SciMedCentral

Annals of Biometrics & Biostatistics

Editorial

A Celebration of Statistics and its Applications

Wenjiang Fu*

Department of Epidemiology and Biostatistics, Michigan State University, USA

Corresponding author

Wenjiang Fu, Department of Epidemiology and Biostatistics, Michigan State University, USA, Email: fuw@msu.edu

Submitted: 31 October 2013

Accepted: 02 November 2013

Published: 04 November 2013

This year, statisticians worldwide are celebrating the International Year of Statistics 2013. Statistics, as a unique subject in data science, has grown enormously during the past fifty years, in particular the most recent twenty years when information technology advances rapidly. Today statistical applications cover a wide range of areas in quantitative sciences in which data are frequently collected, including biology, genetics and genome science, business, economics, finance, management, engineering, health, medicine, and sport science. Recent advances in information technology provide unprecedented opportunities for statisticians to collaborate with investigators in other fields, to make scientific discoveries using statistical modeling and analytical approaches, to solve real world problems in studies of climate change, health and environment, genome research, policymaking, social surveys, social networks, and to decipher complex data using powerful computational tools that would otherwise be impossible to understand.

The journal Annals of Biometrics and Biostatistics is born at this very exciting moment when the world is embracing big data and celebrating the International Year of Statistics. Launching a new journal in this exciting field, the editors and publisher would like to take the opportunity to present the most challenging statistical issues to the readers, to encourage publication of statistical innovations equipped with solid techniques and fascinating applications, and to ensure the best editorial service in manuscript submission and review. Specifically, the journal would like to encourage works in the following frameworks.

1) Articles that critically review current methods and those that develop novel methods are equally important and encouraged. Articles that challenge current methods with solid justification are particularly welcomed, in particular when current methods are not capable of handling data issues that have recently emerged.

2) Scientific controversies are often caused by misunderstandings and misinterpretations, and clarifications are always in high demand. Scientific debate is always encouraged to foster a scientific environment based on the principle that every opinions deserve to be published as long as the work is supported by mathematical statistics theory, equipped with solid statistical techniques with justification and follows stringent scientific logic.

In the first years, the journal would like to entertain the following research areas.

1) Variable selection and dimension reduction; 2) Sparsity and sparse estimation; 3) Computational methods for high dimensional data; 4) Modeling high throughput genome data (e.g. microarray, sequencing and other technology) ; 5) Modeling imaging and other high-dimensional and highly correlated data; 6) Modeling climate change data and other big data; and 7) Complex data analysis and seemingly unsolvable data problems.

In particular, we would like to emphasize the following topics, which have received great attention during the past twenty years.

VARIABLE SELECTION

Although variable selection has always been a primary task in statistical research, it has received greater attention than ever before since the Lasso method was developed [1], which started a new era of variable selection through a shrinkage and regularization approach with desirable statistical properties [2]. Current methods of variable selection through the penalty and regularization approach not only make it easy-to-implement using highly efficient computational methods, such as the LARS [3], but also make it a masterpiece of art [4-6]. For example, the oracle properties of variable selection [4] guarantees, that a variable selection procedure selects variables as if the true model is known. Methods for variable selection with extra high dimensional data are also available, making variable selection a good resource to many quantitative scientists [7].

SPARSITY AND SPARSE ESTIMATION

The advanced information technology generates many large data sets, such as internet streaming and social network data, genome data and stock market data, etc. One typical characteristic of these large data sets is high dimensionality, often of the order of tens of thousands or even millions. Because of the large size and high dimensionality of these data, a thorough understanding of the relationship among many quantities is extremely difficult and challenging. A common approach taken by many quantitative scientists is to assume a sparse data structure and further conduct statistical modeling based on the sparsity assumption [8-11]. Models using Dantzig selector [12], non-convex penalties [7] and others provide a promising approach to compressive sensing for sparsity and signal extraction from big data. These methods are expected to change the landscape of big data studies.

TWO EXAMPLES OF COMPLEX DATA AND SEEMINGLY UNSOLVABLE PROBLEMS

Some statistical problems are challenging not only because they are intriguing and difficult to model or estimate but also because they are confusing and seem to be unsolvable. Here we demonstrate with two examples and would like to invite more work in these areas.

Affymetrix microarray

The microarray technology has produced large amount of data in studies of genome of animals, humans and plants. One of the popular microarray technologies is the Affymetrix array, which uses perfect match and mismatch probe sequences to measure DNA or RNA sequence abundance using probe intensities and to annotate gene expressions and DNA structural variants, such as the single nucleotide polymorphisms (SNPs). By design, mismatch sequences are believed to produce low background intensity and perfect match sequences to produce high signal intensity. Hence a subtraction of the mismatch intensity from the perfect match intensity within each perfect match - mismatch sequence pair would yield a positive value, which would represent the true value of the genetic signal. However, it has been reported that among 30% of the perfect match-mismatch sequence pairs, the mismatch sequence produces higher intensity than the corresponding perfect match sequence [12-14]. Apparently, the previous assumption that perfect match sequence produces higher intensity than the corresponding mismatch is invalid. Although a number of studies have made similar observations, many investigators prefer to believe that mismatch sequence in the array design is of no practical use [15,16], even though it has been found later that such mismatch sequences can be very useful in annotating DNA structural variants and estimating DNA copy numbers [17] if probe sequences are properly modeled by studying nucleotide binding affinity. Such findings imply that one needs to study further and improve the models to make good use of data rather than wasting a major chunk of it because some models cannot produce meaningful results with the data. Currently, modeling nucleotide binding affinity still remains an active research topic and a thorough understanding of such mismatch phenomenon is expected to influence the design of other biotechnologies in the future [18].

Age-period-cohort models

For many years, biostatisticians, demographers and sociologists have been puzzled by a seemingly unsolvable statistical regression problem. In studies of disease incidence or mortality rate of chronic diseases (e.g. cancer, stroke, diabetes, etc) or social event rate (e.g. crime rate, illiteracy, belief in religion, long term disability) in a city or geographic region, age and birth cohort are believed to play a major role in the development of diseases or events due to aging (nature) and nurturing, respectively. Meanwhile, the effect of period (calendar year) is always of interest for the purpose of monitoring and policymaking for public health or safety, etc. However, since the age-year specific rate is tabulated in a rectangular table with rows being the consecutive age groups (5 year intervals, say) and columns being the consecutive periods (5 calendar year intervals, say), the regression of the disease rate or event rate on the fixed effects of age (row), period (column) and birth cohort (diagonal) encounters an identification problem simply because the regression design matrix is singular and has a rank of 1-less than its full value. The singular design matrix yields multiple estimators that have the same fitted value, but each of them presents a temporal trend in age effects, period effects, and cohort effects that vary from estimator to estimator, when plotted against age, calendar year and birth year. This is well known as the identifiability problem in the age-period-cohort (APC) models, which has been studied extensively ever since 1970s. Given the existence of multiple estimators, it has been unanimously agreed upon that it is extremely difficult, if not impossible, to identify a special estimable function, which would completely determine the temporal trend and provide unbiased parameter estimation, thus solve the identifiability problem [19]. However, recent works in this area have taken new approaches and discovered promising routes, potentially leading to a resolution [20-23]. These novel approaches have a common characteristic in that they study the problem from a different angle using modern statistical techniques, such as smoothing method, principal component analysis, and partial least-squares method. Although these methods have not decisively resolved the identifiability problem, a recent surge of publications using the intrinsic estimator method [21] demonstrates that this method has been accepted by more and more researchers in this community. Although opponents have presented evidence against this method, a fair debate about the identifiability problem will surely help to clarify confusions and misinterpretations over the last thirty years. Since more and more applications have been found with the APC models, including studies in epidemiology, public health, sociology, demography, economics, and finance, a final resolution of the identifiability problem will be welcomed by the entire scientific community, which can be achieved only through public and candid debate following stringent scientific logic with support by mathematical statistics theory and well developed statistical techniques.

The editors and publisher truly hope that the Annals of Biometrics and Biostatistics will foster a scientific environment for statisticians and quantitative scientists to develop, test and validate novel statistical models, methods and frameworks, to address challenging statistical issues and to provide statistical support to the quantitative science community.

REFERENCES

1. Tibshirani R. Regression Shrinkage and Selection via the Lasso. J Roy Stat Soc. 1996; 58: 267–288.
2. Knight K, Fu WJ. Asymptotics of Lasso-type estimators. Ann Stat. 2000; 28: 1356-1378.
3. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Amer Statist Assoc. 2001; 96: 1348-1360.
4. Zou H. The Adaptive Lasso and its oracle properties. J Amer Statist Assoc. 2006; 101: 1418-1429.
5. Zhang CH. Nearly unbiased variable selection under the minimax concave penalty. Ann Statist. 2010; 38: 894-942.
6. Efron N, Morgan PB. Practitioner influence on contact lens prescribing in the UK. Cont Lens Anterior Eye. 2009; 32: 185-186.
7. Bickel P, Levina E. Regularized estimation of large covariance matrices.

Ann Statist. 2008; 36: 199-227.

8. Bickel P, Levina E. Covariance regularization by thresholding. Ann Statist. 2008; 36: 2577-2604.

9. El Karoui N. Operator norm consistent estimation of large-dimensional sparse covariance matrices. Ann Statist. 2008; 36: 2717-2756.

10. Lam C, Fan J. PROFILE-KERNEL LIKELIHOOD INFERENCE WITH DIVERGING NUMBER OF PARAMETERS. Ann Stat. 2008; 36: 2232-2260.

11. Candès E, Tao T. The Dantzig selector: statistical estimation when *p* is much larger than *n*. Ann Statist. 2007; 35: 2313-2351.

12. Naef F, Lim DA, Patil N, Magnasco M. DNA hybridization to mismatched templates: a chip study. Phys Rev E Stat Nonlin Soft Matter Phys. 2002; 65: 040902.

13. Bolstad BM, Irizarry RA, Gautier L, Wu Z. Preprocessing High-density Oligonucleotide Arrays. In Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Ed. Gentleman R, Carey V, Huber W, Irizarry RA, Dudoit S. Springer 2005. New York.

14. Seringhaus M, Rozowsky J, Royce T, Nagalakshmi U, Jee J, Snyder M, et al. Mismatch oligonucleotides in human and yeast: guidelines for probe design on tiling microarrays. BMC Genomics. 2008; 9: 635.

15. Rabbee N, Speed TP. A genotype calling algorithm for affymetrix SNP arrays. Bioinformatics. 2006; 22: 7-12.

16. Carvalho B, Bengtsson H, Speed TP, Irizarry RA. Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. Biostatistics. 2007; 8: 485-499.

17. Wan L, Sun K, Ding Q, Cui Y, Li M, Wen Y, et al. Hybridization modeling of oligonucleotide SNP arrays for accurate DNA copy number estimation. Nucleic Acids Res. 2009; 37: e117.

18. Alves NJ, Stimple SD, Handlogten MW, Ashley JD, Kiziltepe T, Bilgicer B. Small-molecule-based affinity chromatography method for antibody purification via nucleotide binding site targeting. Anal Chem. 2012; 84: 7721-7728.

19. Kupper LL, Janis JM, Karmous A, Greenberg BG. Statistical age-period-cohort analysis: a review and critique. J Chronic Dis. 1985; 38: 811-830.

20. Heuer C. Modeling of time trends and interactions in vital rates using restricted regression splines. Biometrics. 1997; 53: 161-177.

21. Fu WJ. Ridge estimator in singular design with application to age-period-cohort analysis of disease rates. Comm Stat - Theo Meth. 2000; 29: 263-278.

22. Fu WJ. A smoothing cohort model in age-period-cohort analysis with applications to homicide arrest rates and lung cancer mortality rates. Soc Meth Res. 2008; 36: 327-361.

23. Tu YK, Davey Smith G, Gilthorpe MS. A new approach to age-period-cohort analysis using partial least squares regression: the trend in blood pressure in the Glasgow Alumni cohort. PLoS One. 2011; 6: e19401.

Cite this article

Fu W (2013) A Celebration of Statistics and its Applications. Ann Biom Biostat 1(2): 1008.