Two Optimal Measurements of Normality for Finite Set of Discrete Data

Ray-Ming Chen

Abstract—When one applies statistical models, he usually makes the assumption that the data or error come from a normal distribution. In order to verify or to estimate the parameters of the normal distribution, the typical approaches would be normality tests and model selections. In this article, we come up with two methods that could fit the data via normal distribution and measure the fitness. This would serve an alternative for model selection. Unlike statistical approaches - parametric or non-parametric statistics - we use approximation approaches to measure the optimal similarity between a given data set and its induced normal distributions and then search for the optimal normal distribution that has the best similarities. The degree of similarity is defined by two approaches: the overlapped area and the $arccos$ function. The idea is to look at the patterns between data set induced step function and sampled normal distributions in the form of approximated step probability density functions. Our analytical approach could measure the degree of normality, or the similarity with normal distributions, which could then used in pioneering findings for related statistical inferences.

Keywords—Normality, Step Functions, Sampling, Similarity, arccos

I. Introduction

There are two main issues regarding the normality of a distribution: normality test and estimation of parameters for normal distribution (or model selection). There are many theories in dealing with normality test: QQ-plot, Bayesian statistics, frequencist test (Razali, 2011), entropy test (Vasicek, 1976). There are also other statistical approaches to test the normality of a set of data (P., 2001; Jean, 2003). Whether a set of data could be fitted via a normal distribution is an important issue. It has wide applications and complications. However, normality test does not give us the exact parameters of the distribution. Our concern lies in estimation of the exact parameters of the normal distribution that fits the data most. The well known approach for such problem is Maximum Likelihood Estimation, which also has a wide application (D. 2019). As for the optimal estimation of the parameters, the optimal parameter for mean is the sampled mean and the optimal parameter for variance is the population variance (John, 1997; Russell, 2011). It seems the problem has been solved. However, if one checks its priori assumption: the observed data are all assumed to be independent distribution, he would know the unrealistic part. Such assumption is over-simplified. On the other hand, if one takes the dependency between random variables into consideration, then the optimal parameters for the set of multivariate normal distribution could only be obtained through a lengthy numerical iterative computational algorithms - for example, Gradient descent method, Newton–Raphson method, Davidon–Fletcher–Powell formula and so on (Nocedal, 2006; Fletcher, 1987). All these algorithms too complex and have less intuitive meaning. In this article, we put up with a method that could reduce such complex and enhance the intuitive sense of normality. The advantage for our approach is that it is easy to apply and much intuitive in interpretation. The disadvantages are it is easily disturbed by the data distribution and a sound justification of the goodness of fit. As for the measurement of normality, we put forward two measurements that could directly measure the normality, or the degree of similarity between a data set and normal distributions. The idea lies in treating the given data set as part of the approximation processes (M., 1981). If one looks at Riemann integral (Halsey, 2017), the way we calculate the integral of an function is applying approximation via step functions. Here we adopt the same algorithm by assigning a step function for the given data and by sampling the normal distribution to yield some finite set of points and then use these points to form step functions. Then we compare the similarity between the data set induced step function and the sample-induced step functions. We could then choose the normal distribution that would yield the optimal similarity for the given data set. The guidelines of the whole mechanism could refer to Section VI. Henceforth, there are several characteristics of this article.

- Firstly, unlike MLE or other statistical method, our estimation of parameters for a normal distribution does not depend on the assumption that each samples are independent;
- Secondly, we directly offer two methods to compute the similarities between data and the fitted normal distribution;
- Thirdly, our approach offers a much intuitive interpretation of the estimation of the parameters.
- Lastly, it could be used to observe the independence of data when one compares the computed normal distribution and the MLE model selection.

II. Maximal and Minimal Functions

A. Notations and Basic Definitions

Suppose

$$\vec{v} = (v_1, v_2, ..., v_{m-1}, v_m, v_{m+1}),$$

$$\vec{w} = (w_1, w_2, ..., w_{n-1}, w_n, w_{n+1})$$

Ray-Ming Chen is with the Department of Mathematics and Statistics, Baise University, Guangxi Province, China (e-mail:baotaoxi@163.com).

Manuscript received April 10, 2020; revised May 30, 2020.

are two ascending vectors, i.e.,

$$v_1 < v_2 < \cdots < v_{m-1} < v_m < v_{m+1},$$

$$w_1 < w_2 < \cdots < w_{n-1} < w_n < w_{n+1}.$$

For any vector $\vec{k}$, we use $|\vec{k}|$ to denote the length of $\vec{k}$, for example, $|\vec{v}| = m+1$. Let $\Delta\vec{v}_i = v_{i+1} - v_i$. Define first-order difference vector of $\vec{v}$ by

$$\Delta\vec{v} = (\Delta\vec{v}_1, \Delta\vec{v}_2, \cdots \Delta\vec{v}_{m-1}, \Delta\vec{v}_m)$$

For example, if $\vec{v} = (1, 1.2, 2.6, 3.0, 3.4, 4.1)$, then

$$\Delta\vec{v} = (0.2, 1.4, 0.4, 0.4, 0.7).$$

Let $\bullet$ denote the Euclidean inner product. For any vector $\vec{k}$, we use $\vec{k}_j$ to denote the j-th element of $\vec{k}$. For any arbitrary real finite set $\mathbb{D} \subseteq \mathbb{R}$, we use $av(\mathbb{D})$ to denote its ascending vector. For example if

$$\mathbb{D} = \{-1, -5, 0, 8, -6.2, 1.6, 2.8, 0.1\},$$

then $av(\mathbb{D}) = (-6.2, -5, -1, 0, 0.1, 1.6, 2.8, 8)$. If $\mathbb{D}$ has some repetitions, we use its ascending multi-set (Blizard,1991;Singh, 2007) (or $MS(\mathbb{D})$) form

$$\mathbb{D} \equiv (\alpha_1)^{i_1}(\alpha_2)^{i_2} \cdots (\alpha_p)^{i_p},$$

where $\alpha_1, \alpha_2, \cdots \alpha_p \in \mathbb{R}$ are the elements with the relation $\alpha_1 < \alpha_2 < \cdots < \alpha_p$, and where $i_1, i_2, \cdots i_p \in \mathbb{N}$ are the multiplicities of the elements.

Example II.1. Suppose a data set

$$\mathbb{D} = \{-2.1, 3.2, 8, -2.1, 0, -2.1, 3.2, -4, 2.4, -6, 2.4, 10, 0, 1\}.$$

Then $\mathbb{D} \equiv (-6)^1(-4)^1(-2.1)^3(0)^2(1)^1(2.4)^2(3.2)^2(8)^1(10)^1$.

Let $S(\vec{v})$ denote the unordered set of $\vec{v}$, i.e.,

$$S(\vec{v}) = \{v_1, v_2, ..., v_{m-1}, v_m, v_{m+1}\}.$$

Let $Min_K(i)$ denote the i-th minimal element in a set K.

Example II.2. Suppose a given set

$$K = \{12, 3, 9, 1, 8, 0, 10, 6\}$$

Then $Min_K(1) = 0, Min_K(2) = 1, Min_K(3) = 3, Min_K(4) = 6, Min_K(5) = 8, Min_K(6) = 9, Min_K(7) = 10, Min_K(8) = 12$.

Let $H = S(\vec{v}) \cup S(\vec{w})$. Let $|H|$ denote the size of set H.

B. Step and Step Probability Matrix

Definition II.1. (vector intersection) Define a vector

$$\vec{v} \wedge \vec{w} = (min_H(1), min_H(2), \cdots), min_H(|H|-1), min_H(|H|)).$$

Example II.3. Suppose there are two ascending vectors

$$\vec{v} = (-4.1, -3.3, -1.7, 0.9, 1.6, 2.8),$$

$$\vec{w} = (-5.1, -3.3, -3.1, -2.1, 1.9, 2.1, 2.8, 4.2).$$

Then one could compute the intersection via $\vec{v} \wedge \vec{w} = (-5.1, -4.1, -3.3, -3.1, -2.1, -1.7, 0.9, 1.6, 1.9, 2.1, 2.8, 4.2)$. Moreover,

$$\Delta\vec{v} = (0.8, 1.6, 2.6, 0.7, 1.2),$$

$$\Delta\vec{w} = (1.0, 0.8, 0.2, 1.0, 0.4, 2.6, 0.7, 0.3, 0.2, 0.7, 1.4).$$

Next, we associate each finite matrix M with a step function and vice versa as follows:

Definition II.2. (Step Probability Matrix) M is a Step Probability Matrix (or SPM) with size m if and only if

$$M = \begin{bmatrix} \vec{v} \\ \vec{r} \end{bmatrix} = \begin{bmatrix} v_1 & v_2 & \cdots & v_{m-1} & v_m & v_{m+1} \\ r_1 & r_2 & \cdots & r_{m-1} & r_m & 0 \end{bmatrix},$$

where $\vec{v}$ is an ascending vector; where each $r_1, r_2, \cdots r_m \geq 0$; and where

$$\Delta\vec{v} \bullet (r_1, r_2, \cdots r_{m-1}, r_m) = 1.$$

Each step probability matrix M is identified with a probability density function $f_M : \mathbb{R} \to [0, 1]$ as follows:

$$f_M(x) = \begin{cases} 0 & \text{if } x \in (-\infty, v_1); \\ r_1 & \text{if } x \in [v_1, v_2); \\ r_2 & \text{if } x \in [v_2, v_3); \\ \vdots & \vdots \\ r_{m-1} & \text{if } x \in [v_{m-1}, v_m]; \\ r_m & \text{if } x \in [v_m, v_{m+1}); \\ 0 & \text{if } x \in [v_{m+1}, +\infty). \end{cases}$$

We use $D_f = \vec{v}$ to represent the endpoints of the intervals in the domain of f_M (or f, if M is understood from the context). Moreover, throughout this article, we use a step matrix M and its probability density function f_M interchangeably.

Definition II.3. (Algorithm) Given a finite set of real-valued data $\mathbb{D} \equiv (\alpha_1)^{i_1}(\alpha_2)^{i_2} \cdot (\alpha_p)^{i_p}$, one could convert it into a step probability matrix $f \in SPM_{p+1}$ via

$$f = \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_p & \alpha_{p+1} \\ \frac{i_1}{\Delta\vec{\alpha}\bullet\vec{i}} & \frac{i_2}{\Delta\vec{\alpha}\bullet\vec{i}} & \vdots & \frac{i_p}{\Delta\vec{\alpha}\bullet\vec{i}} & 0 \end{bmatrix},$$

where $\vec{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_p, \alpha_{p+1})$ and $\vec{i} = (i_1, i_2, \cdots, i_p)$ and where $\alpha_{p+1} = \alpha_p + \phi(\alpha_1, \alpha_2, \cdots, \alpha_p, i_1, i_2, \cdots, i_p)$ for some positive real function ϕ. An explicit ϕ could refer to Definition IV.1.

Definition II.4. Let SPM_{m+1} denote the set of all the step probability matrices with size $m+1$, i.e., $SPM_{m+1} = \{\begin{bmatrix} \vec{\alpha} \\ \vec{r} \end{bmatrix} : |\vec{\alpha}| = |\vec{r}| = m+1; \alpha_1 < \alpha_2 < \cdots < \alpha_m < \alpha_{m+1}; r_1, r_2, \cdots r_m \geq 0, r_{m+1} = 0; \Delta\vec{\alpha} \bullet (r_1, r_2, \cdots r_{m-1}, r_m) = 1\}$.

Let SPM denote the set of all the finite step probability matrices, i.e., $SPM = \bigcup_{m=1}^{\infty} SPM_m$. This set basically represents all the candidates of approximated probability density function via step functions.

C. Some Proofs

Claim 1. $f = \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_p & \alpha_{p+1} \\ \frac{i_1}{\Delta\vec{\alpha}\bullet\vec{i}} & \frac{i_2}{\Delta\vec{\alpha}\bullet\vec{i}} & \vdots & \frac{i_p}{\Delta\vec{\alpha}\bullet\vec{i}} & 0 \end{bmatrix} \in SPM_{p+1}$.

Proof: It suffices to show

$$\Delta\vec{\alpha} \bullet (\frac{i_1}{\Delta\vec{\alpha} \bullet \vec{i}}, \frac{i_2}{\Delta\vec{\alpha} \bullet \vec{i}}, \cdots \frac{i_p}{\Delta\vec{\alpha} \bullet \vec{i}}) = 1.$$

This follows immediately from the definition. ■

Definition II.5. (Step Matrix) M is a Step Matrix (or SM) with size m if and only if

$$M = \begin{bmatrix} \vec{v} \\ \vec{r} \end{bmatrix} = \begin{bmatrix} v_1 & v_2 & \cdots & v_{m-1} & v_m & v_{m+1} \\ r_1 & r_2 & \cdots & r_{m-1} & r_m & 0 \end{bmatrix},$$

where $\vec{v}$ is an ascending vector and each $r_1, r_2, \cdots r_m \geq 0$.

Similarly, let SM_{m+1} denote the set of all the step matrices with size $m+1$ and SM denote the set of all the finite step matrices. Obviously, $SPM_{m+1} \subseteq SM_{m+1}$ and $SPM \subseteq SM$. The area of a given step function f is defined in the following:

Definition II.6. Define $Area : SM \to \mathbb{R}^+$ by

$$Area(\begin{bmatrix} v_1 & v_2 & \cdots & v_{m-1} & v_m & v_{m+1} \\ r_1 & r_2 & \cdots & r_{m-1} & r_m & 0 \end{bmatrix})$$
$$= (\Delta v_1, \Delta v_2. \cdots, \Delta v_{m-1}, \Delta v_m) \bullet (r_1, r_2, \cdots, r_{m-1}, r_m).$$

Example II.4. Supppose a step function

$$f = \begin{bmatrix} -3.2 & -2.1 & 0.7 & 1.5 & 3.9 & 5.0 \\ 0.02 & 0.10 & 0 & 0.02 & 0.08 & 0 \end{bmatrix}.$$

Then by the above definition, $Area(f) =$

$$(1.1, 2.8, 0.8, 2.4, 1.1) \bullet (0.02, 0.10, 0, 0.02, 0.08) = 0.438.$$

Hence this f is not a probability density function. Of course, it could be converted into a probability function by normalization as long as $Area(f) < \infty$.

Let $f, g \in SM$ be arbitrary. We define two operators $max, min : SM \times SM \to SM$ as follows:

Definition II.7. Define $max(f,g) : \mathbb{R} \to \mathbb{R}^+$ by

$$max(f,g)(x) := max\{f(x), g(x)\};$$

similarly, define $min(f,g) : \mathbb{R} \to \mathbb{R}^+$ by

$$min(f,g)(x) := min\{f(x), g(x)\}.$$

Observe that neither max nor min operation is closed over SPM. Nonetheless both are closed over SM, which is sufficient to our purpose.

Claim 2. *max and min are closed over SM.*

Proof: Let $f, g \in SM$ be arbitrary. Then both D_f and D_g are finite, i.e., the ranges of both functions are also finite. Henceforth, the ranges of both $max(f,g)$ and $min(f,g)$ are finite, i.e., $max(f,g), min(f,g) \in SM$. ■

Definition II.8. (vector-based floor function) For each ascending vector $\vec{v}$, we define a floor function $\lfloor \rfloor_{\vec{v}} : [v_1, v_m] \to S(\vec{v})$ by $\lfloor x \rfloor_{\vec{v}} = v_j$ iff only if $v_j \leq x < v_{j+1}$.

For example, if $\vec{v} = (-3.5, -2.1, 0.8, 1.9, 2.9, 5.8)$, then $\lfloor -0.9 \rfloor_{\vec{v}} = -2.1$ and $\lfloor 5.5 \rfloor_{\vec{v}} = 2.9$.

Lemma II.1. *(max, min functions) Suppose $f, g \in SM$. Suppose*

$$D_f = (v_1, v_2, \cdots v_m, v_{m+1}),$$
$$D_g = (w_1, w_2, \cdots w_n, w_{n+1}),$$
$$D_f \wedge D_g = (t_1, t_2, \cdots, t_h).$$

Then one has $max(f,g) =$

$$\begin{bmatrix} t_1 & t_2 & \cdots & t_{h-1} & t_h \\ max(f,g)(t_1) & max(f,g)(t_2) & \cdots & max(f,g)(t_{h-1}) & 0 \end{bmatrix}]$$

and $min(f,g) =$

$$\begin{bmatrix} t_1 & t_2 & \cdots & t_{h-1} & t_h \\ min(f,g)(t_1) & min(f,g)(t_2) & \cdots & min(f,g)(t_{h-1}) & 0 \end{bmatrix}.$$

Proof: By Claim 2, we have $max(f,g), min(f,g) \in SM$. Since each t_i reflects the change of difference between f and g, one only has to calculate the values at each t_i as the result claims. ■

For the convenience of computation, we further derive an explicit formula as follows:

Corollary 1. $max(f,g) =$

$$\begin{bmatrix} t_1 & \cdots & t_{h-1} & t_h \\ \{f(\lfloor t_1 \rfloor), g(\lfloor t_1 \rfloor)\}^* & \cdots & \{f(\lfloor t_{h-1} \rfloor), g(\lfloor t_{h-1} \rfloor)\}^* & 0 \end{bmatrix},$$

and $min(f,g) =$

$$\begin{bmatrix} t_1 & \cdots & t_{h-1} & t_h \\ \{f(\lfloor t_1 \rfloor), g(\lfloor t_1 \rfloor)\}_* & \cdots & \{f(\lfloor t_{h-1} \rfloor), g(\lfloor t_{h-1} \rfloor)\}_* & 0 \end{bmatrix},$$

where the superscribed star denotes the maximum function and the subscribed star denotes the minimum function.

Proof: By Lemma II.1, we could derive that

$$max(f,g)(t_i) = max\{f(t_i), g(t_i)\} = max\{f(\lfloor t_i \rfloor), g(\lfloor t_i \rfloor)\}$$

and similarly, we have

$$min(f,g)(t_i) = min\{f(t_i), g(t_i)\} = min\{f(\lfloor t_i \rfloor), g(\lfloor t_i \rfloor)\}.$$

Hence the result follows immediately. ■

Example II.5. Suppose a given real multi-set

$$\mathbb{D}_1 \equiv (-2.4)^1(-1.8)^3(0.5)^2(1.3)^1(2.6)^2(4.5)^1(6.4)^1,$$
$$\mathbb{D}_2 \equiv (-3.2)^2(-2.1)^2(-1.7)^3(0.5)^1(2.1)^4(3.2)^1(4.3)^3(5.4)^2$$
$$(6.2)^2(7.1)^1(7.8)^2.$$

Then $\vec{i} = (1, 3, 2, 1, 2, 1, 1, 0)$. For simplicity, let us take

$$\phi(\alpha_1, \alpha_2, \cdots, \alpha_7, i_1, i_2, \cdots, i_7) = \alpha_7 - \alpha_6 = 1.9.$$

One has $\vec{\alpha} = (-2.4, -1.8, 0.5, 1.3, 2.6, 4.5, 6.4, 8.3)$. By the algorithm mentioned in Remark II.3, we could associate $\mathbb{D}_1$ with a step probability matrix f_1 as follows:

$$f_1 = \begin{bmatrix} -2.4 & -1.8 & 0.5 & 1.3 & 2.6 \\ 0.0556 & 0.1667 & 0.1111 & 0.0556 & 0.1111 \end{bmatrix}$$
$$\begin{bmatrix} 4.5 & 6.4 & 8.3 \\ 0.0556 & 0.0556 & 0 \end{bmatrix};$$

and similarly, we could associate $\mathbb{D}_2$ with a step probability matrix f_2 as follows:

$$f_2 = \begin{bmatrix} -3.2 & -2.1 & -1.7 & 0.5 & 2.1 & 3.2 & 4.3 \\ 0.078 & 0.078 & 0.118 & 0.039 & 0.157 & 0.039 & 0.118 \end{bmatrix}$$

$$\begin{bmatrix} 5.4 & 6.2 & 7.1 & 7.8 & 8.5 \\ 0.078 & 0.078 & 0.039 & 0.078 & 0 \end{bmatrix}.$$

$$D_{f_1} = (-2.4, -1.8, 0.5, 1.3, 2.6, 4.5, 6.4, 8.3)$$

$$D_{f_2} = (-3.2, -2.1, -1.7, 0.5, 2.1, 3.2, 4.3, 5.4, 6.2, 7.1, 7.8, 8.5).$$

$D_{f_1} \wedge D_{f_2} =$
$(-3.2, -2.4, -2.1, -1.8, -1.7, 0.5, 1.3, 2.1, 2.6, 3.2, 4.3, 4.5, 5.4, 6.2, 6.4, 7.1, 7.8, 8.3, 8.5).$

$$max(f_1, f_2) = \begin{bmatrix} -3.2 & -2.4 & -2.1 & -1.8 & -1.7 \\ 0.078 & 0.078 & 0.078 & 0.1667 & 0.1667 \end{bmatrix}$$

$$\begin{bmatrix} 0.5 & 1.3 & 2.1 & 2.6 & 3.2 & 4.3 & 4.5 \\ 0.1111 & 0.0556 & 0.157 & 0.157 & 0.1111 & 0.118 & 0.118 \end{bmatrix}$$

$$\begin{bmatrix} 5.4 & 6.2 & 6.4 & 7.1 & 7.8 & 8.3 & 8.5 \\ 0.078 & 0.078 & 0.078 & 0.0556 & 0.078 & 0.078 & 0 \end{bmatrix}$$

and

$$min(f_1, f_2) = \begin{bmatrix} -3.2 & -2.4 & -2.1 & -1.8 & -1.7 \\ 0 & 0.0556 & 0.0556 & 0.078 & 0.118 \end{bmatrix}$$

$$\begin{bmatrix} 0.5 & 1.3 & 2.1 & 2.6 & 3.2 & 4.3 & 4.5 \\ 0.039 & 0.039 & 0.0556 & 0.1111 & 0.039 & 0.1111 & 0.0556 \end{bmatrix}$$

$$\begin{bmatrix} 5.4 & 6.2 & 6.4 & 7.1 & 7.8 & 8.3 & 8.5 \\ 0.0556 & 0.0556 & 0.0556 & 0.039 & 0.0556 & 0 & 0 \end{bmatrix}.$$

III. Sampling a Normal Distribution

In order to compare the similarities between a data-derived SM and a normal distribution, we need to sample the normal distribution. Given a finite set of numerical data $\mathbb{D}$, we would like to find the optimal normal distribution that would fit best for this data. If one looks at Figure 1, there are two step functions f and g, where f is a $\mathbb{D}_1$-induced step matrix and g is a $\mathbb{D}_2$-induced step matrix. Obviously f fits better than g does with respect to standard normal distribution $Norm(x, 0, 1)$. On the other hand, the right-hand-side figure shows that with proper sampling, a step probability density function could be applied to approximate the standard normal distribution.

$$g(x) = \begin{cases} 0 & \text{if } x \in (-\infty, -2.4); \\ 0.2 & \text{if } x \in [-2.4, -1.3); \\ 0.3 & \text{if } x \in [-1.3, -1.1); \\ 0.35 & \text{if } x \in [-1.1, 0.1); \\ 0.2 & \text{if } x \in (0.1, 1.8); \\ 0.35 & \text{if } x \in [1.8, 2.5); \\ 0.25 & \text{if } x \in [2.5, 3); \\ 0 & \text{if } x \in [3, \infty). \end{cases}$$

$$f(x) = \begin{cases} 0 & \text{if } x \in (-\infty, -2); \\ N(-2, 0, 1) = 0.0540 & \text{if } x \in (-2, -1.5); \\ N(-1.5, 0, 1) = 0.1295 & \text{if } x \in [-1.5, -0.8); \\ N(-0.8, 0, 1) = 0.2897 & \text{if } x \in [-0.8, 0.4); \\ N(0.4, 0, 1) = 0.3683 & \text{if } x \in [0.4, 1); \\ N(1, 0, 1) = 0.2420 & \text{if } x \in (1, 1.4); \\ N(1.4, 0, 1) = 0.1497 & \text{if } x \in [1.4, 2); \\ N(2, 0, 1) = 0.0540 & \text{if } x \in [2, 2.5); \\ N(2.5, 0, 1) = 0.0175 & \text{if } x \in [2.5, 2.7); \\ N(2.7, 0, 1) = 0.0175 & \text{if } x \in (2.7, 3); \\ 0 & \text{if } x \in [3, \infty). \end{cases}$$

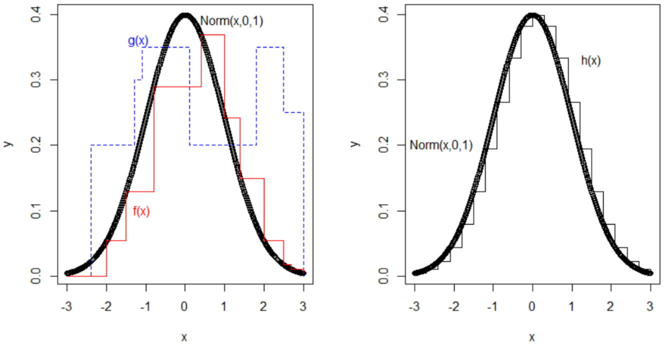

Fig. 1. Approximating Step Functions

Given a normal distribution $Norm(x, \mu, \sigma)$ (or $\rho(x, \mu, \sigma)$), we would like to approximate ρ via a step function f_P^ρ, where $P = (P_1, P_2, ..., P_n)$ is a finite partition of $\mathbb{R}$. Since 99.7 of the data lies in the range of $\pm 3\sigma$, we could set the step function $P_1 = -3\sigma$, and $P_n = 3\sigma$ and let $\Delta P_i \equiv \gamma = \frac{6\sigma}{n}$ for all $1 < i < n$, where n could be any arbitrary natural number. Now we characterize ρ via f_P^ρ as follows: $f_P^\rho = \begin{bmatrix} -3\sigma & -3\sigma + \gamma & \cdots \\ \rho(-3\sigma, \mu, \sigma) & \rho(-3\sigma + \gamma, \mu, \sigma) & \cdots \end{bmatrix}$ $\begin{bmatrix} -3\sigma + (n-1)\gamma & 3\sigma & 3\sigma + \gamma \\ \rho(-3\sigma + (n-1)\gamma, \mu, \sigma) & \rho(3\sigma, \mu, \sigma) & 0 \end{bmatrix}$ One example is $h(x)$ in Figure 1, in which $n = 20, \sigma = 1$ and $\gamma = \frac{6\sigma}{n} = 0.3$.

$$h(x) = \begin{cases} 0 & \text{if } x \in (-\infty, -3); \\ N(-3, 0, 1) = 0.0044 & \text{if } x \in (-3, -2.7); \\ N(-2.7, 0, 1) = 0.0104 & \text{if } x \in [-2.7, -2.4); \\ N(-2.4, 0, 1) = 0.0224 & \text{if } x \in [-2.4, -2.1); \\ & . \\ & . \\ & . \\ N(2.4, 0, 1) = 0.0224 & \text{if } x \in [2.4, 2.7); \\ N(2.7, 0, 1) = 0.0104 & \text{if } x \in [2.7, 3); \\ 0 & \text{if } x \in [3, \infty); \end{cases}$$

where $\sigma = 1, n = 20$ and $\gamma = \frac{6\sigma}{20}$.

One might argue that strictly speaking $f_P^\rho \notin SPM$. This is true, since the induced step function could only capture 99.7

percent, and not 100 percent as $n \to \infty$ (or $|P| \to \infty$). These step functions indeed could be regarded approximately belong to SPM.

IV. Probability Density Function

In Remark II.3, we have given an algorithm to compute a data-induced step probability function. These functions would be used to match the similarity between them and the sampled step function in Section III. Suppose $\mathbb{D} \equiv MS(\mathbb{D}) = (-6)^1(-4)^1(-2)^2(-1)^1(0)^3(3)^2(5)^1(9)^1(12)^1$ (as shown in Figure **??**). By the distinct elements in $\mathbb{D}$, we form a basic vector $B(\mathbb{D}) = (-6, -4, -2, -1, 0, 3, 5, 9, 12)$ in this case. To our aim, we need to specify and justify the chosen form of ϕ. ϕ basically controls the duration of a probability. Since it is the last point in the data and should not affect too much the whole distribution of data. In order to diminish its affect on other points, we decide such duration based on the concept of average duration as follows:

Definition IV.1. (Algorithm) Suppose a finite set of data $\mathbb{D} \equiv b_1^{i_1} b_2^{i_2} \cdots b_n^{i_n}$. Then it is converted into a step function via

$$f_{\mathbb{D}} = \begin{bmatrix} b_1 & b_2 & \cdots & b_{n-1} \\ \dfrac{i_1}{\sum\limits_{k=1}^{n} i_k \cdot \Delta b_k} & \dfrac{i_2}{\sum\limits_{k=1}^{n} i_k \cdot \Delta b_k} & \cdots & \dfrac{i_{n-1}}{\sum\limits_{k=1}^{n} i_k \cdot \Delta b_k} \end{bmatrix}$$

$$\begin{bmatrix} b_n & b_{n+1} \\ \dfrac{i_n}{\sum\limits_{k=1}^{n} i_k \cdot \Delta b_k} & 0 \end{bmatrix}, \text{ where} b_{n+1} = b_n + \frac{b_n - b_1}{\sum\limits_{j=1}^{n-1} i_j} \cdot i_n$$

.

Remark 1. The duration between b_n and b_{n+1} is decided by $\frac{b_n - b_1}{\sum_{j=1}^{n} i_j} \cdot i_n$, where $\frac{b_n - b_1}{\sum_{j=1}^{n} i_j}$ is the average probability duration. This mechanism is aiming at reducing the impact of the last data point on other points.

Example IV.1. Suppose $\mathbb{D}$ is given as above. Since $b_{n+1} = b_n + \frac{12-(-6)}{1+1+2+1+3+2+1+1} \cdot 1 = 12 + \frac{18}{12} \cdot 1 = 13.5$, one has $\sum\limits_{k=1}^{9} i_k \cdot \Delta B(\mathbb{D})_k = (1,1,2,1,3,2,1,1,1) \bullet (2,2,1,1,3,2,4,3,1.5) = 28.5$. Hence $f_{\mathbb{D}} =$

$$\begin{bmatrix} -6 & -4 & -2 & -1 & 0 & 3 \\ 0.0351 & 0.0351 & 0.0702 & 0.0351 & 0.1053 & 0.0702 \end{bmatrix}$$
$$\begin{bmatrix} 5 & 9 & 12 & 13.5 \\ 0.0351 & 0.0351 & 0.0351 & 0.000 \end{bmatrix}.$$

This step function could be shown in Figure 2.

Example IV.2. Suppose a radar receives signals from 6:00 AM to 7:00 AM at different length of time intervals as shown in Table IV.2, where time stands for the observation time, Δb_n denotes the length of the observation time, ac. denotes the accumulated length of observation time, n. denotes the number of signals received.

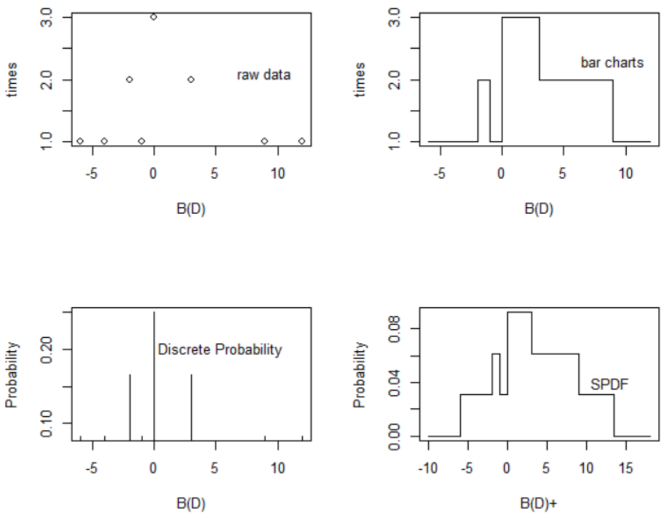

Fig. 2. $MS(\mathbb{D})$ and its Step Probability Density Function (SPDF) $f_{\mathbb{D}}$

time	Δb_n	ac.	n.	i_n	$\Delta b_n \cdot i_n$
6:00-6:04	4	4	3	0	0
6:05-6:11	7	11	2	3	21
6:12-6:13	2	13	1	2	4
6:14-6:21	8	21	1	1	8
6:22-6:27	6	27	1	1	6
6:28-6:38	11	38	6	1	11
6:39-6:48	10	48	9	6	60
6:49-6:53	5	53	3	9	45
6:54-6:58	5	58	4	3	15
6:59-7:00	2	60	2	4	8
hypothetical	3.73	63.73	0	2	7.46
sum					185.46

Then this set of data could be represented by $\mathbb{D}$ by $\mathbb{D} \equiv (0)^0(4)^3(11)^2(13)^1(21)^1(27)^1(38)^6(48)^9(53)^3(58)^4(60)^2$. Then $b_{11} = 60 + \frac{56}{30} \cdot 2 = 63.73$, i.e., $\Delta b_{10} = 3.73$. This data could be associated with an probability density function $f_{\mathbb{D}}$ by $f_{\mathbb{D}} = \begin{bmatrix} 0 & 4 & 11 & 13 & 21 & 27 & 38 & 48 & 53 & 58 & 60 & 63.73 \\ 0 & \frac{3}{k} & \frac{2}{k} & \frac{1}{k} & \frac{1}{k} & \frac{1}{k} & \frac{6}{k} & \frac{9}{k} & \frac{3}{k} & \frac{4}{k} & \frac{2}{k} & 0 \end{bmatrix}$, where $k = 185.46$. The SPDF of $f_{\mathbb{D}}$ (or f) can be visualized by Figure 3.

V. Solution Set: Means and Variances

Now we need to specify the solution set for finding the optimal normal distribution and the given data set $\mathbb{D}$. Let $\vec{v} = B(\mathbb{D})$. Let $m = |\vec{v}|$. Let $mean(\vec{v})$ denote the mean of $S(\vec{v})$. Let μ denote the candidates for the mean of normal distribution. We assume the mean is located between v_1 and v_m and the variance is located between $S_{\vec{v}}^2 = \dfrac{\sum\limits_{i=1}^{m}(v_i - mean(\vec{v}))^2}{m-1}$ and $\dfrac{\sum\limits_{i=1}^{m}(v_i - v_1)^2}{m-1}$.

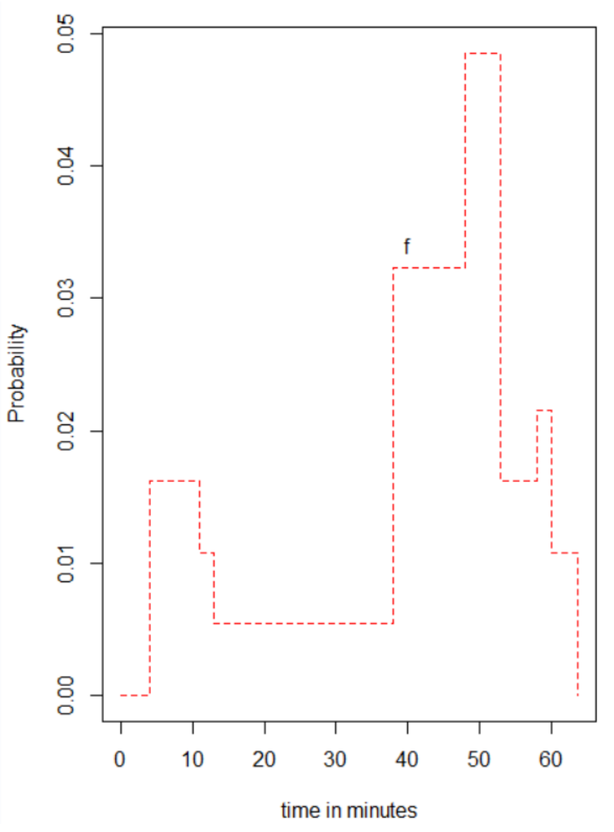

Fig. 3. SPDF for Radar Signals in 60 minutes

Lemma V.1. $\frac{\sum_{i=1}^{m}(v_i - mean(\vec{v}))^2}{m-1} \leq \frac{\sum_{i=1}^{m}(v_i - v_1)^2}{m-1}$.

Proof: Since $\sum_{i=1}^{m}(v_i - mean(\vec{v}))^2 - \sum_{i=1}^{m}(v_i - \mu)^2 = -m \cdot (v_1 - \mu)^2 \leq 0$, the result follows. ■

Based on this, we could delimit the range for the solution set.

Definition V.1. The solution set induced by $\mathbb{D}$ is defined by $SS_{\mathbb{D}} = \{Norm(x, \mu, \sigma) : min(\mathbb{D}) \leq \mu \leq max(\mathbb{D}), \frac{\sum_{i=1}^{m}(v_i - mean(\vec{v}))^2}{m-1} \leq \sigma \leq \frac{\sum_{i=1}^{m}(v_i - v_1)^2}{m-1}\}$

Due to the unique properties of normal distribution, there is no analytical approach to find the optimal solution. We use the numerical computation to approximate the optimal solution.

Example V.1. Suppose $\mathbb{D} \equiv MS(\mathbb{D}) = (-6)^1(-4)^1(-2)^2(-1)^1(0)^3(3)^2(5)^1(9)^1(12)^1$. Suppose $\vec{v} = B(\mathbb{D})$. Then $m = |\vec{v}| = 9$ and $mean(\vec{v}) = 1.7778$ $S_{\vec{v}}^2 = 35.9444$, and $\frac{\sum_{i=1}^{m}(v_i - v_1)^2}{8} = 104$. Hence the solution set $SS_{\mathbb{D}} = \{Norm(x, \mu, \sigma) : -6 \leq \mu \leq 12, \text{ and } 35.9444 \leq \sigma \leq 104\}$.

Latter on, $SS_{\mathbb{D}}$ would be simulated via Monte Carlo method.

VI. Algorithms and Measurements

A. Algorithms

Before we proceed further, let us summarize the algorithms for finding the optimal normality for a finite set of numerical data $\mathbb{D}$.

1) Convert $\mathbb{D}$ into an ascending multi-set form (or $MS(\mathbb{D})$) (ref. Section II-A);
2) Construct $\mathbb{D}$-induced step probability density function $f_{\mathbb{D}}$ (ref. Section IV);
3) Construct $\mathbb{D}$-induced solution set $SS_{\mathbb{D}}$ (ref. Section V);
4) (Looping over $SS_{\mathbb{D}}$) Apply Monte Carlo Method to generate μ and σ in order to yield $Norm(x, \mu, \sigma)$ for sufficiently large times (ref. Section V);
5) Generate the approximated step function $f_{\mathbb{D}}^{\rho}$ based on the chosen normal distribution $\rho \equiv Norm(x, \mu, \sigma)$ (ref. Section III);
6) Compute $max(f_{\mathbb{D}}, f_{\mathbb{D}}^{\rho})$ and $min(f_{\mathbb{D}}, f_{\mathbb{D}}^{\rho})$ (ref. Section II);
7) Apply the optimal criteria to decide the fittest normal distribution for $\mathbb{D}$ (ref. Section VI-B,VI-C).

The computational complexity mainly depends on two parts: sorting the given n data value in ascending order and counting of the multiplicities; and applying the Monte Carlo method. The complexity of the first part would be $O(n^2)$, while the second part depends on the users' predetermined sampling times which could also be decided if one adopts some criteria to build up the threshold.

B. First Measurement

Definition VI.1. Define $Fit1st : SM \times SM \rightarrow [0, 1]$ by

$$Fit1st(f, g) := \frac{Area(min(f, g))}{Area(max(f, g))},$$

in particular $Fit1st(f_{\mathbb{D}}, f_{\mathbb{D}}^{\rho(\mu,\sigma)}) = \frac{Area(min(f_{\mathbb{D}}, f_{\mathbb{D}}^{\rho(\mu,\sigma)}))}{Area(f_{\mathbb{D}}, f_{\mathbb{D}}^{\rho(\mu,\sigma)})}$.

If the value $Fit1st(f_{\mathbb{D}}, f_{\mathbb{D}}^{\rho(\mu,\sigma)})$ is close to 1, then f and g have higher overlapped area and thus higher similarity.

Definition VI.2. (optimal solution) We call $\rho^* \equiv Norm(x, \mu^*, \sigma^*)$ is a $\mathbb{D}$-fittest solution (or the first type normality of $\mathbb{D}$) if and only if $Fit1st(f_{\mathbb{D}}, \rho^*) \leq Fit1st(f_{\mathbb{D}}, g)$ for all $g \in SS_{\mathbb{D}}$.

C. Second Measurement

Definition VI.3. Define $Fit2nd : SM \times SM \rightarrow [0, \frac{\pi}{2}]$ by

$$Fit2nd(f, g) := arccos(\frac{\vec{A}_f \bullet \vec{A}_g}{||\vec{A}_f|| \cdot ||\vec{A}_g||}),$$

in particular $Fit2nd(f_{\mathbb{D}}, f_{\mathbb{D}}^{\rho(\mu,\sigma)}) = arccos(\frac{\vec{A}_f \bullet \vec{A}_g}{||\vec{A}_f|| \cdot ||\vec{A}_g||})$

If the value $Fit2nd(f, g)$ is close to 0, then f and g have higher similarity.

Definition VI.4. (optimal solution) We call $\rho^* \equiv Norm(x, \mu^*, \sigma^*)$ is a $\mathbb{D}$-fittest solution ((or the second type normality of $\mathbb{D}$)) if and only if $Fit2nd(f_{\mathbb{D}}, \rho^*) \leq Fit2nd(f_{\mathbb{D}}, g)$ for all $g \in SS_{\mathbb{D}}$.

This measurement is the usual inner product version in the form of two vectors: maximal function and minimal function. The relation of these two functions is then derived (J., 1981; A., 2016).

VII. Experimental Results

Now we demonstrate how to find the fittest normal distribution. We apply R programming (R x64 3.6.1) to run this experiment. Let us continue the data in Example V.1. Suppose $\mathbb{D} \equiv MS(\mathbb{D}) = (-6)^1(-4)^1(-2)^2(-1)^1(0)^3(3)^2(5)^1(9)^1(12)^1$. After searching the set SS by running 10000 times, we obtain an approximately optimal fitting for D with optimal $\mu = 3.8056$ and optimal standard variance $\sigma = 3.4426$. The optimal measurement for $Fit1st$ is 0.5610218. The fitting between these two: $D-$induced step function and $Norm(x, 3.8056, 3.4426)$ is shown in Figure 4.

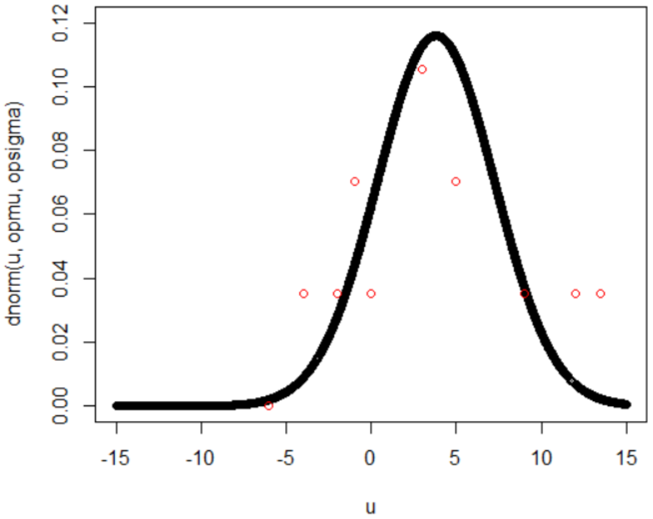

Fig. 4. $\mathbb{D}-$induced step function and its optimal fitting

By the same approach, we could also run an experiment for the second measurement. Here we leave it for the interested readers. Those parameters, including the size of SS, could be trained by machine learning or algorithms (Christpher, 2006; Marc, 2019)

VIII. Problems and Future Works

In this article, we adopt numerical algorithms to estimate the parameters of normal distribution given a set of discrete data. It seems there are still some room to be improved or explored. For example, whether there exists an analytical solution for such estimation? If one could put them into one analytical form, it might save computation time and space. When adopting this numerical approach, we have to resort to Monte Carlo method. The second problem would be whether there exists a way to decide a threshold of experimental times in which the precision of the approximation is very close to optimal. In the future work, we will extend our research to multivariate data type and devise an efficient algorithm that would produce the optimal multivariate normal distribution.

IX. Conclusions

Data fitting and its trend has been an important issue in theoretical and applied data analysis or mathematics. It has wide applications in various fields. In this article, we put forward two approaches for measuring the normality between a given data set and normal distributions. We could then find the optimal normal distribution that could fit the data most. The idea lies in the approximation for an integral of a function via step functions. A discrete data set is treated as part of the approximation process. One uses these discrete points to generate its related step function and then uses this data set to set the solution set for the normal distributions that would be the target that this given data is approximating to. Then one, based on Monte Carlo approach, generates the potential values of the solution set. Each element in solution set is a pair of mean and standard variance. We could then, based on this mean and standard variance, form its step functions, which is also approximately a probability density function. Then we measure the similarities between the data set induced step function and all the step function induced in the solution set. Then one finds the best mean and standard variance that could fit the data set most. Our approach gives an analytical solution for fitting a finite set of numerical data and could be coupled with statistical measurements to solve some decision problems. In the future research, we would combine the approaches used in regression with our methods. In addition, we will devise some statistics to test our methods for data fitting. Lastly, we could expand this research and try to derive the analytical solutions for this fitting method - that might reduce our computational time.

X. Data Availability Statement

All data generated or analysed during this study are included in this published article.

XI. acknowledgment

This article is funded by the Humanities and Social Science Research Planning Fund Project under the Ministry of Education of China (Grant No. 20XJAGAT001).

References

[1] A. Nagoor Gani, K. Kannan, A. R. Manikandan, Inner Product over Fuzzy Matrices, Journal of Mathematics, 2016.

[2] Christpher M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[3] Halsey Royden and Patrick Fitzpatrick, Real Analysis. 4th ed., Pearson , 2017.

[4] J. P. Antoine, K. Gustafson, Partial inner product spaces and semi-inner product spaces, Advances in Mathematics, Volume 41, Issue 3, 1981.

[5] John Aldrich, R. A. Fisher and the Making of Maximum Likelihood 1912– 1922, Statistical Science, 1997, Vol.12, No.3, 162-176.

[6] Jean D. Gibbons, Subhabrata Chakraborti, Nonparametric Statistical Inference, 4th ed., Marcel Dekker Inc., New York, 2003.

[7] Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong. , Mathematics for Machine Learning, Vol. 27, No. 3, Cambridge University Press, 2019.

[8] M. J. D. Powell, Approximation Theory and Methods 1st ed., Cambridge University Press, 1981.

[9] P. Sprent and N. C. Smeeton, Applied Nonparametric Statistical Methods, 3rd ed.,Chapman and Hall, 2001.

[10] Raymond J. Hickey, A Note on the Measurement of Randomness, Journal of Applied Probability, Vol. 19, No. 1, 1982.

[11] Blizard, Wayne D., The Development of Multiset Theory, Modern Logic Vol. 1, No., 1991.

[12] Singh D., Ibrahim, A. M., Yohanna T., Singh J. N., An overview of the applications of multisets, Novi Sad Journal of Mathematics. Vol. 37 , No. 2, 2007.

[13] Russell B. Millar, Maximum Likelihood Estimation and Inference: With Examples in R, SAS and ADMB, Hoboken: Wiley, 2011.

[14] Nocedal, Jorge; Wright, Stephen J., Numerical Optimization. New York: Springer, 2006.

[15] Fletcher, R. . Practical Methods of Optimization. New York: John Wiley & Sons. 1987.

[16] Razali, Nornadiah; Wah, Yap Bee. Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests. Journal of Statistical Modeling and Analytics. 2 (1). 2011.

[17] Vasicek, Oldrich. A Test for Normality Based on Sample Entropy. Journal of the Royal Statistical Society. Series B. 38 (1). 1976.

[18] D.Rodrigues, J.Billeter, D.Bonvin. Maximum-likelihood estimation of kinetic parameters via the extent-based incremental approach. Computers & Chemical Engineering. Volume 122, 2019.