Statistical outlier analysis on mouse electrophysiology data

Zahra Asiri[1], Dr. Pedro Maia[2],Dr.Yuan Bo Peng. [3]

University of Texas at Arlington, Department of Mathematics, Behavioral Neuroscience and Neurophysiology Lab

ABSTRACT

This study addresses outlier detection challenges in Local Field Potentials (LFPs) from rodent brain recordings, crucial for accurate scientific inferences. Using unsupervised techniques, it introduces a Cumulative Outlier Score (COS) method to distinguish genuine brain signals from experimental artifacts. Data preparation involves collaborative domain expertise, followed by exploratory analysis and unsupervised outlier detection using the PyOD library. The hypothesis posits that genuine outliers yield intermediate COS, while artifacts exhibit higher scores. Validation involves analyzing data from different experimental groups and assessing the impact of removing high and intermediate COS points on group separability. Anticipated outcomes include outlier profiles for each data point and cumulative outlier score computation, refining understanding of LFP-related variables and enhancing outlier differentiation. The study contributes to outlier detection in electrophysiology and offers a pedagogical platform for students to engage with practical data management, exploratory analysis, and advanced outlier detection techniques.

BACGROUND

Outlier detection is a statistical analysis technique applied to identify abnormal observations in data. These are the points that significantly deviate from the majority, often indicative of measurement errors, system faults, or novel discoveries. Outliers can profoundly affect the outcomes of data-driven decisions, making their accurate detection vital across fields such as finance, healthcare, and cybersecurity.

Traditionally, understanding and applying outlier detection methods required advanced statistical knowledge, limiting accessibility to those with specialized education. However, the democratization of data science compels us to create tools and resources that are inclusive of all learning backgrounds.

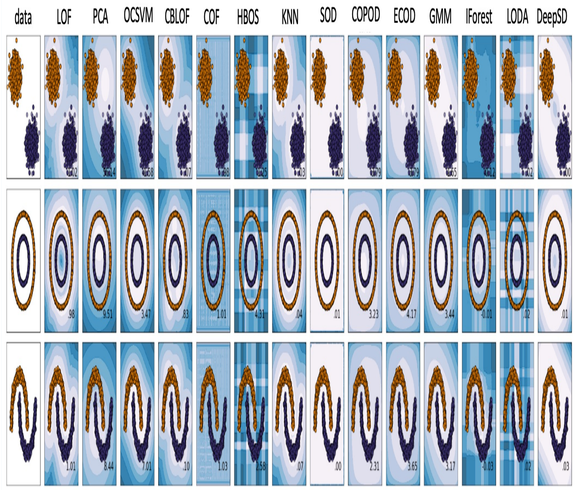
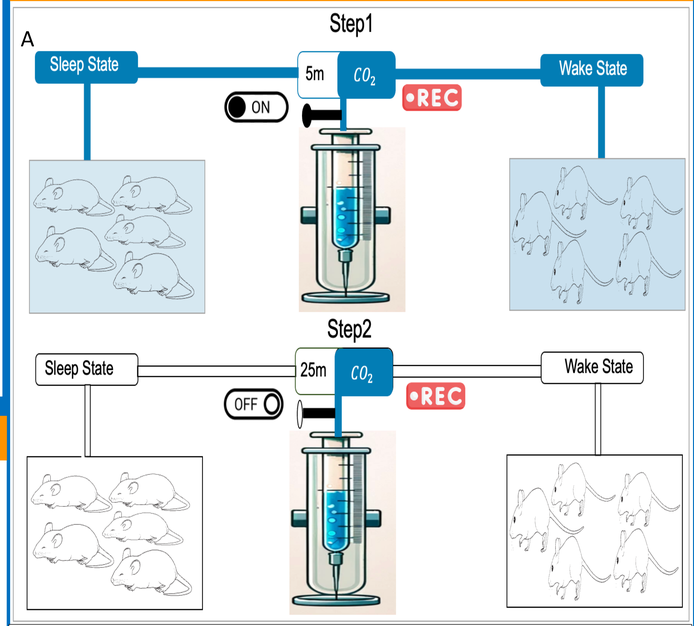


Figure 1:This grid of plots illustrates fourteen outlier detection techniques applied to three unique data configurations. Outliers are marked in orange, and inliers in blue, showcasing the algorithms' ability to discern between them. Below each plot, the numbers reflect the detection sensitivity or the specific score thresholds used by each technique. This comparative display serves as a practical reference for understanding the nuances and effectiveness of various outlier detection strategies.

Experimental setup



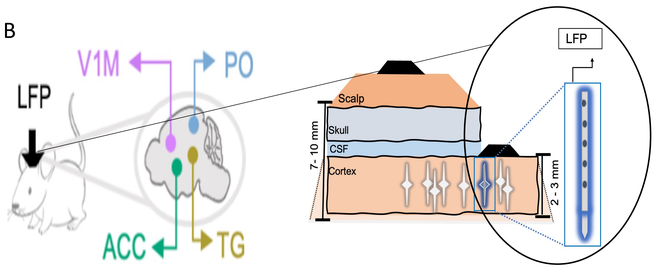


Figure 2:A. Experimental Setup - Step 1 & Step 2 This part of the image illustrates a two-step experimental process involving animals, presumably mice, in different states of sleep and wakefulness. In Step 1, the animals are shown in a sleep state, and some sort of mechanism (possibly to administer CO2) is turned 'ON' for a duration of 5 minutes, which is followed by a recording phase as the animals enter a wake state. In Step 2, the same mechanism is shown in the 'OFF' position for 25 minutes, and again there is a recording phase as the animal's transition to a wake state.

B. Brain Activity Recording Setup This section demonstrates a procedure for recording brain activity, labeled as LFP (Local Field Potential), from a mouse. Different brain regions are indicated by colored arrows (V1M, PO, ACC, TG), suggesting the locations where neural activity is monitored. The diagram shows the depth of insertion for the LFP recording device through the scalp, skull, CSF (Cerebrospinal Fluid), and cortex of the mouse's brain.

Research Goal

The research goal stated as follows:

1.Address the variability in outlierness scoring by different Outlier Detection (OD) methods by proposing a Combined Outlier Score (COS) that is robust and unifies these different scoring approaches.

2.Improve the statistical separability between distinct groups by proposing the removal of outliers with a high COS. This suggests that the presence of outliers can significantly impact the statistical differences between groups, and their removal could lead to clearer distinctions.

Outlier Detection (OD) Methods

different methods for detecting outliers in data. Each method uses unique criteria to determine what makes a data point an outlier, such as their statistical rarity, distance from other points or clusters, or position in a probability distribution. Methods include statistical models like GMM, dependency models like COPOD, projection-based methods like LODA, and isolation techniques like IForest. Each method offers a distinct approach to identifying data points that deviate significantly from the norm.

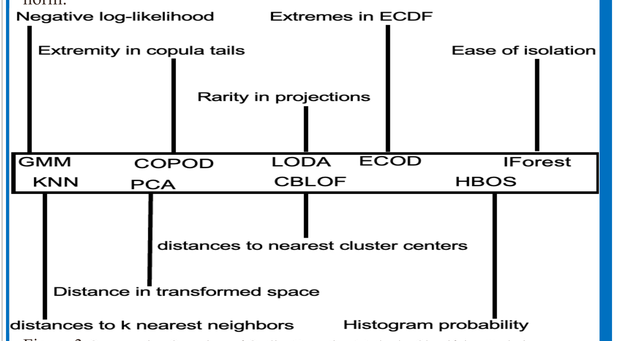


Figure 3:Comparative Overview of Outlier Detection Methods: Classifying Techniques Based on Negative Log-Likelihood, Tail Extremity, Rarity in Projections, Extreme ECDF Values, and Isolation Criteria.

This visualization allows for immediate, side-by-side comparisons of each algorithm's effectiveness, offering insights into their detection patterns, and sensitivity to different types of data distributions.

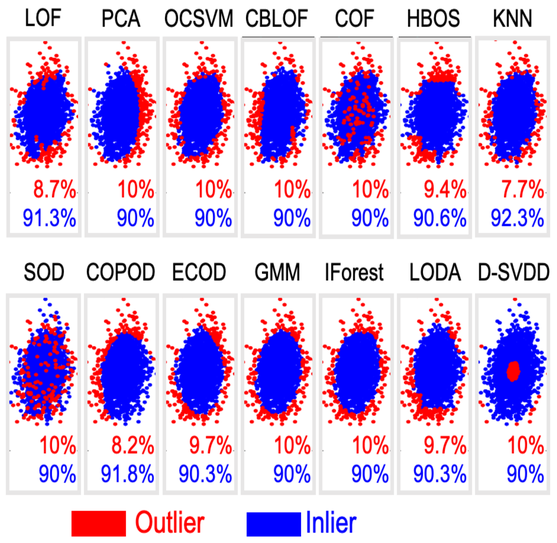


Figure 4:This grid of plots illustrates the performance of twelve different outlier detection algorithms on the electrophysiology data. Each plot demonstrates the outliers (red points) identified by a respective algorithm against the normal data points (inliers, blue points). The percentage values indicate the proportion of data points classified as outliers by each method. This visual comparison highlights the variance in sensitivity and specificity among the algorithms, providing a snapshot of their diverse detection capabilities.

Combined Outlier Score (COS) in Multimodal Detection Techniques

The Combined Outlier Score (COS) is presented as an innovative approach that synthesizes the diverse outlier detection measures from multiple established methods into a single cohesive score. This innovation aims to enhance the reliability and effectiveness of outlier detection by leveraging the strengths of each method to compensate for their individual weaknesses. COS simplifies the decision-making process in identifying outliers by providing a unified score rather than multiple conflicting ones addressing the research goal 1. This toy example, while not representing real data, showcases the clarity COS brings to outlier detection in a straightforward, understandable manner, differentiating it from the complexity of real-world data analysis.

Data point	GMM	COPOD	LODA	ECOD	IForest	KNN	PCA	CBLOF	HBOS	COS
x_1	0.91	0.39	0.92	0.05	0.70	0.10	1	0.66	0.02	0.52
x_2	0.65	0.61	0.96	0.15	0.70	0.20	0.61	0.72	0.06	0.51
x_3	0.26	0.98	0.03	0.03	0.17	0.20	0.98	0.93	0.09	0.40
:	:	:	:	:	:	:	:	:	:	:
x_n	0.27	0.77	1	0.06	1	0.30	0.77	0.45	0.02	0.51

Data point	GMM	COPOD	LODA	ECOD	IForest	KNN	PCA	CBLOF	HBOS	COS
y_1	0.32	0.92	0.26	0.10	0.64	0.53	0.19	1	0.24	0.46
y_2	0.92	0.25	0.07	0.30	0.34	0.20	0.25	0.03	0.24	0.28
y_3	1	0.35	0.19	0.87	0.40	0.20	0.35	0.15	0.56	0.45
:	:	:	:	:	:	:	:	:	:	:
y_m	0.09	0.56	0.46	0.12	1	0.40	0.56	0.44	0.38	0.44

Figure 5:Our toy example is a simplified data set consisting of two series of data points,x_1, x_2, x_3,..., x_n and y_1, y_2, y_3,..., y_m , each scored by different outlier detection algorithms like GMM, COPOD, and IForest. These scores indicate the likelihood of each data point being an outlier.

The idea of addressing the research goal 2 is represented in the figure highlights the effect of different actions on the separability of two groups. Separability refers to how distinct the groups are from each other, which is crucial in statistical analysis, especially in classification tasks.

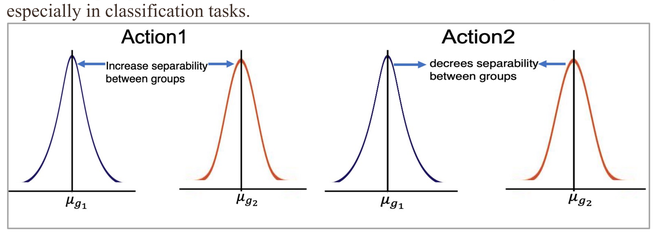


Figure 6: Effects of Actions on Group Separability: The left plot, titled 'Action1', shows increased separability between two groups with means μg1 and μg2, demonstrated by the distinct peaks of their distribution curves. The right plot, titled 'Action2', illustrates decreased separability where the overlap between the two groups is significant, leading to less distinct group definitions. This visual comparison underscores the importance of actions that can either enhance or diminish the clarity with which groups can be distinguished in data analysis.

Results

Our t-test results confirm that Action 1 ($p = 0.001$) successfully increases the separability of groups, underlining important biological differences. On the flip side, Action 2's nonsignificant result ($p = 0.89$) suggests that neglecting outliers can lead to an overlap of groups, masking key biological distinctions.

Action	t-Statistics	P-value
1	3.43	0.001
2	5.09	0.89

Table 1:Statistical Results of Group Separability: The table shows t-statistics and p-values for two actions. Action 1 yields a significant result ($p = 0.001$), indicating enhanced group distinction. Action 2 shows no significant effect ($p = 0.89$), suggesting outliers may obscure group differences.

Conclusion and Recommendation

Action 1 significantly clarifies group differences, validating its biological impact.
Prioritize outlier removal to maintain data integrity and reveal true biological patterns

References

- Wang, Zhen and Peng, Yuan B .Multi-region local field potential signatures in response to the formalin-induced inflammatory stimulus in male rats.Elsevier,2022.
- Charu C Aggarwal and Charu C Aggarwal. An introduction to outlier analysis. Springer, 2017.