Stat. Meth. & Appl. (2008) 17:167–181
DOI 10.1007/s10260-007-0079-z

ORIGINAL ARTICLE

Some challenges for statistics

A. C. Davison

Accepted: 18 October 2007 / Published online: 11 December 2007

Abstract The paper gives a highly personal sketch of some current trends in statistical inference. After an account of the challenges that new forms of data bring, there is a brief overview of some topics in stochastic modelling. The paper then turns to sparsity, illustrated using Bayesian wavelet analysis based on a mixture model and metabolite profiling. Modern likelihood methods including higher order approximation and composite likelihood inference are then discussed, followed by some thoughts on statistical education.

Keywords Bayesian inference · Composite likelihood · Likelihood asymptotics · Metabolite profiling · Mixture model · Sparsity · Statistical education · Stochastic model · Wavelet regression

1 Introduction

Statistical methodology develops largely in response to the demands of the society in which it is rooted, and the scientific challenges of the age. Thus it is difficult to discuss challenges for statistics without reference to the major preoccupations of our time. On

This paper is based on a lecture given at a ceremony to inaugurate the new building of the Department of Statistics at the Università Ca' Foscari, Venice, in September 2006. Some of the work described was performed in collaborations with Christophe Ancey, Alessandra Brazzale, Gaëlle Messerli, Vahid Partovi Nia, Nancy Reid and Sam Zeeman. The author thanks members of the Venice department for their generous hospitality and Christophe Ancey, Nicola Sartori, Victor Panaretos, Vahid Partovi Nia and referees for their helpful comments. The work was supported by the Swiss National Science Foundation.

A. C. Davison (✉)
Institute of Mathematics, School of Basic Sciences, Ecole Polytechnique Fédérale de Lausanne, STAT-IMA-FSB-EPFL, Station 8, 1015 Lausanne, Switzerland
e-mail: anthony.davison@epfl.ch
URL: http://stat.epfl.ch

a grand scale, a partial list might include the collapse of some nation states, population growth, health risks due to spread of infectious diseases and, looming behind these, the elephant in the living-room—environmental change and its consequences. On a smaller scale, there seems to be consensus that the main scientific and engineering challenges for the next quarter-century will be to understand the working of the genome and to use this knowledge for the general good, and to develop technologies that will enable us to live sustainably rather than wildly beyond what our planet can afford.

Statistics is woven from problems and data from substantive disciplines, from the mathematical ideas used to construct stochastic models intended to extract information from the data and, increasingly, from computing technologies, which provide environments within which mathematical ideas may be turned into statistical tools. These threads provide the warp for this paper, which first briefly discusses the abundance of data now available, before turning to stochastic modelling. In Sect. 4 the topic of sparsity is illustrated using Bayesian wavelet analysis and metabolite profiling, before a discussion in Sect. 5 of two topics in likelihood inference, higher order asymptotics and composite likelihood. Section 6 gives some thoughts on the university teaching of statistics, followed by a brief conclusion. Efron (2003) gives a wide perspective on the past and future of statistics, based on the same three threads.

The paper has no pretence to be inclusive: even if I were competent to write an overview of current statistical thought, there is not the space to do so. Rather this is a cartoon of some topics of current interest.

2 Data

One of the most striking changes of recent years is the increasing abundance of data. During my doctoral work around 25 years ago I studied a database of around 13 million numbers, but this was then exceptionally large; those of my fellow students whose research stemmed directly from an applied problem had data sets consisting of a few hundred or perhaps a few thousand numbers. Nowadays there is such a profusion of data that it is difficult not to feel overwhelmed. The main reason is the steady drop in price and increase in capacity of electronic equipment: for example, almost any laboratory can now record images using a webcam or video camera, high-throughput genomic analysis has become the norm in the biological sciences, and the resulting sea of data can be cheaply stored on a large hard disc and made available over the internet. Domains that have profited from this include:

- the biosciences, with the availability of huge quantities of genomic, proteomic, metabolomic and other-'omic' data;
- chemistry, through gas chromatography/mass spectrometry and related techniques, and in the increasingly detailed understanding of molecular interactions, protein folding and the like;
- physics—for example, detection of new elementary particles at installations such as CERN depends on the extraction of a few unusual events from a mind-boggling number of observed fission tracks;
- forensic science, which is increasingly a probabilistic enterprise;
- finance, now an enormous industry based largely on stochastic models;

– commerce, where transaction modelling and credit scoring have transformed, not always for the better, how banks and retailers interact with their clients;
– transport engineering, as traffic flow on major routes can be regulated based on real time data and stochastic models of communication networks; and
– environmental monitoring using dense networks of cheap sensors, to gain an unprecedented level of detail on conditions at the earth-atmosphere interface—see for example http://sensorscope.epfl.ch.

One result is the increasing quantification of many domains of knowledge, as data replace speculation. In some countries a side-effect of this is that studying the mathematical sciences seems to be becoming more attractive: quite apart from the challenge of problem-solving afforded by a numerate degree, the prospect of a well-paid and interesting career is a strong incentive for students.

In some domains it is now possible to have *all* the data. Databases held by banks on their customers, for example, may contain every transaction ever performed, and may be used to determine which customers are likely to become credit risks. Of course the mere availability of such data does not make it useful; the economic conditions that last year made bankruptcy more likely for one type of customer may not be those that cause difficulties for another type of customer next year, so prediction is, as usual, fraught with difficulties. Changes due to this so-called *database drift* raise questions about the appropriateness of highly sophisticated discrimination methods (Hand 2006).

In other domains the number of sampling units remains limited because of financial or other constraints, but the number of measurements on each unit is now much larger than hitherto. The numbers of patients who may be recruited into a clinical trial remains limited by practical considerations, but advances in technology make it possible to have several hundred thousand single nucleotide polymorphisms (SNPs) on each. This raises questions about the comparability of such data, especially when observations from different centres are combined; for example, combination of microarray data from different laboratories can pose serious problems of interpretation in clinical studies.

Elsewhere, for example in social science, budget limitations mean that the quantity of data cannot grow without damaging its quality, though there is scope for linking together existing databases, for example on employment histories and health—so-called *data fusion* is an important topic for national statistics institutes, as well as in many scientific domains.

Underlying this discussion are traditional statistical issues for design of investigations: the choice and construction of sampling plans with appropriate randomisation, replication and blocking and the importance of recognising selection and other forms of sampling bias. It is unfortunate—a traditional complaint of statisticians—that these topics remain under appreciated, to the point that many costly investigations are compromised by poor design.

3 Stochastic modelling

The huge quantities of data now available give correspondingly vast scope for modelling. A simple illustration is afforded by Fig. 1, taken from an experiment conducted to understand the basis of sediment transport and erosion in river beds

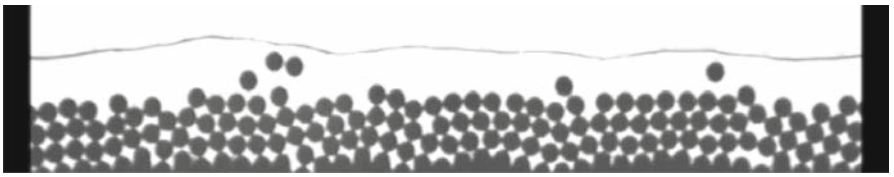

Fig. 1 Still image of beads in motion in an experimental stream (Böhm et al. 2004). The beads are entrained by the flow of the water from left to right, and their positions are recorded 130 times per second. Afterwards the numbers in various states of motion can be determined using imaging software

(Böhm et al. 2004). The figure shows a single frame from a video lasting 60 s, with around 130 frames taken each second—8,000 frames in all. Two parallel sheets of glass are placed 6.5 mm apart, with a base inclined at an angle of around 10° from the horizontal; water flows between them, from left to right in the figure; and beads of diameter 6 mm are released into the flowing water, which then entrains them. The beads can only form a single layer, as transversal movement is impossible. Some beads form a bed, some roll along the surface of the bed, and others bounce along in the direction of the moving water. Image processing software can be used to count the numbers of stationary, rolling and bouncing beads, allowing theories about sediment transport to be tested empirically. A simple surprisingly successful model is an immigration–birth–death process in which immigration corresponds to the arrival of particles from the left, births occur when particles in the image are set in motion, and deaths occur when particles stop or leave the observation frame on the right. This stochastic model seems to describe the behaviour of the beads better than the traditional continuum approach, though there is still room for improvement (Ancey et al. 2008).

More sophisticated examples could be taken from almost any quantitative journal. Some that spring to mind are modelling of the water balance between soil and vegetation (Porporato and Rodríguez-Iturbe 2005), the use of Lévy and other heavy-tailed processes in finance (Barndorff-Nielsen et al. 2001), shape statistics applied to single molecules (Kou et al. 2005; Panaretos 2006), quantum statistical inference (Barndorff-Nielsen et al. 2003), modelling of epidemics with partly unobserved data (Panaretos 2007; Isham 2005), and spatial point process models of rainfall (Cox and Isham 1988).

In these and other applications there is a tension between conceptual modelling of how the data might arise, based on a few key elements, and the detailed representation of component processes, which renders the model less tractable and thus perhaps less useful as an aid to understanding. Clearly the level of detail to be included depends on the goal of the exercise: a major conceptual advance may stem from a simple model involving only the main processes, whereas an activity such as short-term local weather forecasting may require consideration of many processes of atmospheric physics. In the second case as in many others an alternative is the use of predictive 'black boxes' such as neural nets, random forests, support vector machines, and the like, as forcefully advocated by Breiman (2001), though that article should not be read in isolation from the subsequent discussion by Cox and Efron.

4 Sparsity

4.1 Generalities

Many modern applications involve the extraction of a signal from a noisy environment. Examples are finding the gene or combination of genes responsible for a congenital disorder, cleaning a biomedical image, and identifying SPAM emails. Often we seek a sparse representation of the signal, as just a few components, and aim to compress the data as tightly as possible without losing its essence. In many cases the signal will contain many similar elements (genes, pixels, letters), and we seek a good ensemble estimate, which may be provided by shrinkage of its elements towards a common value. Approaches to this in different contexts include the lasso and related estimators (Tibshirani 1996; Efron et al. 2004), mixed modelling (McCulloch and Searle 2001) and Markov random fields (Isham 1981; Clifford 1990; Chellappa and Jain 1993); see also Hastie et al. (2001). Below we briefly outline one technique for building models with sparse elements and give two examples of its use.

4.2 Bayesian wavelet analysis

Figure 2 shows data for which a sparse representation seems necessary. The upper left hand panel shows a transect taken using nuclear magnetic resonance (NMR) imaging. There is a clear if irregular signal with several spikes obscured by homoscedastic noise. The upper right hand panel shows an orthogonal transformation of the data in terms of wavelets, the details of which are unimportant here.

To model this, we follow others such as Abramovich et al. (1998) in supposing that the data may be treated as the realisation of a $n \times 1$ vector $X = \mu + \varepsilon$, where μ represents the signal and the elements of ε are independent normal variables with zero mean and variance σ^2. Let $Y = W^{\mathrm{T}}X$, where W is an $n \times n$ orthogonal matrix representing the wavelet transformation; thus $W^{\mathrm{T}}W = WW^{\mathrm{T}} = I_n$. The matrix W is chosen so that $\theta = W^{\mathrm{T}}\mu$ should be sparse, that is, most elements of θ are small or even zero. Wavelets are known to have this property for a wide variety of functions; in other contexts one might choose other orthogonal transformations of the data, such as discrete Fourier series. In each case, the idea is to choose an operator that 'kills' small elements of Y by setting them to zero, and then to estimate the signal by applying the inverse transformation to the shrunken coefficients, yielding the estimator $\tilde{\mu} = W\{\mathrm{kill}(W^{\mathrm{T}}X)\}$.

A possible prior model is that the coefficients $\theta_1, \ldots, \theta_n$ are drawn independently from the mixture

$$\theta \sim \begin{cases} 0, & \text{with probability } 1-p, \\ \mathcal{N}(0, \tau^2), & \text{with probability } p, \end{cases}$$

or equivalently that

$$\pi(\theta) = (1-p)\delta(\theta) + p\tau^{-1}\phi(\theta/\tau), \quad \theta \in \mathbb{R},$$

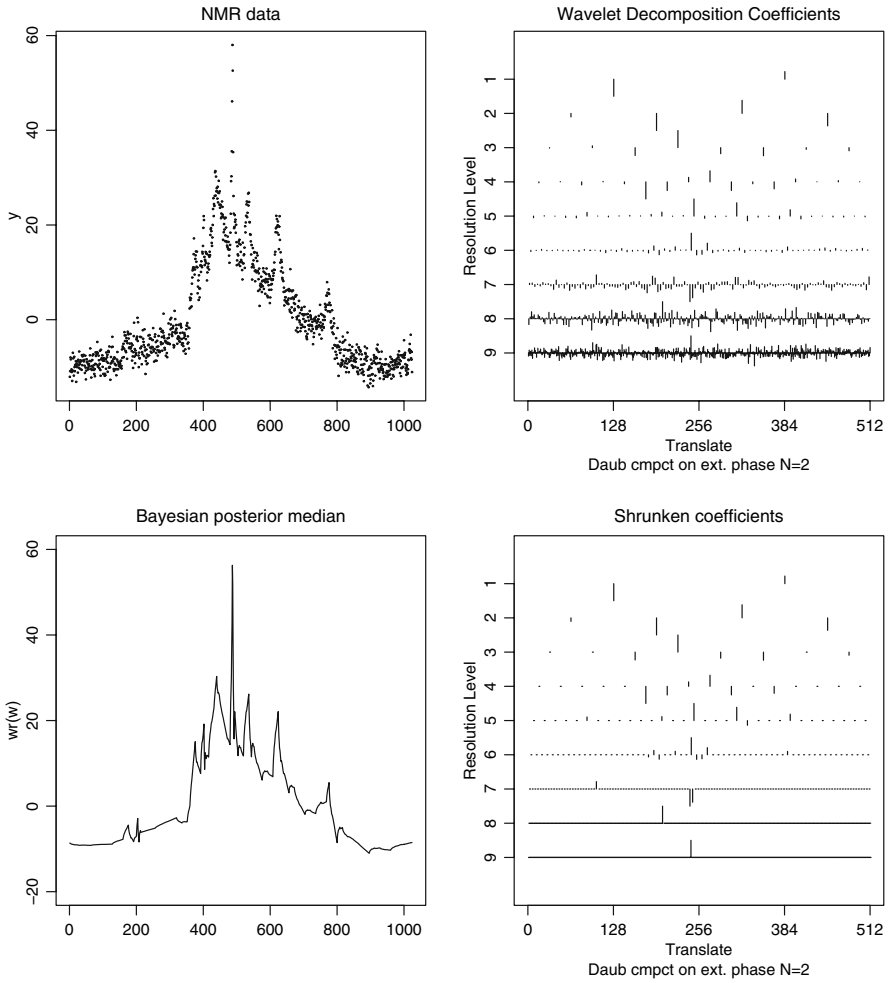

Fig. 2 Nuclear magnetic resonance data. *Left top* data transect, with $n = 1{,}024$ values translated to have average zero. *Right top* coefficients of the wavelet decomposition of the transect. *Left bottom* Bayesian posterior median reconstruction of the transect. *Right* Bayesian posterior median wavelet coefficients

where $\delta(\cdot)$ is the delta function putting unit mass at $\theta = 0$, and $\phi(\cdot)$ represents the standard normal density. Conditional on $\theta_1, \ldots, \theta_n$, we take the elements of Y to be independent normal variables with means θ_j and variance σ^2. Then the θ_j are independent conditional on the data and

$$\pi(\theta \mid y) = (1 - p_y)\delta(\theta) + p_y b^{-1} \phi\left(\frac{\theta - ay}{b}\right), \quad \theta \in \mathbb{R}, \tag{1}$$

where

$$a = \tau^2/(\tau^2 + \sigma^2), \quad b^2 = 1/(1/\sigma^2 + 1/\tau^2),$$

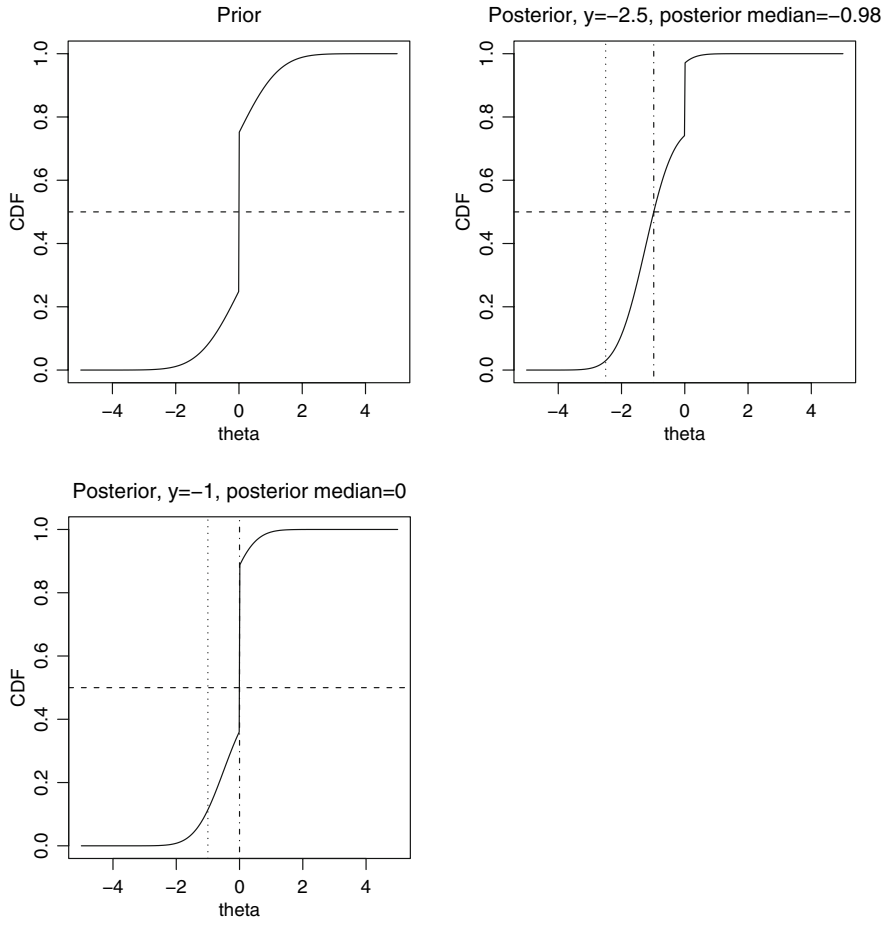

Fig. 3 Bayesian soft thresholding: prior distribution of θ (*top left*), and posterior cumulative distribution functions when $p = 0.5, \sigma = \tau = 1$, and $y = -2.5$ (*top right*), and $y = -1$ (*bottom*). *Horizontal line* (*dashes*): cumulative probability=0.5. *Vertical line* (*dots*): observation y. *Vertical line* (*dot-dash*): posterior median $\tilde{\theta}$

and

$$p_y = \frac{p(\sigma^2+\tau^2)^{-1/2}\phi\{y/(\sigma^2+\tau^2)^{1/2}\}}{(1-p)\sigma^{-1}\phi(y/\sigma) + p(\sigma^2+\tau^2)^{-1/2}\phi\{y/(\sigma^2+\tau^2)^{1/2}\}}$$

is the posterior probability that $\theta \neq 0$.

The posterior mean is the most obvious summary of (1), but if a sparse solution is sought it is better to take the posterior median of (1), that is, the value $\tilde{\theta}$ that satisfies $\Pr(\theta \leq \tilde{\theta} \mid y) = 0.5$. As Fig. 3 shows, this performs a form of soft shrinkage (Donoho and Johnstone 1994): if y is sufficiently close to 0, then $\tilde{\theta} = 0$, and otherwise $\tilde{\theta}$ lies between y and 0.

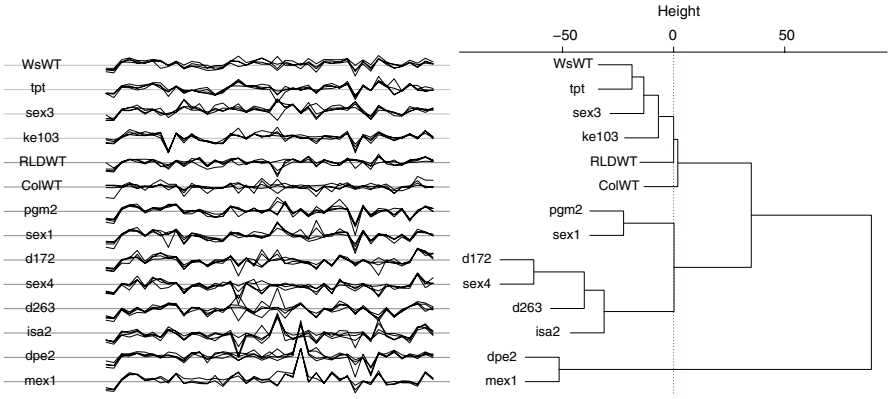

Fig. 4 Metabolic profile data. The *left part* shows four replicate metabolic profiles for each of 14 plant phenotypes, with measurements on 42 metabolites. The right part shows a hierarchical clustering tree obtained using empirical Bayes estimation of a mixture model; the scale 'Height' is log marginal likelihood below zero, and minus log marginal likelihood above zero. The goal is to classify the unknown phenotypes `d172` and `d263` to the known ones

The unknown parameters p, σ, τ may be estimated by maximizing the marginal likelihood based on $y_1, \ldots, y_n$, noting that the y_j are independent conditional on these parameters, with density

$$f(y \mid p, \sigma, \tau) = (1-p)\sigma^{-1}\phi(y/\sigma) + p(\sigma^2+\tau^2)^{-1/2}\phi\{y/(\sigma^2+\tau^2)^{1/2}\},$$

The resulting estimates for the data shown in Fig. 2 are $\tilde{p} = 0.04$, $\tilde{\sigma} = 2.1$, and $\tilde{\tau} = 52.1$, and the corresponding shrunken coefficients and reconstructed signal are shown in the lower panels of the figure. Almost all the coefficients have been set to zero, yet the reconstruction picks out most of the salient features of the data.

The approach above is simple enough to be explained to an undergraduate audience, yet sufficiently powerful to be of real use in complex problems. It can be improved by replacing the point mass/normal mixture with a point mass/Laplace mixture; this too allows analytical calculations and has excellent frequentist properties (Johnstone and Silverman 2005).

4.3 Metabolic profiling

Another use of mixture models such as that outlined in the previous section is in statistical analysis of metabolic profiles. The data shown on the left of Fig. 4 are taken from an experiment performed by Gaëlle Messerli of the Institute of Plant Sciences at the ETH Zürich, and colleagues, in which gas chromatography/mass spectrometry was used to compare metabolic profiles of varieties of the plant *Arabidopsis thaliana* (Messerli et al. 2007). The main idea was that mutations affecting distinct metabolic or signalling pathways may have similar phenotypes, so the screening of traits such as the metabolic profile may allow discrimination of mutants of interest, in this case those

with known deficiencies in starch metabolism. Four replicates were obtained of each of a number of profiles taken from leaves of plants, some having known anomalies with their metabolic pathways, and others having anomalies of unknown origin. The purpose of statistical analysis of the profiles was to identify the known mutants closest to each of the unknown mutants, the aim being to use this information to assess which parts of the metabolic pathway could be responsible for the starch deficiencies observed in the unknown mutants.

Statistically this problem boils down to simultaneous hierarchical cluster and discriminant analysis. The tool used was a mixture model whose main components are a point mass and a normal distribution, like that described in Sect. 4.2, simple enough to allow analytical computation of the marginal likelihood of any given partition of the different profiles, and yet complex enough to allow the incorporation of different levels of variation: between leafs within profiles, between profiles within mutants, between mutants within elements of the partition, and between elements of the partition. The mixture model ensures a potentially sparse representation of the data, as only those metabolites that show substantial variation are used for the clustering. The hyperparameters may be estimated by marginal maximum likelihood, and a cluster tree may be estimated using agglomerative clustering (Heard et al. 2006; Lau and Green 2008). The result, shown in the right part of Fig. 4, agrees well with classical approaches to clustering: there are four main clusters, that at the top of the figure containing the wild types, the next containing two known types, the third grouping the unknowns `d263` and `d172` with the known types `isa2` and `sex4`, and the last containing the highly unusual profiles `mex1` and `dpe2`. The log marginal likelihood increases from around -80 when every profile is attributed to a single cluster to a maximum value of zero, and then decreases to a value of around -80 when there is a single cluster; in order to provide a tree of the usual form the sign of the log marginal likelihood has been changed to the right of zero on the plot.

This approach thus seems to provide plausible clustering trees without the Markov chain Monte Carlo computations required by other Bayesian approaches, but is fully probabilistic. Moreover it has the added advantage of ordering the metabolites in terms of their usefulness for the clustering.

4.4 Comments

Similar approaches based on mixture models have a wide range of other applications, such as detection of gene expression in microarrays (Lönnstedt and Speed 2002; Bhowmick et al. 2006). Although they are powerful tools for use with high-dimensional datasets, they raise questions: which are good general approaches to dealing with sparse high-dimensional data, and on what basis should we judge them? How should one perform inference for the resulting model-selected estimate? For those approaches where Markov chain Monte Carlo must be used, the perennial issue of convergence rears its ugly head, while it would be good to have reliable simulation approaches that can be used by non-experts without tuning.

5 Likelihood

5.1 Higher order asymptotics

One concept that would be on almost every statistician's list of core topics is likelihood, which with its many variants forms the basis of inference in an increasing variety of situations, including those where semiparametric models are used (Bickel et al. 1993; Murphy and van der Vaart 2000; Owen 2001; Rotnitzky 2005). Likelihood theory has been investigated for over 80 years, but is still capable of further development. Below I touch on just two aspects to which statisticians from the Veneto region have made valuable contributions.

Asymptotic arguments in statistics typically generate distributional approximations to be used with finite samples, for example to set confidence intervals or perform tests based on maximum likelihood estimates or likelihood ratio statistics. Likelihood asymptotics were initiated by Fisher in the 1920s (Fisher 1922, 1925), and remain the most widely used inferential tool on our workbench. It is less well-known that Fisher (1934) also suggested the basis of a more refined theory, without developing this beyond some special cases. Over the last 30 years prominent theoretical statisticians have made a major effort to develop this theory, which is now useable in a variety of applications. Consider a random sample of size n from a regular statistical model whose log likelihood $\ell(\psi, \lambda)$ depends on a scalar interest parameter ψ and a vector nuisance parameter λ. Let $\widehat{\theta} = (\widehat{\psi}, \widehat{\lambda})$ denote the overall maximum likelihood estimator and write the partially maximised estimator as $\widehat{\theta}_\psi = (\psi, \widehat{\lambda}_\psi)$, where $\widehat{\lambda}_\psi$ is the maximum likelihood estimator of λ with ψ held fixed. Then one basis for inference on ψ is the likelihood root

$$r(\psi) = \text{sign}(\widehat{\psi} - \psi) \left[2 \left\{\ell(\widehat{\theta}) - \ell(\widehat{\theta}_\psi)\right\}\right]^{1/2},$$

also sometimes called the signed deviance or the signed likelihood ratio statistic. Classical likelihood theory implies that the distribution of $r(\psi)$ is standard normal with error of order $n^{1/2}$; that is

$$\Pr\{r(\psi) \le r; \psi\} = \Phi(r) \left\{1 + O(n^{-1/2})\right\}, \quad \text{as } n \to \infty, \tag{2}$$

where the probability on the left is computed under the true model and Φ denotes the standard normal cumulative distribution function. As usual with asymptotic arguments in applied mathematics, the notion that n becomes large is simply a technical device used to generate approximations for use with finite n.

Expression (2) is a first order asymptotic approximation which can be used to test the hypothesis $\psi = \psi_0$, by computing the tail probability associated with $r(\psi_0)$, or to obtain the limits $(\psi^\alpha, \psi^{1-\alpha})$ of a $(1 - 2\alpha)$ confidence interval for ψ as the solutions to the equation $r(\psi) = \pm z_\alpha$, where z_α is the α quantile of the standard normal distribution. It is a remarkable fact that under essentially the same regularity conditions, and for a wide variety of continuous response models, replacement of $r(\psi)$ in (2) by the modified likelihood root $r^*(\psi) = r(\psi) + r(\psi)^{-1} \log\{v(\psi)/r(\psi)\}$

reduces the order of error to $n^{-3/2}$; in many cases the resulting tests and confidence intervals are essentially exact even for $n \approx 5$. The quantity $v(\psi)$ appearing in the definition of $r^*(\psi)$ depends on the model, but it may be computed explicitly in wide generality. For discrete response models the approximation error becomes $O(n^{-1})$ and the computation of $v(\psi)$ may be a little more complicated. There are close connections to related ideas such as modified profile likelihoods (Sartori 2003), which are widely used in practice.

The extensive theoretical literature on higher order procedures is summarised in the books of Barndorff-Nielsen and Cox (1994), Pace and Salvan (1997) and Severini (2000), and a recent review is provided by Reid (2003). Their application is illustrated in Brazzale et al. (2007), which makes use of an `R` package bundle `hoa` (for higher order asymptotics), and gives numerous practical examples. This may be viewed as a culmination of work by Brazzale and Bellio, among others, who have laboured to make these approximations widely available (Bellio 1999; Brazzale 1999, 2000; Bellio and Brazzale 1999, 2001, 2003).

The ideas described above extend easily to independent but non-identically distributed responses, but the literature contains little discussion of higher order asymptotics for dependent data or for non-regular problems—though see Castillo and López-Ratera (2006). Both these topics and the intriguing connections with Bayesian inference based on matching priors remain to be explored more thoroughly.

5.2 Composite likelihood

Statistical inference for parametric statistical models is ideally performed using the likelihood function, but this is unavailable or difficult to compute for many complex models. It may then be natural to use a composite likelihood function (Lindsay 1988), based on subsets $\mathcal{A}_1, \ldots, \mathcal{A}_m$ of the data for which the densities are available or are more readily computed; often these subsets will most naturally arise from consideration of marginal or conditional densities. The corresponding composite marginal log likelihood is

$$\ell_{\mathcal{A}}(\theta) = \sum_{j=1}^{m} \log f(y_{\mathcal{A}_j}; \theta)$$

in a natural notation. A simple example arises in analysis of a time series $y_1, \ldots, y_n$, where it may be tempting to replace the full likelihood

$$f(y_1, \ldots, y_n; \theta) = f(y_1; \theta) \prod_{j=2}^{n} f(y_j \mid y_1, \ldots, y_{j-1}; \theta)$$

by (Azzalini 1983)

$$f(y_1, \ldots, y_n; \theta) = f(y_1; \theta) \prod_{j=2}^{n} f(y_j \mid y_{j-1}; \theta)$$

if the one-step transition densities are easily computed; this amounts to behaving as if the process were first order Markov. Under regularity conditions analogous to those needed for the limiting normality of the usual maximum likelihood estimator, and if θ is identifiable from the marginal densities contributing to $\ell_{\mathcal{A}}$, the maximum composite likelihood estimator $\tilde{\theta}$ has a limiting normal distribution as $n \to \infty$, with mean θ and covariance matrix of sandwich form estimable by $J(\tilde{\theta})^{-1}K(\tilde{\theta})J(\tilde{\theta})^{-1}$, where $J(\theta)$ and $K(\theta)$ are are the observed information and squared score statistic corresponding to $\ell_{\mathcal{A}}$. Often the efficiency of $\tilde{\theta}$ relative to the maximum likelihood estimator is surprisingly high, though care must be taken not to allow m to become too large; if so the efficiency can drop, and in extreme cases $\tilde{\theta}$ may become inconsistent (Cox and Reid 2004).

Composite likelihood may be seen as the basis for the estimating equation

$$\frac{\partial \ell_{\mathcal{A}}(\theta)}{\partial \theta} = 0,$$

but has the advantage over arbitrarily-defined estimating equations of stemming from a well-defined objective function. It would be natural to base inference for components of θ on the corresponding profile log composite likelihood, but unfortunately the usual limiting Chi-squared result does not apply to the composite likelihood ratio statistic, whose distribution is a sum of differently scaled χ^2_1 variables. Varin (2008) reviews composite likelihood inference and gives many further references. It would be very interesting to compare composite likelihood with other approaches to inference, where such comparison is possible. A further topic worth investigating is to what extent Bayesian inference is feasible using composite likelihoods.

6 Training statisticians

One essential role of a university statistics department is to attract young people, and to educate them to be future leaders and users of our subject. An enormous amount has been written on statistical education, and I touch on just one aspect. The relation with mathematics and with other substantive subjects is crucial: what do we want our students to know and to be able to do? A strong mathematical background seems important for statistical theory, but computational skills are now key to much methodology, while substantive knowledge seems increasingly needed for applied work. If we are to train young people able and eager to take on future challenges, we must try and balance these three aspects.

The construction or revision of a curriculum typically involves deciding what can safely be left out, rather than what should be put in, so it is important to identify which skills and topics are core, and which are merely desirable. The core develops over time: 25 years ago the kernel of a regression course was the linear model, analysis of variance and non-linear and generalized linear modelling. Today it might also include generalized linear mixed models; nonparametric regression including local likelihood, spline smoothing and wavelets; the lasso and related approaches to sparse modelling; neural nets; classification and regression trees; survival data analysis; support vector machines and radial basis functions; and random trees and

forests. The course might also touch on more theoretical topics such as cross-validation and boosting (Freund and Schapire 1997; Bühlmann and Hothorn 2006). As this list is intended to make clear, continual addition to the core is infeasible, at least in the long term; rather the central elements need to be emphasised, with some key examples, so that new techniques can quickly be placed into a mental map of the domain.

One attempt to define a core is set out in Davison (2003), but I am well aware of its limitations. From this viewpoint key inferential topics are likelihood, estimating functions and Bayesian inference; key methods topics are regression in some of its varieties, study design including sampling and experimental design, multivariate statistics and simulation; and key models include a variety of stochastic processes, particularly Markov and point processes and time series.

Mathematical topics that seem essential for a statistician include real analysis, geometry, linear algebra, discrete mathematics, probability—perhaps with a dash of measure—and basic stochastic processes. Beyond this functional and numerical analysis are increasingly important for many statistical applications, with the necessary background of algebra and topology, and stochastic geometry and stochastic calculus also seem valuable. Further beyond this lie statistical applications of almost any conceivable domain of mathematics, examples being number theory, computational algebra, Riemannian geometry and algebraic geometry (Hall 2005). At some point however a balance has to be struck—although it is possible to see much of modern smoothing as an exercise in the geometry of reproducing kernel Hilbert spaces (Pearce and Wand 2006; Wahba 1990; Gu 2002), it is moot whether this is essential for the beginner.

Apart from mathematics, a reasonable background in computational science including programming skills has become essential, and knowledge of optimisation, algorithmics and complexity are increasingly needed to make progress in major applications.

The range of possible applications is so diverse that contact with more than a small subset seems impossible. Apart from the obvious, such as good communication skills and technical competence, key attributes for collaboration are a willingness to ask basic questions and to query assumptions, and respect for the insights given by different viewpoints. A strong element of project work, blending acquisition of statistical methodology with immersion in applications, helps to develop these.

7 Conclusion

These are exciting times for statistics and for statisticians. Important problems require solutions that take proper account of uncertainty, based on data of novel types, computational tools are readily available, and new mathematical models can be applied in imaginative ways. We are sometimes exhorted to ‘think globally, but act locally’, and in this spirit I congratulate the statisticians at the Università Ca’ Foscari on their beautiful new building, and look forward to the further contributions to be made from there.

References

Abramovich F, Sapatinas T, Silverman BW (1998) Wavelet thresholding via a Bayesian approach. J Roy Stat Soc B 60:725–749

Ancey C, Davison AC, Böhm T, Jodeau M, Frey P (2008) Entrainment and motion of coarse particles in a shallow water stream down a steep slope. J Fluid Mech 595:83–114

Azzalini A (1983) Maximum likelihood estimation of order *m* for stationary stochastic processes. Biometrika 70:381–387

Barndorff-Nielsen OE, Cox DR (1994) Inference and asymptotics. Chapman & Hall, London

Barndorff-Nielsen OE, Mikosch T, Resnick SI (2001) Lévy processes: theory and applications. Birkhäuser Verlag, Basel

Barndorff-Nielsen OE, Gill RD, Jupp PE (2003) On quantum statistical inference (with discussion). J Roy Stat Soc B 65:775–816

Bellio R (1999) Likelihood Asymptotics: Applications in Biostatistics. PhD Thesis, Department of Statistical Science, University of Padova

Bellio R, Brazzale AR (1999) On the implementation of approximate conditional inference. Stat Appl 11:251–271

Bellio R, Brazzale AR (2001) A computer algebra package for approximate conditional inference. Stat Comput 11:17–24

Bellio R, Brazzale AR (2003) Higher-order asymptotics unleashed: Software design for nonlinear heteroscedastic models. J Computat Graphical Stat 12:682–697

Bhowmick D, Davison AC, Goldstein DR, Ruffieux Y (2006) A Laplace mixture model for the identification of differential expression in microarrays. Biostatistics 7:630–641

Bickel PJ, Klassen CAJ, Ritov Y, Wellner JA (1993) Efficient and adaptive estimation for semiparametric models. Johns Hopkins University Press, Baltimore

Böhm T, Ancey C, Frey P, Reboud J-L, Ducottet C (2004) Fluctuations of the solid discharge of gravity-driven particle flows in a turbulent stream. Phys Rev E 69:061307

Brazzale AR (1999) Approximate conditional inference in logistic and loglinear models. J Computat Graphical Stat 8:653–661

Brazzale AR (2000) Practical Small-Sample Parametric Inference. PhD Thesis, Department of Mathematics, Swiss Federal Institute of Technology, Lausanne

Brazzale AR, Davison AC, Reid N (2007) Applied asymptotics: case studies in small sample statistics. Cambridge University Press, Cambridge

Breiman L (2001) Statistical modeling: the two cultures (with discussion). Stat Sci 16:199–231

Bühlmann P, Hothorn T (2006) Boosting algorithms: regularization, prediction and model fitting. http://stat.ethz.ch/buhlmann/bibliog.html.

Castillo JD, López-Ratera A (2006) Saddlepoint approximation in exponential models with boundary points. Bernoulli 12:491–500

Chellappa R, Jain A (eds) (1993) Markov random fields: theory and application. Academic, New York

Clifford P (1990) Markov random fields in statistics. In: Grimmett GR, Welsh DJA (eds) Disorder in physical systems: a volume in honour of John M. Hammersley. Clarendon Press, Oxford. pp 19–32

Cox DR, Isham VS (1988) A simple spatial-temporal model of rainfall. Proc Roy Soc Lond A 415:317–328

Cox DR, Reid N (2004) A note on pseudolikelihood constructed from marginal densities. Biometrika 91:211–221

Davison AC (2003) Statistical models. Cambridge University Press, Cambridge

Donoho DL, Johnstone IM (1994) Ideal spatial adaptation by wavelet shrinkage. Biometrika 81:425–455

Efron B (2003) The statistical century. In: Panaretos J (ed) Stochastic musings: perspectives from the pioneers of the late 20th century. Laurence Erlbaum, Florence. pp 31–46

Efron B, Hastie TJ, Johnstone IM, Tibshirani RJ (2004) Least angle regression (with discussion). Ann Stat 32:407–499

Fisher RA (1922) On the mathematical foundations of theoretical statistics. Philos Trans Roy Soc Lond A 222:309–368

Fisher RA (1925) Theory of statistical estimation. Proc Cambridge Philos Soc 22:700–725

Fisher RA (1934) Two new properties of mathematical likelihood. Proc Roy Soc Lond A 144:285–307

Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55:119–139

Gu C (2002) Smoothing spline ANOVA models. Springer, New York

Hall P (2005) On non-parametric statistical methods. In: Davison AC, Dodge Y, Wermuth N (eds) Celebrating statistics: papers in honour of Sir David Cox on his 80th birthday. Clarendon Press, Oxford. pp 137–150
Hand DJ (2006) Classifier technology and the illusion of progress (with discussion). Stat Sci 21:1–34
Hastie TJ, Tibshirani RJ, Friedman JH (2001) The elements of statistical learning: data mining, inference, and prediction. Springer, New York
Heard NA, Holmes CC, Stephens DA (2006) A quantitative study of gene regulation involved in the immune response of Anopheline mosquitoes: an application of Bayesian hierarchical clustering of curves. J Am Stat Assoc 101:18–29
Isham V (1981) An introduction to spatial point processes and Markov random fields. Int Stat Rev 49:21–43
Isham VS (2005) Stochastic models for epidemics. In: Davison AC, Dodge Y, Wermuth N (eds) Celebrating statistics: papers in honour of Sir David Cox on his 80th birthday. Clarendon Press, Oxford. pp 27–54
Johnstone IM, Silverman BW (2005) Empirical Bayes selection of wavelet thresholds. Ann Stat 33:1700–52
Kou SC, Xie XS, Liu JS (2005) Bayesian analysis of single-molecule experimental data (with discussion). Appl Stat 54:469–506
Lau JW, Green PJ (2008) Bayesian model based clustering procedures. Journal of Computational and Graphical Statistics p. (to appear)
Lindsay BG (1988) Composite likelihood methods. Contemporary Math 80:220–241
Lönnstedt I, Speed TP (2002) Replicated microarray data. Stat Sinica 12:31–46
McCulloch CE, Searle SR (2001) Generalized, linear, and mixed models. Wiley, New York
Messerli G, Partovi Nia V, Trevisan M, Kolbe A, Schauer N, Geigenberger P, Chen J, Davison AC, Fernie A, Zeeman SC (2007) Rapid classification of phenotypic mutants of *Arabidopsis* via metabolite fingerprinting. Plant Physiol 143:1484–1492
Murphy SA, van der Vaart AW (2000) On profile likelihood (with discussion). J Am Stat Assoc 95:449–485
Owen AB (2001) Empirical likelihood. Chapman & Hall/CRC, Boca Raton
Pace L, Salvan A (1997) Principles of statistical inference from a neo-fisherian perspective. World Scientific, Singapore
Panaretos VM (2006) The diffusion of radon shape. Adv Appl Prob 38:320–335
Panaretos VM (2007) Partially observed branching processes for stochastic epidemics. J Math Biol 54:645–668
Pearce ND, Wand MP (2006) Penalized splines and reproducing kernel methods. Am Stat 60:233–240
Porporato A, Rodríguez-Iturbe I (2005) Stochastic soil moisture dynamics and vegetation response. In: Davison AC, Dodge Y, Wermuth N (eds) Celebrating Statistics: papers in honour of Sir David Cox on his 80th birthday. Clarendon Press, Oxford. pp 55–72
Reid N (2003) Asymptotics and the theory of inference. Ann Stat 31:1695–1731
Rotnitzky A (2005) On semiparametric inference. In: Davison AC, Dodge Y, Wermuth N (eds) Celebrating statistics: papers in honour of Sir David Cox on his 80th birthday. Clarendon Press, Oxford. pp 115–136
Sartori N (2003) Modified profile likelihoods in models with stratum nuisance parameters. Biometrika 90:533–549
Severini TA (2000) Likelihood methods in statistics. Clarendon Press, Oxford
Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Roy Stat Soc B 58:267–288
Varin C (2008) On composite marginal likelihoods. Statistics (to appear)
Wahba G (1990) Spline models for observational data. CBMS-NSF regional conference series in applied mathematics. SIAM, Philadelphia