

Recent Advances in Statistical Theory and Applications

Augustine Wong ^{1,*} and Xiaoping Shi ² ¹ Department of Mathematics and Statistics, York University, 4700 Keele Street, Toronto, ON M3J 1P3, Canada² Department of Computer Science, Mathematics, Physics and Statistics, University of British Columbia, Kelowna, BC V1V 1V7, Canada; xiaoping.shi@ubc.ca

* Correspondence: august@yorku.ca

Complex data pose unique challenges for data processing. Moreover, traditional inferential methods need to be modified or extended in order to fully utilize information provided by the complex data so as to obtain accurate inference. The present Special Issue of *Entropy*, entitled *Recent Advances in Statistical Theory and Applications*, has captured some recent progress in clustering, change point inference, multiple sample tests, generalized linear model, and machine learning. All the papers included in this Special Issue are motivated by complex data that arise from stochastic processes, time series, and medical images.

Cluster analysis is a statistical method for processing data. The aim is to classify data into groups or clusters based on how closely associated they are. Measuring the closeness of data can vary, and it depends on many factors such as the researchers' preferences and the type of data being analyzed. (contribution 1) considered the problem of obtaining an accurate clustering of unlabelled data. The common approach in solving this issue is ensemble clustering, which aims to combine sets of base clustering to obtain a better and more stable clustering and hence improve clustering accuracy. In this paper, the authors proposed a divergence-based locally weighted ensemble clustering with dictionary learning. Experimental results suggested that the proposed method is a promising method for ensemble clustering. (contribution 2) considered the problem of clustering mixed data that contain both continuous and discrete data, which arise frequently in medical and biological studies. It is an interesting topic of concern as most of the clustering methods found in the literature mainly target continuous data or discrete data alone. In this paper, using a weighted modified chi-squared test, the authors proposed a nonparametric clustering method to detect clustering patterns. Simulation studies showed that the proposed method outperforms AutoClass, which is the benchmark clustering method for mixed data. (contribution 3) examined data obtained from medical images, since an accurate and efficient clustering method will enable the early detection of possible tumors. It is well-known that the Euclidean distance-based Fuzzy C-means method is optimized to detect spherical structural clusters and does not perform well with high dimensional data. In this paper, the authors demonstrated that the Mahalanobis distance-based Fuzzy C-means method significantly outperformed the Euclidean distance-based Fuzzy C-means method.

A change point is a location or time at which observations obey two different models: before and after. Detecting change points is an interesting and important problem and has been studied by many authors. (contribution 4) incorporates prior information about the location of the change point to the current CUSUM-based statistics. The resulting class of weighted CUSUM statistics was shown to detect the change point with high accuracy. (contribution 5) combined neural network regression and location-scale CUSUM methods to detect the change point in autoregressive moving average models. Extensive simulations showed that the proposed method performs well.

Multiple sample tests exist frequently in statistical literature. The simplest example of a multiple sample test is the test of homogeneity of means using one-way ANOVA. However, when the underlying populations are not independently normally distributed, exact results



Citation: Wong, A.; Shi, X. Recent Advances in Statistical Theory and Applications. *Entropy* **2023**, *25*, 1661. <https://doi.org/10.3390/e25121661>

Received: 7 November 2023

Accepted: 5 December 2023

Published: 15 December 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

may not be available. In relation to this, (contribution 6) assumed the populations are independently distributed as gamma distributions. The problem of the testing homogeneity of means is of great research interest, especially because gamma distribution is frequently used in survival data and engineering statistics. Exact results, however, were not available in this study. The author proposed an improved log-likelihood ratio test instead, and simulation results showed an exceptional accuracy for the proposed method. (contribution 7) focused on the homogeneity test that evaluates whether two multivariate samples come from the same distribution. In this paper, the authors proposed an improved test statistic based on data depth, which could be Mahalanobis depth, spatial depth, projection depth, etc. The advantage of the data depth approach is that it is free of strong distributional assumptions; therefore, the depth can provide a ranking of multivariate data. The simulation results conducted demonstrate the superior performance of the proposed tests.

A stochastic process is a collection of random variables indexed by time. Therefore, it is often used to model data with a dependence structure. In (contribution 8), a general statistical framework is provided for using active information to quantify the amount of pre-specified external knowledge that an algorithm uses in order to reach a certain target. By iterating a Metropolis–Hastings type of the Markov chain, the authors were able to compute the algorithm’s active information under the equilibrium and non-equilibrium of the Markov chain, with or without stopping when the targeted set of fine-tuned states has been reached. (contribution 9) applied a temporal self-exciting point process model to the terror data. Moreover, by using a combination of simulation and the random forest method, an accurate prediction of the number of terror events was obtained.

Some types of time series data pose instability and long-term unpredictability, leading to challenges in accurate estimation and inference. (contribution 10) considered the mildly explosive autoregressive process, which can be used to test the explosive behavior of economic growth. In particular, the authors performed theoretical studies on the properties of this model with strong mixing errors. The results are applied to data from the NASDAQ composite index. The subject matter of (contribution 11) is the chaotic time series model, which has the characteristics of internal randomness, nonlinearity, and long-term unpredictability. The authors investigated phase space reconstruction, model training, and model selection in a chaotic time series. Numerical results show that the proposed model performs very well in multi-step predictions.

A generalized linear model is a flexible generalization of an ordinary linear regression model by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value. Hence, the response variable is no longer restricted to being normally distributed. The author of (contribution 12) studied the problem of bilinear regression and extended it to the problem of inductive matrix completion, where the response matrix contains missing data. The author applied the quasi-Bayesian method to address the problem of bilinear regression, and then adapted this approach to the problem of inductive matrix completion. An efficient gradient-based sampling algorithm designed to sample from the quasi-posterior distribution was ultimately developed in the paper. (contribution 13) derived the asymptotic normality of the subsampling M-estimator through Fisher information. Moreover, the asymptotic properties of subsampling estimators of an unbounded generalized linear model with non-natural links were also studied.

Lastly, (contribution 14) applied the machine learning method to improve the accuracy in predicting winter precipitation. The methodology was applied to winter precipitation data collected in Northern China from 1997 to 2018. The results show that the proposed method outperformed the benchmark method.

The collection of articles in this Special Issue of *Entropy* illustrated that challenges arise from analyzing complex data, and new methodologies are proposed to solve the associated problems. We hope that the articles included in this Special Issue will provide readers with beneficial information and helpful ideas for their own research, and, most importantly, that readers enjoy reading these articles.

Author Contributions: A.W. and X.S. wrote the paper. All authors have read and agreed to the published version of the manuscript.

Funding: Wong's work was supported by the NSERC Discovery Grant RGPIN 2017-05179. Dr. Shi's work was supported by NSERC Discovery Grant RGPIN 2022-03264, the Interior Universities Research Coalition and the BC Ministry of Health, and the University of British Columbia Okanagan (UBC-O) Vice Principal Research in collaboration with UBC-O Irving K. Barber Faculty of Science.

Acknowledgments: As Guest Editors, we would like to thank all the anonymous peer reviewers who evaluated the submissions, as well as the authors who contributed to the Special Issue on *Statistical Foundations of Entropy*. We would also like to express our gratitude to the editorial staff of *Entropy* for their fruitful collaboration.

Conflicts of Interest: The authors declare no conflict of interest.

List of Contributions

1. Xu, J.; Wu, J.; Li, T.; Nan, Y. Divergence-Based Locally Weighted Ensemble Clustering with Dictionary Learning and $L_{2,1}$ -Norm. *Entropy* **2022**, *24*, 1324. [[CrossRef](#)] [[PubMed](#)]
2. Xu, Y.; Gao, X.; Wang, X. Nonparametric Clustering of Mixed Data Using Modified Chi-Squared Tests. *Entropy* **2022**, *24*, 1749. [[CrossRef](#)] [[PubMed](#)]
3. Krasnov, D.; Davis, D.; Malott, K.; Chen, Y.; Shi, X. Fuzzy C-Means Clustering: A Review of Applications in Breast Cancer Detection. *Entropy* **2023**, *25*, 1021. [[CrossRef](#)] [[PubMed](#)]
4. Shi, X.; Wang, X.-S.; Reid, N. A New Class of Weighted CUSUM Statistics. *Entropy* **2022**, *24*, 1652. [[CrossRef](#)] [[PubMed](#)]
5. Ri, X.; Chen, Z.; Liang, Y. Detecting Structural Change Point in ARMA Models via Neural Network Regression and LSCUSUM Methods. *Entropy* **2023**, *25*, 133. [[CrossRef](#)] [[PubMed](#)]
6. Wong, A. Comparing Several Gamma Means: An Improved Log-Likelihood Ratio Test. *Entropy* **2023**, *25*, 111. [[CrossRef](#)] [[PubMed](#)]
7. Shi, X.; Zhang, Y.; Fu, Y. Two-Sample Tests Based on Data Depth. *Entropy* **2023**, *25*, 238. [[CrossRef](#)] [[PubMed](#)]
8. Díaz-Pachón, D.A.; Hössjer, O. Assessing, Testing and Estimating the Amount of Fine-Tuning by Means of Active Information. *Entropy* **2022**, *24*, 1323. [[CrossRef](#)] [[PubMed](#)]
9. Wang, S.; Wang, X.; Li, C. Modeling Terror Attacks with Self-Exciting Point Processes and Forecasting the Number of Terror Events. *Entropy* **2023**, *25*, 1011. [[CrossRef](#)] [[PubMed](#)]
10. Liu, X.; Li, X.; Gao, M.; Yang, W. Mildly Explosive Autoregression with Strong Mixing Errors. *Entropy* **2022**, *24*, 1730. [[CrossRef](#)] [[PubMed](#)]
11. Qiao, M.; Liang, Y.; Tavares, A.; Shi, X. Multilayer Perceptron Network Optimization for Chaotic Time Series Modeling. *Entropy* **2023**, *25*, 973. [[CrossRef](#)] [[PubMed](#)]
12. Mai, T.T. From Bilinear Regression to Inductive Matrix Completion: A Quasi-Bayesian Analysis. *Entropy* **2023**, *25*, 333. [[CrossRef](#)] [[PubMed](#)]
13. Teng, G.; Tian, B.; Zhang, Y.; Fu, S. Asymptotics of Subsampling for Generalized Linear Regression Models under Unbounded Design. *Entropy* **2023**, *25*, 84. [[CrossRef](#)] [[PubMed](#)]
14. Lang, Z.; Wen, Q.H.; Yu, B.; Sang, L.; Wang, Y. Forecast of Winter Precipitation Type Based on Machine Learning Method. *Entropy* **2023**, *25*, 138. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.