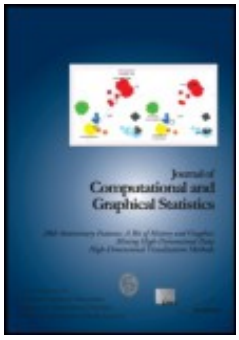

Journal of computational and graphical statistics, ISSN 1061-8600
Volume 25, number 2 (june 2016)

Abstract
SiZer (significant zero crossing of the derivatives) is a multiscale smoothing method for exploring trends, maxima, and minima in data. In this article, a regression spline version of SiZer is proposed in a nonparametric regression setting by the fiducial method. The number of knots for spline interpolation is used as the scale parameter of the new SiZer, which controls the smoothness of estimate. In the construction of the new SiZer, multiple testing adjustment is made to control the row-wise false discovery rate (FDR) of SiZer. This adjustment is appealing for exploratory data analysis and has potential to increase the power. A special map is also produced on a continuous scale using *p*-values to assess the significance of features. Simulations and a real data application are carried out to investigate the performance of the proposed SiZer, in which several comparisons with other existing SiZers are presented.

Abstract
In this article, we analyze approximate methods for undertaking a principal components analysis (PCA) on large datasets. PCA is a classical dimension reduction method that involves the projection of the data onto the subspace spanned by the leading eigenvectors of the covariance matrix. This projection can be used either for exploratory purposes or as an input for further analysis, for example, regression. If the data have billions of entries or more, the computational and storage requirements for saving and manipulating the design matrix in fast memory are prohibitive. Recently, the Nyström and column-sampling methods have appeared in the numerical linear algebra community for the randomized approximation of the singular value decomposition of large matrices. However, their utility for statistical applications remains unclear. We compare these approximations theoretically by bounding the distance between the induced subspaces and the desired, but computationally infeasible, PCA subspace. Additionally we show empirically, through simulations and a real data example involving a corpus of emails, the trade-off of approximation accuracy and computational complexity.

Abstract
In this article, novel joint semiparametric spline-based modeling of conditional mean and volatility of financial time series is proposed and evaluated on daily stock return data. The modeling includes functions of lagged response variables and time as predictors. The latter can be viewed as a proxy for omitted economic variables contributing to the underlying dynamics. The conditional mean model is additive. The conditional volatility model is multiplicative and linearized with a logarithmic transformation. In addition, a cube-root power transformation is employed to symmetrize the lagged response variables.

Using cubic splines, the model can be written as a multiple linear regression, thereby allowing predictions to be obtained in a simple manner. As outliers are often present in financial data, reliable estimation of the model parameters is achieved by trimmed least-square (TLS) estimation for which a reasonable amount of trimming is suggested. To obtain a parsimonious specification of the model, a new model selection criterion corresponding to TLS is derived. Moreover, the (three-parameter) generalized gamma distribution is identified as suitable for the absolute multiplicative errors and shown to work well for predictions and also for the calculation of quantiles, which is important to determine the value at risk. All model choices are motivated by a detailed analysis of IBM, HP, and SAP daily returns. The prediction performance is compared to the classical generalized autoregressive conditional heteroskedasticity (GARCH) and asymmetric power GARCH (APGARCH) models as well as to a nonstationary time-trend volatility model. The results suggest that the proposed model may possess a high predictive power for future conditional volatility.

Penalized Fast Subset Scanning P. 382-404

Skyler Speakman, Sriram Somanchi, Edward McFowland III & Daniel B. Neill

Abstract

We present the penalized fast subset scan (PFSS), a new and general framework for scalable and accurate pattern detection. PFSS enables exact and efficient identification of the most anomalous subsets of the data, as measured by a likelihood ratio scan statistic. However, PFSS also allows incorporation of prior information about each data element's probability of inclusion, which was not previously possible within the subset scan framework. PFSS builds on two main results: first, we prove that a large class of likelihood ratio statistics satisfy a property that allows additional, element-specific penalty terms to be included while maintaining efficient computation. Second, we prove that the penalized statistic can be maximized exactly by evaluating only $O(N)$ subsets. As a concrete example of the PFSS framework, we incorporate "soft" constraints on spatial proximity into the spatial event detection task, enabling more accurate detection of irregularly shaped spatial clusters of varying sparsity. To do so, we develop a distance-based penalty function that rewards spatial compactness and penalizes spatially dispersed clusters. This approach was evaluated on the task of detecting simulated anthrax bio-attacks, using real-world Emergency Department data from a major U.S. city. PFSS demonstrated increased detection power and spatial accuracy as compared to competing methods while maintaining efficient computation.

Parameter Expanded Algorithms for Bayesian Latent Variable Modeling of Genetic Pleiotropy Data P. 405-425

Lizhen Xu, Radu V. Craiu, Lei Sun & Andrew D. Paterson

Abstract

Motivated by genetic association studies of pleiotropy, we propose a Bayesian latent variable approach to jointly study multiple outcomes. The models studied here can incorporate both continuous and binary responses, and can account for serial and cluster correlations. We consider Bayesian estimation for the model parameters, and we develop a novel MCMC algorithm that builds upon hierarchical centering and parameter expansion techniques to efficiently sample from the posterior distribution. We evaluate the proposed method via extensive simulations and demonstrate its utility with an application to an association study of various complication outcomes related to Type 1 diabetes.

Online Variational Bayes Inference for High-Dimensional Correlated Data P. 426-444

Sylvie (Tchumtchoua) Kabisa, David B. Dunson & Jeffrey S. Morris

Abstract

High-dimensional data with hundreds of thousands of observations are becoming commonplace in many disciplines. The analysis of such data poses many computational challenges, especially when the observations are correlated over time and/or across space. In this article, we propose flexible hierarchical regression models for analyzing such data that accommodate serial and/or spatial correlation. We address the computational challenges involved in fitting these models by adopting an approximate inference framework. We develop an online variational Bayes algorithm that works by incrementally reading the data into memory one portion at a time. The performance of the method is assessed through simulation studies.

The methodology is applied to analyze signal intensity in MRI images of subjects with knee osteoarthritis, using data from the Osteoarthritis Initiative.

s-CorrPlot: An Interactive Scatterplot for Exploring Correlation P. 445-463

Sean McKenna, Miriah Meyer, Christopher Gregg & Samuel Gerber

Abstract

The degree of correlation between variables is used in many data analysis applications as a key measure of interdependence. The most common techniques for exploratory analysis of pairwise correlation in multivariate datasets, like scatterplot matrices and clustered heatmaps, however, do not scale well to large datasets, either computationally or visually. We present a new visualization that is capable of encoding pairwise correlation between hundreds of thousands variables, called the s-CorrPlot. The s-CorrPlot encodes correlation spatially between variables as points on scatterplot using the geometric structure underlying Pearson's correlation. Furthermore, we extend the s-CorrPlot with interactive techniques that enable animation of the scatterplot to new projections of the correlation space, as illustrated in the companion video in supplementary materials. We provide the s-CorrPlot as an open-source R package and validate its effectiveness through a variety of methods including a case study with a biology collaborator.

Estimation Stability With Cross-Validation (ESCV) P. 464-492

Chinghway Lim & Bin Yu

Abstract

Cross-validation (CV) is often used to select the regularization parameter in high-dimensional problems. However, when applied to the sparse modeling method Lasso, CV leads to models that are unstable in high-dimensions, and consequently not suited for reliable interpretation. In this article, we propose a model-free criterion ESCV based on a new *estimation stability* (ES) metric and CV. Our proposed ESCV finds a smaller and locally ES-optimal model smaller than the CV choice so that it fits the data and also enjoys estimation stability property. We demonstrate that ESCV is an effective alternative to CV at a similar easily parallelizable computational cost. In particular, we compare the two approaches with respect to several performance measures when applied to the Lasso on both simulated and real datasets. For dependent predictors common in practice, our main finding is that ESCV cuts down false positive rates often by a large margin, while sacrificing little of true positive rates. ESCV usually outperforms CV in terms of parameter estimation while giving similar performance as CV in terms of prediction. For the two real datasets from neuroscience and cell biology, the models found by ESCV are less than half of the model sizes by CV, but preserves CV's predictive performance and corroborates with subject knowledge and independent work. We also discuss some regularization parameter alignment issues that come up in both approaches.

Sparse Penalized Forward Selection for Support Vector Classification P. 493-514

Subhashis Ghosal, Bradley Turnbull, Hao Helen Zhang & Wook Yeon Hwang

Abstract

We propose a new binary classification and variable selection technique especially designed for high-dimensional predictors. Among many predictors, typically, only a small fraction of them have significant impact on prediction. In such a situation, more interpretable models with better prediction accuracy can be obtained by variable selection along with classification. By adding an ℓ_1-type penalty to the loss function, common classification methods such as logistic regression or support vector machines (SVM) can perform variable selection. Existing penalized SVM methods all attempt to jointly solve all the parameters involved in the penalization problem altogether. When data dimension is very high, the joint optimization problem is very complex and involves a lot of memory allocation. In this article, we propose a new penalized forward search technique that can reduce high-dimensional optimization problems to one-dimensional optimization by iterating the selection steps. The new algorithm can be regarded as a forward selection version of the penalized SVM and its variants. The advantage of optimizing in one dimension is that the location of the optimum solution can be obtained with intelligent search by exploiting convexity and a piecewise linear or quadratic structure of

the criterion function. In each step, the predictor that is most able to predict the outcome is chosen in the model. The search is then repeatedly used in an iterative fashion until convergence occurs. Comparison of our new classification rule with ℓ_1-SVM and other common methods show very promising performance, in that the proposed method leads to much leaner models without compromising misclassification rates, particularly for high-dimensional predictors.

Bayesian Variable Selection on Model Spaces Constrained by Heredity Conditions P. 515-535

Daniel Taylor-Rodriguez, Andrew Womack & Nikolay Bliznyuk

Abstract

This article investigates Bayesian variable selection when there is a hierarchical dependence structure on the inclusion of predictors in the model. In particular, we study the type of dependence found in polynomial response surfaces of orders two and higher, whose model spaces are required to satisfy weak or strong heredity conditions. These conditions restrict the inclusion of higher-order terms depending upon the inclusion of lower-order parent terms. We develop classes of priors on the model space, investigate their theoretical and finite sample properties, and provide a Metropolis–Hastings algorithm for searching the space of models. The tools proposed allow fast and thorough exploration of model spaces that account for hierarchical polynomial structure in the predictors and provide control of the inclusion of false positives in high posterior probability models.

Fast Hamiltonian Monte Carlo Using GPU Computing P. 536-548

Andrew L. Beam, Sujit K. Ghosh & Jon Doyle

Abstract

In recent years, the Hamiltonian Monte Carlo (HMC) algorithm has been found to work more efficiently compared to other popular Markov chain Monte Carlo (MCMC) methods (such as random walk Metropolis–Hastings) in generating samples from a high-dimensional probability distribution. HMC has proven more efficient in terms of mixing rates and effective sample size than previous MCMC techniques, but still may not be sufficiently fast for particularly large problems. The use of GPUs promises to push HMC even further greatly increasing the utility of the algorithm. By expressing the computationally intensive portions of HMC (the evaluations of the probability kernel and its gradient) in terms of linear or element-wise operations, HMC can be made highly amenable to the use of graphics processing units (GPUs). A multinomial regression example demonstrates the promise of GPU-based HMC sampling. Using GPU-based memory objects to perform the entire HMC simulation, most of the latency penalties associated with transferring data from main to GPU memory can be avoided. Thus, the proposed computational framework may appear conceptually very simple, but has the potential to be applied to a wide class of hierarchical models relying on HMC sampling. Models whose posterior density and corresponding gradients can be reduced to linear or element-wise operations are amenable to significant speed ups through the use of GPUs. Analyses of datasets that were previously intractable for fully Bayesian approaches due to the prohibitively high computational cost are now feasible using the proposed framework.

Direction-Projection-Permutation for High-Dimensional Hypothesis Tests P. 549-569

Susan Wei, Chihoon Lee, Lindsay Wichers & J. S. Marron

Abstract

High-dimensional low sample size (HDLSS) data are becoming increasingly common in statistical applications. When the data can be partitioned into two classes, a basic task is to construct a classifier that can assign objects to the correct class. Binary linear classifiers have been shown to be especially useful in HDLSS settings and preferable to more complicated classifiers because of their ease of interpretability. We propose a computational tool called direction-projection-permutation (DiProPerm), which rigorously assesses whether a binary linear classifier is detecting statistically significant differences between two high-dimensional distributions. The basic idea behind DiProPerm involves working directly with the one-dimensional projections of the data induced by binary linear classifier. Theoretical properties of DiProPerm are studied under the HDLSS asymptotic regime whereby dimension diverges to infinity while

sample size remains fixed. We show that certain variations of DiProPerm are consistent and that consistency is a nontrivial property of tests in the HDLSS asymptotic regime. The practical utility of DiProPerm is demonstrated on HDLSS gene expression microarray datasets. Finally, an empirical power study is conducted comparing DiProPerm to several alternative two-sample HDLSS tests to understand the advantages and disadvantages of each method.

Soft Null Hypotheses: A Case Study of Image Enhancement Detection in Brain Lesions P. 570-588

Haochang Shou, Russell T. Shinohara, Han Liu, Daniel S. Reich & Ciprian M. Crainiceanu

Abstract

This work is motivated by a study of a population of multiple sclerosis (MS) patients using dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) to identify active brain lesions. At each visit, a contrast agent is administered intravenously to a subject and a series of images are acquired to reveal the location and activity of MS lesions within the brain. Our goal is to identify the enhancing lesion locations at the subject level and lesion enhancement patterns at the population level. We analyze a total of 20 subjects scanned at 63 visits (∼30Gb), the largest population of such clinical brain images. After addressing the computational challenges, we propose possible solutions to the difficult problem of transforming a qualitative scientific null hypothesis, such as "this voxel does not enhance," to a well-defined and numerically testable null hypothesis based on the existing data. We call such procedure "soft null" hypothesis testing as opposed to the standard "hard null" hypothesis testing. This problem is fundamentally different from: (1) finding testing statistics when a quantitative null hypothesis is given; (2) clustering using a mixture distribution; or (3) setting a reasonable threshold with a parametric null assumption.

Bayesian Ising Graphical Model for Variable Selection P. 589-605

Zaili Fang & Inyoung Kim

Abstract

In this article, we propose a new Bayesian variable selection (BVS) approach via the graphical model and the Ising model, which we refer to as the "Bayesian Ising graphical model" (BIGM). The BIGM is developed by showing that the BVS problem based on the linear regression model can be considered as a complete graph and described by an Ising model with random interactions. There are several advantages of our BIGM: it is easy to (i) employ the single-site updating and cluster updating algorithm, both of which are suitable for problems with small sample sizes and a larger number of variables, (ii) extend this approach to nonparametric regression models, and (iii) incorporate graphical prior information. In our BIGM, the interactions are determined by the linear model coefficients, so we systematically study the performance of different scale normal mixture priors for the model coefficients by adopting the global-local shrinkage strategy. Our results indicate that the best prior for the model coefficients in terms of variable selection should place substantial weight on small, nonzero shrinkage. The methods are illustrated with simulated and real data.

Tweedie's Compound Poisson Model With Grouped Elastic Net P. 606-625

Wei Qian, Yi Yang & Hui Zou

Abstract

Wei Qian is Assistant Professor, School of Mathematical Sciences, Rochester Institute of Technology, Rochester, NY 14623 (E-mail: *wxqsma@rit.edu*). Yi Yang is Assistant Professor, Department of Mathematics and Statistics, McGill University, Canada (E-mail: *yi.yang6@mcgill.ca*) Hui Zou is Professor of Statistics, School of Statistics, University of Minnesota, Minneapolis, MN 55455 (E-mail: *zouxx019@umn.edu*).

Tweedie's compound Poisson model is a popular method to model data with probability mass at zero and nonnegative, highly right-skewed distribution. Motivated by wide applications of the Tweedie model in various fields such as actuarial science, we investigate the grouped elastic net method for the Tweedie model in the context of the generalized linear model. To efficiently compute the estimation coefficients, we devise a two-layer algorithm that embeds the blockwise majorization descent method into an iteratively reweighted least square strategy. Integrated with the strong rule, the

proposed algorithm is implemented in an easy-to-use R package HDtweedie, and is shown to compute the whole solution path very efficiently. Simulations are conducted to study the variable selection and model fitting performance of various lasso methods for the Tweedie model. The modeling applications in risk segmentation of insurance business are illustrated by analysis of an auto insurance claim dataset.

Parallel Variational Bayes for Large Datasets With an Application to Generalized Linear Mixed Models P. 626-646

Minh-Ngoc Tran, David J. Nott, Anthony Y. C. Kuk & Robert Kohn

Abstract

The article develops a hybrid variational Bayes (VB) algorithm that combines the mean-field and stochastic linear regression fixed-form VB methods. The new estimation algorithm can be used to approximate any posterior without relying on conjugate priors. We propose a divide and recombine strategy for the analysis of large datasets, which partitions a large dataset into smaller subsets and then combines the variational distributions that have been learned in parallel on each separate subset using the hybrid VB algorithm. We also describe an efficient model selection strategy using cross-validation, which is straightforward to implement as a by-product of the parallel run. The proposed method is applied to fitting generalized linear mixed models. The computational efficiency of the parallel and hybrid VB algorithm is demonstrated on several simulated and real datasets.