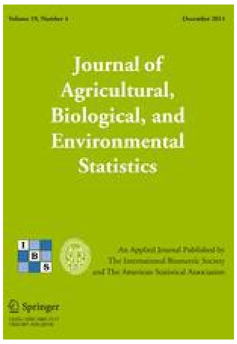

Journal of agricultural, biological, and environmental statistics, ISSN 1085-7117
Volume 20, number 1 (march 2015)

Estimating Abundance from Counts in Large Data Sets of Irregularly Spaced Plots using Spatial Basis Functions

P. 1-27

Jay M. Ver Hoef - John K. Jansen

Abstract

Monitoring plant and animal populations is an important goal for both academic research and management of natural resources. Successful management of populations often depends on obtaining estimates of their mean or total over a region. The basic problem considered in this paper is the estimation of a total from a sample of plots containing count data, but the plot placements are spatially irregular and non-randomized. Our application had counts from thousands of irregularly spaced aerial photo images. We used change-of-support methods to model counts in images as a realization of an inhomogeneous Poisson process that used spatial basis functions to model the spatial intensity surface. The method was very fast and took only a few seconds for thousands of images. The fitted intensity surface was integrated to provide an estimate from all unsampled areas, which is added to the observed counts. The proposed method also provides a finite area correction factor to variance estimation. The intensity surface from an inhomogeneous Poisson process tends to be too smooth for locally clustered points, typical of animal distributions, so we introduce several new overdispersion estimators due to poor performance of the classic one. We used simulated data to examine estimation bias and to investigate several variance estimators with overdispersion. A real example is given of harbor seal counts from aerial surveys in an Alaskan glacial fjord.

Analysing Mark–Recapture–Recovery Data in the Presence of Missing Covariate Data Via Multiple Imputation

P. 28-46

Hannah Worthington - Ruth King

Abstract

We consider mark–recapture–recovery data with additional individual time-varying continuous covariate data. For such data it is common to specify the model parameters, and in particular the survival probabilities, as a function of these covariates to incorporate individual heterogeneity. However, an issue arises in relation to missing covariate values, for (at least) the times when an individual is not observed, leading to an analytically intractable likelihood. We propose a two-step multiple imputation approach to obtain estimates of the demographic parameters. Firstly, a model is fitted to only the observed covariate values. Conditional on the fitted covariate model, multiple “complete” datasets are generated (i.e. all missing covariate values are imputed). Secondly, for each complete dataset, a closed form complete data likelihood can be maximised to obtain estimates of the model parameters which are subsequently combined to obtain an overall estimate of the parameters. Associated standard errors and 95% confidence intervals are obtained using a nonparametric bootstrap. A simulation study is undertaken to assess the performance of the proposed two-step approach. We apply the method to data collected on a well-studied population of Soay sheep and compare the results with a Bayesian data augmentation approach. Supplementary materials accompanying this paper appear on-line.

Bayesian Nonparametric Models of Circular Variables Based on Dirichlet Process Mixtures of Normal Distributions P. 47-64

Gabriel Nuñez-Antonio

Abstract

This article introduces two new Bayesian nonparametric models for circular data based on Dirichlet process (DP) mixtures of normal distributions. The first model is a projected DP mixture of bivariate normals and the second approach is based on a wrapped DP mixture of normal distributions. We show how to carry out inference for these models based on a slice sampling scheme and introduce an approach to estimating a variant of the deviance information criterion which is appropriate in the context of latent variable models. Our models are then compared with both simulated and real data examples.

Assessing Assay Variability of Pesticide Metabolites in the Presence of Heavy Left-Censoring P. 65-82

Haiying Chen - Sara A. Quandt - Dana Boyd Barr

Abstract

Assessing assay variability for field samples in environmental research is challenging, since a quantitative assay is typically constrained by a lower limit of detection. The purpose of this paper is to compare three parametric models for assessing assay variability using duplicate data subject to heavy left-censoring. Efron information criterion (EIC) and Bayesian information criterion (BIC) are used to aid in model selections. Distributional parameter estimates are obtained using maximum likelihood estimation for bivariate lognormal, bivariate zero-inflated lognormal, and bivariate 3-component mixture models. We illustrate a practical application using duplicate pesticide data from the Community Participatory Approach to Measuring Farmworker Pesticide Exposure (PACE3) study. Furthermore, a simulation study is conducted to empirically evaluate the performance of the three models. The results from PACE3 indicate that the bivariate zero-inflated lognormal model is fairly competitive based on EIC or BIC. Further, total variability for the lognormal component can be decomposed into between-subject and within-subject variance based on this model. Assay variability estimates such as within-subject coefficient variation, minimum detectable change, and probability of *k*-fold difference can be easily derived under the bivariate zero-inflated lognormal model. Additionally, the assay variability is rather large for the PACE3 data. Therefore, apparent longitudinal change in pesticide exposure should be examined cautiously in the context of substantial assay variability.

Supplementary materials accompanying this paper appear online.

Maximum Pairwise Pseudo-likelihood Estimation of the Covariance Matrix from Left-Censored Data P. 83-99

Michael P. Jones - Sarah S. Perry

Abstract

Toxicological studies often depend on laboratory assays that have thresholds below which environmental pollutants cannot be measured with accuracy. Exposure levels below this limit of detection may well be toxic and hence it is vital to use data analytic methods that handle such left-censored data with as little estimation bias as possible. In an on-going study for which our methodology is developed, levels of residential exposure to polychlorinated biphenyls (PCBs) and the interrelationships of their subtypes (congeners) are characterized. In any given sample many of the congeners may fall below the detection limit. The main problem tackled in this paper is estimation of mean exposure levels and corresponding covariance and correlation matrices for a large number of potentially left-censored measures that have very low bias and are computationally feasible. The proposed methods are likelihood based, using marginal likelihoods for means and variances and pairwise pseudo-likelihoods for correlations and covariances. In the simple bivariate case, head-to-head comparisons show the proposed methods to be computationally more stable than ordinary maximum likelihood estimates (MLEs) and still maintain comparable bias. When the number of variables is much larger than 2, the proposed methods are far more computationally feasible than MLE. Furthermore, they exhibit much less bias when compared to popular imputation procedures. Analysis of the PCB data uncovered interesting correlational structures.

Characterization of Weighted Quantile Sum Regression for Highly Correlated Data in a Risk Analysis Setting

P. 100-120

Caroline Carrico - Chris Gennings

Abstract

In risk evaluation, the effect of mixtures of environmental chemicals on a common adverse outcome is of interest. However, due to the high dimensionality and inherent correlations among chemicals that occur together, the traditional methods (e.g. ordinary or logistic regression) suffer from collinearity and variance inflation, and shrinkage methods have limitations in selecting among correlated components. We propose a weighted quantile sum (WQS) approach to estimating a body burden index, which identifies “bad actors” in a set of highly correlated environmental chemicals. We evaluate and characterize the accuracy of WQS regression in variable selection through extensive simulation studies through sensitivity and specificity (i.e., ability of the WQS method to select the bad actors correctly and not incorrect ones). We demonstrate the improvement in accuracy this method provides over traditional ordinary regression and shrinkage methods (lasso, adaptive lasso, and elastic net). Results from simulations demonstrate that WQS regression is accurate under some environmentally relevant conditions, but its accuracy decreases for a fixed correlation pattern as the association with a response variable diminishes. Nonzero weights (i.e., weights exceeding a selection threshold parameter) may be used to identify bad actors; however, components within a cluster of highly correlated active components tend to have lower weights, with the sum of their weights representative of the set.

Robust Joint Non-linear Mixed-Effects Models and Diagnostics for Censored HIV Viral Loads with CD4 Measurement Error

P. 121-139

Dipankar Bandyopadhyay - Luis M. Castro

Abstract

Despite technological advances in efficiency enhancement of quantification assays, biomedical studies on HIV RNA collect viral load responses that are often subject to detection limits. Moreover, some related covariates such as CD4 cell count may be often measured with errors. Censored non-linear mixed-effects models are routinely used to analyze this type of data and are based on normality assumptions for the between-subject and within-subject random terms. However, derived inference may not be robust when the underlying normality assumptions are questionable, especially in presence of skewness and heavy tails. In this article, we address these issues simultaneously under a Bayesian paradigm through joint modeling of the response and covariate processes using an attractive class of skew-normal independent densities. The methodology is illustrated using a case study on longitudinal HIV viral loads. Diagnostics for outlier detection is immediate from the MCMC output. Both simulation and real data analysis reveal the advantage of the proposed models in providing robust inference under non-normality situations commonly encountered in HIV/AIDS or other clinical studies.

Estimation of General Multistage Models From Cohort Data

P. 140-155

Perry de Valpine - Jonas Knape

Abstract

Many systems involve progression through a series of distinct stages, such as disease or developmental stages. In ecological studies, often individuals such as small arthropods cannot be marked, so data are collected on cohort development. Multistage models for unmarked cohort data use a distribution for each stage duration and possibly stage-specific mortality rates. We generalize previous models and present computational methods for smoothed maximum likelihood estimation. The general model allows arbitrary distribution assumptions, stage-specific mortality, unobserved stages, and correlations between stage durations using Gaussian copulas. Monte Carlo integration of the stage distributions is used to approximate the probabilities needed for the likelihood. We establish a heuristic smoothing step for the simulated probabilities that yields a smooth approximate likelihood surface. For the case of classic grasshopper cohort data, we demonstrate AIC model selection to determine which among past arbitrary constraints are actually justified by the data. Finally, we demonstrate how estimates of stage distribution parameters depend on the unknown stage correlations.

Pseudo-likelihood Estimation of Multivariate Normal Parameters in the Presence of Left-Censored Data

P. 183-204

Heather J. Hoffman - Robert E. Johnson

Abstract

Environmental data often include left-censored values reported to be less than some limit of detection (LOD). While simple imputation of a specific value such as LOD/2 is common, maximum likelihood methods accounting for censoring provide alternate ways of analyzing such data. Concentration levels of contaminants in water, for example, are typically modeled with normal or lognormal distributions. Corresponding maximum likelihood estimates (MLEs) of means and variances in univariate analyses can be obtained from standard software packages; however, multivariate analyses may be more appropriate when multiple measurements come from the same entity. For example, measures of several dissolved trace metals may be derived from freshwater stream samples. In less-polluted areas, one or more of these measures fall below the LOD. An index of overall contamination may be formed as a linear function of these measures. The desire to estimate this index led to the need to estimate the parameters in the presence of nondetects, which led to our proposed method. We propose a pseudo-likelihood method utilizing pairs of variables that provides MLEs of mean and unstructured covariance parameters corresponding to a multivariate normal or lognormal distribution in the presence of left-censored data. In conducting hypothesis tests and estimating functions of MLEs with standard errors, we apply this method to trace metal concentration data collected from freshwater streams across Virginia.