Technometrics, ISSN 0040-1706
Volume 63, number 2 (may 2021)

Abstract

In recent years, measurement or collection of heterogeneous sets of data such as those containing scalars, waveform signals, images, and even structured point clouds, has become more common. Statistical models based on such heterogeneous sets of data that represent the behavior of an underlying system can be used in the monitoring, control, and optimization of the system. Unfortunately, available methods mainly focus on the scalars and profiles and do not provide a general framework for integrating different sources of data to construct a model. This article addresses the problem of estimating a process output, measured by a scalar, curve, image, or structured point cloud by a set of heterogeneous process variables such as scalar process setting, profile sensor readings, and images. We introduce a general multiple tensor-on-tensor regression approach in which each set of input data (predictor) and output measurements are represented by tensors. We formulate a linear regression model between the input and output tensors and estimate the parameters by minimizing a least square loss function. To avoid overfitting and reduce the number of parameters to be estimated, we decompose the model parameters using several basis matrices that span the input and output spaces, and provide efficient optimization algorithms for learning the basis and coefficients. Through several simulation and case studies, we evaluate the performance of the proposed method. The results reveal the advantage of the proposed method over some benchmarks in the literature in terms of the mean square prediction error. Supplementary materials for this article are available online.

Abstract

Motivated by brain connectome datasets acquired using diffusion weighted magnetic resonance imaging (DWI), this article proposes a novel generalized Bayesian linear modeling framework with a symmetric tensor response and scalar predictors. The symmetric tensor coefficients corresponding to the scalar predictors are embedded with two features: low-rankness and group sparsity within the low-rank structure. Besides offering computational efficiency and parsimony, these two features enable identification of important “tensor nodes” and “tensor cells” significantly associated with the predictors, with characterization of uncertainty. The proposed framework is empirically investigated under various simulation settings and with a real brain connectome dataset. Theoretically, we establish that the posterior predictive density from the proposed model is “close” to the true data generating density, the closeness being measured by the Hellinger distance between these two densities, which scales at a rate very close to the finite dimensional optimal rate of n−1/2n−1/2, depending on how the number of tensor nodes grow with the sample size. The theoretical results with proofs are provided in the supplementary materials which are available online.

Adaptive Convex Clustering of Generalized Linear Models With Application in Purchase Likelihood Prediction P. 171-183

Shuyu Chu, Huijing Jiang, Zhengliang Xue & Xinwei Deng

Abstract

In the pricing of customized products, it is challenging to accurately predict the purchase likelihood of potential clients for each personalized request. The heterogeneity of customers and their responses to the personalized products leads to very different purchase behavior. Thus, it is often not appropriate to use a single model to analyze all the pricing data. There is a great need to construct distinctive models for different data segments. In this work, we propose an adaptive convex clustering method to perform data segmentation and model fitting simultaneously for generalized linear models. The proposed method segments data points using the fused penalty to account for the similarity in model structures. It ensures that the data points sharing the same model structure are grouped into the same segment. Accordingly, we develop an efficient algorithm for parameter estimation and study its consistency properties in estimation and clustering. The performance of our approach is evaluated by both numerical examples and case studies of real business data.

Fast Robust Correlation for High-Dimensional Data P. 184-198

Jakob Raymaekers & Peter J. Rousseeuw

Abstract

The product moment covariance matrix is a cornerstone of multivariate data analysis, from which one can derive correlations, principal components, Mahalanobis distances and many other results. Unfortunately, the product moment covariance and the corresponding Pearson correlation are very susceptible to outliers (anomalies) in the data. Several robust estimators of covariance matrices have been developed, but few are suitable for the ultrahigh-dimensional data that are becoming more prevalent nowadays. For that one needs methods whose computation scales well with the dimension, are guaranteed to yield a positive semidefinite matrix, and are sufficiently robust to outliers as well as sufficiently accurate in the statistical sense of low variability. We construct such methods using data transformations. The resulting approach is simple, fast, and widely applicable. We study its robustness by deriving influence functions and breakdown values, and computing the mean squared error on contaminated data. Using these results we select a method that performs well overall. This also allows us to construct a faster version of the DetectDeviatingCells method (Rousseeuw and Van den Bossche 2018) to detect cellwise outliers, which can deal with much higher dimensions. The approach is illustrated on genomic data with 12,600 variables and color video data with 920,000 dimensions. Supplementary materials for this article are available online.

Bivariate Functional Quantile Envelopes With Application to Radiosonde Wind Data P. 199-211

Gaurav Agarwal & Ying Sun

Abstract

The global radiosonde archives contain valuable weather data, such as temperature, humidity, wind speed, wind direction, and atmospheric pressure. Being the only direct measurement of these variables in the upper air, they are prone to errors. Therefore, a robust analysis and outlier detection of radiosonde data is essential. Among all the variables, the radiosonde winds, which consist of wind speed and direction, are particularly challenging to analyze. In this article, we treat the wind profiles as bivariate functional data across several pressure levels. Since the bivariate distribution of the components of radiosonde winds at a given pressure level is not Gaussian but instead skewed and heavy-tailed, we propose a set of robust quantile methods to characterize the distribution as well as an outlier detection procedure to identify both magnitude and shape outliers. The proposed methods provide an informative visualization tool for bivariate functional data. We also introduce two methods of predicting this bivariate distribution at unobserved pressure levels. In our simulation study, we show that our methods are robust against different types of outliers and skewed data. Finally, we apply our methods to radiosonde wind data to illustrate our proposed quantile analysis methods for visualization, outlier detection, and prediction.

A Component-Position Model, Analysis and Design for Order-of-Addition Experiments P. 212-224

Jian-Feng Yang, Fasheng Sun & Hongquan Xu

Abstract

An order-of-addition experiment is a kind of experiment in which the response is affected by the addition order of materials or components. In many situations, performing the full design with all possible permutations of components is unaffordable, especially when the number of components is larger than four. We introduce a component-position model for analyzing data from such experiments and study associated problems. We further propose a new type of design, called component orthogonal array, as a fraction of the full design for order-of-addition experiments. It is shown that component orthogonal arrays have the same D-efficiency as the full design under our proposed model. Component orthogonal arrays also perform well under the existing pairwise ordering model. Two drug combination experiments are used to show the effectiveness of the proposed model and designs. Supplementary materials for this paper are available online.

The Reconstruction Approach: From Interpolation to Regression P. 225-235

Shifeng Xiong

Abstract

This article introduces an interpolation-based method, called the reconstruction approach, for nonparametric regression. Based on the fact that interpolation usually has negligible errors compared to statistical estimation, the reconstruction approach uses an interpolator to parameterize the regression function with its values at finite knots, and then estimates these values by (regularized) least squares. Some popular methods including kernel ridge regression can be viewed as its special cases. It is shown that the reconstruction idea not only provides different angles to look into existing methods, but also produces new effective experimental design and estimation methods for nonparametric models. In particular, for some methods of complexity O(n3)O(n3), where n is the sample size, this approach provides effective surrogates with much less computational burden. This point makes it very suitable for large datasets. Supplementary materials for this article are available online.

Sequential Optimization in Locally Important Dimensions P. 236-248

Munir A. Winkel, Jonathan W. Stallrich, Curtis B. Storlie & Brian J. Reich

Abstract

Optimizing an expensive, black-box function f(·)f(·) is challenging when its input space is high-dimensional. Sequential design frameworks first model f(·)f(·) with a surrogate function and then optimize an acquisition function to determine input settings to evaluate next. Optimization of both f(·)f(·) and the acquisition function benefit from effective dimension reduction. Global variable selection detects and removes input variables that do not affect f(·)f(·) across the input space. Further dimension reduction may be possible if we consider local variable selection around the current optimum estimate. We develop a sequential design algorithm called *sequential optimization in locally important dimensions* (SOLID) that incorporates global and local variable selection to optimize a continuous, differentiable function. SOLID performs local variable selection by comparing the surrogate's predictions in a localized region around the estimated optimum with the p alternative predictions made by removing each input variable. The search space of the acquisition function is further restricted to focus only on the variables that are deemed locally active, leading to greater emphasis on refining the surrogate model in locally active dimensions. A simulation study across multiple test functions and an application to the Sarcos robot dataset show that SOLID outperforms conventional approaches. Supplementary materials for this article are available online.

Abstract
In the fabrication of artificial soft tissues, novel biomaterials with the required properties are obtained by appropriately adjusting process parameters during material synthesis. One key step in finding the desired material is understanding the relationship between the process parameters and the material properties, and time-course experiments are typically conducted for this purpose. This article proposes a constrained varying-coefficient modeling method for such data in which expert knowledge is properly accommodated in the model estimation to make the modeling practically meaningful. The proposed model has a semiparametric structure and incorporates expert knowledge in the form of constraints on model coefficients. Estimation algorithms based on a smoothing spline and a weighted smoothing spline are also provided. Finally, the proposed method is compared with existing methods in a case study and a numerical study.

Abstract
Ridge regression was introduced to deal with the instability issue of the ordinary least squares estimate due to multicollinearity. It essentially penalizes the least squares loss by applying a ridge penalty on the regression coefficients. The ridge penalty shrinks the regression coefficient estimate toward zero, but not exactly zero. For this reason, the ridge regression has long been criticized of not being able to perform variable selection. In this article, we proposed a new variable selection method based on an individually penalized ridge regression, a slightly generalized version of the ridge regression. An adaptive version is also provided. Our new methods are shown to perform competitively based on simulation and a real data example.