Allotaxonometry

Last updated: 2024/09/26, 08:35:21 EDT

Principles of Complex Systems, Vols. 1, 2, & 3D CSYS/MATH 6701, 6713, & a pretend number, 2024–2025

Prof. Peter Sheridan Dodds

Computational Story Lab | Vermont Complex Systems Center Santa Fe Institute | University of Vermont

 \circledcirc Δ icensed under the Creative Commons Attribution 4.0 International \vec{Z}

"Allotaxonometry and rank-turbulence divergence: A universal instrument for comparing complex systems"

Dodds et al., EPJ Data Science, **12**, 1–42, 2023. [5]

 8 6 FIG. 1. Allotaxonograph comparing 2-gram usage in the first and second half of Jane Austen Austen Austen Australia . Histogram on the left: We bin all non-zero along the bottom of each axis (lighter gray and lighter blue). Contour lines indicate where probability-turbulence divergence ³/4,τ , showing a mixture of rare and common 2-grams. Ranked list on probability-turbulence divergence, and other probability "Probability-turbulence divergence: A tunable allotaxonometric instrument for comparing heavy-tailed categorical distributions" Dodds et al., $, 2020$. [6]

Basic science = Describe + Explain:

- � Dashboards of single scale instruments helps us understand, monitor, and control systems.
- � Archetype: Cockpit dashboard for flying a plane
- � Okay if comprehendible.
- � Complex systems present two problems for dashboards:
	- 1. Scale with internal diversity of components: We need meters for every species, every company, every word.
	- 2. Tracking change: We need to re-arrange meters on the fly.
- � Goal—Create comprehendible, dynamically-adjusting, differential dashboards showing two pieces:¹
	- 1. 'Big picture' map-like overview,
	- 2. A tunable ranking of components.

¹See the lexicocalorimeter \mathbb{Z}

Baby names, much studied: [12]

How to build a dynamical dashboard that helps sort through a massive number of interconnected time series?

The PoCSvers Allotaxonometry 8 of 70 A plenitude of distances Rank-turbulen divergence Probability-turbulence divergence Exploration Nutshell Referenc

For language, Zipf's law has two scaling regimes: [19]

The PoCSverse Allotaxonometry 5 of 70 A plenitude of distances Rank-turbulence divergence Probabilityturbulence div Explorations Nutshell Reference

$$
f\sim \left\{ \begin{array}{l} r^{-\alpha}\ {\rm for}\, r\ll r_{\rm b},\\ r^{-\alpha'}\ {\rm for}\, r\gg r_{\rm b}, \end{array} \right.
$$

When comparing two texts, define Lexical turbulence as flux of words across a frequency threshold:

$$
\phi \sim \left\{ \begin{array}{l} f_{\rm thr}^{-\mu} \: {\rm for} \: f_{\rm thr} \ll f_{\rm b}, \\ f_{\rm thr}^{-\mu'} \: {\rm for} \: f_{\rm thr} \gg f_{\rm b}, \end{array} \right.
$$

,

Estimates: $\mu \simeq 0.77$ and $\mu' \simeq 1.10,$ and $f_{\rm b}$ is the scaling break point.

$$
\phi \sim \left\{ \begin{array}{l} r^{\nu} = r^{\alpha \mu'} \text{ for } r \ll r_{\rm b}, \\ r^{\nu'} = r^{\alpha' \mu} \text{ for } r \gg r_{\rm b}. \end{array} \right.
$$

Estimates: Lower and upper exponents $\nu \simeq 1.23$ and $\nu' \simeq 1.47$.

m 2017/08/19 Divergence contribution $\delta D^{\rm R}_{\alpha}$. 1 0.5 0 0.5 1 $\frac{a=1/3}{1/4-1/2}$ D^2 = (0, 1.0, 1 = 0.4%) $\propto \sum \left| \frac{1}{\mu^{1/3}} - \frac{1}{\mu^{1/3}} \right|$ $100,000$

Goal—Understand this:

Outline

Explorations

Nutshell

References

A plenitude of distances

Rank-turbulence divergence

Probability-turbulence divergence

The PoCSverse Allotaxonometry Site (papers, examples, code):

1 of 70 A plenitude of distances Rank-turbulence divergence Probabilityturbulence divergence Exploration Nutshell References

The PoCSverse Allotaxonometry 2 of 70 A plenitude of distances Rank-turbulenc divergence Probabilityturbulence divergence
sederlence divergence Exploration Nutshell References

http://compstorylab.org/allotaxonometry/ �

Foundational papers:

Balances:

Top bar (optional)—Total size:

- � Relative balance of system sizes.
- � Examples: Total number of words in a book, total number of individuals in an ecology.

Middle bar—Types:

� Fraction of types in each system as a percentage of the union of types from both systems.

Bottom bar—Exclusive types:

- � Types that are present in one system only are 'exclusive types'.
- $\mathbf{\Omega}^{(1)}$ -exclusive and $\Omega^{(2)}$ -exclusive indicate which system an exclusive type belongs to.
- � Percentage of exclusive types in a system relative to that system's total number of types.

So, so many ways to compare probability distributions:

Entropy ²⁰¹⁰, *¹²* ¹⁵⁴² $\frac{1}{\alpha_1}$ $\frac{1}{\alpha_2}$ $\overline{}$ **−** ^α([−] 1) , α ⁶= 0,¹ ¹ [−] ^u ⁺ ^u logu, α = 1. 25.7 $\frac{12.7}{12.7}$ dµ(x). **Contract Contract Contract** ^u ¹[−]^α + (^α [−] 1)^u [−] ^α ^α(^α [−] 1) , α ⁶= 0, ¹,, α = 1. dµ(x). and the contract of ^u ^u + 1 ²^u ^α [−] ^u [−] ^α ¹[−] ^u . . ^α(^α [−] 1) , α ⁶= 0,¹ ¹ [−] ^u ²[−] ^u log \blacksquare 2u , α and α ^u [−] ¹ ² ⁺ $...$ log \cdots ²^u , α = 1. __ (1) (1) (1) ^u + 12 ^α[−]¹[−] ¹# , α = 1. "Families of Alpha- Beta- and Gamma- Divergences: Flexible and Robust Measures of Similarities" \vec{C} Cichocki and Amari,

Several distance measures listed in Table 2 facilitate the *^L*1, more precisely the absolute difference. The eqn (5), which is widely used in ecology [11], is known as *Sørensen* distance [12] or *Bray-Curtis* [2,4,13]. When it is used for comparing two pdfs, it is nothing but the *^L*1 divided by 2. *Gower* distance [14] in the eqn (6) scales the vector space into the normalized space and then uses the *^L*1. Since the pdf is already normalized space, *Gower* distance is the *^L*1 divided by *d*. Other *L*1 family distances that are non-proportional to the *^L*1 include *Soergel* and *Kulczynski* distances given in the eqns (8) [4] and (9) [2] respectively. At first glance, *Canberra* metric given in the eqn (10) [2,15] resembles *Sørensen* but normalizes the absolute difference of the individual level. It is known to be very sensitive to small changes near zero [15]. The eqn (11) [2], attributed to *Lorentzian*, also contains the absolute difference and the natural logarithm is applied. 1 is added to guarantee

Zago de la

---Contractor of

..... **CONTRACTOR** $-$ -11 ^α(^α [−] 1) ^Z part of the United \sim \sim ²^p ^α [−] ^p⁺ ^α ^p [−] ^q ² Z^p log **2** Pr − 22.000 \sim \sim \sim ²log \sim \sim -2 **p** − q 22.23

. .

ZOO

¹

^Z

ĒŤ. A couple of thousand years ago, Euclid stated that the shortest distance between two points is a line and thus the eqn (1) is predominantly known as Euclidean distance. It was often called Pythagorean metric since it is derived from the Pythagorean Theorem. In the late 19th century, Hermann Minkowski considered the city block distance [9]. Other names for the eqn (2) include rectilinear distance, taxicab norm, and Manhattan distance. Hermann also generalized the formulae (1) and (2) to the eqn (3) which is coined after Minkowski. When *p* goes to infinite, the eqn (4) can be derived and it is called the chessboard distance in 2D, the minimax approximation, or the Chebyshev distance named after Pafnuty Lvovich Chebyshev [10]. **Table 2.** *^L*1 family 5. Sørensen [∑] [∑] == ⁺ [−] ⁼ *dii idii isor PQPQd*11() [|] |(5)

^α(^α [−] 1) ^Z $p = 1$ pa + 1 + a(p− q) Z^p log ^p

12 of 70

distances

Entropy, **12**, 1532-1568, 2010. [2] "Comprehensive survey on distance/similarity

measures between probability density functions" Sung-Hyuk Cha,

F. International Journal of Mathematical Models and Methods in Applied Sciences, **1**, 300–307, 2007. [1]

� Comparisons are distances, divergences �, similarities, inner products, fidelities ...

60ish kinds of comparisons grouped into 10 families

A worry: Subsampled distributions with very heavy tails *^d* ions with very h eavy tails

Shannon tried to slow things down in 1956:

"The bandwagon" $\mathbb Z$

turbulence divers $\mathcal{L}_{\rm 1}$ is sharper to be $\mathcal{L}_{\rm 2}$ guistics, fundamental physics, economics, the theory and our critical thresholds should be raised. Authors should submit only their best efforts, and these only a thoroughly scientific attitude can we achieve real

The PoCSverse Allotaxonometry 14 of 70 A plenitude of distances Rank-turbule divergence Probabilityturbulence dive Exploration Nutshell Referen

Claude E Shannon, IRE Transactions on Information Theory, 2, 3, Explorations 1956. [16] a measurement provides the basis for an empirical estimate of

divergences in probability and information theory fields \mathcal{R} . C_1 computing the distance between two pdf's can be regarded as \mathcal{L}_1

- \bullet "Information theory has ... become something of a scientific bandwagon." ecome something of a scientific
- While ... information theory is indeed a valuable tool ... [it] is certainly no panacea for the communication engineer or ... for anyone else.
- ²⁶ A few first rate research papers are preferable to a large number that are poorly conceived or half-finished."

 \bullet No tunability Minkowski considered the city block distance [9]. Other

 $* L_1$ family \supset {Intersectoin (13), Wave Hedges (15), Czekanowski (16), Ruzicka (21), Tanimoto (23), etc}.

(1) is predominantly known as Euclidean distance. It was

The PoCSverse Allotaxonometry 16 of 70 A plenitude of distances Rank-turbulence divergence Probabilityturbulence div Exploration

D_{eferen}

Nutshell

outence urvergence
plorations

 $% \mathcal{L}_{1} = \mathcal{L}_{2}$ for energy (13), \mathcal{L}_{2}

� Shannon's Entropy:

$$
H(P) = \langle \log_2 \frac{1}{p_{\tau}} \rangle = \sum_{\tau \in R_{1,2;\alpha}} p_{\tau} \log_2 \frac{1}{p_{\tau}}
$$
 (1)

Allotaxo A plenitude of distances Rank-turbu divergence Probability-

Exploration Nutshell References

References

� Kullback-Liebler (KL) divergence:

$$
D^{KL} (P_2 || P_1) = \left\langle \log_2 \frac{1}{p_{2,\tau}} - \log_2 \frac{1}{p_{1,\tau}} \right\rangle_{P_2}
$$

=
$$
\sum_{\tau \in R_{1,2;\alpha}} p_{2,\tau} \left[\log_2 \frac{1}{p_{2,\tau}} - \log_2 \frac{1}{p_{1,\tau}} \right]
$$

=
$$
\sum_{\tau \in R_{1,2;\alpha}} p_{2,\tau} \log_2 \frac{p_{1,\tau}}{p_{2,\tau}}.
$$
 (2)

- � Problem: If just one component type in system 2 is not present in system 1, KL divergence = ∞ .
- � Solution: If we can't compare a spork and a platypus directly, we create a fictional spork-platypus hybrid.
- \triangle New problem: Re-read solution.

The PoCSverse Allotaxonometry 20 of 70 distances divergence Probability-Explorations Nutshell � Jensen-Shannon divergence (JSD): [9, 7, 13, 1] JS (¹ ∣∣ 2) = 1 ²KL (¹ ∣∣ ¹ 2 [¹ + 2]) + ¹ ²KL (² ∣∣ ¹ 2 [¹ + 2]) = 1 ² [∑] ∈1,2; (1, log² 1, 1 2 [1, + 2,] + 2, log² 2, 1 2 [1, + 2,]) . (3)

- � Involving a third intermediate averaged system means JSD is now finite: $0 \le D^{J\tilde{S}}(P_1 || P_2) \le 1.$
- \bullet Generalized entropy divergence: [2]

$$
\begin{split} D^{\mathrm{AS2}}_{\alpha}\left(P_{1} \parallel P_{2}\right)= \\ \frac{1}{\alpha(\alpha-1)}\sum_{\tau \in R_{1,2;\alpha}}\left[\left(p_{\tau,1}^{1-\alpha}+p_{\tau,2}^{1-\alpha}\right)\left(\frac{p_{\tau,1}+p_{\tau,2}}{2}\right)^{\alpha}-\left(p_{\tau,1}+p_{\tau,2}\right)\right]. \end{split} \tag{4}
$$

Produces JSD when $\alpha \to 0$.

Desirable rank-turbulence divergence features:

- 1. Rank-based.
- 2. Symmetric.
- 3. Semi-positive: $D_{\alpha}^{R}(\Omega_1 \parallel \Omega_2) \geq 0$.
- 4. Linearly separable, for interpretability.
- 5. Subsystem applicable: Ranked lists of any principled subset may be equally well compared (e.g., hashtags on Twitter, stock prices of a certain sector, etc.).
- 6. Turbulence-handling: Suited for systems with rank-ordered component size distribution that are heavy-tailed.
- 7. Scalable: Allow for sensible comparisons across system sizes.
- 8. Tunable.
- 9. Story-finding: Features 1–8 combine to show which component types are most 'important'

Some good things about ranks:

- � Working with ranks is intuitive
- � Affords some powerful statistics (e.g., Spearman's rank correlation coefficient)
- � Can be used to generalize beyond systems with probabilities

A start:

 $\frac{1}{r_{\tau,1}} - \frac{1}{r_{\tau,2}}$. (5)

- � Inverse of rank gives an increasing measure of 'importance'
- � High rank means closer to rank 1

We introduce a tuning parameter:

� We assign tied ranks for components of equal 'size'

∣ 1

� Issue: Biases toward high rank components

∣ 1

The PoCSvers Allotaxonometry 26 of 70 A plenitude of

$$
\frac{1}{\left[r_{\tau,1}\right]^{\alpha}}-\frac{1}{\left[r_{\tau,2}\right]^{\alpha}}\bigg|^{1/\alpha}\,.
$$

- \triangle As $\alpha \to 0$, high ranked components are increasingly dampened
- � For words in texts, for example, the weight of common words and rare words move increasingly closer together.
- \bullet As $\alpha \to \infty$, high rank components will dominate.
- � For texts, the contributions of rare words will vanish.

The PoCSverse

A plenitude of distances Rank-turbulence divergence Probabilityturbulence diverge Exploration Nutshell Referenc

 \bullet The limit of $\alpha \to 0$ does not behave well for

� The leading order term is:

$$
\left(1-\delta_{r_{\tau,1}r_{\tau,2}}\right)\alpha^{1/\alpha}\left|\ln\frac{r_{\tau,1}}{r_{\tau,2}}\right|^{1/\alpha},\,
$$

which heads toward ∞ as $\alpha \to 0$.

- � Oops.
- \bullet But the insides look nutritious:

 $\left| \ln \frac{r_{\tau,1}}{r_{\tau,2}} \right|$

is a nicely interpretable log-ratio of ranks.

The PoCSverse Allotaxonometry 25 of 70 A plenitude of distances

> Rank-turbulence divergence Probabilityturbulence div Exploration Nutshell References

 (6)

distances Probability-

The PoCSverse Allotaxonometry 27 of 70 A plenitude of distances

Exploration Nutshell **References**

 (7)

Some reworking:

$$
\delta D^{\rm R}_{\alpha,\tau}(R_1 \parallel R_2) \propto \frac{\alpha+1}{\alpha} \left| \frac{1}{\left[r_{\tau,1}\right]^\alpha} - \frac{1}{\left[r_{\tau,2}\right]^\alpha} \right|^{1/(\alpha+1)}.
$$
 (8) $\frac{\text{Rinkrurbulence}}{\text{divergence}} \left|\frac{\text{Rinkrurbulence}}{\text{Probability}}\right|$

- � Keeps the core structure.
- \triangle Large α limit remains the same.
- $\leftrightarrow \alpha \rightarrow 0$ limit now returns log-ratio of ranks.
- \bullet Next: Sum over τ to get divergence.
- � Still have an option for normalization.

Rank-turbulence divergence:

$$
D_{\alpha}^{\rm R}(R_1 \parallel R_2) = \frac{1}{\mathcal{N}_{1,2;\alpha}} \sum_{\tau \in R_{1,2;\alpha}} \delta D_{\alpha,\tau}^{\rm R}(R_1 \parallel R_2) \tag{9}
$$

Normalization:

- � Take a data-driven rather than analytic approach to determining $\mathcal{N}_{1,2;\alpha}$.
- \bigotimes Compute $\mathcal{N}_{1,2;\alpha}$ by taking the two systems to be disjoint while maintaining their underlying Zipf distributions.
- \bullet Ensures: $0 \leq D_{\alpha}^{\mathbb{R}}(R_1 \, \| \, R_2) \leq 1$
- � Limits of 0 and 1 correspond to the two systems having identical and disjoint Zipf distributions.

General normalization:

The PoCSverse Allotaxonometry 28 of 70 A plenitude of

Exploration Nutshell R eferen

The PoCSverse

A plenitude of distances Rank-turbulence divergence Probability-turbulence divergence Exploration Nutshell Reference

The PoCSverse Allotaxonometry 30 of 70 A plenitude of distances Rank-turbulence divergence Probability-turbulence divergence Exploration Nutshell Reference

- \bullet Iif the Zipf distributions are disjoint, then in $\Omega^{(1)}$'s merged ranking, the rank of all $\Omega^{(2)}$ types will be $r = N_1 + \frac{1}{2}N_2$, where N_1 and N_2 are the number of distinct types in each system.
- \bullet Similarly, $\Omega^{(2)}$'s merged ranking will have all of $\Omega^{(1)}$'s types in last place with rank $r = N_2 + \frac{1}{2}N_1$.

\bullet The normalization is then:

$$
\mathcal{N}_{1,2;\alpha} = \frac{\alpha+1}{\alpha} \sum_{\tau \in R_1} \left| \frac{1}{[r_{\tau,1}]^{\alpha}} - \frac{1}{[N_1 + \frac{1}{2}N_2]^{\alpha}} \right|^{1/(\alpha+1)} + \frac{\alpha+1}{\alpha} \sum_{\tau \in R_2} \left| \frac{1}{[N_2 + \frac{1}{2}N_1]^{\alpha}} - \frac{1}{[r_{\tau,2}]^{\alpha}} \right|^{1/(\alpha+1)}.
$$
\n(11)

Allotaxonometry 29 of 70

Limit of
$$
\alpha \to 0
$$
:

$$
D_0^R(R_1 \parallel R_2) = \sum_{\tau \in R_{1,2;\alpha}} \delta D_{0,\tau}^R = \frac{1}{\mathcal{N}_{1,2;0}} \sum_{\tau \in R_{1,2;\alpha}} \left| \ln \frac{r_{\tau,1}}{r_{\tau,2}} \right|, \exp\left(\frac{r_{\text{probability}}}{r_{\text{hydrogen}}} \right)
$$

where

$$
\mathcal{N}_{1,2;0} = \sum_{\tau \in R_1} \left| \ln \frac{r_{\tau,1}}{N_1 + \frac{1}{2} N_2} \right| + \sum_{\tau \in R_2} \left| \ln \frac{r_{\tau,2}}{\frac{1}{2} N_1 + N_2} \right|.
$$
 (13)

 \bullet Largest rank ratios dominate.

 ∞ :

Probability-turbulence divergence:

$$
D_{\alpha}^{\rm p}(P_1 \mid\mid P_2) = \frac{1}{\mathcal{N}_{1,2;\alpha}^{\rm p}} \frac{\alpha+1}{\alpha} \sum_{\tau \in R_{1,2;\alpha}} \left| \left[p_{\tau,1} \right]^{\alpha} - \left[p_{\tau,2} \right]^{\alpha} \right|^{1/(\alpha+1)}
$$
(16)

- \bullet For the unnormalized version ($\mathcal{N}^{\text{P}}_{1,2;\alpha}$ =1), some troubles return with 0 probabilities and $\alpha \rightarrow 0$.
- \bullet Weep not: ${\mathcal N}_{1,2;\alpha}^{\mathbf{p}}$ will save the day.

The PoCSverse Allotaxonometry 36 of 70 A plenitude of distances Rank-turbulenc divergence

.

Probability-turbulence divergence

Exploration Nutshell **References**

Rank-turbulence divergence:

Summing over all types, dividing by a normalization prefactor $\mathcal{N}_{1,2;\alpha}$ we have our prototype:

$$
D_{\alpha}^{R}(R_{1} \parallel R_{2}) = \frac{1}{\mathcal{N}_{1,2;\alpha}} \frac{\alpha+1}{\alpha} \sum_{\tau \in R_{1,2;\alpha}} \left| \frac{1}{[r_{\tau,1}]}^{\alpha} - \frac{1}{[r_{\tau,2}]}^{\alpha} \right|^{1/(\alpha+1)}.
$$
\n(10)

$$
Limit of \alpha \rightarrow
$$

$$
D_{\infty}^{R}(R_{1} \| R_{2}) = \sum_{\tau \in R_{1,2;\alpha}} \delta D_{\infty,\tau}^{R} \lim_{\tau \to 0} \sum_{\text{trabilistic} \atop \text{trichations}} \left(1 - \delta_{r_{\tau,1}r_{\tau,2}}\right) \max_{\tau} \left\{\frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}}\right\}. \quad (14)
$$

where

$$
\mathcal{N}_{1,2;\infty} = \sum_{\tau \in R_1} \frac{1}{r_{\tau,1}} + \sum_{\tau \in R_2} \frac{1}{r_{\tau,2}}.
$$
 (15)

� Highest ranks dominate.

The PoCSverse

Normalization:

With no matching types, the probability of a type present in one system is zero in the other, and the sum can be split between the two systems' types:

$$
\mathcal{N}_{1,2;\alpha}^{\mathrm{p}} = \frac{\alpha+1}{\alpha} \sum_{\tau \in R_1} \left[p_{\tau,1} \right]^{\alpha/(\alpha+1)} + \frac{\alpha+1}{\alpha} \sum_{\tau \in R_2} \left[p_{\tau,2} \right]^{\alpha/(\alpha+1)} \tag{17}
$$

$$
\begin{array}{c|c}\n & \dots & \\
\downarrow & \searrow & \\
\downarrow & \searrow & \downarrow\n\end{array}
$$

$$
\left| \frac{r_{\tau,1}}{1 - \frac{1}{2}} \right| + \sum_{n=1}^{\infty} \left| \ln \frac{r_{\tau,2}}{1 - \frac{1}{2}} \right|.
$$
 (13)

$$
\sum_{\substack{\text{decomputation}\\\text{divergence}\\\text{Probability.}}\\\text{Pr}_{\tau,2}}\left|\ln\frac{r_{\tau,1}}{r_{\tau,2}}\right|,\qquad\substack{r_{\text{coulodining}}\\\text{trivialness}\\}\text{Explortions}}\right|
$$

turbulence divergence

The PoCSverse Allotaxonometry 32 of 70 A plenitude of distances

$$
\sum_{R_{1,2;\alpha}} \delta D^{\rm R}_{0,\tau} = \frac{1}{\mathcal{N}_{1,2;0}} \sum_{\tau \in R}
$$

$$
(\mathcal{O}_{\mathcal{A}})^{\mathcal{O}_{\mathcal{A}}}
$$

Limit of $\alpha = 0$ for probability-turbulence divergence

 \triangle if both $p_{\tau,1} > 0$ and $p_{\tau,2} > 0$ then

$$
\lim_{\alpha \to 0} \frac{\alpha + 1}{\alpha} \left| \left[p_{\tau,1} \right]^{\alpha} - \left[p_{\tau,2} \right]^{\alpha} \right|^{1/(\alpha+1)} = \left| \ln \frac{p_{\tau,2}}{p_{\tau,1}} \right|.
$$
 (18)

 \bullet But if $p_{\tau,1} = 0$ or $p_{\tau,2} = 0$, limit diverges as $1/\alpha$.

Rank-turbulen divergence Probability-turbulence divergence Exploration Nutshell

Type contribution ordering for the limit of α =0

- � In terms of contribution to the divergence score, all exclusive types supply a weight of $1/(N_1 + N_2)$. We can order them by preserving their ordering as $\alpha \to 0$, which amounts to ordering by descending probability in the system in which they appear.
- � And while types that appear in both systems make no contribution to $D_0^p(P_1 \| P_2)$, we can still order them according to the log ratio of their probabilities.
- � The overall ordering of types by divergence contribution for α =0 is then: (1) exclusive types by descending probability and then (2) types appearing in both systems by descending log ratio.

$$
D^{\rm p}_{\infty}(P_1\,\|\,P_2)=\frac{1}{2}\sum_{\tau\in R_{1,2;\infty}}\left(1-\delta_{p_{\tau,1},p_{\tau,2}}\right)\max\left(p_{\tau,1},p_{\tau,2}\right) \label{eq:pp}
$$
 where

Limit of $\alpha = \infty$ for probability-turbulence divergence

$$
\mathcal{N}_{1,2;\infty}^{\mathrm{P}} = \sum_{\tau \in R_{1,2;\infty}} \left(p_{\tau,1} + p_{\tau,2} \right) = 1 + 1 = 2. \tag{22}
$$

(21)

distances

Nutshell Reference

Combine these cases into a single expression:

$$
D_0^{\rm p}(P_1 \, \| \, P_2) = \frac{1}{(N_1 + N_2)} \sum_{\tau \in R_{1,2;0}} \left(\delta_{p_{\tau,1},0} + \delta_{0,p_{\tau,2}} \right). \tag{20} \tag{20} \tag{21} \sum_{\text{subendial} \atop \text{Roficences}} \frac{1}{\sigma_{\text{subendial}}} \left(\delta_{p_{\tau,1}} + \delta_{p_{\tau,2}} \right) \cdot \sum_{\text{subendial} \atop \text{Roficences}} \frac{1}{\sigma_{\text{subendial}}} \left(\delta_{p_{\tau,1}} + \delta_{p_{\tau,2}} \right) \cdot \sum_{\text{subendial} \atop \text{Roficences}} \frac{1}{\sigma_{\text{subendial}}} \left(\delta_{p_{\tau,1}} + \delta_{p_{\tau,2}} \right)
$$

- \bullet The term $\left(\delta_{p_{\tau,1},0} + \delta_{0,p_{\tau,2}} \right)$ returns 1 if either $p_{\tau,1} = 0$ or $p_{\tau,2} = 0$, and 0 otherwise when both $p_{\tau,1} > 0$ and $p_{\tau,2} > 0$.
- � Ratio of types that are exclusive to one system relative to the total possible such types,

Probability-turbulence divergence Connections for PTD:

> $\hat{\mathbf{B}}$ $\alpha = 0$: Similarity measure Sørensen-Dice coefficient [4, 17, 10], F_1 score of a test's accuracy $^{\left[18,\,15\right] }.$

 $\hat{\Phi}$ $\alpha = 1/2$: Hellinger distance [8] and Mautusita distance [11].

 \triangle $\alpha = 1$: Many including all $L^{(p)}$ -norm type constructions.

 $\hat{\mathbf{\infty}} \ \alpha = \infty$: Motyka distance^[3].

Limit of α =0 for probability-turbulence divergence

� Normalization:

$$
\mathcal{N}_{1,2;\alpha}^{\mathrm{p}} \to \frac{1}{\alpha} \left(N_1 + N_2 \right). \tag{19}
$$

 \bullet Because the normalization also diverges as $1/\alpha$, the divergence will be zero when there are no exclusive types and non-zero when there are exclusive types.

The PoCSverse Allotaxonometry 38 of 70 A plenitude of distances Rank-turbulence

divergence Probability-turbulence divergence Exploration Nutshell Reference

The PoCSverse Allotaxonometr 39 of 70 A plenitude of distances Rank-turbule divergence

FIG. 8. Rank-turbulence divergence allotaxonograph [34] of word rank distributions in the incel vs random comment corpora. The rank-rank histogram on the left shows the density of words by their rank in the incel comments corpus against their rank in the random comments corpus. Words at the top of the diamond are higher frequency, or lower rank. For example, the word "the" appears at the highest observed frequency, and thus has the lowest rank, 1. This word has the lowest rank in both corpora, so its coordinates lie along the center vertical line in the plot. Words such as "women" diverge from the center line because their rank in the incel corpus is higher
than in the random corpus. The top 40 words with greatest divergence contribution are shown on the right comparison, nearly all of the top 40 words are more common in the incel corpus, so they point to the right. The word that has the most notable change in rank from the random to incel corpus is "women", the object of hatred

The PoCSverse

distances

Flipbooks for PTD:

\bullet Jane Austen:

Pride and Prejudice, 1-grams Pride and Prejudice, 2-grams Pride and Prejudice, 3-grams

\rightarrow Social media:

Twitter, 1-grams $\boxplus \textcolor{red}{\mathbf{C}}$ Twitter, 2-grams Twitter, 3 -grams $\boxplus \heartsuit$

\bullet Ecology:

Barro Colorado Island

Flipbooks for RTD:

\clubsuit Twitter:

allotaxonometer-flipbook-1-rank-div.pdf�� allotaxonometer-flipbook-2-probability-div.pdf allotaxonometer-flipbook-3-gen-entropy-div.pdf

\clubsuit Market caps:

allotaxonometer-flipbook-4-marketcaps-6years-rank-div.pdf��

\clubsuit Baby names:

allotaxonometer-flipbook-5-babynames-girls-50years-rank-div.pdf $\boxplus\vec$ allotaxonometer-flipbook-6-babynames-boys-50years-rank-div.pdf

Baby girl names over time relative to 1950 $\boxplus \textcircled{\;}$ Baby boy names over time relative to $1950 \overline{11}$

> The PoCSverse Allotaxonometry 60 of 70 A plenitude of Claims, exaggerations, reminders: � Needed for comparing large-scale complex systems:

distances Rank-turbulence divergence

Probability-turbulence divergence

Explorations

� Many measures seem poorly motivated and largely unexamined (e.g., JSD).

� Of value: Combining big-picture maps with ranked lists.

Comprehendible, dynamically-adjusting, differential

- � Online tunable versions of rank-turbulence divergence now exist:
	- \bullet App version: https://allotaxp.vercel.app/ \bullet
	- \bigcirc Observable version:

dashboards.

https://observablehq.com/@jstonge/allotaxonometer-4-all Github: https://github.com/jstonge/allotaxp C

� Future: Probability-turbulence divergence plus many other instruments.

 Ω_1 : Twitter on 2020/03/12 $\frac{1}{4}$ $\frac{1}{1/4}$ $\frac{1}{1/2}$ $\frac{1}{3/4}$ $\frac{1}{3}$ $\begin{array}{l} D_{\gamma_1}^{\mathbb{P}}(\Omega_1\,||\,\Omega_2)=0.88\\ =\frac{1}{2}\sum_{\nu}\Big(1-\delta_{\mu_{\nu}^{\mathbb{U}},\mu_{\nu}^{\mathbb{U}}} \end{array}$

� Google books:

Nutshell allotaxonometer-flipbook-7-google-books-onegrams-rank-div.pdf allotaxonometer-flipbook-8-google-books-bigrams-rank-div.pdf allotaxonometer-flipbook-9-google-books-trigrams-rank-div.pdf

Code:

https://gitlab.com/compstorylab/allotaxonometer

The PoCSvers Allotaxonometry 62 of 70 A plenitude of distances Rank-turbulence divergence

Probabilityturbulence divergence

Explorations Nutshell Referenc

The PoCSverse Allotaxonometry 63 of 70 A plenitude of distances Rank-turbule divergence Probability-turbulence divergence Exploration

Nutshell **References**

References I

[1] S.-H. Cha.

Comprehensive survey on distance/similarity measures between probability density functions. International Journal of Mathematical Models and Methods in Applied Sciences, 1:300-307, 2007. pdf \Box

[2] A. Cichocki and S.-i. Amari. Families of Alpha- Beta- and Gamma- divergences: Flexible and robust measures of similarities. Entropy, 12:1532–1568, 2010. pdf�

[3] M.-M. Deza and E. Deza. Dictionary of Distances. Elsevier, 2006.

References II

[4] L. R. Dice.

Measures of the amount of ecologic association between species. Ecology, 26:297–302, 1945.

[5] P. S. Dodds, J. R. Minot, M. V. Arnold, T. Alshaabi, J. L. Adams, D. R. Dewhurst, T. J. Gray, M. R. Frank, A. J. Reagan, and C. M. Danforth. Allotaxonometry and rank-turbulence divergence: A universal instrument for comparing complex systems. EPJ Data Science, 12(37):1–42, 2023. Available online at https://arxiv.org/abs/2002.09770. pdf�

References III

[6] P. S. Dodds, J. R. Minot, M. V. Arnold, T. Alshaabi, J. L. Adams, A. J. Reagan, and C. M. Danforth. Probability-turbulence divergence: A tunable allotaxonometric instrument for comparing heavy-tailed categorical distributions, 2020. Available online at https://arxiv.org/abs/2008.13078.pdf \neg

[7] D. M. Endres and J. E. Schindelin. A new metric for probability distributions. IEEE Transactions on Information theory, 2003. pdf \Box

[8] E. Hellinger. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen.

Journal für die reine und angewandte Mathematik (Crelles Journal), 1909(136):210-271, 1909. pdf \Box

References IV

The PoCSverse Allotaxonometry 64 of 70 A plenitude of distances Rank-turbulence divergence Probabilityturbulence divergen Exploration Nutshell References

The PoCSverse Allotaxonometry 65 of 70 A plenitude of distances Rank-turbulence divergence Probabilityturbulence divergence Exploration Nutshell References

- [9] J. Lin. Divergence measures based on the Shannon entropy. IEEE Transactions on Information theory, 37(1):145–151, 1991. pdf C
- [10] J. Looman and J. B. Campbell. Adaptation of Sørensen's k (1948) for estimating unit affinities in prairie vegetation. Ecology, $41(3):409-416$, 1960. pdf \mathbb{Z}^{\bullet}
- [11] K. Matusita et al.

Decision rules, based on the distance, for problems of fit, two samples, and estimation. The Annals of Mathematical Statistics, 26(4):631–640, 1955. $pdfG$

References V

- [12] R. Munroe. How To: Absurd Scientific Advice for Common Real-World Problems. Penguin, 2019.
- [13] F. Osterreicher and I. Vajda. A new class of metric divergences on probability spaces and its applicability in statistics. Annals of the Institute of Statistical Mathematics, 55(3):639–653, 2003.
- [14] E. A. Pechenick, C. M. Danforth, and P. S. Dodds. Is language evolution grinding to a halt? The scaling of lexical turbulence in English fiction suggests it is not. Journal of Computational Science, 21:24-37, 2017. pdf
- [15] Y. Sasaki.

The truth of the f-measure, 2007.

The PoCSverse Allotaxonometry References VI

[16] C. E. Shannon.

turbulence divergence The bandwagon. IRE Transactions on Information Theory, 2(1):3, 1956. $pdfC$

Explorations Nutshell References

The PoCSverse Allotaxonometry 68 of 70 A plenitude of distances Rank-turbulen divergence

Explorations Nutshell References

67 of 70 A plenitude of distances Rank-turbulence divergence Probability-

> A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons.

[18] C. J. Van Rijsbergen. Information retrieval. Butterworth-Heinemann, 2nd edition, 1979. Allotaxor.
69 of 70 A plenitude of distances Rank-turbulence divergence Probabilityturbulence diverg Explorations Nutshell

References

The PoCSverse Allotaxonometry

[17] T. Sorensen.

Videnski Selskab Biologiske Skrifter, 5:1–34, 1948.

Physical Review E, 91:052811, 2015. pdf

66 of 70 A plenitude of distances Rank-turbulence divergence Probabilityt toomomy
turbulence divergen Exploration Nutshell References

The PoCSverse Allotaxonometry