

New Developments in Biostatistics and Bioinformatics

Editors

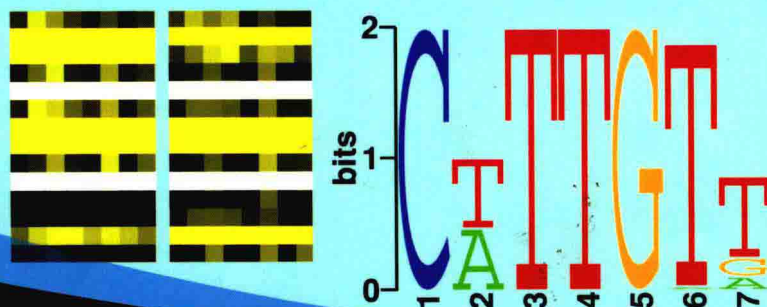
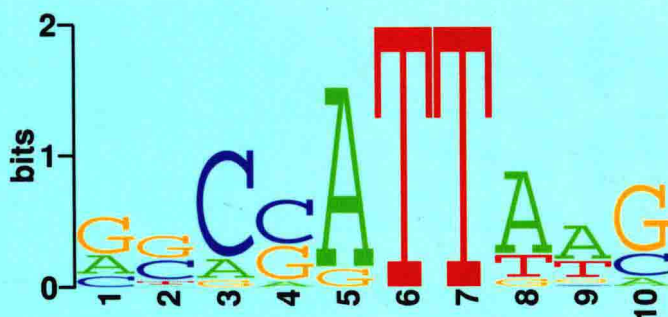
Jianqing Fan

Xihong Lin

Jun S. Liu

Volume 1

Frontiers of Statistics



New Developments in Biostatistics and Bioinformatics

Editors

Jianqing Fan

Princeton University, USA

Xihong Lin

Harvard University, USA

Jun S. Liu

Harvard University, USA

Volume 1

Frontiers of Statistics



Higher Education Press



World Scientific

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI

Published by

World Scientific Publishing Co. Pte. Ltd.

5 Toh Tuck Link, Singapore 596224

USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

NEW DYNAMICS BETWEEN CHINA AND JAPAN IN ASIA

How to Build the Future from the Past?

Copyright © 2010 by World Scientific Publishing Co. Pte. Ltd.

All rights reserved. This book, or parts thereof, may not be reproduced in any form or by any means, electronic or mechanical, including photocopying, recording or any information storage and retrieval system now known or to be invented, without written permission from the Publisher.

For photocopying of material in this volume, please pay a copying fee through the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, USA. In this case permission to photocopy is not required from the publisher.

ISBN-13 978-981-4313-66-7

ISBN-10 981-4313-66-1

Typeset by Stallion Press

Email: enquiries@stallionpress.com

Printed in Singapore.

FRONTIERS OF STATISTICS

FRONTIERS OF STATISTICS

The book series provides an overview on the new developments in the frontiers of statistics. It aims at promoting statistical research that has high societal impacts and offers a concise overview on the recent developments on a topic in the frontiers of statistics. The books in the series intend to give graduate students and new researchers an idea where the frontiers of statistics are, to learn common techniques in use so that they can advance the fields via developing new techniques and new results. It is also intended to provide extensive references so that researchers can follow the threads to learn more comprehensively what the literature is, to conduct their own research, and to cruise and lead the tidal waves on the frontiers of statistics.

SERIES EDITORS

Jianqing Fan

Frederick L. Moore'18 Professor of Finance,
Director of Committee of Statistical Studies,
Department of Operation Research and
Financial Engineering,
Princeton University, NJ 08544, USA.

Zhiming Ma

Academy of Math and Systems Science,
Institute of Applied Mathematics,
Chinese Academy of Science,
No.55, Zhong-guan-cun East Road,
Beijing 100190, China.

EDITORIAL BOARD

Tony Cai, University of Pennsylvania

Min Chen, Chinese Academy of Science

Zhi Geng, Peking University

Xuming He, University of Illinois at Urbana-Champaign

Xihong Lin, Harvard University

Jun Liu, Harvard University

Xiao-li Meng, Harvard University

Jeff Wu, Georgia Institute of Technology

Heping Zhang, Yale University

Jianqing Fan

Department of Operation Research and
Financial Engineering
Princeton University

Jun Liu

Department of Statistics
Harvard University

Xihong Lin

Department of Biostatistics of the School of
Public Health
Harvard University

Copyright © 2009 by

Higher Education Press

4 Dewai Dajie, 100011, Beijing, P.R. China and

World Scientific Publishing Co Pte Ltd

5 Toh Tuch Link, Singapore 596224

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage and retrieval system, without permission in writing from the Publisher.

ISBN 13 : 978-981-283-743-1

ISBN 10 : 981-283-743-4

ISSN 1793-8155

Printed in P. R. China

Preface

The first eight years of the twenty-first century has witnessed the explosion of data collection, with relatively low costs. Data with curves, images and movies are frequently collected in molecular biology, health science, engineering, geology, climatology, economics, finance, and humanities. For example, in biomedical research, MRI, fMRI, microarray, and proteomics data are frequently collected for each subject, involving hundreds of subjects; in molecular biology, massive sequencing data are becoming rapidly available; in natural resource discovery and agriculture, thousands of high-resolution images are collected; in business and finance, millions of transactions are recorded every day. Frontiers of science, engineering, and humanities differ in the problems of their concerns, but nevertheless share a common theme: massive or complex data have been collected and new knowledge needs to be discovered. Massive data collection and new scientific research have strong impact on statistical thinking, methodological development, and theoretical studies. They have also challenged traditional statistical theory, methods, and computation. Many new insights and phenomena need to be discovered and new statistical tools need to be developed.

With this background, the Center for Statistical Research at the Chinese Academy of Science initiated the conference series “International Conference on the Frontiers of Statistics” in 2005. The aim is to provide a focal venue for researchers to gather, interact, and present their new research findings, to discuss and outline emerging problems in their fields, to lay the groundwork for future collaborations, and to engage more statistical scientists in China to conduct research in the frontiers of statistics. After the general conference in 2005, the 2006 International Conference on the Frontiers of Statistics, held in Changchun, focused on the topic “Biostatistics and Bioinformatics”. The conference attracted many top researchers in the area and was a great success. However, there are still a lot of Chinese scholars, particularly young researchers and graduate students, who were not able to attend the conference. This hampers one of the purposes of the conference series. However, an alternative idea was born: inviting active researchers to provide a bird-eye view on the new developments in the frontiers of statistics, on the theme topics of the conference series. This will broaden significantly the benefits of statistical research, both in China and worldwide. The edited books in this series aim at promoting statistical research that has high societal impacts and provide not only a concise overview on the recent developments in the frontiers of statistics, but also useful references to the literature at large, leading readers truly to the frontiers of statistics.

This book gives an overview on recent development on biostatistics and bioinformatics. It is written by active researchers in these emerging areas. It is intended

to give graduate students and new researchers an idea where the frontiers of biostatistics and bioinformatics are, to learn common techniques in use so that they can advance the fields via developing new techniques and new results. It is also intended to provide extensive references so that researchers can follow the threads to learn more comprehensively what the literature is and to conduct their own research. It covers three important topics in biostatistics: Analysis of Survival and Longitudinal Data, Statistical Methods for Epidemiology, and Bioinformatics, where statistics is still advancing rapidly today.

Ever since the invention of nonparametric and semiparametric techniques in statistics, they have been widely applied to the analysis of survival data and longitudinal data. In Chapter 1, Jianqing Fan and Jiancheng Jiang give a concise overview on this subject under the framework of the proportional hazards model. Nonparametric and semiparametric modeling and inference are stressed. Dongling Zeng and Jianwen Cai introduce an additive-accelerated rate regression model for analyzing recurrent event in Chapter 2. This is a flexible class of models that includes both additive rate model and accelerated rate models, and allows simple statistical inference. Longitudinal data arise frequently from biomedical studies and quadratic inference function provides important approaches to the analysis of longitudinal data. An overview is given in Chapter 3 on this topic by John Dziak, Runze Li, and Annie Qiu. In Chapter 4, Yi Li gives an overview on modeling and analysis of spatially correlated data with emphasis on mixed models.

The next two chapters are on statistical methods for epidemiology. Amy Laird and Xiao-Hua Zhou address the issues on study designs for biomarker-based treatment selection in Chapter 5. Several trial designs are introduced and evaluated. In Chapter 6, Jinbo Chen reviews recent statistical models for analyzing two-phase epidemiology studies, with emphasis on the approaches based on estimating-equation, pseudo-likelihood, and maximum likelihood.

The last four chapters are devoted to the analysis of genomic data. Chapter 7 features protein interaction predictions using diverse data sources, contributed by Yin Liu, Inyoung Kim, and Hongyu Zhao. The diverse data sources information for protein-protein interactions is elucidated and computational methods are introduced for aggregating these data sources to better predict protein interactions. Regulatory motif discovery is handled by Qing Zhou and Mayetri Gupta using Bayesian approaches in Chapter 8. The chapter begins with a basic statistical framework for motif finding, extends it to the identification of *cis*-regulatory modules, and then introduces methods that combine motif finding with phylogenetic footprint, gene expression or ChIP-chip data, and nucleosome positioning information. Cheng Li and Samir Amin use single nucleotide polymorphism (SNP) microarrays to analyze cancer genome alterations in Chapter 9. Various methods are introduced, including paired and non-paired loss of heterozygosity analysis, copy number analysis, finding significant altered regions across multiple samples, and hierarchical clustering methods. In Chapter 10, Evan Johnson, Jun Liu and Shirley Liu give a comprehensive overview on the design and analysis of ChIP-chip data on genome tiling microarrays. It spans from biological background and ChIP-chip experiments to statistical methods and computing.

The frontiers of statistics are always dynamic and vibrant. Young researchers

are encouraged to jump into the research wagons and cruise with tidal waves of the frontiers. It is never too late to get into the frontiers of scientific research. As long as your mind is evolving with the frontiers, you always have a chance to catch and to lead next tidal waves. We hope this volume helps you getting into the frontiers of statistical endeavors and cruise on them thorough your career.

Jianqing Fan, Princeton

Xihong Lin, Cambridge

Jun Liu, Cambridge

August 8, 2008

Contents

Preface

Part I Analysis of Survival and Longitudinal Data

Chapter 1 Non- and Semi- Parametric Modeling in Survival Analysis

Jianqing Fan, Jiancheng Jiang..... 3

1	Introduction	3
2	Cox's type of models	4
3	Multivariate Cox's type of models	14
4	Model selection on Cox's models	24
5	Validating Cox's type of models	27
6	Transformation models	28
7	Concluding remarks	30
	References	30

Chapter 2 Additive-Accelerated Rate Model for Recurrent Event

Donglin Zeng, Jianwen Cai..... 35

1	Introduction	35
2	Inference procedure and asymptotic properties	37
3	Assessing additive and accelerated covariates	40
4	Simulation studies	41
5	Application	42
6	Remarks	43
	Acknowledgements	44
	Appendix	44
	References	48

Chapter 3 An Overview on Quadratic Inference Function Approaches for Longitudinal Data

John J. Dziak, Runze Li, Annie Qu..... 49

1	Introduction	49
2	The quadratic inference function approach	51
3	Penalized quadratic inference function	56
4	Some applications of QIF	60
5	Further research and concluding remarks	65
	Acknowledgements	68

References.....	68
Chapter 4 Modeling and Analysis of Spatially Correlated Data	
<i>Yi Li</i>	73
1 Introduction	73
2 Basic concepts of spatial process	76
3 Spatial models for non-normal/discrete data.....	82
4 Spatial models for censored outcome data	88
5 Concluding remarks.....	96
References.....	96
 Part II Statistical Methods for Epidemiology	
 Chapter 5 Study Designs for Biomarker-Based Treatment Selection	
<i>Amy Laird, Xiao-Hua Zhou</i>	103
1 Introduction	103
2 Definition of study designs	104
3 Test of hypotheses and sample size calculation	108
4 Sample size calculation.....	111
5 Numerical comparisons of efficiency	116
6 Conclusions.....	118
Acknowledgements	121
Appendix	122
References.....	126
 Chapter 6 Statistical Methods for Analyzing Two-Phase Studies	
<i>Jinbo Chen</i>	127
1 Introduction	127
2 Two-phase case-control or cross-sectional studies	130
3 Two-phase designs in cohort studies.....	136
4 Conclusions.....	149
References.....	151
 Part III Bioinformatics	
 Chapter 7 Protein Interaction Predictions from Diverse Sources	
<i>Yin Liu, Inyoung Kim, Hongyu Zhao</i>	159
1 Introduction	159
2 Data sources useful for protein interaction predictions.....	161
3 Domain-based methods	163
4 Classification methods	169

5	Complex detection methods.....	172
6	Conclusions.....	175
	Acknowledgements.....	175
	References.....	175

Chapter 8 Regulatory Motif Discovery: From Decoding to Meta-Analysis

	<i>Qing Zhou, Mayetri Gupta.....</i>	179
1	Introduction.....	179
2	A Bayesian approach to motif discovery.....	181
3	Discovery of regulatory modules.....	184
4	Motif discovery in multiple species.....	189
5	Motif learning on ChIP-chip data.....	195
6	Using nucleosome positioning information in motif discovery.....	201
7	Conclusion.....	204
	References.....	205

Chapter 9 Analysis of Cancer Genome Alterations Using Single Nucleotide Polymorphism (SNP) Microarrays

	<i>Cheng Li, Samir Amin.....</i>	209
1	Background.....	209
2	Loss of heterozygosity analysis using SNP arrays.....	212
3	Copy number analysis using SNP arrays.....	216
4	High-level analysis using LOH and copy number data.....	224
5	Software for cancer alteration analysis using SNP arrays.....	229
6	Prospects.....	231
	Acknowledgements.....	231
	References.....	231

Chapter 10 Analysis of ChIP-chip Data on Genome Tiling Microarrays

	<i>W. Evan Johnson, Jun S. Liu, X. Shirley Liu.....</i>	239
1	Background molecular biology.....	239
2	A ChIP-chip experiment.....	241
3	Data description and analysis.....	244
4	Follow-up analysis.....	249
5	Conclusion.....	254
	References.....	254

Subject Index.....	259
---------------------------	------------

Author Index.....	261
--------------------------	------------

Part I

Analysis of Survival and Longitudinal Data

Chapter 1

Non- and Semi- Parametric Modeling in Survival Analysis *

Jianqing Fan [†] Jiancheng Jiang [‡]

Abstract

In this chapter, we give a selective review of the nonparametric modeling methods using Cox's type of models in survival analysis. We first introduce Cox's model (Cox 1972) and then study its variants in the direction of smoothing. The model fitting, variable selection, and hypothesis testing problems are addressed. A number of topics worthy of further study are given throughout this chapter.

Keywords: Censoring; Cox's model; failure time; likelihood; modeling; nonparametric smoothing.

1 Introduction

Survival analysis is concerned with studying the time between entry to a study and a subsequent event and becomes one of the most important fields in statistics. The techniques developed in survival analysis are now applied in many fields, such as biology (survival time), engineering (failure time), medicine (treatment effects or the efficacy of drugs), quality control (lifetime of component), credit risk modeling in finance (default time of a firm).

An important problem in survival analysis is how to model well the conditional hazard rate of failure times given certain covariates, because it involves frequently asked questions about whether or not certain independent variables are correlated with the survival or failure times. These problems have presented a significant challenge to statisticians in the last 5 decades, and their importance has motivated many statisticians to work in this area. Among them is one of the most important contributions, the proportional hazards model or Cox's model and its associated partial likelihood estimation method (Cox, 1972), which stimulated

*The authors are partly supported by NSF grants DMS-0532370, DMS-0704337 and NIH R01-GM072611.

[†]Department of ORFE, Princeton University, Princeton, NJ 08544, USA, E-mail: jqfan@princeton.edu

[‡]Department of Mathematics and Statistics, University of North Carolina, Charlotte, NC 28223, USA, E-mail: jjiang1@uncc.edu

a lot of works in this field. In this chapter we will review related work along this direction using the Cox's type of models and open an academic research avenue for interested readers. Various estimation methods are considered, a variable selection approach is studied, and a useful inference method, the generalized likelihood ratio (GLR) test, is employed to address hypothesis testing problems for the models. Several topics worthy of further study are laid down in the discussion section.

The remainder of this chapter is organized as follows. We consider univariate Cox's type of models in Section 2 and study multivariate Cox's type of models using the marginal modeling strategy in Section 3. Section 4 focuses on model selection rules, Section 5 is devoted to validating Cox's type of models, and Section 6 discusses transformation models (extensions to Cox's models). Finally, we conclude this chapter in the discussion section.

2 Cox's type of models

Model Specification. The celebrated Cox model has provided a tremendously successful tool for exploring the association of covariates with failure time and survival distributions and for studying the effect of a primary covariate while adjusting for other variables. This model assumes that, given a q -dimensional vector of covariates $\mathbf{Z}$, the underlying conditional hazard rate (rather than expected survival time T),

$$\lambda(t|\mathbf{z}) = \lim_{\Delta t \rightarrow 0+} \frac{1}{\Delta t} P\{t \leq T < t + \Delta t | T \geq t, \mathbf{Z} = \mathbf{z}\}$$

is a function of the independent variables (covariates):

$$\lambda(t|\mathbf{z}) = \lambda_0(t)\Psi(\mathbf{z}), \quad (2.1)$$

where $\Psi(\mathbf{z}) = \exp(\psi(\mathbf{z}))$ with the form of the function $\psi(\mathbf{z})$ known such as $\psi(\mathbf{z}) = \beta^T \mathbf{z}$, and $\lambda_0(t)$ is an unknown baseline hazard function. Once the conditional hazard rate is given, the condition survival function $S(t|\mathbf{z})$ and conditional density $f(t|\mathbf{z})$ are also determined. In general, they have the following relationship:

$$S(t|\mathbf{z}) = \exp(-\Lambda(t|\mathbf{z})), \quad f(t|\mathbf{z}) = \lambda(t|\mathbf{z})S(t|\mathbf{z}), \quad (2.2)$$

where $\Lambda(t|\mathbf{z}) = \int_0^t \lambda(t|\mathbf{z})dt$ is the cumulative hazard function. Since no assumptions are made about the nature or shape of the baseline hazard function, the Cox regression model may be considered to be a semiparametric model.

The Cox model is very useful for tackling with censored data which often happen in practice. For example, due to termination of the study or early withdrawal from a study, not all of the survival times $T_1, \dots, T_n$ may be fully observable. Instead one observes for the i -th subject an event time $X_i = \min(T_i, C_i)$, a censoring indicator $\delta_i = I(T_i \leq C_i)$, as well as an associated vector of covariates $\mathbf{Z}_i$. Denote the observed data by $\{(\mathbf{Z}_i, X_i, \delta_i) : i = 1, \dots, n\}$ which is an i.i.d. sample from the population $(\mathbf{Z}, X, \delta)$ with $X = \min(T, C)$ and $\delta = I(T \leq C)$. Suppose that

the random variables T and C are positive and continuous. Then by Fan, Gijbels, and King (1997), under the Cox model (2.1),

$$\Psi(x) = \frac{E\{\delta|\mathbf{Z} = \mathbf{z}\}}{E\{\Lambda_0(X)|\mathbf{Z} = \mathbf{z}\}}, \quad (2.3)$$

where $\Lambda_0(t) = \int_0^t \lambda_0(u) du$ is the cumulative baseline hazard function. Equation (2.3) allows one to estimate the function Ψ using regression techniques if $\lambda_0(t)$ is known.

The likelihood function can also be derived. When $\delta_i = 0$, all we know is that the survival time $T_i \geq C_i$ and the probability for getting this is

$$P(T_i \geq C_i|\mathbf{Z}_i) = P(T_i \geq X_i|\mathbf{Z}_i) = S(X_i|\mathbf{Z}_i),$$

whereas when $\delta_i = 1$, the likelihood of getting T_i is $f(T_i|\mathbf{Z}_i) = f(X_i|\mathbf{Z}_i)$. Therefore the conditional (given covariates) likelihood for getting the data is

$$L = \prod_{\delta_i=1} f(X_i|\mathbf{Z}_i) \prod_{\delta_i=0} S(X_i|\mathbf{Z}_i) = \prod_{\delta_i=1} \lambda(X_i|\mathbf{Z}_i) \prod_i S(X_i|\mathbf{Z}_i), \quad (2.4)$$

and using (2.2), we have

$$\begin{aligned} L &= \sum_{\delta_i=1} \log(\lambda(X_i|\mathbf{Z}_i)) - \sum_i \Lambda(X_i|\mathbf{Z}_i) \\ &= \sum_i \delta_i \log(\lambda(X_i|\mathbf{Z}_i)) - \sum_i \Lambda(X_i|\mathbf{Z}_i). \end{aligned} \quad (2.5)$$

For proportional hazards model (2.1), we have specifically

$$L = \sum_i \delta_i \log(\lambda_0(X_i)\Psi(\mathbf{Z}_i)) - \sum_i \Lambda_0(X_i)\Psi(\mathbf{Z}_i). \quad (2.6)$$

Therefore, when both $\psi(\cdot)$ and $\lambda_0(\cdot)$ are parameterized, the parameters can be estimated by maximizing the likelihood (2.6).

Estimation. The likelihood inference can be made about the parameters in model (2.1) if the baseline $\lambda_0(\cdot)$ and the risk function $\psi(\cdot)$ are known up to a vector of unknown parameters β (Aitkin and Clayton, 1980), i.e.

$$\lambda_0(\cdot) = \lambda_0(\cdot; \beta) \quad \text{and} \quad \psi(\cdot) = \psi(\cdot; \beta).$$

When the baseline is completely unknown and the form of the function $\psi(\cdot)$ is given, inference can be based on the partial likelihood (Cox, 1975). Since the full likelihood involves both β and $\lambda_0(t)$, Cox decomposed the full likelihood into a product of the term corresponding to identities of successive failures and the term corresponding to the gap times between any two successive failures. The first term inherits the usual large-sample properties of the full likelihood and is called the partial likelihood.