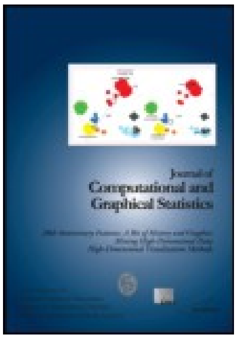

Journal of computational and graphical statistics, ISSN 1061-8600
Volume 26, number 1 (2017)

Regression Models for Multivariate Count Data P. 1-13

Yiwen Zhang, Hua Zhou, Jin Zhou & Wei Sun

Abstract

Data with multivariate count responses frequently occur in modern applications. The commonly used multinomial-logit model is limiting due to its restrictive mean-variance structure. For instance, analyzing count data from the recent RNA-seq technology by the multinomial-logit model leads to serious errors in hypothesis testing. The ubiquity of overdispersion and complicated correlation structures among multivariate counts calls for more flexible regression models. In this article, we study some generalized linear models that incorporate various correlation structures among the counts. Current literature lacks a treatment of these models, partly because they do not belong to the natural exponential family. We study the estimation, testing, and variable selection for these models in a unifying framework. The regression models are compared on both synthetic and real RNA-seq data. Supplementary materials for this article are available online.

Regularized Principal Component Analysis for Spatial Data P. 14-25

Wen-Ting Wang & Hsin-Cheng Huang

Abstract

In many atmospheric and earth sciences, it is of interest to identify dominant spatial patterns of variation based on data observed at p locations and n time points with the possibility that $p > n$. While principal component analysis (PCA) is commonly applied to find the dominant patterns, the eigenimages produced from PCA may exhibit patterns that are too noisy to be physically meaningful when p is large relative to n. To obtain more precise estimates of eigenimages, we propose a regularization approach incorporating smoothness and sparseness of eigenimages, while accounting for their orthogonality. Our method allows data taken at irregularly spaced or sparse locations. In addition, the resulting optimization problem can be solved using the alternating direction method of multipliers, which is easy to implement, and applicable to a large spatial dataset. Furthermore, the estimated eigenfunctions provide a natural basis for representing the underlying spatial process in a spatial random-effects model, from which spatial covariance function estimation and spatial prediction can be efficiently performed using a regularized fixed-rank kriging method. Finally, the effectiveness of the proposed method is demonstrated by several numerical examples.

Sufficient Dimension Reduction and Variable Selection for Large-*p*-Small-*n* Data With Highly Correlated Predictors P. 26-34

Haileab Hilafu & Xiangrong Yin

Abstract

Sufficient dimension reduction (SDR) is a paradigm for reducing the dimension of the predictors without losing regression information. Most SDR methods require inverting the covariance matrix of the predictors. This hinders their use in the analysis of contemporary datasets where the number of predictors exceeds the available sample size and the predictors are highly correlated. To this end, by incorporating the seeded SDR idea and the sequential dimension-reduction framework, we propose a SDR method for high-dimensional data with correlated predictors. The performance of the proposed method is studied via extensive simulations. To demonstrate its use, an application to microarray gene expression data where the response is the production rate of riboflavin (vitamin B_2) is presented.

Variational Approximations for Generalized Linear Latent Variable Models P. 35-43

Francis K. C. Hui, David I. Warton, John T. Ormerod, Viivi Haapaniemi & Sara Taskinen

Abstract

Generalized linear latent variable models (GLLVMs) are a powerful class of models for understanding the relationships among multiple, correlated responses. Estimation, however, presents a major challenge, as the marginal likelihood does not possess a closed form for nonnormal responses. We propose a variational approximation (VA) method for estimating GLLVMs. For the common cases of binary, ordinal, and overdispersed count data, we derive fully closed-form approximations to the marginal log-likelihood function in each case. Compared to other methods such as the expectation-maximization algorithm, estimation using VA is fast and straightforward to implement. Predictions of the latent variables and associated uncertainty estimates are also obtained as part of the estimation process. Simulations show that VA estimation performs similar to or better than some currently available methods, both at predicting the latent variables and estimating their corresponding coefficients. They also show that VA estimation offers dramatic reductions in computation time particularly if the number of correlated responses is large relative to the number of observational units. We apply the variational approach to two datasets, estimating GLLVMs to understanding the patterns of variation in youth gratitude and for constructing ordination plots in bird abundance data. R code for performing VA estimation of GLLVMs is available online. Supplementary materials for this article are available online.

A Marginal Sampler for σ-Stable Poisson–Kingman Mixture Models P.44-53

María Lomelí, Stefano Favaro & Yee Whye Teh

Abstract

We investigate the class of σstable Poisson–Kingman random probability measures (RPMs) in the context of Bayesian nonparametric mixture modeling. This is a large class of discrete RPMs, which encompasses most of the popular discrete RPMs used in Bayesian nonparametrics, such as the Dirichlet process, Pitman–Yor process, the normalized inverse Gaussian process, and the normalized generalized Gamma process. We show how certain sampling properties and marginal characterizations of σstable Poisson–Kingman RPMs can be usefully exploited for devising a Markov chain Monte Carlo (MCMC) algorithm for performing posterior inference with a Bayesian nonparametric mixture model. Specifically, we introduce a novel and efficient MCMC sampling scheme in an augmented space that has a small number of auxiliary variables per iteration. We apply our sampling scheme to a density estimation and clustering tasks with unidimensional and multidimensional datasets, and compare it against competing MCMC sampling schemes. Supplementary materials for this article are available online.

Optimally Adjusted Mixture Sampling and Locally Weighted Histogram Analysis P. 54-65

Zhiqiang Tan

Abstract

Consider the two problems of simulating observations and estimating expectations and normalizing constants for multiple distributions. First, we present a self-adjusted mixture sampling method, which accommodates both adaptive serial tempering and a generalized Wang–Landau algorithm. The set of distributions are combined into a labeled mixture, with the mixture weights depending on the initial estimates of log normalizing constants (or free energies). Then, observations are generated by Markov transitions, and free energy estimates are adjusted online by stochastic approximation. We propose two stochastic approximation schemes by Rao–Blackwellization of the scheme commonly used, and derive the optimal choice of a gain matrix, resulting in the minimum asymptotic variance for free energy estimation, in a simple and feasible form. Second, we develop an offline method, locally weighted histogram analysis, for estimating free energies and expectations, using all the simulated data from multiple distributions by either self-adjusted mixture sampling or other sampling algorithms. This method can be computationally much faster, with little sacrifice of statistical efficiency, than a global method currently used, especially when a large number of distributions are involved. We provide both theoretical results and numerical studies to demonstrate the advantages of the proposed methods.

Imposing Minimax and Quantile Constraints on Optimal Matching in Observational Studies

P. 66-78

Paul R. Rosenbaum

Abstract

Modern methods construct a matched sample by minimizing the total cost of a flow in a network, finding a pairing of treated and control individuals that minimizes the sum of within-pair covariate distances subject to constraints that ensure distributions of covariates are balanced. In aggregate, these methods work well; however, they can exhibit a lack of interest in a small number of pairs with large covariate distances. Here, a new method is proposed for imposing a minimax constraint on a minimum total distance matching. Such a match minimizes the total within-pair distance subject to various constraints including the constraint that the maximum pair difference is as small as possible. In an example with 1391 matched pairs, this constraint eliminates dozens of pairs with moderately large differences in age, but otherwise exhibits the same excellent covariate balance found without this additional constraint. A minimax constraint eliminates edges in the network, and can improve the worst-case time bound for the performance of the minimum cost flow algorithm, that is, a better match from a practical perspective may take less time to construct. The technique adapts ideas for a different problem, the bottleneck assignment problem, whose sole objective is to minimize the maximum within-pair difference; however, here, that objective becomes a constraint on the minimum cost flow problem. The method generalizes. Rather than constrain the maximum distance, it can constrain an order statistic. Alternatively, the method can minimize the maximum difference in propensity scores, and subject to doing that, minimize the maximum robust Mahalanobis distance. An example from labor economics is used to illustrate. Supplementary materials for this article are available online.

Sampling for Conditional Inference on Contingency Tables

P. 79-87

Robert D. Eisinger & Yuguo Chen

Abstract

We propose new sequential importance sampling methods for sampling contingency tables with given margins. The proposal for each method is based on asymptotic approximations to the number of tables with fixed margins. These methods generate tables that are very close to the uniform distribution. The tables, along with their importance weights, can be used to approximate the null distribution of test statistics and calculate the total number of tables. We apply the methods to a number of examples and demonstrate an improvement over other methods in a variety of real problems. Supplementary materials are available online.

Circulant Embedding of Approximate Covariances for Inference From Gaussian Data on Large Lattices

P. 88-97

Joseph Guinness & Montserrat Fuentes

Abstract

Recently proposed computationally efficient Markov chain Monte Carlo (MCMC) and Monte Carlo expectation–maximization (EM) methods for estimating covariance parameters from lattice data rely on successive imputations of values on an embedding lattice that is at least two times larger in each dimension. These methods can be considered exact in some sense, but we demonstrate that using such a large number of imputed values leads to slowly converging Markov chains and EM algorithms. We propose instead the use of a discrete spectral approximation to allow for the implementation of these methods on smaller embedding lattices. While our methods are approximate, our examples indicate that the error introduced by this approximation is small compared to the Monte Carlo errors present in long Markov chains or many iterations of Monte Carlo EM algorithms. Our results are demonstrated in simulation studies, as well as in numerical studies that explore both increasing domain and fixed domain asymptotics. We compare the

exact methods to our approximate methods on a large satellite dataset, and show that the approximate methods are also faster to compute, especially when the aliased spectral density is modeled directly. Supplementary materials for this article are available online.

An Inversion-Free Estimating Equations Approach for Gaussian Process Models P. 98-107

Mihai Anitescu, Jie Chen & Michael L. Stein

Abstract

One of the scalability bottlenecks for the large-scale usage of Gaussian processes is the computation of the maximum likelihood estimates of the parameters of the covariance matrix. The classical approach requires a Cholesky factorization of the dense covariance matrix for each optimization iteration. In this work, we present an estimating equations approach for the parameters of zero-mean Gaussian processes. The distinguishing feature of this approach is that no linear system needs to be solved with the covariance matrix. Our approach requires solving an optimization problem for which the main computational expense for the calculation of its objective and gradient is the evaluation of traces of products of the covariance matrix with itself and with its derivatives. For many problems, this is an $O(n\log n)$ effort, and it is always no larger than $O(n^2)$. We prove that when the covariance matrix has a bounded condition number, our approach has the same convergence rate as does maximum likelihood in that the Godambe information matrix of the resulting estimator is at least as large as a fixed fraction of the Fisher information matrix. We demonstrate the effectiveness of the proposed approach on two synthetic examples, one of which involves more than 1 million data points.

Bayesian and Maximum Likelihood Estimation for Gaussian Processes on an Incomplete Lattice P. 108-120

Jonathan R. Stroud, Michael L. Stein & Shaun Lysen

Abstract

This article proposes a new approach for Bayesian and maximum likelihood parameter estimation for stationary Gaussian processes observed on a large lattice with missing values. We propose a Markov chain Monte Carlo approach for Bayesian inference, and a Monte Carlo expectation-maximization algorithm for maximum likelihood inference. Our approach uses data augmentation and circulant embedding of the covariance matrix, and provides likelihood-based inference for the parameters and the missing data. Using simulated data and an application to satellite sea surface temperatures in the Pacific Ocean, we show that our method provides accurate inference on lattices of sizes up to 512 × 512, and is competitive with two popular methods: composite likelihood and spectral approximations.

Bayesian Model Assessment in Joint Modeling of Longitudinal and Survival Data With Applications to Cancer Clinical Trials P. 121-133

Danjie Zhang, Ming-Hui Chen, Joseph G. Ibrahim, Mark E. Boye & Wei Shen

Abstract

Joint models for longitudinal and survival data are routinely used in clinical trials or other studies to assess a treatment effect while accounting for longitudinal measures such as patient-reported outcomes. In the Bayesian framework, the deviance information criterion (DIC) and the logarithm of the pseudo-marginal likelihood (LPML) are two well-known Bayesian criteria for comparing joint models. However, these criteria do not provide separate assessments of each component of the joint model. In this article, we develop a novel decomposition of DIC and LPML to assess the fit of the longitudinal and survival components of the joint model, separately. Based on this decomposition, we then propose new Bayesian model assessment criteria, namely, ΔDIC and ΔLPML, to determine the importance and contribution of the longitudinal (survival) data to the model fit of the survival (longitudinal) data. Moreover, we develop an efficient Monte Carlo method for computing the conditional predictive ordinate statistics in the joint modeling setting. A simulation study is conducted to examine the empirical performance of the proposed criteria and the proposed methodology is further applied to a case study in mesothelioma. Supplementary materials for this article are available online.

Computationally Efficient Changepoint Detection for a Range of Penalties

P. 134-143

Kaylea Haynes, Idris A. Eckley & Paul Fearnhead

Abstract

In the multiple changepoint setting, various search methods have been proposed, which involve optimizing either a constrained or penalized cost function over possible numbers and locations of changepoints using dynamic programming. Recent work in the penalized optimization setting has focused on developing an exact pruning-based approach that, under certain conditions, is linear in the number of data points. Such an approach naturally requires the specification of a penalty to avoid under/over-fitting. Work has been undertaken to identify the appropriate penalty choice for data-generating processes with known distributional form, but in many applications the model assumed for the data is not correct and these penalty choices are not always appropriate. To this end, we present a method that enables us to find the solution path for all choices of penalty values across a continuous range. This permits an evaluation of the various segmentations to identify a suitable penalty choice. The computational complexity of this approach can be linear in the number of data points and linear in the difference between the number of changepoints in the optimal segmentations for the smallest and largest penalty values. Supplementary materials for this article are available online.

Principal Nested Spheres for Time-Warped Functional Data Analysis

P. 144-151

Qunqun Yu, Xiaosun Lu & J. S. Marron

Abstract

There are often two important types of variation in functional data: the horizontal (or phase) variation and the vertical (or amplitude) variation. These two types of variation have been appropriately separated and modeled through a domain warping method (or curve registration) based on the Fisher–Rao metric. This article focuses on the analysis of the horizontal variation, captured by the domain warping functions. The square-root velocity function representation transforms the manifold of the warping functions to a Hilbert sphere. Motivated by recent results on manifold analogs of principal component analysis, we propose to analyze the horizontal variation via a principal nested spheres approach. Compared with earlier approaches, such as approximating tangent plane principal component analysis, this is seen to be an efficient and interpretable approach to decompose the horizontal variation in both simulated and real data examples.

Interweaving Markov Chain Monte Carlo Strategies for Efficient Estimation of Dynamic Linear Models

P. 152-159

Matthew Simpson, Jarad Niemi & Vivekananda Roy

Abstract

In dynamic linear models (DLMs) with unknown fixed parameters, a standard Markov chain Monte Carlo (MCMC) sampling strategy is to alternate sampling of latent states conditional on fixed parameters and sampling of fixed parameters conditional on latent states. In some regions of the parameter space, this standard data augmentation (DA) algorithm can be inefficient. To improve efficiency, we apply the interweaving strategies of Yu and Meng to DLMs. For this, we introduce three novel alternative DAs for DLMs: the scaled errors, wrongly scaled errors, and wrongly scaled disturbances. With the latent states and the less well known scaled disturbances, this yields five unique DAs to employ in MCMC algorithms. Each DA implies a unique MCMC sampling strategy and they can be combined into interweaving and alternating strategies that improve MCMC efficiency. We assess these strategies using the local level model and demonstrate that several strategies improve efficiency relative to the standard approach and the most efficient strategy interweaves the scaled errors and scaled disturbances. Supplementary materials are available online for this article.

How Many Communities Are There? P. 171-181

D. Franco Saldaña, Yi Yu & Yang Feng

Abstract

Stochastic blockmodels and variants thereof are among the most widely used approaches to community detection for social networks and relational data. A stochastic blockmodel partitions the nodes of a network into disjoint sets, called communities. The approach is inherently related to clustering with mixture models; and raises a similar model selection problem for the number of communities. The Bayesian information criterion (BIC) is a popular solution, however, for stochastic blockmodels, the conditional independence assumption given the communities of the endpoints among different edges is usually violated in practice. In this regard, we propose composite likelihood BIC (CL-BIC) to select the number of communities, and we show it is robust against possible misspecifications in the underlying stochastic blockmodel assumptions. We derive the requisite methodology and illustrate the approach using both simulated and real data. Supplementary materials containing the relevant computer code are available online.

Efficient Computation of the Joint Sample Frequency Spectra for Multiple Populations P. 182-194

John A. Kamm, Jonathan Terhorst & Yun S. Song

Abstract

A wide range of studies in population genetics have employed the sample frequency spectrum (SFS), a summary statistic which describes the distribution of mutant alleles at a polymorphic site in a sample of DNA sequences and provides a highly efficient dimensional reduction of large-scale population genomic variation data. Recently, there has been much interest in analyzing the joint SFS data from multiple populations to infer parameters of complex demographic histories, including variable population sizes, population split times, migration rates, admixture proportions, and so on. SFS-based inference methods require accurate computation of the expected SFS under a given demographic model. Although much methodological progress has been made, existing methods suffer from numerical instability and high computational complexity when multiple populations are involved and the sample size is large. In this article, we present new analytic formulas and algorithms that enable accurate, efficient computation of the expected joint SFS for thousands of individuals sampled from hundreds of populations related by a complex demographic model with arbitrary population size histories (including piecewise-exponential growth). Our results are implemented in a new software package called *momi* (MOran Models for Inference). Through an empirical study, we demonstrate our improvements to numerical stability and computational complexity.

An Augmented ADMM Algorithm With Application to the Generalized Lasso Problem P. 195-204

Yunzhang Zhu

Abstract

In this article, we present a fast and stable algorithm for solving a class of optimization problems that arise in many statistical estimation procedures, such as *sparse fused lasso over a graph*, *convex clustering*, and *trend filtering*, among others. We propose a so-called augmented *alternating direction methods of multipliers* (ADMM) algorithm to solve this class of problems. Compared to a standard ADMM algorithm, our proposal significantly reduces the computational cost at each iteration while maintaining roughly the same overall convergence speed. We also consider a new varying penalty scheme for the ADMM algorithm, which could further accelerate the convergence, especially when solving a sequence of problems with tuning parameters of different scales. Extensive numerical experiments on the sparse fused lasso problem show that the proposed algorithm is more efficient than the standard ADMM and two other existing state-of-the-art specialized algorithms. Finally, we discuss a possible extension and some interesting connections to two well-known algorithms. Supplementary materials for the article are available online.

Fast Tree Inference With Weighted Fusion Penalties P. 205-216

Julien Chiquet, Pierre Gutierrez & Guillem Rigaill

Abstract

Given a dataset with many features observed in a large number of conditions, it is desirable to fuse and aggregate conditions that are similar to ease the interpretation and extract the main characteristics of the data. This article presents a multidimensional fusion penalty framework to address this question when the number of conditions are large. If the fusion penalty is encoded by an ℓ_q-norm, we prove for uniform weights that the path of solutions is a tree that is suitable for interpretability. For the ℓ_1 and ℓ_∞-norms, the path is piecewise linear and we derive a homotopy algorithm to recover exactly the whole tree structure. For weighted ℓ_1-fusion penalties, we demonstrate that distance-decreasing weights lead to balanced tree structures. For a subclass of these weights that we call "exponentially adaptive," we derive an $\mathcal{O}(n\log(n))\mathrm{O}(n\log(n))$ homotopy algorithm and we prove an asymptotic oracle property. This guarantees that we recover the underlying structure of the data efficiently both from a statistical and a computational point of view. We provide a fast implementation of the homotopy algorithm for the single feature case, as well as an efficient embedded cross-validation procedure that takes advantage of the tree structure of the path of solutions. Our proposal outperforms its competing procedures on simulations both in terms of timings and prediction accuracy. As an example we consider phenotypic data: given one or several traits, we reconstruct a balanced tree structure and assess its agreement with the known taxonomy. Supplementary materials for this article are available online.

Discrete Approximation of a Mixture Distribution via Restricted Divergence P. 217-222

Christian Röver & Tim Friede

Abstract

Mixture distributions arise in many application areas, for example, as marginal distributions or convolutions of distributions. We present a method of constructing an easily tractable discrete mixture distribution as an approximation to a mixture distribution with a large to infinite number, discrete or continuous, of components. The proposed DIRECT (divergence restricting conditional tesselation) algorithm is set up such that a prespecified precision, defined in terms of Kullback–Leibler divergence between true distribution and approximation, is guaranteed. Application of the algorithm is demonstrated in two examples. Supplementary materials for this article are available online.

Accurate Small Tail Probabilities of Sums of iid Lattice-Valued Random Variables via FFT P. 223-229

Huon Wilson & Uri Keich

Abstract

Accurately computing very small tail probabilities of a sum of independent and identically distributed lattice-valued random variables is numerically challenging. The only general purpose algorithms that can guarantee the desired accuracy have a quadratic runtime complexity that is often too slow. While fast Fourier transform (FFT)-based convolutions have an essentially linear runtime complexity, they can introduce overwhelming roundoff errors. We present sisFFT (segmented iterated shifted FFT), which harnesses the speed of FFT while retaining control of the relative error of the computed tail probability. We rigorously prove the method's accuracy and we empirically demonstrate its significant speed advantage over existing accurate methods. Finally, we show that sisFFT sacrifices very little, if any, speed when FFT-based convolution is sufficiently accurate to begin with. Supplementary material is available online.