STAT.Sarween Ahmed Maroof@MSc2024.Non-Parametric

by Sarween Ahmad

Submission date: 28-Dec-2023 01:13AM (UTC+0200)
Submission ID: 2265180850
File name: STAT.Sarween_Ahmed_Maroof_MSc2024.Non-Parametric_-_Sami_Obed.pdf (587.69K)
Word count: 3505
Character count: 20202

Salahaddin University-Erbil

College of Administration and Economics

High Education: M.Sc.

Department: Statistics and Informatics

Subject: Non-Parametric

A Review Article about:

"Kernel Density Estimation"

Prepared

Sarween Ahmed Maroof

sarweenahmed293@gmail.com

Supervised

Assit. Prof.Dr. Nazeera Sdeeq Kareem

Academic year (2023-2024)

Abstract

Non-parametric techniques provide a resilient substitute for parametric methodologies, particularly in cases where the actual distribution of the data remains uncertain or when the data strays from the assumptions of parametric approaches. These techniques depend on a reduced number of assumptions, rendering them appropriate for a diverse array of scenarios, including situations involving intricate or non-uniform data structures.

The main objective of this inquiry is to present the Kernel Density Estimation, a widely used technique in statistical analysis, through an examination of various academic articles. To achieve this goal, this research paper undertakes a comprehensive analysis of eleven scholarly papers published between 2004 and 2018. By doing so, this investigation classifies eleven published studies from diverse academic journals that have employed the Kernel Density Estimation using different methodologies and backgrounds. Moreover, there has been a recent resurgence of interest in the computation of the non-parametric test of the Kernel Density Estimation, particularly in relation to the inference of estimating non-parametric test.

A non-parametric statistical method for estimating a random variable's probability density function (PDF) is called kernel density estimation (KDE). It is especially helpful in cases where the data's underlying distribution is complicated or unclear. By inserting a kernel, or smooth, symmetric function, at each data point and adding them up to get an approximation of the density, KDE offers a continuous, smooth representation of the data distribution.

To summarize, Kernel Density Estimation is a potent method that provides a flexible and easy way to comprehend the underlying distribution of random variables. It can be used to estimate probability density functions from data.

Key words: Non-parametric test, KDE, Testing Goodness-Of-Fit, Rank.

Introduction

To determine a random variable's probability density function (PDF), a statistical technique known as kernel density estimation (KDE) is utilized a collection of observed data points. Since KDE is non-parametric, it can capture the shape of the data distribution with more freedom than parametric approaches, which presuppose a certain functional form for the underlying distribution.

KDE's main objective is to present the data fluidly and continuously while providing insights into the underlying patterns and trends. When it is difficult to rely on conventional parametric models due to an unknown or complex true distribution of the data, this strategy is especially helpful.

The basic method by which KDE functions is to assign a kernel function to every data point and then sum these functions to provide a smoothed estimate of the PDF. Each data point's contribution to the total density is influenced by the kernel function, which functions as a weighting mechanism. The bandwidth is one of the most important KDE parameters since it controls the kernel's width and, in turn, the smoothness of the density estimate that is produced. Selecting the right bandwidth is crucial because it controls the trade-off between generating a smooth, generic representation and precisely capturing local fluctuations in the input.

Applications for Kernel Density Estimation can be found in many domains, such as signal processing, machine learning, statistics, and data analysis. It is frequently used as a key element in more intricate statistical procedures as well as for exploratory data analysis and visualization.

Kernel Density Estimation

In the field of statistics, kernel density estimation (KDE) refers to the application of kernel smoothing to approximate the probability density of a random variable, which fundamentally employs a non-parametric approach. This method heavily relies on the use of kernels as weights to estimate the probability density function.

Let (X1, X2, ..., Xn) denote a set of independent and identically distributed samples drawn from a univariate distribution characterized by an unknown density function f at any given point x. Our main focus is on estimating the shape of this function f. To achieve this goal, we utilize the kernel density estimator.

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where xi is a fixed position, $h > 0$ is a smoothing parameter known as the bandwidth, and K is the kernel, a non-negative function. The kernel that possesses a subscript "h" is referred to as the scaled kernel. It is precisely defined as Kh(x) = 1/h K(x/h). In order to optimize the estimator, it is intuitively desirable to select "h" to be as small as the data permits. Nevertheless, a delicate balancing act must be undertaken between the bias of the estimator and its variance.

Estimating the probability distribution is achieved through the utilization of Kernel Density Estimation (KDE). The underlying rationale of KDE can be succinctly described as follows: as the frequency of data points within a sample increases in the vicinity of a particular location, the probability of an observation transpiring at said location is correspondingly elevated.

Review Article

Chen, S., Hong, X. and Harris, C.J, (2004) this study proposed an efficient building algorithm for generating estimates of the density of sparse kernels. An orthogonal forward regression is used to guarantee the density construction's computational efficiency, and the procedure gradually reduces the leave-one-out test score.

To further impose sparsity, a native regularization technique is naturally included in the density creation process. The fact that the suggested algorithm is entirely automated and doesn't require the user to provide any criteria to end the density creation process is an extra benefit. This contrasts with a current state-of-the-art technique for kernel density estimation that uses the support vector machine (SVM) and requires the user to define several crucial algorithm parameters. There are several instances provided to show that the suggested approach can efficiently

provide a very spare kernel density estimate that is as accurate as the optimal Parzen window density estimate for the complete sample. Additionally, our experimental findings show that when it comes to testing accuracy and sparsity, the suggested technique performs comparably with the SVM method for estimating kernel density.

de Freitas, N. (2005) Improved Fast Gauss Transform, and Dual-Tree approaches for quick Kernel Density Estimation (KDE). Analyzed these methods' performance in terms of CPU time and memory use, taking into account the size, dimension, acceptable error, and structure (or "clumsiness") of the data set. In the literature, this is the first comparison of many methods. The results are surprising and refute a number of popular molds about these techniques. Investigators who are thinking about quick fixes for KDE issues might benefit from the findings. Along the way, we offer an IFGT algorithm parameter-selection regime and a corrected error bound. In his presentation, he provided an initial analysis of the foremost techniques employed for Kernel Density Estimation (KDE). Our analysis encompassed the manipulation of not only the quantity of interrelated points, N, but also the data's inherent structure, the desired level of accuracy, and the dimensionality of the state space. The findings suggest that the effectiveness of the fast methods is contingent upon the presence of structure within the kernel matrix. The Fast Gauss Change, Improved Fast Gauss Transform, and Dual-Tree approaches for quick Kernel Density Estimation (KDE) are tested, and the results are shown. In our specific studies, we analyses how well these methods perform in a data set that is superior to the Anchors Hierarchy.

Cao, R. and Lugosi, G. (2005) Investigations are conducted into testing protocols built on the idea of decreasing the L1 space amid a kernel density estimate and any density in the proposed class. He provides over-all non-asymptotic limits for the test's power. It shows that two important factors determining the test's success are the attentiveness of the data-dependent leveling factor and the "size" of the proposed class of densities. For testing basic hypotheses, translation/scale classes, and symmetry, among other specific instances, no asymptotic performance

restraints and consistency are demonstrated. Additionally, simulations are run to contrast the method's behavior with the L2 density-based method because of Fan and the Kolmogorov-Smirnov test check into a few fundamental characteristics of tests using the suggested format in this part. Under general hypotheses on the class F and the factor of smoothing, they developed a number of findings.

Duong, T., Cowling, A., Koch, I. and Wand, M.P, (2008). estimated of multivariate kernel density yields data structure information. A method for determining whether features, including local extrema, are statistically significant is called feature significance. This research presents a methodology that combines hypothesis testing for modal areas with kernel density derivative estimators to determine article import in d-dimensional information. Distributional characteristics are provided for the slope and curvature estimators, and pointwise test statistics are obtained. A unique three-dimensional data visualization complements the theoretical background. Applications to real data sets illustrate the successful performance of tests based on the kernel curvature estimators to find modal zones. Corresponding testing using kernel gradient estimators can improve these results. For one- and two-dimensional data, local extrema, hills, valleys, and vertical slopes are measured features of significance. Significant modal areas or local maxima effectively capture significant aspects for three- and top-dimensional data. Our real-world data examples highlight the usefulness of curvature-based tests in identifying important modal areas that may be presented in a form that is simple to understand. The comprehension of important structure in data can be improved using gradient-based testing of significant gradient areas.

Langlois, T.J., Fitzpatrick, B.R., Fairclough, D.V., Wakefield, C.B., Hesp, S.A., Mclean, D.L., Harvey, E.S. And Meeuwig, J.J.,(2012) By comparing the length-frequency acquired for three heavily fished species using a particular application of the Kernel Density Estimate (KDE) method and the well-known Kolmogorov–Smirnov (KS) test, they evaluated the biases and selectivity of stereo-BRUVS and line fishing. The outcomes of the KS and KDE tests did not differ from one another; yet, KDE offered a data-driven technique for estimating length-

frequency data. to a probability function, which serves as a practical means of characterizing and evaluating any variations between length-frequency samples. A surprising resemblance was discovered between the length-frequency distributions obtained from fishery-independent line fishing surveys of three exploited teleost's and those obtained from stereo-BRUVS. In contrast to line fishing, there was no indication of any bias or skew towards smaller fish using stereo-BRUVS, which defies our hypothesis.

Oliveira-Santos, L.G.R., Zucco, C.A. and Agostinelli, C. (2012) An essential component of physical behaviour and ecology is movement. Still, there haven't been many developments in this area of analysis, and most analyses of activity data use categorical or linear techniques that only provide a limited number of conclusions. The model he presented in this paper is a round, unceasing, nonparametric model of a conditional circular kernel function. This model can be utilized to estimate the activity range of a species, as well as the overlap in activities across different species. In our study, they analyze the impact of the model parameters (specifically, the density isopleth and kernel smoothing parameter) on the estimations of the activity range and activity overlap of animals in the Pantanal wetlands of Brazil, using a dataset from camera traps. To provide an example, also provide a full activity case study of two native peccary species and feral pigs coexisting together in the Pantanal. An essential component of animal behaviour and ecology is activity. However, there haven't been many analytical advancements in this area, and the most common methods for analyzing activity data are categorical or linear approaches. We want to improve our understanding of animal activity by applying our model to ecology. The Supplementary Data contains R codes for activity overlap and activity ranges, which are previously included in the package "circular."

Lahr, H. (2014) This study points out advances in methods for detecting disruptions in scaled earnings or earnings prediction error distributions, which are used to test for earnings management. The suggested test process tackles the crucial issue of bandwidth choice by means of a bootstrap test to endogamies the choice

phase, whereas previous systems use preselected bandwidths for kernel density estimation and histogram creation. Rather than assuming an accurate reference distribution, the bootstrap procedure's primary improvement over earlier approaches is that it produces a reference distribution that is globally indistinguishable from the practical distribution. This process provides an easy approach for locating and testing a local discontinuity while restricting the researcher's degrees of freedom. He estimated profits, earnings changes, and earnings prediction using the bootstrap density estimation method. mistakes made by US businesses between 1976 and 2010. The significance thresholds from previous research are drastically lowered, frequently to negligible amounts. Analysts forecast mistakes and are unable to identify discontinuities, though previous research has shown discontinuities that can be explained by a straightforward rounding method. Contrasts between various bandwidths Previous research has demonstrated a considerable trend towards significant outcomes, according to by criteria for selection and kernel functions. For the normal kernel function, which is most frequently used in the literature, this bias is very pronounced.

Pavia, J.M. (2015) The most often used method for determining the goodness-of-fit of a section to an incessant random distribution has been to measure the modification between the empirical cumulative distribution function and the null hypothesis cumulative distribution function using either L1- or L2-norms. This work (i) presents the GoF Kernel package, and (ii) suggests geometric. test, a novel implementation of the test that quantifies the difference between a sample, and (iii) conducts a substantial simulation exercise to evaluate the sensitivity and calibration of the tests described above as well as the Fan's test which is also included in the GoF Kernel package. The GoF Kernel package includes a few functions that R users may also find useful in adding to dgeometric test and fan test: For bounded random variables, reflected extends density enables the production of consistent kernel density estimates, while random.

Gonzalez, R., Huang, B. and Lau, E. (2015). One of the primary factors causative to PCA's effectiveness is its capacity to identify issues and provide some indication of their location. Nevertheless, Gaussian models are used by PCA and the T-test, which might not be appropriate for detecting process errors. An earlier version of this method used independent component analysis (ICA) for dimension reduction and kernel density estimation for anomaly detection, like PCA, classifies the location of the issues based on linear data-driven methodologies, yet without the presumptions of Gaussian. When used in conjunction with multivariate kernel density estimation, the dimension reduction approach is useful for non-linear connections involving non-Gaussian variables. A comparison is made between the performance of Bayesian networks, ICA, and PCA using data from an industrial size location. They compared methods for dimension reduction and statistical anomaly criteria in this article. PCA/ICA and Bayesian networks were examined for dimension reduction. In an online environment, the Bayesian network technique demonstrated a notable benefit. When making split-second judgements, the Bayesian dimension reduction approach makes it possible to use engineering knowledge in defining the connections, including causality.

The T-test used in PCA and the kernel density estimation (KDE) used in ICA were evaluated for statistical anomaly criteria. and solutions for Bayesian networks. The KDE approach demonstrated lower and more stable false positive rates, providing frights with the more dependable and consistent performance that operators want.

Chen, Y.C. (2017), A moderate overview to kernel density estimation (KDE) and the latest developments in confidence groups and regular/topological features are given in this lesson. He starts by going over the fundamental characteristics of KDE, such as bandwidth selection, density derivative estimates, and convergence rate under different metrics.

He discussed recent developments in the use of KDE to infer geometric and topological characteristics of a density function. Lastly, we show how to estimate a receiver operating characteristic curve and a cumulative distribution function using KDE. At the end, we include R solutions that are important to this topic. We

went over the fundamental characteristics of KDE and its uses in predicting the underlying density function structures in this session. Even though KDE has been extensively researched since its launch in the 1960s, there are still unresolved issues that demand more research. Here, we quickly go over a few unresolved issues pertaining to the course content.

Forte, J.P., Brilha, J., Pereira, D.I. and Nolasco, M.(2018) Over the past 20 years, there has been an increase in the body of research pertaining to the idea of geodiversity. Since it examines the relationships between abiotic factors, the quantification of three-dimensional designs of geodiversity appears to be one of the most exciting fields of natural diversity study. This last factor may be very important for both biodiversity conservation efforts and territory administration. This work's primary goal was to create a novel geographic information system (GIS) process based on centered analysis, compute a geodiversity index using kernel density, and evaluate its use in two cities with different area surfaces and geological settings. The suggested approach is an improvement over before reported techniques based on a landscape-scale spatial grid structure. The method's results demonstrate that lithology and geomorphology are the primary factors influencing the index and that it is feasible to create a spatial geodiversity standard that accurately captures the spatial variance of normal abiotic components on both territories. Furthermore, the testing processes have shown that this technology is applicable at multiple levels and in any geological and geomorphological context. It is also an important instrument for land use planning.

Conclusion

To sum up, Kernel Density Estimation provides a flexible and easy way to comprehend the underlying distribution of random variables. It is a strong and adaptable method for estimating probability density functions from observed data.

In addition, Kernel Density Estimation (KDE) is an effective and flexible statistical method that does not require predetermined distribution assumptions for estimating probability density functions. Because of its non-parametric characteristics, it works especially effectively in scenarios with complicated or uncertain underlying data distributions. KDE offers a continuous and fluid depiction of the data via the use of bandwidth settings and kernel functions, facilitating a detailed comprehension of the data's underlying structure. Numerous domains, including statistics, data analysis, machine learning, and signal processing, have found extensive use for KDE. It is a crucial tool in exploratory data analysis and visualization because of its capacity to provide insights into the distribution of random variables. Furthermore, KDE is an essential component of more sophisticated statistical techniques, such as non-parametric regression.

Nonetheless, careful consideration of the choice of kernel and bandwidth characteristics is necessary for the proper use of KDE. Because of the subjectivity of this selection process and its sensitivity to outliers, careful consideration is needed to balance capturing local variability in the data with producing a smooth, generalizable estimate. Large datasets can also provide difficulties because to the computational complexity of KDE, forcing practitioners to base their decisions on the particulars of their data.

In the decision, Kernel Density Estimation is still a strong and popular method for figuring out the underlying distribution of data. Its adaptability and capacity to manage non-standard data situations make it a vital methodology for statisticians, data scientists, and researchers looking for a reliable method, to density estimation.

References

1. Chen, S., Hong, X. and Harris, C.J., 2004. Sparse kernel density construction using orthogonal forward regression with leave-one-out test score and local regularization. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 34(4), pp.1708-1717.
2. Chen, Y.C., 2017. A tutorial on kernel density estimation and recent advances. Biostatistics & Epidemiology, 1(1), pp.161-187.
3. Cao, R. and Lugosi, G., 2005. Goodness-of-fit tests based on the kernel density estimator. Scandinavian Journal of Statistics, 32(4), pp.599-616.
4. de Freitas, N., 2005. Empirical Testing of Fast Kernel Density Estimation Algorithms (No. UBC TR-2005-03).
5. Duong, T., Cowling, A., Koch, I. and Wand, M.P., 2008. Feature significance for multivariate kernel density estimation. Computational Statistics & Data Analysis, 52(9), pp.4225-4242.
6. Forte, J.P., Brilha, J., Pereira, D.I. and Nolasco, M., 2018. Kernel density applied to the quantitative assessment of geodiversity. Geoheritage, 10, pp.205-217.
7. Gonzalez, R., Huang, B. and Lau, E., 2015. Process monitoring using kernel density estimation and Bayesian networking with an industrial case study. ISA transactions, 58, pp.330-347.
8. Langlois, T.J., Fitzpatrick, B.R., Fairclough, D.V., Wakefield, C.B., Hesp, S.A., McLean, D.L., Harvey, E.S. and Meeuwig, J.J., 2012. Similarities between line fishing and baited stereo-video estimations of length-frequency: novel application of kernel density estimates. PLoS One, 7(11), p.e45973.
9. Lahr, H., 2014. An improved test for earnings management using kernel density estimation. European Accounting Review, 23(4), pp.559-591.
10. Oliveira-Santos, L.G.R., Zucco, C.A. and Agostinelli, C., 2013. Using conditional circular kernel density functions to test hypotheses on animal circadian activity. Animal Behaviour, 85(1), pp.269-280.

11. Pavia, J.M., 2015. Testing goodness-of-fit with the kernel density estimator: GoFKernel. Journal of Statistical Software, 66, pp.1-27.

STAT.Sarween Ahmed Maroof@MSc2024.Non-Parametric

ORIGINALITY REPORT

20%	8%	17%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	L.G.R. Oliveira-Santos, C.A. Zucco, C. Agostinelli. "Using conditional circular kernel density functions to test hypotheses on animal circadian activity", Animal Behaviour, 2013 Publication	2%
2	Langlois, Timothy J., Benjamin R. Fitzpatrick, David V. Fairclough, Corey B. Wakefield, S. Alex Hesp, Dianne L. McLean, Euan S. Harvey, and Jessica J. Meeuwig. "Similarities between Line Fishing and Baited Stereo-Video Estimations of Length-Frequency: Novel Application of Kernel Density Estimates", PLoS ONE, 2012. Publication	2%
3	Ruben Gonzalez, Biao Huang, Eric Lau. "Process monitoring using kernel density estimation and Bayesian networking with an industrial case study", ISA Transactions, 2015 Publication	2%
4	repositorium.sdum.uminho.pt Internet Source	1%

5	RICARDO CAO. "Goodness-of-fit Tests Based on the Kernel Density Estimator", Scandinavian Journal of Statistics, 12/2005 Publication	1 %
6	gipeyop.uv.es Internet Source	1 %
7	Artur Gramacki. "Nonparametric Kernel Density Estimation and Its Computational Aspects", Springer Science and Business Media LLC, 2018 Publication	1 %
8	Ruben Gonzalez, Fei Qi, Biao Huang. "Process Control System Fault Diagnosis: A Bayesian Approach", Wiley, 2016 Publication	1 %
9	Henry Lahr. "An Improved Test for Earnings Management Using Kernel Density Estimation", European Accounting Review, 2014 Publication	1 %
10	Submitted to University of Edinburgh Student Paper	1 %
11	Submitted to Universiti Teknologi Malaysia Student Paper	1 %
12	oro.open.ac.uk Internet Source	1 %

13	en.wikipedia.org Internet Source	1 %
14	Duong, T.. "Feature significance for multivariate kernel density estimation", Computational Statistics and Data Analysis, 20080515 Publication	1 %
15	core.ac.uk Internet Source	<1 %
16	libuwspaceprd02.uwaterloo.ca Internet Source	<1 %
17	www2.uaem.mx Internet Source	<1 %
18	centaur.reading.ac.uk Internet Source	<1 %
19	www.econstor.eu Internet Source	<1 %
20	Alain Desgagné, Pierre Lafaye de Micheaux, Alexandre Leblanc. "Test of Normality Against Generalized Exponential Power Alternatives", Communications in Statistics - Theory and Methods, 2013 Publication	<1 %
21	www.informatik.uni-wuerzburg.de Internet Source	<1 %

22	Xin Peng, Yang Tang, Wenli Du, Feng Qian. "Online Performance Monitoring and Modeling Paradigm Based on Just-in-Time Learning and Extreme Learning Machine for a Non-Gaussian Chemical Process", Industrial & Engineering Chemistry Research, 2017 Publication	<1 %
23	Jiusun Zeng, Shihua Luo, Jinhui Cai, Uwe Kruger, Lei Xie. "Nonparametric Density Estimation of Hierarchical Probabilistic Graph Models for Assumption-Free Monitoring", Industrial & Engineering Chemistry Research, 2017 Publication	<1 %
24	John F. Rudge. "Finding peaks in geochemical distributions: A re-examination of the helium-continental crust correlation", Earth and Planetary Science Letters, 2008 Publication	<1 %
25	Xianghui Ning, Fugee Tsung. "A density-based statistical process control scheme for high-dimensional and mixed-type observations", IIE Transactions, 2012 Publication	<1 %
26	ebin.pub Internet Source	<1 %

27	Santiago Carrillo Menéndez, Bertrand Kian Hassani. "Expected Shortfall Reliability—Added Value of Traditional Statistics and Advanced Artificial Intelligence for Market Risk Measurement Purposes", Mathematics, 2021 Publication	<1%
28	Xiaolu Chen, Jing Wang, JingLin Zhou. "Probability Density Estimation and Bayesian Causal Analysis Based Fault Detection and Root Identification", Industrial & Engineering Chemistry Research, 2018 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		