Technometrics, ISSN 0040-1706
Volume 60, number 2 (may 2018)

Detecting Deviating Data Cells P. 135-145

Peter J. Rousseeuw & Wannes Van Den Bossche

Abstract

A multivariate dataset consists of *n* cases in *d* dimensions, and is often stored in an *n* by *d* data matrix. It is well-known that real data may contain outliers. Depending on the situation, outliers may be (a) undesirable errors, which can adversely affect the data analysis, or (b) valuable nuggets of unexpected information. In statistics and data analysis, the word outlier usually refers to a row of the data matrix, and the methods to detect such outliers only work when at least half the rows are clean. But often many rows have a few contaminated cell values, which may not be visible by looking at each variable (column) separately. We propose the first method to detect deviating data cells in a multivariate sample which takes the correlations between the variables into account. It has no restriction on the number of clean rows, and can deal with high dimensions. Other advantages are that it provides predicted values of the outlying cells, while imputing missing values at the same time. We illustrate the method on several real datasets, where it uncovers more structure than found by purely columnwise methods or purely rowwise methods. The proposed method can help to diagnose why a certain row is outlying, for example, in process control. It also serves as an initial step for estimating multivariate location and scatter matrices.

From Least Squares to Signal Processing and Particle Filtering P. 146-160

Nozer D. Singpurwalla, Nicholas G. Polson & Refik Soyer

Abstract

De facto, signal processing is the interpolation and extrapolation of a sequence of observations viewed as a realization of a stochastic process. Its role in applied statistics ranges from scenarios in forecasting and time series analysis to image reconstruction, machine learning, and the degradation modeling for reliability assessment. This topic, which has an old and honorable history dating back to the times of Gauss and Legendre, should therefore be of interest to readers of *Technometrics*. A general solution to the problem of filtering and prediction entails some formidable mathematics. Efforts to circumvent the mathematics has resulted in the need for introducing more explicit descriptions of the underlying process. One such example, and a noteworthy one, is the Kalman filter model, which is a special case of state space models or what statisticians refer to as dynamic linear models. Implementing the Kalman filter model in the era of "big and high velocity non-Gaussian data" can pose computational challenges with respect to efficiency and timeliness. Particle filtering is a way to ease such computational burdens. The purpose of this article is to trace the historical evolution of this development from its inception to its current state, with an expository focus on two versions of the particle filter, namely, the propagate first-update next and the update first-propagate next version. By way of going beyond a pure review, this article also makes transparent the importance and the role of a less recognized principle, namely, the *principle of conditionalization*, in filtering and prediction based on Bayesian methods. Furthermore, the article also articulates the philosophical underpinnings of the filtering and prediction set-up, a matter that needs to be made explicit, and Yule's decomposition of a random variable in terms of a sequence of innovations.

Efficient Sparse Estimate of Sufficient Dimension Reduction in High Dimension P. 161-168

Xin Chen, Wenhui Sheng & Xiangrong Yin

Abstract

In this article, we propose a new efficient sparse estimate (ESE) in sufficient dimension reduction using distance covariance. Our method is model-free and does not need any kernel function or slicing selection. Moreover, it can naturally deal with multivariate response scenarios, making it appealing in a modified sequential algorithm that targets the large p small n problems. Compared with screening procedures that only use marginal utility, our method can extract more useful information from the data and is capable of determining the size of the selected submodel automatically while most of screening procedures cannot. Under mild conditions, based on manifold theories and techniques, it can be shown that our method would perform asymptotically as well as if the true irrelevant predictors were known, which is referred to as the oracle property. Extensive simulation studies and two real data examples demonstrate the effectiveness and efficiency of the proposed approach. It is remarkable that the analysis in cardiomyopathy microarray data reveals distinct and interesting findings. Supplemental materials for this article are available online.

Phase I Monitoring of Spatial Surface Data from 3D Printing P. 169-180

Yangyang Zang & Peihua Qiu

Abstract

In recent years, 3D printing gets more and more popular in manufacturing industries. Quality control of 3D printing products thus becomes an important research problem. However, this problem is challenging due to the facts that (i) the surface of a product from 3D printing can have arbitrary shape, even when the 3D printing process is in-control, (ii) surface observations of the product obtained from a laser scanner may not have regularly spaced locations, and (iii) the overall geometric positions of 3D printing products might be all different, making proper comparison among different products difficult. In this article, we propose a Phase I control chart for monitoring products from 3D printing that addresses all these challenges. Numerical studies show that it works well in practice.

Real-Time Monitoring of High-Dimensional Functional Data Streams via Spatio-Temporal Smooth Sparse Decomposition P. 181-197

Hao Yan, Kamran Paynabar & Jianjun Shi

Abstract

High-dimensional data monitoring and diagnosis has recently attracted increasing attention among researchers as well as practitioners. However, existing process monitoring methods fail to fully use the information of high-dimensional data streams due to their complex characteristics including the large dimensionality, spatio-temporal correlation structure, and nonstationarity. In this article, we propose a novel process monitoring methodology for high-dimensional data streams including profiles and images that can effectively address foregoing challenges. We introduce spatio-temporal smooth sparse decomposition (ST-SSD), which serves as a dimension reduction and denoising technique by decomposing the original tensor into the functional mean, sparse anomalies, and random noises. ST-SSD is followed by a sequential likelihood ratio test on extracted anomalies for process monitoring. To enable real-time implementation of the proposed methodology, recursive estimation procedures for ST-SSD are developed. ST-SSD also provides useful diagnostics information about the location of change in the functional mean. The proposed methodology is validated through various simulations and real case studies. Supplementary materials for this article are available online.

Resolution Adaptive Fixed Rank Kriging P. 198-208

ShengLi Tzeng & Hsin-Cheng Huang

Abstract

The spatial random effects model is flexible in modeling spatial covariance functions and is computationally efficient

for spatial prediction via fixed rank kriging (FRK). However, the model depends on a class of basis functions, which if not selected properly, may result in unstable or undesirable results. Additionally, the maximum likelihood (ML) estimates of the model parameters are commonly computed using an expectation-maximization (EM) algorithm, which further limits its applicability when a large number of basis functions are required. In this research, we propose a class of basis functions extracted from thin-plate splines. The functions are ordered in terms of their degrees of smoothness with higher-order functions corresponding to larger-scale features and lower-order ones corresponding to smaller-scale details, leading to a parsimonious representation of a (nonstationary) spatial covariance function with the number of basis functions playing the role of spatial resolution. The proposed class of basis functions avoids the difficult knot-allocation or scale-selection problem. In addition, we show that ML estimates of the random effects covariance matrix can be expressed in simple closed forms, and hence the resulting FRK can accommodate a much larger number of basis functions without numerical difficulties. Finally, we propose to select the number of basis functions using Akaike's information criterion, which also possesses a simple closed-form expression. The whole procedure, involving no additional tuning parameter, is efficient to compute, easy to program, automatic to implement, and applicable to massive amounts of spatial data even when they are sparsely and irregularly located. Proofs of the theorems and an R package *autoFRK* are provided in supplementary materials available online.

Gaussian Process Modeling of a Functional Output with Information from Boundary and Initial Conditions and Analytical Approximations P. 209-221

Matthias HY Tan

Abstract

A partial differential equation (PDE) models a physical quantity as a function of space and time. These models are often solved numerically with the finite element (FE) method and the computer output consists of values of the solution on a fine grid over the spatial and temporal domain. When the simulations are time-consuming, Gaussian process (GP) models can be used to approximate the relationship between the functional output and the computer inputs, which consists of boundary and initial conditions. The Dirichlet boundary and initial conditions give the functional output values on parts of the space-time domain boundary. Although this information can help improve prediction of the output, it has not been used to construct GP models. In addition, analytical solutions of the PDE derived by simplifying the PDE can often be obtained, which can help further improve performance of the GP model. This article proposes a Karhunen–Loève (KL) expansion-based GP model that satisfies the Dirichlet boundary and initial conditions almost surely, and effectively uses information from analytical approximations to the PDE solution. Real examples demonstrate the improved prediction performance achieved by using these sources of prior information. Supplementary materials for this article are available online.

Semiparametric Models for Accelerated Destructive Degradation Test Data Analysis P. 222-234

Yimeng Xie, Caleb B. King, Yili Hong & Qingyu Yang

Abstract

Accelerated destructive degradation tests (ADDT) are widely used in industry to evaluate materials' long-term properties. Even though there has been tremendous statistical research in nonparametric methods, the current industrial practice is still to use application-specific parametric models to describe ADDT data. The challenge of using a nonparametric approach comes from the need to retain the physical meaning of degradation mechanisms and also perform extrapolation for predictions at the use condition. Motivated by this challenge, we propose a semiparametric model to describe ADDT data. We use monotonic B-splines to model the degradation path, which not only provides flexible models with few assumptions, but also retains the physical meaning of degradation mechanisms (e.g., the degradation path is monotonic). Parametric models, such as the Arrhenius model, are used for modeling the relationship between the degradation and the accelerating variable, allowing for extrapolation to the use condition. We develop an efficient procedure to estimate model parameters. We also use simulations to validate the developed procedures and demonstrate the robustness of the semiparametric model under model misspecification. Finally, the

proposed method is illustrated by multiple industrial applications. This article has online supplementary materials.

Inference on the Gamma Distribution P. 235-244

Bing Xing Wang & Fangtao Wu

Abstract

This study develops inferential procedures for a gamma distribution. Based on the Cornish–Fisher expansion and pivoting the cumulative distribution function, an approximate confidence interval for the gamma shape parameter is derived. The generalized confidence intervals for the rate parameter and other quantities such as mean are explored. The proposed generalized inferential procedures are extended to construct prediction limits for a single future measurement and for at least p of m measurements at each of r locations. The performance of the proposed procedures is evaluated using Monte Carlo simulation. The simulation results show that the proposed procedures are very satisfactory. Finally, three real examples are used to illustrate the proposed procedures. Supplementary materials for this article are available online.

Band Depth Clustering for Nonstationary Time Series and Wind Speed Behavior P. 245-254

Laura L. Tupper, David S. Matteson, C. Lindsay Anderson & Luckny Zephyr

Abstract

We explore the behavior of wind speed over time, using a subset of the Eastern Wind Dataset published by the National Renewable Energy Laboratory. This dataset gives modeled wind speeds over three years at hundreds of potential wind farm sites. Wind speed analysis is necessary to the integration of wind energy into the power grid; short-term variability in wind speed affects decisions about usage of other power sources, so that the shape of the wind speed time series becomes as important as the overall level. To assess differences in intra-day time series, we propose a functional distance measure, the band distance, which extends the band depth of López-Pintado and Romo. This measure emphasizes the shape of time series or functional observations relative to other members of a dataset and allows clustering of observations without reliance on pointwise Euclidean distance. We show a method for adjusting for seasonal effects in wind speed, and use these standardizations as input for the band distance. We demonstrate the utility of the new method in simulation studies and an application to the MOST power grid algorithm, where the band distance improves reliability over standard methods at a comparable cost.

Model Calibration With Censored Data P. 255-262

Fang Cao, Shan Ba, William A. Brenneman & V. Roshan Joseph

Abstract

The purpose of model calibration is to make the model predictions closer to reality. The classical Kennedy–O'Hagan approach is widely used for model calibration, which can account for the inadequacy of the computer model while simultaneously estimating the unknown calibration parameters. In many applications, the phenomenon of censoring occurs when the exact outcome of the physical experiment is not observed, but is only known to fall within a certain region. In such cases, the Kennedy–O'Hagan approach cannot be used directly, and we propose a method to incorporate the censoring information when performing model calibration. The method is applied to study the compression phenomenon of liquid inside a bottle. The results show significant improvement over the traditional calibration methods, especially when the number of censored observations is large. Supplementary materials for this article are available online.