac{1}{d}+1, N-l)$$

References

1. J. Beirlant, E. J. Dudewicz, Györfi L., and E. C. van der Meulen, "Nonparametric entropy estimation: an overview," *International J. Math. Stat. Sci.*, , no. 6, pp. 17–39, 1997.
2. E. J. Harner, H. Singh, S. Li, and J. Tan, "Computational challenges in computing nearest neighbor estimates of entropy for large molecules," in *Computing Science and Statistics*, 2003, p. 35.
3. Jonathan D. Victor, "Binless strategies for estimation of information from neural data," *Physical Review E*, vol. 66, no. 5, pp. 051903(15), Nov. 2002.
4. Paul Viola and William M. Wells III, "Alignment by maximization of mutual information," *International Journal of Computer Vision*, , no. 2, pp. 137–154, 1997.
5. J.P.W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: A survey," *IEEE Transactions on Medical Imaging*, vol. 22, no. 8, pp. 986–1004, Aug. 2003.
6. F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Transactions on Medical Imaging*, vol. 16, no. 2, pp. 187–198, Apr. 1997.
7. Josien P. W. Pluim, J. B. Antoine Maintz, and Max A. Viergever, "Image registration by maximization of combined mutual information and gradient information," *IEEE Transactions Med. Imag.*, vol. 19, no. 8, Aug. 2000.
8. D. Rueckert, M. J. Clarkson, D. L. G. Hill, and D. J. Hawkes, "Non-rigid registration using higher-order mutual information," in *Proceedings of SPIE Medical Imaging 2000: Image Processing*, 2000, pp. 438–447.
9. D. B. Russakoff, C. Tomasi, T. Rohlfing, and C. R. Jr. Maurer, "Image similarity using mutual information of regions," in *Proceedings of the 8th European Conference on Computer Vision (ECCV)*, T. Pajdla and J. Matas, Eds. May 2004, number 3023 in LNCS, pp. 596–607, Springer.
10. Mert R. Sabuncu and Peter J. Ramadge, "Spatial information in entropy-based image registration.," in *WBIR*, James C. Gee, J. B. Antoine Maintz, and Michael W. Vannier, Eds. 2003, vol. 2717 of *Lecture Notes in Computer Science*, pp. 132–141, Springer.
11. Georges A. Darbellay and Igor Vajda, "Estimation of the information by an adaptive partitioning of the observation space," *IEEE Transactions on Information Theory*, vol. 45, no. 4, pp. 1315–1321, May 1999.
12. Erik G. Miller, "A new class of entropy estimators for multi-dimensional densities," in *Proceedings of ICASSP2003*, 2003.
13. David W. Scott, Ed., *Multivariate Density Estimation : Theory, Practice, and Visualization*, Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 1992.
14. L. F. Kozachenko and N. N. Leonenko, "On statistical estimation of entropy of random vector," *Probl. Inf. Trans.*, vol. 23, no. 9, 1987, (in Russian).

15. A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, , no. 69 066138, 2004.
16. H. Singh, N. Misra, V. Hnizdo, A. Fedorowicz, and E. Demchuk, "Nearest neighbor estimates of entropy," *American journal of mathematical and management sciences*, vol. 23, no. 3–4, pp. 301–321, 2003.
17. Jan Kybic, "Incremental updating of nearest neighbor-based high-dimensional entropy estimation," in *ICASSP2006*, Pierre Duhamel and Luc Vandendorpe, Eds., Toulouse, France, May 2006, pp. III–804, IEEE, DVD proceedings.
18. Paul B. Callahan and S. Rao Kosaraju, "A decomposition of multi-dimensional point-sets with applications to k-nearest-neighbors and n-body potential fields," in *Proceedings 24th Annual AMC Symposium on the Theory of Computing*, 1992, pp. 546–556.
19. M. Smid, "Closest-point problems in computational geometry," 1997, To appear in: Handbook on Computational Geometry, edited by J.-R. Sack, North Holand, Amsterdam.
20. Jeffrey S. Beis and David G. Lowe, "Shape indexing using approximate nearest-neighbour search in high-dimensional spaces," in *Proceedings of Conference on Computer Vision and Pattern Recognition*, June 1997, pp. 1000–1006.
21. F. P. Preparata and M. I. Shamos, *Computational geometry: An introduction*, Texts and Monographs in Computer Science. Springer-Verlag, 1985.
22. Robert Sedgewick, *Algorithms*, Addison Wesley, 1989.
23. A. O. Hero, B. Ma, O. Michel, and J. Gorman, "Applications of entropic spanning graphs," *IEEE Signal Proc. Magazine*, vol. 19, no. 5, pp. 85–95, sep 2002.
24. Huzefa Neemuchwala, Alfred Hero, and Paul Carson, "Image matching using alpha-entropy measures and entropic graphs," *Signal Process.*, vol. 85, no. 2, pp. 277–296, 2005.
25. Radim Šára, "A modification of Kozachenko-Leonenko entropy estimator for quantized data," Unpublished notes, 2006.
26. Arnold Neumaier, "MINQ — general definite and bound constrained indefinite quadratic programming.," 1998, http://www.mat.univie.ac.at/∼neum/software/minq/.
27. K. A. Johnson and J. A. Becker, "The whole brain atlas," http://www.med.harvard.edu/AANLIB/.
28. Juan D. García-Arteaga, Jan Kybic, and Wenjing Li, "Elastic image registration for movement compensation in digital colposcopy," in *BIOSIGNAL: Analysis of Biomedical Signals and Images*, Jiří Jan, Jiří Kozumplík, and Ivo Provazník, Eds., Brno, Czech Republic, June 2006, EURASIP, pp. 236–238, VUTIUM Press.