It is not our data that are non-euclidean, but only our models

Fred L. Bookstein

University of Vienna, University of Washington

The fault, dear Brutus, is not in our stars,
But in ourselves, that we are underlings.
—Shakespeare, *Julius Caesar*, I, ii

Summary: Claiming an empirical problem to be non-euclidean is often an error of collegial strategy rather than a strategy or scholium *per se*. Most original data in the natural sciences come from machines built on euclidean principles, or should. Statistical pattern analyses would do well to acknowledge this common origin in the course of whatever formalisms they follow.

The tourist visiting Vienna today is likely to spend an hour in the glittering *Kunstkammer* (cabinet of curiosities), a collection of 500 years of Habsburg *tchotchkes* just opened to the public inside the Kunsthistorische Museum. The most enchanting of these objects are the representations of the human face and figure, and the primary characteristic of these representations is the extraordinary range of geometrical scales they span, from miniatures to life size. Across the plaza at the Naturhistorische Museum, the *piece de resistance* of the anthropological collection is the Venus of Willendorf, a 25-millennium-old sculpture of the caricature of a woman whose price would be beyond rubies had rubies been known back then: huge fat reserves, and an explicitly carven vulva (together connoting great fertility), and an expensive hairdo. This sculpture is only 11 cm high. Still, we do not view it as the representation of an extremely small woman; we view it instead as the extremely small representation of an ordinary-sized woman (as best we can tell from buried remains). Art, then, is scale-free, especially the art of human representations. Yet our models for explaining the origin of these forms in organismal biology (evolution, development, function, disease, or therapeutic interventions) are intrinsically scale-bound models, which is to say, models in our ordinary Euclidean spaces.

Consider, for instance, a recently published method (Bookstein, 2012, 2013a) for linking geometric morphometrics to biomechanics when both are studying the same form variation. I showed you last year how in a concocted example (a set of cantilever beams all the same length, varying in taper, all strained by the same vertical load at the free end) I could predict strain energy under load almost perfectly from the principal components of Procrustes shape of the unloaded configuration. But to achieve that nearly perfect regression I had to arrange the situation so that Centroid Size hardly varied. After all, in the fundamental equations of these approaches, squared Procrustes distance from a starting form goes as the variance of the eigenvalues of the affine derivative , whereas strain energy goes as the sum of their squared differences from 1.0. (In other words, whereas the Procrustes distance of a rescaled form from the original is zero, the strain energy can be arbitrarily large.) These are not proportional, not even approximately, unless there is no size scaling. Otherwise, we simply could not relate Procrustes analyses of shape to realistic biomechanical analyses of strain. The bridging formalism must be set in form space, not in shape space.

A similar antinomy faces us when we attempt to map a functional index, such as the lever arm of a muscle, onto the shape space for a landmarkconfiguration.Functional indices can be indices of shape — that is, dimensionless — or they can scale with size or any power of size. To map a measure of shape onto shape space is simply a matter of taking its differential (see, e.g., Bookstein, 1986); but obviously a measure that scales with dimension 1 in size cannot be usefully mapped onto shape space. For instance, let u and v be two measures of form that have dimension cm^2 as functions of all the difference vectors $P_i - P_j$ of the landmarks' Cartesian coordinates $P_1, \ldots, P_k$, such as their squared lengths or their dot products in pairs.Then a formula like u/v will map onto shape space sensibly, whereas u and v separately cannot be expected to.

Another issue arises in connection with the shape space formalism as it applies to multiple groups or processes (such as growth) that extend over a considerable diameter.I talked about this in my contribution to the 2009 LASR proceedings. The matrix

$$J = \begin{pmatrix} \delta & 0 & \delta & 0 & \ldots & \delta & 0 \\ 0 & \delta & 0 & \delta & \ldots & 0 & \delta \\ -y_1 & x_1 & -y_2 & x_2 & \ldots & -y_p & x_p \\ x_1 & y_1 & x_2 & y_2 & \ldots & x_p & y_p \end{pmatrix}, \qquad \delta = 1/\sqrt{p},$$

orthonormal by rows, represents the instructions for getting to shape space from Cartesian pairs distributed as an isotropic Gaussian diffusion (of small variance) around a form $\mu = (x_1, y_1), (x_2, y_2), \ldots, (x_p, y_p)$ vectorized as $(x_1, y_1, x_2, y_2, \ldots, x_p, y_p)$ with $\Sigma x_i = \Sigma y_i = 0$, $\Sigma(x_i^2 + y_i^2) = 1$. Projecting out J gets us down to a rotation of Procrustes shape coordinates in tangent space. The projection as a function of μ is not isometric, meaning that the covariances of an initially isotropic Gaussian around different mean forms $\mu_1,\ \mu_2$ differ by relative eigenvalues in the ratios of $1 \pm |\mu_1 - \mu_2|$ in directions $\mu_1 \pm \mu_2$, whereas in the covering space — the space of forms, not shapes — the two embedded metrics are identical. I.e., the covering distributions (diffusions) around different means are exactly the same, but their normalizations into shape space have different geometries of projection. This is unhelpful for any science in which the shapes *per se* are not the objects of the appropriate subject-area theory. With size in the data, the representation is isotropic; once size is projected out, it ceases to be so. The situation is not like that of the "compositional data" methods for which were developed by Aitchison (1986) a generation ago — data that fundamentally arrive wholly without any dimension of scale. Rather, we *had* a scale, but we dismissed the information it bore. That is statistical malpractice.

As another example, consider the beautiful computational approach of Doug Theobald and Deborah Wuttke that was presented at this meeting in the same year of 2009. Theobald (2009) and Theobald and Wuttke (2008) were concerned with modeling the variation of a biological molecule under conditions of spatial anisotropy, namely, much more variation in certain parts of the molecule (free ends, basically) than in others (the core of interacting sites, which needed to be conservative). They invented a fully Bayesian method of modeling via hyperpriors for these distributions, and showed that with the aid of these hyperpriors they could much more accurately estimate the true mean of the underlying stable part of the protein. Such a computation, however, would not be meaningful if set in shape space. Change of size is not an option for a real molecule, so that the molecules being averaged can reasonably be subjected only to the isometric transformations of translation and rotation; they should not be rescaled. Though it was via words like "shape" that findings like these are reported, in fact the analysis had to go forward in terms of form, not in terms of shape.

That the call for contributions for this conference named "shape space" as one of its intended

domains is not the statement of the problem but one of its symptoms. Casting an investigation into shape space can actually interfere with knowledge discovery when it blurs the contingent character of the size-standardization built into shape space formalisms. The art historian might be willing to let size be decoupled from shape, but the natural scientist cannot afford to do so. The submersion of ordinary euclidean space that is our formalism of "shape space" is a reduction to equivalence classes that are often based on no good scientific reasoning. We are only now discovering how much harm this reduction does to underlying notions of physical or biophysical causality on which explanations in the corresponding sciences ultimately rest. Today, nearly *every* interesting paper about applications of shape space to organismal biology must consider allometry, and yet every such paper is inherently a refutation of the geometry Kendall originally suggested for it.

In other words, that shape space is non-euclidean is mainly up to the statistician's discretion, not Nature's — a discretion that has often been abused. In my current presentations of shape findings the thrust is always to embed them in the theoretically coherent euclidean spaces (local registration rules) from which they might plausibly have come. The more one projects out, the riskier the resulting shape space as the domain for any sort of scientific inference. Then analyses in Procrustes projective space, for instance, must be even more fraught than those in Kendall's spaces, as the origin of such data sets in equivalence classes that discard crucial factual data is at root a failure of instrumentation that irrevocably colors what would otherwise be more realistic (i.e., whiter) noise models. The statistician should not be offering to cope passively with such blunders, but instead should be demanding that the data be generated in the form of proper euclidean locations to begin with, even if this requires more expensive machines. We can wait.

Other sciences have analogous problems. One famous example from astronomy, for example, deals with the large-scale structure of the universe. For centuries the "celestial sphere" was just that, a metric sphere on which positions of stars and galaxies (forgive the anachronism) were projected. In these directional data every hint of the missing coordinate (radius from the earth) was gone. Patterns would emerge on this sphere, but there never was a corresponding constitutive theory. Then, beginning in the 1930's, it began possible to estimate that third coordinate, true distance from the earth, by exploiting the Hubble theory of red shifts. First a dozen galaxies were given back their third coordinate in this way, and then hundreds, and now, millions. The result was a revolution in our understanding of the universe: the emergence of large-scale structures at a scale of hundreds of millions of light-years, such as the "Great Wall" of galaxies reported by Geller and Huchra (1989). The physical universe is now understood to be a remarkable fabric of sheets and voids of galaxies, their form the ultimately physical embodiment of conditions soon after the Big Bang. Now that the universe is euclidean again, we can compute its statistics correctly.

Many other examples are like the astronomers', progress coming from the restoration of access to coordinates that had previously been obliterated by imperfect instruments or sheer human obstinacy. Röntgen's original imagery was two-dimensional, with the third dimension missing, but the discovery of inversion formulas for the sinogram permitted the reconstruction of absorption or emission densities back in our familiar three-dimensional euclidean space once again (the CT scans and MR scans that now fill our journals). The current state-of-the-art version of that image, the diffusion tensor image, is likewise a reduced representation of a euclidean tensor field, with six parameters per point of space, and so on.

Ultimately the reason for this privileging of the euclidean is the origin of our data in machines (Wigner, 1960). Machines are built by engineers and calibrated to accord with the laws

of physics, which (in our small neighborhood of spacetime) are ultimately euclidean. But this trope also traces the sums of squares that characterize our toolkit of statistical reasoning as a whole: the exponents of the Gaussian distributions that generate Fisher information, the Maxwell–Boltzmann distributions that describe physical noise.

In the example I discussed first, the interplay between form distance and strain energy, each one is euclidean on its own terms. It is from theinteraction between the two metrics that scientifically interesting patterns and predictions arise. In other settings, the issue of setting a physical scale needs to be replaced by the scientific insight that in some domains signal processing requires a complete invariance regarding scaling, not at the level of data but at the level of spatially localized feature analysis. In this connection, the best current approaches to principal components of form, in my view, are the explicitly self-similar calculations of Mardia et al. (2006) (exemplified in Bookstein (2007) in an application to the cortical midplane in schizophrenia). The method searches not for maxima of variance over linear combinations, the argument of Pearson's old argmax, but instead for maxima of coefficients of variation of locally supported ratios.

Other conflicts of metric arise just as naturally from the superposition of image information over structure as quantified by other means: for instance, the statistical analysis of deformable medical images bearing a gray scale, such as functional brain images. Both spaces are euclidean separately, and when represented that way their interaction is that of a tensor product; but when one is used to register the other, the geometry of their combined space becomes complex and unvisualizable (Bookstein, 2001), and, at present, every approach to this problem of fused data structures leads to a different pattern analysis. If there is to be a non-euclidean space for the representation of deformable medical images, we do not yet know how it should be built. An even deeper issue applies to the connection between embryology and evolution, which consists in analysis of the same image data set by more than one euclidean metric, more than one set of contrasts, at the same time.

I argue, therefore, contrary to the philosophy of statistics implicit in the call for this year's LASR, that to claim a problem is non-euclidean, at least, a problem arising in the natural sciences, might at root be simple human error. Perhaps it was an error by the scientist, but more likely it was an error of collaboration between the scientist and the statistician — an error of collaborative strategy, in which the statistician took perhaps too much pleasure in the subtlety of the manifolds that underlie these models without sufficiently respecting the machines that supplied the data to begin with, and the engineers who designed those machines. In such cases the role of the statistician ought to be to orchestrate the *return* to euclidean geometry, if that is the underlying physical model, by the most expeditious route. Contradicting the collaborator's insistence that the problem lives on a sphere, or a torus, or a network, our job is to reconstruct the originally euclidean operation of some machine, and thereby to explore the origin of the ultimate statistical similes, the sums of squares that constitute our version of information and its uncertainty. Our conferences need to remind today's new generation of statisticians about the centrality of the old euclidean models of machines. And in-between the scientific sessions, the visitors should be urged to visit the art museums, where they will be reminded that their task is to bridge the natural world to the symbolic world of pattern analysis with the least turbulence, the least deformation.

Acknowledgements. Preparation of these remarks was supported in part by research grant DEB–1019583 from the U. S. National Science Foundation to J. Felsenstein and F. Bookstein, University of Washington. As always I am grateful to Kanti Mardia and the other LASR over-

seers for these opportunities to speak in my most heartfelt, iconoclastic voice year after year.

References

Aitchison, J. *The Statistical Analysis of Compositional Data.* Chapman and Hall, 1986.

Bookstein, F.L. Size and shape spaces for landmark data in two dimensions. *Statistical Science* 1:181–242, 1986.

Bookstein, F.L. "Voxel-based morphometry" should never be used with imperfectly registered images. *NeuroImage* 14:1454–1462, 2001.

Bookstein, F.L. Morphometrics and computed homology: an old theme revisited. Pp. 69–81 in N. MacLeod, ed., *Proceedings of a Symposium on Algorithmic Approaches to the Identification Problem in Systematics,* Museum of Natural History, London, 2007.

Bookstein, F.L. For isotropic offset normal shape distributions, covariance distance is proportional to Procrustes distance. In A. Gusnanto et al., eds., *Proceedings of the 2009 Leeds Annual Statistical Research Workshop,* University of Leeds, 2009, pp. 47–51.

Bookstein, F.L. Speculations on two open problems in contemporary biometrics. In K. V. Mardia et al., eds., *New Statistics and Modern Natural Sciences,* University of Leeds, 2012, pp. 43–47.

Bookstein, F.L. *Reasoning and Measuring: Numerical Inference in the Sciences.* Cambridge University Press, to appear, 2013.

Bookstein, F.L. Allometry for the twenty-first century. *Biological Theory* 7:10-25, 2013a.

Bookstein, F.L. The relation between geometric morphometrics and functional morphology. Invited article for *Journal of Anatomy,* 2014.

Geller, M.J., and J. P. Huchra. Mapping the universe. *Science* 246:897–903, 1989.

Mardia, K. V., F.L. Bookstein, J. T. Kent, and C. R. Meyer. Intrinsic random fields and image deformations. *Journal of Mathematical Imaging and Vision* 26:59–71, 2006.

Theobald, D.L. A nonisotropic Bayesian approach for superpositioning multiple macromolecules. In A. Gusnanto et al., eds., *Statistical Tools for Challenges in Bioinformatics,* University of Leeds, 2009, pp. 55–59.

Theobald, D.L., and D. S. Wuttke. Accurate structural correlations from maximum likelihood superpositions. *PLoS Computational Biology* 4(2):e43, 2008.

Wigner, E. The unreasonable effectiveness of mathematics in the natural sciences. *Communications in Pure and Applied Mathematics* 13:1–14, 1960.