Chemical Formula Ratios, Scale Estimation and Distributional Hypothesis Testing for Medicinal Plant Extracts

Jeff Cromwell, PhD

Mathematical Learning Space Research Portfolio Spring Semester 2021: (DRAFT) Preprint 66

1 Abstract

Dimensionality reduction and generalizations are common strategies in the reductionist scientific paradigm for research advancement. Here the projection of two vectors into a single vector based on the ratio of polynomials with their scales and measures are provided for a set of chemical compounds from the organic element alphabet of medicinal plants. For four ratios H/C, H/O, H/N and O/N the pearson types are 1,3,4,and 1 respectively with positive correlation H/C, H/O and H/N and O/N. Based on the scale design model the linear combination of the parameters are: (Intercept) -1.1134 H.C -0.3002 H.O 0.0307 H.N -0.0145 O.N 0.2306

Keywords: H/C, H/O, H/N, O/N ratios, Pearson Distribution, Polynomials, Medicinal herbs

2 Introduction

The relationship between two molecules at different scales with the A and not A categoification is fundamental molecular biology approach where the rule of seperation provides insight into differences between in the relationship. In this relationship A to not A, the $\frac{A}{notA}$ is also an arrangement that reduces the dimensionality from 2 to 1. In the case of four elements each of these fractions can be partitioned, i.e. carbon, hydrogen, nitrogen and oxygen into its own alphabet. Each of these fractions or ratios can be viewed independently and together as group for structural classification. Here in this note, four ratios are constructed and maxmimum likehood estimation used with the Pearson distribution to identified the type of pearson disribution with each of ratio as well as correlation tests with each two vector combination for significance testing. For example, in probability theory, Transforms (function of a random variable) and Combinations (function of several variables) are the measures and scales in knowledge aquistion and accumulation. [1]

3 Distributions and Scales

The pearson family of distribution include types 1 -7 wth the following relationships with other distributions. [1] [300] [301]

1. Beta distribution (type I)
2. Continuous uniform distribution (limit of type I)
3. Chi-squared distribution (type III)
4. Exponential distribution (type III)
5. Gamma distribution (type III)
6. Cauchy distribution (type IV)
7. Normal distribution (limit of type I, III, IV, V, or VI)
8. Inverse-chi-squared distribution (type V)
9. Inverse-gamma distribution (type V)
10. F-distribution (type VI)
11. Beta prime distribution (type VI)
12. Student's t-distribution (type VII, which is the non-skewed subtype of type IV)

Multiplying the variable by any positive real constant yields a scaling of the original distribution where some are self-replicating i.e. scaling yields the same family of distributions, albeit with a different parameter: normal distribution, gamma distribution, Cauchy distribution, exponential distribution, Erlang distribution, Weibull distribution, logistic distribution, error distribution, power-law distribution, Rayleigh distribution. [1] [300] [301]

1. The sum of n Bernoulli (p) random variables is a binomial (n, p) random variable.
2. The sum of n geometric random variables with probability of success p is a negative binomial random variable with parameters n and p.
3. The sum of n exponential (β) random variables is a gamma (n, β) random variable.
4. If the exponential random variables have a common rate parameter, their sum has an Erlang distribution, a special case of the gamma distribution.
5. The sum of the squares of N standard normal random variables has a chi-squared distribution with N degrees of freedom.

The distribution of the sum of independent random variables is the convolution of their distributions. Suppose Z is the sum of n independent random variables $X_1, \dots, X_n X_1, \dots, X_n$ each with probability mass functions $f_{X_i}(x)$ Then

$$Z = \sum_{i=1}^{n} X_i \tag{1}$$

If it has a distribution from the same family of distributions as the original variables, that family of distributions is said to be closed under convolution. Convolution of Bernoulli distributions [1] The convolution of two independent identically distributed Bernoulli random variables is a binomial random variable. That is, in a shorthand notation, [1] $\sum_{i=1}^{2} \text{Bernoulli}(p) \sim \text{Binomial}(2, p)$

let $X_i \sim \text{Bernoulli}(p), \quad 0 < p < 1, \quad 1 \leq i \leq 2$ and define $Y = \sum_{i=1}^{2} X_i$. let "Z" denote a generic binomial random variable: $Z \sim \text{Binomial}(2, p)$

X_1 and X_2 are independent, [300] [301]

[*] The Mathematical Learning Space Research Portfolio
Email address: http://mathlearningspace.weebly.com/ (Jeff Cromwell, PhD)

$$
\begin{aligned}
\mathbb{P}[Y = n] &= \mathbb{P}\left[\sum_{i=1}^{2} X_i = n\right] \qquad (2)\\
&= \sum_{m\in\mathbb{Z}} \mathbb{P}[X_1 = m] \times \mathbb{P}[X_2 = n - m] \qquad (3)\\
&= \sum_{m\in\mathbb{Z}} \left[\binom{1}{m} p^{m} (1-p)^{1-m}\right]\left[\binom{1}{n-m} p^{n-m} (1-p)^{1-n+m}\right] \qquad (4)\\
&= p^{n} (1-p)^{2-n} \sum_{m\in\mathbb{Z}} \binom{1}{m}\binom{1}{n-m} \qquad (5)\\
&= p^{n} (1-p)^{2-n} \left[\binom{1}{0}\binom{1}{n} + \binom{1}{1}\binom{1}{n-1}\right] \qquad (6)\\
&= \binom{2}{n} p^{n} (1-p)^{2-n} = \mathbb{P}[Z = n] \qquad (7)
\end{aligned}
$$

Here, we used the fact that $\binom{n}{k} = 0$ for ”k”>”n” in the last but three equality, and of Pascal’s rule in the second last equality. Medicinal herbs can be placed in two groups: Group A and Group B. [1] [401]

3.1. Group A

1. Fennel [EV:E00057]
2. Eleuthero [EV:E00180]
3. Artichoke [EV:E00756]
4. German chamomile [EV:E00781]
5. Burdock root [EV:E00797]
6. Butterbur [EV:E00800]
7. Feverfew [EV:E00805]
8. Milk thistle [EV:E00822]
9. Bearberry [EV:E00058]
10. Cranberry [EV:E00775]
11. Comfrey [EV:E00779]
12. Cat’s claw [EV:E00773]
13. Pau D’Arco [EV:E00799]
14. Sage [EV:E00783]
15. Thyme [EV:E00787]
16. Peppermint [EV:E00810]
17. Lavender [EV:E00827]
18. Rosemary [EV:E00834]
19. Olive [EV:E00767]
20. Devil’s claw [EV:E00792]
21. Cayenne papper [EV:E00014]
22. Pumpkin seed [EV:E00802]

3.2. Group B

1. Licorice [EV:E00027]
2. Senna [EV:E00122]
3. St. John’s wort [EV:E00784]
4. Flaxseed [EV:E00446]
5. Willow [EV:E00814]
6. C01451 Salicin
7. Hemp [EV:E00162]
8. Mulberry [EV:E00819]
9. Rose hip [EV:E00833]
10. Cloves [EV:E00018]
11. Tea tree [EV:E00791]
12. Evening promrose [EV:E00761]
13. Neem [EV:E00793]
14. Bitter orange peel [EV:E00031]
15. Guarana [EV:E00771]
16. Horse chestnut [EV:E00812]
17. Goldenseal [EV:E00777]
18. Saffron [EV:E00003]
19. Lemon grass [EV:E00830]
20. Turmeric [EV:E00069]
21. Ginger [EV:E00115]
22. Garlic [EV:E00769]
23. Ginkgo [EV:E00760]
24. Ephedra [EV:E00160]
25. Iceland moss [EV:E00757]
26. Maitake [EV:E00815]

For the 343 plant compounds, the following molecular formulas are presented in Table 1. [401]

1	C10H12O	C10H12O	C21H22O5	C42H68O13	C48H78O17
2	C15H16O3	C12H18O2	C12H14O2	C16H14O4	C17H16O6
3	C11H10O5	C48H82O18	C42H72O14	C47H80O17	C47H74O18
4	C48H76O19	C41H70O12	C47H74O18	C17H24O9	C35H60O6
5	C34H46O18	C16H12O7	C15H10O6	C43H42O22	C16H32O2
6	C13H10O	C15H20O	C17H28	C15H18O3	C59H94O29
7	C24H32O6	C57H92O28	C59H94O29	C24H32O6	C57H92O28
8	C21H26N2O4	C15H26O2	C17H26O10	C12H16O7	C12H16O7
9	C7H6O2	C16H16O5	C53H78O17	C53H78O17	C46H56N4O10
10	C46H58N4O9	C16H22O10	C16H20O9	C16H20O9	C21H22N2O2
11	C21H22N2O2	C21H22N2O2	C29H40N2O4	C15H14O6	C17H24O10
12	C22H28N2O4	C29H40N2O4	C22H26O12	C10H20O	C10H20O
13	C10H20O	C21H18O11	C10H18O	C10H14O	C29H36O15
14	C15H22O9	C21H20O11	C21H20O11	C15H22O10	C17H23NO3
15	C17H23NO3	C18H27NO3	C18H27NO3	C5H11NO2	C17H23NO3
16	C18H30O2	C7H7NO2	C42H62O16	C16H12O4	C21H20O9
17	C15H24N2O	C16H12O5	C42H38O20	C42H38O20	C42H40O19
18	C42H40O19	C16H14O5	C42H62O16	C10H8O4	C70H104O32
19	C75H112O36	C70H104O32	C17H21NO4	C14H16O9	C20H34O2
20	C16H32O2	C25H24O6	C19H28O11	C25H38O16	C20H27NO11
21	C27H30O16	C20H27NO11	C20H27NO11	C15H26O	C15H10O6
22	C41H28O27	C41H30O26	C41H30O26	C19H17N3O	C10H16
23	C10H16	C10H16	C20H18NO4	C27H32O14	C16H25NO
24	C28H34O15	C11H16N2O2	C21H30O6	C22H32O6	C24H34O7
25	C22H28O6	C19H24NO3	C15H23NO2	C33H40O15	C47H76O17
26	C41H66O13	C47H76O16	C19H23NO4	C21H22NO4	C37H41N2O6
27	C17H19NO3	C20H19NO5	C34H47NO11	C30H48O5	C10H8O4
28	C20H18NO4	C27H44O7	C15H10O4	C42H38O20	C42H38O20
29	C23H28O11	C9H10O3	C30H50O5	C30H48O4	C30H46O5
30	C8H8O4	C4H8N2O3	C39H64O13	C19H18O11	C44H64O24
31	C41H64O13	C39H62O12	C46H72O17	C44H70O16	C57H94O28
32	C50H80O24	C8H10O3	C21H22O9	C8H13NO2	C15H24O
33	C31H52O	C38H70O4	C12H20O2	C15H18O2	C21H20O6
34	C17H26O4	C19H30O4	C21H34O4	C10H18O	C17H26O4
35	C19H30O4	C21H34O4	C10H18O	C10H18O	C27H42O3
36	C27H45NO3	C22H25NO6	C21H22O11	C24H32O7	C10H18O
37	C9H8O	C10H16O	C9H8O	C18H18O2	C17H19NO3
38	C11H14O2	C12H22O2	C20H30O2	C47H51NO14	C10H15NO
39	C10H15NO4	(C12H18O9)n	C28H44O	C31H50O3	C26H34O6
40	C6H12O6	C6H12O6	CaCO3	CaSO4	(CaSO4)2. H2O
41	C10H12O	C10H14O	C10H18O	C10H18O	C10H16
42	C10H12O	C8H8O3	C10H20O	C10H20O	C10H12O
43	C12H20O2	C10H18O	C10H18O	C12H20O2	C10H18O
44	C10H18O	C12H20O2	C10H14O	C15H26O	C10H14
45	C10H16	C10H16O	C10H14O	C10H14O	C10H18O
46	C10H14O	C10H14	C10H16O	C10H18O	C10H18O
47	C14H12O2	C9H10O2	C10H20O	C10H20O	C10H20O
48	C10H20O	C10H16	C10H18O	C41H30O26	C10H12O2
49	C10H12O2	C10H16	C10H16	C10H16	C10H18O
50	C10H18O	C10H18O	C10H16	C10H16	C10H16
51	C15H26O	C10H16	C12H20O2	C15H24O	C14H22O
52	C10H18O	C10H18O	C10H18O	C12H20O2	C10H18O
53	C15H24	C10H12O	C10H12O	C9H8O	C10H18O
54	C10H18O	C9H8O	C10H12O2	C9H10O2	C10H16
55	C10H16	C10H16	C10H16O	C15H24	C10H16
56	C15H24	C10H18O	C10H18O	C10H18O	C10H16
57	C10H16	C10H12O	C17H24O9	C35H60O6	C34H46O18
58	C25H24O12	C15H26O	C14H16	(C12H20O10)n	C20H28O3
59	C15H20O3	C25H22O10	C12H16O7	C6H8O6	C4H6N4O3
60	C21H24N2O4	C15H14O3	C10H16O	C10H14O	C10H20O
61	C12H20O2	C18H16O8	C25H32O13	C24H30O11	C18H27NO3
62	C29H48O	C42H62O16	C42H38O20	C42H38O20	C42H40O19
63	C42H40O19	C30H16O8	C18H30O2	C18H30O2	C13H18O7
64	C16H32O2	C6H13NO4	C6H8O6	C41H30O26	C10H18O
65	C18H32O2	C18H34O2	C10H16	C8H10N4O2	C55H86O24
66	C20H18NO4	C21H21NO6	C44H64O24	C10H16O	C10H16O
67	C21H20O6	C17H26O4	C19H30O4	C21H34O4	C6H11NO3S
68	C20H24O9	C20H24O10	C20H24O11	C20H24O10	C20H24O10
69	C15H18O8	C10H15NO	C6H10O5	C10H12O	C10H12O
70	C21H22O5	C42H68O13	C48H78O17	C15H16O3	C12H18O2

Table 2 has the vocabulary for the molecular forumulas in Table 1. [401]

1	Anethole	Anethole	Notopterol	Saikosaponin	Saikosaponin
2	Osthol	Cnidilide	Ligustilide	Imperatorin	yakangelicol
3	Fraxidin	Ginsenoside	Ginsenoside	Chikusetsusaponin	Chikusetsusaponin
4	Chikusetsusaponin	Chikusetsusaponin	Chikusetsusaponin	Syringin	Eleutheroside
5	Eleutheroside	Capillarisin	Luteolin	Carthamin	Hexadecanoic
6	Atractylodin	Atractylone	Aplotaxene	alpha-Santonin	Platycodin
7	Magnosalin	Platycodin	Platycodin	Magnosalin	Platycodin
8	Lonicerin	ornyl	Loganin	Arbutin	Arbutin
9	enzoate	Shikonin	Condurango	Condurango	Vincristine
10	Vinblastine	Swertiamarin	Gentiopicrin	Gentiopicrin	Strychnine
11	Strychnine	Strychnine	Emetine	(+)-Catechin	Geniposide
12	Rhyncophylline	Emetine	Catalposide	(-)-Menthol	(-)-Menthol
13	(-)-Menthol	aicalin	(+)-Menthone	Perillyl	Forsythiaside
14	Aucubin	Plantaginin	Plantaginin	Catalpol	L-Hyoscyamine
15	L-Hyoscyamine	Capsaicin	Capsaicin	etaine	Atropine
16	Punicic	N-Methylnicotinate	Glycyrrhizinate	Formononetin	Puerarin
17	Matrine	Obtusifolin	Sennoside	Sennoside	Sennoside
18	Sennoside	razilin	Glycyrrhizinate	Scopoletin	Senegin
19	Onjisaponin	Senegin	Cocaine	ergenin	Plaunotol
20	Hexadecanoic	Morusin	Zizybeoside	Zizybeoside	Amygdalin
21	Multinoside	Amygdalin	Amygdalin	Nerolidol	Kaempferol
22	Geraniin	Eugeniin	Eugeniin	Evodiamine	Limonene
23	Limonene	Limonene	erberine	Naringin	alpha-Sanshool
24	Hesperidin	Pilocarpine	Nigakilactone	Nigakilactone	Nigakilactone
25	Quassin	Lotusine	Nupharidine	Icariin	Akeboside
26	Akeboside	Akeboside	Sinomenine	Palmatine	Tubocurarine
27	Morphine	Protopine	Aconitine	Cimigenol	Anemonin
28	erberine	20-Hydroxyecdysone	Chrysophanol	Sennoside	Sennoside
29	Paeoniflorin	Paeonol	Alisol	Alisol	Alisol
30	Homogentisate	Asparagine	Timosaponin	Mangiferin	Crocin
31	Ophiopogonin	Ophiopogonin	Ophiopogonin	Ophiopogonin	Sibiricoside
32	Sibiricoside	Vanillyl	arbaloin	Arecoline	Cyperol
33	Cylindrin	Coixenolide	alpha-Terpinyl	Curzerenone	Curcumin
34	[6]-Gingerol	[8]-Gingerol	(10)-Gingerol	orneol	[6]-Gingerol
35	[8]-Gingerol	(10)-Gingerol	1,8-Cineole	1,8-Cineole	Diosgenin
36	Peimine	Colchicine	Astilbin	Schizandrin	(-)-orneol
37	Cinnamaldehyde	(+)-Camphor	Cinnamaldehyde	Magnolol	(R,S)-Coclaurine
38	Methyleugenol	Decanoyl	Abietate	Paclitaxel	Ephedrine
39	Kainic	Agarose	Ergosterol	Eburicoic	Cinobufagin
40	D-Fructose	D-Glucose	Calcium	Calcium	Calcium
41	Anethole	(+)-(S)-Carvone	(+)-Linalool	(-)-Linalool	beta-Pinene
42	Anethole	Methyl	(-)-Menthol	(-)-Menthol	Estragole
43	Linalyl	(+)-Linalool	(-)-Linalool	Linalyl	(+)-Linalool
44	(-)-Linalool	Linalyl	Carvacrol	Patchouli	p-Cymene
45	Limonene	Thujone	(-)-Carvone	(+)-(S)-Carvone	1,8-Cineole
46	Thymol	p-Cymene	(-)-Camphor	1,8-Cineole	1,4-Cineole
47	enzyl	enzyl	beta-Citronellol	(-)-Citronellol	beta-Citronellol
48	(-)-Citronellol	alpha-Pinene	1,8-Cineole	Eugeniin	Eugenol
49	Eugenol	alpha-Pinene	Limonene	Limonene	1,8-Cineole
50	(+)-Linalool	(-)-Linalool	Limonene	Limonene	Limonene
51	beta-Eudesmol	Limonene	Linalyl	alpha-Santalol	alpha-Irone
52	Citronellal	Citronellal	Geraniol	alpha-Terpinyl	1,8-Cineole
53	Zingiberene	Anethole	Anethole	Cinnamaldehyde	(+)-Linalool
54	(-)-Linalool	Cinnamaldehyde	Eugenol	enzyl	alpha-Pinene
55	Sabinene	Limonene	Thujone	Thujopsene	Limonene
56	Thujopsene	alpha-Terpineol	beta-Terpineol	gamma-Terpineol	alpha-Pinene
57	beta-PineneAMedicinal	Anethole	Syringin	Eleutheroside	Eleutheroside
58	1,3-Dicaffeoylquinic	(-)-alpha-isabolol	Chamazulene	(2,1-beta-D-Fructosyl)n	Petasin
59	Parthenolide	Silymarin	Arbutin	Ascorbate	Allantoin
60	Mitraphylline	Lapachol	Thujone	Thymol	(-)-Menthol
61	Linalyl	Rosmarinate	Oleuropein	Harpagoside	Capsaicin
62	Stigmasterol	Glycyrrhizinate	Sennoside	Sennoside	Sennoside
63	Sennoside	Hypericin	(6Z,9Z,12Z)-Octadecatrienoic	(9Z,12Z,15Z)-Octadecatrienoic	Salicin
64	Hexadecanoic	Deoxynojirimycin	Ascorbate	Eugeniin	Terpineol-4
65	Linoleate	(9Z)-Octadecenoic	Limonene	Caffeine	Aescin
66	erberine	(+)-Hydrastine	Crocin	Geranial	cis-Citral
67	Curcumin	[6]-Gingerol	[8]-Gingerol	(10)-Gingerol	Alliin
68	Ginkgolide	Ginkgolide	Ginkgolide	Ginkgolide	Ginkgolide
69	ilobalide	Ephedrine	Lichenin	Anethole	Anethole
70	Notopterol	Saikosaponin	Saikosaponin	Osthol	Cnidilide

4 Results

For the following (a) C (b) H (c) O and (d) N for n=343, (a) Pearson type 1 where a 4.031078 and b 3.219709 and location 0.330137 and scale 1.935159 for N=343 (b) Pearson type 3 shape 1.023708 location 1.036424 scale 7.344729. (c) Pearson type 4 m 3.549792 nu -2.853269 location 9.143311 scale 17.35644 and (d) Pearson type 1 a 0.4568822 b 1.647233 location 0.3333333 and scale 15.36983. [1]

Based on the fingerprints, the jarvisPatrick with the nearestNeighbors cutoff=0.6 and the similarity method="Tanimoto", k=2 amd mode="b", the second cluster is provided in Table 1. [401]

1	Ligustilide	Syringin	Carthamin	Atractylodin
2	Atractylone	Lonicerin	Loganin	Arbutin
3	Arbutin	Catalpol	L-Hyoscyamine	L-Hyoscyamine
4	Capsaicin	Capsaicin	Glycyrrhizinate	Matrine
5	Obtusifolin	Glycyrrhizinate	Scopoletin	Eugeniin
6	Eugeniin	Evodiamine	Limonene	Limonene
7	Limonene	Hesperidin	Nigakilactone	Nigakilactone
8	Nigakilactone	Palmatine	Morphine	Cimigenol

Figure 1 has Complexity for Limonene and Thujone Structural Simularities. [401]

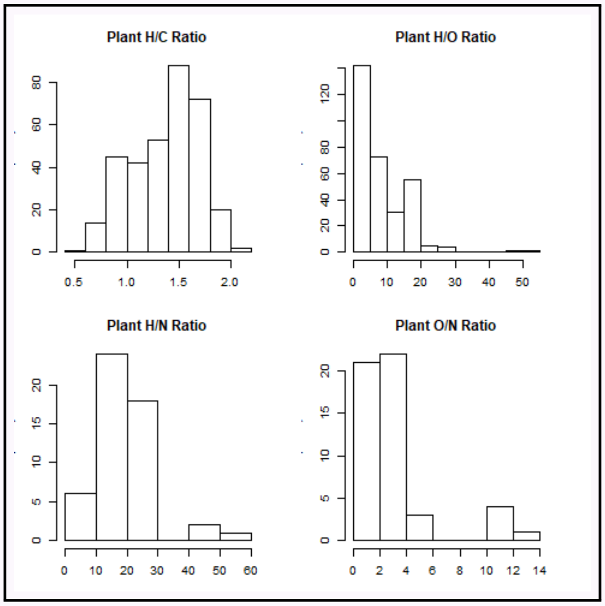

Figure 1: (a) Hydrogen carbon ratio (b) Hydrogen Oxygen (c) Hydrogen Nitrogen and (d) Oxygen Nitrogen Ratios for Plants with N=343

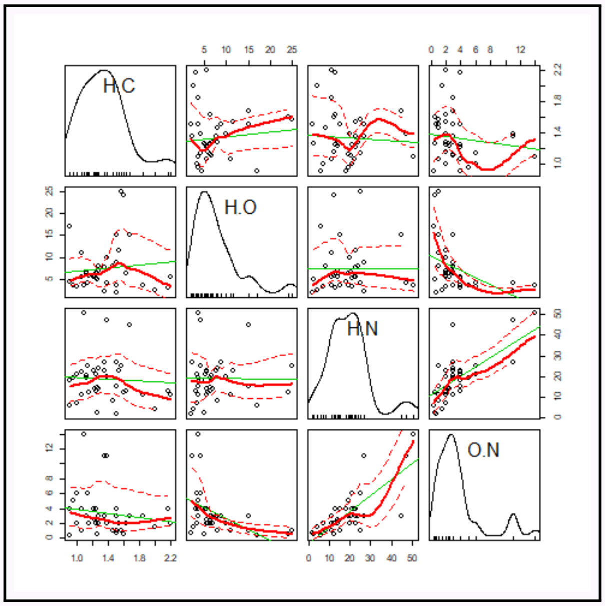

Figure 2: Correlation and Densities for (a) Hydrogen carbon ratio (b) Hydrogen Oxygen (c) Hydrogen Nitrogen and (d) Oxygen Nitrogen Ratios for Plants with N=343

For Pearson's product-moment correlation of H/C and H/O t = 12.173, df = 308, p-value < 2.2e-16 with 95 percent confidence interval: 0.4896198 0.6406540 and cor 0.5699314. for H/O and H/N t = -0.050605, df = 49, p-value = 0.9598 and 95 percent confidence interval: -0.2822505 0.2688903 cor -0.007229101. For H/N and O/N, t = 6.9384, df = 49, p-value = 8.263e-09 with 95 percent confidence interval: 0.5315076 0.8203991 cor 0.7039765.

4.1. *Scale Estimation with Ratio Measures*

For Pearson type 4 distribution with m=2, nu=2, location=1, scale=2, the scale is H/C, H/O and H/N and O/N. Based on the scale design of (Intercept) -1.1134 H.C -0.3002 H.O 0.0307 H.N -0.0145 O.N 0.2306 with Table 1 scale estimation with the Model EDV= H.C + H.O + H.N + O.N with summary information of the cofficients.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.1134	0.8808	-1.26	0.2126
H.C	-0.3002	0.5563	-0.54	0.5920
H.O	0.0307	0.0455	0.67	0.5032
H.N	-0.0145	0.0279	-0.52	0.6043
O.N	0.2306	0.1051	2.19	0.0333

5 Conclusions

Dimensionality reduction and generalizations are common strategies in the reductionist scientific paradigm for research advancement. Here the projection of two vectors into a single vector based on the ratio of polynomials with their scales and measures are provided for a set of chemical compounds from plants. For four ratios H/C, H/O, H/N and O/N the pearson types are 1,3,4,and 1 respectively with positive correlation H/C, H/O and H/N and O/N. Based on the scale design of (Intercept) -1.1134 H.C -0.3002 H.O 0.0307 H.N -0.0145 O.N 0.2306 the prediction accuracy for MSE is 5.339157.

6 References

[1] Hattori, M., Okuno, Y., Goto, S., and Kanehisa, M.; Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. J. Am. Chem. Soc. 125, 11853-11865 (2003).

[2] Kotera, M., Okuno, Y., Hattori, M., Goto, S., and Kanehisa, M.; Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. J. Am. Chem. Soc. 126, 16487-16498 (2004).

[3] Muto, A., Kotera, M., Tokimatsu, T., Nakagawa, Z., Goto, S., and Kanehisa, M.; Modular architecture of metabolic pathways revealed by conserved sequences of reactions. J. Chem. Inf. Model. 53, 613-622 (2013).

[4] Yamanishi, Y., Hattori, M., Kotera, M., Goto, S., and Kanehisa, M.; E-zyme: predicting potential EC numbers from the chemical transformation pattern of substrate-product pairs. Bioinformatics 25, i79-i86 (2009).

[5] Oh, M., Yamada, T., Hattori, M., Goto, S., and Kanehisa, M.; Systematic analysis of enzyme-catalyzed reaction patterns and prediction of microbial biodegradation pathways. J. Chem. Inf. Model. 47, 1702-1712 (2007).

[6] Moriya, Y., Shigemizu, D., Hattori, M., Tokimatsu, T., Kotera, M., Goto, S., and Kanehisa, M.; PathPred: an enzyme-catalyzed metabolic pathway prediction server. Nucleic Acids Res. 38, W138-W143 (2010).

[7] KCF-S: KEGG Chemical Function and Substructure for improved interpretability and prediction inchemical bioinformatics

[8] Muto, A., Kotera, M., Tokimatsu, T., Nakagawa, Z., Goto, S., and Kanehisa, M.; Modular architecture of metabolic pathways revealed by conserved sequences of reactions. J. Chem. Inf. Model. 53, 613-622 (2013).

[9] Kanehisa, M.; Chemical and genomic evolution of enzyme-catalyzed reaction networks. FEBS Lett. 587, 2731-2737 (2013).

[10] Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M.; Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 42, D199–D205 (2014).

[20] Wikipedia contributors. "Oxide." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 16 Jan. 2021. Web. 16 Mar. 2021.

[100] Wikipedia contributors. "Relationships among probability distributions." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 14 Sep. 2020. Web. 16 Mar. 2021.

[300] Wikipedia contributors. "Pearson distribution." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 6 May. 2021. Web. 13 Aug. 2021.

[301] Wikipedia contributors. "Bernoulli distribution." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 13 Jun. 2021. Web. 13 Aug. 2021.

[400] Kanehisa, Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K.; KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 45, D353-D361 (2017).

[401] Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M.; KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44, D457-D462 (2016).

[402] Kanehisa, M. and Goto, S.; KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30 (2000).

[803] Cromwell, J. "Mathematical Learning Space Research Portfolio" Mathematical Learning Space Research Portfolio, http://mathlearningspace.weebly.com/ 8 3 2021. Web. 3 Aug. 2021.

[1000] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[1001] Grün B and Hornik K (2011). “topicmodels: An R Package for Fitting Topic Models.” Journal of Statistical Software, *40*(13), pp. 1-30. doi: 10.18637/jss.v040.i13 (URL: http://doi.org/10.18637/jss.v040.i13).

[1002] Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S and Matsuo A (2018). “quanteda: An R package for the quantitative analysis of textual data.” Journal of Open Source Software, *3*(30), pp. 774.doi: 10.21105/joss.00774 (URL: http://doi.org/10.21105/joss.00774), ⟨URL: https://quanteda.io⟩.

[1003] Ingo Feinerer and Kurt Hornik (2017). tm: Text Mining Package. R package version 0.7-1. https://CRAN.R-project.org/package=tm

[1004] Ingo Feinerer, Kurt Hornik, and David Meyer (2008). Text Mining Infrastructure in R. Journal of Statistical Software 25(5): 1-54. URL: http://www.jstatsoft.org/v25/i05/.

[1005] Sarkar, Deepayan (2008) Lattice: Multivariate Data Visualization with R. Springer, New York. ISBN 978-0-387-75968-5

[1006] Luna A et al. rcellminer: Exploring Molecular Profiles and Drug Response of the NCI-60 Cell Lines in R. Bioinformatics. 2015 Dec. http://www.ncbi.nlm.nih.gov/pubmed/26635141