# New Developments in Biostatistics and Bioinformatics

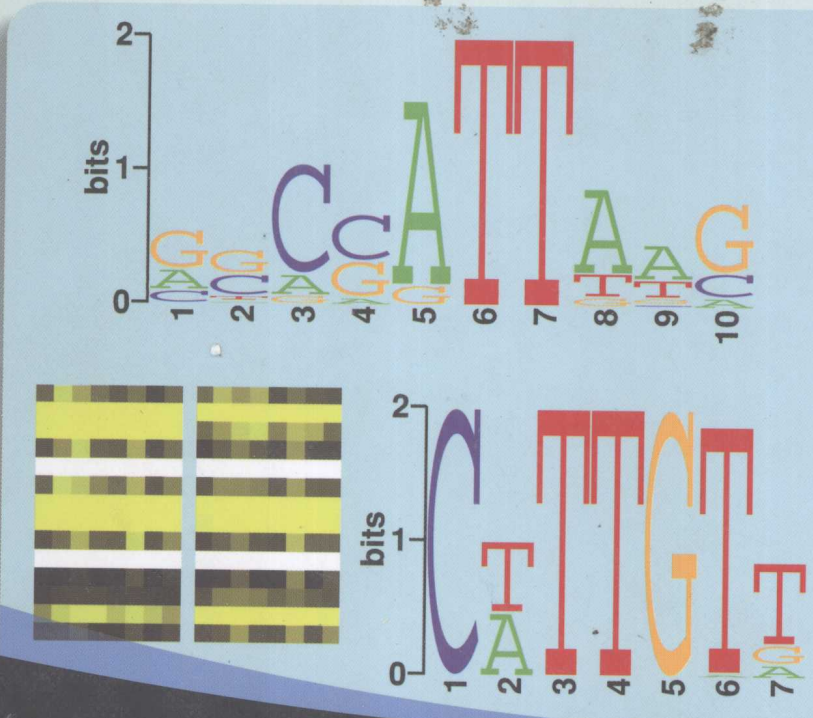
# 生物统计学和生物信息学最新进展

Editors

**Jianqing Fan**
**Xihong Lin**
**Jun S. Liu**

Volume 1

Frontiers of Statistics




高等教育出版社
Higher Education Press

# New Developments in Biostatistics and Bioinformatics

生物统计学与生物信息学最新进展

Jianqing Fan
Xihong Lin
Jun S. Liu

Volume 1

高等教育出版社

# New Developments in Biostatistics and Bioinformatics

# 生物统计学和生物信息学最新进展


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

**Jianqing Fan**
Department of Operation Reasearch 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

**图书在版编目(CIP)数据**

生物统计学和生物信息学最新进展 = New Developments in Biostatistics and Bioinformatics: 英文 / 范剑青, 林希虹, 刘军主编. —北京: 高等教育出版社, 2008.12
(统计前沿 / 范剑青, 马志明主编)
ISBN 978-7-04-024755-8

Ⅰ. 生… Ⅱ. ①范…②林…③刘… Ⅲ. ①生物统计-文集-英文②生物信息论-文集-英文 Ⅳ. Q-332 Q811.4-53

中国版本图书馆 CIP 数据核字 (2008) 第 183634 号



**策划编辑** 王丽萍 **责任编辑** 王丽萍 **封面设计** 张 楠
**责任印制** 陈伟光

| | | | |
|---|---|---|---|
| **出版发行** | 高等教育出版社 | **购书热线** | 010-58581118 |
| **社 址** | 北京市西城区德外大街 4 号 | **免费咨询** | 800-810-0598 |
| **邮政编码** | 100120 | **网 址** | http://www.hep.edu.cn |
| **总 机** | 010-58581000 | | http://www.hep.com.cn |
| | | **网上订购** | http://www.landraco.com |
| **经 销** | 蓝色畅想图书发行有限公司 | | http://www.landraco.com.cn |
| **印 刷** | 涿州市星河印刷有限公司 | **畅想教育** | http://www.widedu.com |
| **开 本** | 787 × 1092 1/16 | | |
| **印 张** | 17.75 | **版 次** | 2009 年 1 月第 1 版 |
| **字 数** | 355 000 | **印 次** | 2009 年 1 月第 1 次印刷 |
| **插 页** | 4 | **定 价** | 49.00 元 |



**物料号 24755-00**

# Preface

The first eight years of the twenty-first century has witted 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

# 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 $T$),

$$\lambda(t|\mathbf{z}) = \lim_{\Delta t \to 0+} \frac{1}{\Delta t} P\{t \leqslant T < t + \Delta t | T \geqslant t, \mathbf{Z} = \mathbf{z}\}$$

is a function of the independent variables (covariates):

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

where $\Psi(\mathbf{z}) = \exp(\psi(\mathbf{z}))$ with the form of the function $\psi(\mathbf{z})$ known such as $\psi(\mathbf{z}) = \boldsymbol{\beta}^T\mathbf{z}$, 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:

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

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 $T_1, \cdots, T_n$ 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 $\delta_i = I(T_i \leqslant C_i)$, as well as an associated vector of covariates $\mathbf{Z}_i$. Denote the observed data by $\{(\mathbf{Z}_i, X_i, \delta_i) : i = 1, \cdots, n\}$ which is an i.i.d. sample from the population $(\mathbf{Z}, X, \delta)$ with $X = \min(T, C)$ and $\delta = I(T \leqslant C)$. 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),

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

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 $\Psi$ 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 $T_i \geqslant C_i$ and the probability for getting this is

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

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

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

and using (2.2), we have

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

For proportional hazards model (2.1), we have specifically

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

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 $\boldsymbol{\beta}$ (Aitkin and Clayton, 1980), i.e.

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

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 $\boldsymbol{\beta}$ 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.

The partial likelihood can also be derived from counting process theory (see for example Andersen, Borgan, Gill, and Keiding 1993) or from a profile likelihood in Johansen (1983). In the following we introduce the latter.

**Example 1** [The partial likelihood as profile likelihood; Fan, Gijbel, and King (1997)] Consider the case that $\psi(\mathbf{z}) = \psi(\mathbf{z};\boldsymbol{\beta})$. Let $t_1 < \cdots < t_N$ denote the ordered failure times and let $(i)$ denote the label of the item failing at $t_i$. Denote by $R_i$ the risk set at time $t_i-$, that is $R_i = \{j : X_j \geqslant t_i\}$. Consider the least informative nonparametric modeling for $\Lambda_0(\cdot)$, that is, $\Lambda_0(t)$ puts point mass $\theta_j$ at time $t_j$ in the same way as constructing the empirical distribution:

$$\Lambda_0(t;\theta) = \sum_{j=1}^{N} \theta_j I(t_j \leqslant t). \tag{2.7}$$

Then

$$\Lambda_0(X_i;\theta) = \sum_{j=1}^{N} \theta_j I(i \in R_j). \tag{2.8}$$

Under the proportional hazards model (2.1), using (2.6), the log likelihood is

$$\begin{aligned}\log L = \sum_{i=1}^{n} [&\delta_i \{\log \lambda_0(X_i;\theta) + \psi(Z_i;\boldsymbol{\beta})\} \\ &-\Lambda_0(X_i;\theta)\exp\{\psi(Z_i;\boldsymbol{\beta})\}].\end{aligned} \tag{2.9}$$

Substituting (2.7) and (2.8) into (2.9), one establishes that

$$\begin{aligned}\log L = &\sum_{j=1}^{n} [\log \theta_j + \psi(Z_{(j)};\boldsymbol{\beta})] \\ &-\sum_{i=1}^{n}\sum_{j=1}^{N} \theta_j I(i \in R_j)\exp\{\psi(Z_i;\boldsymbol{\beta})\}.\end{aligned} \tag{2.10}$$

Maximizing $\log L$ with respect to $\theta_j$ leads to the following Breslow estimator of the baseline hazard [Breslow (1972, 1974)]

$$\hat{\theta}_j = \Big[\sum_{i \in R_j} \exp\{\psi(Z_i;\boldsymbol{\beta})\}\Big]^{-1}. \tag{2.11}$$

Substituting (2.11) into (2.10), we obtain

$$\max_{\lambda_0} \log L = \sum_{i=1}^{n} \Big(\psi(\mathbf{Z}_{(i)};\boldsymbol{\beta}) - \log\Big[\sum_{j \in R_i} \exp\{\psi(\mathbf{Z}_j;\boldsymbol{\beta})\}\Big]\Big) - N.$$

This leads to the log partial likelihood function (Cox 1975)

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Big(\psi(\mathbf{Z}_{(i)};\boldsymbol{\beta}) - \log\Big[\sum_{j \in R_i} \exp\{\psi(\mathbf{Z}_j;\boldsymbol{\beta})\}\Big]\Big). \tag{2.12}$$

An alternative expression is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n}\Big(\psi(\mathbf{Z}_{(i)};\boldsymbol{\beta}) - \log\Big[\sum_{j=1}^{n} Y_j(X_i)\exp\{\psi(\mathbf{Z}_j;\boldsymbol{\beta})\}\Big]\Big),$$

where $Y_j(t) = I(X_j \geqslant t)$ is the survival indicator on whether the $j$-th subject survives at the time $t$.

The above partial likelihood function is a profile likelihood and is derived from the full likelihood using the least informative nonparametric modeling for $\Lambda_0(\cdot)$, that is, $\Lambda_0(t)$ has a jump $\theta_i$ at $t_i$. ◇

Let $\hat{\boldsymbol{\beta}}$ be the partial likelihood estimator of $\boldsymbol{\beta}$ maximizing (2.12) with respect to $\boldsymbol{\beta}$. By standard likelihood theory, it can be shown that (see for example Tsiatis 1981) the asymptotic distribution $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is multivariate normal with mean zero and a covariance matrix which may be estimated consistently by $(n^{-1}I(\hat{\boldsymbol{\beta}}))^{-1}$, where

$$I(\boldsymbol{\beta}) = \int_0^T \Big[\frac{S_2(\boldsymbol{\beta},t)}{S_0(\boldsymbol{\beta},t)} - \Big(\frac{S_1(\boldsymbol{\beta},t)}{S_0(\boldsymbol{\beta},t)}\Big)^{\otimes 2}\Big]\, dN(t)$$

and for $k = 0, 1, 2$,

$$S_k(\boldsymbol{\beta},t) = \sum_{i=1}^{n} Y_i(t)\psi'(\mathbf{Z}_i;\boldsymbol{\beta})^{\otimes k}\exp\{\psi(\mathbf{Z}_i;\boldsymbol{\beta})\},$$

where $N(t) = I(X \leqslant t, \delta = 1)$, and $\mathbf{x}^{\otimes k} = 1, \mathbf{x}, \mathbf{x}\mathbf{x}^T$, respectively for $k = 0, 1$ and 2.

Since the baseline hazard $\Lambda_0$ does not appear in the partial likelihood, it is not estimable from the likelihood. There are several methods for estimating parameters related to $\Lambda_0$. One appealing estimate among them is the Breslow estimator (Breslow 1972, 1974)

$$\hat{\Lambda}_0(t) = \int_0^T \Big[\sum_{i=1}^{n} Y_i(s)\exp\{\mathbf{Z}_i^T\hat{\boldsymbol{\beta}}\}\Big]^{-1}\Big\{\sum_{i=1}^{n} dN_i(s)\Big\}, \tag{2.13}$$

where $N_i(s) = I(X_i \leqslant s, \delta_i = 1)$.

**Hypothesis testing**. After fitting the Cox model, one might be interested in checking if covariates really contribute to the risk function, for example, checking if the coefficient vector $\boldsymbol{\beta}$ is zero. More generally, one considers the hypothesis testing problem

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0.$$

From the asymptotic normality of the estimator $\hat{\boldsymbol{\beta}}$, it follows that the asymptotic null distribution of the Wald test statistic

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T I(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$