Technometrics, ISSN 0040-1706
Volume 61, number 4 (november 2019)

Abstract
The empirical quantiles of independent data provide a good summary of the underlying distribution of the observations. For high-dimensional time series defined in two dimensions, such as in space and time, one can define empirical quantiles of all observations at a given time point, but such time-wise quantiles can only reflect properties of the data at that time point. They often fail to capture the dynamic dependence of the data. In this article, we propose a new definition of empirical dynamic quantiles (EDQ) for high-dimensional time series that mitigates this limitation by imposing that the quantile must be one of the observed time series. The word *dynamic* emphasizes the fact that these newly defined quantiles capture the time evolution of the data. We prove that the EDQ converge to the time-wise quantiles under some weak conditions as the dimension increases. A fast algorithm to compute the dynamic quantiles is presented and the resulting quantiles are used to produce summary plots for a collection of many time series. We illustrate with two real datasets that the time-wise and dynamic quantiles convey different and complementary information. We also briefly compare the visualization provided by EDQ with that obtained by functional depth. The R code and a vignette for computing and plotting EDQ are available at https://github.com/dpena157/HDts/.

Abstract
There has been extensive work on data depth-based methods for robust multivariate data analysis. Recent developments have moved to infinite-dimensional objects, such as functional data. In this work, we propose a notion of depth, the total variation depth, for functional data, which has many desirable features and is well suited for outlier detection. The proposed depth is in the form of an integral of a univariate depth function. We show that the novel formation of the total variation depth leads to useful decomposition associated with shape and magnitude outlyingness of functional data. Compared to magnitude outliers, shape outliers are often masked among the rest of samples and more difficult to identify. We then further develop an effective procedure and visualization tools for detecting both types of outliers, while naturally accounting for the correlation in functional data. The outlier detection performance is investigated through simulations under various outlier models. Finally, the proposed methodology is demonstrated using real datasets of curves, images, and video frames.

Abstract
Multivariate data are typically represented by a rectangular matrix (table) in which the rows are the objects (cases) and the columns are the variables (measurements). When there are many variables one often reduces the dimension by

principal component analysis (PCA), which in its basic form is not robust to outliers. Much research has focused on handling rowwise outliers, that is, rows that deviate from the majority of the rows in the data (e.g., they might belong to a different population). In recent years also cellwise outliers are receiving attention. These are suspicious cells (entries) that can occur anywhere in the table. Even a relatively small proportion of outlying cells can contaminate over half the rows, which causes rowwise robust methods to break down. In this article, a new PCA method is constructed which combines the strengths of two existing robust methods to be robust against both cellwise and rowwise outliers. At the same time, the algorithm can cope with missing values. As of yet it is the only PCA method that can deal with all three problems simultaneously. Its name MacroPCA stands for PCA allowing for Missingness And Cellwise & Rowwise Outliers. Several simulations and real datasets illustrate its robustness. New residual maps are introduced, which help to determine which variables are responsible for the outlying behavior. The method is well-suited for online process control.

Profile Extrema for Visualizing and Quantifying Uncertainties on Excursion Regions: Application to Coastal Flooding P. 474-493

Dario Azzimonti, David Ginsbourger, Jérémy Rohmer & Déborah Idier

Abstract

We consider the problem of describing excursion sets of a real-valued function f, that is, the set of inputs where f is above a fixed threshold. Such regions are hard to visualize if the input space dimension, d, is higher than 2. For a given projection matrix from the input space to a lower dimensional (usually 1, 2) subspace, we introduce profile sup (inf) functions that associate to each point in the projection's image the sup (inf) of the function constrained over the pre-image of this point by the considered projection. Plots of profile extrema functions convey a simple, although intrinsically partial, visualization of the set. We consider expensive to evaluate functions where only a very limited number of evaluations, n, is available, for example, $n<100d$, and we surrogate f with a posterior quantity of a Gaussian process (GP) model. We first compute profile extrema functions for the posterior mean given n evaluations of f. We quantify the uncertainty on such estimates by studying the distribution of GP profile extrema with posterior quasi-realizations obtained from an approximating process. We control such approximation with a bound inherited from the Borell-TIS inequality. The technique is applied to analytical functions ($d = 2$, 3) and to a five-dimensional coastal flooding test case for a site located on the Atlantic French coast. Here f is a numerical model returning the area of flooded surface in the coastal region given some offshore conditions. Profile extrema functions allowed us to better understand which offshore conditions impact large flooding events.

Spatial Signal Detection Using Continuous Shrinkage Priors P. 494-506

An-Ting Jhuang, Montserrat Fuentes, Jacob L. Jones, Giovanni Esteves, Chris M. Fancher, Marschall Furman & Brian J. Reich

Abstract

Motivated by the problem of detecting changes in two-dimensional X-ray diffraction data, we propose a Bayesian spatial model for sparse signal detection in image data. Our model places considerable mass near zero and has heavy tails to reflect the prior belief that the image signal is zero for most pixels and large for an important subset. We show that the spatial prior places mass on nearby locations simultaneously being zero, and also allows for nearby locations to simultaneously be large signals. The form of the prior also facilitates efficient computing for large images. We conduct a simulation study to evaluate the properties of the proposed prior and show that it outperforms other spatial models. We apply our method in the analysis of X-ray diffraction data from a two-dimensional area detector to detect changes in the pattern when the material is exposed to an electric field.

Mixture of Regression Models for Large Spatial Datasets P. 507-523

Karen Kazor & Amanda S. Hering

Abstract

When a spatial regression model that links a response variable to a set of explanatory variables is desired, it is unlikely that the same regression model holds throughout the domain when the spatial domain and dataset are both large and complex. The locations where the trend changes may not be known, and we present here a mixture of regression models approach to identifying the locations wherein the relationship between the predictors and the response is similar; to estimating the model within each group; and to estimating the number of groups. An EM algorithm for estimating this model is presented along with a criterion for choosing the number of groups. Performance of the estimators and model selection are demonstrated through simulation. An example with groundwater depth and associated predictors generated from a large physical model simulation demonstrates the fit and interpretation of the proposed model. R code is provided in the supplementary materials that simulates the scenarios tested herein; implements the method; and reproduces the groundwater depth results. Supplementary materials for this article are available online.

Central Composite Experimental Designs for Multiple Responses With Different Models P. 524-532

Wilmina M. Marget & Max D. Morris

Abstract

Central composite designs (CCDs) are widely accepted and used experimental designs for fitting second-order polynomial models in response surface methods. However, these designs are based only on the number of explanatory variables being investigated. In a multiresponse problem where prior information is available in the form of a screening experiment or previous process knowledge, investigators often know which factors will be used in the estimation of each response. This work presents an alternative design based on CCDs that allows main effects to be aliased for factors that are not related to the same response. This results in fewer required runs than current designs, saving investigators both time and money, by taking this prior information into account. R-package "DoE.multi.response" is included as a supplement for constructing these designs.

Optimal Experimental Design in the Presence of Nested Factors P. 533-544

Peter Goos & Bradley Jones

Abstract

A common occurrence in practical design of experiments is that one factor, called a nested factor, can only be varied for some but not all the levels of a categorical factor, called a branching factor. In this case, it is possible, but inefficient, to proceed by performing two experiments. One experiment would be run at the level(s) of the branching factor that allow for varying the second, nested, factor. The other experiment would only include the other level(s) of the branching factor. It is preferable to perform one experiment that allows for assessing the effects of both factors. Clearly, the effect of the nested factor then is conditional on the levels of the branching factor for which it can be varied. For example, consider an experiment comparing the performance of two machines where one machine has a switch that is missing for the other machine. The investigator wants to compare the two machines but also wants to understand the effect of flipping the switch. The main effect of the switch is conditional on the machine. This article describes several example situations involving branching factors and nested factors. We provide a model that is sensible for each situation, present a general method for constructing appropriate models, and show how to generate optimal designs given these models.

Analysis-of-Marginal-Tail-Means (ATM): A Robust Method for Discrete Black-Box Optimization P. 545-559

Simon Mak & C. F. Jeff Wu

Abstract

We present a new method, called analysis-of-marginal-tail-means (ATM), for effective robust optimization of discrete black-box problems. ATM has important applications in many real-world engineering problems (e.g., manufacturing

optimization, product design, and molecular engineering), where the objective to optimize is black-box and expensive, and the design space is inherently discrete. One weakness of existing methods is that they are not robust: these methods perform well under certain assumptions, but yield poor results when such assumptions (which are difficult to verify in black-box problems) are violated. ATM addresses this by combining both rank- and model-based optimization, via the use of marginal tail means. The trade-off between rank- and model-based optimization is tuned by first identifying important main effects and interactions from data, then finding a good compromise which best exploits additive structure. ATM provides improved robust optimization over existing methods, particularly in problems with (i) a large number of factors, (ii) unordered factors, or (iii) experimental noise. We demonstrate the effectiveness of ATM in simulations and in two real-world engineering problems: the first on robust parameter design of a circular piston, and the second on product family design of a thermistor network.