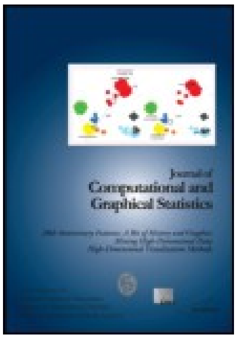

Journal of computational and graphical statistics, ISSN 1061-8600
Volume 25, number 1 (march 2016)

Efficient Implementations of the Generalized Lasso Dual Path Algorithm — P. 1-27

Taylor B. Arnold & Ryan J. Tibshirani

Abstract

We consider efficient implementations of the generalized lasso dual path algorithm given by Tibshirani and Taylor in 2011 Tibshirani, R.J., Taylor, J. (2011), The Solution Path of the Generalized Lasso, Annals of Statistics, 39, 1335–1371.[CrossRef], [Web of Science ®]. We first describe a generic approach that covers any penalty matrix *D* and any (full column rank) matrix *X* of predictor variables. We then describe fast implementations for the special cases of trend filtering problems, fused lasso problems, and sparse fused lasso problems, both with *X* = *I* and a general matrix *X*. These specialized implementations offer a considerable improvement over the generic implementation, both in terms of numerical stability and efficiency of the solution path computation. These algorithms are all available for use in the genlasso R package, which can be found in the CRAN repository.

Confidence Areas for Fixed-Effects PCA — P. 28-48

Julie Josse, Stefan Wager & François Husson

Abstract

Principal component analysis (PCA) is often used to visualize data when the rows and the columns are both of interest. In such a setting, there is a lack of inferential methods on the PCA output. We study the asymptotic variance of a fixed-effects model for PCA, and propose several approaches to assessing the variability of PCA estimates: a method based on a parametric bootstrap, a new cell-wise jackknife, as well as a computationally cheaper approximation to the jackknife. We visualize the confidence regions by Procrustes rotation. Using a simulation study, we compare the proposed methods and highlight the strengths and drawbacks of each method as we vary the number of rows, the number of columns, and the strength of the relationships between variables.

Case-Specific Random Forests — P. 49-65

Ruo Xu, Dan Nettleton & Daniel J. Nordman

Abstract

Random forest (RF) methodology is a nonparametric methodology for prediction problems. A standard way to use RFs includes generating a global RF to predict all test cases of interest. In this article, we propose growing different RFs specific to different test cases, namely case-specific random forests (CSRFs). In contrast to the bagging procedure in the building of standard RFs, the CSRF algorithm takes weighted bootstrap resamples to create individual trees, where we assign large weights to the training cases in close proximity to the test case of interest a priori. Tuning methods are discussed to avoid overfitting issues. Both simulation and real data examples show that the weighted bootstrap resampling used in CSRF construction can improve predictions for specific cases. We also propose a new case-specific variable importance (CSVI) measure as a way to compare the relative predictor variable importance for predicting a particular case. It is possible that the idea of building a predictor case-specifically can be generalized in other areas.

Merging Mixture Components for Clustering Through Pairwise Overlap

P. 66-90

Volodymyr Melnykov

Abstract

Finite mixture models are well known for their flexibility in modeling heterogeneity in data. Model-based clustering is an important application of mixture models, which assumes that each mixture component distribution can adequately model a particular group of data. Unfortunately, when more than one component is needed for each group, the appealing one-to-one correspondence between mixture components and groups of data is ruined and model-based clustering loses its attractive interpretation. Several remedies have been considered in literature. We discuss the most promising recent results obtained in this area and propose a new algorithm that finds partitionings through merging mixture components relying on their pairwise overlap. The proposed technique is illustrated on a popular classification and several synthetic datasets, with excellent results.

Sufficient Dimension Reduction via Distance Covariance

P. 91-104

Wenhui Sheng & Xiangrong Yin

Abstract

We introduce a novel approach to sufficient dimension-reduction problems using distance covariance. Our method requires very mild conditions on the predictors. It estimates the central subspace effectively even when many predictors are categorical or discrete. Our method keeps the model-free advantage without estimating link function. Under regularity conditions, root-*n* consistency and asymptotic normality are established for our estimator. We compare the performance of our method with some existing dimension-reduction methods by simulations and find that our method is very competitive and robust across a number of models. We also analyze the Auto MPG data to demonstrate the efficacy of our method. Supplemental materials for this article are available online.

Assessing the Calibration of High-Dimensional Ensemble Forecasts Using Rank Histograms

P. 105-122

Thordis L. Thorarinsdottir, Michael Scheuerer & Christopher Heinz

Abstract

Any decision-making process that relies on a probabilistic forecast of future events necessarily requires a calibrated forecast. This article proposes new methods for empirically assessing forecast calibration in a multivariate setting where the probabilistic forecast is given by an ensemble of equally probable forecast scenarios. Multivariate properties are mapped to a single dimension through a prerank function and the calibration is subsequently assessed visually through a histogram of the ranks of the observation's preranks. Average ranking assigns a prerank based on the average univariate rank while band depth ranking employs the concept of functional band depth where the centrality of the observation within the forecast ensemble is assessed. Several simulation examples and a case study of temperature forecast trajectories at Berlin Tegel Airport in Germany demonstrate that both multivariate ranking methods can successfully detect various sources of miscalibration and scale efficiently to high-dimensional settings. Supplemental material in form of computer code is available online.

Accounting for Time Series Errors in Partially Linear Model With Single- or Multiple-Runs

P. 123-143

Chunming Zhang, Yu Han & Shengji Jia

Abstract

This article concerns statistical estimation of the partially linear model (PLM) for time course measurements, which are temporally correlated and allow multiple-runs for repeated measurements to enhance experimental accuracy without extending the number of time points within each trial. Such features arise naturally from biomedical data, for example, in brain fMRI, and call for special treatment beyond classical methods in either a purely nonparametric regression model or a PLM with independent errors. We develop a stepwise procedure for estimating the parametric and nonparametric components of the multiple-run PLM and making inference for parameters of interest, adaptive to either single- or multiple-run, in the presence of error temporal dependence. Simulation study and real fMRI data applications illustrate the computational simplicity and effectiveness of the proposed methods. Supplementary material for this article is available online.

Robust Autocorrelation Estimation P. 144-166

Christopher C. Chang & Dimitris N. Politis

Abstract

In this article, we introduce a new class of robust autocorrelation estimators based on interpreting the sample autocorrelation function as a linear regression. We investigate the efficiency and robustness properties of the estimators that result from employing three common robust regression techniques. We discuss the construction of robust autocovariance and positive definite autocorrelation estimates, and their application to AR model fitting. We perform simulation studies with various outlier configurations to compare the different estimators.

Covariance Partition Priors: A Bayesian Approach to Simultaneous Covariance Estimation for Longitudinal Data P. 167-186

J. T. Gaskins & M. J. Daniels

Abstract

The estimation of the covariance matrix is a key concern in the analysis of longitudinal data. When data consist of multiple groups, it is often assumed the covariance matrices are either equal across groups or are completely distinct. We seek methodology to allow borrowing of strength across potentially similar groups to improve estimation. To that end, we introduce a covariance partition prior that proposes a partition of the groups at each measurement time. Groups in the same set of the partition share dependence parameters for the distribution of the current measurement given the preceding ones, and the sequence of partitions is modeled as a Markov chain to encourage similar structure at nearby measurement times. This approach additionally encourages a lower-dimensional structure of the covariance matrices by shrinking the parameters of the Cholesky decomposition toward zero. We demonstrate the performance of our model through two simulation studies and the analysis of data from a depression study. This article includes Supplementary Materials available online.

Statistically and Computationally Efficient Estimating Equations for Large Spatial Datasets P. 187-208

Ying Sun & Michael L. Stein

Abstract

For Gaussian process models, likelihood-based methods are often difficult to use with large irregularly spaced spatial datasets, because exact calculations of the likelihood for n observations require $O(n^3)$ operations and $O(n^2)$ memory. Various approximation methods have been developed to address the computational difficulties. In this article, we propose new, unbiased estimating equations (EE) based on score equation approximations that are both computationally and statistically efficient. We replace the inverse covariance matrix that appears in the score equations by a sparse matrix to approximate the quadratic forms, then set the resulting quadratic forms equal to their expected values to obtain unbiased EE. The sparse matrix is constructed by a sparse inverse Cholesky approach to approximate the inverse covariance matrix. The statistical efficiency of the resulting unbiased EE is evaluated both in theory and by numerical studies. Our methods are applied to nearly 90,000 satellite-based measurements of water vapor levels over a region in the Southeast Pacific Ocean.

Nonparametric Estimation for Self-Exciting Point Processes—A Parsimonious Approach P. 209-224

Feng Chen & Peter Hall

Abstract

There is ample evidence that in applications of self-exciting point-process models, the intensity of background events is often far from constant. If a constant background is imposed that assumption can reduce significantly the quality of statistical analysis, in problems as diverse as modeling the after-shocks of earthquakes and the study of ultra-high frequency financial data. Parametric models can be used to alleviate this problem, but they run the risk of distorting inference by misspecifying the nature of the background intensity function. On the other hand, a purely nonparametric approach to analysis leads to problems of identifiability; when a nonparametric approach is taken, not every aspect of the model can be identified from data recorded along a single observed sample path. In this article, we suggest overcoming this difficulty by using an approach based on the principle of parsimony, or Occam's razor. In particular, we suggest taking the point-process intensity to be either a constant or to have maximum differential entropy, in cases where there is not sufficient empirical evidence to suggest that the background intensity function is more complex than those models. This approach is seldom, if ever, used for nonparametric function estimation in other settings, not least because in those cases more data are typically available. However, our "ontological parsimony" argument is appropriate in the context of self-exciting point-process models. Supplementary materials are available online.

Laplace Variational Approximation for Semiparametric Regression in the Presence of Heteroscedastic Errors P. 225-245

Bruce D. Bugbee, F. Jay Breidt & Mark J. van der Woerd

Abstract

Variational approximations provide fast, deterministic alternatives to Markov chain Monte Carlo for Bayesian inference on the parameters of complex, hierarchical models. Variational approximations are often limited in practicality in the absence of conjugate posterior distributions. Recent work has focused on the application of variational methods to models with only partial conjugacy, such as in semiparametric regression with heteroscedastic errors. Here, both the mean and log variance functions are modeled as smooth functions of covariates. For this problem, we derive a mean field variational approximation with an embedded Laplace approximation to account for the nonconjugate structure. Empirical results with simulated and real data show that our approximate method has significant computational advantages over traditional Markov chain Monte Carlo; in this case, a delayed rejection adaptive Metropolis algorithm. The variational approximation is much faster and eliminates the need for tuning parameter selection, achieves good fits for both the mean and log variance functions, and reasonably reflects the posterior uncertainty. We apply the methods to log-intensity data from a small angle X-ray scattering experiment, in which properly accounting for the smooth heteroscedasticity leads to significant improvements in posterior inference for key physical characteristics of an organic molecule.

Joint Modeling of Multiple Network Views P. 246-265

Isabella Gollini & Thomas Brendan Murphy

Abstract

Latent space models (LSM) for network data rely on the basic assumption that each node of the network has an unknown position in a *D*-dimensional Euclidean latent space: generally the smaller the distance between two nodes in the latent space, the greater their probability of being connected. In this article, we propose a variational inference approach to estimate the intractable posterior of the LSM. In many cases, different network views on the same set of nodes are available. It can therefore be useful to build a model able to jointly summarize the information given by all the network views. For this purpose, we introduce the latent space joint model (LSJM) that merges the information given by multiple network views assuming that the probability of a node being connected with other nodes in each network view is explained by a unique latent variable. This model is demonstrated on the analysis of two datasets: an excerpt of 50 girls from "Teenage Friends and Lifestyle Study" data at three time points and the *Saccharomyces cerevisiae* genetic and physical protein–protein interactions. Supplementary materials for this article are available online.

A Pivotal Allocation-Based Algorithm for Solving the Label-Switching Problem in Bayesian Mixture Models

P. 266-283

Han Li & Xiaodan Fan

Abstract

In Bayesian analysis of mixture models, the label-switching problem occurs as a result of the posterior distribution being invariant to any permutation of cluster indices under symmetric priors. To solve this problem, we propose a novel relabeling algorithm and its variants by investigating an approximate posterior distribution of the latent allocation variables instead of dealing with the component parameters directly. We demonstrate that our relabeling algorithm can be formulated in a rigorous framework based on information theory. Under some circumstances, it is shown to resemble the classical Kullback-Leibler relabeling algorithm and include the recently proposed equivalence classes representatives relabeling algorithm as a special case. Using simulation studies and real data examples, we illustrate the efficiency of our algorithm in dealing with various label-switching phenomena. Supplemental materials for this article are available online.

Algorithms for Envelope Estimation

P. 284-300

R. Dennis Cook & Xin Zhang

Abstract

Envelopes were recently proposed as methods for reducing estimative variation in multivariate linear regression. Estimation of an envelope usually involves optimization over Grassmann manifolds. We propose a fast and widely applicable one-dimensional (1D) algorithm for estimating an envelope in general. We reveal an important structural property of envelopes that facilitates our algorithm, and we prove both Fisher consistency and $\sqrt{n}$-consistency of the algorithm. Supplementary materials for this article are available online.

Computational Aspects of Optional Pólya Tree

P. 301-320

Hui Jiang, John Chong Mu, Kun Yang, Chao Du, Luo Lu & Wing Hung Wong

Abstract

Optional Pólya tree (OPT) is a flexible nonparametric Bayesian prior for density estimation. Despite its merits, the computation for OPT inference is challenging. In this article, we present time complexity analysis for OPT inference and propose two algorithmic improvements. The first improvement, named limited-lookahead optional Pólya tree (LL-OPT), aims at accelerating the computation for OPT inference. The second improvement modifies the output of OPT or LL-OPT and produces a continuous piecewise linear density estimate. We demonstrate the performance of these two improvements using simulated and real date examples.