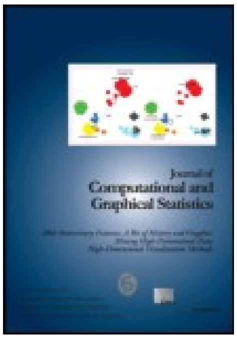

Journal of computational and graphical statistics, ISSN 1061-8600
Volume 26, number 3 (september 2017)

Letter-Value Plots: Boxplots for Large Data P. 469-477

Heike Hofmann, Hadley Wickham & Karen Kafadar

Abstract

Boxplots are useful displays that convey rough information about the distribution of a variable. Boxplots were designed to be drawn by hand and work best for small datasets, where detailed estimates of tail behavior beyond the quartiles may not be trustworthy. Larger datasets afford more precise estimates of tail behavior, but boxplots do not take advantage of this precision, instead presenting large numbers of extreme, though not unexpected, observations. Letter-value plots address this problem by including more detailed information about the tails using “letter values,” an order statistic defined by Tukey. Boxplots display the first two letter values (the median and quartiles); letter-value plots display further letter values so far as they are reliable estimates of their corresponding quantiles. We illustrate letter-value plots with real data that demonstrate their usefulness for large datasets. All graphics are created using the R package lvplot, and code and data are available in the supplementary materials.

Model Choice and Diagnostics for Linear Mixed-Effects Models Using Statistics on Street Corners P. 478-492

Adam Loy, Heike Hofmann & Dianne Cook

Abstract

The complexity of linear mixed-effects (LME) models means that traditional diagnostics are rendered less effective. This is due to a breakdown of asymptotic results, boundary issues, and visible patterns in residual plots that are introduced by the model fitting process. Some of these issues are well known and adjustments have been proposed. Working with LME models typically requires that the analyst keeps track of all the special circumstances that may arise. In this article, we illustrate a simpler but generally applicable approach to diagnosing LME models. We explain how to use new visual inference methods for these purposes. The approach provides a unified framework for diagnosing LME fits and for model selection. We illustrate the use of this approach on several commonly available datasets. A large-scale Amazon Turk study was used to validate the methods. R code is provided for the analyses. Supplementary materials for this article are available online.

Designing Modular Software: A Case Study in Introductory Statistics P. 493-500

Eric Hare & Andee Kaplan

Abstract

Modular programming is a development paradigm that emphasizes self-contained, flexible, and independent pieces of functionality. This practice allows new features to be seamlessly added when desired, and unwanted features to be removed, thus simplifying the software's user interface. The recent rise of web-based software applications has presented new challenges for designing an extensible, modular software system. In this article, we outline a framework for designing such a system, with a focus on reproducibility of the results. We present as a case study a Shiny-based web application called intRo, that allows the user to perform basic data analyses and statistical routines. Finally, we highlight some challenges we encountered, and how to address them, when combining modular programming concepts with reactive programming as used by Shiny. Supplementary material for this article is available online.

Group-Wise Principal Component Analysis for Exploratory Data Analysis P. 501-512

José Camacho, Rafael A. Rodríguez-Gómez & Edoardo Saccenti

Abstract

In this article, we propose a new framework for matrix factorization based on principal component analysis (PCA) where sparsity is imposed. The structure to impose sparsity is defined in terms of groups of correlated variables found in correlation matrices or maps. The framework is based on three new contributions: an algorithm to identify the groups of variables in correlation maps, a visualization for the resulting groups, and a matrix factorization. Together with a method to compute correlation maps with minimum noise level, referred to as missing-data for exploratory data analysis (MEDA), these three contributions constitute a complete matrix factorization framework. Two real examples are used to illustrate the approach and compare it with PCA, sparse PCA, and structured sparse PCA. Supplementary materials for this article are available online.

Quantifying the Uncertainty of Contour Maps P. 513-524

David Bolin & Finn Lindgren

Abstract

Contour maps are widely used to display estimates of spatial fields. Instead of showing the estimated field, a contour map only shows a fixed number of contour lines for different levels. However, despite the ubiquitous use of these maps, the uncertainty associated with them has been given a surprisingly small amount of attention. We derive measures of the statistical uncertainty, or quality, of contour maps, and use these to decide an appropriate number of contour lines, which relates to the uncertainty in the estimated spatial field. For practical use in geostatistics and medical imaging, computational methods are constructed, that can be applied to Gaussian Markov random fields, and in particular be used in combination with integrated nested Laplace approximations for latent Gaussian models. The methods are demonstrated on simulated data and an application to temperature estimation is presented.

One-Step Estimator Paths for Concave Regularization P. 525-536

Matt Taddy

Abstract

The statistics literature of the past 15 years has established many favorable properties for sparse diminishing-bias regularization: techniques that can roughly be understood as providing estimation under penalty functions spanning the range of concavity between ℓ_0 and ℓ_1 norms. However, lasso ℓ_1-regularized estimation remains the standard tool for industrial Big Data applications because of its minimal computational cost and the presence of easy-to-apply rules for penalty selection. In response, this article proposes a simple new algorithm framework that requires no more computation than a lasso path: the path of one-step estimators (POSE) does ℓ_1 penalized regression estimation on a grid of decreasing penalties, but adapts coefficient-specific weights to decrease as a function of the coefficient estimated in the previous path step. This provides sparse diminishing-bias regularization at no extra cost over the fastest lasso algorithms. Moreover, our gamma lasso implementation of POSE is accompanied by a reliable heuristic for the fit degrees of freedom, so that standard information criteria can be applied in penalty selection. We also provide novel results on the distance between weighted-ℓ_1 and ℓ_0 penalized predictors; this allows us to build intuition about POSE and other diminishing-bias regularization schemes. The methods and results are illustrated in extensive simulations and in application of logistic regression to evaluating the performance of hockey players. Supplementary materials for this article are available online.

Tensor Decomposition With Generalized Lasso Penalties

P. 537-546

Oscar Hernan Madrid-Padilla & James Scott

Abstract

We present an approach for penalized tensor decomposition (PTD) that estimates smoothly varying latent factors in multiway data. This generalizes existing work on sparse tensor decomposition and penalized matrix decompositions, in a manner parallel to the generalized lasso for regression and smoothing problems. Our approach presents many nontrivial challenges at the intersection of modeling and computation, which are studied in detail. An efficient coordinate-wise optimization algorithm for PTD is presented, and its convergence properties are characterized. The method is applied both to simulated data and real data on flu hospitalizations in Texas and motion-capture data from video cameras. These results show that our penalized tensor decomposition can offer major improvements on existing methods for analyzing multiway data that exhibit smooth spatial or temporal features.

Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression

P. 547-557

Congrui Yi & Jian Huang

Abstract

We propose an algorithm, semismooth Newton coordinate descent (SNCD), for the elastic-net penalized Huber loss regression and quantile regression in high dimensional settings. Unlike existing coordinate descent type algorithms, the SNCD updates a regression coefficient and its corresponding subgradient simultaneously in each iteration. It combines the strengths of the coordinate descent and the semismooth Newton algorithm, and effectively solves the computational challenges posed by dimensionality and nonsmoothness. We establish the convergence properties of the algorithm. In addition, we present an adaptive version of the “strong rule” for screening predictors to gain extra efficiency. Through numerical experiments, we demonstrate that the proposed algorithm is very efficient and scalable to ultrahigh dimensions. We illustrate the application via a real data example. Supplementary materials for this article are available online.

Monitoring Joint Convergence of MCMC Samplers

P. 558-568

Douglas VanDerwerken & Scott C. Schmidler

Abstract

We present a diagnostic for monitoring convergence of a Markov chain Monte Carlo (MCMC) sampler to its target distribution. In contrast to popular existing methods, we monitor convergence to the joint target distribution directly rather than a select scalar projection. The method uses a simple nonparametric posterior approximation based on a state-space partition obtained by clustering the pooled draws from multiple chains, and convergence is determined when the estimated posterior probabilities of partition elements under each chain are sufficiently similar. This framework applies to a wide variety of problems, and generalizes directly to non-Euclidean state spaces. Our method also provides approximate high-posterior-density regions, and a characterization of differences between nonconverged chains, all with little additional computational burden. We demonstrate this approach on applications to sampling posterior distributions over R*p*, graphs, and partitions. Supplementary materials for this article are available online.

Penalized Nonparametric Scalar-on-Function Regression via Principal Coordinates

P. 569-578

Philip T. Reiss, David L. Miller, Pei-Shien Wu & Wen-Yu Hua

Abstract

A number of classical approaches to nonparametric regression have recently been extended to the case of functional predictors. This article introduces a new method of this type, which extends intermediate-rank penalized smoothing to scalar-on-function regression. In the proposed method, which we call *principal coordinate ridge regression*, one regresses the response on leading principal coordinates defined by a relevant distance among the functional predictors, while applying a ridge penalty. Our publicly available implementation, based on generalized additive modeling software, allows for fast optimal tuning parameter selection and for extensions to multiple functional predictors, exponential family-valued responses, and mixed-effects models. In an application to signature verification data, principal coordinate ridge regression, with dynamic time warping distance used to define the principal coordinates, is shown to outperform a functional generalized linear model. Supplementary materials for this article are available online.

ThrEEBoost: Thresholded Boosting for Variable Selection and Prediction via Estimating Equations

P. 579-588

Ben Brown, Christopher J. Miller & Julian Wolfson

Abstract

Most variable selection techniques for high-dimensional models are designed to be used in settings, where observations are independent and completely observed. At the same time, there is a rich literature on approaches to estimation of low-dimensional parameters in the presence of correlation, missingness, measurement error, selection bias, and other characteristics of real data. In this article, we present ThrEEBoost (*Thr*esholded *EEBoost*), a general-purpose variable selection technique which can accommodate such problem characteristics by replacing the gradient of the loss by an estimating function. ThrEEBoost generalizes the previously proposed EEBoost algorithm (Wolfson 2011 Wolfson, J. (2011), "EEBoost: A General Method for Prediction and Variable Selection Based on Estimating Equations," *Journal of the American Statistical Association*, 106, 296–305.[Taylor & Francis Online], [Web of Science ®], [Google Scholar]) by allowing the number of regression coefficients updated at each step to be controlled by a thresholding parameter. Different thresholding parameter values yield different variable selection paths, greatly diversifying the set of models that can be explored; the optimal degree of thresholding can be chosen by cross-validation. ThrEEBoost was evaluated using simulation studies to assess the effects of different threshold values on prediction error, sensitivity, specificity, and the number of iterations to identify minimum prediction error under both sparse and nonsparse true models with correlated continuous outcomes. We show that when the true model is sparse, ThrEEBoost achieves similar prediction error to EEBoost while requiring fewer iterations to locate the set of coefficients yielding the minimum error. When the true model is less sparse, ThrEEBoost has lower prediction error than EEBoost and also finds the point yielding the minimum error more quickly. The technique is illustrated by applying it to the problem of identifying predictors of weight change in a longitudinal nutrition study. Supplementary materials are available online.

Formal Hypothesis Tests for Additive Structure in Random Forests

P. 1195-1211

Lucas Mentch & Giles Hooker

Abstract

While statistical learning methods have proved powerful tools for predictive modeling, the black-box nature of the models they produce can severely limit their interpretability and the ability to conduct formal inference. However, the natural structure of ensemble learners like bagged trees and random forests has been shown to admit desirable asymptotic properties when base learners are built with proper subsamples. In this work, we demonstrate that by defining an appropriate grid structure on the covariate space, we may carry out formal hypothesis tests for both variable importance and underlying additive model structure. To our knowledge, these tests represent the first statistical tools for investigating the underlying regression structure in a context such as random forests. We develop notions of total and partial additivity and further demonstrate that testing can be carried out at no additional computational cost by estimating the variance within the process of constructing the ensemble. Furthermore, we propose a novel extension of these testing procedures using random projections to allow for computationally efficient testing procedures that retain high power even when the grid size is much larger than that of the training set.

Finding Singular Features P. 598-609

Christopher Genovese, Marco Perone-Pacifico, Isabella Verdinelli & Larry Wasserman

Abstract

We present a method for finding high density, low-dimensional structures in noisy point clouds. These structures are sets with zero Lebesgue measure with respect to the *D*-dimensional ambient space and belong to a *d* < *D*-dimensional space. We call them "singular features." Hunting for singular features corresponds to finding unexpected or unknown structures hidden in point clouds belonging to RD. Our method outputs well-defined sets of dimensions *d* < *D*. Unlike spectral clustering, the method works well in the presence of noise. We show how to find singular features by first finding ridges in the estimated density, followed by a filtering step based on the eigenvalues of the Hessian of the density. The code for plotting all the figures, with the corresponding plots, and the data files used in the article, are in the folder SupplementaryDocument.zip that can be find at the *http://www.stat.cmu.edu/larry/singular.*

High-Dimensional Multivariate Time Series With Additional Structure P. 610-622

Michael Schweinberger, Sergii Babkin & Katherine B. Ensor

Abstract

High-dimensional multivariate time series are challenging due to the dependent and high-dimensional nature of the data, but in many applications there is additional structure that can be exploited to reduce computing time along with statistical error. We consider high-dimensional vector autoregressive processes with spatial structure, a simple and common form of additional structure. We propose novel high-dimensional methods that take advantage of such structure without making model assumptions about how distance affects dependence. We provide nonasymptotic bounds on the statistical error of parameter estimators in high-dimensional settings and show that the proposed approach reduces the statistical error. An application to air pollution in the USA demonstrates that the estimation approach reduces both computing time and prediction error and gives rise to results that are meaningful from a scientific point of view, in contrast to high-dimensional methods that ignore spatial structure. In practice, these high-dimensional methods can be used to decompose high-dimensional multivariate time series into lower-dimensional multivariate time series that can be studied by other methods in more depth. Supplementary materials for this article are available online.

Regularized Estimation of Piecewise Constant Gaussian Graphical Models: The Group-Fused Graphical Lasso P. 1248-1271

Alexander J. Gibberd & James D. B. Nelson

Abstract

The time-evolving precision matrix of a piecewise-constant Gaussian graphical model encodes the dynamic conditional dependency structure of a multivariate time-series. Traditionally, graphical models are estimated under the assumption that data are drawn identically from a generating distribution. Introducing sparsity and sparse-difference inducing priors, we relax these assumptions and propose a novel regularized M-estimator to jointly estimate both the graph and changepoint structure. The resulting estimator possesses the ability to therefore favor sparse dependency structures and/or smoothly evolving graph structures, as required. Moreover, our approach extends current methods to allow estimation of changepoints that are grouped across multiple dependencies in a system. An efficient algorithm for estimating structure is proposed. We study the empirical recovery properties in a synthetic setting. The qualitative effect of grouped changepoint estimation is then demonstrated by applying the method on a genetic time-course dataset. Supplementary material for this article is available online.

Modeling Time-Varying Effects With Large-Scale Survival Data: An Efficient Quasi-Newton Approach

P. 635-645

Kevin He, Yuan Yang, Yanming Li, Ji Zhu & Yi Li

Abstract

Nonproportional hazards models often arise in biomedical studies, as evidenced by a recent national kidney transplant study. During the follow-up, the effects of baseline risk factors, such as patients' comorbidity conditions collected at transplantation, may vary over time. To model such dynamic changes of covariate effects, time-varying survival models have emerged as powerful tools. However, traditional methods of fitting time-varying effects survival model rely on an expansion of the original dataset in a repeated measurement format, which, even with a moderate sample size, leads to an extremely large working dataset. Consequently, the computational burden increases quickly as the sample size grows, and analyses of a large dataset such as our motivating example defy any existing statistical methods and software. We propose a novel application of quasi-Newton iteration method to model time-varying effects in survival analysis. We show that the algorithm converges superlinearly and is computationally efficient for large-scale datasets. We apply the proposed methods, via a stratified procedure, to analyze the national kidney transplant data and study the impact of potential risk factors on post-transplant survival. Supplementary materials for this article are available online.

Approximate Bayesian Computation and Model Assessment for Repulsive Spatial Point Processes

P. 646-657

Shinichiro Shirota & Alan E. Gelfand

Abstract

In many applications involving spatial point patterns, we find evidence of inhibition or repulsion. The most commonly used class of models for such settings are the Gibbs point processes. A recent alternative, at least to the statistical community, is the determinantal point process. Here, we examine model fitting and inference for both of these classes of processes in a Bayesian framework. While usual MCMC model fitting can be available, the algorithms are complex and are not always well behaved. We propose using approximate Bayesian computation (ABC) for such fitting. This approach becomes attractive because, though likelihoods are very challenging to work with for these processes, generation of realizations given parameter values is relatively straightforward. As a result, the ABC fitting approach is well-suited for these models. In addition, such simulation makes them well-suited for posterior predictive inference as well as for model assessment. We provide details for all of the above along with some simulation investigation and an illustrative analysis of a point pattern of tree data exhibiting repulsion. R code and datasets are included in the supplementary material.

A Skewed and Heavy-Tailed Latent Random Field Model for Spatial Extremes

P. 658-670

Behzad Mahmoudian

Abstract

This article develops Bayesian inference of spatial models with a flexible skew latent structure. Using the multivariate skew-normal distribution of Sahu et al., a valid random field model with stochastic skewing structure is proposed to take into account non-Gaussian features. The skewed spatial model is further improved via scale mixing to accommodate more extreme observations. Finally, the skewed and heavy-tailed random field model is used to describe the parameters of extreme value distributions. Bayesian prediction is done with a well-known Gibbs sampling algorithm, including slice sampling and adaptive simulation techniques. The model performance—as far as the identifiability of the parameters is concerned—is assessed by a simulation study and an analysis of extreme wind speeds across Iran. We conclude that our model provides more satisfactory results according to Bayesian model selection and predictive-based criteria. R code to implement the methods used is available as online supplementary material.

A Generalized Smoother for Linear Ordinary Differential Equations P. 671-681

Michelle Carey, Eugene G. Gath & Kevin Hayes

Abstract

Ordinary differential equations (ODEs) are equalities involving a function and its derivatives that define the evolution of the function over a prespecified domain. The applications of ODEs range from simulation and prediction to control and diagnosis in diverse fields such as engineering, physics, medicine, and finance. Parameter estimation is often required to calibrate these theoretical models to data. While there are many methods for estimating ODE parameters from partially observed data, they are invariably subject to several problems including high computational cost, complex estimation procedures, biased estimates, and large sampling variance. We propose a method that overcomes these issues and produces estimates of the ODE parameters that have less bias, a smaller sampling variance, and a 10-fold improvement in computational efficiency. The package *GenPen* containing the Matlab code to perform the methods described in this article is available online.

Precision Matrix Estimation With ROPE P. 682-694

M. O. Kuismin, J. T. Kemppainen & M. J. Sillanpää

Abstract

It is known that the accuracy of the maximum likelihood-based covariance and precision matrix estimates can be improved by penalized log-likelihood estimation. In this article, we propose a ridge-type operator for the precision matrix estimation, ROPE for short, to maximize a penalized likelihood function where the Frobenius norm is used as the penalty function. We show that there is an explicit closed form representation of a shrinkage estimator for the precision matrix when using a penalized log-likelihood, which is analogous to ridge regression in a regression context. The performance of the proposed method is illustrated by a simulation study and real data applications. Computer code used in the example analyses as well as other supplementary materials for this article are available online.

A Cross-Entropy Approach to the Estimation of Generalized Linear Multilevel Models P. 695-708

Marco Bee, Giuseppe Espa, Diego Giuliani & Flavio Santi

Abstract

In this article, we use the cross-entropy method for noisy optimization for fitting generalized linear multilevel models through maximum likelihood. We propose specifications of the instrumental distributions for positive and bounded parameters that improve the computational performance. We also introduce a new stopping criterion, which has the advantage of being problem-independent. In a second step we find, by means of extensive Monte Carlo experiments, the most suitable values of the input parameters of the algorithm. Finally, we compare the method to the benchmark estimation technique based on numerical integration. The cross-entropy approach turns out to be preferable from both the statistical and the computational point of view. In the last part of the article, the method is used to model the probability of firm exits in the healthcare industry in Italy. Supplemental materials are available online.

Penalized Estimation in Large-Scale Generalized Linear Array Models P. 709-724

Adam Lund, Martin Vincent & Niels Richard Hansen

Abstract

Large-scale generalized linear array models (GLAMs) can be challenging to fit. Computation and storage of its tensor product design matrix can be impossible due to time and memory constraints, and previously considered design matrix free algorithms do not scale well with the dimension of the parameter vector. A new design matrix free algorithm is proposed for computing the penalized maximum likelihood estimate for GLAMs, which, in particular, handles nondifferentiable penalty functions. The proposed algorithm is implemented and available via the R package glamlasso. It combines several ideas—previously considered separately—to obtain sparse estimates while at the same time efficiently exploiting the GLAM structure. In this article, the convergence of the algorithm is treated and the performance of its implementation is investigated and compared to that of glmnet on simulated as well as real data. It is shown that the computation time for glamlasso scales favorably with the size of the problem when compared to glmnet. Supplementary materials, in the form of R code, data and visualizations of results, are available online.

Link Prediction for Partially Observed Networks

P. 725-733

Yunpeng Zhao, Yun-Jhong Wu, Elizaveta Levina & Ji Zhu

Abstract

Link prediction is one of the fundamental problems in network analysis. In many applications, notably in genetics, a partially observed network may not contain any negative examples, that is, edges known for certain to be absent, which creates a difficulty for existing supervised learning approaches. We develop a new method that treats the observed network as a sample of the true network with different sampling rates for positive (true edges) and negative (absent edges) examples. We obtain a relative ranking of potential links by their probabilities, using information on network topology as well as node covariates if available. The method relies on the intuitive assumption that if two pairs of nodes are similar, the probabilities of these pairs forming an edge are also similar. Empirically, the method performs well under many settings, including when the observed network is sparse. We apply the method to a protein–protein interaction network and a school friendship network.

One-Step Generalized Estimating Equations With Large Cluster Sizes

P. 734-737

Stuart Lipsitz, Garrett Fitzmaurice, Debajyoti Sinha, Nathanael Hevelone, Jim Hu & Louis L. Nguyen

Abstract

Medical studies increasingly involve a large sample of independent clusters, where the cluster sizes are also large. Our motivating example from the 2010 Nationwide Inpatient Sample (NIS) has 8,001,068 patients and 1049 clusters, with average cluster size of 7627. Consistent parameter estimates can be obtained naively assuming independence, which are inefficient when the intra-cluster correlation (ICC) is high. Efficient generalized estimating equations (GEE) incorporate the ICC and sum all pairs of observations within a cluster when estimating the ICC. For the 2010 NIS, there are 92.6 billion pairs of observations, making summation of pairs computationally prohibitive. We propose a one-step GEE estimator that (1) matches the asymptotic efficiency of the fully iterated GEE; (2) uses a simpler formula to estimate the ICC that avoids summing over all pairs; and (3) completely avoids matrix multiplications and inversions. These three features make the proposed estimator much less computationally intensive, especially with large cluster sizes. A unique contribution of this article is that it expresses the GEE estimating equations incorporating the ICC as a simple sum of vectors and scalars.

Statistically Efficient Thinning of a Markov Chain Sampler

P. 738-744

Art B. Owen

Abstract

It is common to subsample Markov chain output to reduce the storage burden. Geyer shows that discarding $k - 1$ out of every k observations will not improve statistical efficiency, as quantified through variance in a given computational budget. That observation is often taken to mean that thinning Markov chain Monte Carlo (MCMC) output cannot improve statistical efficiency. Here, we suppose that it costs one unit of time to advance a Markov chain and then $\theta > 0$ units of time to compute a sampled quantity of interest. For a thinned process, that cost θ is incurred less often, so it can be advanced through more stages. Here, we provide examples to show that thinning will improve statistical efficiency if θ is large and the sample autocorrelations decay slowly enough. If the lag ℓ □ 1 autocorrelations of a scalar measurement satisfy $\rho_{\ell} > \rho_{\ell+1} > 0$, then there is always a $\theta < \infty$ at which thinning becomes more efficient for averages of that scalar. Many sample autocorrelation functions resemble first order AR(1) processes with $\rho_{\ell} = \rho^{|\ell|}$ for some $-1 < \rho < 1$. For an AR(1) process, it is possible to compute the most efficient subsampling frequency k. The optimal k grows rapidly as ρ increases toward 1. The resulting efficiency gain depends primarily on θ, not ρ. Taking k

= 1 (no thinning) is optimal when $\rho \leq 0$. For $\rho > 0$, it is optimal if and only if $\theta \leq (1 - \rho)^2/(2\rho)$. This efficiency gain never exceeds $1 + \theta$. This article also gives efficiency bounds for autocorrelations bounded between those of two AR(1) processes. Supplementary materials for this article are available online.