Directional Data Analysis for Line Segments

Yoshitomo Akimoto, Takenori Sakumura, and Toshinari Kamakura

***Abstract*—We will derive the LM test statistic for the conditional von Mises distribution for detecting non-uniformity of line segments spread on the two dimensional plane, which has relatively high performance even for small samples as the extension of V-test to half circle. We illuminate the performance of this test by conducting simulation studies and also apply to Japanese active faults. Finally, we will propose a new non-hierarchical clustering method based on angular dispersions.**

***Index Terms*—angular data, Rayleigh test, LM test, von Mises distribution, k-means.**

I. Introduction

IN Japan, we have so many earthquakes including smaller ones. On occasion of the 2011 Tohoku earthquake and the Great Hanshin earthquake, we had the great loss of human life in these disaster. The precise prediction of earthquakes has been expected for many years, but it is very difficult to include the estimations of locations and magnitudes. In the recent studies on the earthquakes, active faults are useful for estimating locations and magnitudes of earthquakes. Active faults are the discontinuity of strata which has the distortion of each rock plane. If the active fault breaks the stress, the accumulated geological energy will cause the earthquakes. In this article we investigate the distribution of active faults focusing on angles of faults devising the new test statistic with high performance even for small samples.

The research on active faults and earthquakes has been studied for a long time. Especially the investigation into active faults and outbreak probabilistic model is also studied and discussed in Japan. We use the active fault database collected by National Institute of Advanced Industrial Science and Technology[1]. The table I shows the specifications of the active faults in Japan. We should explore the property of active faults for anti-disaster from the statistical view-point.

TABLE I
SPECIFICATIONS OF THE ACTIVE FAULTS IN JAPAN

Number of faults named	559
Number of the sum of elements by segment	3069
Total length [km]	10803

II. Testing for directional uniformity

In the first place, we shall be aware of dangerousness of the active faults in the same directions, and we are interested in testing uniformity of the directions of faults against the uniform directions. Mardia[2] described the Rayleigh test when mean direction is given based on von Mises distribution ([3], [4]), whose density function is expressed by the following:

$$f(\theta;\mu,\kappa) = \frac{1}{2\pi I_0(\kappa)}\exp(\kappa\cos(\theta-\mu)).$$

Where κ and μ discribe the parameter of concentration and mean direction respectively. The score statistic is is obtained using the convenient parameter transformation,

$$\boldsymbol{\omega} = (\kappa\cos\mu, \kappa\sin\mu)^T.$$

Uniformity is corresponding to $\kappa = 0$, and the transformation gives rise to $\boldsymbol{\omega} = \mathbf{0}$, which can give us very simple score statistics free from any evaluation of parameter estimation.

$$S_1 \quad = \quad 2n\overline{R}^2.$$

Another test is a just variant of the above test statistic, which is also known as Rayleigh test or V-test ([5], [6], [7], [8], [2], [9], [10], [11], [12], [13]):

$$S_2 \quad = \quad \frac{2}{n}\left\{\sum_{i=1}^{n}\cos(\mu-\theta_i)\right\}^2.$$

The parameter μ included in the statistic S_2 is replaced by the specified direction θ_0. In our study we replaced θ parameter by theoretical mean value of uniform distribution on $(0, 2\pi)$, $\mathrm{E}(\theta_i) = \pi$. In our experiences S_1 statistic does not have good performance in that it does not assure that α-critical levels and high powers against alternatives (small κ) especially for small samples. In this article we will propose the new test statistic based on the LM test and investigate the behavior of this statistic and apply this test to active faults data in Japan. The normal von Mises distribution is defined on $(0, 2\pi)$, but our directions of active faults must be the distribution defined on $(0, \pi)$. We consider the following conditional distribution,

$$f(\theta_i;\mu,\kappa) \quad = \quad \frac{\frac{1}{2\pi I_0(\kappa)}\exp(\kappa\cos(\theta_i-\mu))}{\int_0^{\pi}\frac{1}{2\pi I_0(\kappa)}\exp(\kappa\cos(\theta-\mu))d\theta} \quad (1)$$
$$(0 \le \theta, \mu < \pi, \quad \kappa \ge 0)$$

Tedious calculations of likelihoods give us the following new test statistic:

$$S_3 \quad = \quad \frac{\left\{-\frac{2n}{\pi}\sin\hat{\mu} + \sum_{i=1}^{n}\cos(\hat{\mu}-\theta_i)\right\}^2}{\frac{n}{2} - \frac{4n}{\pi^2}\sin^2\hat{\mu}}.$$

Here we note that the theoretical mean value $\pi/2$ is preferable for $\hat{\mu}$ to test uniformity on $(0, \pi)$.

Manuscript received July 23, 2015, revised July 28, 2015. This work was supported by JSPS KAKENHI Grant Number 26240003.

Y. Akimoto is with Graduate School of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 1128551, Japan. e-mail: sky.a68@gmail.com

T. Sakumura is with Department of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 1128551, Japan. e-mail: sakumura@indsys.chuo-u.ac.jp

T. Kamakura is with Department of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 1128551, Japan. e-mail: kamakura@indsys.chuo-u.ac.jp

ISBN: 978-988-14047-2-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Fig. 1. The active faults in Japan

III. Small sample behavior of the test statistics

We study the small sample behavior of the test statistics based on the random segment spread on the two-dimensional space. The test statistic S_3 is basically explain to the half angular space $(0, \pi)$. For comparison of these statistics we generate the uniform random numbers on $(0, \pi)$, and we use the $2 \times \theta_i$ for S_1 and S_2 and θ_i for S_3, because the former two statistics are ones designed for testing for the uniformity of $(0, 2\pi)$. In our study we need just test statistic defined on half circle $(0, \pi)$. The Table II and Table III show that our proposed statistics conditioned on the half circle based on the conditional distribution has moderately good performance. We can say that the performance of the test statistics are as follows:

$$S_1 \prec S_3 \approx S_2.$$

However, as we conventionally use the test statistics S_1 and S_2 defined on full circle, and we must use the S_3 from the conditioned distribution.

TABLE II
The nominal and the actual levels of significance (random sample size of n from the conditional von Mises distribution $\kappa = 0$)

n		1%	2.5%	5%	10%
3	proposed A	0.005	0.027	0.067	0.142
	proposed B	0.000	0.001	0.016	0.141
	proposed C	0.007	0.017	0.035	0.097
	Rayleigh test	0.000	0.000	0.001	0.106
	V-test	0.000	0.072	0.174	0.298
5	proposed A	0.015	0.035	0.067	0.128
	proposed B	0.001	0.014	0.060	0.156
	proposed C	0.007	0.020	0.046	0.100
	Rayleigh test	0.001	0.016	0.043	0.095
	V-test	0.028	0.076	0.144	0.274
10	proposed A	0.012	0.027	0.053	0.107
	proposed B	0.008	0.028	0.065	0.135
	proposed C	0.009	0.023	0.048	0.100
	Rayleigh test	0.007	0.021	0.046	0.098
	V-test	0.032	0.078	0.147	0.264
20	proposed A	0.009	0.024	0.049	0.099
	proposed B	0.011	0.030	0.062	0.124
	proposed C	0.009	0.024	0.049	0.100
	Rayleigh test	0.008	0.023	0.048	0.099
	V-test	0.034	0.079	0.146	0.261
50	proposed A	0.010	0.025	0.049	0.099
	proposed B	0.011	0.029	0.058	0.113
	proposed C	0.010	0.025	0.050	0.100
	Rayleigh test	0.009	0.024	0.049	0.100
	V-test	0.035	0.081	0.146	0.259
100	proposed A	0.010	0.025	0.050	0.100
	proposed B	0.011	0.028	0.054	0.107
	proposed C	0.010	0.025	0.050	0.100
	Rayleigh test	0.010	0.025	0.050	0.100
	V-test	0.036	0.081	0.147	0.259

[A] the proposed statistics S_3 (the width of sample range for $\hat{\mu}$).
[B] the proposed statistics S_3 (the sample mean for $\hat{\mu}$).
[C] the proposed statistics S_3 (the theoretical mean for $\hat{\mu}$).

TABLE III
The powers of the test (random sample size of n from the conditional von Mises distribution $\kappa = 1$)

n		1%	2.5%	5%	10%
3	proposed A	0.008	0.123	0.234	0.366
	proposed B	0.000	0.004	0.053	0.277
	proposed C	0.012	0.029	0.054	0.114
	Rayleigh test	0.000	0.000	0.001	0.194
	V-test	0.000	0.139	0.292	0.441
5	proposed A	0.144	0.244	0.338	0.450
	proposed B	0.008	0.069	0.208	0.380
	proposed C	0.015	0.034	0.065	0.123
	Rayleigh test	0.007	0.067	0.145	0.256
	V-test	0.105	0.218	0.336	0.506
10	proposed A	0.249	0.344	0.433	0.537
	proposed B	0.114	0.233	0.353	0.496
	proposed C	0.021	0.045	0.080	0.142
	Rayleigh test	0.100	0.196	0.303	0.442
	V-test	0.249	0.397	0.532	0.676
20	proposed A	0.305	0.401	0.489	0.591
	proposed B	0.310	0.453	0.571	0.689
	proposed C	0.032	0.063	0.106	0.178
	Rayleigh test	0.325	0.470	0.594	0.721
	V-test	0.535	0.683	0.787	0.876
50	proposed A	0.343	0.447	0.540	0.644
	proposed B	0.738	0.830	0.888	0.933
	proposed C	0.068	0.121	0.186	0.281
	Rayleigh test	0.858	0.923	0.957	0.980
	V-test	0.943	0.974	0.988	0.995
100	proposed A	0.401	0.515	0.613	0.716
	proposed B	0.966	0.983	0.991	0.996
	proposed C	0.141	0.227	0.319	0.437
	Rayleigh test	0.997	0.999	1.000	1.000
	V-test	0.999	1.000	1.000	1.000

[A] the proposed statistics S_3 (the width of sample range for $\hat{\mu}$).
[B] the proposed statistics S_3 (the sample mean for $\hat{\mu}$).
[C] the proposed statistics S_3 (the theoretical mean for $\hat{\mu}$).

ISBN: 978-988-14047-2-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

IV. The application to active faults data

As we mentioned in the previous section, the active faults in the same direction may be dangerous. We test the uniformity of direction of line segments approximating the faults by picking up both ends of observed faults. The Figure 2, Table IV and Table V show the $q = 1 - p$-value. Thick color is corresponding to higher value of q. The high q-values indicates possibilities of non-uniformity of directions of faults.

Fig. 2. The non-uniformity of the active faults in Japan

TABLE IV
The Prefectures whose q-value is larger

Prefecture	q-value
Toyama	0.956
Osaka	0.836
Yamanashi	0.455
Fukuoka	0.359
Tokyo	0.351

TABLE V
The Prefectures whose q-value is smaller

Prefecture	q-value
Hokkaido	<2e-16
Yamagata	<2e-16
Akita	<2e-16
Oita	<2e-16
Nagasaki	<2e-16

A. generalized test

In addition, we consider that directional data exist in the restricted range (c_1, c_2). We derived the former conditional distribution function in (2).

$$f(\theta_i;\mu,\kappa) = \frac{\frac{1}{\pi I_0(\kappa)}\exp(\kappa\cos(2\theta_i-\mu))}{\int_{c_1}^{c_2}\frac{1}{\pi I_0(\kappa)}\exp(\kappa\cos(2\theta-\mu))d\theta}. \quad (2)$$
$$(c_1 \le \theta, \mu < c_2, \quad \kappa \ge 0)$$

Therefore, we get LM test statistic LM' is (3).

$$LM' = \frac{\left\{\sum_{i=1}^{n}\cos(\mu-2\theta_i) - \frac{n[\sin(\mu-2c_1)-\sin(\mu-2c_2)]}{2(c_2-c_1)}\right\}^2}{\frac{n\{\sin[2(\mu-2c_1)]-\sin[2(\mu-2c_2)]\}-4c_1+4c_2}{8(c_2-c_1)} - \frac{n\{\sin(\mu-2c_1)-\sin(\mu-2c_2)\}^2}{4(c_2-c_1)^2}}. \quad (3)$$

V. clustering

In this section, we consider the problem on clustering line segments based on the angular data. Here we note that the angular data are obserbed just in $(0, \pi)$, that is we cannot abtain the directions of the segments. Then, we must classify these data using the restricted method of k-means to a half circle. We propose the following method considering the measure of angular distance θ and α.

$$\arg\min_{\alpha_1,\alpha_2,\ldots,\alpha_K}\sum_{i=1}^{n}\min_{j}\left[1-\cos\left(2\theta_i-2\alpha_j\right)\right]. \quad (4)$$

Algorithm 1 directional clustering

Input:
K (number of clusters)
$\Theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$ (directional data of line segment)
Output:
$A = \{\alpha_1, \alpha_2, \ldots, \alpha_k\}$
$l(i)|i = 1, \ldots, n$ (cluster labels of θ_i)

$A \leftarrow$ choose K α_j's from angular data sample Θ as the nodes randomly
repeat
  $A_{previous} \leftarrow A$
  for $i = 1$ to n **do**
    $l(i) \leftarrow \arg\min_{j}\left[1-\cos\left(2\theta_i-2\alpha_j\right)\right]$
  end for
  for $j = 1$ to k **do**
    $A \leftarrow$ mean.direction$(\{\theta_i\}\,|l(i) = j)$ from (5)
  end for
until $A_{previous} \neq A$

Figure 3 shows the result of clustering Japanese active faults.

For comparing our proposed method and normal k-means, we plotted the histogram of angular data colored by the clusters. As illustration, look at the lineament data appeared the circular test book [14]. Our proposed method can handle the connectivity of the 0 and π in the case of the line segments data. Here we note that usual circular data have property the connection of 0 and 2π. Then the histogram Figure 4 is seemed to be separated dark colored one cluster considering the 0-π connectivity. However the standard k-means method results in detecting unusual two clusters owning to disconnection of 0 and π (see Figure 5).

ISBN: 978-988-14047-2-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

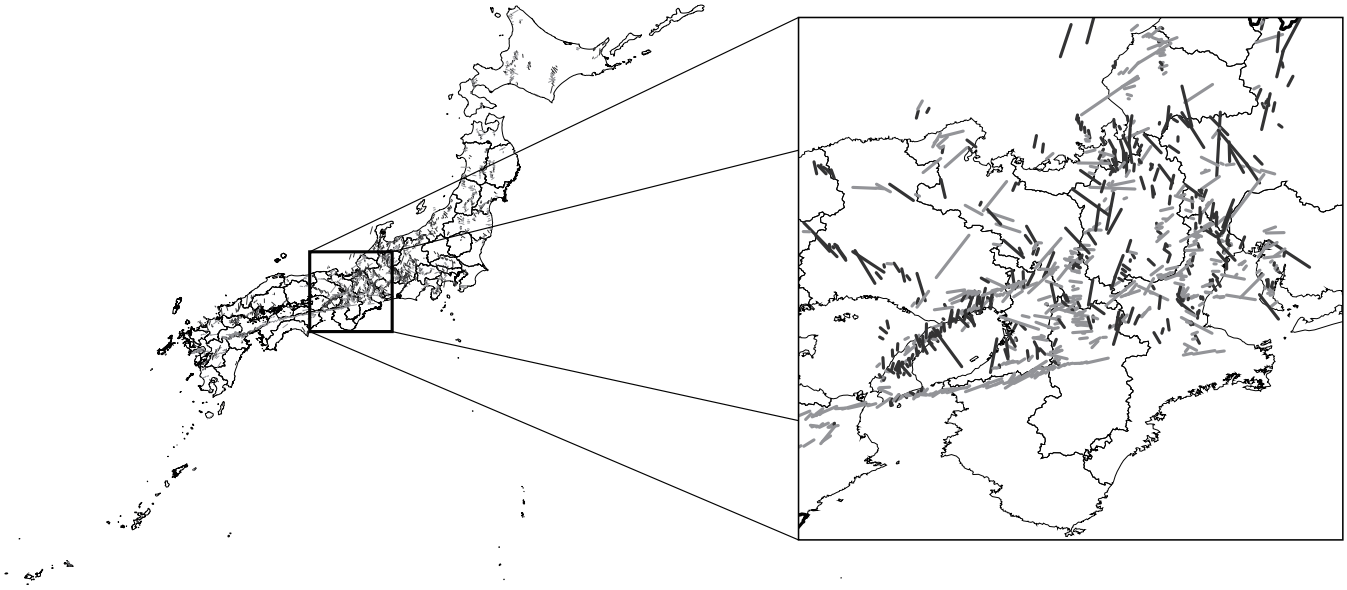

Fig. 3. The result of clustering Japan active faults

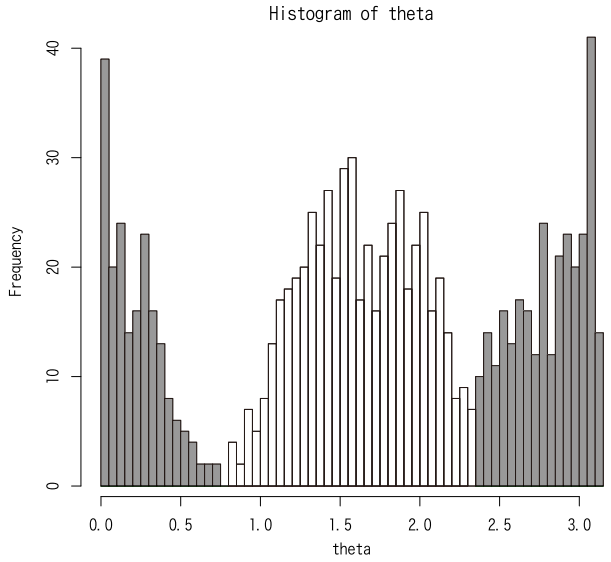

Fig. 4. The result of clustering Fisher's lineament data (proposed method)

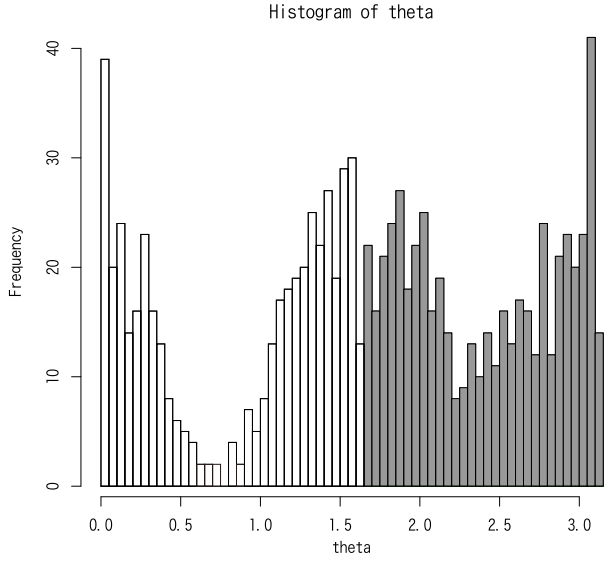

Fig. 5. The result of clustering Fisher's lineament data (normal k-means)

VI. Discussions

We derived the LM test statistic for the conditional von Mises distribution which has relatively high performance even for small samples as the extension of V-test to half circle, we could find the areas of non-uniformity of active faults. We also proposed the new method of clustering technique for angular data. Further investigations will be needed considering locations and length of the active faults.

Appendix A
Derivation of LM test statistic

Likelihood function and Log likelihood function of the conditional von Mises distribution (1) is

$$L(\theta_i;\mu,\kappa)=\prod_{i=1}^{n} f(\theta_i;\mu,\kappa),$$

$$\begin{aligned} l &= \sum_{i=1}^{n}\log f(\theta_i;\mu,\kappa) \\ &= \sum_{i=1}^{n}\left[\kappa\cos(\theta_i-\mu)-\log I_0(\kappa)- \right. \\ &\qquad \left. \log(2\pi)-\log Z(\mu,\kappa)\right], \end{aligned}$$

where I_0 denotes the modified Bessel function of the first kind and order 0, and $Z(\mu,\kappa)$ can be defined by

$$Z(\mu,\kappa)=\int_0^{\pi}\frac{1}{2\pi I_0(\kappa)}\exp(\kappa\cos(\theta-\mu))d\theta.$$

The score statistic with $\kappa = 0$ is calculated using first and second derivatives. Then, $Z(\mu,0)=1/2$, Take partial differentiation of $Z(\mu,0)$ with respect to each parameters μ

ISBN: 978-988-14047-2-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

and κ.

$$\begin{aligned}
\frac{\partial Z(\mu,0)}{\partial \mu} &= 0. \\
\frac{\partial Z(\mu,0)}{\partial \kappa} &= \int_0^\pi \left[\frac{1}{2\pi}\cos(\theta-\mu)\right] d\theta. \\
\frac{\partial^2 Z(\mu,0)}{\partial \mu^2} &= 0. \\
\frac{\partial^2 Z(\mu,0)}{\partial \mu \partial \kappa} &= \int_0^\pi \left[\frac{1}{2\pi}\cdot\sin(\theta-\mu)\right] d\theta. \\
\frac{\partial^2 Z(\mu,0)}{\partial \kappa^2} &= \int_0^\pi \left[\frac{1}{4\pi}+\frac{1}{2\pi}(\cos(\theta-\mu))^2\right] d\theta.
\end{aligned}$$

Thus, substitute each partial differentiation for this result where $A(\kappa) = I_1(\kappa)/I_0(\kappa)$.

$$\begin{aligned}
\frac{\partial l}{\partial \mu} &= \sum_{i=1}^n \left[0\cdot\sin(\theta_i-\mu) - \frac{1}{Z(\mu,0)}\frac{\partial Z(\mu,0)}{\partial \mu}\right] \\
&= 0. \\
\frac{\partial l}{\partial \kappa} &= \sum_{i=1}^n \left[\cos(\theta_i-\mu) - A(0) - \frac{1}{Z(\mu,0)}\frac{\partial Z(\mu,0)}{\partial \kappa}\right] \\
&= \sum_{i=1}^n [\cos(\theta_i-\mu)] - \frac{2n}{\pi}\sin(\mu). \\
\frac{\partial^2 l}{\partial \mu^2} &= \sum_{i=1}^n [-0\cdot\cos(\theta_i-\mu) \\
&\quad -\left(\frac{-1}{Z^2(\mu,0)}\left(\frac{\partial Z(\mu,0)}{\partial \mu}\right)^2\right. \\
&\quad \left.+\frac{1}{Z(\mu,0)}\frac{\partial^2 Z(\mu,0)}{\partial \mu^2}\right)\Bigg] \\
&= 0. \\
\frac{\partial^2 l}{\partial \mu \partial \kappa} &= \sum_{i=1}^n [\sin(\theta_i-\mu) \\
&\quad -\left(\frac{-1}{Z^2(\mu,0)}\frac{\partial Z(\mu,0)}{\partial \kappa}\frac{\partial Z(\mu,0)}{\partial \mu}\right. \\
&\quad \left.+\frac{1}{Z(\mu,0)}\frac{\partial^2 Z(\mu,0)}{\partial \mu \partial \kappa}\right)\Bigg] \\
&= \sum_{i=1}^n [\sin(\theta_i-\mu)] - \frac{2n}{\pi}\cos(\mu). \\
\frac{\partial^2 l}{\partial \kappa^2} &= \sum_{i=1}^n \left[-A'(0) - \left(\frac{-1}{Z^2(\mu,0)}\left(\frac{\partial Z(\mu,0)}{\partial \kappa}\right)^2\right.\right. \\
&\quad \left.\left.+\frac{1}{Z(\mu,0)}\frac{\partial^2 Z(\mu,0)}{\partial \kappa^2}\right)\right] \\
&= \frac{4n}{\pi^2}\sin^2(\mu) - \frac{n}{2}.
\end{aligned}$$

Finally, calculate the score statistic by Lagrange multiplier using Hessian matrix. we call this statistic LM test statistic.

$$\begin{aligned}
H &= \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} = -\begin{bmatrix} \frac{\partial^2 l}{\partial \mu^2} & \frac{\partial^2 l}{\partial \mu \partial \kappa} \\ \frac{\partial^2 l}{\partial \mu \partial \kappa} & \frac{\partial^2 l}{\partial \kappa^2} \end{bmatrix}. \\
J &= \begin{bmatrix} d_1 & d_2 \end{bmatrix} = \begin{bmatrix} \frac{\partial l}{\partial \mu} & \frac{\partial l}{\partial \kappa} \end{bmatrix}.
\end{aligned}$$

$$\mathrm{LM} = d_2^T (H_{22} - H_{12}^t H_{11}^{-1} H_{21})^{-1} d_2.$$

$$S_3 = \frac{\left\{-\frac{2n}{\pi}\sin\hat{\mu} + \sum_{i=1}^n \cos(\hat{\mu}-\theta_i)\right\}^2}{\frac{n}{2} - \frac{4n}{\pi^2}\sin^2\hat{\mu}} \overset{a}{\sim} \chi_1^2.$$

APPENDIX B
DIRECTIONAL STATISTICS

In directional statistics, we cannot calculate arithmetic mean, then the mean direction $\overline{\theta}$ is often used [9]. the mean direction is angle of resultant vector of observations. description of this in complex plane is (5),

$$\overline{\theta} = \arg\left\{\sum_{j=0}^n \cos\theta_j + i\sum_{j=0}^n \sin\theta_j\right\}. \quad (5)$$

Also, $\overline{R}$ that is the length of resultsnt vector $\boldsymbol{R}$ is used for a measure of concentration of a data set.

$$\begin{aligned}
\boldsymbol{R} &= \left(\sum_{i=1}^n \cos\theta_i, \sum_{i=1}^n \sin\theta_i\right). \\
\overline{R} &= \frac{\|\boldsymbol{R}\|}{n}.
\end{aligned}$$

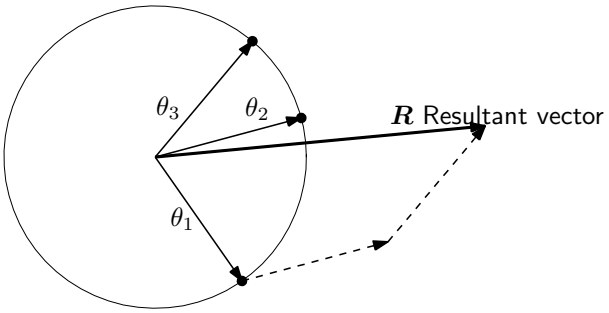

Fig. 6. The resultant vector which indicated the mean direction

REFERENCES

[1] G. S. of Japan of National Institute of Advanced Industrial Science and T. (AIST), "Active fault database of Japan," https://gbank.gsj.jp/activefault/index_e_gmap.html.
[2] K. Mardia, *Statistics of Directional Data*, ser. Probability and Mathematical Statistics a Series of Monographs and Textbooks. Academic Press, 1972.
[3] S. Jammalamadaka and A. Sengupta, *Topics in Circular Statistics*, ser. Series on multivariate analysis. World Scientific, 2001.
[4] R. Von Mises, "Über die " ganzzahligkeit " der atomgewichte und verwandte fragen," *Phys. z*, vol. 19, pp. 490–500, 1918.
[5] E. Batschelet, *Circular Statistics in Biology*, ser. Mathematics in biology. Academic Press, 1981.
[6] J. Davis, *Statistics and Data Analysis in Geology*. Wiley, 2002.
[7] D. Durand and J. A. Greenwood, "Random unit vectors ii: Usefulness of gram-charlier and related series in approximating distributions," *The Annals of Mathematical Statistics*, vol. 28, no. 4, pp. pp. 978–986, 1957.
[8] J. A. Greenwood and D. Durand, "The distribution of length and components of the sum of n random unit vectors," *Ann. Math. Statist.*, vol. 26, no. 2, pp. 233–246, 06 1955.
[9] K. Mardia and P. Jupp, *Directional Statistics*, ser. Wiley Series in Probability and Statistics. Wiley, 2000.
[10] L. R. F.R.S., "Xii. on the resultant of a large number of vibrations of the same pitch and of arbitrary phase," *Philosophical Magazine Series 5*, vol. 10, no. 60, pp. 73–78, 1880.
[11] L. R. O. F.R.S., "Xxxi. on the problem of random vibrations, and of random flights in one, two, or three dimensions," *Philosophical Magazine Series 6*, vol. 37, no. 220, pp. 321–347, 1919.

[12] M. A. Stephens, "Tests for randomness of directions against two circular alternatives," *Journal of the American Statistical Association*, vol. 64, no. 325, pp. 280–289, 1969.
[13] S. M. Swan, A. R. H., "Introduction to geological data analysis," *International Journal of Rock Mechanics and Mining Sciences and Geomechanics Abstracts*, vol. 32, no. 8, p. 387A, 1995.
[14] N. Fisher, *Statistical Analysis of Circular Data*, ser. Statistical Analysis of Circular Data. Cambridge University Press, 1995.

ISBN: 978-988-14047-2-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)