Overview of SARS-Cov-2 RNA Sequences Data Analysis on Tree Spaces

STA 6557
Final Project

Omar Alharthi

Department of Statistics
Florida State University

April 14, 2022

Outline

1. Introduction
2. Motivation
3. Stratified Spaces
4. Tree-Building Methods
5. References

Introduction

What is the difference between SARS-CoV-2 and Covid-19?

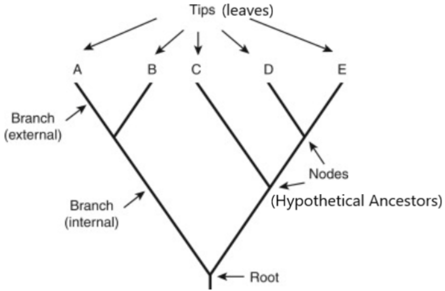
- **SARS-CoV-2** is the virus.
  - stands for **S**evere **A**cute **R**espiratory **S**yndrome **Co**rona**v**irus-**2**.
  - named by The International Committee on Taxonomy of Viruses (ICTV).

- **Covid-19** is the disease.
  - stands for **Co**rona**vi**rus **d**isease 20**19**.
  - named by World Health Organization (WHO).

Introduction

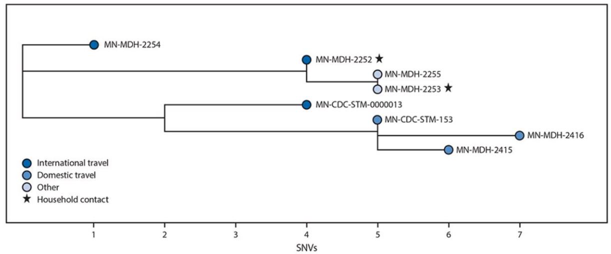
What is a Phylogenetic Tree?

- A phylogenetic tree, also known as a phylogeny, is a diagram that depicts the lines of evolutionary descent of different species, organisms, or genes from a common ancestor.



Motivation

- Trees represent various types of hierarchical relationships between species, organisms, and genes from a common ancestor.
- Phylogenetic analysis is important for clarifying the evolutionary pattern of multigene familiesunderstanding the process of adaptive evolution at the molecular level.
- Deoxyribonucleic acid (DNA) can be used to draw a phylogenetic tree.



Abbreviation: SNV = single nucleotide variant.

Fig.1. Phylogenetic tree shows genetic distance between SARS-Cov-2 specimens (n=8) and exposures histories.

Stratified Space

Stratified Space (Space With A Manifold Stratification) is a metric space $\mathcal{M}$ that admits a filtration $\emptyset = F_{-1} \subseteq F_0 \subseteq F_1 \cdots \subseteq F_n \subset \cdots = M = \cup_i F_i$, By closed subspaces, such that for each $i = 1, \ldots, n, F_i/F_{i-1}$ is empty or is an i-dimensional manifold, called the i-th stratum.
Examples of stratified sample spaces, which are not themselves manifolds include similarity shape spaces (Kendall et.al.(1999)), affine shape spaces (Groisser & Tagare (2009)) and projective shape spaces (Mardia & Patrangenaru (2005)). Spaces of positive semi-definite matrices, which arise as data points in Diffusion Tensor Imaging (Schwartzman et.al. (2008)), and tree spaces (Billera et.al. (2001), Wang & Marron(2007)), are additional examples of stratified sample spaces.

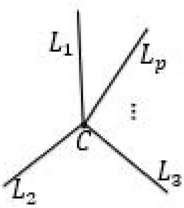
Stratified Space

- **Sticky C.L.T. on Spiders**

To define a Spider, consider an arbitrary nonempty set $K \geq 3$ and, for each of its elements, i, define the ray(leg) $L_i = \{(i, x) : x \in [0, \infty)\}$. The Spider is formed by joining the rays together at the origin.

$$S_K = \{(i, x) : i \in K, x \in [0, \infty)\}$$

where $(i, 0), \ldots, (j, 0)$ for all $i, j \in K$, the equivalence class of all points of the form $(i, 0)$, we denote by 0, named center C.



Stratified Space

- **Sticky C.L.T. on Spiders**

Assume $X_i, i = 1, \ldots, n$ are i.i.d. random objects on a spider S_K, having legs $L_i, i = 1, \ldots, K$ and center C. Further, assume the intrinsic mean μ_I exists and the intrinsic variance is finite. Any probability measure Q on S_K decomposes uniquely as a weighted sum of probability measures Q_K on the legs L_K and an atom Q_0 at C. More precisely, there are nonnegative real numbers $\{w_k\}_{k=0}^{p}$ summing to 1 such that, for any Borel set $A \subseteq S_p$, the measure Q takes the value

$$Q(A) = w_0 Q_0(A \cap C) + \sum_{k=1}^{p} w_k Q_k(A \cap L_k).$$

Stratified Space

Assume $w_0 = 0$ and $x \in L_a$, the Fréchet function is defined as follows,

$$\begin{aligned}
F(x) &= \sum_{i=1, i\neq a}^{K} \int_0^\infty (x+u)^2 w_i Q_i(du) + \int_0^\infty (x-u)^2 w_a Q_a(du) \\
&= x^2 \sum_{i=1}^{K} \int_0^\infty w_i Q_i(du) + 2x[\sum_{i=1, i\neq a}^{K} \int_0^\infty u w_i Q_i(du) \\
&- \int_0^\infty u w_a Q_a(du)] + \sum_{i=1}^{K} \int_0^\infty u^2 w_i Q_i(du) \\
&= x^2 + 2[\sum_{i=1, i\neq a}^{K} v_i - v_a]x + const.
\end{aligned}$$

where $v_i = \int_0^\infty u w_i Q_i(du)$.

Stratified Space

If there exists an unique minimizer for the Fréchet function $F(x)$, the minimizer is called intrinsic mean μ_I (based on the intrinsic distance). The minimizer of the quadratic form is $x^* = v_a - \sum_{i \neq a} v_i$, where $x \in L_a = \{(a, u) : u \in [0, \infty)\}$. Thus, we have three situations:(i)$v_a - \sum_{i \neq a} v_i > 0$ or $v_a > \sum_{i \neq a} v_i$, (ii)$v_a - \sum_{i \neq a} v_i = 0$ or $v_a = \sum_{i \neq a} v_i$ and iii $v_a - \sum_{i \neq a} v_i < 0$ or $v_a < \sum_{i \neq a} v_i$. In case(i), we have μ_I well defined on L_a, then classical C.L.T is applied. In case (ii), we can fold other legs into that half line opposite to L_a then apply C.L.T. and since the negative part is undefined, so the result goes to a positive truncated normal distribution. In case (iii), for any $a \in \{1, \ldots, K\}$, we have $\mu_I = C$, which shows that intrinsic mean μ_I sticks to the center C.

Stratified Space

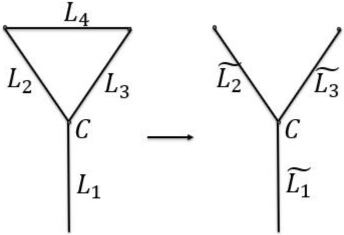
Theorem

1. $v_a > \sum_{i \neq a} v_i$ *for some (unique)* $a \in \{1, \ldots, K\}$, *then* $\mu_I \in L_a$ *and for* n *large enough* $\bar{X}_n \in L_a$ *and* $\sqrt{n}(\bar{X}_n - \mu_I)$ *has asymptotically a normal distribution.*
2. $v_a = \sum_{i \neq a} v_i$ *for some (unique)* $a \in \{1, \ldots, K\}$, *then after folding the legs* $L_i, i \neq a$, *into one half line opposite to* L_a, $\sqrt{n}(\bar{X}_n - \mu_I)$ *has asymptotically a positive truncated normal distribution.*
3. $v_a < \sum_{i \neq a} v_i$ *for all* $a \in \{1, \ldots, K\}$, *then* $/mu_I = C$ *and there is* n_0 *s.t.* $\forall n \geq n_0$, *then* $\bar{X}_n = 0$ *a.s.*

Stratified Space

- **Sticky C.L.T. on Piece-wise Linear Stratified Spaces**

Consider the simplest graph G, which is formed from a spider S_3 with finite legs and connected two legs. Assume $X_i, i = 1, \dots, n$ are i.i.d. r. o's. on G. Denote a weighted sum of probability measures Q_k on the legs L_k and an atom Q_0 at C. To build Fréchet function, we could cut L_4 at a point then bend the two pieces to form new legs $\tilde{L_2}$ and $\tilde{L_3}$, and then rescale legs to unit length. Thus, $X_i, i = 1, \dots, n$ are i.i.d. random objects on G with probability measures $\tilde{Q}_k$ on the legs $\tilde{L_k}$, $\{\tilde{w}_k\}_{k=0}^{p}$ summing to 1.



Stratified Space

Therefore, the problem becomes computing intrinsic mean on unit-length spider S_3. Assume $x \in L_a$, the Fréchet function is defined as follows,

$$
\begin{aligned}
F(x) &= \sum_{i=1, i\neq a}^{3} \int_0^1 (x+u)^2 \tilde{w}_i \tilde{Q}_i(du) + \int_0^1 (x-u)^2 \tilde{w}_a \tilde{Q}_a(du) \\
&= x^2 + 2[\sum_{i=1, i\neq a}^{3} \tilde{v}_i - \tilde{v}_a]x + const.
\end{aligned}
$$

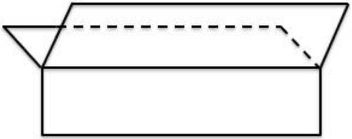
where $\tilde{v}_i = \int_0^1 u \tilde{w}_i \tilde{Q}_i(du)$.
The intrinsic mean μ_I on the graph G would follow a sticky C.L.T., if certain inequalities hold true.

Stratified Space

- **Open Book**

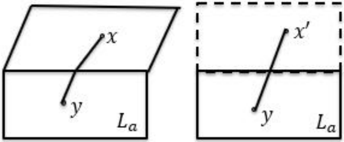
Consider a 2-dimension Spider, which is called Open Book. The Open Book is formed by joining half spaces(leaves), which are defined as $L_i = \{(i, x_1, x_2) : x_1, x_2 \in [0, \infty)\} i = 1, \ldots, K$, together at the spine $S = [0, \infty)$.

$$O_K = \{(i, x_1, x_2) : i \in K, x_1, x_2 \in [0, \infty)\}.$$



Stratified Space

Before discussing the Fréchet function and stickiness of the intrinsic mean, one have to define the distance $d(\mathbf{x}, \mathbf{y})$. If two points $\mathbf{x} = (x_1, x_2), \mathbf{y} = (y_1, y_2)$ are on the same leaf, the distance between two points $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$. If two points $\mathbf{x}, \mathbf{y}$ are on the different leaves, one could replace point $\mathbf{x}$ by $\mathbf{x'}$ onto the half space opposite to L_a, then the distance is defined as $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x'} - \mathbf{y}\|$, where $\mathbf{x'} = (x_1, -x_2)$.



Stratified Space

Assume $X_i, i = 1, \ldots, n$ are i.i.d. random objects on O_k. Denote a weighted sum of probability measure Q_k on the legs L_k and an atom Q_0 at C. Further, assume the intrinsic mean μ_I exists and the intrinsic variance is finite. W.L.O.G., assume $w_0 = 0$ and $x \in L_a$, the Fréchet function is defined as follows,

$$F(\mathbf{x}) = \sum_{i=1, i \neq a}^{K} \int_{O_K} \|\mathbf{x'} - \mathbf{u}\|^2 w_i Q_i(d\mathbf{u}) + \int_{O_K} \|\mathbf{x} - \mathbf{u}\|^2 w_a Q_a(d\mathbf{u})$$

Stratified Space

$$
\begin{aligned}
F(\mathbf{x}) &= \|\mathbf{x}\|^2 - \sum_{i \neq a} 2 \int_o^\infty \langle \mathbf{x'}, \mathbf{u} \rangle w_i Q_i(d\mathbf{u}) - 2 \int_o^\infty \langle \mathbf{x}, \mathbf{u} \rangle w_a Q_a(d\mathbf{u}) \\
&+ \sum_{i=1}^{K} \int_0^\infty \|\mathbf{u}\|^2 w_i Q_i(d\mathbf{u}) \\
&= x_1^2 - 2[\sum_{i=1}^{K} \int_0^\infty u_1 w_i Q_i^{(1)}(du_1)]x_1 + x_2^2 - 2[\int_0^\infty u_2 w_a Q_a^{(2)}(du_2) \\
&- \sum_{i \neq a} \int_0^\infty u_2 w_i Q_i^{(2)}(du_2)]x_2 + const. \\
&= x_1^2 - 2 \sum_{i=1}^{K} v_i^{(1)} x_1 + x_2^2 - 2[v_a^{(2)} - sum_{i \neq a} v_i^{(2)}]x_2 + const.
\end{aligned}
$$

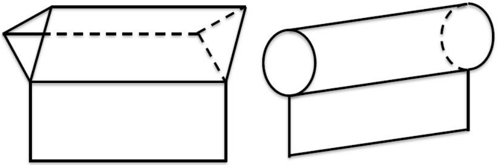
where $v_i^{(j)} = \int_0^\infty u w_i Q_i^{(j)}(du)$.

Stratified Space

To minimize $F(\mathbf{x})$ is to minimize the quadratic form for x_1, x_2 and the solution is $x_1^* = \sum_{i=1}^{K} v_i^{(1)}, x_2^* = v_a^{(2)} - \sum_{i \neq a} v_i^{(2)}$. Since $v_i^{(j)} \geq 0, x_1^*$ is always non-negative. Thus, one could apply C.L.T. on x_1^* . However, for x_2^*, one have to discuss the three situations as did for the spider spaces. Therefore, in this case, if one separate the space into two directions, the stickiness of intrinsic mean would only occur on one direction, which is in a lower dimensional space. If for all a, $v_a^{(2)} < \sum_{i \neq a} v_i^{(2)}$, the intrinsic mean would stick to the spine S, but still has one direction free, mathematically, the intrinsic mean would follow a univariate normal distribution on the spine S.

Stratified Space

Also, we could consider 2D version graphs(see figure below). To discuss the stickiness, we could cut along a line on the top surface, then bend and re-scale to form an open book structure with boundaries. Therefore, the intrinsic mean would stick to the spine if for all a, $v_a^{(2)} < \sum_{i \neq a} v_i^{(2)}$.



Stratified Space

Consider a piece-wise linear stratified space M_K, which is formed by joining n-dimensional half space $L_i = \{(i, x_1, \ldots, x_n) : x_1, \ldots, x_n \in [0, \infty)\} i = 1, \ldots, K$, together at a m-dimensional space $(m < n)$. Let $d = n - m$, the space M_K is defined as

$$M_K = \{(i_1, \ldots, i_d, x_1, \ldots, x_n) : i_1, \ldots, i_d \in K, x_1, \ldots, x_n \in [0, \infty)\}.$$

Stratified Space

Assume $x \in L_a$, the Fréchet function is defined as

$$F(x) = \sum_{k=d+1}^{n} [x_k^2 - 2\sum_{i=1}^{K} v_i^{(k)} x_k] + \sum_{l=1}^{d} [x_l^2 - 2[v_a^{(l)} - sum_{i \neq a} v_i^{(l)}] x_l] + const.$$

Set $E_l = \{$ for some $a \in \{1, \ldots, K\}, v_a^{(l)} \geq \sum_{i \neq a} v_i^{(l)}\}$, then $E_l^c = \{$ for all $a \in \{1, \ldots, K\}, v_a^{(l)} < \sum_{i \neq a} v_i^{(l)}\}, l = 1, \ldots, d$.

1. For all $l = 1, \ldots, d$, E_l occurs, for n large enough, $\bar{X}_n \in L_a$ and $\sqrt{n}(\bar{X}_n - \mu_l) \to N_+$, where $N_+ = max(0, N(0, \Sigma))$.
2. For some $l = 1, \ldots, d, E_l^c$ occurs, μ_i would stick to a lower dimensional spine space and would follow a multivariate normal distribution on the low dimensional spine space.

Tree-Building Methods

Commonly used methods are classified into three major groups:

- Distance Matrix Methods.
    - Unweighted pair-group method using arithmetic averages (UPGMA).
    - Least Squares (LS) Methods (Ordinary - Weighted).
    - Minimum Evolution (ME) Method.
    - Neighbor Joining (NJ) Method.
- Maximum Parsimony (MP) Methods.
    - Unweighted MP.
    - Weighted MP.
- Maximum Likelihood Methods.

References-1

[1] Padron-Regalado, E. (2020, April 23). *Vaccines for SARS-CoV-2: Lessons from Other Coronavirus Strains*. Infectious diseases and therapy. https://pubmed.ncbi.nlm.nih.gov/32328406/.

[2] WHO (2020, March 26). *Origin of SARS-CoV-2*. https://www.who.int/health-topics/coronavirus/origins-of-the-virus

[3] Feragen, A., Nielsen, M., Jensen, E. B. V., du Plessis, A., & Lauze, F. (2014). Geometry and statistics: Manifolds and stratified spaces. *Journal of Mathematical Imaging and Vision*, 50(1), 1-4.

[4] Lau, S. K., Luk, H. K., Wong, A. C., Li, K. S., Zhu, L., He, Z., ... & Woo, P. C. (2020). Possible bat origin of severe acute respiratory syndrome coronavirus 2. *Emerging infectious diseases*, 26(7), 1542.

References-2

[1] E. Miller, M. Owen and J.S. Provan. Polyhedral computational geomtry for averaging metric phylogenetic trees. *Advances in Applied Mathematics*, 68:51-91, 2015.

[2] K. Sturm. Probability measures on metric spaces of nonpositive curvature Heat kernels and analysis on manifolds, graphs, and metric spaces: Lecture notes from a quarter program on heat kernels, random walks, and analysis on manifolds and graphs. *Contemporary Mathematics*, 338 (2003), 357â€".

[3] Centers for Disease Control and Prevention. (2021, April 2). *About Variants of the Virus that Causes COVID-19*. https://www.cdc.gov/coronavirus/2019-ncov/transmission/variant.html.

References-3

[1] Grubaugh, N. D., Petrone, M. E., & Holmes, E. C. (2020). We shouldn't worry when a virus mutates during disease outbreaks. *Nature microbiology*, 5(4), 529-530.

[2] Cristianini, N., & Hahn, M. W. (2006). Introduction to computational genomics: a case studies approach. *Cambridge University Press*.

[3] Sudhir Kumar, Glen Stecher, Michael Li, Christina Knyaz, and Koichiro Tamura (2018) MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Molecular Biology and Evolution* 35:1547-1549.

[4] Basrak, Bojan(2010). Limit theorems for the inductive mean on metric trees. *J. Appl. Probab.* **47**, no. 4, 1136–1149

References-4

[1] Bhattacharya, Rabi N.; Buibas, Marius; Dryden, Ian L.; Ellingson, Leif A.; Groisser, David; Hendriks, Harrie; Huckemann, Stephan; Le, Huiling; Liu, Xiuwen; Marron, James S.; Osborne, Daniel E.; Patrangenaru, Vic; Schwartzman, Armin; Thompson, Hilary W.; Wood, Andrew T. A.(2013) Extrinsic data analysis on sample spaces with a manifold stratification. *Advances in mathematics,*241-251, Ed. Acad. Romane, Bucharest.

[2] Thomas Hotz, Stephan Huckemann, Huiling Le, James S. Marron, Jonathan C. Mattingly, Ezra Miller, James Nolen, Megan Owen, Vic Patrangenaru and Sean Skwerer (2013). Sticky Central Limit Theorems on Open Books. *Annals of Applied Probability,* **23**, 2238–2258.

[3] Billera, Louis J.; Holmes, Susan P.; Vogtmann, Karen(2001). Geometry of the space of phylogenetic trees. *Adv. in Appl. Math.* **27**, no. 4, 733–767.

References-5

[1] Building a phylogenetic tree (phylogeny article). Khan Academy Available at: https://www.khanacademy.org/science/biology/her/tree-of-life/a/building-an-evolutionary-tree. (Accessed: 13th April 2020)

[2] L. Ellingson, V. Patrangenaru, H. Hendriks, P. S. Valentin (2015). CLT on Low Dimensional Stratified Spaces. *Topics in Nonparametric Statistics. Editors: M.G. Akritas, S.N. Lahiri and D. N. Politis*, 227–240. Springer.

[3] Embl-Ebi (n.d.). *What is Phylogenetics?* https://www.ebi.ac.uk/training/online/course/introduction-phylogenetics/what-phylogeny/aspects-phylogenies/.

[4] Educative (n.d.). *What is a graph (data structure)?* https://www.educative.io/edpresso/what-is-a-graph-data-structure.

References-6

[1] Patrangenaru, V., & Ellingson, L. (2015). Nonparametric statistics on manifolds and their applications to object data analysis. *CRC Press*.

[2] Hall, B. G. (2011). Phylogenetic trees made easy: A how to manual (No. 576.88 H174p). Sinauer,.

[3] Page, R. D., & Holmes, E. C. (2009). Molecular evolution: a phylogenetic approach. John Wiley & Sons.

[4] Evolution.berkeley.edu. (n.d.). *Understanding phylogenies; Phylogenetic Trees.* https://evolution.berkeley.edu/evolibrary/article/evo_05.

[5] Huson, D. H., Rupp, R., & Scornavacca, C. (2010). Phylogenetic networks: concepts, algorithms and applications. Cambridge University Press.

Thank you!