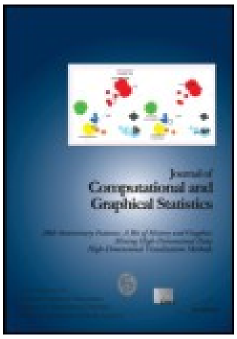

Journal of computational and graphical statistics, ISSN 1061-8600
Volume 24, number 1 (march 2015)

ARTICLES

Advances in Dendrogram Seriation for Application to Visualization

P. 1-25

Denise Earle - Catherine B. Hurley

Abstract

Visualizations of statistical data benefit from systematic ordering of data objects to highlight features and structure. This article concerns ordering via dendrogram seriation based on hierarchical clustering of data objects. It describes DendSer, a general-purpose dendrogram seriation algorithm which when coupled with various seriation cost functions is easily adapted to different visualization settings. Comparisons are made with other dendrogram seriation algorithms and applications are presented. Supplementary materials for this article are available online.

Visualizing Complex Data With Embedded Plots

P. 26-43

Garrett Grolemund - Hadley Wickham

Abstract

This article describes a class of graphs, embedded plots, that are particularly useful for analyzing large and complex datasets. Embedded plots organize a collection of graphs into a larger graphic, which can display more complex relationships than would otherwise be possible. This arrangement provides additional axes, prevents overplotting, and allows for multiple levels of visual summarization. Embedded plots also preprocess complex data into a form suitable for the human cognitive system, which can facilitate comprehension. We illustrate the usefulness of embedded plots with a case study, discuss the practical and cognitive advantages of embedded plots, and demonstrate how to implement embedded plots as a general class within visualization software, something currently unavailable. This article has supplementary material online.

Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation

P. 44-65

Alex Goldstein - Adam Kapelner - Justin Bleich - Emil Pitkin

Abstract

This article presents individual conditional expectation (ICE) plots, a tool for visualizing the model estimated by any supervised learning algorithm. Classical partial dependence plots (PDPs) help visualize the average partial relationship between the predicted response and one or more features. In the presence of substantial interaction effects, the partial response relationship can be heterogeneous. Thus, an average curve, such as the PDP, can obfuscate the complexity of the modeled relationship. Accordingly, ICE plots refine the PDP by graphing the functional relationship between the predicted response and the feature for *individual* observations. Specifically, ICE plots highlight the variation in the fitted values across the range of a covariate, suggesting where and to what extent heterogeneities might exist. In addition to providing a plotting suite for exploratory analysis, we include a visual test for additive structure in the data-generating model. Through simulated examples and real datasets, we demonstrate how ICE plots can shed light on estimated models in ways PDPs cannot. Procedures outlined are available in the R package ICEbox.

Covariance-Guided Mixture Probabilistic Principal Component Analysis (C-MPPCA)

P. 66-83

Chao Han - Scotland Leman - Leanna House

Abstract

To extract information from high-dimensional data efficiently, visualization tools based on data projection methods have been developed and shown useful. However, a single two-dimensional visualization is often insufficient for capturing all or most interesting structures in complex high-dimensional datasets. For this reason, Tipping and Bishop developed mixture probabilistic principal component analysis (MPPCA) that separates data into multiple groups and enables a unique projection per group; that is, one probabilistic principal component analysis (PPCA) data visualization per group. Because the group labels are assigned to observations based on their high-dimensional coordinates, MPPCA works well to reveal homoscedastic structures in data that differ spatially. In the presence of heteroscedasticity, however, MPPCA may still mask noteworthy data structures. We propose a new method called covariance-guided MPPCA (C-MPPCA) that groups subsets of observations based on covariance, not locality, and, similar to MPPCA, displays them using PPCA. PPCA projects data in the dimensions with the highest variances, thus grouping by covariance makes sense and enables some data structures to be visible that were masked originally by MPPCA. We demonstrate the performance of C-MPPCA in an extensive simulation study. We also apply C-MPPCA to a real world dataset. Supplementary materials for this article are available online.

Integrating Data Transformation in Principal Components Analysis

P. 84-103

Mehdi Maadooliat - Jianhua Z. Huang - Jianhua Hu

Abstract

Principal component analysis (PCA) is a popular dimension-reduction method to reduce the complexity and obtain the informative aspects of high-dimensional datasets. When the data distribution is skewed, data transformation is commonly used prior to applying PCA. Such transformation is usually obtained from previous studies, prior knowledge, or trial-and-error. In this work, we develop a model-based method that integrates data transformation in PCA and finds an appropriate data transformation using the maximum profile likelihood. Extensions of the method to handle functional data and missing values are also developed. Several numerical algorithms are provided for efficient computation. The proposed method is illustrated using simulated and real-world data examples. Supplementary materials for this article are available online.

Stable Estimation in Dimension Reduction

P. 104-120

Wenbo Wu - Xiangrong Yin

Abstract

We introduce stable estimation procedures for several aspects of a sufficient dimension-reduction matrix. We first propose a stable method for estimating structural dimension, which only selects the correct directions in the central subspace with no false positive selection. We then provide a Grassmann manifold sparse estimate for the central subspace. By using subsampling, we develop an ensemble method to obtain a stable nonsparse estimate for the central subspace. This ensemble idea is also used to stabilize the choice of the number of slices in sliced inverse methods. Theoretical results are established, and the efficacy of the proposed stable methods is demonstrated by simulation studies and the analysis of Hitters' salary data. Supplementary materials for this article are available online.

High-Dimensional Fused Lasso Regression Using Majorization–Minimization and Parallel Processing

P. 121-153

Donghyeon Yu - Joong-Ho Won - Taehoon Lee - Johan Lim - Sungroh Yoon

Abstract

In this article, we propose a majorization–minimization (MM) algorithm for high-dimensional fused lasso regression (FLR) suitable for parallelization using graphics processing units (GPUs). The MM algorithm is stable and flexible as it can solve the FLR problems with various types of design matrices and penalty structures within a few tens of iterations. We also show that the convergence of the proposed algorithm is guaranteed. We conduct numerical studies to compare our algorithm with other existing algorithms, demonstrating that the proposed MM algorithm is competitive in many settings including the two-dimensional FLR with arbitrary design matrices. The merit of GPU parallelization is also exhibited. Supplementary materials are available online.

A Multiobjective Exploratory Procedure for Regression Model Selection

P. 154-182

Ankur Sinha - Pekka Malo - Timo Kuosmanen

Abstract

Variable selection is recognized as one of the most critical steps in statistical modeling. The problems encountered in engineering and social sciences are commonly characterized by over-abundance of explanatory variables, nonlinearities, and unknown interdependencies between the regressors. An added difficulty is that the analysts may have little or no prior knowledge on the relative importance of the variables. To provide a robust method for model selection, this article introduces the multiobjective genetic algorithm for variable selection (MOGA-VS) that provides the user with an optimal set of regression models for a given dataset. The algorithm considers the regression problem as a two objective task, and explores the Pareto-optimal (best subset) models by preferring those models over the other which have less number of regression coefficients and better goodness of fit. The model exploration can be performed based on in-sample or generalization error minimization. The model selection is proposed to be performed in two steps. First, we generate the frontier of Pareto-optimal regression models by eliminating the dominated models without any user intervention. Second, a decision-making process is executed which allows the user to choose the most preferred model using visualizations and simple metrics. The method has been evaluated on a recently published real dataset on Communities and Crime Within the United States.

Graphical Models for Ordinal Data

P. 183-204

Jian Guo - Elizaveta Levina - George Michailidis - Ji Zhu

Abstract

This article considers a graphical model for ordinal variables, where it is assumed that the data are generated by discretizing the marginal distributions of a latent multivariate Gaussian distribution. The relationships between these ordinal variables are then described by the underlying Gaussian graphical model and can be inferred by estimating the corresponding concentration matrix. Direct estimation of the model is computationally expensive, but an approximate EM-like algorithm is developed to provide an accurate estimate of the parameters at a fraction of the computational cost. Numerical evidence based on simulation studies shows the strong performance of the algorithm, which is also illustrated on datasets on movie ratings and an educational survey.

A Localized Implementation of the Iterative Proportional Scaling Procedure for Gaussian Graphical Models P. 205-229

Ping-Feng Xu - Jianhua Guo - Man-Lai Tang

Abstract

In this article, we propose localized implementations of the iterative proportional scaling (IPS) procedure by the strategy of partitioning cliques for computing maximum likelihood estimations in large Gaussian graphical models. We first divide the set of cliques into several nonoverlapping and nonempty blocks, and then adjust clique marginals in each block locally. Thus, high-order matrix operations can be avoided and the IPS procedure is accelerated. We modify the Swendsen–Wang Algorithm and apply the simulated annealing algorithm to find an approximation to the optimal partition which leads to the least complexity. This strategy of partitioning cliques can also speed up the existing IIPS and IHT procedures. Numerical experiments are presented to demonstrate the competitive performance of our new implementations and strategies.

Learning the Structure of Mixed Graphical Models P. 230-253

Jason D. Lee - Trevor J. Hastie

Abstract

We consider the problem of learning the structure of a pairwise graphical model over continuous and discrete variables. We present a new pairwise model for graphical models with both continuous and discrete variables that is amenable to structure learning. In previous work, authors have considered structure learning of Gaussian graphical models and structure learning of discrete models. Our approach is a natural generalization of these two lines of work to the mixed case. The penalization scheme involves a novel symmetric use of the group-lasso norm and follows naturally from a particular parameterization of the model. Supplementary materials for this article are available online.

Image Denoising by a Local Clustering Framework P. 254-273

Partha Sarathi Mukherjee - Peihua Qiu

Abstract

Images often contain noise due to imperfections in various image acquisition techniques. Noise should be removed from images so that the details of image objects (e.g., blood vessels, inner foldings, or tumors in the human brain) can be clearly seen, and the subsequent image analyses are reliable. With broad usage of images in many disciplines—for example, medical science—image denoising has become an important research area. In the literature, there are many different types of image denoising techniques, most of which aim to preserve image features, such as edges and edge structures, by estimating them explicitly or implicitly. Techniques based on explicit edge detection usually require certain assumptions on the smoothness of the image intensity surface and the edge curves which are often invalid especially when the image resolution is low. Methods that are based on implicit edge detection often use multiresolution smoothing, weighted local smoothing, and so forth. For such methods, the task of determining the correct image resolution or choosing a reasonable weight function is challenging. If the edge structure of an image is complicated or the image has many details, then these methods would blur such details. This article presents a novel image denoising framework based on local clustering of image intensities and adaptive smoothing. The new denoising method can preserve complicated edge structures well even if the image resolution is low. Theoretical properties and numerical studies show that it works well in various applications.

Spatially Weighted Principal Component Analysis for Imaging Classification P. 274-296

Ruixin Guo - Mihye Ahn - Hongtu Zhu Hongtu Zhu

Abstract

The aim of this article is to develop a supervised dimension-reduction framework, called spatially weighted principal component analysis (SWPCA), for high-dimensional imaging classification. Two main challenges in imaging classification are the high dimensionality of the feature space and the complex spatial structure of imaging data. In SWPCA, we introduce two sets of novel weights, including global and local spatial weights, which enable a selective treatment of individual features and incorporation of the spatial structure of imaging data and class label information. We develop an efficient two-stage iterative SWPCA algorithm and its penalized version along with the associated weight determination. We use both simulation studies and real data analysis to evaluate the finite-sample performance of our SWPCA. The results show that SWPCA outperforms several competing principal component analysis (PCA) methods, such as supervised PCA (SPCA), and other competing methods, such as sparse discriminant analysis (SDA).