Quantitative Bio-Science 31(2), 83~84(2012)

A Comment to "A Microbiology Primer for Pyrosequencing"

Stephan Huckemann*

Institut für Mathematische Stochastik Georg-August-Universität Göttingen Goldschmidtstr. 7,
D-37077 Göttingen, Germany

(Received August 10, 2012; Revised October 31, 2012; Accepted November 20, 2012)

ABSTRACT

In view of dimension reduction, aspects of data correlations as well as an intrinsic form of principal component analysis are brought forth. The latter method may be used for assessing the temporal evolution of gut metagenomics in probiotic therapy.

Key words : Dimension reduction, Geodesic PCA, Shape spaces, Temporal evolution

1. The Paper's Scope

The paper under consideration is a very well written and inspiring exposition, gently guiding mathematicians into the exciting field of microbiology by taking them along an underlying example of metagenomic analysis in the human gut. As some of the authors have shown in a separate report ([1]), metagenomic analyses in this specific area are fundamental for the development of effective treatments of contemporary antibiotic therapy related diseases that are often lethal. In particular, unfolding some specific statistical perspectives at the end of the paper opens the stage for a challenging task of developing sophisticated tools that may appropriately shed light onto the highly complex biology responsible for the beneficial effects of specific probiotic therapies. I take special delight in commenting on two possible research directions that came to mind when reading the paper.

2. Dimension Reduction by (Non-Linear) Correlation

Analyses of metagenomics data naturally falls into the realm of "$p \gg n$", the number of parameters (base sequence combinations) being very much larger than the sample size (the number of individuals involved), the statistical treatment of which requires highly skilled methodology. One emphasis has been put by the authors on principal component analysis (PCA) which may be viewed as representations of approximations to the data. With its well known potential beyond descriptive services, the authors also use it for discrimination purposes.

This paradigm implicitly assumes that underlying dependencies can be modelled by linear correlations - certainly a valid approach in the sense of a first order approximation of a truly non-linear relationship caused by the non-linear geometry of the base space, e.g. the space of covariance matrices. An approach for dimension reduction more natural than classical PCA may consist of employing *geodesic PCA* (see [2]) that takes the curvature of the underlying space into account. For the space of covariance matrices, geodesic PCA for *Kendall's size and shape spaces* can be utilized, taking advantage of a proposed modelling by [3] which is further elaborated on in [4]. Essentially, this gives a space of non-negative sectional curvatures which, however, has issues of non-uniqueness. From a practical perspective, guaranteeing uniqueness of intrinsic means, say, it may rewarding to implement geodesic PCA in the more natural geometry of the *universal symmetric space of non-compact*

* Correspondence should be addressed to Dr. Stephan Huckemann, Institut für Mathematische Stochastik, Georg-August-Universität Göttingen Goldschmidtstr. 7, D-37077 Göttingen, Germany. Tel: +49-551-39-13517, Fax: +49-551-39-13505, E-mail: huckeman@math.uni-goettingen.de. Supported by DFG Grant HU 1575/2-1.

type with non-positive non-constant sectional curvatures e.g. [5, Chapter XII]. To the knowledge of the author of this note, this has not been done yet. In fact, it appears mathematically quite challenging.

Speaking of correlations, second order structure may be understood in this context. RNA folding is only possible at loci where base sequences correlate. Not all correlating sequences, however, lead to physical folding, and as the authors of the paper illustrate, often there are more than one possible foldings. Since different folding patterns correspond to different biological functionality, it might be interesting to investigate more closely correlation patterns between sequence correlations on the one side and geometrical folding on the other side.

In addition to this "double" correlation phenomenon, folding may also be viewed as a sudden drop in data dimensionality. This raises the question to what extent the presence of singular strata influences the statistical data analysis - the base space is a manifold with singularities which form lower dimensional strata that are again manifolds with singularities. Means on thus stratified spaces can be *manifold stable* allowing for classical asymptotic analysis, as is the case in G-spaces (cf. [6]). They may also hit or even stick to singularities as is the case in some negative curvature scenarios (cf. [7]), giving a very non-Euclidean asymptotic behavior.

3. The Temporal Evolution of Diversity

In addition to assessing a single status of metagenomic diversity one might also aim at a deeper understanding of the evolution of such diversity, especially if coming from a considerably less diverse initial state. In other biological systems (e.g. [8]), it has been found that growth can be modelled in few dimensions only; for growth in individuals essentially one dimension suffices and within a suitable geometry, geodesic growth

$$\gamma(t)=\exp_p(t\upsilon)$$

can be assumed (cf. [9]). Although the biological system investigated by the authors of the paper comprises a huge number of individuals, it may be expected that the temporal evolution of competition outcome imposes constraining effects that might be caught by appropriate totally geodesic subspace models extending modelling by single geodesics as in [10]. Very likely, competition accounts for points of non-differentiability. This leads to the challenging task of developing (totally) geodesic change point models on stratified spaces.

4. Conclusion

Papers and tutorials of the type "Mathematics for X" exist in abundance and allow researchers in the field X to apply well established mathematical theory. Tutorials of the other type "X for Mathematicians" are extremely rare. These, however, allow for high profile interaction between researches of X and mathematical theory building. Now that the authors have considerably contributed to opening the field X=microbiology to mathematicians, I anticipate most exciting interdisciplinary research.

References

1. Shahinas D, Silverman M, Sittler T, Chi C, Kim P, Allen-Vercoe E, Weese S, Wong A, Low D, Pillai D. Toward an Understanding of Changes in Diversity Associated with Fecal Microbiome Transplantation Based on 16S rRNA Gene Deep Sequencing. mBio 2012;3(5):e00338-12.
2. Huckemann S, Hotz T, Munk A. Intrinsic shape analysis: Geodesic principal component analysis for Riemannian manifolds modulo Lie group actions (with discussion). Stat Sinica 2010; 20(1):1-100.
3. Dryden I, Koloydenko A, Zhou D. Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. Ann Appl Stat 2009;3(3):1102-1123.
4. Huckemann S. Inference on 3D Procrustes means: Tree boles growth, rankdeficient diffusion tensors and perturbation models. Scand J Stat 2011;38(3):424-446.
5. Lang S. Fundamentals of Differential Geometry. Springer; 1999.
6. Huckemann S. On the meaning of mean shape: Manifold stability, locus and the two sample test. Ann I Stat Math 2012; To appear.
7. Hotz T, Huckemann S, Le H, Marron JS, Mattingly J, Miller E et al. Sticky central limit theorems on open books. preprint, arXiv:1202.4267v1 [math.PR] 2012.
8. Morris R, Kent JT, Mardia KV, Aykroyd RG. A parallel growth model for shape. In Arridge S and Todd-Pokropek A Eds. Proceedings in Medical Imaging Understanding and Analysis. Bristol: BMVA 171-174. 2000.
9. Le H, Kume A. Detection of shape changes in biological features. J Microscopy 2000;200(2):140-147.
10. Huckemann S. Intrinsic inference on the mean geodesic of planar shapes and tree discrimination by leaf growth. Ann Stat 2011;39(2):1098-1124.