**Uncovering the horseshoe effect in microbial analyses**

James T. Morton1,2, Liam Toran3, Anna Edlund4, Jessica L. Metcalf5, Christian Lauber6 Rob Knight1,2

1 Department of Computer Science, University of California San Diego, La Jolla, CA, USA

2 Department of Pediatrics, University of California San Diego La Jolla, CA, USA

3 École Normale Supérieure de Lyon, Department of Mathematics, France

4 J. Craig Venter Institute, Genomic Medicine group, 4120 Capricorn Lane, La Jolla, CA, 92037

5 Department of Animal Sciences, Colorado State University, Fort Collins, CO, USA

6 Nestle Institute of Health Sciences, Lausanne, Switzerland

#Corresponding author:

Rob Knight

9500 Gilman Dr. MC 0763

La Jolla, CA 92093

Phone: 858-246-1184

E-mail: robknight@ucsd.edu

**Abstract word count: 211**

**Main Text word count: 2583**

**Abstract: 150**

**Abstract**

**The horseshoe effect is a phenomenon that has long intrigued ecologists.  Commonly thought to be an artifact of dimensionality reduction, multiple techniques were developed to unravel this phenomenon and simplify interpretation. Here, we provide evidence that horseshoes arise as a consequence of distance metrics that saturate - a familiar concept in other fields but new to microbial ecology.   This saturation property loses information about community dissimilarity, simply because it cannot discriminate between samples that do not share any common features. The phenomenon illuminates niche differentiation in microbial communities and indicates species turnover along environmental gradients.  Here we propose a rationale to the observed horseshoe effect from multiple dimensionality reduction techniques applied to simulations, soil samples, and samples from postmortem mice.** **An intuitive understanding of this phenomenon allows for the targeting of niche differentiation patterns from high-level ordination plots.**

**Importance**

Horseshoe effect is often considered an artifact of dimensionality reduction. We show that it's not true in the case for microbiome data and in fact, horseshoes can help analysts discover microbial niches across environments.

**Introduction**

Ecological datasets, particularly those observed in microbiome studies, are typically sparse and high-dimensional, frustrating most conventional statistical techniques.  Many numerical ecology software packages make use of distance-based statistics by calculating the distance between ecological communities, to compare various ecosystems to each other over space and time.  One of the most common exploratory analysis techniques is ordination, where the distances between the communities are embedded into a Euclidean space, and then visualized via Principal Components Analysis (PCA) (1). A widely used extension of this technique, where the distance metric can be varied, is called Principal Coordinates Analysis (PCoA) (1).

One phenomenon that commonly occurs in datasets containing ecological gradients is the horseshoe effect or Guttman effect (2).  This phenomenon is typified by a linear gradient that appears as a curve in ordination space. The horseshoe effect, or its relative the arch effect (3) is observed using multiple types of ordinations, including Principal Components Analysis, Principal Coordinates Analysis, Non-Metric Multidimensional Scaling, Correspondence Analysis, and many others (1).  In 1982, the prevailing view of the horseshoe effect arose, when it was described by Gauch as a mathematical artifact that obscures the underlying ecological gradient.  Soon thereafter, Detrending Correspondence Analysis (3) was invented to unbend the horseshoe using reciprocal averaging. Since then, detrending has become a commonly applied practice to ordinations in ecological datasets.   Although these detrending techniques appear to provide a more intuitive visualization, they have been criticized as providing a distorted perspective of the underlying data, relying on many parameter settings that cannot be chosen in a principled way, and obscuring true underlying patterns in the data (4).

From previous studies, it was shown that horseshoes can arise from band tables (5)(6).  These tables consist of highly dense, non-zero values along the diagonal of the table, and sparse values everywhere else.  This pattern can be apparent when the rows and columns are sorted in the proper order.  ­Although the idea that band tables lead to horseshoes is not a new idea, it is commonly misunderstood how this concept applies to microbial analyses.  Here we provide some intuition behind the mathematical structure of horseshoes.

In Figure 1a, we show a simulated band table, where each vertical band is represented by a sample, and contains 10 non-zero values.  In typical microbiome datasets, these values could reflect OTU or species counts; for simplicity, here we will to refer to them as species counts, although this concept can also be generalized to multiple data types, such as gene counts, metabolite abundances.  Each sample in the table is shifted by 1 row, creating the band effect.  When PCA is applied directly to this table, the first 2 eigenvectors yield a horseshoe pattern (Figure 1b).  Here, the band table is parameterized with a band size of 10, since each sample has exactly 10 non-zero values.

For close local points, the Euclidean distance grows linearly along the gradient (Figure 1c).  However, after a certain point, the distance completely saturates. This property has been previously noted with Euclidean distance (3).  The overlap between the first sample in the band table, and sample 10 and beyond disappears, and the distance between these samples is maximized. This can yield unintuitive properties, sample 10 could be less dissimilar than sample 1 compared to sample 20.  For instance, sample 10 could represent a medium pH environment, sample 1 could represent an acidic environment and sample 20 could represent a basic environment. Sample 1 is expected to a substantially more different microbial community to Sample 20 than Sample 10. The acidophiles that found in Sample 1 are typically not found in basic environments.  But as far as Euclidean distance is concerned, sample 10 is just as dissimilar to sample 1 as sample 20, just because there are no common bacteria shared between these samples.  It is apparent that the saturation property of Euclidean distance does not capture all of the information about community dissimilarity along a gradient, simply because it cannot discriminate between samples that do not share any common features.  Once the distance is saturated, all samples that do not overlap lie within a ball of radius B where B is the band size and the first point is the center of the ball as shown in Figure 1d.

This saturation property has been suggested to give rise to horseshoes in previous studies in other fields (5), and is an unintuitive property that can confound ecological interpretations if not understood properly.  This property also restricts the possible trajectories of samples in the feature space, and gradients cannot be represented by linear trajectories in the real space (Supplemental proof 1). This means that communities in the original high dimensional space do not arrange into linear trajectories in the first place, and when projected to lower dimensions do not fall into linear trajectories.  These trajectories are what we refer to as horseshoes. The horseshoe phenomenon is analogous to the familiar concept of saturation in molecular evolution, where two randomly evolving sequences saturate at 75% DNA sequence identity (assuming equal nucleotide frequencies), even if infinite time has elapsed (7). Consequently, distances that reflect a higher degree of molecular change need to be corrected for multiple substitutions in order to recover the molecular clock-like behavior obtained when comparing more similar sequences. This is why corrections according to models such as Jukes-Cantor or the Kimura 2-parameter model are required to obtain distances for reconstructing better phylogenetic trees. Analogous distance corrections are needed in microbial ecology for reconstructing better relationships among microbial communities (8).

It is important to note that horseshoes do not only arise from PCA, but also arise in PCoA with a variety of distance metrics.  Arch effects have plagued every multidimensional reduction technique we have applied to a wide range of microbial ecology datasets (9). In the following case studies, we’ll show that these distance metrics also have the saturation property.  In addition, if a distance doesn’t have this saturation property, there won’t be an observed horseshoe artifact (Figure S1).

*Case Study 1 - 88 Soils*

In this study, 88 soil samples were obtained from multiple locations across the United States having varying levels of pH (10).  The V4 region of the 16S rRNA gene (16S) within each organism was amplified and sequenced using 454 pyrosequencing to obtain relative abundances of microbial taxa.  A matrix representing abundance values for each taxonomic unit per soil sample was used as input in correspondence analysis (CA) (11). The resulting ordination showed clear separation of the communities based on pH (Figure 2a), which led to the same conclusion that pH is a major driving factor in soil biogeography, i.e. pH has major impacts on the distribution of bacterial taxonomic units in soil (10).  The CA analysis in Figure 2a also shows the classic horseshoe shape.   Here we revisited this study, to better understand the horseshoe shape behind this dataset.

To test the effect of another commonly used distance metric on the sample distribution, we analyzed the same soil dataset applying Chi Squared distance (Figure 2b). Similar to what was observed with Euclidean distance, which was applied in the simulation, the Chi Squared distance increased sharply at pH 3 and 4, but began to saturate at pH of 5. Also the band table similar to what we have observed in Figure 1a can be obtained when sorting the OTU table by sample pH and the mean pH of the samples that the OTUs were observed in (Equation 1).  While the diagonal isn’t completely dense, there are more non-zero values compared to the corners of the heatmap. In line with the findings from the original study, this pattern is likely representative of niche differentiation of OTUs with respect to pH. The organisms that thrive in low pH environments tend not to exist in high pH environments and vice versa.  Low pH and high pH samples are shown in Figure 3c to have few overlapping species, a pattern not observed in the original study as membership was evaluated at coarser levels of taxonomic resolution (10).

*Case Study 2 - Post Mortem Mice Study*

In this study, 120 mice were sacrificed and allowed to decompose on soil. Mice were destructively sampled over approximately 8 weeks (12). 16S sequencing libraries were generated from total DNA extracted from swabs of the skin on the head, and relative abundance values were calculated for each bacterial OTU. A relative abundance matrix was generated for each library and used as input in PCA. This analysis generated a clear horseshoe (Figure 2d) using unweighted UniFrac distance (13), with a gradient with respect to the time since death, possibly reflecting a changing skin microbiome during decomposition of the mouse carcass. When the samples were sorted by time since death using a similar strategy as noted above, a band table emerges (Figure 2f). Also, the unweighted UniFrac distance analysis appears to have the same saturation property as observed previously with Euclidean distance and Chi-squared distance.  It is important to note that highest possible UniFrac distance is 1, suggesting that this distance metric can also be saturated.  In Figure 4b, while the distance hasn’t completely saturated, these distances are quickly approaching the theoretical maximal UniFrac distance.

The striking changes in microbial communities during decomposition are associated with dramatic environmental biochemical changes, including increased pH, ammonia, and total nitrogen, all measured in soil beneath the mice. Correspondingly, microbial communities are predicted to increase in gene abundance of important nitrogen cycling pathways such as amino acid degradation (e.g. glutamate dehydrogenase, lysine decarboxylase, ornithine decarboxylase) and nitrate reduction (e.g. nitrate and nitrite reductase). Bacterial taxa in the families Chromatiaceae (OTU 46026, 4482362) and Rhizobiaceae (OTU 4301099) are involved in nitrogen metabolism and become abundant as mouse bodies progress through the stages of decomposition (e.g. Fresh, Active Decay, Advanced Decay).  As shown in Figure S1, all of these OTUs peak at specific timepoints. The two Chromatiaceae OTUs peak during Active Decay (bloating and purge of fluids) at 15 days of decomposition. The Rhizobiaceae OTU peaks during Advanced Decay (sinking and sagging flesh) at 30 days of decomposition and when pH, ammonia, and total nitrogen were measured at their highest levels (12).

To further validate if saturation leads to horseshoes, a new distance metric EMBAD (Earth Mover Band Aware Distance) was engineered to be non-saturating as a proof of concept (Supplemental Methods).  This distance metric uses prior knowledge about the ordering of the band table, and is determined by calculating the flow between two samples. As shown in Figure S1a, sample 1 and sample 2 each have 4 species proportions.  To calculate the distance between sample 1 and sample 2, the probability mass of species 1 and species 2 needs to be shuffled over to species 3 and species 4.  This concept is analogous to computing maximum flow along a pipe, and can be calculated using Earth Mover’s distance (14–16)

For the 88 soils (Figure S1b), the EMBAD was applied to the pH sorted table. Therefore, even if two samples are not overlapping, samples closer together will have a smaller distance than samples further apart in the gradient.  This is because the distance is defined to be not saturating and explicitly accounts for the pH gradient.  The same strategy was employed for the postmortem interval mice (Figure S1c), sorting the table by decomposition days. The PCoA plots resulting from these applications of EMBAD suggest that a non-saturating distance metric could remove the horseshoe effect from lower dimensional projections of these abundances.  This provides further evidence that this saturation property could explain the the horseshoe phenomenon.

For the 88 soils study a PERMANOVA test investigating the difference between soils with a pH less than 3, and soils with a pH greater than 8.  With the EMBAD distance metric the PERMANOVA gave a pseudo F-statistic of 650.5 and a p-value of 0.0003, which has a much larger effect size compared to the original Chi-squared distance metric with a pseudo F-statistic of 3.8 and a p-value of 0.0004 with 9999 permutations.  A similar trend was observed in the post mortem interval mice study when testing the first decomposition day to the last decomposition day using PERMANOVA.  The EMBAD distance metric had a pseudo F-statistic of 439.8 and a p-value of 0.0001 with 9999 permutations, which has a larger effect size than the Unifrac distance metric, which had a pseudo F-statistic of 25.5 and a p-value of 0.0001. This method is relieved from misinterpretations of data due to horseshoes and arches and facilitates the interpretation of taxonomic units along biologically significant gradients that reflect the selective pressure of these factors on the distribution of microbes.

In light of the benefits of engineering a non-saturating distance metric, the EMBAD distance metric requires the gradient to known a priori. Generalizing this approach in the absence of known gradients is a difficult problem would require an exhaustive using known algorithms. Specifically, this problem falls under the category of NP-hard problems (Supplemental Proof 2). In the 88 soils study and the post mortem mice study, we were fortunate to be able to infer the underlying band table with known metadata.

The band patterns we observe here are probably very common in ecology studies investigating species distribution patterns across spatial or temporal gradients. The pattern confirms microbial ecological fundamentals, i.e.  bacteria have acquired unique adaptations to the environment and occupy either a broad range or very specific niches. In our case studies of the 88 soils - and the postmortem mice we confirmed that by using a band table pattern analysis approach, bacterial species show different adaptations to pH and bacterial diversity changes over time during decomposition of mice carcasses. The band pattern approach we apply here represents an additional method to visualize differences between microbial communities.

Based on our observations here, the horseshoe effect appears in dimensionality reduction techniques due to the saturation property of distance metrics.  While we have only tested a few distance metrics, it is suspected that a vast majority of these distance metrics exhibit the same property, which would also explain why horseshoes are encountered so frequently across many different fields. The saturation property has also been observed in multiple other fields, and other studies from different disciplines have come to similar conclusions (5). In addition, multiple techniques such as self organizing maps (17) and local linear embedding (18) attempt to avoid this saturation phenomenon by focusing on distances of nearby points.

In spite of the saturation property of distance metrics, identifying horseshoes is still highly useful for identifying patterns concerning niche differentiation. Properly understanding this phenomenon will motivate the use and development of algorithms such as biclustering (19) to uncover underlying band patterns. These insights will in effect allow us to identify microbial niches along different environmental conditions.

**Materials and Methods**

All analyses can be found under<https://github.com/knightlab-analyses/horseshoe-analyses>

The mean gradient used for the 2 case studies was calculated as follows.

**Equation 1**

Where is the proportion of OTU in sample , is the mean gradient of OTU , and is the sample gradient at sample . This calculation can be found in the gneiss package under the function **mean\_niche\_estimator.** The function used to sort the tables in Figure 2c used **niche\_sort.** In the 88 soils study, the table was sorted by sample pH and the mean pH of the samples that the organisms were observed in. In the post mortem mice study, the table was sorted by the days of decomposition and the mean day of the samples of that the organisms were observed in.

The heatmap in Figure 2f and the abundances in Figure S2 were normalized using the centre log ratio transformation given by the following equation

**Equation 2**

Where is the geometric mean and is the average of the log transformed values. A pseudocount of 1 is added to all of the counts to prevent logarithms of zero occurring.

Analyses were performed using Scipy, Numpy, Matplotlib, Seaborn, Scikit-bio and Gneiss.

**Acknowledgments**

We would like to thank Noah Fierer on his input on the 88 soils analysis, and Dan Knights for discussion on the EMBAD metric. We’d also like to acknowledge Amnon Amir and Tomasz Kosciolek on their insights into the horseshoe effect.

**Funding information**

JTM was funded by NSF GRFP DGE-1144086. JLM and RK were funded by the Office of Justice Programs National Institute of Justice Award NIJ-2011-DN-BX-K533.

**References**

1. **Legendre P**, **Legendre L**. 2003. Numerical Ecology.

2. **Taxonomy P**. 2002. Resemblance coefficients and the horseshoe effect in principal coordinates analysis **83**:3331–3343.

3. **Hill MO**, **Gauch HG**. 1980. Detrended Correspondence Analysis: An Improved Ordination Technique. Vegetatio.

4. **Ejrnaes R**. 2000. Can we trust gradients extracted by Detrended Correspondence Analysis? J Veg Sci **11**:565–572.

5. **Disconis P**, **Goel S**, **Holmes S**. 2008. Horseshoes in multidimensional scaling and local kernel methods **2**:777–807.

6. **Camiz S**. 2016. The Guttman effect : Its interpretation and a new redressing method.

7. **Smith JM**, **Smith NH**. 1996. Synonymous nucleotide divergence: What is “saturation”? Genetics **142**:1033–1036.

8. **Tourasse NJ**, **Gouy M**. 1997. Evolutionary distances between nucleotide sequences based on the distribution of substitution rates among sites as estimated by parsimony. Mol Biol Evol **14**:287–298.

9. **Kuczynski J**, **Liu Z**, **Lozupone C**, **McDonald D**, **Fierer N**, **Knight R**. 2010. Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. Nat Methods **7**:813–819.

10. **Lauber CL**, **Hamady M**, **Knight R**, **Fierer N**. 2009. Pyrosequencing-Based Assessment of Soil pH as a Predictor of Soil Bacterial Community Structure at the Continental Scale **75**:5111–5120.

11. **Greenacre M**. 1984. Theory and Applications of Correspondence Analysis.

12. **Metcalf JL**, **Xu ZZ**, **Weiss S**, **Lax S**, **Treuren W Van**, **Hyde ER**, **Song SJ**, **Amir A**, **Larsen P**, **Sangwan N**, **Haarmann D**, **Humphrey GC**, **Ackermann G**, **Thompson LR**, **Lauber C**, **Bibat A**, **Nicholas C**, **Gebert MJ**, **Petrosino JF**, **Reed SC**, **Gilbert JA**, **Lynne AM**, **Bucheli SR**, **Carter DO**, **Knight R**. 2016. Mammalian Corpse Decomposition. Science (80- ) **351**:158–162.

13. **Lozupone C**, **Knight R**. 2005. UniFrac : a New Phylogenetic Method for Comparing Microbial Communities. Appl Environ Microbiol **71**:8228–8235.

14. **Werman M**. 2009. Fast and Robust Earth Mover ’ s Distances. Proc 12th IEEE Int Conf Comput Vis - ICCV ’09 460–467.

15. **Pele O**, **Werman M**. 2008. A linear time histogram metric for improved SIFT matching. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) **5304 LNCS**:495–508.

16. **Levina E**, **Bickel P**. 2001. The earth mover’s distance is the Mallows distance: some insights from statistics. Eighth IEEE Int Conf Comput Vis 251–256.

17. **Kohonen T**. 1982. Self-organized formation of topologically correct feature maps. Biol Cybern **43**:59–69.

18. **Saul LK**, **Roweis ST**. An Introduction to Locally Linear Embedding.

19. **Pontes B**, **Giráldez R**, **Aguilar-Ruiz JS**. 2015. Biclustering on expression data: A review. J Biomed Inform **57**:163–180.

**Figure legends**

**Figure 1**: **(a)** A band table where the y axis encodes for individual OTUs and the x axis encodes for samples. Blocks that are colored black have a value of 1/10 while blocks that are colored white have a value of 0. **(b)** The first 2 components from a PCA of the band table, yielding the typical horseshoe shape. **(c)** The Euclidean distance from the point 0 to all of the other points. **(d)** An illustration of distance saturation property.

**Figure 2:** **(a)** Correspondence analysis of 88 soils. **(b)** Distance saturation of chi-squared metric, plotting the chi squared distance of the first sample versus all of the other samples. **(c)** Heatmap of log transformed OTU counts from the 88 soils with the samples sorted by pH and the OTUs sorted by mean pH. **(d)** Principal Coordinates Analysis of unweighted UniFrac distance. **(e)** UniFrac distance of a samples from the last time point versus all of the samples. **(f)** Heatmap of centred log ratio transformed (Equation 2) OTU counts sorted by harvest days.

**Figure S1: (a)** The description of a the EMBAD metric, engineered to be non-saturating. **(b)** A PCoA of the EMBAD distance applied to the 88 soils. **(c)** A PCoA of the EMBAD distance applied to the Post Mortem Interval Mice.

**Figure S2:** The centre log ratio (Equation 2) transformed abundances of Rhizobiaceae (OTU 4301099) and Chromatiaceae (OTU 46026, 4482362) versus time.

**Supplemental Proof 1:** A mathematical proof that N distinct samples separated by a minimum Euclidean distance cannot exist on a linear trajectory.

Supplemental Proof 2: Finding an optimal distance metric to explain a horseshoe without a known sample ordering is NP-hard.