Biometrika, ISSN 0006-3444
Volume 102, number 4 (december 2015)

Optimal multiple testing under a Gaussian prior on the effect sizes

P.753-766

Edgar Dobriban, Kristen Fortney, Stuart K. Kim, Art B. Owen

Abstract

We develop a new method for large-scale frequentist multiple testing with Bayesian prior information. We find optimal *p*-value weights that maximize the average power of the weighted Bonferroni method. Due to the nonconvexity of the optimization problem, previous methods that account for uncertain prior information are suitable for only a small number of tests. For a Gaussian prior on the effect sizes, we give an efficient algorithm that is guaranteed to find the optimal weights nearly exactly. Our method can discover new loci in genome-wide association studies and compares favourably to competitors. An open-source implementation is available.

Strong control of the familywise error rate in observational studies that discover effect modification by exploratory methods

P. 767-782

Jesse Y. Hsu, José R. Zubizarreta, Dylan S. Small, Paul R. Rosenbaum

Abstract

An effect modifier is a pretreatment covariate that affects the magnitude of the treatment effect or its stability. When there is effect modification, an overall test that ignores an effect modifier may be more sensitive to unmeasured bias than a test that combines results from subgroups defined by the effect modifier. If there is effect modification, one would like to identify specific subgroups for which there is evidence of effect that is insensitive to small or moderate biases. In this paper, we propose an exploratory method for discovering effect modification, and combine it with a confirmatory method of simultaneous inference that strongly controls the familywise error rate in a sensitivity analysis, despite the fact that the groups being compared are defined empirically. A new form of matching, strength-*k* matching, permits a search through more than *k* covariates for effect modifiers, in such a way that no pairs are lost, provided that at most *k* covariates are selected to group the pairs. In a strength-*k* match, each set of *k* covariates is exactly balanced, although a set of more than *k* covariates may exhibit imbalance. We apply the proposed method to study the effects of the earthquake that struck Chile in 2010.

Consistent testing for recurrent genomic aberrations

P. 783-796

V. Walter, F. A. Wright, A. B. Nobel

Abstract

We consider the detection and identification of recurrent departures from stationary behaviour in genomic or similarly arranged data containing measurements at an ordered set of variables. Our primary focus is on departures that occur only at a single variable, or within a small window of contiguous variables, but involve more than one sample. This encompasses the identification of aberrant markers in genome-wide measurements of DNA copy number and DNA methylation, as well as meta-analyses of genome-wide association studies. We propose and analyse a cyclic shift-based procedure for testing recurrent departures from stationarity. Our analysis establishes the consistency of cyclic shift *p*-values for datasets with a fixed set of samples as the number of observed variables tends to infinity, under the assumption that each sample is an independent realization of a stationary Markov chain. Our results apply to any test statistic satisfying a simple invariance

condition

Direct estimation of the mean outcome on treatment when treatment assignment and discontinuation compete

P. 797-807

Xin Lu and, Brent A. Johnson

Abstract

Several authors have investigated the challenges of statistical analyses and inference in the presence of early treatment termination, including a loss of efficiency in randomized controlled trials and a connection to dynamic regimes in observational studies. Popular estimation strategies for causal estimands in dynamic regimes lend themselves to studies where treatment is assigned at a finite number of points and the extension to continuous treatment assignment is nontrivial. We re-examine this from a different perspective and propose a new estimator for the mean outcome of a target treatment length policy that does not involve a treatment model. Because this strategy avoids modelling the treatment assignment mechanism, the estimator works for both discrete and continuous treatment length data and eschews bias and imprecision that arise as a result of coarsening continuous time data into intervals. We show how the competition of treatment length assignment and terminating event lead to a competing risks problem. We exemplify the direct estimator through numerical studies and the analysis of two real datasets. When all modelling assumptions for both the direct and inverse weighted estimators are correct, our simulation studies suggest that the direct estimator is more precise.

Bayesian inference for partially observed stochastic differential equations driven by fractional Brownian motion

P. 809-827

A. Beskos, J. Dureau, K. Kalogeropoulos

Abstract

We consider continuous-time diffusion models driven by fractional Brownian motion. Observations are assumed to possess a nontrivial likelihood given the latent path. Due to the non-Markovian and high-dimensional nature of the latent path, estimating posterior expectations is computationally challenging. We present a reparameterization framework based on the Davies and Harte method for sampling stationary Gaussian processes and use it to construct a Markov chain Monte Carlo algorithm that allows computationally efficient Bayesian inference. The algorithm is based on a version of hybrid Monte Carlo simulation that delivers increased efficiency when used on the high-dimensional latent variables arising in this context. We specify the methodology on a stochastic volatility model, allowing for memory in the volatility increments through a fractional specification. The method is demonstrated on simulated data and on the S&P500/VIX time series. In the latter case, the posterior distribution favours values of the Hurst parameter smaller than 1/2, pointing towards medium-range dependence.

Shared kernel Bayesian screening

P. 829-842

Eric F. Lock, David B. Dunson

Abstract

This article concerns testing for equality of distribution between groups. We focus on screening variables with shared distributional features such as common support, modes and patterns of skewness. We propose a Bayesian testing method using kernel mixtures, which improves performance by borrowing information across the different variables and groups through shared kernels and a common probability of group differences. The inclusion of shared kernels in a finite mixture, with Dirichlet priors on the weights, leads to a simple framework for testing that scales well for high-dimensional data. We provide closed asymptotic forms for the posterior probability of equivalence in two groups and prove consistency under model misspecification. The method is applied to DNA methylation array data from a breast cancer study, and compares favourably to competitors when Type I error is estimated via permutation.

Singular value shrinkage priors for Bayesian prediction P. 843-854

Takeru Matsuda, Fumiyasu Komaki

Abstract

We develop singular value shrinkage priors for the mean matrix parameters in the matrix-variate normal model with known covariance matrices. Our priors are superharmonic and put more weight on matrices with smaller singular values. They are a natural generalization of the Stein prior. Bayes estimators and Bayesian predictive densities based on our priors are minimax and dominate those based on the uniform prior in finite samples. In particular, our priors work well when the true value of the parameter has low rank.

Efficient inference and simulation for elliptical Pareto processes P. 855-870

Emeric Thibaud, Thomas Opitz

Abstract

Recent advances in extreme value theory have established ℓ-Pareto processes as the natural limits for extreme events defined in terms of exceedances of a risk functional. In this paper we provide methods for the practical modelling of data based on a tractable yet flexible dependence model. We introduce the class of elliptical ℓ-Pareto processes, which arise as the limits of threshold exceedances of certain elliptical processes characterized by a correlation function and a shape parameter. An efficient inference method based on maximizing a full likelihood with partial censoring is developed. Novel procedures for exact conditional and unconditional simulation are proposed. These ideas are illustrated using precipitation extremes in Switzerland.

Nonparametric methods for group testing data, taking dilution into account P. 871-887

A. Delaigle, P. Hall

Abstract

Group testing methods are used widely to assess the presence of a contaminant, based on measurements of the concentration of a biomarker, for example to test the presence of a disease in pooled blood samples. The test would be perfect if it produced a positive result whenever the contaminant was present, and a negative result otherwise. However, in practice the test is always at least somewhat imperfect, for example because it is sensitive to the proportion of contaminated items in the group, rather than to the sheer existence of one or more contaminated items. We develop a nonparametric method for accommodating this dilution effect. Our approach allows us to estimate, under minimal assumptions, the probability $m(x)$ that an item is contaminated, conditional on the value x of an explanatory variable, and to estimate the probability, q, that an individual chosen at random is disease free, and the specificity Sp, and the sensitivity Se, of the test. These are all ill-posed problems, where poor convergence rates are usually encountered, but despite this, our estimators of q, Sp and Se are root-N consistent, where N denotes the total number of individuals in all the groups, and our estimator of $m(x)$ converges at the rate it would enjoy if q, Sp and Se were known.

A new specification of generalized linear models for categorical responses P. 889-906

J. Peyhardi, C. Trottier, Y. Guédon

Abstract

Many regression models for categorical responses have been introduced, motivated by different paradigms, but it is difficult to compare them because of their different specifications. In this paper we propose a unified specification of regression models for categorical responses, based on a decomposition of the link function into an inverse continuous cumulative distribution function and a ratio of probabilities. This allows us to define a new family of reference models for nominal responses, comparable to the families of adjacent, cumulative and sequential models for ordinal responses. A new equivalence between cumulative and sequential models is shown. Invariances under permutations of the categories are studied for each family of models. We introduce a reversibility property that distinguishes adjacent and cumulative models from sequential models. The new family of reference models is tested on three benchmark classification datasets.

Diagnostic measures for the Cox regression model with missing covariates

P. 907-923

Hongtu Zhu, Joseph G. Ibrahim, Ming-Hui Chen

Abstract

We investigate diagnostic measures for assessing the influence of observations and model misspecification on the Cox regression model when there are missing covariate data. Our diagnostics include case-deletion measures, conditional martingale residuals, and score residuals. The Q-distance is introduced to examine the effects of deleting individual observations on the estimates of finite- and infinite-dimensional parameters. Conditional martingale residuals are used to construct goodness-of-fit statistics for testing misspecification of the model assumptions. A resampling method is developed to approximate the *p*-values of the goodness-of-fit statistics. We conduct simulation studies to evaluate our methods, and analyse a real dataset to illustrate their use.

General weighted optimality of designed experiments

P. 925-935

J. W. Stallings, J. P. Morgan

Abstract

The standard approach to finding optimal experimental designs employs conventional measures of design efficacy, such as the *A*, *E*, and *D*-criterion, that assume equal interest in all estimable functions of model parameters. This paper develops a general theory for weighted optimality, allowing precise design selection according to expressed relative interest in different functions in the estimation space. The approach employs a very general class of matrix-specified weighting schemes that produce easily interpretable weighted optimality criteria. In particular, for any set of estimable functions, and any selected corresponding weights, analogs of standard optimality criteria are found that guide design selection according to the weighted variances of estimators of those particular functions. The results are applied to solve the *A*-optimal design problem for baseline factorial effects in unblocked experiments.

Designing dose-finding studies with an active control for exponential families

P. 937-950

H. Dette, K. Kettelhake, F. Bretz

Abstract

Optimal design of dose-finding studies with an active control has only been considered in the literature for regression models with normally distributed errors and known variances, where the focus is on estimating the smallest dose that achieves the same treatment effect as the active control. This paper discusses such dose-finding studies from a broader perspective. We consider a general class of optimality criteria and models arising from an exponential family. Optimal designs are constructed for several situations and their efficiency is illustrated with examples.

Locally optimal designs for errors-in-variables models

P. 951-958

M. Konstantinou, H. Dette

Abstract

We consider the construction of optimal designs for nonlinear regression models when there are measurement errors in the covariates. Corresponding approximate design theory is developed for maximum likelihood and least-squares estimation, with the latter leading to nonconcave optimization problems. Analytical characterizations of the locally D-optimal saturated designs are provided for the Michaelis–Menten, $E\max$ and exponential regression models. Through concrete applications, we illustrate how measurement errors in the covariates affect the optimal choice of design and show that the locally D-optimal saturated designs are highly efficient for relatively small misspecifications of the parameter values.

Space-filling properties of good lattice point sets P. 959-966

Yongdao Zhou, Hongquan Xu

Abstract

We study space-filling properties of good lattice point sets and obtain some general theoretical results. We show that linear level permutation does not decrease the minimum distance for good lattice point sets, and we identify several classes of such sets with large minimum distance. Based on good lattice point sets, some maximin distance designs are also constructed.

Optimal two-level choice designs for any number of choice sets P. 967-973

Rakhi Singh, Feng-Shun Chai, Ashish Das

Abstract

For two-level choice experiments, we obtain a simple form of the information matrix of a choice design for estimating the main effects, and provide D- and MS-optimal paired choice designs with distinct choice sets under the main effects model for any number of choice sets. It is shown that the optimal designs under the main effects model are also optimal under the broader main effects model. We find that optimal choice designs with a choice set size two often outperform their counterparts with larger choice set sizes.

Changepoint estimation: another look at multiple testing problems P. 974-980

Hongyuan Cao, Wei Biao Wu

Abstract

We consider large scale multiple testing for data that have locally clustered signals. With this structure, we apply techniques from changepoint analysis and propose a boundary detection algorithm so that the clustering information can be utilized. Consequently the precision of the multiple testing procedure is substantially improved. We study tests with independent as well as dependent p-values. Monte Carlo simulations suggest that the methods perform well with realistic sample sizes and show improved detection ability compared with competing methods. Our procedure is applied to a genome-wide association dataset of blood lipids.

On the validity of the pairs bootstrap for lasso estimator P. 981-987

L. Camponovo

Abstract

We study the validity of the pairs bootstrap for lasso estimators in linear regression models with random covariates and heteroscedastic error terms. We show that the naive pairs bootstrap does not provide a valid method for approximating the distribution of the lasso estimator. To overcome this deficiency, we introduce a modified pairs bootstrap procedure and prove its consistency. Finally, we consider the adaptive lasso and show that the modified pairs bootstrap consistently estimates the distribution of the adaptive lasso estimator.

Score tests for association under response-dependent sampling designs for expensive covariates P. 988-994

Andriy Derkach, Jerald F. Lawless, Lei Sun

Abstract

Response-dependent sampling is widely used in settings where certain variables are expensive to obtain. Estimation has been thoroughly investigated but recent applications have emphasized tests of association for expensive covariates and a response variable. We consider testing and provide easily implemented likelihood score tests for generalized linear models under a broad range of sampling plans. We show that when there are no additional covariates, the score statistics are identical for conditional and full likelihood approaches, and are of the same form as for ordinary random sampling. Applications in genetics are discussed briefly.

Clarifying missing at random and related definitions, and implications when coupled with exchangeability P. 995-1000

Fabrizia Mealli, Donald B. Rubin

Abstract

We clarify the key concept of missingness at random in incomplete data analysis. We first distinguish between data being missing at random and the missingness mechanism being a missing-at-random one, which we call missing always at random and which is more restrictive. We further discuss how, in general, neither of these conditions is a statement about conditional independence. We then consider the implication of the more restrictive missing-always-at-random assumption when coupled with full unit-exchangeability for the matrix of the variables of interest and the missingness indicators: the conditional distribution of the missingness indicators for any variable that can have a missing value can depend only on variables that are always fully observed. We discuss implications of this for modelling missingness mechanisms.