Technometrics, ISSN 0040-1706
Volume 59, number 4 (november 2017)

Abstract
Large-scale data analysis problems have become increasingly common across many disciplines. While large volume of data offers more statistical power, it also brings computational challenges. The orthogonalizing expectation–maximization (EM) algorithm by Xiong et al. is an efficient method to deal with large-scale least-square problems from a design point of view. In this article, we propose a reformulation and generalization of the orthogonalizing EM algorithm. Computational complexity and convergence guarantees are established. The reformulation of the orthogonalizing EM algorithm leads to a reduction in computational complexity for least-square problems and penalized least-square problems. The reformulation, named the GOEM (generalized orthogonalizing EM) algorithm, can incorporate a wide variety of convex and nonconvex penalties, including the lasso, group lasso, and minimax concave penalty penalties. The GOEM algorithm is further extended to a wider class of models including generalized linear models and Cox's proportional hazards model. Synthetic and real data examples are included to illustrate its use and efficiency compared with standard techniques. Supplementary materials for this article are available online.

Abstract
Partial least squares (PLS) is a prominent solution for dimension reduction and high-dimensional regressions. Recent prevalence of multidimensional tensor data has led to several tensor versions of the PLS algorithms. However, none offers a population model and interpretation, and statistical properties of the associated parameters remain intractable. In this article, we first propose a new tensor partial least-squares algorithm, then establish the corresponding population interpretation. This population investigation allows us to gain new insight on how the PLS achieves effective dimension reduction, to build connection with the notion of sufficient dimension reduction, and to obtain the asymptotic consistency of the PLS estimator. We compare our method, both analytically and numerically, with some alternative solutions. We also illustrate the efficacy of the new method on simulations and two neuroimaging data analyses. Supplementary materials for this article are available online.

Abstract
Group sparse approaches to regression modeling are finding ever increasing utility in an array of application areas. While group sparsity can help assess certain data structures, it is desirable in many instances to also capture element-wise sparsity. Recent work exploring the latter has been conducted in the context of l_2/l_1 penalized regression in the form of the sparse group lasso (SGL). Here, we present a novel model, called the sparse group elastic net (SGEN),

which uses an l_∞/l_1/ridge-based penalty. We show that the l_∞-norm, which induces group sparsity is particularly effective in the presence of noisy data. We solve the SGEN model using a coordinate descent-based procedure and compare its performance to the SGL and related methods in the context of hyperspectral imaging in the presence of noisy observations. Supplementary materials for this article are available online.

Split-Plot and Multi-Stratum Designs for Statistical Inference P. 446-457

Luzia A. Trinca & Steven G. Gilmour

Abstract

It is increasingly recognized that many industrial and engineering experiments use split-plot or other multi-stratum structures. Much recent work has concentrated on finding optimum, or near-optimum, designs for estimating the fixed effects parameters in multi-stratum designs. However, often inference, such as hypothesis testing or interval estimation, will also be required and for inference to be unbiased in the presence of model uncertainty requires pure error estimates of the variance components. Most optimal designs provide few, if any, pure error degrees of freedom. Gilmour and Trinca (2012 Gilmour, S. G., and Trinca, L. A. (2012), "Optimum Design of Experiments for Statistical Inference" (with discussion), *Applied Statistics*, 61, 345–401.[Crossref], [Web of Science ®], [Google Scholar]) introduced design optimality criteria for inference in the context of completely randomized and block designs. Here these criteria are used stratum-by-stratum to obtain multi-stratum designs. It is shown that these designs have better properties for performing inference than standard optimum designs. Compound criteria, which combine the inference criteria with traditional point estimation criteria, are also used and the designs obtained are shown to compromise between point estimation and inference. Designs are obtained for two real split-plot experiments and an illustrative split–split-plot structure. Supplementary materials for this article are available online.

Bayesian Design of Experiments Using Approximate Coordinate Exchange P. 458-470

Antony M. Overstall & David C. Woods

Abstract

The construction of decision-theoretical Bayesian designs for realistically complex nonlinear models is computationally challenging, as it requires the optimization of analytically intractable expected utility functions over high-dimensional design spaces. We provide the most general solution to date for this problem through a novel approximate coordinate exchange algorithm. This methodology uses a Gaussian process emulator to approximate the expected utility as a function of a single design coordinate in a series of conditional optimization steps. It has flexibility to address problems for any choice of utility function and for a wide range of statistical models with different numbers of variables, numbers of runs and randomization restrictions. In contrast to existing approaches to Bayesian design, the method can find multi-variable designs in large numbers of runs without resorting to asymptotic approximations to the posterior distribution or expected utility. The methodology is demonstrated on a variety of challenging examples of practical importance, including design for pharmacokinetic models and design for mixed models with discrete data. For many of these models, Bayesian designs are not currently available. Comparisons are made to results from the literature, and to designs obtained from asymptotic approximations. Supplementary materials for this article are available online.

Robust Parameter Designs in Computer Experiments Using Stochastic Approximation P. 471-483

Weijie Shen

Abstract

Robust parameter designs are widely used to produce products/processes that perform consistently well across various conditions known as noise factors. Recently, the robust parameter design method is implemented in computer experiments. The structure of conventional product array design becomes unsuitable due to its extensive number of

runs and the polynomial modeling. In this article, we propose a new framework robust parameter design via stochastic approximation (RPD-SA) to efficiently optimize the robust parameter design criteria. It can be applied to general robust parameter design problems, but is particularly powerful in the context of computer experiments. It has the following four advantages: (1) fast convergence to the optimal product setting with fewer number of function evaluations; (2) incorporation of high-order effects of both design and noise factors; (3) adaptation to constrained irregular region of operability; (4) no requirement of statistical analysis phase. In the numerical studies, we compare RPD-SA to the Monte Carlo sampling with Newton–Raphson-type optimization. An "Airfoil" example is used to compare the performance of RPD-SA, conventional product array designs, and space-filling designs with the Gaussian process. The studies show that RPD-SA has preferable performance in terms of effectiveness, efficiency and reliability.

Phase I Distribution-Free Analysis of Multivariate Data P. 484-495

Giovanna Capizzi & Guido Masarotto

Abstract

In this study, a new distribution-free Phase I control chart for retrospectively monitoring multivariate data is developed. The suggested approach, based on the multivariate signed ranks, can be applied to individual or subgrouped data for detection of location shifts with an arbitrary pattern (e.g., isolated, transitory, sustained, progressive, etc.). The procedure is complemented with a LASSO-based post-signal diagnostic method for identification of the shifted variables. A simulation study shows that the method compares favorably with parametric control charts when the process is normally distributed, and largely outperforms other multivariate nonparametric control charts when the process distribution is skewed or heavy-tailed. An R package can be found in the supplementary material.

Statistical Process Control for Latent Quality Characteristics Using the Up-and-Down Test P. 496-507

Dongdong Xiang, Fugee Tsung & Xiaolong Pu

Abstract

In many applications, the quality characteristic of a product is continuous but unobservable, for example, the critical electric voltage of electro-explosive devices. It is often important to monitor a manufacturing process of a product with such latent quality characteristic. Existing approaches all involve specifying a fixed stimulus level and testing products under that level to collect a sequence of response outcomes (zeros or ones). Appropriate control charts are then applied to the collected binary data sequence. However, these approaches offer limited performance. Moreover, the collected dataset provides little information for troubleshooting when an out-of-control signal is triggered. To overcome these limitations, this article introduces the up-and-down test for collecting data and proposes a new control chart based on this test. Numerical studies show that the proposed chart is able to detect any shifts effectively and is robust in many situations. Finally, an example involving real manufacturing data is given to demonstrate the use of our proposed chart.

A Vine Copula Model for Predicting the Effectiveness of Cyber Defense Early-Warning P. 508-520

Maochao Xu, Lei Hua & Shouhuai Xu

Abstract

Internet-based computer information systems play critical roles in many aspects of modern society. However, these systems are constantly under cyber attacks that can cause catastrophic consequences. To defend these systems effectively, it is necessary to measure and predict the effectiveness of cyber defense mechanisms. In this article, we investigate how to measure and predict the effectiveness of an important cyber defense mechanism that is known as *early-warning*. This turns out to be a challenging problem because we must accommodate the *dependence* among certain four-dimensional time series. In the course of using a dataset to demonstrate the prediction methodology, we

discovered a new *nonexchangeable* and *rotationally symmetric* dependence structure, which may be of independent value. We propose a new vine copula model to accommodate the newly discovered dependence structure, and show that the new model can predict the effectiveness of early-warning more accurately than the others. We also discuss how to use the prediction methodology in practice.

Hierarchical Spatially Varying Coefficient Process Model P. 521-527

Heeyoung Kim & Jaehwan Lee

Abstract

The spatially varying coefficient process model is a nonstationary approach to explaining spatial heterogen-eity by allowing coefficients to vary across space. In this article, we develop a methodology for generalizing this model to accommodate geographically hierarchical data. This article considers two-level hierarchical structures and allow for the coefficients of both low-level and high-level units to vary over space. We assume that the spatially varying low-level coefficients follow the multivariate Gaussian process, and the spatially varying high-level coefficients follow the multivariate simultaneous autoregressive model that we develop by extending the standard simultaneous autoregressive model to incorporate multivariate data. We apply the proposed model to transaction data of houses sold in 2014 in a part of the city of Los Angeles. The results show that the proposed model predicts housing prices and fits the data effectively.

Minimum Distance Estimation for the Generalized Pareto Distribution P. 528-541

Piao Chen, Zhi-Sheng Ye & Xingqiu Zhao

Abstract

The generalized Pareto distribution (GPD) is widely used for extreme values over a threshold. Most existing methods for parameter estimation either perform unsatisfactorily when the shape parameter k is larger than 0.5, or they suffer from heavy computation as the sample size increases. In view of the fact that $k > 0.5$ is occasionally seen in numerous applications, including two illustrative examples used in this study, we remedy the deficiencies of existing methods by proposing two new estimators for the GPD parameters. The new estimators are inspired by the minimum distance estimation and the *M*-estimation in the linear regression. Through comprehensive simulation, the estimators are shown to perform well for all values of k under small and moderate sample sizes. They are comparable to the existing methods for $k < 0.5$ while perform much better for $k > 0.5$.