Supplimentary Materials of Comparative Probabilistic Analysis of Earthquake and Volcano Eruption: Novel Empirical Models & Directional Distributional Resemblances

Prithwish Ghosh, Debashis Chatterjee, Amlan Banerjee, Shiladri Shekhar Das

October 2023

The data generated for the analysis in this paper has been provided also as supplementary material.

1 Some Results regarding Regression Models

Degree of Polynomial	R^2 Value
1	0.01288
2	0.01675
3	0.01701
4	0.04834
5	0.04898
6	0.1018
7	0.1597
8	0.1891
9	0.2481
10	0.264

Table 1: This table gives us the R^2 values for different degrees of Polynomials. This is done concerning the median, where we used the Median of the Earthquake's Magnitude nearby vs the Median of the Volcano's Duration

2 Sinusoidal Regression

Sinusoidal regression is a statistical technique used to model data that exhibits periodic or oscillatory behavior. It is beneficial for analyzing phenomena that follow a sinusoidal pattern, such as seasonal trends, waveforms, or cyclic processes [5].

3 Kolmogorov Smirnov Test

The Kolmogorov-Smirnov (KS) test is a nonparametric statistical test used to assess whether a sample follows a particular probability distribution or if two samples come from the same underlying

Degree of Polynomial	R^2 Value
1	0.09941
2	0.09991
3	0.161
4	0.161
5	0.161
6	0.161
7	0.341
8	0.341
9	0.341
10	0.341

Table 2: This table gives us the R^2 values for different degrees of Polynomials. This is done concerning the mean, where we used the absolute mean of the Earthquake's Magnitude nearby vs the mean of the Volcano's Duration for the observed Earthquake count near the Volcano. The threshold distance is 100 km and three years.

distribution. It is based on the maximum vertical difference between the empirical distribution function (EDF) of the sample(s) and the cumulative distribution function (CDF) of the reference distribution.

For a one-sample KS test, the null hypothesis (H_0) is that the sample follows the reference distribution, while the alternative hypothesis (H_1) is that it does not. The KS statistic quantifies the maximum discrepancy between the EDF and the CDF. A smaller KS statistic suggests a better fit for the reference distribution.

In a two-sample KS test, the null hypothesis is that the two samples are drawn from the same distribution, and the test assesses whether the two samples significantly differ. The KS statistic helps determine the maximum difference between the two EDFs, indicating how dissimilar the two samples are.

The KS test is a powerful tool for comparing datasets, and it does not assume any specific distribution for the data. However, it becomes less reliable with smaller sample sizes [1].

4 Anderson-Darling Test for Multivariate Normality

The Anderson-Darling test is a statistical test used to assess whether a dataset follows a multivariate normal distribution. It extends the Anderson-Darling test for univariate normality to the multivariate case.

The null hypothesis (H_0) of the Anderson-Darling test is that the data is multivariate normally distributed, while the alternative hypothesis (H_1) is that the data departs significantly from multivariate normality.

The test statistic is computed based on the sample data and compared to critical values from the Anderson-Darling distribution. The null hypothesis is rejected if the test statistic exceeds the critical value, indicating non-multivariate normality.

The Anderson-Darling test is instrumental when dealing with multivariate data, such as data with multiple dependent variables, as it assesses normality across all dimensions simultaneously.

In practice, the Anderson-Darling test for multivariate normality is often implemented using statistical software or specialized libraries, as it involves complex calculations.

5 Cramér-von Mises Test for Multivariate Normality

The Cramér-von Mises test is a statistical test used to assess whether a given dataset follows a multivariate normal distribution. Multivariate normality is an essential assumption in many statistical analyses.

The Cramér-von Mises test for multivariate normality is based on the Cramér-von Mises statistic, which measures the discrepancy between the empirical distribution function (EDF) of the data and the EDF of a multivariate normal distribution with the same mean and covariance matrix. The test procedure can be summarized as follows:

1. Calculate the Cramér-von Mises statistic, which measures the squared difference between the two EDFs.
2. Under the null hypothesis that the data is multivariate normally distributed, the Cramér-von Mises statistic follows a chi-squared distribution.
3. Compare the calculated Cramér-von Mises statistic to the chi-squared critical value to determine whether the data is consistent with multivariate normality.

The results of the Cramér-von Mises test can be interpreted as follows:

- If the calculated Cramér-von Mises statistic is small and does not exceed the critical value, there is evidence to support the assumption of multivariate normality.
- If the calculated statistic is significant and exceeds the critical value, there is evidence to reject the assumption of multivariate normality.

The Cramér-von Mises test is a valuable tool for assessing the multivariate normality of data, which is an essential assumption in various statistical analyses. It helps researchers determine whether their data can be adequately modeled by a multivariate normal distribution.

6 Directional Statistical Preliminaries

We state some essential definitions and properties of certain directional statistical distributions and directional statistical tests in the Supplementary file.

6.1 Von Mises Distribution

A circular random variable θ follows the Von Mises distribution, also known as the Circular Normal Distribution if it is characterized by the probability density function (pdf): [6]

$$f(\theta; \mu, k) = \frac{1}{2\pi I_0(k)} e^{k \cos\theta(\theta-\mu)}, \tag{1}$$

In this equation, θ lies in the range $[0, 2\pi)$, μ is constrained to $[0, 2\pi)$, and $(k > 0)$. The normalizing constant $I_0(k)$ is the modified Bessel function of the first kind and order zero, given by:

$$\frac{1}{2\pi} \int_1^{2\pi} \exp(k \cos\theta) d\theta = \sum_{r=0}^{\infty} \left(\frac{k}{2}\right)^{2r} \left(\frac{1}{r!}\right)^2 \tag{2}$$

To determine the cumulative distribution of the circular normal or the Von Mises Distribution, we integrate the pdf, resulting in the following cumulative distribution function (cdf):

$$F(\theta) = \frac{1}{2\pi I_0(k)} \left(\theta I_0(k) + 2 \sum_{p=1}^{\infty} \frac{I_p(k) \sin p(\theta - \mu)}{p} \right), \tag{3}$$

where θ is confined to the interval $[0, 2\pi)$.

6.2 Von Mises-Fisher distribution

The von Mises-Fisher distribution is the natural extension of the von Mises distribution on the unit circle to the hypersphere of higher dimensions (sphere in our case). It is an important isotropic distribution for directional data and statistics. The von Mises-Fisher distribution is a distribution on the surface of a sphere. It has two parameters: the mean direction and the concentration (analogous to a normal distribution's mean and standard deviation). Its distribution, in terms of the point $\vec{x} = \{x_1, x_2, x_3\}$ on a circle of unit length is

$$f(\vec{x}; \vec{\mu}, \kappa) = C_3(\kappa) \exp\left(\kappa \vec{x} \cdot \vec{\mu}\right). \tag{4}$$

Here κ is the concentration, μ (a unit vector) is the mean direction, $\vec{x}$ is the random unit vector, and $C_3(\kappa)$ is the normalization coefficient, which can be shown to be dependent only on κ and the dimension (3, in case of the sphere). For details, we refer to [3], [7].

6.3 Wrapped Uniform Distribution

In the Wrapped Uniform Distribution, the total probability is uniformly spread out along the circumference of a circle. This distribution yields the Uniform Circular Distribution, characterized by a constant density given by: [6]

$$f(\theta) = \frac{1}{2\pi} \tag{5}$$

All directions on the circle are equally likely in this distribution, leading to its alternative names, such as the isotropic or random distribution.

6.4 Watson Test

In this study, we primarily employed Watson-type tests to examine whether the positional data adhere to either a Von Mises Distribution or a Circular Uniform Distribution.

[8] introduced a statistic for directional data, similar to the Kolmogorov-Smirnov nonparametric test, to assess the goodness-of-fit of one-sample and two-sample data concerning the uniform distribution or Von Mises Distribution. Watson's statistic is defined as follows:

$${W_n}^2 = \int_0^{2\pi} \left[(F_n - F) - \int_0^{2\pi} (F_n - F)\, dF \right]^2 dF \tag{6}$$

Where W_n represents Watson's statistic, $F_n(\alpha)$ denotes the empirical distribution function, which is based on the ordered observations $\alpha_{(1)} \cdots \alpha_{(n)}$ of a sample of independent and identically distributed variables $\alpha_1, \alpha_2, \cdots, \alpha_n$ drawn from the distribution $F(\alpha)$. F represents the actual distribution function, i.e., $F = F_0(\alpha)$. An alternative representation of Watson's statistic is given by:

$$W_n{}^2 = \sum_{i=1}^{n} \left[\left(U_{(i)} - \frac{i - \frac{1}{2}}{n} \right) - \bar{U} - \frac{1}{2} \right]^2 + \frac{1}{12n}. \tag{7}$$

Here, $U_i = F(\alpha_i)$, and the Cramer-Von-Mises statistic can be viewed as the "second moment" of $(F_n - F)$. Watson's statistic resembles the expression for "variance" in certain aspects.

Using Watson tests, we can examine whether the given positional data aligns with the expected distributions, aiding our analysis of the underlying patterns in the data.

7 A spherical Mixture model based Statistical Approach

Formally, a mixture model corresponds to the mixture distribution representing the probability distribution of observations in the overall population. The Gaussian mixture model is commonly extended to fit a vector of unknown parameters (here, Von Mises Fisher Distribution).

we define the density for directional parametric mixture distribution $p(x|\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_K)$ represent appropriate parameter set, as follows.

$$p(\boldsymbol{\theta}) = \sum_{i=1}^{K} \phi_i F(x|\theta_i). \tag{8}$$

- K = number of mixture components
- N = number of observations
- $\theta_{i=1\ldots K}$ = parameter of distribution of observation associated with component i
- $\phi_{i=1\ldots K}$ = mixture weight, i.e., the prior probability of a particular component i
- Φ = K-dimensional vector set $\phi_{1\cdots K}$; sum to 1
- $z_{i=1\ldots N}$ = component of observation i
- $x_{i=1\ldots N}$ = observation i
- $F(x|\theta)$ = probability distribution of an observation, parametrized on θ
- $z_{i=1\ldots N}$ = Categorical(ϕ)
- $x_{i=1\ldots N}|z_{i=1\ldots N} = F(x_{i=1\ldots N}|\theta_{z_i})$

In our paper, the i th vector component is characterized by $F(x|\theta_i)$ as either von Mises-Fisher distribution with weights ϕ_i, parameter θ_i as means $\boldsymbol{\mu_i}$ and concentration matrices $\boldsymbol{\kappa_i}$, or as circular uniform.

For fitting the mixture Von Mises Fisher distribution to both datasets, we take the help of R package **movMF** [2].

8 Polynomial Regression

Polynomial regression is a type of regression analysis that models the relationship between the independent variable (x) and the dependent variable (y) as an n-th degree polynomial. It is an extension of simple linear regression used when the relationship between variables is not linear.

The polynomial regression model can be represented as:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n + \varepsilon$$

where y is the dependent variable, x is the independent variable, $\beta_0, \beta_1, \ldots, \beta_n$ are the regression coefficients, and ε represents the error term. [4]

Polynomial regression allows us to capture more complex relationships between variables. The choice of the degree (n) of the polynomial depends on the data and the nature of the relationship. Common degrees include linear ($n = 1$), quadratic ($n = 2$), and cubic ($n = 3$) polynomials.

Polynomial regression can be visualized by fitting a curve to the data points. Higher-degree polynomials can lead to overfitting, so it's essential to choose the degree carefully. Regularization techniques like ridge regression and lasso regression can be used to prevent overfitting.

In practice, polynomial regression is implemented in statistical software and data analysis tools, allowing you to fit and evaluate polynomial models.

References

[1] Vance W Berger and YanYan Zhou. Kolmogorov–smirnov test: Overview. *Wiley statsref: Statistics reference online*, 2014.

[2] Kurt Hornik and Bettina Grün. movmf: An r package for fitting mixtures of von mises-fisher distributions. *Journal of Statistical Software*, 58(10):1–31, 2014.

[3] Kanti V Mardia, Peter E Jupp, and KV Mardia. *Directional statistics*, volume 2. Wiley Online Library, 2000.

[4] Eva Ostertagová. Modelling using polynomial regression. *Procedia Engineering*, 48:500–506, 2012.

[5] Barry G Quinn. Estimating the number of terms in a sinusoidal regression. *Journal of time series analysis*, 10(1):71–75, 1989.

[6] Mardia K. V. *Statistics of Directional Data.* Academic Press, New York, 1972.

[7] Geoffrey S Watson. Distributions on the circle and sphere. *Journal of Applied Probability*, 19(A):265–280, 1982.

[8] Stanley Wheeler and Geoffrey S Watson. A distribution-free two-sample test on a circle. *Biometrika*, pages 256–257, 1964.