Sankhyā : The Indian Journal of Statistics
2008, Volume 70-B, Part 1, pp. 1-17

Applied Statistics and the Indianness of Indian Data

Jayanta K. Ghosh
Indian Statistical Institute, Kolkata, INDIA and Purdue University, West Lafayette, USA
Małgorzata Bogdan
Wrocław University of Technology, Wrocław, Poland
Tapas Samanta
Indian Statistical Institute, Kolkata, INDIA

Abstract

We provide an overview of current and possible future directions of Applied Statistics and explore choice of topics (e.g., Bioinformatics), paradigms and methodology. We also provide a tentative assessment of special problems arising in analysis of Indian data. Some priorities and need for special care in applied work are suggested, specially in the Indian context.

Keywords and phrases. Statistical paradigms, algorithms, cross-validation, model selection, Bioinformatics, Indian data

1 Introduction

It is a matter of great pleasure that *Sankhyā*, the Indian Journal of Statistics, will again be issued in two series – Series A for theory and Series B for applications. Hopefully, both series will welcome methodological papers, some of general interest and some targeted to particular interdisciplinary studies. One would expect Series A will have theoretical justification of proposed methods, whereas in Series B the stress will be on their computational development and application to real as well as simulated data. We feel sure both series will maintain high quality. This may require going beyond regular submissions and inviting special papers or papers for discussion.

Our focus in this note is on Series B. In the past Series B has been guided by two different but related goals. It is devoted to Applied Statistics in a general way as part of its commitment to international readers but it has also

played a historical role in nurturing Economic Statistics and Econometrics in the Indian context as well as critical analysis or interpretation of Indian statistical data in other areas. It is the second aspect which is captured by drawing attention to the Indianness of Indian data in Section 4, through a few nontechnical but important examples. Equally important are the general issues in Applied Statistics, reviewed so well in Cox (2007) on a similar occasion last year, and discussed in Section 2 of this article. We have also benefitted from Rao (1997) and Lehmann's reminiscences (Lehmann (2008)) on his own life, his intellectual development as a statistician and that of Classical Statistics.

2 Applied Statistics

Statistics is the science of understanding data and making decisions in the face of variability. A great variety of scientific or social questions which can be addressed by data analysis and the corresponding variety in the structure of data sets have always been the major source of the development of our field. An important role of Applied Statistics has become even more apparent in recent years, when a rapid development of computer technology as well as measurement techniques leads to construction of large data bases which need to be searched for patterns. The process of extracting information from such data bases requires a close cooperation between the specialists in the field, computer scientists and statisticians. While in the past Statistics was often used to test scientific hypotheses, now hypotheses are often formulated based on the analysis of such large data sets. These recent developments emphasize the importance of Statistics for the changing world and create new exciting possibilities and challenges for statisticians.

Luckily, the progress in computer technology aids us in constructing new powerful tools which help to address these new challenges. In particular, new computational power helped the development of nonparametric methods for density and function estimation, resampling techniques and model selection methods. Markov Chain Monte Carlo methods revolutionized Bayesian Statistics and greatly enhanced the range of its application. Moreover, Statistics has been enriched by a wide variety of algorithmic methods, which led to a new statistical paradigm – Statistical Learning or Machine Learning (see, e.g., Breiman (2001a) or Hastie et al. (2001)).

While neutral with respect to the chosen paradigm of an analyst, Cox (2007) makes useful comments on them. Most of the paper seems to assume

classical statistical analysis based on "explicit probability models", and some sort of testing of goodness of fit. But there is also reference to "Bayesian formulations" which "may allow the insertion of additional information which may open the route to bolder speculation ...". He also refers to descriptive or algorithmically specified methods of Breiman (2001a) and suggests that in this case also there should be some validation, however informal, based on the data. Typically, in such cases, validation is actually based on some form of prediction or cross-validation, both of which are so intuitively appealing and doable without any strong model assumptions.

The dynamic development of Statistics inevitably leads to some controversies between the supporters of classical and new approaches. A good example is Breiman (2001a) and the following discussion. Specifically, Breiman (2001a) has argued that Statistical Learning should become the only paradigm for Statistics, replacing model based approaches of Classical Statistics and Bayesian Analysis. Quite aptly, in a slightly different context, Ritov (2007) calls this a "post-modernist" criticism of the older paradigms. The other side of the case is presented well by Cox (2001) and Efron (2001) in their discussion of Breiman (2001a), and to some extent by Candes and Tao (2007) in their reply to Ritov and other discussants of their paper. Ritov (2007) and Candes and Tao (2007) present arguments on both sides very clearly in the context of choosing relevant covariates from among many in regression problems. Interestingly, both sides base their cases on real life applications.

The positive sign is that in spite of these controversies, the major paradigms within Statistics, namely, Classical, Bayesian and Machine Learning, currently coexist well. Nature may not permit a unified theory and real life examples often defy too rigid boundaries, formulations or solutions.

While the development of computer technology brings new opportunities for statisticians, it also creates some dangers. In the previous years a new statistical methodology had to satisfy certain standards, which were usually verified based on theoretical calculations. However, mathematical proofs often require simplified assumptions, which are rarely satisfied fully in the complicated real world. Unfortunately, the complexity of real life problems and the related complexity of new statistical methodology is sometimes used as an excuse for neglecting a proper verification of proposed methods. It often occurs that newly introduced statistical procedures are verified on just a few simulated or real data sets, which does not allow one to judge their general properties. It seems that there is an urgent need for developing new

standards for verification of statistical methodology, as well as for a thorough research on currently used methods of verification.

A good example is the popular and intuitive method of cross-validation, where a part of the data is used for validation of prediction based on the remaining part of the data. The properties of cross-validation turn out to be somewhat more strongly dependent on the proportion of observations selected for the test set than is commonly realized. As illustrated by theoretical results included, e.g., in Shao (1993), Yang (2007) or Chakrabarti and Ghosh (2007), the decision on this proportion should depend on the class of statistical models under consideration. In particular, according to our recent simulation studies (Bogdan, Ghosh and Żak-Szatkowska, unpublished), standard cross validation methods may lead to a substantial underestimation of the prediction error when applied together with all subsets model selection. Moreover, under sparsity (i.e., when only a small proportion of a large number of potential explanatory variables or parameters in a model is useful for explaining the response), the models selected by leave-one-out, four fold or two fold cross-validation are overly complicated and their prediction properties are significantly worse than the prediction properties of models chosen by other, more restrictive, model selection criteria (for some simulations illustrating the prediction properties of popular model selection criteria under sparsity see, e.g., Chakrabarti and Ghosh (2007) or Bogdan et al. (2008b)). As suggested by Yang (2007), "Further research on practical methods for properly choosing the data splitting proportion can be valuable for successful applications of cross validation."

Another danger lies in the fact that the fascination with new technical tools and statistical methods sometimes overshadows the purpose of the statistical analysis. Citing Cox (2007), "The literature of our field is ... inevitably dominated by detailed issues of methodology", while the success of statistical modeling "often hinges more on scientific sense than on technical mastery of complex methods". We would only add that not all applications are scientific, some would be related to government policy or evaluation of consequences of a policy; these would typically involve analysis of what is called Official Statistics. In such cases an understanding of the full context and common sense would replace "scientific sense".

Cox also points to the importance of "an intimate union between subject-matter and statistical aspects of an investigation", which is essential for defining objectives and choice of design and generally enriches analysis, conclusion and interpretation.

The importance of defining the objectives can be illustrated using examples from Testing or Model Selection. Many questions in these fields are ill-posed in the sense that not all relevant components of the problem are well-specified. Typically, the loss from wrong inference or prediction isn't specified well even though answers in a testing problem would depend radically on whether one takes a 0-1 loss, or a loss appropriate for prediction of future observations. Even in the case of a 0-1 loss, the presence or absence of an indifference zone (where one is willing to treat the different hypotheses or models as equally acceptable) or different designation of important alternatives to the model or hypothesis being tested can make a lot of difference. The Lindley paradox, which seems to indicate a fundamental conflict between Bayesian and classical testing of a sharp null hypothesis, is at least partly due to different choices of important alternatives, Ghosh et al. (2005). These issues also play a crucial role in sensitivity (Rubin (1971)). In regression problems, even the "future" to be predicted is ill-posed with many possible valid choices, see, for example, Barbieri and Berger (2004). In model selection, BIC is a good rule for low dimensional nested models with a 0-1 loss – i.e., one assumes the true model is in the model space and the loss is zero if and only if the true model is selected. On the other hand, AIC is a good model selection rule for high dimensional prediction problems in function estimation when the true model is the most complex model or not in the model space. Their penalty for complexity in a model arises in different ways and have rather different goals. In the case of all subsets model selection, both should be penalized further. For more details on some of these thorny issues see, e.g., Chakrabarti and Ghosh (2006, 2007). The choice of the loss function is also essential for the appropriate choice of the regularization parameter for LASSO (see, e.g., Meinshausen and Bühlman (2006)). In all such ill-posed problems the analyst needs to state clearly not only his or her choice but also the reasons.

3 Some Emerging Areas in Applied Statistics

Molecular Biology and Genetics are among the fields which in recent years had the largest impact on the development of Statistics. The rapid progress in the technology and the growing mass of data which need to be analyzed helped to establish Bioinformatics as a separate discipline and statisticians play an important role in constructing the set of tools for this field of research. Probably the best known example is the analysis of microarrays, which allows us to simultaneously investigate the expression of

thousands of genes. Microarrays triggered an explosion of new statistical concepts and methods related to high dimensional multiple testing, clustering, classification and dimension reduction (see, e.g., Wit and McClure (2004), Efron et al. (2001), Efron (2003) and Storey (2007)).

QTL mapping, whose aim is locating genes responsible for quantitative traits, is a less well-known area of Statistical Genetics. However, research in this field has helped to identify problems with standard model selection methods when applied to sparse multidimensional regression (Broman and Speed (2002), Bogdan et al. (2004, 2008a,b)). The problem of high dimensionality becomes even more apparent when locating expression-level QTL, i.e. genetic loci influencing the expression of other genes (West et al. (2007)). Researchers working in this area combine the genotype information with the information obtained from microarray studies. Similar problems of high dimensional model selection and multiple testing need to be addressed when designing appropriate tools for the quickly developing field of association mapping. Here the important genes are identified based on the general populations and the map of many thousands or even millions of genetic markers. Another interesting area related to association or QTL mapping is searching for high order gene interactions. The research in this field gave birth to a new promising statistical methodology called "logic regression" (see Ruczinski et al. (2003) or Schwender and Ickstadt (2008)).

One of the most important and quickly developing fields of Bioinformatics is Proteomics, the discipline dealing with the large-scale study of proteins. Proteomics is used, e.g., to identify biomarkers for disease diagnostics and to identify new drugs. High dimensional multiple testing, model selection and classification problems related to the analysis of protein mixtures by the mass-spectrometry are even more challenging than the related problems for microarrays (see, e.g., Nesvizhskii et al. (2007)). The need of cooperation between biologists and statisticians already at the level of the study design is well recognized within this field (see, e.g., Riter et al. (2005)).

Important problems in analyzing all types of Bioinformatics data include classification of individuals or plants into different groups, e.g. corresponding to a disease status, and a selection of variables, often called biomarkers, that help in classification. Such classification, as well as classification of genes according to their possible effects on different types of tumors have already led to major improvements in treatments. As mentioned above, the statistical analysis of data in Bioinformatics poses challenging technical problems

due to a large dimension and a relatively small sample size. Modern methods from Statistical Learning, like bagging, boosting or random forests, have been particularly useful for improving accuracy of classification because they are not model based. Rather they rely on generating additional pseudo data based on some form of perturbations of the learning data set and aggregating the corresponding predictors based on some form of voting (see e.g. Schapire (1990), Freund (1995), Breiman (1996, 1998, 2001b)). Some successful applications of bagging, boosting and random forests in Bioinformatics and Biometrics are presented e.g. in Long and Vega (2003), Huang et al. (2004), Seligson et. al (2005), Park et al. (2007), Politis (2008) or Schwender and Ickstadt (2008).

Another challenging task of Bioinformatics is discovering gene regulatory networks, i.e., a collection of genes which together regulate some function. Identification of key elements of such networks is important for understanding biological processes and may lead to the development of new drugs. Regulatory networks can be represented by graphs and are a natural field of application of graphical models. In recent years graphical models of conditional probability have made Bayesian Networks and Expert Systems a popular diagnostic and decision making tool, see, e.g., Cowell et al. (1999). Graphical models of dependency have also been applied successfully to classical multivariate analysis, Cox and Wermuth (1996) and Andersson and Perlman (1998). These methods of graphical modeling make use of graph theory to decompose a complex graph representing a joint distribution into a more easily computable product of suitable conditional and marginal distributions. Pioneering contributions in this area were made by Pearl, whose contributions have also led to inference about causes, Pearl (1988). Attempts to discover regulatory networks based on microarray data led to the development of new methods of modeling sparse covariance structures and construction of sparse Gaussian graphical models (see e.g Dobra et al. (2003)). For more theoretical work on sparse graphical models within the framework of classical model selection see Meinshausen and Bühlmann (2006). Sophisticated methods of data analysis are also required to combine the information from microarray studies, DNA sequence data and chromatin immunoprecipitation (ChIP-chip) experiments (see, e.g., Conlon et al. (2003) or Keles (2007)).

As the big pharmaceutical industries enter the Indian market and start research through microarrays or conduct clinical trials, there will be new possibilities of a grand unification of Bioinformatics, clinical trials and classical

survival analysis for life threatening diseases like cancer. However, as discussed in Sen (2008), there still remain serious statistical issues, which need to be resolved before the potential of Bioinformatics can be fully exploited in medicine and clinical sciences.

Unfortunately, there may also be some abuse of Statistics and unethical treatment of the vulnerable human beings who will readily be associated with clinical trials. Hopefully this is not inevitable and can be avoided by a vigilant scientific community.

This short review of developments in Bioinformatics is by no means complete. Our main goal was to illustrate a large variety of serious statistical problems which arise when dealing with modern biological data. Also, Bioinformatics is just one example of the quickly developing fields of science or technology that are a source of new problems. Similar, and sometimes even more challenging, statistical questions are posed by the data of modern astrophysics (see, e.g., Genovese et al. (2004)), or by the analysis of results produced by some diagnostic tools in medicine, like, e.g., functional magnetic resonance imaging (see, e.g., Vedel Jensen and Thorarinsdottir (2007)).

The quick progress of technology creates a need for the development of tools which would allow integration of information collected from different data sources. This seems to be a natural field of application of Bayesian methodology, with prior distributions incorporating existing knowledge. However, the task of eliciting prior distributions is by no means easy and requires a close cooperation and understanding between cognitive scientists, the specialists in the field and statisticians (see, e.g., O'Hagan (1998) or Oakley and O'Hagan (2007)).

Somewhat contrary to these positive developments and future possibilities, a recent panel discussion in Technometrics (Steinberg et al. (2008)) sounds a cautionary note about diminishing importance of applications of Statistics in many manufacturing industries. This is an area where samples are still small or moderate, risks are high and major statistical breakthroughs, after Taguchi designs, have been rare. Taguchi's innovative application of Mahalanobis D^2 (Taguchi and Rajesh (2000)) or data mining of the large data bases of these industries remain unutilized. Six sigma has been important but it is not a statistical methodological innovation. The SQC&OR unit of ISI had shown in the late eighties that a form of exploratory analysis of such a large data base is very promising. But it was not developed into an easily applicable algorithm. Computer experiments, see, e.g., Bayarri et

al. (2007), also seem a promising open area. We would also draw attention to Dasgupta and Mandal (2008).

4 Indianness of Indian Data and Problems

The statistical problems of India can often be rather different from apparently similar problems in countries like the U.S. or U.K. They can be different for historical reasons, or entirely different social norms, or just because of different stages of development. We call this rich variety of reasons the "Indianness" of the problems. We would illustrate these ideas with several important problems, most of which are contemporary but a few are from the history of the Indian Statistical Institute (ISI). We begin with the latter.

The Indian revenue system is very old. It dates back to the time of Chandragupta Maurya and Chanakya (about 300 B.C.). It reached a new stable form by the time of the Moghul Emperor Akbar, which was inherited by the British empire in India. These facts lay at the roots of an important methodological dispute between Professor P.C. Mahalanobis and Professor P.V. Sukhatme, one representing the ISI and the other the Indian Agricultural Statistics Research Institute (IASRI).

Mahalanobis had a general philosophy that in a developing country, one of the best ways of collecting data was through sample surveys, carried out by well-trained field investigators. His confidence was based on the crop surveys in the nineteen thirties in undivided Bengal, at least one of which is often cited. This is the jute survey, where the survey estimate and the complete census estimates were later compared with a more reliable estimate based on trade statistics. The survey estimate was better.

In the early fifties this view led to a dispute with Sukhatme who thought crop estimates based on estimates by patwaris (village accountants) could be just as good (Ghosh et al. (1999)).

Adhikari (1990) points out in a paper on Sociology of Science that at least part of the different approaches is due to the very different revenue collection systems in most of India, including Maharastra and that in areas like undivided Bengal where the British introduced a new administrative arrangement for revenue collection which abolished the role of patwaris. It seems both Mahalanobis and Sukhatme were unaware of these historical differences.

A much sadder example follows. At a convocation in March 2006, the Governor of West Bengal began his speech by quoting from newspapers a couple of days ago. The headlines of the newspapers were about a trivial fashion show and parade of models in Mumbai. None but one carried the news that suicide of farmers in India continued on its rising spiral and had crossed a benchmark. The fascination with trivia seems to be a cultural price for India's unprecedented economic growth, but what can explain the indifference of a democracy to the death of peasants? We haven't seen any serious study either in the media or elsewhere in the Government reports before 2008. The recent waiver of loan to farmers by the Government is a good step but not a substitute for a sustainable strategy. Perhaps the farmers need to be made aware of new risks and uncertainties and financial instruments to reduce uncertainty, elimination of middlemen and some support in a really bad year. All governments in all countries do that. An economist in Chennai has made three important points – that the deaths are taking place outside the Gangetic belt, that the crops involved are not the cereals, which still have a government support system and thirdly, the subsidies to the farmers have apparently declined during the current economic growth. Recently, there have been suicides even in the Gangetic belt caused by a surplus of crop and resulting price crash.

Very recently, the India Development Report 2008, published by the Indira Gandhi Institute of Development Research, provides a realistic picture of both India's rapid economic growth and the condition of extreme stress of about half the people of India, especially its agricultural communities, which is possibly partly induced by the economic reform. Reddy and Mishra (2008) note heavy dependence of farmers on informal sources of finance and a steep decline of the share of credit to agriculture from institutional sources. Interest rates charged by informal sources are quite high compared to institutional credit and are not affordable due to low productivity levels. Growing production and marketing risks, an institutional vacuum and a steep increase in the cost of farming have led to severe stress on Indian farmers. This is evident from the large number of farmer suicides in the country. A recent report of the UN Economic and Social Commission for Asia and the Pacific has pointed to 86,922 suicide deaths of Indian farmers during 2001-05 (The Statesman, Kolkata, 30 March, 2008, p.7).

Unfortunately, there are basic uncertainties even about the deaths. Reddy and Mishra (2008) look into the available sources of data and their limitations. The main official source of data on suicide deaths is the police record provided by the National Crime Records Bureau (NCRB), Ministry of Home

Affairs, which is limited to routine reporting of suicides and may not fully reflect the stress on farmers. It is also likely to be under-reported for several reasons, see Reddy and Mishra (2008, p.47). Profession-wise NCRB data are available only from 1995. Most of the suicides are among male farmers and the available data show that suicide mortality rate (SMR) for male farmers has been rising steeply since 1995, while SMR for male non-farmers has been more or less stable. The second source of data is press reports based on farmer suicides which may have the bias of linking all suicide deaths by farmers to farming related causes. The third source is the official data made available by the state government. It suffers from underestimation as it is directly linked to compensation paid by the government. Many state governments are still in a state of denial. They deny that the deaths are caused by crop failure or crashing prices. As noted by Reddy and Mishra (2008), the number of farmer suicides reported by the state governments are much lower than even the official NCRB data.

These studies notwithstanding, a rigorous statistical analysis of the problem of farmer suicides still remains to be done.

Estimation of poverty in India has attracted many experts from India and abroad. We are familiar with Tendulkar's work on reduction in poverty in the early days of liberalization as well as the work of Angus Deaton and Jean Dreze during the later stages of liberalization (Tendulkar (1996) and Deaton and Dreze (2002)). Both were of the opinion that poverty has diminished, confirming our general impression.

But there is a puzzling aspect of these measuements by different groups of economists. Nearly all of them search for a suitable price index for the rural and urban poor to deflate the current prices of cereals. Nearly all of them come up with their own favourite version of the price index. Given such a lack of standardization in an important component of the estimate, how can we really depend on these estimates? Again, we see no discussion in this lack of standardization or simple remedies. For example, the NSS can collect appropriate price statistics and construct a price index series.

Provision of productive employment, employment security and favourable working conditions for all in the labour force are essential for reduction of poverty. We do not see studies of poverty borrowing strength from related data on employment. As economists often point out, agriculture is still the dominant occupation in India with a high share in the total workforce but with a declining share in the gross domestic product (GDP). The share of

workforce in agriculture is about 56% to 57% but the contribution of agriculture to GDP is only about 20% (Radhakrishna and Chandrasekhar (2008)). This is one of the main obstacles to reduction of poverty. After agriculture the biggest source of employment is the so called informal sector, where people are mostly self-employed and operate an auto-rickshaw or run a small shop. Removal of poverty would seem to need some careful study of these sectors along with incidence of poverty. Shouldn't the study of poverty be more meaningful if combined with a study of employment of the poor and their income? Also needed is a study of the middle class prosperity and the fast growing service sector. We know so little about this new catalyst in India's economic growth.

We end with a remark of Rajeeva Karandikar, a distinguished probabilist and well-known election forecaster, made in connection with election forecasts in India, that methodologies developed in the West are just not applicable in India, that one must recognize these differences and be ready to develop new tools that do justice to Indian reality. In the case of election forecasts, one must cope with great swings in many more constituencies than in the West and also swings in between the forecast and election.

Applied Statistics as practiced in India still suffers from unrecognized bias in data collection as well as lack of estimates of uncertainty arising from precise probability models or sampling. Uncertainty is rarely measured. Assumptions for common models developed elsewhere are not always examined carefully. Hypotheses are rarely tested formally. Possibility of a question being ill-posed is not carefully examined. Usually the discussions are by subject matter experts rather than professional statisticians or econometricians. They are examples of analytical discussion using empirical data, but rarely venturing into formal statistical inference. Hopefully, editors of Series B will choose judiciously from them and, if necessary, help quantify uncertainties in the conclusion. This would help nurture a school of good, modern applied statistical work relevant for Indian problems.

India is a great democracy, and a most improbable success in the eyes of the West – a country that is almost a subcontinent, as diverse as Europe, whose unities are more to be found in its historical tradition of pluralism and tolerance than a common religion or language.

A sign of the self-confidence of the young but strong Indian democracy is the recently passed Right to Information Act. One of us has argued elsewhere (Ghosh (2000)) that Statistics goes well with democracy while dictators fear

Statistics as much as they fear freedom. But our responsibilities as statisticians analyzing Indian data require that we recognize India's great diversity, its complex societies, and non-western norms and values which make the statistical tasks specially challenging.

Statistical articles on incomplete or biased data obtained in difficult circumstances need not mean lowering standards of the journal. We are reminded of, e.g., the Presidential address in JASA, March 2007 (Scheuren (2007)).

5 Concluding Remarks

The progress in science and technology will inevitably continue and so will the need for the development of new, reliable methods of data analysis. The recognition of an ever increasing role of Applied Statistics is reflected by an increasing number of journals publishing research in this area. We believe that the long tradition and the quality of *Sankhyā* will allow Series B to successfully join this effort and become an important area of presentation of applied statistical methodology. Hopefully Series B will encourage some interaction between statisticians and computer scientists in problems of common interest, like classification, pattern recognition and feature or variable selection.

Acknowledgements. We thank Professor P.K. Sen for inviting us to write this article and Professor Ayanendranath Basu for sharing some of his views with us.

References

ADHIKARI, B.P. (1990). Social construction of the statistical estimation of crop yield. Paper presented at the *XII World Congress of Sociology of the International Sociological Association*, Madrid, Spain.

ANDERSSON, S.A. and PERLMAN, M.D. (1998). Normal linear regression models with recursive graphical Markov structure. *J. Multivariate Anal.* **66**, 133–187.

ANTONIAK, C. E. (1974). Mixtures of Dirichlet Processes With Applications to Nonparametric Problems., *Ann. Statist,*, **2**, 1152–1174.

BARBIERI, M.M. and BERGER, J.O. (2004). Optimal predictive model selection. *Ann. Statist.* **32**, 870-897.

BAYARRI, M. J., WALSH, D., BERGER, J. O, CAFEO, J., GARCIA-DONATO, G., LIU, F., PALOMO, J., PARTHASARATHY, R. J., PAULO R. and SACKS J. (2007). Computer model validation with functional output. *Ann. Statist.* **35**, 1874-1906.

Bogdan, M., Ghosh, J. K. and Doerge, R. W. (2004). Modifying the Schwarz Bayesian Information Criterion to Locate Multiple Interacting Quantitative Trait Loci. *Genetics*, **167**, 989-999.

Bogdan, M., Frommlet, F., Biecek, P., Cheng, R., Ghosh, J. K. and Doerge, R. W. (2008a). Extending the Modified Bayesian Information Criterion (mBIC) to dense markers and multiple interval mapping, *Biometrics*, doi: 10.1111/j.1541-0420.2008.00989.x.

Bogdan, M., Ghosh, J.K. and Żak-Szatkowska, M. (2008b). Selecting explanatory variables with the modified version of Bayesian Information Criterion, *Quality and Reliability Engineering International*, **24**, 627-641.

Breiman, L. (1996). Bagging predictor. *Machine Learning* **24**, 123-140.

Breiman, L. (1998). Arcing classifiers. *Ann. Statist.*, **26**, 801-849.

Breiman, L. (2001a). Statistical modeling: The two cultures (with discussion). *Statist. Science.*, **16**, 199-231.

Breiman, L. (2001b). Random Forests. *Machine Learning* **45**, 5-32.

Broman, K. W. and Speed, T. P. (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. *J. Roy. Statist. Soc. Ser. B* **64** (4), 641–656.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n (with discussion). *Ann. Statist.* **35**, 2313-2404.

Chakrabarti, A. and Ghosh, J.K. (2006). Optimality of AIC in inference about Brownian Motion. *Ann. Inst. Statist. Math.* **58**, 1-20.

Chakrabarti, A. and Ghosh, J.K. (2007). Some aspects of Bayesian model selection for prediction (with discussion). In: Bernardo, J.M. et al. (eds) *Bayesian Statistics*, **8**.

Conlon E. M., Liu X. S., Lieb J. D, Liu J. S. (2003). Integrating regulatory motif discovery and genome-wide expression analysis, *Proc. Nat. Acad. Sci. USA* **100**, 3339-3344.

Cowell, R. G., Dawid, A. P., Lauritzen S. L. and Spiegelhalter D. J. (1999). *Probabilistic Networks and Expert Systems.* Springer-Verlag, Berlin-Heidelberg-New York.

Cox, D.R. (2001). Discussion of Leo Breiman's paper "Statistical modeling: The two cultures" *Statist. Sci.*, **16**, 216-217.

Cox, D.R. (2007). Applied statistics: A review. *Ann. Appl. Statist.*, **1**, 1-16.

Cox, D.R. and Wermuth, N. (1996). *Multivariate Dependencies: Models, Analysis and Interpretation.* Chapman and Hall, London.

Dasgupta, T. and Mandal, A. (2008). Estimation of process parameters to determine the optimum diagnosis interval for control of defective items. *Technometrics* **50**, 167-181.

Deaton, A. and Dreze, J. (2002). Poverty and inequality in India – A re-examination. *Economic and Political Weekly*, September 7, 3729-3748.

Dobra, A., Jones B., Hans C., Nevins J. and West, M. (2003). Sparse graphical models for exploring gene expression data. *J. Mult. Anal.*, **90**: 196-212.

Efron, B. (2001). Discussion of Leo Breiman's paper "Statistical modeling: The two cultures". *Statist. Sci.*, **16**, 218-219.

EFRON, B. (2003). Robbins, Empirical Bayes and Microarrays. *Ann. Statist.*, **31**, 366-378.

EFRON, B., TIBSHIRANI, R., STOREY, J.D. and TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.*, **96**, 1151-1160.

FREUND, Y. (1995). Boosting a weak learning algorithm by majority. *Information and Computation*, **121**, 256-285.

GENOVESE, C. R., MILLER, C. J., NICHOL, R. C., ARJUNWADKAR, M., WASSERMAN, L. (2004). Nonparametric Inference for the Cosmic Microwave Background, *Statistical Science*, **19**, 308-321.

GHOSH, J.K. (2000). Statistical science – An emerging paradigm for inference, decision and policy. Sukhatme Memorial Lecture (unpublished).

GHOSH, J.K., MAITI, P., RAO, T.J. and SINHA, B.K. (1999). Evolution of Statistics in India. *International Statistical Review*, **67**, 13-34.

GHOSH, J.K., PURKAYASTHA, S. and SAMANTA, T. (2005). Role of P-values and other measures of evidence in Bayesian analysis. In: Dey, D.K. and Rao, C.R. (eds) *Handbook of Statistics* **25**, Bayesian Thinking: Modeling and Computation, 151-170.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001) *The elements of statistical learning: Data Mining, Inference, and Prediction.* Springer: New York.

HUANG, Y., CAI, J., JI, L. and LI, Y. (2004). Classifying G-protein coupled receptors with bagging classification tree, *Computational Biology and Chemistry*, **28**, 275-280.

KELES, S. (2007). Mixture modeling for genome-wide localization of transcription factors. *Biometrics*, **63**, 10-21.

LEHMANN, E.L. (2008). *Reminiscences of a Statistician: The Company I Kept.*, Springer, New York.

LONG, P. M. and VEGA, V. B. (2003). Boosting and microarray data. *Machine Learning*, **52** :31-44.

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.*, **34**, 1436-1462.

NESVIZHSKII, A. I., VITEK, O. and AEBERSOLD, R. (2007) Analysis and validation of proteomic data generated by tandem mass spectrometry *Nature Methods*, **4**, 787 – 797.

OAKLEY, J. E. and O'HAGAN, A. (2007). Uncertainty in prior elicitations: a nonparametric approach. *Biometrika*, **94**, 427-441.

O'HAGAN, A. (1998). Eliciting expert beliefs in substantial practical applications. *The Statistician*, **47**, 21-35.

PARK, J., WILBUR, J. D., GHOSH, J. K., NAKATSU, C. H. and ACKERMAN C. (2007). Selection of Binary Variables and Classification by Boosting. *Comm. Statist. - Simul. and Comput.*, **36**, 855 - 869.

PEARL, J. (1988), *Probabilistic Reasoning in Intelligent Systems*, Revised Second Printing, San Mateo, CA: Morgan Kaufmann.

POLITIS, D. N. (2008). Bagging Multiple Comparisons from Microarray Data. In *Bioinformatics Research and Applications*, *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 492-503.

Radhakrishna, R. and Chandrasekhar, S. (2008). Growth: achievements and distress. Overview of *India Development Report 2008* edited by Radhakrishna, R., Indira Gandhi Institute of Development Research. Oxford University Press, New Delhi.

Rao, C.R. (1997). Statistics and Truth: Putting Chance to Work. World Scientific, Singapore.

Reddy, D.N. and Mishra, S. (2008). Crisis in agriculture and rural distress in post-reform india. In: Radhakrishna, R. (ed) *India Development Report 2008*, 40-53, Indira Gandhi Institute of Development Research. Oxford University Press, New Delhi.

Riter, L. S., Vitek, O., Gooding, K. M., Hodge, B. D. and Julian, R. K. Jr. (2005). Statistical design of experiments as a tool in mass spectrometry. *Journal of Mass Spectrometry*, **40**, 565-579.

Ritov, Y. (2007). Discussion of "The Dantzig selector: Statistical estimation when p is much larger than n". *Ann. Statist.*, **35**, 2313-2404.

Rubin, H. (1971). A decision-theoretic approach to the problem of testing a null hypothesis. In *Statistical decision theory and related topics (Proc. Sympos., Purdue Univ., Lafayette, Ind., 1970)*, pp 103–108. Academic Press, New York.

Ruczinski, I., Kooperberg, C., LeBlanc, M. (2003). Logic Regression. *Journal of Computational and Graphical Statistics*, **12**, 475-511.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, **5**, 197-227.

Scheuren, F. (2007). The *Pro Bono* Statistician : Presidential address. *J. Amer. Statist. Assoc.*, **102**, 1-6.

Schwender, H. and Ickstadt, K. (2008). Identification of SNP interactions using logic regression. *Biostatistics*, **9**, 187-198.

Seligson, D.B., Horvath, S., Shi, T., Yu, H., Tze, S., Grunstein, M. and Kurdistani S. K. (2005). Global histone modification patterns predict risk of prostate cancer recurrence. *Nature*, **435**, 1262-1266.

Sen, P. K. (2008). Burden of bioinformatics in medical research: Statistical perspectives and controversies. *J. Statist. Plann. Inference*, **138**, 450-463.

Shao, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.*, **88**, 486-494.

Steinberg, D.M., Bisgaard, S., Doganaksoy, N., Fisher, N., Gunter, B., Hahn, G., Keller-McNulty, S., Kettenring, J., Meeker, W.Q., Montgomery, D.C. and Wu, C.F.J. (2008). The Future of Industrial Statistics: A Panel Discussion. *Technometrics* **50**, 103-127.

Storey, J. D. (2007). The optimal discovery procedure: A new approach to simultaneous significance testing. *J. Roy. Statist. Soc. Ser. B*, **69**, 347-368.

Taguchi, G. and Rajesh, J. (2000). New trends in multivariate diagnosis. *Sankhyā* (Ser. B), **62**, 233–248.

Tendulkar, S.D. (1996). Social welfare, social deprivation and economic growth: some reflections on the Indian experience. In: Chattopadhyay, M., Maiti, P. and Rakshit, M. (eds) *Planning and Economic Policy in India – Evaluation and lessons for the future*, 111-141, Sage Publications, New Delhi.

Vedel Jensen, E. B. and Thorarinsdottir, T. L. (2007). A Spatio-Temporal Model for Functional Magnetic Resonance Imaging Data - with a View to Resting State Networks, *Scand. J. Statist.*, **34**, 587-614.

West, M. A. L., Kim, K., Kliebenstein, D. J., van Leeuwen, H., Michelmore, R. W., Doerge, R. W. and St.Clair D. A. (2007). Global eQTL Mapping Reveals the Complex Genetic Architecture of Transcript Level Variation in Arabidopsis. *Genetics*, **175**, 1441-1450.

Wit, E., McClure J. (2004). *Statistics for Microarrays: Design, Analysis and Inference*, Wiley & Sons, Chichester, UK.

Yang, Y. (2007). Consistency of cross validation for comparing regression procedures. *Ann. Statist*, **35**, 2450-2473.

Jayanta K. Ghosh
Department of Statistics
Purdue University
West Lafayette, IN 47907, USA
E-mail: jayanta@isical.ac.in

Małgorzata Bogdan
Institute of Mathematics and Computer Science
Wrocław University of Technology
50-370 Wrocław, Poland
E-mail: Malgorzata.Bogdan@pwr.wroc.pl

Tapas Samanta
Applied Statistics Division
Indian Statistical Institute
203, B. T. Road
Kolkata 700108
INDIA
E-mail: tapas@isical.ac.in

Paper received July 2008; revised September 2008.