A Multivariate Rank-Based Test via Random Representation and Stochastic Lexical Ordering (Draft)

Yanqi Zhang[1]

[1]yz3871@columbia.edu
[1]yanqizhang2019@gmail.com

1/4/2024

Abstract

A combinatorial algorithm that computes a similarity score between two sets of multidimensional vectors through random string representations. The algorithm is discussed in the context of theory of normal numbers and stochastic ordering.

1 Introduction

The Cantor diagonal method is often employed as tool for showing contradictions. Here I use the diagonal method for constructive purposes and show that families of random diagonal bijections have interesting number-theoretical and probabilistic properties and substructures. Those properties can be exploited to define and compute multidimensional similarity scores or ranks.

Given a Euclidean vector, we can represent it in different bases (for example, base-2, base-3). Moreover, as in the Cantor diagonal function, every finite dimensional vector subspace can also be embedded in the open unit interval. In this paper, I explore the possibilities of measuring the "distance" between two data sets that can be only obtained through alternative representations of Euclidean spaces. The combinatorial algorithm described in the text is an example that manifests this idea. The results reveal intuitive but intriguing interactions between representations of the real number system and probabilistic sampling.

2 Problem Description

Let $\mathrm{X} = \{\mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_n}\}$ and $\mathrm{Y} = \{\mathbf{y_1}, \mathbf{y_2}, \dots, \mathbf{y_m}\}$ be sets of vectors in $\mathbb{R}^D$, and let $\mathrm{W} \triangleq \mathrm{X} \cup \mathrm{Y}$. Assume that X and Y are two independent samples drawn from distributions F and G respectively, i.e. $\mathbf{x_i} \sim F$ and $\mathbf{y_j} \sim G$ for each $\in$ X and each $\mathbf{y_j} \in$ Y. We also assume that X and Y are both pairwise independent. Moreover, assume F and G satisfy the following conditions:

(a) F and G are Riemann-integrable.

(b) F and G are continuously differentiable.

(c) The sample spaces Ω_F and Ω_G are relatively compact and convex in the extended Euclidean space $\mathbb{R}^D_*$.

The question is how to design a statistic $\mathrm{U}(W)$ based on the observations X and Y, to respond to a statement about the distributions F and G (e.g., $H_0 : F = G$). When $X, Y \in \mathbb{R}$, the Wilcoxon rank-sum test does the job with minimal restrictions on the distribution under the null by formulating the alternative as H_a: X is stochastically greater or stochastically less than Y [1]. However, a multivariate test with a rejection criterion entirely analogous to the Wilcoxon test may not be robust, as the usual stochastic order is merely a partial order[2]. When X and Y are random vectors, it becomes increasingly unlikely that the differences between their distributions F and G can be captured by a partial order if $F \neq G$ is true. Nevertheless, it is still possible to devise nonparametric rank-based tests for multivariate distributions. Deb and Sen[3][4] proposed multivariate rank-based nonparametric tests via a notion of rank defined through optimal transport. Chatterjee[5] proposed a test of independence through representation in base-2; the method shares similarities with the algorithm introduced in this paper.

In the following section, I build a measure of association that could be an alternative high-dimensional analog to the one-dimensional rank-sum statistic. It is achieved by formulating an alternative hypothesis that is based on stochastic lexical ordering detached from the usual Euclidean topology, and can be assessed via a finite combinatorial algorithm. I will also detail the intuition behind the algorithm as well as some available theory regarding the algorithm's analytical properties and connections to number theory. To justify my idea, I invoke Borel's theory of normal numbers, and extend the notion of normal space to finite dimensional vector spaces.

3 The Algorithm

Base selection step $\{\beta_0, ..., \beta_P\}$ must be:

(1) Strictly increasing, and $\beta_0 \geq 2$. (1)

(2) $\beta_i^n \neq \beta_j^m \quad \forall n, m \in \mathbb{Z}^+, \text{ if } i \neq j.$ (∗) (2)

Note that in implementation, B can either be deterministic or selected via sampling, and may depend on the data sets. A mixture of bases can diversify the representations of X and Y. The reason we do this is because, the digit distributions of a real number can differ considerable across different bases. Given a set of real vectors, certain patterns are easier to detect when it is represented in some bases than others. Condition (2) guarantees that we would not end up with some $\beta \in$ B that

Algorithm 1 L-D Algorithm

```
Input: $\mathrm{X} = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}\} \subset \mathbb{R}^D$, $\mathrm{Y} = \{\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_m}\} \subset \mathbb{R}^D$, $\{\beta_0, ..., \beta_P\}$, $T$, $(\mathrm{dist}_{\beta_\tau}, R_{\beta_\tau})_{\tau=0}^P$
Sample From Data Sets
$\mathrm{W} = \mathrm{X} \cup \mathrm{Y}$
$\mathrm{X}' = \mathrm{sample}(\mathrm{X},\ n')$
$\mathrm{Y}' = \mathrm{sample}(\mathrm{Y},\ m')$
$\mathrm{X}^{\mathrm{s}} = \mathrm{sample}(\mathrm{W},\ n')$
$\mathrm{Y}^{\mathrm{s}} = \mathrm{sample}(\mathrm{W},\ m')$
for $\tau = 0, \ldots, P$ do
    Base Expansion
    $\beta = \beta_\tau$
    $\mathrm{X}' = \psi_\beta(\mathrm{X}')$
    $\mathrm{Y}' = \psi_\beta(\mathrm{Y}')$
    $\mathrm{X}'_{\mathrm{W}} = \psi_\beta(\mathrm{X}'_{\mathrm{W}})$
    $\mathrm{Y}'_{\mathrm{W}} = \psi_\beta(\mathrm{Y}'_{\mathrm{W}})$
    for $t = 0, \ldots, T$ do
        Choose a Family of Diagonal via Permutations
        $\Pi = \pi_1, \ldots, \pi_r = \mathrm{sample}(S_{2\times D},\ r)$
        $\mathcal{D}^\Pi = (\mathcal{D}_1, \ldots, \mathcal{D}_r) = (\pi_1 \circ \mathcal{D}_0, \ldots, \pi_r \circ \mathcal{D}_0)$
        Compute Lexical Rank Scores
        $U_1 = (\sum_{\mathbf{x}\in\mathrm{X}'} \sum_{\mathbf{y}\in\mathrm{Y}'} I(\mathcal{D}_i(\mathbf{x}[t],\ \mathbf{x}[t+1:]) \succeq \mathcal{D}\rangle(\mathbf{y}[t],\ \mathbf{y}[t+1:]))_{i=1}^r$
        $U_2 = (\sum_{\mathbf{x}\in\mathrm{X}^{\mathrm{s}}} \sum_{\mathbf{y}\in\mathrm{Y}^{\mathrm{s}}} I(\mathcal{D}_i(\mathbf{x}[t],\ \mathbf{x}[t+1:]) \succeq \mathcal{D}_i(\mathbf{y}[t],\ \mathbf{y}[t+1:]))_{i=1}^r$
        Compare Rank Scores
        if $\mathrm{dist}_\beta(U_1, U_2) > R_\tau$ then
            return Reject = TRUE
        else if $t = T$ then
            return Reject = FALSE
        else
            $t = t + 2$
            Repeat
        end if
    end for
    if Reject = FALSE or $\tau = P$ then
        return Reject = FALSE
    else
        $\tau = \tau + 1$
    end if
end for
```

gives us information already given by $\mathrm{B}\backslash\{\beta\}$ (see the next section for discussions on normal numbers).

Diagonalization Step Let $s = (s_1, ..., s_D)$ and $\omega = (\omega_1, ..., \omega_D)$ be tuples of strings. In the family of diagonal functions, the strings $s_1, ..., s_D$ and $\omega_1, ..., \omega_D$ are treated as atomic elements.Let $*$ be the concatenation operator. Define a diagonal function $\mathcal{D}_0$:

$$\mathcal{D}_0(s, \omega) \triangleq s_1 * s_2 * s_3 * ...s_{D-1} * s_D * \omega_1 * \omega_2 * \omega_3 * ...\omega_{D-1} * \omega_D.$$

Note that $\mathcal{D}_0$ is only one member of the diagonal function family possible on s and ω. That is why we can choose more by obtaining rearrangement functions $\pi_1, ..., \pi_r$ sampled from $S_{2\times D}$. Note that the algorithm recursively generates new (unused) string representations of the data, in order to detect anomalies in relative word frequencies. There are many ways to generate injective string representations in this context. For example, suppose we have a string $\mathtt{w}$ and a real vector $\mathbf{w}$, then the string $\mathtt{w}$ is a string representation of the vector $\mathbf{w}$ in base-β if and only if there exists a finite permutation function π such that $\mathtt{w} = \pi \circ \mathcal{C}^{\dim(\mathbf{w})}(\mathbf{w}^{(\beta)})$, where $\mathcal{C}^{\dim(\mathbf{w})}$ denotes the canonical Cantor diagonal function with $\mathbb{R}^{\dim(\mathbf{w})}$ as its domain.

The diagonalization/permutation step is a less greedy way to streamline and add randomness to the representation generating process. I use the binary operator $\mathcal{D}_0$, instead of the canonical Cantor diagonal function, to minimize the total number of possible representations. The block partitioning of data ensures the prioritization of ”more significant” representations over ”less significant” representations.

Lexical Ordering under Permutations

I'll fill this up soon. Standard lexical order is used.

4 Normality of Real Numbers

4.1 A motivating example

Consider a simple but somewhat extreme numerical example. Suppose we are presented with two independent samples stored in matrices S and S′. Each row represents the rounded base-10 expansion of an observation. The observations are pair-wise independent. We are asked to judge whether $\mathbf{S}$ and $\mathbf{S}'$ come from the same underlying distribution in $\mathbb{R}^3$.

$$\mathbf{S} = \begin{bmatrix} 6.118672 & 6.844465 & 11.70361 \\ 6.305066 & 7.637517 & 11.95928 \\ 5.807376 & 7.477909 & 10.19678 \\ 6.068215 & 6.292909 & 12.31989 \\ 5.434508 & 6.769660 & 10.64101 \end{bmatrix}$$

$$\mathbf{S}' = \begin{bmatrix} 5.719117 & 8.333660 & 10.92247 \\ 5.803110 & 6.796560 & 10.50057 \\ 6.954881 & 8.338618 & 11.83061 \\ 6.207886 & 7.116304 & 10.75560 \\ 6.008887 & 7.115457 & 11.61501 \end{bmatrix}$$

Define $d_{sup}(\mathbf{M})$ as the supremum of the Euclidean distances between rows of matrix $\mathbf{M}$, and let $\|\mathbf{M}\|_2$ denote its spectral norm of . It can be verified that $d_{sup}(\mathbf{S}) \approx 2.445$, $d_{sup}(\mathbf{S}') \approx 2.340$, and $d_{sup}(\begin{bmatrix} \mathbf{S} \\ \mathbf{S}' \end{bmatrix}) \approx 2.498$. The spectral norms are $\|\mathbf{S}\|_2 \approx 32.715$ and $\|\mathbf{S}'\|_2 \approx 33.081$. The Euclidean similarity between $\mathbf{S}$ and $\mathbf{S}'$ suggests that they might be samples from the same distribution. Employing a sophisticated spatial distance-based test on larger samples, which mirror the geometric similarities between $\mathbf{S}$ and $\mathbf{S}'$, may still lead to the conclusion that the underlying distributions are identical. However, intuitively, judging only by the visual presentation the matrices in numerical form, $\mathbf{S}$ and $\mathbf{S}'$ do not appear to be homogeneous. In the second column of $\mathbf{S}'$, "7" is always ensued by "11" if "7" is the first letter that occurs before the decimal point. And in the first column of $\mathbf{S}'$, the strings "11" and "88" occurs at a much higher frequency than in the first column of $\mathbf{S}$, indicating the possible presence of local attractors in the vicinity. However, these minor discrepancies in local density can go easily undetected by a distance measure as they can be buried under substructures that are considered more interesting geometrically, especially in high dimensions.

Overall, for given column, the cumulative frequencies of certain combinations of digits change at a different rate in $\mathbf{S}$ than in $\mathbf{S}'$, as we move to the right. But if $\mathbf{S}$ and $\mathbf{S}'$ are indeed from the same distribution, intuitively, we would expect the cumulative frequencies to converge to the same limit if the limit exists. In fact, in the case of the standard one-dimensional uniform distribution, the relative frequency of every possible string does converge (and observation made by Émile Borel when he was working on the representation of real numbers). Let $X \sim U(0,1)$, and let $0.X_1X_2X_3...$ be the infinite decimal expansion of x in base β. Borel conjectured that (1) $\{X_j\}_{j=1}^{\infty}$ is a collection of independent variables; and (2) $X_i \overset{\text{i.i.d.}}{\sim} \text{unif}\{0,...,\beta-1\}$.[6]. An extension of this phenomenon, the notion of number normality classifies real numbers by asymptotic relative frequencies of strings.

The following subsection offers clarifications of notations aimed at facilitating a discussion on number normality and and its generalization to finite vector spaces. This discussion will eventually lead to the formulation of an alternative hypothesis, Cantor's diagonal method, and ultimately my algorithm

4.2 Preliminaries

Let x be a real number, for any $\beta \in \mathbb{Z}^+$ and $\beta \geq 2$, then x can be expressed as $x = \text{sgn}(x) \times \sum_{k=-\infty}^{J} a_k \times \beta^k$, for some $a_k \in \{0,...,\beta-1\}$. If x is not a β-adic rational, then x has a unique representation in base-β. If x is a β-adic rationals, we opt for the standard expansion that ends with infinite 0's to represent x to ensure that the representation is unique for every real number. For example, we identify the integer 1 with $1.\bar{0}$ not $0.\bar{9}$. Let $x \overset{(\beta)}{\equiv} \text{sgn}(x)\, a_{-J}a_{-J+1}\ldots a_{-1}a_0.a_1a_2a_3\ldots$ be the unique representation of x in base-β. Define an injective function $\psi_\beta(\cdot)$ that maps x to a string.

$$\begin{aligned} \psi_\beta(x) \;&\triangleq\; x^{(\beta)} \;\triangleq\; \mathrm{a}_{-\mathrm{L}}\ \mathrm{a}_{-\mathrm{L}+1}\ \ldots\ \mathrm{a}_{-1}\ \mathrm{a}_0\ \mathrm{a}_1\ \mathrm{a}_2 \ldots \\ \mathrm{a}_\mathrm{k} &\mapsto (\mathrm{sgn}(x), a_k) \text{ for each } k \leq J \end{aligned}$$

Throughout this document, I use $[\ \cdot\ ]$ to subset indexed strings and tuples. For example,

$$x^{(\beta)}[0:1] = \psi_\beta(x)[0:1] = \mathrm{a}_0\ \mathrm{a}_1.$$

Here are some more examples that illustrate this notation system.

$$\begin{aligned} -\frac{2}{3} &\overset{(2)}{\equiv} -0.\overline{10} \\ (-\frac{2}{3})^{(2)}[k] &= \text{”-1” for every positive odd integer } k \\ (-\frac{2}{3})^{(2)}[1:3] &= \text{”-1-0-1”} \end{aligned}$$

Since every finite real vector or matrix also have a unique representation in base-β, the injective function ψ_β can be extended to those cases. Let $\mathbf{v}$ be a vector in $\mathbb{R}^D$ and let $\mathbf{W}$ be a matrix in $\mathbb{R}^{N\times D}$. Define the vector $\mathbf{v}^{(\beta)}$ component-wise as $(\mathbf{v}^{(\beta)})_j \triangleq (\mathbf{v}_j)^{(\beta)}$. Similarly, define the matrix $\mathbf{W}^{(\beta)}$ entry-wise as $(\mathbf{W}^{(\beta)})_{ij} \triangleq (\mathbf{W}_{ij})^{(\beta)}$. ψ_β maps a real matrix or vector to a matrix or vector whose entries are strings.

$$\begin{aligned} \psi_\beta(\mathbf{v}) \;&\triangleq\; \mathbf{v}^{(\beta)} \\ \psi_\beta(\mathbf{W}) \;&\triangleq\; \mathbf{W}^{(\beta)} \end{aligned}$$

The operator $[\ \cdot\ ]$ is still used to subset entries of a string matrix or string vector: $\mathbf{v}_j^{(\beta)}[d] \triangleq (\mathbf{v}_j)^{(\beta)}[d]$ and $\mathbf{W}_{ij}^{(\beta)}[k] \triangleq (\mathbf{W}_{ij})^{(\beta)}[k]$.

4.3 Normal numbers and normal vectors

So far we have discussed the unique representations of real objects and clarified on notations. In this subsection, I will discuss some fundamental results regarding normal numbers, which will lay the groundwork for generalizations and and serves as an inspiration for an alternative hypothesis.

Let $\beta \geq 2$ be a positive integer. Let x be a real number. Suppose x has unique base-β expansion give by

$$x \overset{(\beta)}{\equiv} \mathrm{sgn}(x)\, a_{-J}a_{-J+1}\ldots a_{-1}a_0.a_1a_2a_3\ldots$$

where $a_k \in \{0, ..., \beta-1\}$ for all $t \geq -J$. Then x is said to be *simply normal in base-β* if

$$\lim_{N\to\infty} \frac{\#\{1 \leq t \leq N : a_t = c\}}{N} = \frac{1}{\beta} \qquad \text{for all letters } c \in \{0, \ldots, \beta-1\}.$$

In other words, x is simply normal if and only if every possible letter occurs at a equal frequency in its base-β decimal expansion. Let s be a sequence of integers such that every element in s is from the alphabet $\{0, ..., \beta - 1\}$. Denote the length of s by $\mathrm{len}(s)$. We say x is *normal in base-β* if for any such finite, non-empty sequence s,

$$\lim_{N\to\infty} \frac{\#\left\{1 \leq t \leq N - \mathrm{len}(s) + 1 : s = a_t a_{t+1} ... a_{t+\mathrm{len}(s)-1}\right\}}{N} = \frac{1}{\beta^{\mathrm{len}(s)}}.$$

That is to say, x is simply normal in base-β if and only if the letter arrangement in $x^{(\beta)}$ becomes completely random asymptotically. Moreover, the number x is said to be *simply normal* if it is simply normal in base-β for every $\beta \geq 2$. Similarly, x is said to be *normal* if it is normal in base-β for every $\beta \geq 2$. Suppose the string representation of x base-β is given by the string

$$x^{(\beta)} = \mathrm{a_{-J}\ a_{-J+1}\ \ldots\ a_1\ a_0\ a_1\ a_2 \ldots}.$$

Denote the sub-string of $x^{(\beta)}$ up to and including the letter indexed by t as follows:

$$x^{(\beta)}[\,:t] \triangleq \begin{cases} \mathrm{a_{-J} \ldots a_t} & \text{if } t \geq -J \\ \text{""} & \text{if } t < -J \end{cases}$$

Let $f_t^{(\beta)}(x, s)$ denote the relative frequency of the finite string s in $x^{(\beta)}[\,:t]$. It is immediately clear that x is normal in base-β if and only if

$$\lim_{t\to-\infty} f_t^{(\beta)}(x, s) = \beta^{-\mathrm{len}(s)}$$

for any s entirely composed of letters in $\{\mathrm{sgn}(x)0, \ldots, \mathrm{sgn}(x)(\beta - 1)\}$ and such that $1 \leq \mathrm{len}(s) < \infty$ Furthermore, for any pair (ϵ, k) such that $\epsilon \in (0, 1)$ and $k \in \mathbb{Z}^+$, x is a normal number if and only if there exists a positive integer T such that

$$\left| f_T^{(\beta)}(x, \omega) - \beta^{-k} \right| < \epsilon$$

for any string ω of length k composed entirely of letters in $\{\mathrm{sgn}(x)0, \ldots, \mathrm{sgn}(x)(\beta - 1)\}$.

From the definition, it is obvious that non-normal numbers exist: rational numbers cannot be normal in any base. And there are numbers that are simply normal in one base but not simply normal in another base. For example $\frac{2}{3}$ is simply normal in base-2 but not simply normal in base-10. There are only a few confirmed cases of numbers normal in a certain base. The Champernowne's constant 0.123456789101112..., constructed by concatenating all positive integers in ascending order, was conjectured and proved to be normal in base-10 by Champernowne himself [7]. Copeland and Erdős proved that 0.235711131719232931374143…, which is constructed by concatenating all positive integers in ascending order, is also normal in base-10. That number is now referred to as the Copeland and Erdős constant[8]. Yet there is not a unified way to verify whether a number is normal in a given base. As of today, it is still contested whether $\sqrt{2}$ is normal in base-2.

The set of normal numbers and the set of non-normal numbers are both dense in the reals. In terms of cardinality, the set of non-normal numbers are uncountable. However, the real line is almost entirely populated by normal numbers from a measure theoretical perspective, as conjectured and partially proven by Borel[6]. Formally, the bewildering 1909 Borel's theorem on normal numbers states the following.

Theorem 4.3.1. (Borel). *The set of non-normal numbers has Lebesgue measure 0.*

Let $\mathcal{N}_\beta$ denote the set of all real numbers that normal in base-b. And define $\mathcal{N} \triangleq \cap_{\beta=2}^{\infty} \mathcal{N}_\beta$. One consequence of Borel's theorem, as mentioned in a previous sub-section, is that suppose $X \sim \text{unif}(0, 1)$, then $\mathrm{P}\left\{X \in \mathcal{N}\right\} = 1$. Inspired by these findings, I propose a definition for a normal vector in a similar fashion that will be immediately useful in the next section.

Definition 4.3.2. A real vector $\mathbf{v} = [v_1, \ldots, v_d]^\top$ is said to be a normal vector in base-β if for any $1 \leq i \leq d$, we have $v_i \in \mathcal{N}_\beta$. And $\mathbf{v}$ is said to be a normal vector if for any $1 \leq i \leq d$, we have $v_i \in \mathcal{N}$.

One straightforward result is:

Corollary 4.3.1. *The set of all non-normal d-dimensional real vectors has Lebesgue measure 0 for any $d \in \mathbb{Z}^+$.*

Proof. Let $\mathcal{P}^m$ be the set of all non-normal vectors in $\mathbb{R}^m$. Suppose $\lambda(\mathcal{P}^n) = 0$. Then $\lambda(\mathcal{P}^{n+1}) \leq (n+1) \cdot \lambda(\mathbb{R} \times \mathcal{P}^n) = 0$, because $\mathcal{P}^n$ is measure zero. And since Theorem 3.3.1., $\lambda(\mathcal{P}^1) = 0$, by induction, $\mathcal{P}^d$ is measure zero for any positive integer d. Note that there are number of ways to build a more constructive proof. I think some cases such as the base-2 case can be proved through multidimensional Rademacher functions. I will look into it and fill it in if I have time.

There is a number of results in normal numbers that can be directly useful for this project. A large proportion of the vintage analytical results was produced by by Erdős and friends, and they are highly relevant[8][9] [10]. Many are from the perspectives of discrepancy theory[11] and polynomial theory[12]. Many newer results have focused on computability. This one seems particularly interesting[13]. For a general survey, see [14].

5 Alternative Hypothesis

Having defined the notion of a normal vector, we can add a fourth condition to our list of assumptions.

(d) F must satisfy the condition that, if random variable $X \sim F$,

$$\mathrm{P}\left\{X \in \mathcal{N}^D\right\} = 1.$$

Let $\text{dict}(\beta)$ be the finite dictionary induced by base-β. Let $\omega = (\omega_1, \ldots, \omega_D)$ such that $\omega_1, ...\omega_D \in$

dict(β). Define a random variable $f_t^{(\beta)}(X, \omega)$ as:

$$f_t^{(\beta)}(X, \omega) = \begin{bmatrix} f_t^{(\beta)}(\mathrm{x}_1, \omega_1) \\ f_t^{(\beta)}(\mathrm{x}_2, \omega_2) \\ \vdots \\ f_t^{(\beta)}(\mathrm{x}_\mathrm{D}, \omega_D) \end{bmatrix}.$$

From condition (d), we can build the following alternative hypothesis:

H_{α_1} : There exists a positive integer $\beta \geq 2$ such that, for some $t \in \mathbb{Z}$,
and some $s = (s_1, \ldots, s_D)$ with $s_1, \ldots, s_D \in$ dict(β),
such that len(s_i) $\neq 0$ for some i,
one of the following is true:
(1) $f_t^{(\beta)}(X, s) \overset{\text{st}}{\prec} f_t^{(\beta)}(Y, s)$
(2) $f_t^{(\beta)}(X, s) \overset{\text{st}}{\succ} f_t^{(\beta)}(Y, s)$.

We may reject H_0, if for some words $(s_1, ..., s_d)$, there exists a significant discrepancy between the empirical relative frequencies of these words in two data sets, i.e., reject the null if the difference between $\hat{f}_t^{(\beta)}(X, s)$ and $\hat{f}_t^{(\beta)}(Y, s)$ crosses some threshold.

Let $\mathring{\mathbf{X}}$ and $\mathring{\mathbf{Y}}$ be the respective finite (rounded) versions of the real data matrices $\mathbf{X}^{(\beta)}$ and $\mathbf{Y}^{(\beta)}$. And let $\mathring{\mathbf{W}} \triangleq \begin{bmatrix} \mathring{\mathbf{X}} \\ \mathring{\mathbf{Y}} \end{bmatrix}$. Let J be the left-most index and let $-K$ be the right-most index of the expansion of every entry $\widetilde{w}_{ij}^{(\beta)}$. The data matrices are formatted in such a way that the previously non-existing parts of each entry are filled with "0" strings.

Let $\omega = (\omega_1, ..., \omega_d)$ and $\omega' = (\omega'_1, ..., \omega'_d)$ be tuples of strings. Define $h(\omega, \omega')$ to be the function that outputs a vector:

$$h(\omega, \omega') = \begin{bmatrix} \mathrm{count}(\omega_1, \omega'_1) \cdot \mathrm{len}^{-1}(\omega_1) \\ \mathrm{count}(\omega_2, \omega'_2) \cdot \mathrm{len}^{-1}(\omega_2) \\ \vdots \\ \mathrm{count}(\omega_d \omega'_d) \cdot \mathrm{len}^{-1}(\omega_d) \end{bmatrix}.$$

Define the $\hat{f}_t^{(\beta)}(W_i, s) \triangleq h(\mathring{\mathbf{w}}_i[\,:t], s) \in \mathbb{R}^D$. In other words, the vector $\hat{f}_t^{(\beta)}(W_i, s) \in \mathbb{R}^D$ represents the empirical relative frequency of how often $(s_1, ..., s_D)$ occurs in the observation $(\mathring{w}_{i1}, ..., \mathring{w}_{iD})$, up to and including the the t-th digit. .

Let $\mathcal{B} = \{B_i\}_{i=1}^{n+m}$ be a sequence of D-dimensional hypercube, such that $B_i \subset [0,1]^D$ and $f_t^{(\beta)}(W_i, s) \in B_i$. Let $\mathcal{P}_t(\mathbf{X}) = \{\hat{f}_t^{(\beta)}(X_i, s)\}_{i=1}^{n}$ and $\mathcal{P}_t(\mathbf{Y}) = \{\hat{f}_t^{(\beta)}(Y_i, s)\}_{i=1}^{m}$. Define a discrepancy:

$$\mathcal{D}_{\circ t}^{(\beta)}(\mathbf{X}, \mathbf{Y}, s, \mathcal{B}) \triangleq \sup_{B \in \mathcal{B}} \left|\lambda_D^{-1}(B)\right| \cdot \left| \frac{A(B; \mathcal{P}_t(\mathbf{X}))}{n} - \frac{A(B; \mathcal{P}_t(\mathbf{Y}))}{m} \right|.$$

$A(B; \mathcal{P})$ denotes the number of points in $\mathcal{P}$ that fall into the ball B, and λ_D denotes the D-dimensional Lebesgue measure. Suppose x and y are normal numbers with the same sign, then we already know that for any pair (ϵ, k) such that $\epsilon \in (0,1)$ and $k \in \mathbb{Z}^+$, there exists some T such that

$$\left| f_T^{(\beta)}(x, \omega) - f_T^{(\beta)}(y, \omega) \right| < 2\epsilon \qquad (\star)$$

for any string ω of length k that shares the same sign as x and y. Suppose x and y are actually samples drawn from a common one-dimensional distribution, it is not too unreasonable to assume there is another bound induced by the common distribution. Therefore, fix the j-th column, we devise a decision rule based on the discrepancy between $\{\hat{f}_{-J}^{(\beta)}(\mathbf{X}_{\cdot j}, \omega_j), ..., \hat{f}_{K}^{(\beta)}(\mathbf{X}_{\cdot j}, \omega_j)\}$ and $\{\hat{f}_{-J}^{(\beta)}(\mathbf{Y}_{\cdot j}, \omega_j), ..., \hat{f}_{K}^{(\beta)}(\mathbf{Y}_{\cdot j}, \omega_j)\}$ for some string ω_j. i.e., reject the null hypothesis if the there is a word ω_j, such that the relative frequencies of ω_j in $\mathbf{Y}_{\cdot j}$ changes at a different rate than the relative frequencies of ω_j in $\mathbf{X}_{\cdot j}$. Similarly to the discrepancy between $\mathcal{P}_t(\mathbf{X})$ and $\mathcal{P}_t(\mathbf{Y})$ defined above, we can define a discrepancy between $\mathcal{R}_j(\mathbf{X})$ and $\mathcal{R}_j(\mathbf{Y})$. Let $\mathcal{V} = \{V_i\}_{i=1}^{n+m}$ be a sequence of hypercubes in $\mathbf{R}^{J+K+1}$, such that $V_i \subset [0,1]^{J+K+1}$ and $\mathcal{R}_j(\mathbf{w}_i) = \{\hat{f}_{J}^{(\beta)}(\mathbf{W}_{ij}, \omega_j), ..., \hat{f}_{-K}^{(\beta)}(\mathbf{W}_{ij}, \omega_j)\} \in V_i$. Define a discrepancy between as follows.

$$\mathcal{D}_{j\circ}^{(\beta)}(\mathbf{X}, \mathbf{Y} | \omega, \mathcal{V}) \triangleq \sup_{V \in \mathcal{V}} \left|\lambda_{J+K+1}^{-1}(V)\right| \cdot \left| \frac{A(V; \mathcal{R}_j(\mathbf{X}))}{n} - \frac{A(V; \mathcal{R}_j(\mathbf{Y}))}{m} \right|.$$

Ideally, the test statistics are justified if we can define some connected sets Λ_1 and Λ_2 to serve as the rejection regions under the null, such that if the null is true, then,

$$P\left(\sup_{t \in \{J, \ldots, -K\}} \left(\mathcal{D}_{\circ t}^{(\beta)}(\mathbf{X}, \mathbf{Y} | s, \mathcal{B}) \right) \in \Lambda_1 \right) \leq \alpha$$

or

$$P\left(\sup_{j \in \{1, \ldots, D\}} \left(\mathcal{D}_{j\circ}^{(\beta)}(\mathbf{X}, \mathbf{Y} | s, \mathcal{B}) \right) \in \Lambda_2 \right) \leq \alpha.$$

This would enable us to reject the null hypothesis if an estimated quantity exceeds its associated threshold. The primary task then is to construct thresholds for the discrepancy statistics or other analogous statistics that measure sequence divergence.

6 Approach 1: Can we directly estimate digit distribution from empirical data?

One possible approach is to derive a good inference of $\mathbb{E}\left[X^{(\beta)}\right]$ and $\mathbb{E}\left[Y^{(\beta)}\right]$by first explicitly estimating each $\mathbb{E}\left[X_{\cdot j}^{(\beta)}[k]\right]$ and $\mathbb{E}\left[Y_{\cdot j}^{(\beta)}[k]\right]$ where $1 \leq d \leq D$ and $-J \leq k \leq K$. Suppose there exists a triplet of integers (a, t, j) such that t is non-negative and $1 \leq j \leq D$, and such that we are gain a good approximation of the true mean of every individual $X_{\cdot j}^{(\beta)}[a], X_{\cdot j}^{(\beta)}[a-1], ..., X_{\cdot j}^{(\beta)}[a-t]$ and $Y_{\cdot j}^{(\beta)}[a], Y_{\cdot j}^{(\beta)}[a-1], ..., Y_{\cdot j}^{(\beta)}[a-t]$. Then we can approximate $\mathbb{E}\left[X_{\cdot j}^{(\beta)}[a : a-t]\right]$ and $\mathbb{E}\left[Y_{\cdot j}^{(\beta)}[a : a-t]\right]$ by concatenating the individual estimations, and reject if the two quantities diverge. The same rational can be applied if we want to reject the null of if the sub-strings $\{\hat{\mathbb{E}}\left[X_{\cdot j}^{(\beta)}[b_j]\right]\}_{j=1}^{D} \not\equiv \{\hat{\mathbb{E}}\left[Y_{\cdot j}^{(\beta)}[b_j]\right]\}_{j=1}^{D}$ for some index set $\{b_j\}_{j=1}^{D}$.

A similar rationale can be applied if can reject the null by comparing empirical (atomic) digit distributions, as the estimating joint distributions can be more unrealistic . Let $\hat{g}_{j,k}^{(\beta)}$ denote the empirical distribution of $X_{\cdot j}^{(\beta)}[k]$, and let $\hat{l}_{j,k}^{(\beta)}$ denote the empirical distribution $Y_{\cdot j}^{(\beta)}[k]$. For fixed j, define sequences empirical distributions $\hat{g}_{j,\circ}^{(\beta)} \triangleq \{\hat{g}_{j,k}^{(\beta)}\}_{k=-J}^{K}$ and $\hat{l}_{j,\circ}^{(\beta)} \triangleq \{\hat{l}_{j,k}^{(\beta)}\}_{k=-J}^{K}$. Let $D_{\mathrm{KL}}(P \| Q)$ denote the Kullback–Leibler divergence between discrete probability distributions P and Q. We can measure the discrepancy between the sequences based the Kullback–Leibler divergence of the each individual component:

$$\mathcal{D}_{KL}(\hat{g}_{j,\circ}^{(\beta)}, \hat{l}_{j,\circ}^{(\beta)}) \;\triangleq\; \mathbf{v} \cdot \mathrm{KL}(\hat{g}_{j,\circ}^{(\beta)} \| \hat{l}_{j,\circ}^{(\beta)})$$

where $\mathbf{v}$ is a real vector, and $\mathrm{KL}(\hat{g}_{j,\circ}^{(\beta)} \| \hat{l}_{j,\circ}^{(\beta)})$ is vector defined by $[\mathrm{KL}(\hat{g}_{j,\circ}^{(\beta)} \| \hat{l}_{j,\circ}^{(\beta)})]_k \triangleq D_{\mathrm{KL}}(\hat{g}_{j,k}^{(\beta)} \| \hat{l}_{j,k}^{(\beta)})$. Ideally, we can reject the null if for some j, the divergence crosses a threshold.

However, an algorithm may or may not achieve a good estimation of the atomic digit distributions or the atomic means, depending on the underlying distribution F, the base β to represent the data, and also specifically how to interrelation between each possible pair $(X_{\cdot j}^{(\beta)}[1'], ..., X_{\cdot j'}^{(\beta)}[k'])$ is encoded in F.

Let us assume F satisfies conditions (a) through (d), and data sets $\mathrm{X}_{\mathbb{R}}$ and $\mathrm{Y}_{\mathbb{R}}$ are drawn according to the assumptions defined in the problem description section. For simplicity, let us also assume the case where $\mathrm{X} = \{\mathrm{x}_1, ..., \mathrm{x}_\mathrm{n}\}$ and $\mathrm{Y} = \{\mathrm{y}_1, ..., \mathrm{y}_\mathrm{m}\}$ are the base-2 representations of $\mathrm{X}_{\mathbb{R}}$ and $\mathrm{Y}_{\mathbb{R}}$. And suppose the digits are aligned by index, i.e., each entry starts with the $-J$-th digit and ends with the K-th digit, and the digits that were not in the native unique expansions are filled with "0" strings. Let $q(\mathrm{x}_{\mathrm{od}} \mid \mathrm{X}_{\mathbb{R}})$ denote the conditional distribution of $\mathrm{x}_{\mathrm{od}} = [\mathrm{x}_{\mathrm{od}}[-J], ..., \mathrm{x}_{\mathrm{od}}[K]]^T$ given $\mathrm{X}_{\mathbb{R}}$, if $1 \leq d \leq D$. For convenience, write $q(\mathrm{x}_{\mathrm{od}} \mid \mathrm{X}_{\mathbb{R}})$ as $q(\mathrm{x}_{\mathrm{od}})$, and write the marginal conditional distribution of every atomic $\mathrm{x}_{\mathrm{od}}[t]$ as $q(\mathrm{x}_{\mathrm{od}}[t])$. Note that each $\mathrm{x}_{\mathrm{od}}[t]$ here is a binary variable. If each marginal distribution is known, then the MPM inference will decide $\hat{\mathrm{x}}_{od}[t] = 0$ if and only if $q(\mathrm{x}_{\mathrm{od}}[t]) = 0 \leq q(\mathrm{x}_{\mathrm{od}}[t]) = 1$.

Suppose we have an algorithm $\mathcal{A}$ that computes $q^*(\mathrm{x}_{\mathrm{od}}[t])$ and assigns $\mathrm{x}^*_{\mathrm{o}d}[t]$ by the MPM rule, the question is thus under what circumstances will $\hat{\mathrm{x}}_{\mathrm{od}}[t]$ converge to $\mathbb{E}\left[X^{(2)}_{\mathrm{o}d}[t]\right]$ in probability. Suppose F belongs to a family of distributions S, such that the relations between members of S can be specified by a geometric structure (e.g. if S is the exponential family then, then $F, G \in S$ are points on the manifold S). Define a sub-manifold $M^*_d \triangleq \{p(x) \in S \mid \mathbb{E}_p[x] = \mathrm{x}^*_{\mathrm{o}d}\}$. It is immediate that $\mathrm{x}^*_{\mathrm{o}d}$ is a good approximation if $q(\mathrm{x}_{\mathrm{od}}) \in M^*_d$. From an information-geometric point of view, whether $\mathrm{x}^*_{\mathrm{o}d}$ is a good estimation of the true mean depends on the e-curvature of the sub-manifold M^*_d in S, and the e-curvature is influenced by the graph $G_{\mathrm{o}d}$ of $\{\mathrm{x}_{\mathrm{od}}[-J], ..., \mathrm{x}_{\mathrm{od}}[K]\}$ the encodes the interconnections of the binary variables in F[15].

In the case that the underlying distribution X is the one-dimensional standard uniform distribution, if the dataset we are displayed with is a representation in some base β that is not a power of 3, I think it is very feasible to get a good estimate of the distribution of $X[t]$ for at least one or a few t's, provided that the encoding/decoding precision is high enough and the dataset size is large enough (see **Proposition Idea**).

Proposition Idea: Let B be an arbitrary finite, nonempty, nonsingleton Borel subset of $\mathbb{N}_0$. Order the elements of B to produce a strictly increasing sequence: $B \equiv \{b_1, \ldots, b_k\}$. Let $N \geq b_k$ be a positive integer. Enumerate the Borel subsets of $\{0, \ldots, N\}$ as $B'_1, \ldots, B'_L$, and for each $1 \leq r \leq L$, let C_r be the set of digits indexed by B'_r. Let $\mathcal{U}$ denote the standard uniform distribution. Suppose that for some positive integer $\beta \geq 1$, $\mathcal{U}$ can be parameterized to induce a manifold S that can be expressed in the form:

$$S = \left\{p(x; \boldsymbol{\theta}, \boldsymbol{v}) \mid \boldsymbol{\theta} \in \mathbb{R}^N, \boldsymbol{v} \in \mathbb{R}^L\right\},$$

and such that (1) S is geodesically convex when the proximity between points on S is measured by Kullback-Leibler divergence, and (2) when $\mathcal{U}$ is written in this form, the interactions within each digit clique C_r in base-β, is encoded in $\mathcal{U}$ only through a simple polynomial c_r.

Suppose $(\mathcal{U},\ \beta)$ satisfied the condition specified right above. Let $\boldsymbol{w} = (w_1, \ldots, w_n)$ be a tuple of random variables drawn i.i.d. from $\mathcal{U}$, and let $x^{(\beta)}[t]^*_{(n)}$ be the equilibrium point of the BP algorithm for estimating $\mathbb{E}\left[x^{(\beta)}[t]\right]$ from $\boldsymbol{w}$. Then we have

$$x^{(\beta)}[t]^*_{(n)} \xrightarrow{P} \frac{\beta - 1}{2}$$

for every positive integer t between 1 and N.

In the generic case (i.e., the binary variables do not represent digit expansion of a normal number), we can impose bounds on the e-curvature of M^*_d via simple restrictions on the form that F

can take. For example, if each clique in the graph $\{\mathrm{x}_{\circ\mathrm{d}}[-J], ..., \mathrm{x}_{\circ\mathrm{d}}[K]\}$ is represented in F as a simple finite polynomial, then $q(\mathrm{x}_{\circ\mathrm{d}})$ will lie in the sub-manifold M_d^* computed through backpropagation(BP) if $G_{\circ d}$ is a tree. In this generic case, it is relatively straightforward to obtain a rejection rule through theoretical derivation or sampling. However, the problem we have here is that the the restrictions on $G_{\circ 1}, ..., G_{\circ D}$ have to be consistent the earlier assumption that $\mathrm{P}\left\{X \in \mathcal{N}^D\right\} = 1$, that is, X is almost always a normal vector. This is not necessarily to find a a set of restrictions on the graphs that do all three: (1) actually models reality, (2) induces properties that preserve the number-theoretical normality of the random vector, and (3) results in useful constraints on the rate of change in relative frequencies, for example, by further characterizing the inequality ($\star$).

To be continued...

References

1. Mann HB and Whitney DR. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. The Annals of Mathematical Statistics 1947;18:50–60.
2. Multivariate Stochastic Orders. In: *Stochastic Orders*. Ed. by Shaked M and Shanthikumar JG. New York, NY: Springer New York, 2007:265–322. DOI: 10.1007/978-0-387-34675-5_6. URL: https://doi.org/10.1007/978-0-387-34675-5_6.
3. Deb N and Sen B. Multivariate Rank-based Distribution-free Nonparametric Testing using Measure Transportation. 2019. arXiv: 1909.08733 [math.ST].
4. Huang Z and Sen B. A Kernel Measure of Dissimilarity between M Distributions. 2022. arXiv: 2210.00634 [math.ST].
5. Chatterjee S. A survey of some recent developments in measures of association. 2023. arXiv: 2211.04702 [stat.ME].
6. Borel E. Les probabilités dénombrables et leurs applications arithmétiques. Rend. Circ. Mat. Palermo 1909;27:247–71.
7. Champernowne DG. The Construction of Decimals Normal in the Scale of Ten. Journal of the London Mathematical Society 1933;s1-8:254–60.
8. Copeland AH and Erdős P. Note on normal numbers. Bull. Amer. Math. Soc. 1946;52:857–60.
9. Fan S. The Copeland-Erdős Theorem on Normal Numbers. Accessed: [Your Access Date Here]. Unknown. URL: https://math.dartmouth.edu/~stevefan/papers/The%20Copeland-Erdos%20Theorem%20on%20Normal%20Numbers.pdf.
10. Erdős P and Rényi A. On a new law of large numbers. Journal d'Analyse Mathématique 1970;23:103–11.
11. Nakai Y and Shiokawa I. Discrepancy Estimates for a Class of Normal Numbers. Acta Arithmetica 1992;62:271–84.

12. Bailey DH and Crandall RE. On the Random Character of Fundamental Constant Expansions. The Journal's Name 2001;Volume Number:Page Range.

13. Bourke C, Hitchcock JM, and Vinodchandran NV. Entropy Rates and Finite-State Dimension. Theoretical Computer Science 2005;349:392–406.

14. Harman G. One Hundred Years of Normal Numbers. Publisher Address: Publisher Name, 2001. DOI: `DOINumber`. URL: `URL%20to%20Publication`.

15. Ikeda S, Tanaka T, and Amari Si. Stochastic Reasoning, Free Energy, and Information Geometry. Neural Computation 2004;16:1779–810.