J Math Imaging Vis (2014) 50:1–4
DOI 10.1007/s10851-014-0504-5

Geometry and Statistics: Manifolds and Stratified Spaces

Aasa Feragen · Mads Nielsen · Eva Bjørn Vedel Jensen · Andrew du Plessis · François Lauze

Published online: 25 March 2014

1 Introduction

Statistics and machine learning typically take place in linear spaces. When data points are described by fixed-dimensional vector measurements, analysis takes place in a Euclidean space, and when data is analyzed using kernels, analysis takes place in a reproducing kernel Hilbert space.

Geometry enters statistics and machine learning when explicit models for particular data classes are needed. *Object-oriented data analysis* [1,2] where the data objects are explicitly modeled are powerful for avoiding overfitting to noise for data with low signal-to-noise ratios or small sample sizes. Ultrasound, cryo electron microscopy or DWI images are examples of data where explicit modeling will often lead to improved predictive models. Explicit data modeling also allows us to solve **object-valued problems**, or problems whose answers belong to the same class of objects as the data. Examples of object-valued problems are *what is the average object in a population of objects?* or *visualize the variation along the first principal component*. Object-oriented data analysis through specialized data models often leads to non-linear data spaces, whose geometry influences data analysis.

Manifolds and *stratified spaces* are large families of non-linear geometric spaces used for mathematical modeling of real data. When data is modeled in such spaces, standard operations such as interpolation, averaging, principal components or hypothesis testing are no longer straightforward or even necessarily well defined. This special issue aims to capture the state-of-the-art in statistics on manifolds and stratified spaces, drawing on scientific communities as diverse as mathematical statistics, geometry, image analysis and computational biology.

2 Manifold Statistics

Shapes are classical examples of objects whose variation exhibits nonlinear behaviour. The history of shape statistics starts as early as 1917 with D'Arcy Thompson's book *On Growth and Form* [3], pioneering the study of shapes through their deformations. D'Arcy Thomson cast the problem of analyzing the growth and form of biological shapes as a study of deformations of point sets and conformal maps.

Two main schools of thought have emerged from D'Arcy Thompson's work.

2.1 Pattern Theory and Its Descendants: Analyzing Data Through Its Deformations

In the 1970–1980s, U. Grenander developed the *General Pattern Theory* [4], studying both shapes and more general patterns through their deformations, as represented by group actions. In this formulation the analysis is transferred to the domain of Lie groups and homogeneous spaces. The princi-

A. Feragen (✉) · M. Nielsen · F. Lauze
Department of Computer Science, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen, Denmark
e-mail: aasa@diku.dk

M. Nielsen
e-mail: madsn@diku.dk

F. Lauze
e-mail: francois@diku.dk

E. B. Vedel Jensen · A. du Plessis
Department of Mathematics, Aarhus University, Ny Munkegade 118, building 1535, room 328, 8000 Aarhus C, Denmark
e-mail: eva@imf.au.dk

A. du Plessis
e-mail: matadp@imf.au.dk

ples of Grenander's pattern theory are the foundation of modern research topics such as Log-Demons [5] and Large Deformation Diffeomorphic Metric Mappings (LDDMM) [6,7]; see also [8] for a detailed presentation.

In *Efficient Parallel Transport of Deformations in Time Series of Images: from Schild's to Pole Ladder*, Lorenzi and Pennec [9] present the *pole ladder*, a numerical method for parallel transport on manifolds, applicable to longitudinal image registration in the Log-Demons and LDDMM frameworks.

In *Multivariate tensor-based morphometry with a right-invariant Riemannian distance on $GL+(n)$*, Zacur et al. [10] use a right-invariant metric on $GL^+(n)$ for voxel-based statistics on the Jacobian of the registration diffeomorphism from a template in any given voxel.

2.2 Explicit Manifold Parametrization of Data

Also in the 1970–1980s, Kendall developed a statistical model of shapes parametrized by finite point sets [11]. In Kendall's work, shape spaces are metric spaces obtained from representing a shape in $\mathbb{R}^n$ as a finite, ordered family of points in $\mathbb{R}^n$, modulo the action of the group of direct similarities (scaling and special Euclidean transformations). In low dimension ($n \leq 2$), the resulting orbit space is a Riemannian manifold, while in higher dimensions, the metric space obtained possesses singularities and in that sense is a stratified space. The direct representation and comparison of shapes as found in Kendall's model have since been used widely in the statistical and applied communities both for analyzing shapes and more general data. In many cases, data is modeled as residing on a Riemannian manifold of finite or infinite dimension, and the generalization to such spaces of statistical properties and corresponding numerical algorithms has thus become an active area of research.

In *Intrinsic Polynomials for Regression on Riemannian Manifolds*, Hinkle et al. [12] develop a theory of intrinsic polynomial curves for regression on Riemannian manifolds, in which polynomial curves are characterized by the vanishing of the fixed higher-order covariant derivative.

In *Density Estimators of Gaussian Type on Closed Riemannian Manifolds*, Bates and Mio [13] prove consistency results for density estimators of Gaussian type on closed, connected Riemannian manifolds.

In their paper *Overview of the Geometries of Shape Spaces and Diffeomorphism Groups* [14], Bruveris, Bauer and Michor give a comprehensive overview of the state-of-the-art on the geometry of infinite-dimensional shape spaces. Here, a smooth shape is represented as a Riemannian manifold embedded or immersed in Euclidean space, which is a very general, natural and compelling definition of shape. The paper gives an overview of diffeomorphism groups, their actions on mapping spaces and the resulting shape spaces. In particular, geometric spaces of metrics on shape spaces are studied, with applications to LDDMM.

3 Stratified Statistics

Many types of real-life data are not modeled well as residing on a manifold. For instance, factoring out a group of similarities can lead to a non-smooth orbit space, as is the case with Kendall's shape space in dimensions ≥ 3.The paper *On Means and Their Asymptotics: Circles and Shape Spaces* by Huckemann and Hotz [15] surveys current results and open questions related to the effect of singular strata on means and their asymptotics in circles, shape spaces and quotients of proper Lie group actions on Riemannian manifolds.

While singular spaces have played a part in shape analysis since the introduction of Kendall's shape space, the term *stratification* has not been used for that long. A stratified space is a union of smooth manifolds, meeting in a "controlled" way [16]. Stratified spaces lend themselves well to modeling data with variable topology, such as weighted trees [17–23] or graphs [24]. Approaches to stratified data spaces include the thesis of Bendich [25], who addressed the inverse problem of estimating stratified data spaces from data using persistent homology. Recently, the 2010-11 SAMSI working group *Data Analysis on Sample Spaces with a Manifold Stratification* made significant contributions to developing a statistical theory for stratified spaces [26,27].

One important open question in stratified statistics is how dimensionality reduction should be defined. Some approaches model the first principal component as a geodesic optimizing a least squares cost function [19,22]. However, it is unclear both how well a geodesic can describe data in a stratified space, how such principal components might be computed, and how to pass to the second principal component. The latter question is also not fully solved in the case of manifold data spaces.

The paper *Backwards Principal Component Analysis and Principal Nested Relations* by Marron and Damon [28], suggests a new approach to PCA/dimensionality reduction, where dimensions are peeled off in a backwards fashion through a series of nested relations. This definition of PCA does not assume a particular data space, only that the principal component candidates can be defined as level sets of a function.

Sampling is a fundamental tool for solving optimization problems, but sampling in stratified spaces is a difficult open question because analogues of distributions such as Gaussians do not translate trivially. In the paper *Diffusion on some simple stratified spaces*, Nye and White [29] develop a theory for diffusion processes on simple stratified spaces.

In the paper *Tree-oriented analysis of brain artery structure* by Skwerer et al. [30], tree-space statistics are used to

analyze a set of leaf-labeled blood vessel trees from the Circle of Willis, which is topologically very heterogeneous. The authors investigate the effect of heterogeneous topologies on mean trees and intrinsic dataset geometry as captured by minimum spanning trees. Moreover, multidimensional scaling is used to transfer the data to a Euclidean space where further statistical analysis is performed.

4 Estimation of Shape Properties

The previous sections focus on the nonlinear nature of data spaces and the effect of data space geometry on statistical measurements. Often, however, the geometric properties of the data objects themselves are highly descriptive, and lead to new mathematical questions.

The paper *Equi-Affine Invariant Geometry for Shape Analysis* by Raviv et al. [31] studies local descriptors of shapes in $3D$. Each shape is represented as a 2-dimensional manifold in $\mathbb{R}^3$ with an equi-affine invariant Riemannian (pseudo-) metric derived from the second fundamental form. The new metric is used for shape operations such as Voronoi tesselation, shape retrieval and symmetry testing using the heat kernel associated to the new Riemannian (pseudo-) metric.

The paper *Stable Length Estimates of Tube-like Shapes* by Pausinger and Edelsbrunner [32] proposes stable length estimates for tubular structures such as river networks, blood vessels or dendrites. An integral geometric estimate of length for tube-like shapes is deeloped, and persistent homology is used to separate out the stable part of the estimate.

References

1. Wang, H., Marron, J.: Object oriented data analysis: sets of trees. Ann. Stat. **35**(5), 1849–1873 (2007)
2. Marron, J., Alonso, A.: Overview of object oriented data analysis. Biom. J. (2014)
3. Thompson, D., Bonner, J.: On Growth and Form. Canto, Cambridge University Press, Cambridge, MA (1992)
4. Grenander, U.: General Pattern Theory: A Mathematical Study of Regular Structures. Oxford Mathematical Monographs, Clarendon (1993)
5. Arsigny, V., Commowick, O., Pennec, X., Ayache, N.: A log-euclidean framework for statistics on diffeomorphisms. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) Medical Image Computing and Computer-Assisted Intervention MICCAI 2006, Lecture Notes in Computer Science, vol. 4190, pp. 924–931. Springer, Berlin (2006)
6. Dupuis, P., Grenander, U.: Variational problems on flows of diffeomorphisms for image matching. Q. Appl. Math. **LVI**(3), 587–600 (1998)
7. Trouvé, A.: Diffeomorphisms groups and pattern matching in image analysis. Int. J. Comput. Vision **28**(3), 213–221 (1998)
8. Younes, L.: Shapes and Diffeomorphisms, 1st edn. Springer, Berlin (2010)
9. Lorenzi, M., Pennec, X.: Efficient parallel transport of deformations in time series of images: from schild's to pole ladder. J. Math. Imaging Vision (2013). doi:10.1007/s10851-013-0470-3
10. Zacur, E., Bossa, M.N., Olmos, S.: Multivariate tensor-based morphometry with a right-invariant riemannian distance on $gl + (n)$ (2013). doi:10.1007/s10851-013-0479-7
11. Kendall, D.G.: Shape manifolds, procrustean metrics, and complex projective spaces. Bull. London Math. Soc. **16**(2), 81–121 (1984). doi:10.1112/blms/16.2.81
12. J. Hinkle, P.F., Joshi, S.: Intrinsic polynomials for regression on Riemannian manifolds (2013). doi:10.1007/s10851-013-0489-5
13. Bates, J., Mio, W.: Density estimators of gaussian type on closed riemannian manifolds (2013). doi:10.1007/s10851-013-0460-5
14. Bauer, M., Bruveris, M., Michor, P.: Overview of the geometries of shape spaces and diffeomorphism groups. 38 (2013). doi:10.1007/s10851-013-0490-z
15. Huckemann, S., Hotz, T.: On means and their asymptotics: circles and shape spaces (2013). doi:10.1007/s10851-013-0462-3
16. Pflaum, M.: Analytic and geometric study of stratified spaces, lecture notes in mathematics, 1768 (2001)
17. Billera, L., Holmes, S., Vogtmann, K.: Geometry of the space of phylogenetic trees. Adv. Appl. Math. **27**(4), 733–767 (2001)
18. Owen, M., Provan, J.: A fast algorithm for computing geodesic distances in tree space. ACM/IEEE Trans. Comput. Biol. Bioinf. **8**, 2–13 (2011)
19. Feragen, A., Owen, M., J., P., Wille, M., Thomsen, L., Dirksen, A., de Bruijne, M.: Tree-space statistics and approximations for large-scale analysis of anatomical trees. In: IPMI (2013)
20. Feragen, A., Lo, P., de Bruijne, M., Nielsen, M., Lauze, F.: Towards a theory of statistical tree-shape analysis. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 2008–2021 (2013)
21. Feragen, A., Hauberg, S., Nielsen, M., Lauze, F.: Means in spaces of tree-like shapes. In: ICCV (2011)
22. Nye, T.: Principal components analysis in the space of phylogenetic trees. Ann. Stat. **39**(5), 2716–2739 (2011)
23. Bacak, M.: A novel algorithm for computing the Fréchet mean in Hadamard spaces. Preprint, http://arxiv.org/abs/1210.2145 (2012)
24. Jain, B.J., Obermayer, K.: Structure spaces. JMLR **10**, 2667–2714 (2009)
25. Bendich, P., Cohen-Steiner, D., Edelsbrunner, H., Harer, J., Morozov, D.: Inferring local homology from sampled stratified spaces. In: FOCS, pp. 536–546 (2007)
26. Miller, E., Owen, M., Provan, J.S.: Averaging metric phylogenetic trees. Preprint, http://arxiv.org/abs/1211.7046 (2012)
27. Hotz, T., Huckemann, S., Le, H., Marron, J., Mattingly, J., Miller, E., Nolen, J., Patrangenary, V., Skwerer, S.: Sticky central limit theorems on open books. Ann. Appl. Probab. **23**(6), 2238–2258 (2010)
28. Damon, J., Marron, J.: Backwards principal component analysis and principal nested relations. J. Math. Imaging Vision (2013). doi:10.1007/s10851-013-0463-2
29. Nye, T., White, M.: Diffusion on some simple stratified spaces (2013). doi:10.1007/s10851-013-0457-0
30. Skwerer, S., Bullitt, E., Huckeman, S., Miller, E., Oguz, I., Owen, M., Patrangenaru, V., Provan, S., Marron, J.: Tree-oriented analysis of brain artery structure (2013). doi:10.1007/s10851-013-0473-0
31. Raviv, D., Bronstein, A., Bronstein, M., Waisman, D., Sochen, N., Kimmel, R.: Equi-affine invariant geometry for shape analysis. 1–20 (2013). doi:10.1007/s10851-013-0467-y
32. Edelsbrunner, H., Pausinger, F.: Stable length estimates of tube-like shapes. J. Math. Imaging Vision (2013). doi:10.1007/s10851-013-0468-x

Aasa Feragen received her PhD in mathematics from the University of Helsinki, Finland, in 2010. She is currently an Associate Professor at the Department of Computer Science, University of Copenhagen, Denmark. Her research interests include the mathematical modeling of structured data such as trees and networks, and the development of statistical and machine learning methodology for their analysis, with applications in medical image analysis.

Mads Nielsen received his PhD in computer science in 1995 from the University of Copenhagen. In 1995–96 was post doc at Utrecht University and The University of Copenhagen, where he was assistant professor 1997–99. He became associate professor in 1999 and professor in 2002 at the IT University of Copenhagen, where he headed the Image Analysis Group. In 2007 he became professor and head of the Image Group, University of Copenhagen. He was general chair of MICCAI 2006 and is a member of the editorial board of IJCV and JMIV.

Eva Bjørn Vedel Jensen received her doctoral degree in 1987 with a thesis in stereology. She has been at Aarhus University since 1979, first as associate professor, then docent, research professor, and full professor since 2003. She has led a number of research centers and networks, including the Marie Curie training network "Advanced Medical Imaging and Spatial Statistics" 2002–2006, the T.N. Thiele Center for Applied Mathematics in the Sciences 2005–2009 and the Centre for Stochastic Geometry and Advanced Bioimaging 2010–2015. She has received several research prices, including The Villum Kann Rasmussen Annual Award for Technical and Scientific Research 2009.

Andrew du Plessis received his PhD from the University of Liverpool in 1974 with a thesis in Singularity Theory. He joined the University of Aarhus in 1977 where he has stayed since as associate professor in mathematics. Andrew is known for his work on topological stability and topological finite determinacy, as well as a set of seminar notes on Mather's topological stability theorem written together with C.G. Gibson, K. Wirthmueller and E.J.N. Looijenga, which remains a key reference in singularity theory today.

François Lauze studied Mathematics in France, University of Nice-Sophia-Antipolis where he was awarded a PhD in Algebraic Geometry in 1994. He spent some years in Burkina Faso, West Africa, where he taught Mathematics at the University of Ouagadougou. He then moved to Denmark and engaged in yet another PhD, awarded in 2004 at the IT University of Copenhagen, including work on Variational Methods for Motion Compensated Inpainting and motion recovery. He has since worked with variational and geometric methods for Image Analysis.