Bioinformatics, 2023, pp. 1–5

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Application Note

APPLICATION NOTE

pikun: An Information Theoretic Toolkit for Analysis and Visualization of Species Delimitation Metric Space

Jeet Sukumaran[1,*]

[1]Biology, San Diego State University, Street, Postcode, State, Country
[*]Corresponding author. jsukumaran@sdsu.edu
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Summary `pikun` is a Python package for analyzing and visualizing species delimitation models in an information theoretic framework that, in addition to classic measures of information such as the entropy and mutual information [16], provides for the calculation of the variance of information criterion [11], a true metric or distance function for species delimitation models that is aligned with the lattice of partitions. The metric as well as other more specialized statistical properties of the variance of information distance statistic calculated here that provide insightful interpretive benefits as well as computational ones when used to understand species delimitation model space.
Availability and implementation `pikun` is available under the MIT license from its public repository: https://github.com/jeetsukumaran/pikun. `pikun` is written in pure Python [20], and depends on the following libraries: "NumPy" [7], "SciPy" [22], "PANDAS" [19, 10], "Matplotlib"[8], "Seaborn"[23].

Key words: species delimitation, information theory, statistics, modeling

Introduction

The field of species delimitation, i.e., the identification of species units in nature [18], has made major advances, including extending machine learning approaches [15] to incorporate natural history [3], hierarchically modeling the speciation process over a primary population-level structuring processes [17], or modeling other biological criteria [4]. However, despite these advances, the field lacks a true metric space to compare and relate various species delimitation models. Some useful indexes have been described previously, such as the taxonomic index of congruence (C_{tax}) of [14] or the match ratio of [1] (implemented in [5]), but these are not true metrics (e.g., having the property of triangle inequality [2]) and some are only defined for cases in which one species delimitation model is nested within the other [14], which makes intuitive interpretation of the distances challenging when more than two species delimitation models are involved, as well as less ideal in supporting inference or optimization algorithms in machine learning [11].

pikun: Information Theoretic Species Delimitation Metrics

This paper describes `pikun`[1], a Python package for the analysis and visualization of species delimitation models in an information theoretic framework that provides a true distance or metric space for these models. The species delimitation models being analyzed may be generated by any inference package, such as BPP [6], SNAPP [9], DELINEATE [17] etc., or constructed based on taxonomies or classifications based on conceptual descriptions in literature, geography, folk taxonomies, etc. Regardless of source or basis, each species delimitation model can be considered a *partition* of taxa or lineages [17], and thus can be represented as a simple JSON list of lists, or otherwise in a dedicated and widely-supported data exchange format for species delimitation models, "`SPART-XML`" [13], both of which `pikun` accepts as input formats. Further specialized input formats include the comprehensive results file from a DELINEATE [17] analysis, in which case it is able to include the extra information that this approach provides (e.g., the conditional and unconditional probabilities of different species delimitation models under various constraints).

`pikun` provides a range of univariate information theory statistics for each individual model in the input set (e.g., the entropy [16]), as well as bivariate statistics (e.g., the mutual information, joint entropy, [16]) for each distinct pair of these models. In addition, `pikun` reports distances between every pair of species delimitation models based on these information theoretic measures: the variance of information [11] and the normalized joint variation of information distance [21]. These distances are *true* metrics

[1] "sparrowhawk" in the Kumeyaay language. San Diego State University is built on Kumeyaay land.

© The Author 2023. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

on the species distribution models that have additional properties that are useful both in interpretation of these distances and in computational applications.

The Variance of Information Partition Distance

Every species delimitation model can be represented as a *partition* of a set of lineages into a set of mutually exclusive and jointly comprehensive subsets [13]. As such, the *variance of information* criterion of [11] which provides a true distance function for partitions, also establishes a metric space for species delimitation models. The variance of information of two partitions, ψ^u, ψ^v, $VI(\psi^u, \psi^v)$, is defined as [11]:

$$VI(\psi^u, \psi^v) = H(\psi^u) + H(\psi^v) - 2I(\psi^u, \psi^v). \tag{1}$$

$H(\psi)$ is the entropy of partition ψ which divides n elements into K subsets, with the kth subset having n_k elements, and is given by [11]:

$$H(\psi) = -\sum_{k=1}^{K} p(k) \log p(k), \tag{2}$$

where $p(k)$ is the probability of subset or cluster k, which is given in this approach by the cardinality of the subset as a proportion of the entire set: $p(k) = \frac{n_k}{n}$.

$I(\psi^u, \psi^v)$, on the other hand, is the mutual information of partitions ψ^u and ψ^v, and is given by:

$$I(\psi^u, \psi^v) = \sum_{k_i=1}^{K_i} \sum_{k_j=1}^{K_j} p(k_i, k_j) \log \frac{p(k_i, k_j)}{p(k_i)p(k_j)}, \tag{3}$$

where $p(k_i, k_j)$ is the joint probability of subset $\psi^u_{k_i}$ from partition ψ^u and subset $\psi^v_{k_j}$ from partition ψ^v. This is given by the proportion of the size of the intersection of subsets to the entire dataset:

$$p(k_i, k_j) = \frac{|\psi^u_i \cap \psi^v_j|}{n}. \tag{4}$$

Usefulness as a Species Delimitation Partition Metric

The information theoretic-based distances provided by `pikun` are very flexible. For example, different organizations of different populations into sets of species may be analyzed together from multiple different inferences or publications, even if the numbers of individuals, populations, or genes sampled across these sources vary. Furthermore, there are no constraints on the relationships between the partitions considered, whether they are nested, hierarchical, parallel, incompatible, or have different levels of structuring in the inference models. Detailed discussion of the statistical characteristics and properties of this statistic are given in [11]. Here I present a conceptual description of the properties of this metric that make it useful for analysis of differences between species delimitation models.

A true metric or distance function, such as the variance of information, has the properties of non-negativity, symmetry, and the triangle equality [2]. These properties are useful in aligning with human intuition when interpreting values, as well as preferable in statistical or computational terms due to benefits in algorithm and data structure design or scaling up comparisons [11].

The benefits of the variance of information as a distance for species delimitation partitions, however, go beyond its metric properties. In addition to being a true metric, the variance of information is aligned with the lattice of the set of partitions. Formally, a partition lattice is an ordered set of partitions, where the order is defined by the refinement of the partitions. We can represent the set of all partitions of a particular set of elements as a *partition lattice* using the refinement order (a partition U is defined as a refinement of a partition V if every block in U is a subset of a block in V: we say that U is finer than V, or, equivalently, U is coarser than V). Conceptually, a partition lattice provides geometrical representation of all possible divisions of a set of elements, from the most granular, where each element forms its own subset, to the most general, where all elements belong to a single subset. This allows for the understanding of relationships between different partitions, offering insights into how small changes in groupings can lead to new partitions, and how these partitions are (or are not) nested within each other. A metric on partitions that is actually aligned along the lattice of partition sets has a number of advantages for interpretation of the similarities and dissimilarities between different partitions, as well as for computation, as neighbors in the metric space correspond to refinements in the partition lattice [12]. Such a metric takes into account the hierarchical nature of partitions, recognizing that partitions related through a series of refinements are closer to each other than partitions that differ more fundamentally. That is, under this metric, the nearest neighbor of a species delimitation partition can be obtained by "lumping" the two smallest "species" units of that partition, and vice versa via "splitting". This alignment with the partition lattice is not only useful for intuition, but may also have potential to provide benefits in designing inference or stochastic sampling algorithms (e.g., MCMC moves in partition space).

Another useful property of this metric is n-invariance, i.e., the distance between two partitions is independent of the absolute number of elements [11]. This allows comparisons as well as combinations of distances of partitions of data sets of different sizes. For example, two independent studies of the system may report distances between species delimitation models, with the first study organizing dozens of lineages into species units while the other may involve hundreds or even greater. Despite this disparity in the number of elements being organized into higher-level groups or clusters, the distances between these models can be compared, as the critical quantity in is the relative proportion of the entire respective sets represented by the nominal species units in each partition [11]. As described by [11], the distances can in fact be *combined* (added or subtracted) if it makes sense to do so, regardless of data set size, which also can be seen to provide potential benefits for optimization or inference algorithms as well.

The Normalized Variance of Information Partition Distance

The upper bound of this metric is given by $2 \log K$, where K is the maximum number of subsets in a partition, or, equivalently, $log(n)$, where n is the number of elements [11]. In some applications, it might be useful to have this normalized to the range $[0, 1]$. Unfortunately, while it is straight-forward enough to normalize by these known upper bounds, this will preclude comparison of distances across different datasets (in particular, datasets of different sizes), which, as I discussed above, is a desirable property. Instead, `pikun` provides for the calculation of the normalized variance of information following [21], which retains the metric properties of the variance of information.

Example: Exploratory Discovery Analysis

`pikun` was run on the results of a DELINEATE inference on a dataset of *Lionepha* beetles available on the DELINEATE website). The analysis assigned known species identities ("constraints") to all but 13 of the 44 putative population lineages. The constraints left 22,436 possible partitions of the population lineages that would not result in species blocks that violated the constraints, and the DELINEATE analysis reports each of these and their associated probabilities, listed in descending order of probabilities. To map all these partitions into a metric space would require 251687048 distinct comparisons. While with advances in computing this may be possible within our lifetimes, a more restricted set of partitions that excluded the long tail of those with low probabilities would probably just as insightful for understanding the system. `pikun` provides the option to limit the analysis to a specified number of partitions, and this was used to select the 1000 most probable partitions, collectively constituting approximately 0.85 in cumulative probability.

Figure 1 is a visualization of the results of the `pikun` distance analysis that illustrates some of the advantages of the variation of information criterion as a distance metric on species delimitation models. This shows a heatmap of the distances between partitions, with partitions listed on the x and y axes in terms of their binned log-scaled probabilities, and the normalized variance of distance between the partitions as the "heat" or the "elevation" indicated by the shade, with blue regions representing shorter distances and red larger. As can be seen, the partitions with mutual high probabilities (i.e., the upper-right area of the plot) have relatively smaller distances to each other compared to partitions with mutual lower probabilities (i.e., the lower-left area of the plot). As discussed above, one of the properties of the VI metric is its alignment to the lattice of partitions, which, among other things results in partitions that are refinements of each other being closer to each other. Based on this we can surmise that, in this particular analysis, the species delimitation partition models with higher probabilities are generally disagreeing on the finer-scale splitting of species, in contrast to the lower probability models, which may be conflicting in more fundamental ways (e.g., non-nested). Conversely, one can imagine other datasets might result in regions of species delimitation partition space where, conversely, clusters of strongly supported models had deeper or more fundamental differences from one another. This would serve as the motivation or bases for more directed statistical analysis at the identified regions of partition space of interest. This sort of interpretive potential demonstrates some of the particular advantages of the variance of information as a metric on species delimitation models.

Acknowledgements

- NSF-DEB 1937725 "COLLABORATIVE RESEARCH: Phylogenomics, spatial phylogenetics and conservation prioritization in trapdoor spiders (and kin) of the California Floristic Province".

References

1. Dirk Ahrens, Tomochika Fujisawa, Hans-Joachim Krammer, Jonas Eberle, Silvia Fabrizi, and Alfried P. Vogler. Rarity and Incomplete Sampling in DNA-Based Species Delimitation. *Systematic Biology*, 65(3):478–494, May 2016.
2. Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. 2006.
3. Shahan Derkarabetian, James Starrett, and Marshal Hedin. Using natural history to guide supervised machine learning for cryptic species delimitation with genetic data. *Frontiers in Zoology*, 19(1):8, February 2022.
4. Jordan Douglas and Remco Bouckaert. Quantitatively defining species boundaries with more efficiency and more biological realism. *Communications Biology*, 5(1):755, 2022. Publisher: Nature Publishing Group UK London.
5. Jacques Ducasse, Visotheary Ung, Guillaume Lecointre, and Aurélien Miralles. LIMES: a tool for comparing species partition. *Bioinformatics (Oxford, England)*, 36(7):2282–2283, April 2020.
6. Tomáš Flouri, Xiyun Jiao, Bruce Rannala, and Ziheng Yang. Species tree inference with BPP using genomic sequences and the multispecies coalescent. *Molecular biology and evolution*, 35(10):2585–2593, 2018. Publisher: Oxford University Press.

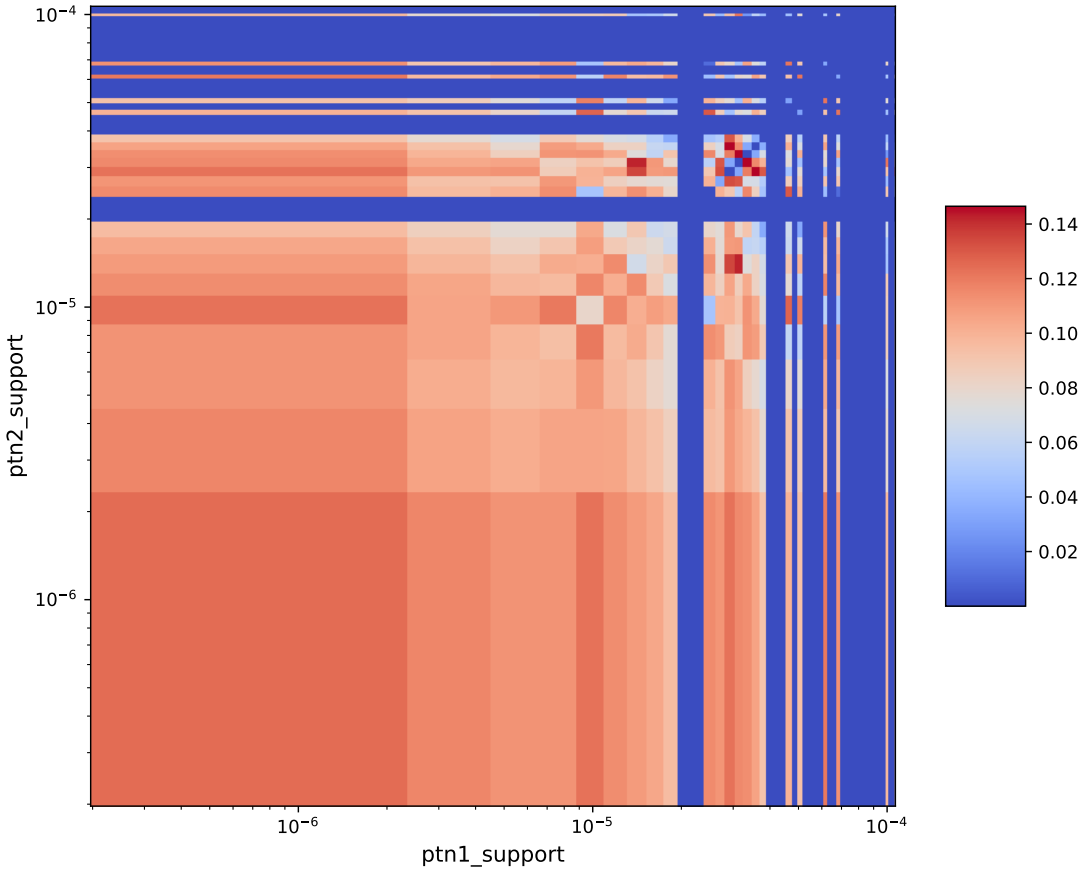

Fig. 1. Heat map showing the distances between different species delimitation models under the variance of information criterion [11] in relation to their support (probability). The x- and y-axes represent partitions ordered by their log-scaled probabilities. The color gradient represents the distance between partitions, with blue indicating closer proximity (smaller distance) and red indicating greater separation (larger distance). Note how partitions with mutual high probabilities (upper-right area) are more similar to each other than to partitions with mutual lower probabilities (lower-left area). The alignment of the variance of information metric with the lattice of partitions means that partitions that are nestings or refinements of each other have relatively smaller distances, and this allows us to interpret this plot to suggest that, in this particular analysis, the high-probability species delimitation models tend to differ in terms of how the resolve the granularity of some of the species blocks, rather than broader or more fundamental disagreements.

7. Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. Publisher: Springer Science and Business Media LLC.
8. J. D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. Publisher: IEEE COMPUTER SOC.
9. Adam D. Leaché and Remco R. Bouckaert. Species trees and species delimitation with SNAPP: a tutorial and worked example. In *Workshop on Population and Speciation Genomics, Český Krumlov*, 2018.
10. Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010.
11. Marina Meilă. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, May 2007.
12. Marina Meilă. Comparing clusterings: an axiomatic view. In *Proceedings of the 22nd international conference on Machine learning*, ICML '05, pages 577–584, New York, NY, USA, August 2005. Association for Computing Machinery.
13. Aurélien Miralles, Jacques Ducasse, Sophie Brouillet, Tomas Flouri, Tomochika Fujisawa, Paschalia Kapli, L. Lacey Knowles, Sangeeta Kumari, Alexandros Stamatakis, Jeet Sukumaran, Sarah Lutteropp, Miguel Vences, and Nicolas Puillandre. SPART: A

versatile and standardized data exchange format for species partition information. *Molecular Ecology Resources*, 22(1):430–438, 2022. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.13470.

14. Aurélien Miralles and Miguel Vences. New Metrics for Comparison of Taxonomies Reveal Striking Discrepancies among Species Delimitation Methods in Madascincus Lizards. *PLOS ONE*, 8(7):e68242, July 2013. Publisher: Public Library of Science.
15. Jingwen Pei, Chong Chu, Xin Li, Bin Lu, and Yufeng Wu. CLADES: A classification-based machine learning method for species delimitation from population genetic data. *Molecular ecology resources*, 18(5):1144–1156, 2018. Publisher: Wiley Online Library.
16. C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948. Conference Name: The Bell System Technical Journal.
17. Jeet Sukumaran, Mark T. Holder, and L. Lacey Knowles. Incorporating the speciation process into species delimitation. *PLOS Computational Biology*, 17(5):e1008924, May 2021. Publisher: Public Library of Science.
18. Jeet Sukumaran and L. Lacey Knowles. Multispecies coalescent delimits structure, not species. *Proceedings of the National Academy of Sciences*, 114(7):1607–1612, February 2017. Publisher: Proceedings of the National Academy of Sciences.
19. The Pandas Development Team. pandas-dev/pandas: Pandas, June 2023.
20. Guido Van Rossum and Fred L. Drake. *Python 3 reference manual*. CreateSpace, Scotts Valley, CA, 2009.
21. Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *Journal of Machine Learning Research*, 11(95):2837–2854, 2010.
22. Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.
23. Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021. Publisher: The Open Journal.