Technometrics

A Journal of Statistics for the Physical, Chemical and Engineering Sciences

ASQ and American Statistical Association

Technometrics, ISSN 0040-1706
Volume 60, number 4 (november 2018)

Variable Selection for the Prediction of C[0,1]-Valued Autoregressive Processes using Reproducing Kernel Hilbert Spaces P. 139-153

Beatriz Bueno-Larraz & Johannes Klepsch

Abstract

A model for the prediction of functional time series is introduced, where observations are assumed to be continuous random functions. We model the dependence of the data with a nonstandard autoregressive structure, motivated in terms of the reproducing kernel Hilbert space (RKHS) generated by the auto-covariance function of the data. The new approach helps to find relevant points of the curves in terms of prediction accuracy. This dimension reduction technique is particularly useful for applications, since the results are usually directly interpretable in terms of the original curves. An empirical study involving real and simulated data is included, which generates competitive results.

Assessing Tuning Parameter Selection Variability in Penalized Regression P. 154-164

Wenhao Hu, Eric B. Laber, Clay Barker & Leonard A. Stefanski

Abstract

Penalized regression methods that perform simultaneous model selection and estimation are ubiquitous in statistical modeling. The use of such methods is often unavoidable as manual inspection of all possible models quickly becomes intractable when there are more than a handful of predictors. However, automated methods usually fail to incorporate domain-knowledge, exploratory analyses, or other factors that might guide a more interactive model-building approach. A hybrid approach is to use penalized regression to identify a set of candidate models and then to use interactive model-building to examine this candidate set more closely. To identify a set of candidate models, we derive point and interval estimators of the probability that each model along a solution path will minimize a given model selection criterion, for example, Akaike information criterion, Bayesian information criterion (AIC, BIC), etc., conditional on the observed solution path. Then models with a high probability of selection are considered for further examination. Thus, the proposed methodology attempts to strike a balance between algorithmic modeling approaches that are computationally efficient but fail to incorporate expert knowledge, and interactive modeling approaches that are labo

On Data Integration Problems With Manifolds P. 165-175

Mark V. Culp, Kenneth J. Ryan, Prithish Banerjee & Michael Morehead

Abstract

This article focuses on data integration problems where the predictor variables for some response variable partition into known subsets. This type of data is often referred to as multi-view data, and each subset of the predictors is called a data view. Accounting for data views can add practical value in terms of both interpretation and predictive performance. Many existing approaches for multi-view data rely on view-agreement principles, strong smoothness assumptions, or regularization penalties. The former approaches can be sensitive to modest noise in the response or predictor variables, while the latter approach is linear and can usually be out-performed. We develop semiparametric data integration methods to span key tradeoffs including the bias-variance tradeoff on prediction error, the possibility that the data may be fully viewed with no appreciable view relationships, and the use of sparse anchor point methods to detect and use manifolds (i.e., possibly nonelliptical structures) within views if they enhance performance. Theoretical results help justify the new technique, and its effectiveness and computational feasibility are demonstrated empirically. This new semiparametric methodology is available for public use through the supplemental R package mvltools.

Convex Bidirectional Large Margin Classifiers

P. 176-186

Zhengling Qi & Yufeng Liu

Abstract

Classification problems are commonly seen in practice. In this article, we aim to develop classifiers that can enjoy great interpretability as linear classifiers, and at the same time have model flexibility as nonlinear classifiers. We propose convex bidirectional large margin classifiers to fill the gap between linear and general nonlinear classifiers for high-dimensional data. Our method provides a new data visualization tool for classification of high-dimensional data. The obtained bilinear projection structure makes the proposed classifier very interpretable. Additional shrinkage to approximate variable selection is also considered. Through analysis of simulated and real data in high-dimensional settings, our method is shown to have superior prediction performance and interpretability when there are potential subpopulations in the data.

Inference for Errors-in-Variables Models in the Presence of Systematic Errors with an Application to a Satellite Remote Sensing Campaign

P. 187-201

Bohai Zhang, Noel Cressie & Debra Wunch

Abstract

Motivated by a satellite remote sensing mission, this article proposes a multivariable errors-in-variables (EIV) regression model with heteroscedastic errors for relating the satellite data products to similar products from a well-characterized but globally sparse ground-based dataset. In the remote sensing setting, the regression model is used to estimate the global divisor for the satellite data. The error structure of the proposed EIV model comprises two components: A random-error component whose variance is inversely proportional to sample size of underlying individual observations which are aggregated to obtain the regression data, and a systematic-error component whose variance remains the same as the underlying sample size increases. In this article, we discuss parameter identifiability for the proposed model and obtain estimates from two-stage parameter estimation. We illustrate our proposed procedure through both simulation studies and an application to validating measurements of atmospheric column-averaged CO_2 from NASA's Orbiting Carbon Observatory-2 (OCO-2) satellite. The validation uses coincident target-mode OCO-2 data that are temporally and spatially sparse and ground-based measurements from the Total Carbon Column Observing Network (TCCON) that are spatially sparse but more accurate.

Gaussian Process Modeling Using the Principle of Superposition

P. 202-218

Matthias H. Y. Tan & Guilin Li

Abstract

Partial differential equation (PDE) models of physical systems with initial and boundary conditions are often solved numerically via a computer code called the simulator. To study the dependence of the solution on a functional input, the input is expressed as a linear combination of a finite number of basis functions, and the coefficients of the bases are varied. In such studies, Gaussian process (GP) emulators can be constructed to reduce the amount of simulations required from time-consuming simulators. For linear initial-boundary value problems (IBVPs) with functional inputs as additive terms in the PDE, initial conditions, or boundary conditions, the IBVP solution is theoretically a linear function of the coefficients conditional on all other inputs, which is a result called the principle of superposition. Since numerical errors cause deviation from linearity and nonlinear IBVPs are widely solved in practice, we generalize the result to account for nonlinearity. Based on this generalized result, we propose mean and covariance functions for building GP emulators that capture the approximate conditional linear effect of the coefficients. Numerical simulations demonstrate the substantial improvements in prediction performance achieved with the proposed emulator. Matlab codes for reproducing the results in this article are available in the online supplement.

Constructing Two-Level Designs by Concatenation of Strength-3 Orthogonal Arrays

P. 219-232

Alan R. Vazquez, Peter Goos & Eric D. Schoen

Abstract

Two-level orthogonal arrays of N runs, k factors, and a strength of 3 provide suitable fractional factorial designs in situations where many of the main effects are expected to be active, as well as some two-factor interactions. If they consist of $N/2$ mirror image pairs, these designs are fold-over designs. They are called even and provide at most $N/2 - 1$ degrees of freedom to estimate interactions. For $k < N/3$ factors, there exist strength-3 designs that are not fold-over designs. They are called even-odd designs and they provide many more degrees of freedom to estimate interactions. For $N \leqslant 48$, attractive even-odd designs can be extracted from complete catalogs of strength-3 orthogonal arrays. However, for larger run sizes, no complete catalogs exist. To construct even-odd designs with $N > 48$, we develop an algorithm for an optimal concatenation of strength-3 designs involving $N/2$ runs. Our approach involves column permutations of one of the concatenated designs, as well as sign switches of the elements of one or more columns of that design. We illustrate the potential of the algorithm by generating two-level even-odd designs with 64 and 128 runs involving up to 33 factors, because this allows a comparison with benchmark designs from the literature. With a few exceptions, our even-odd designs outperform or are competitive with the benchmark designs in terms of the aliasing of two-factor interactions and in terms of the available degrees of freedom to estimate two-factor interactions.

Bayesian Analysis of Accumulated Damage Models in Lumber Reliability

P. 233-245

Chun-Hao Yang, James V. Zidek & Samuel W. K. Wong

Abstract

Wood products that are subjected to sustained stress over a period of long duration may weaken, and this effect must be considered in models for the long-term reliability of lumber. The damage accumulation approach has been widely used for this purpose to set engineering standards. In this article, we revisit an accumulated damage model and propose a Bayesian framework for analysis. For parameter estimation and uncertainty quantification, we adopt approximation Bayesian computation (ABC) techniques to handle the complexities of the model. We demonstrate the effectiveness of our approach using both simulated and real data, and apply our fitted model to analyze long-term lumber reliability under a stochastic live loading scenario.

Material Degradation Modeling and Failure Prediction Using Microstructure Images

P. 246-258

Wujun Si, Qingyu Yang & Xin Wu

Abstract

Degradation data, frequently along with low-dimensional covariate information such as scalar-type covariates, are widely used for asset reliability analysis. Recently, many high-dimensional covariates such as functional and image covariates have emerged with advances in sensor technology, containing richer information that can be used for degradation assessment. In this article, motivated by a physical effect that microstructures of dual-phase advanced high strength steel strongly influence steel degradation, we propose a two-stage material degradation model using the material microstructure image as a covariate. In Stage 1, we show that the microstructure image covariate can be reduced to a functional covariate while statistical properties of the image are preserved up to the second order. In Stage 2, a novel functional covariate degradation model is proposed, based on which the time-to-failure distribution in terms of degradation level passages is derived. A penalized least squares estimation method is developed to obtain the closed-form point estimator of model parameters. Analytical inferences on interval estimation of the model parameters, the mean degradation levels, and the distribution of the time-to-failure are also developed. Simulation studies are implemented to validate the developed methods. Physical experiments on dual-phase advanced high strength steel are designed and conducted to demonstrate the proposed model. The results show that a significant improvement is achieved for material failure prediction by using material microstructure images compared with multiple benchmark models.

Efficient Integration of Sufficient Dimension Reduction and Prediction in Discriminant Analysis

P. 259-272

Xin Zhang & Qing Mai

Abstract

Sufficient dimension reduction (SDR) methods are popular model-free tools for preprocessing and data visualization in regression problems where the number of variables is large. Unfortunately, reduce-and-classify approaches in discriminant analysis usually cannot guarantee improvement in classification accuracy, mainly due to the different nature of the two stages. On the other hand, envelope methods construct targeted dimension reduction subspaces that achieve dimension reduction and improve parameter estimation efficiency at the same time. However, little is known about how to construct envelopes in discriminant analysis models. In this article, we introduce the notion of the envelope discriminant subspace (ENDS) as a natural inferential and estimative object in discriminant analysis that incorporates these considerations. We develop the ENDS estimators that simultaneously achieve sufficient dimension reduction and classification. Consistency and asymptotic normality of the ENDS estimators are established, where we carefully examine the asymptotic efficiency gain under the classical linear and quadratic discriminant analysis models. Simulations and real data examples show superb performance of the proposed method.