Contents

Colouring and breaking sticks, pairwise coincidence losses and clustering expression profiles

P. Green[1]

[1] University of Bristol

Abstract

We consider methods for Bayesian model-based clustering of gene expression profiles, that is, measurements of expression levels of a large number of genes, typically from microarray assays, across a number of different experimental conditions and/or biological subjects. We follow a familiar approach using Dirichlet-process-based models to cluster the genes implicitly, but depart from standard practice in several ways. First, we incorporate regression on covariate information at the condition/subject level by modelling regression coefficients, not the expectations of the data directly. More importantly, we replace the Dirichlet process by one of a richer family of models, generated from a stick-colouring-and-breaking construction, under which cluster identities are not exchangeable: this allows modelling a 'background' cluster, for example. Finally, we take a decision-theoretic approach to find an optimal clustering, using a pairwise coincidence loss function. This is joint work with John Lau at Bristol.

Monotone smoothing: application of a compound Poisson-Gamma process to modelling radiocarbon-dated depth chronologies

J. Haslett[1] and A. Parnell[1]

[1] Trinity College Dublin

Abstract

We propose a new and simple continuous Markov monotone stochastic process and use it for Bayesian monotone smoothing. The process is piecewise linear, based on additive independent Gamma increments arriving in a Poisson fashion. A special case allows very simple conditional simulation of sample paths given known values of the process. We take advantage of a re-parameterisation involving the Tweedie distribution to provide efficient MCMC computation. The motivating problem is the establishment of a chronology for samples taken from lake sediment cores; that is, the attribution of a set of dates to samples of the core given their depths, knowing that the depth-age relationship is monotone. The chronological information arises from radiocarbon (^{14}C) dating at a subset of depths. We use the process to model the stochastically varying sedimentation rate.

Keywords: Monotone smoothing, Radiocarbon dating, Tweedie distribution, Compound Poisson-Gamma distribution.

1 Introduction

Monotone smoothers arise in many diverse applications. Our interest lies in modelling the monotone relationship between age and depth in sediment cores for lakes or peat. Biological proxies in such cores provide information on past climates; see Haslett et al (2006). Typically, data (such as multivariate counts of distinguishable pollen taxa) are available at each of n depths in a core. For a subset of these, at depths d_j for $j = 1, \ldots, m$, dating information is also available in the form of suitably qualified laboratory estimates $x_j \pm \tau_j$ (in calendar years before present). The challenge is to estimate the true calendar dates - θ_i at all depths d_i for $i = 1, \ldots, n$. A chronology is the set of estimates (θ_i, d_i). It will necessarily be monotone

as younger ages must occur at shallower depths. Uncertainty arises in the calibration of the ^{14}C ages, and in stochastic interpolation for the $n - m$ depths at which we have no age information. We present a new and simple Bayesian method for sampling smooth monotonic chronologies given the data.,

2 Methods

We introduce our procedure in the context of a bivariate renewal process (Hunter, 1974) and discuss the conditions for mean-square continuous chronologies. We then specialise to the simpler Compound Poisson-Gamma (CPG) process, and thence to an alternative parameterisation via the Tweedie distribution. We consider properties of the process conditioned on sample points and the likelihood of a sample of points under the model.

3 Results

The CPG model is implemented at two sites; Sluggan Moss, Co. Antrim and Glendalough, Co. Wicklow. Both sites have important consequences for past climate change in Ireland. Our findings demonstrate that there is a considerable under-estimate of uncertainty compared with previous model estimates of their chronologies.

References

Haslett, et al (2006) Bayesian palaeoclimate reconstruction. *Journal of the Royal Statistical Society, Series A 169*(3), 395–438. 47, 4, 607-615.

Hunter, J. (1974). Renewal theory in two dimensions. *Advances in Applied Probability* 6 (2), 220-221.

A grade of membership model for rank data

C. Gormley[1] and B. Murphy[2]

[1] University College Dublin

[2] Trinity College Dublin

Abstract

The grade of membership (GoM) model is a 'soft' clustering model in that it allows observations have partial membership of each of the homogeneous extreme profiles which constitute a population. The GoM model is used to identify voting blocs within the Irish electorate and to examine the partial membership of voters of these voting blocs. Specialized rank data models are incorporated to account for the ranked nature of Irish votes.

Keywords: soft clustering, rank data, Bayesian inference.

1 Introduction

Voters from the 1997 Irish presidential electorate are examined to highlight voting blocs and to determine the mechanisms which motivate voter preferences. In addition, a soft clustering of the electorate is achieved such that each voter has an associated 'mixed membership' parameter which describes the likelihood of their membership of each voting bloc.
The Irish electorate in particular is examined as Irish elections use an electoral system where voters rank some or all of the candidates in order of preference. Thus the lower preferences of the voters contain information which must be exploited when making inferences about the electorate.

2 Methods

The Plackett-Luce model for rank data (Plackett, R.L. (1975)) exploits the information contained in the ranked voter preferences. The grade of membership model (Erosheva, E.A.(2003)) is incorporated with the Plackett-Luce model to estimate the characteristics of voting blocs in the electorate. In addition a mixed membership vector for each voter is estimated.
Model fitting is performed within the Bayesian paradigm subsequent to the imputation of latent variables. An MCMC (Metropolis within Gibbs)

sampler is employed to estimate model parameters. Ideas from the MM algorithm (Lange et al. (2000)) are used to construct valid and tractable proposal distributions for the Metropolis step of the algorithm. Issues such as label switching and model comparison also require attention.

3 Results

The grade of membership model was fitted to polled voters from the 1997 Irish presidential electorate. Results suggest party politics play a large role in influencing voter preferences, even in an election where little political power is traditionally involved. Voting blocs characterized by different voting preferences are identified. Soft clustering of the voters results from examining the mixed membership parameters — Figure 1 shows density estimates of the mixed membership parameters for two randomly sampled voters.

FIGURE 1. Density estimates of the mixed membership parameters for two randomly sampled voters.

References

Erosheva, E.A. (2003). Bayesian Estimation of the Grade of Membership Model. *Bayesian Statistics 7.*

Lange, K. et al. (2000). Optimization Transfer Using Surrogate Objective Functions *Journal of Computational and Graphical Statistics* 9, 1, 1-20.

Plackett, R.L. (1975). The Analysis of Permutations *Applied Statistics* 24, 2, 193–202.

Genetic Mixed-Effect Models for Twin Lifetime Data under LTRC

I. Ha[1], Y. Lee[2], and Y. Pawitan[3]

[1] Daegu Haany University, Gyeongsan, Korea

[2] Seoul National , Korea

[3] Karolinska Institute, Stockholm, Sweden.

Abstract

Twin studies are the most widely-used methods for quantifying the contribution of genetic and environmental factors on traits such as lifespan or disease susceptibility. In this talk we propose a genetic mixed-effect model for twin survival data which are subject to left truncated and right censored (LTRC) due to limited period of observation, and develop a new h-likelihood (hierarchical-likelihood) procedure, leading to a simple and fast computation for analyzing large survival data sets. We apply the methodology to the genetic analysis of lifetime data in Swedish Twin Register.

Keywords: Environment effect, Genetic effect, h-likelihood, LTRC.

Introduction

To do the genetic study based on twin survival data, the data on monozygotic (MZ) and dizygotic (DZ) twins are required and they have been analyzed using proportional hazards (PH) frailty models, which allow to separate the effects of genetic and environment (Yashin et al, 1999). As an alternative, as in the accelerated failure-time models, the mixed-effect models have been proposed, which give robust results against various misspecifications about the model assumptions (Ha, Lee & Song, 2002). For analyzing of the Swedish twin lifetime data (The old cohort) we propose a genetic mixed-effect model, which allows general fixed predictors and random components to capture genetic and environmental effects. The model handles LTRC problems, which often occur in the data collection on twin study. For the model inference marginal likelihood require intractable integration. We avoid the problem using the h-likelihood, giving a statistically efficient and simple unified framework for various random-effect models (Lee & Nelder, 1996, 2001; Ha et al., 2002).

The proposed model

Let T_{ij} be the lifetime for the j-th member of the i-th twin pair. T_{ij} are only partially observed due to LT and RC variables (L_{ij}, F_{ij}), which are assumed to be independent of the T_{ij}'s. Let g_{ij}, c_{ij} and e_{ij} be the random genetic (G), shared environment (C) and unshared environmental (E) components, respectively. We propose the mixed-effect model (GCE model) with two random effects (g_{ij} & c_{ij}): for $i = 1, 2, \ldots, q$ and $j = 1, 2$,

$$\log T_{ij} = x_{ij}^T \beta \ + \ g_{ij} \ + \ c_{ij} \ + \ e_{ij}, \tag{1}$$

where x_{ij} is a vector of covariates corresponding to fixed effects β, and $g_{ij} \sim N(0, \sigma_g^2)$, $c_{ij} \sim N(0, \sigma_c^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$ are mutually independent error components. Following Pawitan et al. (2004), if the i-th twin pair is MZ, it is assumed that $\text{corr}(g_{i1}, g_{i2}) = 1$ & $\text{corr}(c_{i1}, c_{i2}) = 1$, and if it is DZ $\text{corr}(g_{i1}, g_{i2}) = 0.5$ & $\text{corr}(c_{i1}, c_{i2}) = 1$. Note that *heritability* is given by $h_g^2 = \sigma_g^2/(\sigma_g^2 + \sigma_c^2 + \sigma_e^2)$, which measures the importance of genetics relative to other factors in explaining the variability of a trait in a population (Sham, 1998).

Results and Discussion

As a result of combined analysis (MZ & DZ), for the males, the AIC chooses the GE model without C-component as the best model, with estimated heritability $\widehat{h_g^2} = 26\%$. For the females, the GE model is also best, with $\widehat{h_g^2} = 21\%$. By using the AIC, Yashin et al. (1999) also chose a genetic frailty model corresponding to the GE model as the final model. However, in frailty models the estimates of dispersion parameters can be sensitive to mis-specification of random-effect distribution (Xue, 2001; Ha & Lee, 2005).

References

Ha et al. (2002). Hierarchical likelihood approach for mixed linear models with censored data. *Lifetime Data Analysis*, 8, 163-176.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion). *JRSS B*, 58, 619-678.

Pawitan et al. (2004). Estimation of genetic and environmental factors for binary traits using family data. *Statistics in Medicine*, 23, 449-465.

Yashin et al. (1999). Half of the variation in susceptibility to mortality is genetic: findings from Swedish twin survival data. *Behavior Genetics*, 29, 11-19.

Trials and Observations. Combining Evidence.

C. Walsh[1]

[1] Trinity College Dublin

Abstract

Meta-analyses of Randomised Controlled Trials should be what physicians base their clinical decisions on. Or should they? A simple examination of the way in which such analyses can be presented in a confusing fashion demonstrates that care is needed in interpreting the outcome of interest.
This study started life as a straightforward Bayesian analysis of observational data. However, during the process of eliciting a prior, a meta-analysis of RCTs which equivocated about the treatment effect was uncovered.
The reason for the unclear interpretation in Fillipini et al (2003) was because of the large between study variation introduced into the random effects meta-analysis because of unreasonable assumptions.
By combining information from an observational study, it is clear that the caution expressed by the authors of the original meta analysis was less than useful for clinical practitioners. A sensitivity analysis, allowing for different levels of 'evidence' in the hierarchy was carried out.
The background, analysis and clinical interpretation as outlined in O'Rourke et al (2007) is provided during this talk.

Keywords: Prior Elicitation, Evidence Synthesis, Observational Studies, Modified Likelihood, Bayesian Inference.

References

Fillipini et al (2003). Interferons in relapsing remitting multiple sclerosis: a systematic review. *The Lancet.* 361, p545-552.

O'Rourke K, Walsh, C and Hutchinson, M (2007). Outcome of beta-interferon treatment in relapsing-remitting multiple sclerosis: a Bayesian analysis. *J Neurology.* Accepted Mar 2007.

Industrial and applied statistics using the SAS system

T. Fitzpatrick[1]

[1] SAS Ireland, Dublin

Abstract

In a world where the traditional bases of competitive advantage have largely evaporated, how do you separate your companys performance from the pack? Use advanced analytics to make better decisions and extract the maximum value form your business. This is competing on analytics. This presentation will focus on the current business trends that are pushing companies to look to at advanced analytics for that ever elusive competitive edge.

Estimating optimal dynamic treatment regimes from observational studies: can we hope to succeed?

A. Rodnitzky[1]

[1] Harvard School of Public Health

Abstract

Dynamic treatment regimes are set rules for sequential decision making based on patient covariate history. Observational studies are well suited for the investigation of dynamic treatment regimes because of the variability in treatment and clinic visit timing found in them. This variability exists because different physicians make different decisions in the face of similar patient histories. However, the analysis poses several difficult challenges: i) methods for estimation of treatment effects have to appropriately control for high dimensional time dependent confounders (i.e. time varying risk factors that predict future treatment); standard multivariate longitudinal regression methods which adjust for time dependent risk factors generally yield biased estimators and cannot be used, ii) the determination of an optimal treatment strategy is a high dimensional sequential decision problem; the set of potential dynamic regimes from which to search for the optimal may be very large, iii) the optimal treatment strategy depends on the frequency of the occasions at which decisions can be made, i.e. the clinic visits; yet in observational studies the timing of clinic visits is severely confounded with time varying risk factors and iv) long follow-up observational studies of chronic diseases suffer severely from drop-out. In this talk we examine possible estimation strategies to address these issues and the assumptions under which these strategies should yield valid inference. The estimation, from a large French database, of the optimal CD4 count value at which to start highly anti-retroviral therapy in HIV+ patients is used to illustrate the discussion.

Statistical analysis of genome-wide case-control studies of genotype-phenotype associations

D. Clayton[1]

[1] Wellcome Trust

Abstract

Initial attempts to map genes implicated in human disease centred around linkage analysis in multiply affected families. Although this was successful in discovering genes responsible for simple, or "Mendelian" diseases, it became evident that this approach lacked the power necessary to detect the smaller relative risks attributable to disease susceptibility genes for more common conditions. For some years progress was slow, as investigators sought to adapt, and there was growing pessimism in some quarters. However, a combination of large case collections and technological advance have led to the possibility of very large scale case-control studies, involving thousands of subjects measured at hundreds of thousands of genetic loci. In recent months several of these have reported and it seems clear that the field will take a major step forward. This background will be reviewed, and statistical problems in the design and analysis of the new "genome-wide" studies will be outlined, using the example of the Wellcome Trust Case-Control Consortium (WTCCC) — a recently completed study involving around 17,000 subjects measured for 500,000 genetic polymorphisms.

A comparative study of nonparametric derivative estimators

J. Einbeck[1] and J. Newell[2]

[1] Durham University

[2] National University of Ireland, Galway

Abstract

Several papers published in the mid-nineties, particulary originating from the local polynomial smoothing community, gave the impression that the entire issue of nonparametric derivative estimation is solved, and as a result the research activity about this topic stalled to some extent. This is unfortunate, as many open questions remain, and most problems are still treated rather cursorily in the literature. Typically derivative estimates are calculated as "by-products" from a local polynomial or spline fit. However, these estimates often suffer from boundary effects and are very sensitive to outliers. Apart from this, the local polynomial estimators suffer from a systematic downward bias. This article is intended to re-establish research interest in derivative estimation, and to guide the user who needs to work with one of the available packages.
Keywords: Derivatives; Kernels; Splines

Motivation

Nonparametric estimation of derivatives is important in a variety of disciplines. Specifically, when considering a regression problem of type $y = m(x_i) + e_i$, one is often not interested in $m(\cdot)$ itself, but rather in the relative change dm/dx of m when increasing or decreasing x by a small value dx. An important special case is when x represents time, in which the 1st derivative of m has the interpretation of a speed, and the 2nd derivative of an acceleration, which is for example of interest in the analysis of growth curves. However, the importance of estimating derivatives goes far beyond the end in itself. Often one relies on asymptotic approximations in order to obtain bias and variance estimates, confidence intervals, optimal bandwidths, etc., and these expressions usually involve derivatives of $m(\cdot)$, which are normally unknown and have to be estimated. A further field of application for derivative estimators are change point problems. For instance,

when analyzing blood lactate data of elite athletes, one is interested in the workload at which the lactate level suddenly rises, which can be detected by finding the maximum of the second derivative (Newell et al., 2005).

On nonparametric derivative estimation

There are two main approaches to nonparametric derivative estimation. Consider firstly local polynomials of degree p. The estimator of the j^{th} derivative $m^{(j)}(x)$ $(0 < j \leq p)$ at point x is given by $\hat{m}^{(j)}(x) = j!\hat{\beta}_j(x)$ according to Taylor's theorem, where $\hat{\beta}_j(x)$ is obtained by minimizing $\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)\left(y_i - \sum_{j=0}^{p} \beta_j(x)(x_i - x)^j\right)^2$ in terms of the vector $(\beta_0(x), \ldots, \beta_p(x))$ (Fan & Gijbels, 1996). Thereby K is a kernel function and h the bandwidth controlling the degree of smoothing. Derivative estimators of this type have been implemented in the R functions `locfit` (contained in the homonymous package) and `locpoly` in package **KernSmooth**. Secondly, in spline smoothing, the usual way of estimating derivatives is to take the derivatives of the spline estimate. In other words, if $\hat{m}(x)$ is an estimate of $m(x)$, one considers $\frac{d^j}{dx^j}\hat{m}(x)$ as an estimator of $m^{(j)}(x)$. Several authors have pursued this idea, mostly in connection with penalization. Concretely, Heckman & Ramsay (2000) consider a minimization problem of type $\sum_{i=1}^{n}(y_i - m(x_i))^2 + \lambda P(m)$, where $P(\cdot)$ is some nonnegative valued operator penalizing high curvature of m. Their approach is implemented in the function `smooth.Pspline` in R package **pspline**, and a variant of it using numerical methods is provided in R function `D1D2` (**sfsmisc**).

Comparison of available routines

We compare the methods and R routines mentioned above using simulated and real data sets. Based on these results, and theoretical considerations, we conclude that the local polynomial estimators suffer from a systematic downward smoothing bias. The spline based methods perform somewhat better, but there is a general lack of *robust* derivative estimators, and smoothing parameter selection tools are mostly not satisfactory or not available. Examples of where derivative estimation is required are provided using sports science and environmental data.

References

Fan, J., and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. London: Chapman & Hall.

Heckman, N.E., and Ramsay, J.O. (2000). Penalized regression with model-based penalties. *The Canadian Journal of Statistics*, **28**, 241–258.

Newell, J., Einbeck, J., Madden, N., and McMillan, K. (2005). Model free endurance markers based on the second derivative of blood lactate curves. In: Francis et al. (Eds), *Proc. of the 20th IWSM*, 357–364, Sydney.

Functional Data Analysis and the Linear Mixed Effect Model

N. Coffey[1], K. Hayes[1], O. Donoghue[1] and A. Harrison[1]

[1] University of Limerick

Abstract

The development of sophisticated data collection tools has resulted in the production of high dimensional data. Functional data analysis (FDA) [5, 8] is an emerging statistical methodology used to analyse such data. Standard FDA methods include functional principal components analysis (FPCA), functional regression etc. Such methods explore the characteristic behaviour of the data but do have some limitations, e.g.replicate functions are smoothed as if the data are independent. However, replicates on the same individual are expected to be more similar than between individuals. This information is not used in FDA. Functional data can be viewed as a special case of longitudinal data and hence the `lme` model [2, 6] is explored as an alternative method for analysing such data. This model naturally incorporates nesting/grouping structures and longitudinal designs. The connection between smoothing methods and the `lme` model [1, 7, 9] can be exploited to smooth the data in the same step. Such connections are best made in [7], however the links between FDA methods and the `lme` model are not explored. We aim to examine FDA methods via the `lme` model and develop it to incorporate longitudinal designs for functional data.

Keywords: Longitudinal design, smoothing, P-splines.

1 Introduction

Functional data typically arises as a sequence of discrete measurements and involves the use of rapidly developing smoothing methods to determine the underlying smooth functions. Regression splines estimate basis coefficients via least squares but choice of location and number of knots K is a complex problem. Smoothing splines set $K = n$ (number of observations) and control the fit via a penalty term. Here, a smoothing parameter λ must be chosen. Both methods have computational difficulties when K is very large. P-splines [4, 9] allow for adaptive knot selection and K is chosen to be large but $<< n$, thus reducing the dimensionality of the problem. Representing P-splines as a `lme` model [1] ensures that λ can be chosen via RE(ML).

2 Methods

The data analysed were collected by Dr. Orna Donoghue and Dr. Andrew J. Harrison from the University of Limerick, [3]. We use P-splines and the linear truncated power basis of the form $g(t_j) = \beta_0 + \beta_1 t_j + \sum_{k=1}^{K} u_k (t_j - \kappa_k)_+ = \mathbf{X}\beta + \mathbf{Zu}$, where $\beta = (\beta_0, \beta_1)'$, $\mathbf{u} = (u_1, \ldots, u_K)'$,

$$\mathbf{X} = \begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_n \end{pmatrix} \quad \text{and} \quad \mathbf{Z} = \begin{pmatrix} (t_1 - \kappa_1)_+ & \cdots & (t_1 - \kappa_K)_+ \\ \vdots & \ddots & \vdots \\ (t_n - \kappa_1)_+ & \cdots & (t_n - \kappa_K)_+ \end{pmatrix}.$$

Then $\mathbf{y} = \mathbf{g} + \varepsilon = \mathbf{X}\beta + \mathbf{Zu} + \varepsilon$, $\mathbf{u} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I}_K)$ and $\varepsilon \sim N(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_n)$.

3 Results

The most general model for subject i is $y_{ij} = \mu(t_{ij}) + f_i(t_{ij}) + \varepsilon_{ij}$, $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2 \mathbf{R})$, $\mu(t_{ij}) = \beta_0 + \beta_1 t_{ij} + \sum_{k=1}^{K} u_k (t_{ij} - \kappa_k)_+ = \mathbf{X}\beta_\mu + \mathbf{Zu}_\mu$, $u_k \sim N(0, \sigma_u^2)$, $f_i(t_{ij}) = a_{i1} + a_{i2} t_{ij} + \sum_{k=1}^{K} \nu_{ik} (t_{ij} - \kappa_k)_+ = \mathbf{X}\beta_{f_i} + \mathbf{Zu}_{f_i}$, $(a_{i1}, a_{i2})' \sim N(0, \Sigma)$, $\nu_{ik} \sim N(0, \sigma_\nu^2)$. $\mu(t)$ is the population mean function and $f_i(t)$ are the subject-specific deviations from $\mu(t)$. Then $\mathbf{y} = \mathbf{X}\beta_\mu + \mathbf{Zu}_\mu + \mathbf{X}\beta_{f_i} + \mathbf{Zu}_{f_i} + \varepsilon$. An interaction model was fitted to our data, estimating subject mean functions and replicate-specific deviations from these means.
Using the lme model to estimate population average and subject-specific curves results in very good fits. Incorporating the smoothing step into the estimation procedure creates a strong link between the lme model and the regularisation approach implemented in FDA methodology.
FPCA determined that the presence of orthotics increases the range of motion in the Achilles tendon angle of injured subjects. We aim to carry out a similar analysis via the lme model and compare the results.

References

[1] Brumback, et al. (1999). Comment on Variable Selection and Function Estimation in Additive Nonparametric Regression Using a Data-Based Prior. *Journal of the American Statistical Association.* 94, 794-797.

[2[Diggle, et al. (1995). *Analysis of Longitudinal Data.* Oxford University Press.

[3] Donoghue, et al. (2007). Functional Data Analysis of Running Kinematics in Achilles Tendon Injury Subjects. *Journal of Medicine and Science in Sports and Exercise.* Reviewed.

[4] Eilers and Marx (1996). Flexible Smoothing with B-splines and Penalties. *Statistical Science.* 11, 89-102.

[5] Ferraty F and Vieu P (2006). *Nonparametric Functional Data Analysis.* Springer, New York.

[6] Laird N M and Ware J H (1982). Random-effects Models for Longitudinal Data. *Biometrics.* 38, 963-974.

[7] Speed T (1991). Comment on That BLUP Is a Good Thing. *Statistical Science.* 6, 15-51.

[8] Ramsay J O and Silverman B W (2005). *Functional Data Analysis.* Springer, New York.

[9] Ruppert, et al. (2003). *Semiparametric Regression.* Cambridge University Press, USA.

Wavelet Regression for Lack-of-Fit Tests in Semiparametric Mixed Models

H. Ding[1], G. Claeskens[1] and M. Jansen[1]

[1] Catholic University of Leuven, Belgium

Abstract

In this paper we study the asymptotic distribution of restricted likelihood ratio tests in mixed linear models with a fixed and finite number of random effects. In particular we study the testing power of wavelets for testing a hypothesized parametric model within a mixed model framework.

Keywords: Lack-of-fit test, likelihood ratio test, mixed models, penalization, restricted maximum likelihood, variance components, wavelets.

Introduction

The main aim of this paper is to construct test statistics based on wavelets for testing a parametric null model against a nonparametric alternative model. Our simulations show that the wavelet based lack-of-fit tests outperform the competitor based on penalized regression splines in several situations with a mixed model framework. A second result in this paper is that the asymptotic distribution that we obtain holds in general mixed models (not necessarily using wavelets).

Methods

Under the null hypothesis the parametric model is

$$H_0 : Y = \beta_0 + \beta_1 x + \ldots + \beta_q x^q + \varepsilon.$$

In matrix notation, a nonparametric lack-of-fit test contrasts this null model with a semiparametric alternative model of the form $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon}$. The design matrices of fixed and random effects are given by

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & \ldots & x_1^q \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \ldots & x_n^q \end{bmatrix}, \quad \boldsymbol{Z} = \begin{bmatrix} \psi_1(x_1) & \ldots & \psi_{K_n}(x_1) \\ \vdots & \ddots & \vdots \\ \psi_1(x_n) & \ldots & \psi_{K_n}(x_n) \end{bmatrix}$$

where $\psi_k, k = 1, ..., K_n$, are wavelet basis functions, and ε_i are independent identically distributed $N(0, \sigma^2_\varepsilon)$. The random effects $u_k \sim N(0, \sigma^2_u)$ are i.i.d. independent from the ε_k. Testing H_0 against the two-sided alternative that the conditional mean response has any different structure in the mixed model representation is equivalent with testing the now one-sided hypothesis $H_0 : \sigma^2_u = 0$ versus $H_a : \sigma^2_u > 0$. The mixed model formulation dramatically reduces the dimensionality of the testing problem. We then employ the profile restricted log-likelihood ratio test. We denote the restricted log-likelihood of the data under the alternative model H_a as $\mathcal{L}(\lambda)$; under H_0, we set $\lambda = 0$. Then our test statistic is $\mathcal{R}_n = 2\{\mathcal{L}(\hat{\lambda}) - \mathcal{L}(0)\}$. The test statistic has a mixture distribution, which is *not* always composed of chi^2 components, see our paper for more details.

Results

In a simulation study we investigated the power properties of the wavelet based lack-of-fit test statistic $\mathcal{R}_n$ and the spline based test statistic $\mathcal{R}_{ns}$. Simulated power curves are shown using the critical values obtained from the empirical and bootstrapped distributions under the null hypothesis. The wavelet-based tests often clearly outperforms the similar spline-based tests, especially for non-smooth alternative models.

FIGURE 1. Simulated power curves of wavelet and spline based tests for the Blocks and Bumps alternative functions, using critical values from both the empirical and bootstrapped distribution.

References

Claeskens G., Ding H. and Jansen M. (2007). Lack-of-fit tests in semiparametric mixed models. *Submitted*

Statistical Characterisation of Errors in Estimation of Haemodynamic Parameters from Bolus Tracking with Dynamic Magnetic Resonance Imaging.

N. Fitzgerald[1], F. O'Sullivan[1], G. Newman [2], D. O'Mahony[3] and N. O'Donovan[4]

[1] University College Cork

[2] University of Wisconsin, USA

[3] Cork University Hospital

[4] Victoria Hospital, Cork

Abstract

The most common type of stroke occurs when blood flow to an area of the brain is interrupted. Being able to identify the location and extent of damage caused is of diagnostic utility. The recent advancement of perfusion CT/MRI scans has made this possible in an easily accessible manner. During the scan a bolus of contrast agent is injected and monitored as it passes in the bloodstream through the brain. Indicator dilution theory [1] allows us to measure the haemodynamic parameters flow, volume and mean transit time. The central equation behind this theory is

$$C_T(t) = \int_0^t C_p(s)R(t-s)ds, \tag{1}$$

where C_T is the amount of contrast agent observed in a tissue region, C_p is the concentration of contrast agent observed entering through a main artery and R is the residue function. By reconstruction of R we can estimate the parameters flow, volume and mean transit time. This type of problem is an ill-posed inverse problem and best solved by deconvolution. We attempt to adopt a robust method of regularisation by identifying the optimal order of smoothing and selecting the regularisation parameter (λ) by a generalised cross validation criterion [5]. Results indicate an improvement on the truncated singular value decomposition (TSVD)[3] method. We also examine theoretical convergence of the method as a function of increasing signal to noise ratio (SNR).

FIGURE 1. Estimation of the residue function using regularisation, TSVD and the actual solution

Methods

The discretization of (1) yields the linear model

$$Y = [XR]_t + \epsilon_t, \quad \text{with} \epsilon_t \approx N(0, \sigma),$$

where

$$\begin{aligned} C_T &= [XR]_t \\ R_t &= R(t) \\ X_{t-s} &= \begin{cases} C_p(t-s)\Delta s, & s \leq t \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

We consider regularised estimates [5] which minimise

$$l_\lambda^m(R) = ||Y - XR||^2 + \lambda ||L_m R||^2$$

Here $||Y - XR||^2$ is the residual sum of squares and $||LmR||^2$ a penalty with L_m a discretized m-th derivative ($\lambda > 0$). The residue reconstruction is given by the formula

$$\hat{R} = (X^T X + \lambda L_m^T L_m)^{-1} X^T Y$$

The regularisation parameter λ is selected using a modified GCV criterion. Theoretical MSE convergence is well modelled by

$$\begin{aligned} MSE &\propto \sigma^{-2\beta} \\ \log(MSE) &= c - \beta \log(\sigma), \end{aligned}$$

where λ depends on properties of input, residue and order of smoothing (m).

FIGURE 2. Observed and theoretical rates of decrease of MSE estimate of blood parameters: Flow, Volume and MTT

Results

During simulations we compare MSE results to evaluate the quality of residue function estimation. Results of the simulations determined the optimal order of smoothing and a suitable GCV selection criterion. These solutions based on regularisation indicate improvement on the TSVD technique. We also investigate behaviour of error in estimation of flow, volume and MTT over a range of SNR. The observed rate of decrease is determined and compared to theoretical values.

References

[1] Meier P., Zierler K. L. (1962) Theoretical basis of indicator-dilution methods for measuring flow and volume. *Circ. Res.* 10, 393-407.

[2] Nychka D., Cox D.D. (1989 Convergence Rates for Regularized Solutions of Integral Equations from Discrete Noisy Data. *The Annals of Statistics* 17, 556-572.

[3] Ostergaard L, Weisskoff R.M., Chesler D.A., Gyldensted C. and Rosen B.R. (1996) High resolution measurement of cerebral blood flow using intravascular tracer bolus passages. I. Mathematical approach and statistical analysis. *Magn. Reson. Med* 36, 715-725.

[4] O'Sullivan F. (1995) A study of Least Squares and Maximum Likelihood for Image reconstruction in Positron Emission Tomography. *The Annals of Statistics* 23, 1267-1300.

[5] Wahba, G. Spline Models for Observational Data *CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59.* SIAM, Philadelphia.

Supported in part by the Irish Health Research Board.

Modelling uncertainty in phylogenetic trees constructed from distance matrices

W. Gilks[1]

[1] University of Leeds

Abstract

Phylogenetics is the study of evolutionary trees linking present-day species. A popular approach to estimating phylogenetic trees involves the construction of a distance matrix between extant species, typically based on DNA or protein sequence data. The advantage of this approach is that methods are generally simple and quick and avoid the raft of assumptions demanded by the more sophisticated approach of full-probability parametric modelling. However, distance-matrix methods fail to formally account for uncertainty in the phylogenetic reconstruction (although some methods implicitly exploit a variance model in a limited way).

We present a new agglomerative phylogenetic method in which only first and second distance moments are modelled, incorporating both tree-distortion and measurement noise. A stochastic implementation of our method allows full uncertainty in the phylogenetic reconstruction to be explored. Simulations reveal how information is lost or gained as we attempt to reconstruct increasingly ancient parts of a phylogeny.

Using Gene Expression Microarrays to Find Interactions

A. Canty[1] and S. Bashir[1]

[1] McMaster University, Hamilton, Ontario

Abstract

We examine the use of regular and robust linear models to look for interactions in designed gene expression microarray experiments. Our methods use resampling based approaches building on the Significance Analysis of Microarrays approach. The methods are applied to two sets of data from a study of Type 1 Diabetes conducted at the Hospital for Sick Children in Toronto. In one of these experiments we were interested in finding locus-locus interactions on gene expression and in the other we were interested in sex by genotype interactions. We also conduct a simulation study to examine the performance of the methods.

Keywords: Gene Expression Microarray, Significance Analysis of Microarrays, Robust Regression, Resampling.

Introduction

Type 1 Diabetes is a complex disease in which many genetic loci as well as environmental factors influence susceptibility to the disease. One major goal in researching this disease is to discover how the various genetic loci, and how genetic and environmental factors, interact to increase or decrease resistance to the disease. We will introduce a rodent model which can be used to examine some factors influencing Type 1 Diabetes. Gene expression microarrays can be used with this model and designed experiments to allow us to look for interactions between various genetic and environmental factors.

Methods

The rodent model which we shall consider is based on the Non-Obese Diabetec (NOD) and Non-Obese Resistant (NOR) strains. These are inbred strains of mice which are genetically very similar. In fact they are identical

by descent in 88% of the genome. Crucially they are identical by descent in the region of the chromosome containing the major genetic contributor to Type 1 Diabetes so this model allows us to examine some of the other genetic contributors. The 12% of the genome in which the NOD and NOR differ includes 4 regions linked to Type 1 diabetes. As well as the parental NOD and NOR strains, a number of congenic strains have been produced enabling us to look at how these regions affect diabetes resistance.
Microarrays can be used as an exploratory tool to measure gene expression and hence to find genes which differ in their expression in different conditions. Over the past decade many methods such as Significance Analysis of Microarrays (SAM, Tusher et al 2001) have been developed to examine this type of question. The current generation of microarrays allow gene expression to be measured for thousands of genes simultaneously. The question of coping with multiple testing is therefore a major problem. Due to cost considerations, the number of independent observations is generally quite small and the distribution of the expression measurements for a given gene is often non-normal. It is therefore common to use permutation or resampling based approaches. The basic method involves comparing the observed ordered test statistics against their expected values under permutations. For a given set of significant genes, the statistics based on the permutations can also be used to estimate the false discovery rate.
We discuss extensions of this basic methodology to look for interactions. Outliers are a common feature of microarray experiments so we examine how robust linear models can be used. As with SAM, we use resampling methods to approximate the null distribution of the test statistics. For factorial experiments the resampling can be done based on residuals from a fitted model satisfying the null hypothesis of no interaction. For the regular linear model residual resampling is a well known technique but there are complications with robust estimation so most bootstrap approaches for robust regression use case resampling. This is not appropriate for experiments in which the design matrix is fixed so we consider a number of residual resampling schemes for robust regression. The q-value approach (Storey, 2002) is used to determine genes which show evidence of an interaction. The q-value for a given gene is defined as the minimum false discovery rate when that gene is declared significant. This enables us to select a set of genes satisfying a limit on the false discovery rate.

References

Storey J.S. (2002). A Direct Approach to False Discovery Rates *JRSS B* 64, 3, 479-498.

Tusher V, Tibshirani R and Chu G (2001). Significance Analysis of Microarrays Applied to Transcriptional Responses to Ionizing Radiation. *Proc. Natn. Acad. Sci. USA* 98, 5116-5121

Mixture-Regression Cluster Model applied to Longitudinal Microarray Experiments

E. Holian[1] and J. Hinde[2]

[1] University of Limerick

[2] National University of Ireland, Galway

Abstract

The aim of this work is to explore various statistical techniques to identify genes which contribute to some change in phenotype level. For example, the response of fish kept under stressful conditions for various lengths of time. We aim to assess the level of *differential* expression of each gene in the tissue samples and also attempt to model the expression patterns of genes over time, not only to classify genes by similarities in expression patterns, but also to model these patterns as specified functions.

Keywords: Microarray; Longitudinal; Mixtures; Regression; Random effects.

Introduction and Background

In agriculture, particularly in fish farming, it is useful to be able to monitor animal health. Research being carried out on rainbow trout by the National Diagnostics Centre, Galway, aims to uncover a protein indicator which when present in the water occupied by the fish indicates signs of stress. Proteins are formed from the information held in the fish DNA genome, by a gene becoming *expressed*, so finding which gene or genes become expressed during stress will lead to researchers finding the relevant protein indicator of stress. Samples of liver tissue were extracted from rainbow trout fish exposed to confinement stress for varying lengths of time, at times 2,6,24,168 and 504 hours of stress, these times represented by $t, t \in 1:5$, and labelled as tissue variety *treatment*. Samples of liver tissue from unstressed fish left for the same period of time in a neighbouring tank are also included in the analysis, labelled as *control* samples.
Genetic expression for a tissue sample is measured on a glass slide known as a microarray. One microarray can measure the expression of thousands of genes simultaneously, the position of the measurement, also referred to

as a *spot*, on the array will identify which gene this measurement belongs to, referred to as the *probe*.
Let the measured expression for a probe at spot $s \in [1:21168]$ of the 80 arrays the expression was measured on,
In order to remove some experimental variation, a series of corrections are applied to the measured expressions, a process referred to as *normalisation*.
The differential expression profile between treatment and control for each probe s can be calculated as $\boldsymbol{Y}_s$, with elements Y_{st} for $t = [1, 5]$. Clustering these probes into groups of similarity may give some indication as to genes that co-regulate in the production of proteins in response to exposure to stress.

Formulation of Cluster Model and Estimation

The proposed Mixture-Regression Cluster Model is developed to model *and* cluster the genes into groups according to their expressions measured over time. This model is similar to that of the multivariate normal mixture model in that clusters are identified by the EM algorithm but is adapted to incorporate the flexibility of regression curves to fit the trends. In this way, additional features such as covariates, random effects and correlation structures can be incorporated into the model while potentially offering a considerable saving on the number of parameters required to model the trends.
For a particular cluster $i \in [1:c]$ let the differential response vector be modelled by $\boldsymbol{Y}_s = X_s\boldsymbol{\beta}_i + Z_s\boldsymbol{b}_{is} + \epsilon_{is}$. for fixed effects $\boldsymbol{\beta}_i$, usually the regression curve in time, specified by the design matrix X_s and any optional random effects $\boldsymbol{b}_{is}$ specified by design matrix Z_s. Where the errors have a normal density $\epsilon_{is} \sim N(0, \Sigma_i)$, and the random effects $\boldsymbol{b}_{is}$ have a normal density function $f_i(\boldsymbol{b_{is}}) = \phi(0, D_i)$, then the marginal model for $\boldsymbol{Y}_s$ is normally distributed, $f_i(\boldsymbol{Y}_s) = \phi(X_s\boldsymbol{\beta}_i, V_{is})$ with $V_{is} = Z_s D_i Z_s' + \Sigma_i$.
The full distribution $f(\boldsymbol{Y}_s; \Psi)$ is then a mixture of the clusters so that

$$\boldsymbol{Y}_s \sim \sum_{i=1}^{c} \pi_i N(X_s\boldsymbol{\beta_i}, V_{is}).$$

Let the set of parameters for each of the component densities be denoted by $\boldsymbol{\theta}_i = (\boldsymbol{\beta_i}, V_{is})$ then $\Psi = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_c, \pi_1, \cdots, \pi_{(c-1)})'$ is the vector containing all unknown parameters.
Estimation of these parameters, for a pre-specified number of components, c, can be done via the Expectation-Maximization (EM) algorithm, an iterative procedure which is initiated be a random allocation of probes into the clusters. The M-step estimates the parameters, $\Psi^{(1)}$, using this initial allocation, by a weighted regression using R-subroutines GLS for a model with no random effects and LME fitting a model with a random intercept

or slope. The E-step then updates the allocation ratios using the estimated set of parameters $\Psi^{(1)}$ from the M-step. The iterations continue until there is little difference in the observed log likelihood as calculated in the E-step.

Results and Remarks

Simulations have shown that the mixture-regression model can recover clusters successfully and for each resulting cluster can provide a parametric model for the longitudinal trend followed by probes in the same cluster.
To find the model specification of optimal fit to the data, certain features of the model can be varied and the model refitted. For example re-specifying, the number of clusters c, or re-specifying $X_s\boldsymbol{\beta}_i$ to be polynomials of varying degrees in time, re-specifying $Z_s\boldsymbol{b}_{si}$ to include a random intercept or slope and varying the correlation structure within each cluster Σ_i.
The optimal model is selected so that the log-likelihood is maximised or, if a penalisation for the number of parameters is more desirable, aim to minimise Akaike's Information Criterion AIC, or the more prudent Bayesian Information Criterion BIC.
We show how these procedures were applied to a filtered subset of the fish stress probes resulting in a 17-component mixture. Some discussion will also follow as to how a number of these interesting clusters have proven to be a very useful source of information in understanding the molecular processes tested in these experiments.

Acknowledgments: Special Thanks to colleagues working at the National Diagnostics Centre, and in the Mathematics Department of National University of Ireland, Galway, where the work was carried out.

References

Bowtell, D., and Sambrook, J. (2002). *DNA Microarrays*. Cold Spring Harbor Laboratory Press.

Gentleman, Rossini, Dudoit, and Hornik (2003). The Bioconductor FAQ, *http://www.bioconductor.org*

McLachlan, G.J., and Peel, D. (2000). *Finite Mixture Models*. New York: John Wiley & Sons.

McLachlan, G.J., and Krishnan, T. (1997). *The EM Algorithm and Extensions*. New York: John Wiley & Sons.

Verbeke, G., and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag.

Goldstein, H. (1995). *Multilevel statistical models*. London : E.Arnold; New York : Halsted Press.

Diggle, P.J., Heagerty, P., Liang, K., and Zeger, S.L. (2002). *Analysis of Longitudinal Data*. Oxford University Press.

Cui, X., Kerr, K.K., and Churchill, G.A. (2003). *Transformations for cDNA Microarray Data*. Statistical Applications in Genetics and Molecular Biology, **2(1)**, Article 4.

Wit, E., and McClure, J. (2004). *Statistics for Microarrays: design, analysis and inference*. John Wiley & Sons.

Wolfinger, R.D., Gibson, G., Wolfinger, E.D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R.S (2001). *Assessing Gene Significance from cDNA Microarray Expression Data via Mixed Models*. Journal of Computational Biology, **8(6)**, 625-637.

On the Detection of Selective Sweeps

David Ramsey[1]

[1] University of Limerick

Abstract

A selective sweep occurs when a favourable mutation spreads through the population. A selectively neutral allele which lies close to this mutation tends to spread through the population. This effect is known as hitchhiking and results in the population exhibiting very little genetic variation in the region around the mutation. Unfortunately, other demographic effects can have similar consequences. However, recent rapid advances in the mapping of genomes has provided the necessary data to develop new improved statistical methods for detecting selective sweeps. This talk outlines an approach to detecting multiple selective sweeps on a single chromosome.

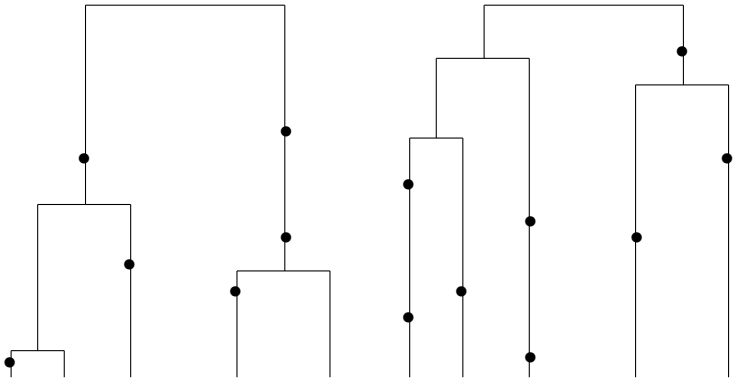
Keywords: Selective sweep, gene mapping, coalescent theory, multiple likelihood tests

Introduction

Recent rapid advances in the mapping of genomes has provided a wealth of data which can aid us in our understanding of evolutionary processes. Genetic maps of a sample do not just contain information about the present population, but indicate how the population has evolved. The variation in the present sample can be described by a coalescent tree (Kingman [1982]). This is a Markov chain model which considers the demographic changes within a population, together with the effects of selection, mutation and crossover. A coalescent tree models the reverse process of evolution back to the most recent common ancestor of the sample. A coalescent occurs when two members of the sample are traced back to their most recent common ancestor. Classical tests for selective waves (e.g. Tajima [1989]) are based on data from a small segment of the genome. It is assumed that these segments are short enough for the effects of crossover to be neglected and the rate of mutations is low enough to neglect the possibility of two mutations at the same locus. Mutations occur on each branch of the coalescent process as a Poisson process with constant rate.

The null hypothesis in classical tests for selective sweeps is that the population size is constant and there is no selection. The alternative is that

a selection sweep has occurred. Typical realisations of the coalescent tree under these hypotheses are presented below (the dots represent mutations):



Typical coalescent under H_0 Typical coalescent under H_1

A segregating site is a locus at which a sample is polymorphic. The number of segregating sites is the number of mutations in the coalescent tree. Under H_1 it is likely that there will be a relatively large number of ancestral lines for a long period of time. This leads to a large number of segregating sites. However, pairs of individuals will tend to differ only at a small number of sites. This relationship is used to test for selection. However, the coalescent tree resulting from e.g. a population bottleneck has similar properties.

When a selective sweep occurs, the action of crossover means that the genetic diversity of the population will be maintained outside a narrow band around the favourable mutation, whereas demographic effects will have a global effect. Genome mapping and advances in coalescent theory have enabled a global approach to detecting selective sweeps. Nielsen et al. (2005) developed an effective test using coalescent theory to predict how a selective sweep decreases variability relative to "background" variability which depends on demographic effects. Their approach assumes that only one sweep occurs on a chromosome, but real data indicate that several sweeps can occur on one chromosome. We consider such models.

References

Kingman, J.F.C. (1982) On the Genealogy of Large Populations. *J. App. Prob.* 19A, 27-43.

Nielsen R., Williamson S., Kim Y., Hubisz M. J, Clark A. G. and Bustamante C. (2005) Genomic scans for selective sweeps using SNP data. *Genome Research,* 15, 1566-1575.

Tajima F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics,* 123, 585-595.

Bayesian Kernel Classification Method for Multinomial data

K. Domijan[1] and S. Wilson[1]

[1] Trinity College Dublin

Abstract

A Bayesian approach to multi-category classification based on reproducing kernel Hilbert space (RKHS) is proposed. The likelihood function is taken to be multinomial logistic and is modelled through a latent variable. The hierarchical model is treated with a fully Bayesian inference procedure. The Bayesian RKHS classifier is aimed at high dimensional data and is able to achieve good classification results without dimension reduction, considerably reducing the manual pre-processing that is usually required.

Keywords: Bayesian inference, classification, multinomial logistic regression, reproducing kernel Hilbert spaces.

Introduction

The training data are n samples $(\mathbf{x}_1, \mathbf{y}_1)$, ..., $(\mathbf{x}_n, \mathbf{y}_n)$ where the predictors $\mathbf{x}_i \in \mathbb{R}^J$ are vectors of feature values and $\mathbf{y}_i = (y_{i1}, ...y_{iK})$ are categorical response variables with $y_{ik} = 1$ if $\mathbf{x}_i$ belongs to a class k and 0 otherwise. The multinomial logistic likelihood for the training data is given by $p(\mathbf{y}|\mathbf{z}) = \prod_{i=1}^{n} \prod_{k=1}^{K} p(y_{ik} = 1|z_{ik})^{y_{ik}}$, where

$$p(y_{ik} = 1|z_{ik}) = \frac{\exp(z_{ik})}{\sum_{l=1}^{K} \exp(z_{il})}. \quad (1)$$

In the RKHS approach z_{ik} are taken to be linear combinations of the reproducing kernel functions: $z_{ik}(\mathbf{x}_i, \beta_k, \theta) = \sum_{l=1}^{n} \beta_{kl} K(\mathbf{x}_i, \mathbf{x}_l|\theta) = \mathbf{K}_i^T \beta_k, \;\; i = 1, ..., n$, where $K(\mathbf{x}_i, \mathbf{x}_l|\theta) = \exp(-\sum_{j=1}^{J} \theta_j (x_{ij} - x_{lj})^2)$, $i, l = 1, ..., n$, is the Gaussian kernel and $\boldsymbol{\beta}_k = [\beta_{1k}, \beta_{2k}, ..., \beta_{nk}]$ is a set of regression parameters corresponding to class k. In order to make the model of full rank $\boldsymbol{\beta}_K$ is set equal to 0. The functions z_{ik}s are re-defined as Gaussian random variables with means $\mathbf{K}_i^T \beta_k$ and variance σ^2 (Mallick et al., 2005).
The prior model is specified as $z_{ik} \sim N(\mathbf{K}_i^T \beta_k, \sigma^2)$, $\beta_k \sim MVN(0, \sigma^2 \mathbf{T}_k^{-1})$, $\sigma^2 \sim IG(\gamma_1, \gamma_2)$, $\tau_{ik} \sim G(\gamma_3, \gamma_4)$ and $\theta_j \sim U(0, a_u)$. $\mathbf{T}_k$ is a matrix with diagonal entries $\tau_{1k}, ..., \tau_{nk}$, G denotes a gamma prior, IG an inverse gamma.

A Metropolis-within-Gibbs algorithm was used for sampling from the posterior. The output from the MCMC is a a set of samples ($\beta^{(m)}$, $\theta^{(m)}$, $\mathbf{z}^{(m)}$, $\sigma^{2\,(m)}$, $\tau^{(m)}$), for $m = 1, ..., M$ iterations from the joint posterior: $p(\beta, \theta, \mathbf{z}, \tau, \sigma^2|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{z}, \beta, \theta, \tau, \sigma^2)p(\mathbf{z}|\beta, \theta, \sigma^2)p(\beta|\tau)p(\theta)p(\tau)p(\sigma^2)$.
Parameters β, σ^2 and τ are block updated directly from their full conditional distributions via Gibbs steps. Parameters θ and $\mathbf{z}$ are sampled using a Metropolis step within the Gibbs algorithm.

Results

Khan et al.(2001) describe gene expression profile data consisting of 83 mRNA microarray slides, divided into a training and testing data set. Each microarray slide corresponds to an individual suffering from one of four tumour types. The total of 2308 genes profiles are reported for each slide. This corresponds to a 4 category classification problem with a large number of features ($J = 2308$) and small number of observations ($n = 83$).
For the same split of data into training and testing sets as used by Khan et al.(2001) and Lee et al.(2004), the algorithm classifies all of the observations correctly, with all of the features utilized in the construction of the kernel matrix. In addition, the algorithm was ran for ten random splits of the data with approximately equal sample sizes in the training and testing data set. The average misclassification error was 0.06 with standard deviation 0.04.

Discussion

The kernel classifier presented here is a fully Bayesian, genuine multicategory extension of the Bayesian binary kernel classifier. It achieves good results with the high dimensional microarray data set. The classifier does not require pre-processing steps to reduce the dimension of the feature space.

References

Khan et al. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7, (6), 673-679.

Lee et al. (2004). Multicategory Support Vector Machines: Theory, and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99, 67-81.

Mallick et al. (2005). Bayesian classification of tumors using gene expression data. *J. Royal Statistical Soc. B*, 67, 219-234.

Statistical Issues in the Assessment of Comorbidity Influence in Medical Studies

V. Kiri[1]

[1] Parexel International

Background

Most patients with a given chronic disease suffer from other diseases. These comorbidities can be important factors in assessment of risks associated with morbidity and mortality of such patients. In a particular chronic disease population, accurate assessment of comorbidity patterns might help identify common diagnostic markers relevant in the aetiology of specific disorders as well as comorbid conditions. Comorbidities may occur more frequently in patients with a particular chronic disease than in others of similar demographic characteristics who are free of the condition. Whilst certain factors may be considered as known risk factors for each of such diseases in the general population, the chronic disease may itself be a risk factor for some of these other diseases. Of course, determining the presence of a combination of diseases is important for clinical practice. In a given chronic disease, such information can influence the quality of life of the patient as well as decisions on treatment. There are many examples and discussions in epidemiology of studies where lack of adequate control for the possible influence of comorbidity has resulted in effect estimates confounded by disease severity.

The recent withdrawal of many high profile drugs on grounds of adverse effects deduced from epidemiologic studies on non-randomized (real-life) observational data have highlighted the role of such studies in drug safety assessment. But compared with randomised controlled trials (RCTs), observational studies are more prone to bias. Never the less, such studies can provide valuable insight on drug safety in rare adverse events, particularly in situations where long-term RCTs may not be feasible. These days, such studies are often initiated as part of either a regulatory requirement or a risk management plan, and they are increasingly being utilised to complement routine pharmacovigilance activities. However, because of the inherent risk of bias associated with these studies, it is universally acknowledged that appropriate design, efficient analytical strategy and appropriate reporting are essential. Unfortunately, in practice, these requirements are often ignored with resulting consequence of contradictory results among studies conducted in different samples of the same population and treatments).

We identify three specific areas that require appropriate handling of comorbidity in clinical research: 1. The common practice in epidemiologic studies of assuming a constant effect for comorbidity, which effectively suggests that the duration of the condition does not influence prognosis. We challenge the assumption by demonstrating the time-dependent nature of the influence of certain comorbidities on patient survival. 2. The common practice in medical studies of measuring comorbid conditions by modelling the number of comorbidities as a covariate. The underlying hypothesis is that risk increases with the number of comorbidities present. However, such an approach does not identify, nor adjust for, combinations of specific comorbidities, which may confer adverse prognosis. We question these approaches by demonstrating that: outcome may not be associated linearly with the comorbidity count, the weights combining a set of binary comorbidities need not be positive (i.e. hypothesis that outcome worsens with increasing comorbidity may be false), and there is value in identifying specific interactions that influence prognosis. 3. Exploration of the relationships between comorbidity and a particular (a) chronic disease and (b) drug using appropriate methodologies. In drug safety assessment of a candidate-signal, two distinct possibilities need exploration: signal-from-disease and/or signal-from-drug relationships. In the healthcare setting (i.e. the primary source of data for most observational studies), the decision to give a particular treatment to a particular patient with a given disease is generally based on patient specific characteristics, the most important of which is disease condition. Thus, confounding by indication/disease severity is a common source of bias. Consequently, failure to properly control for the bias could lead to serious flaws in the study, which for drug safety assessment might result in false signals. The use of propensity scores has been suggested as one of the best analytical approaches for handling this problem on the basis of its theoretical potential for minimising the bias. However, the methodology does not extend to assessment of possible signal-from-disease relationship. We propose a method for assessing the relationship between the candidate-signal and the main disease of interest, so as to facilitate disentanglement of any disease effect from a drug (adverse) effect and minimization of the risk of false signals in pharmacovigilance activities. We demonstrate an application of the method to investigate associations between chronic obstructive pulmonary disease (COPD) and two specific candidates- pneumonia and osteoporosis in the UK general practice research database (GPRD).

Analysis of mismatch negativity data via spatially smooth ANOVA

K. Choudhury[1] and C. Pettigrew[1]

[1] University College Cork

Abstract

Mismatch negativity (MMN) is a fronto-centrally distributed event-related potential that is elicited by any discriminable auditory change. A recent study demonstrated that the MMN responses of language-disordered (aphasic) subjects ($n = 6$) were attenuated compared to normals in response to complex tone stimuli (deviating in duration) and speech stimuli (words and non-words). A 32-channel Neuroscan® Quikcap was used to record the nose-referenced EEG. MMN waveforms were derived from the data following a series of post processing steps [1]. In [1], the analysis was based on MMN amplitudes calculated as the largest negative peak at the maximum electrode (FCZ for tones, FZ for speech stimuli). This data was analysed using 2-way ANOVAs with repeated measures (group and condition factors) and 3-way ANOVAs with repeated measures (group, condition and electrode).

Our reanalysis of the data utilizes data from all electrodes, rather than just one or two, to estimate and test for experimental effects. It is in effect, a two stage analysis. At the first stage, we do a pointwise ANOVA analysis. In this case, the appropriate analysis model is a nested crossed model of the form:

$$y_{ijk} = \mu + \tau_i + \beta_j + \gamma_{k(j)} + (\tau\beta)_{ij} + \epsilon_{ijk}, \tag{1}$$

where y_{ijk} is the (averaged) MMN response. The spatial index x is omitted for brevity. The index $i = 1, 2, 3, 4$ runs over experimental conditions, i.e. respectively Frequency, Duration, Word and Non-word. Similarly $j = 1, 2$ stands for the group of subjects, respectively controls and aphasics. Finally $k = 1, 2, .., 6$ indexes the subjects in each group. In model (1), μ stands for the grand average MMN response, τ_i stands for the main effect of the i-th experimental condition, β_j stands for the main effect of the j-th group, $\gamma_{k(j)}$ stands for the main effect of the k-th subject nested within the j-th group and $(\tau\beta)_{ij}$ is the interaction effect between the i-th experimental condition and the j-th group. Finally, ϵ_{ijk} are i.i.d. mean 0 measurement errors with standard deviation σ. We will assume that the experimental and group effects τ_i and β_j and their interaction $(\tau\beta)_{ij}$ are fixed quantities,

while the subject effects $\gamma_{k(j)}$ are random, i.e. i.i.d. from some common distribution with standard deviation σ_γ. The parameters are subject to the usual linear model constraints. Estimates for this model can be derived using the orthogonality of the effects.
In the next stage, we seek that the estimates of the linear model in be spatially smooth. This is achieved using the following penalized least squares criterion:

$$L_P(\beta) = \sum_{i=1}^{N} [Y(x_i) - X\beta(x_i)]^T [Y(x_i) - X\beta(x_i)] + \sum_{c=1}^{k} \lambda_c \int_D p_c(\beta_c(x))dx, \tag{2}$$

where the linear model is written in the form: $Y = X\beta + \epsilon$ and x_i denotes location of the i-th electrode. We can then show the following:
Theorem: *Let $\hat{\beta}^{LS}(x)$ be an ordinary least squares estimator of the linear model at x. If the least squares criterion permits a parameter-ANOVA decomposition, i.e.*

$$[Y(x) - X\beta(x)]^T [Y(x) - X\beta(x)] = \sum_{c=1}^{K} [l_c^T Y(x) - \beta_c(x)]^2, \qquad \forall x$$

then $\hat{\beta}^{PLS}(x)$, the penalized least squares estimator of β, i.e. the minimiser of (2), can be obtained (component-wise) as: $\hat{\beta}_c^{PLS} = S_c \hat{\beta}_c^{LS}$, where S_c are smoothing operators corresponding to the penalty functions in (2). Furthermore, if the ordinary least squares estimator satisfies a linear constraint of the type: $\sum_{j=1}^{J} c_j \hat{\beta}_j^{LS}(x) = 0$, this constraint will also be satisfied by $\hat{\beta}^{PLS}(x)$, provided we choose $S_1 = S_2 = \ldots = S_J$.
We have developed kernel smoothers for the MMN problem. The smoothing parameter is chosen using ordinary cross-validation. Based on this idea, we can develop smoothed maps of estimated effects for this data set, as follows:

These maps show that MMN values for aphasics are more positive than controls across the brain. Estimates for other factors and methods of inference for such effects will be also presented.

References

(1) Pettigrew, C.M., Murdoch, B.E., Kei, J., Ponton, C.W., Alku, P. and Chenery, H.J. (2005). The mismatch negativity (MMN) response to complex tones and spoken words in individuals with aphasia. *Aphasiology*, 19(2), 131-163.

Developing a 3 state Markov model for Northern Ireland chronic kidney disease patients

A. Rainey[1], A. Marshall[1], K. Cairns[1], M. Quinn[1], G. Savage[1] and D. Fogarty[1]

[1] Queen's University of Belfast

Introduction

Chronic Kidney Disease (CKD) is now recognised as a common condition and is associated with ESRF (End Stage Renal Failure) and premature cardiovascular death. Renal replacement therapy in the UK is rising rapidly, costing over 2% of the total NHS budget. Consequently there exists a great demand for a model that will represent the natural history of this prevalent disease. Glomerular Filtration Rate (GFR) represents an approximate percentage of kidney function. However, this measure is an extremely difficult element to obtain accurately from blood samples. Instead an estimated GFR (eGFR) is more commonly used and obtained from the following MDRD1 equation:

$$eGFR = 186x(y/88.4) - 1.154x(Age) - 0.203x(0.742if female)x(1.210if black) \quad (1)$$

where y is the creatinine; a substance found in the blood and urine. These eGFR results are often collected as part of a patient's blood sample and enable the progression and regression of patient kidney function to be explored through Markov Modelling. This abstract outlines the transition rates and preliminary Markov Model for CKD.

The Theoretical Model

Figure 1 illustrates a three state model devised to represent the flow of CKD patients based on their eGFR as an indicator for kidney function. An eGFR<60mls/min highlights a potential CKD patient thus this criteria was used to define two states within the Markov model, 'No CKD' and 'CKD', in addition to the absorbing state 'Death'.
This system can be expressed as two linear differential equations:

$$\frac{dN}{dS} = Cr_{21} - N(r_{12} + r_{13}) \quad (2)$$

$$\frac{dC}{dS} = Nr_{12} - C(r_{21} + r_{23}), \quad (3)$$

where N and C represent the No CKD and CKD states respectively, r_{ij} defines the transition rate from state i to state j and S is the number of days. Due to their coupled nature, these equations were solved numerically using a function within MATLAB known as 'ode45'. Initial transition estimates will determine a distribution for the length of stay for patients in each state, no CKD and CKD, as shown in Figure 2.

The observed model

In Northern Ireland 2,892,340 routine blood samples from 1st January 2001 - 31st December 2002 were extracted from regional laboratories. Merging these to an enriched General Practice database for those 18 years or older, produced a cohort of 77,615 patients containing 312,120 results. A recently corrected version of the MDRD equation known as the IDMS was used to estimate the GFR for all samples. In order to generate length of stay distributions, the large dataset was sorted by patient, day and CKD state, respectively. An assumption was made for those patients with multiple tests on the same day whereby the worst case state was assumed the 'true' state and all other results for that day were discarded. This reduced the sample size to 296,320 of which 29.8% had an eGFR<60ml/min. A unique code was developed in MATLAB to calculate the length of stay in each state before transition to another state. From this, a distribution of the total length of stay was achieved, as shown in Figure 3.
This research utilizes MATLAB to fit the theoretical model to the observed data thus providing the actual transition rates for the Markov Model for CKD patients.

Conclusion

This study introduces a three compartment model to represent the flow of patients through kidney function states in relation to the unprecedented

extreme growth of CKD, using eGFR as a sound measure. It reveals the techniques exploited in MATLAB in order to obtain the transition rates. Future work will deal with expanding this three compartment model to six states, ergo will accommodate all possible well known stages of CKD. Markov Modelling has a unique ability to handle costs and weighted utilities thus the final product will serve as a useful resource in the world of Nephrology.

References

Levey et al. (1999). A more accurate method to estimate glomerular filtration rate from serum creatinine: A new prediction equation. Modification of Diet in Renal Disease Study Group. *Ann. Intern. Med.*, 130(6), 461-470.

Statistical Issues in Radon Mapping

P. Murphy[1] and C. Organo[2]

[1] University College Dublin

[2] Radiation Protection Institute of Ireland, Dublin

Abstract

The World Health Organisation estimates that between 6% and 15% of all lung cancer deaths per year are caused by exposure to indoor radon. Globally it is estimated that this equates to up to 170,000 cases.
The Radiological Protection Institute of Ireland (RPII) is the organisation tasked by the government with responsibility for radon issues in Ireland. The RPII have estimated that the average radon level in Ireland of $89Bq/m^3$ and their work has also established that Ireland exhibits extreme maxima in the tens of thousands of Bq/m^3. These results place Ireland at the upper end of radon levels in Europe. Indeed it has been estimated that approximately 13% of lung cancer deaths in Ireland are due to indoor radon exposure.
Once a house has been established to have high radon concentrations, mitigation is relatively easy with consequent significant reductions in the risk of death due to lung cancer being eminently feasible. The challenge therefore is to establish with as much accuracy as possible which houses are at risk. Work on producing accurate maps of radon concentrations has been conducted since the 1980s. In 1990 a reference level was established at $200Bq/m^3$ such that remediation is advised in those houses which exceed this reference level. The statistical task involved in producing radon maps is to estimate the proportion of homes P_{RL} in a particular region which exceed this threshold.
Early work in the UK established that radon levels could be modelled using a Log-Normal distribution. Further research has shown that it is possible to achieve an improved fit between the observed data and the theoretical Log-Normal model if one assumes an underlying background Radon Level common to all houses in the population. Empirical evidence in the UK showed that a background level of $4Bq/m^3$ was appropriate there, while research conducted in Ireland indicated that $6Bq/m^3$ was a more suitable level for the Irish background.
This paper will begin by describing three modelling approaches using the Log-Normal and Beta distributions that have been used in the UK and

Ireland to produce point estimates for P_{RL}. We will then report the results of recent work that was conducted in collaboration with the RPII which investigates the effect of outliers on predictions of P_{RL} and computes confidence intervals for P_{RL}. This work also provides for the first time accurate determinations of appropriate sample sizes to be used in surveys for the construction of radon maps.

Keywords: Radon Mapping, Log-Normal Modelling, Sample Size Determination, Simulation.

Parallel Coordinates: Extensions and Variations

C. Hurley[1], R. Oldford[2]

[1] National University of Ireland, Maynooth

[2] University of Waterloo, Canada

Abstract

This talk presents various extensions to parallel coordinate plots. A construction is given where all pairs of variables appear adjacently. An extension to parallel planes is described. A further extension corresponding to connected two or three dimensional scatterplots is discussed.

Keywords: Cable plot, Euler path, Grand tour, Scatterplot interpolation.

Introduction

Parallel coordinate displays are multivariate data displays where variables are assigned to parallel, equispaced axes, observations are plotted on the appropriate axis and lines are drawn connecting observations belonging to each case. These displays show (i) all one-dimensional variable distributions, (ii) bivariate correlation and (iii) some high-dimensional outliers and clusters.
Parallel coordinate displays were introduced by Inselberg(1985) and later developed by Wegman(1990). Interestingly, Friendly and Denis (2004) trace the origins of parallel coordinate displays to the work of d'Ocagne(1885). Wegman and co-authors (1991) have also proposed a number of extensions to parallel coordinates, notably the parallel coordinate grand tour which rotates the axes in R^n and more recently, the smoothed grand tour,
In this talk we present further variations and extensions of parallel coordinate displays.

All-pairs

Parallel coordinate are a compact alternative to scatterplot matrices for displaying multivariate data. The penalty for this compactness is that for

n variable data, only $n-1$ of the pairwise relationships are shown whereas the scatterplot matrix shows all $\binom{n}{2}$ bivariate relationships.
We investigate parallel coordinate displays showing all $\binom{n}{2}$ bivariate relationships. Such a display uses (at least) $\binom{n}{2}+1$ axes when n is odd, and $\binom{n}{2}+n/2$ axes when n is even and so is appropriate when n is not too large. Finding these sequences of variable axes amounts to constructing an Euler trail on the complete graph with n nodes.

Cables

We also describe a three dimensional version of parallel coordinate displays where pairs of variables are assigned to parallel, equispaced planes. Line segments are drawn connecting the observations belonging to each case. The resulting structure is a collection of "cables" living in R^3 which we explore using rotation. Since four variables are assigned to each pair of parallel planes, each section of a cable plot shows a four dimensional subspace of the n-dimensional data space.
Dispensing with parallel axes and planes, we could just draw line segments connecting the sequence of pairs (or triples) of observations belonging to each case. This can be thought of as a sequence of overlaid two-dimensional (or three-dimensional) scatterplots, where line segments connect the coordinates of a case. Here we have gained an extra dimension for data coordinates, but sequence position is no longer displayed explicitly as position along an axis.

References

D'Ocagne, M. (1885), *Coordonnées parallèles et axiales: Méthode de transformation géométrique et procédé nouveau de calcul graphique déduits de la considération des coordonnées paralleles.* Paris: Gauthier-Villars

Inselberg, A. (1985), "The plane with parallel coordinates", *The Visual Computer*, 1, 69-91.

item [Moustafa, R and Wegman, E.] (2006), "Multivariate continuous data- parallel coordinates", *Graphics of Large Datasets: visualizing a million'*, eds A. Unwin, M. Theus, H. Hofman, Springer, 143–155.

Wegman, E.J. (1990), "Hyperdimensional data analysis using parallel coordinates", *Journal of the American Statistical Association*, 85, 664-675.

Wegman, E.J. (1991), "The grand tour in k-dimensions", *Computing Science and Statistics: Proceedings of the 22nd Symposium in the Interface*, 127-136.

A model system for experiments on competition for site occupancy

C. Brophy[1], I. Fagerli[2], S. Duodo[2], M. S.[2] and J. Connolly[1]

[1] University College Dublin

[2] University of Tromsø, Norway.

Abstract

Nodulation competitiveness is a key factor affecting the ability of legume plants to fix nitrogen from the atmosphere. We developed a system of experimentation and a model of competition for occupancy of a fixed number of sites based on multinomial baseline category logit random effects models. We tested the nodulation competitiveness of three strains of Mesorhizobium loti (called A, B and C). Comparing the initial proportion of each strain present to the final proportion of nodules occupied by each strain, we found that strain C out-competed A and B at almost all initial proportion mixtures.

Keywords: Multinomial Logit Models, Random Effects, Offset, simplex design, Occupancy Competition, Rhizobia, Legume, Nodulation Competitiveness.

Introduction

Rhizobium is a bacterium that can invade the roots of legumes. There it can form nodules and fix atmospheric nitrogen. On each root there are a limited number of invasion sites and once a site has been invaded by a bacterium it cannot be invaded by another. An experiment was carried out to test the competitiveness of three strains of Mesorhizobium loti, A, B and C in occupying these sites. Each root in the study was inoculated with a solution (the inoculum) containing a mixture of each of the three strains. Each root gives a trinomial response, the total number of invasion sites occupied by each of the three strains. The primary question is whether the observed occupancy of invasion sites by strains reflects their proportions in the inoculum.

Methods

The simplex design used in this experiment included seven different initial mixtures of the three strains (Table 1) by two overall cell densities in the

Mixture #	A	B	C
1	0.85	0.07	0.08
2	0.15	0.74	0.11
3	0.13	0.08	0.79
4	0.41	0.26	0.33
5	0.55	0.35	0.10
6	0.52	0.07	0.41
7	0.14	0.39	0.48

TABLE 1. The initial proportion of each strain present in the innoculum. This is a simplex design with seven mixtures.

inoculum. Each mixture by cell density combination was replicated 5 times, giving 70 'clusters' in total. After 4 weeks the number of nodules occupied by each strain was counted in each cluster. We analysed the data using a variant of the multinomial baseline category random effects model with the following notation for the i-th cluster. The response is (n_{iA}, n_{iB}, n_{iC}), the number of sites occupied by strain A, B and C respectively. The multinomial parameters in the model are $(\pi_{iA}, \pi_{iB}, \pi_{iC})$, the expected proportion of nodules occupied by strain A, B and C respectively. The proportions of each strain present initially in the inoculum are (P_{iA}, P_{iB}, P_{iC}). If all strains are equally competitive (a null hypothesis of a competition model) $\pi_{iA}/P_{iA} = \pi_{iB}/P_{iB} = \pi_{iC}/P_{iC}$ or $\ln[(\pi_{iA}/P_{iA})/(\pi_{iC}/P_{iC})] = 0$ and similarly for B vs C. If the null hypothesis is not true then we wish to see what design factors affect the relative competitiveness of the strains. We modify the usual multinomial baseline category model (Agresti 2002) to allow for the initial proportions and include a random cluster effect (Hartzel et al. 2001) to give the linear predictor for A vs C as

$$\ln\left(\frac{\pi_{iA}/P_{iA}}{\pi_{iC}/P_{iC}}\right) = \beta_A P_{iA} + \beta_B P_{iB} + \beta C P_{iC} + \beta_D Dens_i + u_i,$$

where $i = 1, .., 70; Dens_i = 0$ (1) if the i-th cluster had low (high) initial cell density; u_i is a random effect for the i-th cluster and the β's are the regression coefficients. Model [1] can be rewritten as

$$\ln(\pi_{iA}/\pi_{iC}) = \beta_A P_{iA} + \beta_B P_{iB} + \beta_C P_{iC} + \beta_D Dens_i + \ln(P_{iA}/P_{iC}) + u_i,$$

where $\ln(P_{iA}/P_{iC})$ enters the fitting as an offset. A similar model compares strains B and C and the random terms in the two models may have different variances and non-zero covariance. Other explanatory variables can be readily added. We fitted the models using the NLMIXED procedure in SAS. Combining the two models we predicted the proportion of nodules infected by each strain for each mixture by cell density combination.

Results

Interaction terms between P_{iA}, P_{iB} and P_{iC} were included in the final models. We found that C was the most competitive strain at both densities (Fig. 1). It completely out-competed both strain A and B at high density and was dominant at almost all mixtures at low density.

Discussion

The methods developed here are applicable to many situations where there is competition for occupancy of a limited number of sites and generalise

readily to more than three components. The inclusion of the offset discounts for the initial proportion of each competitor initially present, allowing competition to be readily assessed. Including the random effect from the model allows for possible overdispersion due to variation in proportions from replicate to replicate.

References

(1) Agresti A. (2002). *Categorical data analysis* John Wiley. New Jersey.

(2) Hartzel J., A. Argesti and B. Caffo (2001). Multinomial logit random effects models. *SM* 1, 81-102.

A study of socio-economic status bias in the Quarterly National Household Survey

Ó. Burke[1], P. Murphy[1]

[1] University College Dublin

Abstract

There is much literature to indicate that a relationship exists between an individual's socio-economic status (SES) and their willingness to participate in surveys. Studies on SES bias among non-respondents point to a so-called "middle class effect" wherein respondents from higher socio-economic groups are more inclined to respond than those from lower socio-economic groups.,
The aim of this study, conducted with the Central Statistics Office, was to determine if the socio-economic status (SES) of an individual has a significant effect on whether or not the individual responds in a survey.,
In the 2002 Quarter 3 Quarterly National Household Survey (QNHS), participants were asked for their permission to conduct a further survey on electoral behaviour. Information is available for all participants in the QNHS and so this provides us with details of both the respondents and non-respondents for this supplementary survey on electoral behaviour.,
It was shown, using logistic regression, that the social class, principle economic status and level of education of an individual do not have significant effects on the individual's response.,
The results of this study were based on a very large scale data set and provide evidence of a departure among Irish respondents from the traditional behaviour described in the literature.,

On Modelling Correlated Binary Co-morbidities

S. Conde[1] and G. MacKenzie[1]

[1] University of Limerick

Keywords: Co-morbidity index; binary data; hierarchical log-linear model.

Introduction

A co-morbidity is a coexisting (or additional) medical condition co-occurring with a primary disease of interest. In phase four studies, for example, when patients are on medication, the scientific interest is often in outcome - recurrence or death. Then, the burden of co-morbidity may be an important contributory determinant of outcome - one which is often overlooked in headline reporting attributing adverse events erroneously to the original treatment.
A number of solutions have been proposed in the medical literature. For example Charlson (1987) developed a Co-morbidity Index (a CCI) based on all patients admitted to the New York Hospital-Cornell Medical Center during a 1-month period in 1984. It comprises a linear combination of the co-morbidities with (age-adjusted) weights derived from a multivariate proportional hazards model of mortality. More recently Davis (1996) working with patients on dialysis derived another score based on clinical insight into the role of co-morbidity.
The construction of such indices (or so-called *risk-scores*) by divers methods is common in the medical literature and a fundamental concern is the optimality of such techniques. Below, we criticise classical methods of CCI construction and propose alternative methods of analysing multivariate binary co-morbidities, especially when p is large.

Model formulation

Given p binary co-morbidities we consider a p-dimensional contingency table with exactly $n = 2^p$ cells. Let n_j be the observed frequency (the count) in the j-th. cell, $j = 1, \ldots, n$, where the cells are ordered lexicographically in Fortran major order and we have the bijective mapping $j \mapsto (i_1, \ldots, i_p)$ with each $i_1, \ldots, i_p$ taking the value 0 (absent) or 1 (present), MacKenzie

& O'Flaherty (1982). Then our basic model is the usual log-linear model for contingency tables in which:

$$E(N_j) = \mu_j = \exp(\mathrm{a}_\mathrm{j}'\theta) \tag{1}$$

where N_j is the random variable denoting the number in the j-th. cell, a_j' is the j-th. row of the $(n \times n)$ saturated design matrix, A, and θ is the $(n \times 1)$ vector of unknown parameters measuring the influence of the constant, main effects and interactions on the response.

Paradigms & Problems

Our adoption of this framework is predicated on the need to address some open problems in different, but related, modelling areas. For example, much original log-linear modelling was formulated in a model development environment dating back to the 1970's where $p = 10$ was considered very large. Then, today's data-mining paradigm was not envisaged and the original ideas have become ossified in legacy code in the major software packages. Accordingly, one objective of the current research is to relax these constraints by developing a new package in R. Yet another challenge is the ability to address the analysis of sparse, high-dimensional, contingency tables which might arise, for example, in thresholded micro-array data. The ability to search within these high dimensional spaces efficiently and so identify the model best supported by the data is a key objective of this research. Such searches may be facilitated by sacrificing high order interaction terms, replacing them by random effects terms instead, thereby extending the model class from a GLM to a GLMM.

References

Birch, M.W. (1963). Maximum Likelihood in three-way contingency tables. *JRSS B.* **25**, 220-233.

Charlson, M.E., Pompei, P., Ales, K.L., MacKenzie, C.R. (1987). , A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J. Chron. Dis.* **40**, 5, 373-383.

Goodman, L.A.(1971). The Analysis of Multidimensional Contingency Tables: Stepwise Procedures and Direct Estimation Methods for Building Models for Multiple Classifications. *Technometrics.* **13**, 1, 33-61.

Haberman, S.J. (1972). Algorithms AS 51: Log-linear Fit for Contingency Tables. *Applied Stats.* **21**, 2, 218-225.

O'Flaherty, M. and MacKenzie, G. (1982). Direct Simulation of Nested Fortran DO-LOOPS. *Statistical Algorithms. Applied Statistics.* **31**, 1.

Estimation of the parameters of the truncated negative binomial distribution with application to counts of neurites emanating from brain cells treated with growth factors

P. Deacon[1] and K. Choudhury[1]

[1] University College Cork

Introduction

The Negative Binomial Distribution is encountered in many practical problems involving count data such as actuarial, ecological and biological data. In this study, the data was gathered from counts of neurites emanating from brain cells which had been treated with growth factors. Neurites are branches extending from the cell body which are too small to be determined as an axon or a dendrite. The types of brain cells included glial cells, nerve cells which do not carry nerve impulses, and astrocytes, which are a subgroup of glial cells. They act as support cells to neurons (Diagram A) and their functions include digestion of parts of dead neurons, manufacturing myelin, and providing physical and nutritional support for neurons. The promotion of growth in such cells is important in the treatment of diseases such as Parkinson's Disease and Multiple Sclerosis.
The data consists of counts of Primary, Secondary, Tertiary and Quaternary branches of cells in a control group, and three treatment groups.

Methodology

On examination of the distribution of the data for Primary branches, taking into account the censoring of zero cases and disproportionality of the variance to the mean, the zero-truncated Negative Binomial was considered to be the more suitable distribution, rather than the Poisson, to which to fit the data. In this study, the following form of the Complete Negative Binomial with parameters r and p was taken:
The probability of x failures before reaching the r-th success is

$$P(X = x) = \begin{pmatrix} x + r - 1 \\ r - 1 \end{pmatrix} p^r (1 - p)^x$$

for $x = 0, 1, 2, \ldots$ and $0 \leq p \leq 1$.

Simulation

Estimates for r and p by Maximum Likelihood prove difficult as the equation for r is not in closed form and previous literature has presented methods for dealing with this problem [Simon].

Simulation was used to compare Maximum Likelihood to Method of Moments estimates using the same random samples from both the Complete and Truncated Negative Binomial distribution. Sample sizes ranged from 100 to 1,090 and values of r and p were taken to produce random data in keeping with the real data. The Mean Squared Error for $\hat{p}$, denoted $mse(\hat{p})$, was found and the natural logarithm of $mse(\hat{p})$ was plotted against the natural logarithm of its corresponding sample size n. Using the equation $mse(\hat{p}) = \frac{\sigma^2}{n}$ and taking logarithms on each side: $\ln[mse(\hat{r})] = 2ln(\sigma) - \ln(n)$

Thus, $\ln[mse(\hat{r})]$ was regressed on $\ln(n)$ and the intercept of the regression line was taken as an estimate for $2\ln(\sigma_{\hat{r}})$. Likewise an estimate for $2\ln(\sigma_{\hat{p}})$ was found.

In the application of these methods to the count of Primary neurites, it was found that the use of Method of Moments estimates as starting values for maximum likelihood estimation was not realistic, as the equations produced negative values for r estimates, and values of p greater than one. Random samples taken for the above table were re-examined and it was found that in some of the random samples where $r = 1, p = 0.5$, similar results had been found. The deduction that there are constraints on the method of moment equations is under investigation. Surface plots of the log likelihood functions of Primary neurites are presently to be used instead of moments to estimate starting values for maximum likelihood estimation.

		$\sigma^2_{\hat{r}}$ by MLE	$\sigma^2_{\hat{p}}$ by MLE	$\sigma^2_{\hat{r}}$ by MME	$\sigma^2_{\hat{p}}$ by MME
$r = 1, p = 0.5$	Complete	0.0369	0.002	2.49	0.593
	Truncated	0.08	0.126	15.3	5.79
$r = 2, p = 0.4$	Complete	0.0526	0.00057	1.05	0.454
	Truncated	5.393	0.016	137.6	0.675
$r = 3, p = 0.7$	Complete	1.347	0.0079	8.27	0.00812
	Truncated	18.35	0.046	88.1	0.0113
$r = 4, p = 0.8$	Complete	34.37	0.0739	219.76	0.3496
	Truncated	52.34	0.149	68.03	0.925

TABLE 1. Comparison of asymptotic variances of estimates

References

Rider P. R. (1955). Truncated binomial and negative binomial distributions. *Journal of the American Statistical Association.* 50, No. 271, 877-883.

Ross G. J. S. and Preece D. A. (1985). The negative binomial distribution. *The Statistician.* 34, 323-336.

Simon L. J. (1985). Fitting Negative Binomial Distributions by the method of maximum likelihood. *PCAS XLVIII,* 45-53.

Non-traditional statistical process control for commercial irradiation

J. Donovan[1] and E. Murphy[2]

[1] Institute of Technology Sligo

[2] University of Limerick.

Abstract

Commercial irradiation typically involves applying electron beam or gamma radiation to a product with a view to sterilizing the product and killing any bacteria present. This is extremely important in the medical device industry where packaged products or pallets require irradiation between a minimum and maximum dose. As packaged product provides a level of shielding it is difficult to monitor such processes to ensuring that all products on the pallet received a dose greater than the minimum and yet less than the maximum. This difficulty was solved for both the gamma ray and electron beam radiation facilities by the use of non-traditional Statistical Process Control (SPC) techniques. By combining standardized charts and group charts, a single control chart could accommodate both minimum and maximum doses while effectively monitoring the entire process. In addition, guidance tables were developed the allowed operations to determine a window of suitable dosages that would simultaneously satisfy the irradiation requirements of different groups of products.
Keywords: Statistical Process Control, irradiation, standardized charts.

Introduction

The initial step in planning an irradiation process for a new product involves performing a dose mapping exercise. The aim of this is to characterise the process variability for the product and to determine and identify the locations where the minimum and maximum doses, D_{min} and D_{max} exist. This exercise results in the determination of a Dose Uniformity Ratio (DUR) that relates dosage received at any location to the applied radiation dose. In fact it is possible that the location where the minimum dose is found cannot be used for routine dosimetry measurements, as it will involve unacceptable unpacking or de-palletising of shippable product. A measure of process variability referred to as p is determined from the dose

mapping exercise that that includes calibration uncertainty, dose mapping uncertainty and dosimeter reproducibility.

Methods

The level of quality (p) is specified for the process and a target level of radiation dose called D_{mean}. This value of D_{mean} incorporates the influence of the dose uniformity ratio and allows an operation widow of acceptable values of D_{mean}.

Typically p will be 0.001 indicating that 99.9% of all doses will be within the specified sterilization dosage range. The use of an operating window allows the planning and incorporation of numerous products inside the sterilization chamber without having to devote the test to an individual product type. Having selected appropriate values of D_{mean} for each product or group of products it is now possible to monitor the process using dosimeters. If the process is under statistical control, the measured doses at the minimum dose location will be centred on D_{mean} and will exhibit a spread consistent with the variability expected from the relevant components of uncertainty σ_p. For each product, or group of products, the dose at the monitoring location(s) D_{mon} (corresponding to D_{mean} as calculated from the dose mapping) is determined. The plot point on the control chart D_{plot} is then calculated using: $D_{plot} = (D_{meas} - D_{mon})/\sigma_p$ where D_{meas} is the measured dosage at the monitoring location(s).

Results

Guidance tables were developed that assisted in the determination of the D_{mean} operating process window for various values of σ_p. Standardised group charts allowed the monitoring of all products and multiple simultaneous dosimeter measurements on the same SPC chart. The standardised control limits were set at $\pm 3.5\sigma_p$ to avoid unnecessary false alarms in the process. Both the minimum and maximum measured doses were plotted on the same chart.

Discussion

This control charting method has proven very effective in the monitoring and control of commercial radiation facilities and has been accepted by the European panel on gamma and electron irradiation as the method of monitoring and controlling irradiation processes in these facilities. The use of the operating window for D_{mean} has meant that numerous products can

be irradiated simultaneously with a common value of D_{mean} that meets the minimum and maximum specified doses of all the products.

Acknowledgements

The authors gratefully acknowledge the support of Dr Peter Sharpe of the National Physical Laboratory, UK and the European Panel on Gamma and Electron Irradiation who funded this research.

Time series forecasting using neuro-fuzzy methods

K. Flanagan[1] and S. Mulligan[1]

[1] Dublin Institute of Technology

Abstract

This paper is concerned with time series forecasting using the adaptive neuro-fuzzy inference system (ANFIS). In this study, historical daily closing values for up to one year of the share prices for five companies quoted on the Irish stock exchange are used, with daily forecasts for up to eighteen days ahead being produced using each of the forecasting methods. The forecasts using ANFIS are compared with three statistical techniques: namely exponential smoothing, ARIMA (autoregressive integrated moving average) models and a random walk model. It is found that the ANFIS produces the best short term forecasts while the ARIMA methodology is the best method for longer term horizons.

Introduction

A brief description of the ANFIS is provided in the methods section. An outline of the results is also presented. The technique follows that of Jang (1993).

Methods

Jang (1993) found that three historical values could be used for the creation of a forecast. These are normally in the pattern shown in the table where the capital letters indicate the actual historical values of the data provided with, for example, $X(1)$ being the oldest value while the lowercase letters indicate the three values used to create the forecast $\hat{y}_i$. Each of these inputs x_i is provided with a number of membership functions to determine the degree to which an input possesses a particular property. In this study, two membership functions will be used for all inputs.

x_1	x_2	x_3	$\hat{y}_i$
X(1)	X(3)	X(5)	X(7)
X(2)	X(4)	X(6)	X(8)

A choice must be made as to the type of membership function. The generalised bell membership function will be used in this study as it was also used by Jang (1993).

$$O_{1ij} = A_{ij}(x_i) = \left(1 + \left|\frac{x_i - c}{a}\right|^{2b}\right)^{-1}$$

where x_i is the input value, $A_{ij}(x_i)$ is the membership value output for x_i and a, b and c are parameters yet to be determined.
These memberships now need to be aggregated into an overall figure for the various combinations of input. The fuzzy AND operator is used for this aggregation which in this case is taken to be multiplication. The output rules are then normalised by division so that their total is one. A weighed average of the input values x_i is taken so that it reflects the combination required to produce the forecast.

$$f_m = p_{m1}x_1 + p_{m2}x_2 + p_{m3}x_3 + p_{m4} \text{ where } m = 1, 2, \ldots, 8$$

where p_{m1}, p_{m2}, p_{m3} and p_{m4} are parameters yet to be determined. The normalised values are then combined with the weighted average. The forecast $\hat{y}_i$ is obtained by adding the normalised combinations.
The estimation phase can begin with initial values being provided for each of the fifty parameters. Historical data can be used to train the model so that over a given number of training runs, a neural network will determine the optimum values for each of the fifty parameters.

Results

A comparison of the average margin of error for each of the forecasts is calculated and the symmetric mean absolute percentage error is used.

$$\text{s}MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{(y_i + \hat{y}_i)/2}$$

where y_i is the actual value for period i, $\hat{y}_i$ is the forecast for that period and n is the number of forecasts. For one period ahead forecasts, the ANFIS had an average error of 0.98% while the ARIMA had an error of 1.1%. ANFIS had the lowest margin of error for the averages up to eight periods ahead, after which the ARIMA models had the smallest margin of error.

References

Jang J.-S. R (1993). ANFIS: Adaptive-Network-based Fuzzy Inference Systems *Transactions on Systems, Man, and Cybernetics* 23(3), 665-685.

Development of a Model for the Eruption of First Permanent Molars to guide Fissure Sealing Programmes

E. Flannery[1], 2, F. O'Sullivan[1] and H. Whelton[2]

[1] University College Cork

[2] Univeristy Dental School and Hospital, Cork

Introduction

Oral health services are provided free of charge to primary school children in the Republic of Ireland. The Health Services Executive Dental Service is responsible for the provision of these oral health services in the Republic of Ireland and target children for dental services according to school class. The routine examination and treatment services they provide include the placement of fissure sealants. A fissure sealant is a plastic resin which, if applied at the right time i.e. between the time of eruption of the teeth and the development of decay to the pits and fissures of posterior (back) permanent teeth, protects these vulnerable surfaces from decay. Although there are variations across administrative areas, most children are screened in second class in primary school (aged approximately 8 years) when any first permanent molars that are erupted and free from dental caries (decay) are sealed with a fissure sealant.

Analysis of data from the North South Survey of Children's Oral Health 2002, has shown that the eruption of the first permanent molars occurs over a wide range from approximately 4.5 years to approximately 8 years. There is also a lot of variation in the actual ages of children in second class. The targeting of children for fissure sealing by school class is inefficient because of individual variations in eruption patterns and caries risk. When examined in Second Class for the North South Survey of Children's Oral Health, many children had unerupted first permanent molars, whilst these teeth were already decayed for many more. Thus a more individualised approach is required to identify children and teeth at the optimum time for placement of sealants.

Methods

Preliminary analysis has confirmed that research is needed to develop a model for the eruption of first permanent molars in the population and to identify the best time to target children for sealing of their teeth seal in order to reduce the risk of decay to these children.
We use a Probit function to model the proportion of a first permanent molar that has erupted for an individual.
i.e. $P(t)$ = Proportion of tooth erupted at time t for an individual

$$P(t) = \Phi\left(\frac{t-\mu}{\sigma_e}\right).$$

For a sample of N individuals the eruption proportion at time t will be $\Phi\left(\frac{t-\mu_i}{\sigma_e}\right)$ where $\mu_i \sim N(\mu_p, \sigma_p)$. We can use this model to predict the optimal time to seal a particular tooth for an individual.
The model for the individual can then be expanded to model eruption of the tooth for the population. If $\mu \sim N(\mu_p, \sigma_p)$, then the expected proportion of erupted teeth in the population at time t is

$$E\left[\Phi\left(\frac{t-\mu}{\sigma_e}\right)\right] = \overline{P}(t) = \int_{-\infty}^{\infty} \Phi\left(\frac{t-s}{\sigma_e}\right)\phi\left(\frac{s-\mu_p}{\sigma_p}\right) ds.$$

Using Fourier theory $\overline{P}(t)$ can be simplified to the following:

$$\overline{P}(t) = \Phi\left[\frac{t-\mu_p}{\sqrt{\sigma_e^2+\sigma_p^2}}\right].$$

The observed number of erupted teeth for individuals aged t in a sample of size N will be Binomial $B(N, \overline{P}(t))$. Thus a probit analysis will allow the values of μ_p and $\sqrt{\sigma_e^2+\sigma_p^2}$ to be estimated.

Results

Using data from the North South Survey of Children's Oral Health we were able to estimate the values of μ_p and $\sqrt{\sigma_e^2+\sigma_p^2}$.
Simulated analyses using the model for the population, $\overline{P}(t))$, with the addition of an error term, provided results consistent with literature from a longitudinal study of the eruption of first permanent molars [1].

Discussion

We have developed a model for the eruption of first permanent molars in children that is consistent with published results on tooth eruption from

longitudinal studies. The model for the eruption of first permanent molars for the population can now be used to predict the optimal age to target children fissure sealing.

References

Ekstrand K. R., Christiansen J., Christiansen M.E. (2003). Time and duration of eruption of first and second permanent molars: a longitudinal investigation. *Community Dent Oral Epidemiol.* 31(5), 344-350.

Non-Homogeneous Markov Models for Healthcare Systems Modelling

S. McClean[1], L. Garg[1], B. Meenan[1] and P. Millard[2]

[1] University of Ulster

[2] University of Westminster, London

Abstract

In previous work, Markov chain models have been used for healthcare systems, where the states in hospital are described as phases, such as acute, rehabilitation, or long-stay and likewise social care in the community may be described using phases such as dependent, convalescent, or nursing home. This allows us to adopt a unified approach to health and community care, rather than focusing on the improvement of part of the system to the possible detriment of other components. Here, this approach has been extended to show how the non-homogeneous Markov framework can be used to obtain various metrics of interest and extract patient pathways.

Keywords: Healthcare System modelling, Markov Models, Healthcare Performance Monitoring, Interesting Patient Pathways

Introduction

Escalating proportions of elderly people is creating a problem with regard to maintaining the quality of care well in to older age. A systems approach to healthcare planning is therefore needed, where we model hospital and community care patients using phase-type distributions (Faddy and McClean, 1999). In addition, covariates may be incorporated into the models, thus further increasing their ability to describe complex healthcare processes. In order to model the whole patient-care system, we then describe stay in hospital and stay in the community as two separate phase type-distributions with transient states being phases of care in hospital and the community respectively and death being an absorbing state. A non-homogeneous Markov representation is used to incorporate time-dependent covariates thus improving realism of the model. The approach is illustrated using data on geriatric patients from an administrative database of a London hospital.

The Non-homogenous Markov model

Faddy and McClean (2005) described the Coxian Phase-type model with four hospital states and three community states. Time dependent covariates, such as age, are described via a time-heterogeneous Markov model where the parameters are updated every time a patient makes a transition. This can be achieved by having the transition rates depend log-linearly on covariates. Such dependency parameters have previously been estimated by maximum likelihood (Faddy and McClean, 1999). The update time can then be represented by mean time to make the corresponding transition from hospital to community or vice versa. In order to implement time-dependence in the covariates we update the values of parameters every time there is a discharge or admission to hospital with the new updated age and updated year as covariates. Also, for each admission (or re-admission) and each discharge, we calculate the expected total time spent in hospital and in the community respectively. This expected total time is then used to update the age and the year after each admission (or re-admission) to hospital and each discharge to the community. Using this approach, various patient pathways can also be extracted.

Results

We now obtain results for the time-heterogeneous Markov model using parameters estimated from Faddy and McClean (2005). We here compare our previous results presented in McClean et al. (2006) with the results from our new approach, which incorporates time-dependent covariates, namely age and year. We find that incorporating time dependent covariates slightly increases the mean and variance of the number of admissions to hospital and number of discharges. There is a substantial decrease in the number of days spent in hospital and the community respectively, corresponding to higher death (absorbing) probabilities in all cases. This is to be expected as older patients are more likely to die during a spell in hospital or back in the community than younger patients. Our new non-homogeneous Markov model is therefore likely to be more realistic than our previous approach.

References

M.Faddy, S.McClean (1999) Analysing data on lengths of stay of hospital patients using phase-type distributions *Applied Stochastic Models in Business and Industry*, 15:311-317

M.Faddy, S.McClean (2005) Markov Chain Modelling for Geriatric Patient Care *Methods of Information in Medicine* 369-373

S.McClean, M.Faddy, P.Millard (2006) Using markov models to assess the performance of a health and community care system *CMBS* 777-782

A mixture distributions approach to in vivo correlation modelling of a dual component drug delivery system

C. Gaynor[1], S. Rossenu[2], A. Vermeulen[2], A. Dunne [1] and A. Cleton[2]

[1] University College Dublin

[2] Johnson and Johnson Pharmaceutical Development, Beerse, Belgium

Introduction

A predictive mathematical model describing the relationship between the measurements made in a laboratory, e.g. the rate at which a tablet dissolves, and observations made on humans, such as blood concentrations of a drug, is known as an in vitro - in vivo correlation (IVIVC). Substantial effort and resources go into the development of these models which can be used to accurately predict an in vivo response from in vitro observations. The procedure routinely employed when attempting to establish an IVIVC for an extended release (ER) drug product involves the study of a number of formulations of the dosage form of interest, each with a different release rate. A sample of the ER dosage units of interest from each formulation are dissolved in vitro and the fraction which has been dissolved is recorded at a series of time points for each dosage unit. Further ER dosage units from each formulation are administered to a number of human subjects, ideally as part of a crossover study, and their plasma drug concentration-time profiles are determined. A unit reference dose, which may take the form of a solution, IV injection or an immediate release (IR) tablet, is administered to each of the subjects and their plasma drug concentrations are measured over time. The resulting data are used to develop a model which can be used to predict the in vivo plasma drug concentration profile of the ER formulation using in vitro dissolution data only.
ER dosage units (usually tablets or capsules) are formulated to release a drug slowly over time. They are, therefore, administered less frequently than traditional dosage forms and result in prolonged and more consistent drug concentrations in the blood. For chronic treatments, such a formulation is desirable as it is a more convenient dosage regimen and may result in better patient compliance. Following the administration of many such ER formulations, however, it can take the drug a significant period of time

to reach the required therapeutic level. As a consequence, an initial loading dose (e.g. IR formulation) may be administered prior to the ER maintenance dose to ensure that patients remain in the effective exposure range during the complete dosing interval. Formulations, which combine these two components i.e. a loading and an extended release dose, are becoming increasingly popular and, therefore, require the development of extensions to existing IVIVC modelling techniques. A non-linear mixed effects modelling approach to establishing IVIVCs, reported by O'Hara et al. [2], was modified to describe this dual component drug delivery system and further adapted to allow the simultaneous analysis of more than one formulation.
Galantamine is a reversible cholinesterase inhibitor [3] and has been approved for marketing in immediate and extended release tablet forms. It has shown consistent efficacy on cognition, global functioning and activities of daily living in patients with mild to moderate Alzheimer's disease [4, 5, 6]. A once-daily extended release capsule, which incorporates an immediate release element to act as a loading dose, was developed for chronic treatment. Data collected during a study conducted with four different formulations was used to illustrate the use of this extended IVIVC model.

Methods

The fraction of each capsule dissolved at any given time was modelled as a mixture of two distributions. The proportion initially released as a loading dose was modelled according to a standard two compartment pharmacokinetic model commonly used to describe immediate release data, while the remaining extended release component, and the relationship between in vitro and in vivo dissolution were modelled using a modified version of the method described by O'Hara et al. [2]. The established IVIVC model was used to predict plasma drug concentration profiles for each subject separately.
In order to establish an IVIVC model's validity, its accuracy of prediction must be evaluated. The United States Food and Drug Administration (FDA) recommend that formulations with three or more release rates be used to develop the model. Assessment of the model's ability to describe these data is referred to as internal validation. The model must be able to predict the area under the plasma concentration/time curve (AUC) and the peak plasma concentration (Cmax) to within acceptable limits as set out by the FDA [1]. Data collected using dosage units from a fourth or subsequent formulation which was not used in the model development stage may also be used to similarly assess external predictability. The AUC and Cmax were calculated for both the observed and predicted profiles and were used to calculate the percentage prediction error.

Results

Observed and predicted plasma drug concentration profiles for each formulation averaged across subjects are illustrated in Figure 1. The corresponding percentage prediction errors are given in Table 1.

Conclusion and Discussion

A mixture distribution based model was developed in order to describe a dual component drug delivery system. This model successfully captured both the loading dose and ER elements of a capsule. The technique was tested on four formulations of galantamine and an IVIVC model, which meets the FDA criteria for internal and external predictability, was established.

FIGURE 1. Observed data and predicted profiles for each formulation averaged across subjects

FORMULATION	% PE Cmax	% PE AUC
Medium	-3.51254	-4.51084
Slow	-5.60758	-2.78782
External	-2.99568	-6.85654
Fast	-6.32688	-8.05706

TABLE 1. percentage prediction error for each formulation

References

(1) FDA's Guidance for Industry (1997). *Extended release oral dosage forms: development, evaluation, and application of in vitro/in vivo correlations.*

(2) O'Hara T., Hayes S., Davis J., Devane J., Smart T. and Dunne A. (2001). In vivo-in vitro (IVIVC) modeling incorporating a convolution step. *J. Pharmacokinetics and pharmacodynamics*, 28, 277-298.

(3) Schrattenholz A., Pereira E.F.R., Roth U., Weber K.H., Albuquerque E.X., Maelicke A. (1996). Agonist responses of neuronal nicotinic acetylcholine receptors are potentiated by a novel class of allosterically acting ligands. *Mol. Pharmacol.*, 49, 1-6.

(4) Tariot P.N., Solomon P.R., Morris J.C. et al. (2000). A 5-month, randomized, placebo-controlled trial of galantamine in AD. *Neurology*, 54,2269-2276.

(5) Raskind M.A., Peskind E.R., Wessel T. et al. (2000). Galantamine in AD: A 6-month randomized, placebo-controlled trial with a 6-month extension. *Neurology*, 54,2261-2268.

(6) Wilcock G.K., Lilienfeld S., Gaens E. (2000). Efficacy and safety of galantamine in patients with mild to moderate Alzheimer's disease: multicentre randomized controlled trial. *Br. Med. J.*, 321, 1-7.

Sensitivity Analysis of an Early Detection Technique for Field Failures

B. Honari[1], J. Donovan[1], T. Joyce[2] and E. Lisay, Jr.[3]

[1] Institute of Technology Sligo

[2] Alcatel-Lucent Technologies, Dublin

[3] Lucent Technologies, Westford Ma, USA

Abstract

A new technique for detecting changes in field failure data was developed and presented by B. Honari et al. in RAMS 2007 [1]. This detection technique is based on Statistical Process Control charts. It is imperative that a method be available for the reliability engineers to effectively monitor and detect reliability changes and to understand if the reliability of the product has changed significantly from what was expected. In this paper we focus on sensitivity analysis for the proposed detection technique.

Introduction

The monitoring of changes in the pattern of failure field data is an important requirement of any reliability management program. Systematic monitoring of field failures also ensures that adverse changes in the production process are identified quickly in order to avoid facing serious reliability and warranty problems. These kinds of problems are mainly caused by unanticipated failure modes, unknown changes in raw material, changes in operating environmental conditions, etc. It is imperative that a method be available for the reliability engineers to effectively monitor and detect reliability changes and to understand if the reliability of the product has changed significantly from what was expected. The early detection of reliability problems through analysis of field data will save the manufacturer large amounts of money in warranty costs, improve product quality while also retaining customer goodwill. While the manufacturing company warranty policy is based on the type II error in this hypothesis test, most of the academic published works deal with the type I error. This paper develops some studies about type II error in the formally formulated hypothesis test for reliability deterioration detection as above.

Problem Formulations

The problem of detection of reliability deterioration can be formulated as a multiple-parameter hypothesis test like:

$$\begin{aligned} H_0 : &\quad \lambda_1 \leq \lambda_1^0, \lambda_2 \leq \lambda_2^0, \ldots, \lambda_M \leq \lambda_M^0 \\ H_a : &\quad \lambda_1 > \lambda_1^0, \lambda_2 > \lambda_2^0, \ldots, \lambda_M > \lambda_M^0, \end{aligned}$$

in which λ_i represents the return rate for units of i-th shipment period, λ_i^0 represents the reference value for λ_i which can be found from the historical records, and M is the pre-specified number of future periods to be monitored. While the manufacturing company warranty policy is based on the type II error in this hypothesis test, most of the academic published works deal with the type I error. This paper develops some studies about type II error in the formally formulated hypothesis test for reliability deterioration detection as above.

Methods

In this paper it is proposed that the units shipped in each calendar month are analyzed separately. Control limits are established for the product based on the known statistical distribution of its field failures. For example, suppose the product conforms to the lognormal distribution with a location of μ and a scale of σ. Based on the cumulative distribution function (CDF) of this lognormal distribution we can determine the expected proportion of units that will be returned each successive month after shipment. This expected proportion is known as $\overline{p}$ and will vary from month to month. As $\overline{p}$ changes each month, new limits are derived for the monthly returns consistent with Equation 1

$$\overline{p} \pm 3\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}. \tag{1}$$

In this equation n represents the number of units successfully operating. Initially n represents the number of units shipped in a particular month but is reduced each month to take account the number of units that remain functioning in the field. Symmetrical limits are set at the mean ± 3 standard deviations as this approximates to a normal distribution. The failure data is assumed to follow a lognormal distribution with $\mu = 4, \sigma = 1.5$. To be similar to values that have been observed in the real life data. For the sake of simplicity it is assumed that the number of units shipped in the particular month of interest is 10,000. The CDF values for the first 60 months have been determined and a portion of these months is shown in Table 1. The limits are devised for each month based on the expected proportion of failures for that month. The limits for number of failures are also devised based on the limits for proportions and the number of working units.

Month	Proportion	No. of units at the beginning of the month	Lower limit on predicted proportion	Upper limit on predicted proportion
0	0.00383	10000	0.0019769	0.0056831
1	0.00991	9962	0.0069344	0.0128896
2	0.01280	9863	0.0094035	0.0161945
3	0.01417	9735	0.0105781	0.0177659
4	0.01479	9593	0.0110909	0.0184851
5	0.01499	9445	0.0112373	0.0187387

TABLE 1. Excel sheet of simulated data

The next step is to simulate the failure times of the 10,000 shipped units. As the detection limits are set at $n \pm 3 \times standarddeviations$, it is possible that a false alarm could occur and 0.135% of the months could naturally fall outside each limit. The results are shown in Figure 1.

Results

Now by changing the distribution parameters, we simulate the type II error as defined as follows for the proposed detection technique.
Type II Error = P(Not detecting the changes in the pattern of field retune Data | There is a deterioration in field reliability) Simulation results for estimating the value of type II error for the various changes are described.

Discussion

In this paper we studied and analyzed the type II error for the hypothesis test of detection problem as sensitivity of the proposed early detection

technique for field failures. The results can be used to develop a warranty policy considering the possible misleading observations.

References

(1) Honari B. and Donovan J. (2007). Early Detection of Reliability Changes For A Non-Poisson Life Model Using Field Failure Data. *RAMS 2007 Proceedings*

(2) Nelson L. S. (1994). A control chart for parts-per-million nonconforming items. *Journal of Quality Technology*, 39(3), 239-240.

(3) Wu H. and Meeker W. Q.] (2002). Early detection of reliability problems using information from Warranty Databases. *Technometrics*,

44(2), 120-133.

(4) Montgomery D. C. (2005). *Introduction to Statistical Quality Control*, John Wiley & Sons, New York.

Acknowledgements

This work was supported by Science Foundation Ireland under SFI grant 03/CE3/1405.

Statistical significance in the Analysis of Gene Expression Data

S. O'Neill[1], J. Huang[1], K. O'Sullivan[1] and C. von Gertten [2]

[1] University College Cork

[2] Karolinska Institute, Sweden

Keywords: Gene expression data, error rates, False Discovery Rate (FDR)

1 Introduction

The analysis of gene expression data must consider many thousands of statistical tests simultaneously. Each test has a certain probability of reaching an incorrect inference, necessitating a control on the error rates. Using standard p-values to measure significance does not account for multiplicity of testing, but applying an adjusted correction produces results that are excessively conservative. Recently, the false discovery rate (FDR) was proposed to measure statistical significance in gene expression study. We compared performance of p-values, adjusted p-values and FDR in identifying differentially expressed genes for real microarray data.

2 Methods

FDR is defined as the expected proportion of false positives among genes declared to be differentially expressed. The estimated FDR is calculated using an R-package OCplus. The data was collected from a two factor factorial design. The factors investigated were type of injury (Spreading Disease brain injury (SD) or Traumatic Brain Injury (TBI)) and type of treatment (treatment or no treatment). The goal of the study was to identify genes which were differentially expressed. The focus was on determining genes which are differentially expressed between the two treatment levels for each type of injury and also between the types of injury for each of the two treatment levels. Sixteen rats were randomly assigned to one of the four groups. During the course of the experiment one of the rats died. Information on 1400 genes was collected from each of the remaining 15 rats.

3 Results

Comparison of types of injury for each of the two treatment levels and also the two treatment levels for each type of injury were conducted separately. The numbers of genes identified to be differentially expressed by using P-value, the Bonferroni correction and FDR were present in Table 1.
¿From Table 1 we see that using the standard p-values resulted in the largest number of differentially expressed genes. The Bonferroni correction did not identify any differentially expressed genes,

TABLE 1. The number of genes identified to be differentially expressed by using P-value < 0.05, the Bonferroni correction (P-value $< 0.05/18838$) and FDR < 0.05. Comparison of types of injury for each of the two treatment levels and also the two treatment levels for each type of injury was conducted separately.

	p-value	Bonferroni	FDR
TBI non-treated vs SD non-treated	2919	0	54
TBI treated vs SD treated	2926	0	2
TBI non-treated vs TBI treated	2308	0	89
SD non-treated vs SD treated	2308	0	89

4 Discussion

Analysis of gene expression data must consider tens of thousands of statistical tests simultaneously. FDR provided more realistic controls over standard approaches of controlling for multiplicity of testing.

Image-Based Recovery of an Input Function for Kinetic Analysis of a Cerebral Glucose Utilization based on FDG-PET Scanning.

J. Kirrane[1], F. O'Sullivan[1], 2, M. Muzi[2] and A. Spence[2]

[1] University College Cork

[2] University of Washington, Seattle

Abstract

Quantitation of dynamic PET studies via kinetic modelling analysis usually requires the time-course of the radio-tracer in the arterial blood as the input function. While the input can be obtained by direct arterial sampling, there are situations where it is inconvenient or impractical to obtain this information in practice. Hence image-based extraction of an arterial blood input can be of interest. In the case that the blood pools in the field of view are limited, such as cerebral PET studies with FDG, this is a difficult task. Here we develop and investigate a statistical approach based on a flexible parametric representation of the input function and a localized 2-component mixture model representation of blood signal attenuation (recovery) and contamination (spillover) processes. Segmentation is used to extract approximate blood pools and also to identify the relevant sources of spillover contamination. This technique is applied to data from a series of 12 human cerebral PET studies with ^{18}F-Fluorodeoxyglucose (FDG). With this approach the regional accuracy of kinetic quantitation obtained from the image-derived input function is assessed in comparison to corresponding results obtained from radially sampled arterial input. The results are quite promising.

Supported in part by HRB (Ireland) and NIH (US).

References

Phelps et al. (1979). Tomographic measurement of local cerebral glucose metabolic rate in humans with [F-18]2-Fluoro-2-deoxy-D-glucose: validation of method *J. Nucl. Med.* 38, 1161-1168.

Chen et al. (1998). Noninvasive quantification of the cerebral metabolic rate for glucose using positron emission tomography, FDG, the Patlak Method and an image-derived input function *", Journal of Cerebral Blood Flow & Metabolism.* 18, 716-723.

De Geus-Oei L. et al. (2006). Comparison of image-derived and arterial input functions for estimating the rate of glucose metabolism in therapy-monitoring 18F-FDG PET studies. *J. Nucl. Med.* 47(6), 945-949.

Modelling N_20 Emissions from Irish Grasslands

F. Leonard[1], N. Quinn[1], K. Richards[2] and D. Fay[2]

[1] Waterford Institute of Technology

[2] Teagasc Environmental Research Centre, Wexford

Abstract

One of the pressing challenges of the 21st century is to efficiently utilise the worlds natural resources while curtailing environmental damage. The moral imperative to act is reinforced by a variety of internationally agreed legislation such as the Kyoto protocol, one of the battle fronts in the agricultural sector centres around pollution caused by nitrogen losses to the environment from grazed grassland systems.

Nitrogen exists in such grassland systems in different forms or pools controlled by a number of natural soil microbial processes which convert the nitrogen from one form to another thereby inducing flows or fluxes between these pools. Nitrogen can be removed from the system as a product (meat or milk) but it can also be lost in drainage water (leaching) which leads to water pollution or to the atmosphere as a green house gas (denitrification and volatilisation). The challenge to farmers and their advisors is to devise sustainable management strategies (i.e. of fertiliser application) which minimise losses and maximise utilisation. However, the interactive effect of such factors as climate, soil type and management policy on the various nitrogen converting processes which ultimately dictate the extent of the losses prohibit the devising of a simple one size fits all grassland management strategy.

One successful approach to the problem is the NCYCLE model has been developed in the UK and adapted for Irish grasslands. It involves modelling the fluxes between the various nitrogen pools in such systems. One of the reasons for the success of this approach is a consequence of it employing sub models to deal with each individual flux in the system. Thus it can readily be tailored to grasslands systems which have different soil types, climates and management systems. (Of course some of these quantities, fertiliser application for example, need not be modelled as they can be

quantified directly). The NCYCLE model makes a mass balance assumption. That is to say, the total (inorganic) nitrogen in the soil is unchanged from the beginning to the end of the grazing season. All fluxes apart from the losses dues to leaching and denitrification are either measured directly or modelled. The mass balance constraint permits a cumulative total for these two to be calculated once all other fluxes are estimated. To determine what proportion of the total may be attributed to denitrification, a matrix of coefficients (representing denitrification proportion) is used which incorporates information on both the soil texture and the drainage status. These coefficients are constant values and are independent of time and of instances of fertiliser application.

A recent study undertaken at Johnstown castle measured emissions (denitrification) from a range of Irish grassland soil types. The investigation involved simulating animal urine deposits as well as fertiliser application on the three different soil types. Data collected showed large N_2O emission spikes following fertiliser and urine application, the magnitude of which depended on the time of application, soil type and antecedent climatic conditions. These results suggest that a more sophisticated denitrification sub model of NCYCLE might be developed. Ideally such a model would be formulated in terms of independent parameters which would have meaningful interpretation in terms of soil type, climate etc. which would enhance its applicability.

Breast Cancer Survival Analysis and Local Health Authority League Tables

J. Lynch[1] and G. MacKenzie[1]

[1] University of Limerick

Keywords: Breast Cancer, Survival Analysis, League Tables, Covariate Adjustment.

Introduction

Using a relative survival approach, Coleman (1999) reported that North Staffordshire Local Health Authority (LHA) was ranked last of 99 LHAs in England & Wales with respect to breast cancer survival. His method made no allowance for case-mix. We re-analyse an augmented dataset from the West Midlands of England, including North Staffordshire, by more traditional methods and report on the resulting *case-mix* adjusted league table.

Data and Methods

The population data analysed comprise 15,516 incident cases of cancer of the female breast diagnosed in the West Midlands, UK, between 1991-1995 and followed-up to the end of 2001. We compare the Kaplan-Meier (KM), non-parametric maximum likelihood, estimator with the proportional hazard (PH) model of Cox (1972) and a Gamma frailty variant (Hougaard, 1994). Because some of the covariates studied do not obey the PH assumption we also adopted the non-PH Generalised Time-Dependent Logistic Model (MacKenzie 1996, 1997) and the Gamma frailty variant discussed by Blagojevic, MacKenzie & Ha (2003).

Results

In the Cox analysis, stage and treatment were the the most important of ten statistically significant covariates employed. Some 100 bootstrap samples generated from the original dataset (re-sampled to respect the proportion censored) confirmed the stability of variable selection. We have computed League tables based on West Midlands and LHA-specific values

TABLE 1. All-cause 5-year Survival League tables

LHA	$\hat{S}_{KM}(t=5)$	LHA	$\hat{S}^{*}_{PH}(t=5\|\bar{x})$	LHA	$\hat{S}^{**}_{PH}(t=5\|\bar{x})$
Solihull	0.71	Unkn	0.783	Unkn	0.779
Worcester	0.68	Birm	0.763	Cov	0.765
Hereford	0.68	Her	0.753	Birm	0.762
Shropshire	0.68	Cov	0.752	Her	0.752
Warwick	0.68	War	0.732	War	0.735
Wolver	0.67	Sand	0.727	Wolv	0.735
Coventry	0.67	Shrop	0.727	Shrop	0.723
SStaff	0.65	Sol	0.727	Sand	0.722
Walsall	0.65	Wolv	0.722	Dud	0.720
Unknown	0.65	Worc	0.716	Sol	0.717
Birm	0.65	Dud	0.706	Wal	0.714
Dudley	0.65	Wal	0.706	NStaff	0.711
Sandwell	0.63	NStaff	0.696	Worc	0.711
NStaff	0.58	SStaff	0.686	SStaff	0.689

* $\hat{S}^{**}_{PH}(t=5|\bar{x})$ is the stratified Cox model employing a separate baseline hazard function for each Health Authority and $\bar{x}$ is the West Midlands mean.

for both baseline hazard functions and covariate means. The interpretation of League tables is fraught with difficulty and should be treated with due caution (Goldstein et al, 1996).

Discussion

While much work remains to be done, the preliminary findings suggest that covariate adjustment is required for valid interpretation. Fitting frailty models in R is a rather slow process and the standard errors of the frailty variance parameter are not produced in standard R output. In the paper we shall compare the findings obtained by fitting non-PH models, discuss the value of frailty in this context and comment on the process of obtaining adjusted survival curves.

References

Coleman et al (1999)Cancer Survival in the Health Authorities of England up to 1998; Patients diagnosed 1991-1993 and followed up to 31 December 1998 - A report prepared for the National Health Services Executive under contract with the National Centre for Health Outcomes Development, London, The Office for national Statistics.

Cox DR (1972) Regression models and life-tables (with discussion). *J. R. Statist. Soc.* B, **34**, 187-220.

Goldstein H; Spiegelhalter D.J. (1996) League Tables and Their Limitations. *J. R. Statist. Soc.* **159**, 385-443.

MacKenzie, G. (1996) Regression models for survival data: the generalised time dependent logistic family. *JRSS Series D*, 45, 21-34.

MacKenzie, G. (1997) On a non-proportional hazards regression model for repeated medical random counts.

Hougaard, P. (1994). Heterogeneity Models of Disease Susceptibility, with Applications to Diabetic Nephropathy. populations. *Biometrics*, 50, 1178-1188. *Statistics in Medicine*, 16, 1831-1843.

Blagojevic M., MacKenzie G. and Ha I.D. (2003) A Comparison of non-PH & PH - Gamma frailty models. *IWSM 2003,pp 39-43.*

An assessment of spatial heterogeneity in the boundary of human sarcoma imaged with FDG-PET

K. McKeown[1], F. O'Sullivan[1] , J. Eary[2], M. Janes[2] and J. O'Sullivan[1]

[1] University College Cork

[2] University of Washington Medical Center, Seattle

Abstract

The use of positron emission tomography (PET) scanning in the analysis of fluoro-deoxyglucose (FDG) utilisation in human sarcoma has lead to advancements in patient care through the prediction of overall patient survival and disease progression. Previous work on the statistical significance of spatial heterogeneity as a predictor of patient survival was examined through the deviation of the FDG utilisation from a unimodal elliptically contoured spatial pattern. Although this elliptical heterogeneity measure was shown to be a strong prognostic indicator of time to death [O'Sullivan et al. (2003)], the later incorporation of boundary information into this definition showed potential for improved risk prediction [O'Sullivan et al. (2005)]. Motivated by these findings, this current research focuses solely on the distribution of the FDG uptake in the boundary of the tumor and investigates its prognostic potential through the development of a new measure of heterogeneity. The computation of the new measure utilises data obtained from a set of 226 sarcoma patients imaged between August 1994 and February 2004 at the University of Washington Medical Center, for which full follow-up information exists for each patient, to evaluate this technique. The work examines the prognostic utility of boundary information in relation to other measures that are currently available.

Introduction

Sarcomas are a cancer of the connective or supportive tissue and are characterised as heterogeneous in nature. This characteristic gives rise to the increased desire to develop and utilise measures of heterogeneity in exploring new approaches to aid the prediction of tumor behaviour. This approach

would not be applicable to various other tumor classes such as brain tumors due to their size, compact form and fast growing nature. Treatment for all patients in this study involves surgical resection. A compact mass can easily be removed but with a diffuse mass such as that in Figure 1 B, it is inevitable that part of the tumor will remain after surgery. The characteristics of that left behind will have a strong influence on patient survival and so the tumor boundary heterogeneity should ultimately have the strongest association with patient survival.

The selected images in Figure 1 illustrate the diversity that exists in the sarcoma population, not only in terms of tumor size, shape and location but also in terms of the variation in the distribution of the FDG uptake. The FDG uptake portrayed in image A is compact throughout the tumor and while C is also compact in nature it is localised near the centre, therefore portraying different characteristics on the boundary. Images B and D are diffuse tumors and hence are spatially heterogeneous both on and within the boundary. The latter two tumors are more likely to be partially left behind after resection and therefore potentially pose more of a threat to patient survival. An alternative heterogeneity measure is developed in this study to assess boundary activity and its ability to predict tumor progression and overall patient survival.

Methods

Heterogeneity within the tumor itself is thought to be a major factor contributing to the failure of standard clinical measures in treating sarcoma patients. Ignoring activity within the tumor a technique to locate the centre of mass of the boundary is developed and the deviation of the FDG uptake on the tumor boundary from this centre of mass is examined. As the single-voxel boundary defined by a solid line in Figure 1 may not be an exact replica of the tumor boundary removed during surgery, a more stable boundary definition is used incorporating several voxels in the neighbourhood of the boundary. In developing the measure of heterogeneity specific to the boundary a distance component which measures the path distances between voxels, evaluated through a shortest path algorithm, is combined with a directional component representing the direction of a voxel from the centre of mass of the tumor boundary.

Results

We explore the ability of the tumor boundary heterogeneity to predict patient survival and disease progression. A Cox proportional hazards survival analysis is carried out, and to facilitate comparison with previous work includes standard prognostic factors as well as alternative PET-based measures. Preliminary results using the boundary coefficient of variation as a measure of heterogeneity show no improvements in risk prediction on that previously recorded. Thus a more refined characterisation of the spatial pattern of boundary heterogeneity may be required.

References

O'Sullivan et al. (2003). A statistical measure of tissue heterogeneity with application to 3D PET sarcoma data. *Biostatistics,* 4(3), 433-448.

O'Sullivan et al. (2005). Incorporation of tumor shape into an assessment of spatial heterogeneity for human sarcomas imaged with PET. *Biostatistics,* 6(2), 293-301.

Supported in part by NIH (CA 065537)

Central Composite Design Applied to Drug Production

K. O'Sullivan[1], S. O'Neill[1], J. O'Mullane[1], J. Huang[1], M. Rea[2] and D. Cadogan[2]

[1] University College Cork

[2] Wexport Ltd., Cork

Keywords: response surface methodology, central composite design, optimization, enzymatic depolymerisation

Introduction

Tinzaparin, a low molecular weight Heparin, is a drug widely used for the prevention and/or treatment of blood clots. Tinzaparin compared to normal Heparin is a more effective anticoagulant resulting in a reduced risk of bleeding and a longer duration of action. Tinzaparin is derived from Heparin by enzymatic depolymerisation using Heparinase. Two factors of this system were evaluated; process factors A and B. The aim of this study was to determine the combination of these two factors that optimises an outcome variable in the manufacture of Tinzaparin while satisfying the required manufacturing specifications.

Methods

Response surface methodologies (RSM) are experimental procedures employed to identify factor settings that optimise a response. Specifically, a rotatable central composite design was used. Five levels of each factor were chosen with 22 experimental runs being performed. A second-order model permitted the evaluation of linear, quadratic and interaction effects of the factors on the outcome variable. Three-dimensional surface plots and contour plots were drawn for illustration. The specifications of the additional parameters were assessed by examining the relationship between the process factors A and B, and each of the parameters. Each model was evaluated in terms of the optimal factor settings and the required specifications.

Results

The analysis of the second-order model indicated that a linear-model provided a good fit to the data (Table 1).

Source	DF	SS	F	P
Regression	2	3.975	4.12	0.033
Linear	2	3.975	4.12	0.033
Residual Error	19	9.156		
Lack-of-it	6	1.743	0.51	0.791
Pure Error	13	7.412		
Total	21	13.131		

TABLE 1. ANOVA results for the first-order response surface model.

Response surface and contour plots were derived from the linear model (Figure 1).
These plots indicate that low levels of process factor A produced optimal values of the outcome variable, while process factor B had negligible effect. This was confirmed from the tests of the individual factors (P-values for process factors A and B were 0.010 and 0.956 respectively). However, the manufacturing specifications necessitated that these factors had to be constrained. Subsequent analysis revealed that one constraint (K) dominated and that reducing process factor A must occur with an increase in process factor B so as to meet required specifications.

Conclusion

This experimentation and analysis demonstrated that although process factor A had a significant impact on optimisation of the outcome variable, low levels of process factor A produced optimal values of the outcome variable

(this did not depend on process factor B), it was not possible to reduce process factor A without increasing process factor B so as to maintain manufacturing specifications. Constraint K had the greatest impact on process factor settings and process factor A is constrained to at least 93 to meet manufacturing specifications. Transference from laboratory to the production process had begun with a lower level of process factor A and a higher level of process factor B than typically used by the plant.

Bayesian methods for analysing relative sea level data

A. Parnell[1] and C. Anderson[2]

[1] Trinity College Dublin

[2] University of Sheffield, UK

Abstract

We propose a method for estimating relative sea-level (RSL) curves and their associated uncertainties. More specifically, we estimate the changing state of sea level in the Humber estuary, UK, over the course of the last 10,000 years. These estimates are obtained through Bayesian methods involving Gaussian processes. Our method allows sea-level scientists the opportunity to accurately gauge the uncertainty in sea level, whilst accounting for many sources of uncertainty in the data.

Keywords: Bayesian modelling, Gaussian process, Sea level.

Introduction

Predictions of future sea level rise show large parts of the Ireland and other areas of the world to be under threat (Rahmstorf, 2007). As with all predictions, the associated models rely on past data to learn about model features. It is therefore vital to gain a good understanding of past changes in sea level. Current research into former sea level is concentrated in two areas: firstly, that of creating physical models which attempt to separate and deterministically evaluate important components of sea-level change (eg Peltier, 1998); secondly, those concerned with examining the uncertainties in sediment data (known as *index points*) from samples located at or near former coastlines (Shennan and Horton, 2002). We focus on the latter class of models, which, traditionally, are used to estimate the parameters of the former. The uncertainty structure present in the index points is somewhat complex, and is unsuitable for traditional techniques such as linear regression (although this has been attempted). We present a method for interpolating between index points and for accounting for each of the sources of uncertainty.

Methods

We consider models of the form:

$$Y = X\beta + \zeta + \Gamma + \Sigma \tag{1}$$

where Y is RSL elevation in metres, X is a (stochastic) design matrix containing calibrated radiocarbon dates and locations applied to parameter vector β, ζ is a spatio-temporal Gaussian process, Γ is a set of shift parameters to allow for different types of index point, and Σ is a zero-mean i.i.d. Gaussian error vector with known variances. Individually, we treat the elevation of each index point, y_i, as conditionally independent given RSL, $\eta(\theta_i, \Lambda)$, (with $\eta(\theta_i, \Lambda) = X\beta + \zeta$), at date θ_i, location Λ and standard deviation σ_i:

$$y_i|\eta(\theta_i, \Lambda), \sigma_i \sim N(\eta(\theta_i, \Lambda), \sigma_i^2) \text{ for } i = 1, \ldots, n. \tag{2}$$

Results

RSL curves are obtained for a number of locations around the Humber Estuary, UK. In particular, the model allows for varying uncertainty in the rate of RSL change. It also allows predictions of RSL in calendar years, after allowing for the uncertainty due to the calibration of radiocarbon dates. Finally, with some further analysis on the Gaussian process output, we are able to ascertain some important geological information on the nature of the collected data.

References

Peltier, W. (1998). The inverse problem for mantle viscosity. *Inverse Problems*, 14, 441478.

Rahmsdorf, S. (2007). A Semi-Empirical Approach to Projecting Future Sea-Level Rise. *Science*, 315, 368-370.

Shennan, I. and Horton, B. (2002). Holocene land and sea level changes in Great Britain. *Journal of Quaternary Science*, 17, 511526.

Gaussian Approximation Techniques

M. Salter-Townshend[1] and J. Haslett[1]

[1] Trinity College Dublin

Abstract

We address some new developments in the forward model stage of an ongoing palaeoclimate reconstruction project via analysis of pollen data. In particular, following the work of Rue et al (2007) we look at deterministic approximations to posteriors as an alternative to Markov Chain Monte Carlo (MCMC). We also look at the synergy of zero-inflation methodology that we are developing with this framework.
Keywords:Bayesian inference, Gaussian Markov Random Fields, zero-inflation, spatial process, compositional and counts data.

Introduction

Haslett et al (2006) reported on the reconstruction of the palaeoclimate. The essential science is that pollen found in lake sediment reflects ancient vegetation, which in turn reflects the ancient climate. Thus changes in the pollen composition with sediment depth reflect changes in the climate with past time.

The reconstruction task may be stated simply: given a forward model for the random variation in pollen counts (built on a modern data set including climate measurements) construct the posterior distribution of climate, given additionally one or more count vectors taken from sediment, and thus corresponding to unknown fossil climate.

The computational challenge lies in the high dimensionality of the number of points in space and time at which reconstructions of the vegetation response to climate (forward model) are sought. This is in fact the main crux of the current methodology which achieves such reconstructions via MCMC simulation with a very large number (order 10^4) of variables.

Zero-inflation of the counts data in the forward model is an added complication for which we seek a parsimonious approach.

Methods

Latent parameters describing our response surfaces are modelled a priori as Gaussian Markov Random Fields (GMRFs; see Rue and Held (2005)). Neighbours of points on a lattice are specified via a graph. The surfaces are Gaussian and furthermore are Markov; given its neighbours, each point is conditionally independent of all non-neighbouring points. This makes the precision matrix very sparse and some properties of GMRFs now become very useful. Additionally, the posterior can be well approximated with a Gaussian distribution by matching the mode and the curvature at the mode.
Zero-inflation is modelled with the addition of a single extra parameter. Response when present and probability of potential presence are modelled as two functions of a single underlying spatial process.

Results

Results that show highly accurate approximations to posteriors of response surfaces can be made using these deterministic approximation techniques within minutes using existing optimised C libraries. Compare this to sampling via MCMC (conducted on an 8 node Beowulf cluster using specialised and parallelised software) which is typically run for more than a week - at which point convergence is still not assured - and the benefits are obvious.

Discussion

Deterministic approximations can be conducted many orders of magnitude quicker than stochastic MCMC methods. Furthermore, MCMC is itself an approximation. Chains must be run for extremely long times to detect any error in the deterministic approximations. The techniques also allow for computationally expensive tasks such as cross validation to be performed conveniently.

References

Haslett, et al (2006). Bayesian Palaeoclimate Reconstruction *JRSS A.* 169, 3, 1-36

Rue, et al (2007). Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations *Available as a technical report at http://www.math.ntnu.no/preprint/statistics/2007/S1-2007.pdf*

Rue H and Held L (2005). Gaussian Markov Random Fields: Theory and Applications *Monographs on Statistics and Applied Probability* Vol 104

Optimal Choice of λ in Reconstruction of Wave Height Fields from Light Transmission Data

M. Samanta[1], K. Choudhury[1] and F. O'Sullivan[1]

[1] University College Cork

Abstract

We have been developing statistical methods for the study of wind-wave dynamics with the collaboration of researchers from IRPHE, Marseille. Their experiments capture measurements of video sequences of slope images in the x- and y- directions (blue and red slope respectively) from which the weighted height field can be constructed. In a discretized form the slope images, (Z_x, Z_y), can be expressed in the form of a linear model:

$$Z_x = D_x h + W_x^{-1/2}\epsilon_x; \qquad Z_y = D_y h + W_y^{-1/2}\epsilon_y, \tag{1}$$

where D_x and D_y are divided difference matrices, h is the discretized wave height field, W_x and W_y are the weights and ϵ_x and ϵ_y are the noises in the x and y directions respectively. The method of regularization is applied for recovery of the wave height image from these data. A weighted penalised objective function corresponding to model (1) has been formulated.

$$l_\lambda(h) = (Z_x - D_x h)' W_x (Z_x - D_x h) + (Z_y - D_y h)' W_y (Z_y - D_y h) + \lambda \|h\|^2, \tag{2}$$

where λ is the regularisation parameter.
An iterative algorithm based on Gauss-Jacobi was previously developed (Samanta, M. et. al. CASI'05).
Optimal reconstruction of height is based on the choice of optimal λ. It could be decided on the basis of RSS (Residual Sum of Squares). But the problem with RSS is that it is not satisfactory as a model selector as it doesn't give an idea how a learner will do when it is asked to make new predictions for data it has not already seen. So Cross Validation technique is introduced and for computation simplicity GCV (Generalized Cross Validation) (Craven and Wahba, 79) is adopted

$$GCV(\lambda) = \frac{RSS(\lambda)/N}{[1 - \text{Trace}(H(\lambda))]^2},$$

where

$$H(\lambda) = \begin{bmatrix} D_x \\ D_y \end{bmatrix} [D_x' D_x + D_y' D_y + \lambda I]^{-1} [D_x' D_y'].$$

Therefore,

$$Trace(H(\lambda)) = \text{Trace}\left(\begin{bmatrix} D_x \\ D_y \end{bmatrix} [D_x' D_x + D_y' D_y + \lambda I]^{-1} [D_x' D_y'] \right). \quad (3)$$

Trace($H(\lambda)$) is computed using equation (3) for 7 different dimensions (4X5, 5X6, 5X10, 10X15, 15X20, 20X30, 40X50) and shown in Fig 1(a).
RSS(λ) and then Trace($H(\lambda)$) are calculated to compute GCV(λ). Unlike small dimension it's a difficult task to compute trace for large dimension data (e.g., 254X350) which happens to be our case. So our primary task is to express the trace as a function of dimension (υ) and λ. Wahba (1990) has already provided an expression for Trace($H(\lambda)$) with eigenvalues (λ_υ) as

$$\text{Trace}(H(\lambda)) = \sum_{\upsilon=1}^{n-1} \frac{\lambda_\upsilon}{\lambda_\upsilon + \lambda}. \quad (4)$$

Trace($H(\lambda)$) is computed for different dimensions using λ_υ of $[D_x' D_x + D_y' D_y]$ and plotted in Fig. 1(b). Fig. 1(a) and Fig. 1(b) are exactly same which implies it would be enough if we can express λ_υ as a function of υ only. Therefore, using υ and λ_υ, Trace($H(\lambda)$) and finally GCV(λ) may be calculated.
Analyses and results of a detailed simulation study for obtaining the functional form will be presented at the conference.

References

Samanta, M., Roy Choudhury, K and O'Sullivan, F. (2005). Weighted Reconstruction of Wave Height Fields from Light Transmission Data *CASI*.

Wahba G. (1990). *Spline models in statistics.* CBMS-NSF Regional Conference Series, SIAM.

Craven P. and Wahba G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik,* 31, 377-403.

RiboSort: an R package for rapid classification and preliminary analysis of microbial community profiles

U. Scallan[1], A. Liliensiek[1] and J. Connolly[1]

[1] University College Dublin

Abstract

Community fingerprinting techniques, such as T-RFLP and ARISA, have commonly been applied to studies of microbial ecology, significantly increasing our understanding of the role and diversity of bacteria in the environment (Anderson & Cairney 2004). Automatic sequencers are used to carry out these analyses and generate microbial community fingerprints for each sample. Data produced by the sequencer is rarely in a format suitable for statistical analysis, and the process of manually sorting and manipulating profiles into the desired format is a tedious and time-consuming operation. We describe a computer package that works directly on multiple sequencer output files to produce a single spreadsheet of classified profiles ready for statistical analysis. In addition to saving time, the use of this program can improve the accuracy and consistency of classification by eliminating human error. RiboSort also provides tools for an initial exploratory analysis of the data, including the straightforward and speedy creation of Multi-dimensional Scaling plots to compare samples. The functionality of RiboSort is illustrated with an example of data from the European Biodioversity study, COST 852. Our presentation also highlights the benefits of the Package functionality in R, and briefly describes how to go about creating a Package.

Keywords: R Package, Data Classification, Microbial Community Fingerprints, Multi-dimensional Scaling.

References

I. C. Anderson and J. W. G. Cairney (2004). Diversity and ecology of soil fungal communities: increased understanding through the application of molecular techniques. *Environmental Microbiology* 6, No 8, 769-779.

One Sided Classification

D. Toher[1,2], G. Downey[1] and B. Murphy[2]

[1] Ashtown Food Research Centre, Teagasc, Dublin

[2] Trinity College Dublin

Abstract

In food science classification problems, the main focus of the investigation can be the identification and classification of a single group. The total number of groups within the data may be unknown – the analyst being solely interested in differentiation between one group and all others. As such, we develop a model whereby a single group of observations is modelled using model-based classification techniques and all other observations are modelled using a Poisson distribution.

Keywords: Food Science, Model-based Classification, One-sided Classification, Variable Selection.

1 Introduction

When examining the veracity of food labelling food scientists must classify samples into two basic categories: *"label correct"* or *"label incorrect"*. To adequately describe the *"label incorrect"* group challenging because there a large number of reasons why a product may be mislabelled (*e.g.* incorrect region of origin or incorrect ingredient information). Even once the source of the mislabelling is identified, there are many different possibilities of what the product might be. For example, if the analyst was able to determine that the region of origin was incorrectly specified on the product, trying to identify the actual region of origin remains problematic. Fully characterising all possibilities for mislabelling and contamination of foods is unpractical, therefore we propose a model that requires only the group of interest (or what the product is labelled to be) to be fully characterised and where observations not belonging to this group require less information.

2 Methods

In model-based discriminant analysis, the model is fitted to data $\mathbf{x}_n$ where $n = 1, 2, \ldots, N$ and labels $\mathbf{l}_n$ where $l_{ng} = 1$ if observation n belongs to

group g and 0 otherwise. Given that there exist G groups and that the probability of an observation belonging to group g is p_g:

$$Data \sim \prod_{g=1}^{G} p_g N(\mu_g, \Sigma_g)$$

Allowing $\theta_g = (\mu_g, \Sigma_g)$, the resulting likelihood function is,

$$\mathcal{L}_{\text{disc}}(p_1, p_2, \ldots, p_G; \theta_1, \theta_2, \ldots, \theta_G | \mathbf{x}, \mathbf{l}) = \prod_{n=1}^{N} \prod_{g=1}^{G} [p_g f(\mathbf{x}_n | \theta_g)]^{l_{ng}} . \quad (1)$$

Using a Gaussian distribution to describe observations within the single group of interest and assuming the other observations can be adequately described by Poisson noise:

$$Data \sim p_g N(\mu, \Sigma) + p_o / V \quad (2)$$

where p_g is the probability of belonging to the single group, $p_o = 1 - p_g$ and V is an estimate of the volume of the data.
Modelling the single group of interest as a mixture of Gaussian distributions, the data are then described as follows:

$$Data \sim p_g f_g(x|\theta_g) + p_o f_o(x|\theta_o) \quad (3)$$

where $f_o(x|\theta_o) = 1/V$ and $f_g(x|\theta_g)$ is a mixture of Gaussian distributions. The number of Gaussian distributions required to describe the single group of interest is unknown and thus must be estimated from the data.
Using the EM algorithm on the complete data log-likelihood functions of the equations (2,3) and a greedy stepwise variable selection method we apply this methodology to a number of different food science datasets. Brier's score is used as a measure of performance, quantifying the accuracy and certainty of predicted group memberships.

References

Bensmail, C. and Celeux, G. (1996). Regularized Gaussian discriminant analysis through eigenvalue decomposition. *JASA.* 91, 1743-1748.

Brier, G. W (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review.* 78,1-3.

Dempster, A. P. et al (1977). Maximum likelihood for incomplete data via the EM algorithm (with discussion). *JRSS B.* 39, 1-38.