Technometrics, ISSN 0040-1706
Volume 58, number 4 (november 2016)

Abstract
We consider regression scenarios where it is natural to impose an order constraint on the coefficients. We propose an order-constrained version of ℓ_1-regularized regression (Lasso) for this problem, and show how to solve it efficiently using the well-known pool adjacent violators algorithm as its proximal operator. The main application of this idea is to time-lagged regression, where we predict an outcome at time t from features at the previous K time points. In this setting, it is natural to assume that the coefficients decay as we move farther away from t, and hence the order constraint is reasonable. Potential application areas include financial time series and prediction of dynamic patient outcomes based on clinical measurements. We illustrate this idea on real and simulated data.

Abstract
A new sparse PCA algorithm is presented, which is robust against outliers. The approach is based on the ROBPCA algorithm that generates robust but nonsparse loadings. The construction of the new ROSPCA method is detailed, as well as a selection criterion for the sparsity parameter. An extensive simulation study and a real data example are performed, showing that it is capable of accurately finding the sparse structure of datasets, even when challenging outliers are present. In comparison with a projection pursuit-based algorithm, ROSPCA demonstrates superior robustness properties and comparable sparsity estimation capability, as well as significantly faster computation time.

Abstract
Distance covariance and distance correlation have been widely adopted in measuring dependence of a pair of random variables or random vectors. If the computation of distance covariance and distance correlation is implemented directly accordingly to its definition then its computational complexity is $O(n^2)$, which is a disadvantage compared to other faster methods. In this article we show that the computation of distance covariance and distance correlation of real-valued random variables can be implemented by an $O(n\log n)$ algorithm and this is comparable to other computationally efficient algorithms. The new formula we derive for an unbiased estimator for squared distance covariance turns out to be a U-statistic. This fact implies some nice asymptotic properties that were derived before via more complex methods. We apply the fast computing algorithm to some synthetic data. Our work will make distance correlation applicable to a much wider class of problems. A supplementary file to this article, available online, includes a Matlab and C-based software that realizes the proposed algorithm.

A Distribution-Free Multivariate Control Chart P. 448-459

Nan Chen, Xuemin Zi & Changliang Zou

Abstract

Monitoring multivariate quality variables or data streams remains an important and challenging problem in statistical process control (SPC). Although the multivariate SPC has been extensively studied in the literature, designing distribution-free control schemes are still challenging and yet to be addressed well. This article develops a new nonparametric methodology for monitoring location parameters when only a small reference dataset is available. The key idea is to construct a series of conditionally distribution-free test statistics in the sense that their distributions are free of the underlying distribution given the empirical distribution functions. The conditional probability that the charting statistic exceeds the control limit at present given that there is no alarm before the current time point can be guaranteed to attain a specified false alarm rate. The success of the proposed method lies in the use of data-dependent control limits, which are determined based on the observations online rather than decided before monitoring. Our theoretical and numerical studies show that the proposed control chart is able to deliver satisfactory in-control run-length performance for any distributions with any dimension. It is also very efficient in detecting multivariate process shifts when the process distribution is heavy-tailed or skewed. Supplementary materials for this article are available online.

Self-Starting Monitoring Scheme for Poisson Count Data With Varying Population Sizes P. 460-471

Xiaobei Shen, Kwok-Leung Tsui, Changliang Zou & William H. Woodall

Abstract

In this article, we consider the problem of monitoring Poisson rates when the population sizes are time-varying and the nominal value of the process parameter is unavailable. Almost all previous control schemes for the detection of increases in the Poisson rate in Phase II are constructed based on assumed knowledge of the process parameters, for example, the expectation of the count of a rare event when the process of interest is in control. In practice, however, this parameter is usually unknown and not able to be estimated with a sufficiently large number of reference samples. A self-starting exponentially weighted moving average (EWMA) control scheme based on a parametric bootstrap method is proposed. The success of the proposed method lies in the use of probability control limits, which are determined based on the observations during rather than before monitoring. Simulation studies show that our proposed scheme has good in-control and out-of-control performance under various situations. In particular, our proposed scheme is useful in rare event studies during the start-up stage of a monitoring process. Supplementary materials for this article are available online.

Online Updating of Computer Model Output Using Real-Time Sensor Data P. 472-482

Huijing Jiang, Xinwei Deng, Vanessa López & Hendrik F. Hamann

Abstract

Data center thermal management has become increasingly important because of massive computational demand in information technology. To advance the understanding of the thermal environment in a data center, complex computer models are extensively used to simulate temperature distribution maps. However, due to management policies and time constraints, it is not practical to execute such models in a real time fashion. In this article, we propose a novel statistical modeling method to perform real-time simulation by dynamically fusing a *base*, steady-state solution of a computer model, and *real-time* thermal sensor data. The proposed method uses a Kalman filter and stochastic gradient descent method as computational tools to achieve real-time updating of the base temperature map. We evaluate the performance of the proposed method through a simulation study and demonstrate its merits in a data center thermal management application. Supplementary materials for this article are available online.

Pairwise Meta-Modeling of Multivariate Output Computer Models Using Nonseparable Covariance Function P. 483-494

Yongxiang Li & Qiang Zhou

Abstract

Gaussian process (GP) is a popular method for emulating deterministic computer simulation models. Its natural extension to computer models with multivariate outputs employs a multivariate Gaussian process (MGP) framework. Nevertheless, with significant increase in the number of design points and the number of model parameters, building an MGP model is a very challenging task. Under a general MGP model framework with nonseparable covariance functions, we propose an efficient meta-modeling approach featuring a pairwise model building scheme. The proposed method has excellent scalability even for a large number of output levels. Some properties of the proposed method have been investigated and its performance has been demonstrated through several numerical examples. Supplementary materials for this article are available online.

Computer Experiments With Both Qualitative and Quantitative Variables P. 495-507

Hengzhen Huang, Dennis K. J. Lin, Min-Qian Liu & Jian-Feng Yang

Abstract

Computer experiments have received a great deal of attention in many fields of science and technology. Most literature assumes that all the input variables are quantitative. However, researchers often encounter computer experiments involving both qualitative and quantitative variables (BQQV). In this article, a new interface on design and analysis for computer experiments with BQQV is proposed. The new designs are one kind of sliced Latin hypercube designs with points clustered in the design region and possess good uniformity for each slice. For computer experiments with BQQV, such designs help to measure the similarities among responses of different level-combinations in the qualitative variables. An adaptive analysis strategy intended for the proposed designs is developed. The proposed strategy allows us to automatically extract information from useful auxiliary responses to increase the precision of prediction for the target response. The interface between the proposed design and the analysis strategy is demonstrated to be effective via simulation and a real-life example from the food engineering literature. Supplementary materials for this article are available online.

A Note on Foldover of 2^{k-p} Designs With Column Permutations P. 508-512

William Li & Dennis K. J. Lin

Abstract

Foldover is a commonly used follow-up strategy in experimental designs. All existing foldover designs were constructed by reversing the sign of columns of the initial design. We propose a new methodology by allowing the permutation of columns in foldover. Focusing on resolution IV designs, we show that almost all designs are better than existing results with respect to the minimum aberration criterion. While augmenting a design by a foldover with column permutations may result in a nonregular combined design, the proposed designs all have a resolution of 4.5 or higher, for which no two-factor interaction is fully aliased with any other two-factor interactions.

Augmenting the Unreturned for Field Data With Information on Returned Failures Only P. 513-523

Zhi-Sheng Ye & Loon-Ching Tang

Abstract

Field data are an important source of reliability information for many commercial products. Because field data are often collected by the maintenance department, information on failed and returned units is well maintained. Nevertheless, information on unreturned units is generally unavailable. The unavailability leads to truncation in the lifetime data. This study proposes a data-augmentation algorithm for this type of truncated field return data with returned failures available only. The algorithm is based on an idea to reveal the hidden unobserved lifetimes.

Theoretical justifications of the procedure for augmenting the hidden unobserved are given. On the other hand, the algorithm is iterative in nature. Asymptotic properties of the estimators from the iterations are investigated. Both point estimation and the information matrix of the parameters can be directly obtained from the algorithm. In addition, a by-product of the algorithm is a nonparametric estimator of the installation time distribution. An example from an asset-rich company is given to demonstrate the proposed met

Technometrics, ISSN 0040-1706
Volume 58, number 4 (november 2016)