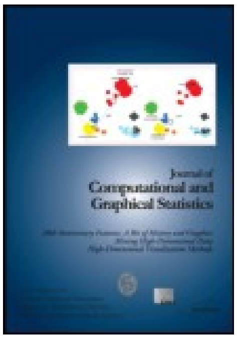

Journal of computational and graphical statistics, ISSN 1061-8600
Volume 26, number 2 (2017)

Clusters Beat Trend!? Testing Feature Hierarchy in Statistical Graphics

P. 231-242

Susan VanderPlas & Heike Hofmann

Abstract

Graphics are very effective for communicating numerical information quickly and efficiently, but many of the design choices we make are based on subjective measures, such as personal taste or conventions of the discipline rather than objective criteria. We briefly introduce perceptual principles such as preattentive features and gestalt heuristics, and then discuss the design and results of a factorial experiment examining the effect of plot aesthetics such as color and trend lines on participants' assessment of ambiguous data displays. The quantitative and qualitative experimental results strongly suggest that plot aesthetics have a significant impact on the perception of important features in data displays. Supplementary materials for this article are available online.

Path Boxplots: A Method for Characterizing Uncertainty in Path Ensembles on a Graph

P. 243-252

Mukund Raj, Mahsa Mirzargar, Robert Ricci, Robert M. Kirby & Ross T. Whitaker

Abstract

Graphs are powerful and versatile data structures that can be used to represent a wide range of different types of information. In this article, we introduce a method to analyze and then visualize an important class of data described over a graph—namely, ensembles of paths. Analysis of such path ensembles is useful in a variety of applications, in diverse fields such as transportation, computer networks, and molecular dynamics. The proposed method generalizes the concept of *band depth* to an ensemble of paths on a graph, which provides a center-outward ordering on the paths. This ordering is, in turn, used to construct a generalization of the conventional boxplot or whisker plot, called a *path boxplot*, which applies to paths on a graph. The utility of path boxplot is demonstrated for several examples of path ensembles including paths defined over computer networks and roads. Supplementary materials for this article are available online.

The Torgegram for Fluvial Variography: Characterizing Spatial Dependence on Stream Networks

P. 253-264

Dale L. Zimmerman & Jay M. Ver Hoef

Abstract

We introduce a graphical diagnostic called the Torgegram for characterizing spatial dependence among observations of a variable on a stream network. The Torgegram consists of four component empirical semivariograms, each one corresponding to a particular combination of flow-connectedness within the network and model type (tail-up/tail-down). We show how an overall strategy for fluvial variography can be based on a careful examination of the Torgegram. An analysis of water temperature data from a stream network within the Columbia River basin of the northwest United States illustrates the diagnostic value of the Torgegram as well as its limitations. Additional uses and extensions of the Torgegram are discussed.

Semiparametric Bayesian Regression via Potts Model

P. 265-274

Alejandro Murua & Fernando A. Quintana

Abstract

We consider Bayesian nonparametric regression through random partition models. Our approach involves the construction of a covariate-dependent prior distribution on partitions of individuals. Our goal is to use covariate information to improve predictive inference. To do so, we propose a prior on partitions based on the Potts clustering model associated with the observed covariates. This drives by covariate proximity both the formation of clusters, and the prior predictive distribution. The resulting prior model is flexible enough to support many different types of likelihood models. We focus the discussion on nonparametric regression. Implementation details are discussed for the specific case of multivariate multiple linear regression. The proposed model performs well in terms of model fitting and prediction when compared to other alternative nonparametric regression approaches. We illustrate the methodology with an application to the health status of nations at the turn of the 21st century. Supplementary materials are available online.

Regression Adjustment for Noncrossing Bayesian Quantile Regression

P. 275-284

T. Rodrigues & Y. Fan

Abstract

A two-stage approach is proposed to overcome the problem in quantile regression, where separately fitted curves for several quantiles may cross. The standard Bayesian quantile regression model is applied in the first stage, followed by a Gaussian process regression adjustment, which monotonizes the quantile function while borrowing strength from nearby quantiles. The two-stage approach is computationally efficient, and more general than existing techniques. The method is shown to be competitive with alternative approaches via its performance in simulated examples. Supplementary materials for the article are available online.

Identifying Mixtures of Mixtures Using Bayesian Estimation

P. 285-295

Gertraud Malsiner-Walli, Sylvia Frühwirth-Schnatter & Bettina Grün

Abstract

The use of a finite mixture of normal distributions in model-based clustering allows us to capture non-Gaussian data clusters. However, identifying the clusters from the normal components is challenging and in general either achieved by imposing constraints on the model or by using post-processing procedures. Within the Bayesian framework, we propose a different approach based on sparse finite mixtures to achieve identifiability. We specify a hierarchical prior, where the hyperparameters are carefully selected such that they are reflective of the cluster structure aimed at. In addition, this prior allows us to estimate the model using standard MCMC sampling methods. In combination with a post-processing approach which resolves the label switching issue and results in an identified model, our approach allows us to simultaneously (1) determine the number of clusters, (2) flexibly approximate the cluster distributions in a semiparametric way using finite mixtures of normals and (3) identify cluster-specific parameters and classify observations. The proposed approach is illustrated in two simulation studies and on benchmark datasets. Supplementary materials for this article are available online.

Combining Functional Data Registration and Factor Analysis

P. 296-305

Cecilia Earls & Giles Hooker

Abstract

We extend the definition of functional data registration to encompass a larger class of registration models. In contrast to traditional registration models, we allow for registered functions that have more than one primary direction of variation. The proposed Bayesian hierarchical model simultaneously registers the observed functions and estimates the two primary factors that characterize variation in the registered functions. Each registered function is assumed to be predominantly composed of a linear combination of these two primary factors, and the function-specific weights for each observation are estimated within the registration model. We show how these estimated weights can easily be used to classify functions after registration using both simulated data and a juggling dataset. Supplementary materials for this article are available online.

Locally Sparse Estimator for Functional Linear Regression Models P. 306-318

Zhenhua Lin, Jiguo Cao, Liangliang Wang & Haonan Wang

Abstract

A new locally sparse (i.e., zero on some subregions) estimator for coefficient functions in functional linear regression models is developed based on a novel functional regularization technique called “fSCAD.” The nice shrinkage property of fSCAD allows the proposed estimator to locate null subregions of coefficient functions without over shrinking nonzero values of coefficient functions. Additionally, a roughness penalty is incorporated to control the roughness of the locally sparse estimator. Our method is theoretically sounder and computationally simpler than existing methods. Asymptotic analysis reveals that the proposed estimator is consistent and can identify null subregions with probability tending to one. Extensive simulations confirm the theoretical analysis and show excellent numerical performance of the proposed method. Practical merit of locally sparse modeling is demonstrated by two real applications. Supplemental materials for the article are available online.

Sparse Functional Dynamical Models—A Big Data Approach P. 319-329

Ela Sienkiewicz, Dong Song, F. Jay Breidt & Haonan Wang

Abstract

Nonlinear dynamical systems are encountered in many areas of social science, natural science, and engineering, and are of particular interest for complex biological processes like the spiking activity of neural ensembles in the brain. To describe such spiking activity, we adapt the Volterra series expansion of an analytic function to account for the point-process nature of multiple inputs and a single output (MISO) in a neural ensemble. Our model describes the transformed spiking probability for the output as the sum of kernel-weighted integrals of the inputs. The kernel functions need to be identified and estimated, and both local sparsity (kernel functions may be zero on part of their support) and global sparsity (some kernel functions may be identically zero) are of interest. The kernel functions are approximated by B-splines and a penalized likelihood-based approach is proposed for estimation. Even for moderately complex brain functionality, the identification and estimation of this sparse functional dynamical model poses major computational challenges, which we address with big data techniques that can be implemented on a single, multi-core server. The performance of the proposed method is demonstrated using neural recordings from the hippocampus of a rat during open field tasks. Supplementary materials for this article are available online.

Grouped Functional Time Series Forecasting: An Application to Age-Specific Mortality Rates P. 330-343

Han Lin Shang & Rob J. Hyndman

Abstract

Age-specific mortality rates are often disaggregated by different attributes, such as sex, state, and ethnicity. Forecasting age-specific mortality rates at the national and sub-national levels plays an important role in developing social policy. However, independent forecasts at the sub-national levels may not add up to the forecasts at the national level. To address this issue, we consider reconciling forecasts of age-specific mortality rates, extending the methods of Hyndman et al. in 2011 Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., and Shang, H. L. (2011), “Optimal Combination Forecasts for Hierarchical Time Series,” *Computational Statistics and Data Analysis*, 55, 2579–2589. [Crossref], [Web of Science ®], [Google Scholar] to functional time series, where age is considered as a continuum. The grouped functional time series methods are used to produce point forecasts of mortality rates that are aggregated appropriately across different disaggregation factors. For evaluating forecast uncertainty, we propose a bootstrap method for reconciling interval forecasts. Using the regional age-specific mortality rates in Japan, obtained from the Japanese Mortality Database, we investigate the one- to ten-step-ahead point and interval forecast accuracies between the independent and grouped functional time series forecasting methods. The proposed methods are shown to be useful for reconciling forecasts of age-specific mortality rates at the national and sub-national levels. They also enjoy improved forecast accuracy averaged over different disaggregation factors. Supplementary materials for the article are available online.

A Semiparametric Two-Sample Hypothesis Testing Problem for Random Graphs P. 344-354

Minh Tang, Avanti Athreya, Daniel L. Sussman, Vince Lyzinski, Youngser Park & Carey E. Priebe

Abstract

Two-sample hypothesis testing for random graphs arises naturally in neuroscience, social networks, and machine learning. In this article, we consider a semiparametric problem of two-sample hypothesis testing for a class of latent position random graphs. We formulate a notion of consistency in this context and propose a valid test for the hypothesis that two finite-dimensional random dot product graphs on a common vertex set have the same generating latent positions or have generating latent positions that are scaled or diagonal transformations of one another. Our test statistic is a function of a spectral decomposition of the adjacency matrix for each graph and our test procedure is consistent across a broad range of alternatives. We apply our test procedure to real biological data: in a test-retest dataset of neural connectome graphs, we are able to distinguish between scans from different subjects; and in the *C. elegans* connectome, we are able to distinguish between chemical and electrical networks. The latter example is a concrete demonstration that our test can have power even for small-sample sizes. We conclude by discussing the relationship between our test procedure and generalized likelihood ratio tests. Supplementary materials for this article are available online.

Sparse Steinian Covariance Estimation P. 355-366

Brett Naul & Jonathan Taylor

Abstract

We consider a new method for sparse covariance matrix estimation which is motivated by previous results for the so-called Stein-type estimators. Stein proposed a method for regularizing the sample covariance matrix by shrinking together the eigenvalues; the amount of shrinkage is chosen to minimize an unbiased estimate of the risk (UBEOR) under the entropy loss function. The resulting estimator has been shown in simulations to yield significant risk reductions over the maximum likelihood estimator. Our method extends the UBEOR minimization problem by adding an ℓ_1 penalty on the entries of the estimated covariance matrix, which encourages a sparse estimate. For a multivariate Gaussian distribution, zeros in the covariance matrix correspond to marginal independences between variables. Unlike the ℓ_1-penalized Gaussian likelihood function, our penalized UBEOR objective is convex and can be minimized via a simple block coordinate descent procedure. We demonstrate via numerical simulations and an analysis of microarray data from breast cancer patients that our proposed method generally outperforms other methods for sparse covariance matrix estimation and can be computed efficiently even in high dimensions.

High-Dimensional Mixed Graphical Models P. 367-378

Jie Cheng, Tianxi Li, Elizaveta Levina & Ji Zhu

Abstract

While graphical models for continuous data (Gaussian graphical models) and discrete data (Ising models) have been extensively studied, there is little work on graphical models for datasets with both continuous and discrete variables (mixed data), which are common in many scientific applications. We propose a novel graphical model for mixed data, which is simple enough to be suitable for high-dimensional data, yet flexible enough to represent all possible graph structures. We develop a computationally efficient regression-based algorithm for fitting the model by focusing on the conditional log-likelihood of each variable given the rest. The parameters have a natural group structure, and sparsity in the fitted graph is attained by incorporating a group lasso penalty, approximated by a weighted lasso penalty for computational efficiency. We demonstrate the effectiveness of our method through an extensive simulation study and apply it to a music annotation dataset (CAL500), obtaining a sparse and interpretable graphical model relating the continuous features of the audio signal to binary variables such as genre, emotions, and usage associated with particular songs. While we focus on binary discrete variables for the main presentation, we also show that the proposed methodology can be easily extended to general discrete variables.

Penalized Versus Constrained Generalized Eigenvalue Problems P. 379-387

Irina Gaynanova, James G. Booth & Martin T. Wells

Abstract

We investigate the difference between using an ℓ_1 penalty versus an ℓ_1 constraint in generalized eigenvalue problems arising in multivariate analysis. Our main finding is that the ℓ_1 penalty may fail to provide very sparse solutions; a severe disadvantage for variable selection that can be remedied by using an ℓ_1 constraint. Our claims are supported both by empirical evidence and theoretical analysis. Finally, we illustrate the advantages of the ℓ_1 constraint in the context of discriminant analysis and principal component analysis. Supplementary materials for this article are available online.

Composite Likelihood Inference in a Discrete Latent Variable Model for Two-Way "Clustering-by-Segmentation" Problems P. 388-402

Francesco Bartolucci, Francesca Chiaromonte, Prabhani Kuruppumullage Don & Bruce G. Lindsay

Abstract

We consider a discrete latent variable model for two-way data arrays, which allows one to simultaneously produce clusters along one of the data dimensions (e.g., exchangeable observational units or features) and contiguous groups, or segments, along the other (e.g., consecutively ordered times or locations). The model relies on a hidden Markov structure but, given its complexity, cannot be estimated by full maximum likelihood. Therefore, we introduce a composite likelihood methodology based on considering different subsets of the data. The proposed approach is illustrated by simulation, and with an application to genomic data.

Programming With Models: Writing Statistical Algorithms for General Model Structures With NIMBLE P. 403-413

Perry de Valpine, Daniel Turek, Christopher J. Paciorek, Clifford Anderson-Bergman, Duncan Temple Lang & Rastislav Bodik

Abstract

We describe NIMBLE, a system for programming statistical algorithms for general model structures within R. NIMBLE is designed to meet three challenges: flexible model specification, a language for programming algorithms that can use different models, and a balance between high-level programmability and execution efficiency. For model specification, NIMBLE extends the BUGS language and creates model objects, which can manipulate variables, calculate log probability values, generate simulations, and query the relationships among variables. For algorithm programming, NIMBLE provides functions that operate with model objects using two stages of evaluation. The first stage allows specialization of a function to a particular model and/or nodes, such as creating a Metropolis-Hastings sampler for a particular block of nodes. The second stage allows repeated execution of computations using the results of the first stage. To achieve efficient second-stage computation, NIMBLE compiles models and functions via C++, using the Eigen library for linear algebra, and provides the user with an interface to compiled objects. The NIMBLE language represents a compilable domain-specific language (DSL) embedded within R. This article provides an overview of the design and rationale for NIMBLE along with illustrative examples including importance sampling, Markov chain Monte Carlo (MCMC) and Monte Carlo expectation maximization (MCEM). Supplementary materials for this article are available online.

Adaptive Sequential Monte Carlo for Multiple Changepoint Analysis P. 414-423

Nicholas A. Heard & Melissa J. M. Turcotte

Abstract

Process monitoring and control requires the detection of structural changes in a data stream in real time. This article introduces an efficient sequential Monte Carlo algorithm designed for learning unknown changepoints in continuous time. The method is intuitively simple: new changepoints for the latest window of data are proposed by conditioning only on data observed since the most recent estimated changepoint, as these observations carry most of the information about the current state of the process. The proposed method shows improved performance over the current state of the art. Another advantage of the proposed algorithm is that it can be made adaptive, varying the number of particles according to the apparent local complexity of the target changepoint probability distribution. This saves valuable computing time when changes in the changepoint distribution are negligible, and enables rebalancing of the importance weights of existing particles when a significant change in the target distribution is encountered. The plain and adaptive versions of the method are illustrated using the canonical continuous time changepoint problem of inferring the intensity of an inhomogeneous Poisson process, although the method is generally applicable to any changepoint problem. Performance is demonstrated using both conjugate and nonconjugate Bayesian models for the intensity. Appendices to the article are available online, illustrating the method on other models and applications.

Efficient Computation of Bayesian Optimal Discriminating Designs P. 424-433

Holger Dette, Roman Guchenko & Viatcheslav B. Melas

Abstract

An efficient algorithm for the determination of Bayesian optimal discriminating designs for competing regression models is developed, where the main focus is on models with general distributional assumptions beyond the "classical" case of normally distributed homoscedastic errors. For this purpose, we consider a Bayesian version of the Kullback–Leibler (KL). Discretizing the prior distribution leads to local KL-optimal discriminating design problems for a large number of competing models. All currently available methods either require a large amount of computation time or fail to calculate the optimal discriminating design, because they can only deal efficiently with a few model comparisons. In this article, we develop a new algorithm for the determination of Bayesian optimal discriminating designs with respect to the Kullback–Leibler criterion. It is demonstrated that the new algorithm is able to calculate the optimal discriminating designs with reasonable accuracy and computational time in situations where all currently available procedures are either slow or fail.

Adaptive, Delayed-Acceptance MCMC for Targets With Expensive Likelihoods P. 434-444

Chris Sherlock, Andrew Golightly & Daniel A. Henderson

Abstract

When conducting Bayesian inference, delayed-acceptance (DA) Metropolis–Hastings (MH) algorithms and DA pseudo-marginal MH algorithms can be applied when it is computationally expensive to calculate the true posterior or an unbiased estimate thereof, but a computationally cheap approximation is available. A first accept-reject stage is applied, with the cheap approximation substituted for the true posterior in the MH acceptance ratio. Only for those proposals that pass through the first stage is the computationally expensive true posterior (or unbiased estimate thereof) evaluated, with a second accept-reject stage ensuring that detailed balance is satisfied with respect to the intended true posterior. In some scenarios, there is no obvious computationally cheap approximation. A weighted average of previous evaluations of the computationally expensive posterior provides a generic approximation to the posterior. If only the k-nearest neighbors have nonzero weights then evaluation of the approximate posterior can be made computationally cheap provided that the points at which the posterior has been evaluated are stored in a multi-dimensional binary tree, known as a KD-tree. The contents of the KD-tree are potentially updated after every computationally intensive evaluation. The resulting adaptive, delayed-acceptance [pseudo-marginal] Metropolis–Hastings algorithm is justified both theoretically and empirically. Guidance on tuning parameters is provided and the methodology is applied to a discretely observed Markov jump process characterizing predator–prey interactions and an

ODE system describing the dynamics of an autoregulatory gene network. Supplementary material for this article is available online.

Divide-and-Conquer With Sequential Monte Carlo

P. 445-458

F. Lindsten, A. M. Johansen, C. A. Naesseth, B. Kirkpatrick, T. B. Schön, J. A. D. Aston & A. Bouchard-Côté

Abstract

We propose a novel class of Sequential Monte Carlo (SMC) algorithms, appropriate for inference in probabilistic graphical models. This class of algorithms adopts a divide-and-conquer approach based upon an auxiliary tree-structured decomposition of the model of interest, turning the overall inferential task into a collection of recursively solved subproblems. The proposed method is applicable to a broad class of probabilistic graphical models, *including* models with loops. Unlike a standard SMC sampler, the proposed divide-and-conquer SMC employs multiple independent populations of weighted particles, which are resampled, merged, and propagated as the method progresses. We illustrate empirically that this approach can outperform standard methods in terms of the accuracy of the posterior expectation and marginal likelihood approximations. Divide-and-conquer SMC also opens up novel parallel implementation options and the possibility of concentrating the computational effort on the most challenging subproblems. We demonstrate its performance on a Markov random field and on a hierarchical logistic regression problem. Supplementary materials including proofs and additional numerical results are available online.

FFT-Based Fast Computation of Multivariate Kernel Density Estimators With Unconstrained Bandwidth Matrices

P. 459-462

Artur Gramacki & Jarosław Gramacki

Abstract

The problem of fast computation of multivariate kernel density estimation (KDE) is still an open research problem. In our view, the existing solutions do not resolve this matter in a satisfactory way. One of the most elegant and efficient approach uses the fast Fourier transform. Unfortunately, the existing FFT-based solution suffers from a serious limitation, as it can accurately operate only with the constrained (i.e., diagonal) multivariate bandwidth matrices. In this article, we describe the problem and give a satisfactory solution. The proposed solution may be successfully used also in other research problems, for example, for the fast computation of the optimal bandwidth for KDE. Supplementary materials for this article are available online.

An Efficient Implementation of the EMICM Algorithm for the Interval Censored NPMLE

P. 463-467

Clifford Anderson-Bergman

Abstract

The EMICM algorithm is an established method for computing the interval-censored NPMLE, a generalization of the Kaplan Meier curves for interval censored data. The novel contribution in this work is an efficient implementation, allowing each iteration to be computed in linear time. Using simulated data, it is shown that this new implementation is significantly faster than alternative EMICM implementations or other competing algorithms, allowing for analyses of datasets orders of magnitude larger than previously available.