Research Statement

Subhrajyoti Roy

PhD Candidate,
Interdisciplinary Statistical Research Unit,
Applied Statistics Division,
Indian Statistical Institute, Kolkata

Principal Information Researcher,
SysCloud Technologies Pvt. Ltd.
8th Floor, ISprout, Sohini Tech Park,
Nanakramguda Rd, Financial District, Hyderabad

Research Interests

My research interests lie in robust statistics, high-dimensional data analysis, and spatio-temporal data analysis, with a strong focus on developing methods that are theoretically sound, computationally efficient, and practically impactful. Additionally, I am interested in exploring different applications of these methodologies in diverse domains such as video and image processing, epidemiology, and environmental sciences. A recurring theme in my current research work is leveraging robust statistical frameworks to handle real-world datasets that are often contaminated with noise and outliers, high-dimensional, or exhibit intricate temporal and spatial dependencies.

Research Contributions

My doctoral research has been centered around the development of robust matrix factorization methods using minimum divergence estimators. Minimum divergence estimators have a natural appeal to parametric statistical inference problems. Many of these estimators have strong robustness properties, in fact, as we show in [10], such estimators often have an asymptotic breakdown point free of data dimension. This property makes them extremely attractive in developing robust techniques for analyzing high-dimensional data. As an example, I introduced novel algorithms for Robust Singular Value Decomposition (rSVD) [9] and Robust Principal Component Analysis (rPCA) [8] based on the Minimum Density Power Divergence Estimator (MDPDE). I also show that these methods are embarrassingly parallel and scale to arbitrarily high-dimensional data, and exhibit strong robustness even under high contamination. Theoretical guarantees like equivariance, consistency, and asymptotic normality also establish the soundness of these estimators as illustrated in [8]. For reproducibility and broader dissemination of my work, I have published an R package `rsvdpd` [5] and a python package `decompy` [6].

There are multiple applications of the proposed robust matrix factorization techniques across different domains. In [9], I demonstrate one such application in developing a background modelling and foreground estimation technique on video surveillance data, which provides a simple unsupervised approach to solving this problem in contrast to the supervised state-of-the-art deep learning models. Further applications of the matrix factorization techniques based on minimum distance estimators include dimensionality reduction, matrix completion, image and video watermarking, latent semantic indexing, fraud detection, etc.

My master's dissertation was focused on the development of a robust and trustworthy dimensionality reduction technique [7]. In this, I proposed multiple criteria for measuring the reliability of any dimensionality reduction technique, and a novel method that adheres to these criteria. The algorithm combined multiple techniques from various fields, such as Voronoi tessellation from computational geometry and multidimensional scaling from machine learning. The resulting algorithm was found to be competitive with state-of-the-art algorithms like tSNE and UMAP for several benchmark datasets.

In non-parametric statistical inference, spatio-temporal modelling of the data is another research avenue that I have a keen interest in. This began with my research on the analysis of the environmental data on air pollution patterns of my hometown, Kolkata, India [11]. This research combined exploratory insights on environmental challenges in urban areas and the detection of temporal and spatial patterns using nonparametric smoothing techniques. Another collaborative research contribution of mine is the development of a nonparametric quantile regression modelling framework for spatio-temporal data [2]. Beyond the usual theoretical guarantees like consistency, asymptotic normality, and derivation of simultaneous confidence intervals, we also show an application for estimating quantile levels of daily energy demands across different households throughout the years. This application has a broader impact on designing the supply chain of the energy markets. I have also worked on designing a novel Rough-Fuzzy Change Point Detection (roufcp) algorithm [1], which is particularly useful in the detection of gradual changepoints in time-series data. By combining rough set theory and fuzzy logic with nonparametric smoothing techniques, this work provides potential applications in detecting gradual changes in data streams, language patterns, and environmental and climatology data.

As a part of the multidisciplinary effort, I have also contributed towards developing a generalized epidemiological model incorporating dynamic changes to the population due to migratory movements, and the presence

of asymptomatic cases. This model, as shown in [3], has provided useful insights into prediction of a second wave for the spread of infectious diseases like COVID-19. Additionally, one of my research on ordinal response models [4] demonstrates how minimum density power divergence estimator can be used to provide robust and efficient estimators for solving classification problems.

Future Research Plans

Building onto my previous works, I aim to further advance the theoretical and computational aspects of robust statistics (based on minimum divergence estimation) and their applications. I plan to generalize my robust PCA and SVD methodologies to tensor decomposition, which has direct applications in hyperspectral imaging, neuroimaging, and explaining the structure of convolutional neural networks.

Due to my industrial experience as the Principal Information Researcher at SysCloud, I am also familiar with the advances in deep learning, specifically in deep generative models. One of my research interests is to explore the application of various statistical divergence measures in these deep generative networks (such as in score based generative modelling for Stable Diffusions), which may open up new directions of researches for enabling robustness and fairness guarantees for these neural network models.

With the rise of the Internet of Things (IoT) and sensor networks, spatio-temporal data is becoming increasingly prevalent. I plan to expand my works on nonparametric and spatiotemporal analysis with more generalized frameworks. One of my ongoing projects involves incorporating infinite dimensional historical covariates in a spatiotemporal model and supports irregularly spaced distribution of the observed points of space and time.

In addition to all of the above, I am always eager to collaborate with domain experts to apply my methodologies to novel areas such as climate science, finance, public health, etc. Additionally, I would continue to theoretically explore the limitations and various generalizations of divergence based estimators. In one of my current projects, we are exploring the properties of a generalized class of divergence that includes many popular divergences, such as the density power divergence, logarithmic super divergence and (ϕ, γ)-divergence.

By bridging theoretical advancements with practical applications, my future research endeavours will continue to address pressing challenges in robustness in statistics, machine learning and generative models, while contributing to their foundational understanding.

References

[1] R. Bhaduri, S. Roy, and S. K. Pal. Rough-fuzzy cpd: a gradual change point detection algorithm. *Journal of Data, Information and Management*, 4(3):243–266, 2022.

[2] S. Deb, C. Neves, and S. Roy. Nonparametric quantile regression for spatio-temporal processes, 2024.

[3] A. Ghatak, S. Singh Patel, S. Bonnerjee, and S. Roy. A generalized epidemiological model with dynamic and asymptomatic population. *Statistical Methods in Medical Research*, 31(11):2137–2163, 2022.

[4] A. Pyne, S. Roy, A. Ghosh, and A. Basu. Robust and efficient estimation in ordinal response models using the density power divergence. *Statistics*, 58(3):481–520, 2024.

[5] S. Roy. *rsvddpd: Robust Singular Value Decomposition using Density Power Divergence*, 2021. R package version 1.0.0.

[6] S. Roy. *decompy: Python package for robust algorithms of matrix decomposition and analysis*, 2024. Python package version 1.1.1.

[7] S. Roy. Trustworthy dimensionality reduction, 2024.

[8] S. Roy, A. Basu, and A. Ghosh. Robust principal component analysis using density power divergence. *Journal of Machine Learning Research*, 25(324):1–40, 2024.

[9] S. Roy, A. Ghosh, and A. Basu. Robust singular value decomposition with application to video surveillance background modelling. *Statistics and Computing*, 34(5):178, 2024.

[10] S. Roy, A. Sarkar, A. Ghosh, and A. Basu. Asymptotic breakdown point analysis for a general class of minimum divergence estimators, 2023.

[11] S. Roy, D. Sengupta, K. Rudra, and U. S. Saha. Analysis of pollution patterns in regions of kolkata. *Calcutta Statistical Association Bulletin*, 72(2):133–170, 2020.