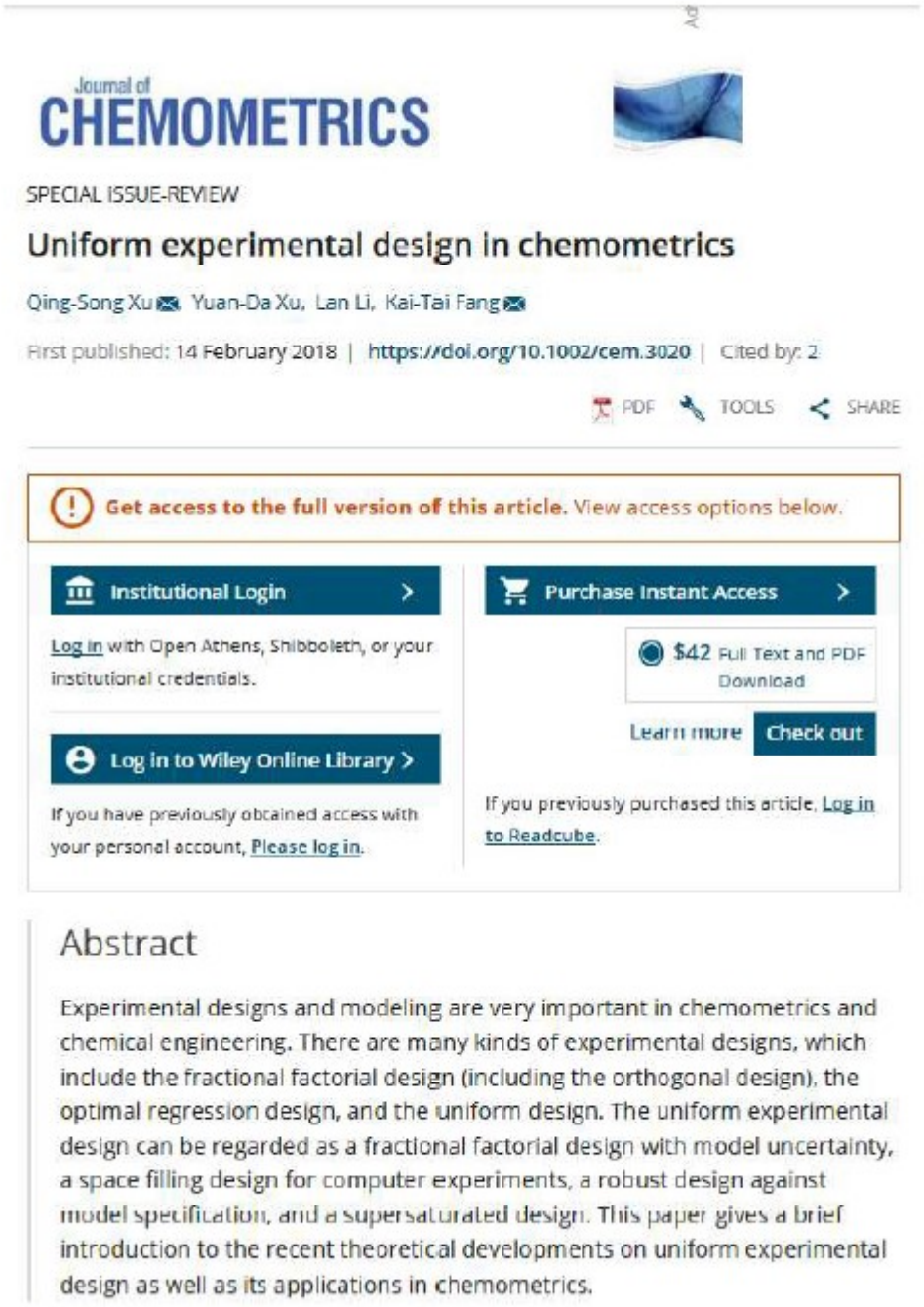
近三年统计专业团队发表文章首页扫描



Journal of CHEMOMETRICS

SPECIAL ISSUE-REVIEW

Uniform experimental design in chemometrics

Qing-Song Xu, Yuan-Da Xu, Lan Li, Kai-Tai Fang

First published: 14 February 2018 | https://doi.org/10.1002/cem.3020 | Cited by: 2

PDF TOOLS SHARE

Get access to the full version of this article. View access options below.

Institutional Login

Log in with Open Athens, Shibboleth, or your institutional credentials.

Log in to Wiley Online Library

If you have previously obtained access with your personal account, Please log in.

Purchase Instant Access

$42 Full Text and PDF Download

Learn more Check out

If you previously purchased this article, Log in to Readcube.

Abstract

Experimental designs and modeling are very important in chemometrics and chemical engineering. There are many kinds of experimental designs, which include the fractional factorial design (including the orthogonal design), the optimal regression design, and the uniform design. The uniform experimental design can be regarded as a fractional factorial design with model uncertainty, a space filling design for computer experiments, a robust design against model specification, and a supersaturated design. This paper gives a brief introduction to the recent theoretical developments on uniform experimental design as well as its applications in chemometrics.

Metrika (2018) 81:307–336
https://doi.org/10.1007/s00184-018-0644-5

New results on quaternary codes and their Gray map images for constructing uniform designs

A. M. Elsawah[1,2] · Kai-Tai Fang[2,3]

Received: 6 July 2017 / Published online: 5 February 2018

Abstract The research of developing efficient methodologies for constructing optimal experimental designs has been very active in the last decade. Uniform design is one of the most popular approaches, carried out by filling up experimental points in a determinately uniform fashion. Applications of coding theory in experimental design are interesting and promising. Quaternary codes and their binary Gray map images attracted much attention from those researching design of experiments in recent years. The present paper aims at exploring new results for constructing uniform designs based on quaternary codes and their binary Gray map images. This paper studies the optimality of quaternary designs and their two and three binary Gray map image designs in terms of the uniformity criteria measured by: the L_{cc}, wrap-around, symmetric, centered and mixture discrepancies. Strong relationships between quaternary designs and their two and three binary Gray map image designs are obtained, which can be used for efficiently constructing two-level designs from four-level designs and vice versa. The significance of this work is evaluated by comparing our results to the existing literature.

Keywords Coding theory · Gray map images · Quaternary design · Binary design · Uniform design · Uniformity criteria

✉ A. M. Elsawah
a_elsawah85@yahoo.com; amelsawah@uic.edu.hk; a.elsawah@zu.edu.eg

[1] Department of Mathematics, Faculty of Science, Zagazig University, Zagazig 44519, Egypt

[2] Division of Science and Technology, BNU-HKBU United International College, Zhuhai 519085, China

[3] The Key Lab of Random Complex Structures and Data Analysis, The Chinese Academy of Sciences, Beijing, China

JOURNAL OF APPLIED STATISTICS
https://doi.org/10.1080/02664763.2018.1545013

Taylor & Francis
Taylor & Francis Group

A catalog of optimal foldover plans for constructing U-uniform minimum aberration four-level combined designs

A. M. Elsawah [a,b] and Kai-Tai Fang [b,c]

[a] Faculty of Science, Department of Mathematics, Zagazig University, Zagazig, Egypt; [b] Division of Science and Technology, BNU-HKBU United International College, Zhuhai, People's Republic of China; [c] The Key Lab of Random Complex Structures and Data Analysis, The Chinese Academy of Sciences, Beijing, People's Republic of China

ABSTRACT

The foldover plan is a transformation map for adding a foldover design to the initial design, thus resulting in a combined design which can be used for breaking the links between aliased effects. This paper discusses the optimality of foldover plans for four-level designs via the most common criteria: the generalized word-length pattern (GWLP), Lee discrepancy (LD), wrap-around discrepancy (WD), centered discrepancy (CD) and mixture discrepancy (MD). We prove that the LD, WD and GWLP are equivalent for: any initial design and the corresponding foldover designs; any foldover design and its complementary foldover design; and any combined design and its complementary combined design. However, these interesting properties do not necessarily take place in the cases of the CD and MD. New analytical expressions and lower bounds of these discrepancies are given for initial and combined designs, which can be used as benchmarks for constructing uniform designs. For illustration of the usage of our theoretical results, a catalog of optimal foldover plans for constructing U-uniform minimum aberration four-level combined designs that involve $2 \leq m \leq 10$ factors with $8 \leq n \leq 52$ run is tabulated, which can be used for investigating either qualitative or quantitative factors.

ARTICLE HISTORY
Received 21 May 2017
Accepted 1 November 2018

KEYWORDS
Foldover plan; Foldover design; combined design; optimal foldover plan; discrepancy; generalized word-length pattern

MSC
62K05; 62K15

1. Introduction

A symmetric full factorial experiment (FuFE) is an experiment whose $(n; s^m)$-design consists of m factors, each with s levels, and whose runs n take on all possible combinations of levels across all factors, i.e. $n = s^m$. A FuFE delivers better information about the system under study and permits the researcher to estimate all possible effects of the input variables on the response variables. Because n grows exponentially with m, FuFEs are often too difficult to use for real-life projects. A fractional factorial experiment (FrFE) reduces efforts by using a fraction of a FuFE in terms of n and resources. A FrFE is constructed by carefully choosing a fraction of n of a FuFE. The fraction is chosen so as to achieve the

CONTACT A. M. Elsawah a_elsawah85@yahoo.com, amelsawah@uic.edu.hk, a.elsawah@zu.edu.eg Faculty of Science, Department of Mathematics, Zagazig University, Zagazig 44519, Egypt; Division of Science and Technology, BNU-HKBU United International College, Zhuhai 519085, People's Republic of China

Comput Stat (2018) 33:933–965
https://doi.org/10.1007/s00180-017-0778-3

ORIGINAL PAPER

Choice of optimal second stage designs in two-stage experiments

A. M. Elsawah[1,2]

Received: 19 January 2017 / Accepted: 7 November 2017 / Published online: 17 November 2017

Abstract In real-life projects, in order to obtain precious information about the process, we often partition the experiment into two stages with equal size. The main purpose of this article is to study how to choose the first stage experimental designs (FSED) and the second stage experimental designs (SSED) to construct uniform or at least good approximation to uniform (GATU) two-stage experimental designs (TSED) that involve a mixture of $\omega_1 \geq 1$ factors with $\mu_1 \geq 2$ levels and $\omega_2 \geq 1$ factors with $\mu_2 \geq 2$ levels whether regular or nonregular. Through theoretical justification, this paper proves that the SSED is uniform (GATU) if and only if the FSED is uniform (GATU), the TSED is uniform (GATU) if and only if its corresponding complementary TSED is uniform (GATU), and the TSED is uniform or at least GATU if and only if the FSED is uniform.

Keywords Second stage design · Second stage map · Two-stage design · Uniform design · Optimal design · Complementary design

1 Introduction

Two-stage sequential experimentation is often essential to obtain valuable information about the system under consideration by increasing the precision of factorial effect estimates. Two-stage experiments have the ability of estimating the main effects of the factors and the significant interactions (lower order) between them, thus they have been

✉ A. M. Elsawah
a_elsawah85@yahoo.com; amelsawah@uic.edu.hk ; a.elsawah@zu.edu.eg

[1] Department of Mathematics, Faculty of Science, Zagazig University, Zagazig 44519, Egypt

[2] Division of Science and Technology, BNU-HKBU United International College, Zhuhai 519085, China

Journal of Systems Science and Complexity
April 2019, Volume 32, Issue 2, pp 681–708 | Cite as

Designing Uniform Computer Sequential Experiments with Mixture Levels Using Lee Discrepancy

Authors | Authors and affiliations

A. M. Elsawah

Article
First Online: 12 January 2019

Abstract

Computer experiments are constructed to simulate the behavior of complex physical systems. Uniform designs have good performance in computer experiments from several aspects. In practical use, the experimenter needs to choose a small size uniform design at the beginning of an experiment due to a limit of time, budget, resources, and so on, and later conduct a follow up experiment to obtain precious information about the system, that is, a sequential experiment. The Lee distance has been widely used in coding theory and its corresponding discrepancy is an important measure for constructing uniform designs. This paper proves that all the follow up designs of a uniform design are uniform and at least two of them can be used as optimal follow up experimental designs. Thus, it is not necessary that the union of any two uniform designs yields a uniform sequential design. Therefore, this article presents a theoretical justification for choosing the best follow up design of a uniform design to construct a uniform sequential design that involves a mixture of $\omega \geq 1$ factors with $\beta k \geq 2$, $1 \leq k \leq \omega$ levels. For illustration of the usage of the proposed results, a closer look is given at using these results for the most extensively used six particular cases, three symmetric and three asymmetric designs, which are often met in practice.

Keywords

Computer experiment | Lee discrepancy | Lee distance | lower bound | sequential design | uniform design

J Indian Soc Probab Stat (2018) 19:359–377
https://doi.org/10.1007/s41096-018-0049-9

RESEARCH ARTICLE

Asymptotic Theory of Dual Generalized Order Statistics from Heterogeneous Population

A. M. Elsawah[1,2] · Fatma Essawe[3,4] · Hui Zhao[3]

Accepted: 7 July 2018 / Published online: 16 July 2018

Abstract

The outcomes of several real-life experiments arise in descending order. Dual generalized order statistics (DGOS) have been introduced as a unification of several models of descendingly ordered random variables like reversed ordered order statistics, lower k-records and lower Pfeifer records. The asymptotic theory (AT) proceeds by assuming that it is possible (in principle) to keep collecting additional data, so that the sample size grows infinitely. Under this assumption, many results can be obtained that are unavailable for samples of finite size. The AT is widely used in various statistical approaches, such as ordered random variables, time series models, estimation, testing hypotheses and so on. While the AT of DGOS from homogeneous population, i.e., all of the data points come from the same distribution, has been soundly investigated, no research has been devoted to this problem for heterogeneous population, i.e., the data points come from more than one distribution. This paper gives a closer look at the AT of DGOS based on data from a finite mixture of distributions normalized by the same continuous strictly monotonic sequence or a mixture of continuous strictly monotonic sequences.

Keywords Dual generalized order statistics · Asymptotic theory · Homogeneous population · Heterogeneous population · Normalization

Mathematics Subject Classification 62G30 · 60F05 · 62E20

✉ A. M. Elsawah
a_elsawah85@yahoo.com; amelsawah@uic.edu.hk; a.elsawah@zu.edu.eg

1 Department of Mathematics, Faculty of Science, Zagazig University, Zagazig 44519, Egypt

2 Division of Science and Technology, BNU-HKBU United International College, Zhuhai 519085, China

3 Faculty of Mathematics and Statistics, Central China Normal University, Wuhan 430079, China

4 Department of Statistics and Computer, Faculty of Applied Science, Read Sea University, Port Sudan, Sudan

ProbStat Forum, Volume 11, July 2018, Pages 104–116
ISSN 0974-3235

ProbStat Forum is an e-journal.
For details please visit www.probstat.org.in

Extreme value theory of mixture generalized order statistics

A M Elsawah[a], Gajendra K Vishwakarma[b], Zhongquan Tan[c]

[a]*Department of Mathematics, Faculty of Science, Zagazig University, Zagazig 44519, Egypt and Division of Science and Technology, BNU-HKBU United International College, Zhuhai 519085, China.*
[b]*Department of Applied Mathematics, Indian Institute of Technology, Dhanbad 826004, India.*
[c]*College of Mathematics, Physics and Information Engineering, Jiaxing University, Jiaxing 314001, China.*

Abstract. Most statistical approaches assume that all of the data points come from the same distribution. However, in real-life applications the data points come from more than one distribution with no information to identify which observation goes with which distribution. In such cases, the classical extreme value theory can not help us. In this paper, we investigate the extreme value theory of data from more than one distribution based on generalized order statistics under continuous strictly monotone normalization. This paper investigates the asymptotic behaviors of upper and lower extremes generalized order statistics based on a random sample drawn from a finite mixture of distributions normalized by the same continuous strictly monotonic sequence or a mixture of continuous strictly monotonic sequences. For illustration of the usage of our theoretical results, these asymptotic behaviors under linear and power normalization are studied with examples.

1. Introduction

1.1. Generalized order statistics

Let $\{X_j : j \in N\}$ be a sequence of independent and identically distributed random variables with common probability density function f and distribution function $\mathcal{F}$. If the first n random variables are arranged in ascending order of magnitude and written as $X_{1:n} < X_{2:n} < \ldots < X_{n:n}$, we call them ordinary order statistics. Generalized order statistics have been introduced by Kamps (1995) as a unification of several models of ascendingly ordered random variables. The generalized order statistics $X_{1:n}^{(\tilde{m},k)}$, $X_{2:n}^{(\tilde{m},k)}$,..., $X_{n:n}^{(\tilde{m},k)}$ are defined by their probability density function, which is given on the cone $\{(x_1,...,x_n) : x_0 = \mathcal{F}^{-1}(0) < x_1 \leq ... \leq x_n < \mathcal{F}^{-1}(1) = x^0\}$ as follows

$$f_{1,2,...,n:n}^{(\tilde{m},k)}(x_1,...,x_n) = kf(x_n)(1-\mathcal{F}(x_n))^{k-1}\prod_{i=1}^{n-1}\theta_i f(x_i)(1-\mathcal{F}(x_i))^{m_i},$$

where $\theta_1,...,\theta_n$ are defined by $\theta_n = k > 0$ and $\theta_j = k + n - j + \sum_{i=j}^{n-1} m_i > 0$, $j = 1,2,...,n-1$ and $\tilde{m} = (m_1,...,m_{n-1}) \in \Re^{n-1}$.

2010 *Mathematics Subject Classification.* 62G30, 60F05, 62E20.
Keywords. Mixture distributions; Extreme value theory; Generalizes order statistics; Linear normalization; Power normalization; Domains of attraction.
Received: 31 December 2017; Revised: 03 April 2018, Re-revised 07 May 2018; Accepted: 13 June 2018.
Elsawah-UIC Grants (Nos: R201409, R201712 and R201810) & the Zhuhai Premier Discipline Grant and Tan-National Science Foundation of China (No. 11501250) & Natural Science Foundation of Zhejiang Province of China (No. LQ14A010012).
Email address: a_elsawah86@yahoo.com, amelsawah@uic.edu.hk, a.elsawah@zu.edu.eg (A M Elsawah)

Received: 25 May 2018 | Revised: 18 October 2018 | Accepted: 23 October 2018

DOI: 10.1002/bimj.201800148

RESEARCH PAPER

Biometrical Journal

Tuning model parameters in class-imbalanced learning with precision-recall curve

Guang-Hui Fu[1] | Lun-Zhao Yi[2] | Jianxin Pan[3]

[1]School of Science, Kunming University of Science and Technology, Kunming, P. R. China

[2]Yunnan Food Safety Research Institute, Kunming University of Science and Technology, Kunming, P. R. China

[3]School of Mathematics, The University of Manchester, Manchester, UK

Correspondence:
Jianxin Pan, School of Mathematics, The University of Manchester, Manchester, M13 9PL, UK.
Email: jianxin.pan@manchester.ac.uk

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 11761041, 21465016, 21775058

Abstract

An issue for class-imbalanced learning is what assessment metric should be employed. So far, precision-recall curve (PRC) as a metric is rarely used in practice as compared with its alternative of receiver operating characteristic (ROC). This study investigates the performance of PRC as the evaluating criterion to address the class-imbalanced data and focuses on the comparison of PRC with ROC. The advantages of PRC over ROC on assessing class-imbalanced data are also investigated and tested on our proposed algorithm by tuning the whole model parameters in simulation studies and real data examples. The result shows that PRC is competitive with ROC as performance measurement for handling class-imbalanced data in tuning the model parameters. PRC can be considered as an alternative but effective assessment for preprocessing (such as variable selection) skewed data and building a classifier in class-imbalanced learning.

KEYWORDS

class imbalance, measurement, parameter tuning, precision-recall curve, receiver operating characteristic

1 | INTRODUCTION

Class-imbalanced data arise from many fields of recent scientific discoveries such as rare disease diagnostics and medical imaging identification where the prevalence is typically very low (Guo et al., 2017). To date, class-imbalanced learning is a relatively new challenge and there are still many questions to be further investigated. In this study, we focus on how to use precision-recall curve (PRC) as an assessment measurement for statistical modeling of class-imbalanced data.

So far, the widely used assessment measures for classification are classification accuracy and receiver operating characteristic (ROC) plot. However, both of them are problematic in the presence of class imbalance. Classification accuracy is the percentage of true results among the total number of cases in a testing set. A classifier that maximizing the accuracy is ineffective for extremely class-imbalanced data. For example, in the case where 99% of all the data are from one class and the rest 1% belong to the other class. A classifier can produce an accuracy of 99% even though it simply predicts all the data as the majority. Another disadvantage of accuracy is that it always assumes that the classifier will operate on data drawn from the same distribution as the training data. However, the class imbalance present in the training set is not always the one that is encountered throughout the operating life of the classifier. The information related to the class imbalance of data is often not known (Provost, 2000).

ROC is often considered as the standard measurement for assessing the performance of a classifier in many fields. Its properties and associated indices have extensively been studied (Fawcett, 2006; Ma & Huang, 2005; Zhou et al., 2012). However, ROC curve tends to provide an overly optimistic view of the performance of an algorithm for class-imbalanced data (Davis & Goadrich, 2006), and it is insensitive to the change in class distribution. For example, if the proportion of positive to negative instances changes in a test set, ROC curve will not change. But, the change in class distribution often leads to change of the true and false positive rates (Webb & Ting, 2005). Realizing its disadvantages in dealing with class-imbalanced data, precision-recall curve has become a basis for assessing classification methods on class-imbalanced data (Ozenne, Subtil, & Maucort-Boulch, 2015;

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 19, 2018, accepted April 20, 2018, date of publication April 26, 2018, date of current version May 24, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2830350

A Tighter Set-Membership Filter for Some Nonlinear Dynamic Systems

ZHIGUO WANG[1], XIAOJING SHEN[1], YUNMIN ZHU[1], AND JIANXIN PAN[2]
[1]Department of Mathematics, Sichuan University, Chengdu 610064, China
[2]School of Mathematics, The University of Manchester, Manchester M13 9PL, U.K.

Corresponding author: Xiaojing Shen (shenxj@scu.edu.cn)

This work was supported in part by NSFC under Grant 61673282, in part by the open research funds of BACC-STAFDL of China under Grant 2015afdl010, and in part by PCSIRT under Grant PCSIRT16R53.

ABSTRACT In this paper, we propose a tighter set-membership filter for some nonlinear dynamic systems by using an analytic method and a boundary sampling technique. The nonlinear dynamic systems can be linearized about the current estimate, then the remainder term is bounded in real time by an optimization ellipsoid, other than *a priori* remainder bound. For a 2-D radar system and a quadratic system, some regular properties can be derived for the remainder term, which helps us obtain a tighter bounding ellipsoid to cover the remainder. Moreover, the prediction step and the measurement update step are derived based on the recent optimization method and the on-line bounding ellipsoid of the remainder, so that a tighter set-membership filter can be achieved. The numerical examples demonstrate the effectiveness of the proposed filter.

INDEX TERMS Nonlinear dynamic systems, set-membership filter, randomization, semi-infinite optimization, target tracking.

I. INTRODUCTION

Filtering techniques for dynamic systems are widely used in practiced fields such as target tracking, signal processing, automatic control, computer vision [1]–[4]. The Kalman filter is a fundamental tool for solving a broad class of filtering problems with linear dynamic systems. It is well known that the extended Kalman filter (EKF) can be used to handle the nonlinear dynamic systems, which is based on local linear approximation of the nonlinear system with the higher order term ignored. Most recently, [5] proposes a box particle filter to analyze interval data using interval analysis and constraint satisfaction techniques. The advantage of the box particle filter over the standard particle filter is its reduced computational complexity [6]. However, most of Monte Carlo filtering techniques [7], [8] assume that the probability density functions of the state noise and measurement noise are known.

Actually, when the underlying probabilistic assumptions are not realistic (e.g., the main perturbation may be deterministic), it is more natural to assume that the state noise and measurement noise are unknown but bounded [9], then [10] proposed set-membership estimation technique. The idea of propagating bounding ellipsoids (or boxes, polytopes, simplexes, parallelotopes, and polytopes) for systems with bounded noises has also been extensively explored, for example, see recent papers [11]–[15] and references therein. Most of these methods concentrate on the linear dynamic systems.

The set-membership filtering for nonlinear dynamic systems is considered to be a difficult problem. Based on ellipsoid-bounded, fuzzy-approximated or Lipschitz-like nonlinearities, several studies have been made [16]–[18]. These studies assume that the ellipsoid bounds, the coefficients of fuzzy-approximation or Lipschitz constants are known before filtering, limiting their use in real-time implementation. For example, for a typical nonlinear dynamic system in a radar, the bounds of the remainder depend on the past estimates, so that they cannot be obtained before filtering. References [19] and [20] develop nonlinear set-membership filters that can estimate the bounding ellipsoid of nonlinearities in real-time, which are called the extended set-membership filter (ESMF) and set-valued nonlinear filter (SVNF), respectively. Actually, if the remainder is bounded by using a tighter ellipsoid and some recent advanced optimization techniques for filtering, we should be able to derive a tighter set-membership filtering for the nonlinear dynamic system.

In order to guarantee the on-line usage of the set membership filter for nonlinear dynamic systems, the nonlinear dynamic systems are linearized about the current estimate,

International Journal of Heat and Mass Transfer 124 (2018) 761–768

Contents lists available at ScienceDirect

International Journal of Heat and Mass Transfer

journal homepage: www.elsevier.com/locate/ijhmt

Extraction and evolution of bubbles attributes in a two-phase direct contact evaporator

Qingtai Xiao [a,b], Qin Gao [c], Jing Zhang [d], Jianxin Xu [a,d,*], Jianxin Pan [e,*], Hua Wang [a,f,*]

[a] State Key Laboratory of Complex Nonferrous Metal Resources Clean Utilization, Kunming University of Science and Technology, Kunming 650093, PR China
[b] School of Mathematical and Statistical Sciences, The University of Texas Rio Grande Valley, Edinburg, TX 78541, USA
[c] Faculty of Science, Kunming University of Science and Technology, Kunming 650500, PR China
[d] Quality Development Institute, Kunming University of Science and Technology, Kunming 650093, PR China
[e] School of Mathematics, University of Manchester, Manchester M13 9PL, UK
[f] Faulty of Metallurgical and Energy Engineering, Kunming University of Science and Technology, Kunming 650093, PR China

ARTICLE INFO

Article history:
Received 30 November 2017
Received in revised form 9 March 2018
Accepted 1 April 2018
Available online 6 April 2018

Keywords:
Direct-contact evaporation
Dispersion and distribution
Bubbles uniformity
Shape feature parameters
Image analysis

ABSTRACT

Understanding the bubble regimes is a fundamental step toward conducting heat transfer enhancement. The non-invasive measurement of mixing inside a direct-contact heat transfer process, using a direct video imaging technology, provides powerful opportunities for characterising the visual observations of the phenomena and quantifying the process complexities previously. Experimental bubble shape feature parameters were obtained by means of the photographic recording technique for a direct-contact evaporator. Four design factors with three levels respectively were analysed for the mixing system that involves the exchange of heat between two immiscible fluids (continuous and dispersed phases). Using the Ripley's *K* function, new results are presented for two-phase flow mixing which can distinguish differences in the mixing behavior of dispersed phase. In all cases considered, quantitative comparisons of the evolution curves representing different experimental conditions were conducted with reported experimental data. Following the local mixing curve, the current results can also be processed to provide the mixing time found to be in good agreement with available data. The relationship between shape feature parameters of bubbles and volumetric heat transfer coefficient was found to be highly independent on experimental design parameters.

1. Introduction

Although boiling is a complex process, it is a very efficient mode of heat transfer in various heat exchange systems [1]. Direct-contact boiling evaporators are widely employed as the chemical reactors in the industrial processes [2,3]. The mixing quality and the critical parameters (such as mixing uniformity of bubbles, droplets, or particles, local bubble size distribution, gas holdup, interfacial area, etc.) in reactor design and control have a great influence on the performance of contactors [4,5]. Direct imaging technology is an effective and convenient method for the estimation of those critical parameters [6,7]. To the best of our knowledge, many researchers have carried out this study thoroughly using numerous techniques including Doppler anemometry techniques [8], tomographic techniques [9], invasive probe techniques, and direct imaging techniques [10]. Additionally, image analysis with advanced mathematical methods, regarded as a normal practice, is gaining importance for object identification [11,12].

On the one hand, mixing uniformity of objects (bubbles, droplets, or particles) has been considered [13,14]. The purpose of mixing is to obtain homogeneous [15]. It has a decisive impact on the overall performance of reaction processes. There is therefore an increased desire for measuring and comparing mixing performance [16,17]. The efficient evaluation of mixing uniformity is also required in the boiling heat transfer process which is one of the most efficient kinds of heat transfer processes widely used in numerous engineering systems. The resulting improvement heat transfer performance is believed on how uniformly the discrete phase is mixed into the continuous phase. The bubble detection problem is not easy to solve because the bubbles are not transparent when imaging inside a same industrial process, which causes the bubble appearance to vary. For granular materials, the nonuniformity of porosity distribution within a specimen was evaluated

* Corresponding authors at: State Key Laboratory of Complex Nonferrous Metal Resources Clean Utilization, Kunming University of Science and Technology, Kunming 650093, PR China (J. Xu, H. Wang) and School of Mathematics, University of Manchester, Manchester M13 9PL, UK (J. Pan).
E-mail addresses: qingtaixiao2016@kmust.edu.cn (Q. Xiao), xujianxina@163.com (J. Xu), jianxin.pan@manchester.ac.uk (J. Pan), wanghua65@163.com (H. Wang).

https://doi.org/10.1016/j.ijheatmasstransfer.2018.04.002
0017-9310/

Scandinavian Journal of Statistics Theory and Applications

Scandinavian Journal of Statistics
doi: 10.1111/sjos.12314

Joint Modelling of Survival and Longitudinal Data with Informative Observation Times

HONGSHENG DAI
Department of Mathematical Sciences, University of Essex
JIANXIN PAN
School of Mathematics, University of Manchester

ABSTRACT. In this paper, we consider the joint modelling of survival and longitudinal data with informative observation time points. The survival model and the longitudinal model are linked via random effects, for which no distribution assumption is required under our estimation approach. The estimator is shown to be consistent and asymptotically normal. The proposed estimator and its estimated covariance matrix can be easily calculated. Simulation studies and an application to a primary biliary cirrhosis study are also provided.

Key words: Cox model, informative observation times, log-normal distribution, longitudinal data, multistate models

1. Introduction

The motivation for this paper arose from a primary biliary cirrhosis (PBC) study (Murtaugh *et al.*, 1994). The PBC is a chronic, fatal, but rare liver disease characterized by inflammatory destruction of the small bile ducts within the liver, which eventually leads to cirrhosis of the liver. Patients often present abnormalities in their blood tests, such as elevated and gradually increased serum bilirubin. The research interest is to study how the drug D-penicillamine affects event times and how the patterns of time courses of bilirubin levels affect death due to PBC. Patients in this study will have their blood tests roughly at six months, one year and annually thereafter. Longitudinal measurements (such as bilirubin levels) will be collected at these time points. These predetermined time points are independent of the longitudinal measurements; however, some longitudinal observations may be observed at an 'extra' visit, which is often undertaken unexpectedly because of worsening medical condition. Therefore, such an observation time point is informative to the longitudinal measurement. For survival events, a patient in this study may experience a single event, death/transplant (or censoring), or may experience a death/transplant (or censoring) event and an extra visit to clinic (implying worsening medical condition).

Multiple event models such as multistate models (Andersen and Keiding, 2002; Meira-Machado *et al.*, 2009) are suitable for modelling the extra-visit event and death event. To incorporate the effects of longitudinal measurements, we consider a joint analysis of multiple event models for the survival data and linear mixed effect models for the longitudinal measurements, where the dependency on the informative observation time points is also considered. The sub-models are joint via a common biomarker process. Such joint models for longitudinal data and survival events have been well developed, when the observation times for longitudinal data are non-informative. Henderson *et al.* (2000) demonstrated the advantage of using a joint

JOURNAL OF STATISTICAL THEORY AND PRACTICE
2018, VOL. 12, NO. 1, 12–22
http://dx.doi.org/10.1080/15598608.2017.1298062

A semiparametric mixture regression model for longitudinal data

Tapio Nummi[a], Janne Salonen[b], Lasse Koskinen[c], and Jianxin Pan[d]

[a]Faculty of Natural Sciences, University of Tampere, Tampere, Finland; [b]Research Department, The Finnish Centre for Pensions, Helsinki, Finland; [c]Faculty of Management, University of Tampere, Tampere, Finland; [d]School of Mathematics, The University of Manchester, Manchester, United Kingdom

ABSTRACT

A normal semiparametric mixture regression model is proposed for longitudinal data. The proposed model contains one smooth term and a set of possible linear predictors. Model terms are estimated using the penalized likelihood method with the EM algorithm. A computationally feasible alternative method that provides an approximate solution is also introduced. Simulation experiments and a real data example are used to illustrate the methods.

ARTICLE HISTORY
Received 16 November 2016
Accepted 19 February 2017

KEYWORDS
Curve clustering; EM algorithm; finite mixtures; growth curves

AMS SUBJECT CLASSIFICATION
62G05; 62B99; 62J07

1. Introduction

Modeling of longitudinal data has been of special interest in statistics during recent decades. Depending on the context, several approaches have been used: multivariate analysis, linear and generalized linear mixed and mixture models, structural equation models, Bayesian methods, quantile regression, and so on. For comprehensive summaries of different approaches to longitudinal data analysis we can refer to Fitzmaurice et al. (2011) and Diggle et al. (2013), for example.

In our approach the focus is on the situation, where the studied population is not completely homogeneous over time, but is instead comprised of groups of individuals with the same kind of mean developmental profiles. One approach to understanding such heterogeneity is to apply the theory of finite mixtures (FM). Nagin (1999; 2005) and Jones et al. (2001) apply the generalized linear models theory to FM with the assumption that observations within a given mixture are independent. A further extension is to take some model parameters (e.g., polynomial coefficients) as random variables or (latent factors); see, for example, Muthen and Khoo (1998). These random terms can then be used for modeling the correlation of the observations within a component mixture. The other kind of mixture regression application arises if part of the random model parameters arise from a mixture distribution (see, e.g., Verbeke and Lesaffre 1996).

The focus in the present study is especially on modeling the mean within the mixture using semiparametric regression techniques (Nummi et al. 2011; Nummi et al. 2013). The mean consists of one time-dependent smooth term and a set of linear predictors that may or may not depend on time. Model terms are estimated using the penalized likelihood method with the EM algorithm. This study also introduces a computationally feasible alternative that provides an

CONTACT Tapio Nummi tan@uta.fi Faculty of Natural Sciences, University of Tampere, Tampere, FIN-33014, Finland.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/ujsp.

Physica A 500 (2018) 162–176

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

A deterministic and stochastic model for the system dynamics of tumor–immune responses to chemotherapy

Xiangdong Liu [a], Qingze Li [b], Jianxin Pan [b,*]

[a] Donlinks School of Economics and Management, University of Science and Technology Beijing, Beijing, 100083, China
[b] School of Mathematics, The University of Manchester, Oxford Road, Manchester, M13 9PL, UK

HIGHLIGHTS

- A deterministic and a stochastic tumor–immune model are constructed.
- The basic dynamical properties are investigated in the deterministic model.
- The CTMC model is harnessed to estimate the extinction probability of tumor cells.
- Numerical simulations are performed to confirm the obtained theoretical results.

ARTICLE INFO

Article history:
Received 21 February 2017
Received in revised form 3 January 2018

Keywords:
Chemotherapy
Deterministic model
Stochastic model
Tumor

ABSTRACT

Modern medical studies show that chemotherapy can help most cancer patients, especially for those diagnosed early, to stabilize their disease conditions from months to years, which means the population of tumor cells remained nearly unchanged in quite a long time after fighting against immune system and drugs. In order to better understand the dynamics of tumor–immune responses under chemotherapy, deterministic and stochastic differential equation models are constructed to characterize the dynamical change of tumor cells and immune cells in this paper. The basic dynamical properties, such as boundedness, existence and stability of equilibrium points, are investigated in the deterministic model. Extended stochastic models include stochastic differential equations (SDEs) model and continuous-time Markov chain (CTMC) model, which accounts for the variability in cellular reproduction, growth and death, interspecific competitions, and immune response to chemotherapy. The CTMC model is harnessed to estimate the extinction probability of tumor cells. Numerical simulations are performed, which confirms the obtained theoretical results.

1. Introduction

Cancer treatment is a major public health problem in most parts of the world. It causes the highest mortality rate in economically developed countries and the second highest mortality rate in developing countries [1]. In 2012, there were about 14.1 million new cancer cases and 8.2 million deaths in the world based on the GLOBOCAN estimates. The occurrence of cancer is increasing as a result of population aging and growth, leading to an increasing prevalence of established risk factors such as smoking, overweight, physical inactivity, and changing reproductive patterns associated with urbanization and economic development [2].

* Corresponding author.
E-mail addresses: xdliu@ustb.edu.cn (X. Liu), jianxin.pan@manchester.ac.uk (J. Pan).

https://doi.org/10.1016/j.physa.2018.02.118
0378-4371/

Journal of Alloys and Compounds 737 (2018) 448–455

Contents lists available at ScienceDirect

Journal of Alloys and Compounds

journal homepage: http://www.elsevier.com/locate/jalcom

An original technique for quantifying the flow-field characteristics in an electrodeposition process of Zn-SiO_2 with Fe

Qingtai Xiao [a, b], Jianxin Pan [c], Yunying Fan [d], Jianxin Xu [a, c, e, *], Hua Wang [a, b, **]

[a] State Key Laboratory of Complex Nonferrous Metal Resources Clean Utilization, Kunming University of Science and Technology, Kunming 650093, PR China
[b] Faulty of Metallurgical and Energy Engineering, Kunming University of Science and Technology, Kunming 650093, PR China
[c] School of Mathematics, The University of Manchester, Manchester M13 9PL, UK
[d] School of Material Science and Engineering, Kunming University of Science and Technology, Kunming 650093, PR China
[e] Quality Development Institute, Kunming University of Science and Technology, Kunming 650093, PR China

ARTICLE INFO

Article history:
Received 24 August 2017
Received in revised form
21 November 2017
Accepted 10 December 2017
Available online 12 December 2017

Keywords:
Flow-field characteristics
Composite electro deposition
Uniformity measure
Plating parameters
Zn-Fe-SiO_2

ABSTRACT

The purposed of this article is to introduce a novel approach (uniformity measure, *U*) based on entropy theory for measuring the micron-particle blend homogeneity of aqueous electrolytes, which applies directly to the imaging data of flow field and does not require contacting and disturbing it. Effectiveness of the new method has been illustrated on synthetic imaging data. To verify the feasibility of our method for real experimental data, we analyze the flow-field images from electrodeposited Zn-Fe-SiO_2 composite coatings process. The numerical results showed that the potential of proposed method was demonstrated successfully to quantitatively establishes association between plating parameters and flow field characteristics. The possible recommendations are to monitor the deposition of micro-particles during the composite electrodeposition processes and to apply this technique for studying a variety of multi-phase mixing problems in which assessment of uniformity is required.

1. Introduction

As one of the most significant metallic material surface finishing technologies and metal-based composite material preparation technologies, composite electrodeposition plating technology is widely adopted to prepare the new chemical materials with satisfactory performance [1–4]. In particular, it is one of most commonly practiced industrial techniques for the fabrication of zinc coatings which are widely used for the corrosion protection of ferrous materials, acting both as a physical barrier from the surrounding corrosive environment and as a self-sacrificial anodic protective layer [5–8]. There are two basic types of Zn and Zn alloy plating baths currently available: acid and alkaline type [9–11]. Blend homogeneity is key to composite electrodeposition, and researchers put substantial resources into inspiring flow field to develop mixing uniformity. Zinc is a well-known sacrificial coating material for iron and co-deposition of suitable particles is of interest for further improving its corrosion protection performance [12–14]. Electrodeposition of zinc-iron alloys is of practical importance since they have better corrosion resistance and mechanical properties than pure zinc coating [15,16]. Although some of the open research focused on electroplating technological parameters and electrochemical theory, the number of publications which address the quantification of flow-field characteristics of electrolyte solution is very limited [17].

Studies show that the mixing quality of electrolyte solution and the electrochemical reaction on the surface of are important for appraising performance of composite electrodeposition. However, the reports have not been completely able to provide a quantitative interpretation of mixture. Khan et al. (2011) reported a detailed study of Zn-SiO_2 nanocomposite coatings deposited from a zinc sulfate ($ZnSO_4$) solution at pH = 3 [13]. Shahri et al. (2013) prepared a new nanocomposite coatings by means of the conventional electrodeposition in chloride solution containing different concentrations of hexagonal boron nitride particles [18]. Xia et al. (2013) investigated the microstructure of Ni-AlN composite

* Corresponding author. State Key Laboratory of Complex Nonferrous Metal Resources Clean Utilization, Kunming University of Science and Technology, Kunming 650093, PR China.
** Corresponding author. State Key Laboratory of Complex Nonferrous Metal Resources Clean Utilization, Kunming University of Science and Technology, Kunming 650093, PR China.
E-mail addresses: qingtaixiao2016@kmust.edu.cn (Q. Xiao), jianxin.pan@manchester.ac.uk (J. Pan), yunyingfan7739@sina.com (Y. Fan), jianxin.xu@manchester.ac.uk, xujianxina@163.com (J. Xu), wanghua65@163.com (H. Wang).

https://doi.org/10.1016/j.jallcom.2017.12.098
0925-8388/

Energy-Efficient Priority-Based Scheduling for Wireless Network Slicing

Qing Wang*, Jing Fu†, Jingjin Wu*‡, Bill Moran§, Moshe Zukerman*
*Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong SAR, P. R. China
†School of Mathematics and Statistics, The University of Melbourne, Victoria, Australia
‡Division of Science and Technology, BNU-HKBU United International College, Zhuhai, Guangdong, P. R. China
§Department of Electrical and Electronic Engineering, The University of Melbourne, Victoria, Australia
Email: qwang99-c@my.cityu.edu.hk; jing.fu@unimelb.edu.au; jingjinwu@uic.edu.hk;
wmoran@unimelb.edu.au; m.zu@cityu.edu.hk

***Abstract*—Wireless network slicing is a promising technology for next-generation networks to provide tailored on-demand services to mobile users. We consider a scheduling policy for wireless network slicing with the aim to maximize the energy efficiency of the network defined as the ratio of long-run average throughput of user requests to the long-run average power consumption. This gives rise to a problem of extremely high computational complexity which prevents direct application of conventional optimization techniques. We propose a scalable priority-based policy, referred to as *the Most Energy-Efficient Resource First* (MEERF). MEERF is proved to be asymptotically optimal in the special case appropriate for a local wireless environment with highly dense user population and exponentially distributed service time requirement. The robustness of MEERF to different service time distributions is demonstrated by extensive simulations. We present numerically the effectiveness of MEERF by comparing it with benchmark policies in a more general network with potentially geographically distributed users and infrastructures. The results show that MEERF outperforms the benchmark policies in most of our experiments and achieves up to 52% improvement in terms of energy efficiency.**

***Index Terms*—wireless network slicing, energy-efficient networking, priority-based policies.**

I. Introduction

In recent years, traffic in wireless networks has experienced explosive growth as a result of widespread usage of mobile communication devices and increasing popularity of mobile multimedia applications. Meanwhile, mobile service providers (MSPs) are expected to simultaneously support vertical mobile applications and their diverse requirements, such as ultra-low latency, densely distributed users, high scalability, and high reliability [1]. The fifth generation mobile networks (5G) have been proposed to address these issues and are expected to be put into service around 2020 [2]. Compared to traditional mobile networks (4G or earlier) that were mainly designed for offering a specific type of service such as voice, text or Internet access, 5G has significantly higher capacity (at least 1000 times more than 4G) that enables it to support various emerging applications such as e-health, virtual reality, and automatic monitoring of unmanned devices [3].

Wireless network slicing (WNS) is a technique proposed recently to provide specialized and dedicated virtual networks (*slices*) to users with customized requirements [4]. All slices created are built on a common physical network infrastructure with the help of network function virtualization and software defined networking technologies [5]. The main advantages of WNS are high flexibility and low cost. MSPs can provide their users on-demand tailored services for a series of specific scenarios without investing in additional physical infrastructures.

Because of the multitude and diversity of requests from various applications, an efficient *scheduling policy* is essential to manage network resources and allocate these resources to users effectively. However, it is usually difficult to derive an optimal scheduling policy in practical situations, as conventional dynamic optimization techniques are usually computationally prohibitive for networks of practical size.

Existing scheduling policies for WNS have mainly focused on achieving better Quality of Service (QoS) of networks (e.g., [6]–[8]). Meanwhile, energy efficiency, as another important issue in wireless communication, has been largely ignored in WNS studies. In fact, improving energy efficiency of wireless networks can significantly reduce the operating cost of MSPs and have a positive impact on the environment [9].

Our aim for this paper is to propose a computationally efficient dynamic scheduling policy for WNS, such that users can be provided with appropriate network resources when requesting a slice. In particular, the policy focuses on maximizing energy efficiency of the network, defined as the ratio of long-run average throughput to long-run average power consumption [9].

Our network model incorporates several features in recent development of wireless communications. Specifically, we consider a heterogeneous wireless network consisting of different *communication nodes* (CNs), e.g. cellular base stations or WiFi access points, in a densely populated area, where the fast growing number of mobile users that require a larger number of CNs and higher CN capacities [2]. Each CN has different power consumption profile and maximum capacity [10]. Connections between users and CNs are established based on the Orthogonal Frequency-Division Multiple Access (OFDMA) scheme, in which orthogonal channels are allocated to users in both frequency and time domains [11], [12].

The main contributions of this paper are summarized as follows:

Statistical Papers
https://doi.org/10.1007/s00362-019-01143-6

REGULAR ARTICLE

Sharp lower bounds of various uniformity criteria for constructing uniform designs

A. M. Elsawah[1,2] · Kai-Tai Fang[1,3] · Ping He[1] · Hong Qin[4,5]

Received: 20 January 2019 / Revised: 14 October 2019

Abstract
Several techniques are proposed for designing experiments in scientific and industrial areas in order to gain much effective information using a relatively small number of trials. Uniform design (UD) plays a significant role due to its flexibility, cost-efficiency and robustness when the underlying models are unknown. UD seeks its design points to be uniformly scattered on the experimental domain by minimizing the deviation between the empirical and theoretical uniform distribution, which is an NP hard problem. Several approaches are adopted to reduce the computational complexity of searching for UDs. Finding sharp lower bounds of this deviation (discrepancy) is one of the most powerful and significant approaches. UDs that involve factors with two levels, three levels, four levels or a mixture of these levels are widely used in practice. This paper gives new sharp lower bounds of the most widely used discrepancies, Lee, wrap-around, centered and mixture discrepancies, for these types of designs. Necessary conditions for the existence of the new lower bounds are presented. Many results in recent literature are given as special cases of this study. A critical comparison study between our results and the existing literature is provided. A new effective version of the fast local search heuristic threshold accepting can be implemented using these new lower bounds. Supplementary material for this article is available online.

Keywords Balanced design · Uniform design · Discrepancy · Lower bound

Mathematics Subject Classification 62K05 · 62K15

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00362-019-01143-6) contains supplementary material, which is available to authorized users.

✉ A. M. Elsawah
a_elsawah85@yahoo.com; amelsawah@uic.edu.hk; a.elsawah@zu.edu.eg

Extended author information available on the last page of the article

Published online: 02 November 2019

COMMUNICATIONS IN STATISTICS—SIMULATION AND COMPUTATION®
2019, VOL. 48, NO. 2, 458–471
https://doi.org/10.1080/03610918.2017.1385813

Representative points for location-biased datasets

Zong-Feng Qi[a], Yong-Dao Zhou[b,c], and Kai-Tai Fang[d]

[a]The State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System, Luoyang, China; [b]Institute of Statistics, Nankai University, Tianjin, China; [c]College of Mathematics, Sichuan University, Chengdu, China; [d]Division of Science and Technology, BNU-HKBU United International College, Zhuhai, China

ABSTRACT

Representative points (RPs) are a set of points that optimally represents a distribution in terms of mean square error. When the prior data is location biased, the direct methods such as the *k*-means algorithm may be inefficient to obtain the RPs. In this article, a new indirect algorithm is proposed to search the RPs based on location-biased datasets. Such an algorithm does not constrain the parameter model of the true distribution. The empirical study shows that such algorithm can obtain better RPs than the *k*-means algorithm.

ARTICLE HISTORY
Received 5 February 2017
Accepted 25 September 2017

KEYWORDS
Good lattice point set; Kernel estimator; Randomized likelihood sampling; Representative point

MATHEMATICS SUBJECT CLASSIFICATION
62E17; 62D05

1. Introduction

A complex equipment such as military radar should do type approval test before the large-scale production. The type approval test can verify the performance of the equipment such as the condition in which the equipment works well or loses efficacy, and the consistence with the performance parameters provided by manufacturer. Usually, the department for organizing type approval test is different with the manufacturer, who may provide some dataset of prior experiments of the equipment for the reference of the stage of type approval test.

Let the prior data be $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, $i = 1, \ldots, n$, where $\boldsymbol{x}_i$ is the ith independent variable, $\boldsymbol{y}_i$ is the response, and n is often very large. One may directly use the full prior data to estimate the unknown distribution of the response and the model between $\boldsymbol{x}$ and $\boldsymbol{y}$. However, the prior experiments may not be well designed, for example, most of the data locate at a partial region of the experimental domain. One possible reason is that the manufacturer may only show the dataset with goodness of the equipment and cut down the dataset with some shortcomings. For example, the true distribution is the uniform distribution on $[0, 1]$, while most of the prior data locate at $[0, 0.5]$ or the empirical distribution of prior dataset is much different with the true distribution. Such prior dataset is called as *location-biased data*. Usually, the researchers may have some prior knowledge of the prior dataset, that is, whether it is location biased or not. Then, it is necessary to use some methods to obtain the true information based on the prior data. One useful way is to find some typical cases of the true distribution, that is, representative points (RPs), for reducing the effect of the location bias. In this article, a new method for searching RPs is proposed for the dataset whether it is location biased or not.

CONTACT Yong-Dao Zhou ydzhou@nankai.edu.cn Institute of Statistics, Nankai University, Tianjin 300071, China.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lssp

中国科学 : 数学　2019 年　第 49 卷　第 7 期: 1009 ~ 1020
SCIENTIA SINICA Mathematica
论 文

FM- 代表点

周永道[1*], 方开泰[2,3]

1. 南开大学统计研究院, 天津 300071;
2. 北京师范大学香港 - 浸会大学联合国际学院理工科技学部, 珠海 519085;
3. 中国科学院数学与系统科学研究院应用数学研究所, 北京 100190

E-mail: ydzhou@nankai.edu.cn, ktfang@uic.edu.hk

收稿日期: 2017-08-10; 接受日期: 2018-03-21; 网络出版日期: 2019-04-24; * 通信作者
国家自然科学基金 (批准号: 11471229)、中国科学院数学与系统科学研究院科研基金和珠海市优势学科基金资助项目

摘要 本文提出一种寻找连续性随机变量代表点的新准则: FM- 准则, 该准则在前 $n-1$ 样本矩等于相应的总体矩的约束条件下, 最小化经验分布函数与总体分布之间差异的 L_2- 范数. 本文证明该准则对很多分布都是有意义的. 当约束条件不满足时, 该准则被推广至伪 FM- 准则. 一些例子表明, FM- 代表点比其他类型的代表点更优.

关键词 F- 偏差　均方误差　矩　代表点

MSC (2010) 主题分类 62E10

1 引言

在实际应用中, 经常要求用一个离散分布来近似某个连续概率分布. 例如, Fang 和 He [1] 用代表点方法寻找服装的最优尺寸. Flury [2]、Flury 和 Tarpey [3] 用主成分点决定瑞士军队的防毒面具的最优尺寸和形状, 以及函数型数据的最优子集等.

给定某连续型随机变量 X 的概率密度函数 $f(x)$ 和相应的分布函数 $F(x)$, 我们寻找一个离散型随机变量 ξ, 使其分布函数 $G(x)$ 能尽可能多地保留 $F(x)$ 的性质. 寻找 ξ 的一个基本想法是, 选择一些代表点, 并对每个点赋以一定的权重. 文献中有许多构造方法, 如 Monte Carlo 方法、括号 - 中位数法[4]、推广的 Pearson-Tukey 法 [5]、数论方法 [6] 和均方误差法 [1,2,7]. Monte Carlo 方法考虑一个随机样本 $\{\xi_1,\ldots,\xi_n\}$, 并且每个点有相同的权重; 我们把相应的代表点记为 MC-RP (Monte Carlo-representative point). 括号 - 中位数法是把连续分布函数 $F(x)$ 分成一些相同的区间, 并对每个区间取其括号 - 中位数为代表点. 推广的 Pearson-Tukey 法选择三个点作为代表点, 即中位数、0.05 和 0.95 分位数, 且其权重分别为 0.63、0.185 和 0.185. 对于一维情形, 数论方法选择 $\{\xi_i = F^{-1}((2i-1)/(2n)), i = 1,\ldots,n\}$ 作为 n 个代表点, 其中 $F^{-1}(x)$ 为 $F(x)$ 的逆函数, 记相应的代表点为 NTM-RP (number-theoretic

英文引用格式: Zhou Y D, Fang K T. FM-criterion for representative points (in Chinese). Sci Sin Math, 2019, 49: 1009–1020, doi: 10.1360/SCM-2017-0529

Metrika
https://doi.org/10.1007/s00184-019-00713-w

The main effect confounding pattern for saturated orthogonal designs

Yuxuan Lin[1] · Kai-Tai Fang[1,2]

Received: 6 August 2018

Abstract
In this paper, we propose a criterion "the main effect confounding pattern (MECP)" for comparing projection designs based on saturated symmetric orthogonal designs. Some studies for $L_9(3^4)$, $L_{27}(3^{13})$ and $L_{16}(4^5)$ are given. They show that the new criterion MECP is mostly consistent with the criteria: the generalized word-length pattern and the discrepancies CD and MD. Moreover, the MECP can provide more information about statistical performance in the classification for projection designs than the other criteria. Hence, designs with the best projection MECP may perform better in the view of confounding. The MECP provides a way to find the best main effect arrangement for the experimenter. We also prove that all the geometrically equivalent $L_n(f^s)$ designs have the same WD/CD/MD discrepancy values.

Keywords Main effect confounding pattern · Orthogonal design · Generalized word-length pattern · Centered L_2-discrepancy · Mixture discrepancy · Isomorphism

1 Introduction

Isomorphism of experimental designs has played an important role in the study of factorial designs. Let $\mathcal{L}_n(f^s)$ be the set of the symmetric orthogonal designs $L_n(f^s)$, where n represents the number of runs (rows), s the number of factors (columns), and f the number of levels of each factor. A design is called a U-type design if all levels of each factor appear equally often and denoted as $U(n, f^s)$. The set of all $U(n, f^s)$

✉ Yuxuan Lin
yuxuanlin@uic.edu.hk

Kai-Tai Fang
ktfang@uic.edu.hk

1 Division of Science and Technology, BNU-HKBU United International College, Zhuhai 519085, China

2 The Key Lab of Random Complex Structures and Data Analysis, The Chinese Academy of Sciences, Beijing, China

Published online: 27 March 2019

Statistical Papers
https://doi.org/10.1007/s00362-019-01143-6

REGULAR ARTICLE

Sharp lower bounds of various uniformity criteria for constructing uniform designs

A. M. Elsawah[1,2] · Kai-Tai Fang[1,3] · Ping He[1] · Hong Qin[4,5]

Received: 20 January 2019 / Revised: 14 October 2019

Abstract

Several techniques are proposed for designing experiments in scientific and industrial areas in order to gain much effective information using a relatively small number of trials. Uniform design (UD) plays a significant role due to its flexibility, cost-efficiency and robustness when the underlying models are unknown. UD seeks its design points to be uniformly scattered on the experimental domain by minimizing the deviation between the empirical and theoretical uniform distribution, which is an NP hard problem. Several approaches are adopted to reduce the computational complexity of searching for UDs. Finding sharp lower bounds of this deviation (discrepancy) is one of the most powerful and significant approaches. UDs that involve factors with two levels, three levels, four levels or a mixture of these levels are widely used in practice. This paper gives new sharp lower bounds of the most widely used discrepancies, Lee, wrap-around, centered and mixture discrepancies, for these types of designs. Necessary conditions for the existence of the new lower bounds are presented. Many results in recent literature are given as special cases of this study. A critical comparison study between our results and the existing literature is provided. A new effective version of the fast local search heuristic threshold accepting can be implemented using these new lower bounds. Supplementary material for this article is available online.

Keywords Balanced design · Uniform design · Discrepancy · Lower bound

Mathematics Subject Classification 62K05 · 62K15

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00362-019-01143-6) contains supplementary material, which is available to authorized users.

✉ A. M. Elsawah
a_elsawah85@yahoo.com; amelsawah@uic.edu.hk; a.elsawah@zu.edu.eg

Extended author information available on the last page of the article

Published online: 02 November 2019

STATISTICS
2019, VOL. 53, NO. 6, 1357–1385
https://doi.org/10.1080/02331888.2019.1688816

Constructing optimal projection designs

A. M. Elsawah [a,b], Yu Tang[c] and Kai-Tai Fang[a,d]

[a]Division of Science and Technology, BNU-HKBU United International College, Zhuhai, People's Republic of China; [b]Department of Mathematics, Faculty of Science, Zagazig University, Zagazig, Egypt; [c]School of Mathematical Sciences, Soochow University, Suzhou, People's Republic of China; [d]The Key Lab of Random Complex Structures and Data Analysis, The Chinese Academy of Sciences, Beijing, People's Republic of China

ABSTRACT

The early stages of many real-life experiments involve a large number of factors among which only a few factors are active. Unfortunately, the optimal full-dimensional designs of those early stages may have bad low-dimensional projections and the experimenters do not know which factors turn out to be important before conducting the experiment. Therefore, designs with good projections are desirable for factor screening. In this regard, significant questions are arising such as whether the optimal full-dimensional designs have good projections onto low dimensions? How experimenters can measure the goodness of a full-dimensional design by focusing on all of its projections?, and are there linkages between the optimality of a full-dimensional design and the optimality of its projections? Through theoretical justifications, this paper tries to provide answers to these interesting questions by investigating the construction of optimal (average) projection designs for screening either nominal or quantitative factors. The main results show that: based on the aberration and orthogonality criteria the full-dimensional design is optimal if and only if it is optimal projection design; the full-dimensional design is optimal via the aberration and orthogonality if and only if it is uniform projection design; there is no guarantee that a uniform full-dimensional design is optimal projection design via any criterion; the projection design is optimal via the aberration, orthogonality and uniformity criteria if it is optimal via any criterion of them; and the saturated orthogonal designs have the same average projection performance.

ARTICLE HISTORY
Received 4 January 2019
Accepted 31 October 2019

KEYWORDS
Projection; level permutations; optimal projection designs; Hamming distance; orthogonality; uniformity; aberration; moment aberration

2010 MATHEMATICS SUBJECT CLASSIFICATIONS
62K05; 62K15

1. Introduction

Design of experiments is becoming ubiquitous in engineering, science, industry and many real-world problems for studying complex phenomena and investigating the relationship between inputs affecting an experiment and its outputs. The significant problem experimenters may face is the selection of efficient designs for their experiments, which reduce the experimental cost and provide more efficient information about the behaviour of the

CONTACT A. M. Elsawah a_elsawah85@yahoo.com, amelsawah@uic.edu.hk Division of Science and Technology, BNU-HKBU United International College, Zhuhai 519085, People's Republic of China; Department of Mathematics, Faculty of Science, Zagazig University, Zagazig 44519, Egypt

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
https://doi.org/10.1080/03610918.2019.1628988

Some interesting behaviors of good lattice point sets

A. M. Elsawah[a,b], Kai-Tai Fang[b,c], and Yu Hui Deng[b]

[a]Department of Mathematics, Faculty of Science, Zagazig University, Zagazig, Egypt; [b]Division of Science and Technology, BNU-HKBU United International College, Zhuhai, China; [c]The Key Lab of Random Complex Structures and Data Analysis, The Chinese Academy of Sciences, Beijing, China

ABSTRACT

Good lattice point (GLP) sets are sets of points that are uniformly distributed over the domain of interest and thus have good space-filling property. GLP sets are an important kind of points for many applications, such as multidimensional quadrature, simulation, computer experiments, quasi-Monte Carlo techniques and design of experiments. It is a significant issue to study the behaviors of GLP sets. For instance, most real-life applications require that the columns of a GLP set are independent, i.e., the GLP set has full rank. If one or more columns are linearly dependent, there will be more confounding among the main effects and interactions in statistical models and also we can not find the least squares estimation of regression coefficients in the linear regression model. The problem of finding the rank of the GLP set remained unsolved since its inception in 1959. It is desirable to put the first stone for solving this significant problem. Through theoretical justifications, this paper gives some interesting behaviors of the generating vector of any GLP set and shows the independence among its rows and columns by presenting some analytic linkages among these rows and columns. The results of this paper are used not only as a benchmark for constructing full rank GLP sets, but also in various applications of GLP sets.

ARTICLE HISTORY
Received 4 November 2018
Accepted 3 June 2019

KEYWORDS
Good lattice points; Generating vector; Confounding; Rank; Euler function; Number theory

MATHEMATICS SUBJECT CLASSIFICATION
62K05; 62K15; 11Z99

1. Introduction

Good lattice point (GLP) method (cf. Korobov 1959) is a widely used important method for producing sets of points that are evenly distributed over the domain of interest. GLP sets are a significant kind of points for quasi-Monte Carlo, multidimensional quadrature, computer experiments, simulation and design of experiments (cf. Wang and Hickernell 2002; Sloan and Joe 1994; Niederreiter 1992; and Zaremba 1966). GLP sets have good space-filling property (cf. Zhou and Xu 2015; Fang and Wang 1994; and Hua and Wang 1981) and thus can be used to construct space-filling designs for many real-life experiments. Uniform designs (cf. Fang 1980) are a widely used class of space-filling designs. Constructing a uniform design is an NP hard problem, specially for designs with large sizes. Many authors proposed several efficient methods and algorithms for constructing uniform designs (cf. Elsawah et al. 2019 and Fang et al. 2017) among which the GLP method has played an important role (cf. Fang et al. 2018). Most

CONTACT A. M. Elsawah a_elsawah85@yahoo.com; amelsawah@uic.edu.hk; a.elsawah@zu.edu.eg Department of Mathematics, Faculty of Science, Zagazig University, Zagazig 44519, Egypt.

Building some bridges among various experimental designs

Article *in* Journal of the Korean Statistical Society · September 2019 *with* 53 Reads
DOI: 10.1007/s42952-019-00004-0
Cite this publication

A. M. Elsawah
18.7 · United International College

Abstract

Designing their experiments is the significant problem that experimenters face. Maximin distance designs, supersaturated designs, minimum aberration designs, uniform designs, minimum moment designs and orthogonal arrays are arguably the most exceedingly used designs for many real-life experiments. From different perspectives, several criteria have been proposed for constructing these designs for investigating quantitative or qualitative factors. Each of those criteria has its pros and cons and thus an optimal criterion does not exist, which may confuse investigators searching for a suitable criterion for their experiment. Some logical questions are now arising, such as are these designs consistent, can an optimal design via a specific criterion perform well based on another criterion and can an optimal design for screening quantitative factors be optimal for qualitative factors? Through theoretical justifications, this paper tries to answer these interesting questions by building some bridges among these designs. Some conditions under which these designs agree with each other are discussed. These bridges can be used to select a suitable criterion for studying some hard problems effectively, such as detection of (combinatorial/geometrical) non-isomorphism among designs and construction of optimal designs. Benchmarks for reducing the computational complexity are given.

Des. Codes Cryptogr.
https://doi.org/10.1007/s10623-018-0498-x

The vertex-isoperimetric number of the incidence and non-incidence graphs of unitals

Alice M. W. Hui[1] · **Muhammad Adib Surani**[2] · **Sanming Zhou**[2]

Received: 1 December 2017 / Revised: 13 May 2018 / Accepted: 23 May 2018

Abstract We derive upper and lower bounds for the vertex-isoperimetric number of the incidence graphs of unitals and determine its order of magnitude. In the case when a unital contains sufficiently large arcs, these bounds agree and give rise to the precise value of this parameter. In particular, we obtain the exact value of the vertex-isoperimetric number of the incidence graphs of classical unitals and a certain subfamily of BM-unitals. In the case when the maximum size of arcs in the unital is relatively small, we obtain an upper bound for this parameter in terms of the vertex-isoperimetric number of the incidence graph. We also determine the exact value of the vertex-isoperimetric number of the non-incidence graph of any unital.

Keywords Vertex-isoperimetric number · Unital · Incidence graph

Mathematics Subject Classification 05C40 · 05B25

1 Introduction

A *unital* is a 2-$(n^3+1, n+1, 1)$ design for some integer $n \geq 2$. In this paper we derive upper and lower bounds for the vertex-isoperimetric number of the incidence graphs of unitals and

Communicated by D. Ghinelli.

✉ Sanming Zhou
sanming@unimelb.edu.au

Alice M. W. Hui
alicemwhui@uic.edu.hk; huimanwa@hotmail.com

Muhammad Adib Surani
m.surani@student.unimelb.edu.au

[1] Statistics Program, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai, China

[2] School of Mathematics and Statistics, The University of Melbourne, Parkville, VIC 3010, Australia

Published online: 31 May 2018

IEEE Access

Received September 18, 2019, accepted October 8, 2019, date of publication October 15, 2019, date of current version October 28, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2947637

Cost-Efficient Millimeter Wave Base Station Deployment in Manhattan-Type Geometry

MIAOMIAO DONG[1], (Student Member, IEEE), TAEJOON KIM[2], (Member, IEEE), JINGJIN WU[3], (Member, IEEE), AND ERIC W. M. WONG[1], (Member, IEEE)

[1]Department of Electrical Engineering, City University of Hong Kong, Hong Kong
[2]Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS 66045, USA
[3]Department of Statistics, BNU–HKBU United International College, Zhuhai 519087, China

Corresponding author: Miaomiao Dong (mmdong2-c@my.cityu.edu.hk)

This work was supported in part by the Guangdong University Innovation and Enhancement Programme Funds under Grant 2017KQNCX236.

ABSTRACT Urban millimeter wave (mmWave) communications are limited by link outage due to frequent blockages by obstacles. One approach to this problem is to increase the density of base stations (BSs) to achieve macro diversity gains. Dense BS deployment, however, incurs the increased BS installation cost as well as power consumption. In this work, we propose a framework for connectivity-constrained minimum cost mmWave BS deployment in Manhattan-type geometry (MTG). A closed-form expression of network connectivity is characterized as a function of various factors such as obstacle sizes, BS transmit power, and the densities of obstacles and BSs. Optimization that attains the minimum cost is made possible by incorporating a tight lower bound of the analyzed connectivity expression. A low-complexity algorithm is devised to effectively find an optimal tradeoff between the BS density and transmit power that results in the minimum BS deployment cost while guaranteeing network connectivity. Numerical simulations corroborate our analysis and quantify the best tradeoff of the BS density and transmit power. The proposed BS deployment strategies are evaluated in different network cost configurations, providing useful insights in mmWave network planning and dimensioning.

INDEX TERMS Millimeter wave network, connectivity, base station deployment cost, Manhattan-type geometry, lattice process.

I. INTRODUCTION

Next generation cellular networks will be deployed in millimeter wave (mmWave) bands to support the data-intensive fifth generation (5G) broadband use cases in urban areas [1]–[3]. In mmWave bands, large-scale antenna arrays are used at both base stations (BSs) and user equipments (UEs) to generate directional narrow beams in order to overcome the severe pathloss [4]–[6]. Directional transmission enables almost interference-free communications [7], but it also imposes new challenges, as the weak penetration and diffraction of mmWave propagation make the link susceptible to physical blockages. The blockage incurs frequent link outage in highly-obstructed urban areas. This is in contrast with the conventional sub-6GHz systems, where the outage largely results from co-channel interference rather than a physical blockage.

One approach alleviating mmWave link outage is to cover each UE by multiple BSs, i.e., macro diversity techniques [8], [9]. When a link from a BS is blocked, the UE can switch to another unblocked BS to restore its link. An important practical implication of imposing the macro diversity is an increased number of BSs (i.e., dense BS deployment), which will then increase the expenditures on BS installment. Another alternative is to extend the cell coverage by increasing the transmit power of each BS in order to provide sufficient cell overlap. However, this approach incurs large power consumption. Under a certain connectivity requirement, how to resolve the best tradeoff between the BS density and transmit power that minimizes the BS deployment cost is of great interest for 5G network operators.

Link connectivity is hinged upon network geometry. A Poisson point process (PPP) has been verified to be an

The associate editor coordinating the review of this manuscript and approving it for publication was Zhenyu Xiao.

Covariance Matrix Regularization for Banded Toeplitz Structure via Frobenius-Norm Discrepancy

Xiangzhao Cui, Zhenyang Li, Jine Zhao, Defei Zhang, and Jianxin Pan

Abstract In many practical applications, the structure of covariance matrix is often blurred due to random errors, making the estimation of covariance matrix very difficult particularly for high-dimensional data. In this article, we propose a regularization method for finding a possible banded Toeplitz structure for a given covariance matrix A (e.g., sample covariance matrix), which is usually an estimator of the unknown population covariance matrix Σ. We aim to find a matrix, say B, which is of banded Toeplitz structure, such that the Frobenius-norm discrepancy between B and A achieves the smallest in the whole class of banded Toeplitz structure matrices. As a result, the obtained Toeplitz structured matrix B recoveries the underlying structure behind Σ. Our simulation studies show that B is also more accurate than the sample covariance matrix A when estimating the covariance matrix Σ that has a banded Toeplitz structure. The studies also show that the proposed method works very well in regularization of covariance structure.

Keywords Covariance matrix structure · Frobenius norm · Regularization · Toeplitz structure

1 Introduction

Estimation of covariance matrices is important in many application fields including spectroscopy, functional magnetic resonance imaging, text retrieval, gene array, climate study and imaging analysis. This problem has been widely researched in statistics. The traditional "Burg technique", which is to find the maximum likelihood

X. Cui · Z. Li · J. Zhao · D. Zhang
School of Mathematics, University of Honghe, Yunnan, China
e-mail: cxzh1972@126.com

J. Pan (✉)
School of Mathematics, University of Manchester, Manchester, UK
e-mail: Jianxin.Pan@manchester.ac.uk

S. E. Ahmed et al. (eds.), *Matrices, Statistics and Big Data*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-17519-1_9

Received: 31 May 2019 | Revised: 8 July 2019 | Accepted: 18 July 2019

DOI: 10.1002/cem.3177

RESEARCH ARTICLE

WILEY CHEMOMETRICS

LASSO-based false-positive selection for class-imbalanced data in metabolomics

Guang-Hui Fu[1] | **Lun-Zhao Yi**[2] | **Jianxin Pan**[3]

[1]School of Science, Kunming University of Science and Technology, Kunming, China

[2]Faculty of Agriculture and Food, Kunming University of Science and Technology, Kunming, China

[3]School of Mathematics, The University of Manchester, Manchester, UK

Correspondence
Jianxin Pan, School of Mathematics, The University of Manchester, Manchester M13 9PL, UK.
Email: jianxin.pan@manchester.ac.uk

Funding information
the National Natural Science Foundation of China, Grant/Award Number: 11761041, 21465016 and 21775058

Abstract

Feature selection and rebalancing can be seen as two preprocessing ways in class-imbalanced learning. Recently, there have been many research achievements and applications on LASSO-type feature selection, whereas most of them are not directly designed for addressing class-imbalanced data. In this study, we proposed a LASSO-based stable feature selection algorithm for class-imbalanced data analysis, and false-positive selection (FPS) under balanced and imbalanced situations was calculated via selection frequency of each predictor in doing stable selection. The results on simulation studies and real data examples show that class imbalance contributes to avoid overselection caused by LASSO when the data are highly correlated and a lower FPS can be obtained with class-imbalanced data than balanced one in most of cases in the same settings. A statistical explanation was given for this phenomenon. In addition, it does not need to rebalance the class-imbalanced data for performing such LASSO-based feature selection with a stable strategy, and to some degree, intentionally disequilibrating the balanced data could be an alternative strategy to weaken overselection and to perform biomarker identification for finding a few of most important biomarkers.

KEYWORDS

class imbalance, false-positive selection, LASSO-based feature selection, rebalance

1 | INTRODUCTION

Feature selection can be utilized as the preprocessing of selecting a subset of candidate predictors and is gaining popularity in class-imbalanced learning in this "big data" era.[1-4] Because of the inherent complex characteristics of class-imbalanced data, performing feature selection from such data requires new understandings and principles to transform vast amounts of raw data efficiently into information and knowledge representation.[5] Feature selection is a critical step because of the high dimensionality of data across many scientific discoveries, such as quantitative structure activity relationships (QSAR) study[6-8] and spectroscopic analysis.[9-13] There are three categories of feature selection in the context of classification, depending on how these feature selection searches combine with the construction of the classification model: filtering,[14,15] wrapping,[16,17] and embedding.[1,2,4] The filtering method selects high-ranking features based on statistical or information measures, which is independent of the classifier, whereas it ignores the dependencies among features. The wrapping method wraps a search algorithm around the classification model to search the space of all feature subsets. However, wrapping methods are generally computationally intensive as the number of subsets from the feature space grows exponentially as the number of features increases. The embedding method screens out key features while considering the construction of a classifier. Namely, it is integrated in the modeling process and is classifier dependent.[18] However,

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
https://doi.org/10.1080/03610918.2019.1571607

Correlation structure regularization via entropy loss function for high-dimension and low-sample-size data

Chen Chen[a], Jie Zhou[a], and Jianxin Pan[b]

[a]College of Mathematics, Sichuan University, Chengdu, Sichuan, China; [b]School of Mathematics, University of Manchester, Manchester, UK

ABSTRACT

Estimating structured covariance or correlation matrix has been paid more and more attentions in recent years. A recent method based on the entropy loss function was proposed to regularize the covariance structure for a given covariance matrix whose underlying structure may be blurred due to random noises from different sources. However, the entropy loss function considered is very likely to be unavailable in covariance regularization for high-dimension and low-sample-size (HDLSS) data. In this paper, a new discrepancy is proposed for regularizing correlation structure, in which the given correlation matrix (e.g., sample correlation matrix) and the candidate structure in the entropy loss function are both added by the identity matrix multiplied by a constant, so that the problem owing to likely singularity of sample correlation matrix for HDLSS data can be overcome. The candidate correlation structures considered in this paper include tri-diagonal Toeplitz, compound symmetry, AR(1) and banded Toeplitz. The regularized correlation estimates for the first three structures can be obtained by solving one-dimensional optimization problems, while the regularized one for the fourth structure can be computed efficiently using Newton's iteration method. Simulation studies show that the proposed new approach works well, providing a reliable method to regularize the correlation structure for HDLSS data.

ARTICLE HISTORY
Received 30 July 2018
Accepted 14 January 2019

KEYWORDS
Bregman divergence; Correlation matrix estimation; Entropy loss function; High-dimensional correlation matrix; Regularization

1. Introduction

The covariance matrices with certain special structures have emerged widely in various application fields such as signal processing (Pascal et al. 2008; Soloveychik and Wiesel 2014), ultrasound imaging (Asl and Mahloojifar 2012), neuroimaging (Zhou et al. 2016), genetics selection (Meyer 2009), and social science (van der Leeden et al. 1996). It is usually necessary to estimate the structured covariance or correlation matrix whose underlying structure may be blurred by random noise. This problem was studied by many authors in the literature (see, e.g., Kang et al. (2015); Soloveychik et al. (2016); Sun et al. (2015) and references therein). Estimating structured covariance matrix with high-dimension is fundamental in statistics. Lots of existing works in statistics provided various estimators for structured covariance matrices, including the maximum likelihood-based estimators (Wang and Carey 2003; Jennrich and Schluchter 1986; Daniels

CONTACT Jie Zhou jzhou@scu.edu.cn College of Mathematics, Sichuan University, Chengdu, Sichuan, 610064, China

中国科学: 数学 2020年 第50卷 第9期: 1149～1168
SCIENTIA SINICA Mathematica
论 文

统计分布的代表点集及其应用

献给越民义教授 100 华诞

方开泰[1,2*], 贺平[1], 杨骏[1]

1. 北京师范大学 - 香港浸会大学联合国际学院理工科技学部, 珠海 519087;
2. 中国科学院数学与系统科学研究院应用数学研究所, 北京 100190
E-mail: ktfang@uic.edu.hk, heping@uic.edu.hk, jyang5037@gmail.com

收稿日期: 2019-10-15; 接受日期: 2020-03-16; 网络出版日期: 2020-09-14; * 通信作者
珠海市优势学科基金 (批准号: R1050) 和北京师范大学 - 香港浸会大学联合国际学院校内科研基金 (批准号: R201712, R201810, R201912 和 R202010) 资助项目

摘要 用一个离散统计分布来近似一个连续的统计分布 (一维或多维) 一直是统计学研究的核心内容. 显然这个离散统计分布的支撑点集必须有代表性, 故称它们为代表点集, 或简称代表点. 选择代表点可以有不同的考虑, 本文回顾并比较 4 类近似离散统计分布: 随机样本 (独立同分布)、修改的 Monte Carlo 方法、数论方法的样本 (伪 Monte Carlo 方法) 及在最小平方误准则下的代表点集和相应的统计分布. 其中修改的 Monte Carlo 方法是本文新提出的. 本文比较 4 类方法在密度估计和重采样的统计推断中的表现, 其中有一类是改进的自助法. 本文对最小平方误准则下的代表点的性质和数值算法进行了详细回顾, 并且得到一些新结果, 例如, 随机样本的最小平方误准则的统计分布、椭球等高分布代表点的几何结构以及椭球等高分布代表点和主成分的关系.

关键词 统计分布代表点 伪 Monte Carlo 方法 统计推断 正态分布 椭球等高分布 主成分和主成分点

MSC (2010) 主题分类 65C05, 65C50

1 统计分布的代表点

统计分布是统计学用来建模的重要工具, 一个随机变量 X 的分布函数定义为 $F(x)=\mathrm{P}(X\leqslant x)$. 离散的随机变量 X 的分布, 常常表示为

$$\begin{array}{c|c} X & x_1\ x_2\ \cdots\ x_n \\ \hline p & p_1\ p_2\ \cdots\ p_n \end{array}. \tag{1.1}$$

这里 $x_1,\ldots,x_n$ 称为 X 的支撑点并且 $\mathrm{P}(X=x_i)=p_i>0$, $i=1,\ldots,n$. 连续的统计分布有分布密度函数 $p(x)$, 它是一条曲线. 统计学是通过样本来研究总体的科学和艺术. 如果已知总体的分布类型, 需

英文引用格式: Fang K T, He P, Yang J. Sets of representative points of statistical distributions and their applications (in Chinese). Sci Sin Math, 2020, 50: 1149–1168, doi: 10.1360/SSM-2019-0251

Stat Papers (2020) 61:823–849
https://doi.org/10.1007/s00362-017-0963-z

REGULAR ARTICLE

New foundations for designing U-optimal follow-up experiments with flexible levels

A. M. Elsawah[1,2] · **Kai-Tai Fang**[2,3]

Received: 18 January 2017 / Revised: 23 October 2017 / Published online: 14 November 2017

Abstract Follow-up experiments are used extensively to provide precious information about the relationships between inputs and outputs to gain a better understanding of a process or system under study. This article gives a new look at designing optimal follow-up experiments that involve any number of factors with any number of different levels in light of the uniformity behaviour of the corresponding two stage sequential experiments, which are composed of initial experiments and follow-up experiments. Novel analytical expressions and lower bounds of the wrap-around L_2-discrepancy, as a uniformity measure, for sequential experimental designs are proposed for evaluating the optimality of the follow-up experimental designs. Finding equivalent follow-up experimental designs is investigated, which can be used to reduce the computational complexity. Our results show that two stage sequential experimental designs give greater precision than single stage experimental designs with the same size.

Keywords Indicator function · Follow-up experiment · Follow-up map · Sequential experiment · Equivalent design · Optimal sequential experiment

Mathematics Subject Classification 62K05 · 62K15

✉ A. M. Elsawah
a_elsawah85@yahoo.com ; amelsawah@uic.edu.hk ; a.elsawah@zu.edu.eg

[1] Department of Mathematics, Faculty of Science, Zagazig University, Zagazig 44519, Egypt

[2] Division of Science and Technology, BNU-HKBU United International College, Zhuhai 519085, China

[3] The Key Lab of Random Complex Structures and Data Analysis, The Chinese Academy of Sciences, Beijing, China

ARTICLE IN PRESS

Journal of King Saud University – Science xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of King Saud University – Science

journal homepage: www.sciencedirect.com

Original article

An algorithm for outlier detection in a time series model using backpropagation neural network

Gajendra K. Vishwakarma [a,*], Chinmoy Paul [a,b], A.M. Elsawah [c,d]

[a] Department of Mathematics & Computing, Indian Institute of Technology Dhanbad, Dhanbad 826004, India
[b] Department of Statistics, Pandit Deendayal Upadhyaya Adarsha Mahavidyalaya, Eraligol, Karimganj 788723, India
[c] Division of Science and Technology, Beijing Normal University–Hong Kong Baptist University United International College, Zhuhai 519085, China
[d] Department of Mathematics, Faculty of Science, Zagazig University, Zagazig 44519, Egypt

ARTICLE INFO

Article history:
Received 13 March 2020
Revised 18 August 2020
Accepted 11 September 2020
Available online xxxx

Mathematics Subject Classification:
68Wxx

Keywords:
Multivariate outliers
Detection
Neural network
Robust estimate
Time series
Backpropagation algorithm

ABSTRACT

Outliers are commonplace in many real-life experiments. The presence of even a few anomalous data can lead to model misspecification, biased parameter estimation, and poor forecasts. Outliers in a time series are usually generated by dynamic intervention models at unknown points of time. Therefore, detecting outliers is the cornerstone before implementing any statistical analysis. In this paper, a multivariate outlier detection algorithm is given to detect outliers in time series models. A univariate time series is transformed to bivariate data based on the estimate of robust lag. The proposed algorithm is designed by using robust measures of location and dispersion matrix. Feed forward neural network is used for designing time series models. Number of hidden units in the network is determined based on the standard error of the forecasting error. A comparison study between the proposed algorithm and the widely used algorithms is given based on three real-data sets. The results demonstrated that the proposed algorithm outperformed the existing algorithms due to its non-requirement of a priori knowledge of the time series and its control of both masking and swamping effects. We also discussed an efficient method to deal with unexpected jumps or drops on share prices due to stock split and commodity prices near contract expiry dates.

1. Introduction

The detection of outliers or unusual data structures is one of the important tasks in the statistical analysis of time series data as outliers may have a substantial influence on the outcome of an analysis. Appropriate definition of an outlier usually depends on the assumptions about the structure of data and the applied detection method. Hawkins (1980) defined the outlier as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. Barnett and Lewis (1994) indicated that an outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs. Similarly, Johnson (1992) viewed that, an outlier is an observation in a data set which appears to be inconsistent with the remainder of that set of data. There are many definitions of outlier proposed in the literature of time series. Outlier observations in some situations are also referred as anomalies, discordant observations, or contaminants Carreno et al. (2019).

The presence of outliers in a time series has a significant effect on the results of standard procedures of analysis. The consequences may lead to improper model specification, faulty parameter estimation and substandard forecasting. A crucial point here is that any outlier detection technique can at most detect a set of data points having different behavior than the rest of the data and hence, it can be termed as a probable set of outliers. However, it is up an analyst to take various itineraries to come up with a final decision to justify these detected points as outliers. It is probable that a point detected as an outlier has some real facts behind it, e.g., the price of a stock just after the date of stock split with split ratio of 2-for-1 or 3-for-1, which means a stockholder gets two or three shares, respectively, for every share held. In a reverse stock

* Corresponding author.
E-mail addresses: vishwagk@rediffmail.com (G.K. Vishwakarma), chinmoy.gcc@gmail.com (C. Paul), amelsawah@uic.edu.cn (A.M. Elsawah).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksus.2020.09.018
1018-3647/

Please cite this article as: G. K. Vishwakarma, C. Paul and A. M. Elsawah, An algorithm for outlier detection in a time series model using backpropagation neural network, Journal of King Saud University – Science, https://doi.org/10.1016/j.jksus.2020.09.018

frontiers in Psychiatry

ORIGINAL RESEARCH
published: 26 November 2020
doi: 10.3389/fpsyt.2020.585915

The Medium-Term Impact of COVID-19 Lockdown on Referrals to Secondary Care Mental Health Services: A Controlled Interrupted Time Series Study

*Shanquan Chen[1], Rui She[2], Pei Qin[3], Anne Kershenbaum[1,4], Emilio Fernandez-Egea[1,4], Jenny R. Nelder[1], Chuoxin Ma[5], Jonathan Lewis[4], Chaoqun Wang[6] and Rudolf N. Cardinal[1,4]**

[1] Department of Psychiatry, University of Cambridge, Cambridge, United Kingdom, [2] The Jockey Club School of Public Health and Primary Care, The Chinese University of Hong Kong, Hong Kong, China, [3] Department of Biostatistics and Epidemiology, Shenzhan University Health Science Center, Shenzhen, China, [4] Cambridgeshire and Peterborough NHS Foundation Trust, Cambridge, United Kingdom, [5] Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom, [6] College of Public Administration, Central China Normal University, Wuhan, China

OPEN ACCESS

Edited by:
Wulf Rössler,
Charité – Universitätsmedizin Berlin, Germany

Reviewed by:
Elizabeta Blagoja Mukaetova-Ladinska,
University of Leicester, United Kingdom
Michaela Pascoe,
Victoria University, Australia, Australia

***Correspondence:**
Rudolf N. Cardinal
rnc1001@cam.ac.uk

Specialty section:
This article was submitted to Public Mental Health, a section of the journal Frontiers in Psychiatry

Received: 21 July 2020
Accepted: 28 October 2020
Published: 26 November 2020

Citation:
Chen S, She R, Qin P, Kershenbaum A, Fernandez-Egea E, Nelder JR, Ma C, Lewis J, Wang C and Cardinal RN (2020) The Medium-Term Impact of COVID-19 Lockdown on Referrals to Secondary Care Mental Health Services: A Controlled Interrupted Time Series Study. Front. Psychiatry 11:585915. doi: 10.3389/fpsyt.2020.585915

To date, there is a paucity of information regarding the effect of COVID-19 or lockdown on mental disorders. We aimed to quantify the medium-term impact of lockdown on referrals to secondary care mental health clinical services. We conducted a controlled interrupted time series study using data from Cambridgeshire and Peterborough NHS Foundation Trust (CPFT), UK (catchment population ~0.86 million). The UK lockdown resulted in an instantaneous drop in mental health referrals but then a longer-term acceleration in the referral rate (by 1.21 referrals per day per day, 95% confidence interval [CI] 0.41–2.02). This acceleration was primarily for urgent or emergency referrals (acceleration 0.96, CI 0.39–1.54), including referrals to liaison psychiatry (0.68, CI 0.35–1.02) and mental health crisis teams (0.61, CI 0.20–1.02). The acceleration was significant for females (0.56, CI 0.04–1.08), males (0.64, CI 0.05–1.22), working-age adults (0.93, CI 0.42–1.43), people of White ethnicity (0.98, CI 0.32–1.65), those living alone (1.26, CI 0.52–2.00), and those who had pre-existing depression (0.78, CI 0.19–1.38), severe mental illness (0.67, CI 0.19–1.15), hypertension/cardiovascular/cerebrovascular disease (0.56, CI 0.24–0.89), personality disorders (0.32, CI 0.12–0.51), asthma/chronic obstructive pulmonary disease (0.28, CI 0.08–0.49), dyslipidemia (0.26, CI 0.04–0.47), anxiety (0.21, CI 0.08–0.34), substance misuse (0.21, CI 0.08–0.34), or reactions to severe stress (0.17, CI 0.01–0.32). No significant post-lockdown acceleration was observed for children/adolescents, older adults, people of ethnic minorities, married/cohabiting people, and those who had previous/pre-existing dementia, diabetes, cancer, eating disorder, a history of self-harm, or intellectual disability. This evidence may help service planning and policy-making, including preparation for any future lockdown in response to outbreaks.

Keywords: COVID-19/SARS-CoV-2 coronavirus pandemic, lockdown, secondary care mental health services, controlled interrupted time series analysis, comorbidity

Journal of Computational and Applied Mathematics 384 (2021) 113164

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

Journal homepage: www.elsevier.com/locate/cam

An appealing technique for designing optimal large experiments with three-level factors

A.M. Elsawah *

Division of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai 519085, China
Department of Mathematics, Faculty of Science, Zagazig University, Zagazig 44519, Egypt

ARTICLE INFO

Article history:
Received 1 June 2020
Received in revised form 11 August 2020

MSC:
62K05
62K15

Keywords:
Multiple tripling
Hamming distance
Aberration
Power moments
Orthogonality
Uniformity

ABSTRACT

Experimental design is arguably the most commonly used and effective methodology in scientific investigations and industrial applications. Real-world experiments may have hundreds or even thousands of input variables (factors) and thus a large number of observations (experimental runs) is needed to gain a better understanding of the phenomena under the investigation and estimate the most important parameters without bias and with minimum variance. Constructing optimal designs for these large experiments is a significant NP-hard problem investigators may face. This paper gives a new simple efficient technique, called multiple tripling technique, for constructing optimal (in view of distance, aberration, power moments, orthogonality, uniformity) designs for large experiments with three-level factors by multiple tripling of small and simple three-level initial designs. Some logical questions are now arising, such as how to effectively select initial designs to get optimal resulting multiple triple designs, how to measure the optimality of a resulting multiple triple design relative to all the possible designs with the same size, and what is the efficiency of the multiple tripling technique relative to the existing widely used techniques for constructing large three-level designs? Through theoretical and computational justifications, this paper tries to answer these significant questions. Without computational time (no computer search), the multiple tripling technique is used to construct new recommended optimal designs which are better than the existing recommended designs or cannot be constructed by the existing techniques due to their large sizes.

1. Introduction

Design of experiments is arguably the most commonly used and effective tool for understanding the behavior of complex phenomena in industrial and scientific applications by investigating the effect of input variables (factors) on the response variables (outputs). The most significant hard problem experimenters may face is the optimality selection of the experimental runs (experimental designs) which provide useful information about the behavior of the phenomena under the experimentation. An experiment with an optimal design allows more parameters to be estimated with minimum variance and without bias, while an experiment with a non-optimal design needs a greater number of experimental runs to estimate the parameters with the same accuracy as an optimal design. From a practical point of view, constructing

* Correspondence to: Division of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai 519085, China.
E-mail addresses: amelsawah@uic.edu.cn, a_elsawah85@yahoo.com, a.elsawah@zu.edu.eg.

https://doi.org/10.1016/j.cam.2020.113164
0377-0427/

Journal of the Korean Statistical Society (2020) 49:55–81
https://doi.org/10.1007/s42952-019-00004-0

Online ISSN 2005-2863
Print ISSN 1226-3192

RESEARCH ARTICLE

Building some bridges among various experimental designs

A. M. Elsawah[1,2]

Received: 25 September 2018 / Accepted: 11 June 2019 / Published online: 1 January 2020

Abstract

Designing their experiments is the significant problem that experimenters face. Maximin distance designs, supersaturated designs, minimum aberration designs, uniform designs, minimum moment designs and orthogonal arrays are arguably the most exceedingly used designs for many real-life experiments. From different perspectives, several criteria have been proposed for constructing these designs for investigating quantitative or qualitative factors. Each of those criteria has its pros and cons and thus an optimal criterion does not exist, which may confuse investigators searching for a suitable criterion for their experiment. Some logical questions are now arising, such as are these designs consistent, can an optimal design via a specific criterion perform well based on another criterion and can an optimal design for screening quantitative factors be optimal for qualitative factors? Through theoretical justifications, this paper tries to answer these interesting questions by building some bridges among these designs. Some conditions under which these designs agree with each other are discussed. These bridges can be used to select a suitable criterion for studying some hard problems effectively, such as detection of (combinatorial/geometrical) non-isomorphism among designs and construction of optimal designs. Benchmarks for reducing the computational complexity are given.

Keywords Orthogonality · Uniformity · Discrepancy · Aberration · Moment aberration · Hamming distance · Combinatorial isomorphism · Geometrical isomorphism

Mathematics Subject Classification 62K05 · 62K15

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s42952-019-00004-0) contains supplementary material, which is available to authorized users.

✉ A. M. Elsawah
a_elsawah85@yahoo.com; amelsawah@uic.edu.hk; a.elsawah@zu.edu.eg

[1] Division of Science and Technology, BNU-HKBU United International College, Zhuhai 519085, China

[2] Department of Mathematics, Faculty of Science, Zagazig University, Zagazig 44519, Egypt

J Syst Sci Complex (2020) 33: 1–38

Cross-Entropy Loss for Recommending Efficient Fold-Over Technique*

WENG Lin-Chen · ELSAWAH A M · FANG Kai-Tai

DOI: 10.1007/s11424-020-9267-9
Received: 23 September 2019 / Revised: 3 June 2020

Abstract Due to the limited resources and budgets in many real-life projects, it is unaffordable to use full factorial experimental designs and thus fractional factorial (FF) designs are used instead. The aliasing of factorial effects is the price we pay for using FF designs and thus some significant effects cannot be estimated. Therefore, some additional observations (runs) are needed to break the linages among the factorial effects. Folding over the initial FF designs is one of the significant approaches for selecting the additional runs. This paper gives an in-depth look at fold-over techniques via the following four significant contributions. The first contribution is on discussing the adjusted switching levels fold-over technique to overcome the limitation of the classical one. The second contribution is on presenting a comparison study among the widely used fold-over techniques to help experimenters to recommend a suitable fold-over technique for their experiments by answering the following two fundamental questions: Do these techniques dramatically lessen the confounding of the initial designs, and do the resulting combined designs (combining initial design with its fold-over) via these techniques have considerable difference from the optimality point of view considering the markedly different searching domains in each technique? The optimality criteria are the aberration, confounding, Hamming distance and uniformity. Many of these criteria are given in sequences (patterns) form, which are inconvenient and costly to represent and compare, especially when the designs have many factors. The third innovation is on developing a new criterion (dictionary cross-entropy loss) to simplify the existing criteria from

WENG Lin-Chen
Division of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai 519085, China.
ELSAWAH A M (Corresponding author)
Division of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai 519085, China; Department of Mathematics, Faculty of Science, Zagazig University, Zagazig 44519, Egypt. Email: a_elsawah85@yahoo.com; amelsawah@uic.edu.cn; a.elsawah@zu.edu.eg.
FANG Kai-Tai
Division of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai 519085, China; The Key Lab of Random Complex Structures and Data Analysis, Chinese Academy of Sciences, Beijing 100190, China.
*This research was supported by the Beijing Normal University-Hong Kong Baptist University United International College under Grant Nos. R201810, R201912 and R202010, and the Zhuhai Premier Discipline Grant.
° *This paper was recommended for publication by Editor ZHU Liping.*

中国科学: 数学 2020年 第50卷 第5期: 561～570
SCIENTIA SINICA Mathematica
综 述

均匀设计理论与应用

献给方开泰教授 80 华诞

贺平[1], 林共进[2], 刘民千[3], 许青松[4], 周永道[3*]

1. 北京师范大学 - 香港浸会大学联合国际学院理工科技学部, 珠海 519087;
2. Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA;
3. 南开大学统计与数据科学学院, 天津 300071;
4. 中南大学数学与统计学院, 长沙 410083
E-mail: heping@uic.edu.hk, dkl5@psu.edu, mqliu@nankai.edu.cn, qsxu@csu.edu.cn, ydzhou@nankai.edu.cn

收稿日期: 2020-03-03; 接受日期: 2020-03-07; 网络出版日期: 2020-05-06; * 通信作者
国家自然科学基金 (批准号: 11771220 和 11871288)、国家 "万人计划" 科技创新领军人才项目 (批准号: W03020195)、天津市人才发展特殊支持计划 "高层次创新创业团队" 项目 (批准号: TJTZJH-GCCCXCYTD-2-23)、天津市 "131" 创新型人才团队项目、天津市自然科学基金 (批准号: 19JCZDJC31100) 资助项目

摘要 随着科学和技术的发展, 试验涉及的因素越来越多, 它们之间的关系更加复杂, 特别是在高科技的发展中, 面临多因素、非线性和模型未知等复杂性, 因此, 如何科学地组织试验就显得非常重要. 常见的试验设计类型有正交设计、均匀设计和最优设计等. 均匀设计的主要思想是把设计点均匀地散布在试验区域中, 它具有试验点数可灵活选取、对模型稳健、适用于多类试验区域等诸多优点. 本文综述均匀设计理论发展过程、最近进展及应用现状.

关键词 计算机试验 偏差 优化 均匀设计

MSC (2010) 主题分类 62K15

1 引言

试验设计在工业创新中发挥着重要作用. 试验设计在中国的广泛应用肇始于日本统计学家 Genichi Taguchi 于 20 世纪 60 年代到北京大学讲学, 他系统地介绍了有关正交设计的方法. 之后, 北京大学教师把相关内容整理为一本试验设计讲义, 这是一本优秀的普及正交设计的教材. 应用 Taguchi 的方法, 中国科学院数学研究所的统计工作者在指导工厂试验的过程中, 遇到了许多多因素试验的试验区域很大且因素与响应之间存在非线性关系的情形. 那时, 工程师不知道统计学中的科学试验设计和分析, 走了许多弯路, 有些试验甚至做几年都还没有达到目的. 对于这些复杂的问题, 统计学家意识到, Taguchi 的方法需要改进. 他们给出如下建议: (1) 因素的水平可以不止 2 个, 通常可取 3 到 5 个水平;

英文引用格式: He P, Lin D K J, Liu M Q, et al. Theory and application of uniform designs (in Chinese). Sci Sin Math, 2020, 50: 561–570, doi: 10.1360/SSM-2020-0065

中国科学: 数学 2020 年 第 50 卷 第 5 期: 571～584
SCIENTIA SINICA Mathematica
综 述

多元统计分析及其应用

献给方开泰教授 80 华诞

李刚[1], 梁家卷[2*], 潘建新[3], 彭小令[4], 田国梁[5]

1. Department of Biostatistics, Jonathan and Karin Fielding of Public Health, University of California at Los Angeles, Los Angeles, CA 90095-1772, USA;
2. College of Business, University of New Haven, West Haven, CT 06516, USA;
3. Department of Mathematics, The University of Manchester, Manchester M13 9PL, UK;
4. 北京师范大学 - 香港浸会大学联合国际学院理工科技学部, 珠海 519087;
5. 南方科技大学理学院统计与数据科学系, 深圳 518055
E-mail: vli@ucla.edu, jliang@newhaven.edu, jianxin.pan@manchester.ac.uk, xlpeng@uic.edu.hk, tiangl@sustech.edu.cn

收稿日期: 2020-03-09; 接受日期: 2020-03-10; 网络出版日期: 2020-05-07; * 通信作者

摘要 自 20 世纪 50 年代以来, 多元统计的理论、方法及其应用受到了越来越广泛的关注. 国内多元统计方向的研究始于 20 世纪 30 年代末至 40 年代初许宝騄在西南联合大学时期. 现代大数据分析的需要使得古典多元统计方法不能完全有效地解决当前的实际问题. 古典多元统计理论从 20 世纪 70 年代以来已经得到了快速发展, 本文旨在对国内学者在推广古典多元统计理论及其应用方面的工作进行概述, 主要包括: 多元统计分析和广义多元统计、一般对称多元分布、增长曲线模型及其他方向. 广义多元统计是正态假设下的传统统计方法论的推广. 其目的是将传统的统计方法论, 如参数估计、假设检验和统计模型等, 推广到更大的多元分布族. 这个分布族称为椭球等高分布族. 一般对称多元分布构成一个更大的多元统计分布族. 这个分布族包含了椭球等高分布族作为其特例. 增长曲线模型包含了一类统计方法, 它允许考虑个体内部及个体之间随着时间变化时的相关关系. 异常观察点及影响观察点的辨别是增长曲线模型研究的一个重要方向.

关键词 copula l_1- 模对称分布 球对称分布 随机表示 椭球等高分布 占有问题 增长曲线模型 左球矩阵分布

MSC (2010) 主题分类 62H10, 62H99

1 引言

多元统计是研究客观事物中多个变量 (或多个因素) 之间相互依赖的规律性的一个统计学分支. 如果研究个体有多个观察数据, 每个观察数据点可看作高维空间中的一个点. 众多的数据点组成高维

英文引用格式: Li G, Liang J J, Pan J X, et al. Multivariate statistics and its applications (in Chinese). Sci Sin Math, 2020, 50: 571–584, doi: 10.1360/SSM-2020-0071

Chapter 4
A Review of Prof. Kai-Tai Fang's Contribution to the Education, Promotion, and Advancement of Statistics in China

Gang Li and Xiaoling Peng

Abstract As an eminent leader in the field of statistics, Prof. Kai-Tai Fang has made impactful contributions to the application, promotion, education and advancement of statistics in China. Under his leadership, his team had completed some of the China's hallmark industrial projects through novel applications of statistics and developments of new statistical methodologies. He has authored/coauthored a series of best-selling modern statistics textbooks, taught numerous workshops and short courses, and mentored a large number of students. He has been active in promoting scholastic exchanges and organizing national and international statistics conferences. He has also served on the leadership of many national and international statistics organizations and on the editorial boards of many major statistical journals. This article provides a selective review of Prof. Fang's contributions to the education, promotion, and advancement of statistics in China.

4.1 Background

Since the early twentieth century, statistics has seen a flourishing development, and the modern data-centric statistical data science has received extensive recognitions with widespread applications in all industries. In past decades, more and more Chinese statisticians started to show their talents in international statistical academia, and gained unprecedented recognition and attention. As one of the most influential pioneers of statistics in China, Prof. Kai-Tai Fang has dedicated himself to the education, promotion, and advancement of statistics in China during his entire

G. Li
Departments of Biostatistics and Biomedicine, UCLA, Los Angeles, CA 90095-1772, USA
e-mail: vli@ucla.edu

X. Peng (✉)
Division of Science and Technology, BNU-HKBU United International College, Zhuhai 519087, China
e-mail: xlpeng@uic.edu.hk

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_4

Pediatric Neurology 105 (2020) 35–40

Contents lists available at ScienceDirect

Pediatric Neurology

journal homepage: www.elsevier.com/locate/pnu

Original Article

Seizure Characteristics, Outcome, and Risk of Epilepsy in Pediatric Anti-*N*-Methyl-D-Aspartate Receptor Encephalitis

Xin-ping Qu, MD [a, b, c, d, e], Jorge Vidaurre, MD [f], Xiao-ling Peng, PhD [g], Li Jiang, MD [a, b, c, d, e], Min Zhong, MD [a, b, c, d, e], Yue Hu, MD, PhD [a, b, c, d, e, *]

[a] Department of Neurology, Children's Hospital of Chongqing Medical University, Chongqing, China
[b] Ministry of Education Key Laboratory of Child Development and Disorders, Chongqing, China
[c] National Clinical Research Center for Child Health and Disorders, Chongqing, China
[d] China International Science and Technology Cooperation Base of Child Development and Critical Disorders, Chongqing, China
[e] Chongqing Key Laboratory of Pediatrics, Chongqing, China
[f] Division of Pediatric Neurology, Department of Pediatrics, The Ohio State University, Nationwide Children's Hospital, Chongqing, China
[g] Division of Science and Technology, Beijing Normal University-Hongkong Baptist University United International College, Zhuhai, China

ARTICLE INFO

Article history:
Received 31 August 2019
Accepted 4 November 2019
Available online 30 November 2019

Keywords:
Pediatrics
Anti-NMDAR encephalitis
Seizures
EEG
Status epilepticus
Autoimmune encephalitis

ABSTRACT

Background: We identified seizure characteristics, long-term outcome, and predictors of persistent seizures in children with anti-*N*-methyl-D-aspartate receptor (anti-NMDAR) encephalitis.
Method: Data were analyzed from patients with anti-NMDAR encephalitis who presented with seizures at our center between August 2012 and June 2018.
Results: Sixty-two of 86 patients with anti-NMDAR encephalitis experienced seizures. Seizures occurred within two weeks of disease onset in 58 of 62 (93.6%) patients; 36 of 62 (58.1%) had seizures as the initial symptom. Males were more likely to exhibit seizures as the initial symptom ($P = 0.039$). More than a quarter of patients (17 of 62, 27.4%) manifested two or more seizure types. Focal seizures were the most common (46 of 62, 74.2%). Status epilepticus occurred in 27 of 62 (43.5%) patients, and nonconvulsive status epilepticus, in two of 62 (3.2%) patients. No patient developed refractory status epilepticus. No systemic tumors were found. Electroencephalographic abnormalities included background slowing (77.4%), absence of a posterior dominant rhythm (62.9%), interictal epileptic discharges (50.0%), and extreme delta brush (6.5%). In the acute phase, 45 patients (45 of 62, 72.6%) received antiepileptic drugs. Persistent seizures occurred in only five of 62 (8%) patients. On univariate analysis, status epilepticus and combination antiepileptic drug treatment were associated with persistent seizures, but neither independently predicted persistent seizures.
Conclusions: Multiple seizure types may develop at any stage of anti-*N*-methyl-D-aspartate receptor encephalitis. Refractory status epilepticus, systemic tumors, and extreme delta brush in electroencephalography are rare in pediatric patients. Anti-NMDAR encephalitis–associated seizures appear to have good prognosis, without the need for long-term antiepileptic drug treatment.

Introduction

Anti-*N*-methyl-D-aspartate receptor (anti-NMDAR) encephalitis is the most common autoimmune encephalitis in children. This condition is frequently associated with seizures, which occur mostly during the acute phase of illness. The NMDAR, located in the postsynaptic membrane, is an ionotropic excitatory glutamate receptor composed of three subunits, NR1, NR2, and NR3. The cause of seizures in anti-NMDAR encephalitis is yet to be elucidated, but it is generally believed to be due to involvement of the NR1 subunit. Anti-NMDAR antibody binds to the NR1 subunit *N*-terminal extracellular epitope, causing reversible and selective decrease of synaptic NMDA (mainly in the hippocampus) by a mechanism of cross-linking and internalization. This interferes with the signaling of excitatory glutamate, leading to glutamate accumulation and

Conflicts of interest: The authors report no conflicts of interest. The authors take full responsibility for the content and writing of the paper.

* Communications should be addressed to: Hu; Department of Neurology; Children's Hospital of Chongqing Medical University; No.136 Zhongshan 2nd Road, Yu Zhong District, Chongqing 400014, China.
E-mail address: huyue915@163.com (Y. Hu).

https://doi.org/10.1016/j.pediatrneurol.2019.11.011
0887-8994/

Power Consumption and GoS Tradeoff in Cellular Mobile Networks with Base Station Sleeping and Related Performance Studies

Jingjin Wu, *Member, IEEE*, Eric W. M. Wong, *Senior Member, IEEE*, Yin-Chi Chan, *Member, IEEE* and Moshe Zukerman, *Life Fellow, IEEE*

***Abstract*—Mobile network operators usually consider power consumption and Grade of Service (GoS) as two important aspects in the design and planning of modern cellular networks. Base station (BS) sleeping is an effective approach to reduce the power consumption of the network, by switching some of the BSs to a low-power "sleep mode" during off-peak traffic hours. In this paper, we model each BS with sleeping mechanism as an M/G/1/K queue with vacations, and the entire cellular network as a network of such queues, to incorporate practical factors in BS sleeping, such as close-down and startup periods and additional power consumption for activating a sleeping BS. We investigate the power consumption and GoS under three BS sleeping schemes: (1) the isolated scheme, in which each BS switches between active and sleep modes based on its own real-time traffic load, (2) the cooperative scheme, in which selective BSs are switched to long-term sleep and traffic is allowed to overflow from sleeping BSs to nearby active BSs, and (3) the hybrid scheme, in which some BSs are switched to long-term sleep and other BSs switch modes according to their real-time traffic load. A robust, scalable and computationally efficient analytical method is proposed to evaluate GoS metrics, including mean delay and blocking probability, and power consumption under each scheme. We validate the accuracy of the proposed method, demonstrate the trade-off among power consumption, blocking probability and mean delay, and compare the performance of the three schemes via extensive and statistically reliable numerical experiments.**

***Index Terms*—Base station sleeping, performance analysis, teletraffic model, power-performance tradeoff**

I. Introduction

Recently, base station (BS) sleeping has emerged as an effective approach to reduce power consumption in cellular mobile networks [1]. Energy saving is achieved by switching BSs (or certain components of them) to a low power-consuming mode called "sleep mode" during non-busy hours when traffic in the network is relatively low. As BSs consume up to 80% of energy in cellular networks, BS sleeping may reduce a considerable amount of power consumption [2].

BS sleeping belongs to a broad family of approaches aiming at improving the energy efficiency of cellular networks by adjusting the transmitting power of BSs [3]. A BS selected to sleep reduces its transmit power to zero while neighboring active BSs increase their transmit power to maintain coverage. Compared to other power-saving approaches for green cellular networks, including applying renewable energy solutions or upgrading hardware components, BS sleeping can be implemented in existing network infrastructure and is thus considered more cost-effective [2]. On the other hand, switching some BSs to sleep mode leads to a reduction in the network capacity. Therefore, network operators must accurately evaluate the Grade of Service (GoS) metrics and investigate the impact of different BS sleeping strategies on the GoS [2], [4]. In this paper, we provide new methods to evaluate GoS measures such as blocking probability and mean delay. Such methods can be used to obtain accurate numerical values for such measures under various BS sleeping schemes that help us assess the trade-off between power consumption and GoS metrics.

In particular, we consider a cellular network, where each BS is modeled as a single-server queue, fed by arrivals that follow a Poisson process, with a finite buffer size of K and generally distributed service times. This queueing model is known as the M/G/1/K queue. The assumption of Poisson arrivals has been applied for modeling the *Busy Hour Traffic*, which refers to network traffic load during the busiest hour, in existing research on teletraffic models [5]. We will demonstrate that our proposed method can still obtain accurate evaluations when this assumption is relaxed in Section V. The generally distributed service time addresses various factors that may affect the service time of a user request in a cellular network, such as the application type, the amount of data to transmit and the channel condition. We refer to the parameter K, which represents the maximum number of requests that a BS can serve concurrently, as the capacity of the BS. We further consider vacations with startup and close-down times in the queue to model the operation of BS sleeping [6]. Henceforth, we will use the notation M/G/1/K queue to represent the general case of this queue with or without a range of modeling extensions, including vacations, startup and close-down times. In cases where reference to a specific case is important for clarity, we will specify the particular modeling extension that

J. Wu is with the Department of Statistics, BNU-HKBU United International College, Zhuhai, Guangdong, China, and also with the Department of Electrical Engineering, City University of Hong Kong, Hong Kong SAR, China (Email:jj.wu@ieee.org).

E. W. M. Wong, Y. C. Chan and M. Zukerman are with the Department of Electrical Engineering, City University of Hong Kong, Hong Kong SAR, China (Email: eeewong@cityu.edu.hk, ycchan26@cityu.edu.hk, m.zu@cityu.edu.hk).

A preliminary version of this paper was presented at IEEE GLOBECOM 2017, and was published in its Proceedings.

The work described in this paper was partly supported by College Research Grant from BNU-HKBU United International College [UIC-R201911] and a grant from the Innovation and Technology Fund (ITF) of the Hong Kong Special Administrative Region, China [ITS/191/16].

Millimeter-Wave Base Station Deployment Using the Scenario Sampling Approach

Miaomiao Dong, Taejoon Kim, *Senior Member, IEEE*, Jingjin Wu, *Member, IEEE*, and Eric Wing-Ming Wong, *Senior Member, IEEE*

***Abstract*—While the Poisson point process (PPP) has been widely employed to model the user distribution in many network design problems, an existing challenge is that it often reveals inaccuracy in small-cell networks. In this paper, instead of employing PPP, we capture the randomness of user equipment (UE) by collecting many their realizations. Specifically, we focus on the millimeter-wave (mmWave) base station (BS) deployment problem in an urban geometry, based on the application of a scenario sampling approach, previously introduced for large-scale optimization, to quantitatively sample a portion of the UE realizations. Motivated by the scenario sampling, a reduced-scale mmWave BS deployment problem is formulated, whose optimal solution is attained by the proposed low-complexity iterative search algorithm. A required number of samples that guarantee a specified majority of the link quality constraints is analyzed. Simulation results verify the scenario sampling theory and the effectiveness of the proposed algorithm.**

***Index Terms*—Millimeter-wave networks, base station deployment, scenario sampling, large-scale integer linear programming.**

I. Introduction

The ever-growing number of high data rate mobile applications coupled with the promise of connecting billions of user equipments (UEs) is demanding the deployment of millimeter-wave (mmWave) small-cell networks at large scale. Due to the high frequency and directional transmission, millimeter wave (mmWave) links are less susceptible to interference but are more vulnerable to physical blockage caused by obstacles in urban geometry [1]. Moreover, the capacity of each mmWave base station (BS)[1], equipped with hybrid antenna arrays, is strictly limited by the number of radio frequency (RF) chains, in which capacity-limited blockage occurs when all RF chains are occupied by user equipments (UEs).

To mitigate both the physical and capacity-limited blockages, multiple BSs can be installed to cover a region. The more the number of deployed BSs, the more UEs can be served with reduced blockages. However, deploying more BSs comes at the price of higher network deployment and maintenance costs. To mitigate this issue, cost-efficient approaches that minimize the number of deployed BSs while keeping the outage of each UE below a certain threshold, has been considered more recently [2], [3]. However, solution approaches to the latter problems largely depend on the UE deployment statistics in the geometry of interest [4]. While the Poisson point process (PPP) was validated for macro-cellular networks to capture the randomness of UEs, it often reveals inaccuracy in the small-cell dense mmWave networks [5]. Holding both the accuracy and tractability of a stochastic model is often a dilemma because geometry-dependency exponentially adds complexity to network design problems. It is within this context that in the mmWave BS deployment literature, the UE distribution is often ignored [6], [7] or many simplifying assumptions have been made [2], [3]. An alternative approach is to rely on available UE realization data that have been measured at different time scales to sufficiently capture the randomness of UEs[2]. Although a major benefit of leveraging the measurement data is its accuracy, involving all realizations to solve the BS deployment problem is yet computationally prohibitive.

In this work, we propose a scenario sampling approach, previously studied in the context of large-scale convex problems [9] to solve the blockage probability-guaranteed minimum-cost mmWave BS deployment problem in an urban geometry. For each UE realization, we analyze the physical and capacity-limited blockage probabilities and formulate the mmWave BS deployment into a large-scale integer linear problem (ILP), which is non-convex and excessively complex to be directly solved. The scenario sampling approach is then applied to form a small-scale ILP. The global optimality of the samll-scale ILP is achieved by the proposed low-complexity iterative search algorithm. However, any domain reduction introduced by the sampling could lead to substantial suboptimality. Thus, a required number of samples that ensure the optimal solution of the reduced problem satisfies a specified majority of the UE realizations is derived. We perform numerical simulations to corroborate the established analysis as well as effectiveness of the proposed algorithm.

It should be noted that the mmWave BS deployment problem in this work is approached in a link-connectivity point-of-view with the primary focus on providing the initial connectivity to UEs with the blockage tolerance guarantees. Therefore, our major focus is not to describe a physical-layer algorithmic

M. Dong, E. W. M. Wong are with the Department of Electrical Engineering, City University of Hong Kong, Hong Kong, emails: miao4600@163.com, eeewong@cityu.edu.hk. T. Kim is with the Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA, email: taejoonkim@ku.edu. J. Wu is with the Department of Statistics, BNU-HKBU United International College, Zhuhai, Guangdong, China, email: jj.wu@ieee.org.

T. Kim was supported in part by the National Science Foundation (NSF) under CNS 1955561 and AST 2037864. J. Wu was supported in part by BNU-HKBU UIC College Research Grant 201811 and 201911.

[1]BS capacity refers to the maximum number of UEs that a BS can simultaneously serve.

[2]For example, UE placements in a network can be collected by the existing LTE BSs using positioning reference signals as described in [8].

Discrete Mathematics 343 (2020) 111857

Contents lists available at ScienceDirect

Discrete Mathematics

journal homepage: www.elsevier.com/locate/disc

Characterising elliptic solids of $Q(4, q)$, q even

S.G. Barwick [a], Alice M.W. Hui [b,*], Wen-Ai Jackson [a,1]

[a] School of Mathematical Sciences, University of Adelaide, Adelaide, 5005, Australia
[b] Statistics Program, BNU-HKBU United International College, Zhuhai, China

ARTICLE INFO

Article history:
Received 7 December 2018
Received in revised form 4 February 2020
Accepted 5 February 2020
Available online xxxx

Keywords:
Projective geometry
Quadrics
Hyperplanes

ABSTRACT

Let $\mathcal{E}$ be a set of solids (hyperplanes) in PG(4, q), q even, $q > 2$, such that every point of PG(4, q) lies in either 0, $\frac{1}{2}(q^3 - q^2)$ or $\frac{1}{2}q^3$ solids of $\mathcal{E}$, and every plane of PG(4, q) lies in either 0, $\frac{1}{2}q$ or q solids of $\mathcal{E}$. This article shows that $\mathcal{E}$ is either the set of solids that are disjoint from a hyperoval, or the set of solids that meet a non-singular quadric $Q(4, q)$ in an elliptic quadric.

1. Introduction

This article gives a characterisation of the elliptic hyperplanes of a non-singular quadric $Q(4, q)$ of PG(4, q), see [7, Chapter 22] for background on non-singular quadrics. Note that we refer to the hyperplanes of PG(4, q) as solids. There are a number of known characterisations of points, lines, planes and solids related to the non-singular quadric $Q(4, q)$.

In 1956, Tallini [10] characterised sets of points in PG(n, q) by their intersection numbers with lines. A full version of Tallini's result is given in English in [7, Theorem 22.11.13]. We use a specialisation of this characterisation to PG(4, q).

Result 1.1. *A set $\mathcal{A}$ of $q^3 + q^2 + q + 1$ points in PG(4, q), $q \neq 2$, such that every line contains 0, 1, 2 or $q + 1$ points of $\mathcal{A}$ is either a solid or a non-singular quadric.*

Buekenhout [3] characterised a set of points in PG(n, q) by looking at its plane intersections; and a modification specific to PG(4, q) was given by Butler in [4]. Ferri and Tallini [6] and Schillewaert [9] prove characterisations of the pointset of $Q(4, q)$ using plane and solid intersections. Note that it is not possible to characterise quadrics by solid intersection numbers alone due to the existence of quasi-quadrics [5].

When q is odd, there is a polarity associated with $Q(4, q)$ which interchanges planes and lines. Using this polarity, we can dualise these characterisations, giving for example, characterisations of the tangent solids of a non-singular quadric $Q(4, q)$, q odd. Our motivation was to look for a characterisation of the solids of $Q(4, q)$ when q is even. This article gives a characterisation involving the elliptic solids of a non-singular quadric $Q(4, q)$, q even. A characterisation of the hyperbolic solids of a non-singular quadric $Q(4, q)$, q even, is given in [1].

* Corresponding author.
E-mail addresses: susan.barwick@adelaide.edu.au (S.G. Barwick), alicemwhui@uic.edu.hk (A.M.W. Hui), wen.jackson@adelaide.edu.au (W.-A. Jackson).

[1] A.M.W. Hui acknowledges the support of the Young Scientists Fund (Grant No. 11701035) of the National Natural Science Foundation of China.

https://doi.org/10.1016/j.disc.2020.111857
0012-365X/

Designs, Codes and Cryptography
https://doi.org/10.1007/s10623-019-00669-y

Characterising hyperbolic hyperplanes of a non-singular quadric in PG(4, q)

S. G. Barwick[1] · Alice M. W. Hui[2] · Wen-Ai Jackson[1] · Jeroen Schillewaert[3]

Received: 12 November 2018 / Revised: 19 May 2019 / Accepted: 25 July 2019

Abstract
Let $\mathcal{H}$ be a non-empty set of hyperplanes in PG$(4, q)$, q even, such that every point of PG$(4, q)$ lies in either 0, $\frac{1}{2}q^3$ or $\frac{1}{2}(q^3 + q^2)$ hyperplanes of $\mathcal{H}$, and every plane of PG$(4, q)$ lies in 0 or at least $\frac{1}{2}q$ hyperplanes of $\mathcal{H}$. Then $\mathcal{H}$ is the set of all hyperplanes which meet a given non-singular quadric $Q(4, q)$ in a hyperbolic quadric.

Keywords Projective geometry · Quadrics · Hyperplanes

Mathematics Subject Classification 51E20

1 Introduction

This article gives a characterisation of the hyperbolic hyperplanes of a non-singular quadric $Q(4, q)$ of PG$(4, q)$, see [8] for background on non-singular quadrics. Note that we refer to the hyperplanes of PG$(4, q)$ as solids. There are a number of known characterisations of points, lines, planes and solids relating to the non-singular quadric $Q(4, q)$. In 1956, Tallini [10] characterised sets of points in PG(n, q) by their intersection numbers with lines.

Communicated by J. D. Key.

✉ S. G. Barwick
susan.barwick@adelaide.edu.au

Alice M. W. Hui
alicemwhui@uic.edu.hk; huimanwa@gmail.com

Wen-Ai Jackson
wen.jackson@adelaide.edu.au

Jeroen Schillewaert
j.schillewaert@auckland.ac.nz

[1] School of Mathematical Sciences, University of Adelaide, Adelaide 5005, Australia

[2] Statistics Program, BNU-HKBU United International College, Zhuhai, China

[3] Department of Mathematics, University of Auckland, Auckland, New Zealand

Published online: 01 August 2019

Thomas Holgersson
Martin Singull *Editors*

Recent Developments in Multivariate and Random Matrix Analysis

Festschrift in Honour of Dietrich von Rosen

Springer

Chapter 3
Convexity of Sets Under Normal Distribution in the Structural Alloy Steel Standard

Kai-Tai Fang, Zhen Luo, and Yung Liang Tong

Abstract The paper is motivated by the structural alloy steel standard that has been used in China for a long period. This standard indicates the scope of several chemical elements in the steel and requests several mechanical properties for qualification. Fang and Wu (Acta Math Appl Sin 2:132–148, 1979) established the relationships between the percents of the controlled chemical elements and testing mechanical properties by a multivariate regression model, and proposed the algorithm for calculating qualification rate. Moreover, they proved the existence of the optimal chemical element combination. However, the uniqueness of the optimal solution for high dimensional case has been left. This open question is equivalent to showing the convexity of a type of probability sets under multivariate normal distribution. This paper proves that the open question is true.

3.1 Motivation

In 1973, the first author of this paper was invited to join an important project by the Ministry of Metallurgy of China to review the national standard for the structural alloy steel 20CrMnTi. In addition to controlling the carbon element, the structural alloy steel also needs chromium (Cr), manganese (Mn), nickel (Ni), molybdenum (Mo), silicon (Si), and titanium (Ti). The content of these elements must fall into the scope according to the national standards. Furthermore, the five mechanical properties such as strength, elasticity, etc., must exceed a certain threshold. Let

K.-T. Fang (✉)
Division of Science and Technology, BNU-HKBU United International College Zhuhai, Zhuhai, China
e-mail: ktfang@uic.edu.hk

Z. Luo
Pfizer China, Pudong, Shanghai, China
e-mail: zhen.luo@pfizer.com

Y. L. Tong
Department of Mathematics, Georgia Institute of Technology, Atlanta, GA, USA

T. Holgersson, M. Singull (eds.), *Recent Developments in Multivariate and Random Matrix Analysis*, https://doi.org/10.1007/978-3-030-56773-6_3

Taylor & Francis

Communications in Statistics - Simulation and Computation

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

New non-isomorphic detection methods for orthogonal designs

Xiao Ke , Kai-Tai Fang , A.M. Elsawah & Yuxuan Lin

To cite this article: Xiao Ke , Kai-Tai Fang , A.M. Elsawah & Yuxuan Lin (2020): New non-isomorphic detection methods for orthogonal designs, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2020.1844895

To link to this article: https://doi.org/10.1080/03610918.2020.1844895

Semi-Supervised Learning for Fault Identification in Electricity Distribution Networks

Xinyang Li[a], Hongfa Meng[b], Xiaoling Peng*[c]
[a]School of Mathematical Sciences, Ocean University of China, Qingdao, China 266100
[b]Guangzhou Stratac Information Technology Co. Ltd, Gaungzhou, China 511400
[c]Division of Science and Technology, BNU-HKBU United International College, Zhuhai, China 519087

ABSTRACT

The detection and identification of faults in electricity distribution networks is essential in improving the reliability of power supply. After observing many fault current signals we found that: (1) features of many recorded fault electrical signals were unknown or obscure; (2) the fault types of most sample signals had no clear definition, that is, the labeled sample were very limited. In this situation, the semi-supervised support vector machine (S3VM) and SVM active learning were firstly introduced to distinguish the short circuit and grounding in distribution networks. We used wavelet packet analysis to extract features based on energy spectrum as the physical features of electric signals, then some statistical characteristics were also computed and selected to form a mixed feature set. A case study was conducted on a real data set including 72 labeled and 7720 unlabeled electrical signals for fault diagnosis. By performing transductive support vector machine (TSVM) and SVM active learning with mixed features, our experimental results showed that both of the two models can effectively identify the fault types. Meanwhile, the accuracy of TSVM is higher than that of SVM active learning.

Keywords: circuit fault classification, S3VM, wavelet packet energy spectrum analysis, statistical features

1. INTRODUCTION

With the rapid development of China's electrification process, the coverage of power grid in China is getting much wider than before, which puts forward higher requirements for the management of power distribution network. The occurrence of faults in power distribution would affect people's daily life, make the production of enterprises and factories stagnate, which will also have a great impact on the operation safety and reliability of the power grid. Since the fault current is small and unstable, it is not easy to identify ungrounded neutrals fault and single-phase grounding (also known as small current grounding) fault in power distribution grounded by arc suppression coil. Recently, the application of small current grounding fault line selection technology based on fault transient information had largely improved the accuracy rate of fault line selection. However, there are still a high proportion of false positives and omissions in fault identification.

In past years, various machine learning techniques had been employed in fault diagnosis and had acquired good performance, such as BP neural network[1, 2], support vector machine (SVM)[3], improved genetic algorithm[4] and classification and regression tree[5, 6]. However, all these methods are known as supervised learning that can only deal with labeled current signals. That is, the fault type of each sample signal is known in the training set. But in practice, it is infeasible to label the fault type for each fault current since it will require massive manpower and effort. Therefore, although large number of real-time fault signals can be collected by the online monitoring system, most of them are unlabeled that cannot be used in supervised leaning. On the other side, unsupervised learning such as clustering, density estimation and anomaly detection can automatically group unlabeled electrical signals, but the result of grouping will not guarantee a classification of fault types with practical significance. Lie between supervised and unsupervised leaning, semi-supervised learning (SSL)[7] is trained by a small number of labeled instances and a large number of unlabeled instances. For inductive semi-supervised learning, the goal is to predict the labels on future test data, while for transductive semi-supervised learning, the goal is to predict the classes on the unlabeled instances in the training sample.

*xlpeng@uic.edu.cn; phone 86 756 3620623; fax 86 756 3620888

Twelfth International Conference on Signal Processing Systems, edited by Jie Tian, Dahong Qian, Yue Lyu, Kezhi Mao, Proc. of SPIE Vol. 11719, 117190B · © 2021 SPIE
CCC code: 0277-786X/21/$21 · doi: 10.1117/12.2589229

Journal of King Saud University – Science 32 (2020) 3328–3336

Contents lists available at ScienceDirect

Journal of King Saud University – Science

journal homepage: www.sciencedirect.com

Original article

An algorithm for outlier detection in a time series model using backpropagation neural network

Gajendra K. Vishwakarma [a,*], Chinmoy Paul [a,b], A.M. Elsawah [c,d]

[a] Department of Mathematics & Computing, Indian Institute of Technology Dhanbad, Dhanbad 826004, India
[b] Department of Statistics, Pandit Deendayal Upadhyaya Adarsha Mahavidyalaya, Eraligool, Karimganj 788723, India
[c] Division of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai 519085, China
[d] Department of Mathematics, Faculty of Science, Zagazig University, Zagazig 44519, Egypt

ARTICLE INFO

Article history:
Received 13 March 2020
Revised 18 August 2020
Accepted 11 September 2020
Available online 22 September 2020

Mathematics Subject Classification:
68Wxx

Keywords:
Multivariate outliers
Detection
Neural network
Robust estimate
Time series
Backpropagation algorithm

ABSTRACT

Outliers are commonplace in many real-life experiments. The presence of even a few anomalous data can lead to model misspecification, biased parameter estimation, and poor forecasts. Outliers in a time series are usually generated by dynamic intervention models at unknown points of time. Therefore, detecting outliers is the cornerstone before implementing any statistical analysis. In this paper, a multivariate outlier detection algorithm is given to detect outliers in time series models. A univariate time series is transformed to bivariate data based on the estimate of robust lag. The proposed algorithm is designed by using robust measures of location and dispersion matrix. Feed forward neural network is used for designing time series models. Number of hidden units in the network is determined based on the standard error of the forecasting error. A comparison study between the proposed algorithm and the widely used algorithms is given based on three real-data sets. The results demonstrated that the proposed algorithm outperformed the existing algorithms due to its non-requirement of a priori knowledge of the time series and its control of both masking and swamping effects. We also discussed an efficient method to deal with unexpected jumps or drops on share prices due to stock split and commodity prices near contract expiry dates.

1. Introduction

The detection of outliers or unusual data structures is one of the important tasks in the statistical analysis of time series data as outliers may have a substantial influence on the outcome of an analysis. Appropriate definition of an outlier usually depends on the assumptions about the structure of data and the applied detection method. Hawkins (1980) defined the outlier as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. Barnett and Lewis (1994) indicated that an outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs. Similarly, Johnson (1992) viewed that, an outlier is an observation in a data set which appears to be inconsistent with the remainder of that set of data. There are many definitions of outlier proposed in the literature of time series. Outlier observations in some situations are also referred as anomalies, discordant observations, or contaminants Carreno et al. (2019).

The presence of outliers in a time series has a significant effect on the results of standard procedures of analysis. The consequences may lead to improper model specification, faulty parameter estimation and substandard forecasting. A crucial point here is that any outlier detection technique can at most detect a set of data points having different behavior than the rest of the data and hence, it can be termed as a probable set of outliers. However, it is up an analyst to take various itineraries to come up with a final decision to justify these detected points as outliers. It is probable that a point detected as an outlier has some real facts behind it, e.g., the price of a stock just after the date of stock split with split ratio of 2-for-1 or 3-for-1, which means a stockholder gets two or three shares, respectively, for every share held. In a reverse stock

* Corresponding author.
E-mail addresses: vishwagk@rediffmail.com (G.K. Vishwakarma), chinmoy.goc@gmail.com (C. Paul), amelsawah@ucas.ac.cn (A.M. Elsawah).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksus.2020.09.018
1018-3647/

frontiers in Psychiatry

ORIGINAL RESEARCH
published: 26 November 2020
doi: 10.3389/fpsyt.2020.585915

The Medium-Term Impact of COVID-19 Lockdown on Referrals to Secondary Care Mental Health Services: A Controlled Interrupted Time Series Study

*Shanquan Chen[1], Rui She[2], Pei Qin[3], Anne Kershenbaum[1,4], Emilio Fernandez-Egea[1,4], Jenny R. Nelder[1], Chuoxin Ma[5], Jonathan Lewis[4], Chaoqun Wang[6] and Rudolf N. Cardinal[1,4]**

[1] Department of Psychiatry, University of Cambridge, Cambridge, United Kingdom, [2] The Jockey Club School of Public Health and Primary Care, The Chinese University of Hong Kong, Hong Kong, China, [3] Department of Biostatistics and Epidemiology, Shenzhen University Health Science Center, Shenzhen, China, [4] Cambridgeshire and Peterborough NHS Foundation Trust, Cambridge, United Kingdom, [5] Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom, [6] College of Public Administration, Central China Normal University, Wuhan, China

OPEN ACCESS

Edited by:
Wulf Rössler, Charité – Universitätsmedizin Berlin, Germany

Reviewed by:
Elizabeta Blagoja Mukaetova-Ladinska, University of Leicester, United Kingdom
Michaela Pascoe, Victoria University, Australia, Australia

***Correspondence:**
Rudolf N. Cardinal
rnc1001@cam.ac.uk

Specialty section:
This article was submitted to Public Mental Health, a section of the journal Frontiers in Psychiatry

Received: *21 July 2020*
Accepted: *28 October 2020*
Published: *26 November 2020*

Citation:
Chen S, She R, Qin P, Kershenbaum A, Fernandez-Egea E, Nelder JR, Ma C, Lewis J, Wang C and Cardinal RN (2020) The Medium-Term Impact of COVID-19 Lockdown on Referrals to Secondary Care Mental Health Services: A Controlled Interrupted Time Series Study. Front. Psychiatry 11:585915. doi: 10.3389/fpsyt.2020.585915

To date, there is a paucity of information regarding the effect of COVID-19 or lockdown on mental disorders. We aimed to quantify the medium-term impact of lockdown on referrals to secondary care mental health clinical services. We conducted a controlled interrupted time series study using data from Cambridgeshire and Peterborough NHS Foundation Trust (CPFT), UK (catchment population ~0.86 million). The UK lockdown resulted in an instantaneous drop in mental health referrals but then a longer-term acceleration in the referral rate (by 1.21 referrals per day per day, 95% confidence interval [CI] 0.41–2.02). This acceleration was primarily for urgent or emergency referrals (acceleration 0.96, CI 0.39–1.54), including referrals to liaison psychiatry (0.68, CI 0.35–1.02) and mental health crisis teams (0.61, CI 0.20–1.02). The acceleration was significant for females (0.56, CI 0.04–1.08), males (0.64, CI 0.05–1.22), working-age adults (0.93, CI 0.42–1.43), people of White ethnicity (0.98, CI 0.32–1.65), those living alone (1.26, CI 0.52–2.00), and those who had pre-existing depression (0.78, CI 0.19–1.38), severe mental illness (0.67, CI 0.19–1.15), hypertension/cardiovascular/cerebrovascular disease (0.56, CI 0.24–0.89), personality disorders (0.32, CI 0.12–0.51), asthma/chronic obstructive pulmonary disease (0.28, CI 0.08–0.49), dyslipidemia (0.26, CI 0.04–0.47), anxiety (0.21, CI 0.08–0.34), substance misuse (0.21, CI 0.08–0.34), or reactions to severe stress (0.17, CI 0.01–0.32). No significant post-lockdown acceleration was observed for children/adolescents, older adults, people of ethnic minorities, married/cohabiting people, and those who had previous/pre-existing dementia, diabetes, cancer, eating disorder, a history of self-harm, or intellectual disability. This evidence may help service planning and policy-making, including preparation for any future lockdown in response to outbreaks.

Keywords: COVID-19/SARS-CoV-2 coronavirus pandemic, lockdown, secondary care mental health services, controlled interrupted time series analysis, comorbidity

Computational Statistics and Data Analysis 155 (2021) 107110

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Joint generalized estimating equations for longitudinal binary data

Youjun Huang [a], Jianxin Pan [b,*]

[a] Mathematical College, Sichuan University, Chengdu 610065, China
[b] Department of Mathematics, The University of Manchester, Manchester M13 9PL, UK

ARTICLE INFO

Article history:
Received 30 March 2020
Received in revised form 30 September 2020
Accepted 30 September 2020
Available online 8 October 2020

Keywords:
Correlation coefficients
Generalized estimating equations
Joint mean and correlation parameter estimation
Longitudinally correlated binary data

ABSTRACT

Modeling longitudinal binary data is challenging but common in practice. Existing methods on modeling of binary responses take no account of the fact that the correlation coefficient of binary responses must have an upper bound which is smaller than one. Ignoring this fact can lead to incorrect statistical inferences for longitudinal binary data. A novel method is proposed to model the mean and within-subject correlation coefficients for longitudinal binary data, simultaneously, by taking into account the constraints of the upper bounds. By introducing latent normally distributed random variables, the correlation coefficients of binary responses are connected to those for the latent variables, of which the correlation coefficients are modeled accordingly. A joint generalized estimating equation (GEE) method is developed for this purpose and the resulting correlation coefficients are shown to satisfy the constraints. Asymptotic normality of the parameter estimators is derived and simulation studies are made under various scenarios, showing that the proposed joint GEE method works very well even if the working covariance structures are misspecified. For illustration, the proposed method is applied to two real data practices to assess the effects of covariates on the mean and within-subject correlation coefficients.

1. Introduction

Binary responses with repeat measurements are very common in many fields such as medical and biological sciences. Longitudinal normally distributed data have been studied well, while it is very challenging to model longitudinal binary data. A major issue is that the joint distribution of correlated binary data does not have an analytically tractable form. In the literature, various methods were proposed to model the conditional mean and marginal mean of correlated binary responses. The former was mainly studied within the framework of generalized linear mixed model (GLMM). For example, Heagerty (1999) considered a random effect model for longitudinal binary data. Wang and Louis (2003) developed an intercept random effect model with bridge distribution, and Parzen et al. (2011) extended the single random intercept model for longitudinal binary data. Although the GLMM-based approach produces good estimators of the fixed effects and has good predictions of the random effects, the within-subject correlations of binary responses remain unclear as they are analytically intractable. It becomes more difficult to understand how the within-subject correlations are associated with covariates of interest.

The latter was investigated by using generalized estimation equation (GEE) methods, see, e.g., Liang and Zeger (1986). With unnecessary assumption of distribution, the GEE methods directly solve estimating equations in order to obtain the

* Corresponding author.
E-mail address: jianxin.pan@manchester.ac.uk (J. Pan).

https://doi.org/10.1016/j.csda.2020.107110
0167-9473/

Chapter 13
Variable Selection in Joint Mean and Covariance Models

Chaofeng Kou and Jianxin Pan

Abstract In this paper, we propose a penalized maximum likelihood method for variable selection in joint mean and covariance models for longitudinal data. Under certain regularity conditions, we establish the consistency and asymptotic normality of the penalized maximum likelihood estimators of parameters in the models. We further show that the proposed estimation method can correctly identify the true models, as if the true models would be known in advance. We also carry out real data analysis and simulation studies to assess the small sample performance of the new procedure, showing that the proposed variable selection method works satisfactorily.

13.1 Introduction

In longitudinal studies, one of the main objectives is to find out how the average value of the response varies over time and how the average response profile is affected by different treatments or various explanatory variables of interest. Traditionally the within-subject covariance matrices are treated as nuisance parameters or assumed to have a very simple parsimonious structure, which inevitably leads to a misspecification of the covariance structure. Although the misspecification need not affect the consistency of the estimators of the parameters in the mean, it can lead to a great loss of efficiency of the estimators. In some circumstances, for example, when missing data are present, the estimators of the mean parameters can be severely biased if the covariance structure is misspecified. Therefore, correct specification of the covariance structure is really important.

On the other hand, the within-subject covariance structure itself may be of scientific interest, for example, in prediction problems arising in econometrics and finance. Moreover, like the mean, the covariances may be dependent on various explanatory variables. A natural constraint for modelling of covariance structures

C. Kou · J. Pan (✉)
Department of Mathematics, University of Manchester, Manchester, UK
e-mail: ckou@maths.man.ac.uk; jianxin.pan@manchester.ac.uk

T. Holgersson, M. Singull (eds.), *Recent Developments in Multivariate and Random Matrix Analysis*, https://doi.org/10.1007/978-3-030-56773-6_13

中国科学: 数学 2020 年 第 50 卷 第 5 期: 645 ~ 666
SCIENTIA SINICA Mathematica
论 文

具有序列相关结构的混合增长曲线模型

献给方开泰教授 80 华诞

潘雅婷[1], 费宇[1], 倪明明[1], 潘建新[2*]

1. 云南财经大学统计与数学学院, 昆明 650221;
2. Department of Mathematics, The University of Manchester, Manchester M13 9PL, UK
E-mail: Yating-Pan@hotmail.com, feiyukm@aliyun.com, mingming.ni90@gmail.com, Jianxin.Pan@manchester.ac.uk

收稿日期: 2019-05-23; 接受日期: 2019-09-06; 网络出版日期: 2020-05-08; * 通信作者
国家自然科学基金 (批准号: 11871357, 11561071 和 11971421) 和国家重点研发计划 (批准号: 2018YFC0831900) 资助项目

摘要 增长曲线模型是纵向数据分析的重要方法, 它的一个重要假设是组别阵已知, 但这在现实中往往难以满足. 因此, 本文提出全新的混合增长曲线模型处理该问题, 随之将数据分布假设从正态扩展到混合正态. 此外, 新模型中加入了序列相关结构, 为常伴有序列相关的纵向数据提供了合理的分析方法. 本文以数据驱动的方法来搜索最佳模型, 在获得模型参数估计的同时实现了序列相关数据的聚类与拟合. 模拟研究和实际数据分析都说明了所提出的方法具有合理性和有效性.

关键词 增长曲线模型 Gauss 混合模型 序列相关结构 最大期望算法 聚类

MSC (2010) 主题分类 62F99, 62H30, 62J99

1 引言

增长曲线模型 (growth curve model, GCM) [1] 是纵向数据分析的一种重要方法, 它在经济学、生物学和医学等领域有着广泛的应用. 该模型在一般的多元线性模型基础上引入组别设计阵, 实现了对观测个体纵向趋势的组别控制, 其一般定义如下 (参见文献 [2]):

$$\begin{cases} \boldsymbol{Y}_{p\times n} = \boldsymbol{X}_{p\times m}\boldsymbol{B}_{m\times r}\boldsymbol{Z}_{r\times n} + \boldsymbol{\varepsilon}_{p\times n}, \\ \boldsymbol{\varepsilon}_{p\times n} \sim N_{p,n}(\boldsymbol{0}, \boldsymbol{\Sigma}, \boldsymbol{I}_n), \end{cases} \tag{1.1}$$

其中 $\boldsymbol{X}_{p\times m}$ 和 $\boldsymbol{Z}_{r\times n}$ 都是已知设计矩阵, 它们的秩分别为 m 和 r ($m < p$ 且 $r < n$), 回归系数矩阵 $\boldsymbol{B}_{m\times r}$ 未知, 随机误差矩阵 $\boldsymbol{\varepsilon}_{p\times n}$ 的各列是相互独立的 p 元正态向量, 均值向量为 $\boldsymbol{0}$, 未知的 p 维协方差矩阵满足 $\boldsymbol{\Sigma} > \boldsymbol{0}$. 因此, 响应矩阵满足 $\boldsymbol{Y}_{p\times n} \sim N_{p,n}(\boldsymbol{XBZ}, \boldsymbol{\Sigma}, \boldsymbol{I}_n)$, 即 $\boldsymbol{Y}_{p\times n}$ 服从矩阵正态分布, 均值为 $\boldsymbol{XBZ}$, 个体间的协方差为 n 维单位方阵 $\boldsymbol{I}_n$, 即个体相互独立, 而每个个体内 p 次观测的相关关系由正定协方差矩阵 $\boldsymbol{\Sigma}$ 描述.

英文引用格式: Pan Y T, Fei Y, Ni M M, et al. Growth curves mixture model with serial covariance structure (in Chinese). Sci Sin Math, 2020, 50: 645–666, doi: 10.1360/N012019-00145

Fu et al. BMC Bioinformatics (2020) 21:121
https://doi.org/10.1186/s12859-020-3411-3

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Hellinger distance-based stable sparse feature selection for high-dimensional class-imbalanced data

Guang-Hui Fu[1*], Yuan-Jiao Wu[1], Min-Jie Zong[1] and Jianxin Pan[2]

Abstract

Background: Feature selection in class-imbalance learning has gained increasing attention in recent years due to the massive growth of high-dimensional class-imbalanced data across many scientific fields. In addition to reducing model complexity and discovering key biomarkers, feature selection is also an effective method of combating overlapping which may arise in such data and become a crucial aspect for determining classification performance. However, ordinary feature selection techniques for classification can not be simply used for addressing class-imbalanced data without any adjustment. Thus, more efficient feature selection technique must be developed for complicated class-imbalanced data, especially in the context of high-dimensionality.

Results: We proposed an algorithm called sssHD to achieve stable sparse feature selection applied it to complicated class-imbalanced data. sssHD is based on the Hellinger distance (HD) coupled with sparse regularization techniques. We stated that Hellinger distance is not only class-insensitive but also translation-invariant. Simulation result indicates that HD-based selection algorithm is effective in recognizing key features and control false discoveries for class-imbalance learning. Five gene expression datasets are also employed to test the performance of the sssHD algorithm, and a comparison with several existing selection procedures is performed. The result shows that sssHD is highly competitive in terms of five assessment metrics. In addition, sssHD presents limited differences between performing and not performing re-balance preprocessing.

Conclusions: sssHD is a practical feature selection method for high-dimensional class-imbalanced data, which is simple and can be an alternative for performing feature selection in class-imbalanced data. sssHD can be easily extended by connecting it with different re-balance preprocessing, different sparse regularization structures as well as different classifiers. As such, the algorithm is extremely general and has a wide range of applicability.

Keywords: Hellinger distance, Class-imbalance learning, Feature selection, Sparse regularization

Background

Feature selection has recently gained considerable attention in class-imbalance learning due to the high-dimensionality of class-imbalanced data across many scientific disciplines [1–3]. To date, a variety of feature selection methods have been proposed to address high-dimensional data. However, only a small number of them are technically designed to handle the problem of class distribution under a class-imbalance setting [4–7]. Thus, performing feature selection from class-imbalanced data remains a challenging task due to the inherent complex characteristics of such data, and a new understanding or principle is required to efficiently transform vast amounts of raw data into information and knowledge representation [8].

*Correspondence: guanghuifu@kust.edu.cn;ghuifu@126.com
[1] School of Science, Kunming University of Science and Technology, Kunming 650500, People's Republic of China
Full list of author information is available at the end of the article

Received: 11 August 2019 | Revised: 23 April 2020 | Accepted: 30 April 2020

DOI: 10.1002/er.5577

RESEARCH ARTICLE

INTERNATIONAL JOURNAL OF ENERGY RESEARCH WILEY

A novel method for measuring spatial uniformity of irregular boiling bubbles in a direct contact heat exchanger

Qin Wang[1,2] | Junwei Huang[3] | Jianxin Pan[4] | Hua Wang[2,5] | Jianxin Xu[2,5]

[1]Faculty of Management and Economics, Kunming University of Science and Technology, Kunming, China

[2]State Key Laboratory of Complex Nonferrous Metal Resources Clean Utilization, Kunming University of Science and Technology, Kunming, China

[3]Faculty of Mechanical and Electrical Engineering, Yunnan Agriculture University, Kunming, China

[4]School of Mathematics, The University of Manchester, Manchester, UK

[5]Faculty of Metallurgical and Energy Engineering, Kunming University of Science and Technology, Kunming, China

Correspondence
Jianxin Xu, State Key Laboratory of Complex Nonferrous Metal Resources Clean Utilization, Kunming University of Science and Technology, Kunming 650093, China.
Email: xujianxina@163.com

Funding information
China Scholarship Council, Grant/Award Number: 201608535030; National Natural Science Foundation of China, Grant/Award Numbers: 51666006, 51706195; Natural Science Foundation of Yunnan Province, Grant/Award Number: 2017FB093

Summary

As the binary image fails to reflect the density distribution of the bubble, this paper proposes an improved new indicator for accurate measurement of spatial uniformity based on gray image analysis: moment. Compared with the inclination angle method in binary image, the test results show that the two methods have significant positive correlation. The experimental results also indicate that the evolution curve of bubble numbers is consistent with its moment evolution curve, and the moment evolution fitting curve can distinguish the high noise in the experiment. Compared with other methods, mixing performance index (t) obtained by moment evolution fitting curve is correlated with heat transfer performance (Pearson correlation coefficient $a = -0.94$). The variation in characteristic area of the bubble in the direction of bubble growth can characterize the heat transfer performance. The index of dispersion of the global moment and the local moment indicates that the local moment can be used to precisely quantify the spatial uniformity of boiling bubbles. This work could also be applied to study a variety of problems involving spatial uniformity through visualization techniques.

KEYWORDS

direct contact heat exchanger, heat transfer coefficient, image analysis, moment, spatial uniformity

1 | INTRODUCTION

Direct contact heat exchanger refers to a heat exchanger that the heating medium and the heated medium exchange heat through direct contact. Compared with other heat exchangers such as heat pipe, direct contact heat exchanger have the advantages of transferring significant amounts of heat efficiently,[1] simple design, low cost[2] and low heat transfer resistance.[3] Currently, direct contact heat exchangers are widely used in numerous engineering systems, such as water desalination, crystallization, solar energy, power production and chemical industry.[4-8] Therefore, it is so critical to understand the flow and mixing characteristics and the proper design and optimization of the direct contact heat exchanger. The size distribution of bubble groups is an important parameter for predicting and designing the

Qin Wang, Junwei Huang, and Jianxin PanEqually contributed equally to this study.

Chapter 15
Estimation of Covariance Matrix with ARMA Structure Through Quadratic Loss Function

Defei Zhang, Xiangzhao Cui, Chun Li, and Jianxin Pan

Abstract In this paper we propose a novel method to estimate the high-dimensional covariance matrix with an order-1 autoregressive moving average process, i.e. ARMA(1,1), through quadratic loss function. The ARMA(1,1) structure is a commonly used covariance structures in time series and multivariate analysis but involves unknown parameters including the variance and two correlation coefficients. We propose to use the quadratic loss function to measure the discrepancy between a given covariance matrix, such as the sample covariance matrix, and the underlying covariance matrix with ARMA(1,1) structure, so that the parameter estimates can be obtained by minimizing the discrepancy. Simulation studies and real data analysis show that the proposed method works well in estimating the covariance matrix with ARMA(1,1) structure even if the dimension is very high.

Keywords ARMA(1,1) structure · Covariance matrix · Quadratic loss function

15.1 Introduction

Covariance matrix estimation is a fundamental problem in multivariate analysis and time series. Especially, the estimation of high-dimensional covariance matrix is rather challenging. In the literature, many research works were proposed to tackle the problem, such as [1, 3, 8, 9] among many others. However, when the covariance matrix has a certain of structures like order-1 autoregressive moving average, i.e. ARMA(1,1) structure or others, the estimation and regularization were hardly [6]. Recently, Lin et al. [7] proposed a new method to estimate and regularize the high-dimensional covariance matrix. Their idea is summarized as follows. Suppose A is a given $m \times m$ covariance matrix, that is, it is symmetric non-negative definite.

D. Zhang · X. Cui · C. Li
Department of Mathematics, Honghe University, Mengzi 661199, China

J. Pan ( )
Department of Mathematics, University of Manchester, Manchester M13 9PL, UK
e-mail: jianxin.pan@manchester.ac.uk

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_15

SCIENTIFIC REPORTS
nature research

OPEN

Integrative analysis of Mendelian randomization and Bayesian colocalization highlights four genes with putative BMI-mediated causal pathways to diabetes

Qian Liu[1,2], Jianxin Pan[3], Carlo Berzuini[1], Martin K. Rutter[4,5] & Hui Guo[1]✉

Genome-wide association studies have identified hundreds of single nucleotide polymorphisms (SNPs) that are associated with BMI and diabetes. However, lack of adequate data has for long time prevented investigations on the pathogenesis of diabetes where BMI was a mediator of the genetic causal effects on this disease. Of our particular interest is the underlying causal mechanisms of diabetes. We leveraged the summary statistics reported in two studies: UK Biobank (N = 336,473) and Genetic Investigation of ANthropometric Traits (GIANT, N = 339,224) to investigate BMI-mediated genetic causal pathways to diabetes. We first estimated the causal effect of BMI on diabetes by using four Mendelian randomization methods, where a total of 76 independent BMI-associated SNPs ($R^2 \leq 0.001$, $P < 5 \times 10^{-8}$) were used as instrumental variables. It was consistently shown that higher level of BMI (kg/m^2) led to increased risk of diabetes. We then applied two Bayesian colocalization methods and identified shared causal SNPs of BMI and diabetes in genes *TFAP2B*, *TCF7L2*, *FTO* and *ZC3H4*. This study utilized integrative analysis of Mendelian randomization and colocalization to uncover causal relationships between genetic variants, BMI and diabetes. It highlighted putative causal pathways to diabetes mediated by BMI for four genes.

Diabetes is a long term health condition that affects approximately 1 in 11 adults with rapid increase in prevalence worldwide[1]. Elevated BMI in both children and adults has been consistently found causally associated with the risk of diabetes[2–9]. Genome-wide association studies (GWASs) have identified hundreds of genetic variants, in particular, single nucleotide polymorphisms (SNPs) that are associated with both BMI and diabetes[10–14], which have induced investigations on the role of BMI-associated SNPs in the development of diabetes[2,15]. However, there was limited data on the pathogenesis of diabetes where BMI was a mediator of the genetic causal effects on this disease.

Publicly accessible large-scale GWAS summary results provide great resources of integrative analyses of disease pathogeneses[16–19], e.g., Mendelian randomization (MR)[20–22] and colocalization[9,23–26]. MR is designed for estimating causal effect of an exposure on a disease, where exposure associated SNPs are selected as instruments. These instruments are not necessarily causal SNPs due to linkage disequilibrium (LD). Colocalization explores shared causal SNPs of a pair of traits, whether they are exposures, diseases, or exposure and disease. It was not developed for identifying causal relationship between the traits. Thus, the causal questions addressed by the two approaches are different[27]. Each of the approaches alone is insufficient to investigate exposure-mediated genetic causal pathways to a disease. Very recently, frameworks of integrative analysis by combining MR with colocalization have been developed to identify biological mediators in the causal pathways to various clinical outcomes[28–30].

[1]Centre for Biostatistics, School of Health Sciences, The University of Manchester, Manchester, UK. [2]School of Mathematics and Statistics, Xidian University, Xi'an, China. [3]School of Mathematics, Faculty of Engineering and Physical Science, The University of Manchester, Manchester, UK. [4]Division of Endocrinology, Diabetes and Gastroenterology, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, UK. [5]Manchester Diabetes Centre, Central Manchester University NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK. ✉e-mail: hui.guo@manchester.ac.uk

International Journal of Refrigeration 112 (2020) 201–214

Contents lists available at ScienceDirect

International Journal of Refrigeration

journal homepage: www.elsevier.com/locate/ijrefrig

Synergistic effect of flow pattern evolution of dispersed and continuous phases in direct-contact heat transfer process

Jianxin Xu[a,f,1], Fanhan Liu[b], Qingtai Xiao[a,f,1], Junwei Huang[c,1], Yu Fei[e,*], Yunfei Yang[d,**], Yuling Zhai[f,**], Jianxin Pan[g,**], Hua Wang[a,**]

[a] State Key Laboratory of Complex Nonferrous Metal Resources Clean Utilization, Kunming University of Science and Technology, Kunming, Yunnan 650093, China
[b] College of Civil Engineering and Architecture, Jiaxing University, Jiaxing 314001, China
[c] Faculty of Mechanical and Electrical Engineering, Yunnan Agriculture University, Kunming, Yunnan 650100, China
[d] School of Management Engineering and Business, Hebei University of Engineering, Handan 056038, China
[e] School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming 650221, China
[f] Faculty of Metallurgy and Energy Engineering, Kunming University of Science and Technology, Kunming, Yunnan 650093, China
[g] School of Mathematics, The University of Manchester, Manchester M13 9PL, UK

ARTICLE INFO

Article history:
Received 17 September 2019
Accepted 16 November 2019
Available online 30 November 2019

Keywords:
Distribution
Two phase flow
Heat transfer
Synergy
Image analysis

ABSTRACT

The marker-controlled multiple watershed segmentation are achieved to distinguish darker continuous phase in gas-liquid two-phase flow patterns efficiently. Each plot of the Betti numbers β_i is curve-fitted using a four-parameter logistics model, for characterizing mixing effects. Similarity between adjacent pixels can be quantified by the distance. The β_1 of continuous phase decreases linearly at distance ≥ 2, which can be used to determine the threshold for segmentation. Repeated tests with different pixels and methods are conducted to ensure the repeatability and effectiveness of this model. More interestingly, we find that the rapid increase of β_1 of bubbles swarm coincides with the evolution of β_1 of continuous phase, and the median of difference of β_1 between the two phases in the visible window, as a novel metric of flow regime control is obtained and correlated with average volumetric heat transfer coefficient. This dynamic image analysis method, equipped with computational homology, provides the ability of evaluation of the spatial flow structure from a two-dimensional image and can be used to control the process.

Effet synergique de l'évolution de la configuration d'écoulement des phases dispersées et continues dans le processus de transfert de chaleur par contact direct

Mots-clés: Distribution; Écoulement diphasique; Transfert de chaleur; Synergie; Analyse par imagerie

* Corresponding author.
** Co-corresponding Authors.
E-mail addresses: feiyukm@aliyun.com (Y. Fei), hbgcdx2017@163.com (Y. Yang), zhaiyuling00@126.com (Y. Zhai), jianxin.pan@manchester.ac.uk (J. Pan), wanghua65@163.com (H. Wang).
[1] Equally contributing authors.

https://doi.org/10.1016/j.ijrefrig.2019.11.020
0140-7007/

Statistical Papers
https://doi.org/10.1007/s00362-021-01236-1

REGULAR ARTICLE

Properties and generation of representative points of the exponential distribution

Long-Hao Xu[1] · Kai-Tai Fang[1,2] · Ping He[1]

Received: 27 August 2020 / Revised: 27 March 2021 / Accepted: 19 April 2021

Abstract

It is known that the exponential distribution has many nice properties. Graf and Luschgy (2000) pointed out that the mean squared error of the set of representative points of the exponential distribution is fully determined by the smallest representative point. In this paper we concern with the representative points of the exponential distribution and find a number of new interesting properties. A new algorithm is proposed to effectively generate representative points of the exponential distribution. In addition, the performance of representative points of the exponential distribution is evaluated.

Keywords Discrete approximation · Exponential distribution · Mean squared error · Principal points · Representative points

1 Introduction

The problem of selecting a given number of representative points (RPs for short) which retain as much information of the population as possible arises in many situations. It can also be considered as a problem of approximating a continuous distribution by a discrete distribution. Let X be a univariate random variable with probability density function (p.d.f.) $p(x)$ and cumulative distribution function (c.d.f.) $F(x)$. We want to find a discrete random variable Y shown as below to represent the continuous random

✉ Ping He
heping@uic.edu.cn

Long-Hao Xu
893405062@qq.com

Kai-Tai Fang
ktfang@uic.edu.cn

1 Division of Science and Technology, BNU-HKBU United International College, Zhuhai 519085, China

2 The Key Lab of Random Complex Structures and Data Analysis, The Chinese Academy of Sciences, Beijing 100864, China

Published online: 10 May 2021

Statistical Papers (2021) 62:267–307
https://doi.org/10.1007/s00362-019-01089-9

REGULAR ARTICLE

New recommended designs for screening either qualitative or quantitative factors

A. M. Elsawah[1,2] · Kai-Tai Fang[2,3] · Xiao Ke[4,5]

Received: 29 March 2018 / Revised: 16 November 2018 / Published online: 9 February 2019

Abstract

By the affine resolvable design theory, there are 68 non-isomorphic classes of symmetric orthogonal designs involving 13 factors with 3 levels and 27 runs. This paper gives a comprehensive study of all these 68 non-isomorphic classes from the viewpoint of the uniformity criteria, generalized word-length pattern and Hamming distance pattern, which provides some interesting projection and level permutation behaviors of these classes. Selecting best projected level permuted subdesigns with $3 \leq k \leq 13$ factors from all these 68 non-isomorphic classes is discussed via these three criteria with catalogues of best values. New recommended uniform minimum aberration and minimum Hamming distance designs are given for investigating either qualitative or quantitative $4 \leq k \leq 13$ factors, which perform better than the existing recommended designs in literature and the existing uniform designs. A new efficient technique for detecting non-isomorphic designs is given via these three criteria. By using this new approach, in all projections into $1 \leq k \leq 13$ factors we classify each class from these 68 classes to non-isomorphic subclasses and give the number of isomorphic designs in each subclass. Close relationships among these three criteria and lower bounds of the average uniformity criteria are given as benchmarks for selecting best designs.

Keywords Design isomorphism · Orthogonal designs · Level permutation · Projection · Generalized word-length pattern · Hamming distance pattern · Uniformity criteria

Mathematics Subject Classification 62K05 · 62K15 · 94B05

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00362-019-01089-9) contains supplementary material, which is available to authorized users.

✉ Kai-Tai Fang
ktfang@uic.edu.hk

Extended author information available on the last page of the article

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
https://doi.org/10.1080/03610918.2021.1938121

A novel algorithm for generating minimum energy points from identically charged particles in 1D, 2D and 3D unit hypercubes

A. M. Elsawah[a,b], Li Meng Hua[a,c], and Kai-Tai Fang[a,d]

[a]Division of Science and Technology, Beijing Normal University–Hong Kong Baptist University United International College, Zhuhai, China; [b]Department of Mathematics, Faculty of Science, Zagazig University, Zagazig, Egypt; [c]Chinasoft International Company, Shenzhen, China; [d]The Key Lab of Random Complex Structures and Data Analysis, The Chinese Academy of Sciences, Beijing, China

ABSTRACT

Generating minimum energy points (MEPs) is an optimal solution of many real-world problems, such as the selection of best locations for hospitals inside a city that reduce the overcrowding and competition and avoid the less-populated regions. The key idea is considering these locations as charged particles with the same sign (i.e., repel each other) inside a box and distribute these points by minimizing the total electric potential energy (TEPE) among them. The practice demonstrated that most of the existing techniques for generating MEPs are complex, especially for non-mathematicians. Therefore, the greedy algorithm (GreA) is the classical widely used algorithm for its simplicity even though a satisfactory result is not guaranteed. This paper gives a novel algorithm for generating MEPs from identically charged particles in 1D, 2D and 3D unit hypercubes. The results show that the new algorithm distributes the points far away from each other to reduce the TEPE of the generated MEPs more effectively than the GreA. The new algorithm is a significant improvement of the GreA to overcome its unsatisfactory results. Therefore, the new algorithm in its current form or after some improvements is highly recommended to be used instead of the GreA for many different applications.

ARTICLE HISTORY
Received 4 June 2020
Accepted 29 May 2021

KEYWORDS
DEoptim; Electric field lines; Electric potential energy; Equipotential circles; GenSA; Greedy algorithm; Minimum energy points; PSO; Representative points

MATHEMATICAL SUBJECT CLASSIFICATION
1.00940; 1.00950; 1.00990

1. Introduction

Consider the following real-life problem of selecting the best locations to open gas stations in a new city. How to find the best locations that will avoid the less-populated regions and minimize the competition among these gas stations? For solving this significant real-life problem, consider the new city (experimental region) as a box and the locations of the gas stations are points inside this box. To effectively distribute these points inside this box, consider that these points are charged particles with the same sign and the charge represents the experimental response (inversely proportional to the population density). Therefore, these points will repel each other and try to be as far away as possible from each other. The distributed points will take positions inside the box so as to minimize the total potential energy in electrodynamics (Thomson's theorem, see, e.g., Zhou 1999). These points are called minimum energy points (MEPs). MEPs are

CONTACT A. M. Elsawah a_elsawah85@yahoo.com or amelsawah@uic.edu.cn or a.elsawah@zu.edu.eg Division of Science and Technology, Beijing Normal University–Hong Kong Baptist University United International College, Zhuhai 519085, China.

Supplemental data for this article is available online at https://doi.org/10.1080/03610918.2021.1938121

J Syst Sci Complex (2021) 34: 402–439

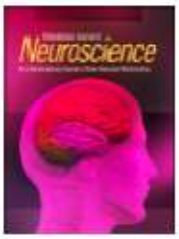
Cross-Entropy Loss for Recommending Efficient Fold-Over Technique*

WENG Lin-Chen · ELSAWAH A M · FANG Kai-Tai

DOI: 10.1007/s11424-020-9267-9
Received: 23 September 2019 / Revised: 3 June 2020

Abstract Due to the limited resources and budgets in many real-life projects, it is unaffordable to use full factorial experimental designs and thus fractional factorial (FF) designs are used instead. The aliasing of factorial effects is the price we pay for using FF designs and thus some significant effects cannot be estimated. Therefore, some additional observations (runs) are needed to break the linages among the factorial effects. Folding over the initial FF designs is one of the significant approaches for selecting the additional runs. This paper gives an in-depth look at fold-over techniques via the following four significant contributions. The first contribution is on discussing the adjusted switching levels fold-over technique to overcome the limitation of the classical one. The second contribution is on presenting a comparison study among the widely used fold-over techniques to help experimenters to recommend a suitable fold-over technique for their experiments by answering the following two fundamental questions: Do these techniques dramatically lessen the confounding of the initial designs, and do the resulting combined designs (combining initial design with its fold-over) via these techniques have considerable differences from the optimality point of view considering the markedly different searching domains in each technique? The optimality criteria are the aberration, confounding, Hamming distance and uniformity. Many of these criteria are given in sequences (patterns) form, which are inconvenient and costly to represent and compare, especially when the designs have many factors. The third innovation is on developing a new criterion (dictionary cross-entropy loss) to simplify the existing criteria from

WENG Lin-Chen
Division of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai 519085, *China*.
ELSAWAH A M (Corresponding author)
Division of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai 519085, *China; Department of Mathematics, Faculty of Science, Zagazig University, Zagazig* 44519, *Egypt*. Email: a_elsawah85@yahoo.com; amelsawah@uic.edu.cn; a.elsawah@fsz.edu.eg.
FANG Kai-Tai
Division of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai 519085, *China; The Key Lab of Random Complex Structures and Data Analysis, Chinese Academy of Sciences, Beijing* 100190, *China*.
*This research was supported by the Beijing Normal University-Hong Kong Baptist University United International College under Grant Nos. R201810, R201912 and R202010, and the Zhuhai Premier Discipline Grant.
° *This paper was recommended for publication by Editor ZHU Liping.*



International Journal of Neuroscience

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ines20

Developmental characteristics of visual evoked potentials to different stimulation in normal children

Ting-Ting Jiang, Li Wang, Hong-Liang Chen, Yu Deng, Xiao-Ling Peng & Yue Hu

To cite this article: Ting-Ting Jiang, Li Wang, Hong-Liang Chen, Yu Deng, Xiao-Ling Peng & Yue Hu (2021): Developmental characteristics of visual evoked potentials to different stimulation in normal children, International Journal of Neuroscience, DOI: 10.1080/00207454.2021.1912039

To link to this article: https://doi.org/10.1080/00207454.2021.1912039

Published online: 07 Jul 2021.

Journal of Statistical Planning and Inference 213 (2021) 262–281

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

Journal homepage: www.elsevier.com/locate/jspi

A scalable surrogate L_0 sparse regression method for generalized linear models with applications to large scale data

Ning Li [a], Xiaoling Peng [b], Eric Kawaguchi [c], Marc A. Suchard [d], Gang Li [e,*]

[a] Department of Medicine Statistics Core, University of California – Los Angeles, CA, USA
[b] Division of Science and Technology, Beijing Normal University - Hong Kong Baptist University United International College, Zhuhai, China
[c] Division of Biostatistics and Epidemiology, Keck School of Medicine, University of Southern California, CA, USA
[d] Departments of Computational Medicine, Biostatistics and Human Genetics, University of California – Los Angeles, CA, USA
[e] Department of Biostatistics, University of California – Los Angeles, CA, USA

ARTICLE INFO

Article history:
Received 6 February 2020
Received in revised form 1 December 2020
Accepted 4 December 2020
Available online 17 December 2020

Keywords:
Generalized linear models
High dimensional massive sample size data
L_0 penalty
Ridge regression
Variable selection

ABSTRACT

This paper rigorously studies large sample properties of a surrogate L_0 penalization method via iteratively performing reweighted L_2 penalized regressions for generalized linear models and develop a scalable implementation of the method for sparse high dimensional massive sample size (sHDMSS) data. We show that for generalized linear models, the limit of the algorithm, referred to as the broken adaptive ridge (BAR) estimator, is consistent for variable selection, enjoys an oracle property for parameter estimation, and possesses a grouping property for highly correlated covariates. We further demonstrate that by taking advantage of an existing efficient implementation of massive L_2-penalized generalized linear models, the proposed BAR method can be conveniently implemented for sHDMSS data. An illustration is given using a large sHDMSS data from the Truven MarketScan Medicare (MDCR) database to investigate the safety of dabigatran versus warfarin for treatment of nonvalvular atrial fibrillation in elder patients.

1. Introduction

L_0-penalized regression has been popularly used in the classical variable selection methods through the well-known information criteria such as Mallow's C_p (Mallows, 1973), Akaike's information criterion (AIC) (Akaike, 1974), the Bayesian information criterion (BIC) (Schwarz et al., 1978; Chen and Chen, 2008), and risk inflation criteria (RIC) (Foster and George, 1994). It directly penalizes the cardinality of a model and has been shown to possess some optimal properties for variable selection and parameter estimation (Shen et al., 2012). On the other hand, L_0-penalization is computationally NP-hard and thus not scalable to high dimensional covariates. It can also be unstable for variable selection (Breiman et al., 1996). The broken adaptive ridge (BAR) method has been recently studied as a scalable surrogate to L_0 penalization for simultaneous variable selection and parameter estimation (Dai et al., 2018a,b; Frommlet and Nuel, 2016; Liu and Li, 2016). Defined as the limit of an iteratively reweighted L_2 penalization algorithm, the BAR estimator has been shown to enjoy the best of both L_0 and L_2 penalizations (Dai et al., 2018a,b) while avoiding their shortcomings. For instance, it is easily scalable to high dimensional covariates and has been shown to be consistent for variable selection, oracle for parameter estimation,

* Corresponding author.
E-mail address: vli@ucla.edu (G. Li).

https://doi.org/10.1016/j.jspi.2020.12.001
0378-3758/

Expert Systems With Applications 184 (2021) 115545

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A hybrid feedforward neural network algorithm for detecting outliers in non-stationary multivariate time series

Gajendra K. Vishwakarma [a,*], Chinmoy Paul [a,b,*], A.M. Elsawah [c,d,*]

[a] Department of Mathematics & Computing, Indian Institute of Technology Dhanbad, Dhanbad 826004, India
[b] Department of Statistics, Pandit Deendayal Upadhyaya Adarsha Mahavidyalaya, Eraligool, Karimganj 788723, India
[c] Division of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai 519085, China
[d] Department of Mathematics, Faculty of Science, Zagazig University, Zagazig 44519, Egypt

ARTICLE INFO

Keywords:
Feedforward neural network
Mahalanobis distance
Non-stationary time series
Outlier detection
Robust estimate

Mathematics Subject Classification:
62M10
68W50
91B84

ABSTRACT

To understand the behavior of complex phenomena, data collection and data analysis are the two basic key issues in this process. The most significant hard problem experimenters may face is the optimality selection of the dataset which provides valuable information about the behavior of the phenomena under the experimentation. An experiment with an optimal dataset allows more significant parameters to be estimated with minimum variance and without bias. Unfortunately, the collected datasets of many real-life experiments are not optimal due to the existence of outliers. An outlier is an observation that deviates significantly from other experimental data points arouse suspicions that it was generated by a different mechanism. The presence of even a few outliers leads to misspecification of model, biased estimation of parameters, and poor forecasts. Therefore, removing the outliers from the collected datasets is the critical and significant step before analyzing the data. This paper gives a hybrid feedforward neural network algorithm for detecting outliers as single points as well as small and large clusters in non-stationary multivariate time series using robust measures of location and dispersion matrix. From various perspectives, the performance of the proposed algorithm is compared with the existing algorithms under different scenarios using simulated datasets.

1. Introduction

In order to understand the behavior of complex phenomena in industrial applications and scientific investigations, data collection and data analysis are the basic two issues in this process. There is no doubt that collecting an optimal experimental dataset for the experiment under the investigation is the cornerstone and the most important and significant issue in this process for the following two simple reasons: (i) there are many effective data analysis software that can be used to analyze the collected experimental dataset, and (ii) an optimal experimental dataset for a given experiment able to capture maximum valuable (accurate) information about its behavior and thus more significant parameters can be estimated without bias and with minimum variance, whereas a non-optimal experimental dataset cannot produce accurate knowledge about the behavior of the experiment, even if effective softwares are used to extract possible information from it (cf. Elsawah, 2021a). Therefore, the optimality collection of experimental datasets, which provide valuable information about the behavior of the phenomena under the experimentation, is the important step in this process (cf. Elsawah, 2021b). From a practical point of view, selecting such optimal experimental datasets is the most challenging and difficult part investigators may face (cf. Elsawah, 2021c).

Unfortunately, the collected experimental datasets of many real-life experiments may have outliers. The presence of even a few outliers may lead to misspecification of model, biased estimation of parameters, and poor forecasts. An outlier is an observation that deviates significantly from other experimental data points arouse suspicions that it was generated by a different mechanism (cf. Hawkins, 1980). Simply, an outlier can be thought of as an experimental observation that does not follow the expected behavior. An outlier arises due to human error, changes in system behavior, variability in the measurement, fraudulent behavior, or/and instrument error. For example: (i) a physical instrument for taking experimental observations may have suffered a temporary malfunction, (ii) there may have been an error in experimental data transmission or/and experimental data transcription, or/and (iii) a selected experimental dataset may have been contaminated with data

* Corresponding authors.
E-mail addresses: vishwagk@rediffmail.com (G.K. Vishwakarma), chinmoy.grc@gmail.com (C. Paul), amelsawah@uic.edu.cn (A.M. Elsawah).

https://doi.org/10.1016/j.eswa.2021.115545
Received 10 March 2020; Received in revised form 21 June 2021; Accepted 1 July 2021
Available online 7 July 2021
0957-4174/

Journal of Computational and Applied Mathematics 384 (2021) 113164

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

Journal homepage: www.elsevier.com/locate/cam

An appealing technique for designing optimal large experiments with three-level factors

A.M. Elsawah *

Division of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai 519085, China
Department of Mathematics, Faculty of Science, Zagazig University, Zagazig 44519, Egypt

ARTICLE INFO

Article history:
Received 1 June 2020
Received in revised form 11 August 2020

MSC:
62K05
62K15

Keywords:
Multiple tripling
Hamming distance
Aberration
Power moments
Orthogonality
Uniformity

ABSTRACT

Experimental design is arguably the most commonly used and effective methodology in scientific investigations and industrial applications. Real-world experiments may have hundreds or even thousands of input variables (factors) and thus a large number of observations (experimental runs) is needed to gain a better understanding of the phenomena under the investigation and estimate the most important parameters without bias and with minimum variance. Constructing optimal designs for these large experiments is a significant NP-hard problem investigators may face. This paper gives a new simple efficient technique, called multiple tripling technique, for constructing optimal (in view of distance, aberration, power moments, orthogonality, uniformity) designs for large experiments with three-level factors by multiple tripling of small and simple three-level initial designs. Some logical questions are now arising, such as how to effectively select initial designs to get optimal resulting multiple triple designs, how to measure the optimality of a resulting multiple triple design relative to all the possible designs with the same size, and what is the efficiency of the multiple tripling technique relative to the existing widely used techniques for constructing large three-level designs? Through theoretical and computational justifications, this paper tries to answer these significant questions. Without computational time (no computer search), the multiple tripling technique is used to construct new recommended optimal designs which are better than the existing recommended designs or cannot be constructed by the existing techniques due to their large sizes.

1. Introduction

Design of experiments is arguably the most commonly used and effective tool for understanding the behavior of complex phenomena in industrial and scientific applications by investigating the effect of input variables (factors) on the response variables (outputs). The most significant hard problem experimenters may face is the optimality selection of the experimental runs (experimental designs) which provide useful information about the behavior of the phenomena under the experimentation. An experiment with an optimal design allows more parameters to be estimated with minimum variance and without bias, while an experiment with a non-optimal design needs a greater number of experimental runs to estimate the parameters with the same accuracy as an optimal design. From a practical point of view, constructing

* Correspondence to: Division of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai 519085, China.
E-mail addresses: amelsawah@uic.edu.cn, a_elsawah85@yahoo.com, a.elsawah@fsc.edu.eg.

https://doi.org/10.1016/j.cam.2020.113164
0377-0427/

Statistical Papers
https://doi.org/10.1007/s00362-020-01221-0

REGULAR ARTICLE

Multiple doubling: a simple effective construction technique for optimal two-level experimental designs

A. M. Elsawah[1,2]

Received: 9 January 2020 / Revised: 15 December 2020 / Accepted: 30 December 2020

Abstract

Design of experiment is an efficient statistical methodology of establishing which input variables are important (have significant effects) in an experiment (process) and the conditions under which these inputs should work to optimize the outputs of that process. Two-level designs are widely used in high-tech industries and manufacturing for productivity and quality improvement experiments. The construction of (nearly) optimal two-level designs for real-life experiments with large number of input variables can be quite challenging. The practice demonstrated that the existing techniques are complex, highly time-consuming, produce limited types of designs, and likely to fail in large experiments (i.e., optimal results are not expected). To overcome these significant problems, this article gives a simple and effective technique for constructing large two-level designs with good statistical properties. To meet practical needs in different fields, the statistical properties of the generated designs by the new technique are investigated from four basic perspectives: minimizing the similarity among the experimental runs, minimizing the aliasing among the input variables, maximizing the resolution, and filling the experimental domain as uniformly as possible. New recommended saturated orthogonal main effect plans and uniform orthogonal arrays of strength three with thousands or even millions of runs and factors are generated via the new technique without recourse to optimization software.

Keywords Multiple doubling · Orthogonal arrays · Minimum aberration designs · Minimum moment aberration designs · Minimum probability Hamming distance designs · Uniform designs

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00362-020-01221-0.

✉ A. M. Elsawah
a_elsawah85@yahoo.com; amelsawah@uic.edu.cn; a.elsawah@zu.edu.eg

1 Division of Science and Technology, Beijing Normal University–Hong Kong Baptist University United International College, Zhuhai 519085, China

2 Department of Mathematics, Faculty of Science, Zagazig University, Zagazig 44519, Egypt

Published online: 01 February 2021

Communications in Mathematics and Statistics
https://doi.org/10.1007/s40304-021-00241-y

Designing Optimal Large Four-Level Experiments: A New Technique Without Recourse to Optimization Softwares

A. M. Elsawah[1,2]

Received: 13 July 2020 / Revised: 27 October 2020 / Accepted: 24 March 2021

Abstract

Experimental design is an effective statistical tool that is extensively applied in modern industry, engineering, and science. It is proved that experimental design is a powerful and efficient means to screen the relationships between input factors and their responses and to distinguish significant and unimportant factor effects. In many practical situations, experimenters are faced with large experiments having four-level factors. Even though there are several techniques provided to design such experiments, the challenge faced by the experimenters is still daunting. The practice has demonstrated that the existing techniques are highly time-consuming optimization procedures, satisfactory outcomes are not guaranteed, and non-mathematicians face a significant challenge in dealing with them. A new technique that can overcome these defects of the existing techniques is presented in this paper. The results demonstrated that the proposed technique outperformed the current techniques in terms of construction simplicity, computational efficiency and achieving satisfactory results capability. For non-mathematician experimenters, the new technique is much easier and simpler than the current techniques, as it allows them to design optimal large experiments without the recourse to optimization softwares. The optimality is discussed from four basic perspectives: maximizing the dissimilarity among experimental runs, maximizing the number of independent factors, minimizing the confounding among factors, and filling the experimental domain uniformly with as few gaps as possible.

Keywords Multiple quadrupling technique · TA algorithm · Augmented design technique · Level permutation technique · Confounding · Hamming distance · Space-filling

✉ A. M. Elsawah
a_elsawah85@yahoo.com; amelsawah@uic.edu.cn; a.elsawah@zu.edu.eg

1 Division of Science and Technology, United International College, Beijing Normal University-Hong Kong Baptist University, Zhuhai 519085, China

2 Department of Mathematics, Faculty of Science, Zagazig University, Zagazig 44519, Egypt

Journal of Agriculture and Food Research 5 (2021) 100163

Contents lists available at ScienceDirect

Journal of Agriculture and Food Research

journal homepage: www.sciencedirect.com/journal/journal-of-agriculture-and-food-research

Modeling and optimization of the effect of abiotic stressors on the productivity of the biomass, chlorophyll and lutein in microalgae *Chlorella pyrenoidosa*

Barathan Balaji Prasath [a], A.M. Elsawah [a,b,*], Zhou Liyuan [a], Karen Poon [a]

[a] Division of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai, 519085, China
[b] Department of Mathematics, Faculty of Science, Zagazig University, Zagazig, 44519, Egypt

ARTICLE INFO

Keywords:
Modeling
Optimization
Microalgae
Biomass
Chlorophyll
Lutein
Chlorella pyrenoidosa

ABSTRACT

Microalgae have been exploited for food, biofuels, animal feed and pharmaceutical products over the last few decades. The microalgae biotechnology has grown and diversified significantly, despite these developments the number of commercially available products are still limited and thus there is a significant need to increase the production. Modeling and optimization are used to simulate the behavior of the studied problems and provide a basis for selecting optimum input settings that optimize the outputs. This paper investigates the effect of the beijerinck solution (BS), phosphate solution (PS) and hunter trace (HT) with different four levels (values) on the biomass productivity (BP), lutein productivity (LP) and total chlorophyll (TC) in microalgae *Chlorella pyrenoidosa*. A closer look at the effect of each level of the input factors on the output factors is given. Various modeling techniques, such as the spline model and Gaussian Kriging model, are investigated to recommend the optimal models for describing the relationships between the input factors and each output factor. The validation and efficiency of the recommended models are studied from various points of view. Based on the recommended models, the predicted values of the BP, LP and TC are given for all the possible level-combinations (conditions) of the BS, PS and HT and the best (optimal) level-combinations of the BS, PS and HT that maximize the BP, LP and TC are recommended.

1. Introduction

Microalgae have been exploited for food, biofuels, animal feed and pharmaceutical products over the last few decades [1,2]. The microalgae biotechnology has grown and diversified significantly, despite these developments the number of commercially available products are still limited. There is still a significant demand in the global market margin to improve the economic yield for large-scale lutein production [3], since the marigold based lutein production is still much lower than the growth rate and biomass yield and it serves as the dominant source nowadays [4]. However, microalgae are often too low for the commercial viability of lutein production. Various cultivation strategies to increase lutein content and lutein productivity in microalgae species have been developed [5,6]. However, these efforts still cannot meet requirements for commercial production.

Moreover, some microalgae deal with the rising irradiance, pH, temperatures, and nitrate comprise of photosynthetic pigments content increased [7,8]. However, the correlation under different conditions of phosphate and trace metals is still unclear. In order to improve the commercial viability of lutein and its related products, it is necessary to improve the culture efficiency of microalgae, emphasizing the cell growth rate and the content of microalgae lutein. One possibility of increasing the lutein content in algae is exposing them to environmental stress (called lutein triggering) before microalgae harvesting and pigment extraction. Lutein triggering may involve algae nutrient deprivation (e.g., starvation of nitrogen), altering the light exposure, heat shock, osmotic pressure, chemical stress, or radiation [9]. Recently, some microalgae culture strategies have been used to enhance biomass and value target metabolites yield and the best of our knowledge; however, minimal efforts have yet been made to evaluate the interactions between different nutrients culture mediums.

This paper investigates the optimization of the culture condition for

* Corresponding author. Division of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai, 519085, China.
E-mail addresses: a.elsawah85@yahoo.com, amelsawah@uic.edu.cn, a.elsawah@zu.edu.eg (A.M. Elsawah).

https://doi.org/10.1016/j.jafr.2021.100163
Received 16 January 2021; Received in revised form 25 March 2021; Accepted 16 May 2021
Available online 26 May 2021

An Optimum Signal Detection Approach to the Joint ML Estimation of Timing Offset, Carrier Frequency and Phase Offset for Coherent Optical OFDM

Xinwei Du, *Member, IEEE*, Tianyu Song, *Member, IEEE*, Yan Li, Ming-Wei Wu, *Member, IEEE*, and Pooi-Yuen Kam, *Member, IEEE*

***Abstract*—Coherent orthogonal frequency-division multiplexing (OFDM) is one of the prime digital modulation techniques for present and future generations of wireless and optical communications. Accurate synchronization is the main obstacle to the implementation of a reliable OFDM receiver. We propose here a joint maximum likelihood (ML) estimator for the timing offset (TO), carrier frequency offset (CFO), and carrier phase offset (CPO) for coherent optical OFDM (CO-OFDM) motivated by the theory of ML signal detection. Our approach starts conceptually with the idea of first computing L replicas of the original OFDM spectrum, where L is a power of two and is assumed sufficiently large. This is done by padding $(L-1)N$ zeros to the end of the N received noisy OFDM samples, where N is the number of OFDM subcarriers. By computing the LN-point discrete Fourier transform (DFT) of these LN time samples, we get L replicas of the DFT of the original OFDM spectrum, where each replica corresponds to the DFT for one hypothesized value of the CFO and the set of possible CFO values is $\{l/L\}_{l=0}^{L-1}$. We select the most probable replica by choosing the one that is at the minimum Euclidean distance from the original OFDM spectrum, which leads to a matched-filtering (MF) operation in the frequency domain in either a blind or a data-aided manner. Building on this MF concept, we then develop a joint ML estimator of the TO and CPO for each hypothesized value of the CFO. The novelty here is that the TO and the CPO are estimated efficiently as the frequency and phase of a complex sinusoid observed in noise, via either a time-domain or a frequency-domain approach. The resulting joint CFO, CPO, and TO estimator is simpler than existing estimators, both conceptually and in implementation. A much simpler sequential approach in which we first decide on the CFO and then perform a joint TO and CPO estimation is also proposed. The performance loss of this sequential approach compared to the optimum joint approach is small at high signal-to-noise ratio (SNR). We obtain the performance of all our estimators via simulations, and show that they perform better when compared with the existing well-known estimators. Finally, we derive the Cramér–Rao lower bounds (CRLB) on the performance of our estimators, and show via simulations that our estimators for high SNR attain these performance lower bounds.**

***Index Terms*—Carrier frequency offset, carrier phase offset, matched filter, maximum likelihood estimation, orthogonal frequency-division multiplexing (OFDM), timing offset.**

I. Introduction

COHERENT optical orthogonal frequency-division multiplexing (CO-OFDM) has attracted much research attention due to its high spectral efficiency and its robustness to the chromatic dispersion (CD) and polarization mode dispersion (PMD) [1]. In a CO-OFDM system, a serial high-rate data stream is split into multiple parallel data streams with lower speed, each of which is modulated onto orthogonal subcarriers. The subcarriers are orthogonally overlapping with one another, leading to high spectral efficiency. However, synchronization errors in timing, frequency and phase can introduce both inter-symbol interference (ISI) and inter-carrier interference (ICI) that lead to severe performance degradation. The timing offset (TO) is caused by the propagation delay or the sample timing offset (frequency difference / phase offset) between the transmitter and receiver, and the carrier frequency offset (CFO) is induced by the frequency offset between the transmitter and receiver laser. The existence of a carrier phase offset (CPO) due to phase noise in the transmitter and receiver oscillators also causes signal constellation rotations in the subcarriers that can lead to bit error rate (BER) performance degradation. Therefore, achieving reliable synchronization is a major challenge for CO-OFDM systems.

Many approaches for CO-OFDM synchronization are available in the literature by now. The most promising approach to

Manuscript received November 21, 2019; revised April 13, 2020, July 3, 2020, and September 14, 2020; accepted November 20, 2020. Date of publication December 4, 2020; date of current version March 16, 2021. This work was supported in part by UIC Start-up Research Fund under Grant R72021107, in part by NSFC under Grant 61571316 and Grant 61971372, and in part by HK RGC GRF under Grant 15200718. *(Corresponding author: Tianyu Song.)*

Xinwei Du is with the Division of Science and Technology, BNU-HKBU United International College, Zhuhai 519087, China (e-mail: xinweidu@uic.edu.cn).

Tianyu Song was with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583, Singapore. He is now with Huawei Technologies, Shenzhen 518129, China (e-mail: song.tianyu@u.nus.edu).

Yan Li is with the School of Electronics and Information Technology, Sun Yat-Sen University, Guangzhou 510275, China (e-mail: liyan329@mail.sysu.edu.cn).

Ming-Wei Wu is with the School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China (e-mail: mingweiwu@ieee.org).

Pooi-Yuen Kam is with the School of Science and Engineering, Chinese University of Hong Kong, Shenzhen 518172, China, and also with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583, Singapore (e-mail: pykam@cuhk.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JLT.2020.3042546.

Digital Object Identifier 10.1109/JLT.2020.3042546

Journal of Combinatorial Theory, Series A 182 (2021) 105476

Contents lists available at ScienceDirect

Journal of Combinatorial Theory, Series A

www.elsevier.com/locate/jcta

Characterising the secant lines of $Q(4,q)$, q even

Susan G. Barwick [a], Alice M.W. Hui [b,*], Wen-Ai Jackson [a], Jeroen Schillewaert [c]

[a] School of Mathematical Sciences, University of Adelaide, Adelaide, 5005, Australia
[b] Statistics Program, BNU-HKBU United International College, Zhuhai, China
[c] Department of Mathematics, University of Auckland, New Zealand

ARTICLE INFO

Article history:
Received 22 October 2019
Received in revised form 15 April 2021
Accepted 22 April 2021
Available online 11 May 2021

Keywords:
Projective geometry
Characterisation
Non-singular quadric

ABSTRACT

We show that a set $\mathcal{A}$ of lines in $\mathrm{PG}(4,q)$, q even, is the set of secant lines of a parabolic (non-singular) quadric if and only if $\mathcal{A}$ satisfies the following three conditions:

(I) every point of $\mathrm{PG}(4,q)$ lies on $0, \frac{1}{2}q^3$ or q^3 lines of $\mathcal{A}$;
(II) every plane of $\mathrm{PG}(4,q)$ contains 0, $\frac{1}{2}q(q+1)$ or q^2 lines of $\mathcal{A}$; and
(III) every hyperplane of $\mathrm{PG}(4,q)$ contains $\frac{1}{2}q^2(q^2+1)$, $\frac{1}{2}q^3(q+1)$ or $\frac{1}{2}q^2(q+1)^2$ lines of $\mathcal{A}$.

* Corresponding author.
E-mail addresses: susan.barwick@adelaide.edu.au (S.G. Barwick), alicemwhui@uic.edu.cn, huimanwa@gmail.com (A.M.W. Hui), wen.jackson@adelaide.edu.au (W.-A. Jackson), j.schillewaert@auckland.ac.nz (J. Schillewaert).

https://doi.org/10.1016/j.jcta.2021.105476
0097-3165/

SCANDINAVIAN ACTUARIAL JOURNAL
https://doi.org/10.1080/03461238.2021.1958917

Banach contraction principle, q-scale function and ultimate ruin probability under a Markov-modulated classical risk model

Zhengjun Jiang

Statistics Programme, Division of Science and Technology, BNU-HKBU United International College, Zhuhai, Guangdong, People's Republic of China

ABSTRACT

Suppose that risk reserves of an insurance company are governed by a Markov-modulated classical risk model with parameters modulated by a finite-state irreducible Markov chain. The main purpose of this paper is to calculate ultimate ruin probability that ruin time, the first time when risk reserve is negative, is finite. We apply Banach contraction principle, q-scale functions and Markov property to prove that ultimate ruin probability is the fixed point of a contraction mapping in terms of q-scale functions and that ultimate ruin probability can be calculated by constructing an iterative algorithm to approximate the fixed point. Unlike Gajek and Rudź (Insurance: Mathematics and Economics, 80 (2018), 45–53), our paper uses q-scale functions to obtain more explicit Lipschitz constant in Banach contraction principle in our case so that proofs of several Lemmas and theorems in their Appendix are unnecessary and some of their assumptions are confirmed in our case.

ARTICLE HISTORY
Received 1 February 2021
Accepted 19 July 2021

KEYWORDS
Ruin probability; Markov-modulated classical risk model; q-scale function; Markov property; Banach contraction principle

MATHEMATICS SUBJECT CLASSIFICATIONS (2020)
91G05; 60K37; 60J70

1. Introduction

Since Cramér (1930) investigated classical risk model, many authors have applied various risk models in risk theory. See e.g. Chapter II in Asmussen & Albrecher (2010) for ultimate ruin probability under a diffusion risk model, Belhaj (2010) for optimal dividend distribution under a jump-diffusion risk model with exponential jump density, Gajek & Rudź (2018) for ultimate ruin probability under a Markov-modulated classical risk model, Jiang & Pistorius (2012) for optimal dividend policy under a Markov-modulated Brownian motion with positive drifts, and Gajek & Kuciński (2017) for the investigation of the following generalization of the ultimate ruin problem: under a spectrally negative Lévy risk model, dividend payments and capital injections are employed to maximize an insurance company's value and the optimal strategy of capital management is described. This quite important research paper reconciles somehow maximizing dividend payments with minimizing ultimate ruin probability.

Particularly, Markov-modulated risk models have become increasingly popular. The motivations for the popularity of Markov-modulated risk model are time-changing parameters and analytical tractability (see e.g. Jiang 2015, 2019, Jiang & Pistorius 2012). Banach contraction principle and fixed point theorem are widely applied in the study of nonlinear differential/integral equations to solve relevant problems in risk theory. For example, Jiang & Pistorius (2012) and Jiang (2015, 2019) apply Banach contraction principle and fluctuation theory in terms of q-scale functions to study optimal

CONTACT Zhengjun Jiang zhengjunjiang@uic.edu.cn, jiang_zhengjun@163.com Statistics Programme, Division of Science and Technology, BNU-HKBU United International College, 2000 Jintong Road, Tangjiawan, Zhuhai, Guangdong 519087, People's Republic of China

symmetry

MDPI

Article

On the Admissibility of Simultaneous Bootstrap Confidence Intervals

Xin Gao [1,*], Frank Konietschke [2,3] and Qiong Li [4]

[1] Department of Mathematics and Statistics,York University, Toronto, ON M3J 1P3, Canada
[2] Berlin Institute of Health, Institute of Biometry and Clinical Epidemiology Berlin and Charité–Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, 10117 Berlin, Germany; frank.konietschke@charite.de
[3] Berlin Institute of Health (BIH), Anna-Louise-Karsch-Straße 2, 10178 Berlin, Germany
[4] BNU-HKBU United International College, Zhuhai 519087, China; qiongli@uic.edu.cn
* Correspondence: xingao@yorku.ca

Abstract: Simultaneous confidence intervals are commonly used in joint inference of multiple parameters. When the underlying joint distribution of the estimates is unknown, nonparametric methods can be applied to provide distribution-free simultaneous confidence intervals. In this note, we propose new one-sided and two-sided nonparametric simultaneous confidence intervals based on the percentile bootstrap approach. The admissibility of the proposed intervals is established. The numerical results demonstrate that the proposed confidence intervals maintain the correct coverage probability for both normal and non-normal distributions. For smoothed bootstrap estimates, we extend Efron's (2014) nonparametric delta method to construct nonparametric simultaneous confidence intervals. The methods are applied to construct simultaneous confidence intervals for LASSO regression estimates.

Keywords: simultaneous confidence interval; bootstrap; admissibility

Citation: Gao, X.; Konietschke, F.; Li, Q. On the Admissibility of Simultaneous Bootstrap Confidence Intervals. *Symmetry* **2021**, *13*, 1212. https://doi.org/10.3390/sym13071212

Academic Editors: John H. Graham and Alessandro Sarracino

Received: 6 May 2021
Accepted: 29 June 2021
Published: 6 July 2021

1. Introduction

In statistical sciences, the computation of simultaneous confidence intervals (SCIs) for parameters of interest plays a major role [1]. For instance, multiple testing problems, regression analysis, and multivariate inference are well known application areas. When the finite sampling distribution of the estimator (here, a vector) under the alternative is known, exact SCIs can be computed. However, that is often not the case, and when the underlying distribution of the estimator is difficult to obtain analytically, bootstrap methods (using drawing with replacement from the data) can be used to approximate its sampling distribution (see, e.g., [2,3], among others). The authors of [4] provide an algorithm to construct SCIs for multivariate estimates, which calculates the number of bootstrap samples that fall outside a confidence region. To achieve the pre-specified level of $(1-\alpha)\times 100\%$, one can trial different values of the lower and upper bounds of the intervals until attaining the target level. In [5], a more efficient way of computing $(1-\alpha)\times 100\%$ upper SCIs was developed by sorting the bootstrap realizations of the multivariate estimates by the maximum rank across all the coordinates. Then, the $(1-\alpha)\times 100\%$ percentile of the maximum rank is determined. The limit of the confidence interval for each coordinate is the value of the estimate sharing the same rank as the $(1-\alpha)$ percentile of the maximum rank. By symmetry, the $(1-\alpha)\times 100\%$ lower SCIs can also be constructed.

In this note, we improve their method by sharpening the limits of the SCIs for each coordinate. It is observed that the value of the estimate sharing the same rank as the $(1-\alpha)\times 100\%$ percentile of the maximum rank is indeed higher than the estimates in the selected $(1-\alpha)\times 100\%$ samples, but the threshold for each coordinate is too conservative. We can tighten them rather than using the same maximum rank cutoff values (see details

Received: 24 October 2019 | Revised: 7 April 2021 | Accepted: 24 May 2021

DOI: 10.1002/sam.11530

WILEY

RESEARCH ARTICLE

Penalized composite likelihood for colored graphical Gaussian models

Qiong Li[1] | Xiaoying Sun[2] | Nanwei Wang[3] | Xin Gao[2]

[1]Division of Science and Technology, BNU-HKBU United International College, Zhuhai, Guangdong China

[2]Department of Mathematics and Statistics, York University, Toronto, Ontario Canada

[3]Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Ontario Canada

Correspondence
Xin Gao, Department of Mathematics and Statistics, York University, Toronto, ON, Canada.
Email: xingao@mathstat.yorku.ca

Funding information
Guangdong Basic and Applied Basic Research Foundation, Grant/Award Number: 2019A1515011238; Natural Sciences and Engineering Research Council of Canada

Abstract

This article proposes a penalized composite likelihood method for model selection in colored graphical Gaussian models. The method provides a sparse and symmetry-constrained estimator of the precision matrix and thus conducts model selection and precision matrix estimation simultaneously. In particular, the method uses penalty terms to constrain the elements of the precision matrix, which enables us to transform the model selection problem into a constrained optimization problem. Further, computer experiments are conducted to illustrate the performance of the proposed new methodology. It is shown that the proposed method performs well in both the selection of nonzero elements in the precision matrix and the identification of symmetry structures in graphical models. The feasibility and potential clinical application of the proposed method are demonstrated on a microarray gene expression dataset.

KEYWORDS

l_1 penalty, model selection, nonconvex minimization, precision matrix estimation

1 | INTRODUCTION

In recent years, undirected graphical models [21] have been playing an important role in statistical inference, which are widely employed to analyze and visualize conditional dependence relationships among variables. In a graphical model of a multivariate distribution, vertices represent random variables and edges encode conditional dependencies among the vertices. Precision matrix estimation and model selection in graphical Gaussian models is equivalent to estimating parameters and identifying zeros in the precision matrix. The increasing availability of large data in different disciplines makes graphical models an excellent tool to capture the conditional structure between component variables. Graphical Gaussian models have been successfully applied in a number of fields such as genetic networks [8], biological networks [29], and financial networks [10].

Colored graphical models are developed by adding symmetry restrictions to the precision matrix of graphical models [19]. As constrainted graphical Gaussian models, colored graphical models can be represented by coloring the associated underlying graphs. The colored edges and vertices are associated with the restricted equal entries in the precision matrix. Colored graphical models are especially useful in revealing the structures of gene regulatory networks [15]. The equalities of the entries in the precision matrix provide a convenient tool to demonstrate the commonalities between the genes which can be represented by the same color vertices in the colored models.

Received: 7 April 2020 | Revised: 19 April 2021 | Accepted: 5 June 2021

DOI: 10.1111/sjos.12545

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

Multistate analysis of multitype recurrent event and failure time data with event feedbacks in biomarkers

Chuoxin Ma | **Jianxin Pan**

Department of Mathematics, The University of Manchester, Manchester, UK

Correspondence
Jianxin Pan, Department of Mathematics, The University of Manchester, Oxford Road, Manchester M13 9PL, UK.
Email: jianxin.pan@manchester.ac.uk

Abstract

In this paper we propose a class of multistate models for the analysis of multitype recurrent event and failure time data when there are past event feedbacks in longitudinal biomarkers. It can well incorporate various effects, including time-dependent and time-independent effects, of different event paths or the number of occurrences of events of different types. Asymptotic unbiased estimating equations based on polynomial splines approximation are developed. The consistency and asymptotic normality of the proposed estimators are provided. Simulation studies show that the naive estimators which either ignore the past event feedback or the measurement errors are biased. Our method has a better coverage probability of the time-varying/constant coefficients, compared to the naive methods. An application to the dataset from the Atherosclerosis Risk in Communities Study, which is also the motivating example to develop the method, is presented.

KEYWORDS

cardiovascular disease, measurement errors, multistate process, past event feedback, semiparametric coefficients

The Annals of Applied Statistics
2021, Vol. 15, No. 3, 1308–1328
https://doi.org/10.1214/21-AOAS1445

MODELING PAST EVENT FEEDBACK THROUGH BIOMARKER DYNAMICS IN THE MULTISTATE EVENT ANALYSIS FOR CARDIOVASCULAR DISEASE DATA

BY CHUOXIN MA[1,*], HONGSHENG DAI[2] AND JIANXIN PAN[1,†]

[1]*Department of Mathematics, The University of Manchester,* [*]*chuoxin.ma@manchester.ac.uk;* [†]*Jianxin.Pan@manchester.ac.uk*
[2]*Department of Mathematical Sciences, University of Essex, hdaia@essex.ac.uk*

In cardiovascular studies we often observe ordered multiple events along disease progression which are, essentially, a series of recurrent events and terminal events with competing risk structure. One of the main interests is to explore the event specific association with the dynamics of longitudinal biomarkers. A new statistical challenge arises when the biomarkers carry information from the past event history, providing feedbacks for the occurrences of future events and, particularly, when these biomarkers are only intermittently observed with measurement errors. In this paper we propose a novel modeling framework where the recurrent events and terminal events are modeled as multistate processes and the longitudinal covariates that account for event feedbacks are described by random effects models. Considering the nature of long-term observation in cardiac studies, flexible models with semiparametric coefficients are adopted. To improve computation efficiency, we develop an one-step estimator of the regression coefficients and derive their asymptotic variances for the computation of the confidence intervals, based on the proposed asymptotically unbiased estimating equation. Simulation studies show that the naive estimators, which either ignore the past event feedbacks or the measurement errors, are biased. Our method achieves better coverage probability, compared to the naive methods. The model is motivated and applied to a dataset from the Atherosclerosis Risk in Communities Study.

1. Introduction. Identifying risk factors associated with the course of cardiovascular disease (CVD) is of great medical interest. A main feature of the practice presents difficulties in the estimation of the association between risk factors and CVD events. That is, a constellation of events of different types are observed from the same subject and the occurrence of the previous events may affect the risk of the subsequent ones through the associated biomarkers. Specifically, a subject may experience, for example, recurrent myocardial infarction (MI) and may be followed by cardiovascular death. Events of different types occurred to the same subject can not be simply assumed to be independent. As pointed out by Kim et al. (2012) and Rogers et al. (2016), the risk of recurrent heart failures and myocardial infarction are associated with that of the fatal CVD events. The successive events are actually a nested series of competing risk events. After occurrence of MI, as long as the subject is still under observation, he/she can possibly encounter another MI or death in the future. The order of the recurrent events and terminal event is informative, as it reflects disease progression. It is expected that risk factors associated with different types of events would be different. The strength of association can also vary between recurrences. Furthermore, it is very likely that CVD occurrences in the past can cause a change in the profile of the associated biomarkers, and the resulting changes of the trajectory will further affect the proneness to new events in the future. When such effect accumulates with time, the feedback from past event history gets

Received October 2019; revised January 2021.
Key words and phrases. Asymptotically unbiased estimating equation, cardiovascular disease, measurement errors, multistate models, ordered multiple event, past event feedback, semiparametric coefficients.

ESC European Society of Cardiology

European Heart Journal Open (2021) **00**, 1–12
doi:10.1093/ehjopen/oeab019

ORIGINAL ARTICLE
Vascular and Cardiac Imaging

Pericoronary and periaortic adipose tissue density are associated with inflammatory disease activity in Takayasu arteritis and atherosclerosis

Christopher Wall[1], Yuan Huang[2], Elizabeth P.V. Le[1], Andrej Corović[1], Christopher P. Uy[3], Deepa Gopalan [4,5], Chuoxin Ma[6], Roido Manavaki[7], Tim D. Fryer[8], Luigi Aloj[7], Martin J. Graves[7], Enrico Tombetti[9], Ben Ariff[5], Paul Bambrough[10], Stephen P. Hoole [10], Rosemary A. Rusk[11], David R. Jayne[12], Marc R. Dweck[13], David Newby[13], Zahi A. Fayad[14], Martin R. Bennett[1], James E. Peters[15], Piotr Slomka [16], Damini Dey [17], Justin C. Mason [3], James H.F. Rudd[1], and Jason M. Tarkin [1,3,*]

[1]Division of Cardiovascular Medicine, Department of Medicine, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge, CB2 2QQ, UK [2]EPSRC Centre for Mathematical Imaging in Healthcare, University of Cambridge, Cambridge, UK; [3]Vascular Sciences, National Heart & Lung Institute, Faculty of Medicine, Imperial College London, Hammersmith Campus, DuCane Road, London, W12 0HS, UK; [4]Department of Radiology, Cambridge University Hospitals NHS Trust, Hills Road, Cambridge, CB2 2QQ, UK; [5]Department of Radiology, Imperial College Healthcare NHS Trust, Hammersmith Hospital, London, W12 0HS, UK; [6]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK; [7]Department of Radiology, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge, CB2 2QQ, UK; [8]Department of Clinical Neurosciences, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge, CB2 2QQ, UK; [9]Department of biomedical Sciences L. Sacco, Universitá degli Studi di Milano, Milan, Italy; [10]Department of Cardiology, Royal Papworth Hospital, Cambridge, UK CB2 0AY, UK; [11]Department of Cardiology, Cambridge University Hospitals NHS Trust, Hills Road, Cambridge, CB2 2QQ, UK; [12]Department of Medicine, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge, CB2 2QQ, UK; [13]Centre for Cardiovascular Science, University of Edinburgh, 47 Little France Crescent, Edinburgh, EH16 4TJ, UK; [14]BioMedical Engineering & Imaging Institute, Icahn School of Medicine at Mt Sinai, Gustave L. Levy Place, New York, NY 10029-5674, USA; [15]Centre for Inflammatory Diseases, Imperial College London, London, UK; [16]Department of Medicine, Cedars-Sinai Medical Center, 8700 Beverly Blvd, Los Angeles, CA 90048, USA and [17]Biomedical Imaging Research Institute, Cedars-Sinai Medical Center, 116 N Robertson Blvd, Los Angeles, CA, 90048, USA

Received 16 June 2021; revised 26 July 2021; editorial decision 4 August 2021; accepted 4 August 2021; online publish-ahead-of-print 6 August 2021

Handling editor: Alessia Gimelli

Aims To examine pericoronary adipose tissue (PCAT) and periaortic adipose tissue (PAAT) density on coronary computed tomography angiography for assessing arterial inflammation in Takayasu arteritis (TAK) and atherosclerosis.

Methods and results PCAT and PAAT density was measured in coronary ($n = 1016$) and aortic ($n = 108$) segments from 108 subjects [TAK + coronary artery disease (CAD), $n = 36$; TAK, $n = 18$; atherosclerotic CAD, $n = 32$; matched controls, $n = 22$]. Median PCAT and PAAT densities varied between groups (mPCAT: $P < 0.0001$; PAAT: $P = 0.0002$). PCAT density was 7.01 ± standard error of the mean (SEM) 1.78 Hounsfield Unit (HU) higher in coronary segments from TAK + CAD patients than stable CAD patients ($P = 0.0002$), and 8.20 ± SEM 2.04 HU higher in TAK patients without CAD than controls ($P = 0.0001$). mPCAT density was correlated with Indian Takayasu Clinical Activity Score ($r = 0.43$, $P = 0.001$) and C-reactive protein ($r = 0.41$, $P < 0.0001$) and was higher in active vs. inactive TAK ($P = 0.002$). mPCAT density above -74 HU had 100% sensitivity and 95% specificity for differentiating active TAK from controls [area under the curve = 0.99 (95% confidence interval 0.97–1)]. The association of PCAT density

* Corresponding author. Tel: +44 (0)1223331504, Email: jt545@cam.ac.uk

Received: 12 August 2020 | Revised: 1 June 2021 | Accepted: 30 June 2021

DOI: 10.1002/sam.11538

WILEY

RESEARCH ARTICLE

Imbalanced classification: A paradigm-based review

Yang Feng[1] | Min Zhou[2] | Xin Tong[3]

[1]Department of Biostatistics, School of Global Public Health, New York University, New York, New York USA

[2]Division of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai, China

[3]Department of Data Sciences and Operations, Marshall School of Business, University of Southern California, Los Angeles, California USA

Correspondence
Xin Tong, Department of Data Sciences and Operations, Marshall School of Business, University of Southern California, Los Angeles, CA 90089, USA.
Email: xint@marshall.usc.edu

Funding information
National Science Foundation, Grant/Award Number: DMS-1554804; US National Institutes of Health, Grant/Award Number: R01 GM 120507

Abstract

A common issue for classification in scientific research and industry is the existence of imbalanced classes. When sample sizes of different classes are imbalanced in training data, naively implementing a classification method often leads to unsatisfactory prediction results on test data. Multiple resampling techniques have been proposed to address the class imbalance issues. Yet, there is no general guidance on when to use each technique. In this article, we provide a paradigm-based review of the common resampling techniques for binary classification under imbalanced class sizes. The paradigms we consider include the classical paradigm that minimizes the overall classification error, the cost-sensitive learning paradigm that minimizes a cost-adjusted weighted type I and type II errors, and the Neyman-Pearson paradigm that minimizes the type II error subject to a type I error constraint. Under each paradigm, we investigate the combination of the resampling techniques and a few state-of-the-art classification methods. For each pair of resampling techniques and classification methods, we use simulation studies and a real dataset on credit card fraud to study the performance under different evaluation metrics. From these extensive numerical experiments, we demonstrate under each classification paradigm, the complex dynamics among resampling techniques, base classification methods, evaluation metrics, and imbalance ratios. We also summarize a few takeaway messages regarding the choices of resampling techniques and base classification methods, which could be helpful for practitioners.

KEYWORDS

binary classification, classical classification (CC) paradigm, cost-sensitive (CS) learning paradigm, imbalance ratio, imbalanced data, Neyman-Pearson (NP) paradigm, resampling methods

1 | INTRODUCTION

Classification is a widely studied type of supervised learning problem with extensive applications. A myriad of classification methods (e.g., logistic regression [LR], support vector machines [SVMs], random forest [RF], neural networks [NNs], boosting), which we refer to as the *base classification methods* in this paper, have been developed to deal with different distributions of data [32]. However, in the case where the classes are of different sizes (i.e., the imbalanced classification scenario), naively applying the existing methods could lead to undesirable results. Some prominent applications include defect detection [4], medical diagnosis [16], fraud detection [66], spam email

Yang Feng and Min Zhou contributed equally to this work.

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
https://doi.org/10.1080/03610926.2021.1970772

Limiting behavior of the gap between the largest two representative points of statistical distributions

Long-Hao Xu[a], Kai-Tai Fang[a,b], and Jianxin Pan[c]

[a]Division of Science and Technology, BNU-HKBU United International College, Zhuhai, China; [b]The Key Lab of Random Complex Structures and Data Analysis, The Chinese Academy of Sciences, Beijing, China; [c]Department of Mathematics, The University of Manchester, Manchester, UK

ABSTRACT

This paper explores the properties of the gap of representative points (RPs) in the sense of minimum mean square error for various univariate statistical distributions. We illustrate the relationship between RPs and doubly truncated mean residual life (DMRL) as well as mean residual life (MRL), which are widely used in survival analysis. The limiting behavior of the gap between the largest two RPs is discussed. In addition, an upper bound of the optimal MSE is given when the univariate random variable X has a domain on finite interval. In simulation studies, the performance of RPs for various distributions is assessed in terms of moment estimation and resampling technique. A brief discussion about the relationship between the tail of the distribution and the gap of RPs is also given.

ARTICLE HISTORY
Received 6 January 2021
Accepted 15 August 2021

KEYWORDS
Discrete approximation; mean residual life; mean square error; principal points; representative points; survival analysis

2020 MATHEMATICAL SUBJECT CLASSIFICATION
62E17

1. Introduction

It is often to request to find a discrete distribution to approximate a given continuous distribution with retaining information as much as possible. Let X be a continuous random variable with probability density function $f(x)$ and $\mathbb{E}(X^2) < \infty$. One wants to use a discrete random variable Y shown in Equation (1) to represent X, where $\mathbb{P}(Y = y_i) = p_i > 0, i = 1, ..., n$ and $y_1 < ... < y_n$.

$$\begin{array}{c|ccc} Y & y_1 & \cdots & y_n \\ \hline p & p_1 & \cdots & p_n \end{array} \tag{1}$$

The concept of representative points in the sense of minimum mean square error (MSE RPs or RPs for short) has been widely used to solve the problem (Cox 1957; Fang and He 1982; Flury 1990). The representative points are also called principal points (Flury 1990, 1993; Tarpey 1995; Tarpey, Li and Flury 1995) and quantizer (Max 1960; Lloyd 1982; Graf and Luschgy 2007) in the literature. There are many applications of RPs in signal compression (Max 1960; Lloyd 1982), cluster analysis (Anderberg 1973), statistical simulation (Fang, Zhou, and Wang 2014; Lemaire, Montes, and Pagès 2020), and numerical integration (Pagès 1998; Pagès and Printems 2003; Pagès 2015).

CONTACT Jianxin Pan Jianxin.Pan@manchester.ac.uk Department of Mathematics, The University of Manchester, Oxford Road, Manchester M13 9PL, UK.

Received: 9 November 2020 | Revised: 16 April 2021 | Accepted: 5 June 2021

DOI: 10.1002/bimj.202000336

Biometrical Journal

RESEARCH ARTICLE

Penalized joint generalized estimating equations for longitudinal binary data

Youjun Huang[1] | **Jianxin Pan**[2]

[1] Mathematical College, Sichuan University, Chengdu, P. R. China

[2] Department of Mathematics, The University of Manchester, Manchester, UK

Correspondence
Jianxin Pan, Department of Mathematics, The University of Manchester, Manchester M13 9PL, UK.
Email: jianxin.pan@manchester.ac.uk

Funding information
National Natural Science Foundation of China, Grant/Award Number: 11871357; Office of the Royal Society, Grant/Award Number: R124683

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract

In statistical research, variable selection and feature extraction are a typical issue. Variable selection in linear models has been fully developed, while it has received relatively little attention for longitudinal data. Since a longitudinal study involves within-subject correlations, the likelihood function of discrete longitudinal responses generally cannot be expressed in analytically closed form, and standard variable selection methods cannot be directly applied. As an alternative, the penalized generalized estimating equation (PGEE) is helpful but very likely results in incorrect variable selection if the working correlation matrix is misspecified. In many circumstances, the within-subject correlations are of interest and need to be modeled together with the mean. For longitudinal binary data, it becomes more challenging because the within-subject correlation coefficients have the so-called Fréchet–Hoeffding upper bound. In this paper, we proposed smoothly clipped absolute deviation (SCAD)-based and least absolute shrinkage and selection operator (LASSO)-based penalized joint generalized estimating equation (PJGEE) methods to simultaneously model the mean and correlations for longitudinal binary data, together with variable selection in the mean model. The estimated correlation coefficients satisfy the upper bound constraints. Simulation studies under different scenarios are made to assess the performance of the proposed method. Compared to existing PGEE methods that specify a working correlation matrix for longitudinal binary data, the proposed PJGEE method works much better in terms of variable selection consistency and parameter estimation accuracy. A real data set on Clinical Global Impression is analyzed for illustration.

KEYWORDS
correlation matrix, joint mean and correlation models, longitudinal binary data, penalized generalized estimating equations, variable selection

1 | INTRODUCTION

In the era of big data, information science technology has been developed rapidly, which greatly reduces the cost of data collection, storage, and transmission. Massive data are generated every day, providing opportunities and challenges for various scientific studies. These data often contain a large amount of complex covariates. Even at the time of information

Acta Mathematicae Applicatae Sinica, English Series
Vol. 37, No. 2 (2021) 214–231
https://doi.org/10.1007/s10255-021-1004-1
http://www.ApplMath.com.cn & www.SpringerLink.com

Acta Mathematicae Applicatae Sinica, English Series

Sphericity and Identity Test for High-dimensional Covariance Matrix Using Random Matrix Theory

Shou-cheng YUAN[1], Jie ZHOU[1,†], Jian-xin PAN[2], Jie-qiong SHEN[3]

[1]College of Mathematics, Sichuan University, Chengdu 610064, China ([†]E-mail: jzhou@scu.edu.cn)
[2]School of Mathematics, University of Manchester, Manchester M13 9PL, UK
[3]School of Computer and Data Engineering, Zhejiang University Ningbo Institute of Technology, Ningbo 315100, China

Abstract This paper addresses the issue of testing sphericity and identity of high-dimensional population covariance matrix when the data dimension exceeds the sample size. The central limit theorem of the first four moments of eigenvalues of sample covariance matrix is derived using random matrix theory for generally distributed populations. Further, some desirable asymptotic properties of the proposed test statistics are provided under the null hypothesis as data dimension and sample size both tend to infinity. Simulations show that the proposed tests have a greater power than existing methods for the spiked covariance model.

Keywords sphericity test; identity test; high-dimensional covariance matrix; spiked model; spectral distribution
2000 MR Subject Classification 62F03; 62F05

1 Introduction

High-dimensional statistical inference problems for covariance matrices are increasingly encountered in many applications such as image processing, stock marketing and genetics. A fundamental problem in such applications is the hypothesis test for covariance matrix when data dimension is much larger than the sample size. With the advancement in computer technology, it is feasible to analyze the high-dimensional data. However, many of the classical multivariate methods may not work properly when the dimension equals or exceeds the sample size. These procedures rely on the classical regime where the sample size tends to infinity while the dimension remains fixed.

Let $X_1, \cdots, X_n$ be independent and identically distributed (i.i.d.) p-dimensional vectors with mean zero and covariance matrix Σ_p. We focus on testing two structures for the population covariance matrix:

1) The sphericity test

$$H_{0a} : \Sigma_p = \sigma^2 \mathrm{I}_p \quad \text{vs.} \quad H_{1a} : \Sigma_p \neq \sigma^2 \mathrm{I}_p; \tag{1.1}$$

2) The identity test

$$H_{0b} : \Sigma_p = \mathrm{I}_p \quad \text{vs.} \quad H_{1b} : \Sigma_p \neq \mathrm{I}_p, \tag{1.2}$$

where σ is an unknown but finite positive constant, and I_p is the $p \times p$ identity matrix. Traditional tests for covariance matrix based on the likelihood ratio (cf. [1]) can not used when the

Manuscript received September 20, 2019. Accepted on November 25, 2020.
This paper is supported by the National Natural Science Foundation of China (Nos. 61374027, 11871357) and by the Sichuan Science and Technology Program (Nos. 2019YJ0122).
[†]Corresponding author.

Journal of Multivariate Analysis 184 (2021) 104739

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Estimation and optimal structure selection of high-dimensional Toeplitz covariance matrix

Yihe Yang [a], Jie Zhou [a], Jianxin Pan [b,*]

[a] College of Mathematics, Sichuan University, Chengdu 610065, China
[b] Department of Mathematics, The University of Manchester, Oxford Road, Manchester M13 9PL, UK

ARTICLE INFO

Article history:
Received 14 April 2020
Received in revised form 14 February 2021
Accepted 14 February 2021
Available online 9 March 2021

AMS 2010 subject classifications:
62H12
62F12
62J10

Keywords:
Covariance matrix
Entropy loss
High-dimension
Nonconvex penalty
Toeplitz covariance structure

ABSTRACT

The estimation of structured covariance matrix arises in many fields. An appropriate covariance structure not only improves the accuracy of covariance estimation but also increases the efficiency of mean parameter estimators in statistical models. In this paper, a novel statistical method is proposed, which selects the optimal Toeplitz covariance structure and estimates the covariance matrix, simultaneously. An entropy loss function with nonconvex penalty is employed as a matrix-discrepancy measure, under which the optimal selection of sparse or nearly sparse Toeplitz structure and the parameter estimators of covariance matrix are made, simultaneously, through its minimization. The cases of both low-dimensional ($p \leq n$) and high-dimensional ($p > n$) covariance matrix estimation are considered. The resulting Toeplitz structured covariance estimators are guaranteed to be positive definite and consistent. Asymptotic properties are investigated and simulation studies are conducted, showing that very high accurate Toeplitz covariance structure estimation is made. The proposed method is then applied to practical data analysis, which demonstrates its good performance in covariance estimation in practice.

1. Introduction

Covariance matrix, which characterizes the pairwise linear correlation of multiple variable, plays an important role in statistical research and application. In financial risk assessment [13], longitudinal data analysis [21], spatial and spatio-temporal data analysis [11] and signal processing [27], for instance, covariance matrix is always indispensable and fundamentally important. With the fast development of technology, datasets generated in various fields become increasingly large and may be high-dimensional, that is, the number of variables is bigger than the sample size (i.e., $p > n$), which seriously challenges standard statistical methods. For example, the sample covariance matrix, the most commonly used estimator of the population covariance matrix, degenerates as the dimension of data is close to, or larger than, the sample size. In addition, most of the available algorithms for high-dimensional covariance estimation are inevitably computationally intensive and numerically instable. In this paper, we focus on a broad class of covariance structures — Toepliz structure, which contains many commonly used covariance structures as special cases, such as order-one moving average (MA(1)), order-one autoregression (AR(1)) and order-one autoregressive and moving average (ARMA(1,1)) among many others. We then propose a novel method that selects the optimal Toeplitz covariance structure and estimates the high-dimensional covariance matrix, simultaneously.

* Corresponding author.
E-mail address: jianxin.pan@manchester.ac.uk (J. Pan).

https://doi.org/10.1016/j.jmva.2021.104739
0047-259X/

Received: 6 May 2020 | Revised: 25 September 2020 | Accepted: 25 September 2020

DOI: 10.1002/bimj.202000129

Biometrical Journal

RESEARCH PAPER

D-optimal designs of mean-covariance models for longitudinal data

Siyu Yi[1,2] | **Yongdao Zhou**[1] | **Jianxin Pan**[3]

[1] School of Statistics and Data Science, LPMC & KLMDASR, Nankai University, P. R. China

[2] College of Mathematics, Sichuan University, Chengdu, P. R. China

[3] Department of Mathematics, University of Manchester, Manchester, UK

Correspondence
Yongdao Zhou, School of Statistics and Data Science, LPMC, and KLMDASR, Nankai University, Tianjin 300071, P. R. China.
Email: ydzhou@nankai.edu.cn

Funding information
National Key R&D Programme of China, Grant/Award Number: 2018YFC0831900

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to their computational complexity.

Abstract

Longitudinal data analysis has been very common in various fields. It is important in longitudinal studies to choose appropriate numbers of subjects and repeated measurements and allocation of time points as well. Therefore, existing studies proposed many criteria to select the optimal designs. However, most of them focused on the precision of the mean estimation based on some specific models and certain structures of the covariance matrix. In this paper, we focus on both the mean and the marginal covariance matrix. Based on the mean–covariance models, it is shown that the trick of symmetrization can generate better designs under a Bayesian D-optimality criterion over a given prior parameter space. Then, we propose a novel criterion to select the optimal designs. The goal of the proposed criterion is to make the estimates of both the mean vector and the covariance matrix more accurate, and the total cost is as low as possible. Further, we develop an algorithm to solve the corresponding optimization problem. Based on the algorithm, the criterion is illustrated by an application to a real dataset and some simulation studies. We show the superiority of the symmetric optimal design and the symmetrized optimal design in terms of the relative efficiency and parameter estimation. Moreover, we also demonstrate that the proposed criterion is more effective than the previous criteria, and it is suitable for both maximum likelihood estimation and restricted maximum likelihood estimation procedures.

KEYWORDS

Bayesian, cost function, D-optimality criterion, sequential number-theoretic optimization (SNTO)

1 | INTRODUCTION

Longitudinal data are very common in practice, for example, the randomized controlled trials in health and medical sciences, the quality control in industry and the growth curve analysis in biological and agriculture. In longitudinal data analysis, measurements are taken from the same subject repeatedly over time. The responses between subjects may be independent, but the repeated measurements within subjects are very likely to be correlated.

It is well known that the misspecification of the covariance structure may lead to a great loss of efficiency of the mean parameter estimators. Also, if the longitudinal data contain certain missing values and/or are not normally distributed, the mean parameter estimators may be biased when the covariance structure is misspecified (Daniels & Zhao, 2003).

