Contents

Abstracts

*speaker

Collaboration, statistics and major research challenges: avoiding boring people

Murray Cameron
CSIRO, Austrailia
e-mail: Murray.Cameron@csiro.au

Abstract

Statistics is a subject that underpins progress in many other fields. it develops through the interchange between theory and practice and the interchange between researchers who approach problems from different backgrounds and perspectives. I will illustrate the value of these interchanges by examples involving statisticians working in Japan, New Zealand and Australia on problems in the modelling of time series and point processes.

Recently, two major figures in molecular biology (James Watson and Craig Venter) have published autobiographies in which they extol the virtues of working on 'big research challenges'. This is consistent with CSIRO's directions in the last 8 years. I will discuss some of the opportunities that arise by following such a path, as well as some of the potential pitfalls for statisticians. I will conclude with a discussion on what future challenges might be.

Clustering species according to their relationship with their physical environment

Ross Darnell
CSIRO, Australia
e-mail: Ross.Darnell@csiro.au

Abstract

The Marine Biodiversity Research Hub is a collaboration between the University of Tasmania, CSIRO, Geoscience Australia, the Australian Institute of Marine Science and Museum Victoria. One of the tasks for the Hub is to to predict functional assemblage patterns on the Continental shelf using data on multiple species. We will present the use of latent class models to aggregate species into assemblages according to the relationship between their presence and physical covariates.

Particle Markov chain Monte Carlo

Arnaud Doucet
Institute of Statistical Mathematics, Japan
e-mail: arnaud@cs.ubc.ca

Abstract

Markov chain Monte Carlo (MCMC) and sequential Monte Carlo (SMC) methods (also known as particle filters) have emerged as the two main tools to sample from high-dimensional probability distributions. Although asymptotic convergence of MCMC algorithms is ensured under weak assumptions, the performance of these algorithms is unreliable when the proposal distributions used to explore the space are poorly chosen and/or if highly correlated variables are updated independently. We show here how it is possible to build efficient high-dimensional proposal distributions using SMC methods. This allows us not only to improve over standard MCMC schemes but also to make Bayesian inference feasible for a large class of statistical models where this was not previously the case. We demonstrate these algorithms on a non-linear state-space model and a Lévy-driven stochastic volatility model.

RAD Biodiversity: Modelling many species' counts together

Scott Foster
CSIRO, Australia
e-mail: scott.foster@csiro.au

Piers Dunstan
CSIRO, Australia

Abstract

Biodiversity is an important topic of ecological research. A common form of data collected to investigate patterns of biodiversity is the number of individuals of each species at a series of locations. These data contain information on the number of individuals (abundance), the number of species (richness) and the relative proportion of each species within the sampled assemblage (evenness). If there are enough sampled locations across an environmental gradient then the data should contain information on how these three attributes of biodiversity change over gradients. We show that the rank abundance distribution (RAD) representation of the data provides a convenient method for quantifying these three attributes constituting biodiversity. We present a statistical framework for modelling RADs and allow their multivariate distribution to vary according to environmental gradients. The method relies on three models: a negative binomial model, a truncated negative binomial model, and a novel model based on a modified Dirichlet-multinomial that allows for a particular type of heterogeneity observed in RAD data. The method is motivated by, and applied to, a large scale marine survey off the coast of Western Australia, Australia. It provides a rich description of biodiversity and how it changes with environmental conditions.

Effects of zoning on exploited fish populations in the Ningaloo Marine Park

Mick Haywood
CSIRO, Australia
e-mail: Mick.Haywood@csiro.au

Abstract

Populations of fish targeted by recreational fishers in the Ningaloo Marine Park, were surveyed in 2006 and 2007 to assess whether populations in pre-existing sanctuary zones (established in 1987) differed from those in areas that were open to fishing. A further aim of the work was to provide baseline data on populations from newly declared sanctuary zones that could be used to assess future trends in protected populations as well as across the park as a whole. Over 900 sites were surveyed over this time using Underwater Visual Census (UVC), with effort focused on 12 sanctuary zones distributed along the length of the park.

Fish assemblage structure showed clear trends with habitat and from north to south. There was also a significant overall difference in fish assemblages inside and outside sanctuary zones. The zoning related patterns appeared to be complex however, and examination of assemblages on a region by region basis showed zoning-related patterns in assemblages at only three sites, where targeted species were among those most likely to explain observed differences in assemblages. Non-target groups, including large grazers (scarids and kyphosids) were also associated with these differences. Among the species most commonly targeted by anglers there was an overall increase in biomass of the spangled emperor (L. nebulosus) of between 0.4 and 2.8 times greater in pre-existing sanctuary zones. These trends in fish biomass were largely driven by the size structure of populations in sanctuary zones-the trend was strongest in the in fish greater than the minimum legal size, consistent with fishing being the factor driving these differences.

Other species commonly targeted by recreational fishers were significantly more common outside sanctuary zones than inside them. The reasons for this are unclear but are likely to be complex, relating to the uneven distribution of habitat among pre-existing sanctuary zones and open areas, habitat preferences of these species, as well as the distribution of fishing effort around the reef. Most of these species are strongly associated with reef slope habitats which have been relatively poorly represented in pre-existing zones. Significant trends in relation to fishing pressure were nevertheless present among many of these species, which included large groupers and sharks, with biomass tending to be significantly lower in areas with higher levels of recreational fishing pressure.

A short history of genome-wide association study (GWAS) and statistical challenges

Naoyuki Kamatani
RIKEN, Japan
e-mail: kamatani@msb.biglobe.ne.jp

Abstract

Statistical genetic approaches for estimating population structures with applications to fisheries populations

Toshihide Kitakado
Tokyo University of Marine Science and Technology, Japan
e-mail: kitakado@kaiyodai.ac.jp

Shuichi Kitada
Tokyo University of Marine Science and Technology, Japan

Hirohisa Kishino
University of Tokyo, Japan

Abstract

For developing better management procedures for fisheries populations, understanding the population structures is one of crucial issues. Statistical genetic approaches have been playing key roles for this purpose.

To study the genetic structures of natural populations, assessing population differentiation among subpopulations is an important step. Samples are often taken from several localities. Sometimes, populations have continuous structures and consist of a large number of subpopulations. In such cases, assuming a hierarchical structure like metapopulation with an infinite-island model is a natural way to estimate the genetic differentiation between subpopulations. Some statistical methods for estimating a measure of population differentiation (say global-Fst) such as the conventional maximum likelihood and pseudo-likelihood methods have been developed. However, these methods may cause underestimation of global-Fst when the number of sampling localities is small. A statistical estimation using an integrated likelihood approach was proposed for estimating global-Fst (Kitakado ea al 2006). Simulation studies demonstrated that the integrated likelihood method outperformed the two methods previously developed. Furthermore, the assumption of hierarchical structure for expressing metapopulation earns benefit to use an empirical Bayes estimation for pairwise-Fst among subpopulations (Kitada et al2007, 2008). This approach also showed better performance compared to conventional methods. The method was applied to Pacific herring population.

Besides the spatial differentiation in genetic composition over areas or localities, different genetic compositions can be observed in different temporal stages such as years, generation and so on. Such temporal changes in genetic composition can be observed in stock enhancement programs. For fisheries stocks that reduce their numbers, artificial release of juveniles is one of possible ways to recover the population levels. In these programs, it is important to assess mixing proportions of released individuals in stocks. For this purpose, genetic stock identification has been applied. The allele frequencies in a composite population are expressed as a mixture of the allele frequencies in the natural and released populations. The estimation of mixing proportions is possible, under successive sampling from the composite population, based on temporal changes in allele frequencies. The natural population is not generally observed. The allele frequencies in the natural population may be estimated from those of the composite population in the preceding year. However, it should be noted that these frequencies could vary between generations due to genetic drift. A method for simultaneous estimation of mixing proportions and genetic drift in a stock enhancement program has been developed (Kitakado et al 2006, 2009). The model of genetic variation

acts a penalty against large genetic drift. The method is illustrated with application to real data on mud crab stocking.

Mixing could also occur unfavorably by introduction of foreign populations; individuals from foreign populations are sometimes released in fishing grounds. Because of difference in environmental conditions in habitats of natural and foreign populations or biological reasons, hybridization cannot be necessarily occurs. However, if the foreign population makes reproduction successfully within their populations, habitat of original population may be subject to be invaded. Therefore, before deep hybridization occurs, the monitoring and detection of mixing of foreign populations are of interest. An example for clam population in Japan is introduced.

References

Kitada, S., Kitakado, T. and Kishino, H. (2007), *Genetics*, **177**, 861–873.

Kitada, S., Kitakado, T. and Kishino, H. (2008), *Fish Genet. Breed. Sci.*, **38**, 41–50 (in Japanese).

Kitakado, T., Kitada, S. and Kishino, H. (2009), (in preparation)

Kitakado, T., Kitada, S., Obata, Y. and Kishino, H. (2006), *Genetics*, **173**, 2063–2072.

Kitakado, T., Kitada, S., Kishino, H. and Skaug, H. J. (2006), *Genetics*, **173**, 2073–2082.

Textile plot: a new LD display of multiple SNP genotype data

Natsuhiko Kumasaka
RIKEN, Japan
e-mail: kumasaka@stat.math.keio.ac.jp

Abstract

Advances in high-throughput genotyping technology enabled us to obtain dense SNP markers of human genome. The existence and distribution of linkage disequilibrium (LD) is being a major topic of interest to find disease susceptibility loci through genome wide association studies, or to reveal underlying historical and biological processes such as selection, mutation, recombination and population history (Pritchard et al. 2001). A graphical representation of LD for multiple SNP genotype data (e.g. Barrett et al. 2005) also becomes an indispensable tool to make inference about underlying genetic variation through LD structure before applying statistical or mathematical models to the data.

The textile plot (Kumasaka et al. 2008) has been proposed as a graphical representation for any high dimensional data. Recent studies of the textile plot applied to genetic data revealed that it can accentuate presence of LD by specific geometrical shapes. That is, the LD between adjacent SNPs is represented by line crossings between adjacent loci so that low-grade line crossing indicates high dependency of alleles between SNPs. Besides, the vertical dispersion of genotypes approximates the structure of pair-wise correlation coefficients for all SNP genotypes.

In addition, the textile plot may also accentuate other genetic features present in the SNP genotype data, such as allele frequencies, haplotype configurations and the deviation of samples from Hardy-Weinberg equilibrium, simultaneously with LD in one display. This talk aims to show the potential usefulness of the textile plot as an aid to the interpretation of such genetic variations among the multiple SNP genotype data.

References

Barrett J. C., Fry B., Maller J., Daly M. J. (2005), Haploview: analysis and visualization of LD and haplotype maps, *Bioinformatics*.

Kumasaka, N., Shibata, R. (2008), High-dimensional data visualisation: the Textile Plot, *Computational Statistics & Data Analysis*, **52**, 3616–3644.

Pritchard, J. and Przeworski, M. (2001), Linkage Disequilibrium in Humans: Models and Data, *American Journal of Human Genetics*, **69**, 1–14.

Understanding water quality measurements

Sarah Lennox
CSIRO, Australia
e-mail: Sarah.Lennox@csiro.au

Abstract

The Queensland Bulk Water Supply Authority (trading as Seqwater) provides and monitors water for the South-East Queensland, Australia region. To help understand the factors that affect the measurement of water quality within the reservoirs data was collected for several sites within a reservoir. This dataset focuses on 7 water quality variables taken at a variety of depths. Issues which arise include the multivariate nature of the problem, non-linear trends in both depth and time, correlations in both depth and time, interactions between these variables, instrument error and heterogeneous variances. Of particular interest to the environmentalists is the diurnal variability.

Statistical challenges for modeling data with many zeros

Mihoko Minami
The Institute of Statistical Mathematics and Keio University, Japan
e-mail: mminami@math.keio.ac.jp

Abstract

In ecological and environmental studies, count data such as the number of animals per unit area or unit effort often contain many zero-valued observations. Analyzing such data without any consideration for excess zeros may produce misleading results. In this talk, first we would like to show a possible consequence of ignoring excess zeros in analysis. Then, we will introduce a new feature extraction method for very non-normal multivariate data, such as multivariate data with many zero-valued observations.

The negative binomial regression model is a commonly used model for count data with over-dispersion relative to a Poisson model. However, fitting negative binomial regression model to data with excess zeros may produce very misleading results. Minami et al. (2007) found that the negative binomial regression model over-estimated temporal trends in species relative abundance. We show that this phenomenon could easily occur with count data that contain excess zeros and we investigate why this happens.

We propose a new feature extraction method which extends principle component analysis (PCA) in the same manner as the generalized linear model extends the ordinary linear regression model. As an example, we analyze multivariate species-size data from the purse-seine fishery in the eastern Pacific Ocean. The objective of this analysis is to explore species associations, and their relationship to environmental factors, as a means of developing options for reducing catch of unmarketable animals. The data contain many zero-valued observations for each variable (combinations of species and size). Thus, as an error distribution we use the Tweedie distribution which has a probability mass at zero and apply Tweedie-generalized PCA (GPCA) method to the data.

The Tweedie-GPCA method can be described as follows. Suppose we want to extract k characteristic features (components) from m dimensional data. PCA finds projections to minimize the mean square reconstruction error. That is, under the model $\boldsymbol{Y} = \boldsymbol{M} + \boldsymbol{E}$ where $\boldsymbol{M}$ is a matrix of rank k, PCA minimizes $\sum_{i,j} E_{ij}^2$. In other words, PCA maximizes the likelihood under the model: $\boldsymbol{Y} = \boldsymbol{M} + \boldsymbol{E}$, $E_{ij} \sim N(0, \sigma^2)$, i.i.d. with a constraint $\mathrm{rank}(\boldsymbol{M}) = k$.

We propose a generalized PCA method that extends classical PCA in the following sense:

1. The rank of matrix $g(\boldsymbol{M})$, rather than $\boldsymbol{M}$ itself, is k where g is a monotone increasing function and $\boldsymbol{M} = \mathrm{E[Y]}$.

2. Y_{ij} independently follows a distribution $f(y; M_{ij}, \sigma^2)$ in the exponential family.

Features are obtained from $g(M)$ using singular value decomposition or independent component analysis. Our proposed method is a likelihood-based method. The proportion of deviance explained can be used as a criteria for choosing the number of features k.

For the fishery data, we assume that Y_{ij} follows a Tweedie distribution and take g to be the natural logarithm (i.e., we assume a log link). With $k = 4$ ($m = 56$), about 70
% of deviance was explained by the model. The first few features for variables (species, size) appear to be associated with abundance of several species that are considered vulnerable to fisheries impacts (e.g., sharks and turtles), and the associated features for sets show spatial pattern that may be related to oceanography. Some features show similar spatial patterns to

those of features obtained from non-metric multidimensional scaling with Sorensen distance. However, the features of GPCA appear to have more coherent spatial structure. These results suggest that GPCA may be a useful tool for identifying areas within the region occupied by the purse-seine fishery with greater occurrence of catch of vulnerable species. These results also suggest that, more generally, the Tweedie-GPCA method shows promise as a tool for studying the impact of fisheries on ecosystems and exploring ecosystem community structure.

References

Hosking L. K., Boyd P. R., Xu C. F., Nissum M., Cantone K., Purvis I. J., Khakhar R., Barnes M. R., Liberwirth U., Hagen-Mann K., Ehm M. G., Riley J. H. (2002), Linkage disequilibrium mapping identifies a 390 kb region associated with CYP2D6 poor drug metabolising activity, *Pharmacogenomics J*, **2**, 165–175.

Weight distribution in trawling data

Mayumi Naka
Keio University, Japan
e-mail: naka@stat.math.keio.ac.jp

Ritei Shibata
Keio University, Japan
e-mail: shibata@math.keio.ac.jp

Abstract

Distribution of individual weight caught in Experiment 2 of the trawling effect experimental survey is investigated. Gamma distribution is a good candidate for the distribution since it is an stable approximation to the stationary distribution of a random coefficient log-nomral growing process. However, a proper grouping of several species is indispensable for the check of goodness of fit to the data since the effective sample size is very small in each species. Such homogeneous species groups were found in the four classes, Crustacean, Demospongiae, Gastropoda and Echinoidea. The estimated gamma parameters reveal effect of trawlings more clearly than several descriptive statistics or a linear regression.

Quantitative modelling of financial risks

Pavel Shevchenko
CSIRO, Australia
e-mail: Pavel.Shevchenko@csiro.au

Abstract

Changes in information technology, globalisation, complex financial products and other factors expose financial industries to new risks. To ensure stability, new international regulatory frameworks (Basel II and Solvency II) have been developed for the banking and insurance industries. Currently, major financial institutions are undertaking quantitative modelling of risk to satisfy these requirements.

CMIS has worked in the area of financial mathematics since 1999. This talk presents statistical models we have developed and applied during recent consulting and research projects. In particular, we discuss modelling operational risk, credit risk, option pricing, modelling commodity prices and associated numerical techniques such as Markov chain Monte Carlo algorithms.

Financial data involve rare events with high impact and often should be supplemented with expert opinions. *Data Science* in this area presents many challenges and different approaches are hotly debated.

Modeling counts in trawling data

Ritei Shibata
Keio University, Japan
e-mail: shibata@math.keio.ac.jp

Yuki Sugaya
Keio University, Japan
e-mail: sugaya@stat.math.keio.ac.jp

Abstract

This is a report on our collaboration with people, Dr Ross Darnell, Mr Mick Haywood, Dr Charris Burridge and others in CSIRO Marine Laboratories, Cleveland. The target data we have chosen for the collaboration is Northern Prawn Fishery Data. The sample survey was designed by Mr Mick Haywood and his group, and undertaken during the period 2002 to 2005. The aim of this survey was to quantify the effect of trawling on seabed fauna in the Northern Prawn Fishery in the north of Australia. However, two simple analyses, one based on a model for successive biomass removal by trawling and the other on a regression model with random effects, were not successful, mainly because the data collected is sparse with many zero catches for most species and their analysis are all based on the number of catches and the biomass per hectre. We have applied a data science approach that started by looking at the raw data. It was found that the Thomas model (used for modeling the number of plants in a region) gave a good explanation of catch size, provided that species are properly selected and the survey region is split into several locally homogeneous sub-regions based on a careful examination of the raw data. The effect of trawling can be clearly explained by the parameters of the Thomas model together with the parameters of the gamma distribution for weights, which is reported by a separate talk by Ms Mayumi Naka. This is still the beginning of our model building process. By continuing our collaboration we hope that it will reveal more significant scientific aspects of the data through model.

Smile curve and local volatility

Ritei Shibata
Keio University, Japan
e-mail: shibata@math.keio.ac.jp

Yuuka Tanizawa
Keio University, Japan
e-mail: yuuka@stat.math.keio.ac.jp

Abstract

Volatility smile is a curve of implied volatility as a function of strike price of an option. The implied volatility is the volatility of an underlying asset derived from market option prices by making use of Black-Scholes formula. It is expected to be a constant if the price of the underlying asset follows a log normal process which Black-Scholes formula relies on. However, the reality does not support this supposition. The implied volatility smiles on strike price and and differently for each period time of the maturity. There have been many attempts to give a good answer to this phenomena but non of them are successful as far as we know. We will show that some of smiles can be explained by introducing a local volatility model for the underlying asset.

Investigating the issues of sampling in marine surveys

Hideyasu Shimadzu
Visiting Scientist CSIRO , Australia
e-mail: Hideyasu.Shimadzu@csiro.au

Ross Darnell
CSIRO , Australia
e-mail: Ross.Darnell@csiro.au

Abstract

Marine samples are collected to ascertain the abundance of many species ranging from fish to sponges. Sampling issues such as time of day, moonphase, season, sampling device and subsampling ratio can generate bias in the recorded measure of abundance. We explore and quantify some of these issues to explain the impact these may have on measures of biodiversity. Specifically, we estimate the bias introduced to the probability of species presence when subsamples are taken from the sample caught by benthic sled.

Circular-circular regression, functional relationship and measurement error models

Kunio Shimizu
Keio University, Japan
e-mail: shimizu@math.keio.ac.jp

Abstract

Regression model of a linear variable on an angular variable is used in environmental and ecological sciences, meteorology and other sciences. Examples include prediction of (1) air quality index by temperature and wind direction, (2) ozone level by temperature, wind speed and wind direction, (3) significant waveheight by wind speed and wind direction, and (4) wind energy by velocity, time and wind direction.

In this paper we have interest in data sets which consist of two angular variables. Examples are (1) wind directions at 6 a.m. and noon, (2) systolic blood pressure peak times, (3) the spawning time of certain fish and the time of low tide, and (4) locations of orthologous genes on archaeal and bacterial genomes. Downs and Mardia (2002) used von Mises distributions as error distributions for their circular-circular regression model, while Kato et al. (2008) proposed the use of wrapped Cauchy distributions. Here we study circular-circular functional relationship and measurement error (structural) models based on wrapped Cauchy distributions.

References

Downs, T. D. and Mardia, K. V. (2002), Circular regression, *Biometrika*, **89**(3), 683–697.

Kato, S., Shimizu, K. and Shieh, G. S-R. (2008), A circular-circular regression model, *Statistica Sinica*, **18**(2), 633–645.

This is joint work with Shogo Kato, Institute of Statistical Mathematics, Tokyo, and Grace Shwu-Rong Shieh, Institute of Statistical Science, Academia Sinica, Taipei.

Bell polynomials in discrete probability distributions

Masaaki Sibuya
Keio University, Japan
e-mail:sibuyam@1986.jukuin.keio.ac.jp

Abstract

In this report we try to find how Bell polynomials can be used to solve the problems in the probability distribution theory. A main problem is the analysis of compound distributions and related random partitions of number.

Some typical problems are solved in terms of the generalized Stirling numbers, a simpler subgroup of Bell polynomials, but the subgroup is not wide enough. Others can be solved by calculating derivatives of generating functions. Then, Bell polynomials appear in Faa di Bruno formula for the chain rule of higher derivatives of a composite function. Compound Poisson is typical and equivalent to gexponential-typeh total and partial Bell polynomials.

Bell polynomials will provide a unified view to use a computer algebra system in statistical applications including data-analysis.of cluster samples.

Likelihood-based method for estimating penetrance and disease susceptibility allele frequency

Yuki Sugaya
Keio University, Japan
e-mail: sugaya@stat.math.keio.ac.jp

Ritei Shibata
Keio University, Japan
e-mail: shibata@math.keio.ac.jp

Abstract

Linkage analysis based on estimated recombination fraction is still useful for detecting a gene associated with a disease from pedigree data. Penetrance and disease susceptibility allele frequency play an important role in due course of maximum likelihood estimation of the recombination fraction. Such values are often assumed to be known a priori. But it is not true in practice and rather subjective values are inevitably used. For example, 1 or 0 for penetrance and universal allele frequency for a disease are used in practice. It is easily understood that the maximum likelihood estimate relies on those values and an objective way of determining those values is desirable. One of such ways is to employ again the maximum likelihood principle for estimating those parameters. Fortunately it is enough to consider the likelihood given phenotypes so that the estimation can be executed apart from estimation of the recombination fraction. Maximisation of the likelihood becomes simpler if it is noted the fact that the likelihood is a polynomial of such parameters. It will be shown by practical examples that the estimation of recombination fraction becomes better than the case when such parameters are subjectively chosen.

Mixed methods for fitting the GEV distribution

Pierre Ailliot
Université de Brest, France

Craig Thomson
Statistics Research Associates Ltd, New Zealand

Peter Thomson
Statistics Research Associates Ltd, New Zealand
e-mail: peter@statsresearch.co.nz

Abstract

The generalised extreme-value (GEV) distribution is widely used for modelling and characterising extremes. It is a flexible 3-parameter distribution that combines three extreme-value distributions within a single framework: the Gumbel, Frechet and Weibull. Common methods used for estimating the GEV parameters are the method of maximum likelihood and the method of L-moments.

This paper generalises the mixed maximum likelihood and L-moments GEV estimation procedures proposed by Morrison and Smith (2002) and derives the asymptotic properties of the resulting estimates. Analytic expressions are given for the asymptotic covariance matrices in a number of important cases, including the estimators proposed by Morrison and Smith (2002). These expressions are verified by simulation and the efficiencies of the various estimators established.

The asymptotic results are compared to those obtained for small samples, and the properties of the various estimators, including constrained maximum likelihood estimators, are considered. The corresponding quantile estimators are also assessed for accuracy and bias. Using simplified constraints for the support of the log-likelihood, computational strategies and graphical tools are developed which lead to computationally efficient, numerically robust, estimation procedures. These methods are also applied to 24-hour annual maximum rainfall at Wellington, New Zealand, over the period 1940-1949 and within each phase of the Interdecadal Pacific Oscillation (IPO).

Keywords: Extremes; GEV distribution; mixed estimation methods; asymptotic properties; small samples; quantile estimation; constrained maximum likelihood.

Challenges analysing modern genomics data

Bill Wilson
CSIRO, Australia
e-mail: Bill.Wilson@csiro.au

Abstract

Our involvement in various projects over the last few years has focused on the generation and analysis of large high quality datasets. Together with the CSIRO Preventative Health Flagship we have acquired large gene expression datasets from samples of colon tissue (normal and diseased) as well as brain tissues from Alzheimerfs affected and unaffected individuals. High quality datasets not only allow us to conduct quality analyses, delivering quality science results, but also allows us to develop our statistical analytical tools as the data collection technologies change in this field.

One recent technological advance has been the decrease in cost and vast increase in capacity of DNA sequencing. High throughput, deep sequencing, or next generation sequencing technologies, are producing vast amounts of data. The analysis of this kind of data presents a unique opportunity to integrate tools used in data collection and management, and genome and statistical analysis. We are currently engaged in activities within CSIRO to advance our skills in the area of analysis of high throughput sequencing data, identifying the novel statistical issues in understanding such complex data.

What can we do for hedge fund return data under the DandD environment?

Daisuke Yokouchi
Hitotsubashi University, Japan
e-mail: yokouchi@ics.hit-u.ac.jp

Ryozo Miura
Hitotsubashi University, Japan
e-mail: rmiura@ics.hit-u.ac.jp

Abstract

This paper discussed analysis of performance data of individual hedge funds by The Center for International Securities and Derivatives Market (CISDM), and design of a DandD client software for helping its analysis. To start with, we organized the data into a DandD instance, and stored the entity of the data into the PostgreSQL database system in order to analyze the data under the DandD environment. Since the DandD environment provides us with a good tool for browsing data in a DandD instance as a Textile Plot, we tried to create its textile plot to understand summary of the data. However, we could not obtain it owing to its data size, over 660000 records. Then, we have been starting with design and implementation of a new DandD client software gICS FinAnalyzer (ICSFA)h for interactively handling a large amount of data by mouse action or revision of SQL query described in the DandD instance, simply grasping the whole picture of a large amount of financial time series data as a textile plot or a parallel coordinate plot, and doing preliminary analysis. By using the ICSFA, we organized characteristics of each strategy of hedge fund returns visually, and with the help of its function for calculating some performance or risk measurement. And likewise, we examined differences between surviving hedge funds and not, and relations between performance of hedge funds and that of stock indices, foreign exchanges, and interest rates.