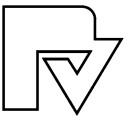
COMPSTAT 2008



COMPSTAT

Proceedings
in Computational Statistics

18th Symposium Held in Porto,
Portugal, 2008

Edited by
Paula Brito

With 128 Figures
and 66 Tables

Physica-Verlag
A Springer Company

Professor Dr. Paula Brito
Faculdade de Economia
Universidade do Porto
Rua Dr. Roberto Frias
4200-464 Porto
Portugal
mpbrito@fep.up.pt

ISBN 978-3-7908-2083-6 e-ISBN 978-3-7908-2084-3

DOI 10.1007/978-3-7908-2084-3

Library of Congress Control Number: 2008932061

Cover design: WMXDesign GmbH, Heidelberg, Germany

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Preface

The 18th Conference of IASC-ERS, COMPSTAT'2008, is held in Porto, Portugal, from August 24th to August 29th 2008, locally organised by the Faculty of Economics of the University of Porto.

COMPSTAT is an initiative of the European Regional Section of the International Association for Statistical Computing (IASC-ERS), a section of the International Statistical Institute (ISI). COMPSTAT conferences started in 1974 in Wien; previous editions of COMPSTAT were held in Berlin (2002), Prague (2004) and Rome (2006). It is one of the most prestigious world conferences in Computational Statistics, regularly attracting hundreds of researchers and practitioners, and has gained a reputation as an ideal forum for presenting top quality theoretical and applied work, promoting interdisciplinary research and establishing contacts amongst researchers with common interests. COMPSTAT'2008 is the first edition of COMPSTAT to be hosted by a Portuguese institution.

Keynote lectures are addressed by Peter Hall (Department of Mathematics and Statistics, The University of Melbourne), Heikki Mannila (Department of Computer Science, Faculty of Science, University of Helsinki) and Timo Teräsvirta (School of Economics and Management, University of Aarhus). The conference program includes two tutorials: "Computational Methods in Finance" by James Gentle (Department of Computational and Data Sciences, George Mason University) and "Writing R Packages" by Friedrich Leisch (Institut für Statistik, Ludwig-Maximilians-Universität). Each COMPSTAT meeting is organised with a number of topics highlighted, which lead to Invited Sessions. The Conference program includes also contributed sessions in different topics (both oral communications and posters).

The Conference Scientific Program Committee includes Paula Brito (University of Porto, Portugal), Helena Bacelar-Nicolau (University of Lisbon, Portugal), Vincenzo Esposito-Vinzi (ESSEC, France), Wing Kam Fung (The University of Hong Kong, Hong Kong), Gianfranco Galmacci (University of Perugia, Italy), Erricos Kontoghiorghes (University of Cyprus, Cyprus), Carlo Lauro (University of Naples Federico II, Italy), Alfredo Rizzi (University "La Sapienza", Roma, Italy), Esther Ruiz-Ortega (University Carlos III, Spain), Gilbert Saporta (Conservatoire National des Arts et Métiers, France), Michael Schimek (Medical University of Graz, Austria), Antónia Turkman (University of Lisbon, Portugal), Joe Whittaker (University of Lancaster, UK), Djamel A. Zighed (University Lumière Lyon 2, France) and Edward Wegman (George Mason University, USA), who were responsible for the Conference Scientific Program, and whom the organisers wish to thank for their invaluable cooperation and permanent availability. Special thanks are also

due to Tomas Aluja, Chairperson of the IASC-ERS and Jaromir Antoch, IASC President, for their continuous support and collaboration.

Due to space limitations, the Book of Proceedings includes keynote speakers' papers and invited sessions speakers' papers only, while the CD-Rom, which is part of it, includes all accepted papers, as well as the tutorials' support texts. The chapters of the Book of Proceedings hence correspond to the invited sessions, as follows:

Keynote
Advances on Statistical Computing Environments
Classification and Clustering of Complex Data
Computation for Graphical Models and Bayes Nets
Computational Econometrics
Computational Statistics and Data Mining Methods for Alcohol Studies (Interface session)
Finance and Insurance (ARS session)
Information Retrieval for Text and Images
Knowledge Extraction by Models
Model Selection Algorithms
Models for Latent Class Detection (IFCS session)
Multiple Testing Procedures
Random Search Algorithms
Robust Statistics
Signal Extraction and Filtering

The papers included in this volume present new developments in topics of major interest for statistical computing, constituting a fine collection of methodological and application-oriented papers that characterize the current research in novel, developing areas. Combining new methodological advances with a wide variety of real applications, this volume is certainly of great value for researchers and practitioners of computational statistics alike.

First of all, the organisers of the Conference and the editors would like to thank all authors, both of invited and contributed papers and tutorial texts, for their cooperation and enthusiasm. We are specially grateful to all colleagues who served as reviewers, and whose work was crucial to the scientific quality of these proceedings. We also thank all those who have contributed to the design and production of this Book of Proceedings, Springer Verlag, in particular Dr. Martina Bihn and Irene Barrios-Kezic, for their help concerning all aspects of publication.

The organisers would like to express their gratitude to the Faculty of Economics of the University of Porto, who enthusiastically supported the Conference from the very start, and contributed to its success, and all people there who worked actively for its organisation. We are very grateful to all

our sponsors, for their generous support. Finally, we thank all authors and participants, without whom the conference would not have been possible.

The organisers of COMPSTAT'2008 wish the best success to Gilbert Saporta, Chairman of the 19th edition of COMPSTAT, which will be held in Paris in Summer 2010. See you there!

Porto, August 2008

Paula Brito
Adelaide Figueiredo
Ana Pires
Ana Sousa Ferreira
Carlos Marcelo
Fernanda Figueiredo
Fernanda Sousa
Joaquim Pinto da Costa
Jorge Pereira
Luís Torgo
Luísa Canto e Castro
Maria Eduarda Silva
Paula Milheiro
Paulo Teles
Pedro Campos
Pedro Duarte Silva

Acknowledgements

The Editors are extremely grateful to the reviewers, whose work was determinant for the scientific quality of these proceedings. They were, in alphabetical order :

Andres M. Alonso
Russell Alpizar-Jara
Tomás Aluja-Banet
Conceição Amado
Annalisa Appice
Helena Bacelar-Nicolau
Susana Barbosa
Patrice Bertrand
Lynne Billard
Hans-Hermann Bock
Carlos A. Braumann
Maria Salomé Cabral
Jorge Caiado
Margarida Cardoso
Nuno Cavalheiro Marques
Gilles Celeux
Andrea Cerioli
Joaquim Costa
Erhard Cramer
Nuno Crato
Guy Cucumel
Francisco De A.T. De Carvalho
José G. Dias
Jean Diatta
Pedro Duarte Silva
Lutz Edler
Ricardo Ehlers
Lars Eldén
Vincenzo Esposito Vinzi
Nuno Fidalgo
Fernanda Otília Figueiredo
Mário Figueiredo
Peter Filzmoser
Jan Flusser
Roland Fried
Fabio Fumarola
Wing K. Fung
Gianfranco Galmacci
João Gama
Ivette Gomes
Esmeralda Gonçalves
Gérard Govaert
Maria Do Carmo Guedes
André Hardy
Nick Heard
Erin Hodgess
Sheldon Jacobson
Alípio Jorge
Hussein Khodr
Guido Knapp
Erricos Kontoghiorghes
Stéphane Lallich
Carlo Lauro
S.Y. Lee
Friedrich Leisch
Uwe Ligges
Corrado Loglisci
Rosaria Lombardo
Nicholas Longford
Donato Malerba
Jean-François Mari
J. Miguel Marin
Leandro Marinho
Geoffrey McLachlan
Paula Milheiro-Oliveira
Isabel Molina Peralta
Yuichi Mori
Irini Moustaki
Maria Pilar Muñoz Gracia
Amedeo Napoli
Manuela Neves
João Nicolau

Monique Noirhomme
M. Rosário de Oliveira
Francesco Palumbo
Rui Paulo
Ana Pérez Espartero
Jorge Pereira
Isabel Pereira
Ana Pires
Mark Plumbley
Pilar Poncela
Christine Preisach
Gilbert Ritschard
Alfredo Rizzi
Paulo Rodrigues
J. Rodrigues Dias
Julio Rodriguez
Fernando Rosado
Patrick Rousset
Esther Ruiz
Gilbert Saporta
Radim Sara
Pascal Sarda
Michael G. Schimek
Lars Schmidt-Thieme
Luca Scrucca
Maria Eduarda Silva
Giovani Silva
Artur Silva Lopes
Carlos Soares
Gilda Soromenho
Fernanda Sousa
Ana Sousa Ferreira
Elena Stanghellini
Milan Studeny
Yutaka Tanaka
Paulo Teles
Valentin Todorov
Maria Antónia Turkman
Kamil Turkman
Antony Unwin
Michel Van De Velden
Maurizio Vichi
Philippe Vieu
Jirka Vomlel
Rafael Weissbach
Joe Whittaker
Peter Winker
Michael Wiper
Djamel A. Zighed

Sponsors

We are extremely grateful to the following institutions whose support contributes to the success of COMPSTAT'2008:

ORGANIZERS:

Contents

Part I

Keynote

Nonparametric Methods for Estimating Periodic Functions, with Applications in Astronomy

Peter Hall

Department of Mathematics and Statistics, University of Melbourne, Melbourne, VIC 3130, Australia, *p.hall@ms.unimelb.edu.au*

Abstract. If the intensity of light radiating from a star varies in a periodic fashion over time, then there are significant opportunities for accessing information about the star's origins, age and structure. For example, if two stars have similar periodicity and light curves, and if we can gain information about the structure of one of them (perhaps because it is relatively close to Earth, and therefore amenable to direct observation), then we can make deductions about the structure of the other. Therefore period lengths, and light-curve shapes, are of significant interest. In this paper we briefly outline the history and current status of the study of periodic variable stars, and review some of the statistical methods used for their analysis.

Keywords: astronomy, curve estimation, light curve, local-linear methods, Nadaraya-Watson estimator, nonparametric regression, periodogram, stars.

1 Introduction

1.1 Periodic variation arising in astronomy

Stars for which brightness changes over time are referred to, unsurprisingly, as variable stars. Some 31,000 such stars are known to exist, and at least another known 15,000 light sources are likely candidates. For many (although not all) such stars, brightness varies in a periodic, or approximately periodic, way. Moreover, stars of this type can often be observed with relatively unsophisticated equipment, for example with small telescopes, binoculars and even with the naked eye. The first variable stars were discovered by direct, unaided observation.

The pulsating star Mira, Latin for "the wonderful," was the first-discovered periodic variable star. It was recorded by David Fabricius, a German minister of religion, in 1596. At first he did not give it much of his attention, but when he noticed the star brighten during 1609 he realised that he had found a new type of light source.

The periodicity of Mira was established by Jan Fokkens Holwarda, a Dutch astronomer, who during 1638 and 1639 estimated the period to be about 11 months. Today we know that the length of the cycle is close to 331

days. For much of its cycle, Mira can be seen unaided. Its brightness varies from about magnitude 2 or 3 up to about 10, and then back again. (On the "magnitude" scale of star brightness, stars of higher magnitude are dimmer, or more difficult to see. Stars of magnitude 8 or larger are not visible to the naked eye.) The relative brightness of Mira, at least for much of its period, would have made it visible to astronomers in classical times.

Variable stars are catalogued into two broad classes — Intrinsic, for which the sources of variability lie within the star itself, and Extrinsic, where the variability comes, in effect, from the star's surface or from outside the star. About 65% of Intrinsic variable stars are "pulsating," and in those cases the brightness varies on account of cyclic expansions and contractions. Mira is of this type; it is a Long-Period Variable star, and stars in this category have periods of between a few days and several years.

Extrinsic variable stars are either Eclipsing Binaries or Rotating Variables. These sources of variation are perhaps the simplest for non-astronomers to understand. In the case of Eclipsing Binaries, one star rotates around the other, and when that star gets between its partner and the observer, the total amount of recorded light is reduced. When the two stars are well separated, as seen by the observer, the total amount of recorded light is maximised. The light emitted by a Rotating Variable star changes through the rotation of material on the star's surface.

This brief account of the nature of variable stars, and more specifically of periodic-variable stars, indicates that we often have only sketchy knowledge of the mechanisms that cause brightness to fluctuate. Even in the case of eclipsing binary stars, for which the nature of the mechanism is relatively clear, the extent of interaction between the two stars may be unknown. For example, mass can be transferred from one star to the other in an eclipsing binary system, although the scale of the transfer may be unclear.

Having a graph of star brightness, as a function of phase during the cycle, can give insight into the nature of these mechanisms within the star, or within the star system. Sometimes an understanding of the mechanisms can be gained for stars that are relatively close, and by comparing their brightness curves with those of distant stars we have an opportunity to gain information about the latter. It is therefore advantageous to have nonparametric estimators of brightness curves, which do not impose mathematical models that dictate the shape of the curve estimates.

1.2 Related literature in astronomy and statistics

Astronomers typically refer to a plot of the mean brightness of a periodic variable star, representing a function of phase during the time duration of a period, as the star's "light curve." Distinctions between the notion of a theoretical light curve, on which we have only noisy data, and an estimate of that curve based on the data, are generally not made. Likewise, the difference between the function on which the true light curve or its estimate are based,

and a graph of that function, is generally not remarked on. These issues should be borne in mind when reading the astronomy literature, and also when interpreting the discussion below.

Ways of explaining the mechanisms that lead to periodic variation in brightness are continuously under development; see Prigara (2007), for instance. Likewise, estimates and interpretations of the curves that represent this variation are constantly becoming available. For example, Norton et al. (2008) present and discuss the light curves of 428 such stars, of which only 68 of had previously been recognised as being of this type. Eyer and Cuypers (2000) predict that the GAIA space mission, expected to be launched by the European Space Agency in 2011, will be able to detect some 18 million variable sources, among them five million classic periodic variable stars and two to three million eclipsing binary systems. Thus, the potential scope of the research problems discussed in this paper is likely to expand rapidly.

Book-length accounts of of variable stars, their properties and their light curves, include those given by Hoffmeister et al. (1985), Sterken and Jaschek (1996), Good (2003), North (2005), Warner (2006) and Percy (2007). The MACHO project, where the acronym stands for MAssive Compact Halo Objects, includes a very large catalogue of light curves for variable stars. See Axelrod et al. (1994) for discussion of statistical issues in connection with the MACHO data.

The astronomy literature on periodic variable stars is sophisticated from a quantitative viewpoint. For example, it includes methodology for discovering light curves that are "outliers" in a catalogue of such curves; see e.g. Protopapas et al. (2006). And it involves automated methodology for identifying periodic variable stars among millions of light sources in the night sky; see e.g. Derue et al. (2002) and Kabath et al. (2007).

There is a large literature on modelling curves in terms of trigonometric series. In statistics and related fields it includes work of Pisarenko (1973), Hannan (1974), Frost (1976), Quinn and Fernandes (1991), Quinn and Thompson (1991), Quinn (1999) and Quinn and Hannan (2001). Many other contributions can be found in the engineering literature. If the number of components is taken large then the methodology essentially amounts to nonparametric curve estimation, and is closely related to approaches discussed below in section 3. Computational and statistical-efficiency issues connected with the estimation of periodic functions are addressed by McDonald (1986) and Bickel et al. (1993, p. 107), respectively.

Early work in astronomy on nonparametric methods for analysing data on periodic variable stars includes contributions from Lafler and Kinman (1965) and Renson (1978). The method most favoured by astronomers for estimating light curves is the periodogram, which was used by statisticians more than a century ago to assess periodicity. Work on formal testing for periodicity includes that of Fisher (1929), Whittle (1954) and Chiu (1989). The theory of periodogram estimation owes much to Walker (1971, 1973) and Hannan

(1973). The periodgram was introduced to astronomy largely through the work of Barning (1963), Deeming (1975), Lomb (1976), Ferraz-Mello (1981) and Scargle (1982). See also Vityazev (1997). For examples of analyses undertaken using this approach, see Waelkens et al. (1998), de Cat and Aerts (2002), DePoy et al. (2004), Lanza et al. (2004), Aerts and Kolenberg (2005), Maffei et al. (2005), Hall and Li (2006) and Shkedy et al. (2004). Bayesian methods were proposed by Shkedy et al. (2007). Alternative techniques include those of Reimann (1994), Hall et al. (2000) and Hall and Yin (2003).

For some variable stars, the fluctuation of brightness is explained well by a model where period and/or amplitude are also functions of time. See, for example, work of Eyer and Genton (1999), Koen (2005), Rodler and Guggenberger (2005), Sterken (2005), Hart, Koen and Lombard (2007) and Genton and Hall (2007).

1.3 Summary

Section 2 provides an account of least-squares methods for inference in the simplest case, where the light curve can reasonably be modelled in terms of a single periodic function. Periodogram-based methods, and inference when the curve is more plausibly a superposition of p different periodic functions, are treated together in section 3. The case of evolving periodic models is addressed in section 4. Our treatment follows lines given in greater detail by Hall et al. (2000), Hall and Yin (2003), Hall and Li (2006) and Genton and Hall (2007).

2 Models and methodology in the case of periodicity based on least squares

2.1 Models for brightness and observation times

Let $g(x)$ denote the "true" value of brightness of the star at time x. A graph of g, as a function of phase, would be called by astronomers the true "light curve" of the star. We make observations Y_i at respective times X_i, where $0 < X_1 \le \cdots \le X_n$, and obtain the data pairs (X_i, Y_i) for $1 \le i \le n$. The model is superficially one of standard nonparametric regression:

$$Y_i = g(X_i) + \epsilon_i \,, \tag{1}$$

where the ϵ_i's, describing experimental error, are independent and identically distributed random variables with zero mean and finite variance. We take g to be a periodic function with period θ; its restriction to a single period represents the light curve. From the data (X_i, Y_i) we wish to estimate both θ and g, making only periodic-smoothness assumptions about the latter.

A range of generalisations is possible for the model (1). For example, we might replace the errors ϵ_i by $\sigma(X_i)\,\epsilon_i$, where the standard deviation $\sigma(X_i)$

is either known, as is sometimes the case with data on star brightness, or accurately estimable. Then, appropriate weights should be incorporated into the series at used to estimate θ; see (3) below. To reflect many instances of real data, the time points X_i should remain separated as n increases, and in particular the standard "infill asymptotics" regime of nonparametric regression is inappropriate here.

Neither should the X_i's be modelled as equally spaced quantities. Indeed, it is straightforward to see that in this case, and for many values of θ (in particular where θ is a rational multiple of the spacing), consistent estimation is not possible.

Realistic mathematical models for the spacings between successive X_j's include the case where they are approximately stochastically independent. One such model is

$$X_j = \sum_{i=1}^{j} V_i\,, \quad 1 \le j \le n\,, \tag{2}$$

where V_1, V_2, ... are independent and identically distributed nonnegative random variables. Clearly there are limitations, however, to the generality of the distribution allowable for V, representing a generic V_i. In particular, if the distribution is defined on an integer lattice, and if θ is a rational number, then identifiability difficulties continue to cause problems.

These problems vanish if we assume that the X_j's are generated by (2), where the distribution of $V > 0$ is absolutely continuous with an integrable characteristic function and that all moments of V are finite. Call this model $(\mathrm{M}_{X,1})$. The fact that the characteristic function should be integrable excludes the case where the X_i's are points of a homogeneous Poisson process, but that context is readily treated separately.

Another class of processes $\mathcal{X}$ is the sequence $X_j = X_j(n) \equiv nY_{nj}$, where $Y_{n1} < \cdots < Y_{nn}$ are the order statistics of a random sample $Y_1, \ldots, Y_n$ from a Uniform distribution on the interval $[0, y]$, say. Call this model $(\mathrm{M}_{X,2})$. Models $(\mathrm{M}_{X,1})$ and $(\mathrm{M}_{X,2})$ are similar, particularly if V has an exponential distribution. There, if $\mathcal{X}(n+1) = \{X_1, \cdots, X_{n+1}\}$ is a sequence of observations generated under $(\mathrm{M}_{X,1})$, if $\mathcal{X}'(n) = \{X'_1, \ldots, X'_n\}$ is generated under $(\mathrm{M}_{X,2})$ with $y = 1$, and if we define $X_{\mathrm{tot}} = \sum_{i \le n+1} X_i$, then $\{X_1/X_{\mathrm{tot}}, \ldots, X_n/X_{\mathrm{tot}}\}$ has the same distribution as $\mathcal{X}'(n)$.

A third class of processes $\mathcal{X}$ is the jittered grid of Akaike (1960), Beutler (1970) and Reimann (1994), where $X_j = j + U_j$, for $j \ge 1$, and the variables U_j are independent and Uniformly distributed on $(-\frac{1}{2}, \frac{1}{2})$. Call this model $(\mathrm{M}_{X,3})$. Each of $(\mathrm{M}_{X,1})$, $(\mathrm{M}_{X,2})$ and $(\mathrm{M}_{X,3})$ has the property that the spacings $X_j - X_{j-1}$ are identically distributed and weakly dependent.

2.2 Least-squares estimation of g and θ

In this section we give an overview of methods for inference. The first step is to construct a nonparametric estimator $\hat{g}(\cdot \,|\, \theta)$ of g on $(0, \theta]$, under the

assumption that the period of g is θ. Next we extend $\hat{g}$ to the real line by periodicity, and, using a squared-error criterion,

$$S(\theta) = \sum_{i=1}^{n} \{Y_i - \hat{g}(X_i \,|\, \theta)\}^2 \,, \tag{3}$$

take our estimator $\hat{\theta}$ of θ to be the minimiser of $S(\theta)$. (We could use a leave-one-out construction of $S(\theta)$, omitting the pair (X_i, Y_i) from the data. While this would give slightly different numerical results, it would not influence first-order asymptotic properties of the method.) Finally, for an appropriate estimator $\tilde{g}(\cdot \,|\, \theta)$ of g under the assumption of period θ, we employ $\hat{g}_0 \equiv \tilde{g}(\cdot \,|\, \hat{\theta})$ to estimate g.

Even if $\hat{g}$ and $\tilde{g}$ are of the same type, for example both local-linear estimators, it is usually not a good idea to take them to be identical. In particular, to ensure approximately optimal estimation of θ, the version $\hat{g}(\cdot \,|\, \theta)$ that we use to define $S(\theta)$ at (3) should be smoothed substantially less than would be appropriate for point estimation of g. In general the function S has multiple local minima, not least because any integer multiple of θ can be considered to be a period of g.

Next we discuss candidates for $\hat{g}$. Under the assumption that the true period of g is θ, the design points X_i may be interpreted modulo θ as $X_i(\theta) = X_i - \theta \lfloor X_i/\theta \rfloor$, for $1 \le i \le n$, where $\lfloor x \rfloor$ denotes the largest integer strictly less than x. Then, the design points of the set of data pairs $\mathcal{Y}(\theta) \equiv \{(X_i(\theta), Y_i), 1 \le i \le n\}$ all lie in the interval $(0, \theta]$. We suggest repeating *ad infinitum* the scatterplot represented by $\mathcal{Y}(\theta)$, so that the design points lie in each interval $((j-1)\theta, j\theta]$ for $-\infty < j < \infty$; and computing $\hat{g}(\cdot \,|\, \theta)$, restricted to $(0, \theta]$, from the data, using a standard second-order kernel method such as a Nadaraya-Watson estimator or a local-linear estimator. In practice we would usually need to repeat the design only in each of $(-\theta, 0]$, $(0, \theta]$ and $(\theta, 2\theta]$, since the effective bandwidth would be less than θ. We define $\hat{g}(\cdot \,|\, \theta)$ on the real line by $\hat{g}(x \,|\, \theta) = \hat{g}(x - \theta \lfloor x/\theta \rfloor \,|\, \theta)$.

In view of the periodicity of g it is not necessary to use a function estimation method, such as local linear, which accommodates boundary effects. Indeed, our decision to repeat design points in blocks of width θ means that we do not rely on the boundary-respecting properties of such techniques. The Nadaraya-Watson estimator, which suffers notoriously from boundary problems but is relatively robust against data sparseness, is therefore a good choice here. The resulting estimator of g is

$$\hat{g}(x \,|\, \theta) = \frac{\sum_i Y_i \, K_i(x \,|\, \theta)}{\sum_i K_i(x \,|\, \theta)} \,, \quad 0 \le x \le \theta \,, \tag{4}$$

where $K_i(x \,|\, \theta) = K[\{x - X_i(\theta)\}/h]$, K is a kernel, h is a bandwidth, and the two series on the right-hand side of (4) are computed using repeated blocks of the data $\mathcal{Y}(\theta)$.

Alternative estimators of g, of slightly lower statistical efficiency than that defined in (4), can be based on the periodogram. This approach tends to be favoured by astronomers, not least because it is readily extended to the case of multiperiodic functions; see section 3.

2.3 Properties of estimators

If g has r bounded derivatives; if the estimator $\hat{g}$ is of rth order, meaning that its asymptotic bias is of size h^r and its variance is of size $(nh)^{-1}$; and if $h = h(n)$ has the property that for some $\eta > 0$, $n^{-(1/2)+\eta} \le h = o(n^{-1/(2r)})$; then $\hat{\theta} = \operatorname{argmin} S(\theta)$ is consistent for θ and, under regularity conditions,

$$n^{3/2}\,(\hat{\theta} - \theta) \to \mathrm{N}\big(0, \tau^2\big) \tag{5}$$

in distribution, where $0 < \tau^2 < \infty$. When $\hat{g}$ is a Nadaraya-Watson or local-linear estimator,

$$\tau^2 = 12\,\sigma^2\,\theta^3\,\mu^{-2}\left\{\int_0^\theta g'(u)^2\,du\right\}^{-1}, \tag{6}$$

where $\sigma^2 = \operatorname{var}(\epsilon_i)$ and $\mu = \lim_{j\to\infty} E(X_j - X_{j-1})$, assumed to be finite and nonzero. Formula (5) implies that $\hat{\theta}$ converges to θ at a parametric rate. In Quinn and Thompson's (1991) parametric analysis of a closely related problem they obtained the same limit theorem for $\hat{\theta}$, albeit with a different value of τ^2.

Formula (6) implies that estimators of period have lower variance when the function g is 'less flat', i.e. when g has larger mean-square average derivative. This accords with intuition, since a perfectly flat function g does not have well-defined period, and more generally, the flatter g is, the more difficult it is to visually determine its period.

If $h \sim Cn^{-1/(2r)}$ for a constant $C > 0$, and $\hat{g}$ is an r'th order regression estimator, then $n^{3/2}\,(\hat{\theta} - \theta)$ remains asymptotically Normally distributed but its asymptotic bias is no longer zero. In the r'th order case, $h = O(n^{-1/(2r)})$ is the largest order of bandwidth that is consistent with the parametric convergence rate, $\hat{\theta} = \theta + O_p(n^{-3/2})$.

This high degree of accuracy for estimating θ means that, if $\tilde{g}(\cdot\,|\,\theta)$ is a conventional estimator of g under the assumption that the period equals θ, then first-order asymptotic properties of $\hat{g}_0 \equiv \tilde{g}(\cdot\,|\,\hat{\theta})$ are identical to those of $\tilde{g}(\cdot\,|\,\theta)$. That is, from an asymptotic viewpoint the final estimator $\hat{g}_0$ behaves as though the true period were known. These results follow by Taylor expansion. For example, if $\tilde{g}(\cdot\,|\,\theta)$ is the Nadaraya-Watson estimator defined at (2.4), but with a different bandwidth h_0 say, satisfying $h_0 \ge n^{-(1/2)+\xi}$ for some $\xi > 0$, then a Taylor-expansion argument shows that for all $\eta > 0$,

$$\tilde{g}(\cdot\,|\,\hat{\theta}) = \tilde{g}(\cdot\,|\,\theta) + o_p\big\{(nh_0)^{-1/2}\big\}\,. \tag{7}$$

The remainder $o_p\big\{(nh_0)^{-1/2}\big\}$ here is of smaller order than the error of $\tilde{g}(\cdot\,|\,\theta)$ about its mean.

3 The case of multiperiodic functions

3.1 Model for g, and issues of identifiability

In some cases the radiation from a star can reasonably be modelled as a superposition of more than one periodic function. To avoid problems of non-identifiability we take g to be representable as

$$g(x) = \mu + \sum_{j=1}^{p} g_j(x)\,, \quad -\infty < x < \infty\,, \tag{8}$$

where μ denotes a constant, g_j is a smooth, nonvanishing, real-valued periodic function with minimal period θ_j, $0 < \theta_1 < \ldots < \theta_p < \infty$, and each g_j is centred by the condition

$$\int_0^{\theta_j} g_j(x)\,dx = 0\,. \tag{9}$$

Therefore, the constant term in any orthogonal expansion of g_j on $[0, \theta_j]$, with respect to an orthonormal system where one of the orthonormal functions is constant, is absorbed into μ at (8). This property will motivate our estimators of $g_1, \ldots, g_p$; see section 3.3 below.

We assume p is known, and address the problem of estimating $\theta_1, \ldots, \theta_p$ and $g_1, \ldots, g_p$ without making parametric assumptions about the latter. Of course, by conducting inference for different values of p one can obtain significant information about its "true" value, but we do not have a satisfactory approach to formally estimating p.

By saying that θ_j is the minimal period of g_j we mean that if g_j is also periodic with period θ' then $\theta_j \le \theta'$. This does not render either the θ_j's or the representation at (8) uniquely defined, however. Indeed, the representation is unique if and only if the periods are "relatively irrational", meaning that θ_i/θ_j is irrational for each $1 \le i < j \le p$. We shall say that the periods are "relatively rational" if each value of θ_i/θ_j is a rational number.

At first sight this suggests an awkward singularity in the statistical problem of conducting inference about g_j and θ_j, as follows. Since each irrational number is approximable arbitrarily closely by rational ones, then so too each statistically identifiable problem can be approximated arbitrarily closely by a non-identifiable one, by slightly altering the periods θ_j and leaving the shape of each g_j essentially unchanged. And since the periods in the approximating problem can be chosen to be relatively rational, then new and quite different representations may be constructed there, involving finite mixtures of periodic functions that are different from those in the relatively irrational form of the problem. This implies that, even if the original mean function g uniquely enjoys the representation at (8), there is an infinity of alternative mean functions that, while being themselves very close to g, have representations, as mixtures of periodic functions, that differ significantly from the unique representation of g.

While this is correct, it does not often hinder statistical analysis of real or simulated data, since the alternative representations involve functions g_j that are either very rough or have very long periods. In such cases the g_j's are often not practically recognisable as periodic functions, and in particular they lead to solutions that usually appear as pathological.

3.2 Period estimators based on the periodogram

Assume that the data pairs (X_i, Y_i) are generated as at (1), but that g is now a multiperiodic function. Least-squares methods can be used to construct estimators of g and of the periods θ_j, but they are awkward to use in practice, at least without appropriate "starting estimators," since the analogue of $S(\theta)$, at (3), has many local extrema. On the other hand, methods based on the periodogram are relatively easy to implement; we describe them below.

Let cs denote either the cosine or the sine function. For any real number ω, define the squared periodogram by

$$A(\omega)^2 \equiv A_{\cos}(\omega)^2 + A_{\sin}(\omega)^2 \,,$$

where $A_{\mathrm{cs}}(\omega) = n^{-1} \sum_i Y_i \,\mathrm{cs}(\omega X_i)$. If $p = 1$, in which case there is a unique period θ, say, then the quantity $\widehat{\omega}$ which produces a local maximum of $A(\omega)$ achieves a local maximum in the vicinity of each value $\omega^{(k)} = 2k\pi/\theta$, where k is any nonzero integer. This property is readily used to estimate θ.

More generally, in the multiperiodic case the periodogram A has its large peaks near points $2k\pi/\theta_j$, for arbitrary integers k and for $j = 1, \ldots, r$. By sorting peak locations into r disjoint sets, for each of which adjacent values are approximately equal distances apart, the values of θ_j may be estimated as before. In either case the estimators converge at the rate $n^{-3/2}$ discussed for the least-squares methods introduced in section 2.

3.3 Estimators of g

Having constructed $\hat{\theta}_1, \ldots, \hat{\theta}_p$ we use orthogonal series methods to develop estimators $\hat{g}_1, \ldots, \hat{g}_p$, as follows. Let $\{\psi_0, \psi_1, \ldots\}$ denote a complete orthonormal sequence of functions on the interval $[0, 1]$, with $\psi_0 \equiv 1$. Extend each function to the real line by periodicity. Given an integer $m \geq 1$, which will play the role of a smoothing parameter; given generalised Fourier coefficients a_{jk} for $1 \leq j \leq p$ and $1 \leq k \leq m$; and given a constant μ; put

$$\tilde{g}(x \,|\, a, \mu) = \mu + \sum_{j=1}^{p} \sum_{k=1}^{m} a_{jk} \, \psi_k(x/\hat{\theta}_j) \,, -\infty < x < \infty \,, \tag{10}$$

where a denotes the parameter vector of length $q = mp$ made up of all values of a_{jk}. Of course, the functions $\psi_k(\cdot/\hat{\theta}_j)$ used in this construction are periodic

with period $\hat{\theta}_j$. The estimator (10) reflects the model (8) and the constraint (9), the latter imposed to help ensure identifiability.

Take $(\hat{a}, \widehat{\mu})$ to be the minimiser of

$$T(a, \mu) = \sum_{i=1}^{n} \{Y_i - \tilde{g}(X_i|a, \mu)\}^2 .$$

In this notation our estimator of g_j is

$$\hat{g}_j(x) = \sum_{k=1}^{m} \hat{a}_{jk}\, \psi_k(x/\hat{\theta}_j) .$$

In practice we recommend taking $\{\psi_j\}$ to be the full trigonometric series: $\psi_0 \equiv 1$ and, for $j \geq 1$,

$$\psi_{2j}(x) = 2^{1/2} \cos(2j\pi x) \quad \text{and} \quad \psi_{2j-1}(x) = 2^{1/2} \sin(2j\pi x) .$$

4 Evolving periodic functions

4.1 Introduction

The notion that star brightness is given by a fixed periodic function, unchanging over time, is of course a simplification. The very mechanisms that produce periodicity are themselves the subject of other mechanisms, which affect their properties and so influence the period and amplitude of the supposedly periodic function. Thus, while the model at (1) might be reasonable in many circumstances, in some instances we should allow for the fact that the characteristics of g will alter over time.

In the sections below we develop models for functions with evolving amplitude and period, and then we combine these to produce a model for g. Finally we use that model to motivate estimators.

4.2 The notions of evolving period and amplitude

Write g_0 for a periodic function with unit period, and let t denote a continuously differentiable, strictly increasing function. Represent time by x, and put $t_x = t(x)$ and $t'_x = t'(x) > 0$. We shall consider the function t to provide a change of time, from x to t_x.

Assume that a function g can be represented as

$$g(x) = g_0(t_x) . \tag{11}$$

We think of g as having period $1/t'_x$ at time x, and in fact for small $u > 0$,

$$g(x + u) = g_0\{t_x + t'_x\, u + o(u)\} .$$

Since the function $d(u) = g_0(t_x + t'_x\, u)$ has period $1/t'_x$, then, if the time-transformation t_{x+u} were to be applied in a linear way for $u > 0$, g would have period $1/t'_x$ at all future times $x+u$. More generally, without the linearity assumption, the function g given by (11) can be considered to have a period $1/t'_x$ that evolves as time, x, increases.

Amplitude can also evolve. If $a > 0$ is a smooth function, representing amplitude; and if we write a_x for $a(x)$; then we might generalise (11) to:

$$g(x) = a_x\, g_0(t_x)\,. \tag{12}$$

Here we could consider g to have period $1/t'_x$, and amplitude $a_x\, g_0(t_x)$, at x.

The concept of evolving amplitude has to be treated cautiously, however. While altering time can change only the distances between successive peaks and troughs in the function g_0, altering both amplitude and time can produce a function which is very different. Any smooth, strictly positive function g can be constructed non-uniquely as at (12), with $a > 0$ representing a smooth amplitude change, $t_x \equiv x$ being the identity transformation, and g_0 denoting any strictly positive, smooth function, periodic or otherwise.

One conclusion to be drawn from this discussion is that, unless amplitude is determined by a relatively simple parametric model; and unless it changes only very slowly over time, relative to the lengths of periods; it can interact too greatly with period to be interpretable independently of period.

It is possible for non-identifiability of g_0 to occur even when $a \equiv 1$ and the function t has a simple parametric form. For example, suppose that, in the particular case $p = 1$, $t_{x+kp} = t_x + k$ for each $x \in [0, 1]$ and each integer $k \geq 1$. Then, since g_0 has period 1, it follows that $g_0(t_{x+k}) = g_0(t_x)$ for each x and each integer k. Therefore, the periodic function $g \equiv g_0(t)$ is representable as either a time-changed version of the function g_0 with unit period, or more directly as the non-time changed function $g_1 \equiv g_0(t)$ with unit period. If we consider this particular time-change function t, and also the identity time-change, to be members of a larger parametric class, $\mathcal{T}$ say, of time-change functions, then there is ambiguity in determining the member s of $\mathcal{T}$ that enables us to represent $g \equiv g_0(t)$ as $g = g_2(s)$ where g_2 has period 1.

4.3 Models for period

We shall interpret (12) as a model for a regression mean, g, where the functions a and t are determined parametrically and g_0 is viewed nonparametrically. In order for (12) to be interpretable in astronomical terms, it is helpful for the models for t to be quite simple. For example, taking $t_x = \theta_2^{-1}\, \log(\theta_1 + \theta_2 x) + \theta_3$, for constants $\theta_1 > 0$, θ_2 and θ_3, implies that $1/t'_x = \theta_1 + \theta_2 x$. In this case the initial period is θ_1, and the period changes linearly with time, with slope θ_2. If we start measuring time at zero when $x = 0$ then we require $\theta_3 = -\theta_2^{-1} \log \theta_1$, and then the model becomes:

$$t_x = \theta_2^{-1}\, \log\left(1 + \theta_1^{-1}\theta_2 x\right). \tag{13}$$

We might refer to (13) as a "linear model," since it results from a linear model for period. Analogously we could describe the model

$$t_x = (\theta_1 \theta_2)^{-1} \left(1 - e^{-\theta_2 x}\right), \tag{14}$$

for which $1/t'_x = \theta_1 \, e^{\theta_2 x}$, as an "exponential model." It is an attractive alternative to the linear model in certain cases, since its period is unequivocally positive.

A time-change function such as

$$t_x = \int_0^x \left(\theta_1 + \theta_2 \, u + \ldots + \theta_k \, u^{k-1}\right)^{-1} du \tag{15}$$

produces a period the evolution of which, in time, is described exactly by a polynomial of degree $k - 1$, and represents a generalisation of the linear model.

It should be appreciated that in models (13)–(15), and in a setting where data are assumed to be observed at an approximately constant rate over a time interval $[0, n]$ of increasing length n, usually only the parameter θ_1, representing period at time $x = 0$, would be kept fixed as n increased. The parameters $\theta_2, \ldots, \theta_k$ would typically decrease to zero as n increased, and in fact would usually decrease at such a rate that $n^{j-1} \, |\theta_j|$ was at least bounded, if not decreasing to zero, for $2 \le j \le k$. This prevents period from changing by an order of magnitude over the observation time-interval. Moreover, if $\theta_1 > 0$ is fixed and $\sup_{1 \le j \le k} \, n^{j-1} \, |\theta_j| \to 0$ as $n \to \infty$, then for all sufficiently large values of n, t_x is strictly monotone increasing on $[0, n]$. In such cases, (15) is asymptotically equivalent to the simpler model,

$$t_x = \theta_1^{-1} \, x + \theta_2 \, x^2 + \ldots + \theta_k \, x^k \,, \quad 0 \le x \le n \,, \tag{16}$$

modulo a reparametrisation. An exponentiated version of (16) is also possible.

4.4 Models for amplitude

Models for the function a_x can be constructed similarly to those for t_x. However, in order to avoid identifiability problems we should insist that $a_x = 1$ at the initial time, so that initial amplitude is incorporated into the function g_0. Bearing this in mind, and taking the initial time to be $x = 0$, potential models include

$$a_x = 1 + \omega_1 \, x + \ldots + \omega_\ell \, x^\ell \,, \quad 0 \le x \le n \,,$$

and its exponentiated form, $a_x = \exp(\omega_1 \, x + \ldots + \omega_\ell \, x^\ell)$.

4.5 Model for data generation

Assume that data $(X_1, Y_1), \ldots, (X_n, Y_n)$ are generated by the model

$$Y_i = a(X_i \,|\, \omega^0) \, g_0\{t(X_i \,|\, \theta^0)\} + \epsilon_i \,,$$

where $a_x = a(x \,|\, \omega)$ and $t_x = t(x \,|\, \theta)$ denote smooth, positive functions determined by finite vectors ω and θ of unknown parameters, ω^0 and θ^0 are the true values of the respective parameters, $t(\,\cdot\,|\,\theta)$ is strictly increasing, $a(\,\cdot\,|\,\omega)$ is bounded away from zero and infinity, and the experimental errors, ϵ_i, have zero mean. For example, $t(\,\cdot\,|\,\theta)$ and $a(\,\cdot\,|\,\omega)$ could be any one of the models introduced in sections 4.3 and 4.4, respectively.

As in section 4.2, g_0 is assumed to be a smooth, periodic function with unit period. Therefore, even if the regression mean, $g(x) = a(x \,|\, \omega)\, g_0\{t(x \,|\, \theta)\}$, were a conventional periodic function, without any amplitude or time change, the period, p say, would be inherited from the time-change function $t(x \,|\, \theta)$, which here would be linear: $t(x \,|\, \theta) = x/p$ and $\theta = p$, a scalar. We shall take $a(0 \,|\, \omega) = 1$ if $x = 0$ is the earliest time-point on our scale, so that amplitude is inherited from g_0.

Similar results, and in particular identical convergence rates of estimators, are obtained for a variety of processes X_i that are weakly stationary and weakly independent. They include cases where the X_i's are (a) points of a homogeneous Poisson process with intensity μ^{-1} on the positive real line; or (b) the values of $[n/\mu]$ (integer part of n/μ) independent random variables, each uniformly distributed on the interval $[0, n]$; or (c) the values within $[0, n]$ of the "jittered grid" data $j\mu^{-1} + V_j$, where the variables V_j are independent and identically distributed on a finite interval. See section 2.1 for discussion of models such as (a), (b) and (c). In each of these cases the average spacing between adjacent data is asymptotic to μ as $n \to \infty$.

4.6 Estimators

To estimate g_0, ω and θ, put

$$\hat{g}_0\big\{t(x \,|\, \theta) \,\big|\, \theta, \omega\big\} = \frac{\sum_i a(X_i \,|\, \omega)^{-1}\, Y_i\, K_i(x \,|\, \theta)}{\sum_i K_i(x \,|\, \theta)}\,,$$

$$S(\theta, \omega) = \sum_i \Big[Y_i - a(X_i \,|\, \omega)\, \hat{g}_0\big\{t(X_i \,|\, \theta) \,\big|\, \theta, \omega\big\}\Big]^2\,,$$

where $K_i(x \,|\, \theta) = K[\{x(\theta) - X_i(\theta)\}/h]$, K is a kernel function, h is a bandwidth,

$$x(\theta) = t(x \,|\, \theta) - \lfloor t(x \,|\, \theta) \rfloor\,, \quad X_i(\theta) = t(X_i \,|\, \theta) - \lfloor t(X_i \,|\, \theta) \rfloor\,,$$

and $\lfloor u \rfloor$ denotes the largest integer strictly less than u.

Let $(\theta, \omega) = (\hat{\theta}, \widehat{\omega})$ be the minimiser of $S(\theta, \omega)$. Then, potentially using, to construct $\hat{g}_0$, a bandwidth different from the one employed earlier, our estimator of g_0 is $\hat{g}_0(\,\cdot\,|\,\hat{\theta},)$. Estimators of the time-change function $t_x = t(x \,|\, \theta^0)$ and amplitude function $a_x = a(x \,|\, \omega^0)$ are given by $\hat{t}_x = t(x \,|\, \hat{\theta})$ and $\hat{a}_x = a(x \,|\, \widehat{\omega})$, respectively. We estimate g, defined at (12), as $\hat{g}(x) = \hat{a}_x\, \hat{g}_0(\hat{t}_x)$.

References

AKAIKE, H. (1960): Effect of timing error on the power spectres of sampled-data. *Ann. Inst. Statist. Math. 11, 145-165.*

AXELROD, T.S. and 18 OTHER AUTHORS (1994): Statistical issues in the MACHO Project. In: G. J. Babu and E. D. Feigelson (Eds.) *Statistical Challenges in Modern Astronomy II*, Springer, New York, 209-224.

AERTS, C. and KOLENBERG, K. (2005): HD 121190: A cool multiperiodic slowly pulsating B star with moderate rotation. *Astronom. Astrophys. 431, 614-622.*

BARNING, F.J.M. (1963): The numerical analysis of the light-curve of 12 Lacertae. *Bull. Astronom. Institutes of the Netherlands 17, 22-28.*

BEUTLER, F.J. (1970): Alias-free randomly timed sampling of stochastic processes. *IEEE Trans. Inform. Theor. IT-16, 147-152.*

BICKEL, P.J., KLAASSEN, C.A.J., RITOV, Y. and WELLNER, J.A. (1993): *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins University Press, Baltimore.

CHIU, S. (1989): Detecting periodic components in a white Gaussian time series. *J. Roy. Statist. Soc. Ser. B 5, 249-259.*

DE CAT, P. and AERTS, C. (2002): A study of bright southern slowly pulsating B stars, II. The intrinsic frequencies. *Astronom. Astrophys. 393, 965-982.*

DEEMING, T.J. (1975): Fourier analysis with unequally-spaced data. *Astrophys. Space Sci. 36, 137-158.*

DEPOY, D.L., PEPPER, J., POGGE, R.W., STUTZ, A., PINSONNEAULT, M. and SELLGREN, K. (2004): The nature of the variable galactic center source IRS 16SW. *Astrophys. J. 617, 1127-1130.*

DERUE, F. and 49 OTHER AUTHORS (2002): Observation of periodic variable stars towards the Galactic spiral arms by EROS II. *Astronom. Astrophys. 389, 149-161.*

EYER, L. and CUYPERS, J. (2000): Predictions on the number of variable stars for the GAIA space mission and for surveys as the ground-based International Liquid Mirror Telescope. In: L. Szabados and D. W. Kurtz (Eds.): *The Impact of Large Scale Surveys on Pulsating Star Research.* ASP Conference Series, vol. 203. Astronomical Society of the Pacific, San Francisco, 71-72.

EYER, L., GENTON, M.G. (1999): Characterization of variable stars by robust wave variograms: an application to Hipparcos mission. *Astron. Astrophys, Supp. Ser. 136, 421-428.*

FERRAZ-MELLO, S. (1981): Estimation of periods from unequally spaced observations. *Astronom. J. 86, 619-624.*

FISHER, R.A. (1929): Tests of significance in harmonic analysis. *Proc. Roy. Soc. London Ser. A 125, 54-59.*

FROST, O.L. (1976): Power-spectrum estimation. In: G. Tacconi (Ed.): *Aspects of Signal Processing with Emphasis on Underwater Acoustics*, Part I. Reidel, Dordrecht, 12-162.

GENTON, M.G. and HALL, P. (2007): Statistical inference for evolving periodic functions. *Roy. Statist. Soc. Ser. B 69, 643-657.*

GOOD, G.A. (2003): *Observing Variable Stars.* Springer, New York.

HALL, P. and LI, M. (2006): Using the periodogram to estimate period in nonparametric regression. *Biometrika 93, 411-424.*

HALL, P., REIMANN, J. and RICE, J. (2000): Nonparametric estimation of a periodic function. *Biometrika 87, 545-557.*

HALL, P. and YIN, J. (2003): Nonparametric methods for deconvolving multiperiodic functions. *J. Roy. Stat. Soc. Ser. B 65, 869-886.*

HANNAN, E.J. (1973): The estimation of frequency. *J. Appl. Prob. 10, 510-519.*

HANNAN, E.J. (1974): Time-series analysis. System identification and time-series analysis. *IEEE Trans. Automatic Control AC-19, 706-715.*

HART, J.D., KOEN, C. and LOMBARD, F. (2004): An analysis of pulsation periods of long-period variable stars. *J. Roy. Statist. Soc. Ser. C 56, 587-606.*

HOFFMEISTER, C., RICHTER, G. and WENZEL, W. (1985): *Variable Stars.* Springer, Berlin.

KABATH, P., EIGMÜLLER, P., ERIKSON, A, HEDELT, P., RAUER, H., TITZ, R. and WIESE, T. (2007): Characterization of COROT Target Fields with BEST: Identification of Periodic Variable Stars in the IR01 Field. *Astronom. J. 134, 1560-1569.*

KOEN, C. (2005): Statistics of O-C diagrams and period changes. In: C. Sterken (Ed.): *The Light-Time Effect in Astrophysics.* ASP Conference Series vol. 335. Astronomical Society of the Pacific, San Francisco, 25-36.

LAFLER, J. and KINMAN, T.D. (1965): An RR Lyrae survey with the Lick 20-inch astrograph II. The calculation of RR Lyrae period by electronic computer. *Astrophys. J. Suppl. Ser. 11, 216-222.*

LANZA, A.F., RODONÒ, M. and PAGANO, I. (2004): Multiband modelling of the Sun as a variable star from VIRGO/SoHO data. *Astronom. Astrophys. 425, 707-717.*

LOMB, N.R. (1976): Least-squares frequency analysis of unequally spaced data. *Astrophys. Space Sci. 39, 447-462.*

MAFFEI, P., CIPRINI, S. and TOSTI, G. (2005): Blue and infrared light curves of the mysterious pre-main-sequence star V582 Mon (KH 15D) from 1955 to 1970. *Monthly Not. Roy. Astronom. Soc. 357, 1059-1067.*

MCDONALD, J.A. (1986): Periodic smoothing of time series. *SIAM J. Sci. Statist. Comput. 7, 665-688.*

NORTH, G. (2005): *Observing Variable Stars, Novae, and Supernovae.* Cambridge University Press, Cambridge, UK.

NORTON, A.J. and 19 OTHER AUTHORS (2008): New periodic variable stars coincident with ROSAT sources discovered using SuperWASP. *Astronom. Astrophys.*, to appear.

PERCY, J.R. (2007): *Understanding Variable Stars.* Cambridge University Press, Cambridge, UK.

PISARENKO, V. (1973): The retrieval of harmonics from a covariance function. *Geophys. J. Roy. Astronom. Soc. 33, 347-366.*

PRIGARA, F.V. (2007): Radial solitary waves in periodic variable stars. Manuscript.

PROTOPAPAS, P., GIAMMARCO, J.M., FACCIOLI, L., STRUBLE, M.F., DAVE, R. and ALCOCK, C. (2006): Finding outlier light curves in catalogues of periodic variable stars. *Monthly Not. Roy. Astronom. Soc. 369, 677-696.*

QUINN, B.G. (1999): A fast efficient technique for the estimation of frequency: interpretation and generalisation. *Biometrika 86, 213-220.*

QUINN, B.G. and FERNANDES, J.M. (1991): A fast efficient technique for the estimation of frequency. *Biometrika 78, 489-497.*