Statistical Anomaly Detection in Turbulent Flow Data

Vivek Mathew

1 Introduction

This report presents a statistical analysis of turbulent flow data, specifically focusing on the identification of rare events or "intermittency." The dataset comprises energy measurements from two distinct eddy sizes, denoted as variables x_1 and x_2. The primary objective was to characterize the statistical behavior of these signals, test the validity of standard Gaussian assumptions, and implement a robust anomaly detection algorithm to isolate extreme physical events from the nominal turbulent fluctuations.

2 Visual Inspection of Feature Space

The initial phase of analysis involved visualizing the data structure in a two-dimensional feature space. Figure 1 displays the scatter plot of the two energy signals.

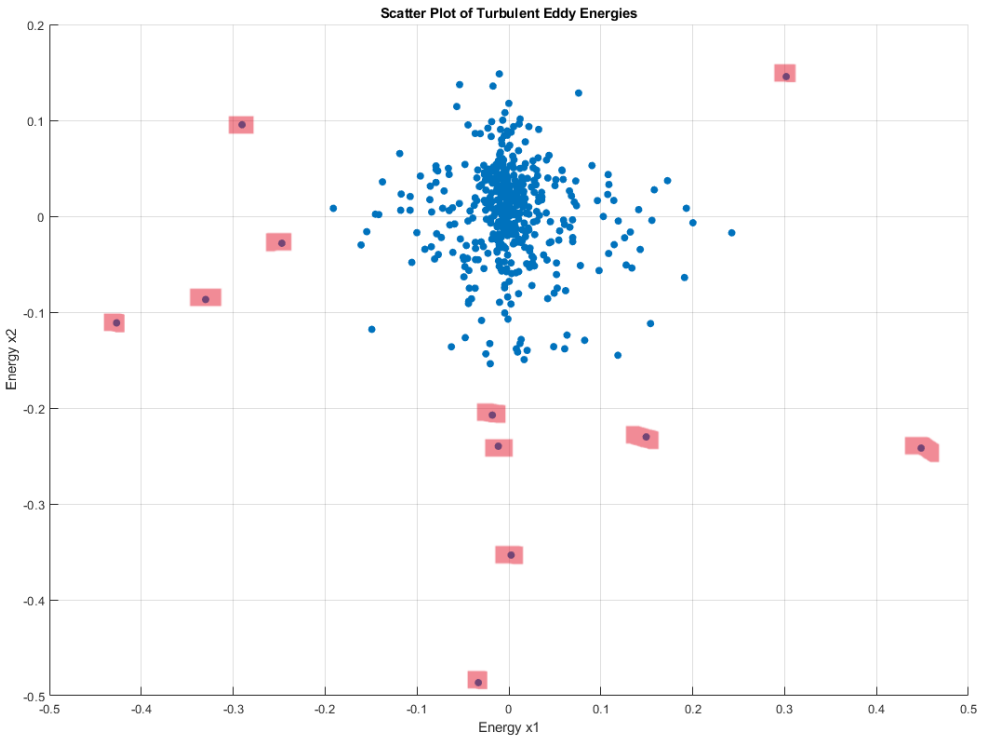

Figure 1: Scatter plot of turbulent eddy energies. The majority of the data forms a dense central cluster representing nominal flow conditions. The red boxes highlight manually identified outliers located at significant distances from the center. The distribution exhibits an elliptical shape, indicating differing variances between the two signals.

Visual inspection reveals that the data is concentrated around the origin, representing the average energy state. "Rare events" appear as outliers significantly removed from this central density. Furthermore, the spatial distribution indicates that the signal x_1 possesses a larger dynamic range (higher variance) than x_2, resulting in a horizontal spread. This suggests that energy fluctuations are more intense in the first eddy scale compared to the second.

3 Statistical Characterization

3.1 Validity of Gaussian Approximation

A critical step in anomaly detection is determining if the data follows a normal distribution. We compared the probability density functions (PDFs) of both signals against theoretical Gaussian curves derived from the data's mean and standard deviation.

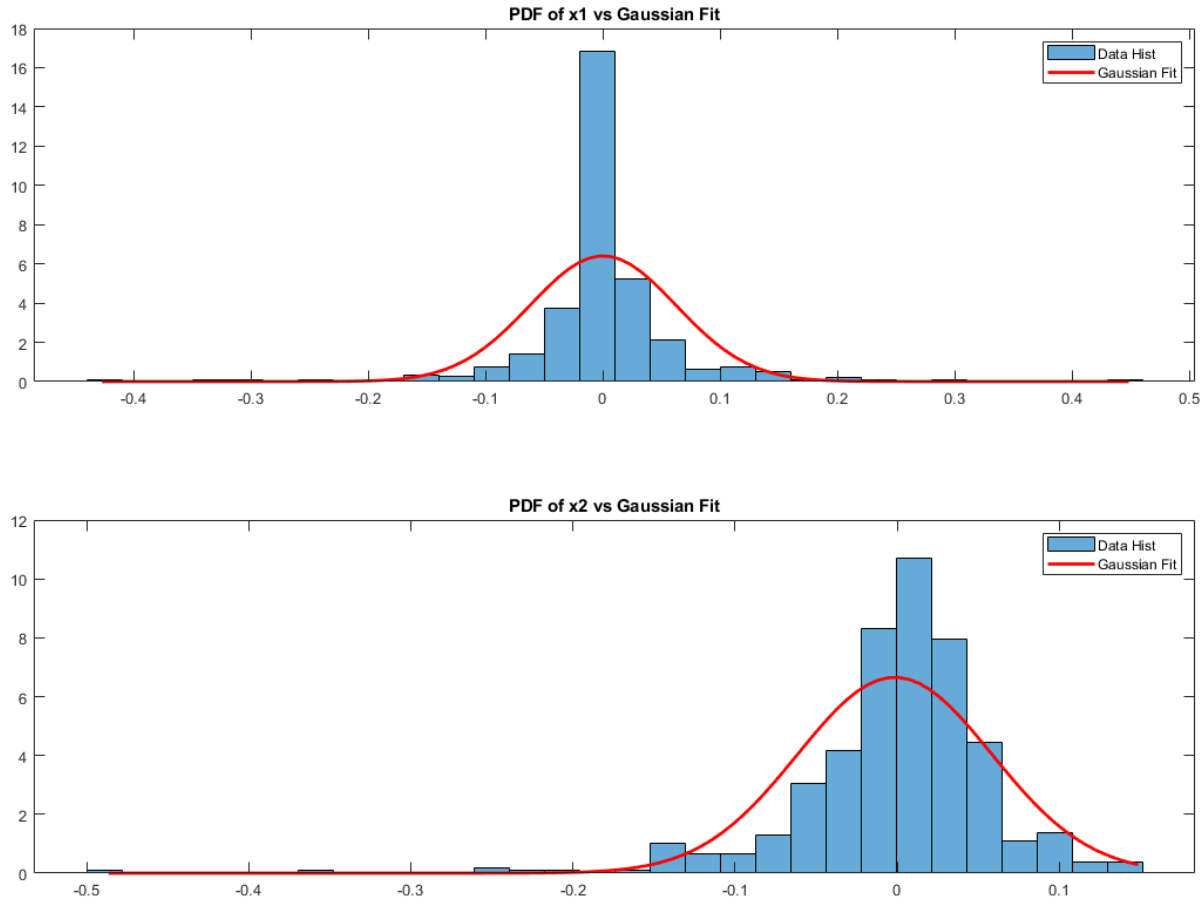

Figure 2: Histograms of signal x_1 (top) and x_2 (bottom) overlaid with theoretical Gaussian fits (red lines). The actual data shows values at the edges that persist well beyond where the Gaussian curve drops to zero.

As illustrated in Figure 2, the Gaussian model is an imperfect approximation for this turbulent dataset. The histograms clearly show that there are data values at the edges of the distribution far from the center that occur much more frequently than the Gaussian curve predicts. For example, in the far negative region of x_2, the Gaussian model predicts a probability effectively of zero, implying such events are impossible. However, the data bars show that these events do exist. Consequently, while the Gaussian model provides a baseline for determining the center of the data, it underestimates the likelihood of the extreme values at the edges.

3.2 Independence Analysis

To justify calculating the joint probability as the product of marginal probabilities, we evaluated the statistical independence of the two signals. We calculated the Pearson correlation coefficient (ρ), which is defined as the covariance of the two variables divided by the product of their standard deviations:

$$\rho = \frac{\text{Cov}(x_1, x_2)}{\sigma_{x_1}\sigma_{x_2}}$$

The covariance term in the numerator (Cov) represents the average of the product of deviations. If x_1 and x_2 tended to rise or fall together, the covariance would be positive. If one rose while the other fell, the covariance would be negative.

Our calculation resulted in a value of -0.0583. Because this value is so close to zero, it indicates that the positive and negative deviations effectively cancel each other out in the summation. This confirms that there is no significant linear relationship between the signals. Therefore, we can treat x_1 and x_2 as statistically independent variables for the purpose of this analysis.

4 Probability-Based Anomaly Detection

4.1 Calculation of Joint Probability and Threshold Sweeping

To quantitatively identify anomalies, we determined the probability density for every individual data point in the feature space. Relying on the statistical independence established in Section 3.2, the joint probability density was calculated as the simple product of the marginal Gaussian probabilities for each variable:

$$P(x_1, x_2) = P(x_1) \cdot P(x_2)$$

Using this method, a specific probability score was assigned to each of the 499 data points.

To determine the appropriate definition of a "rare event," we performed a sensitivity analysis by varying the density threshold (ϵ). After manually sweeping through a range of ϵ values. For each iteration of the loop, the script compared every data point's probability score against the current ϵ and counted the number of points falling below that threshold. This process allowed us to construct a full sensitivity curve.

Figure 3 displays the results of this sweep. As the threshold ϵ increases, the criteria for "rarity" becomes less strict, causing the count of detected anomalies to rise sharply as points from the dense central cluster are included.

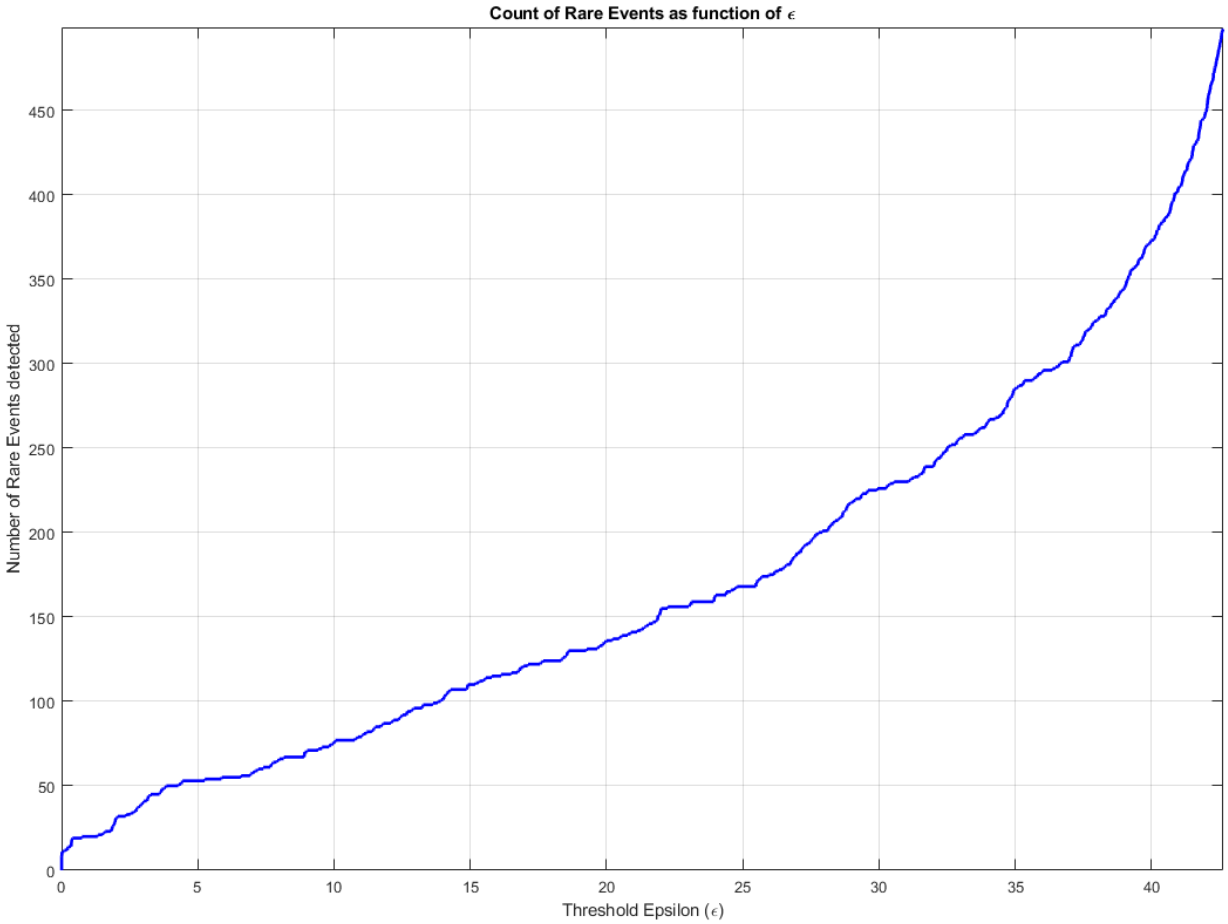

Figure 3: Full sensitivity curve showing the number of detected rare events as a function of the threshold ϵ. The curve illustrates the trade-off between detection sensitivity and specificity.

To select a physically meaningful threshold, we examined the curve for regions of stability. In this context, a stable count over a range of thresholds implies a natural "probability gap" between the true outliers and the nominal data.

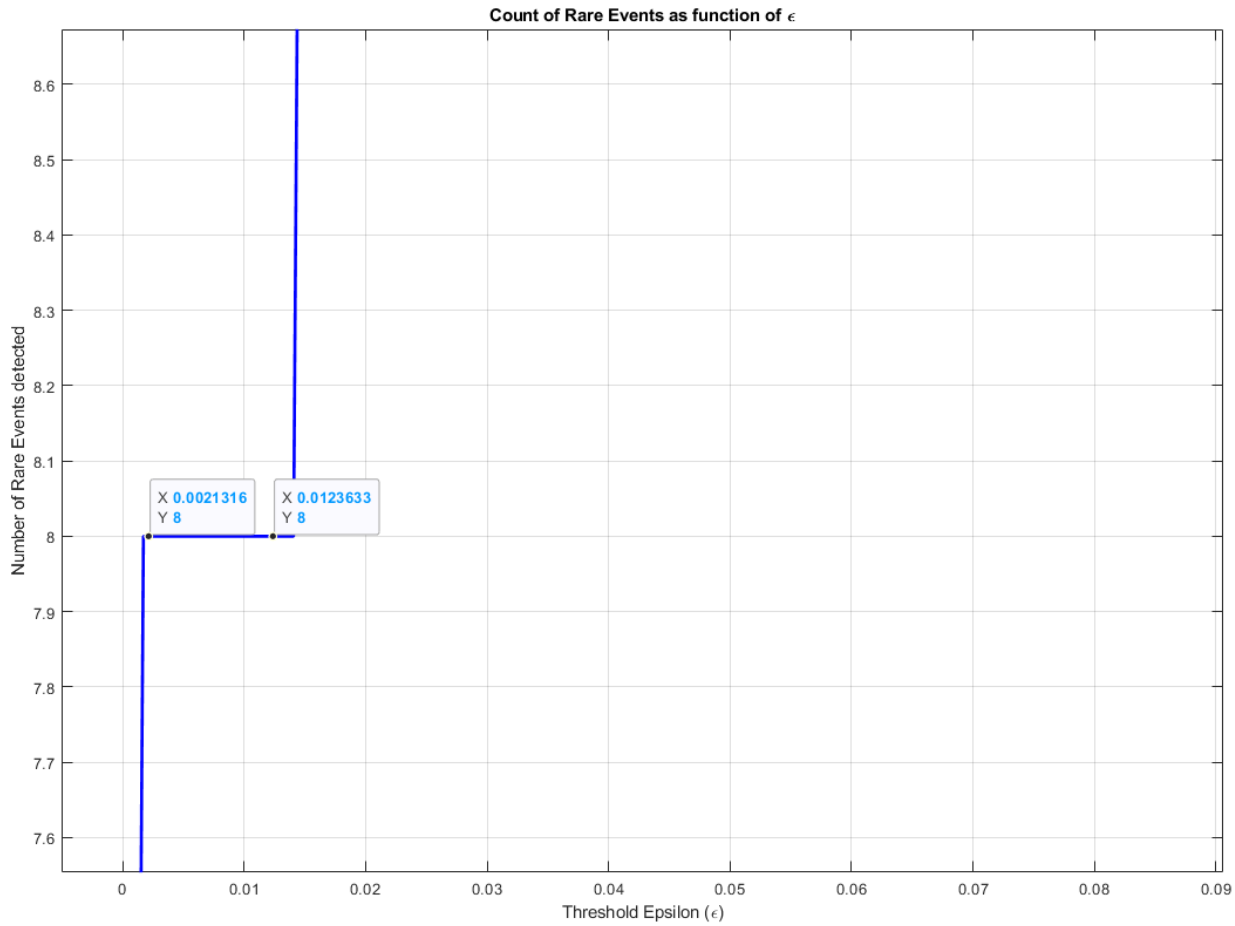

Figure 4: Zoomed view of the sensitivity curve focusing on the low-ϵ region. A distinct stability plateau is observed where the anomaly count remains constant at 8 events across the range $0.002 < \epsilon < 0.012$.

Figure 4 reveals a distinct stability plateau. For threshold values between approximately 0.002 and 0.012, the number of detected events remains constant at exactly 8. This plateau indicates that the 8th most extreme event is separated from the 9th event by a significant difference in probability density. Based on this analysis, we selected a threshold within this stable region ($\epsilon \approx 0.012$) to reliably isolate these 8 distinct anomalies.

5 Conclusion

The final classification results are presented in Figure 5. Based on the stability plateau identified in the sensitivity analysis, code was written to specifically filter and extract the 8 data points possessing the lowest joint probability scores.

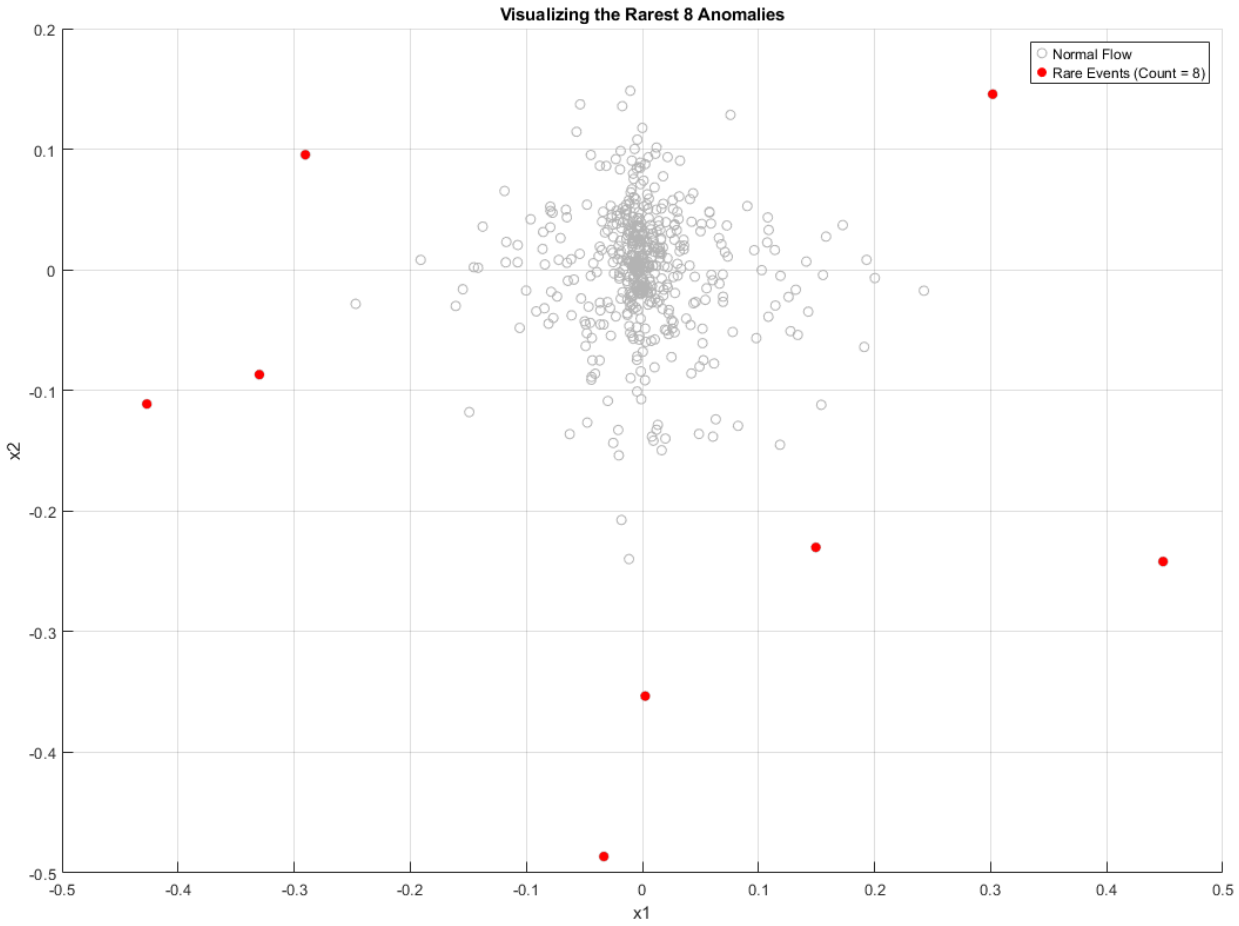

Figure 5: Final visualization of the anomaly detection results. The red points represent the 8 rare events identified via the stability plateau analysis, corresponding to approximately 1.6% of the dataset.

These computationally selected points were compared to the manual outliers identified via visual inspection in Section 2. The algorithmic selection aligned perfectly with the red boxes manually placed in the initial scatter plot. This confirms that the probability-based method, when calibrated using the stability plateau, effectively isolates the most distinct physical anomalies comprising approximately 1.6% of the dataset without arbitrarily including nominal fluctuations.