

An Entropy Estimator Based on Polynomial Regression with Poisson Error Structure

Hideitsu Hino^{1(✉)}, Shotaro Akaho², and Noboru Murata³

¹ University of Tsukuba, 1-1-1 Tennoudai, Tsukuba, Ibaraki 305–8573, Japan
hinohide@cs.tsukuba.ac.jp

² National Institute of Advanced Industrial Science and Technology,
1-1-1 Umezono, Tsukuba, Ibaraki 305–8568, Japan

³ Waseda University, 3-4-1 Ohkubo, Shinjuku-ku, Tokyo 169–8555, Japan

Abstract. A method for estimating Shannon differential entropy is proposed based on the second order expansion of the probability mass around the inspection point with respect to the distance from the point. Polynomial regression with Poisson error structure is utilized to estimate the values of density function. The density estimates at every given data points are averaged to obtain entropy estimators. The proposed estimator is shown to perform well through numerical experiments for various probability distributions.

Keywords: Entropy · Regression · Density estimation · Poisson error structure

1 Introduction

Let X be a p -dimensional random variable with a probability density function (pdf) $f(X)$. The differential entropy [1, 2] of this distribution with pdf $f(x)$ is defined by

$$H(f) = - \int f(x) \ln f(x) dx. \quad (1)$$

We consider estimating the entropy $H(f)$ in non-parametric manner using a set of observations $\mathcal{D} = \{x_i\}_{i=1}^n$, where $x_i, i = 1, \dots, n$ are the independent realizations of X with pdf $f(x)$. There are a large number of non-parametric entropy estimation methods. The simplest approach is firstly estimating the pdf using the observed dataset \mathcal{D} by using, for example the kernel density estimator [3], then substitute the estimate $\hat{f}(x)$ into the definition of the entropy. The entropy can be estimated by numerical integration, though, it is known that numerical integration for multivariate function is unstable and time consuming, it is recommended in [4] to use empirical expectation with respect to \mathcal{D} as

$$\hat{H}(\mathcal{D}) = -\frac{1}{n} \sum_{i=1}^n \ln \hat{f}(x_i). \quad (2)$$

One of the most popular methods for differential entropy estimation is the method based on the k -nearest neighbor method [5–10]. In this work, we derive a nonparametric entropy estimator based on the second order expansion of probability mass function and polynomial regression with Poisson error structure. The proposed method is experimentally shown to work well for estimating the differential entropy of various probability distributions.

2 Preliminary and Notation

We consider the problem of estimating the value $f(z)$ of the probability density function at the *inspection point* $z \in \mathbb{R}^p$ using the set of observation $\mathcal{D} = \{x_i\}_{i=1}^n$. Let the p -dimensional ball with radius ε centered at z be $b(z; \varepsilon) = \{x \in \mathbb{R}^p \mid \|z - x\| < \varepsilon\}$, which has volume $|b(z; \varepsilon)| = c_p \varepsilon^p$, where $c_p = \pi^{p/2} / \Gamma(p/2 + 1)$ is a volume element of the p -dimensional unit ball and $\Gamma(\cdot)$ is the gamma function. The probability mass of the ball is defined by

$$q_z(\varepsilon) = \int_{x \in b(z; \varepsilon)} f(x) dx. \quad (3)$$

Expanding the integrand, we obtain

$$\begin{aligned} q_z(\varepsilon) &= \int_{x \in b(z; \varepsilon)} \{f(x) + (x - z)^\top \nabla f(z) + O(\varepsilon^2)\} dx \\ &= |b(z; \varepsilon)| (f(z) + O(\varepsilon^2)) = c_p \varepsilon^p f(z) + O(\varepsilon^{p+2}). \end{aligned}$$

Assume that the radius ε of the ball is enough small and ignore the second order term. Then, approximating the left hand side of the above equation by the proportion of the number of samples fallen in the ball to the whole sample size n , we obtain a first order approximation of the value of pdf as

$$\hat{f}(z; \varepsilon) = \frac{k_\varepsilon}{nc_p \varepsilon^p}, \quad (4)$$

where k_ε is the number of samples in \mathcal{D} inside the ε -ball [11–13]. Conversely, when we fix the number of sample points from the inspection point to k , we obtain the k -NN density estimator $\hat{f}^{nn}(z; k) = k / (nc_p \varepsilon_k^p)$, where ε_k is the distance between the inspection point to the k -th nearest point. Denoting the values of k -NN estimator at $x_i \in \mathcal{D}$ using $\mathcal{D} \setminus \{x_i\}$, namely, without using $\{x_i\}$, by $\hat{f}_i^{nn}(x_i; k)$, we obtain the k -NN based entropy estimator [6] by

$$\hat{H}^{nn}(\mathcal{D}; k) = - \sum_{i=1}^n \ln \hat{f}_i^{nn}(x_i; k). \quad (5)$$

3 Second Order Method

In our previous work [14], we derived nonparametric entropy estimators based on the second order expansion of the integrand of Eq. (3).

Proposition 1. *The probability mass of the ε -ball around z is expanded as*

$$q_z(\varepsilon) = c_p f(z) \varepsilon^p + \frac{n}{4(p/2+1)} c_p \text{Tr} \nabla^2 f(z) \varepsilon^{p+2} + O(\varepsilon^{p+4}). \quad (6)$$

Approximating the left hand side of Eq. (6) by k_ε/n , and dividing the equation by $c_p \varepsilon^p$, we obtain

$$\frac{k_\varepsilon}{n c_p \varepsilon^p} = f(z) + C \varepsilon^2 + O(\varepsilon^4), \quad (7)$$

where $C = \frac{n \text{Tr} \nabla^2 f(z)}{4(p/2+1)}$. Introducing the *response variable* $Y_\varepsilon = \frac{k_\varepsilon}{n c_p \varepsilon^p}$ and the *explanatory variable* $X_\varepsilon = \varepsilon^2$, and ignoring higher order term with respect to ε , we obtain a linear equation

$$Y_\varepsilon \simeq f(z) + C X_\varepsilon. \quad (8)$$

This Eq. (8) can be regarded as a linear regression model with respect to $(X_\varepsilon, Y_\varepsilon)$. These variables vary with the different values of ε . Taking a set of radii $\mathcal{E} = \{\varepsilon_i\}_{i=1}^m$ and regarding the pairs $\{(X_\varepsilon, Y_\varepsilon)\}_{\varepsilon \in \mathcal{E}}$ observed samples, we can estimate $f(z)$ and C by minimizing the squared error

$$R = \frac{1}{m} \sum_{\varepsilon \in \mathcal{E}} (Y_\varepsilon - f(z) - C X_\varepsilon)^2, \quad (9)$$

which is nothing but the fitting of simple linear model. Namely, the intercept of the linear model is the estimate of the value of the pdf at z . Let $\hat{f}_i^s(x_i)$ be the estimate obtained by solving Eq. (9) without using a sample x_i . Then, by leave-one-out estimate, we obtain a nonparametric entropy estimator

$$\hat{H}^s(\mathcal{D}) = -\frac{1}{n} \sum_{i=1}^n \ln \hat{f}_i^s(x_i), \quad (10)$$

which we call the Simple Regression Entropy Estimator (SRE) [14].

In [14], another entropy estimator is also proposed, by substituting the relation Eq. (8) to the empirical estimate of the differential entropy (2) and fitting linear model. Suppose ε is fixed, and consider Eq. (8) at the inspection point $x_i \in \mathcal{D}$. Here $Y_\varepsilon = \frac{k_\varepsilon}{n c_p \varepsilon^p}$ and $C = \frac{n \nabla^2 f(x_i)}{4(p/2+1)}$ depend on the inspection point x_i , we denote them as Y_ε^i and C^i , respectively. To derive an entropy estimator based on Eq. (2), we consider the minus of the logarithm of $Y_\varepsilon^i = f(x_i) + C^i X_\varepsilon$. By averaging this quantity with respect to all sample points $x_i \in \mathcal{D}$, we obtain

$$\begin{aligned}
-\frac{1}{n} \sum_{i=1}^n \ln Y_\varepsilon^i &= -\frac{1}{n} \sum_{i=1}^n \ln \{f(x_i) + C^i X_\varepsilon\} \\
&= -\frac{1}{n} \sum_{i=1}^n \ln f(x_i) - \frac{1}{n} \sum_{i=1}^n \ln \left(1 + \frac{C^i X_\varepsilon}{f(x_i)}\right) \\
&\simeq -\frac{1}{n} \sum_{i=1}^n \ln f(x_i) - \frac{1}{n} \left(\sum_{i=1}^n \frac{C^i}{f(x_i)}\right) X_\varepsilon.
\end{aligned}$$

The last equation is from the first order Taylor expansion of $\ln(1+x)$. The first term of the above equation is the empirical estimate (2) of the entropy. So, by defining $\bar{Y}_\varepsilon = -\frac{1}{n} \sum_{i=1}^n \ln Y_\varepsilon^i$, $H(\mathcal{D}) = -\frac{1}{n} \sum_{i=1}^n \ln f(x_i)$, and $\bar{C} = -\frac{1}{n} \sum_{i=1}^n \frac{C^i}{f(x_i)}$, we obtain a relationship

$$\bar{Y}_\varepsilon = H(\mathcal{D}) + \bar{C} X_\varepsilon. \quad (11)$$

In the same manner as in SRE, this equation is valid for each of sample points $\{(X_\varepsilon, \bar{Y}_\varepsilon)\}_{\varepsilon \in \mathcal{E}}$. By fitting a linear model, we obtain an entropy estimator $\hat{H}^d(\mathcal{D})$ as the estimated intercept, which is called the Direct Regression Entropy Estimator (DRE). The SRE and DRE take the higher order information of pdf into account, though they do not consider the characteristic of the error structure for the observation, namely, the number of samples within a ball should be treated as a counting variable.

4 Proposed Method

In this section, we derive a novel entropy estimator based on linear regression with a Poisson error structure. The left hand side of the Eq. (6) is approximated by k_ε/n again. Then, multiply both sides of equation by n to obtain

$$k_\varepsilon \simeq c_p n f(z) \varepsilon^p + c_p n \frac{n}{4\Gamma(p/2 + 1)} \text{Tr} \nabla^2 f(z) \varepsilon^{p+2}. \quad (12)$$

This equation is regarded as a regression of k_ε on $(\varepsilon^p, \varepsilon^{p+2})$. Namely, for a certain $\varepsilon > 0$, the explanatory variable is defined by $X = (\varepsilon^p, \varepsilon^{p+2})$ and response variable is defined by $Y = k_\varepsilon$, which are linked by a simple generalized linear model $Y = \beta^\top X$. Since k_ε is a counting data, the Poisson error structure is a natural choice. We note that it is common to adopt the logarithmic link function for Poisson regression as a link function in the generalized linear model, though, in our formulation, the identity link function is natural. The logarithmic link function can avoid negative values for the response variable while the identity link cannot. However, our aim is not in prediction but in fitting to obtain the coefficient as the estimate of pdf value, and the estimated pdfs in all of our experiments were non-negative.

More concretely, for a set of radii $\mathcal{E} = \{\varepsilon_i\}_{i=1}^m$, we calculate the corresponding pairs of explanatory and response variable $\{(Y_i, X_i)\}_{i=1}^m$ where $X_i = (\varepsilon_i^p, \varepsilon_i^{p+2})$ and

Y_i is the number of samples within the ε_i -ball centered at the inspection point. Then, we maximize the likelihood function

$$L(\boldsymbol{\beta}) = \prod_{i=1}^m \frac{e^{-X_i^\top \boldsymbol{\beta}} (X_i^\top \boldsymbol{\beta})^{Y_i}}{Y_i!} \quad (13)$$

of a Poisson distribution with respect to $\boldsymbol{\beta} = (\beta_1, \beta_2)$. The ML estimate $\hat{\beta}_1$ for the first coefficient β_1 is divided by $c_p n$ to obtain the estimate of the pdf value at z as $\hat{f}(z) = \hat{\beta}_1 / (c_p n)$. This procedure is done for each $X \in \mathcal{D}$ and by leave-one-out method, we obtain the proposed entropy estimator in the same manner as in SRE. The notable characteristic of the proposed method, compared to conventional SRE, is in utilizing the Poisson error structure. We call the proposed entropy estimator EPI (Entropy Estimator with Poisson-noise structure Identity-link regression) henceforth.

5 Numerical Experiments

We apply some conventional and proposed entropy estimators for samples from various distributions to see the accuracies of the estimators.

5.1 Univariate Case

We evaluate the performance of entropy estimation by the absolute error

$$AE = |H(f) - \hat{H}(\mathcal{D})| \quad (14)$$

between the ground truth entropy $H(f)$ and the estimates $\hat{H}(\mathcal{D})$. We used the following 15 distributions: (1) Normal, (2) Skewed, (3) Strongly Skewed, (4) Kurtotic, (5) Bimodal, (6) Skewed Bimodal, (7) Trimodal, (8) Claw, (9) 4th Power Exponential, (10) Logistic, (11) Laplace, (12) t with df=5, (13) Mixed t, (14) Exponential, and (15) Cauchy, which are shown in Fig. 1. Details of these distributions are shown in [14]. We compare the proposed method to four existing methods, (a) KDE: pdf is estimated by kernel density estimation, and the estimate $\hat{f}(x)$ is substituted in Eq. (2). The kernel band width is optimized by using the unbiased cross-validation method [15]. (b) kNN: entropy is estimated by the k -NN method [6]. The number k is fixed to $k = 3$ following the empirical results reported in [16]. (c, d) SRE, DRE: the methods explained in Sect. 3.

From the ground truth distributions, we sampled datasets 100 times and perform entropy estimation, and the performance is estimated by average and standard deviations of AE as shown in Table 1. The number of sample in each dataset is 500. We note that we performed the same experiments with different sample sizes, and observed similar tendencies. From this result, it is seen that the proposed method significantly outperforms other four methods in 2 out of 15 distributions, and marks the second best method in 9 methods.

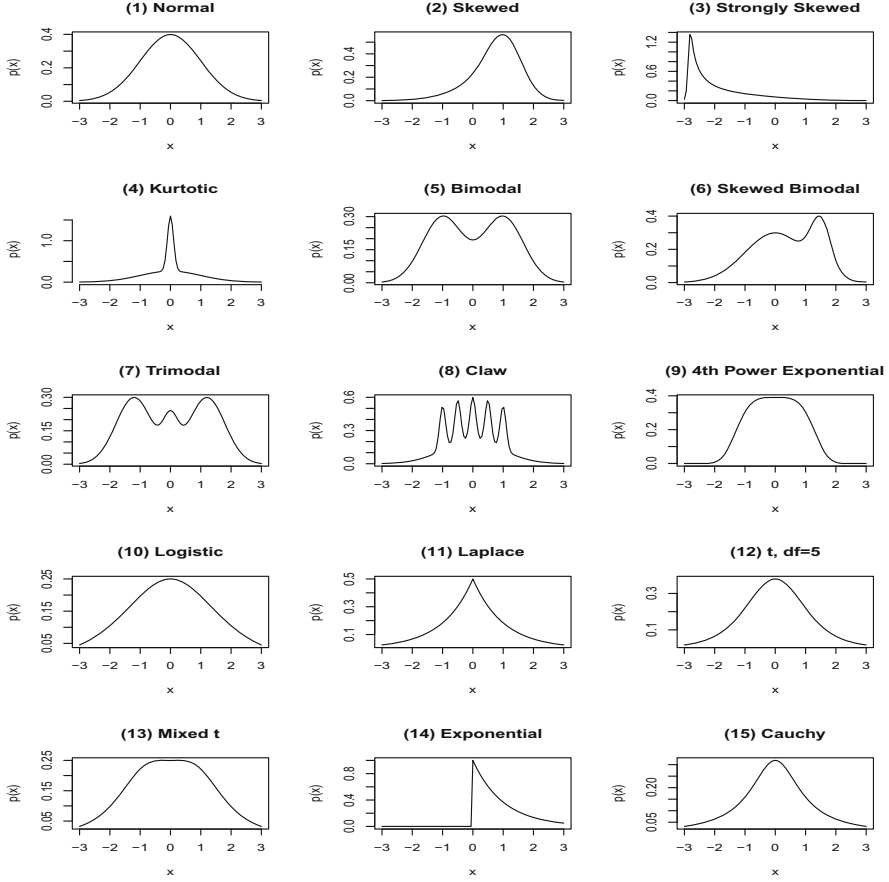


Fig. 1. Plots of 15 probability density functions for generating samples.

5.2 Multivariate Case

For seeing the effect of dimensions on the estimation accuracy, we performed a set of experiments with multidimensional distributions. It is difficult to calculate the ground truth entropy values for general multidimensional distributions, hence we use Gaussian distributions with three different covariance structures:

Isometric: covariance matrix is a p dimensional unit matrix.

Band: diagonal elements of covariance matrix is one, and its upper and lower elements are 0.3.

Full Correlation: Each element of the covariance matrix is set to

$$[\Sigma_p]_{ij} = 0.9^{|i-j|+1}, \quad 0 \leq i, j \leq p. \quad (15)$$

Table 1. Averages of absolute errors of entropy estimations for different seven methods. Sample size n is set to 500. The minimum AE results are shown in boldface, the second best method is shown with †, and when the minimum is statistically significant in t-test with $\alpha = 0.05$ compared to the second best result, the result is shown with *.

	KDE	kNN	SRE	DRE	EPI
Type 1	0.028 (0.0194)	0.040(0.0292)	0.066(0.0334)	0.030(0.0233)	†0.029(0.0192)
Type 2	†0.029(0.0246)	0.039(0.0310)	0.061(0.0329)	0.032(0.0226)	0.027 (0.0239)
Type 3	0.149(0.1891)	†0.035(0.0247)	0.087(0.0295)	0.139(0.0259)	*0.020 (0.0163)
Type 4	0.219(0.2491)	0.040 (0.0313)	0.149(0.0409)	0.088(0.0381)	†0.041(0.0301)
Type 5	0.022(0.0161)	0.033(0.0226)	0.022(0.0161)	0.017 (0.0122)	†0.021(0.0154)
Type 6	0.026(0.0206)	0.037(0.0259)	0.027(0.0200)	0.023 (0.0170)	†0.024(0.0172)
Type 7	0.022(0.0189)	0.032(0.0228)	0.018 (0.0127)	†0.019(0.0144)	0.020(0.0131)
Type 8	0.154(0.1289)	0.038(0.0354)	*0.025 (0.0185)	0.055(0.0289)	†0.036(0.0257)
Type 9	0.022(0.0150)	0.034(0.0254)	0.025(0.0191)	†0.021(0.0159)	0.020 (0.0136)
Type 10	0.053(0.0617)	0.041(0.0318)	0.091(0.0412)	†0.038(0.0229)	*0.033 (0.0256)
Type 11	0.156(0.2368)	0.049(0.0353)	0.131(0.0426)	0.036 (0.0260)	†0.039(0.0322)
Type 12	0.094(0.1175)	†0.041(0.0345)	0.108(0.0411)	0.040 (0.0299)	0.042(0.0304)
Type 13	0.251(0.3148)	0.044(0.0334)	0.086(0.0381)	0.035 (0.0257)	†0.041(0.0327)
Type 14	0.505(0.4753)	0.045 (0.0334)	0.073(0.0465)	0.127(0.0479)	†0.048(0.0361)
Type 15	1.903(1.0509)	0.477(0.0933)	*0.169 (0.0740)	0.509(0.1055)	†0.310(0.1006)

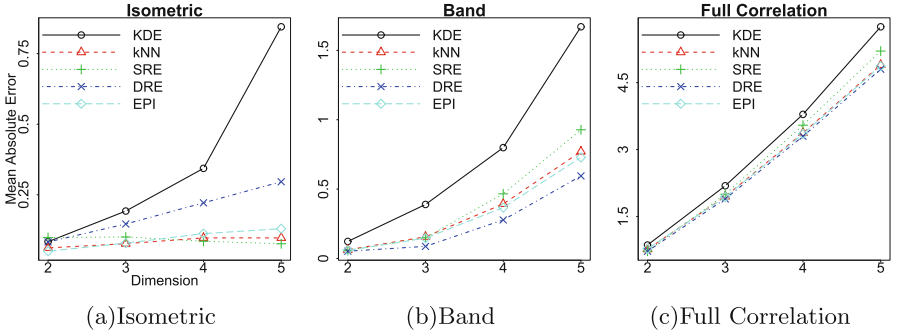


Fig. 2. Averages of absolute errors of entropy estimation when p is varied from 2 to 5. The number of samples is fixed to $n = 300$.

We varied the sample size to $n = 100, 300, 500, 700$, though, we didn't see systematic difference, and we show the case with $n = 300$ in Fig. 2. From Fig. 2, we can see that for all of three distributions, the proposed method shows moderate increase in estimation error as the increase of dimension, and it is a strong candidate of the non-parametric entropy estimator among classical k NN, KDE, and other recently proposed methods.

6 Conclusion

We proposed a non-parametric entropy estimator based on the second order expansion of probability mass function, and polynomial fitting with respect to the distance from the inspection point. By modeling the error structure by Poisson distribution, we obtained comparable or superior estimation accuracies to conventional method. It is also shown that the proposed method works well for multi-dimensional cases. Our future work includes investigation of statistical properties of the proposed estimator, including the optimal choice of the radii \mathcal{E} . We are also planning to apply the proposed estimator to various real-world problems which require accurate entropy estimation.

Acknowledgement. Part of this work was supported by JSPS KAKENHI No. 25120009, 25120011, and 16K16108.

References

1. Cover, T.M., Thomas, J.A.: Elements of information theory. Wiley, Hoboken (1991)
2. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(3), 379–423 (1948)
3. Wand, M.P., Jones, M.C.: Kernel Smoothing. Chapman & Hall/CRC, London (1994)
4. Joe, H.: Estimation of entropy and other functionals of a multivariate density. Ann. Inst. Stat. Math. **41**(4), 683–697 (1989)
5. Kozachenko, L.F., Leonenko, N.N.: Sample estimate of entropy of a random vector. Prob. Inf. Transm **23**, 95–101 (1987)
6. Gorla, M.N., Leonenko, N.N., Mergel, V.V., Novi Inverardi, P.L.: A new class of random vector entropy estimators and its applications in testing statistical hypotheses. J. Nonparametric Stat. **17**(3), 277–297 (2005)
7. Beirlant, J., Dudewicz, E.J., Györfi, L., Meulen, E.C.: Nonparametric entropy estimation: an overview. Int. J. Math. Stat. Sci. **6**, 17–39 (1997)
8. Györfi, L., van der Meulen, E.C.: Density-free convergence properties of various estimators of entropy. Comput. Stat. Data Anal. **5**(4), 425–436 (1987)
9. Paninski, L.: Estimation of entropy and mutual information. Neural Comput. **15**, 1191–1253 (2003)
10. Pérez-Cruz, F.: Estimation of information theoretic measures for continuous random variables. In: NIPS, pp. 1257–1264 (2008)
11. Loftsgaarden, D.O., Quesenberry, C.P.: A nonparametric estimate of a multivariate density function. Ann. Math. Stat. **36**(3), 1049–1051 (1965)
12. Mack, Y.P., Rosenblatt, M.: Multivariate k-nearest neighbor density estimates. J. Multivar. Anal. **9**(1), 1–15 (1979)
13. Moore, D.S., Yackel, J.W.: Consistency properties of nearest neighbor density function estimators. Ann. Stat. **5**(1), 143–154 (1977)
14. Hino, H., Koshijima, K., Murata, N.: Non-parametric entropy estimators based on simple linear regression. Comput. Stat. Data Anal. **89**, 72–84 (2015)
15. Rudemo, M.: Empirical choice of histograms and kernel density estimators. Scand. J. Stat. **9**(2), 65–78 (1982)

16. Khan, S., Bandyopadhyay, S., Ganguly, A.R., Saigal, S., Erickson, D.J., Protopopescu, V., Ostrouchov, G.: Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. *Phys. Rev. E: Stat., Nonlin., Soft Matter Phys.* **76**(2 Pt 2), 026209 (2007)

Neural Information Processing

23rd International Conference, ICONIP 2016, Kyoto,

Japan, October 16–21, 2016, Proceedings, Part II

Akira, H.; Seiichi, O.; Doya, K.; Kazushi, I.; Minho, L.;

Derong, L. (Eds.)

2016, XIX, 739 p. 252 illus., Softcover

ISBN: 978-3-319-46671-2