Technometrics, ISSN 0040-1706
Volume 64, number 3 (august 2022)

A Multifidelity Function-on-Function Model Applied to an Abdominal Aortic Aneurysm P. 279-290

Christoph Striegel, Jonas Biehler, Wolfgang A. Wall & Göran Kauermann

Abstract

In this work, we predict the outcomes of high fidelity multivariate computer simulations from low fidelity counterparts using function-to-function regression. The high fidelity simulation takes place on a high definition mesh, while its low fidelity counterpart takes place on a coarsened and truncated mesh. We showcase our approach by applying it to a complex finite element simulation of an abdominal aortic aneurysm which provides the displacement field of a blood vessel under pressure. In order to link the two multidimensional outcomes we compress them and then fit a function-to-function regression model. The data are high dimensional but of low sample size, meaning that only a few simulations are available, while the output of both low and high fidelity simulations is in the order of several thousands. To match this specific condition our compression method assumes a Gaussian Markov random field that takes the finite element geometry into account and only needs little data. In order to solve the function-to-function regression model we construct an appropriate prior with a shrinkage parameter which follows naturally from a Bayesian view of the Karhunen–Loève decomposition. Our model enables real multivariate predictions on the complete grid instead of resorting to the outcome of specific points.

Fast and Exact Leave-One-Out Analysis of Large-Margin Classifiers P. 291-298

Boxiang Wang & Hui Zou

Abstract

Motivated by the Golub–Heath–Wahba formula for ridge regression, we first present a new leave-one-out lemma for the kernel support vector machines (SVM) and related large-margin classifiers. We then use the lemma to design a novel and efficient algorithm, named “magicsvm,” for training the kernel SVM and related large-margin classifiers and computing the exact leave-one-out cross-validation error. By “magicsvm,” the computational cost of leave-one-out analysis is of the same order of fitting a single SVM on the training data. We show that “magicsvm” is much faster than the state-of-the-art SVM solvers based on extensive simulations and benchmark examples. The same idea is also used to boost the computation speed of the V-fold cross-validation of the kernel classifiers.

A Gaussian Process Emulator Based Approach for Bayesian Calibration of a Functional Input P. 299-311

Zhaohui Li & Matthias Hwai Yong Tan

Abstract

Bayesian calibration of a functional input/parameter to a time-consuming simulator based on a Gaussian process (GP) emulator involves two challenges that distinguish it from other parameter calibration problems. First, one needs to specify a flexible stochastic process prior for the input, and reduce it to a tractable number of random variables. Second, a sequential experiment design criterion that decreases the effect of emulator prediction uncertainty on calibration results is needed and the criterion should be scalable for high-dimensional input and output. In this article, we address these two issues. For the first issue, we employ a GP with a prior density for its correlation parameter as prior for the functional input, and the Karhunen-Loève (KL) expansion of this non-Gaussian stochastic process to reduce its dimension. We show that this prior gives far more robust inference results than a GP with a fixed correlation parameter. For the second issue, we propose the weighted prediction variance (WPV) criterion (with posterior density of the calibration parameter as weight) and prove the consistency of the sequence of emulator-based likelihoods given by the criterion. The proposed method is illustrated with examples on hydraulic transmissivity estimation for groundwater models.

Data-Driven Determination of the Number of Jumps in Regression Curves P. 312-322

Guanghui Wang, Changliang Zou & Peihua Qiu

Abstract

In nonparametric regression with jump discontinuities, one major challenge is to determine the number of jumps in a regression curve. Most existing methods to solve that problem are based on either a sequence of hypothesis tests or model selection, by introducing some extra tuning parameters that may not be easy to determine in practice. This article aims to develop a data-driven new methodology for determining the number of jumps, using an order-preserved sample-splitting strategy together with a cross-validation-based criterion. Statistical consistency of the determined number of jumps by our proposed method is established. More interestingly, the proposed method allows us to move beyond just point estimation, and it can quantify uncertainty of the proposed estimate. The key idea behind our method is the construction of a series of statistics with marginal symmetry property and this property can be used for choosing a data-driven threshold to control the false discovery rate of our method. The proposed method is computationally efficient. Numerical experiments indicate that it has a reliable performance in finite-sample cases. An R package jra is developed to implement the proposed method.

Reliable Post-Signal Fault Diagnosis for Correlated High-Dimensional Data Streams P. 323-334

Dongdong Xiang, Peihua Qiu, Dezhi Wang & Wendong Li

Abstract

Rapid advance of sensor technology is facilitating the collection of high-dimensional data streams (HDS). Apart from real-time detection of potential out-of-control (OC) patterns, post-signal fault diagnosis of HDS is becoming increasingly important in the filed of statistical process control to isolate abnormal data streams. The major limitations of the existing methods on that topic include (i) they cannot achieve reliable diagnostic results in the sense that their performance is highly variable, and (ii) the informative correlation among different streams is often neglected by them. This article elaborates the problem of reliable fault diagnosis for monitoring correlated HDS using the large-scale multiple testing. Under the framework of hidden Markov model dependence, new diagnostic procedures are proposed, which can control the missed discovery exceedance (MDX) at a desired level. Extensive numerical studies along with some theoretical results show that the proposed procedures can control MDX properly, leading to diagnostics with high reliability and efficiency. Also, their diagnostic performance can be improved significantly by exploiting the dependence among different data streams, which is especially appealing in practice for identifying clustered OC streams.

Functional PCA With Covariate-Dependent Mean and Covariance Structure P. 335-345

Fei Ding, Shiyuan He, David E. Jones & Jianhua Z. Huang

Abstract

Incorporating covariates into functional principal component analysis (PCA) can substantially improve the representation efficiency of the principal components and predictive performance. However, many existing functional

PCA methods do not make use of covariates, and those that do often have high computational cost or make overly simplistic assumptions that are violated in practice. In this article, we propose a new framework, called covariate-dependent functional principal component analysis (CD-FPCA), in which both the mean and covariance structure depend on covariates. We propose a corresponding estimation algorithm, which makes use of spline basis representations and roughness penalties, and is substantially more computationally efficient than competing approaches of adequate estimation and prediction accuracy. A key aspect of our work is our novel approach for modeling the covariance function and ensuring that it is symmetric positive semidefinite. We demonstrate the advantages of our methodology through a simulation study and an astronomical data analysis.

Spectral Clustering on Spherical Coordinates Under the Degree-Corrected Stochastic Blockmodel P. 346-357

Francesco Sanna Passino, Nicholas A. Heard & Patrick Rubin-Delanchy

Abstract

Spectral clustering is a popular method for community detection in network graphs: starting from a matrix representation of the graph, the nodes are clustered on a low-dimensional projection obtained from a truncated spectral decomposition of the matrix. Estimating correctly the number of communities and the dimension of the reduced latent space is critical for good performance of spectral clustering algorithms. Furthermore, many real-world graphs, such as enterprise computer networks studied in cyber-security applications, often display heterogeneous within-community degree distributions. Such heterogeneous degree distributions are usually not well captured by standard spectral clustering algorithms. In this article, a novel spectral clustering algorithm is proposed for community detection under the degree-corrected stochastic blockmodel. The proposed method is based on a transformation of the spectral embedding to spherical coordinates, and a novel modeling assumption in the transformed space. The method allows for simultaneous and automated selection of the number of communities and the latent dimension for spectral embeddings of graphs with uneven node degrees. Results show improved performance over competing methods in representing computer networks.

Locally Optimal Design for A/B Tests in the Presence of Covariates and Network Dependence P. 358-369

Qiong Zhang & Lulu Kang

Abstract

A/B test, a simple type of controlled experiment, refers to the statistical procedure of experimenting to compare two treatments applied to test subjects. For example, many IT companies frequently conduct A/B tests on their users who are connected and form social networks. Often, the users' responses could be related to the network connection. In this article, we assume that the users, or the test subjects of the experiments, are connected on an undirected network, and the responses of two connected users are correlated. We include the treatment assignment, covariate features, and network connection in a conditional autoregressive model. Based on this model, we propose a design criterion that measures the variance of the estimated treatment effect and allocate the treatment settings to the test subjects by minimizing the criterion. Since the design criterion depends on an unknown network correlation parameter, we adopt the locally optimal design method and develop a hybrid optimization approach to obtain the optimal design. Through synthetic and real social network examples, we demonstrate the value of including network dependence in designing A/B experiments and validate that the proposed locally optimal design is robust to the choices of parameters. Supplementary materials for this article are available online.

A Statistical Approach to Surface Metrology for 3D-Printed Stainless Steel P. 370-383

Chris J. Oates, Wilfrid S. Kendall & Liam Fleming

Abstract

The improvement of sensing technology enables features of process variables to be collected during the fabrication of products. This article develops an automatic tool for process feature rankings based on these data. Based on the sensing data characteristics and the need of manufacturing system analysis, we propose two rules of the feature ranking scheme: assessing general dependency between each individual process feature and the quality variable, and satisfying a diversity rule. Specifically, we propose a feature ranking scheme based on the sparse distance correlation (SpaDC) that satisfies these two rules. Theoretical properties of the proposed algorithm are investigated. Simulation studies and two real-case studies from semiconductor manufacturing applications demonstrate that the SpaDC method ranks the features effectively given these two ranking rules.

Ranking Features to Promote Diversity: An Approach Based on Sparse Distance Correlation

P. 384-395

Andi Wang, Juan Du, Xi Zhang & Jianjun Shi

Abstract

The improvement of sensing technology enables features of process variables to be collected during the fabrication of products. This article develops an automatic tool for process feature rankings based on these data. Based on the sensing data characteristics and the need of manufacturing system analysis, we propose two rules of the feature ranking scheme: assessing general dependency between each individual process feature and the quality variable, and satisfying a diversity rule. Specifically, we propose a feature ranking scheme based on the sparse distance correlation (SpaDC) that satisfies these two rules. Theoretical properties of the proposed algorithm are investigated. Simulation studies and two real-case studies from semiconductor manufacturing applications demonstrate that the SpaDC method ranks the features effectively given these two ranking rules.

Density Regression with Conditional Support Points

P. 396-408

Yunlu Chen & Nan Zhang

Abstract

Density regression characterizes the conditional density of the response variable given the covariates, and provides much more information than the commonly used conditional mean or quantile regression. However, it is often computationally prohibitive in applications with massive datasets, especially when there are multiple covariates. In this article, we develop a new data reduction approach for the density regression problem using conditional support points. After obtaining the representative data, we exploit the penalized likelihood method as the downstream estimation strategy. Based on the connections among the continuous ranked probability score, the energy distance, the L_2 discrepancy and the symmetrized Kullback–Leibler distance, we investigate the distributional convergence of the representative points and establish the rate of convergence of the density regression estimator. The usefulness of the methodology is illustrated by modeling the conditional distribution of power output given multivariate environmental factors using a large scale wind turbine dataset.

A New Sparse-Learning Model for Maximum Gap Reduction of Composite Fuselage Assembly

P. 409-418

Juan Du, Shanshan Cao, Jeffrey H. Hunt, Xiaoming Huo & Jianjun Shi

Abstract

Natural dimensional variabilities of incoming fuselages affect the assembly speed and quality of fuselage joins in composite fuselage assembly processes. Shape control is critical to ensure the quality of composite fuselage assembly. In current practice, the structures are adjusted to the design shape in terms of the l2 loss for further assembly without considering the existing dimensional gap between two structures. Such practice has two limitations: (a) controlling each fuselage to the design shape may not be the optimal shape control strategy in terms of a pair of incoming fuselages with different incoming dimensions; (b) the maximum gap is the key concern during the fuselage assembly process, so the l∞ loss of gap after control ought to be considered. This article proposes an optimal shape control methodology via the l∞ loss for the composite fuselage assembly process by considering the existing dimensional gap between the incoming pair of fuselages. On the other hand, due to the limitation on the number of available actuators in practice, we face an important problem of finding the best locations for the actuators among many potential locations, which makes the problem a sparse estimation problem. We are the first to solve the optimal shape control in the fuselage assembly process using the l∞ model under the framework of sparse estimation, where we use the l1 penalty to control the sparsity of the resulting estimator. From the statistical point of view, this can be formulated as the l∞ loss based linear regression, and under some standard assumptions, such as the restricted eigenvalue (RE) conditions, and the light-tailed noise, the nonasymptotic estimation error of the l1 regularized l∞ linear model is derived, which meets the upper-bound in the existing literature. Compared to the current practice, the case study shows that our proposed method significantly reduces the maximum gap between two fuselages after shape adjustments.