บทความที่น่าสนใจประจำเดือนเมษายน 2557
สาขาวิทยาศาสตร์และเทคโนโลยี

1	Title:	Association pattern discovery via theme dictionary models
	Author:	Deng, K., Geng, Z. and Liu, J. S.
	Journal:	Journal of the Royal Statistical Society: Series B (Statistical Methodology), Volume 76, Issue 2, pages 319–347, March 2014
	Summary:	Discovering patterns from a set of text or, more generally, categorical data is an important problem in many disciplines such as biomedical research, linguistics, artificial intelligence and sociology. We consider here the well-known 'market basket' problem that is often discussed in the data mining community, and is also quite ubiquitous in biomedical research. The data under consideration are a set of 'baskets', where each basket contains a list of 'items'. Our goal is to discover 'themes', which are defined as subsets of items that tend to co-occur in a basket. We describe a generative model, i.e. the theme dictionary model, for such data structures and describe two likelihood-based methods to infer themes that are hidden in a collection of baskets. We also propose a novel sequential Monte Carlo method to overcome computational challenges. Using both simulation studies and real applications, we demonstrate that the new approach proposed is significantly more powerful than existing methods, such as association rule mining and topic modelling, in detecting weak and subtle interactions in the data.
	Database:	Wiley Online Library

2	Title:	Regularized matrix regression
	Author:	Zhou, H. and Li, L.
	Journal:	Journal of the Royal Statistical Society: Series B (Statistical Methodology), Volume 76, Issue 2, pages 463–483, March 2014
	Summary:	Modern technologies are producing a wealth of data with complex structures. For instance, in two-dimensional digital imaging, flow cytometry and electroencephalography, matrix-type covariates frequently arise when measurements are obtained for each combination of two underlying variables. To address scientific questions arising from those data, new regression methods that take matrices as covariates are needed, and sparsity or other forms of regularization are crucial owing to the ultrahigh dimensionality and complex structure of the matrix data. The popular lasso and related regularization methods hinge on the sparsity of the true signal in terms of the number of its non-zero coefficients. However, for the matrix data, the true signal is often of, or can be well approximated by, a low rank structure. As such, the sparsity is frequently in the form of low rank of the matrix parameters, which may seriously violate the assumption of the classical lasso. We propose a class of regularized matrix regression methods based on spectral regularization. A highly efficient and

	scalable estimation algorithm is developed, and a degrees-of-freedom formula is derived to facilitate model selection along the regularization path. Superior performance of the method proposed is demonstrated on both synthetic and real examples.
Database:	Wiley Online Library

3

Title:	Space–time modelling of extreme events
Author:	Huser, R. and Davison, A. C.
Journal:	Journal of the Royal Statistical Society: Series B (Statistical Methodology), Volume 76, Issue 2, pages 439–461, March 2014
Summary:	Max-stable processes are the natural analogues of the generalized extreme value distribution when modelling extreme events in space and time. Under suitable conditions, these processes are asymptotically justified models for maxima of independent replications of random fields, and they are also suitable for the modelling of extreme measurements over high thresholds. The paper shows how a pairwise censored likelihood can be used for consistent estimation of the extremes of space–time data under mild mixing conditions and illustrates this by fitting an extension of a model due to Schlather to hourly rainfall data. A block bootstrap procedure is used for uncertainty assessment. Estimator efficiency is considered and the choice of pairs to be included in the pairwise likelihood is discussed. The model proposed fits the data better than some natural competitors.
Database:	Wiley Online Library

4

Title:	Entropic Latent Variable Integration via Simulation
Author:	Schennach, S. M.
Journal:	Econometrica, Volume 82, Issue 1, pages 345–385, January 2014
Abstract:	This paper introduces a general method to convert a model defined by moment conditions that involve both observed and unobserved variables into equivalent moment conditions that involve only observable variables. This task can be accomplished without introducing infinite-dimensional nuisance parameters using a least favorable entropy-maximizing distribution. We demonstrate, through examples and simulations, that this approach covers a wide class of latent variables models, including some game-theoretic models and models with limited dependent variables, interval-valued data, errors-in-variables, or combinations thereof. Both point- and set-identified models are transparently covered. In the latter case, the method also complements the recent literature on generic set-inference methods by providing the moment conditions needed to construct a generalized method of moments-type objective function for a wide class of models. Extensions of the method that cover conditional moments, independence restrictions, and some state-space models are also given.
Database:	Wiley Online Library

5

Title:	Competition for a Majority
Author:	Barelli, P., Govindan, S. and Wilson, R.
Journal:	Econometrica, Volume 82, Issue 1, pages 271–314, January 2014
Abstract:	We define the class of two-player zero-sum games with payoffs having mild discontinuities, which in applications typically stem from how ties are resolved. For such games, we establish sufficient conditions for existence of a value of the game, maximin and minimax strategies for the players, and a Nash equilibrium. If all discontinuities favor one player, then a value exists and that player has a maximin strategy. A property called payoff approachability implies existence of an equilibrium, and that the resulting value is invariant: games with the same payoffs at points of continuity have the same value and ε-equilibria. For voting games in which two candidates propose policies and a candidate wins election if a weighted majority of voters prefer his proposed policy, we provide tie-breaking rules and assumptions about voters' preferences sufficient to imply payoff approachability. These assumptions are satisfied by generic preferences if the dimension of the space of policies exceeds the number of voters; or with no dimensional restriction, if the electorate is sufficiently large. Each Colonel Blotto game is a special case in which each candidate allocates a resource among several constituencies and a candidate gets votes from those allocated more than his opponent offers; in this case, for simple-majority rule we prove existence of an equilibrium with zero probability of ties.
Database:	Wiley Online Library

6

Title:	A model-based correction for outcome reporting bias in meta-analysis
Author:	John Copas, Kerry Dwan, Jamie Kirkham, and Paula Williamson
Journal:	Biostatistics, Volume 15 Issue 2 April 2014, pages 370-383
Abstract:	It is often suspected (or known) that outcomes published in medical trials are selectively reported. A systematic review for a particular outcome of interest can only include studies where that outcome was reported and so may omit, for example, a study that has considered several outcome measures but only reports those giving significant results. Using the methodology of the Outcome Reporting Bias (ORB) in Trials study of (Kirkham and others, 2010. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. British Medical Journal 340, c365), we suggest a likelihood-based model for estimating the effect of ORB on confidence intervals and p-values in meta-analysis. Correcting for bias has the effect of moving estimated treatment effects toward the null and hence more cautious assessments of significance. The bias can be very substantial, sometimes sufficient to completely overturn previous claims of significance. We re-analyze two contrasting examples, and derive a simple fixed effects approximation that can be used to give an initial estimate of the effect of ORB in practice.
Database:	Oxford Journals Online

7	Title:	Extending distributed lag models to higher degrees
	Author:	Matthew J. Heaton and Roger D. Peng
	Journal:	Biostatistics, Volume 15 Issue 2 April 2014, pages 398-412
	Abstract:	Distributed lag (DL) models relate lagged covariates to a response and are a popular statistical model used in a wide variety of disciplines to analyze exposure–response data. However, classical DL models do not account for possible interactions between lagged predictors. In the presence of interactions between lagged covariates, the total effect of a change on the response is not merely a sum of lagged effects as is typically assumed. This article proposes a new class of models, called high-degree DL models, that extend basic DL models to incorporate hypothesized interactions between lagged predictors. The modeling strategy utilizes Gaussian processes to counterbalance predictor collinearity and as a dimension reduction tool. To choose the degree and maximum lags used within the models, a computationally manageable model comparison method is proposed based on maximum a posteriori estimators. The models and methods are illustrated via simulation and application to investigating the effect of heat exposure on mortality in Los Angeles and New York.
	Database:	Oxford Journals Online

8	Title:	How allele frequency and study design affect association test statistics with misrepresentation errors
	Author:	Valentina Escott-Price, Mansoureh Ghodsi, and Karl Michael Schmidt
	Journal:	Biostatistics, Volume 15 Issue 2 April 2014, pages 311-326
	Abstract:	We evaluate the effect of genotyping errors on the type-I error of a general association test based on genotypes, showing that, in the presence of errors in the case and control samples, the test statistic asymptotically follows a scaled non-central χ^2 distribution. We give explicit formulae for the scaling factor and non-centrality parameter for the symmetric allele-based genotyping error model and for additive and recessive disease models. They show how genotyping errors can lead to a significantly higher false-positive rate, growing with sample size, compared with the nominal significance levels. The strength of this effect depends very strongly on the population distribution of the genotype, with a pronounced effect in the case of rare alleles, and a great robustness against error in the case of large minor allele frequency. We also show how these results can be used to correct p-values.
	Database:	Oxford Journals Online

9	Title:	The analysis of multivariate longitudinal data: A review
	Author:	Geert Verbeke, Steffen Fieuws, Geert Molenberghs, and Marie Davidian
	Journal:	Statistical methods in medical research, vol. 23 no. 1, February 2014, pages 42-59
	Abstract:	Longitudinal experiments often involve multiple outcomes measured repeatedly within a set of study participants. While many questions can be answered by modeling the various outcomes separately, some questions can only be answered in a joint analysis of all of them. In this article, we will present

	a review of the many approaches proposed in the statistical literature. Four main model families will be presented, discussed and compared. Focus will be on presenting advantages and disadvantages of the different models rather than on the mathematical or computational details.
Database:	SAGE Journals Online

10

Title:	On random sample size, ignorability, ancillarity, completeness, separability, and degeneracy: Sequential trials, random sample sizes, and missing data
Author:	Geert Molenberghs, Michael G Kenward, Marc Aerts, Geert Verbeke, Anastasios A Tsiatis, Marie Davidian, and Dimitris Rizopoulos
Journal:	Statistical methods in medical research, vol. 23 no. 1, February 2014, pages 11-41
Abstract:	The vast majority of settings for which frequentist statistical properties are derived assume a fixed, a priori known sample size. Familiar properties then follow, such as, for example, the consistency, asymptotic normality, and efficiency of the sample average for the mean parameter, under a wide range of conditions. We are concerned here with the alternative situation in which the sample size is itself a random variable which may depend on the data being collected. Further, the rule governing this may be deterministic or probabilistic. There are many important practical examples of such settings, including missing data, sequential trials, and informative cluster size. It is well known that special issues can arise when evaluating the properties of statistical procedures under such sampling schemes, and much has been written about specific areas (Grambsch P. Sequential sampling based on the observed Fisher information to guarantee the accuracy of the maximum likelihood estimator. Ann Stat 1983; 11: 68–77; Barndorff-Nielsen O and Cox DR. The effect of sampling rules on likelihood statistics. Int Stat Rev 1984; 52: 309–326). Our aim is to place these various related examples into a single framework derived from the joint modeling of the outcomes and sampling process and so derive generic results that in turn provide insight, and in some cases practical consequences, for different settings. It is shown that, even in the simplest case of estimating a mean, some of the results appear counterintuitive. In many examples, the sample average may exhibit small sample bias and, even when it is unbiased, may not be optimal. Indeed, there may be no minimum variance unbiased estimator for the mean. Such results follow directly from key attributes such as non-ancillarity of the sample size and incompleteness of the minimal sufficient statistic of the sample size and sample sum. Although our results have direct and obvious implications for estimation following group sequential trials, there are also ramifications for a range of other settings, such as random cluster sizes, censored time-to-event data, and the joint modeling of longitudinal and time-to-event data. Here, we use the simplest group sequential setting to develop and explicate the main results. Some implications for random sample sizes and missing data are also considered. Consequences for other related settings will be considered elsewhere.
Database:	SAGE Journals Online