LASR 2011 — Next Generation Statistics in Biosciences

30th Anniversary Souvenir Issue

Programme, special papers and abstracts

Edited by K.V. Mardia, A. Gusnanto, A.D. Riley & J. Voss

5th to 7th July 2011
The reference ratio method applied to sample compact protein structures. We used the amino acid sequence of ubiquitin. The left picture shows 20 non-compact samples from TorusDBN alone, $g(x)$, which is a probabilistic model of protein structure on a local length scale. The right picture shows 20 compact samples from the reference ratio distribution, $\hat{f}(x)$, which is obtained by combining TorusDBN, $g(x)$, with a suitable normal distribution, $f_1(y)$, over the radius of gyration. In the expressions $x$ denotes the sequence of dihedral angles in the protein backbone, $y = m(x)$ denotes the corresponding radius of gyration and $g_1(y)$ is the marginal for TorusDBN. See p. 55 for details. Figure by Jes Frellsen and Mikael Borg.
Next Generation Statistics in Biosciences

International Conference, held in Leeds, UK, 5-7 July 2011,
The 30th Leeds Annual Statistical Research (LASR) Workshop
Sponsored and organised by the Department of Statistics, University of Leeds.
Sponsored by EPSRC

Proceedings edited by
K.V. Mardia, A. Gusnanto, A.D. Riley & J. Voss
Department of Statistics,
University of Leeds, Leeds LS2 9JT.

Conference Organisers
A. Gusnanto, K.V. Mardia (chairman), J.M. Brennan, J. Voss, A.D. Riley,

Copyright © 2011 K.V. Mardia, A. Gusnanto, A.D. Riley & J. Voss,
Department of Statistics, University of Leeds, U.K.
ISBN 978 0 85316 302 2
Dear Delegate,

Welcome to the 30th anniversary of Leeds Annual Statistical Research Workshop. The history of the Workshop is described at some length on pages 9-15 of our LASR Proceedings of the Silver Jubilee, 2006. Here, we note again that LASR has grown considerably since the first workshop in 1974, and has changed dramatically in that time. What was originally an internal workshop has expanded into a conference of which we at Leeds are proud. We are glad that you are all here in Leeds to help us celebrate another landmark. We have used this attractive new venue (Hinsley Hall) since 2006. This was a monastery converted into a congenial and inspiring conference centre and has really provided a central setting which has kept our interdisciplinary spirit buoyant!

Note that the 30th Anniversary represents a landmark in our contributions to pursuing interdisciplinary (emerging) areas of science: the first decade was Geosciences, the second Image Analysis and third decade (now) Molecular Biology, a great boost for the next generation. There is always great pleasure but also associated pain in interdisciplinary work, and LASR has been unique its ambition and execution. We are taking special steps to encourage the participation of young people, women, and other minorities in science; to encourage new collaborations, we have tried to keep the Workshop informal and to provide opportunities for discussion during the breaks and poster sessions. The LASR mission is to:

• bring together world-class scientists and statisticians;

• encourage cutting-edge interdisciplinary research;

• sustain and develop our workshops;

• and, currently, to showcase new, innovative statistical methods and applications in structural bioinformatics.

Figure 1: David Kendall (left) in 1979, planting the seed for the LASR Workshops.
As part of our celebrations, a special feature of this year’s Proceedings is a series of special articles from eminent figures who could not come but were invited to prepare forward looking articles in Statistics as well as Molecular Biology, and we wish to express our thanks to Sir David Cox, Prof. Bradley Efron, Prof. Michael Levitt, Prof. Simon Phillips, Prof. Terry Speed and Prof. Anna Tramontano for these articles. I am pleased to say that there are many other articles from the participants also giving future directions.

There are some unifying themes in these predictions. Some statistical principles are eternal, for example, randomization, replications, controls, designs, Bayes rule, probabilistic modelling and robustness. But specific methodologies will be driven by new technologies related to large scale data. Technological advances have been producing a flood of large data sets (and this flood will increase); in Life Sciences there are areas such as High Throughput, Genomic and Proteomic. These large data sets have led to massive data analytic problems. They raise many questions, and thousands of estimates or hypothesis tests. Indeed, new difficulties arise and can easily lead to flawed inferences. Bayesian methods have become a key tool for large scale data and regularisation. In this setting Bayesian methods have become increasingly important, especially to regularize problems with large numbers of parameters. Empirical Bayes may have a better future as an approach to statistics that lies somewhere between the frequentist and Bayesian approach. Indeed, modern computer technology has been accompanied by renaissance in Bayesian methods. However, the new statistical methods have to be fast as one is dealing with very large data sets in real applications, and will be more and more in real time. The future of Bioscience is a very attractive area for statisticians interested in contributing to the most challenging project ever undertaken by science, that of understanding ourselves, our past and our future. Holistic Statistics is vital for this exciting future.

![Figure 2: At the 2nd LASR Workshop, on MRF, in 1981.](image)

Another key feature of this Workshop is a celebration of David Kendall’s contributions to shape. David died on 23rd October 2007, and the second day is dedicated to shape in his memory. He was a great source of strength and inspiration for me and many others in ways that no words can explain. The success of our LASR workshops owes a great deal to his contributions. In fact, he suggested in 1979 that the workshops become a regular annual event, and he continued to support our shape workshops over the years. Indeed, the first printed LASR Proceedings (1995) were dedicated jointly to him.
Last year, we lost another friend, Julian Besag, who died on 6th August 2010. He conducted our LASR workshop in 1981, 30 years ago on Markov Random Fields! At that time, the format of the Workshop meant that the speaker had to give an instructional course for 3 days, and he really gave a thorough review. With his workshop here in 2000 and 2006, as usual he had new topics, in both workshops generating a great discussion on both these occasions. A memorial event was held for him in Bristol in April 2011, the website for which is http://www.sustain.bris.ac.uk/JulianBesag/tributes/.

We wish to record our sincerest gratitude to our sponsors the EPSRC for their valuable support to this workshop. Looking forward, we are already planning for LASR 2012. Details are still taking shape, but for now, we hope that you will enjoy this year’s LASR Workshop, and we would greatly appreciate your feedback.

Kanti Mardia
July 2011
<table>
<thead>
<tr>
<th>Conference</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geostatistics (Conference)</td>
<td>1979</td>
</tr>
<tr>
<td>Prof. D.G. Kendall,</td>
<td></td>
</tr>
<tr>
<td>Prof. A. Marechal</td>
<td></td>
</tr>
<tr>
<td>Spatial Processes on Regular Lattices</td>
<td>1980</td>
</tr>
<tr>
<td>Dr. R.J. Martin</td>
<td></td>
</tr>
<tr>
<td>Applied Spatial Processes</td>
<td>1981</td>
</tr>
<tr>
<td>Prof. J. Besag</td>
<td></td>
</tr>
<tr>
<td>Spatial Statistics</td>
<td>1982</td>
</tr>
<tr>
<td>Prof. B.D. Ripley</td>
<td></td>
</tr>
<tr>
<td>Estimation Procedures for Stationary Spatial Processes</td>
<td>1983</td>
</tr>
<tr>
<td>Prof. X. Guyon</td>
<td></td>
</tr>
<tr>
<td>Statistics of Image Processing (Conference)</td>
<td>1984</td>
</tr>
<tr>
<td>Prof. J.V. Kittler</td>
<td></td>
</tr>
<tr>
<td>Bispectral Analysis and Bilinear Time Series Models</td>
<td>1985</td>
</tr>
<tr>
<td>Prof. T. Subba-Rao</td>
<td></td>
</tr>
<tr>
<td>Statistics and Image Processing</td>
<td>1986</td>
</tr>
<tr>
<td>Prof. J. Haslett</td>
<td></td>
</tr>
<tr>
<td>Markov Random Fields</td>
<td>1987</td>
</tr>
<tr>
<td>Prof. H.R. Künsch</td>
<td></td>
</tr>
<tr>
<td>Experimental Design Package</td>
<td>1988</td>
</tr>
<tr>
<td>Dr. A. Baines</td>
<td></td>
</tr>
<tr>
<td>Itô Calculus</td>
<td>1989</td>
</tr>
<tr>
<td>Prof. W.S. Kendall</td>
<td></td>
</tr>
<tr>
<td>Statistical Image Analysis</td>
<td>1990</td>
</tr>
<tr>
<td>Prof. C. Jennison</td>
<td></td>
</tr>
<tr>
<td>Shape Analysis, Landmarks and Biomedical Imaging</td>
<td>1991</td>
</tr>
<tr>
<td>Prof. F.L. Bookstein</td>
<td></td>
</tr>
<tr>
<td>Shape and Image Analysis for Industry and Medicine</td>
<td>1992</td>
</tr>
<tr>
<td>Prof. C.R. Goodall</td>
<td></td>
</tr>
<tr>
<td>Shape Recognition in 2D and 3D Imaging</td>
<td>1993</td>
</tr>
<tr>
<td>Prof. M.I. Miller</td>
<td></td>
</tr>
<tr>
<td>Medical Imaging: Shape and Visualization</td>
<td>1994</td>
</tr>
<tr>
<td>Prof. S. Pizer</td>
<td></td>
</tr>
<tr>
<td>Current Issues in Statistical Shape Analysis</td>
<td>1995</td>
</tr>
<tr>
<td>Prof. F.L. Bookstein</td>
<td></td>
</tr>
<tr>
<td>Prof. J. Gower,</td>
<td></td>
</tr>
<tr>
<td>Prof. P.J. Green,</td>
<td></td>
</tr>
<tr>
<td>Prof. D.G. Kendall</td>
<td></td>
</tr>
<tr>
<td>Image Fusion and Shape Variability Techniques</td>
<td>1996</td>
</tr>
<tr>
<td>Dr. A. Guéziec,</td>
<td></td>
</tr>
<tr>
<td>Prof. V.E. Johnson,</td>
<td></td>
</tr>
<tr>
<td>Prof. J.J. Koenderink,</td>
<td></td>
</tr>
<tr>
<td>Prof. D. Terzopoulos</td>
<td></td>
</tr>
<tr>
<td>The Art and Science of Bayesian Image Analysis</td>
<td>1997</td>
</tr>
<tr>
<td>Prof. U. Grenander</td>
<td></td>
</tr>
<tr>
<td>Prof. V.E. Johnson,</td>
<td></td>
</tr>
<tr>
<td>Prof. M. Titterington</td>
<td></td>
</tr>
<tr>
<td>Spatial-Temporal Modelling and its Applications</td>
<td>1999</td>
</tr>
<tr>
<td>Prof. A. Blake,</td>
<td></td>
</tr>
<tr>
<td>Sir David Cox,</td>
<td></td>
</tr>
<tr>
<td>Prof. D.C. Hogg,</td>
<td></td>
</tr>
<tr>
<td>Prof. C. Taylor</td>
<td></td>
</tr>
<tr>
<td>The Statistics of Directions, Shapes and Images</td>
<td>2000</td>
</tr>
<tr>
<td>Prof. J. Besag,</td>
<td></td>
</tr>
<tr>
<td>Prof. W.S. Kendall,</td>
<td></td>
</tr>
<tr>
<td>Prof. J.V. Kittler,</td>
<td></td>
</tr>
<tr>
<td>Prof. A. Rosenfeld</td>
<td></td>
</tr>
<tr>
<td>Functional and Spatial Data Analysis</td>
<td>2001</td>
</tr>
<tr>
<td>Prof. A. Baddeley,</td>
<td></td>
</tr>
<tr>
<td>Prof. M. Berman,</td>
<td></td>
</tr>
<tr>
<td>Prof. B.W. Silverman,</td>
<td></td>
</tr>
<tr>
<td>Prof. C.K. Wikle,</td>
<td></td>
</tr>
<tr>
<td>Prof. K. Worsley</td>
<td></td>
</tr>
<tr>
<td>The Statistics of Large Datasets: Functional and image data, bioinformatics and data mining</td>
<td>2002</td>
</tr>
<tr>
<td>Prof. G.J. Barton,</td>
<td></td>
</tr>
<tr>
<td>Prof. L. Edler,</td>
<td></td>
</tr>
<tr>
<td>Dr M.J.K Gales,</td>
<td></td>
</tr>
<tr>
<td>Prof. D.J. Hand</td>
<td></td>
</tr>
<tr>
<td>Stochastic Geometry, Biological Structure and Images</td>
<td>2003</td>
</tr>
<tr>
<td>Prof. C. Bystrøff,</td>
<td></td>
</tr>
<tr>
<td>Dr W. Gilks,</td>
<td></td>
</tr>
<tr>
<td>Prof. J. Hein,</td>
<td></td>
</tr>
<tr>
<td>Prof. A. Lesk,</td>
<td></td>
</tr>
<tr>
<td>Prof. D. Stoyan</td>
<td></td>
</tr>
<tr>
<td>Bioinformatics, Images and Wavelets</td>
<td>2004</td>
</tr>
<tr>
<td>Prof. M.J.E. Sternberg,</td>
<td></td>
</tr>
<tr>
<td>Prof. G.P. Nason</td>
<td></td>
</tr>
<tr>
<td>Quantitative Biology, Shape Analysis, and Wavelets</td>
<td>2005</td>
</tr>
<tr>
<td>Prof. P.J. Green,</td>
<td></td>
</tr>
<tr>
<td>Prof. V.E. Johnson,</td>
<td></td>
</tr>
<tr>
<td>Sir J. Kingman,</td>
<td></td>
</tr>
<tr>
<td>Interdisciplinary Statistics and Bioinformatics</td>
<td>2006</td>
</tr>
<tr>
<td>Prof. B.D. Athey,</td>
<td></td>
</tr>
<tr>
<td>Prof. J. Besag,</td>
<td></td>
</tr>
<tr>
<td>Sir D.R. Cox,</td>
<td></td>
</tr>
<tr>
<td>Systems Biology &amp; Statistical Bioinformatics</td>
<td>2007</td>
</tr>
<tr>
<td>Prof. J.M. Goodfellow</td>
<td></td>
</tr>
<tr>
<td>Prof. S. Tavaré,</td>
<td></td>
</tr>
<tr>
<td>Prof. A. Tramontano,</td>
<td></td>
</tr>
<tr>
<td>Prof. G.M. Crippen</td>
<td></td>
</tr>
<tr>
<td>The Art and Science of Statistical Bioinformatics</td>
<td>2008</td>
</tr>
<tr>
<td>Prof. M. Levitt,</td>
<td></td>
</tr>
<tr>
<td>Prof. Z. Yang,</td>
<td></td>
</tr>
<tr>
<td>Statistical Tools for Challenges in Bioinformatics</td>
<td>2009</td>
</tr>
<tr>
<td>Prof. A. Trouvé</td>
<td></td>
</tr>
<tr>
<td>Prof. S. Richardson</td>
<td></td>
</tr>
<tr>
<td>Prof. D. Wild</td>
<td></td>
</tr>
<tr>
<td>Dr. W. Taylor</td>
<td></td>
</tr>
<tr>
<td>Dr. D. Theobald</td>
<td></td>
</tr>
<tr>
<td>High-throughput Sequencing, Proteins and Statistics</td>
<td>2010</td>
</tr>
<tr>
<td>Prof. T. Speed</td>
<td></td>
</tr>
<tr>
<td>Prof. E. Wit</td>
<td></td>
</tr>
<tr>
<td>Prof. G. McLachlan</td>
<td></td>
</tr>
<tr>
<td>High-throughput Sequencing, Proteins and Statistics</td>
<td>2011</td>
</tr>
<tr>
<td>Prof. Y. Moulin</td>
<td></td>
</tr>
<tr>
<td>Prof. J. Gower,</td>
<td></td>
</tr>
<tr>
<td>Prof. P.J. Green,</td>
<td></td>
</tr>
<tr>
<td>Interdisciplinary Statistics and Bioinformatics</td>
<td>2012</td>
</tr>
<tr>
<td>Prof. B.D. Athey,</td>
<td></td>
</tr>
<tr>
<td>Prof. J. Besag,</td>
<td></td>
</tr>
<tr>
<td>Sir D.R. Cox,</td>
<td></td>
</tr>
<tr>
<td>Systems Biology &amp; Statistical Bioinformatics</td>
<td>2013</td>
</tr>
<tr>
<td>Prof. J.M. Goodfellow</td>
<td></td>
</tr>
<tr>
<td>Prof. S. Tavaré,</td>
<td></td>
</tr>
<tr>
<td>Prof. A. Tramontano,</td>
<td></td>
</tr>
<tr>
<td>Prof. G.M. Crippen</td>
<td></td>
</tr>
<tr>
<td>The Art and Science of Statistical Bioinformatics</td>
<td>2014</td>
</tr>
<tr>
<td>Prof. M. Levitt,</td>
<td></td>
</tr>
<tr>
<td>Prof. Z. Yang,</td>
<td></td>
</tr>
<tr>
<td>Statistical Tools for Challenges in Bioinformatics</td>
<td>2015</td>
</tr>
<tr>
<td>Prof. A. Trouvé</td>
<td></td>
</tr>
<tr>
<td>Prof. S. Richardson</td>
<td></td>
</tr>
<tr>
<td>Prof. D. Wild</td>
<td></td>
</tr>
<tr>
<td>Dr. W. Taylor</td>
<td></td>
</tr>
<tr>
<td>Dr. D. Theobald</td>
<td></td>
</tr>
<tr>
<td>High-throughput Sequencing, Proteins and Statistics</td>
<td>2016</td>
</tr>
<tr>
<td>Prof. T. Speed</td>
<td></td>
</tr>
<tr>
<td>Prof. E. Wit</td>
<td></td>
</tr>
<tr>
<td>Prof. G. McLachlan</td>
<td></td>
</tr>
</tbody>
</table>
CONTENTS

1. Timetable 1

2. Posters 4

3. Special Invited Papers 5

4. Abstracts
   Session I 27
   Session II 45
   Session III 77
   Session IV 105
   Session V 119
   Posters 135

5. List of Participants 171

6. Author Index 175
1. TIMETABLE

Tuesday 5th July

9:00-10:00 Registration

10:00-10:30 COFFEE

SESSION I

10:30-10:45 Welcome
   Prof. Michael Arthur, Vice Chancellor

10:45-11:30 (1) Overview of the special invited articles;
(2) some selected topics of next generation statistics
   Kanti V. Mardia

11:30-12:15 Genome-wide protein structure prediction and structure-based function annotations
   Yang Zhang

12:15-12:40 Full Bayesian analysis of the generalized non-isotropic Procrustes problem with scaling
   Douglas L. Theobald* and Kanti V. Mardia

12:45-1:45 LUNCH

SESSION II

2:00-2:45 Exploiting protein family information to predict protein functions and functional networks
   Juan Ranea, Jonathan Lees, Ian Morilla, Jean-Karim Heriche, Zac Hussain, John Shawe-Taylor, Robert Rentzsch, David Lee and Christine Orengo*

2:45-3:10 A statistical view on the reference ratio method
   Kanti Mardia, Jes Frellsen, Mikael Borg, Jesper Ferkinghoff-Borg and Thomas Hamelryck*

3:10-3:20 Overview of posters
   John Kent

3:20-3:45 TEA

3:45-4:10 Metabolic network evolution and operation in unicellular eukaryotes
   David Westhead*, John Whitaker, Thomas Forth and Glenn McConkey

4:10-4:35 Structural bioinformatics: current status and future trends
   Swanand Gore*, Sameer Velankar and Gerard Kleywegt

4:35-5:00 Statistical alignment of multiple protein structures under a dynamics-based model of structural evolution
   Joseph Herman* and Jotun Hein

5:00 Drinks Reception

In multiple author papers, * indicates the presenter(s).
WEDNESDAY 6th JULY

SESSION III

9:15-9:30 David Kendall’s shape legacy
   Kanti V. Mardia

9:30-9:55 Riemannian centres of mass and a new look at the classic central limit theorem
   Wilfrid Kendall

9:55-10:20 Detecting the shape of a configuration from random projections
   Huiling Le

10:20-10:45 Speculations on the next statistical toolkit for complex organized systems
   Fred Bookstein

10:45-11:15 COFFEE

11:15-11:40 Colouring and breaking sticks: random distributions and heterogeneous clustering
   Peter Green

11:40-12:05 Bayesian molecular alignment using random fields
   Irina Czogiel, Ian Dryden* and Christopher Brignell

12:05-12:30 Manifold stability and the central limit theorem for mean shape
   Stephan Huckemann

12:45-1:45 LUNCH

2:00-2:15 Photographs

2:15-3:15 POSTER SESSION

3:15-3:45 TEA

SESSION IV

3:45-4:10 A fresh look at projective shape spaces
   John Kent

4:10-4:35 Pliable template: a null model for shape variation when the mean shape changes
   Christian Klingenberg

4:35-5:00 Shape analysis and skeletal form and function: a personal perspective on the geometric morphometric “revolution”
   Paul O’Higgins

7:00 Drinks reception

7:30 Conference Dinner

In multiple author papers, * indicates the presenter(s).
THURSDAY 7th JULY

SESSION V

9:30-9:55 Data integration in genomics: a new biostatistical challenge  
   Michael Schimek

9:55-9:20 Efficient, non-disruptive local moves for Monte Carlo sampling of proteins  
   Sandro Bottaro*, Wouter Boomsma, Kristoffer Johansson, Christian Andreetta, Thomas Hamelryck, and Jesper Ferkinghoff-Borg

10:20-10:45 Exploring the energy landscapes of protein folding simulations with Bayesian computation  
   Nikolas Burkoff* and David Wild

10:45-11:15 COFFEE

11:15-11:40 “Aggregate maps” for exploratory and confirmatory analysis of nuclear architecture  
   Richard Russell, David Weston*, David Stephens, Niall Adams and Paul Freemont

11:40-12:05 Estimating soil microbial species abundance through metagenomic sequencing - journey into the “unknown-ome”  
   Wally Gilks*, Elisa Loza, Michael Defoin-Platel, Ian Clark and Penny Hirsch

12:05-12:20 Forward-looking review of LASR 2011  
   Kanti V. Mardia

12:30-1:30 LUNCH

In multiple author papers, * indicates the presenter(s).
2. POSTERS

An evolutionary algorithm for stochastic context-free grammar design, with applications to RNA secondary structure prediction
   James Anderson*, Joe Staines, Paula Tataru, Jotun Hein and Rune Lyngsø

Statistical spectroscopic methods in drug metabolism and metabolic profiling
   Toby Athersuch

Testing for equality of distribution for dihedral angles in amino acids
   Maha Bakoben* and Charles Taylor

Mixture models for spherical data with applications to protein bioinformatics
   Philippa Burdett*, Kanti Marida and Stuart Barber

Stochastic modelling in Immunology
   Mark Day

ABC for coronary heart disease policy modelling
   Nathan Green

Modelling threshold violations of air pollution concentrations using multiple change-points Poisson process
   J. Gyarmati-Szabó*, L.V. Bogachev and H. Chen

Insights into protein folding through logistic regression of contact maps
   Kerstin Hommola*, Walter Gilks and Kanti Mardia

Evaluation of tissues surrogacy using gene expressions
   Adetayo Kasim*, Ziv Shkedy, Dan Lin, Suzy Van Sanden, Jose Cortinãs Abrahantes, Hinrich Goehlmann, Luc Bijens, Dani Yekutieli, Jeroen Aerssens, Michael Camilleri and Willem Talloen

Multi-scale analysis of high performance liquid chromatography data
   Jennifer Klapper* and Stuart Barber

Bayesian inference for joint modelling of longitudinally continuous, binary and ordinal events
   Qiuju Li*, Jianxin Pan and John Belcher

Motor neuron disease and motor unit number estimation: the role of the observed data likelihood in Bayesian inference and computation
   A. Pettitt*, C. Drovandi, R. Henderson and P. McCombe

In multiple author papers, * indicates the presenter(s).
3. Special Invited Papers
Evolution of Statistical Principles

David Cox

Nuffield College, Oxford

Professor Mardia and his colleagues are surely to be congratulated on the series of influential conferences that they have organized in Leeds, as well as on their own distinctive contributions. Professor Mardia asked me to contribute a few remarks on the current position of statistical research and the implications for the future. I do so with some nervousness!

My sense is that over recent years the general appreciation of the relevance of statistical thinking has broadened, among scientists in particular and among the general public to a lesser extent. Of course this is not quite the same as appreciation of statisticians, but I think even this may have improved too!

There are, perhaps, two broad challenges to our subject. One is to recognize the great diversity of contexts in which statistical ideas arise, the need for a variety of approaches and indeed to appreciate that no two applications are quite the same. The other is, as far as possible, to see some unity of ideas and to limit the extent to which the subject fragments into separate, even occasionally warring, components.

Like so many areas of work, statistics is very heavily influenced by current technology, in our case by computing, both as it affects data capture as well as data analysis. I recall with some displeasure hours spent reading paper traces to convert the signal level into numerical form as a basis for time series analysis, which itself was a pretty painful process. Nowadays computational time is sometimes still a limit on data analysis; think for example of the ABC algorithm needed for fitting some complex models in genetics and systems biology and perhaps other areas too.

But on the whole computational considerations have drastically transformed the tactics of much if not all of data analysis. In the old days, before embarking on any other than the very simplest forms of analysis one thought and thought again and then possibly computed; remember that inverting an 8 by 8 matrix by hand would take at least a person-day. Now, in many fields at least, it is normal to compute first and think later. This has obviously many advantages. It would be good if one of the few advantages of the older approach, the ability to see the pathways between the data and the final answer and the ability to detect quickly absurd answers, could be protected.

A key question, however, concerns whether the basic principles of applied statistics have been radically changed, rather than been subject to gradual evolutionary processes. In so far as the principles concern question formulation, study design, measurement issues, and the various phases of analysis, leading on to interpretation, I tend to think the broad principles to be fairly stable, although answers to such a wide-ranging question are no doubt strongly age-dependent!

References


A Two-Hundred-and-Fifty-Year Argument

Bradley Efron
Stanford University

Abstract

This is a brief commentary on the theme of the 30th Leeds Annual Statistical Research Workshop, “Next Generation Statistics in Biosciences.” It raises the question of whether objective Bayes results enjoy the favorable properties of Bayesian analysis based on genuine prior distributions. This is pertinent to the Workshop theme since current activity in biomedical statistics has a strong objective Bayesian strain. Two methodologies, empirical Bayes and the bootstrap, are mentioned as connecting links between the frequentist and Bayesian worlds.

I am grateful to Professor Mardia for the chance to add a few brief comments on “Next Generation Statistics,” the pregnant theme of the 30th Anniversary LASR Workshop. Let me begin facing in the wrong direction, back toward the beginning of statistical inference.

The year 2013 will mark the 250th anniversary of Bayes rule. The rule has been influential over the entire period, and controversial over most of it. Its popularity describes a rollercoaster graph: lofty in the time of Laplace, dipping low after Venn’s 1866 Logic of Chance, back up by 1900, and then down again under Fisher and Neyman’s massive influence. Most of the Twentieth Century was down time, especially in applications, but we now see a healthy revival of interest that is certain to influence the Next Generation.

Almost nobody questions Bayes rule in the presence of a genuine prior distribution, or for personal decision-making in the Savage–DeFinetti subjective prior framework. Controversy begins when the rule is applied to scientific problems in which prior information, often absent or untrustworthy, is replaced by “uninformative” priors of the type advocated by Jeffreys (1961). The healthy revival just mentioned is moving very much along Jeffreys’ line. Markov chain Monte Carlo, MCMC, the key technology for computing posterior distributions in complicated circumstances, feasts on conventional uninformative priors.

Here is a simple (but true) story illustrating my point of concern. A physicist friend of mine found out, via sonogram, that she was going to have twin boys, and asked me what was the probability that they would be identical rather than fraternal. Her doctor had told her that only one-third of twin pairs are identical. Starting from this prior distribution, and the fact that a same-sex sonogram is twice as likely for identicals (since they are always same-sex whereas it’s 50-50 for fraternals), Bayes rule gives the answer

\[ \Pr\{\text{Identical}|\text{Same Sex}\} = \frac{1}{2}. \]  

A key assumption here, and the one that makes calculation (1) convincing, is that the doctor’s one-third/two-thirds prior distribution was based on some huge registry of previous twin births. Suppose, though, he had been speaking from limited personal experience, say just three previous twin deliveries, one identical and two fraternal. A conventional Jeffreys analysis might assign \( p \), the population proportion of identical twins, a beta(1,2) distribution, with density

\[ g(p) = 2 \cdot (1 - p) \quad \text{for } 0 \leq p \leq 1, \]  

\[ ^1 \text{Max H. Stein Professor of Statistics and Biostatistics (Health Research and Policy)} \]
obtained by beginning with an improper \( \text{beta}(0,0) \) “Haldane” hyperprior, updated to (2) on the basis of the doctor’s experiences. (Jeffreys actually preferred starting from a \( \text{beta}(\frac{1}{2}, \frac{1}{2}) \) hyperprior.) An entertaining application of Bayes rule starting from (2), which I only got wrong twice, yields \( \Pr\{\text{Identical|Same Sex}\} = \frac{1}{2} \), the same as before.\(^2\)

My point of concern is whether this “1/2” has the same logical force as that in (1). The \( \text{beta}(1,2) \) prior density (2) seems reasonable enough, but in fact there is nothing medical or twin-related about form (2). It yields other interesting conclusions — for instance, that the posterior probability of \( p \) exceeding 0.5 is 0.313 — which to a non-Jeffreysonian might seem suspiciously precise.

There is more at stake here than this little example suggests. Convenience priors of all types, “invariant,” “uninformative,” “reference,” “conjugate,” “objective,” are featured in our journals, most often in complicated data analyses where it is difficult to trace their effect on conclusions. Next Generation Statistics, the subject here, is certain to include a substantial Bayesian component of the objective sort. In order to avoid another dip of the Bayesian rollercoaster we need a deeper understanding of the operational consequences of objective Bayesianism.

Besides its convenience and satisfying output (as seen in the twins example), Bayesian methodology offers an array of tempting theoretical advantages. A two-sample study, for instance, might include an over-ample supply of possible covariates, which have to be winnowed down before making the final comparison. Frequentist methods must take account of the complicated winnowing-down process, while the Bayesian’s prior distribution is presumed to incorporate all such decisions. Similarly, interim looks at the data in a clinical trial require frequentist adjustment of the significance level, while the Bayesian need only consider the final comparison. “Selection bias,” the data-based choice of interesting-looking cases, is no problem at all in the Bayesian world.

All of these properties follow logically and mathematically from Bayes theorem. The concern here is whether they are scientifically valid when a prior distribution, like (2), is not fully anchored in past experience. It seems important to me that objective Bayesian methods, which are certain to play an important Next Generation role, enjoy at least some frequentist support. (Bayarri and Berger (2010) pursue this point much more thoroughly.)

What I am hoping for from the Next Generation are ideas and methods that connect the two worlds, Bayesian and frequentist. Empirical Bayes is an attracting connecting technology, as discussed at length in Efron (2010). Hierarchical Bayes implementations of EB usually begin with uninformative priors at the top level, again raising the question of how far we can trust their formal Bayesian properties. (In this vein, Efron (2009) investigates freedom from selection bias in empirical Bayes estimation models.)

The bootstrap, usually presented in stark frequentist terms, also has its Bayesian connections. Figure 1 concerns an artificial but not completely unrealistic example. Independent normal responses from an unknown regression model \( f(x) \) have been observed,

\[
y_i \sim \mathcal{N}(f(x_i), 1) \quad \text{for } i = 1, 2, \ldots, 101
\]

with the goal of inferring

\[
\theta = f(1),
\]

the regression value at the far right end of the \( x \) scale.

---

\(^2\)The physicist’s twins turned out to be fraternal, not identical. If the doctor were updating his prior from (2), it would now become \( \text{beta}(1,3) \), leading to a lower estimate than 1/2 if he were asked the same question about his next twins delivery.
A sequence of polynomial models

\[ f^{(J)}(x) = \sum_{j=0}^{J} \beta_j x_i^j \quad J = 0, 1, 2, \ldots, 8 \]  

were fit by ordinary least squares, and compared in terms of the \( C_p \) (or AIC) criterion

\[ C_p^{(J)} = \sum_{i=1}^{101} \left( y - \hat{f}^{(J)}(x_i) \right)^2 + 2 \cdot (J + 1). \]  

\( J = 5 \) gave a clear minimum, yielding the 5th-degree polynomial \( \hat{f}(x) \) shown as the solid curve in 1, and the estimate

\[ \hat{\theta} = \hat{f}(1) = 2.25. \]  

The assessment of confidence intervals after model selection is an exemplary Next Generation problem. One approach is via a parametric bootstrap. \( B = 10,000 \) (many more than necessary) bootstrap replicates \( \theta^* \) were generated by each time taking

\[ y_i \sim \mathcal{N}\left( \hat{f}(x_i), 1 \right) \quad \text{for } i = 1, 2, \ldots, 101, \]  

finding the minimum \( C_p \) polynomial \( \hat{f}^*(x) \) based on the \( y^* \)'s, and setting \( \hat{\theta}^* = \hat{f}^*(1) \). The vertical dashed line at the right shows the 90% central percentile interval \( \theta \in [1.13, 3.32] \), with endpoints at the fifth and ninety-fifth percentiles of the 10,000 \( \theta^* \)'s.

Newton and Raftery, in an under-appreciated 1994 paper, show how bootstrap replications, nonparametric in their case, can be reweighted to give Bayes posterior distributions (using a form of importance sampling). The solid vertical line is the central Bayes 90% credible interval \( \theta \in [1.10, 3.30] \), based on a prior distribution following Smith and Spiegelhalter (1980)'s prescription.
The reweighting values, which are mild and easy to compute in this case, give one a nice feeling for the relation between frequentist and Bayesian inferences. I expect, or at least hope, for an upsurge of progress along the Bayes/frequentist boundary in the near future.

References


Next generation statistics in the sciences

Michael Levitt

Department of Structural Biology, Stanford University

I studied physics as an undergraduate and that seemed to prepare me well for a career in computational biology. Slowly I have come to realize that there was a gaping hole in my basic learning: an almost complete lack of training in advanced statistics. It also became clear that overall statistical knowledge is appallingly neglected in general education yet it is the one tool essential to estimate the risks associated with the choices we need to make throughout life.

Thanks to a working knowledge of college mathematics and a passion for computer programming, I have seen myself become increasingly statistical. In all our work, we now think of controls and the need to test alternative hypotheses. Bootstrapping and resampling have become standard tools of my trade, which are likely to grow in importance with the ever increasing abundance of computational resources.

Statistical problems have a way of being relevant to all fields. For me this is seen in the problem of ranking that arises in assessing different computed conformations of a folded protein or RNA molecule to find which is most likely to be real. Ranking is much more widespread particularly in the US where there are lists of the best doctors, the best universities, the best hospitals, etc. Ranking is particularly interesting in that it seems like it should be easy and objective. Alas, when there is more than a single criterion (say wealth and intelligence), it is impossible to have a ranking that is not subjective as one needs an arbitrary scale factor that converts between the two measures. Nevertheless, such ranking is done all the time.

Ranking is also a problem in the academic world, especially in the UK and Europe where groups are assessed on their past achievement often using numerical measures. In the US, the problem is circumvented by manual ranking of all NIH grant proposals but this comes at a massive cost to the research community where an average scientist spends at least as much of their time reviewing grants as they do writing them. It seems that the community needs to help develop proper evaluation tools before these are foisted upon us.

An overriding problem with statistics is that many different models seem plausible yet the results one gets do depend on the model. A model cannot be judged on its outcome but needs to be judged on some other beneficial criterion such as robustness. Any development of new methods for ranking and research evaluation are going to need to be transparent and easily understood to achieve wide-spread acceptance.

References


Thirty years of structural biology and bioinformatics

Simon E.V. Phillips
Research Complex at Harwell, Rutherford Appleton Laboratory
Harwell Oxford, Didcot, Oxon

Biological organisms are the manifestation of the bewildering complexity of their underlying molecular processes. While biological molecules must obey the same laws of physics and chemistry as simpler molecules, this complexity means that our ultimate goal of understanding how organisms and cells work in molecular detail calls for a prodigious volume of experimental data, coupled with its analysis by advanced statistics, mathematics and computation. In fact, a feature of biological macromolecules is their information content, manifested not only in DNA sequence, but also in the three-dimensional structure of proteins, the major workhorses of the cell, and folded RNA molecules. Biology is therefore a fertile area for the application of mathematics and statistics.

Pioneering protein crystallographers in the second half of the twentieth century began the creation of a database of protein structures, mainly by X-ray crystallography with some more recent contributions from NMR spectroscopy, borrowing methods from physics and chemistry, with rapid developments in computation and computer graphics. In 1981, the year of the first LASR meeting, the Protein Data Bank (PDB) contained only 86 structures, but this has now risen to over 72,386 today (13 June 2011), driven by revolutions in molecular biology and genetic engineering, as well as advanced technologies such as synchrotron radiation sources. Over a similar period, the use of purely computational methods to study the dynamics of biological molecules has grown from the seeds planted by a few visionaries, who felt it must be possible to simulate their behaviour given the rules of physics and chemistry, to the powerful field bioinformatics is today. In addition to the intellectual beauty of this understanding, these developments are beginning to reap practical rewards in the design of new drugs and therapies, as well as provide novel engineered reagents for industrial processes. Recent examples are the drug treatments for HIV and pandemic influenza.

LASR meetings have tracked the progress of the field, and also second explosion in biological computation triggered by high-throughput methods in experimental biology to decipher whole genome sequences and chip array methods to follow the behaviour of gene expression and metabolism in cells. The next thirty years promise to be even more exciting, with biologists, chemists, physicists and mathematicians working even more closely to seek a complete understanding of how a cell works. Will we have a complete model of a cell, and the ability to predict its behaviour in molecular detail in time for the sixtieth anniversary of LASR? Perhaps, but what is certain is that we will continue to be astounded at the power of linking across traditional disciplines to understand biology, and the spin-offs in medicine and therapeutic intervention will be spectacular.
Controls

Terry P. Speed

Bioinformatics Division, Walter & Eliza Hall Institute of Medical Research

Controls are important in experiments for biologists. They use many forms of controls, going well beyond the classical “treatment vs control” notion used by statisticians. Statisticians might benefit from learning more about biological controls, and thinking up ways to use data on controls in their analyses.

Terry: I just want to say one word to you. Just one word.
Statistician: Yes, sir.
Terry: Are you listening?
Statistician: Yes, I am.
Terry: Controls.

I don’t know when I first started to become aware of the importance and variety of controls in biological experimentation, but I’m guessing it was almost 25 years ago. If true, then I am a really slow learner, as I shall explain. But I have always believed in the dictum Better late than never, so I’ll suppress my shame and come clean. I do so in the hope that some of my readers might avoid my simple mistakes, and restrict themselves to making interesting mistakes of their own. The main purpose of this note is to suggest that we statisticians should start to appreciate the depth and breadth of the notion of control, and in doing so, broaden our view of experimental design, and try to use more control data in our analyses.

Since my earliest days I have been a biological groupie, always wanting to visit my biological collaborators in their labs, look at their equipment, and generally stick my nose further in than they and many of my statistical colleagues thought was normal. Around 1987 I took this tendency to a new level, and started sitting in on the regular lab meetings of my biological collaborators, sometimes two or three such meetings each week. Let me say right now that I have never progressed beyond this stage. While I have often wished I could conjure up the results of an experiment that I could conceive, that is, behave like a biologist and generate my own biological data, I have always suppressed this urge. My reason for doing so is simple: I like to collaborate with first-rate scientists wherever possible, or perhaps second or third-rate ones, but certainly not tenth or twentieth-rate ones, which is what I would be if I started playing the biologist. At first the lab meetings were mind-numbingly boring, although at times I got some of the pleasure an anthropologist might get trying to understand another culture by observing the behaviour of some of its members. But I wasn’t an anthropologist, I was — well, I didn’t say it quite like this, but I thought — I’m a statistician. I’m here to help you. I didn’t know the meaning of many of the words people used. I had little to no understanding of the assays, reagents, or experimental protocols being used on a daily basis, much less the interpretation of the results of experiments. All of these were matters that occupied a considerable amount of time at lab meetings, for example, discussion of why the amount of Mg or the salt concentration was very important in some particular assay. I started with a little abstract understanding of the new science emerging as a result of what is now termed the recombinant DNA revolution, and of the polymerase chain reaction (PCR) revolution which came slightly later, and I took it from there. For example, I quickly realised that gel electrophoresis was central, so I tried to get on top of that technique.

My general strategy was to try and learn a couple of new things each week. Remember
that back in those days, there was no Google or Wikipedia. I was very reluctant to ask simple
questions of the senior people who permitted me to sit in on their meetings. To learn new
things, I used books, papers, or (my preferred option) graduate students. Just in case it is not
obvious, let me tell you that there is an enormous gap between the knowledge encoded in books
with titles such as Genetics, or Molecular Biology of the Gene (or the Cell), and the day-to-day
activities of a molecular genetics lab. Even reading Molecular Cloning: A Laboratory Manual,
which I dived into at a later date, barely sufficed. Filling that gap was my main task, all the
while hoping some data would appear on which I would be consulted. That was pretty rare, as
I learned.

As mentioned, I was unfamiliar with the standard assays being used, and in particular, I
was quite inexpert with what was a common practice then, and still is, but to a lesser extent:
the interpretation of black marks on images of agarose or polyacrylamide gels. Some of these
marks looked like clean bands, as they were supposed to (see any text book explanation of
Southern blots), but most of them looked like the ink smudges that were such a prominent
feature of my penmanship as a child, or Rorschach tests, which were also familiar to me from
my childhood. But I persevered, and along the way I learned about the critical role of controls
in biological experimentation. Mistake number one was not getting this right at the outset, that
is, not realizing the importance and the variety of controls. I should have asked.

I’ll skip ahead a little now, to my short career in DNA forensics, and give my first definition
(for statisticians) of one kind of biological control. A solid or slab gel is a thin rectangle of
porous material a little like a slice of very fine-grained Swiss cheese, through the holes of
which weakly charged macromolecules are encouraged to travel, with a little help from a voltage
difference at the ends. (Note: the holes were far too small to be seen.) What counted was where
the molecules ended up when the run was declared over, as this is related to their size. Smaller
molecules moved faster than larger ones, and most gels were run with size standards called
ladders. The molecules are generally invisible, and so their final resting place needed to be
visualized with the help of radioactive or fluorescent labels. Radiolabeled gels were developed,
in much the same way as photos were back then. The result would be an image (“autorad”)
showing the collection of bands, smudges and smears to which I have already alluded. Rough
definition number one: a positive control (in this context) is a reagent that leads to a band that
is clearly present on the gel, right where it should be. And rough definition number two is the
converse: a negative control is a reagent that leaves no bands where there should be no bands.
Keep this in mind, and note that while this use of the term control seems rather different from the
usual one (cf Treatment vs Control), there is a connection. You can think of a positive control
as a Treatment whose response is known, and different from that resulting from the Control,
which should show nothing, but more generally is a treatment whose response is known, and in
some sense is standard or baseline.

Let me now mention a moment where the significance of controls became very clear to me.
It was in connection with a certain murder case, where there was blood, a police lab, a scientist
in that lab, and forensic DNA identification based on the blood. One question was whether
there had been cross-contamination of DNA due to poor sample handling procedures. In the
discussion of this question, the behaviour of substrate controls (see any CSI, or Wikipedia)
played a key role, and not only the controls in the case itself, but also those from previous
work, as records were available of the scientist’s past performance in both cases and tests. I can
vividly recall how serious doubts were cast on the DNA identifications the scientist made in the
murder case under discussion, because of issues concerning his substrate controls in that case,
and in many prior instances. The inescapable conclusion was that if a lab scientist wanted to be
taken as reliable, then all his controls should behave properly all the time. On many occasions
since then, I have seen judgements made about scientists, both on the basis of the way they use
controls (how many, and where), and on the behaviour (performance) of their controls.

As statisticians, we tend to think that we know about experimental design. \textit{Randomization, Replication and local control} (now called Blocking) is the mantra introduced by R. A. Fisher in the context of agricultural science, and many of us (myself included) have repeated it many times, perhaps giving the impression that we feel little more is needed to carry out a proper experiment. To be fair, this is often in contexts where little or no attention is being paid to design issues, and so any attention to design is better than none. But I still think we statisticians often behave as though variations on and modest elaborations of these three themes (RRB) suffice for most experimental science. Within our statistical context, control usually means the untreated or the standard treatment, though we are all aware of cases where some particular “treatment” is designated the control.

I found myself making use of a biological notion of control as soon as my first microarray dataset arrived. In the context of a comparative microarray experiment, a (better: one notion of, see below) positive control is a gene which is known (or expected), \textit{a priori}, to be differentially expressed. Conversely, a negative control is a gene known (or expected) not to be differentially expressed. The first microarray analysis in which I was involved concerned liver cells from mice in which a particular gene was knocked out, KO mice, and liver cells from mice which did not have this gene knocked out, so-called wild-type (WT) mice of the same inbred strain. Naturally we hoped and expected that the average expression level of this gene from 8 KO mice would be much less that the average from 8 WT mice, and indeed it was. In fact, the most differentially expressed gene in the experiment was the one that had been knocked out: it was at its normal level in the WT mice, and at a very low level, probably zero apart from some technical variability, in the KO mice. This item of confirmation — that our positive control behaved as a good positive control should — was most gratifying, as it gave us some confidence in our other results.

At this point I should add that a well-conducted microarray experiment typically has many other controls. The one I have just defined might be called an \textit{experiment-wise} control, or, better, a control for our \textit{statistical method}. The idea behind all controls is the allowance for only one variable or condition to change in each comparison, so that we know with greater confidence that any difference we observe is due to the variable or condition of interest. So if we have different variables or conditions, multiple controls are needed, each corresponding to one of the variables or conditions being considered, see \url{http://en.wikipedia.org/wiki/Scientific_control}.

Rather soon after my first microarray analysis I found myself getting asked by others to help them design their microarray experiments. It didn’t take long for me to see that traditional statistical design principles were inadequate for microarray experiments, in which many thousands of measurements are taken on each experimental unit. Not only do traditional concepts of type I error need to be revised, but the notion of power becomes a much more elusive concept. I am sure you are all familiar with the explosion of research on multiple testing that has taken place in the last decade, much stimulated by microarray experiments. Many variations on the notion of false discovery rate have been developed. There is some but much less work on power in this context, as the problem seems harder, and there is less agreement on how to modify the traditional notion. But even more fundamental is the fact that it is often quite hard to tell whether a microarray experiment was a success or a failure. Positive controls as described above can play a valuable role, as my first experience showed. We could easily imagine designing an experiment aimed at “discovering” a positive control, a much more manageable task. Very early on, when I found myself talking with biologists about the design of a microarray study they were planning, the first thing I would ask them is: do you have a good list of positive control genes? Not infrequently I received the answer “No, that’s why we are doing the experiment!” to which I would reply “Then how will we tell whether it was a success?” and their reply would usually
be “Oh, we can check some putative differentially expressed genes afterwards, to see whether the experiment was a success.” I won’t continue describing this discussion, but it can and did go on for quite a while, and the general issues are admirably covered in the book Experimental Design for Biologists by David J Glass, for which a short synopsis — Chapter 19 of the book — is given in the Appendix. This book isn’t specifically aimed at microarray, or more generally, omics researchers, but it contains much that should interest them and us.

Now let me turn briefly to negative controls. It embarrasses me to admit that it was only very recently that I realized the enormous potential value in using negative control genes in the analysis of microarray data. The primary use — already adopted informally by several groups — is for normalization, and for removing unwanted variation such as batch effects. For some details, see http://www.stat.berkeley.edu/tech-reports/800.pdf. I really regret not having realized this over a decade ago. At a time when I was irritating my collaborators with talk about positive controls, I failed to think hard about negative controls. How could I miss something like this? Don’t make my mistake.

There is much more to be said on the use of positive and negative control genes in the analysis not only of microarray data, but of proteomic, metabolomic, fMRI, and a wide range of other kinds of data. I plan to keep working on these issues for a while longer, so watch this space.

My general feeling now is that we statisticians should pay much more attention to the way biologists work, and attempt to incorporate their insights into our thinking, and, wherever possible, our analyses. Not only do biologists have many valuable things to say about controls — Glass discussed no fewer than seven types, each with several different variants — they also attempt to deal with models and causality, two rather contentious topics within our community. They do so in much more down to earth ways than we typically do, but always in ways that impact on their collection of data, which should be of interest to us. I feel certain that I am not the only statistician who has much to learn from them, and I encourage you to begin doing so, perhaps starting by reading Glass’ excellent book, or Wikipedia, or sitting in on a biology lab meeting nearby.

Acknowledgements: I’d like to thank Johann Gagnon-Bartsch, Laurent Jacob and Hui Shen for their invaluable comments on earlier drafts of these remarks.

Appendix

1. Framework. What is the question you want to answer?

2. Inductive space. Decide on what aspects of prior knowledge relate to your question. Read the literature.

3. System. What tools will you use to answer your question?

4. System controls. How do you know your system works? How do you know your system can provide the type of data you require? Is the chosen system well matched to your question, or might a different system be better?

5. Experiment. What are you going to do to answer the question? Be sure that measurements are taken multiple times and that you are measuring the effect in a representative fashion. Study the effect over time and over a range of experimental conditions. Do a
dose-response with any experimental agent. Try to determine the “representative case” for the subject if that is relevant to the question. Consult a statistician and discuss the mode of analysis for your data and how many data points are required.

6. **Establish criteria** for the effect in advance of the experiment.

7. **Negative controls.** What negative controls are required? Is there an “all but X” control available?

8. **System-positive controls.** How do you know that the system is still operational? What positive controls are necessary to prove that the thing you want to measure was actually measurable within the context of the experiment?

9. **Effect-positive controls.** How do you know the effect you want to measure can be produced in your system? What positive controls are necessary to produce the effect?

10. **Assumption controls.** If X is being measured, can something else be measured that has been shown to occur when X occurs? If you think Y has happened as a result of X, is there something else you can measure that has been shown to also happen when Y happens?

11. **Do the experiment.**

12. **Experimentalist controls.** Analyse the data in a blinded fashion.

13. **Repetition.** Repeat the experiment using the same criteria and methodology.

14. **Model building.** What is the answer to your question?

15. **Model check.** Is the answer responsive to the question?

16. **Prediction.** Does the model predict what will happen again? Repeat the experiment.

17. **Extension.** Does the model hold in different circumstances?

18. **Change the system.** Approach the question in another way.

19. **Change the scientist.** See if others can reproduce the effect.

20. **Present the data.** What do others think of the result? What do others think of the interpretation of the result?

21. **Predict the future.** See if the model continues to represent what will happen under various circumstances.

22. **Amend the model** as instances are found in which it is not predictive. Limit your claims as you find limits to your claims.

23. **If the system is reductionist** in nature, realize that and apply the model in a non-reductionist or less-reductionist setting.

24. **A model that is limited but verifiable is superior** to a model that is comprehensive but not predictive.

25. **If an idea cannot be subjected** to experimentation and verification, it is not relevant to your problem.
The future of biology; a very convenient place for computational scientists and statisticians

Anna Tramontano

Department of Physics, Sapienza University of Rome

1 Introduction

In the Citadelle, Antoine de Saint-Exupery wrote: “As for the future, your task is not to foresee it, but to enable it”. Nevertheless, in bioinformatics and computational biology, the speed at which new technologies appear makes it necessary, in order to enable the future, to try and foresee it and to be prepared to face the new data challenges in a timely fashion.

This is not an easy task, and it is almost impossible to predict what waits for us around the corner. However, there is no doubt that a combination of different data and expertise will be instrumental and effective for inferring interesting and useful properties of biological systems, both at the general level and specifically directed towards important biomedical problems. The availability of more and more diverse biological datasets and the increased attention of the biological and medical communities to computational analysis tools have led to an explosion in methods and databases in the last decade or so. This “pilot” worldwide phase of computational biology, which I believe is coming to an end, has been based so far on the availability of, essentially, one single genome for each of a limited number of species and on not completely adequate methodologies for integrating data, essentially due to the intrinsic difficulty of obtaining large scale reliable datasets for expression levels and molecular interactions and to the limited reproducibility of experimental results obtained on different systems with different techniques. But this is rapidly changing and it is important to be prepared for the coming phase of the post-genomic era. In the very near future data on human and pathogen variations will accumulate very rapidly and experimental techniques will improve in terms of reliability and coverage. We have to be ready to face the challenges taking advantage of all the data as soon as they are available, because only by extracting the information in real time we can drive the next generation of experiments and technical developments.

It is essential to keep in mind that in computational biology the availability of more data does not only affect quantity, but also quality. Any new information can be linked and correlated to all the others, adding value and biological insight to it. Similarly, the access to more computational resources does not simply imply that the same things can be done faster, but that intractable problems can suddenly come within our reach. The dimension and scope of the post-genomic data are such that it is becoming more and more essential to join efforts at the worldwide level and to use multidisciplinary approaches. This is also apparent by the continuous emergence of consortia for handling some of the formidable challenges that lay ahead of us (Kidd, et al. (2008); McKusick (2007)).

These worldwide efforts aimed at understanding a biological organism can be mapped to an engineering project where we need to have the parts list of the system, understand their tolerance thresholds, find the assembly instruction and, finally, simulate its behaviour. Within the same analogy, I would like to stress that it is of paramount importance to have sufficient resources to scale up the project from a “prototype” to a “production” phase as early as possible in the process because this can reveal unexpected difficulties and/or opportunities that need to be recognized and addressed.
2 The parts list

After ten years of efforts, the number of genes in the human genome has not yet been established (Eckhardt, et al. (2004)). In the mean time, the picture has become more and more complex as we proceed. The role of epigenetics and of short and long non-coding RNAs is revealing to be more and more important, and one expects more new regulatory mechanisms to be discovered. Determining the identity and function of all of the sequence elements in human DNA is a daunting challenge; even in relatively simple organisms, using the available data to reconstruct the network of logical and physical interactions among their elements is far from being a solved problem.

Because of historical reasons, genomics and proteomics research have been rather separate and the results of the first have mainly been regarded as useful ingredients to pursue the second. It is time to challenge this view. The availability of more mature methods for protein structure analysis and prediction and the improved speed of the computational resources make it feasible now to analyse at the structural level not only the clearly identified molecules identified in a genome, but also all the putative ones and use the results to assess the likelihood that they code for a phenotype. So far gene finding methods have relied on sequence-based methods, but the time is ripe and the computational resources sufficiently fast to allow the structural aspects of the putative products to be included in gene identification methods.

3 The tolerance threshold: Evaluating the effect of evolution and mutations on protein structure and function

Many studies aimed at correlating variations in the genes and the insurgences of diseases are ongoing (Kidd, et al. (2008); McKusick (2007)) and the human variation data are accumulating at an unprecedented speed (it is expected that a human genome will be sequenced in a matter of hours in the very near future). In both cases, one key problem is to analyse the observed variations/mutations in order to distinguish between neutral and pathological ones. A mutation can be causative of pathology for a number of reasons, some are trivial to detect, for example when the mutation is located in an essential part of the products known to be responsible for its function. Most of these mutations are too deleterious for permitting survival and therefore are not observed and it is likely that the majority of those not falling in this latter category have been already discovered. Other, more difficult to detect, effects can be due to destabilization of the protein, to the formation or disruption of an intermolecular complex, or to an impaired ability of the product to reach its final active three-dimensional conformation for thermodynamics or kinetics reasons. It is also the case that a mutation in one gene is deleterious only when combined with mutations in other genes even if there is not direct interaction between them.

At present, there is no method able to reliably distinguish between neutral and deleterious mutations with high accuracy at the molecular level. As data on single genomes of patients and healthy individuals accumulate, more powerful statistical methods are needed to use the information effectively.

4 The assembly instructions: Exploiting the analysis for validating interaction and expression data

Protein-protein and protein-nucleic acids interactions are at the basis of most cellular processes and crucial for many bio-technological applications. During the last few years the development of high-throughput technologies has produced several large-scale protein-interaction data sets for various organisms (Stelzl and Wanker (2006)), while Next Generation Sequencing Techniques such as ChipSeq experiments (Park (2009)) are leading to very long lists of putative protein-nucleic acids interactions. It is important to develop tools for dissecting their content
and analyse the information they encode using data-integration and computational methods.

Interactions can be mediated by the presence of specific features, such as motifs, surface patches and domains. The co-occurrence of these features in proteins interacting with the same protein, for example, are indicative of mutually exclusive interactions and, therefore, can be used for inferring the involvement of the proteins in common biological processes. Similarly, the analysis of the nucleic acids sequences where a given macromolecule binds can allow the identification of patterns leading to the discovery of novel putative binding sites, provided the statistical significance of the matches can be accurately evaluated.

5 The simulation of the system

Systems biology, defined here as the simulation of the behaviour of biochemical pathways, cell compartments, whole cells or whole organs is still at its infancy, but it is important to pursue its goals, even with limited and incomplete data. Data completeness will probably never be achieved in biology, and it is essentially impossible to control all external factors that can influence the behaviour of biological systems because they are highly complex systems. Nevertheless, approximate approaches, where for example each molecule is treated as a non-structured entity and only its concentration and known interaction partners are used in the simulation, have already revealed their power in revealing interesting properties of the system under study.

So far most attempts have relied on very crude and approximate representations of the key players, such as proteins and small metabolites, partly because of the prohibitive computational time required for describing the system and the interactions at a more detailed atomic level, partly because structural information is not available for a sizeable number of the involved molecules.

The question is whether a hybrid approach, where only some of the molecules or parts of the system are modelled with medium-high accuracy, is feasible. The main problems with such an approach are the computational resources needed and the treatment of the “boundary conditions”. In other words, treating the interaction between molecules defined at the molecular level is very time consuming, and it is unclear how to handle the case where a molecule described at the atomic level interacts with an entity for which no information apart from its existence and maybe concentration is available.

6 Conclusions

In conclusion, the future of biology is a very convenient place to be for computational scientists and statisticians interested in contributing to the most challenging (and complex) project ever undertaken by science, that of understanding ourselves, our past, our future, our behaviour and our health.

This unprecedented challenge requires open minded, bold and creative actors who will not be afraid to enter into new areas and to collaborate with their fellow scientists, develop a common language and share their knowledge and intellectual resources.

References


**Further Reading**


4. ABSTRACTS

Session I
Some Selected Topics of Next Generation Statistics

Kanti V. Mardia

Department of Statistics, University of Leeds.

1 Introduction

For this paper we have selected three topics in statistics which might influence future scientific research in certain areas. The first selected topic is shape analysis. The second topic is False Discovery Rate (FDR) and its overall influence on large scale data analysis; the Empirical Bayesian approach takes this topic forward. The third topic is of data assimilation which has influenced climate change. We conclude with some futuristic remarks.

Importantly, there is a committee of the National Science Foundation, USA which is looking into The Mathematical Sciences in 2025. The idea is three fold: "This strategic examination will cover three aspects of the mathematical sciences enterprise: discovery, connections, and community. Here, "discover" refers to basic research at the frontiers of knowledge in mathematics and statistics. "Connections" refers to exploiting research opportunities at boundaries of the mathematical sciences to promote the progress of science, to enhance national security, and to strengthen economic competitiveness. "Community" refers to cultivating a community of researchers, students, and professionals of sufficient breadth, depth, and diversity to sustain the nation’s mathematical sciences enterprise in the 21st century." (AMSTATNEWS, May 2011).

1.1 Floods of Data

First, we draw to the attention of Statisticians the vast topic of Life Science, spanning so many areas (see Figure 1). This area of Science has also led to a flood of large data sets such as High Throughput, Genomic and Proteomic. These large data sets have led to massive data analytic problems. They raise many questions, and thousands of estimates or hypothesis tests; this is not what classical statistical masters had in mind (Efron, 2010). Indeed, new difficulties arise and can easily lead to flawed inferences.

1.2 A Wheel of Research

Following Kitano (2002), we indicate here how statistics dominates the research development in experimental sciences. A general strategy/roadmap which can be labelled a "Wheel of Research" is depicted in Figure 2. We quote Kitano (2002): "A cycle of research begins with the selection of contradictory issues of biological significance and the creation of a model representing the phenomenon. Models can be created either automatically or manually. The model represents a computable set of assumptions and hypotheses that need to be tested or supported experimentally."

Here "biological significance" can be replaced by any other science. Equally, the model could be abstract. "Computational "dry" experiments, such as simulation, on models reveal computational adequacy of the assumptions and hypotheses embedded in each model. Inadequate models would expose inconsistencies with established experimental facts, and thus need to be rejected or modified."

Here the emphasis is on hypothesis testing. "Models that pass this test become subjects of a thorough system analysis where a number of predictions may be made. A set of predictions that can distinguish a correct model among competing models is selected for wet experiments. Successful experiments are those that eliminate inadequate models."
The interaction between wet and dry experiments is highlighted, somewhat in contrast to Bioinformatics where databases play a major role. "Models that survive this cycle are deemed to be consistent with existing experimental evidence. While this is an idealized process of systems biology research, the hope is that advancement of research in computational science, analytical methods, technologies for measurements, and genomics will gradually transform biological research to fit this cycle for a more systematic and hypothesis-driven science."

Surprisingly, the word "Statistics" or "Statistical Science" is left out in this paper though there is uncertainty at each level in this "Hypothesis driven research in Biosciences". The topics “Experimental design, Experiment data analysis, Data and hypothesis driven models, predictions, simulations” all convey a need for statistical tools. A computer science approach could give some answers (but in general could rely on heuristics).

2 The Protein Folding Problem and Shape

If the last century belongs to Physics in Science then this century must belong to Life Sciences with many break-throughs following the DNA and protein work of the 1950’s! If DNA is like a recording system then protein is a playback system. Indeed, proteins are the workhorses of all living systems; they are for example responsible for digesting food and protection against infections. In medicine, proteins are the target of most therapeutic drugs. In biotechnology, proteins catalyze reactions that are very difficult to perform by chemical means. A protein is a complex object in the sense it takes a three–dimensional shape from the one–dimensional sequence of amino acids. One of the central problems of Protein Science is how this folding (a shape in 3–D) takes place, and this property has relevance to protein functionality, drug discovery and evolutionary biology.
Figure 2: Hypothesis driven research in Biosciences (adapted from Kitano, Science 2002).

Figure 3: The folding problem of proteins; figure drawn by Thomas Hamelryck.
2.1 Shape Analysis

Statistical shape analysis is concerned with extracting the shape information contained in a random sample of physical objects, where shape is defined to be all the geometrical information about an object that is invariant under a particular transformation of interest. (A closely related area in certain aspects is Functional Data Analysis; see Ramsay and Silverman, 2005.) Therefore, an important stage in the comparison of the shapes of two or more objects is to align them in some optimal sense, under some geometric transformation, so that the information which remains is the shape information of interest (see, for example, Dryden and Mardia, 1998). Here the objects are molecules, where the bond lengths between atoms should be preserved, so only the rigid body transformations are meaningful. Further, the basic shape here can be summarized by a set of dihedral angles rather than the standard representation by a set of landmarks.

One useful way to understand proteins is through the study of the local structure of protein called protein structure prediction. Given a sequence (fragment), say, of 9 amino acids, how can the local structure (3-D atomic coordinates) be found? The I-sites library contains protein fragments that have a strong sequence signature, and Rosetta gives structure prediction. A statistical solution using generalized Hidden Markov Models with a torus distribution is given by Boomsma et al (2008). The field is still evolving; for example, how can artificial proteins be built (protein design)? How does it all fit in at the cellular level with various interactions? How does it all help in functional modelling and drug discovery? Where are the clues for evolution? See, in particular, the review articles by Congreve et al. (2005) and Blundell et al. (2006) that highlight how vital structural biology and bioinformatics are in drug discovery. A statistical approach has been initiated in Mardia et al (2011). However, in all this forward-looking work, the new statistical methods have to be fast as one is dealing with very large data sets in real applications.

3 False Discovery Rates

We pointed out in Section 1 that various new questions arise from large scale data such as when a query (new) protein is compared with a large data base of proteins to find a similar protein or a set of similar proteins. The search is based on the principle that similar proteins have the same common functionality. Thus we end with the multiple hypotheses testing scenario.

Formally, the problem is to test multiple null hypotheses $H_{01}, H_{02}, ..., H_{0m}$ on the basis of large data sets $X$. One procedure is to use False Discovery Rates (FDR) pioneered by Benjamini and Hochberg (1995). (A more common procedure in Bioinformatics is the use of a so-called e-value which is tailored for a given data base.)

Let $P_1, P_2, ..., P_m$ be the corresponding p-values. For a single null case

$$FDR = \frac{\text{Type 1 error}}{\text{Type 1 error} + \text{Type 2 error}}.$$  

For a given level of significance $\alpha$ and large $m$, the expected FDR $= \alpha/2$ is a less conservative procedure. Efron (2010) has noted some flaws in this approach and put forward a new approach based on Empirical Bayes.

The Bayesian approach to statistics treats unknown parameters as random variables, and prior distributions model information about parameters. In contrast, the classical approach to statistics has no need of prior distributions as it treats unknown parameters as fixed constants. Empirical Bayes is an approach to statistics that lies somewhere between the two. Unknown hyper-parameters in Empirical Bayes are treated as fixed constants (as are the parameters in the classical approach) but in general these are estimated from data unlike in the standard Bayesian
approach. We refer to the new book by Efron (2010) for further details including some real and insightful examples.

4 Climate change and Data Assimilation

We now come to a totally different problem, namely that of Climate Change. Over the past 20 years, there has been a gradual shift away from a purely deterministic approach to weather forecasting towards a more integrated probabilistic approach; see, for examples, Lorenc and Hammon (1988) and Bengtsson et al (2003).

Like any mathematical modelling of a complex dynamic system, the traditional deterministic approach to weather forecasting seeks to solve nonlinear partial differential equations by numerical approximation, integrating forward in time simplified versions of these differential equations. Limitations in this arise due to the deterministic chaotic nature which such geophysical processes exhibit. Data assimilation improves on numerical weather forecasting accuracy through the combination of deterministic modelling and observational data. In short, data assimilation inserts spatially weighted observational data into a deterministic model to constrain the model to more accurately represent the “true” atmospheric state. In data assimilation, the information from new observations is then used to modify the model state, to be as consistent as possible with the observations and the previous information. The utilization of computer models for complex real-world processes requires addressing Uncertainty Quantification (UQ). Thus, corresponding issues range from inaccuracies in the models to uncertainty in the parameters or intrinsic stochastic features.

4.1 Falkland Islands Weather Data

The paper of Quinn et al (2004) (LASR Proceedings) gives such an example for a Falkland island related to weather prediction for landing of airplanes at an airport between two mountain ridges. During 2000 and 2001, an array of 20 automatic weather stations was used to collect high temporal resolution surface data in order to characterise the occurrence of weather phenomena in the Falkland Islands. An example of severe turbulence occurs in the lee of two mountain ridges, each of approximately 600m in height, on East Falkland. This phenomenon is routinely observed from the island airport situated at Mount Pleasant and presents a severe aviation hazard. Figure 4 shows the final sites chosen for the weather stations, with a large number clustered around the airport to capture surface data in this region. Prior to the field experiment it was not known what the optimum positions of sites would be in order to maximize the amount of information captured. The positions of the stations are not topographically equivalent, with some being placed on the top of ridges and others being situated at the bottom of valleys between ridges. Several stations are situated in extremely close proximity to the runway itself (shown in Figure 4).

The 20 automatic weather stations recorded data as 30 second averages over a period of approximately one year, meaning a potential maximum of around 1,051,200 records per station over the duration of the experiment. However, due to the nature of the experiment, some of these stations performed better than others, and data sets between stations range from sporadic to near uninterrupted. The high resolution data assimilation methodology being considered envisages that, via statistical interpretation of the spatial and temporal data, appropriate weightings may be attributed to observations collected at certain weather stations. In turn these weighted observations anticipate the development of statistically derived “nudging constraints” to be incorporated into the three-dimensional weather prediction model based on an extended form of the Navier-Stokes equation to improve its forecast accuracy.
4.2 The Navier-Stokes equation

The Navier-Stokes equation of fluid motion is

$$\rho \left( \frac{\partial \mathbf{v}}{\partial t} + \mathbf{v} \cdot \nabla \mathbf{v} \right) = -\nabla p + \nabla \cdot \mathbf{T} + \mathbf{f}$$  \hspace{1cm} (1)

where \( \mathbf{v} \) is the flow velocity, \( \rho \) is the fluid density, \( p \) is the pressure, \( \mathbf{T} \) is the stress tensor, \( \mathbf{f} \) represents body forces (per unit volume) acting on the fluid and \( \nabla \) is the del operator. This equation has some critical boundary conditions as well as requiring grid selection in practice. We can use a spatio-temporal statistical model for this purpose.

Figure 4: Falkland Island Airport, and the monitoring stations having the two ridges around the stations 12 and 15.

5 Concluding Remarks and Holistic Statistics

5.1 Science and Statisticians

Let us examine the current status of statisticians in fundamental scientific breakthroughs (Terry Speed, LASR 2010):

- Initial: by people in the field (rarely statisticians) who develop methods that work, and whose results are published in the high-profile journals, perhaps mainly by virtue of novelty and relevance;
- Intermediate: by statisticians in the field, perhaps doing rather better than those running the initial phase, perhaps presenting more lasting solutions;
• External: by statisticians from outside the field, hoping to develop new theories and methods, and perhaps making a difference here but, if not, then adding to the corpus.

Perhaps this is somewhat different from the early days of the subject, when statisticians such as Galton and R.A. Fisher played a leading role in interdisciplinary research. It seems that with the floods of large-scale data, computer scientists and statisticians with computing skills have a major part in creating impact.

5.2 Robot Statisticians

Could it be that the “Wheel of Research” will become more automated (see Figure 5)? Already there are dedicated softwares for some areas. A Robot Statistician can operate as follows: it has the input of data and background knowledge, followed by analysis and then suggestions for robotic statistical design. It then leads for consistent hypothesis followed by model/simulation. The output of Results is listed which is then iterated to New Design and finally a coherent hypothesis. Before this happens, there is a tremendous opportunity for statisticians to make a stronger impact in emerging subjects such as Bioinformatics/Life Sciences. But statisticians need to be:

• more open;
• more ready to learn “molecular biology”;
• more computationally aware;
• more ready to understand data banks ...!

Currently, in some sense as a profession, statistics suffers from more and more delusion.

“The bulk of current statistical research appears to be finding exact solutions to wrong problems instead of approximate solutions to right problems” (attrib. John Tukey).

In fact, 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 (see Figure 6);

• Third, a change of attitude is required by statisticians — a **paradigm shift** — for the subject to go forward.

![Figure 6: Holistic Statistics: encompassing several statistical areas.](image)

©K.V.Mardia

*Cure: “Go Holistic!”*

## 6 Acknowledgments

I wish to express my thanks to Chris Fallaize and Zhengzheng Zhang for their help.

## 7 References


Genome-wide protein structure prediction and structure-based function annotations

Yang Zhang

Center for Computational Medicine & Bioinformatics, Department of Biological Chemistry, University of Michigan

The biological function of protein molecules is dictated by the shape of their three-dimensional structures. Is it possible to predict protein structure and function from the amino acid sequence? We developed a new algorithm, I-TASSER, which assembles atomic structure of unknown proteins using fragments excised from experimentally solved structures of other proteins. Functional insights (e.g. ligand-binding affinity, enzyme classification and gene ontology) are then deduced by matching the predicted structure models with the known proteins in protein function libraries. The I-TASSER algorithm was tested for automated protein structure prediction in the communitywide CASP experiments of 2006, 2008 and 2010; it was also tested for protein function annotation in CASP9 in 2010.

In this talk, we first review the recent progress in protein structure prediction including the new developments in ab initio folding and atomic structure refinements since the CASP9 experiment, and show that the protein structure prediction problem can in principle be solved using the current PDB library. Next, we discuss the application of the developed methods to the structural and functional modeling of a number of genomes, including all G-protein coupled receptors (GPCRs) in the human genome, yielding models of which are shown to have correct topology, and Marek’s disease virus, the first success of the computational modeling of a complete viral genome. Finally, we demonstrate how the predicted I-TASSER structure models can be used to annotate the biological function of the proteins and screening drug candidates by matching their global topology and active sites against the existing structure/function/binding databases.

Despite the completeness of protein structures in the PDB, structural connections between most of the non-homologous protein pairs cannot be established by current threading/fold-recognition methods. While the current structural genomics efforts are helpful to bridge the gap, new method developments for sensitive distant-homology detection are demanding. The development of extended template libraries by artificial sequence and structure design may be useful to solve the problems. Work along this line is under progress.

References


7. Schwede T. Assessment of Function Prediction in CASP9 Experiment. 2010 December 5-9; Asilomar Conference Grounds, Pacific Grove, CA.


Full Bayesian analysis of the generalized non-isotropic Procrustes problem with scaling

Douglas L. Theobald and Kanti V. Mardia

1 Department of Biochemistry, Brandeis University
2 Department of Statistics, University of Leeds

1 Introduction

An object’s shape is defined as the geometrical information about an object that remains when translational, rotational, and scaling effects are removed. Hence, shape is invariant under the corresponding similarity transformations. To compare and contrast the shapes of different objects, it is first necessary to superimpose them in some optimal fashion.

An object can often be reduced to a set of \( k \) landmarks (Cartesian points in \( d \) dimensions) and represented as a \( k \times d \) matrix (a configuration of landmarks for a geometrical object). For example, in the classical molecular biology approach, the structures of biological macromolecules are described concisely by the three-dimensional coordinates of their atoms. Structures are conventionally referred to a common reference frame using the statistical optimization method of ordinary least-squares (OLS) (Flower(1999); Dryden and Mardia(1998)). The least squares criterion stipulates that the optimal transformations are those that minimize the sum of squared distances among corresponding landmarks in the objects. OLS assumes that all landmarks have the same variance and are uncorrelated, yet both conditions are frequently violated in real data.

In previous work, we relaxed the assumptions of homoscedasticity and non-correlation by treating the superposition problem within a likelihood framework where multiple structures are distributed normally (Theobald and Wuttke(2006a); Theobald and Wuttke(2006b); Theobald and Wuttke(2008)); some theoretical work on such models appeared in Goodall and Mardia (1993). This ML method effectively accounts for uneven variances and correlations in the landmarks by weighting by the inverse of the covariance matrix. Notably, this work did not address issues of scaling, as the size of a macromolecule is fixed by the physics of chemical bonding.

Simultaneous point estimation of the sample covariance matrix and the translations is generally impossible due to identifiability issues, which has been a significant impediment to a viable non-isotropic Procrustes analysis (Dryden and Mardia(1998); Lele and Richtsmeier(1990); Lele(1993); Lele and Richtsmeier(2001); Glasbey et al.,(1995); Goodall(1991b)). We allow joint identifiability by regularizing the covariance matrix using a hierarchical, empirical Bayes approach in which the eigenvalues of the covariance matrix are treated as variates from an inverse gamma distribution. This hierarchical method is analogous to putting a conjugate inverse Wishart prior on the covariance matrix. An expectation-maximization implementation of this method performs well in practice.

Here we describe a Bayesian extension of this matrix normal model for the generalized non-isotropic Procrustes problem with scaling. Building on previous work, this analysis applies to multiple configurations (as opposed to pairwise Procrustes), uses an arbitrary covariance matrix (as well as a diagonal and isotropic covariance matrix as special cases), and allows for proper conjugate priors for all parameters. Generalizing to full shape analysis with scaling is non-trivial, as the conditional posterior distribution of the scale factors turns out to be a non-standard form (which we christen the halfnormal-gamma). We have developed rejection and Metropolis-Hastings algorithms for simulating from this distribution.
2 Methods

2.1 A matrix normal probability model for the macromolecular superposition problem

Consider \( n \) structures \((X_i, i = 1 \ldots n)\), each with \( k \) labelled landmarks, where each structure is described by a \( k \times d \). We assume that each structure \( X_i \) is normally distributed and is observed in an arbitrary coordinate system (Dryden and Mardia(1998); Goodall(1991a); Goodall and Mardia(1993)). Heterogeneous variances and correlations among the landmarks are described by a \( k \times k \) covariance matrix \( \Sigma \) (isotropic in the \( d \)-dimensional space). Hence each \( X_i \) can be considered to be an arbitrarily scaled, rotated, and translated zero-mean normal matrix displacement

\[
E_i \sim N_{K,D}(0, \Sigma, I)
\]

where \( \beta_i \) is a global scale factor, \( R_i \) is a \( d \times d \) orthogonal rotation matrix, \( t_i \) is a \( d \times 1 \) column vector for the translational offset, and \( 1_K \) denotes the \( k \times 1 \) column vector of ones.

\[
X_i = \frac{1}{\beta_i} (M + E_i) R'_i - 1_K t'_i
\]

2.2 A Procrustes matrix normal likelihood function

The full joint likelihood function for the model given in (1) is obtained from a matrix normal distribution (Dawid(1981)). Define

\[
Y_i = (\beta_i X_i + 1_K t'_i) R_i
\]

then the PDF for the likelihood function is:

\[
p(X|R, t, \beta, M, \Sigma) = C \exp \left( -\frac{1}{2} \sum_{i=1}^{n} \text{tr} \left\{ \left[ Y_i - M \right] \Sigma^{-1} \left[ Y_i - M \right] \right\} \right),
\]

with normalization constant:

\[
C = (2\pi)^{-\frac{kdn}{2}} \left( \prod_{i=1}^{n} \beta_i^{k/d} \right) |\Sigma|^{-\frac{dn}{2}}.
\]

2.3 A Bayesian extension

The likelihood analysis described above does not provide ready estimates of the uncertainty in the estimated parameters. In an earlier presentation at this conference (Theobald, 2009), a Bayesian extension was described allowing for the incorporation of other prior data. For the Bayesian analysis we assume that \( \Sigma, M, R, t, \beta \) are all independent, so that

\[
p(\Sigma, M, R, t, \beta | X) \propto p(X | \Sigma, M, R, t, \beta) p(\Sigma) p(M) p(R) p(t) p(\beta).
\]

We will also assume a hierarchical prior for \( \Sigma \):

\[
p(\Sigma) \propto p(\Sigma | \delta, n) p(\delta).
\]
2.3.1 Conditional probability of the scale factors $\beta$

The conditional probability density function of $\beta_i$ is given by

$$
p(\beta_i | X, \Sigma, M, t_i, R) = C_i \beta_i^{m-1} \exp \left\{ -\frac{\phi_i}{2} \beta_i^2 - \beta_i \gamma_i \right\}, \quad (6)
$$

where $C_i = \frac{2\phi_i}{\Gamma(m) D_{m-\frac{m}{2}}(\frac{\gamma_i}{\sqrt{2}\phi_i})}$, $\phi_i = \beta_i^2 \operatorname{tr}(\tilde{X}_i \Sigma^{-1} \tilde{X}_i)$, $\gamma_i = \beta_i \operatorname{tr}(M \Sigma^{-1} \tilde{X}_i R_i)$, $m = kd + 1$, $\Gamma(m)$ is a “parabolic cylinder function”, a type of confluent hypergeometric function, defined in Gradshteyn and Ryzhik p 1028, section 9.24-9.25 (also described as the Whittaker function in Chapter 19 of Abramowitz and Stegun). This distribution (halfnormal-gamma) has been studied in Mardia et al., (2011), where the isotropic single global scaling case (pairwise superposition) has been treated as well.

References


Session II
Exploiting Protein Family Information to Predict Protein Functions and Functional Networks.

Juan Ranea, Jonathan Lees, Ian Morilla, Jean-Karim Heriche, Zac Hussain, John Shawe-Taylor, Robert Rentzsch, David Lee and Christine Orengo*

University College London, London, UK (UCL)

1 Background: Functional Classification in the CATH database of Domain Structure Superfamilies.

We have been developing methods which exploit homology to infer shared molecular functions and shared biological processes between related proteins. Our approaches benefit from a comprehensive classification of proteins into evolutionary superfamilies in the CATH database (Cuff, Sillitoe et al. 2009). Protocols based on both structure comparisons and searches against sequence profiles and HMMs are used to identify homologues. CATH currently classifies more than 110,000 domain structures and nearly 10 million domain sequences (Lees, Yeats et al. 2010) into 2600 evolutionary superfamilies. For each superfamily a large amount of publicly available information on protein functions (e.g. from KEGG (Kanehisa, Araki et al. 2008), GO (Ashburner, Ball et al. 2000), IntAct (Kerrien, Alam-Faruque et al. 2007)) is integrated into the classification enabling studies into the evolution of protein functions. Typically over 60% of proteins in a genome can be reliably annotated with a CATH domain. Nearly 80% of domain sequences in completed genomes can be assigned to a domain structure superfamily in CATH suggesting that this classification is highly representative of protein families in nature. Although, there is under-representation of membrane associated proteins, structural genomics initiatives in the United States are specifically targeting these proteins with the aim of increasing their representation in the PDB, CATH and other related resources.

Analyses of CATH superfamilies have revealed the extent to which protein functions can diverge within superfamilies. About 100 (<5%) of the domain superfamilies in CATH have been highly duplicated in the genomes and account for nearly one third of all domain sequences in CATH. Many of these large superfamilies are universal to all kingdoms of life and in some of them relatives have diverged considerably in their structures and functions. A recent analysis of the large HUP domain superfamily revealed that relatives were associated with more than 300 GO molecular function terms and were acting in metabolic pathways, protein biosynthesis and signaling pathways.

We have recently developed a computational approach (Lee, Rentzsch et al. 2009) which uses a sequence based profile-profile protocol to perform agglomerative clustering of relatives possessing highly similar molecular functions. These clusters or “functional families”, as they are described, enable more accurate inheritance of function for uncharacterized relatives. We have generated a library of functional family HMMs, for all domain superfamilies in CATH, against which a new sequence relative can be scanned to assign a putative function. Using this approach we have been able to extend the functional annotations of domain sequences in CATH by up to 7-fold in some superfamilies.

This approach has been successfully benchmarked against 8 very large superfamilies in the Structure Function Linkage Database (SFLD) established in the Babbitt laboratory, for which there is a large amount of experimental data characterising the relatives. GeMMA has been shown to perform well and is highly competitive with other approaches which use tree-based methods for sub-classifying functional relatives within superfamilies. We are currently assess-
ing the performance of this approach by means of a blind test being conducted as part of the CAFA Automatic Function Prediction SIG at the ISCB meeting in Vienna this year.

2 Using the CATH Classification and Available Experimental Data to Predict Protein Functional networks

In addition to providing information on molecular functions for domains in CATH, we have recently extended the CATH classification by integrating information on the biological processes and protein interactions in which domain relatives participate. Data has been extracted from a wide range of public resources (e.g. HPRD (Keshava Prasad, Goel et al. 2009), IntAct (Kerrien, Alam-Faruque et al. 2007), MINT (Chatr-aryamontri, Ceol et al. 2007), Reactome (Matthews, Gopinath et al. 2009), BioGrid (Breitkreutz, Stark et al. 2008), DIP (Xenarios, Rice et al. 2000)). However, it is clear that this information still only represents a small proportion of all the interactions in which proteins participate. For example, Stumpf et al. performed simulations on the human interactome which suggested that fewer than 10% of interactions involving human proteins had been experimentally characterised (Stumpf, Thorne et al. 2008). Therefore, to enrich the information on protein networks for human and other model organisms, we have developed a suite of prediction algorithms some of which exploit the CATH protein family information.

For example, the HiPPI protocol (Homology Inferred Protein Protein Interactions) uses CATH superfamilies to inherit protein interactions (Ranea, Morilla et al. 2010). Whilst the CODA algorithm (Co-Occurrence of Domain Associations) uses the protein family information to infer an association between two proteins in one organism if they are found fused together in other organism (Reid, Ranea et al. 2010). In order to resolve the noise caused by multiple copies of a domain superfamily within an organism a weighting scheme was introduced and optimized using benchmark datasets. The phylotuner method (Ranea, Yeats et al. 2007) infers an association between two proteins if the presence and absence of their domain superfamilies in a large set of competed genomes are correlated. As well as exploiting domain family information for predicting protein associations we have also applied approaches based on experimental data. For example the GECO protocol (Gene Expression COrrelations) identifies pairs of proteins with correlated patterns of gene expression under different experimental conditions (Ranea, Morilla et al. 2010). Large amounts of such data is now readily available from the ArrayExpress repository at the EBI, for human and other organisms.

Since all these methods are still very noisy we have integrated them using Fishers weighted statistic to improve the signal (see figure 1) which gives rise to a considerable increase in precision. Recent application of these approaches to predicting protein interactions and associations in human and the model organism yeast, showed a significant increase in the number of proteins predicted to be participating in interactions (See figure 2). Protein functional networks generated using this association information exhibited similar characteristics to those generated using purely experimental data, validating our approach. In addition, we were able to identify highly connected proteins not yet experimentally characterized but predicted to be major player in signaling networks and other regulatory processes. Protein interactions and associations associated with these processes tend to be transient and harder to capture using current experimental technologies.

To further improve the precision of our predictions we included additional, orthogonal approaches in our prediction pipeline. Predictions from the iHOP (Hoffmann and Valencia 2005) and Co-Cite methods, developed in the group of Alfonso Valencia, CNIO, Madrid, which exploit textmining to detect protein interactions and associations were integrated, as were predictions from the machine learning approach (MLNN) developed in the Group of Soren Brunak, Technical University, Denmark. To experimentally verify the performance of the integrated
methods we applied them to predict proteins associated with the mitotic spindle from a large set of putative proteins identified by a proteomics experiment. Experimental testing of 20 proteins from the top of a ranked list generated by our approach involved gene knockouts and co-localisation studies and gave a 75% success rate for our integrated pipeline compared to a 35% success rate for selecting targets by a simple bioinformatics approach using BLAST to recognize putative spindle proteins.

3 Predicting Host Virus Protein Interactions Associated with Herpes Virus (HV)

Primary effusion lymphoma (PEL) is a disease specific to B-Cells. Malignancies of this type occur in immunosuppressed patients and most commonly with AIDS patients. The presence of Kaposi’s Sarcoma Herpes Virus (KSHV) is a near universal requirement for the PEL. The mechanism of host cell transformation by KSHV is only partly understood. Elucidating the human genes targeted for B-Cell transformation to PEL would provide useful steps in developing therapeutic strategies for this highly prevalent disease.

In the current strategy, since this is a largely unknown system (unlike the highly covered spindle system with additional proteomics set), we decided to include experimental datasets (ie from public repositories such as IntAct or the literature) rather than rely fully on ab-initio sequence based predictions. The computational challenge then is how to best leverage the rapidly expanding protein association data.

We assembled 4 different protein-protein association networks for the human genome namely database interactions, text mined interactions, interactions inherited by orthology and Gene Ontology semantic similarity based associations. We also included known KSHV host-pathogen and pathogen-pathogen protein associations. These were collected from databases such as Virus-Mint, and various high throughput studies in the literature. Hence we carried out a kind of in-silico infection of the human genome. This allows us to use the viral genes as known starting positives and search for human proteins which could be associating with them. This may prove to be a valuable approach since most of the genes known to be involved in PEL are viral.

Figure 1: Improvements in the precision of predicting human protein associations obtained by integrating several independent approaches using Fishers weighted statistic.
We used these information rich networks to provide a ranked list of targets to be experimentally verified by the group of Paul Kellam at the Sanger Centre, UK. Because of the way the experiments will be carried out (one or two at a time) and the low level of current knowledge we have decided to use an online learning strategy. In this way a few targets are tested at a time, the results of which are then fed back into the computational strategy to give an improved ranking. The strategy we have chosen to use is the LinRelMKL algorithm (Hussain et al, 2008). This provides some desirable features including reweighting of kernels to find better combinations
and a search strategy that ensures good coverage in the network. The search algorithm will be important for avoiding regions of the network that are activated but irrelevant to our study (e.g., ER metabolism). The strategy (figure 4) requires a starting set of known positives (genes demonstrating the phenotype of interest) and ideally negatives although these can be added in during later rounds. The Kellam group has conducted a meta-analysis of microarray data and have identified a few hundred genes that are upregulated specifically in PEL in comparison to nearby/related cell types. This can be used as an optional orthogonal filtering step to provide extra PEL biological context to the ranking.

Using the known PEL-KSHV transforming genes (LANA,K1,K12,K13, v-flip, vIL6, vIRF1) as our positives and PEL cell culture as a model for the PEL disease we have ranked human genes according to involvement in involved in PEL oncogenesis / transformation. We have also ranked human genes likely to be involved in PEL terminal differentiation, using genes known to maintain the PEL state (PRDM1, IRF4, STAT3, c-FOS, SPI-B, XBP1) as our positives.

We have validated our approach in-silico by benchmarking on pathways in the KEGG database. Figure 5 shows that by integrating the different predictions using kernel based approaches we can significantly increase the number of interactions/associations ranked at the top of the list.

4 The FuncNet and Fun-L Websites providing ranked lists based on predicted interactions/associations

To provide our results to the wider community we have developed a robust and coherent network of services “FuncNet”. This pipeline combines various protein association predictors and
Figure 5: Increase in the number of true positives (red line) identified in a ranked list of proteins by using kernel based approaches to integrate experimental data (pale blue) with predictions based on the gene ontology (dark blue), homology (yellow) and text mining (brown).

enables distributed queries on several different FuncNet prediction methods. FuncNet was developed as part of the EU funded ENFIN network for Systems Biology, headed by Ewan Birney at the EBI and has an ENFIN XML interface and is hence compliant with ENFIN web services and available for querying via ENFIN’s EnVision 2 web portal. Another resource FunL-L (Functional Lists) (http://funl.org) provides ranked lists of protein interactors for query sequences in several model organisms including human.

References


A statistical view on the reference ratio method

Kanti V. Mardia¹, Jes Frellsen², Mikael Borg², Jesper Ferkinghoff-Borg³ and Thomas Hamelryck²∗

¹Department of Statistics, University of Leeds.
²Bioinformatics Centre, University of Copenhagen.
³Department of Electrical Engineering, Technical University of Denmark.

1 Introduction

The recently introduced reference ratio method (Hamelryck et al. (2010)) allows combining distributions over fine grained variables with distributions over coarse grained variables in a meaningful way. This problem is a major bottleneck in the prediction, simulation and design of protein structure and dynamics. Hamelryck et al. introduce the reference ratio method in this context, and show that the method provides a rigorous statistical explanation of the so called potentials of mean force (PMFs). These potentials are widely used in protein structure prediction and simulation, but their physical justification is highly disputed (Thomas and Dill(1996); Ben-Naim(1997); Koppensteiner and Sippl(1998)). The reference ratio method clarifies, justifies and extends the scope of these potentials.

As the reference ratio method is of general relevance for statistical purposes, we present the method here in a general statistical setting. Subsequently, we discuss two example applications of the method. First, we present a simple educational example, where the method is applied to independent normal distributions. Secondly, we reinterpret an example originating from Hamelryck et al.; in this example, the reference ratio method is used to combine a detailed distribution over the dihedral angles of a protein with a distribution that describes the compactness of the protein. Finally, we outline the relation between the reference ratio method and PMFs.

2 Reference ratio method

We start by introducing the reference ratio method using general statistical terms. Let \( f(x) \) be the probability density function (pdf) of \( X \), which is unknown, but

(i) the pdf \( f_1(y) \) of \( Y = m(X) \) is known for \( f(\cdot) \), where \( m(\cdot) \) is a specified many-to-one function, and

(ii) the pdf \( g(x) \) is specified and approximately close to \( f(x) \), in the sense that \( f_2(x|y) \approx g_2(x|y) \) for all practical purposes.

Here, \( f_2(x|y) \) denotes the conditional pdf of \( X \) given \( Y \) for \( f(\cdot) \), and \( g_2(x|y) \) denotes the corresponding conditional pdf for \( g(\cdot) \). Note that these two conditional pdfs are not specified and that their closed form expressions are not necessarily easily expressed. In the work of Hamelryck et al., \( X \) is denoted the fine grained variable and \( Y \) the coarse grained variable due to their functional relation.

Now assume that we want to construct a new density \( \hat{f}(x) \), close to \( f(x) \), such that

(iii) the marginal pdf of \( Y \) for \( \hat{f}(\cdot) \) is equal to \( f_1(y) \) and

(iv) the conditional pdf of \( X \) given \( Y = y \) for \( \hat{f}(\cdot) \) is equal to \( g_2(x|y) \).

In other words \( \hat{f}(x) \) should have the properties that \( \hat{f}_1(y) = f_1(y) \) and \( \hat{f}_2(x|y) = g_2(x|y) \), where \( \hat{f}_1(y) \) and \( \hat{f}_2(x|y) \) respectively denotes the marginal distribution of \( Y \) and the conditional distribution of \( X \) given \( Y \) for \( \hat{f}(\cdot) \). It would be straightforward to construct \( \hat{f}(x) \) if the
conditional pdf $g_2(x|y)$ was known. In particular, generation of samples would be efficient, since we could sample $\tilde{y}$ according to $f_1(\cdot)$ and subsequently sample $\tilde{x}$ according to $g_2(\cdot|y)$, if efficient sampling procedures were available for the two distributions. However, as previously stated $g_2(x|y)$ is generally not known. An approximate solution for sampling could be to approximate the density $g_2(x|y)$ by drawing a large amount of sample according to $g(x)$ and retain those with the required value of $Y$. Obviously, this approach would be intractable for a large sample space.

The solution to the problem was given by Hamelryck et al. in the form of a closed form expression for $f(\cdot)$. The authors showed that the conditions (iii) and (iv) are satisfied for the pdf given by

$$f(x) = \frac{f_1(\tilde{y})}{g_1(y)} g(x),$$

where $\tilde{y} = m(x)$. The construction of the density $\hat{f}(\cdot)$ in the above expression is known as the reference ratio method and the corresponding distribution is denoted the reference ratio distribution (Hamelryck et al., 2010). It is easy to check that $\hat{f}(\cdot)$ is properly normalized, $\int \hat{f}(x) \, dx = 1$, and that the two conditions (iii) and (iv) are satisfied.

Since $Y$ is a function of $X$, the joint pdf, $g_3(x, y)$, of $(X, Y)$ for $g(\cdot)$ is zero for all values $(x, y')$, where $y' \neq m(x)$, and we can write the joint pdf as

$$g_3(x, y) = g(x) \delta(y - m(x)),$$

where $\delta(\cdot)$ is the Dirac delta function. Consequently, we can also express the result from equation (1) as

$$\hat{f}(x) = f_1(\tilde{y}) g_2(x|\tilde{y}).$$

Based on this expression, we can recast the result as follows: the pdf $f(x)$ is unknown, but its marginal density $f_1(y)$ is known and an approximation, $g_2(x|y)$, of $f_2(x|y)$ is indirectly available through $g(x)$ to approximate the density $f(x)$. In the following we will present two applications of the reference ratio method.

### 3 Example with independent normals

The purpose of our first example is purely educational. It is a simple toy example based on independent normal distributions, which simplifies the functional form of the pdfs involved. Let $X = (X_1, X_2)$, where $X_1$ and $X_2$ are independent normals with

$$X_1 \sim \mathcal{N}(\mu, 1) \quad \text{and} \quad X_2 \sim \mathcal{N}(0, 1).$$

Accordingly, the pdf of $X$ is given by

$$f(x) = c \, e^{-\frac{1}{2}(x_1-\mu)^2 - \frac{1}{2}x_2^2},$$

where $x = (x_1, x_2)$ and $c$ is the normalizing constant. For the distribution $g(x)$, which is approximately close to $f(x)$, let $X_1$ and $X_2$ be independently distributed as

$$X_1 \sim \mathcal{N}(0, 1), \quad X_2 \sim \mathcal{N}(0, 1).$$

Consequently the pdf of $X$ is given by

$$g(x) = d \, e^{-\frac{1}{2}x_1^2 - \frac{1}{2}x_2^2},$$

where $d$ is the normalizing constant. Suppose that $Y = m(X) = X_1$. This means that the marginal pdf of $Y$ for $f(\cdot)$ is

$$f_1(y) = C \, e^{-\frac{1}{2}(y-\mu)^2},$$
and for \( g(\cdot) \) the marginal density is
\[
g_1(y) = d' \, e^{-\frac{1}{2} y^2},
\]
where \( c' \) and \( d' \) are the appropriate normalizing constants. Note that \( g(x) \) is only a good approximation to \( f(x) \) for \( \mu \approx 0 \), but for both \( f(\cdot) \) and \( g(\cdot) \) the conditional density of \( X \) given \( Y \) is the same and equal to the pdf of the normal distribution \( \mathcal{N}(0, 1) \).

By applying the ratio method from equation (1), we obtain the expression
\[
\hat{f}(x) = \frac{c' \, e^{-\frac{1}{2}(x_1-\mu)^2} \, d \, e^{-\frac{1}{2} x^2 - \frac{x_1^2}{2}}}{d' \, e^{-\frac{1}{2} x^2}} = c \, e^{-\frac{1}{2}(x_1-\mu)^2 - \frac{1}{2} x^2}.
\]

In this example we observed that \( \hat{f}(\cdot) = f(\cdot) \), which is expected since the conditional distribution of \( X \) given \( Y \) is the same for both \( f(\cdot) \) and \( g(\cdot) \). Accordingly, it is now trivial to check that the marginal distribution of \( Y \) for \( \hat{f}(\cdot) \) is equal to \( f_1(\cdot) \) and that the conditional distribution of \( X \) given \( Y \) is \( g_2(x|y) \), as stated in (iii) and (iv).

Generally, the conditional pdf \( g_2(x|y) \) is only assumed to be approximately equal to \( f_2(x|y) \), which means that \( \hat{f}(\cdot) \) and \( f(\cdot) \) are not guaranteed to be equal. In fact, in most relevant applications of the reference ratio method this conditional distribution is unknown for \( f(\cdot) \). In next section we will consider such an example.

### 4 Sampling compact protein structures

A more realistic application of the reference ratio method is given by Hamelryck et al. In this example the method is used to sample compact proteins structures. We will recount the example here using the notation introduced above. The setup is as follows:

(a) Let \( f(x) \) be an unknown distribution of the dihedral angles \( x = \{(\phi_i, \psi_i) | i = 1, \ldots, n\} \) in a protein with a known sequence of \( n \) amino acids.

(b) Let \( Y = m(X) \) be the radius of gyration \( (r_g) \) of the protein, and assume that \( f_1(y) \) is a normal distribution with \( \mathcal{N}(22 \, \text{Å}, 4 \, \text{Å}^2) \).

(c) The pdf, \( g(x) \), of the approximating distribution is given by TorusDBN, which is a probabilistic model of local protein structure (Boomsma et al. (2008)).

(d) The marginal density \( g_1(y) \) is obtained by sampling from \( g(x) \), which can be done since TorusDBN is a generative model.

The reference ratio method is applied to construct the density \( \hat{f}(\cdot) \), based on the normal distribution over the radius of gyration, \( f_1(y) \), the TorusDBN distribution, \( g(x) \), and the marginal distribution over \( r_g \) for TorusDBN, \( g_1(y) \). It is important to stress that typical samples generated from TorusDBN, \( g(x) \), are unfolded and non-compact, while typical samples from \( \hat{f}(x) \) will be compact as the radius of gyration is controlled by the specified normal distribution. Accordingly, samples from the reference ratio distribution, \( \hat{f}(x) \), are expected to look more like folded structures than samples from \( f(x) \).

Hamelryck et al. test this setup on the protein ubiquitin, which consists of 76 amino acids. Figure 1 shows the distribution over \( y \) (\( r_g \)) obtained by sampling from \( g(x) \) and \( \hat{f}(x) \), respectively. The figure also shows the normal density \( f_1(y) \). We observe that samples from \( g(x) \) have an average radius of gyration around \( 27 \, \text{Å} \), while samples from \( \hat{f}(x) \) indeed have a distribution very near \( f_1(y) \). As expected, samples from \( \hat{f}(x) \) are compact, unlike samples from \( g(x) \). Examples of such samples are shown in figure 2.
A key question here is how can we sample from \( \hat{f}(x) \) efficiently? As described earlier, we would from a generative point of view use equation (2) directly and generate a sample, \( \tilde{x} \), using the two steps:

1. sample \( \tilde{y} \) according to \( f_1(y) \) and
2. sample \( \tilde{x} \) according to \( g_2(x|\tilde{y}) \).

However, a problem lies in step 2, as there is no efficient way to sample from \( g_2(x|y) \); Torus-DBN only allows for efficient sampling from \( g(x) \). One could consider using rejection sampling or the ABC method (Pritchard et al. 1999; Beaumont et al. 2002; Marjoram et al. 2003) for step 2, but both methods would be very inefficient. Hamelryck et al. (2010) have given a highly efficient method, which does not (in principle) involve any approximations. The idea is to use the Metropolis-Hastings algorithm with \( g(x) \) as proposal distribution and \( \hat{f}(x) \) as target distribution. In this case, the probability of accepting a proposed value \( x' \) given a previous values \( x \) becomes

\[
\alpha(x'|x) = \min \left(1, \frac{f_1(y')g_1(y')/g_1(y)}{f_1(y)g_1(y)/g_1(y')} \right)
\]

where \( y = m(x) \) and \( y' = m(x') \). In practice, the proposal distribution in the MCMC algorithm would only change a randomly chosen consecutive subsequence of \( X \) using TorusDBN (see supporting information of Boomsma et al. (2008) for details), as this leads to a higher acceptance rate. It can be shown that the acceptance probability in this case also is given by equation (3).

5 The reference ratio method explains PMFs

Methods for predicting the structure of proteins rely on an energy function or probability distribution that describes the space of possible conformations. One approach to constructing such energies or distributions is to estimate them from a set of experimental determined protein structures. In this case they are called knowledge based potentials.

A subclass of the knowledge based potentials are based on probability distributions over pairwise distances in proteins. These are called potentials of mean force (PMFs) and are loosely based on an analogy with the statistical physics of liquids (Ben-Naim 1997; Koppensteiner and Sippl 1998). The potential of mean force, \( W(r) \), associated with a set of pairwise distances \( r \) is given by an expression of the form

\[
W(r) \propto -\log \frac{f_1(r)}{g_1(r)},
\]

where \( f_1(r) \) is a pdf estimated from a database of known protein structure, and \( g_1(r) \) is the pdf of \( r \) for a so-called reference state. The reference state is typically defined based on physical considerations. The pdf \( f_1(r) \) is constructed by assuming that the individual pairwise distances are conditionally independent, which constitutes a crude approximation. In practice, the potential of mean force is combined with an additional energy function, that is concerned with the local structure of proteins. This additional energy term is typically brought in via sampling from a fragment library (Simons et al. 1997) – a set of short fragments derived from experimental protein structures – or any other sampling method that generates protein-like conformations. From a statistical point of view, this means that the samples are generated according to the pdf

\[
\hat{f}(x) \propto \frac{f_1(r)}{g_1(r)} g(x),
\]
Figure 1: The reference ratio method applied to sampling protein structures with a specified distribution over the radius of gyration ($r_g$). The distribution over $r_g$ for samples from TorusDBN, $g(x)$, is shown as triangles, while the $r_g$-distribution for samples from the ratio distribution, $\hat{f}(x)$, is shown as circles. The pdf, $f_1(y)$, for the desired distribution normal distribution over $r_g$ is shown as a solid line, $\mathcal{N}(22 \text{ Å}, 4 \text{ Å}^2)$. The samples are produced using the amino acid sequence of ubiquitin. The figure is adapted from figure 3 in (Hamelryck et al. (2010)).

where $x$ are the dihedral angles in the protein, $r$ are the pairwise distances implied by $x$, and $g(x)$ is the pdf of the dihedral angles embodied in the sampling method.

In this formulation, it can be seen that PMFs are justified by the reference ratio method; their functional form arises from the combination of the sampling method (which concerns the fine-grained variable) with the pairwise distance information (which concerns the coarse-grained variable). This interpretation of PMFs also provides some surprising new insights. First, $g_1(r)$ is uniquely defined by $g(x)$, and does not require any external physical considerations. Second, if the three involved probability distributions are properly defined, the PMF approach is entirely rigorous and statistically well justified. Third, the PMF approach generalizes beyond pairwise distances to arbitrary coarse-grained variables. In conclusion, the reference ratio method settles a dispute over the validity of PMFs that has been going on for more than twenty years, and opens the way to efficient and well-justified probabilistic models of protein structure.

6 Acknowledgements

The authors acknowledge funding by the Danish Program Commission on Nanoscience, Biotechnology and IT (NaBiIT, project: Simulating proteins on a millisecond time-scale, 2106-06-0009).

References


Ben-Naim, A. (1997). Statistical potentials extracted from protein structures: Are these mean-
Figure 2: The reference ratio method applied to sample compact protein structures. The left picture shows twenty typical, non-compact samples obtained using TorusDBN alone, $g(x)$. The right picture shows twenty typical, compact samples from the reference ratio distribution, $\hat{f}(x)$, which is obtained by combining TorusDBN, $g(x)$, with a suitable normal distribution over the radius of gyration, $f_1(y) = N(22 \text{ Å}, 4 \text{ Å}^2)$. We used the amino acid sequence of ubiquitin.


Metabolic network evolution and operation in unicellular eukaryotes

David R. Westhead*, John W. Whitaker, Thomas Forth and Glenn A. McConkey

Faculty of Biological Sciences
University of Leeds

We have recently developed metaTIGER¹, a WWW resource devoted to the study of metabolic networks and their evolution. Using our metaSHARK software for metabolic annotation of genomes, we have annotated metabolic enzyme functions in over 500 species, including more than 100 eukaryotes. The WWW site has easy to use facilities for viewing and comparing the metabolic networks in different organisms via highlighted pathway images (using information from KEGG) and tables. metaSHARK provides genome annotations using raw nucleic acid sequences as input, predicting gene models using matching of profile hidden Markov models (HMMs), and is therefore able to annotate eukaryotes for which gene predictions are not available. This allows access to species for which only preliminary genome or EST sequence data exists, or for which gene structures are hard to predict owing to lack of suitable training information. We expect these features to be increasingly useful as new sequencing methods allow the sequencing of many more genomes for which detailed manual annotation will be impossible.

The main novelty of the metaTIGER site is the inclusion of significant amounts of evolutionary information. For each metabolic function (defined by Enzyme Commission (E.C.) number) the corresponding enzyme sequences were used to create a maximum likelihood phylogenetic tree, resulting in a comprehensive database of 2,257 trees. These trees were created from just the most conserved parts of the enzyme, and were limited to include only sequences which are very confident hits to the profile/HMM (E value < 10^-30). In addition, for each genome a (sub-)sequence was only associated with an E.C. number if it was the best match to the corresponding profile/HMM within the genome and was not a more confident match to any other profile/HMM. Thus the sequence sets for our trees eliminate the inclusion of paralogs as far as possible, and are designed to give the highest possible quality phylogenetic trees. The site contains facilities for viewing the trees using the state-of-the-art tree viewer iTOL. In addition there are tree query facilities which allow the user to search for trees with particular clade structures, for instance including potential horizontal gene transfers, or to find sequences suitable for concatenation to infer consensus organism phylogenies.

We have recently used the facilities described above to make the most comprehensive survey yet available of the horizontal transfer of metabolic genes between bacteria and unicellular eukaryotes. The 30 eukaryotic species with complete genome sequences considered derive from 10 eukaryotic genera, and reveal significant levels of confidently asserted transfers. The species set contains several significant groups of parasites of human and economic importance, and some of the transfers we find are potentially related to pathogenicity and the adaptation to a parasitic lifestyle.

Finally our research has moved in the direction of systems biology of metabolic networks, where we have been using flux balance analysis alongside experimental measurements on the malaria parasite to provide predictions of growth and behaviour. I will describe some preliminary results in this area.

¹http://www.bioinformatics.leeds.ac.uk/metatiger/
Structural bioinformatics: current status and future directions

Swanand Gore*, Sameer Velankar and Gerard J. Kleywegt

Protein Data Bank in Europe (PDBe)
EMBL-EBI
Wellcome Trust Genome Campus, Hinxton, Cambridge

1 Introduction

Structural biology is a key discipline in basic and applied biological research. It reveals atomic and mechanistic details of biological macromolecules in normal and diseased states and enables researchers to modulate the molecular machinery in a rational manner, through the design of important molecules of practical relevance such as drugs, inhibitors, enzymes, antibodies, pesticides etc. Structural bioinformatics - an umbrella term encompassing many techniques in biocomputing and informatics of macromolecular structures - is an essential component of structural biology due to the quantity and complexity of structural data.

2 An ensemble of sub-disciplines

Structure determination - the core technique in structural biology - is made possible by computational refinement of structure models using experimental data and augmenting it with prior knowledge e.g. of basic covalent geometry. While X-ray crystallography and NMR spectroscopy remain the most popular methods, cryo-electron microscopy (Chiu et al.(2005)) and hybrid experimental techniques (Wanga et al.(2011)) are rapidly becoming more important as the field moves towards studying the structures of large assemblies and molecular machines. This has necessitated the development of more powerful structure refinement procedures (Tang et al.(2007); Adams et al.(2010)). High-throughput crystallography protocols have been developed for structural genomics (Terwilliger et al.(2009)) and fragment-based drug-discovery initiatives (Hajduk and Greer(2007)).

Structure prediction is crucial for proteins whose structures are experimentally intractable and is carried out using techniques such as comparative (homology) modelling, threading and ab-initio methods. Homology models are now available on a genomic scale (Pieper et al.(2004)) produced using new powerful algorithms for mainchain and sidechain modelling (Krivov et al. (2009); Eswar et al.(2006)). New approaches are being developed to model remote homology more reliably (Zhou and Skolnick(2010)) and to include experimental restraints wherever available (Möglich et al.(2005)). Cutting-edge technologies like Rosetta and folding@home have been very successful (Raman et al.(2009); Beberg and Pande(2009); Cooper et al.(2010)) for folding structures in silico where no suitable homologous structures exist. In the related field of molecular dynamics simulations, new methods are being developed towards speed and robustness (Shaw et al.(2010)), to include quantum-mechanical effects (Kamerlin et al.(2009)), and coarse-graining (Hall and Sansom(2009)).

Possible errors in experimental or predicted structures may be detected by validation criteria (Kleywegt(2009); Kleywegt(2000); Chen et al.(2010)) such as the Ramachandran plot, real-space-R value etc. This is becoming more relevant due to a variety of non-expert users taking to structural studies (Kleywegt(2009)).

The classification of structures is an important area of structural bioinformatics, and uses structure-based alignments (Sierk and Kleywegt(2004)). Among the many structure-superposition methods (Novotny et al.(2004); Kolodny et al.(2005)), some differentiate themselves by the
ability to handle challenging cases like partial matches or structural flexibility. Aligned structures generally have evolutionarily related sequences which often form globular domains. Domain is a fundamental concept in structure classification (Ponting and Russell(2002)). Domain classifications like SCOP (Andreeva et al.(2007)) and CATH (Greene et al.(2007)) are popular. Similar domains often exhibit conservation of functionally important residues, e.g. in surface-accessible pockets and active sites. These can be predicted by unusual conservation of residues (Chelliah et al.(2004)) whereas larger conserved surface patches often indicate domain-domain interfaces. In silico docking (Sousa et al.(2006)) is used to design small molecules that optimally fit known or predicted binding sites. Domain-independent structure analysis can sometimes reveal structural propensities, e.g. those of $\alpha$ and $\beta$ secondary structures (Kihara(2005)), or a tendency to remain unstructured (Dosztányi et al.(2005)) as in domain linkers, or recurring patterns of sidechain orientations, structure templates (Tendulkar et al.(2004)) etc.

Effective data organisation is key to analyses in structural bioinformatics. Experimentally determined structural models are available in PDB and mmCIF\(^1\) file formats which define data types and the relationships between them. Efforts to produce molecule- and residue-level mappings between structure and sequence, sequence families, structural domains and source organisms (Velankar et al.(2005)) enable the integration of structure and other biological databases. Various ontologies (Reeves et al.(2008)) have been defined to capture structure-sequence features in a controlled vocabulary.

3 A PDBe perspective

The Worldwide Protein Data Bank (wwPDB; wwpdb.org) is the international consortium, consisting of RCSB PDB and BMRB (USA), PDBe (Europe) and PDBj (Japan). wwPDB is responsible for collecting, annotating, archiving and disseminating 3D biomacromolecular structure data. The EMBL-EBI Protein Data Bank in Europe (Cambridge, UK; pdbe.org) aims to become an integrated structure resource for all of bioscience, focussing on advanced services, ligands, integration, validation and experimental data (Velankar et al.(2010); Velankar and Kleywegt(2011); Velankar et al.(2011)). This effort reflects the realisation that 3D structure data is now used by a wide variety of non-expert scientists who need reliable information delivered in a biological or chemical context that is familiar to them (Velankar and Kleywegt(2011)).

Current advanced services at PDBe include: (a) PDBeMotif (pdbe.org/motif), a service for analysing detailed molecular interactions and correlate them with sequence or structure patterns (Golovin and Henrick(2008)); (b) PDBeFold (pdbe.org/fold), a powerful interactive structure-alignment tool (Krissinel and Henrick(2004)); (c) PDBePISA (pdbe.org/pisa), a quaternary-structure prediction service (Krissinel and Henrick(2007)); (d) PDBeXplore (pdbe.org/browse), a tool that allows browsing and analysis of the structural archive based on familiar chemical and biological classification systems (such EC, CATH and Pfam).

Small-molecule ligands and their interactions with biomacromolecules are important, but they are often poorly determined and annotated. Therefore, PDBe plans to develop relevant services for both structure producers and users, e.g. for analysis, validation (Bruno et al.(2004)) and visualisation. Integration is a keyword in bioinformatics and the joint UniProt/PDBe mapping resource SIFTS (pdbe.org/sifts) will be further enriched with cross-references to other biological data resources. Validation is crucial for identifying reliable structures and regions in PDB entries. The wwPDB partners have convened Validation Task Forces for X-ray crystallography, NMR spectroscopy and Electron cryo-Microscopy and their recommendations will be implemented as part of the new joint structure deposition and annotation system that is currently being developed. Finally, PDBe intends to provide services where the experimental data

---

\(^1\)see http://www.ebi.ac.uk/pdbe/docs/mmcif/mmcif.html
can be visualised and used to assess the reliability of structures or details of structures (such as a binding site).

4 Conclusion

Structural biology and bioinformatics encompass many different techniques and areas, and often these have to be combined to solve a problem. Structure determination is no longer an end in itself, but a key tool for gaining insight into a biological problem. This has created challenges in computation and informatics of macromolecular structure that are currently being addressed by the community working together.

References


A. Beberg and V. S. Pande. Folding@home: lessons from eight years of distributed computing. In IEEE International Symposium on Parallel & Distributed Processing, pages 1–8, 2009.


1 Introduction

Successful models for the evolution of protein sequences have been in existence since the 1960s, and are widely used in a large range of applications, ranging from inference of phylogeny to identification of binding site motifs. The importance of sequence arises largely from its determination of structure, and this structure may be conserved despite major changes in sequence. Indeed, for some families the sequence identity may fall as low as 10\% while still preserving the overall fold to a remarkable degree, such that structure comparison may enable the detection of distant homology that cannot be recognised from sequence alone. However, despite decades of research seeking to bridge the gap between sequence and structure, current methods for studying structural evolution are much less well developed, and the origins of the empirical relationships between sequence and structural divergence first characterised by Chothia and Lesk (1986) remain to a large extent a mystery.

Central to the comparative study of protein structures is the requirement of a measure of similarity in structure space. Most widely used distance metrics treat protein structures as static entities, despite the fact that studies carried out in biological conditions suggest that most proteins exhibit flexibility of some kind. Indeed, many protein functions are mediated through changes in conformation of the structure, and there is increasing evidence to suggest that these conformational changes can in fact be regarded as redistributions of the populations of thermally sampled states (Kern and Zuiderweg, 2003).

Following observations that the space sampled by a set of homologous structures is strikingly similar to the subspace spanned by the low frequency dynamics of the individual proteins (e.g. Leo-Macias et al., 2005), Echave introduced a linear response model—termed the linearly forced elastic network model, or LFENM—to investigate the expected pattern of structural divergence upon mutation. The model stipulates that mutational perturbations are dissipated along energetically favourable directions in a similar fashion to the effects of ligand binding events (Echave and Fernández, 2010). This is an example of the more general fluctuation-dissipation theorem relating relaxation from non-equilibrium states to thermally accessible fluctuations at equilibrium.

Implicit in the extension of this theorem to evolutionary perturbations is the assumption that the folding pathway is not significantly altered by mutation events, since a change of folding pathway could lead the structure from one native state to another without the two being interconvertible in the folded state. This is often assumed in the analysis of mutagenesis experiments such as double mutant cycles, and is reasonable provided that the structures are similar.

Under a model of dynamically mediated structural evolution such as the LFENM, assuming at this stage a star phylogeny between the organisms, we can model a set of homologous structures $\mathcal{X} = \{X^{(1)}, \ldots, X^{(K)}\}$ as if each member of the set were sampled from a common distribution, such that the likelihood of an alignment can be written as a product of the form

$$p(\mathcal{X}|\mathcal{A}, \theta) = \prod_k f(A^{(k)} X^{(k)}|\theta)$$

(1)
where \( f(\cdot) \) describes the common equilibrium distribution of the structures represented by the \( L^{(k)} \times 3 \) matrices \( X^{(k)} \), and \( A^{(k)} \in A \) is an \( L \times L^{(k)} \) binary matrix that maps each residue (row) in the \( k \)th configuration to the corresponding index in the \( L \)-dimensional joint model, with 
\[
L = \sum_k L^{(k)}.
\]
We also impose the restriction that no more than one residue from each structure can map to a given index (i.e. each row of \( A^{(k)} \) contains at most one non-zero entry). The set of maps \( A \) defines an alignment. Ultimately one might wish to use a physically realistic atomistic potential as the function \( f \), but given the approximations involved in the model, such detail would almost certainly be inappropriate, and we may choose instead to examine the utility of more coarse-grain energy functions.

2 Structural alignment

2.1 Rossmann method

First introduced by Rao and Rossmann (1973), and developed further during the 1970s, this approach remains at the core of many commonly used methods of structural alignment, consisting of iterative alternating least squares superposition and alignment based upon the coordinate RMSD (which is closely related to the Procrustes distance):

\[
cRMSD(k, l) = ||A^{(k)} X^{(k)} - A^{(l)} X^{(l)} \hat{R} + \hat{T}||_F / \sqrt{\text{tr}(A^{(k)^T A^{(l)}})}
\]

where \( \hat{T} \) is a translation to bring the centres of mass of the two proteins into superposition, and \( \hat{R} \) is the 3D rotation matrix that minimises the cRMSD. The quantity \( \text{tr}(A^{(k)^T A^{(l)}}) \) in the normalising factor corresponds to the number of aligned residues. Since the \( cRMSD^2 \) is linearly additive over the matched residues in the alignment, then conditional on \( \hat{T} \) and \( \hat{R} \), and given a gap penalty parameter, \( \lambda \), the optimal alignment can be found by dynamic programming using algorithms such as Needleman-Wunsch, minimising a score of the form 
\[
cRMSD^2 + \lambda(L - \text{tr}(A^{(k)^T A^{(l)}})).
\]

The basic Rossmann algorithm for pairwise alignment proceeds by iterating between superposition and alignment steps until convergence is achieved.

A variant of the Rossmann method can be seen to be equivalent to an EM scheme (Kent et al., 2010), and this method (or variations thereof) is implemented in a number of well-known structural alignment programs, such as STRUCTAL, LSQMAN, SSM, and MAMMOTH. The superposition step is also known as Procrustes registration, and can be generalised to multiple structures, as implemented in MUSTANG and MAMMOTH Mult.

2.2 Extensions to Bayesian framework

One of the problems with many of the aforementioned methods is that the scoring schemes are not based on probabilistic models, producing a point estimate with little information of the uncertainty surrounding this estimate. A common approach is to fit the scores to an extreme value distribution, such that empirical \( p \)-values can be calculated, giving a rough measure of significance, although there are a number of issues associated with such approaches. Another shortcoming of most existing methods is that the choice of parameters such as gap opening and gap extension penalties may have a large effect on the output, and is usually left to the user, who may have no good intuition as to how to specify these parameters.

Over the last twenty years, probabilistic evolutionary models have been successfully developed for protein sequences to address such problems. More recently, several methods (e.g. Green and Mardia, 2006; Rodriguez and Schmidler, submitted) have been developed for addressing the structural alignment problem within a statistical framework. These methods essentially use the Rossmann method, albeit within a Bayesian probabilistic model. Since the likelihood in such models involves maximisation or integration over rotations, it is invariant to
rotations of the data. In the case of Green and Mardia (2006), the integration is carried out through an MCMC procedure.

However, the above methods generally assume that the location of each residue is distributed with independent, isotropic variance. In the unweighted Procrustes schemes mentioned in the previous section, a high variance at a particular region of the protein will therefore be spread out over the whole structure by the least-squares rotation, giving a misleading picture of any covariance structure within the two conformations (Walker, 2000). Similarly, in the Bayesian procedure, the posterior induced on rotations is affected equally by the deviation at each residue.

Since the covariance structure is critically important to any model of dynamically constrained evolution, it is essential that we allow for interresidue correlations in any alignment scheme. Various methods have been proposed to account for heteroskedasticity and correlated dynamics when superposing protein structures, for example using weighted Procrustes, either with a theoretical estimate of the covariance matrix as the weights, or estimating the covariance and superposition jointly from the data (Theobald and Wuttke, 2008).

However, it can be shown (Kent and Mardia, 1997) that only if the covariance matrix is equal to the identity matrix (a highly unrealistic scenario for protein molecules) will superposition-based methods yield a consistent estimator for the mean structure; such methods will, therefore, in general lead to biased estimators of the covariance. Indeed, any method that relies on superposition will encounter significant problems in situations where it is impossible to simultaneously line up all the corresponding regions, for example in the case of an open and closed conformation of a protein such as calmodulin that undergoes a major hinge motion. This problem can be addressed by splitting the protein up into domains or structural units and superposing these separately, inferring breakpoints between domains either manually (e.g. Lesk and Chothia, 1980) or through automated methods (Mechelke and Habeck, 2010; Schmidler, submitted). However, such methods inevitably lose information about the interdomain correlation structure, and are highly sensitive to the choice of what constitutes a superposable unit.

### 2.3 Distance-based scoring functions

While distance-based methods have the advantage that there is no need for superposition, there is the complication that distance-based scoring functions are generally non-local, and therefore not linearly additive over the residues. This makes optimisation of such scores difficult, with all existing implementations using a heuristic at some point during the procedure. Although the class of distance-based methods includes some of the most successful structural alignment algorithms, including programmes such as SSAP (Taylor and Orengo, 1989), DALI (Holm and Sander, 1993), and CE (Shindyalov and Bourne, 1998), the heuristics involved make the significance of resulting alignments difficult to determine, and convergence properties are often far from obvious.

### 2.4 Reflection size-and-shape

As discussed by Goodall and Mardia (1993), and more recently by Dryden et al. (2008), the size-and-shape (up to reflection) of a configuration can be represented by the centred Gram matrix \( G(X) \), or a projection onto the column space thereof. Goodall and Mardia (1993) focus on the case in which \( X \) is distributed about a mean configuration \( M \), with diagonal covariance, allowing the rotations to be integrated out analytically, yielding a hypergeometric distribution for \( G \). However, this marginal distribution is difficult to work with directly, due to the presence of the hypergeometric function. Nevertheless, as argued by Prompers and Brüschweiler (2002), for any three-dimensional object free to rotate in space, the population mean of the coordinate matrix \( X \) (without superposition) will in fact be zero, such that we need only consider the central case, which is simply a Wishart distribution on \( G \). It should be noted that, in the case
of proteins, since amino acids are found in only one enantiomer, and helices are therefore all right-handed, reflection invariance is unlikely to lead to erroneous ascription of homology.

3 Results

We have recently developed a model of this type for multiple structures taken from a joint distribution of the form in equation (1), with $f$ taken to be a zero-mean Gaussian with unknown covariance. Integrating out this covariance, we obtain a marginal posterior for the alignment. If we consider the alignment as specifying a particular model, then we can regard the ratios between posteriors as Bayes factors, and the alignment problem can be thought of as model selection.

Simulating from this model using MCMC, we can generate samples of alignments for any set of structures. Using a recently developed method for maximum posterior decoding (MPD) in the space of alignment columns (Herman et al., submitted), it is then possible to generate a single representative alignment that minimises one of a family of loss functions similar to that proposed by Green and Mardia (2006).

As an example here we show results for a set of eighteen structures, comprising seven known homologous pairs or triplets and one outlier, taken from a variety of structural classes, determined by either NMR or crystallography. In the case of NMR we work with an ensemble of models, which provides additional information regarding the uncertainty in the structures. Aligning all against all and computing Bayes factors for the MPD alignment with respect to the unaligned model, we obtain a measure of homology between each of the structures (shown in the figure above). Aside from one false positive (1OYC) and one pair of false negatives (1BXD, 1B3Q), the method successfully clusters the data into the sets of known homologues.

References


Session III
Riemannian centres of mass and a new look at the classic central limit theorem

Wilfrid S. Kendall

In the survey article Kendall and Le (2010), Huiling Le and I discussed various aspects of statistical shape theory. In particular our interest was caught by recent work (Bhattacharya and Patrangenaru (2003); Bhattacharya and Patrangenaru (2005); Bhattacharya and Bhattacharya (2008)) on central limit theorems for Riemannian centres of mass, also called Fréchet means. The result was Kendall and Le (2011). In my talk I plan to use illustrations and examples to present this work; the following extended abstract is a rather drier succinct summary of the main points.

Recall that a Fréchet mean for a random variable $X$ is defined as a minimizer $E[X]$ of the “energy function” $x \mapsto E\left[\frac{1}{2}\text{dist}(x, X)^2\right]$. In case $X$ is a real-valued random variable (or even a vector-valued random variable) of finite variance, it is a common exercise set for first-year undergraduates to identify the minimizer as the conventional statistical expectation $E[X]$. More generally, if $X$ takes values in a curved Riemannian manifold or metric space (such as a shape space) then $E[X]$ is no longer uniquely defined, though it will be unique if the distribution of $X$ is confined to a region with geometry comparable to that of a small hemisphere (Kendall (1991)). The notions of Fréchet mean and their generalizations have found significant applications in pure probability and analysis, but also in applied statistics and particularly in the theory of shape. For example, as soon as one fixes on a notion of distance between curves, then it makes sense to talk about the Fréchet mean(s) of a random curve.

Interest in shape theory has focussed on empirical Fréchet means (minimizers $E_e[X_1, \ldots, X_n]$ of $\Phi_n(x) = \sum_{i=1}^n \frac{1}{2}\text{dist}(x, X_i)^2$ for an i.i.d. random sample $X_1, \ldots, X_n$), algorithms for computing such means (Le (1998); Le (2001); Le (2004); Groisser (2004)), and consideration of limit theory and asymptotic distribution theory (Bhattacharya and co-workers, as noted above). Huiling Le and I posed ourselves the question, what would the central limit theory look like if one relaxed the requirement of identical distribution in favour of some Lindeberg-like condition. The conclusion involves an unexpected and intriguing perspective on the classic central limit theorem.

To set the scene for this perspective, we first introduce a convenient notation: if $X_1, X_2, \ldots$ are independent manifold-valued random variables taking values in a Riemannian manifold $\mathbb{M}$ then we write

$$\phi_n(o) = \sum_{i=1}^n E\left[\frac{1}{2}\text{dist}(o, X_i)^2\right]$$

for the “aggregate energy” of $X_1, \ldots, X_n$ at a reference point $o \in \mathbb{M}$. In the vector-valued zero-mean case this would yield half the trace of the variance-covariance matrix of $X_1 + \ldots + X_n$, and this gives the right intuition for $\phi_n$.

The natural setting for a manifold central limit theorem is the following: $X_1, X_2, \ldots$ are independent manifold-valued random variables, each with the same Fréchet mean $E[X_i] = o$ for a fixed reference point $o$. If the distributions of the $X_i$ are non-identical then we will need a condition of Lindeberg type to control variation: for each $\varepsilon > 0$,

$$\frac{1}{\phi_n(o)} \sum_{i=1}^n E\left[\frac{1}{2}\text{dist}(o, X_i)^2 ; \frac{1}{2}\text{dist}(o, X_i)^2 > \varepsilon \phi_n(o)\right] \to 0.$$  

(Here the notation $E[X; A]$, for random variable $X$ and event $A$, is a convenient way to represent the expectation of $X$ multiplied by the indicator random variable of $A$; thus the expectations in...
(2) only contribute when the $X_i$ are relatively far away from the location $x$. This condition is appealingly similar to the condition for the most general possible weak law of large numbers for non-negative random variables (Chow and Teicher (2003), Chapter 10, Theorem 1, Corollary 2); with modest extra conditions it implies a weak law of large numbers for the empirical Fréchet mean $E_e[X_1, \ldots, X_n]$ as $n \to \infty$.

For a scalar central limit theorem it would suffice to require the scalar specialization of (2); indeed the classic Lindeberg-Feller theorem shows that this is essentially necessary as well as sufficient. However in the manifold case we require further conditions in order to control the relationship between curvature and $X_1, X_2, \ldots$. Rather than go into details about these (which are enumerated in Kendall and Le, 2011, Theorem 4), we rehearse the intuition underlying the proof of the manifold central limit theorem, which is that the result follows by using Newton’s root-finding method to reduce to the vector-valued case.

Neglecting details concerning such things as cut-loci, one sees this by noting that $E_e[X_1, \ldots, X_n]$ must attain the minimum of $\Phi_n(x) = \sum_{i=1}^n \frac{1}{2} \text{dist}(x, X_i)^2$, and hence should be a zero of the gradient vector field

$$\operatorname{grad}_x \frac{1}{n} \sum_{i=1}^n \left( -\frac{1}{2} \text{dist}(x, X_i)^2 \right) = -\frac{1}{n} \sum_{i=1}^n \text{dist}(x, X_i) \operatorname{grad}_x \text{dist}(x, X_i).$$

(The tangent vector $Y_i(x) = -\text{dist}(x, X_i) \operatorname{grad}_x \text{dist}(x, X_i)$ points along the geodesic from $x$ to $X_i$, so it is intuitively reasonable that the above sum of gradients should point towards $E_e[X_1, \ldots, X_n]$.) Following the theme of Newton’s method, we then consider the first-order Taylor expansion about $x = o$ for $\operatorname{grad}_x \text{dist} \sum_{i=1}^n \left( -\frac{1}{2} \text{dist}(x, X_i)^2 \right)$; we find that

$$\operatorname{grad}_x \sum_{i=1}^n \left( -\frac{1}{2} \text{dist}(x, X_i)^2 \right) \approx -\sum_{i=1}^n \text{dist}(o, X_i) \operatorname{grad}_o \text{dist}(o, X_i) - H_n \gamma_x'(0) + \text{error}. \quad (3)$$

Here $\gamma_x$ is the geodesic which starts at $o$ and reaches $x$ at time 1; $H_n$ is the Hessian, or matrix of second-order covariant derivatives, of $\Phi_n$ at $o$. In particular, if $\gamma$ is a geodesic with initial tangent vector $\gamma'(0)$ at $\gamma(0) = o$ then $\langle H_n \gamma'(0), \gamma'(0) \rangle = \left[ \frac{d^2}{dt^2} \Phi_n(\gamma(t)) \right]_{t=0}$. We write “$\approx$” rather than $=$ in (3) because the tangent vectors on the left-hand side are located at $x$, while on the right hand side they are located at $o$; therefore full sense of (3) requires the notion of stochastic parallel transport, suppressed here for clarity of exposition.

Now the left hand side of (3) vanishes at $x = E_e[X_1, \ldots, X_n]$. Setting $\gamma_n = \gamma_{E_e[X_1, \ldots, X_n]}$, and suppressing the error term, we find from (3) that

$$\gamma_n'(0) \approx -H_n^{-1} \sum_{i=1}^n \text{dist}(o, X_i) \operatorname{grad}_o \text{dist}(o, X_i) = -H_n^{-1} \sum_{i=1}^n Y_i(o).$$

But $\gamma_n'(0)$ can be viewed as locating $E_e[X_1, \ldots, X_n]$ in the geometrically natural “exponential coordinates”. So if we can control the error term and the matrix inverse $H_n^{-1}$ (using the further conditions on the interaction of the geometry of $M$ with the random sample) then we can achieve a central limit theorem for $E_e[X_1, \ldots, X_n]$ in a natural coordinate system so long as we can derive a central limit theorem for the essentially vector-valued sum $\sum_{i=1}^n Y_i(o)$. (This follows by a Slutsky theorem argument, since under re-scaling and appropriate conditions the matrix inverse then converges weakly to a constant.)
But now there arises an intriguing problem. It is a direct geometric fact arising from its
definition that $\|Y_i(o)\| = \text{dist}(o, X_i)$. Therefore the independent zero-mean vectors $Y_i = Y_i(o)$
hinherit a Lindeberg-like condition from (2), namely

$$\frac{1}{\phi_n(o)} \sum_{i=1}^{n} \mathbb{E} \left[ \frac{1}{2} \|Y_i\|^2 ; \frac{1}{2} \|Y_i\|^2 > \varepsilon \phi_n(o) \right] \to 0 ,$$

(4)

where we can write $\phi_n(o) = \sum_{i=1}^{n} \mathbb{E} \left[ \frac{1}{2} \|Y_i\|^2 \right]$. But (4) controls only the lengths of the $Y_i$,
not the individual coordinates. There are easy examples satisfying (4) and yet not satisfying a
central limit theorem. For example, consider the oscillating distributional behaviour of normalized
partial sums of a sequence of independent two-dimensional random variables, switching
between $(N(0,1), 0)$ and $(0, N(0,1))$ in successive runs of lengths $1, 2, 4, \ldots, 2^n, \ldots$.

Thus (4) cannot deliver a central limit theorem on its own. However the example is one in
which the partial sum distribution remains bivariate normal, even though the variance-covariance
matrix does not converge to a limit. And it turns out that this is a good hint of the power of (4); it
cannot guarantee convergence in distribution but does guarantee convergence to zero of an
appropriate distance between the normalized partial sum and a matching multivariate normal
distribution. This distance is the so-called “truncated Wasserstein distance”, a convenient way
of converting statements about weak convergence into statements about distances between dis-
tributions. The actual proof follows the classical central limit argument closely; indeed the
analogue of the Feller converse to the Lindeberg theorem holds; under a natural condition (4)
is also necessary for the partial sum distribution to converge to matched normality. Hence
Riemannian centres of mass have led to a new insight about classic CLT!

References

asymptotic distribution and curvature. Proceedings of the American Mathematical Society 136(08),
2959–2967.

Bhattacharya, R. N. and V. Patrangenaru (2003). Large sample theory of intrinsic and extrinsic

Bhattacharya, R. N. and V. Patrangenaru (2005). Large sample theory of intrinsic and extrinsic

Chow, Y. S. and H. Teicher (2003). Probability theory: independence, interchangeability, mar-


Molchanov (Eds.), New Perspectives in Stochastic Geometry, Chapter 10, pp. 348–373.

and non-identically distributed manifold-valued random variables. Brazilian Journal of


Detecting the shape of a configuration from random projections

H. Le
University of Nottingham, UK

Towards the end of the 1980’s, one of DG Kendall’s main research interests was to visualise spaces of 3-D shapes, and so to understand the 3-D shapes themselves, using the singular-values decomposition of matrices. In my talk, I shall continue along these lines, but looking from a slightly different angle.

Statistical analysis of 3-D shapes is naturally more complicated than that for 2-D shapes. On the other hand, it is often the case that data we have on a 3-D configuration is in terms of a range of 2-D projections. Thus it may be useful in some cases to use 2-D projections of 3-D configurations to determine or estimate the original 3-D shapes. Although it is possible to determine the shape of a 3-dimensional configuration by a finite number of projections as long as the directions of those projections are known, the deduction is no longer straightforward when the directions of the projections are random or unknown. The issue of determining the shape of a configuration by random projections is equivalent to the question of the extent to which one can determine the shape of the configuration by its projections onto a pre-fixed plane after the configuration itself has been randomly rotated. One motivation for the latter point of view comes, for example, from biophysics: in studying the structure of biological molecules, biophysicists use electron microscopes to view planer images of single particles. However, the size of the particle, together with its movement, makes it impossible to control the direction of the projection.

If we denote the pre-size-and-shape of the configuration, i.e. the centred configuration, in \( \mathbb{R}^m \) with \( k > m \) labelled vertices by an \( m \times (k - 1) \) matrix \( X \) then, up to reflection, the size-and-shape of the configuration is determined by \( X^\top X \). It is shown in (Panaretos, V.M., (2009)) that

\[
\int_{SO(m)} (P_*RX)^\top (P_*RX) \Psi(dR) = \frac{m - 1}{m} X^\top X,
\]

where \( P_* \) denotes the projection onto a pre-fixed plane and \( \Psi \) the normalised Haar measure, i.e. uniform measure, on \( SO(m) \). Thus, the law of large numbers enables us to estimate the size-and-shape, and so the shape, of \( X \) up to reflection.

To answer any further statistical inference questions, we may need to know the more detailed statistical behaviour of the shape, or size-and-shape, of the projection of the configuration after being randomly rotated. For this, we can investigate the corresponding induced shape distribution.

There are two features involved in determining such an induced shape distribution: its supporting domain, consisting of the shapes of all possible projections of the given configuration, and its Radon-Nikodym derivative on this domain. The equivalence between the projection onto a pre-fixed hyperplane after random rotation and the random projection (without rotation) implies that, for a generic configuration, the subset consisting of all shapes of the projections onto a pre-fixed hyperplane of a randomly rotated configuration in \( \mathbb{R}^m \) is diffeomorphic, say via a diffeomorphism \( \phi \), with \( S^{m-1} \), the set of all possible directions in which to project. To proceed further, we recall that the shape of a configuration with pre-size-and-shape matrix \( X \) is determined by the two factors, \( \Lambda/\|\Lambda\| \) and \( S \), in the singular-values decomposition \( X = R\Lambda S \),
where $R \in SO(m)$, $\Lambda = \text{diag}\{\lambda_1, \cdots, \lambda_m\}$ and $S^T = (s_1, s_2, \cdots, s_m)$ is a $(k-1) \times m$ matrix with the column vectors $s_i$ being the unit eigenvectors of $X^TX$ corresponding to eigenvalues $\lambda_i^2$. The eigenvalues $\lambda_i^2$ are usually taken in non-increasing order and the $\lambda_i$ all non-negative except possibly for $\lambda_m$ when $k = m + 1$.

We start with the case $m = 2$, the corresponding 2-D problem. That is, we assume that the configuration to be estimated lies in $\mathbb{R}^2$ and that the data we have are its 1-D random projections. Hence, the shapes of random projections of the 2-D configuration lie in the shape space $\Sigma^1_2$, which is known to be the sphere $S^{k-2}$. In this case, the set of the size-and-shapes of all possible random projections of the configuration with pre-size-and-shape $X$ can be described by the points of $\mathbb{R}^{k-1}$ on the ellipse with major and minor axes $\lambda_1s_1$ and $\lambda_2s_2$ respectively. From this, it follows that the support of the induced distribution is just the embedded circle $S^4$ of $S^{k-2}$ spanned by $s_1$ and $s_2$, the two columns of $S^T$. Note that this domain alone does not uniquely determine the matrix $S$ appearing in the singular-values decomposition of $X$: $S$ and $TS$, $T \in SO(2)$, determine the same great circle in $S^{k-2}$. To complete the determination of $S$, and hence $X$, we need more detailed information on the induced distribution itself.

The Radon-Nikodym derivative, with respect to the volume element on $S^1$, of the induced distribution at $\phi(v)$, the shape of the projection onto the line orthogonal to $v$ of the configuration with pre-size-and-shape $X$, is obtained in (Panaretos, V.M., (2006)) and is given by

$$\frac{1}{2\pi} \frac{||P(v)||^2}{\lambda_1 \lambda_2},$$

where $P(v) = (I - vv^T)X$, the projection of $X$ onto the line orthogonal to $v$, and so $||P(v)||$ is the size of the projected configuration. The density has the maximum value $\lambda_1/|\lambda_2|$ and the minimum value $|\lambda_2|/\lambda_1$, achieved when $v = \pm s_1$ and $v = \pm s_2$ respectively. These two extreme values of the Radon-Nikodym derivative determine the two eigenvalues of $XX^T/||XX^T||$. Hence, it is possible to recover all the information on the shape, up to reflection, of the initial given configuration from the support and the extreme values, together with the locations where they are achieved, of this density function. Note also that, when the shape of $X$ is most regular, i.e. when $\lambda_1 = |\lambda_2|$, the induced distribution is uniform on the circle $\phi(S^1)$; while, as the shape of $X$ tends to a degenerate collinear shape, the induced distribution tends to become concentrated at the two points $\pm s_1$. Thus, the more regular the shape, the more all its possible projections become equally likely to be observed. However, the density function is bimodal so that a given configuration and its reflection will determine the same induced distribution for the shape of their random projections.

Returning to the case $m = 3$ in which our main interest lies, it can be shown that the set of the size-and-shapes of all possible projections consists of all $2 \times (k-1)$ matrices whose rows $t_1^T$ and $t_2^T$ are such that the vectors $t_1$ and $t_2$ of $\mathbb{R}^{k-1}$ are orthogonal to each other and lie on the ellipsoid having axes $\lambda_1s_1$, $\lambda_2s_2$, $\lambda_3s_3$ with the restriction that $t_1$ lies on the ellipse with axes $\lambda_1s_1$ and $\lambda_3s_3$. This, after normalisation, determines the support $\phi(S^2)$ of the induced shape distribution in $\Sigma^k_2$.

It is shown in (Le, H. and Barden, D. (2010)) that, with respect to the volume element on $\phi(S^2)$, the Radon-Nikodym derivative at $\phi(v)$, the shape of the projection onto the plane with normal $v$ of the configuration with pre-size-and-shape $X$, is given by

$$\frac{1}{4\pi} \frac{||P(v)||^4}{\lambda_1^2 \lambda_2^2 + \lambda_1^2 \lambda_3^2 + \lambda_2^2 \lambda_3^2 - \lambda_1(v)^2 \lambda_2(v)^2},$$

where $\lambda_1(v)^2$ and $\lambda_2(v)^2$ are the two non-zero eigenvalues of $P(v)P(v)^T$ and $P(v) = (I - vv^T)X$, the projection of $X$ onto the plane with normal $v$ uniformly distributed on $S^2$. The
maximum and minimum values of this Radon-Nikodym derivative are respectively $(\lambda_1^2 + \lambda_2^2)/\lambda_3^2$ and $(\lambda_2^2 + \lambda_3^2)/\lambda_1^2$. These two extreme values are again sufficient to determine the three eigenvalues of $XX^\top/\|X\|^2$. These maximum and minimum values are reached when $v$ is normal to the planes spanned by $\{s_1, s_2\}$ and $\{s_2, s_3\}$ respectively. The density function also has saddle points occurring when $v$ is normal to the plane spanned by $\{s_1, s_3\}$, where the density function takes the value $(\lambda_1^2 + \lambda_3^2)/\lambda_2^2$. When $\lambda_2 = |\lambda_3|$, respectively $\lambda_1 = \lambda_2$, these saddle points coincide with the maximum points, respectively the minimum points. These locations clearly suffice to determine, up to sign, the rows of the matrix $S$ in the singular-values decomposition of $X$ so that for a generic $X$, the density of the induced distribution contains all the information required to determine the shape of $X$ up to reflection. Again, when the shape of $X$ is most regular, i.e. when $XX^\top$ has three identical eigenvalues, the induced distribution is uniform on $\phi(S^2)$.

Similar results hold for the size-and-shape of the projection of a randomly rotated configuration, where the support is implied by the above discussion and the corresponding Radon-Nikodym derivative of the induced distribution is given in (Le, H. and Barden, D. (2010)).

**References**


Speculations on the next statistical toolkit for complex organized systems

Fred L. Bookstein
University of Vienna and University of Washington

Summary. The flat cases-by-variables “data matrix $X$” is the classic notational underlay for the conventional multivariate analyses arising in linear models and analogous contexts where variables have no intrinsic semantics of their own. This flat matrix notation is ill-suited to 21st-century statistics, which already centers and will continue to center on applications to complex organized systems such as geotectonics, morphogenesis, physiology, and similar contexts in which extensive prior knowledge is encoded a-priori in quantitative characterizations of the variables jointly. Now is a good time, and LASR is a good venue, for creative brainstorming along these lines toward the appropriate replacement lexicons for observed patterns. This speculative essay, the twentieth or so in my series of LASR provocations, offers an idiosyncratic personal selection of possibilities.

Introduction

As Ludwig Wittgenstein famously said, “Die Grenzen meiner Sprache bedeuten die Grenzen meiner Welt” — “The limits of my language are the limits of my world.” This applies as trenchantly to statisticians as to philosophers: the limits of statistical data analysis are set in great part by limitations built into the language by which we report the outcomes of our pattern searches there. Put another way, the rhetoric, meaning persuasive style, in which the scientist reports the results of his collaborative statistical analyses, is dependent on the rhetoric, meaning grammatical craft, by means of which the collaborating statistician has posed the questions (the pattern searches) in the first place. That rhetoric has been fatally hamstrung by the early 20th-century formalism of the “data matrix $X$,” cases by variables, in that nearly all of the a-priori structuring is assigned to the cases facet of the design (sampling frames, covariance structures) and hardly anything to the meaning of the variables. There are many and diverse prototypes for an effort, here in the early 21st century, to balance the investment more equitably. Hence this speculative essay.

Statistics as the scientific semantics of empirical patterns

Mathematics, the cliché goes, is the language of science. Actually, it is not that language, but only the syntax of that language, the rules that allow a symbolic construal of the empirical patterns uncovered by the inherited perceptual machinery of the mind. For the semantics (the treatment of meaning), we need to turn to information theory and statistics, specifically, statistical data analysis. But that would not be today’s academic statistics, which has exhausted the implications of both its 19th-century origins in natural philosophy and its 20th-century impetuses in experimental design and the algebra of covariance matrices. We need instead a semantics — a linguistics, really — for situations like the following.

Systematic, stable phenomena with complicated ranges. Among the triumphs of 20th-century science are the following two textbook-standard plots, one conceived in the first half of the century, the other in the second half.
The plot at left is the \textit{Hertzsprung–Russell} diagram, the joint distribution of stellar spectral color by absolute luminosity. It led directly to S. Chandrasekhar’s model of stellar evolution, which lies at the root of most of today’s cosmology. The plot beneath it runs at about 20 orders of magnitude more precision: it is, of course, the \textit{Ramachandran plot} of the parameters $\phi$ and $\psi$ describing the local geometry of a protein backbone in space. The issue I am raising regarding both of these plots, along with several famous others (e.g., the Periodic Table), is that we lack a proper language for the way such evidence about organized constraints on reality constrains our quantitative explanations in turn. Issues of dynamics are omitted from the upper figure, for instance, and issues of autocorrelation along the backbone from the lower.

Other aspects of controlled vocabulary are equally familiar, at least to the applied statistician of the natural sciences. The first attempt at a formal semantics of measurement in physics, for instance, was James Clerk Maxwell’s, the same man who effectively argued the reality of the Gaussian distribution for components of the velocity of gas molecules. In a paper remarkably in advance of its time, Maxwell (1871) sketched the types of quantities that would underlie all of classical physics and its extensions into the nearby sciences right up through the present. We have only today begun to work out how to adapt the Maxwell notation to the now-overwhelming flow of data from just one context, the physiological (Cook et al., (2011)). An analogous development (Munck, (2005)) characterizes the semantics for reports of patterns in the chemometric \textit{spectra} returned by the laboratory machines that reify the variation of living materials at the chemical level.

\textit{The role of the origin of data.} One of the most unfortunate anachronisms left over from the heyday of classical multivariate analysis is the notation $N(\mu, \Sigma)$, which subtly encourages us to imagine an average and the variation of real systems around that average as if there were no constraints on either by virtue of the other. Ever since the first promulgation of shape coordinates (Bookstein, 1986, 1991) we have understood that in morphometrics, at least, it is essential to understand the entanglement between $\mu$ and $\Sigma$. Both in classic landmark-driven morphometrics and in its modern extension to curves, surfaces, and tensor image contents, morphometric reports of variations in a selected tangent space to one or another Riemannian shape manifold (from metrics like Procrustes distance, strain energy, or image energy) go forward in the context of the space of possible shape reports \textit{at the touching point of the tangent space selected}, usually a sample average. In one simple example, a pattern of changes of Procrustes shape coordinates, or of proportions among pairwise distances between landmarks, that is a uniform shear for one starting configuration is a deformation of arbitrarily high complexity when referred to a range of different starting forms instead. As another example, the relation between Procrustes distance and bending energy crucially depends on landmark spacing (Bookstein, (1991); Bookstein and Green, (1993)).
Those are prototypes for what we need more generally: a syntax for pattern claims on diagrams like those above (or the less colorful equivalent in Procrustes principal coordinates) that combines assertions about choice of variables and about the patterns they reveal in a flexible joint language, originally visual, springing from the vicissitudes of the underlying machines and their output. The origin of morphometric localizations in the technology of imaging machines is so familiar to us (see, e.g., Weber and Bookstein, (2011)) that we have overlooked how crucial is the analogous formal task in the wider world of pattern analysis for complex organized systems more generally.

The language of measurements: An assortment of responses

In the remaining space I sketch some tricks and approaches that this statistical data analyst often exploits to circumvent the limitations of the tools in which he was trained. Any of these are among the possible replacement candidates for the core pattern engines to serve over the next few decades. I will mention three: geometries of measurement spaces, symmetry-breaking as a particularly fruitful style of empirical description, and (a major current concern of mine) the construction of a bridge between statistical methodology and computational continuum mechanics, namely, the representation of uncertainty in the inputs to a finite-element analysis, and the effect of this uncertainty upon the outputs and the associated risks. The quality that distinguishes all these from their classical analogues is an emphasis on the issue of choice of measurements, not the mere construction of linear composites of the haphazard quantities that already happen to be available.

Geometries of measurement spaces. The earliest development making it clear that geometric morphometrics could not be contained within classical multivariate theory may have been my observation of the 1980’s that any average shape of a triangle of landmarks establishes an involution between sets of shape measurements at 90° in the corresponding dual space, and also an involution between distance-ratios and angles as descriptors of these systems (see Bookstein, 1991, Chap. 6). The ratio of generalized interlandmark distances affording the greatest change of proportions corresponds to distances that start and end at 90°, while the displacement of the shape change vector by a 90° rotation interchanges the role of distance and angle in the preceding description: now the distance-ratio is invariant while the angle changes fastest. In this wise morphometrics inverts the usual temporal relationship between data analysis and variable construction: analysis first, with delineation of variables much later, just prior to publication. An earlier version of these structuring ideas was implicit in the factor-analysis approach of Louis Guttman (cf. Guttman, (1966)), and they have made an appearance in multivariate calibration as well (Bookstein et al., (1996)).

More recently our attention has been drawn to the role of the natural metric \( \Sigma \log^2 \lambda \) for covariance matrices (Mitteroecker and Bookstein, (2009)), where the \( \lambda \)'s are the relative eigenvalues, and the curious circumstance that patterns characterizing these covariance distances are affine-invariant even though patterns of the underlying probability models (e.g., sphericity) are not. On the other hand, we still have no good candidate for a pattern engine that can handle the simplest quantitative descriptors of cell cycles (cf. Tyson et al., (2003)) — not their graphs and surely not their parametric dynamics.

Symmetries of systems, symmetries of measurement. I limit myself here to two interesting examples. In one, illustrated below, we see a geodesic (of image-based strain energy) between two shapes, a Helvetica A and a Helvetica B (Wirth et al., (2011)). The example is unexpected in that while both endpoints of the geodesic are characterized by bilateral symmetry, the axes are different: for the A, vertical; for the B, horizontal. It appears that the geodesic lies close to the submanifold of bilateral symmetry for the A at one end, and close
to that for the B at the other end. Only for a short length in the middle does the asymmetry stray far from the symmetry characterizing the nearer endpoint. This suggests an interaction between the Riemannian distance metric here (a generalized strain) and the subspaces BL introduced by Mardia et al., 2000, to formalize the Procrustes morphometrics of asymmetry in the same situation. Strain energy may be anisotropic in Procrustes distance within vs. normal to this subspace of symmetric forms. Such an anisotropy would bear profound implications for computer image analysis as well as human perceptual psychology.

A similar search for geometries of symmetry transformations would reward many other flavors of empirical data analysis. For instance, studies of vertebrate embryogenesis are typically driven by images from the surface of a sphere. One natural normalization group is the rotation group, but there are others, including the Möbius group (a representation of the linear fractional group of the complex plane into which the surface of the sphere can be mapped by stereographic projection). Like the rotation group, the Möbius can be infinitesimalized; unlike the rotation group, it permits the setting of landmarks at the poles (the one or two fixed points) of the transformation. The success of the Procrustes metric for normalization of “landmark shape” may be only one example of a wider variation of normalization groups corresponding to different empirical image analyses.

Information compression and representation of variance and uncertainty in finite element analyses. This last theme, a current interest of mine (Weber, Bookstein, and Strait, (2011)), is badly overdue for mathematical-statistical treatment in depth. Nowhere in today’s communities of computational mechanics and biomechanics do I find a formalism for the effect of variances or uncertainties of form upon uncertainties of the computed deformations of form when both are reported in terms of the locations of nodes in a finite-element decomposition of some solid material. Recent advances in computational mechanics are promising. The method of isogeometric analysis, for instance (Hughes et al., (2005)), seems like a fine analogue of principal warps (eigenfunctions of the bending energy of the thin-plate spline) for the representation of large-scale variation in shape subspaces requiring relatively few parameters. The reports can be by displacement fields or by strain fields (tensors), along with the associated physical energy. One technical issue is of tracing the implications of measured uncertainties of location for the solutions of the partial differential equations for which those locations serve as boundary conditions; another is the handling of coordinates within the tangent spaces of curving boundaries (which differs radically between current morphometric practice and current finite-element methodology). A typical practical application might be the assessment of material properties and their heterogeneity by strain-based measurements. Another, the prediction of the risk of fracture of a particular geriatric proximal femur, along with the distribution of probable locations of such a fracture, is already of considerable medical interest.

Concluding remarks

If statistics is to regain its status as one important bridge between mathematics and the natural sciences, we have some work to do upon foundations. We were launched well in the 19th century, first with the Maxwell-Boltzmann physics of normal distributions, then with the origins of correlation and regression in biometry. But our 20th-century multivariate statistics has headed off in directions that, in the context of the history of ideas treated more broadly, have generally proven intellectual dead ends. The curricula through which I was led in graduate school, emphasizing probability spaces, significance testing, and the algebra of covariance matrices, have little in common with the better formal structures encoding today’s most pressing
questions about empirical patterns and their dynamics, causes, or effects. These questions rely on preparation of the measurement spaces even more carefully than the handling of the data once measured.

The new foundation for applied statistics will emphasize the origin of measurements in the performance of machines, including the nature of spaces of measurements and their constraints, the pattern analyses by which we distinguish signal from noise and meaning from meaninglessness in these contexts of organized complexity, and the rhetoric by which all this is converted from arithmetic into qualitative summaries for our typical scientific and general audiences, for purposes that range from explanation through medical treatment planning and onward to risk management and public policy. We should reinvent statistical data analysis following on hints from the best 21st-century pattern analyses from the natural sciences, including the information sciences, in the same way that earlier versions of the core followed on hints from the less complex pattern analyses of the 19th and 20th centuries. An attention to the interplay between measurements and statistics — between numbers and reasons — is seriously overdue.

**Literature Cited**


Colouring and breaking sticks: random distributions and heterogeneous clustering

Peter J. Green
School of Mathematics, University of Bristol

1 Introduction
The purpose of this work is four-fold: to remind some Bayesian nonparametricians gently that closer study of some probabilistic literature might be rewarded, to encourage probabilists to think that there are statistical modelling problems worth of their attention, to point out to all another important connection between the work of John Kingman and modern statistical methodology (the role of the coalescent in population genetics approaches to statistical genomics being the most important example), and finally to introduce a modest generalisation of the Dirichlet process.

The most satisfying basis for statistical clustering of items of data is a probabilistic model, which usually takes the form of a mixture model, broadly interpreted. In most cases, the statistical characteristics of each cluster or mixture component are the same, so that cluster identities are a priori exchangeable. We will introduce a class of simple mixture models in which clusters are of different categories, or colours as we shall call them, with statistical characteristics that are constant within colours, but different between colours. Thus cluster identities are exchangeable only within colours.

2 A coloured Dirichlet process
In many domains of application, practical considerations suggest that the clusters in the data do not have equal standing; the most common such situation is where there is believed to be a ‘background’ cluster, and one or several ‘foreground’ clusters, but more generally, we can imagine there being several classes of cluster, and our prior beliefs are represented by the idea that cluster labels are exchangeable within these classes, but not overall. It would be common, also, to have different beliefs about cluster-specific parameters within each of these classes.

We present a variant on standard mixture/cluster models of the kinds we have already discussed, aimed at modelling this situation of partial exchangeability of cluster labels. We stress that it will remain true that, a priori, item labels are exchangeable, and that we have no prior information that particular items are drawn to particular classes of cluster; the analysis is to be based purely on the data \( \{Y_i\} \).

We will describe the class of a cluster henceforth as its ‘colour’. To define a variant on the DP in which not all clusters are exchangeable:

1. for each ‘colour’ \( k = 1, 2, \ldots \), draw \( G_k \) from a Dirichlet process \( \text{DP}(\theta_k, G_{0k}) \), independently for each \( k \);
2. draw weights \( (w_k) \) from the Dirichlet distribution \( \text{Dir}(\gamma_1, \gamma_2, \ldots) \), independently of the \( G_k \);
3. define \( G \) on \( \{k\} \times \Omega \) by \( G(k, B) = w_k G_k(B) \);
4. draw colour–parameter pairs \( (k_i, \phi_i) \) i.i.d. from \( G \), \( i = 1, 2, \ldots, n \).

This process, denoted \( \text{CDP}(\{\gamma_k, \theta_k, G_{0k}\}) \), is a Dirichlet mixture of Dirichlet processes (with different base measures), \( \sum_k w_k \text{DP}(\theta_k, G_{0k}) \), with the added feature that the the colour
of each cluster is identified (and indirectly observed), while labelling of clusters within colours is arbitrary.

It can be defined by a ‘stick-breaking-and-colouring’ construction:

1. colour segments of the stick using the Dirichlet(\{\gamma_k\})-distributed weights;
2. break each coloured segment using an infinite sequence of independent Beta(1, \theta_k) variables \(V_{jk}\);
3. draw \(\phi_{jk}^* \sim G_{0k}\), i.i.d., \(j = 1, 2, \ldots; k = 1, 2, \ldots\);
4. define \(G_k\) to be the discrete distribution putting probability \((1 - V_{1k})(1 - V_{2k}) \cdots (1 - V_{j-1,k})V_{jk}\) on \(\phi_{jk}^*\).

Note that in contrast to other elaborations to more structured data of the Dirichlet process model, in which the focus is on nonparametric analysis and more sharing of information would be desirable, here, where the focus is on clustering, we are content to leave the atoms and weights within each colour completely uncoupled \(a \text{ priori}\).

References


Bayesian molecular alignment using random fields

Irina Czogiel\textsuperscript{1}, Ian L. Dryden\textsuperscript{*2} and Christopher J. Brignell\textsuperscript{3}

\textsuperscript{1} Max Planck Institute for Molecular Genetics, Germany
\textsuperscript{2} Department of Statistics, University of South Carolina, USA
\textsuperscript{3} School of Mathematical Sciences, University of Nottingham, UK.

1 Introduction

David Kendall’s pioneering work in the 1970s and 1980s led to his path-breaking paper on shape spaces for landmark data (Kendall, (1984)). This classic paper and his subsequent work (e.g. Kendall, 1989; Kendall et al., 1999) continues to be an inspiration for a generation of inter-disciplinary researchers\textsuperscript{1} in the field of shape analysis.

The task of comparing unlabeled marked point sets has been of recent interest, e.g. Rangarajan et al. (1997), Taylor et al. (2003), Green and Mardia (2006), Dryden et al. (2007), Schmidler (2007), Ruffieux and Green (2009) and Mardia et al. (2010). Our method does not aim to model point correspondences, but rather the objects are compared by assuming a common underlying reference field which gives rise to the spatial distribution of the marks.

One example where the alignment of unlabeled marked point sets is of practical importance comes from the fields of structural bioinformatics and chemoinformatics, where it is of great interest to align molecules. However, the task is often very difficult. Our motivating application is a dataset analyzed by Dryden et al. (2007) comprising 31 steroid molecules which bind to the corticosteroid binding globulin (CBG) receptor (e.g. see Figure 1). For each molecule, the Cartesian coordinates of the atom positions, as well as the associated van der Waals radii, and the partial atomic charge values at the atom positions are provided. The dataset of atom coordinates and partial charges was constructed by Jonathan Hirst and James Melville (School of Chemistry, University of Nottingham). These 31 steroids are the same molecules described by Wagener et al. (1995), although the constructed co-ordinates here are not identical.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{aldosterone.png}
\caption{The first steroid in the dataset: aldosterone}
\end{figure}

The steroids fall into three activity classes with respect to their binding activity to the CBG receptor (Good et al., (1993)), and the main objective in this application is to compare the molecules in order to obtain the common features in each of the three groups and to examine whether these features are associated with the type of binding activity.

\textsuperscript{1}ILD: I feel honoured and privileged to have had David Kendall as my external examiner for my PhD, and I am forever grateful for his encouragement and kindness at all times.
We consider a simple model under which spatial prediction of a reference field is carried out using the observed marks in each configuration.

2 Random fields and kriging

The starting point for our model is an underlying reference random field \( \{ Z(x) : x \in \mathbb{R}^m \} \), which is assumed to be second-order stationary with a constant mean \( E(Z(x)) = \mu \) and a positive definite covariance function \( \sigma(h) = \text{Cov}(Z(x), Z(x + h)) = \sigma(-h) \).

The basic inference problem we consider is formulated as follows: if we are given the two marked point sets \( A \) and \( B \) with \( z^A \) recorded at locations \( \{x^A_1, \ldots, x^A_{k_A}\} \) and \( z^B \) recorded at locations \( \{x^B_1, \ldots, x^B_{k_B}\} \) can we measure how similar they are, taking into account an unknown \( 1 \)-\( 1 \) and onto transformation \( \Phi \) from \( B \) to \( A \)? The method involves aligning the point sets by estimating the transformation parameters in \( \Phi \) (e.g. a rotation matrix and translation vector in a rigid body transformation).

In order to predict the underlying reference random field from each point set we consider simple kriging (e.g. Cressie, 1993, pp. 110) which assumes the mean field \( \mu = 0 \). For the steroid molecules with partial charge or van der Waals radius marks it makes sense to fix \( \mu = 0 \), so that a long way from the molecular skeleton the predicted field is zero. Given the observed values in \( z \), the corresponding system of equations for minimizing the prediction mean square error has the solution \( u = \Sigma^{-1} \sigma \), and the predicted value for \( Z(x_0) \) is given by \( \hat{Z}(x_0) = \sigma(x_0)^T \Sigma^{-1} z = u^T z \), where \( \sigma(x_0) = (\sigma(x_1 - x_0), \ldots, \sigma(x_k - x_0))^T \) and \( (\Sigma)_{ij} = \sigma(x_i - x_j), 1 \leq i, j \leq k \).

For a general location \( x \) this yields the predicted field

\[
\hat{Z}(x) = z^T \Sigma^{-1} \sigma(x) = \sum_{i=1}^{k} w_i \sigma(x_i - x) ,
\]

where the weight vector \( w = (w_1, \ldots, w_k)^T = \Sigma^{-1} z \) is optimal in terms of minimizing the PMSE if the underlying assumptions are met. Using (1) and based on the observed data vectors \( z^A \) and \( z^B \) we can obtain a different prediction of the underlying reference random field from each of the two marked point sets \( A \) and \( B \), and the resulting predicted fields \( \hat{Z}_A(x) \) and \( \hat{Z}_B(x) \) then need to be compared.

3 Function similarity and the Kernel Carbo Index

In order to measure the similarity of the predicted fields \( \hat{Z}_A(x) \) and \( \hat{Z}_B(x) \) we require a metric space where the notion of similarity can be defined by means of the corresponding inner product. A commonly used metric space for functions is the space of Lebesgue square-integrable functions \( L_2 \), where the inner product has the form

\[
\langle f, g \rangle_{L_2} = \int f(x) g(x) \, dx .
\]

Based on (2) an intuitive measure of similarity between two functions \( f \) and \( g \) can be formulated which does not depend on the scales of \( f \) and \( g \), i.e.

\[
R_{fg} = \frac{\int f(x) g(x) \, dx}{\left( \int f(x)^2 \, dx \right)^{1/2} \left( \int g(x)^2 \, dx \right)^{1/2}} = \frac{\langle f, g \rangle_{L_2}}{\langle f, f \rangle_{L_2}^{1/2} \langle g, g \rangle_{L_2}^{1/2}} ,
\]

and so \( R_{fg} = 1 \) if \( f = cg \), where \( c > 0 \) is a positive constant, and \( R_{fg} = -1 \) if \( c < 0 \). Note that \( R_{fg} \) is a generalization of Pearson’s correlation coefficient for comparing two functions. Also note that, in general, calculation of \( R_{fg} \) would involve numerical integration over the domain, which may be computationally demanding.
An alternative metric space for functions is a reproducing kernel Hilbert space (RKHS) that, for a given reproducing kernel, can easily be constructed and is much simpler and quicker to use in practice. This alternative is very useful for our model because the covariance function $\sigma$ of the reference random field can be viewed as a reproducing kernel on $\mathbb{R}^m \times \mathbb{R}^m$ due to the properties of a general covariance function (e.g. symmetric and positive definite). Hence, the corresponding RKHS exists and can be written as $\mathcal{H}_\sigma = \{ f \mid f(x) = \sum_{i=1}^{k_A} \alpha_i \sigma(x_i^A - x) \}$. In this space the inner product of $f(x) = \sum_{i=1}^{k_A} \alpha_i \sigma(x_i^A - x) \in \mathcal{H}_\sigma$ and $g(x) = \sum_{j=1}^{k_B} \beta_j \sigma(x_j^B - x) \in \mathcal{H}_\sigma$ has the form
\[
<f, g >_{\mathcal{H}_\sigma} = \sum_{i=1}^{k_A} \sum_{j=1}^{k_B} \alpha_i \beta_j \sigma(x_i^A - x_j^B),
\]
which can be evaluated without expensive numerical integration.

Note that we can view the kriging predictor (1) as a member of $\mathcal{H}_\sigma$, and hence we can use the RKHS inner product $<\ldots>_{\mathcal{H}_\sigma}$ to measure the similarity between the predicted fields of $A$ and $B$. Let $\hat{Z}_A(x) = \sum_{i=1}^{k_A} w_i^A \sigma(x_i^A - x)$ and $\hat{Z}_B(x) = \sum_{j=1}^{k_B} w_j^B \sigma(\Phi(x_j^B) - x)$ denote the predicted fields of the marked point sets $A$ and $B$ in the relative position defined by $\Phi$. The similarity measure we use has the form
\[
C_{AB}(\phi) = \frac{<\hat{Z}_A, \hat{Z}_B >_{\mathcal{H}_\sigma}}{||\hat{Z}_A||_{\mathcal{H}_\sigma} ||\hat{Z}_B||_{\mathcal{H}_\sigma}},
\]
where $||\hat{Z}_M||_{\mathcal{H}_\sigma} = <\hat{Z}_M, \hat{Z}_M >_{\mathcal{H}_\sigma}$ ($M \in \{A, B\}$), and $\phi$ denotes the parameter vector of the unknown transformation $\Phi$. The numerator term measures the “overlap” of the fields (in a certain relative position) whereas the denominator is a transformation invariant normalizing constant which ensures that $C_{AB}(\phi) \in [-1, 1]$. Optimizing (3) with respect to the transformation parameters yields the “Kernel Carbo Index”
\[
C(A, B) = \sup_{\phi} C_{AB}(\phi) = \sup_{\phi} \frac{<\hat{Z}_A, \hat{Z}_B >_{\mathcal{H}_\sigma}}{||\hat{Z}_A||_{\mathcal{H}_\sigma} ||\hat{Z}_B||_{\mathcal{H}_\sigma}},
\]
in which configuration $B$ is transformed (by the relative transformation function $\Phi$) to be as similar as possible to configuration $A$.

In many applications it is of interest to match parts of objects rather than the entire configurations. Our steroids application is one such example because only a part of each molecule may fit into the binding pocket of the common receptor. We therefore also consider mask parameters which are indicator functions, signifying whether or not each atom should be used in the field comparison.

Note that the optimization in (4) is not straightforward in practice due to local maxima, especially when the mask parameters are also included. As an approximation to using the Kernel Carbo index in (4) we propose a Bayesian model where the likelihood is an increasing function of (3) and find the value of the similarity index (3) at the maximum a posteriori (MAP) estimates of the transformation and mask parameters. In Czogiel et al. (2011) the Bayesian methodology is applied to 2D and 3D data and suitable prior models and covariance functions based on the Matérn family are explored. The steroids data are analyzed and are available in the supplementary material to Czogiel et al. (2011).

References


Manifold stability and the central limit theorem for mean shape

Stephan F. Huckemann

Institute for Mathematical Stochastics, Georg-August Universität Göttingen

Abstract

In the literature there is an overwhelming number of concepts for a “mean” on shape spaces. First we classify those into three fundamental types. Then we introduce manifold stability which is essential to apply a standard two sample test on a non-manifold shape space. For sample means on Kendall’s shape spaces, we give the proof to establish manifold stability and illustrate consequences for the discrimination of non-concentrated shapes. We conclude with a discussion of the result for general shape spaces and population means.

1 Three Fundamental Types of Means

Suppose that a $X_1, \ldots, X_n \overset{i.i.d.}{\sim} X$ are random elements on a Riemannian manifold $M$ which is embedded in a Euclidean space $\mathbb{R}^m$. Then the orthogonal projection $\Phi : \mathbb{R}^m \to M$ is well defined except possibly for a set of Lebesgue measure zero (Bhattacharya and Patrangenaru (2003)). If $d$ denotes the intrinsic distance on $M$ due to the Riemannian structure, $\| \cdot \|$ the Euclidean norm, $d\Phi$ the projection to the tangent space and $E$ the classical expectation we have the three fundamental types of possibly set-valued means:

<table>
<thead>
<tr>
<th>intrinsic</th>
<th>extrinsic</th>
<th>residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\arg\min_{p \in M} E{d(p, X)^2}$</td>
<td>$\arg\min_{p \in M} E{|p - X|^2}$</td>
<td>$\arg\min_{p \in M} E{|d\Phi(p - X)|^2}$</td>
</tr>
</tbody>
</table>

In particular, extrinsic means are as unique as the orthogonal projection is since they are equal to the set given by $\Phi(E(X))$, (Hendriks and Landsman (1996); Bhattacharya and Patrangenaru (2003)). Moreover on spheres, residual means are identical to the set of dominating eigenvectors of $XX^T$, (Jupp (1988)).

In the modelling of shape, a compact Lie group $G$ (e.g. rotations) acts isometrically on $M$ giving the quotient shape space

$$\Sigma = M/G = \{[p] : p \in M\}, \quad [p] = \{gp : g \in G\}.$$  

With $g_p \in \arg\min_{g \in G} \|p - gp\|$ define the following distances on the quotient

$$d^{(i)}(\Sigma)([p], [p']) := \min_{g \in G} d(p, gp'), \quad d^{(e)}(\Sigma)([p], [p']) := \|p - gp'\|, \quad d^{(z)}(\Sigma)([p], [p']) := \|d\Phi_p - gp'\|$$

to obtain the corresponding means on the quotient:

<table>
<thead>
<tr>
<th>intrinsic</th>
<th>Ziezold</th>
<th>Procrustean</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\arg\min_{p \in \Sigma} E{d^{(i)}(\Sigma)([p], [X])^2}$</td>
<td>$\arg\min_{p \in \Sigma} E{d^{(z)}(\Sigma)([p], [X])^2}$</td>
<td>$\arg\min_{p \in \Sigma} E{d^{(e)}(\Sigma)([p], [X])^2}$</td>
</tr>
</tbody>
</table>

Most of the means in the literature are of one of these fundamental types, some of which are reported in Table 1. The so called partial Procrustes means on $\Sigma = S\Sigma_h^k$ are of all three types because in this case, all three of the above distances agree with one another.

1 supported by DFG HU 1575/2-1
Table 1: Classifying many concepts of means found in the literature.

<table>
<thead>
<tr>
<th>Concept in the literature</th>
<th>Fundamental Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Center of gravity (Kobayashi and Nomizu (1969); Kendall (1977))</td>
<td>Intrinsic</td>
</tr>
<tr>
<td>Crude residuals (Jupp (1988))</td>
<td>Residual</td>
</tr>
<tr>
<td>Full Procrustes mean on $\Sigma_k$ (Gower (1975))</td>
<td>Procrustean</td>
</tr>
<tr>
<td>Mean figure (Ziezold (1994))</td>
<td>Ziezold</td>
</tr>
<tr>
<td>Mean location (Hendriks and Landsman (1996))</td>
<td>Extrinsic</td>
</tr>
<tr>
<td>Partial Procrustes mean on $S\Sigma_m$ (Dryden and Mardia (1998))</td>
<td>All</td>
</tr>
<tr>
<td>Schoenberg mean on $R\Sigma_m$ (Bandulasiri et al (2009))</td>
<td>Extrinsic</td>
</tr>
<tr>
<td>Veronese Whitney mean of $\Sigma_2^k = CP^{k-2}$ (Bhattacharya and Patrangenaru (2003))</td>
<td>Extrinsic, Procrustean</td>
</tr>
</tbody>
</table>

2 Manifold Stability and the Central Limit Theorem

In general, a quotient space $\Sigma = M/G$ of a Riemannian manifold modulo a compact Lie group $G$ acting isometrically contains a dense manifold part $\Sigma^* = M^*/G$, $M^*$ is open and dense in $M$ and a possibly non-void singular part $\Sigma^0 = \Sigma \setminus \Sigma^* = M^0/G$, $M^0 = M \setminus M^*$. With the isometry group $I_p = \{ g \in G : gp = p \}$ at $p \in M$ we restrict ourselves to the case that $I_p = \{ e \}$ for $p \in M^*$ while $I_p \neq \{ e \}$ for $p = M^0$.

If $\mu_n$ is any of the above sample means and $\mu$ a corresponding unique population mean then $\mu_n \to \mu$ a.s. (Ziezold (1977); Bhattacharya and Patrangenaru (2003)) Further, if $\mu \in \Sigma^*$ is unique, then under suitable conditions, there is a Central Limit Theorem

$$\sqrt{n}(\phi(\mu_n) - \phi(\mu)) \overset{d}{\to} \mathcal{N}(0, \Sigma_\phi)$$

with a covariance matrix $\Sigma_\phi$ depending on a local chart $\phi$ (Bhattacharya and Patrangenaru (2005); Hendriks and Landsman (1998); Huckemann (2010a)). Hence, for a one-sample test for a specific mean shape on the manifold part, it may be assumed that sample means eventually lie on the manifold part as well. A similar two-sample test, however, requires the following property of manifold stability.

**Definition.** A mean shape enjoys manifold stability if the mean shape is assumed on the manifold part for any random shape assuming the manifold part with non-zero probability.

**Example 2.1.** Consider $M = S^2 = \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 = 1\}$ with the isometric action of $S^1 = \{ e^{i\phi} : \phi \in [0, \pi) \}$ given by $e^{i\phi}(x, y, z) = (x \cos \phi, y \sin \phi, z)$. Then

$$\Sigma = S^2/S^1 \cong [-1, 1] = \{-1\} \cup \{0\} \cup \{1\} \cup \Sigma^*.$$ 

As depicted in Figure 1, suppose that the distribution of $X$ on $S^2$ is uniform on $\{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 = 1\}$ with total weight $1/3$ combined with a point mass at the north pole $(0, 0, 1)$ of weight $2/3$. Then the set of residual means is given by the north and south pole. Hence the set of Procrustean means of $[X]$ is precisely the singular part $\Sigma^0$.

In consequence we have the following.

**Remark 2.2.** In general Procrustean means are not manifold stable.

3 Manifold Stability for Samples on Kendall’s Shape Space

Recall that Kendall’s shape space $\Sigma_k = S^k_m/SO(m)$ is the unit sphere $S^k_m$ in the space of $m \times (k-1)$ matrices modulo the special orthogonal group $SO(m)$. Every $m \times (k-1)$ matrix
Proof. Assume that $g(X_j) \neq 0$ for at least one $j \in \{1, \ldots, n\}$ with any $X_j \in [\Sigma_j] \in (\Sigma^k_m)^\ast$ and $p^\ast \in [p^\ast]$. Moreover, w.l.o.g. we may assume that $g_{p^\ast}X_j = X_j$ ($j = 1, \ldots, n$) as well as that there is $e \neq g \in G$ such that $0 \neq d(p^\ast, gX_1) = d(p^\ast, X_1)$, i.e. $\langle p^\ast, X_1 - gX_1 \rangle = 0$, as well as $gX_1 \neq X_1$. Then using differentiation for the following first line and Lagrange minimization under constraining conditions for the latter two, obtain at once

$$p^\ast \in \text{argmin}_{p \in \mathcal{S}_m^k} \frac{1}{n} \sum_{j=1}^n d(X_j, p)^2 \quad \Rightarrow \quad \frac{1}{n} \sum_{j=1}^n (X_j - p^\ast \langle X_j, p^\ast \rangle) \frac{\arccos \langle p^\ast, X_j \rangle}{\|X_j - p^\ast \langle X_j, p^\ast \rangle\|} = 0,$$

$$p^\ast \in \text{argmin}_{p \in \mathcal{S}_m^k} \frac{1}{n} \sum_{j=1}^n \|X_j - p\|^2 \quad \Rightarrow \quad \frac{1}{n} \sum_{j=1}^n (X_j - p^\ast \langle X_j, p^\ast \rangle) = 0,$$

$$p^\ast \in \text{argmin}_{p \in \mathcal{S}_m^k} \frac{1}{n} \sum_{j=1}^n (1 - \langle X_j, p \rangle^2) \quad \Rightarrow \quad \frac{1}{n} \sum_{j=1}^n \langle X_j, p^\ast \rangle (X_j - p^\ast \langle X_j, p^\ast \rangle) = 0.$$

By hypothesis, in every line $X_1$ can be replaced by $gX_1$ without changing the value of the sums. In the first two lines, this yields the contradiction $X_1 = gX_1$. In the third line we only have this contradiction unless $\langle p^\ast, X_1 \rangle = 0$. This yields the assertion. \hfill \Box

4 Discrimination Made Difficult for Full Procrustes Means

Here we consider on $\Sigma = \Sigma^4_3$ two maximally remote shapes $\sigma_1$ (1D) and $\sigma_2$ (2D) as depicted in Figure 2. Note that $\sigma_1 \in \mathcal{S}^d$ while $\sigma_2 \in \Sigma^\ast$. Below we report the percentages of successful discrimination of the following two groups.

- Group one consists six noisy versions of $\sigma_1$.
- Group two consists of one noisy version of $\sigma_1$ and five noisy versions of $\sigma_2$.

Multivariate normal isotropic noise with zero mean and standard deviation 0.01 has been added independently to every landmark. The discrimination has been based on a two-sample test of level 0.05 using tangent space coordinates obtained by orthogonal projection (for intrinsic and Ziezold means) and full Procrustes residuals (for full Procrustes means). 

Figure 1: Left: uniform distribution on a quarter of the equator and a heavy point mass at the north pole (black) together with the two residual means at the poles (blue), unique extrinsic mean (red) and below the unique intrinsic mean (magenta). Right: distribution (black) and means projected to the quotient giving two Procrustean means (blue) and a unique Ziezold mean (red) and below a unique intrinsic mean (magenta).
Figure 2: Two degenerate tetrahedral configurations in $\mathbb{R}^3$ the shapes of which are maximally remote in Kendall’s shape space: $\sigma_1$ (pink) and $\sigma_2$ (grey). Corresponding landmarks are shown as balls in the same color.

<table>
<thead>
<tr>
<th>Intrinsic means</th>
<th>Ziezold means</th>
<th>full Procrustes means</th>
</tr>
</thead>
<tbody>
<tr>
<td>76 %</td>
<td>74 %</td>
<td>25 %</td>
</tr>
</tbody>
</table>

Table 2: Simulated power of a discrimination test of two groups using a two-sample test based on the respective means.

As shown in Table 2, the discrimination power of a two-sample test based on Procrustes means can be much smaller than that of intrinsic and Ziezold means, if the Procrustes mean is attained on or close to the non-manifold part $\Sigma^0$.

5 Discussion

With greater effort, namely introducing a horizontal lifting in optimal position on a twisted product (general shape spaces can be described by this extension of the concept of a fiber bundle), excluding only very artificial settings hardly occurring in practice, the result derived here for sample means can be extended to intrinsic and Ziezold population means on general shape spaces, thus fully justifying a two-sample test based on these means, (Huckemann (2010b)). Under slightly more restrictive side conditions, e.g. that a random shape has a non-vanishing density w.r.t. the projection of the spherical volume, manifold stability can also be established for full Procrustes means, (Huckemann (2011)).

Curiously, the general result applied to the finite dimensional subspaces exhausting the quotient shape space of closed planar curves with arbitrary initial point introduced by Zahn and Roskies (1972) and further studied by Klassen et al. (2004), gives that the shape of the circle, since it is a singularity, can never be an intrinsic shape mean of non-circular curves.

References


Session IV
A fresh look at projective shape spaces

John T. Kent
Department of Statistics, University of Leeds

1 Introduction

Similarity shape analysis is a well-developed subject. For a $k \times p$ matrix $X$ giving the coordinates of $k$ points in $p$ dimensions, the (similarity) shape of $X$ consists of the information in $X$ that is invariant under $p$-dimensional similarity transformations, that is, translation, rotation and scaling. There is also a variant of the definition in which reflection is also allowed. In either case a similarity shape can be described as an equivalence class of configurations under a certain group of transformations. Much effort has been expended in this setting to find useful coordinate systems to represent shape and describe statistical variability.

By changing the group of transformations, it is possible to extend the concept of shape. Examples include affine shape, not studied here, and projective shape. Projective shape is a classic branch of mathematics dealing with the features in a camera view of an object that are invariant to the position of the camera. The subject is fundamental in computer vision (e.g. Faugeras and Luong, (2001); Hartley and Zisserman, (2000)).

Statistical work to represent and compare projective shapes has tended to focus on projective invariants (Goodall and Mardia, (1999)) or the use of projective frames (Mardia and Patrangenaru, (2005)), an analogue of Bookstein coordinates. However, a more promising general approach involves Procrustes methods based on “Tyler-standardization” (Mardia and Kent, (2006); Kent and Mardia, (2007)).

The purpose of this paper is to move beyond these earlier approaches by taking a closer look at the definition of projective shape. By considering different types of camera, it is possible to distinguish four different variants of projective shape based on four variants of the group of projective transformations (Section 2). A simple illustration based on four collinear points is given in Section 3.

First we set up the notation. Consider a $k \times m$ matrix $U_0$ representing an underlying object of $k$ points or landmarks in $m$ dimensions, of rank $m$, and let

$$U = [U_0 \ 1_k]$$

be the $k \times p$ dimensional matrix, $p = m + 1$, obtained from $U_0$ by appending a column of ones. A “scene” $W(k \times p)$ is obtained by linearly embedding $U_0$ in $\mathbb{R}^p$,

$$W = 1_k \alpha^T + U_0 B_0^T = U B^T,$$

where the nonsingular $p \times p$ matrix $B$ is obtained from $B_0$ by appending the vector $\alpha$.

Next, following Mardia and Kent (2006) and Kent and Mardia (2007), imagine a camera with focal point at the origin, with a spherical film of unit radius. The spherical film image can be represented by a $k \times p$ matrix $Y$ with $i$th row

$$y_i = w_i/||w_i||, \quad i = 1, \ldots, k.$$

Then projective shape consists of the information in the film image that is invariant under the position of the camera. Mathematically, it is more convenient to keep the focal point fixed at the origin and to let the matrix $B$ vary.
2 Types of camera

The new contribution in this paper is to observe that it is helpful to characterize the camera by two properties: (a) oriented vs. unoriented, and (b) directional vs. axial.

(a) In an oriented camera we know the side of the scene that the camera lies on. That is, mathematically we know whether $\det(B)$ is positive or negative. Conversely, for an unoriented camera, the sign of $\det(B)$ is unknown.

(b) For a directional camera, we declare that the points $\{y_i, i = 1, \ldots, k\}$ form the image of the scene on the film. For an axial camera, we declare that the unsigned directions or axes $\{\pm y_i, i = 1, \ldots, k\}$ form the image of the scene on the film.

Thus there are 4 types of camera, which can be abbreviated as OD, UD, OA, UA. Here “O” and “U” stand for oriented and unoriented; “D” and “A” stand for directional and axial. There is less information about the scene available from an axial camera than from a directional camera. Similarly there is less information about the scene from an unoriented camera than from an oriented one.

Classically, the subject of projective shape analysis has focused on the UA camera. However, from a practical point of view, the OD camera is often more relevant since we usually know which side of the object the camera is located and the physical construction of a camera enables us to determine the direction from the camera to the object.

Since it is not possible from the point $y_i$ on the film to tell how far the point $w_i$ is from the origin, the projective shape of a configuration $U$ is defined as the equivalence class of configurations $Z$,

$$Z = DUB^T,$$

under arbitrary choices for $D(k \times k)$ diagonal nonsingular and $B(p \times p)$ nonsingular. In the oriented case the restriction $\det(B) > 0$ is added; in the unoriented case either sign for $\det(B)$ is allowed. In the directional case the diagonal elements of $D$ are restricted to be positive, $d_i > 0$, $i = 1, \ldots, k$; in the axial case the diagonal elements $d_i$ can take either sign.

In Mardia and Kent (2006) and Kent and Mardia (2007), it was shown that, subject to mild regularity conditions, it is possible to make particular choices $B_T$ (with $\det(B) > 0$) and $D_T$ (diagonal, with positive elements), so that $X = D_T U B_T^T$ is “Tyler-standardized”; that is,

$$\text{diag}(XX^T) = I_k, \quad X^T X = (k/p)I_p.$$

Thus the rows of $X$ are unit vectors and the columns are orthonormal up to a factor $k/p$. The existence and construction of this standardized representation follows from earlier work of Tyler in the area of robustness.

From the point of view of an OD camera, the matrix $X$ is determined up to rotation, so that $X$ is effectively a “pre-shape” in a suitable similarity shape space. Hence Procrustes methods from similarity shape analysis can be used to compare different OD projective shapes.

3 Example: four collinear points

To illustrate the effects of different types of camera, we explore in detail the simplest nontrivial case of $k = 4$ landmarks on a line; that is, $m = 1$, or equivalently, $p = 2$. A general OD film image consists of 4 points on the circle lying within an open semi-circle. Write a general film image $Y$ in angular form as $\phi = (\phi_1, \phi_2, \phi_3, \phi_4)$, where the 4 angles lie within an open semi-circle. In the “regular” case in which the 4 points are distinct, the angular representation of the Tyler-standardized version $X$ takes the form $\theta = (-\delta/2, \delta/2, \pi/2 - \delta/2, \pi/2 + \delta/2)^T$ up to an arbitrary angular shift and up to the ordering of the points, where $0 < \delta < \pi/2$ (Mardia...
and Kent, (2006); Kent and Mardia, (2007)). There are $4! = 24$ possible orderings. Thus the regular part of OD projective shape space contains 24 line segments each of length $\pi/2$.

In passing we note that under the OA version of projective shape space, we identify cyclic shifts in the landmark ordering (leaving $24/4 = 6$ distinct line segments); under the UD version we identify reflections (leaving $24/2 = 12$ distinct line segments); and under the UA version we identify both cyclic shifts and reflections (leaving $24/8 = 3$ distinct line segments). The UA version was presented in Mardia and Kent (2006) and Kent and Mardia (2007), and the space of UA projective shapes was viewed as either a spherical triangle or a planar triangle.

Figure 1 shows a typical OD Tyler-standardized circular film image. The solid line in the underlying scene contains four landmarks, each marked by “X”. The OD film image is marked by the diamonds on the circular film. In the OA film image, only the axes demarcated by the diamonds and triangles can be identified. For the UD and UA unoriented cameras, the image here and its reflection about the solid line cannot be distinguished.

The description of OD projective shape space is completed by including several partially singular cases. These include (a) a double-pair coincidence: $\phi_1 = \phi_2 < \phi_3 = \phi_4$, (b) a one-sided single pair coincidence, $\phi_1 < \phi_2 < \phi_3 = \phi_4$, and (c) two-sided single pair coincidence: $\phi_1 < \phi_2 = \phi_3 < \phi_4$. It turns out that these cases are identified with one another in classic projective geometry (UA projective shape space) because a suitable version of the cross-ratio reduces to 0 in all cases, but they are partly distinguished from one another in the OD projective shape space.

4 Conclusion

Projective shape is a subtle subject with many curious properties and limiting results. In particular, it has been particularly challenging to construct plausible statistical models, though some progress have been made in the setting of projective frames (Mardia and Patrangenaru, (2005)). The recognition of different types of camera and the use of Tyler-standardized coordinates provides a clear geometric picture of projective shape, but the construction of suitable statistical models in this setting remains an open task.

References


Pliable template: a null model for shape variation when the mean shape changes

Christian Peter Klingenberg
Faculty of Life Sciences, University of Manchester

1 Introduction

The shapes of most organisms are integrated—their parts vary in a coordinated manner, for developmental, functional and evolutionary reasons (e.g. Klingenberg, (2008)). Because variation is the raw material for evolution by selection or random drift, understanding the patterns of integrated variation of shape is important for evolutionary studies. But variation is also of central importance in other contexts, such as medical research. Analyses of the patterns of variation of organsimal shapes are therefore important for biological studies.

Morphometric studies have usually considered the differences in mean shapes between some groups and the shape variation around the respective group means as two separate questions. This unlikely to be a biologically realistic view. The changes of the mean shapes among several groups of organism result from changes in the developmental system by which the organisms build their shapes. As consequence of the changes in development that are responsible for evolution of differences in the mean shape among groups, it is likely that the propensity to vary about the group mean shapes has changed in concert with those mean shapes themselves. Accordingly, morphometric studies need tools to consider changes in means and variation around the means together.

This paper informally presents a null model, called the pliable template, which starts to address this problem. The model views the mean shape as providing a sort of template for the variation around it. Changes in the mean shape therefore are transforming this template, so that the patterns of variation around the mean are expected to undergo a corresponding change. For instance, it is plausible that the patterns of variation for neck vertebrae of the giraffe are “stretched-out” versions of the patterns found in the okapi (a related species with a “normal”-length neck), just as the average shape of a neck vertebra in the giraffe is mostly a stretched-out version of the corresponding vertebra of the okapi. The pliable template model provides a null expectation of how the variation around a mean might change passively as a result of changes in the mean shape.

The pliable template model is related to D’Arcy Thompson’s classical method of transformation grids (e.g. Thompson, (1961)). As in the transformation grids, the pliable template views the anatomical structure as being embedded in a sort of matrix that is deformed when the shape of the structure changes. Accordingly, it is possible to use the thin-plate spline (Bookstein, (1989)) as the basis for implementing the pliable template model. The model can be applied to both contrasts between groups (e.g. males in females in different species, patients and controls in different populations) and patterns of variation in different samples (e.g. patterns of morphological integration in different species). I present work in progress on this model and its implementation.

2 Outline of the model and implementation

The main idea behind the model is that a change in the mean shape causes a change in the patterns of variation around the mean. The mean shape acts as a sort of template for variation. Because the mean shape itself can change, it is a template that is pliable. The motivation for this model is biological, based on considerations of the developmental system that produces the
structures under study and is therefore responsible for both the average shape and the variation around that average. To make the idea amenable to morphometric analysis, the model must be expressed in terms of the tools of statistical shape analysis.

The central idea is that the change of the mean shape affects variation in the surroundings of the landmark positions in the mean shape. Because the thin-plate spline (Bookstein, 1989) provides a smooth interpolation from the changes of landmark positions to the entire plane or three-dimensional space in which the landmark configuration is situated, it is suitable for implementing the pliable template model. In brief, the thin-plate spline is used to “unwarp” the mean shapes of two or more groups to a common consensus shape, such as the overall Procrustes mean shape for all configurations. The thin-plate spline for each group is then applied to all the landmark configurations in that group, and the desired comparisons can be carried out on the modified data.

This approach can be used in different contexts: to study differences between subgroups or to analyze variation within groups by comparing their covariance matrices.

2.1 Differences between subgroups

Comparisons of differences between subgroups within groups involve questions such as whether sex dimorphism is the same in different species or whether the differences between patients and healthy controls are the same in different ethnic groups. To address these questions in the framework of the pliable template model, the thin-plate spline is used to provide an unwarping from the mean shapes of the main groups (species, ethnic group etc.) to the grand mean shape. Differences between the subgroups within each group are then computed from these unwarped data, and the differences can be statistically evaluated by bootstrap or permutation tests (e.g. Manly, 2007).

2.2 Comparison of covariance matrices

The covariance matrices to be compared can be computed from the unwarped data for each group. For computing covariance matrices from the asymmetric component of shapes with bilateral symmetry (Mardia et al., 2000; Klingenberg et al., 2002), the same unwarping procedure can be used, provided the symmetric group mean shape is added to the asymmetry values of each specimen.

For some types of covariance matrices of great biological interest, such as genetic covariance matrices (e.g. Lynch and Walsh, 1998), individual observations are not available. Because the thin-plate spline is nearly linear in the close neighbourhood of the landmarks, it is possible to use a linear approximation of the unwarping transformation in these cases (genetic covariance matrices are computed from variation within populations, where shape variation is small, so that the linear approximation will usually be appropriate).

The statistical comparison of the resulting covariance matrices can use principal component analysis and models derived from it to identify and evaluate aspects of the covariance structure that are shared or that differ between groups (Flury, 1988, Boik, 2002). These methods have not been applied to shape analysis so far, and considerable work may be needed to adapt them to that context.

3 Outlook

Comparisons of covariance patterns among species have mostly been based on statistical models such as common principal components (Flury, 1988), which do not take changes in the means into consideration when comparing covariance matrices in biological studies (e.g. Steppan et al., 2002). The pliable template model is a first attempt at incorporating changes in mean shape into comparisons of morphometric variation among groups.
It is clear that the model is not biologically realistic, and so it is desirable that other models or entire frameworks of models are developed for more systematic comparisons of morphometric variation. Common principal components and the models related to it (Flury, (1988); Boik, (2002)) provide such frameworks in a general multivariate context, but not specifically for shape analysis.

The pliable template model will be implemented in the *MorphoJ* software (Klingenberg, (2011)) to make it available to a broad range of users. This model may contribute to a growing interest in approaches that combine considerations from evolutionary and developmental biology with statistical shape analysis (e.g. Klingenberg, (2010)).

**References**


Shape analysis and skeletal form and function: a personal perspective on the geometric morphometric ‘revolution’

Paul O’Higgins
The Centre for Anatomical and Human Sciences, Hull York Medical School, The University of York

It is a pleasure to have the opportunity to join this workshop to celebrate 30 years of LASR. I am not a statistician, but rather a user of the intellectual product of statisticians in my biological work. I am very grateful for the role these workshops have played in the development of the methods of “statistical shape analysis” (Dryden and Mardia, (1998)) or “geometric morphometrics” (GM; Rohlf and Marcus, (1993)) which initially brought me into contact with Professors Mardia and Kent at Leeds, and later Professor Bookstein, then at Michigan. It is surprising and pleasing to see the impact these methods. A Google search of the terms "geometric morphometrics" OR "statistical shape analysis" for each of the years 1992-2010 shows the rising numbers of publications using these approaches (Figure 1).

Mostly these are applications to biology, and key themes in this context include genetics, evolution and comparative form and function. A major driver in the uptake of these methods has been the development and release of software tools, first two dimensional and later three dimensional, that have facilitated uptake of these methods by biologists. I have been involved in the development of two of these tools, with the latest being the EVAN toolkit. This arose from an EU wide collaboration, EVAN (European Virtual Anthropology, 2006-9), a research training network supported by EU FP6 Marie Curie Actions. The focus of the resulting software tool has been on future applications and as such it has moved away from the emphasis of earlier tools on the analysis of variations among landmark configurations based on (few) identifiable equivalent points towards descriptions and analyses of surfaces using semilandmarks, providing sophisticated tools for warping and for analysis and visualisation of variations among surfaces.

In the last ten years my own interests have moved away from applications of GM to skull growth and towards the analysis of form and function in skeletal material, especially in primates. There is a good reason for this; GM methods are well suited to analyses of patterns of growth but to understand causes, the impact of function is key. In this regard GM analyses of covariances between skeletal form and functionally relevant variables provide insights into how form relates to function. But, to understand the physical meanings of such relationships and to assess in detail how aspects of skeletal form, form variability and growth relate to function it is necessary to understand how loads are borne, how skeletal elements deform and how varying loads cause varying deformations. To these ends, in collaboration with engineers and computer scientists, I have recently focussed on the development of tools to facilitate simulation of skeletal loading. These are specifically tailored for skeletal biologists and implement the standard engineering approach of finite elements analysis (FEA) in ways that specifically
suit the modelling of skeletal elements with applied muscle loads and constraints. FEA results in maps of stress and strain of skeletal elements which are of specific use in interpreting the consequences for the particular skeletal element of the particular loading applied. Thus, regions of high or low strain may predict bone deposition or resorption during growth or adaptation and of particularly concentrated high strain, fracture. As such this approach pervades biomedical engineering studies of e.g. the design of medical implants to replace joints.

However, in my own work I have different needs from those of the biomedical engineers in that I wish to compare skeletal deformations under different loadings and eventually assess how different skeletons vary in their responses to loads. Thus, many skeletal analyses exist in the literature that relate form to biological adaptations and so to function, but it would be a significant advance to base functional comparisons directly on the mechanical behaviour of bones. To these ends my most recent explorations have turned to how GM tools might combine with FEA.

FEA first computes the deformation of the whole mesh based on the specified material properties, loads and constraints, and then relative deformations throughout the mesh are quantitatively assessed by computing strains. FEA outputs the displacements of the nodes of the underlying mesh and these completely describe how the nodes move relative to each other and how the mesh as a whole is translated, rotated and scaled. Strains are usually computed to assess relative deformations throughout the mesh but GM offers an interesting alternative; to treat these relative deformations as the vector between unloaded and loaded shapes in Kendall’s shape space and the difference in centroid size. The result is an alternative representation of the resulting deformation in terms of Procrustes or Form (Procrustes plus size) distances. Empirically, form distances behave rather well in relation to applied loads and isometric scaling. Thus doubling the load doubles the distance, isometrically scaling the specimen by doubling lengths, results in a halving of the distance and halving of stiffness (Young’s modulus) doubles distance. Where multiple deformations are being compared, plots of principal components of form are particularly informative. The overall approach is similar to that of geometric motion analysis (Slice, (1999); O’Higgins et al., (2002); Adams and Cerney, (2007)) in that the aim is to compare changes in form rather than forms themselves.

Although potentially useful, there are several underlying issues with this approach that require addressing. Thus, when working with a single specimen and assessing e.g. the effects of different loads and constraints, point equivalences are known absolutely; the nodes of the mesh or anatomical landmarks on the mesh are the same from loadcase to loadcase. Thus GM analyses can be carried out at any scale between that of the mesh nodes and that of (the much sparser) anatomical landmarks. This does not apply, however, when comparing deformations between specimens. The identification of equivalences between them is limited to a few anatomical landmarks and as Bookstein has pointed out (Virtual anthropology meets biomechanics workshop, Vienna 2010) the methods of sliding of semi landmarks cannot be applied at all to improve the situation. This means that in comparing deformations due to loading between different exemplars of the same skeletal element, the resolution and the scale of this comparison is inevitably limited by the density of equivalent points. This is considerably less than that of the points defining the FE mesh. The comparison of deformations between different objects is also complicated by the fact that the differences among specimens are likely to be much greater than those due to loading. Steps have to be taken to deal with the deformations alone. A further issue arises in interpretation of results in that while transformation grids may prove a useful visual device in understanding the results of GM analyses of deformations from FEA, meaningful strains cannot be computed from them.

Besides comparing the results of loading simulations of skeletal material there are other important connections between GM and FEA on the input side, in making models of skeletal form.
for subsequent FEA. Thus, FEA uses a single specimen and most (if not all) published applications of FEA to interpreting skeletal function in a zoological context use single specimens, ignoring variation. Returning to the introduction of this abstract, recent software tools for GM, including the EVAN toolkit, offer powerful tools for warping surface models. By applying such warpings to FEA models it is feasible to alter their geometry in ways that allow exploration of variability. Thus, while it may take many weeks to segment and build a model of a skeletal structure from CT scans it takes moments to warp this model to represent e.g. the mean or limits of variation of a sample. Additionally GM analyses of co-variations between form and function might for instance result in a regression model of form on some interesting biomechanical quantities such as muscle forces. FEA then offers the prospect of detailed mechanical simulation of the resulting warped models in order to understand why form varies in this particular way. There are a few caveats and cautions to bear in mind in carrying out analyses of warped FE models, particularly in relation to the effects on function of warping internal bony anatomy but initial work using solid models is producing promising results and pointers in this regard.

Thus, GM methods have played a central role in my academic life and look to set to continue to do so. More than 25 years ago as a PhD student in Leeds, I was fortunate to encounter Kanti Mardia and John Kent at a time when Fred Bookstein was at the beginning of his work in shape analysis. They came together first at Leeds for what turned out to be a very long exploration of statistical shape analysis and I was privileged to be there to witness their debates and to work alongside and learn from another PhD student with specialist skills in this area, Ian Dryden. That exploration has for me led to an academic lifetime of studies of skeletal size and shape variability. As my chart shows it was a lonely place in biology for some years but in recent times the methods of GM have come to be regarded as the standard approach in biology for the landmark based comparison of form. Likewise I have a feeling that the application of GM tools to the study of skeletal function via links to FEA will eventually yield new approaches that will eventually impact on comparative biomechanics.

References


Session V
Data integration in genomics: a new biostatistical challenge

Michael G. Schimek
Medical University of Graz, Austria

1 Introduction

Current research in molecular biology and medicine typically produces a large amount of data from rather small studies, many of them addressing similar questions. As a direct result, there is a strong demand for methods that combine or integrate diverse findings. Public data repositories make such information easily accessible. The central problem is the need for data to be combined, that differ in important aspects such as source (e.g. platform in a microarray experiment), quantification (e.g. of gene expression), scale, and study size. Moreover, often only subsets of unknown size of these data are relevant for the purpose of integration. Most recently, rank-based methods have emerged as useful tools for combining such information.

In this presentation, we describe a new approach to deal with two or more rankings of the same set of objects (e.g. genes) that combines the estimation of the lengths of highly conforming sublists with their stochastic aggregation. The goal is to obtain a much smaller set of informative common objects in a new rank order. The input lists can be of large or huge size, with irregular or incomplete rankings due to random and missing assignments. A moderate deviation-based inference procedure and a cross-entropy Monte Carlo technique are used to handle the combinatorial complexity of the task. Two alternative distance measures are modified to accommodate truncated list information. Finally, we illustrate the outlined approach, applying it on multiple microarray data. It is implemented in the R package TopKLists.

2 Assumptions

Let us have \( \ell \) input lists representing rank positions of the same set of \( N \) objects. The rank assignment mechanisms operate independently of each other (for each list). The ranking of objects is from 1 to \( N \), for highest to lowest without ties, but missing assignments are allowed.

We assume a discrete space \( O \) that contains all \( N \) objects, denoted by \( o_i, i = 1, \ldots, N \). Since all objects \( o \) can be associated with a unique label \( i = 1, \ldots, N \), \( O \) can be viewed without loss of generality as a list \( O = \{1, 2, \ldots, N\} \). Let us denote the rank of element \( o_i \) in \( O \) by \( R(i) \) under a particular assignment. Then a permutation of \( O, \tau(O) = \{1, 2, \ldots, N\} \), such that \( R(\tau(i)) \leq R(j) \) for any \( i < j \) is a complete ranking of the items in \( O \). We refer to \( \tau(O) \) as a full ranked list, and to \( \tau_i(i) \) as the rank of object \( o_i \) under the assignment mechanism \( \tau \).

In genomic data integration, a full ranked list is neither desirable nor available. Instead, one is only interested in a partial list (sub-space) \( O' \subset O \) of length \( k \). Without loss of generality, we assume that the partial ranked list \( \tau(O') = \{o'_1, o'_2, \ldots, o'_k\} \) is ordered according to their ranks such that \( R(o'_i) < R(o'_j) \) for \( i < j \). It is implicitly assumed that all the items that are in \( O \) but not in \( O' \) are ranked lower than \( k \) (i.e. have indices \( k+1, k+2, \ldots, N \)).

3 Moderate deviation-based top-\( k \) list estimation

In Hall and Schimiek (2010) a nonparametric inference method for the truncation of paired ranked lists was developed. The associated iterative algorithm allows estimating the length, \( k \), of a top-\( k \) list in the presence of irregular and missing assignments. Overlap of rank positions in two input lists is represented by a sequence of indicators, where \( I_j = 1 \) if the ranking, given by the second assessor to the object ranked \( j \) by the first assessor, is not more than \( \delta \)
index positions distant from \( j \), and otherwise \( I_j = 0 \). The variables \( I_j \) are assumed to follow a Bernoulli random distribution. This implies independence, which is motivated by \( k \ll N \) and a strong random contribution due to irregular assignments in real data. However, the above mentioned authors could prove that their theoretical results obtained under the assumption of complete independence also apply to the situation of \( m \)-dependence.

For the Bernoulli random variables \( I_1, \ldots, I_N \), it is assumed that \( p_j \geq \frac{1}{2} \) for each \( j < j_0 \), and \( p_j = \frac{1}{2} \) for \( j \geq j_0 \), and in addition, a “general decrease” of \( p_j \) for increasing \( j \) that need not be monotone. The index \( j_0 \) is the rank position where the consensus information of the two lists, representing the same set of objects, degenerates into noise (degradation of information). The estimation of \( j_0 - 1 = \tilde{k} \) is achieved via a moderate deviation-based approach.

In theoretical analysis of the probability that an estimator, computed from a pilot sample size \( \nu \), exceeds a value \( z \), the deviation above \( z \) is said to be a moderate deviation if its associated probability is polynomially small as a function of \( \nu \), and to be a large deviation if the probability is exponentially small in \( \nu \). In regular cases, the values of \( z = z_\nu \) that are associated with moderate deviations are

\[
z_\nu \equiv \left( C' \nu^{-1} \log \nu \right)^{1/2},
\]

where \( C > \frac{1}{4} \). The null hypothesis \( H_0 \) that \( p_k = \frac{1}{2} \) for \( \nu \) consecutive values of \( k \), versus the alternative \( H_1 \) that \( p_k > \frac{1}{2} \) for at least one of the values of \( k \), is rejected if and only if \( \tilde{p}_j^+ - \frac{1}{2} > z_\nu \).

The quantities \( \hat{p}_j^+ \) and \( \hat{p}_j^- \) represent estimates of \( p_j \) computed from the \( \nu \) data pairs \( I_m \) for which \( m \) lies immediately to the right of \( j \), or immediately to the left of \( j \), respectively. Under \( H_0 \), the variance of \( \hat{p}_j^\pm \) equals \((4\nu)^{-1}\), hence we can evaluate the above inference procedure in practice. However, apart from the pilot sample size \( \nu \) and the constant \( C \), inference results also depend on the distance \( \delta \).

4 Stochastic rank aggregation of multiple lists

We cannot aggregate rank positions in a stochastic manner without an optimization criterion which itself is specific to the choice of a distance measure. Measures conforming with the so-called generalized Kemeny guidelines are most appropriate for this task. The well-known Kendall’s \( \tau \) distance is one of them.

Let us have rankings \( \tau_1, \tau_2, \ldots, \tau_r \) (these are usually from truncated lists) as input. Let \( O' = \bigcup_{l=1}^{r} O'_l \) and \( \tau \) be the consensus ranking with respect to \( O' \), assuming that the list lengths \( k_l \)'s are fixed. Then our goal is to find an estimate of \( \tau \) (i.e. an ordered subset of \( O' \)) that minimizes the sum of weighted distances between \( \tau \) and each of the lists \( \tau_l \). We seek \( \tau^* \) such that

\[
\tau^* = \arg \min_{\tau} \left\{ \sum_{l=1}^{r} w_l d(\tau, \tau_l), \tau \subset O' \right\},
\]

where \( w = (w_1, w_2, w_3, \ldots, w_r) \) is a weight vector that can be used to specify prior information on the relative importance of the input lists, and \( d \) is a distance measure. Note, the ranked lists can be of different lengths and from different spaces.

When \( w_l = 1/\ell \) for all \( l \) and \( d(\tau, \tau_l) = K(\tau, \tau_l) \), denoting Kendall’s \( \tau \) distance, the estimate \( \tau^* \) in equation (1) reduces to the Kemeny optimal aggregation (Schimek, Lin and Wang, (2012)). The actual computation of the optimal aggregation of full lists of size \( N \), or even partial lists when \( k \) is large, constitutes a severe combinatorial problem. To overcome this obstacle, Markov chain (MC) approaches have been devised (see e.g. Deconde et al., (2006)). Consensus rankings (majority preferences) between pairs of items across lists are formed. The assumption that assessors continuously compare pairs of alternatives during their decision process leads naturally to a MC representation. A decision matrix characterizes the potential transitions between alternative decisions. The limiting equilibrium distribution represents the global assessment of
all objects. An advantage of MC approaches is that they do not require all lists to comprise the same objects. A drawback is the associated computational effort.

A more recent approach to solve (1) is cross-entropy Monte Carlo (CEMC) introduced by Rubinstein (1997) for estimating probabilities of rare events in complex stochastic networks and then followed up with complicated combinatorial optimization problems. Lin and Ding (2009) could extend the CEMC approach by the introduction of an Order Explicit Algorithm (OEA). The orders of the objects in the optimal list are explicitly given in the probability matrix \( v \). Taking advantage of this fact, Lin and Ding’s algorithm is much faster, thus permitting for those list lengths that are required for the integration of genomic data.

Let us assume a random matrix \( X = (X_{jr})_{N \times k} \) with each component variable \( X \) taking the values 0 or 1, and with the constraints of its columns summing up to 1 and its rows summing up to at most 1. This implies that each realization of \( X, x \), uniquely determines an ordered list of length \( k \) by the position of 1’s in each column from left to right. The length \( k \) of the aggregated top-\( k \) list can be any number not exceeding the size of the union of the full lists, but usually much smaller than \( N \). Let \( v = (p_{jr})_{N \times k} \) denote the corresponding probability matrix (each column sums to 1). For each column variable, \( X_r = (X_{1r}, X_{2r}, \ldots, X_{Nr}) \), a multinomial distribution with sample size 1 and probability vector \( v_r = (p_{1r}, p_{2r}, \ldots, p_{Nr}) \) under the constraints of the joint column variables is assumed.

Any realization \( x \) of \( X \) uniquely determines the corresponding top-\( k \) candidate list without reference to the probability matrix \( v \). That is, \( A = f(x) = (x_{jr}, x_{jr} = 1, j = 1, 2, \ldots, N, r = 1, 2, \ldots, k) \). The 1’s in each of the \( k \) columns make up the top-\( k \) list, in that order. Given the 1-to-1-correspondence between \( A \) and \( x \), finding \( A^* \) is equivalent to finding \( x^* \) that minimizes \( \Phi \{ f(x) \} \).

Using CEMC, \( x^* \) can be obtained by iteratively updating the parameter matrix \( v \) such that, iteration by iteration, \( P_{v}(x) \) will place more and more of its probability mass on the \( x \)’s that are in the “neighbourhood” of \( x^* \). Loosely speaking, \( x \) is called a neighbour of \( x^* \) if the corresponding value of the objective function, \( y = \Phi \{ f(x; v) \}, \) is close to the minimum \( y^* \). Let \( v \) be the current estimate of the parameter matrix. The next parameter update \( v' \) is chosen to minimize the cross entropy \( CE(Q^*, P_{v'}) \) between the distributions \( P_{v'} = P_{v'}(x) \) and \( Q^* \), where \( Q^* \) (see below) is the ideal but unobtainable importance sampling distribution for estimating the rare probability \( b = P_{v} [ \Phi \{ f(x; v) \} \leq y] \),

\[
Q^*(x) = \frac{I[y \leq \Phi \{ f(x; v) \}] \log P_{v'}(x)}{b}.
\]

Minimizing \( CE(Q^*, P_{v'}) \) is equivalent to maximizing

\[
\sum_{x} \{ I[\Phi \{ f(x; v) \} \leq y] \log P_{v'}(x) \} P_{v}(x) = \mathbb{E}_{v} [ I[y \leq \Phi \{ f(x; v) \} \leq y] \log P_{v'}(x) ],
\]

which is now free from the probability \( b \) to be estimated.

Suppose \( x_i = (x_{ijr})_{N \times k}, i = 1, 2, \ldots, m, \) is a sample drawn from \( P_{v}(x) \) with the current parameter specification \( v \) and the corresponding candidate top-\( k \) lists denoted as \( \tau_i = f(x_i), i = 1, 2, \ldots, m \). Then

\[
\nu_{new} = \arg \max_{v'} \left\{ \frac{1}{m} \sum_{i=1}^{m} I[y \leq \Phi \{ f(x_i; v) \}] \log P_{v'}(x_i) \right\}
\]  
\[
= \left[ \frac{\sum_{i=1}^{m} I[\Phi(\tau_i) \leq y]}{\sum_{i=1}^{m} I[\Phi(\tau_i) \leq y]} \right]_{j=1,\ldots,N;r=1,\ldots,k},
\]

can be used in the update for the next parameter matrix \( v' \). In addition, the threshold value \( y \) can also be updated iteratively. Equation (2) respectively (3) lead to the construction of a sequence,
which converges to a value $y_\infty$ close to $y^*$ (Margolin, (2005)). Similarly, $v_0, v_1, \ldots$, converges to $v_\infty$, with the corresponding $P_{v_\infty}(x)$ placing most of its probability mass on the $x$’s that satisfy $\Phi\{f(x; v)\} \leq y_\infty$ (Lin and Ding, (2009)).

5 An application

A topic of current interest in molecular science is the integration of results obtained from numerous microarray experiments. Real-time quantitative reverse transcription polymerase chain reaction (RT-qPCR) is a widely used laboratory technology for the precise analysis of gene expression. Normalization of gene expression measurements is crucial and is based on one or several control genes. The expression of these genes must be stable (with overall high expression and low variability) across different types of tissue and conditions. They are essential for molecular scientists because these special genes are associated with the maintenance of primary cellular functions and therefore often called *housekeeping* genes.

![Figure 1: Global estimate $\hat{j}_0$ for the three lists $\tau_1$, $\tau_2$, and $\tau_3$ in dependence of $\delta$](image)

Popovici et al. (2009) aimed at the identification of relevant control genes, by mining microarray gene expression data belonging to various types of cancer tissue. They proposed a method for the ranking of genes which was applied to each individual data set. Taking into account the intensity of gene expression and its variability across the samples, the best candidates were ranked in the top range of these lists. Subsequently, a rank product score (Breitling et al., (2004)) was applied in order to aggregate lists, providing a final list of genes with a new ranking. However, the issue of estimating the index for list truncation in order to obtain only relevant genes remained unsolved. This aspect is tackled in our example we describe below.

For the purpose of demonstration, we analysed only a subset of ranked lists, representing three common types of cancer, breast ($\tau_1$), prostate ($\tau_2$), and colon ($\tau_3$), from a total of eight lists in Popovici et al. (2009), each of length $N = 10,000$. Executing the inference procedure for all pairwise combinations of the three input lists produced overall estimates $\hat{k}^*$ depending on the distance parameter $\delta$ and the pilot sample size $\nu$. Quite important is the choice of $\delta$. Hall and Schimek (2010) introduced the so-called $\Delta$-plot which displays the decrease of discordance of two lists as a function of increasing $\delta$. For $\tau_1$ vs. $\tau_2$ as well as $\tau_1$ vs. $\tau_3$ (figures not displayed), the respective plots indicate an adequate range of $\delta$-values between 10 and 100. As the study goal was the identification of control genes for RT-qPCR, a small set had been required and we decided for $\delta = 10$. 

124
Table 1: Aggregation results comprising 40 genes based on truncated gene lists of length 65. Columns represent lists obtained from different methods.

<table>
<thead>
<tr>
<th>Method Rank</th>
<th>CEMC</th>
<th>MC</th>
<th>BORDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CALM2</td>
<td>CALM2</td>
<td>CALM2</td>
</tr>
<tr>
<td>2</td>
<td>RPL37</td>
<td>RPL9</td>
<td>HNRP1</td>
</tr>
<tr>
<td>3</td>
<td>RPL27A</td>
<td>RPL37</td>
<td>RPL27A</td>
</tr>
<tr>
<td>4</td>
<td>RPL30</td>
<td>RPL30</td>
<td>RPL30</td>
</tr>
<tr>
<td>5</td>
<td>RPL9</td>
<td>RPL27A</td>
<td>RPL9</td>
</tr>
<tr>
<td>6</td>
<td>RPS24</td>
<td>RPS13</td>
<td>NACA</td>
</tr>
<tr>
<td>7</td>
<td>NACA</td>
<td>NACA</td>
<td>RPL37</td>
</tr>
<tr>
<td>8</td>
<td>RPL27</td>
<td>HNRP1A</td>
<td>RPS11</td>
</tr>
<tr>
<td>9</td>
<td>RPS29</td>
<td>RPL24</td>
<td>MORF4L2</td>
</tr>
<tr>
<td>10</td>
<td>RPS13</td>
<td>RPS11</td>
<td>LAPTMA4</td>
</tr>
<tr>
<td>11</td>
<td>LAPTMA4</td>
<td>RPL27</td>
<td>RPS24</td>
</tr>
<tr>
<td>12</td>
<td>RPS11</td>
<td>RPS24</td>
<td>RPS13</td>
</tr>
<tr>
<td>13</td>
<td>HNRP1A</td>
<td>LAPTMA4</td>
<td>RPS29</td>
</tr>
<tr>
<td>14</td>
<td>RPL24</td>
<td>RPL38</td>
<td>RPS27A</td>
</tr>
<tr>
<td>15</td>
<td>HNRPK</td>
<td>RPS29</td>
<td>RPL24</td>
</tr>
<tr>
<td>16</td>
<td>ACTG1</td>
<td>MRFA1</td>
<td>RPL27</td>
</tr>
<tr>
<td>17</td>
<td>RPS27A</td>
<td>RPLP1</td>
<td>RPL31</td>
</tr>
<tr>
<td>18</td>
<td>RPL38</td>
<td>MORF4L2</td>
<td>RPL38</td>
</tr>
<tr>
<td>19</td>
<td>SRP14</td>
<td>SRP14</td>
<td>RPLP1</td>
</tr>
<tr>
<td>20</td>
<td>RPS15</td>
<td>HNRPK</td>
<td>RPL6</td>
</tr>
<tr>
<td>21</td>
<td>RPL31</td>
<td>RPL6</td>
<td>RPS17</td>
</tr>
<tr>
<td>22</td>
<td>CD63</td>
<td>RPS17</td>
<td>RPS12</td>
</tr>
<tr>
<td>23</td>
<td>RPL6</td>
<td>DYNLL1</td>
<td>CD63</td>
</tr>
<tr>
<td>24</td>
<td>FAU</td>
<td>RPL31</td>
<td>ACTG1</td>
</tr>
<tr>
<td>25</td>
<td>RPS17</td>
<td>RPS12</td>
<td>RPS15</td>
</tr>
<tr>
<td>26</td>
<td>EEF1G</td>
<td>RPS27A</td>
<td>MRFA1</td>
</tr>
<tr>
<td>27</td>
<td>RPLP1</td>
<td>CD63</td>
<td>UBB</td>
</tr>
<tr>
<td>28</td>
<td>MRFA1</td>
<td>ACTG1</td>
<td>RPL32</td>
</tr>
<tr>
<td>29</td>
<td>UBB</td>
<td>USMG5</td>
<td>FAU</td>
</tr>
<tr>
<td>30</td>
<td>USMG5</td>
<td>EEF1G</td>
<td>HNRPK</td>
</tr>
<tr>
<td>31</td>
<td>RPL32</td>
<td>CLTC</td>
<td>EEF1G</td>
</tr>
<tr>
<td>32</td>
<td>MORF4L2</td>
<td>UBB</td>
<td>SRP14</td>
</tr>
<tr>
<td>33</td>
<td>MYL6</td>
<td>RPS15</td>
<td>DYNLL1</td>
</tr>
<tr>
<td>34</td>
<td>EIF4A2</td>
<td>RPL39</td>
<td>RPL19</td>
</tr>
<tr>
<td>35</td>
<td>GABARAP</td>
<td>RPL32</td>
<td>CLTC</td>
</tr>
<tr>
<td>36</td>
<td>RPL11</td>
<td>RPS20</td>
<td>USMG5</td>
</tr>
<tr>
<td>37</td>
<td>PTGES3</td>
<td>RPS4X</td>
<td>GABARAP</td>
</tr>
<tr>
<td>38</td>
<td>RPL19</td>
<td>RPL19</td>
<td>RPL35</td>
</tr>
<tr>
<td>39</td>
<td>NDUF5</td>
<td>FAU</td>
<td>ATPF1</td>
</tr>
<tr>
<td>40</td>
<td>DYNLL1</td>
<td>GABARAP</td>
<td>RPL11</td>
</tr>
</tbody>
</table>

To prepare for the data integration, we truncated the three gene lists to the length of 65, a global value resulting from several runs of the Hall and Schimek (2010) algorithm (parameters were $\nu = 10$, $C = 0.251$, and $\delta = 10$). For the purpose of illustration, the estimation results are given in Figure 1 for a wide range of $\delta$-values. Relevant in practice are only those results obtained for rather small $\delta$'s.

Next, the truncated lists were aggregated by means of the CEMC algorithm as put forward in Lin and Ding (2009). For optimization, Kendall’s $\tau$-distance was applied. The obtained results from this sophisticated stochastic approach were finally compared to the two classical ones, Borda and MC. Borda’s method, the oldest known aggregation technique (Borda, (1781)), and
its variants are based on a function of the rank positions across all lists. Typical aggregation functions are the arithmetic mean, the geometric mean, and the median (here the latter was adopted).

The final aggregated lists comprise 40 consolidated genes in new rank orders and are shown in Table 1. When space considerations are taken into account, the differences between the three methods are minor. The majority of identified genes, independent of the aggregation method, are known housekeeping genes associated with different cell functions. The basic requirement of control genes for RT-qPCR is their stable gene expression, defined by a large mean and a small standard deviation. All the applied algorithms selected such genes.

In conclusion, owing to the data-driven selection of partial lists, a robust set of top-ranked genes could be identified with the desired properties. All the described methods are available in the R package TopKLists (Schimek et al., 2011). Simulation evidence and other genomics applications can be found in Schimek, Myšičková and Budinská (2010).

References


Acknowledgement
I wish to thank the members of the core software development team of TopKLists, Eva Budinská, Karl Kugler, and Shili Lin for their input.
Efficient, non-disruptive local moves for Monte Carlo sampling of proteins

Sandro Bottaro *,1, Wouter Boomsma1,2, Kristoffer E. Johansson 3, Christian Andreetta 3, Thomas Hamelryck 3, and Jesper Ferkinghoff-Borg 1

1 DTU Elektro, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark
2 Department of Chemistry, University of Cambridge, Cambridge, UK
3 Bioinformatics Centre, Department of Biology, University of Copenhagen, Ole Maaloes Vej 5, 2200 Copenhagen N, Denmark

Molecular dynamics (MD) is the common choice of simulation technique for studying protein dynamics in its biological active state (native state). A major bottleneck of this approach is that MD relies on time-steps of \( \approx 10^{-15} \) seconds, whereas structural transitions relevant for protein function occur at \( 10^{-3} - 10^0 \) seconds. These transitions are therefore inaccessible to traditional MD-simulations. Based on tools from differential geometry, we have found a novel analytical solution to a classical problem in protein geometry, known as the chain closure problem. This solution allows us to devise a new kinetic algorithm, CRISP - Concerted Rotations Involving Self-consistent Proposals - (Bottaro et al (2003)), which greatly enhances sampling efficiency of protein structures compared to the current state of the art MC methodologies (Ulmschneider J, Jorgensen W (2003), Smith C, Kortemme T (2008)). Studying the native state of ubiquitin, a key protein to many cellular signaling networks, we demonstrate that CRISP reproduces the full structural heterogeneity of the ensemble known from experiments. This includes functionally important conformational transitions, which are inaccessible to existing state-of-the-art simulation techniques.

References


Exploring the energy landscapes of protein folding simulations with Bayesian computation

Nikolas S. Burkoff* and David L. Wild

Systems Biology Centre,
University of Warwick, Coventry, UK

Nested sampling is a technique developed to explore probability distributions localised in an exponentially small area of the parameter space. The algorithm provides both posterior samples and an estimate of the evidence (marginal likelihood) of the model. Previous applications of the algorithm have yielded large efficiency gains over other sampling techniques, including parallel tempering. In this work we apply the nested sampling algorithm to the problem of protein folding in a Gō-type force field of empirical potentials that were designed to stabilize secondary structure elements in room-temperature simulations. A topological analysis of the posterior samples is performed to produce energy landscape charts, which give a high level description of the potential energy surface for the protein folding simulations. These charts provide qualitative insights into both the folding process and the nature of the model and the force field used.

We demonstrate the method by conducting folding simulations on a number of small proteins which are commonly used for testing protein folding procedures: protein G, the SH3 domain of Src tyrosine kinase and chymotripsin inhibitor 2. We compare our results for the protein G to those obtained using parallel tempering with the same model. The topology of the protein molecule emerges as a major determinant of the shape of the energy landscape. The nested sampling algorithm also provides an efficient way to calculate free energies and the expectation value of thermodynamic observables at any temperature, through a simple post-processing of the output.


"Aggregate maps" for exploratory and confirmatory analysis of nuclear architecture

Richard A. Russell\textsuperscript{1,3}, David J. Weston\textsuperscript{*1}, David A. Stephens\textsuperscript{4}
Niall M. Adams\textsuperscript{1} and Paul S. Freemont\textsuperscript{2}

\textsuperscript{1}Department of Mathematics, \textsuperscript{2}Centre for Structural Biology, Imperial College, South Kensington, London, United Kingdom.
\textsuperscript{3}NIHR Biomedical Research Centre for Ophthalmology, Moorfields Eye Hospital NHS, Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom.
\textsuperscript{4}Department of Mathematics and Statistics, McGill University, Canada.

1 Introduction

Learning about object configuration in the study of nuclear architecture is a challenging problem. Typical approaches to this problem use techniques, such as radial analysis, on individual images (Shiels et al. (2007)). This approach is vulnerable to low statistical power in cases where the object of interest occurs infrequently in single images. Furthermore it is not straightforward to interpret results from a collection of cells, so analysed, especially where there is significant variation between those cells. It is to address these issues that we propose ‘Aggregate Maps’ (Russell et al. (2011)).

An aggregate map for a collection of cells is constructed by fusing the images of individual cells. The approach, which uses standard methods for image registration, is as follows. First we extract regularly spaced landmarks from the nuclear boundary of each cell. We take advantage of the ovoid-like shape of the nucleus to put landmarks for each pair of cells into correspondence, from which we can construct the Procrustes mean shape for the nucleus boundary of the cell collection. An aggregate map for objects of interest is produced by mapping the objects from each cell onto the mean shape using a thin plate spline.

We demonstrate the effectiveness and robustness of our approach by constructing a simulator to produce synthetic instances of a cell nucleus shape that contains objects following a known spatial process. We show under what conditions we can reconstruct the spatial pattern. Real data, consisting of PML nuclear bodies in mammalian fibroblasts in a number of experimental conditions, is also analysed.

References


Estimating soil microbial species abundance through metagenomic sequencing - journey into the “unknown-ome”.

Wally Gilks⁺,¹, Elisa Loza², Michael Defoin-Platel², Ian Clark² and Penny Hirsch²

¹ Department of Statistics, University of Leeds
² Rothamsted Research, Harpenden

Soil microbial species are mostly unknown as they are generally unculturable. Their genomes may therefore be collectively called the 'unknown-ome'. One way to learn about soil microbes is through metagenomic sequencing of soil samples. This produces vast numbers of short sequence fragments, randomly sampled from the indwelling microbial community. These fragments are too short for reliable genome assembly, there being so many unknown species and so few reference genomes, mostly of distantly related non-soil species. We would like to estimate microbial species abundance in these samples — a formidable challenge since we do not know what species exist or how to recognise them. We describe an approach to this problem which combines statistical sampling theory with phylogenetic inference.
Posters
An evolutionary algorithm for stochastic context-free grammar design, with applications to RNA secondary structure prediction.

James W. J. Anderson1, Joe Staines2, Paula Tataru3, Jotun Hein1 and Rune Lyngsø1

1Department of Statistics, University of Oxford, 
2Department of Computer Science, University College London, 
3Bioinformatics Research Center, Aarhus University.

1 Introduction

Stochastic Context-Free Grammars (SCFGs) have been used widely in modelling RNA secondary structure. They were motivated by the use of Hidden Markov Models (HMMs) in protein modelling (Krogh et al., (1993)). What was lacking in HMMs though, was the ability to model long range interactions which are necessary to provide an effective model for RNA secondary structure. Thus, SCFGs, as generalisations of HMMs which can model long range interactions, were used to model RNA secondary structure (Sakakibara et al. (1994), Lefebvre (1996)).

Whilst much effort has been put into extending the SCFG model to describe features such as pseudoknots (Brown (1996)) or evolutionary history (Knudsen & Hein, (1999)), considerably less has been put into the SCFG design itself. This was initially considered by Dowell and Eddy (2004), where nine SCFGs were tested for their ability to predict RNA secondary structure. These were all hand-designed SCFGs (including the SCFG from Knudsen and Hein (1999), denoted hereafter KH99, which performed the best out of the nine), and it is the approach used in this paper that motivated two main questions addressed here.

Firstly, are there any better SCFGs than the ones used in Dowell and Eddy (2004)? There is little reason to suggest that there are not stronger SCFGs out there, and finding these would improve the quality of RNA secondary structure predictions. Similarly, if it can be shown that the method of automatic SCFG design is effective for this problem, other bioinformatics applications which use SCFGs will benefit.

The second question concerns SCFG ambiguity. A SCFG is said to be ambiguous if there exists more than one derivation for any string (Reeder et al. (2005)). Ambiguity is generally considered undesirable as it allows suboptimal predictions. Dowell and Eddy (2004) tested two SCFGs which were ambiguous, and these performed poorly in the evaluation. However, it is quite possible that the performance of these two ambiguous SCFGs was due to SCFG design and not to ambiguity. Therefore, there might be ambiguous SCFGs which perform well.

2 Methods

It was decided to use an evolutionary algorithm to explore the space of SCFGs. The size of the space of SCFGs with $n$ non-terminal variables is $O(2^n^3)$, so intelligent searching methods are necessary. An evolutionary algorithm also allows features of SCFGs which do not have any immediate use to maintain and potentially be of use later.

Firstly, a normal form must be decided upon for the SCFG to enable training and testing algorithms to be implemented. Whilst Chomsky Normal Form (Chomsky (1959)) might be a
desirable form, it was felt that a new normal form might better relate to RNA secondary structure whilst still encapsulating all SCFGs. The following ‘double emission normal form’ was used, allowing the following rule types:

\[ T \rightarrow UV \]
\[ T \rightarrow . \]
\[ T \rightarrow (U) \]

Here ‘.’ represents an unpaired nucleotide, and corresponding parenthesis paired nucleotides.

The initial population was then taken to be a selection of SCFGs with two non-terminal variables, which could then increase in functionality through breeding and mutation. The following SCFGs were used:

\[ S \rightarrow \{ SS \mid SZ \mid ZS \mid ZZ \mid (S) \} \]
\[ Z \rightarrow . \]

with probabilities obtained through maximum likelihood estimation. Breeding and mutations were then designed to allow the SCFG to develop additional production rules and non-terminal variables.

The fitness function contained many features which would reward SCFGs for predicting RNA secondary structure effectively. In particular, sensitivity and positive predictive value of predictions, mountain metric of predictions (Moulton et al. (2000)), and the difference and ratio of probabilities of true structure and predicted structure were all considered. Small complexity penalties were also added to SCFGs with a large number of production rules or non-terminal variables, as these were heuristically poor.

Data was taken from RNAStrand (Andronescu et al. (2008)), giving a filtered data set which contained 339 sequences. This was split up into three subsets, a training set (for SCFG parameter inference), a test set (to evaluate the prediction performance within the evolutionary algorithm), and an evaluation set (to evaluate the SCFGs produced by the evolutionary algorithm). The data set used in Dowell and Eddy (2004) was also used to benchmark the grammars produced in the evolutionary search. The search was also augmented by two other searches, one a brute force search of SCFGs with at most two non-terminal variables, and the other a local search of KH99 to determine whether slight modifications of KH99 would improve prediction quality.

3 Results

In the search many strong SCFGs were found. Seven SCFGs are particularly strong (see Table 1), although many were presented in the search. In this sense the space was found to have many local optima, in that the search would very often converge to a strong SCFG. Methods were taken to ensure that the search heuristic was a strong one, and this was seen in how the SCFGs were at the level of the best of the hand-designed SCFGs, KH99. The benchmarking on the Dowell and Eddy (2004) data showed that the SCFGs were strong predictors of RNA secondary structure, and were not overtrained. The results from the brute force and local search can be seen in Figure 1. These support the idea that intelligent searching methods are required by examining the landscape of the space of SCFGs. For small SCFGs, they lack the ability to generate complicated structure. When moving around the space, many movements that might be made have little or no change in the SCFG prediction quality, seen in the local search of KH99. Heuristically, the landscape has lots of ‘cliffs’.
Table 1: The sensitivities, PPV, and ambiguity of grammars GG1-GG7 and KH99′.

<table>
<thead>
<tr>
<th>Grammar</th>
<th>Sensitivity</th>
<th>PPV</th>
<th>Ambiguity</th>
<th>Completeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>KH99′</td>
<td>0.496</td>
<td>0.479</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>GG1</td>
<td>0.503</td>
<td>0.480</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>GG2</td>
<td>0.505</td>
<td>0.481</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>GG3</td>
<td>0.474</td>
<td>0.421</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>GG4</td>
<td>0.474</td>
<td>0.454</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>GG5</td>
<td>0.469</td>
<td>0.467</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>GG6</td>
<td>0.486</td>
<td>0.468</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>GG7</td>
<td>0.526</td>
<td>0.479</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Figure 1: Results of the brute force search (left) and local search of KH99 (right). For the brute force search, 16384 SCFGs were searched, and none had sensitivity stronger than 0.25, demonstrating how larger SCFGs with more structural features are needed. The local search tested 32 SCFGs with a single rule added to KH99, and 496 with two rules added. The results are extremely concentrated around the original sensitivity, many having exactly the same predictive ability.

4 Conclusions

In this investigation the space of SCFGs was searched using an evolutionary algorithm to look for stronger grammars for RNA secondary structure prediction. Furthermore, whether or not the SCFGs were ambiguous was considered, to determine whether the practical implications of SCFG ambiguity were as bad as has been suggested.

Many strong SCFGs were found but, given the sensitivity to the data, it is hard to say which is the best performer of the seven candidates. The SCFGs found were consistent performers on the Dowell and Eddy (2004) data set but their ordering was not. Despite this, it is clear that SCFG design is effective with this method.

Of interest was that all of the SCFGs found were ambiguous. This demonstrates how ambiguity can affect some SCFGs in greater ways than others. This suggests the need for a measure of ambiguity, which can perhaps give some idea of how much the predictions will suffer. However, in SCFG design, it is clear that one might not spend the huge time and effort on ensuring
unambiguity.

References


Statistical spectroscopic methods in drug metabolism and metabolic profiling

Toby Athersuch

MRC-HPA Centre for Environment & Health, Imperial College London, UK

Statistical methods for helping characterise and make sense of spectroscopic data are of growing importance in biological research. The quantity of data describing biological objects and systems produced on a daily basis continues to increase, going hand-in-hand with developments in high-throughput techniques, assay multiplexing, miniaturisation using nanofluidic devices, and, more commonly, with improvements in instrumental resolution and sensitivity. Trying to organize and interrogate this data to extract the information contained within is therefore a task of increasing complexity, and there is a need to develop methods that efficiently combine data for rapid analysis to help understand and interpret chemical and biochemical processes. Thus, the use of efficient statistical techniques to integrate and directly investigate spectroscopic data has gained prominence in recent years.

A number of statistical techniques have been developed from the basis of methods for generalized two-dimensional correlation spectroscopy, summarised in (1). In metabonomics, the most commonly used have been statistical total correlation spectroscopy (STOCSY), statistical heterospectroscopy (SHY), and related techniques for the interpretation of spectroscopic data, which has been demonstrated previously (2-11), and can be applied with relative ease to datasets containing a sufficient number of high quality spectra.

As described in Cloarec et al. (2005) (2), central to the STOCSY approach is the calculation of the correlation matrix \( C \) that can be obtained from \( n \) set(s) of spectra with \( v \) spectral variables:

\[
C = \frac{1}{n} - 1 \times \frac{1}{n} \times (1/n - 1)X_1^tX_2
\]

where \( X_1 \) and \( X_2 \) correspond to \( n \times v_1 \) and \( n \times v_2 \) matrices respectively, containing the spectral intensity data.

If the spectral sets are identical, an autocorrelation matrix is obtained, and indicates the internal correlation structure of the spectral variables. Where the \( X_1 \) and \( X_2 \) matrices describe different spectral data for the sample samples, the correlation matrix will reflect correlations between the responses in the different analytical platforms/experimental measurements. Therefore, the matrix \( C \) will describe both inter- and intra-metabolite correlations where present, and can provide information on molecular structure and connectivity, as well as biochemical relationships. The strength of the correlation will rely upon the inherent detection properties of the technique, collinearity of signals arising from the same molecule, and be influenced by spectral overlap and noise interferences. Other than on the diagonal, the correlation coefficient in such a matrix (where \( X_1 = X_2 \)) will be \( \leq 1 \).

The two most commonly-used spectroscopic platforms used in metabolism and metabonomics studies are high-resolution nuclear magnetic resonance spectroscopy (NMR) and mass spectrometry (MS). NMR is an inherently quantitative technique and in the context of metabonomics, NMR was used as the exemplar technique to show how statistical correlations between variables can aid spectral interpretation to great effect.

Electrospray ionization mode MS is also commonly used in metabonomics. However, the relative response of different analytes (or fragments generated in the analysis) is prone to modulation by numerous factors (relating to both the instrument and to the sample matrix), and
therefore additional care must be taken to ensure that correlations revealed using statistical methods are meaningful.

Statistical techniques can help augment these existing physical analytical techniques by highlighting correlations between spectra, either in similar samples, or from the same analytical run. In some ways, statistical spectroscopy can be considered as a mathematical equivalent of the editing experiments that are used in traditional spectroscopy. NMR spectroscopists commonly implement pulse sequences to edit spectra such that they contain information pertaining to a particular factor of interest such as molecular connectivity or rate of diffusion. Analogously, mass spectrometrists may employ editing techniques (e.g. MSn experiments) in their analysis to select particular species of interest for further investigation. It has been shown that the correlation maps that are generated from $^1$H NMR data can be edited by only showing those areas that also have a correlation to another variable. This was described for heteronuclear correlations between parallel $^1$H and $^{19}$F NMR spectra in the HET-STOCSY approach, an equivalent of the HETCOR NMR experiment (6). Additionally, editing of the $^1$H-$^1$H autocorrelation matrix using correlations to the $^{19}$F nucleus provided a statistical equivalent of the 3D-HSQC-TOCSY NMR experiment (X-STOCSY) (6).

The STOCSY approach can aid visualization for human interpretation of spectral data, and this has been exploited to help infer biochemical pathway connectivity (5), explore reaction kinetics (6), and to enable clustering of spectral variables to reduce complexity (7) Statistical recoupling of variables (SRV) has been used in combination, to give rise to R-STOCSY (8) and the ranking of the correlations between the various spectral variables exploited to allow iterative, less complicated analysis to be performed using I-STOCSY (9).

Integration of spectral data from different platforms — coined statistical hetereospectroscopy (SHY) — using statistical methods has been demonstrated. Typically, profiling of biofluids using MS is conducted in hyphenation with a separation technique to improve resolution, sensitivity and reduce spectral distortions (e.g. ion suppression) that occur with the analysis of multiple analytes simultaneously. Hyphenation with chromatography presents additional opportunities to implement statistical methods for more efficient analysis, through the separation of spectral components based on minor differences in elution time, despite these being overlapped (virtual chromatographic resolution enhancement) (10).

In this contribution, the methodology used for applying these statistical tools will be outlined, with the factors that have been examined to date (including molecular connectivity, biochemical pathway associations, reaction kinetics, etc) described along with the necessary experimental designs shown. The application of these tools in the context of metabolite identification, metabolic profiling, and deconvolution of metabolic pathways, will be described. Specifically, examples relating to the utility of these approaches in for efficient data integration in drug metabolism studies will be given.

The array of these tools is rapidly expanding, and in this contribution, the strengths, weaknesses and potential applications of these methods are presented. The scope for further development in this area will be suggested, with reference to improved pathway analysis tools, and integration of spectroscopic data with that generated by other high-throughput, information rich platforms used for systems biology research. One important point is that such techniques can be applied to datasets where the samples themselves are no longer available, and therefore physical separation is not possible. Therefore, these statistical tools may be useful in enhancing analyses of legacy data.

It is hoped that this contribution will provide an opportunity for other applications of statistical spectroscopy to be discussed as the basis of future chemometrics-based collaborations between bioscientists and statisticians.
References


Testing for equality of distributions for dihedral angles in amino acids

Maha Bakoben and Charles Taylor

Department of Statistics, University of Leeds

Abstract

This study project focuses on the goal of testing for equality of bivariate distributions of backbone dihedral angles \( (\phi, \psi) \) in amino acids. The assumption of equal bivariate distributions is assessed by the Energy test which is based on a measure of distance between observations and implemented on the pooled sample. The study supports predictive modelling and determination of protein structure.

Introduction

Amino acids are molecules that are the main composition of peptides and proteins. Peptides are short chains of two or more amino acids; while proteins or polypeptides are long chains of hundreds of amino acids (J. Stenesh, 1998). Mutations in the amino acid sequence result in production of variant types of these chains. Twenty types of amino acids were identified as the common structural components of proteins (J. Stenesh, 1998). The main elements of amino acids are Carbon, Hydrogen, Oxygen and Nitrogen. These elements are linked together to form backbone atoms of proteins. The backbone conformation can be specified by the dihedral angles \( \phi \) and \( \psi \). The dihedral angle \( \phi \) connect the \( C_\alpha - N \) bond to the adjacent bond, and \( \psi \) connect the \( C_\alpha - C \) bond to the adjacent bond (Voet et al., 2011).

Predictions of proteins from secondary structures have been extensively examined in research studies using variant statistical methods (Chou and Fasman, 1974; Lim, 1974; Stolorz et al., 1991; Chandonia and Karplus, 1995; Nageswara Rao et al., 2010). Protein models were built on the assumption of single backbone covalent geometry (Berkholz et al., 2011). However, Berkholz et al. (2011) and other studies invalidate this assumption. Berkholz et al. (2011) conducted analysis on the covalent lengths and bond angles of peptides to visualize the \( \phi, \psi \) - dependent variations using kernel regression approach.

Considering the distribution of dihedral angles \( (\phi, \psi) \) for Arginine and Tryptophan in Figure 1, it seems that the bivariate distributions in the two amino acids are reasonably similar. However, a more accurate test of comparing multivariate distributions is applied in this study. We aim to confirm the independent bivariate distribution of the peptide backbone dihedral angles \( (\phi, \psi) \) using a non-parametric test of independent known as energy test. We will begin with an insight on the procedure of energy test and proceed with its application in comparing the dihedral distributions of \( \phi \) and \( \psi \).

Method

The analysis was implemented on protein data set which consists of 73,136 observations collected by X-ray crystallography. This is a "cleaned-up" subset of the top 500 proteins from the Kinemage database. The dihedral angles \( \phi \) and \( \psi \) were measured in degrees and for the purpose of analysis they have been converted to radians.

The method used is the energy test for equality of distribution. It was firstly proposed by Szekely et al. (2004), it is a new non-parametric test for equal distribution for two or more multivariate distributions. Energy test is a simulation-based version, which can be used for multivariate data. The test has the power that it can be applied when the underlying distribu-
Figure 1: The bivariate distribution of the backbone dihedral angles $\phi$, $\psi$ for the amino acids Arginine (R) and Tryptophan (W)

...tions are unknown (Szekey,2004). Although it is very useful for high dimensional data, we are concerned in the two-sample test for comparing the bivariate distribution of the dihedral angles for each pair of amino acids.

Suppose $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ are independent random vectors of the dihedral angles $(\phi, \psi)$, for amino acids $A_1$ and $A_2$. An appropriate measure of distance between two bivariate angles $X_1$ and $X_2$ was quantified as follow

$$\| X_1 - X_2 \| = \sqrt{(1 - \cos(\phi_1 - \phi_2))^2 + (1 - \cos(\psi_1 - \psi_2))^2}$$

For each pair of amino acids $(A_i, A_j)$; $i \neq j$, with sample sizes $n_i$ and $n_j$ respectively. We test the null hypothesis of equal bivariate distributions for dihedral angles $(\phi, \psi)$ between $A_i$ and $A_j$. $H_0 : F_i(\phi, \psi) = F_j(\phi, \psi)$ vs $H_1 : F_i(\phi, \psi) \neq F_j(\phi, \psi)$

The distance between the two sample distributions is measured by the e-statistic:

$$\varepsilon_{A_1,A_2} = \frac{n_1n_2}{n_1 + n_2} \left( \frac{2}{n_1n_2} \sum_{i=1}^{n_1} \sum_{m=1}^{n_2} || X_i - Y_m || - \frac{1}{n_1^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} || X_i - X_j || - \frac{1}{n_2^2} \sum_{l=1}^{n_2} \sum_{m=1}^{n_2} || Y_l - Y_m || \right)$$

A positive constant value of $\varepsilon$ is expected when sample size tends to infinity under the null hypothesis of equal distribution, while $\varepsilon$ tends to infinity under the alternative hypothesis (M. Rizzo, 2008). We reject the null hypothesis of equal distribution if $\varepsilon > c_\alpha$. Where $\alpha$ is chosen between $(0, 1)$, and $c_\alpha$ is constant such that $\lim_{n \to \infty} p(\varepsilon_n > c_\alpha) = \alpha$.

The procedure of the energy test is based on the bootstrap approach on the pooled sample that have the same properties of the data to obtain a distribution free test procedure (Szekly,2004).
Results

Due to large sample sizes (e.g., $n_i = 8000$), we selected random samples of size 1000 from each type of amino acids. The results of energy test suggest that some amino acids could have different bivariate distributions of dihedral angles. However, it has been found that 27 pairs out of 190 of amino acids have similar distributions of the dihedral angles, which in some cases are more similar than others. The $\varepsilon_{nl}$ test is highly significant for differences in the bivariate distributions of the dihedral angles of amino acids Asparagine, Aspartic, Glycine, Proline and Valine from other types of amino acids. While the test is not significant at $\alpha = 0.05$ when bivariate distributions of Cysteine and Phenylalanine are compared with some set of amino acids. Table 1 gives the two-sample test results showing the degree of similarity in the bivariate distribution of the dihedral angles the 20 amino acids. A small value of e-statistic is indication of similar distributions.

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>ALA</th>
<th>ARG</th>
<th>ASN</th>
<th>ASP</th>
<th>CYS</th>
<th>GLN</th>
<th>GLU</th>
<th>GLY</th>
<th>HIS</th>
<th>ILE</th>
<th>LEU</th>
<th>LYS</th>
<th>MET</th>
<th>PHE</th>
<th>PRO</th>
<th>SER</th>
<th>THR</th>
<th>TRY</th>
<th>TYR</th>
<th>VAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALA</td>
<td>0.0</td>
<td>8.7</td>
<td>38.2</td>
<td>13.9</td>
<td>79.2</td>
<td>1.5</td>
<td>0.3</td>
<td>330.3</td>
<td>42.2</td>
<td>86.6</td>
<td>18.2</td>
<td>13.2</td>
<td>14.6</td>
<td>66.5</td>
<td>98.6</td>
<td>35.0</td>
<td>62.2</td>
<td>22.4</td>
<td>71.3</td>
<td>122.6</td>
</tr>
<tr>
<td>ARG</td>
<td>8.7</td>
<td>0.0</td>
<td>22.1</td>
<td>11.1</td>
<td>35.4</td>
<td>4.9</td>
<td>8.1</td>
<td>301.7</td>
<td>12.6</td>
<td>41.3</td>
<td>2.9</td>
<td>0.5</td>
<td>0.9</td>
<td>27.2</td>
<td>77.6</td>
<td>8.7</td>
<td>25.2</td>
<td>3.2</td>
<td>30.2</td>
<td>66.7</td>
</tr>
<tr>
<td>ASN</td>
<td>38.2</td>
<td>22.1</td>
<td>0.0</td>
<td>11.0</td>
<td>42.5</td>
<td>26.1</td>
<td>36.7</td>
<td>225.7</td>
<td>19.7</td>
<td>52.4</td>
<td>21.7</td>
<td>17.3</td>
<td>20.9</td>
<td>41.7</td>
<td>84.1</td>
<td>26.8</td>
<td>49.1</td>
<td>23.8</td>
<td>40.7</td>
<td>78.3</td>
</tr>
<tr>
<td>ASP</td>
<td>13.9</td>
<td>11.1</td>
<td>11.0</td>
<td>0.0</td>
<td>59.8</td>
<td>6.8</td>
<td>12.3</td>
<td>308.3</td>
<td>27.2</td>
<td>66.9</td>
<td>12.6</td>
<td>10.6</td>
<td>13.5</td>
<td>52.4</td>
<td>83.2</td>
<td>29.5</td>
<td>54.3</td>
<td>20.0</td>
<td>54.1</td>
<td>104.4</td>
</tr>
<tr>
<td>CYS</td>
<td>79.2</td>
<td>35.4</td>
<td>42.5</td>
<td>59.8</td>
<td>0.0</td>
<td>65.6</td>
<td>77.5</td>
<td>288.2</td>
<td>8.3</td>
<td>1.3</td>
<td>24.5</td>
<td>28.5</td>
<td>27.2</td>
<td>11.0</td>
<td>74.9</td>
<td>9.5</td>
<td>4.5</td>
<td>17.9</td>
<td>5.9</td>
<td></td>
</tr>
<tr>
<td>GLN</td>
<td>1.5</td>
<td>4.9</td>
<td>26.1</td>
<td>6.8</td>
<td>65.6</td>
<td>0.0</td>
<td>1.0</td>
<td>318.7</td>
<td>12.6</td>
<td>41.3</td>
<td>2.9</td>
<td>0.5</td>
<td>0.9</td>
<td>27.2</td>
<td>77.6</td>
<td>8.7</td>
<td>25.2</td>
<td>3.2</td>
<td>30.2</td>
<td>66.7</td>
</tr>
<tr>
<td>GLU</td>
<td>0.3</td>
<td>8.7</td>
<td>36.7</td>
<td>12.3</td>
<td>77.5</td>
<td>1.0</td>
<td>0.0</td>
<td>334.7</td>
<td>39.7</td>
<td>86.2</td>
<td>17.7</td>
<td>12.3</td>
<td>13.6</td>
<td>64.9</td>
<td>103.6</td>
<td>33.4</td>
<td>59.9</td>
<td>21.4</td>
<td>69.5</td>
<td>121.4</td>
</tr>
<tr>
<td>GLY</td>
<td>330.3</td>
<td>301.7</td>
<td>225.7</td>
<td>308.3</td>
<td>288.1</td>
<td>318.7</td>
<td>334.7</td>
<td>0.0</td>
<td>268.8</td>
<td>331.3</td>
<td>330.5</td>
<td>293.5</td>
<td>302.1</td>
<td>302.3</td>
<td>389.1</td>
<td>279.5</td>
<td>317.0</td>
<td>297.5</td>
<td>293.5</td>
<td>341.0</td>
</tr>
<tr>
<td>HIS</td>
<td>42.2</td>
<td>12.6</td>
<td>19.7</td>
<td>27.2</td>
<td>8.3</td>
<td>30.7</td>
<td>397.7</td>
<td>268.8</td>
<td>0.0</td>
<td>13.3</td>
<td>8.6</td>
<td>3.8</td>
<td>5.6</td>
<td>81.9</td>
<td>1.5</td>
<td>6.3</td>
<td>4.3</td>
<td>5.8</td>
<td>26.9</td>
<td></td>
</tr>
<tr>
<td>ILE</td>
<td>86.6</td>
<td>41.3</td>
<td>52.4</td>
<td>66.9</td>
<td>1.3</td>
<td>73.8</td>
<td>86.2</td>
<td>331.3</td>
<td>13.3</td>
<td>0.0</td>
<td>27.8</td>
<td>34.6</td>
<td>31.9</td>
<td>2.4</td>
<td>70.0</td>
<td>12.8</td>
<td>6.4</td>
<td>22.2</td>
<td>1.5</td>
<td>3.8</td>
</tr>
<tr>
<td>LEU</td>
<td>18.2</td>
<td>2.9</td>
<td>21.7</td>
<td>12.6</td>
<td>24.5</td>
<td>12.3</td>
<td>17.7</td>
<td>330.5</td>
<td>8.6</td>
<td>27.8</td>
<td>0.0</td>
<td>2.1</td>
<td>2.2</td>
<td>18.0</td>
<td>59.5</td>
<td>5.7</td>
<td>17.6</td>
<td>2.4</td>
<td>20.2</td>
<td>50.5</td>
</tr>
<tr>
<td>LYS</td>
<td>13.2</td>
<td>0.5</td>
<td>17.3</td>
<td>10.6</td>
<td>28.5</td>
<td>7.7</td>
<td>12.3</td>
<td>293.5</td>
<td>8.3</td>
<td>34.6</td>
<td>2.1</td>
<td>0.0</td>
<td>0.4</td>
<td>21.7</td>
<td>74.1</td>
<td>5.6</td>
<td>20.6</td>
<td>1.9</td>
<td>23.9</td>
<td>58.0</td>
</tr>
<tr>
<td>MET</td>
<td>14.6</td>
<td>0.9</td>
<td>20.9</td>
<td>13.5</td>
<td>27.2</td>
<td>9.1</td>
<td>13.6</td>
<td>302.1</td>
<td>7.8</td>
<td>31.9</td>
<td>2.2</td>
<td>0.4</td>
<td>0.0</td>
<td>19.8</td>
<td>79.8</td>
<td>5.1</td>
<td>18.3</td>
<td>1.3</td>
<td>22.0</td>
<td>54.4</td>
</tr>
<tr>
<td>PHE</td>
<td>66.5</td>
<td>27.2</td>
<td>41.7</td>
<td>52.4</td>
<td>1.1</td>
<td>54.8</td>
<td>64.9</td>
<td>302.3</td>
<td>5.6</td>
<td>2.4</td>
<td>18.0</td>
<td>21.7</td>
<td>19.8</td>
<td>0.0</td>
<td>76.6</td>
<td>5.0</td>
<td>1.2</td>
<td>11.8</td>
<td>0.2</td>
<td>9.1</td>
</tr>
<tr>
<td>PRO</td>
<td>98.6</td>
<td>77.6</td>
<td>84.1</td>
<td>83.2</td>
<td>74.9</td>
<td>93.9</td>
<td>103.6</td>
<td>389.1</td>
<td>81.9</td>
<td>70.0</td>
<td>59.5</td>
<td>74.1</td>
<td>79.8</td>
<td>76.6</td>
<td>0.0</td>
<td>73.4</td>
<td>86.6</td>
<td>73.5</td>
<td>79.7</td>
<td>93.8</td>
</tr>
<tr>
<td>SER</td>
<td>35.0</td>
<td>8.7</td>
<td>26.8</td>
<td>29.5</td>
<td>9.4</td>
<td>26.5</td>
<td>334.7</td>
<td>279.5</td>
<td>1.5</td>
<td>12.8</td>
<td>5.7</td>
<td>5.6</td>
<td>5.1</td>
<td>73.4</td>
<td>0.0</td>
<td>5.1</td>
<td>1.4</td>
<td>6.8</td>
<td>27.0</td>
<td></td>
</tr>
<tr>
<td>THR</td>
<td>62.2</td>
<td>25.2</td>
<td>49.1</td>
<td>54.3</td>
<td>4.5</td>
<td>51.7</td>
<td>59.9</td>
<td>317.0</td>
<td>6.3</td>
<td>6.4</td>
<td>17.6</td>
<td>20.6</td>
<td>18.3</td>
<td>1.2</td>
<td>86.6</td>
<td>5.1</td>
<td>0.0</td>
<td>10.5</td>
<td>2.5</td>
<td>13.2</td>
</tr>
<tr>
<td>TYR</td>
<td>22.4</td>
<td>3.2</td>
<td>23.8</td>
<td>20.0</td>
<td>17.9</td>
<td>16.0</td>
<td>214.0</td>
<td>295.7</td>
<td>4.3</td>
<td>22.2</td>
<td>2.4</td>
<td>1.9</td>
<td>1.3</td>
<td>11.8</td>
<td>73.5</td>
<td>1.4</td>
<td>10.5</td>
<td>0.0</td>
<td>13.9</td>
<td>41.0</td>
</tr>
<tr>
<td>VAL</td>
<td>71.3</td>
<td>30.2</td>
<td>40.7</td>
<td>54.1</td>
<td>0.5</td>
<td>58.4</td>
<td>69.5</td>
<td>293.5</td>
<td>5.8</td>
<td>1.5</td>
<td>20.2</td>
<td>23.9</td>
<td>22.0</td>
<td>0.2</td>
<td>79.7</td>
<td>6.8</td>
<td>2.5</td>
<td>13.9</td>
<td>0.0</td>
<td>7.6</td>
</tr>
</tbody>
</table>

Table 1: Two sample e-statistics for 20 amino acids

The degree of similarity of the bivariate distributions for $\phi$ and $\psi$ between amino acids is represented in the multidimensional scaling plot (Figure 2). The non-metric MDS suggests three groups of clusters of amino acids have similar bivariate distributions; first group consists of GLN, GLU and ALA. The second group includes MET, LYS, ARG, TRP and SER. The third group consists of CYS, TYR, PHE and ILE. Whereas, the degree of similarity between HIS, LER and THR and the three clusters is comparatively lower.

Conclusion

Our results are compatible with the assumption of variant backbone covalent geometry as a function of $\phi$ and $\psi$ in protein structure. The degrees of similarities between dihedral angles distributions in amino acids were precisely quantified by e-statistics. This study would help in discriminate between amino acids. It would suggest further studies on identifying the underly-
Figure 2: Multidimensional Scaling plot for bivariate distribution of the backbone dihedral angles $\phi$ and $\psi$ in amino acids, based on e-statistic. Dashed line connects amino acids that are significantly different at $\alpha = 0.05$. Solid lines connects amino acids that are significantly different at $\alpha = 0.01$.

ing bivariate distribution of the backbone dihedral angles in amino acids, kernel estimator might be an appropriate method to use.

References


Mixture models for spherical data with applications to protein bioinformatics

Philippa Burdett*, Kanti Mardia and Stuart Barber

Department of Statistics, University of Leeds

1 Introduction

For over forty years it has been known that a protein spontaneously folds into a compact structure determined by its amino acid sequence; hydrogen bonds play a key role in folding a protein into its 3-D structure. We focus on a probabilistic approach which seeks to attack protein structure prediction using statistical shape analysis by defining the hydrogen bond as four atomic coordinates in 3-D space in a representation equivalent to that of Kortemme et al (2003).

2 Shape Analysis

Data on 11653 hydrogen bonds between secondary structures in proteins have been made available by our collaborating group at Copenhagen University. The data consist of four distances, three angles, and two categorical variables. Driven by these measurements and previous work by Paulsen (2009) a new ‘Euclidean-Latitude’ representation of the hydrogen bond is introduced. The new model reduces the original 12 variables to 4 shape variables, incorporating 3 spherical coordinates and a co-latitude.

We show that the joint distribution of the spherical coordinates form a ‘shell’ on the surface of the sphere with the depth of the shell determined by the length of the hydrogen bond. This distribution is dependent on the secondary structures and the separation distance between the interacting pair. We propose a flexible model for the distribution of hydrogen bonds by conditioning on the hydrogen bond length and using the EM algorithm to fit a mixture of Fisher-Bingham 5 parameter distributions using the full MLEs of the parameters. This is an extension of the method of Peel et al (2001) which used the moments estimators of the parameters given in Kent (1982).

References


My poster looks at continuum (lattice-free) and inherently stochastic models of immune cellular interactions, using the simplest hypothesis, that cells follow Brownian paths. The timescale for a cell to explore a volume such as a lymph node is $L^2/D$, where $L$ is the radius of the region and $D$ the diffusivity of a cell. The average time a cell spends in a volume with an exit is proportional to $L^3/aD$, where $a$ is the radius of the exit. The mean time before a cell encounters a zone of attraction with radius $b$ around, for example, an antigen presenting cell, is proportional to $L^3/bD$.

The poster also explores a stochastic model of interactions between effector T cells and Regulatory T cells, which play a key role in controlling autoimmunity.
ABC for coronary heart disease policy modelling

Nathan Green

North West Institute for BioHealth Informatics, University of Manchester

1 Introduction

Common chronic diseases, such as cardiovascular diseases, are a major public health problem and consume large amounts of scarce healthcare resources (Scarborough et al., (2010)). In order to use limited resources to best effect, there is a need for robust evidence-based decision making when considering different medical and public health approaches, especially when transparency and accountability are of central importance. Thus, we consider how commissioners and providers might explore the potential impacts of different policy options via a flexible and accessible decision-support tool. In particular, we shall present a coronary heart disease (CHD) simulation model.

Previously, we have used discrete event simulation and survival analysis methods, combined with a novel application of simulated annealing. However, this approach does not enable the manipulation of the uncertainty that is present in the model, data and expert opinion. An appreciation of this uncertainty is important and can be explicitly accounted for with a Bayesian formulation. Unfortunately, in particularly complex environments, such as this, it is hard to write down an analytical form of the likelihood function that underlies Bayesian inference.

A recent advance in this field called Approximate Bayesian computation (ABC), e.g. Sisson et al., (2007), addresses this difficulty by enabling one to proceed without analytically specifying or evaluating the likelihood distribution. This is achieved through the use of computer simulation models that stochastically simulate measurements for a given set of parameter values. The incorporation of the various types of data can be performed in a more formal way, providing greater confidence to the conclusions and ultimately, improved decision making.

2 Previous Model

The model consists of a specified directed graph with unique hazard functions on each edge. Denote a hazard function by \( h(t) \). The formal definition for \( h(t) \) is a limiting probability,

\[
h(t) = \lim_{\delta \downarrow 0} \frac{\Pr[t \leq T \leq t + \delta | T \geq t]}{\delta}.
\]

The hazard functions are used to determine the transitions between states for each individual through time, in the presence of competing risks. The collections of competing hazards, indexed by integers 1 to \( K \) have corresponding latent failure times \( (T_1, T_2, \ldots, T_K) \). The first failure occurs at time \( T^* = \min\{T_i : i = 1, 2, \ldots, K\} \) – this corresponds to an individual leaving the state they are currently in to a neighbouring state. Statistics are produced from the literature to describe stepwise discrete distributions, which we shall assume can be transformed to approximate the respective true underlying hazard functions of the model. The aim of the algorithm is to fit a ‘best’ model for forward simulation of patient cases. A simulated annealing algorithm is used to fit the model to the data.

In the spirit of an ABC approach, we replaced the computation of the likelihood with a comparison between the observed data and the simulated data. The data generated from the model to use in place of the likelihood is the hazard function value at the observed data time using some \( \theta \). Analogous to approximating the posterior distribution in a Bayesian context, we estimate a global maximum \( \hat{\theta} \). Thus, it is natural to extend the simulated annealing approach to a Bayesian formulation.
3 Bayesian Formulation

In a Bayesian formulation, we are interested in making inference about the unknown model parameter $\theta$ through the posterior distribution $p(\theta|data)$ using Bayes’ Theorem, $P(\theta|data) = (p(data|\theta)p(\theta))/p(data)$, where $p(\theta)$ is the prior distribution. Usually, the functional form of the likelihood $p(data|\theta)$ is known and may be evaluated directly. In our situations, where the likelihood is computationally intractable, this is avoided by using model simulation in place of likelihood evaluation via an ABC approach. The subsequent model will be formalised in a full Bayesian setting using an ABC-MCMC sampler (Toni et al., 2009).

Table 1: Generic ABC-MCMC step-by-step algorithm description

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Initialise $i = 1$ and $\theta_i$.</td>
</tr>
<tr>
<td>2</td>
<td>Generate a candidate parameter vector $\theta_i^\dagger$, from some proposal distribution $\theta_i^\dagger \sim q(\theta</td>
</tr>
<tr>
<td>3</td>
<td>Simulate a dataset from the model, described by a conditional distribution $g(\phi</td>
</tr>
<tr>
<td>4</td>
<td>Set $\theta_{i+1} = \theta_i^\dagger$, with probability $\alpha = \min\left{1, \frac{p(\theta_i^\dagger, \phi_i^\dagger)}{p(\theta_i, \phi_i)}I(\rho(\phi_i^\dagger, \phi_0) \leq \epsilon)\right}$, otherwise set $\theta_{i+1} = \theta_i$.</td>
</tr>
</tbody>
</table>

In Table 1 Step 4, the evaluation of the likelihood is replaced by $I(\rho(\phi_i^\dagger, \phi_0) \leq \epsilon)$. Generally, the distance function $\rho$ is the Euclidean distance between summary statistics, e.g. means, calculated from the simulated and observed data points. For a competing risks model, using simple mean failure times along each edge will give bias results due to right censoring by the failure times on the competing edges. Approaches to address this include imputing the latent failure times or choosing a distance measure that is robust to censoring. Candidates include, the mean survival time obtained as the area under the Kaplan-Meier estimate of the survival curve, $\hat{\mu} = \int \tilde{S}(t) dt$, where $\tilde{S}(t)$ is the Kaplan-Meier estimator, or alternatively, the Kaplan-Meier estimator of the median.

When the data consist of two sets of right censored failure times, a widely used method in survival analysis for comparison is the Cox proportional hazards model (Cox (1972)). We shall use the following hazard ratio in the distance function,

$$e^{\beta Z_i} h_0 / h_0 = e^{\beta Z_i} = C,$$

where $h_0$ is the observed data hazard and $Z_i$ is the indicator covariate denoting observed or simulated times. The coefficient $\beta$ can be estimated using a standard partial likelihood approach. The distance measure equation in Table 1 is thus replaced by $I(|C - 1| \leq \epsilon)$.

4 Example: A Basic Competing Risks Model

A proof-of-principle model was implemented to demonstrate the methodology. A simple competing risks graph was used, shown in Figure 1. This represents key elements of the more general CHD model in Buchan et al. (2010). The population of individuals may be divided into males and females. Clearly, state A is the only starting state and states B and C are mutually exclusive (absorbing) sink states.

We ran the algorithm for 10,000 iterations and with a 100 iteration thinning period. We considered only males for simplicity. The true parameter values for the underlying mixture hazard functions, defined by $h(t) = \text{Weibull}(k, \lambda)$, were $k_B = 4.4, \lambda_B = 9.4$ and $k_C = 3, \lambda_C = 6.7$. Denote the full set of parameters $\theta = \{k_B, \lambda_B, k_C, \lambda_C\}$. For simplicity, we assume that all individuals have starting age 65. The observed data were produced by running the model
for known parameter values and recording the failure times $T^*$. We used the Cox proportional hazards approach. A proposal scheme was chosen that updates all components at once with a random walk with normally distributed jump sizes. In future, it may be more preferable to block together certain correlated components. Multiple processing cores were exploited to parallelise the software implementation and speed-up the run-time.

The results, represented by marginal posterior distributions for all 4 hyperparameters, are shown in Figure 2. We can see that the posterior modes are close to the true values in all cases. Note that the posterior distributions appear to be positively skewed.

5 Discussion

In many public health problems the raw data may not exist, but rather a collection of summary statistics from literature. This secondary data should be used with care with an understanding of how these figures have been derived. The demonstration of the ABC method above is flexible enough to be extended to this context with appropriate modifications.

Error exists in the estimation of the model statistics due to the sample size used. The smaller the number of the runs, the larger the variance on the estimate. There is error both on the observed data and the simulated output. Since the synthetic data is only calculated once, we can afford to use a longer run. However, the ABC summary estimates need to be calculated at each iteration and so there is a trade-off between the accuracy of the estimates and the run time.

There is also a trade-off between providing enough information to the model, to be able to identify a sufficiently small family of models to make forward simulation useful, and over-
specifying the model with conflicting information such that no model fit will satisfy this data. The idea of an envelope constraint is of interest to epidemiologists. That is, a value that is an upper or lower bound on what is expected to be observed rather than a more precise point value.

6 Conclusions

Previously, the model employed a simulated annealing algorithm to fit the model to the available data. However, in this context, the uncertainty due to the model fit is not represented and as such there is no indication of how good a given model is. Further, the model outputs used in prediction will have no measure of uncertainty about there accuracy to real life. Any policy decisions would be made assuming that the model outputs are correct with certainty- a potentially dangerous assumption.

7 Future Work

An ABC approach brings with it extra computational cost. There are a number of approaches that may address this issue.

Even the simple example given above, including males and females separately, requires 16 parameters and for a much larger model this number may become unwieldy. When the necessary computation power is not available, some form of parameter reduction is advisable. One approach is to fix some of the parameter values in advance- appropriate when values are known a priori to a sufficiently degree of certainty or the model outputs are insensitive to their value. Alternatively, hierarchical prior distributions may be used, where the same prior distribution is used for multiple edges. Finally, it may be possible to identify nuisance parameters and simply ignore them in the inference.

Further, we may subdivide the graph and approximate some of the process. That is, divide the network into connected subgraphs and perform a standard ABC. We can take the posterior distributions for the subset of parameter values and approximate these by a Gaussian distribution using some kernel density estimation. These approximated distributions can be used to feed-in to the subsequent subgraph as inputs.

References


Modelling threshold violations of air pollution concentrations using multiple change-points Poisson process

J. Gyarmati-Szabó*,1,2, L.V. Bogachev1 and H. Chen2

1 Department of Statistics, University of Leeds
2 Institute for Transport Studies, University of Leeds

1 Introduction

Due to harmful effects of air pollution on human health and the environment, air quality is a matter of worldwide concern amongst scientists, policy makers and public alike. The UK National Air Quality Strategy has determined statutory standards expressed in terms of threshold violations for eight main air pollutants, including nitrogen dioxide (NO₂), nitrogen monoxide (NO) and carbon monoxide (CO). In order to monitor and assess the efficacy of past and future policies, it is important to develop adequate statistical methods to measure the impact of regulations on the dynamics of various pollutants.

To date, a number of statistical methods have been developed to model violations of the air quality standards. The most common approach is based on extreme value theory. An alternative method is to model threshold exceedances using non-homogeneous Poisson processes (Smith and Shively, 1995; Achcar et al., 2010). This approach is motivated by the asymptotic theory of stationary point processes stating that the series of occurrences of high threshold exceedances can be approximated by a Poisson process. For nonstationary data, the approximating Poisson process has to be nonhomogeneous, with the intensity rate depending on time.

Estimation of the unknown rate function is facilitated by choosing certain parametric classes (see Achcar et al., 2010, and references therein). However, such models are not suitable if the rate may change abruptly (e.g., following an updated air quality action plan). To capture this type of variation in the exceedance dynamics, Achcar et al. (2010) built a non-homogeneous Poisson model allowing one to handle up to three change-points using complicated iterative updates of the prior information.

In the present paper, we propose a non-homogeneous Poisson model assuming for simplicity that the rate is a step function but with no prior restriction on the number of change-points. An efficient reversible jump Markov chain Monte Carlo (MCMC) estimator is employed to automatically determine the number and locations of change-points, as well as the step heights of the rate function. These techniques are applied to NO₂, NO and CO data collected in the Leeds city centre as daily concentration maxima in 1993–2009.

2 Model preprocessing

2.1 Thresholding

To set a specific threshold in order to extract the statistics of exceedances, it might seem natural to use the air quality standards; however, this produces rather scarce statistics of the resulting exceedances (e.g., 0.11% for NO₂). Instead, a quantile thresholding has been applied at the 90th empirical percentile, leading to the specific threshold values of 96 µg/m³ (NO₂), 185 µg/m³ (NO) and 2.1 mg/m³ (CO). Cumulative counts of the resulting exceedance data for all three pollutants are shown in Fig. 1.
2.2 Missing values

Our concentration level data contain a noticeable proportion of missing values (6.6%, 5.5% and 9.9% in the NO\textsubscript{2}, NO and CO data, respectively), which considerably decreases the exceedance statistics available and therefore may adversely affect the estimation of the unknown rate function. One can improve estimation by compensating for the omissions. In this work, we employed a two-sided moving average estimator with the window size ±65 days. We first imputed raw concentration values drawn independently according to their estimated (univariate) distribution, and then applied thresholding.

2.3 Deseasonalisation

Varying meteorological conditions lead to yearly oscillations (seasonality) in the concentration data (e.g., due to photochemical reactions between NO\textsubscript{2} and NO driven by the solar radiation). Yearly cycles were confirmed by computing the autocorrelation at lag 365 days and analysing the estimated spectral density. Deseasonalisation may be based on fitting a regression model for the log-transformed data (Cox and Lewis, 1966). Note that possible deviations from stationarity are likely to occur on a scale much slower than the annual variability, hence the regression estimation may be expected to give satisfactory results; indeed, the deseasonalised data had no visible periodic patterns.

2.4 Declusterisation

The time series of threshold exceedances by air pollution concentrations is likely to involve dependence manifested in clustering of consecutive exceedances, which may render Poisson model unsuitable. However, due to a relatively short range of such correlations, the dependence problem can be rectified by a suitable thinning, or “declusterisation”, achieved by retaining one point from each cluster. We deployed the nonparametric Wald–Wolfowitz runs test of independence applied to a two-valued data sequence representing threshold exceedances versus non-exceedances. The runs test accepted independence on all intervals between the estimated change-points, with the corresponding \( p \)-values above 0.10.

3 Results for the air pollution data

Estimation of the rate function \( \lambda(t) \) was carried out using a reversible jump MCMC algorithm adapted from Green (1995). As a result of the MCMC performance, it was concluded that the MCMC sampler spent most of its time in parametric states corresponding to one, two and four change-points for NO\textsubscript{2}, NO and CO, respectively. Their locations and the step heights of the rate function \( \lambda(t) \) were estimated via the modes and medians, respectively, of the posterior marginal densities. The graph of the step function \( \hat{\lambda}(t) \) (with estimated change-points and step heights) is shown in Fig. 1, along with an “integral” estimate \( \bar{\lambda}(t) \) of the posterior mean rate (calculated by averaging 5000 sample step rate functions drawn from the posterior distribution). Transition areas for \( \hat{\lambda}(t) \) near the change-points are typically quite narrow, confirming a stepwise nature of the rate function \( \lambda(t) \). Fig. 1 also shows the cumulative graphs simulated via Poisson processes with rates \( \hat{\lambda}(t) \) and \( \bar{\lambda}(t) \), contrasted with the observed cumulative plot. As could be anticipated, the integrated (cumulative) estimates demonstrate a much better fit to the data.

To validate the employed MCMC sampler, we simulated data sets from non-homogeneous Poisson processes with the estimated step function \( \bar{\lambda}(t) \) as a rate function and again applied MCMC to see if there were any significant differences in the estimates. The results (not shown here) demonstrated an excellent match of the posterior estimates for the simulated data sets with those obtained earlier for the observed data. Furthermore, following the method of posterior predictive simulation (Gilks et al., 1996), we replicated the counting processes and compared them to the corresponding cumulative plots (not shown). Due to random fluctuations, devia-
MCMC results for each pollutant: estimated rate functions (left axes) and the corresponding cumulative plots for data sets simulated via non-homogeneous Poisson process (right axes). Colour-coded plots correspond to the posterior mean rate $\bar{\lambda}(t)$ (red) and a step rate $\hat{\lambda}(t)$ function estimated from the posterior distribution (blue). Black dotted graphs show the cumulative plots of observed threshold exceedances. Dashed vertical lines indicate the locations of estimated change-points.

The accuracy can be improved by simulating Poisson processes conditioned to have the same number of occurrences as in the real data. Conditional replicates (not presented here) showed an excellent fit with the observed plot, being also consistent with the sample mean plot for data sets simulated with a posterior estimate.

4 Conclusion

In this paper, we have fitted a non-homogeneous Poisson model to the observed series of threshold exceedance occurrences in time, extracted from the daily pollution concentration data. The unknown rate function was assumed to be a step function, with an arbitrary number of possible change-points. Statistical estimation was carried out using a reversible jump MCMC algorithm adapted from Green (1995). Our results have demonstrated the computational and statistical efficiency of this method. The application of non-homogeneous Poisson processes with multiple change-points may provide a suitable modelling framework in the air quality management. In particular, these methods may be instrumental in assessing the impact of environmental actions on the dynamical patterns of potentially hazardous pollutants.

Acknowledgements

J. Gyarmati-Szabó was supported by an EPSRC Doctoral Training Grant, the Strategic Fund of the Institute for Transport Studies and a Postgraduate Research Scholarship of the School of Mathematics (University of Leeds). L. V. Bogachev was partially supported by a Leverhulme Research Fellowship.

References


Insights into protein folding through logistic regression of contact maps

Kerstin Hommola*, Walter R. Gilks, Kanti V. Mardia

Department of Statistics, University of Leeds

We gain insight into generic principles of protein folding by studying aggregated contact maps of single-domain proteins belonging to the four major SCOP classes (all-α, all-β, α/β, α + β). Our approach is completely independent of protein amino-acid sequence and focuses entirely on the positions of contacting residues.

We quantify contact patterns in two specific regions of aggregated contact maps through logistic regression models. The first region represents contacts of residues aligned to the N-terminus with subsequent residues, and the second region contacts of residues aligned to the C-terminus with previous residues. Thus, our analysis is completely symmetric with respect to the chain termini. The models for each of the two regions of interest contain factors for the positions of contacting residues as well as factors describing parallel and anti-parallel β-strand contact patterns.

We observe a striking asymmetry between N-aligned and C-aligned contacts in protein chains belonging to the α/β SCOP class. The N-aligned region shows a strong propensity for parallel contacts between the first few residues and residues further along the sequence, whereas the last few C-aligned residues do not show strong patterns of any kind. This N-terminal dominance could indicate cotranslational folding. The other classes do not show this asymmetry, but reveal predominantly anti-parallel β-strand patterns (all-β class), mixed patterns (α + β class) or no distinct patterns (all-α class).
Evaluation of tissues surrogacy using gene expressions.

Adetayo Kasim\textsuperscript{*1}, Ziv Shked\textsuperscript{2}, Dan Lin\textsuperscript{2}, Suzy Van Sanden\textsuperscript{3}, Jose Cortin\~nas Abrahantes\textsuperscript{4}, Hinrich Goehlmann\textsuperscript{3}, Luc Bij\textsuperscript{3}nens\textsuperscript{3}, Dani Yekutieli\textsuperscript{5}, Jeroen Aerssens\textsuperscript{3}, Michael Camilleri\textsuperscript{6}, Willem Talloen\textsuperscript{3}

\textsuperscript{1} Wolfson Research Institute, Durham University, United Kingdom
\textsuperscript{2} Interuniversity Institute of Biostatistics and Statistical Bioinformatics, Hasselt University, Belgium
\textsuperscript{3} Janssen Pharmaceutical Companies of Johnson and Johnson, Beerse, Belgium.
\textsuperscript{4} European Food Safety Authority, Parma, Italy
\textsuperscript{5} Department of statistics and operational research, Tel Aviv University, Israel
\textsuperscript{6} Mayo Clinic College of Medicine, Rochester MN, USA

1 Introduction

In order to define gene signatures for a disease, the disease-affected tissue is often profiled for gene expressions. When the disease-affected tissue is not readily available, it has recently been shown that disease associated gene signatures could also be identified by profiling an alternative tissue. Achiron and Gurevich (2006) profiled peripheral blood mononuclear cells (PBMC) instead of brain biopsy to investigate gene signatures for multiple sclerosis. Similarly, Le-Niculescu et al. (2008) identified biomarkers for mood disorders using gene expression from the blood. They concluded that blood biomarkers may offer an unexpectedly informative window into brain functioning and disease state. For hepatitis C, Huang et al. (2008) also profiled peripheral blood mononuclear cells to identify gene signatures associated with response to interferon plus ribavirin combination therapy in patients with chronic hepatitis C.

In this paper, we investigate tissues surrogacy between colon and rectum tissues in order to define gene signatures for irritable bowel syndrome (IBS). The colon tissue is considered as the disease-associated tissue and the rectum tissue as its surrogate, given that abnormalities in bowel function (diarrhea, and/or constipation) are a clinically important characteristic for IBS. We investigate the proportion of overlapping differentially expressed genes and prediction accuracy of the disease-associated gene signatures based on expression profiling of both tissues. We also quantify the association between disease effects on both tissues. This study is based on the previous study by Aerssens et al.(2008) and consisted of 34 IBS patients and 24 healthy controls.

2 Differential Gene Signatures for IBS

The differentially expressed genes between the IBS patients and the healthy controls are identified based on gene specific linear models defined as

\[
\begin{pmatrix}
X^A_{ij} \\
X^B_{ij}
\end{pmatrix}
\sim
N\left(\begin{pmatrix}
\mu_{A_j} + \alpha_j Z_i \\
\mu_{B_j} + \beta_j Z_i
\end{pmatrix}; \Sigma_j\right),
\]

(1)

where \(X^A_{ij}\) represents the expression level of \(j\) gene and subject \(i\) in the colon tissue and \(X^B_{ij}\) represents the expression level of the \(j\)th gene and subject \(i\) in the rectum tissue. We denote the
gene-specific disease effects by $\alpha_j$ and $\beta_j$ for the colon and rectum tissues, respectively. Note that $Z_i$ is 1 if subject $i$ is IBS patient and 0 otherwise. The model based covariance matrix $\Sigma_j$ is structured as

$$
\Sigma_j = \begin{pmatrix}
\sigma_{AA_j} & \sigma_{AB_j} \\
\sigma_{BA_j} & \sigma_{BB_j}
\end{pmatrix}.
$$

(2)

After adjusting for multiple testing, the colon tissue resulted in 30 and 72 differentially expressed genes at FDR of 5% and 10%, respectively. From the rectum tissue, no gene was found to be differentially expressed. Also, there is very little overlapping in the lists of top genes from both tissues.

3 Disease Level Surrogacy between the Tissues

Although, no gene was found to be differentially expressed from the rectum tissue, Figure 1 suggests strong association between the estimates of disease effects from the colon and the rectum tissues. The disease level surrogacy seeks to investigate whether disease effects from the colon tissue can be predicted by the disease effect on the rectum tissue.

![Figure 1: Relationship between estimated disease effects from the colon ($\hat{\alpha}_j$) and rectum tissues ($\hat{\beta}_j$).](image)

Table 1 confirms strong associations between disease effects on the colon and the rectum tissues as suggested by in Figure 1. This implies that gene expression from the rectum tissue may contain information about IBS.

<table>
<thead>
<tr>
<th>Top K</th>
<th>LM</th>
<th>RF</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.90 [0.80; 0.95]</td>
<td>0.90 [0.76; 0.97]</td>
<td>0.75 [0.35; 0.94]</td>
</tr>
<tr>
<td>100</td>
<td>0.90 [0.83; 0.94]</td>
<td>0.93 [0.87; 0.97]</td>
<td>0.83 [0.62; 0.93]</td>
</tr>
<tr>
<td>200</td>
<td>0.86 [0.81; 0.90]</td>
<td>0.92 [0.88; 0.96]</td>
<td>0.84 [0.72; 0.92]</td>
</tr>
<tr>
<td>300</td>
<td>0.84 [0.80; 0.88]</td>
<td>0.91 [0.88; 0.94]</td>
<td>0.83 [0.73; 0.90]</td>
</tr>
<tr>
<td>400</td>
<td>0.82 [0.77; 0.86]</td>
<td>0.88 [0.84; 0.92]</td>
<td>0.81 [0.72; 0.87]</td>
</tr>
<tr>
<td>500</td>
<td>0.81 [0.77; 0.85]</td>
<td>0.88 [0.85; 0.91]</td>
<td>0.81 [0.73; 0.86]</td>
</tr>
</tbody>
</table>
4 Classification and Class Prediction

In this section, we develop gene signatures for IBS based on gene expression levels from both the colon and rectum tissues using the following approaches.

- gene selection and classification based on gene expression from the colon tissue (T/T)
- gene selection and classification based on gene expression from the rectum tissue (S/S)
- gene selection based on gene expression from the colon tissue, but classification based on gene expression from the rectum tissue (T/S)
- known gene signatures with classification based on gene expression from the rectum tissue (A/S)

Based on the argument by Ruschhaupt et al. (2004) that classification algorithms should be disentangled from the biology in order to establish what amount of the observed discrimination can be attributed to biological differences, a summarized overview of the results obtained from different approaches is shown in Figure 2. The box plot shows the overall misclassification errors averaged over all the feature selections methods, classification algorithms, and the different size of genes. This finding suggests that an overall significant improvement in the predictive power of gene expression from the rectum tissue can be obtained if a more relevant set of genes, such as genes identified based on the colon tissue are used for classification.

![Box plot showing misclassification errors](image-url)

*Figure 2: Overview of misclassification errors from classification and class prediction for IBS patients and healthy controls.*

5 Discussion

In line with the recent development in microarray experiments, where alternative or surrogate tissues are profiled to investigate gene signatures for diseases, we have shown in the case of
irritable bowel syndrome that gene expression levels from the rectum tissue could discriminate between IBS patients and healthy controls if relevant gene sets are used.

References


Multi-scale analysis of high performance liquid chromatography data

Jennifer H Klapper* and Stuart Barber

Department of Statistics, University of Leeds

Chromatography is a process used to separate a chemical sample into constituent components, which is widely used in chemical analysis. High Performance Liquid Chromatography (HPLC) experimentation produces chromatograms, a type of spectral data comprising a collection of sequentially recorded numbers of ions which arrive at the detector coupled with the corresponding mass-to-charge ratio. Each individual constituent creates a ‘peak’ in a data trace and from this the constituent can be identified and subsequently quantified. Statistically, the challenge is to locate these peaks from noisy data and estimate the area under each peak.

There are many problems which confound the analysis of chromatograms. Variations in conditions and instrumentation as well as the presence of background noise, amongst other factors, cause problems in analysis. Traditionally Fourier transforms played a major role in the analysis of this type of spectral data. However, recent developments in wavelet methods allow practitioners to decompose complex signals, including those produced by HPLC, into components with different frequencies. We therefore implement a wavelet based approach to HPLC data analysis. We continue previous work by Walls et al (2007) which first implemented vaguelette-wavelet (VW) methods to improve the accuracy of peak location. Firstly, we propose a new method for location of the peak end times and in doing so we include two tuning parameters which define the sensitivity of the peak detection procedure, experimenting to find the optimal values of these parameters. To further test the accuracy of our method we experiment to find the lower limit of detectability. This is a point of interest to practitioners using HPLC as it is useful to know how small a peak can be and still be reliably detected. Finally, we propose and experiment as to whether further integrating VW methods into our methodology to test whether this increases the accuracy of peak quantification.

References

Bayesian inference for joint modelling of longitudinally continuous, binary and ordinal events

Qiuju Li 1*, Jianxin Pan 1 and John Belcher 2

1 School of Mathematics, University of Manchester
2 Arthritis Research UK Primary Care Centre, Keele University

1 Introduction

Joint pain and stiffness are a common problem in older people. In the past decade, there were clinical studies trying to investigate the pattern of the occurrence in older people, including the North Staffordshire Osteoarthritis Project (NorStOP) for the local area population with age \( \geq 50 \) years. The NorStOP study has been lasting for six years. In additional to the subject baseline information such as age, gender, alcohol and income status, body mass index (BMI), pain inference (PI) degree and depression status are collected longitudinally. In fact, for each patient the BMI, PI and depression status are measured at year 0, 3 and 6. Among these longitudinal data, the BMI are continuous while the PI are ordinal data, taking a possible value from \( \{1, 2, 3, 4, 5\} \), and the depression status are binary data. In the NorStOP study, longitudinal patterns of these three variables over time are of the primary interest.

For longitudinal data, Laird and Ware (1982) introduced the linear mixed model for continuous data, and Breslow and Clayton (1993) proposed the generalized linear mixed model for discrete responses. Alternatively, the generalized estimating equations (GEE), proposed by Liang and Zeger (1986), can be used to analyze such either continuous or discrete longitudinal data. In the literatures, people only focus on the analysis for a single longitudinal outcome variable, i.e., separate analysis, even if there are several associated longitudinal processes which may have different types of outcomes like the NorStOP study. Naturally, as these longitudinal processes are from the same subject there are a kind of associations between these longitudinal outcomes. Ignoring such natural associations may lead to biased estimates for parameters in the models for these longitudinal outcomes. In this study, we demonstrate the importance of joint modelling for different type outcomes of longitudinal processes. We also propose a new method to incorporate the association between different type outcomes of longitudinal processes into statistical models through random effects.

2 Models and methods

Motivated by the longitudinal processes from NorStOP study, we propose a joint random-effects model for the three events, that is, longitudinal continuous, ordinal and binary data.

For each subject, besides covariates \( X_i \), we have three types of outcomes, \((Y_i, Z_i, T_i)\), \(i = 1, 2, ..., m\), where \( Y_i = (Y_{i1}, Y_{i2}, ..., Y_{im})' \), \( Z_i = (Z_{i1}, Z_{i2}, ..., Z_{im})' \), and \( T_i = (T_{i1}, T_{i2}, ..., T_{im})' \) are longitudinally continuous, binary and ordinal data, respectively. We assume \( Z_{ij} \in \{0, 1\}, T_{ij} \in \{1, 2, ..., K, K \geq 3\} \), and \( Y_{ij} \) is the measurement of subject \( i \) at visit \( j \). We propose to use the linear mixed model or the generalized linear mixed model to analyze the different outcomes.
Since \((Y_i, Z_i, T_i)\) comes from one subject, we therefore propose to use the following models

\[
Y_{ij} = X'_{ij} \beta_1 + d'_{ij} u_{i1} + \epsilon_{ij}, \quad i = 1, 2, ..., m
\]

\[
P(Z_{ij} = 1) = \Phi \left( X'_{2ij} \beta_2 + d'_{2ij} u_{i2} \right)
\]

\[
P(T_{ij} \leq k) = \Phi \left( \alpha_k - X'_{3ij} \beta_3 - d'_{3ij} u_{i3} \right), \quad k = 1, 2, ..., K - 1
\] (1)

To jointly model the different outcomes for the three longitudinal processes. Note that \((Y_i, Z_i, T_i)\) are assumed to be mutually independent provided that \(u_i\) is given. In the above models, \( \beta_1 p_{3 \times 1}, \beta_{2 p_{2 \times 1}}, \beta_{3 p_{3 \times 1}}, G(q_1 + q_2 + q_3) \times 1 \) and \( \sigma^2 \) are unknown regression coefficients or variance component parameters, \( u_i \) are random effects which are assumed to be independent with \( \epsilon_i \). We define \( D_{1i} = (d'_{i11}, d'_{i12}, ..., d'_{i1m})' \), and similarly we have \( D_{2i} \) and \( D_{3i} \). \( \Phi(\cdot) \) is the CDF of standard normal distribution, which is chosen as the link function for the models for both binary and ordinal responses. For the ordinal data model, we assume \( \alpha_1 \leq \alpha_2 \leq ... \leq \alpha_{K-1} \) to reflect the ordinal feature. To make the model identifiable, we set \( \alpha_1 = 0 \) and have an intercept in \( \beta_3 \).

In clinical trials, it is quite often that we may have some prior information about trials and we want to add this information to real data analysis. Thus, for the joint models in (1) we make statistical inferences within the Bayesian framework.

For the joint models in (1), we may wish to find the ML estimators of all the parameters through marginalizing the likelihood function. However, it is quit challenging due to the analytically intractable likelihood function for the generalized linear mixed models. Within the Bayesian framework, it is not easy yet to find the likelihood function or posterior joint density of all parameters, but it becomes much easier if we use the data augmentation approach. Similar to Albert and Chib (1993), for the binary response data we can introduce the latent variables \( (\lambda_{1i}^1, \lambda_{1i}^2, ..., \lambda_{1i}^{q_1}) \), where \( \lambda_{1i}^1 \) are independent samples from \( N(X_{2i} \beta_2 + D_{2i} u_{i2}, I_{n_i}) \) provided that \( u_{i2} \) are given. We take \( Z_{ij} = 1 \) if \( \lambda_{1i}^j > 0 \) otherwise \( Z_{ij} = 0 \). Similarly, for the ordinal response data we also introduce the latent variables \( (\lambda_{1i}^2, \lambda_{1i}^3, ..., \lambda_{1i}^{q_2}) \), where \( \lambda_{1i}^2 \) are independent samples from \( N(X_{3i} \beta_3 + D_{3i} u_{i3}, I_{n_i}) \) by assuming \( u_{i3} \) are given. We take \( T_{ij} = k \) if \( \alpha_{k-1} < \lambda_{1i}^2 \leq \alpha_k \), where \( -\infty = \alpha_0 < \alpha_1 \leq ... \leq \alpha_K = \infty \), and \( \alpha_1 = 0 \).

For the regression coefficients, we choose normal distributions as the prior. For the variance parameters, we may choose the inverse-Gamma or inverse-Wishart distribution. We then apply Gibbs sampling method to simulate random samples from the posterior joint density of all parameters. Note that if \( u_i \) are multivariate random-effects, the proposed method may be computationally intensive. It is also our interest to resolve this issue in this project study.

3 Numerical results

Below, we present some simulation results for the joint models where the association between models does exist. We first study the performance of the models by ignoring the model association. In other words, we model the three longitudinal events separately even if the longitudinal processes are associated. This can provide evidence on the price we have to pay if the models we use are not correct. We then model the three longitudinal outcomes, simultaneously, by using the random effects models in (1). For easy computation, we start with only taking
the continuous measurements $Y_i$ as longitudinal, but not the binary and ordinal measurements. Similar model was considered by Wang, Wang, and Wang (2000) where no ordinal model was involved. In our simulation setup, only random intercept is included into the three models. In the models for the binary and ordinal responses, two covariates are considered. The sample size is taken as 500. 100 simulation samples are generated.

Simulation results are summarized in the table and figures below. It shows that when the three events outcomes are really associated, the separate modelling strategy, i.e., ignoring the model association, can result in biased estimates for the regression coefficients, especially for the models for the binary and ordinal responses. The proposed joint modelling method lead to very less biased or even unbiased estimates of the regression coefficients. This shows the evidence that it is really important to do jointly modelling when multiple longitudinal precesses are involved.

We are currently in modelling of the NorStOP data using the proposed approach and will report the details in the full paper.

<table>
<thead>
<tr>
<th>Par.</th>
<th>True Value</th>
<th>Separate analysis</th>
<th>Joint Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Estimator(sd)</td>
<td>Bias</td>
</tr>
<tr>
<td>$\beta_{21}$</td>
<td>0.2</td>
<td>0.1897 (0.04668)</td>
<td>0.01032</td>
</tr>
<tr>
<td>$\beta_{22}$</td>
<td>-1.0</td>
<td>-0.8895 (0.07732)</td>
<td>0.11052</td>
</tr>
<tr>
<td>$\beta_{31}$</td>
<td>0.3</td>
<td>0.2165 (0.02676)</td>
<td>0.08348</td>
</tr>
<tr>
<td>$\beta_{32}$</td>
<td>-1.0</td>
<td>-0.7200 (0.05811)</td>
<td>0.28002</td>
</tr>
</tbody>
</table>

**Table 1. Part of Simulation Results**

**Acknowledgements**

This research is financially supported by a NIHR grant and by a PhD studentship from the School of Mathematics, University of Manchester.
References


Motor neuron disease and motor unit number estimation: the role of the observed data likelihood in Bayesian inference and computation

A. Pettitt*, C. Drovandi, R. Henderson and P. McCombe

Queensland University of Technology, Royal Brisbane and Womens Hospital, University of Queensland

Introduction

Measuring the progress of motor neuron diseases (MND) such as amyotrophic lateral sclerosis (ALS) (or Lou Gehrig’s disease), is a challenge for clinical neurologists. Motor unit number estimation (MUNE) is an attempt to assess loss of functional motor units. MUNE can also be used to measure the success of potential therapies such as stem cell implants, (2).

An aim is to develop a reliable, repeatable and fast real-time method for estimating the number of motor units, $N$, that serve a given muscle so that the method can be used in hospital clinics. (5) and (4) present a Bayesian MUNE method based on data obtained from a stimulus response curve which is the graph of the compound muscle action potential (CMAP) obtained using surface electrodes from repeated stimulation of a nerve at stimulus intensities ranging from baseline to supramaximal; see Figure 1.

Figure 1: Data from an ALS patient and predicted motor unit firing patterns.

(5) uses known physiology of motor neurons and a fixed $N$ model.(4) calculates a posterior distribution for $N$ using RJMCMC. This method can suffer from poor mixing requiring extremely long runs. We seek an approach that has a thorough within fixed $N$ model exploration of the possibly multimodal posterior and is robust for varying $N$ model comparisons.

Here, we examine methods of inference that use the observed data likelihood. First, the observed data likelihood can improve RJMCMC mixing. Second, we investigate the approach of
(1) to make inferences for $N$ using the posterior distribution of the observed data log-likelihood. We explore if the above approach, which could produce computationally fast results, can reliably approximate the output of our properly mixing RJMCMC. (3) notes that the (1) method targets the product of posteriors, implying that simulations are drawn from the posterior of each model separately rather than the Bayesian approach of simulating over the joint parameter/model space. It is unclear that such a marginal approach is reliable for non-toy examples (3).

The unknown individual motor unit firing indicators, see Figure 1, in the complete data likelihood, have to be marginalized over in order to obtain the observed data likelihood. For relatively small $N$, exact computation is not even possible and we suggest an approximation.

For two patient data sets, we find that the (1) approach would lead to inferences for the number of motor units, $N$, which favors large values of $N$ (with a substantial amount of uncertainty) depending upon the upper value of the range of possible values of $N$. In contrast, the posterior distribution of $N$, found using RJMCMC and the joint distribution of $(N, \Theta)$, is concentrated around values which are highly plausible to neurologists. It would appear, as (6) argue, that to make unbiased inferences for $N$, simulation of values from the joint distribution of $(N, \Theta)$ is required. Furthermore, in all the cases we found that the DIC, BIC and AIC are unable to provide inferences that are in agreement for both data sets with the reversible jump approach.

**Acknowledgments** We thank the Australian Research Council for financial support, patients and staff of RBWH, Matt Armstrong for software development, Dr Fusun Baumann for data collection and Dr Gareth Ridall.

**References**


5. List of participants

N. Adams, University College London, UK
   email: n.adams@imperial.ac.uk

F. Al-Ashwali, Department of Statistics, University of Leeds, UK
   email: mmfsaa@maths.leeds.ac.uk

J. Anderson, Department of Statistics, University of Oxford, UK
   email: anderson@stats.ox.ac.uk

C. Andreetta, Bioinformatics Centre, University of Copenhagen, Denmark
   email: chrandr@binfo.ku.dk

T. Athersuch, MRC-HPA Centre for Environment and Health, Imperial College London, UK
   email: toby.athersuch@imperial.ac.uk

M. Bakoben, Department of Statistics, University of Leeds, UK
   email: mmmab@leeds.ac.uk

S. Barber, Department of Statistics, University of Leeds, UK
   email: stuart@maths.leeds.ac.uk

S. BenFarag, Department of Statistics, University of Leeds, UK
   email: sauad_o@yahoo.com

F.L. Bookstein, Department of Anthropology, University of Vienna, Vienna, Austria
   email: flb@stat.washington.edu

S. Bottaro, DTU Medicine and Technology, Denmark
   email: sbo@elektro.dtu.dk

C. Bowman, University of Leeds, UK
   email: clivebowman@me.com

P. Burdett, Department of Statistics, University of Leeds, UK
   email: philippa.belfield@googlemail.com

N. Burkoff, Warwick Systems Biology Centre, University of Warwick, UK
   email: N.S.Burkoff@warwick.ac.uk

M. Day, Department of Applied Mathematics, University of Leeds, UK
   email: jhs5msd@maths.leeds.ac.uk

I. Dryden, Department of Statistics, University of South Carolina, USA
   email: dryden@mailbox.sc.edu

C. Fallaize, Department of Statistics, University of Leeds, UK
   email: chrisf@maths.leeds.ac.uk

J. Frellsen, The Bioinformatics Centre, University of Copenhagen, Denmark
   email: frellsen@binfo.ku.dk

W. Gilks, Department of Statistics, University of Leeds, UK
   email: wally.gilks@maths.leeds.ac.uk

S. Gore, European Bioinformatics Institute, Cambridge, UK
   email: swanand@ebi.ac.uk
## 6. Author index

<table>
<thead>
<tr>
<th>Author</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>N.M. Adams</td>
<td>131</td>
</tr>
<tr>
<td>J. Aerssens</td>
<td>160</td>
</tr>
<tr>
<td>J.W.J. Anderson</td>
<td>137</td>
</tr>
<tr>
<td>C. Andreetta</td>
<td>127</td>
</tr>
<tr>
<td>T. Athersuch</td>
<td>141</td>
</tr>
<tr>
<td>M. Bakoben</td>
<td>??</td>
</tr>
<tr>
<td>S. Barber</td>
<td>149, 164</td>
</tr>
<tr>
<td>J. Belcher</td>
<td>165</td>
</tr>
<tr>
<td>L. Bijnens</td>
<td>160</td>
</tr>
<tr>
<td>L.V. Bogachev</td>
<td>155</td>
</tr>
<tr>
<td>F.L. Bookstein</td>
<td>87</td>
</tr>
<tr>
<td>W. Boomsma</td>
<td>127</td>
</tr>
<tr>
<td>M. Borg</td>
<td>55</td>
</tr>
<tr>
<td>S. Bottaro</td>
<td>127</td>
</tr>
<tr>
<td>C.J. Brignell</td>
<td>95</td>
</tr>
<tr>
<td>P. Burdett</td>
<td>149</td>
</tr>
<tr>
<td>N.S. Burkoff</td>
<td>129</td>
</tr>
<tr>
<td>M. Camilleri</td>
<td>160</td>
</tr>
<tr>
<td>H. Chen</td>
<td>155</td>
</tr>
<tr>
<td>I. Clark</td>
<td>133</td>
</tr>
<tr>
<td>J. Cortinãs</td>
<td>160</td>
</tr>
<tr>
<td>D. Cox</td>
<td>7</td>
</tr>
<tr>
<td>I. Czogiel</td>
<td>95</td>
</tr>
<tr>
<td>M. Day</td>
<td>150</td>
</tr>
<tr>
<td>M. Defoin-Platel</td>
<td>133</td>
</tr>
<tr>
<td>C. Drovandi</td>
<td>169</td>
</tr>
<tr>
<td>I.L. Dryden</td>
<td>95</td>
</tr>
<tr>
<td>B. Efron</td>
<td>9</td>
</tr>
<tr>
<td>J. Ferkinghoff-Borg</td>
<td>55, 127</td>
</tr>
<tr>
<td>T. Forth</td>
<td>63</td>
</tr>
<tr>
<td>P.S. Freemont</td>
<td>131</td>
</tr>
<tr>
<td>J. Frellsen</td>
<td>55</td>
</tr>
<tr>
<td>W.R. Gilks</td>
<td>133, 159</td>
</tr>
<tr>
<td>H. Goehlmann</td>
<td>160</td>
</tr>
<tr>
<td>S. Gore</td>
<td>65</td>
</tr>
<tr>
<td>N. Green</td>
<td>151</td>
</tr>
<tr>
<td>P.J. Green</td>
<td>93</td>
</tr>
<tr>
<td>J. Gyarmati-Szabó</td>
<td>155</td>
</tr>
<tr>
<td>T. Hamelryck</td>
<td>55, 127</td>
</tr>
<tr>
<td>J. Hein</td>
<td>71, 137</td>
</tr>
<tr>
<td>R. Henderson</td>
<td>169</td>
</tr>
<tr>
<td>J-K. Heriche</td>
<td>47</td>
</tr>
<tr>
<td>J.L. Herman</td>
<td>71</td>
</tr>
<tr>
<td>P. Hirsch</td>
<td>133</td>
</tr>
<tr>
<td>K. Hommola</td>
<td>159</td>
</tr>
<tr>
<td>S.F. Huckemann</td>
<td>99</td>
</tr>
<tr>
<td>Z. Hussain</td>
<td>47</td>
</tr>
<tr>
<td>K.E. Johansson</td>
<td>127</td>
</tr>
<tr>
<td>A. Kasim</td>
<td>160</td>
</tr>
<tr>
<td>W.S. Kendall</td>
<td>79</td>
</tr>
<tr>
<td>J.T. Kent</td>
<td>107</td>
</tr>
<tr>
<td>J. Klapper</td>
<td>164</td>
</tr>
<tr>
<td>G.J. Kleywegt</td>
<td>65</td>
</tr>
<tr>
<td>C.P. Klingenberg</td>
<td>111</td>
</tr>
<tr>
<td>H. Le</td>
<td>83</td>
</tr>
<tr>
<td>D. Lee</td>
<td>47</td>
</tr>
<tr>
<td>J. Lees</td>
<td>47</td>
</tr>
<tr>
<td>M. Levitt</td>
<td>13</td>
</tr>
<tr>
<td>Q. Li</td>
<td>165</td>
</tr>
<tr>
<td>D. Lin</td>
<td>160</td>
</tr>
<tr>
<td>E. Loza</td>
<td>133</td>
</tr>
<tr>
<td>R. Lyngsø</td>
<td>71, 137</td>
</tr>
<tr>
<td>K.V Mardia</td>
<td>29, 41, 55, 149, 159</td>
</tr>
<tr>
<td>P. McCombe</td>
<td>169</td>
</tr>
</tbody>
</table>
G.A. McConkey 63
I. Morilla 47
P. O’Higgins 115
C. Orengo 47
J. Pan 165
A. Pettitt 169
S.E.V. Phillips 15
J. Ranea 47
R. Rentzsch 47
R.A. Russell 131
M.G. Schimek 121
J. Shawe-Taylor 47
Z. Shkedey 160
T.P. Speed 17
J. Staines 137
D.A. Stephens 131
M. Talloen 160
P. Tataru 137
C.C. Taylor ??
D.L. Theobald 41
A. Tramontano 23
S. Van Sanden 160
S. Velankar 65
D.R. Westhead 63
D.J. Weston 131
J.W. Whitaker 63
D.L. Wild 129
D. Yekutieli 160
Y. Zhang 39