I will assume that every genomic or genetic DNA sequence is of the form, $b_1 b_2 \cdots b_n$, where $\forall i \in \{1, \dots, n\}$ $b_i \in B := \{A, C, G, T\}$.

$$B^+ = \bigcup_{n \in \mathbb{N}_{>0}} B^n \quad , \quad V_B = sp\{e_A, e_C, e_G, e_T\} \text{ where } e^1\vec{1} = e_A,\ e^2\vec{1} = e_C,\ e^3\vec{1} = e_G,\ e^4\vec{1} = e_T.$$

(V_B over $\mathbb{Q}$ or $\mathbb{R}$ or $\mathbb{C}$)

This defines coordinates.

I am using $m \approx 2.3$ million observations from GISAID. Each observation provides information on the following variables:

Collection YYYY
Collection MM
Collection DD
→ Time (cts)
(Sojo's Suggested Transform)

Continent
Country
Providence/Division/State
City
→ N° from Equator
→ E° from Prime Meridian
(0: Precision: 2 ft)

Host (vector of indicators)
Age ($\mathbb{N}$)
Sex (vector of indicators)

Possible Bases in Position 1 — θ^1 — $\{0,1\}^4$
Possible Bases in Position 2 — θ^2 — $\{0,1\}^4$
⋮
Possible Bases in Position ℓ — $\theta^\ell = \theta^{\#\theta} = \theta^*$ — $\{0,1\}^4$

↑ This is a variable Provided

This is when I focus on the simplification where we observe $\theta = (\theta^1, \theta^2, \dots, \theta^*)$ (over time). Its interesting, the sequence part of the "data" (format) is really only a sequence when the read was "perfect" otherwise we are provided a parameter.

Onto the Models!

A fun Model!

$$\mathcal{F} = \operatorname{span}_{\{0,1\}}\left\{ e_B = e_{B_1} \otimes \cdots \otimes e_{B_n} \mid B \in \mathcal{B}^{\times n},\ n \in \mathbb{N} \right\} = \left\{ c_B e_B \mid B \in \mathcal{B}^{\times n},\ n \in \mathbb{N},\ c_B \in \{0,1\} \right\}, \quad \Omega = \bigcup_{F \in \mathcal{F}} F$$

(field of coeffs.)

$$\mathcal{H} = \left\{ \theta \in \{0,1\}^{\#\mathcal{A} \times n} \mid n \in \mathbb{N} \right\} = \left\{ (), \begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{pmatrix}1\\0\\0\end{pmatrix}, \ldots, \begin{pmatrix}1\\1\\1\end{pmatrix}, \left(\begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{pmatrix}0\\0\\0\end{pmatrix}\right), \left(\begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}0\\0\\0\end{pmatrix}\right), \ldots \right\}$$

$$\mathcal{M}_\otimes = \left(\bigoplus_{n \in \mathbb{N}} V_B^{\otimes n} \right)^* = \bigcup_{n \in \mathbb{N}} L(\underbrace{V_B, \ldots, V_B}_{n \text{ times}}; \mathbb{R})$$

(dual)

$P(\cdot \mid \bullet) = P_\bullet : \mathcal{H} \to \mathcal{M}_\otimes$ given by:

$$\xi = \left(\begin{matrix}\xi_A^1\\ \xi_C^1\\ \xi_G^1\\ \xi_T^1\end{matrix}, \begin{matrix}\xi_A^2\\ \xi_C^2\\ \xi_G^2\\ \xi_T^2\end{matrix}, \ldots, \begin{matrix}\xi_A^\#\\ \xi_C^\#\\ \xi_G^\#\\ \xi_T^\#\end{matrix} \right) \longmapsto \left(\frac{\bigotimes \xi}{|\xi^1| \cdots |\xi^\#|} \right)^*$$

(dual)

(i.e. $P_\bullet := {}^*\bullet \otimes \left(\frac{\theta^1}{|\theta^1|}, \ldots, \frac{\theta^\#}{|\theta^\#|} \right)$)

In particular:

$$\frac{\bigotimes \xi}{|\xi^1| \cdots |\xi^\#|} = \frac{\sum_{B \in \mathcal{B}^{\#\mathcal{A}}} \xi^1_{B_1} \otimes \cdots \otimes \xi^\#_{B_{\#\mathcal{A}}}}{|\xi^1| \cdots |\xi^\#|} \overset{\text{notation}}{=} \frac{\sum_{B \in \mathcal{B}^{\#\mathcal{A}}} \xi_B}{|\xi^1| \cdots |\xi^\#|}$$

and the number of B such that $\xi^1_{B_1} = 1$ and $\xi^2_{B_2} = 1$ and ... and $\xi^\#_{B_{\#\mathcal{A}}} = 1$ is $|\xi^1| \cdots |\xi^\#|$

So $P_\xi = \frac{1}{|\xi^1| \cdots |\xi^\#|} \sum_{B \in \bigcap_{F \in \mathcal{F}} F} \delta^B$ satisfies $P_\xi(\Omega) = |\xi^1| \cdots |\xi^\#| \frac{1}{|\xi^1| \cdots |\xi^\#|} = 1$.

(dual basis element; $\otimes \xi \in \mathcal{F}$)

So now we have a Random Probability Measure !!

It would be completely Random if the following condition is satisfied (Kingman Completely Rand. Measures)

If $A_1, A_2, \ldots, A_n$ mutually disjoint, then $P_\xi(A_1), P_\xi(A_2), \ldots, P_\xi(A_n)$ are independent. RVs.

I don't believe this is the case because you can take a Partition $\bigcup_{i=1}^n A_i = \Omega$, the joint distribution is not independent because $P_\xi(A_n) = 1 - \sum_{i=1}^{n-1} P_\xi(A_i)$.

Is this condition of Kingman's well-defined? With Respect to what measure should one verify this claim with?

Also this doesn't seem to be true of any probability measure...

Anyways if you want "Completely Random Measures" in our context just don't normalize $\otimes \theta$ in the defn of $P_\bullet$. Id rather get probability measures than do this.

Lastly, one could consider Random Variables as elements of $\operatorname{span}_{\mathbb{K}}\{e_B \mid B \in \mathcal{B}^{\#\mathcal{A}},\ n \in \mathbb{N}\}$, where $\mathbb{K}$ is a field.

Another "Model"!

$\Theta = \{\theta \in \mathbb{R}^{B^+} \mid |\theta|_1 = 1\}$ Note: B^+ must be orderable for this to be valid: Partially order by length and then Lexicographically.

$\mathcal{M}_\Theta = \bigcup_{n \in \mathbb{N}} L(V_{B_1}, \dots, V_{B^n}; [0,1])$

$P_\bullet$ defined by: $(f_\omega)_{\omega \in B^+} \longmapsto \sum_{\omega \in B^+} f_\omega \delta^\omega$

More Random Probability Measures!

What happens if we assume sequences are exchangeable?

Suppose we suppose a DNA sequence is exchangeable & define $P_\bullet$ s.t. P_θ exchangeable $\forall \theta$.

Then, our Probability Distribution, $P_\theta = \sum_{n \in \mathbb{N}} P_\theta^{(n)}$, must satisfy:

$g P_\theta^{(n)} = P_\theta^{(n)} \quad \forall g \in G^{(n)}$ ← Symmetric group on n letters.

$P_\theta^{(n)} \in L(X_1, \dots, X_n; [0,1])$

Since $\{e_A, e_C, e_G, e_T\} \otimes \dots \otimes \{e_A, e_C, e_G, e_T\}$

$= \{e_\omega \mid \omega \in B^n\}$

$= \{g(e_A^{\otimes n_1} \otimes e_C^{\otimes n_2} \otimes e_G^{\otimes n_3} \otimes e_T^{\otimes n_4}) \mid g \in G^{(n)}, (n_1, n_2, n_3, n_4) \in \mathbb{N}^4, n_1 + n_2 + n_3 + n_4 = n\}$

$= \bigcup_{\substack{(n_1,n_2,n_3,n_4) \in \mathbb{N}^4 \\ n_1+n_2+n_3+n_4=n}} G^{(n)}(e_A^{\otimes n_1} \otimes e_C^{\otimes n_2} \otimes e_G^{\otimes n_3} \otimes e_T^{\otimes n_4})$

is a basis for the Domain of $P_\theta^{(n)}$, it is defined by the vals it takes on this basis.

So $P_\theta^{(n)} = \sum_{\substack{\alpha \in \mathbb{N}^4 \\ |\alpha| = n}} \sum_{e_\omega \in G^{(n)}(e_A^{\otimes \alpha_1} \otimes \dots \otimes e_T^{\otimes \alpha_4})} P_\theta^{(n)}(e_\omega) = \sum_{\substack{\alpha \in \mathbb{N}^4 \\ |\alpha| = n}} |G^{(n)}(e_A^{\otimes \alpha_1} \otimes \dots \otimes e_T^{\otimes \alpha_4})| \, P_\theta^{(n)}(e_A^{\otimes \alpha_1} \otimes \dots \otimes e_T^{\otimes \alpha_4})$

$= \sum_{\substack{\alpha \in \mathbb{N}^4 \\ |\alpha| = n}} \alpha_1! \alpha_2! \alpha_3! \alpha_4! \, P_\theta^{(n)}(e_A^{\otimes \alpha_1} \otimes \dots \otimes e_T^{\otimes \alpha_4}) \overset{\text{By Dim=Dim}}{\cong} \sum_{\substack{\alpha \in \mathbb{N}^4 \\ |\alpha| = n}} \alpha! \, c_{\theta,\alpha}^{(n)} z_A^{\alpha_1} z_C^{\alpha_2} z_G^{\alpha_3} z_T^{\alpha_4} = \sum_{\substack{\alpha \in \mathbb{N}^4 \\ |\alpha| = n}} \alpha! \, c_{\theta,\alpha} z^\alpha$

This is not an equality.

In Particular, $P_\theta^{(n)}(\Omega) = \sum_{\substack{\alpha \in \mathbb{N}^4 \\ |\alpha| = n}} \alpha! \, c_{\theta,\alpha}^{(n)}$

Since $P_\theta = \sum_{n \in \mathbb{N}} P_\theta^{(n)}$ is a probability Measure: $1 = P_\theta(\Omega) = \sum_{n \in \mathbb{N}} \sum_{\substack{\alpha \in \mathbb{N}^4 \\ |\alpha| = n}} \alpha! \, c_{\theta,\alpha}^{(n)}$

This follows from concluding we have the isomorphism in yellow for each n, since n was arbitrary.

Thus: Every exchangeable Probability Measure is isomorphic to some Polynomial (with coeffs in $\text{Range}(P_\theta)$).

All of the requirements of this polynomial follow from those on the exchangeable Prob. Measure.

To call a probability density or distribution exchangeable is to assert that it is G-invariant. Since exchangeable probability distributions are equivalent to a polynomials, it is natural to want an irreducible representation. These polynomials are not constrained to any degree, so Lets Consider them as elements of $\mathbb{R}[z_A, z_C, z_G, z_T]$, which has a G-invariant subspace $\{c\, z_A^n z_C^n z_G^n z_T^n \mid c \in \mathbb{R}, n \in \mathbb{Z}\} = \langle z_A^n z_C^n z_G^n z_T^n \mid n \in \mathbb{Z} \rangle = \langle z_A z_C z_G z_T \rangle$. We can take care of this By taking the Polynomial Ring in 3 variables, e.g. $\mathbb{R}[z_A - z_C, z_C - z_G, z_G - z_T]$.

Interestingly, since the original polynomial did not change sign under the application of a transposition, these parameterizations don't either. This all terms are of even degree. or $\mathbb{R}[z_A - z_T, z_C - z_T, z_G - z_T]$.

"Model" for exchangeable Sequences

$$\textcircled{H} = \{\theta \in \mathbb{R}^{\mathbb{N}^4} \mid |\theta| = 1\} \text{ or } \{\theta \in \mathbb{C}^{\mathbb{N}^4} \mid |\theta| = 1\}$$

$$\mathcal{M}_{\textcircled{H}} = \left\{ \sum_{\alpha \in \mathbb{N}^4} c_\alpha z^\alpha \mid c_\alpha \geq 0 \right\} = \mathbb{R}_+[z_A, z_C, z_G, z_T]$$

P. given by $\theta \mapsto \sum_{\alpha \in \mathbb{N}^4} \theta_\alpha z^\alpha$

It would be appropriate to let: $\text{Sort}(\theta) \sim GEM(\eta)$

(This is more of a fact than a modelling choice).

A simple thing that should be incorporated.

$$\textcircled{H} = \{\theta \in \mathbb{R}^{\infty} \mid |\theta| = 1\}$$

$$\mathcal{M}_{\textcircled{H}} = \bigcup_n L(v_{\beta_1}, \ldots, v_{\beta_n}; [0,1])$$

$$\theta \longmapsto \left(\sum_{n \in \mathbb{N}} \theta_n \otimes 1^{\otimes n} \right)^*$$

This is just a distribution over lengths of seqs. Discrete normal?

Any (mixture) of a Location/Scale dist. should do. I guess I think of it this way because $\mathbb{P} = \sum \mathbb{P}^{(n)}$ so Empirical Length distribution is. $(\|\mathbb{P}^{(n)}\|, \|\mathbb{P}^{(n+1)}\|, \ldots, 0, \ldots)$

For Messiness

observed Param

$$\Phi = \left\{ \left(\alpha, (\phi_\eta)_{\substack{\eta \neq \beta \\ |\eta| \leq l}} \right) \;\middle|\; \alpha = \sum_i^{\#\theta} 1_{|\theta^i|=1}\, \theta^i,\ \beta = \sum_i^{\#\theta} 1_{|\theta^i|\neq 1}\, \theta^i,\ l = \#\theta \right\}$$

$$\mathcal{M}_\Phi = \mathbb{R}[z]$$

$$(\alpha, \phi) \longmapsto z^\alpha \left(\sum \phi_\eta z^\eta \right)$$

Torosion ? Lag?

$$\mathcal{H} = \bigcup_{n \in \mathbb{N}} \{0,1\}^{\#\mathcal{G} \times n}$$

$$\mathcal{M}_\mathcal{H} = L(\mathbb{N}; V_B^{\otimes \mathcal{G}})$$

$$\theta \longmapsto \left(\theta^{\max(1,i-4)} \otimes \dots \otimes \theta^i \right)_{i=1}^{\#\theta} = \left(S\left(\theta^{\max(1,i-4)} \otimes \dots \otimes \theta^i\right) \oplus A\left(\theta^{\max(1,i-4)} \otimes \dots \otimes \theta^i\right) \right)_{i=1}^{\#\theta}$$

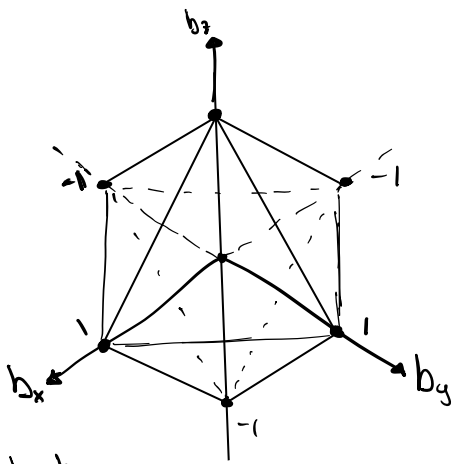
Some $\mathbb{R}[z]$ Some $\mathbb{R}[G_B]$

Anything on the closed simplex (w/ e_A, e_C, e_G, e_T as the extreme points)

$$\mathcal{H} = \left\{ \alpha = (\alpha^1, \dots, \alpha^n) = \begin{pmatrix} \alpha_x^1 & \dots & \alpha_x^n \\ \alpha_y^1 & \dots & \alpha_y^n \\ \alpha_z^1 & & \alpha_z^n \end{pmatrix} \in [-1,1]^{3 \times n} \;\middle|\; \forall i \in \{1, \dots, n\}\ |\alpha^i|_1 \leq 1,\ n \in \mathbb{N} \right\}$$

$$\mathcal{M}_\mathcal{H} = \bigcup_{n \in \mathbb{N}} L(\tilde{V}_B, \dots \tilde{V}_B; \mathbb{R}) \quad \text{where } \tilde{V}_B = \text{span}\{\underbrace{b_A - b_C}_{:= b_x}, \underbrace{b_C - b_G}_{:= b_y}, \underbrace{b_G - b_T}_{:= b_z}\} \text{ (say).}$$

$$\alpha \longmapsto \bigotimes \left(\alpha_x^1 b_x + \alpha_y^1 b_y + \alpha_z^1 b_z, \dots, \alpha_x^n b_x + \alpha_y^n b_y + \alpha_z^n b_z \right)^*$$



Any Distribution on $\alpha^i_{b_1,b_2}$ induces one on G_B.
It Probably makes most sense to let $\alpha^i_{b_1,b_2}$ be exponential or gamma or Laplace (w/ location 0). i like this.

Prob distributions on $\alpha^i_{b_1,b_2}$, say νf, tells us the effect of sign changes, ie. how $f(x)$ compares to $f(-x)$.
This induces a distribution on the transposition $(b_1\ b_2)$, which has representation, $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, a 2×2 real matrix or a complex number.
$\text{Log}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} \frac{\pi i}{2} & -\frac{\pi i}{2} \\ -\frac{\pi i}{2} & \frac{\pi i}{2} \end{pmatrix}$ means that real multiplicative relations have additive Complex analogues.
What other fun approaches are there using Complex variables?

Some facts about Distributions taking $t \in G_{\mathbb{R}}$ as an argument.

Gamma(shape κ, scale θ) $\Rightarrow f(x;\kappa,\theta) = \frac{1}{\Gamma(\kappa)\theta^\kappa} x^{\kappa-1} e^{-x/\theta} \Rightarrow f(t;\kappa,\theta) = \frac{t^{\kappa-1} e^{-t/\theta}}{\Gamma(\kappa)\theta^\kappa}$

(i require this $\in \mathbb{Z}$)

$= \frac{e^{-1/\theta}}{\Gamma(\kappa)\theta^\kappa} t^{\kappa-1} e^{t} = f(t^{\kappa-1}; 0, \theta)$

Also for $c \in \mathbb{R}_{>0}$ $f(ct;\kappa,\theta) = f((ct)^{\kappa-1}; 0, \theta)$

Inverse Gamma(shape κ, scale β) $\Rightarrow f(x;\kappa,) = \frac{\beta^\kappa}{\Gamma(\kappa)} x^{-\kappa-1} e^{-\frac{\beta}{x}}$

$\Rightarrow f(t) = \frac{\beta^\kappa}{\Gamma(\kappa)} t^{-\kappa-1} e^{-\frac{\beta}{t}} = \frac{\beta^\kappa}{\Gamma(\kappa)} (t^{-1})^{\kappa+1} e^{-\beta t^{-1}} = \text{gamma}(t^{-1}, \kappa+2, \frac{1}{\beta})$

$= \text{gamma}(t^{-\kappa-2}, 0, \frac{1}{\beta})$

Chi-Squared ($\kappa \in \mathbb{N}_+$) $\Rightarrow f(x,\kappa) = \frac{1}{\Gamma(\frac{\kappa}{2}) 2^{\kappa/2}} x^{\frac{\kappa}{2}-1} e^{-\frac{x}{2}}$

$\Rightarrow f(x, 2\kappa) = \text{Gamma}(x; \kappa, 2) \Rightarrow \chi^2_{2\kappa}(t) = \text{Gamma}(t^{-\kappa}; 0, 2)$

I have sort of been assuming we may treat our data in time as a stochastic process. We can do this by:

extending the random Probability Measure formulation to each point in time $t_1, t_2, \dots, t_n$ by defining $\mathbb{P}_{t_i} \in T_{t_i}(\bigoplus_n V_B^{\otimes n})^* \simeq (T_{t_i}(\bigoplus_n V_B^{\otimes n}))^*$

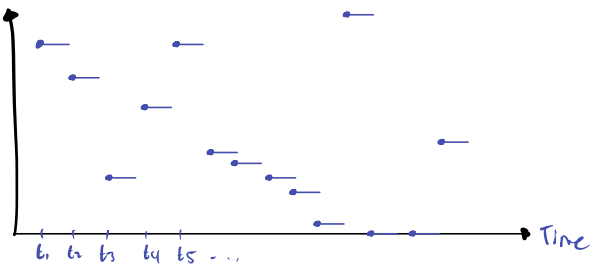
ie. the random measure determined by the data at a point in time t_i is an element of the dual of the tangent space at t_i, which I think is called the cotangent space.

I would prove this but it follows from letting Data take values in $\mathbb{R} \times \bigoplus_n V^{\otimes n}$ where $\mathbb{R}$ is our proxy for time (taking $\mathbb{C}$ could be useful and is appropriate because Polynomials and Permutations are holomorphic). It would be more reflective of the data to take "time" as $\{2014, 2020, 2021\} \times \{1, \dots, 12\} \times \{1, \dots, 31\}$. This perspective is valuable since some data points do not provide MM or DD.

Empirical distribution function up to time T: $\mathbb{F}_T := \frac{\sum_{t \le T} \mathbb{P}_t}{\| \sum_{t \le T} \mathbb{P}_t \|}$

Picture:

A function space $\rightarrow M_\oplus$



Kingman's theoretical approach to genetics problems is motivated by:

> A geneticist who examines a sample of n gametes from a population of size N, and whose experimental techniques do not permit a labelling of the alleles he can distinguish, will have as data a member of ϖ_n.

"ϖ_n" is the set of all partitions of n, i.e. integer partitions of n

In notation: $\varpi_n := \{(a_1, \dots, a_n) \in \mathbb{N}^n \mid \sum_{j=1}^{n} j a_j = n\}$

And a_1 is interpreted as the number of alleles represented once in the sample, a_2 alleles are represented twice in the sample, ..., a_n alleles are represented n times in the sample.

So if we observe a sample (of size n) and denote the set of observed alleles as $S = \{s_1, \dots\}$ and f_{s_i} frequency of some s_i as f_{s_i}, then we may determine $(a_1, \dots, a_n)$ by:

$$a_1 = \sum_{s \in S} 1_{f_s = 1}, \quad a_2 = \sum_{s \in S} 1_{f_s = 2}, \quad \dots, \quad a_n = \sum_{s \in S} 1_{f_s = n}$$

I am not sure what kinds of experimental techniques result(ed) in this kind of raw data. Anyways once one accepts this notion of a partition, there is absolutely no relation to the underlying subject or applied setting (unless this is reflective of the observation generating mechanism).

Regardless, modern techniques can provide arrays of sequences where each element of a sequence is a subset of nucleotides. Subsets of $B = \{A, C, G, T\}$ are observed because the sequencing technology cannot always determine the exact nucleotide at a location, but in these instances it can sometimes "narrow down" which nucleotides could be at that location. So the typical observed genomic or genetic sequence, α, of length $\#\alpha = \ell$, is an element of $\underbrace{2^B \times \dots \times 2^B}_{\ell \text{ times}} = (2^B)^{\times \ell}$. So each sequence may be identified with a subset of exact sequences because

Set theory gives us: $\overbrace{2^B \times \dots \times 2^B}^{\ell \text{ times}} \simeq 2^{\overbrace{B \times \dots \times B}^{\ell \text{ times}}}$. For example,

$$\alpha = (\{A, C\}, \{G, T\}, \{T\}) \simeq \{(A, G, T), (A, T, T), (C, G, T), (C, T, T)\} = \times \alpha$$

Cartesian Product.

It is natural to wonder what distribution our data was drawn from (if there is one) or what the limiting distribution of our data is (if there is one).

Kingman takes a really neat approach to this when the sample is an element of ϖ_n (the set of integer partitions). The sample, a, is an element of ϖ_n because a_i denotes the number of alleles represented i times in the sample, so the sample size is $\sum_i i a_i$, which equals n since $a \in \varpi_n$. Kingman provides a notion a consistent family of distributions, called a "partition structure"

$P = (P_1, P_2, P_3, \ldots) = (P_n \in \Pi_n)_{n \in \mathbb{N}}$ is a partition structure if for all $m < n$, P_m may be obtained by random sampling from P_n (Π_n denotes the set of probability distributions on ϖ_n).

Kingman then provides a construction for some Partition Structures

A very large class of partition structures can be constructed in the following way. Consider an infinite population divided into a countable number of sub-populations labelled (say) by colours 1, 2, 3, ..., and suppose that the proportion of the population which is of colour i is x_i, where

$$x_i \geqslant 0, \quad \sum_{i=1}^{\infty} x_i = 1. \tag{2.8}$$

For $n \geqslant 1$ and $\pi \in \varpi_n$, let $P_n(\pi)$ be the probability that a random sample of size n has a_1 colours represented once, a_2 twice, and so on. Thus $P_n(\pi)$ is a function

$$P_n(\pi) = \phi_\pi(x) \tag{2.9}$$

of $x = (x_1, x_2, \ldots)$, given explicitly by

$$\phi_\pi(x) = n! \prod_{j=1}^{n} (j!)^{-a_j} \sum x_1^{\nu_1} x_2^{\nu_2} \ldots, \tag{2.10}$$

where the sum extends over all sequences $(\nu_1, \nu_2, \ldots)$ of integers $0 \leqslant \nu_i \leqslant n$ such that, for $1 \leqslant j \leqslant n$, exactly a_j of the ν_i are equal to j.

As before, it is clear (and easy to check directly using (2.6)) that (2.9) defines a partition structure for any x satisfying (2.8). Even more general structures can be constructed by allowing x to be random. Thus let Δ be the set of all sequences x satisfying (2.8), which may conveniently be topologized as a subset of the compact metrisable cartesian product $[0, 1]^\infty$ whose elements are sequences x satisfying $0 \leqslant x_i \leqslant 1$. Then ϕ_π is a continuous function from Δ to $[0, 1]$, and if μ is any probability measure on (the Borel subsets of) Δ,

$$P_n(\pi) = \int_\Delta \phi_\pi \, \mathrm{d}\mu \tag{2.11}$$

defines a member P_n of Π_n. Because (2.9) satisfies (2.7), we have

$$\begin{aligned} (\sigma_{mn} P_n)(\rho) &= \sum_{\pi \in \varpi_n} \int_\Delta \phi_\pi \, \mathrm{d}\mu \, S(\pi, \rho) \\ &= \int_\Delta \Big(\sum_{\pi \in \varpi_n} \phi_\pi S(\pi, \rho) \Big) \mathrm{d}\mu \\ &= \int_\Delta \phi_\rho \, \mathrm{d}\mu = P_m(\rho), \end{aligned}$$

so that (2.11) defines a partition structure for any choice of μ.

Results of Watterson (1976) show that the Ewens structure is itself expressible in the form (2.11). To do this, μ is taken as the *Poisson–Dirichlet distribution* $\mathscr{PD}(\theta)$

As statisticians this is a compelling structure. However, this approach is far too abstract when the genomic or genetic data is not an integer partition. This is the case.

I propose a genomic method first.

⓪ Let: $\mathbb{K}$ be a field.

$$V_{YYYY} = \text{span}_{\mathbb{K}}\{e_{2019}, e_{2020}, e_{2021}\}$$
$$V_{MM} = \text{span}_{\mathbb{K}}\{e_{\text{Jan}}, \dots, e_{\text{Dec}}\}$$
$$V_{DD} = \text{span}_{\mathbb{K}}\{e_{01}, \dots, e_{31}\}$$
$$V_{\text{Continent}} = \text{span}_{\mathbb{K}}\{e_{\text{antartica}}, \dots, e_{\text{southamerica}}\}$$
$$V_{\text{Country}} = \text{span}_{\mathbb{K}}\{e_{\text{ctry}} \mid \text{ctry} \in \text{Countries in the world}\}$$
$$V_{\text{City}} = \text{span}_{\mathbb{K}}\{e_{\text{city}} \mid \text{city} \in \text{Cities in the world}\}$$
$$V_{\text{Host}} = \text{span}_{\mathbb{K}}\{e_h \mid h \in \text{Hosts}\}$$
$$V_{\text{Age}} = \text{span}_{\mathbb{K}}\{e_{n\text{years}} \mid n \in \{0, \dots, 100\}\}$$ ← CHECK
$$V_{\text{SEX}} = \text{span}_{\mathbb{K}}\{e_{\text{Male}}, e_{\text{Female}}\}$$
$$V_B = \text{span}_{\mathbb{K}}\{e_A, e_C, e_G, e_T\}$$

Require: $\|e_x\|_1 = 1 \quad \forall x \in \{2019, 2020, 2021, \dots, A, C, G, T\}$

① Let: Each observation, X, be an element of

$$V_{\text{Info}} := V_{YYYY} \times V_{MM} \times V_{DD} \times V_{\text{Continent}} \times V_{\text{Country}} \times V_{\text{City}} \times V_{\text{Host}} \times V_{\text{Age}} \times V_{\text{SEX}} \times \bigcup_{l \in \mathbb{N}_+} V_B^{\times l}$$

$$= \bigcup_{l \in \mathbb{N}_+} V_{YYYY} \times V_{MM} \times V_{DD} \times V_{\text{Continent}} \times V_{\text{Country}} \times V_{\text{City}} \times V_{\text{Host}} \times V_{\text{Age}} \times V_{\text{SEX}} \times V_B^{\times l}$$ (× distributes over ∪)

$$= \bigcup_{l \in \mathbb{N}} V_{YYYY} \times V_{MM} \times V_{DD} \times V_{\text{Continent}} \times V_{\text{Country}} \times V_{\text{City}} \times V_{\text{Host}} \times V_{\text{Age}} \times V_{\text{SEX}} \times \overbrace{V_B \times \dots \times V_B}^{l \text{ times}}$$ (How V_B^0 works and notations $B_1 = B^{\times 1}$)

$$= \left(V_{YYYY} \times V_{MM} \times V_{DD} \times V_{\text{Continent}} \times V_{\text{Country}} \times V_{\text{City}} \times V_{\text{Host}} \times V_{\text{Age}} \times V_{\text{SEX}} \times \{1\}\right) \cup \left(\bigcup_{l \in \mathbb{N}_+} V_{YYYY} \times V_{MM} \times V_{DD} \times V_{\text{Continent}} \times V_{\text{Country}} \times V_{\text{City}} \times V_{\text{Host}} \times V_{\text{Age}} \times V_{\text{SEX}} \times V_{B_1} \times \dots \times V_{B_l}\right)$$

Require: $\|X_u\|_1 = 1$ if $\dim(V_u)$ unbounded, for $u \in \text{Vars} = \{YYYY, MM, DD, \text{Continent}, \text{Country}, \text{City}, \text{Host}, \text{Age}, \text{SEX}, B_1, B_2, \dots\}$

Note: $V_{\text{Info}} \cong \bigcup_{l \in \mathbb{N}} V_{YYYY} \times MM \times DD \times \text{Continent} \times \text{Country} \times \text{City} \times \text{Host} \times \text{Age} \times SEX \times B^{\times l}$

② Denote the data by $\mathbb{X}$ and the sample size by m.

Observe: $\mathbb{X} \in V_{\text{Info}}^{\times m}$, so $\mathbb{X} = (X^1, X^2, \dots, X^m)$ where $X^i \in V_{\text{Info}}$ for $i \in \{1, \dots, m\}$

Note: $\#\mathbb{X} = m$ & It is possible that $|\{X^i \mid i \in \{1, \dots, m\}\}| \neq m$

③ The empirical density of an observation is $(\otimes x)^*$.
If $|x|=1$, then $(\otimes x)^*$ is the empirical distribution of the observation.

④ The average Density of the observations $\underline{x}$ is $\frac{1}{\#\underline{x}}\sum_{i=1}^{\#\underline{x}}(\otimes \underline{x}^i)^*$
If $|\underline{x}^i|=1 \ \forall i=1,\dots,\#\underline{x}$, then $\frac{1}{\#\underline{x}}\sum_{i=1}^{\#\underline{x}}(\otimes \underline{x}^i)^*$ is the empirical d.f., Denoted $\mathbb{F}$.

④.1 If you would like, define a fn. $\omega: V_{\text{info}} \to \mathbb{R}$ and then the density w.r.t. ω is $\sum_{i=1}^{\#\underline{x}} \omega(\underline{x}^i)(\otimes \underline{x}^i)^*$.

⑤ A process of $\mathbb{F}$ is defined by a non-empty subset $J \subset \text{Vars}$, and the isomorphism $\mathbb{F} \simeq V_J \otimes V^*_{\text{Vars}\setminus J}$. This satisfies the requirements of a vector field given by the basis of $V_{j_1}\otimes\dots\otimes V_{j_{|J|}}$ where the points are the indicies of the basis. The value of this vector field at a point is an element of $V^*_{\text{Vars}\setminus J} = L(V_{\text{Vars}\setminus J}; \mathbb{K})$, in particular $\mathbb{F}_\alpha$ where $\alpha \in \bigtimes_{j\in J} \text{Vals}(j)$ may be defined by $\mathbb{F}_\alpha = \frac{\partial}{\partial x_\alpha}\mathbb{F}^*$ which really just amounts to the sum of duals of basis elements whose indicies are equal to α when projected onto the variables of J. Weight these however you please.

In epidemiological settings it is probably desireable to accomodate genomic and non-genomic data at the same time because we do not get genomic data from every infected individual. More broadly, we may wish to incorporate population data with genomic data.

We can achieve both with the above formalism. By construction we have permitted a genomic observation to have 0 nucleotide bases, or equivalently, the sequence observed is in $B^{\times 0} = \{()\}$. The tensor product maps this into the field $\mathbb{K}$. Letting it map to $1_{\mathbb{K}}$ is really the only sensible thing to do since we'd hope that $\otimes B^{\times l} = \otimes B^{\times 0} \times B^{\times l}$. In which case, a observation be interpreted as population obs $\otimes$ genomic obs where population obs Lives in a covariate vector space over $\mathbb{K}$ and the genomic obs lives in $\otimes B^{\times l}$ for some $l \in \mathbb{N} = \{0,1,\dots\}$.

In retrospect we probably want to express some certainty or uncertainty about what the true measure is, if there is one. If we have reason to believe our observations are drawn from a "true" distribution that changes over time or with other variables (observed or not), then expressing uncertainty over an indexed collection of measures is also of interest. We can achieve this by expressing priors over the Θs in the various "models" above. I recognize these are very explicit notions of ambiguity and that some more flexibility or ambiguity given by a fewer number of parameters might offer useful ways to "turn knobs" when we are restricted to a number of hands.

Below are some ideas:

Let $\theta_b^n \sim \begin{cases} 1 & \text{w.p. } x_b \\ 0 & \text{w.p. } 1-x_b \end{cases}$ $\forall n \in \{1, \dots, \#\Theta\}$ $\forall$ observed Θ

What if we take $\theta_b^n \sim f(x) = \frac{\theta}{(1-x)^{\theta-1}} dx, \theta > 0$

Let $\theta_b^n \sim \text{Erlang}(k, \lambda)$ CDF: $F(x) = 1 - \sum_{j=0}^{k-1} \frac{1}{j!} e^{-\lambda x} (\lambda x)^j$

then θ

Kingman is also desires a "Partition Structure", which he defines as a consistent family of partition structures $(P_1, P_2, \ldots)$ such that when $m < n$, P_m may be obtained from P_n by randomly sampling m things from P_n without replacement. (I need to mention Kingmans view of a potion vs. ours). We can provide this By Hypergeometric Distribution:

$$P(m_1, \ldots, m_n) = \frac{\prod_{i=1}^{n} \binom{n_i}{m_i}}{\binom{n}{k}} = \frac{\prod \frac{n_i!}{m_i!(n_i - m_i)!}}{\frac{n!}{m!}} = \frac{m!}{n!} \prod \frac{n_i!}{m_i!(n_i - m_i)!}$$

So if we start with P_n, Kingman desires $\frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}} P_n \propto P_m$ where $\underline{n} - \underline{m} = \alpha$

we can provide him this:

Let $P_m = \left(\frac{m!}{n!} \prod \frac{1}{(n_i - m_i)!} \right) \partial^{\alpha} P_n$

$$= \frac{(n - |\alpha|)!}{n!} \cdot \frac{1}{\alpha!} \partial^{\alpha} P_n$$

$$= \frac{(n - |\alpha|)!}{n! \, \alpha!} \partial^{\alpha} P_n$$

$$\bigcup_{l \in \mathbb{N}} \mathbb{N}^{\times l} = \bigcup_{l \in \mathbb{N}} \bigcup_{\substack{\phi \in \mathbb{N}^{\times l} \\ i \leq j \Rightarrow \phi_i \leq \phi_j}} G_{\phi}$$

Some Geometric Models

Data is $X = (X^1, \dots, X^n)$

Supp up to & including $j-1$ is $H_{j-1} = \{x^i \mid i \le j-1\}$

Bernoulli Sequence (Not Exchangeable), $\theta > 0$

$$\left.\begin{array}{ll} X^j \notin H_{j-1} & \text{w/ prob. } \frac{\theta}{\theta + j - 1} \\ X^j \in H_{j-1} & \text{w/ prob. } \frac{j-1}{\theta + j - 1} \end{array}\right\} \quad \mathbb{P}_X = \prod_{j=1}^{\#X} \frac{\theta \mathbb{1}(x^j \notin H_{j-1}) + (j-1)\mathbb{1}(x^j \in H_{j-1})}{\theta + j - 1}$$

note: $\mathbb{1}(x^j \notin H_{j-1}) = \prod_{i<j} \mathbb{1}(x^i \neq x^j)$

Weird Bernoulli Sequence (Exchangeable), $\theta > 0$

$$\left.\begin{array}{ll} X^j \notin H_{j-1} & \text{w/ density } \frac{\theta}{\theta + j - 1} \\ X^j \in H_{j-1} & \text{w/ density } \frac{1}{\theta + j - 1} \end{array}\right\} \quad f_X = \prod_{j=1}^{\#X} \frac{\theta \mathbb{1}(x^j \notin H_{j-1}) + \mathbb{1}(x^j \in H_{j-1})}{\theta + j - 1} = \frac{\theta^{|H_{\#X}|}}{\theta \cdot \dots \cdot (\theta + \#X - 1)}$$

So the R.V. is Really $|H_{\#X}| = |\{x^j \mid j = 1, \dots, \#X\}|$ = number of unique observations.

Now, $\sum_{u=1}^{\#X} \frac{\theta^u}{\theta \cdot \dots \cdot (\theta + \#X - 1)} \neq 1$ if $\#X \neq 2$

However, $\sum_{g \in G_{\#X}} \frac{\theta^{|c(g)|}}{\theta \cdot \dots \cdot (\theta + \#X - 1)} = 1$ where $c(g) \vdash n$ is the cycle decomposition of g

$\lambda \vdash n$ iff. $\lambda \in \mathbb{N}^n$ & $\sum_{i=1}^{n} i\lambda_i = n$

So $|\lambda|$ is number of parts aka $|c(g)|$ is number of cycles.

λ_i may be interpreted as "number of unique vals occuring i times in $\#X$"

or

"number of classes of size i"

$\sum_{g \in G_{\#X}} \frac{\theta^{|c(g)|}}{\theta \cdot \dots \cdot (\theta + \#X - 1)}$ Seems like a good Place to start to find the d.f.

$$1 = \sum_{g \in G_{\#X}} \frac{\theta^{|c(g)|}}{\theta \cdot \dots \cdot (\theta + \#X - 1)} = \sum_{u=1}^{\#X} \sum_{\substack{g \in G_{\#X} \\ |c(g)| = u}} \frac{\theta^{|c(g)|}}{\theta \cdot \dots \cdot (\theta + \#X - 1)} = \sum_{u=1}^{\#X} \sum_{\lambda \vdash \#X} \sum_{g \in G_{\#X}} \frac{\theta^{|c(g)|}}{\theta \cdot \dots \cdot (\theta + \#X - 1)} = \sum_{u=1}^{\#X} \sum_{\substack{\lambda \vdash \#X \\ |\lambda| = u}} \frac{|\{g \in G_{\#X} \mid c(g) = \lambda\}| \, \theta^{|\lambda|}}{\theta \cdot \dots \cdot (\theta + \#X - 1)}$$

$$= \sum_{u=1}^{\#X} \sum_{\substack{\lambda \vdash \#X \\ |\lambda| = u}} \frac{\frac{\#X!}{\prod_{i=1}^{\#X} \lambda_i! \, i^{\lambda_i}} \theta^{|\lambda|}}{\theta \cdot \dots \cdot (\theta + \#X - 1)} \quad \text{which gives} \quad \mathbb{P}(|H_{\#X}| = u) = \frac{\sum_{\substack{\lambda \vdash \#X \\ |\lambda| = u}} \theta^u}{\theta \cdot \dots \cdot (\theta + \#X - 1)} = \frac{\left(\frac{\#X!}{\prod_{i=1}^{\#X} \lambda_i! \, i^{\lambda_i}}\right) \theta^{|\lambda|}}{\theta \cdot \dots \cdot (\theta + \#X - 1)}$$

This also gives us a density over Partitions:

$$\mathbb{P}(c(g)) = \frac{\frac{\#X!}{\prod_{i=1}^{\#X} \lambda_i! \, i^{\lambda_i}} \theta^{|\lambda|}}{\theta \cdot \dots \cdot (\theta + \#X - 1)}$$

The last 2 Formulas can be found in Johnson, Kotz, Balakrishnan CH41, I just wanted to derive them from $\sum_{g \in G_{\#X}} \frac{\theta^{|c(g)|}}{\theta \cdot \dots \cdot (\theta + \#X - 1)} = 1$.

A Different Approach

It is much more intuitive to speak about the unique vals and their respective frequencies, e.g. "the sequence _____ occurs __ times in the sample". We can work from this perspective with a k-tople, α, where

$$\alpha = (\alpha_{s_1}, \dots, \alpha_{s_j}, \dots, \alpha_{s_n})$$

(α_{s_j}: frequency of s_j in the sample; slot for sequence s_j)

In Particular, if $X = (X^1, \dots, X^n)$ is our data, a n-tople of sequences, then $\alpha_{s_j} = \sum_{i=1}^{n} 1_{x^i = s_j}$.

This approach should also be compelling because α is the tople of coefficients on basis elements which are indexed by sequences, in notation:

$$\sum_{i=1}^{n} \otimes X^i = \sum_{j=1}^{k} \alpha_{s_j} e_{s_j} \quad \text{and} \quad \left(\sum_{i=1}^{n} \otimes X^i\right)^* = \sum_{j=1}^{k} \alpha_{s_j} \delta^{s_j}$$

If SEQs is the set of observed sequences, ie. the basis of the support of $\sum_{i=1}^{n} \otimes X^i$, the above expressions are: $\sum_{s \in SEQS} \alpha_s e_s$ and $\sum_{s \in SEQS} \alpha_s \delta^s$

note: $SEQS = H_{\#x} = H_n$

Motivation complete.

Returning to the weird bernoulli seq with $X^j \notin H_{j-1}$ w/ density $\frac{\theta}{\theta + \delta - 1}$. $X^j \in H_{j-1}$ w/ density $\frac{1}{\theta + j - 1}$

The density function: $f_X = \frac{\theta^{|H_{\#x}|}}{\theta \cdots (\theta + \#X - 1)} = \frac{\theta^{\#\alpha}}{\theta \cdots (\theta + |\alpha| - 1)}$

The d.f. of k given n (same as before): $\frac{\sum_{\lambda \vdash n, |\lambda| = u} \theta^u}{\theta \cdots (\theta + n - 1)} = \frac{P_{n,u}\, \theta^u}{\theta \cdots (\theta + n - 1)}$ where $P_{n,u}$ is the number of partitions of n with u parts.

The d.f. of α: $P(\alpha) = \frac{|\alpha|!}{\prod_{i=1}^{|\alpha|} \left(\sum_{j=1}^{\#\alpha} 1_{\alpha_j = i}\right)! \; i^{\left(\sum_{j=1}^{\#\alpha} 1_{\alpha_j = i}\right)}} \cdot \frac{\theta^{\#\alpha}}{\theta \cdots (\theta + |\alpha| - 1)}$

$$= \frac{|\alpha|!}{\prod_{i=1}^{|\alpha|} \left(\sum_{j=1}^{\#\alpha} 1_{\alpha_j = i}\right)! \prod_{i=1}^{|\alpha|} i^{\left(\sum_{j=1}^{\#\alpha} 1_{\alpha_j = i}\right)}} \cdot \frac{\theta^{\#\alpha}}{\theta \cdots (\theta + |\alpha| - 1)} = \frac{|\alpha|!}{|Aut(\alpha)| \prod_{j=1}^{\#\alpha} \alpha_j} \cdot \frac{\theta^{\#\alpha}}{\theta \cdots (\theta + |\alpha| - 1)}$$

($\prod_{i=1}^{|\alpha|} \left(\sum_{j=1}^{\#\alpha} 1_{\alpha_j = i}\right)! \; i^{\left(\sum_{j=1}^{\#\alpha} 1_{\alpha_j = i}\right)}$: over values in α, $\prod$ (number of entries in α equal to i)! $= |Aut(\alpha)|$ and $\prod_{i=1}^{|\alpha|} i^{(\text{number of times } i \text{ appears in } \alpha)} = \prod_{j=1}^{\#\alpha} \alpha_j$; over values of α)

$$= \frac{1}{|Aut(\alpha)|} \frac{|\alpha|!}{\theta \cdots (\theta + |\alpha| - 1)} \prod_{j=1}^{\#\alpha} \frac{\theta}{\alpha_j} = \frac{1}{|Aut(\alpha)|} \cdot \frac{|\alpha|!}{\theta \cdots (\theta + |\alpha| - 1)} \Pi\left(\frac{\theta}{\alpha}\right) = \frac{1}{|Aut(\alpha)|} \cdot \frac{\theta + |\alpha|}{|\alpha|^{\theta}} B(\theta, |\alpha| + 1) \Pi\left(\frac{\theta}{\alpha}\right)$$

($\frac{\theta}{\alpha_j}$: just notation; $\Pi\left(\frac{\theta}{\alpha}\right)$: eulerian integral of the second kind; B: Beta Fn.)

The last line is just 3 ways to write the same thing. I will provide more on these as well as some other formulas for the integer partition version. Just so we're all on the same page:

	Data	Integer Partitions	Type Frequencies
Variable used	X	λ	α
Sample Size (n)	$\#X$	$\sum_{i=1}^{\#\lambda} i\lambda_i$	$\lvert\alpha\rvert$
Number of unique elements (k)	$\lvert\{x^i \mid i\in\{1,\dots,\#X\}\}\rvert$	$\lvert\lambda\rvert$	$\#\alpha$
Where variable "Lives", Given n	$\left(\bigoplus_{\ell\in\mathbb{N}} V_B^{\otimes\ell}\right)^{\otimes n}$	$\mathbb{N}^{\times n}$	$\hat{\bigcup_{\ell=1}} \mathbb{N}_s^{\times\ell}$
Where variable "Lives", Given k	$\left(\bigoplus_{\ell\in\mathbb{N}} V_B^{\otimes\ell}\right)^{\otimes n}$	$\bigcup_{n\geq k} \mathbb{N}^{\times n}$	$\mathbb{N}_s^{\times k}$
Where variable "Lives", Given $n \& k$	$\left(\bigoplus_{\ell\in\mathbb{N}} V_B^{\otimes\ell}\right)^{\otimes n}$	$\mathbb{N}^{\times n}$	$\mathbb{N}_s^{\times k}$
Where Variable "Lives", unknown $n \& k$	$\tilde{\bigcup_{n=1}} \left(\bigoplus_{\ell\in\mathbb{N}} V_B^{\otimes\ell}\right)^{\otimes n}$	$\tilde{\bigcup_{n=1}} \mathbb{N}^{\times n}$	$\tilde{\bigcup_{k=1}} \mathbb{N}_s^{\times k}$

... Lots of $\mathbb{N}$.

If we have perfect (i.e. every read is perfect.) sequences, we are in business!

This is not the case.

If we are willing to take $V_{2^B\setminus\{\phi\}}$ as the underlying vector space and let our random measures live in $\bigcup_{\ell\in\mathbb{N}} V_{2^B\setminus\{\phi\}}^{\otimes\ell}$, we are in business. This is not really reflective of actual real-life sequences and doesn't provide a distribution over true sequences :. So this does not seem like the move. (we could provide a connection between $2^B\setminus\{\phi\}$ and B to solve this).

The Problem is: So long as we normalize each observation α will live in $\mathbb{Q}$. Even if we were willing to settle for working with λ, Kingman is of no help here.

The "It is Possible" Solution: Given the observation Parameter $\theta = \begin{pmatrix} \theta_A^1 & \theta_A^{\#} \\ \theta_T^1 & \theta_T^{\#} \end{pmatrix}$ from the begining of this doc, construct the random probability measure as we did, i.e. $\left(\frac{\otimes\theta}{|\theta^1|\cdot\dots\cdot|\theta^{\#}|}\right)^*$ and then multiply it by 12^m where m is the maximum length of a sequence in the sample. We pick 12 because it is $LCM(\{1,2,3,4\}) = LCM(\text{Possible Vals of } |\theta^i|)$. Each observation gets equal weight and now every coefficient is in $\mathbb{N}$ so $\alpha\in\mathbb{N}_s^{\times k}$ and $\lambda\in\mathbb{N}^{\times n}$.

This is computationally exhausting because m is big.

Question: Can our nice formula $\frac{1}{|\mathrm{Aut}(\alpha)|}\cdot\frac{|\alpha|!}{\theta^{\uparrow}:(\theta+|\alpha|-1)}\Pi(\frac{\alpha}{r})$ hold over $\mathbb{Q}$?

note: $|\alpha|$ will always be in $\mathbb{N}$ because $|\alpha|$ is the number of obs.

Intuition: The formula doesn't require anything of the elements of α, so why shouldn't it hold?

Logic: It was derived with the assumption that $\alpha\in\mathbb{N}^{*K}$. In particular From the formula: $\frac{|\alpha|!}{\prod_{i=1}^{|\alpha|}(\sum_{j=1}^{n}1_{\alpha_j=i})!\; i^{(\sum_{j=1}^{n}1_{\alpha_j=i})}}\cdot\frac{\theta^{\#\alpha}}{\theta^{\uparrow}:(\theta+|\alpha|-1)}$

This term may be derived via a counting argument

Note: There might be a theorem or something or someone might know the answer. Do you know / have any ideas? I don't know, but gave it a shot.

Suppose: $\alpha\in\mathbb{Q}_{>}^{*4}$ and that $|\alpha|\in\mathbb{N}_{>}$.

Let: $r=\gcd(\alpha_1,\dots,\alpha_n)$ **notation:** $\alpha^i=\sum_{j=1}^{n}1_{\alpha_j=i}$, $[a:x:b]=\{a,a+x,a+2x,\dots,b\}$

By assumption $\frac{\alpha}{r}\in\mathbb{N}_{>}$.

The term in Question is: $\frac{|\alpha|!}{\prod_{i=1}^{|\alpha|}\alpha^i!\; i^{\alpha^i}}$. This fails over $\mathbb{Q}$ by many counter examples.

The term of interest is: $\frac{|\alpha|!}{\prod_{i}^{\uparrow}\alpha^i!\; i^{\alpha^i}}$, i.e. can we define it over $\mathbb{Q}$?

Pick your product!

Guess: $\frac{|\alpha|!}{\prod_{i\in\mathbb{Q}_{>}}\alpha^i!\; i^{\alpha^i}}=\frac{|\alpha|!}{|\mathrm{Aut}(\alpha)|\prod_{i\in\mathbb{Q}_{>}}i^{\alpha^i}}=\frac{|\alpha|!}{|\mathrm{Aut}(\alpha)|\,\Pi(\alpha)}$

This looks good. How does the result compare with that of $\frac{\alpha}{r}$ and is this correspondence coherent and does it share the behaviours of the definitions over $\mathbb{N}$?

$$\frac{|\frac{\alpha}{r}|!}{\prod_{i\in\mathbb{Q}_{>}}(\frac{\alpha^i}{r})!\; i^{(\frac{\alpha}{r})^i}}=\frac{|\frac{\alpha}{r}|!}{|\mathrm{Aut}(\alpha)|\prod_{i\in\mathbb{Q}_{>}}i^{(\frac{\alpha}{r})^i}}=\frac{|\frac{\alpha}{r}|!}{|\mathrm{Aut}(\alpha)|\,\Pi(\frac{\alpha}{r})}=\frac{(\frac{|\alpha|}{r})(\frac{|\alpha|-r}{r})(\frac{|\alpha|-2r}{r})\cdots(\frac{|\alpha|-(|\alpha|-1)r}{r})}{|\mathrm{Aut}(\alpha)|\;\frac{\Pi(\alpha)}{r^{\#\alpha}}}=\frac{\frac{|\alpha|\cdot(|\alpha|-r)\cdots(|\alpha|-(|\alpha|-1)r)}{r^{|\alpha|}}}{|\mathrm{Aut}(\alpha)|\;\frac{\Pi(\alpha)}{r^{\#\alpha}}}$$

$$=\frac{|\alpha|\cdot(|\alpha|-r)\cdots(|\alpha|-(|\alpha|-1)r)}{|\mathrm{Aut}(\alpha)|\,\Pi(\alpha)\,r^{|\alpha|-\#\alpha}}$$

Check: $\beta\in\mathbb{N}_{>}^{*K}$ $\frac{|r\beta|!}{\prod_{i\in\mathbb{Q}_{>}}(r\beta)^i!\; i^{(r\beta)^i}}=\frac{|r\beta|!}{|\mathrm{Aut}(\beta)|\prod_{i\in\mathbb{Q}_{>}}i^{(r\beta)^i}}=\frac{|r\beta|!}{|\mathrm{Aut}(\beta)|\,\Pi(r\beta)}=\frac{(r|\beta|)!}{|\mathrm{Aut}(\beta)|\,r^{\#\beta}\,\Pi(\beta)}=$ For another day

Following up on the "it is possible way":

Let $\alpha \in \mathbb{Q}_s^{\times n}$ be observed.

So: $\beta := 12^m \alpha \in \mathbb{N}_s^{\times n}$

Then: $\mathbb{P}(\beta) = \frac{|\beta|!}{\prod_{i=1}^{|\beta|} \beta^i! \, i^{\beta^i}} = \frac{|\beta|!}{|\mathrm{Aut}(\beta)| \prod_{i=1}^{|\beta|} i^{\beta^i}} = \frac{|\beta|!}{|\mathrm{Aut}(\beta)| \, \pi(\beta)} = \frac{|12^m \alpha|!}{|\mathrm{Aut}(\alpha)| \, \pi(12^m \alpha)}$

$= \frac{|12^m \alpha|!}{|\mathrm{Aut}(\alpha)| \, 12^{mk} \pi(\alpha)} = \frac{\prod_{i=1}^{|12^m \alpha|} i}{|\mathrm{Aut}(\alpha)| \, 12^{mk} \pi(\alpha)} = \frac{\prod_{i=1}^{12^m n} i}{|\mathrm{Aut}(\alpha)| \, 12^{mk} \pi(\alpha)} = \Pi$

Cluster Coefficient θ Model:

$n = 61{,}065{,}458$ genetic observations $= |\alpha|$

$k = 2{,}805{,}095 = \#\alpha$

$\max(\alpha) = 2183654$

Maybe Let:

$\Delta = \mathbb{Q}_{>0}^{\times 4}$ gives us $\sum_{\lambda \in \Delta} c_\lambda x^\lambda$