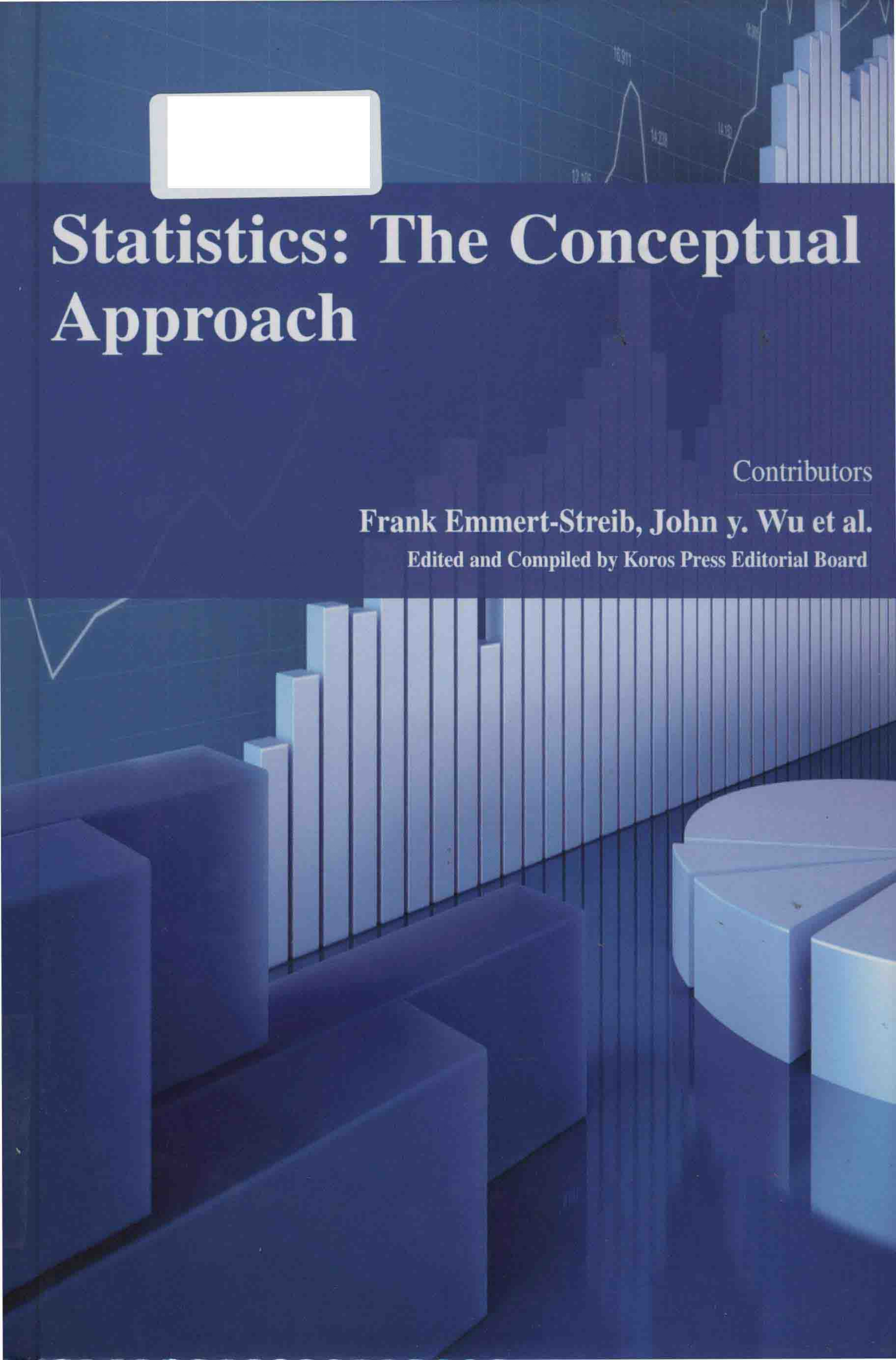

Statistics: The Conceptual Approach

Contributors

Frank Emmert-Streib, John y. Wu et al.

Edited and Compiled by Koros Press Editorial Board

Statistics: The Conceptual Approach

Contributors

Frank Emmert-Streib, John y. Wu et al.

KOROS PRESS

KOROS PRESS LIMITED

Statistics: The Conceptual Approach

Contributors: Frank Emmert-Streib, John y. Wu et al.

Published by Koros Press Limited

www.korospress.com

United Kingdom

Notice

Contributors, whose names have been given on the book cover, are not associated with the Publisher. The editors and the Publisher have attempted to trace the copyright holders of all material reproduced in this publication and apologise to copyright holders if permission has not been obtained. If any copyright holder has not been acknowledged, please write to us so we may rectify.

Reasonable efforts have been made to publish reliable data. The views articulated in the chapters are those of the individual contributors, and not necessarily those of the editors or the Publisher. Editors and/or the Publisher are not responsible for the accuracy of the information in the published chapters or consequences from their use. The Publisher accepts no responsibility for any damage or grievance to individual(s) or property arising out of the use of any material(s), instruction(s), methods or thoughts in the book.

Statistics: The Conceptual Approach

ISBN: 978-1-78569-109-6

British Library Cataloguing in Publication Data
A CIP record for this book is available from the British Library

Printed in the United Kingdom

Statistics: The Conceptual Approach

List of Contributors

Wang Fang Xue
Computational Biology and Machine Learning, Center for Cancer Research and Cell Biology, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, Belfast, United Kingdom

Hui Li
Computational Biology and Machine Learning, Center for Cancer Research and Cell Biology, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, Belfast, United Kingdom

Frank Emmert-Streib
Computational Biology and Machine Learning, Center for Cancer Research and Cell Biology, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, Belfast, United Kingdom

John y. Wu
Pathology Department, Visalia Pathology Medical Group, Visalia, USA

Diana Bílková
University of Economics, Faculty of Informatics and Statistics, Department of Statistics and Probability, Prague, Czech Republic
Department of Information Technology and Analytical Methods, University of Business, Prague, Czech Republic

Timur Zubayraev
Faculty of Computational Mathematics and Cybernetics, Moscow State University, Moscow, Russia

Mohan Delampady
Statistics and Mathematics Unit, Indian Statistical Institute, Bangalore, India

Giovanni Girone
Faculty of Economics, University of Bari, Bari, Italy

Antonella Nannavecchia
Faculty of Economics, University of Bari, Bari, Italy

Amir Nobari
School of Engineering, University of Liverpool, the Quadrangle, Brownlow Street, Liverpool, L69 3GH, United Kingdom

Huajiang Ouyang
School of Engineering, University of Liverpool, the Quadrangle, Brownlow Street, Liverpool, L69 3GH, United Kingdom

Paul Bannister
Jaguar Land Rover, Abbey Road, Whitley, Coventry, CV3 4LF, United Kingdom

Aleksandar Mijatović
Department of Mathematics, Imperial College London, United Kingdom

Martijn Pistorius,
Department of Mathematics, Imperial College London, United Kingdom

F. Brouers
Department of Chemical Engineering, Liege University, Liege, Belgium

Preface

Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data. A conceptual approach, built around common issues and problems rather than statistical techniques, allows to understand the conceptual nature of statistical procedures and to focus more on cases and examples of analysis. This book contains nine chapters. The purpose of first chapter is to introduce a new measure of complexity we call statistic complexity that is not only different to all other complexity measures introduced so far, but also connects directly to statistics, specifically, to statistical inference. Second chapter describe about statistics–it's utility or risks. The L-moments and TL-moments as an alternative tool of statistical data analysis has been presented in third chapter. Fourth chapter deals with asymptotic analysis for U-statistics and its application to Von Mises statistics. The aim of fifth chapter is to discuss the minimum description length methods in Bayesian model selection. The distribution of the concentration ratio for samples from a uniform population has been described in sixth chapter. Seventh chapter gives the details on statistics of complex eigen-values in friction-induced vibration. Eight chapter focuses on asymptotic independence of three statistics of maximal segmental scores. Statistical foundation of empirical isotherms has been outlined in the last chapter.

Contents

Chapter 1

STATISTIC COMPLEXITY: COMBINING KOLMOGOROV COMPLEXITY WITH AN ENSEMBLE APPROACH

Frank Emmert-Streib

Computational Biology and Machine Learning, Center for Cancer Research and Cell Biology, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, Belfast, United Kingdom

ABSTRACT

The evaluation of the complexity of an observed object is an old but outstanding problem. In this paper we are tying on this problem introducing a measure called *statistic complexity*.

INTRODUCTION

Complex systems is the study of interactions of simple building blocks that result in a collective behavior or properties absent in the elementary components of the system itself. Due to the fact that this problem does not fit into one of the traditional research fields, it is connected to various of these, for instance physics, biology, chemistry or econometrics [1]–[5]. Many measures, properties or characteristics of a multitude of different complex systems from these fields has been studied to date [6]–[8], however, the *complexity* of an object may have received the most attention. This property of complex systems has fascinated generations of scientists [9]–[11]trying to quantify such a notation. Very coarsely speaking, *an object is said to be 'complex' when it does not match patterns regarded as simple*, as López-Ruiz et al. [12] describe it in their article. Over the last decades, many approaches have been suggested to define the complexity of an object quantitatively [9], [11], [13]–[19]. An intrinsic problem with such a measure is that there are various ways to perceive and, hence, characterize complexity leading to complementing complexity measures [20]. For example, Kolmogorov complexity [9], [11],[21] is based on algorithmic information theory considering objects as individual symbol strings, whereas the measures *effective measure complexity* (EMC) [17], *excess entropy* [22],*predictive information* [23] or *thermodynamic depth* [18] relate objects to random variables and are ensemble based. Interestingly, despite considerable differences among all these complexity measures $\mathcal{M}$ they all have in common that they assign a complexity value to each individual object x' under consideration, $C_{\mathcal{M}}(x')$. In this paper we will assume that x' corresponds to a string sequence of a certain length and its components assume values from a certain domain, e.g., $\mathcal{A} = \{0, 1\}$ or $\mathcal{A} = [0, 1]$. It is of importance to note that there is a conceptually different measure recently introduced by Vitányi et al. that evaluates the complexity *distance* among two objects x' and x'' instead of their absolute values. This measure is called the *normalized compression distance* (NCD) [24], $NCD(x', x'')$, and is based on Kolmogorov complexity[10].

The purpose of this paper is to introduce a new measure of complexity we call *statistic complexity* that is not only different to all other complexity measures introduced so far, but also connects directly to statistics, specifically, to statistical inference [25], [26]. More precisely, we introduce a complexity measure with the following properties. First, the measure is bivariate comparing two objects, corresponding to pattern generating processes, on the basis of the*normalized compression distance* with each other. Second, this measure provides the quantification of an error that could have encountered by comparing samples of finite size from the underlying processes. Hence, the *statistic complexity* provides a statistical quantification of the statement 'X is similarly complex as Y'.

This paper is organized as follows. In the next section we describe the general problem in more detail and introduce our complexity measure. Then we present numerical results and provide a discussion. We finish with conclusions and an outlook.

METHOD

Currently, a commonly acknowledged, rigorous mathematical definition of the complexity of an object is not available. Instead, when complexity measures are suggested they are normally assessed by their behavior with respect to three qualitative patterns, namely simple, random (chaotic) and complex patterns. Qualitatively, a complexity measure is considered *good* if: (1) the complexity of simple and random objects is less than the complexity value of complex objects [17], (2) the complexity of an object does not change if the system size changes. For example, Kolmogorov complexity has the desireable property to remain unchanged if the system size doubles, i.e., $C_K(x)=C_K(xx)$, however, it cannot distinguish random from complex pattern because in both cases the compressibility of an object is low resulting in high values of C_K. We want to add a third property to the above criteria: (3) A complexity measure should quantify the uncertainty of the complexity value. As motivation for this property we just want to mention that there is a crucial difference between an observed object x' and its generating process X [23]. If the complexity of X should be assessed, based on the observation x' only, this assessment may be erroneous. This error may stem from the limited (finite) size of observations. Also, the possibility of measurement errors would be another source derogating the ability of an error-free assessment. For this reason, the major objective of this article is to introduce a complexity measure possessing all three properties listed above that assesses the complexity classes of the underlying processes instead of individual objects.

We start by pointing out that criteria (1) provides a relative statement connecting different objects. That means the complexity of an object is always related to the complexity of another object [20] leading to relative statements like 'X is similarly complex as Y'. Hence, a numerical value $C(X)$ without knowledge of any other complexity value for other objects has no meaning at all. For reasons of mathematical rigor, we propose to include this implicit reference point into a proper definition of complexity. This implies that a fundamental complexity measure needs to be bivariate,

$C(X, Y)$, instead of univariate comparing two processes X and Y. As a side note, we remark that all complexity measures suggested so far we are aware of are univariate measures [13], [14], [16]-[18], [22], [23] with respect to the context set above, except for the normalized compression distance (NCD) [24], [27]. However, a practical problem of the NCD is that Kolmogorov complexity, on which it is based, is not computable but only upper semi-computable [27]. Li et al. introduced in [27] a normalized and universal metric called normalizedinformation distance (NID) which can be approximated by,

$$NCD(x,y) = \frac{C(xy) - min\{C(x), C(y)\}}{max\{C(x), C(y)\}}, \quad (1)$$

the normalized compression distance [27]. Here, $C(x)$ denotes the compression size of string x and $C(xy)$ the compression size of the concatenated stings x and y. Practically, the quantities $C()$ are obtained by compressors like gzip or bzip2, see [28], [29] for details.

Criteria (3) of a complexity measure stated above acknowledges the fact that an assessment of an object›s complexity cannot be without uncertainty or error in case only finite information about this object is available. That means, for a complexity measure to be applicable to real objects (rather than pure mathematical ones) it has to be statistic in order to deal appropriately with incomplete information. Based on these considerations, the *statistic complexity* measure we suggest is defined by the following procedure visualized in Fig. 1:

- Estimate the empirical distribution function $\hat{F}_{X.X}$ (We indicate estimated entities by $\hat{F}$and refer to the ensemble by F.) of the normalized compression distance from $n_1$samples, $S^{n_1}_{X.X} = \{x_i = \text{NCD}(x', x'') | x', x'' \sim X\}^{n_1}_{i=1}$, from objects x' and x'' of size m generated by process X (Here $x \sim X$ means that x is generated (or drawn) from process (distribution) X.).
- Estimate the empirical distribution function $\hat{F}_{X.Y}$ of the normalized compression distance from n_2 samples, $S^{n_2}_{X.Y} = \{y_i = \text{NCD}(x', y') | x' \sim X, y' \sim Y\}^{n_2}_{i=1}$, from objects x' and y' of size m from two different processes, X and Y.
- Determine $T = \sup_x |\hat{F}_{X.X}(x) - \hat{F}_{X.Y}(x)|$ and $p = \text{Prob}(T \leq t)$.
- Define, $C_S\left(S^{n_1}_{X.X}, S^{n_2}_{X.Y} | X, Y, m, n_1, n_2\right) := p$, as *statistic complexity*

This procedure corresponds to a two-sided, two-sample Kolmogorov-Smirnov (KS) test [30],[31] based on the normalized compression distance [24], [27] obtaining distances among observed objects. The *statistic complexity* corresponds to the p-value of the underlying null hypotheses, $H_0 : F_{X,X} = F_{X,Y}$, and, hence, assumes values in $[0,1]$. The null hypothesis is a statement about the null distribution of the test statistic $T = \sup_x |\hat{F}_{X,X}(x) - \hat{F}_{X,Y}(x)|$, and because the distribution functions are based on the normalized compression distances among objects x' and x'', drawn from the processes X and Y, this leads to a statement about the distribution of normalized compression distances. Hence, verbally, H_0 can be phrased as 'in average, the compression distance of objects from X to objects from Y equals the compression distance of objects only taken from X'. It is important to emphasize that this equality holds in *average* and, thus needs to be connected to two ensembles X and Y. If the alternative hypothesis, $H_1 : F_{X,X} \neq F_{X,Y}$, is true this equality does no longer hold implying differences in the underlying processes X and Y, leading to differences in the NCDs. From the formulation of the hypotheses, tested by the *statistic complexity*, it is apparent that we are following closely the guiding principle expressed by López-Ruiz et al. [12] as cited at the beginning of this paper, because C_S is intrinsically a comparative measure. As a side note regarding the choice of the null hypothesis we want to remark that substituting F_{XY} with F_{YY}may encounter problems in cases where the complexity value of objects in Y is systematically shifted compared to the complexity value of objects in X. In this case, the distributions $F_{XX}$and F_{YY} could be similar, although, the complexity of elements in X and Y are different. Practically, this may correspond to a pathological case rarely encountered in practice, however, conceptually, such a null hypothesis is apparently less stringent.

Regarding the notation and interpretation of the above procedure it is important to note the following. First, the entities x and y refer to values of the NCD. For example, $x = \text{NCD}(x', x'')$ whereas x' and x'' are observable objects that are identically and independently (iid) generated from a process X, $x', x'' \sim X$. Because x' and x'' are generated from the same process X, the resulting distribution function $F_{X,X}$ is only indexed by this process. The y entities are obtained similarly, however, in this case x' and y' are objects generated from two *different* processes, namely $x' \sim X$ and $y' \sim Y$

. For this reason the distribution function is indexed by these two processes, $F_{X,Y}$. Second, we use the notation, $x' \sim X$, to indicate that x' is generated from a process X, but also that x' is drawn from X. The first meaning is clear if thinking of X as a model for a complex system, e.g., a cellular automata or a stochastic process. The latter emphasizes the fact that such a process, even if deterministic, becomes random with respect to, e.g., random initial conditions and, hence, effectively is a stochastic process. Third, for reasons of conceptual simplicity we require all objects to have the same size m. This condition may be relaxed to allow objects of varying sizes but it may require additional technical consideration. On a technical note, the above defined *statistic complexity* has the very desirable property that the power reaches asymptotically 1 for $n_1 \rightarrow \infty$ and $n_2 \rightarrow \infty$ [32]. This means, for infinite many observations the error of the test to falsely accept the null hypotheses when in fact the alternative is true becomes zero. This limiting property is important to hold, because in this case all information about the system is available and, hence, it would be implausible if for such circumstances no error-free decision could be achieved. Formally, this property can be stated as $p \rightarrow 0$ for $n_1 \rightarrow \infty$ and $n_2 \rightarrow \infty$. Finally, we would like to note that despite the fact that *statistic complexity*is a statistical test, it borrows part of its strength from the NCD respectively Kolmogorov complexity on which this is based on. Hence, it unites various properties from very different concepts.

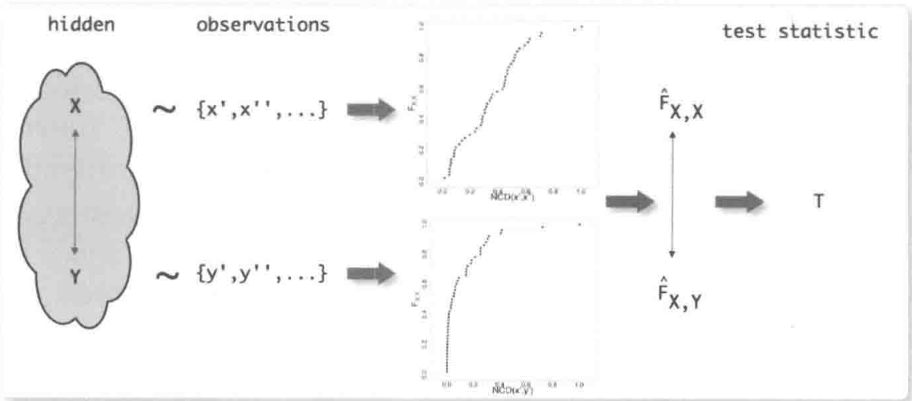

Figure 1. Visualization of the problem and the construction of the test statistic from observations.