Holistic statistics and contemporary life sciences

Kanti V. Mardia

Department of Statistics, University of Leeds

1 Introduction

We begin here with my general philosophical quote:
**“Statistics without science is incomplete,
Science without statistics is imperfect."** - Mardia

If the last century in Science belongs to Physics then this century must belong to Life Sciences with many break-throughs starting from the DNA! Indeed, there has been a general stock taking by scientists and humanists with the beginning of the new millennium. Statisticians have not been the exception to this stock taking and there is now a stronger trend towards interdisciplinary research which breaks down the walls between traditional disciplines. In this new trend, it is worth noting what change in attitude is needed in statistical research and its dissemination. What computing technology will have an effect on our endeavours? How statistics will fit into the broad future aspirations of science, humanity and technology? What major challenges lie in unravelling the mystery of consciousness? I give here a personal view on these scenarios through examples from Life Sciences.

There is evidence of a great desire for the new and important ideas that drive our times where the highlight is on **life sciences and computer sciences** (see, for examples, Brockman, 2003 and Mardia and Gilks, 2005). Revolutionary developments include in molecular biology, genetic engineering, nanotechnology, artificial intelligence, complex adaptive systems, expert systems, the human genome, cellular automata, consciousness, immortality, ...

An organism’s DNA, including its genes, holds almost all the information required for its development and function. Human understanding of this information is at an early stage, but is accumulating rapidly due to new high-throughput forms of experimentation. This has led to large and rapidly expanding databases of DNA sequence, and related databases of the structure and function of biomolecules such as proteins. Bioinformatics is concerned with the development of these databases, and tools for deciphering and exploiting the information they contain.

With all the excitement generated by gene sequences, it is easy to forget that the primary purpose of most genes is to code for proteins. The proteins are biological macromolecules that are of primary importance to all living organisms. If gene sequencing is like the recording of music, then proteins are like the playback.

We emhasize why statisticians need a shift in paradigm if they want to help in resolving, for example, the nobel-prize-calibre problem of protein folding!

2 Holistic statistics

First we consider what one can mean by holistic science. The term holistic science has been used as a category encompassing a number of scientific research fields (http://en.wikipedia.org/wiki/Holism_in_science). The term may not have a precise definition. Fields of scientific research considered potentially holistic do however have certain things in common. First, they are multidisciplinary. Second, they are concerned with the behavior of complex systems. Third, they recognize feedback within systems as a crucial element for understanding their behavior.

The Santa Fe Institute, a centre of holistic scientific research in the United States, expresses it like this: “The two dominant characteristics of the SFI research style are commitment to a multidisciplinary approach and an emphasis on the study of problems that involve complex interactions among their constituent parts."

The area of life sciences is vast and has all kind of new topics such as proteomics, genomics, bioinformatics, system biology and now the new emerging field of synthetic biology. On the other hand, statistical methodology is also vast encompassing so many different specialist areas which could be relevant to life sciences. At least in protein structure, we have our experience that the two subjects of “shape analysis and directional statistics", which are not standard topics of statistics but can revolutionise protein structure. We will give in this paper a few examples.

Mardia and Gilks (2005) have identified three themes for statistics in the 21st century.

- First, statistics should be viewed in the **broadest way** for scientific explanation or prediction of any phenomenon.
- Second, the future of statistics lies in a **holistic approach** to interdisciplinary research.
- Third, a change of attitude is required by statisticians - **a paradigm shift** - for the subject to go forward.

Thus there is a tremendous opportunity for statistics to make a stronger impact in the subject of Bioinformatics/Life Sciences, but statisticians need to be more **open**, more ready to **learn** "molecular biology", more computationally **aware**, and more ready to understand **data banks**.

Holistic statistics is in contrast to the parable of six blind men examining an elephant where the blind men can comprehend only its parts so to understand what kind of the object the elephant is, we must look at it from all sides (Anekaantvaad principle of the Jain logic). Holistic statistics aims to work in totality - researching all the relevant subjects and dealing with the multi-facet aspects of “truth"! Thus it has to be interdisciplinary!!

3 Shape analysis

3.1 Many faces of protein and protein folding

Proteins are the work horses of all living systems. Unfortunately, the proteome is much more complicated than the genome. Furthermore, the basis of protein function is not only chemistry (as with genes) but also shapes! Indeed, there are various ways to look at a protein (a top-down view is as follows)

1. 3D-Coordinates of atoms (**tertiary “structure"**).
2. Equivalently, a set of **dihedral angles**/conformational angles.
3. A **curve** in 3-D with 500-2000 atoms along it, folded as a ball (**backbone/main chain**).
4. The **backbone** (the curve) having “branches" sticking out (branches = **side chains**).
5. Various shapes (sub-curves) are linked together along the curve as **helices, loops, strands** . . . (**secondary structure**).
6. A sequence of amino acids (out of 20) arranged linearly (**primary structure**).

   In addition, one can view it as

7. Surface in 3-D,
8. Dynamic object, . . .

There are many geometrical constraints due to chemistry of the atoms and their interaction with other proteins. Protein folding problem is the **central problem** in computational biology. The question is how the amino acids sequence encodes a (compact) **3D shape, or fold**. One can visualise a protein as snakes which are transferred into a basket where not only they curl in a particular way but also allow each other to breathe. Of course this is a very simple analogy.

3.2 Definitions

3.2.1 Labelled shape/form

Just to remind (see, for example, Dryden and Mardia, 1998) that objects are everywhere - natural and man made. The study of their shape is inherently non-Euclidean and recent advancements in the fields of shape statistics have been motivated by image analysis, morphometrics, etc., and most recently by statistical bioinformatics.

Shape (similarity) is the description of objects after ignoring changes in location, scale, and rotation where objects can be described in terms of, for example, by landmarks.

For proteins, **form** is really meaningful; in form one ignores changes in **location, rotation** but **not scale**. The major developments in statistical shape analysis has been mainly for similarity shape but now form analysis is picking up, see for example, Theobald and Wuttke (2006).

3.2.2 Unlabelled shape analysis

The problem of unlabelled shape analysis is as follows. Here the aim is to align two or more configuration points in space, and to make inference about the geometrical transformations, which maps one configuration to another. Thus there is additional complication of matching only subsets of configurations. Possibly there may also be partial labelling information which will constrain the matching. The aim is to find the largest common point set under a plausible model.

3.2.3 Historical comments

Unknown to statisticians, RMSD (Root Mean Squares Derivation) has been used by bioinformaticians which has a flavour of Procrustes analysis; the two communities have been working independently until very recently. In fact the history of RMSD goes quite back. It at least starts from Green (1952) and Kabsch (1978), but it was popularised by John Gower in statistics which led to the first book appearance in Mardia et al (1979). Note that Kabsch (1978) is the usual reference used in crystallography/bioinformatics, whereas Green (1952) is the standard reference in statistics. The term Procrustes was introduced by Hurley and Catell (1962). A graphic description is given in the frontispage of Gower and Dijksterhuis (2004):

“Procrustes (the subduer), son of Poseidon, kept an inn benefiting from what he claimed to be a wonderful all-fitting bed. He lopped off excessive limbage from tall guests and either flattened short guests by hammering or stretched them by racking. The victim fitted the bed perfectly but, regrettably, died. To exclude the embarrassment of an initially exact-fitting guest, variants of the legend allow Procrustes two, different-sized, beds. Ultimately, in a crackdown on robbers and monsters, the young Theseus fitted Procrustes to his own bed."
For further historical details, see Gower and Dijksterhuis (2004, Table 1.1, p.5).

3.3 Homology

Protein structure prediction is one of the important applications of bioinformatics. The amino acid sequence of a protein can be easily determined from the sequene on the gene that codes for it. In the vast majority of cases, this sequence uniquely determines a structure (3D information) in its native environment. Knowledge of this structure is vital in understanding the function of the protein. One example of this is the similar protein homology between haemoglobin in humans and the haemoglobin in legumes (leghaemoglobin). Both serve the same purpose of transporting oxygen in the organism. Though both of these proteins have completely different amino acid sequences, their protein structures are virtually identical, which reflects their near identical purposes.

Homology is used to determine which parts of a protein are important in structure formation and interaction with other proteins by “matching" protein A whose function is known, to “protein B" whose function is unknown. Phylogenetic information is also critical. In homology modelling, this information is used to predict the structure of a protein once the structure of a homologous protein is known. This currently remains the only way to predict protein structures reliably (see also Section 4.3 below).

3.4 The form model

3.4.1 Labelled form

Let X be a random configuration matrix (k landmarks in m dimensions) then the perturbation model for “form" (shape-size) with observations $X_1, \ldots, X_n$ is

$$X_i = (\mu + E_i)\Gamma_i + 1\gamma_i^T, \quad i = 1, \ldots, n$$

where Γ_i is a rotation matrix, γ_i is a translation vector and E_i is an error distribution (with zero mean) and μ is the mean form. E_i can have for example multivariate Gaussian matrix distribution (ie. for stacked E_i to assume the full multivariate normal distribution). Goodall and Mardia (1993) have studied this model fully with the distributional properties (cf. Goodall, 1991). The simplest form will have an isotropic covariance function for E. Next stage could be a tensor structure for the covariance matrix but it is arguable whether this tensor structure can have a better robustness property then the isotropic covariance matrix when the model is false (cf. Theobald and Wuttke, 2006). For planar shape analysis, a model (complex Bingham quartic distribution) which allows full covariance matrix is given by Kent et al (2006).

3.4.2 Unlabelled form

A Bayesian hierarchical model (Green and Mardia, 2006) is defined as follows:

$$x_j \sim N_d(\mu_{\xi_j}, \sigma^2 I), \quad Ay_k + \tau \sim N_d(\mu_{\eta_k}, \sigma^2 I)$$

where μ are hidden variables, A is a rotation matrix, τ is a translation vector and M is the matching matrix. We are led to the joint posterior density $p(M, A, \tau, \sigma, x, y)$, prop. to

$$|A|^n p(A)p(\tau)p(\sigma) \times \prod_{j,k:M_{jk}=1} \left(\frac{\rho\phi(\{x_j - Ay_k - \tau\}/\sigma\sqrt{2})}{\lambda(\sigma\sqrt{2})^d} \right)$$

where $\phi(.)$ is the normal density for $N_d(0, I_d)$ and ρ/λ is the matching parameter. This forms the basis of inference about M, A, τ and σ^2; MCMC implementation has been developed. The model can be connected to the RMSD which is analogous to the relationship between the least squares and the Gauss-Markov model (see Mardia et al, 2007a). For an extension to multiple configurations, see Ruffieux and Green (2008). Another model with “coffin bin" formulation is given in Kent et al (2008a); their implementation is through the EM algorithm.

3.5 The two mysterious proteins

In LASR 2007, we presented (Mardia, 2007) structures of two mysterious proteins by a drug company without knowing what they were. The aim was to give our assessment on how similar these were. From the GM algorithm (Green and Mardia, 2006), we concluded that these were very similar. Indeed, out of 551 (x) and 552 (y) points (C_α atoms), only one from x did not match y. The posterior probability was extremely high. In fact, Anna Tramontano solved the mystery.

The two proteins are cyclooxygenase (COX). The first from **sheep**, the second from **mouse**. So these are expected to be very similar!! COX is an enzyme (EC 1.14.99.1) that is responsible for information of important biological mediators called prostanoids. Pharmacological inhibition of COX can provide relief from the symptoms of inflammation and pain; this is the method of action of well-known drugs such as aspirin and ibuprofen.

4 Directional statistics

4.1 The Ramachandran plot

The Ramachandran plot gives the scatter plot of the dihedral angles ϕ and ψ for alpha helices, beta strands, loops, etc. of proteins (Ramachandran et al, 1963), ie. it is really not simply a plot but it analyses the empirical distribution of the secondary structures of protein. Their importance is summarized by Rose (2001) as follows:

> **"No biochemistry textbook is complete without a phi, psi-plot ...** This plot ranks alongside the double helix and the alpha-helix among fundamentals of structural biochemistry".

These (ϕ, ψ) lie on a torus and statistical models on a torus (ϕ, ψ) started from Mardia (1975) but the work in this area continues to grow (see Section 4.2.1). However, until very recently directional statisticians were not aware of the Ramachandran plot!

4.2 Directional statistics

Directional statistics deals in particular with angular observations (directions), eg. we can consider wind bearing as points on the circle of unit radius with centre at the origin (see, for example, Mardia and Jupp, 2000). The first step in statistics is to define sensible mean and deviation. Then the next step is to find a plausible normal type model. In this case there are traps if we do not allow the circular space eg. arithmetic mean of 1^0 and 359^0 is 180 which shows clearly that linear methods are not applicable. The subject has been developing fast with new distributions on torus and their applications.

4.2.1 A multivariate distribution on the Torus

The bivariate case has been described in this Proceeding (Kent et al, 2008b). We define a multivariate angular distribution (Mardia et al, 2008) which is an extension of the sine model (Singh et al, 2002) as follows. The probability density function of $\boldsymbol{\Theta}^T = (\Theta_1, \Theta_2, \ldots, \Theta_p)$ is given by

$$\{T(\boldsymbol{\kappa}, \boldsymbol{\Lambda})\}^{-1} \exp\{\boldsymbol{\kappa}^T c(\boldsymbol{\theta}, \boldsymbol{\mu}) + \frac{1}{2} s(\boldsymbol{\theta}, \boldsymbol{\mu})^T \Lambda \quad s(\boldsymbol{\theta}, \boldsymbol{\mu})\},$$

where $-\pi < \theta_i \leq \pi, -\pi < \mu_i \leq \pi, \kappa_i \geq 0, -\infty < \lambda_{ij} < \infty, \boldsymbol{\kappa}^T = (\kappa_1, \ldots, \kappa_p)$,

$$c(\boldsymbol{\theta}, \boldsymbol{\mu})^T = (\cos(\theta_1 - \mu_1), \ldots, \cos(\theta_p - \mu_p)), \quad s(\boldsymbol{\theta}, \boldsymbol{\mu})^T = (\sin(\theta_1 - \mu_1), \ldots, \sin(\theta_p - \mu_p)),$$

and $(\boldsymbol{\Lambda})_{ij} = \lambda_{ij} = \lambda_{ji}, \quad i \neq j, \quad \lambda_{ii} = 0$, with $\{T(\boldsymbol{\kappa}, \boldsymbol{\Lambda})\}^{-1}$ a normalizing constant. We call this the multivariate von Mises density. Note that for $p = 1$, this is a univariate von

Mises density and for $p = 2$, this density corresponds to the bivariate sine model. For large concentrations in the circular variables, we have ($\boldsymbol{\mu} = \mathbf{0}$ without any loss of generality).

$$\boldsymbol{\Theta} = (\Theta_1, \Theta_2, \ldots, \Theta_p)^T \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}^{-1}), \quad \text{where } (\boldsymbol{\Sigma}^{-1})_{ii} = \kappa_i,\ (\boldsymbol{\Sigma}^{-1})_{ij} = -\lambda_{ij},\ i \neq j.$$

Some bioinformatics applications are given in Mardia et al (2007b) and Mardia et al (2008).

4.3 Protein structure prediction

The aim is to find the native state of a protein given its amino acid sequence. The state-of-the-art method for enforcing local structure today is due to the Baker group (eg. Bystroff and Baker, 1998). The procedure is to select representative short fragments from the PDB (Protein Data Bank)and build new structures by merging together fragments. All constructed proteins have reasonable backbone angles while significantly reducing the conformational search space. However, the previous work uses discrete represenation for the angles, and the construction does not seem to lead to statistical interpretations.

Our main idea (Boomsma et al, 2008) is to input fragments and to use a plausible statistical model with a continuous bivariate torus model to output the full range of backbone angles ϕ and ψ; this output of (ϕ, ψ) for a given input sequence of amino acids turns out to be plausible. The method, for example, allows (1) simulation of structures given a sequence, (2) drawing of consistent samples of part a structure and (3) calculation of the probability of a resampled segment, compared to the original. In fact some of these ideas are used in a program called Phaistos (`http://sourceforge.net/projects/phaistos/`) which is a collection tools for protein structure prediction. It currently features the FB5DBN (Hamelryck et al, 2006) and TorusDBN (Boomsma et al, 2008) models, which makes it possible to sample protein structures compatible with a given amino acid and/or secondary structure sequence.

5 Discussion

It is expected that any cross-fertilization should lead to a better science and we have given some examples in protein bioinformatics (for RNA structure, see Frellsen et al, 2008). Reflecting back in my joint book on shape analysis, we missed out any reference to life sciences including RMSD; similarly for my joint book on directional statistics, we missed out the Ramachandran plot though these books have been published in 1998 and 2000 respectively. However, with our collaborative activities including those with the Copenhagen School, we hope to develop these statistical subjects further motivated in particular by protein structure, and by synthetic biology in general.

Acknowledgements

I wish to thank Wouter Boomsma, Clive Bowman, Jes Frellsen, Peter Green, Thomas Hamelryck, John Kent, Vysaul Nyirongo and Anna Tramontano for the material in this paper.

References

Boomsma, W., Mardia, K.V., Taylor, C.C., Ferkinghoff-Borg, J., Krogh, A. and Hamelryck, T. (2008). A generative, probabilistic model of local protein structure. *PNAS*, in press.

Brockman, J. (ed) (2003) *The New Humanists - Science at the Edge*, New York: Barnes and Noble.

Bystroff, C. and Baker, D. (1998). Prediction of local structure in proteins using a library of sequence-structure motifs. *J. Mol. Biol.*, **281**, 565–577.

Dryden, I.L. and Mardia, K.V. (1998). *Statistical Shape Analysis*, Wiley.

Frellsen, J., Moltke, I., Thiim, M., Mardia, K.V., Ferkinghoff-Borg, J. and Hamelryck, T. (2008). A probabilistic model of local RNA 3-D structure. Submitted.

Goodall, C.R. (1991). Procrustes methods in the statistical analysis of shape (with discussion). *Journal of the Royal Statistical Society, Series B*, **53**, 285–339.

Goodall, C.R. and Mardia, K.V. (1993). Multivariate aspects of shape theory. *Ann.Statist.*, **21**, 848–866.

Gower, J.C. and Dijksterhuis, G.B. (2004). *Procrustes Problems*. Oxford University Press.

Green, B.F. (1952). The orthogonal approximation of an oblique structure in factor analysis. *Psychometrika*, **17**, 429–440.

Green, P.J. and Mardia, K.V. (2006) Bayesian alignment using hierarchical models, with applications in protein bioinformatics. *Biometrika*, **93**, 235–254.

Hamelryck, T., Kent, J., Krogh, A. (2006). Sampling realistic protein conformations using local structural bias, *PLoS Comput. Biol.*, **2**, 1121–1133.

Hurley, J.R. and Cattell, R.B. (1962). The Procrustes program: producing direct rotation to test a hypothesized factor structure. *Beh. Sci.*, **7**, 258–262.

Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Cryst.*, **A34**, 827–828.

Kent, J.T., Mardia, K.V. and McDonnell, P. (2006). The complex Bingham quartic distribution and shape analysis. *Journal of Royal Statistical Society Series B*, **68**, 747–765.

Kent, J.T., Mardia, K.V. and Taylor, C.C. (2008a). Bioinformatics and the problem of matching unlabelled configurations. Submitted.

Kent, J.T., Mardia, K.V. and Taylor, C.C. (2008b). Modelling strategies for bivariate circular data. In: The Art and Science of Statistical Bioinformatics. Leeds, Leeds University Press.

Mardia, K.V. (1975). Statistics of directional data (with discussion). *J. Roy. Statist. Soc. Ser. B.*, **37**, 349–393.

Mardia, K.V. (2007). On some recent advancements in applied shape analysis and directional statistics. In S. Barber, P.D. Baxter, & K.V.Mardia (eds), *Systems Biology & Statistical Bioinformatics*, pp.9–17 . Leeds, Leeds University Press.

Mardia, K.V. and Gilks, W. (2005). Meeting the statistical needs of 21st-century science. *Significance*, **2**, 162–165.

Mardia, K.V., Hughes, G., Taylor, C.C. and Singh, H. (2008). Multivariate von Mises distribution with applications to bioinformatics. *Canadian Journal of Statistics*, **36**, 99–109.

Mardia, K.V. and Jupp, P.E. (2000). *Directional Statistics*. Wiley.

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979). *Multivariate Analysis*, Academic Press, New York.

Mardia, K.V., Nyirongo, V.B., Green, P.J., Gold, N.D. and Westhead, D.R. (2007a). Bayesian refinement of protein functional site matching. *BMC Bioinformatics*, **8:257**. http://www.biomedcentral.com/qc/1471-2105/8/257.

Mardia, K.V., Taylor, C.C., and Subramaniam, G.K. (2007b) Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data. *Biometrics*, **63**, 505–512.

Ramachandran, G.N., Ramakrishnan, C. and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *Molecular Biology*, **7**, 95–99.

Rose, G.D. (2001). In Memoriam: Professor G.N. Ramachandran (1922-2001). *Protein Science*, **10**, 1691–1692.

Ruffieux, Y. and Green, P.J. (2008). Alignment of multiple configurations using hierarchical models. Submitted http://www.stats.bris.ac.uk/ peter/papers/MAlign.pdf

Singh, H., Hnizdo, V., Demchuk, E. (2002). Probabilistic model for two dependent circular variables. *Biometrika*, **89**, 719–723.

Theobald, D.L. and Wuttke, D.S. (2006). Empirical Bayes hierarchical models for regularizing maximum likelihood estimation in the matrix Gaussian Procrustes problem. *Proceedings of the National Academy of Sciences*, **103**, 18521–18527.