Technometrics, ISSN 0040-1706
Volume 57, number 4 (november 2015)

Abstract
We propose a new class of models providing a powerful unification and extension of existing statistical methodology for analysis of data obtained in mixture experiments. These models, which integrate models proposed by Scheffé and Becker, extend considerably the range of mixture component effects that may be described. They become complex when the studied phenomenon requires it, but remain simple whenever possible. This article has supplementary material online.

Abstract
Multivariate polynomials are increasingly being used to construct emulators of computer models for uncertainty quantification. For deterministic computer codes, interpolating polynomial metamodels should be used instead of noninterpolating ones for logical consistency and prediction accuracy. However, available methods for constructing interpolating polynomials only provide point predictions. There is no known method that can provide probabilistic statements about the interpolation error. Furthermore, there are few alternatives to grid designs and sparse grids for constructing multivariate interpolating polynomials. A significant disadvantage of these designs is the large gaps between allowable design sizes. This article proposes a stochastic interpolating polynomial (SIP) that seeks to overcome the problems discussed above. A Bayesian approach in which interpolation uncertainty is quantified probabilistically through the posterior distribution of the output is employed. This allows assessment of the effect of interpolation uncertainty on estimation of quantities of interest based on the metamodel. A class of transformed space-filling design and a sequential design approach are proposed to efficiently construct the SIP with any desired number of runs. Simulations demonstrate that the SIP can outperform Gaussian process (GP) emulators. This article has supplementary material online.

Abstract
Robust parameter design with computer experiments is becoming increasingly important for product design. Existing methodologies for this problem are mostly for finding optimal control factor settings. However, in some cases, the objective of the experimenter may be to understand how the noise and control factors contribute to variation in the response. The functional analysis of variance (ANOVA) and variance decompositions of the response, in addition to the mean and variance models, help achieve this objective. Estimation of these quantities is not easy and few methods are able to quantity the estimation uncertainty. In this article, we show that the use of an orthonormal polynomial model of

the simulator leads to simple formulas for functional ANOVA and variance decompositions, and the mean and variance models. We show that estimation uncertainty can be taken into account in a simple way by first fitting a Gaussian process model to experiment data and then approximating it with the orthonormal polynomial model. This leads to a joint normal distribution for the polynomial coefficients that quantifies estimation uncertainty. Supplementary materials for this article are available online.

Optimal Sliced Latin Hypercube Designs P. 479-487

Shan Ba, William R. Myers & William A. Brenneman

Abstract

Sliced Latin hypercube designs (SLHDs) have important applications in designing computer experiments with continuous and categorical factors. However, a randomly generated SLHD can be poor in terms of space-filling, and based on the existing construction method that generates the SLHD column by column using sliced permutation matrices, it is also difficult to search for the optimal SLHD. In this article, we develop a new construction approach that first generates the small Latin hypercube design in each slice and then arranges them together to form the SLHD. The new approach is intuitive and can be easily adapted to generate orthogonal SLHDs and orthogonal array-based SLHDs. More importantly, it enables us to develop general algorithms that can search for the optimal SLHD efficiently.

Constructing General Orthogonal Fractional Factorial Split-Plot Designs P. 488-502

Bagus Sartono, Peter Goos & Eric Schoen

Abstract

While the orthogonal design of split-plot fractional factorial experiments has received much attention already, there are still major voids in the literature. First, designs with one or more factors acting at more than two levels have not yet been considered. Second, published work on nonregular fractional factorial split-plot designs was either based only on Plackett–Burman designs, or on small nonregular designs with limited numbers of factors. In this article, we present a novel approach to designing general orthogonal fractional factorial split-plot designs. One key feature of our approach is that it can be used to construct two-level designs as well as designs involving one or more factors with more than two levels. Moreover, the approach can be used to create two-level designs that match or outperform alternative designs in the literature, and to create two-level designs that cannot be constructed using existing methodology. Our new approach involves the use of integer linear programming and mixed integer linear programming, and, for large design problems, it combines integer linear programming with variable neighborhood search. We demonstrate the usefulness of our approach by constructing two-level split-plot designs of 16–96 runs, an 81-run three-level split-plot design, and a 48-run mixed-level split-plot design. Supplementary materials for this article are available online.

The Spatial LASSO With Applications to Unmixing Hyperspectral Biomedical Images P. 503-513

Daniel V. Samarov, Jeeseong Hwang & Maritoni Litorja

Abstract

Hyperspectral imaging (HSI) is a spectroscopic method that uses densely sampled measurements along the electromagnetic spectrum to identify the unique molecular composition of an object. Traditionally HSI has been associated with remote sensing-type applications, but recently has found increased use in biomedicine, from investigations at the cellular to the tissue level. One of the main challenges in the analysis of HSI is estimating the proportions, also called abundance fractions of each of the molecular signatures. While there is great promise for HSI in the area of biomedicine, large variability in the measurements and artifacts related to the instrumentation has slow adoption into more widespread practice. In this article, we propose a novel regularization and variable selection method called the spatial LASSO (SPLASSO). The SPLASSO incorporates spatial information via a graph Laplacian-based penalty to help improve the model estimation process for multivariate response data. We show the strong performance of this approach on a benchmark HSI dataset with considerable improvement in predictive accuracy over

the standard LASSO. Supplementary materials for this article are available online.

Informative Sensor and Feature Selection via Hierarchical Nonnegative Garrote P. 514-523
Kamran Paynabar, Judy Jin & Matthew P. Reed

Abstract
Placing sensors in every station of a process or every element of a system to monistor its state or performance is usually too expensive or physically impossible. Therefore, a systematic method is needed to select important sensing variables. The method should not only be capable of identifying important sensors/signals among multistream signals from a distributed sensing system, but should also be able to extract a small set of interpretable features from the high-dimensional vector of a selected signal. For this purpose, we develop a new hierarchical regularization approach called hierarchical nonnegative garrote (NNG). At the first level of hierarchy, a group NNG is used to select important signals, and at the second level, the individual features within each signal are selected using a modified version of NNG that possesses good properties for the estimated coefficients. Performance of the proposed method is evaluated and compared with other existing methods through Monte Carlo simulation. A case study is conducted to demonstrate the proposed methodology that can be applied to develop a predictive model for the assessment of vehicle design comfort based on the tested drivers' motion trajectory signals. This article has supplementary material online.

Matrix Discriminant Analysis With Application to Colorimetric Sensor Array Data P. 524-534
Wenxuan Zhong & Kenneth S. Suslick

Abstract
With the rapid development of nano-technology, a "colorimetric sensor array" (CSA) that is referred to as an optical electronic nose has been developed for the identification of toxicants. Unlike traditional sensors that rely on a single chemical interaction, CSA can measure multiple chemical interactions by using chemo-responsive dyes. The color changes of the chemo-responsive dyes are recorded before and after exposure to toxicants and serve as a template for classification. The color changes are digitalized in the form of a matrix with rows representing dye effects and columns representing the spectrum of colors. Thus, matrix-classification methods are highly desirable. In this article, we develop a novel classification method, matrix discriminant analysis (MDA), which is a generalization of linear discriminant analysis (LDA) for the data in matrix form. By incorporating the intrinsic matrix-structure of the data in discriminant analysis, the proposed method can improve CSA's sensitivity and more importantly, specificity. A penalized MDA method, PMDA, is also introduced to further incorporate sparsity structure in discriminant function. Numerical studies suggest that the proposed MDA and PMDA methods outperform LDA and other competing discriminant methods for matrix predictors. The asymptotic consistency of MDA is also established. R code and data are available online as supplementary material.

Malware Detection Using Nonparametric Bayesian Clustering and Classification Techniques P. 535-546
Yimin Kao, Brian Reich, Curtis Storlie & Blake Anderson

Abstract
Computer security requires statistical methods to quickly and accurately flag malicious programs. This article proposes a nonparametric Bayesian approach for classifying programs as benign or malicious and simultaneously clustering malicious programs. The analysis is based on the dynamic trace (DT) of instructions under the first-order Markov assumption. Each row of the trace's transition matrix is modeled using the Dirichlet process mixture (DPM) model. The DPM model clusters programs within each class (malicious or benign), and produces the posterior probability of being a malware which is used for classification. The novelty of the model is using this clustering algorithm to improve the classification accuracy. The simulation study shows that the DPM model outperforms the elastic net logistic (ENL) regression and the support vector machine (SVM) in classification performance under most of the scenarios, and also outperforms the spectral clustering method for grouping similar malware. In an analysis of real malicious and benign

programs, the DPM model gives significantly better classification performance than the ENL model, and competitive results to the SVM. More importantly, the DPM model identifies clusters of programs during the classification procedure which is useful for reverse engineering.

Confidence Regions and Intervals for Meta-Analysis Model Parameters P. 547-558

Andrew L. Rukhin

Abstract

This article obtains confidence regions for the heteroscedastic, one-way random effects model's parameters the heteroscedastic, one-way random effects model. The confidence regions are based on canonical representations of the restricted and profile likelihood functions in terms of independent normal random variables and χ^2 random variables. These regions provide conservative confidence intervals for the common mean and heterogeneity variance. Mathematical details and the R code are available online as supplementary material.

An Exact Confidence Set for a Maximum Point of a Univariate Polynomial Function in a Given Interval P. 559-565

Fang Wan, Wei Liu, Yang Han & Frank Bretz

Abstract

Construction of a confidence set for a maximum point of a function is an important statistical problem which has many applications. In this article, an exact $1 - \alpha$ confidence set is provided for a maximum point of a univariate polynomial function in a given interval. It is shown how the construction method can readily be applied to many parametric and semiparametric regression models involving a univariate polynomial function. Examples are given to illustrate this confidence set and to demonstrate that it can be substantially narrower and so better than the only other confidence set available in the statistical literature that guarantees $1 - \alpha$ confidence level.

A Nonparametric Kernel Approach to Interval-Valued Data Analysis P. 566-575

Yongho Jeon, Jeongyoun Ahn & Cheolwoo Park

Abstract

This article concerns datasets in which variables are in the form of intervals, which are obtained by aggregating information about variables from a larger dataset. We propose to view the observed set of hyper-rectangles as an empirical histogram, and to use a Gaussian kernel type estimator to approximate its underlying distribution in a nonparametric way. We apply this idea to both univariate density estimation and regression problems. Unlike many existing methods used in regression analysis, the proposed method can estimate the conditional distribution of the response variable for any given set of predictors even when some of them are not interval-valued. Empirical studies show that the proposed approach has a great flexibility in various scenarios with complex relationships between the location and width of intervals of the response and predictor variables.

A Semiparametric Software Reliability Model for Analysis of a Bug-Database With Multiple Defect Types P. 576-585

Vignesh T. Subrahmaniam, Anup Dewanji & Bimal K. Roy

Abstract

Software bug-databases provide an important source of data for assessing the reliability of a software product after its release. Statistical analysis of these databases can be challenging when software usage is unknown, that is, there is no information about the usage, either in the form of a parametric model, or in the form of actual measurements. Reliability metrics, when defined on a calendar time scale, would depend on this unknown and time-dependent usage of the software and hence cannot be estimated. This article proposes a semiparametric analysis that makes use of defect classifications into multiple types to enable estimation of a model without making strict assumptions about the

underlying usage of the software. New reliability metrics whose computation does not depend on the unknown usage of the software have been proposed and methods for estimating them have been developed. The proposed method has been illustrated using data retrieved from the bug-database of a popular scripting language, named Python. This illustration compares reliability of two versions of the software without making assumptions about their unknown usage. This article has supplementary material online.