网络优先数字出版时间:2014-10-24
网络优先数字出版地址:http://www.cnki.net/kcms/doi/10.13656/j.cnki.gxkx.20141024.010.html
广西科学 Guangxi Sciences 2014,21(6):634～651

Application of Semiparametric Technique in Biomedical Fields*

半参数技术在生物医学领域中的应用

YAN Shao-min, WU Guang**
严少敏,吴　光

(State Key Laboratory of Non-food Biomass Enzyme Technology, National Engineering Research Center for Non-food Biorefinery, Guangxi Biomass Industrialization Engineering Institute, Guangxi Key Laboratory of Biorefinery, Guangxi Academy of Sciences, Nanning, Guangxi, 530007, China)
(广西科学院,非粮生物质酶解国家重点实验室,国家非粮生物质能源工程技术研究中心,广西生物质产业化工程院,广西生物炼制重点实验室,广西南宁　530007)

Abstract: Over recent years, semiparametric technique becomes to have more and more applications in biometrics as evidenced in the number of publications per year. Therefore it is necessary and timely to review semiparametric technique's applications in biomedical researches in order that we can get not only a whole picture on its development tendency but also some insights into its applications. In this review, we addressed the application of semiparametric technique in different biomedical fields, including receiver operating characteristic (ROC) analysis, survival and longitudinal analysis, pharmacokinetics and pharmacodynamics, exposure study, genetic study and other research fields. In each section, we started from general views on the background of semiparametric technique, advanced to its applicable circumstances and finished with comparison studies, through which one would have a general view on the background of applications, applicable cases, advantages in comparison with parametric and nonparametric techniques, and covariates in diseases, which can be applicable for semiparametric technique.

Key words: data analysis, exposure study, genetic study, longitudinal analysis, pharmacodynamics, pharmacokinetics, ROC analysis, semiparametric technique, survival analysis

摘要:近年来半参数技术越来越多地应用于生物特征的识别,相关研究成果(出版物)逐年增加。到目前为止,该技术已用于生物医学领域的各种研究,对此有必要对其应用情况进行及时总结,这将有助于对其发展趋势及应用有一个全面了解。本文综述了半参数技术在不同生物医学领域的应用,包括接受者操作特征(ROC)分析、生存和纵向分析、药代动力学和药效学、暴露研究、基因研究及其它研究领域。其中每一部分以对半参数技术背景的一般描述开始,进而介绍其适用情况,最后对不同的研究加以比较,使大家对半参数技术的应用背景、适用情况、与参数和非参数方法相比的优点以及在疾病诊断中的协变量有概念性认识。

关键词:数据分析　暴露研究　遗传研究　纵向分析　药效学　药代动力学　ROC 分析　半参数方法　生存分析
中图分类号:Q332　　**文献标识码**:A　　**文章编号**:1005-9164(2014)06-0634-18

收稿日期:2014-06-30

作者简介:严少敏(1958-),女,博士,研究员,主要从事计算生物学、生物信息学和定量诊断研究。

*广西自然科学基金重点项目(2013GXNSFDA019007),广西科技创新能力与条件建设计划项目(桂科能 12237022)和广西人才小高地建设专项基金项目资助。

**通讯作者:吴　光(1956-),男,博士,研究员,主要从事计算生物学、生物信息学和模型研究,E-mail:hongguanglishibahao@yahoo.com。

Of various biometric techniques, semiparametric (semi-parametric) technique is far less used in

biomedical fields in comparison with parametric and nonparametric techniques. It is not only because semiparametric technique appears somewhat new so that most researchers are not familiar with it but also because researchers do not know how to apply this technique to real-life cases.

Despite of this unfamiliarity and difficulty, semiparametric technique steadily increases its applications in biomedical research as shown in Fig. 1, where a literature search was conducted on June 17, 2013 in PubMed [1], using (i) semi-parametric, (ii) "semi-parametric", (iii) semiparametric and (iv) "semiparametric" as searching keywords. Either semi-parametric or "semi-parametric" search resulted in 427 articles including 28 articles published in 2013, and either semiparametric or "semiparametric" search resulted in 943 articles including 51 articles published in 2013. Undoubtedly, semiparametric technique began its long journey in biomedical research fields and has been drawing more and more attention from biomedical community. As can be seen in Fig. 1, the earliest article found with keyword semiparametric was dated back to 1982 and the earliest article found with keyword semi-parametric was dated back to 1984, but 48 and 112 articles were found with keywords of semi-parametric and semiparametric in 2012. Interestingly, 427 semiparametric articles are not a subgroup of 943 semiparametric articles because only 18 articles exist in both semi-parametric and semiparametric searches.

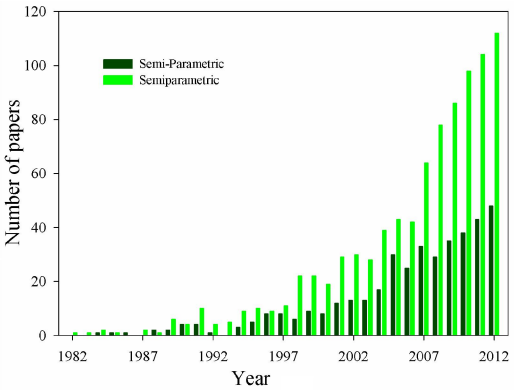

Fig. 1 Number of published papers from 1982 to 2012 searched using semi-parametric including "semi-parametric" and semiparametric including "semiparametric" as keywords in PubMed.

Although it is too early to say that semiparametric technique is quite popular, it would be timely and informative to review how semiparametric technique was applied in biomedical research fields and to summarize these application procedures in order that one can have a whole picture on the application of semiparametric technique in biomedical settings. Yet, a review in this fast-developing field also would give us possible research directions for the application of semiparametric technique.

Over years, comparison was made in studies between semiparametric technique and nonparametric as well as parametric techniques. Moreover, simulation studies were also conducted for such comparison. Therefore, we could be at the position to cite several studies in order to get a clear picture on the application of semiparametric technique in various research fields. In this review, we attempt to summarize general views on the background of applications, applicable cases, advantages in comparison with parametric and nonparametric techniques, and covariates in diseases, which can be applicable for semiparametric technique.

1 Semiparametric technique: Its definition in biomedical context

When checking the entry "semiparametric model" in Wikipedia on June 18, 2013 [2], it says: (i) a parametric model is one in which the indexing parameter is a finite-dimensional vector; (ii) in nonparametric models, the set of possible values of the parameter is a subset of some space, not necessarily finite dimensional; and (iii) in semiparametric models, the parameter has both a finite dimensional component and an infinite dimensional component. It turns out that a similar definition can be found in the entry "semiparametric regression" in Wikipedia. This definition seems to be ambiguous for non-professionals, and one might need to consider this definition in the biomedical context. In plain words, parametric technique assumes that samples for study are normally distributed whereas nonparametric technique does not require this assumption. Thus, this definition implies that semiparametric technique is something in between. However, it appears hard to define what is neither normal distribution nor

non-normal distribution in biomedical settings.

Regardless this ambiguity, we have witnessed the increasing applications of semiparemetric technique in biomedical fields[3]. This means that researchers might not need a deep understanding on very core of semiparametric technique because this request would cost valuable time for studies, but can apply this technique correctly even according to various publications with the use of this technique. Thus, it is necessarily useful to see how researchers had used this technique in recent years in order to get a practical sense on how to use this technique because any application requires familiarity with the relevant scientific context. Through comparison with other techniques, we can really see in which circumstances that semiparametric technique can be used and what the difference is between semiparametric technique and other techniques. Thereafter, we can be able to distinguish the characteristics of semiparametric technique from others and determine whether we can use other techniques to solve the problem and determine specific applicable fields for semiparametric technique.

2 Applied research fields

As semiparametric technique is defined between parametric and nonparametric techniques, thus almost all the analytically parametric and nonparametric techniques in biometrics could be potential fields for the applications of semiparametric technique. These analytically parametric and nonparametric techniques could be summarized as follows.

2.1 ROC analysis

2.1.1 Background

ROC stands receiver operating characteristic, which was developed as early as in World War Ⅱ in signal detection theory[4]. In biomedical research fields, the ROC analysis is mainly used for diagnostic tests, and for evaluation of various testing models. So far, there is a large body of literatures, which describes the utility of ROC in biomedical research fields[5~7]. Technically, the ROC analysis is used when a test has a continuous value, each tested value can be interpreted as an yes/no event. Then a sample population is divided into yes and no groups such as a disease group and a healthy one for each tested value, the true positive rate is counted in disease group and the false positive rate is counted in health group. Hereafter, a graph is made with true positive rate as y-axis and false positive rate as x-axis. With different tested values as a cutoff value, an ROC curve can be constructed. A diagnostic decision can be made based on ROC curve, which should be above the diagonal line that presents random chance. Of various parameters of ROC, *AUC* (area under the curve) is an important indicator to decide performance of a diagnostic test as the larger the *AUC* is, the better the performance is, and *AUC* ranges theoretically from zero to unity. Indeed, it was suggested that the ROC is the best statistical method for assessing performance of tests and models[8,9]. Actually, the ROC analysis is quite popular because even the popular science magazine, Scientific American, has an article to introduce it[10].

2.1.2 Applicable circumstances for semiparametric technique

As a sample population of tested values can be normally or non-normally distributed either in healthy population or in disease one, this leads to parametric technique to calculate *AUC* with the assumption that tested values in healthy and disease populations are normally distributed, and non-parametric technique to calculate *AUC* without the consideration of whether or not tested values in healthy and disease populations are normally distributed[11,12]. Indeed, the assumption of distributions of tested values in both healthy and disease populations is crucial because the estimated association between tested values and disease ones is biased if the assumption of distributions is mis-specified. In this case, semiparametric technique does not consider the distribution of tested values that is similar to nonparametric technique, but computes the parameters in parametric technique that is similar to parametric technique. The second circumstance for the distribution of tested values is that there is a tendency that the higher the tested value is the higher the risk of disease is. This tendency renders a normal distribution of tested values so an ROC curve is a probabilistic quantity directly incorporating monotonici-

ty[13]. The third circumstance related to ROC is the so-called discrimination and association, i. e., whether a diagnostic test can discriminate a disease from tested population and whether a diagnostic test is associated with a disease status[5~7].

In these contexts, at least there are two major reasons for applying semiparametric technique to ROC analysis. The first reason is to more accurately calculate *AUC*. Graphically, an ROC curve is a relationship between true positive rate and false one. If an ROC curve can be accurately described using an equation, then integration would result in an accurate *AUC*. However, it is difficult to apply a cause-consequence relationship to the relationship between true and false positive rates because it is hard to say that a change in true positive rate leads to a change in false positive rate or verse vice. This is different from cause-consequence relationship such as Newton's law, where an object's acceleration would be expected if a force is applied. Thus regression is applied to the relationship between true and false positive rates, and the regression will result in an equation for integration. The problem here is that regression requires its data to be normally distributed, but tested values of some diagnostic tests may not be normally distributed. When tested values in both healthy and disease populations are normally distributed, the regression can be safely applied, which is to calculate the *AUC* using parametric technique. When tested values in either healthy or disease population are not normally distributed, the regression theoretically cannot be applied, and the *AUC* has to be calculated using the sum of areas of rectangles, which approximate to the *AUC*. The latter one is the so-called nonparametric technique to calculate the *AUC*. So it is generally considered that nonparametric technique results in a less accurate *AUC* because it does not generate a smooth ROC curve, while parametric technique results in a more accurate *AUC* than nonparametric technique but its strict restriction limits its application[11,12]. Practically, the *AUC* calculated by using nonparametric technique can be as accurate as by using parametric technique if the interval between any two cutoff values is sufficient small. When tested values are exponentially distributed or distributed in other power-law distributions, the proportional hazard model assumption is met[13], then the application of semiparametric technique for comparing correlated *AUC*s under proportional hazard models[14,15]. For this reason, semiparametric technique is to deal with non-normal distribution of tested values using regression to generate the *AUC*.

The second reason for applying semiparametric technique to ROC analysis is the problem of covariates, which demonstrate the property of monotonicity, for example, age increases as time going on. Therefore it is necessary to determine whether an association between a diagnostic test and a disease is partly due to monotonicity of covariates. This can be more clearly explained by looking at the following equation, for example, $g\{P(D=1 \mid S,Z)\}=f(S)+\beta Z$, where g is a link function, P is probability, D is disease status with 0 for health and 1 for disease, S is a tested value, Z is age, f is an unspecified monotone function, and β is the regression coefficient for age to be estimated[16]. In this general equation, $f(S)$ can be considered as parametric term while βZ can be considered as nonparametric term. In this example, semiparametric technique includes both parametric and nonparametric terms. In other words, semiparametric technique actually is a mixed effects model, which is true because the standard generalized linear mixed effects software, such as SAS[17], can be used. This second reason can furthermore imply that the ROC analysis needs to incorporate both time-varying nature of diagnostic test and the clinical onset time of the disease, i. e., time-dependent ROC curve or whether true positive rate or false one is time-dependent, which results in a covariate-specific ROC curve[18~20] as well as time lag between when the marker is measured and the occurrence of disease[21]. In this context, the ROC analysis would incorporate with censored survival outcomes[22].

2.1.3 Comparison studies

Actually, programs for comparison among parametric, nonparametric and semiparametric techniques are available for certain software[23]. In published comparison studies, there are studies, which demonstrate the advantage of semiparametric tech-

nique over parametric and nonparametric techniques. Zheng and Heagerty[17] used the data of 21 138 patients with 171 306 observations between 1990 and 1998 from the U. S. Cystic Fibrosis Foundation National Patient Registry with ROC curves to determine how well the pulmonary function measurement can distinguish the subjects that progress to death from the subjects that remain alive. In this study, the diagnostic test is the forced expiratory volume as a pulmonary function measure for cystic fibrosis. Since the forced expiratory volume is changeable over time, longitudinal analysis hopes to evaluate whether changes in a forced expiratory volume correlate with death of patients. The authors hoped to extend an ROC curve to incorporate both the time-varying nature of forced expiratory volume and the clinical onset time of the death. Therefore, they defined the true positive rate and false positive rate in ROC as time-dependent functions, and used a semiparametric technique to regress time-dependent ROC curves, and their results showed the semiparametric technique offers greater flexibility by allowing separate model choices for each of the key distributional aspects in comparison with the parametric technique. Heagerty et al endorsed to develop a time-dependent ROC analysis with censored observations[21]. With the examples from flow cytometry data of 1292 women from 1983 to 1992, following-up time censored on May 1, 1997 and modified eligibility criteria for HIV trial of 3257 HIV-negative gay men followed for three semiannual visits, the authors showed that semi-parametric technique is more efficient than Kaplan-Meier methods for estimating ROC curves. Pepe theoretically suggested the advantage of semiparametric techniques over parametric and non-parametric techniques by comparing equations[11].

In other types of ROC analyses, Huang and Pepe[24], and Wan and Zhang[25,26] also showed that semiparametric technique is substantially more efficient than nonparametric technique under a correctly specified model when dealing with risk prediction models in case-control studies with ROC analysis. In clinical applications, Zou et al. studied 100 unenhanced spiral computed tomography (CT) scans of patients with proved obstructing ureteral stones using parametric, nonparametric and semiparametric techniques to compute *AUC* of ROC[27], and their results showed that semiparametric technique generated the largest *AUC* although the authors concluded that parametric techniques is preferred for constructing a smooth ROC curve with available stone-size data derived from spiral computed tomography (CT). Of course, simulation studies also suggest that semiparametric technique is better to work with ROC analysis[18,28] or comparable with parametric technique when the assumption of data distribution is correctly specified for the parametric technique[12,25,26].

Naturally, there are studies, which do not show the advantage of semiparametric technique over nonparametric and parametric techniques. Tang et al. used both nonparametric and semiparametric techniques with simulation of parametric technique to determine which of two diagnostic tests, computed tomography (CT) and positron emission tomography (PET), is better to detect non-small cell lung cancer by comparing weighted *AUC* under ROC curves[15]. Their results showed nonparametric technique is robust to model mis-specification and has excellent finite-sample performance while semi-parametric technique allows survival outcome measurements in the presence of censoring with correct model specification. In a study by Huang and Pepe[29], the authors randomly sampled 250 prostate cancer cases and 250 controls from 5519 cases using parametric ROC model to evaluate prostate-specific antigen as a risk prediction marker for the diagnosis of prostate cancer from the biopsy, and found parametric technique better than semi-parametric one because the latter uses logistic regression to fit a risk model, however, the method based on ROC curve has some desirable properties that logistic regression lacks. For example, the parametric fitting depends on the scale of tested values whereas an ROC curve is rank invariant so it does not matter on what scale the marker is measured. In theoretical front, Gu et al. showed that a Bayesian bootstrap, a fully non-parametric technique, is better than parametric and semiparametric techniques in ROC anal-

ysis when tested values are binormally distributed with simulations and real data[30].

With respect to estimate the standard error of *AUC* in ROC analysis, Hajian-Tilaki and Hanley showed semiparametric and nonparametric techniques produced very close results and the choice between these two techniques should rely on researchers' preferences and practicality[31]. When dealing with repeated diagnostic tests in the same group of patients, it is showed that both semiparametric and parametric techniques yield similar results based on the data of 72 pigmented lesions suspected of being malignant melanoma with diagnostic scoring based on repeatedly measurements of asymmetry, border irregularity, color variation, and diameter, with or without a dermoscope[32].

2.2 Survival and longitudinal analysis

2.2.1 Background

So far, the survival analysis perhaps is the most important research field, where semiparametric technique can be well applied. Survival analysis is so important that it has been fully discussed in various textbooks and statistical books. Graphically, survival analysis is a curve of survival function versus time. The most common curve is the Kaplan-Meier curve, and the log-rank test is applied to comparison between Kaplan-Meier curves. Indeed, the Kaplan-Meier curve with the log-rank test can generally answer questions of the size of fraction of a population that will survive past a certain time point and at what rate they will die.

In survival analysis, both models of the proportional hazards and the accelerated failure time are very popular. The linear relationship is considered between interest covariates and the logarithm of hazard function in the proportional hazards model as well as the logarithm of survival time in the accelerated failure time model. To evaluate the proportional hazards assumption, a rule of thumb is commonly to use the nonparametric Kaplan-Meier survival curve. It is expected that the logarithm of cumulative hazard function is parallel if the proportional hazards assumption holds.

Sometimes, longitudinal studies are considered as survival analysis, which is partially true because the Kaplan-Meier curve, the proportional hazards regression and accelerated failure time regression are used widely in longitudinal study. However, there is a little difference between survival analysis and longitudinal study, that is, the proportional hazards regression and accelerated failure time regression are confined to traditional survival (i. e., single-event) data and time invariant covariates[33]. However, there may be the subjects with dependent failure time if the data might be sampled in clusters. Also, if several events might be potentially experienced in each study, there may be the subject with multivariate (chapters 8～10 in Reference[34]). The data are often collected on occurrence time of a certain event and on repeated measures of a response variable in longitudinal studies.

2.2.2 Applicable circumstances for semiparametric technique

The Kaplan-Meier curve and its log-rank test are most suitable when examined factors can be categorized, for example, treated versus placebo. If the examined factors possess continuous character, then the proportional hazards regression or accelerated failure time regression has to be used. Moreover, if one would like to know whether multiple causes need to be taken into account in survival analysis and how a particular circumstance or characteristic can change the odds of survival, then one needs to consider either the proportional hazards regression or accelerated failure time regression because these two models describe not only a simple relationship between survival function and time, but also many covariates behind survival curve, which would have impact on the survival function.

The hazard function for the Cox proportional hazards regression has the following formulae[35]: $\lambda(t \mid X)=\lambda_0(t)\exp(\beta_1 X_1+\cdots+\beta_p X_p)=\lambda_0(t)\exp(\beta' X)$ where t is time, X_i is the i-th covariate, and β is regression coefficient. The model indicates that the hazard function, $\lambda(t)$, of the failure time conditional on a set of possibly time varying covariates, X, is the product of an arbitrary base-line hazard function, $\lambda_0(t)$, and a regression function of the covariates, X_i. From this equation, we can partly answer the question of which component is semipara-

metric in proportional hazard model. That is various covariates that would have different types of distributions: Some can be normally distributed whereas some can be non-normally distributed. In other words, we have incomplete information on covariates. The mixture of data distributions for various covariates is the essence of semiparametric technique because ordinary regression requires the regressed covariates normally distributed. Actually, Gorfine et al[36]. assumed a gamma distribution for semiparametric with data of breast cancer family control full-term pregnancy as covariate. Although being a flexible model due to its unspecified baseline hazard function, sometimes the proportional hazard model seems inflexible, because the proportional hazard and proportional odds models[37] are linear transformation of the failure time to covariates[38]. On the other hand, the accelerated failure time regression is generally regarded as a parametric model[39], which is that its probability distribution should be specified. Therefore, the accelerated failure time regression works robustly by omitting of covariates, and is less affected by the choice of probability distribution[40,41]. However, semiparametric technique has been added into the accelerated failure time regression as early as 1980s[42~50]. For example, an accelerated failure time partial linear regression can be expressed as follows: $\log T = X^T\beta + f(u) + \varepsilon$, where T is survival time, β is a vector of regression coefficient, X is a vector of covariates, u is a covariate, f is function and εs are independent error terms with a normal distribution[50]. So we can see that the application of semiparametric technique in survival analysis is mainly related to models, in particularly is related to the distribution of baseline hazard function as well as time-changing covariate effects.

Recently, there are huge interests in joint modeling, where the dependence on a common set of random effects is assumed for the failure time and repeated measures. To assess the joint effects of base-line covariates, such models have been used in the analysis of repeated measures to adjust for informative drop-out and on the failure time to study the effects of potentially mis-measured time varying covariates. For analyzing the failure time, the proportional hazards model can be used with normal random effects while for analyzing repeated measures there is the linear mixed model, which is confined to continuous repeated measures. Also, the transformation of the response variable is assumed to be known. However, misspecification of transformation is highly non-robust in random-effects models, therefore semiparametric models (e. g. linear transformation models) are employed for continuous repeated measures, in order to avoid a parametric specification of the distribution or transformation. In addition, a semiparametric linear mixed model was proposed for continuous repeated measures, where the response variable is completely transformed to be unspecified[51].

It requires the conditional distribution of censoring given covariates to be modeled. The context can actually be cast as a missing data problem. It is customary to assume that the data are independent given the parameters. Zeng and Lin proposed a semiparametric linear mixed model with right censored data[52,53], $\tilde{H}(Y_{ij}) = \alpha^T X_{ij} + b_i^T \tilde{X}_{ij} + \varepsilon_{ij}$, where $\tilde{H}$ is an unknown increasing function, α is a set of regression coefficients, b_i is a set of random effects with density $f(b;\gamma)$, $\tilde{X}_{ij}$ is typically a subset of X_{ij}, ε_{ij} is independent errors with density $f\varepsilon$ because of censoring time. Here, if the transformation function, $\tilde{H}$, is specified, this equation reduces to the conventional (parametric) linear mixed model, whereas leaving the form of $\tilde{H}$ unspecified is in agreement with the semi-parametric feature of the transformation models for event times. The authors suggested $\tilde{\Lambda}(y) = \exp\{\tilde{H}(y)\}$, which is somewhat similar to $S(t) = \exp(-\Lambda(t))$.

In longitudinal studies, semi-parametric regression models are highly useful to estimate covariate effects on potentially censored responses, such as repeated measures and failure times. In addition to right censoring, the failure times are shown to left truncation in some applications. To consider both right censoring and left truncation, Zeng and Lin[53] defined $N(t)$ and $R(t)$ as the event number observed

by time t and the at-risk indicator at time t, accommodating general censoring/truncation patterns. On the other hand, clustered failure times arise often in studies on medicine and epidemiology, for instance, disease onset times of twins (with time expressed in terms of age), multiple recurrence times of infections on an individual, or time to blindness for the two eyes within an individual.

2.2.3 Comparison studies

Some studies show the advantage of semiparametric technique over parametric and nonparametric techniques in survival analysis. Zou et al. applied semiparametric technique to the data from 1583 breast cancer patients with exclusion of 3.4% black women[48] and showed that parametric technique, proportional hazards model, could not work for this dataset because it was hard to find a parallel relationship between survival curves and marital status. The equation used by authors for this group of data is $\log T=\beta_1\times\text{regional}+\beta_2\times\text{distan}\ t+\beta_3\times\text{married}+\beta_4\times\text{other}+f(\text{age})+\varepsilon$, which includes covariates with non-normal distributions or with unspecified distributions. Zhang et al.[49] showed that semiparametric technique in mixture cure model is useful in cancer treatment when the therapeutic effect increases gradually from zero, and the authors used the data from 1584 breast diagnosed patients and the model with semi-parametric technique produced the survival curve is closer to the Kaplan-Meier survival curve but with indication of covariates effects.

The consideration in above cited studies was mainly concerned with the distribution of covariates, which is an important aspect for the application of semiparametric technique. In the study by Ghosh et al[50], the concern was given to the distributional assumptions for the values of slopes in survival analysis, and the application of semiparametric Bayesian method could reduce this concern. Thus the distributions are not necessary to be identically normally distributed but can be a mixture of various distributions. Following this, the authors studied 2039 patients diagnosed with distant testicular cancer with age as a covariate, and they concluded that the fitting using semiparametric technique is better than that using parametric technique.

Although many studies with clinical indicators favor semiparametric technique in survival analysis, covariates with complex indicators such as haplotype sometimes produced different results. Scheike et al. compared semiparametric technique with nonparametric technique to investigate possible effects of haplotype in platelet activating factor receptor on cardiovascular events in patients with coronary artery disease including 1268 subjects and with a total of 116 deaths[54], where the difficulty was that the maximum-likelihood estimates of frequencies of 32 possible haplotypes were highly unstable, and it was not possible to assess the uncertainty directly for all 32 haplotypes, therefore the expectation-maximization algorithm was slow in the context of the semiparametric technique. The results suggest that the expectation-maximization algorithm seriously underestimates the variance of the relative risk parameters.

The inclusion of covariates was in fact not limited to clinical characters, for example, Zhao and Hanson used a nonparametric technique to cooperate geographical information as a covariate into analysis of a cohort of 5786 women diagnosed with malignant breast cancer starting in 1989 ending in 1991 with follow-up continued through the end of 2003[55]. What the authors were interested in was how survival changed geographically rather than other time effects such as age, and their study suggested that nonparametric technique improved semiparametric technique in this estimation.

2.3 Pharmacokinetics and pharmacodynamics

2.3.1 Background

Although semiparametric technique was applied to pharmacokinetics and pharmacodynamics as early as the 80s in the 20th century[56], its application is not as many as in other research fields. Even there is a tendency of decreasing the use of semiparametric technique in pharmacokinetics and pharmacodynamics. Indeed, semiparametric technique used in pharmacokinetics and pharmacodynamics is very different from its applications used in other research fields. So far there are several types of applications of semiparametric technique in pharmacokinetics and pharmacodynamics.

As a matter of fact, there are few applications of semiparametric technique in pharmacokinetics-pharmacodynamics, and the definition of semiparametric technique is somewhat different from those in other research fields. The effort to collapse the lag time between blood drug concentration and therapeutic effect appears to be the best case of application of semiparametric technique, but the lack of software package limits the application. From the sense of distributions of covariates, population approach could be considered as a semiparametric technique.

2.3.2 Applicable circumstances for semiparametric technique

The first and most widely application is to address hysteresis between blood drug concentration and drug's therapeutic effect, because it was found very long ago that a drug's therapeutic effect is later than its corresponding blood drug concentration. In other words, a drug's therapeutic effect has yet to reach its maximum when its blood concentration has already reached its peak level. The lag time between blood drug concentration and therapeutic effect leads to the difficulty in coupling pharmacokinetics and pharmacodynamics together, because pharmacokinetics in theory is a compartmental model, which accounts the relationship between blood drug concentration and time, i. e., $C(t) = A_i e^{-a_i t}$ where $C(t)$ is blood drug concentration at given time, and A and a are model parameters, and pharmacodynamics is mainly a sigmoid model or Michaelis-Menten kinetics typed model, which accounts the relationship between blood drug concentration and therapeutic effect, i. e., $\text{Effect} = \frac{\text{Effect}_{\max} C}{K + C}$ where K is model parameter. Therefore, it would appear easy to build a pharmacokinetic-pharmacodynamic model by replacing the blood drug concentration in pharmacodynamic model with the blood drug concentration computed from pharmacokinetic model, thereafter researchers and clinicians would have the relationship between time and drug's therapeutic effect as well as the relationship between time and drug's side effect, and then be able to predict therapeutic and side effects. This is particularly important for anesthetic effect for patients in operations, which was one of focuses of the earliest application of pharmacokinetic - pharmacodynamic models[57,58]. However, due to the lag time between blood drug concentration and drug's therapeutic effect, the timing of maximal therapeutic effect becomes difficult. Thus, an effect compartment was proposed to collapse this lag time because the maximal drug concentration in effect compartment could be later than that in blood[59], however, it still appeared difficult to use a compartmental model with an effect compartment to fit blood drug concentration and drug's therapeutic effect simultaneously. So this is the background for the development of semiparametric technique in pharmacokinetics and pharmacodynamics.

Technically, semiparametric technique only estimates a single parameter, *keo*, which is the elimination rate of drug from effect compartment[60~64], thus it simplified the process of estimating the parameters in compartmental model otherwise it would have a number of model parameters to fit. This is the major application of semiparametric technique in pharmacokinetics and pharmacodynamics. Such an application can be in theory better to deal with the relationship between time and drug's therapeutic effect. However, there is no specialized and user-friendly software although there was a FORTRAN written program[65], which can use semiparametric technique to minimize hysteresis and result in the first-order rate equilibrium constant (*keo*).

The second application is that semiparametric technique could be considered as a combination of parametric approach with nonparametric approach[66,67], where a nonparametric function that is estimated using penalized splines was applied to correlate in vivo and in vitro relationship to drugs. Also, there are several minor applications of semiparametric technique, such as population pharmacokinetics[68], non-compartment approach with pharmacodynamic effect[69], non-steady-state pharmacodynamic data[70] and so on. However, few follow-up studies have been found so far along those lines of research.

2.3.3 Comparison studies

Actually, there are only few studies, which were conducted to compare semiparametric technique with parametric and nonparametric techniques in pharmacokinetics-pharmacodynamics, because there is no substantial amount of publications using semiparametric technique in this field. A study using the population approach[68] compared semiparametric technique with parametric and nonparametric techniques using area under concentration-time (*AUC*), peak concentration and time to peak concentration (T_{peak}), with the noisy population data from a sparsely sampled prospectively designed trial, and the authors found that the semiparametric technique performed as good as or better than standard nonparametric technique. However, if a model is mis-specified, semiparametric technique was superior to a standard parametric technique. Yuan and Yin conducted a study[71], which shows that semiparametric technique can converge a curve estimate to what parametric technique achieves when assumption for parametric technique holds, and can converge a curve estimate to what nonparametric technique achieves when assumption for parametric technique is mis-specified. Another study, which could be considered as a pro study, was to study morphine-6-glucuronide's effect on pupil size[72] with semiparametric technique to estimate the rate constant of transfer from effect compartment to plasma in eight volunteers, and semiparametric technique worked in this study. Although the results from semiparametric technique were similar to the results from the parametric technique, semiparametric technique produced a better description of the data for few animals, where the results differed[60].

2.4 Exposure study

2.4.1 Background

The crucial thing in exposure study is to establish a relationship or a correlation or an association between exposure indicators and a certain disease. In general, this is done with regression because we can get correlation through regression. Although regression does not account for a cause-consequence relationship, it does find many exposure indicators that are associated with a certain disease. Needless to say, some exposure indicators may well grasp the essence of a natural phenomenon such as population, but may not represent the true factor that really matters a disease. There is a hidden indicator or an invisible indicator, which is associated with both measurable indicators and a disease, and this indicator is yet to be found. In such a case, latent variable regression could be a tool to uncover the hidden indicator, which bridges measurable indicators and a disease.

For example, in a study trying to find the association between traffic particles and occurrence of acute myocardial infarction [73], the authors measured nitrogen dioxide, NO_2, and primary particulate matter with aerodynamic diameter less than 2.5 μm, $PM_{2.5}$, and had the cases of acute myocardial infarction and controls. However, the authors did not consider traffic particles directly measurable because NO_2 can have several non-traffic sources. In such a case, latent variable regression is usable. Once again, latent variable regression can include many covariates, which would have very different distributions, and lay the foundation for semiparametric technique[74].

Another thing is that the measurement of pollutants often takes several periods of time, which would generate a time-varying association between pollutants and something measured over a period of time, which is the case suitable for the application of semiparametric technique. Yet, two-stage studies are another development in applying semiparametric technique in exposure studies [75~77].

2.4.2 Applicable circumstance for semiparametric technique

Exposure studies are used to conduct by means of collecting data from case-control analysis at a single time point. This is sufficiently good for a relationship between exposure intensity and numbers of disease cases in a sense of two-dimensional graph, for example. However, cumulative exposure becomes more and more important to evaluate the disease risk, and disease status also changes along the time course. This development necessarily creates a relationship between time-varying exposure intensity and numbers of disease cases, which could be con-

sidered in a sense of three-dimensional graph with a time axis at least. One might suggest that life would be easier if we would have a function describe the relationship between time and exposure intensity, however, some type of regression is still in need because it associates exposure intensity with disease risk. Therefore Bayesian approach was developed to account for the exposure derived from the past[78], which should answer the question of how both exposure history and disease status history impact on the present disease status. Indeed, this gives a feeling of Markov chain probability.

Another applicable direction is still related to a period of time of exposure. A study was conducted to evaluate the short-term effect of diurnal temperature range on emergency room admissions among elderly adults in Beijing[79]. By means of semiparametric technique, the authors used 3-, 6- and 8-day moving average of diurnal temperature range as an indicator to correspond to daily emergency room admissions for different diseases in senior citizens, and analyzed the exposure-effect association relationship between diurnal temperature range and daily emergency room admissions. Actually, there are studies[80,81] along this line of research thought, for example, the association of seasonal grass pollen with childhood asthma emergency department presentations[82], where authors used semiparametric technique to examine a short time series pollen data in terms of daily grass pollen levels with relation to mean daily emergency department attendance for asthma.

An interesting application is to include geographic information into exposure studies[83~85]. In a broad sense, this geographic information includes many pieces of environment information such as landscape[83,84], temperature[86], air quality[87], groundwater[88~90], and socioeconomic position[91]. In fact, not only time-varying variable can be considered using semiparametric technique[73] but also spatial-varying variable can be considered too[92~94]. For instance, certain factors during the first trimester of pregnancy promote cardiac congenital anomaly development[95], and semiparametric technique was used to deal with multiple pollutants and a multivariate cardiac anomaly grouping outcome jointly, i. e., the inclusion of space and time simultaneously in both the locations and the masses. In such a way, the authors identified critical weeks during pregnancy. And the results suggest semiparametric technique is better than nonparametric technique.

2.4.3 Comparison studies

Unfortunately, comparison studies in this field mostly were not conducted to compare the performance among semiparametric, parametric and nonparametric techniques, because they might not be compatible in this research field although there are indeed studies for comparisons related to direct and mixed effect models. In a study investigating acute effects of an exposure to 100 ppm 1-methoxypropanol-2 on the upper airways of human subjects[96], the authors studied 20 volunteers using both semiparametric and parametric techniques to assess residual and period effects caused by the design of study. However, the results seem compatible for both semiparametric and parametric techniques.

2.5 Genetic study

2.5.1 Background

It turns out that the applications of semiparametric technique in genetic studies are the most diverse, not only because genome studies raise many statistic issues, which can be treated with parametric, nonparametric and semiparametric techniques, but also because genomic analysis is still at its early stage of development in exploitation of all the information in genome. On the one hand, semiparametric technique's applications are too many to summarize in a review, as these applications scatter widely in literatures and are difficult to group them into different categories.

2.5.2 Applicable circumstance for semiparametric technique

A somewhat familiar direction appears similar to what is done in exposure study, that is, to find out an association between genetic and environmental factors with a certain disorder. For example, Touchette et al. studied genetic and environmental influences on daytime and nighttime sleep duration in childhood at 6, 18, 30, and 48 months of age[97], which in fact is similar to the abovementioned stud-

ies in previous section[79~82]. To determine genetic factors, the authors used 995 twins (405 being monozygotic and 586 dizygotic), to analyze daytime and nighttime sleep duration trajectories that are a time-varying factor, and semiparametric technique showed that environmental factors affect all daytime sleep duration trajectories while genetic factors mainly influence nighttime sleep duration.

Take a step further, this type of studies advances to explore associations of gene-environment interaction with a certain disease as well as a certain phenotype[98~101]. It is necessary to say that environmental factors can be well beyond what is considered in previous section. For example, demographic, behavioral and nutritional characteristics of mothers could be considered as environmental factors[101]. Actually, one problem with those environmental factors is that they distribute very differently among research subjects, which may lead to various distribution constraints. The combination of genetic and environmental factors can be expressed as follows[102], $\log it\ p(Y=1 \mid G^m, G^c, X)$, where Y is the binary case-control status, G^m and G^c are the genotype pairs of a single-nucleotide polymorphism for mother and child, and X is the vector of environmental variables. Indeed, this approach also considers the gene-environment interaction, which appears

$$\log it\ p(Y=1 \mid G^m, G^c, X) = \beta_0 + \beta_1 G^m + \beta_2 G^c + \beta_3 X + \beta_I G^c \times X,$$

where $\beta_I G^c \times X$ is the gene-environment interaction.

A very important research direction in genetic study is linked to microarray data[103~107], of which a more technically oriented method may relate to the optimal discovery procedure (ODP) that is a statistic designed for large-scale significant testing in analysis of gene expression microarrays[108~110]. This optimal discovery procedure is somewhat based on Neyman-Pearson definition[111]

$$\frac{\text{probability of data under alternative distribution}}{\text{probability of data under null distribution}},$$

therefore the optimal discovery procedure has a similar formula

$$\frac{\text{sum of probability of data}_j \text{ under each true alternative distribution}}{\text{sum of probability of data}_j \text{ under true null distribution}}.$$

In the case of gene expression microarrays, the optimal discovery procedure has a statistic as $\frac{g_{m_0+1}(x)+g_{m_0+2}(x)+\ldots+g_m(x)}{f_1(x)+f_2(x)+\ldots+f_{m_0}(x)}$, where significance test i has null probability density function f_i and alternative density g_i. So one needs to know the distribution associated with each test as well as the null or alternative is true for each test. At this point, semiparametric technique was proposed in order that prior distribution and posterior distribution can be managed [112]

2.5.3 Comparison studies

Ajaz et al analyzed the association of polymorphisms in the methylene tetrahydrofolate reductase gene with the renal cell carcinomas, and found that the combined genotype significantly increased the risk of the tumors. Their study demonstrated that a semiparametric expectation - maximization - based haplotype analysis gave an evident result as that obtained from the combined genotype analysis [113].

In vaccine efficacy trials, an intriguing objective is to estimate the vaccine effect to prevent infection in terms of the genetic distance between the exposing viral strain and the vaccine constructed strain, which can be evaluate by using the continuous mark- specific proportional hazards model. This model can evaluate mark-specific vaccine efficacy with adjustment for covariate effects, however, the missingness of mark variable in failures was not accounted. Sun and Gilbert developed two consistent estimation approaches based on the inverse probability weighted (IPW) of the complete-case estimator, and found that the augmented IPW estimator is more efficient because of its double robustness property [114]. In practice, the auxiliary should be applied to complete cases, thus the use of nonparametric density estimation is needed otherwise a reasonable parametric model is specified. However, the definition of early and later viruses provides a univariate mark that can be well-predicted, which reveals benefits for analyzing the effects of vaccines.

2.6 Other research fields

Simeparametric technique still has applications to different biomedical researches, which however frequently overlap one another. For example, studies on breast cancer could be classified as epidemiologi-

cal studies as well as cancer studies with respect to the research fields of interest although their objective could be the same. Therefore, there are many applications, which need to be mentioned. To minimize the size of this review, we do not detail the applications of semiparamtric technique in biomedical settings with respect to covariates versus diseases. A particular application, which is worth mentioning, is the medical policy, where the costs of medical treatment are the focus[115~122].

3 Takeaway message on semiparametric technique

Until this point in this review, we can observe several advantages of semiparametric technique over parametric and nonparametric techniques. As all these techniques go through a modeling process, we can start our elaborations along modeling line.

3.1 Dataset

Obviously, semiparametric technique does not care whether or not a dataset is normally distributed, even length-biased sampling[123]. This would allow to more pieces of prior information to enter a model, and to make estimation and inference more precise.

3.2 Covariates

The most impressive advantage of semiparametric technique nevertheless is to consider covariates, including time- and space-varying covariates.

3.3 Control cases

For a natural phenomenon, it oftentimes has difficulty to set control study. When using semiparametric technique, the time tendency can be considered. Information about the probability of disease can be incorporated in the estimation procedure to improve quality of parameter estimates, which cannot be done in conventional case-control studies[124].

3.4 Statistical methods with semiparametric technique

So far semiparametric technique is used in many statistical methods, for examples, semiparametric Cox regression model, semiparametric Marshall-Olkin models, semi-parametric Poisson regression, semiparametric accelerated failure time frailty model[125], double-robust semiparametric estimator[126], semiparametric stochastic modeling[127], semiparametric structural equation model [128,129], efficient semiparametric mean-association estimation[130], random effects quantile regression model[131], semiparametric smoothing[132], the semiparametric varying-coefficient partially linear model[133], truncated Dirichlet process[134], Copula model[135], collaborative double robust targeted maximum likelihood estimation[136], overlapping scanning windows[137], SIMEX[138], semiparametric maximum likelihood method[139], the jackknife resampling method[140], etc.

3.5 Software

Actually, software packages have been recently reviewed[141], but a free software package is available: R statistical software Comprehensive R Archive Network (CRAN) mirror site[142~144].

References:

[1] US National Library of Medicine National Institutes of Health. PubMed[EB/OL]. [2013-03-15]. http://www.ncbi.nlm.nih.gov/pubmed.

[2] Wikimedia Foundation Inc. Wikipedia, the free encyclopedia[EB/OL]. [2013-03-15]. http://en.wikipedia.org/wiki/Semiparametric_model.

[3] Ruppert D, Wand M P, Carroll R J. Semiparametric regression during 2003-2007[J]. Electron J Stat, 2009, 3: 1193-1256.

[4] Green D M, Swets J A. Signal Detection Theory and Psychophysics[M]. New York: Wiley, 1966.

[5] Zweig M H, Campbell G. Receiver-operator characteristic plots: A fundamental evaluation tool in clinical medicine[J]. Clinical Chemistry, 1993, 39: 561-577.

[6] Hanley J A. Receiver operating characteristic (ROC) methodology: The state of the art [J]. Critical Reviews in Diagnostic Imaging, 1989, 29: 307-335.

[7] Baker S G. The central role of receiver operating characteristic curves in evaluating tests for the early detection of cancer [J]. Journal of the National Cancer Institute, 2003, 95: 511-555.

[8] Pepe M S. The Statistical Evaluation of Medical Tests for Classification and Prediction [M]. New York: Oxford University Press, 2003.

[9] Zhou X H, Obuchowski N A, McClish D K. Statistical Methods in Diagnostic Medicine [M]. New York: John Wiley & Sons, 2002.

[10] Swets J A, Dawes R M, Monahan J. Better decisions through science[J]. Scientific American, 2000, 283: 82-87.

[11] Pepe M S. An interpretation for the ROC curve and inference using GLM procedures [J]. Biometrics, 2000, 56:352-359.

[12] Colak E, Mutlu F, Bal C, et al. Comparison of semiparametric, parametric, and nonparametric ROC analysis for continuous diagnostic tests using a simulation study and acute coronary syndrome data[J]. Computational and Mathematical Methods in Medicine, 2012, 2012: 698320-698326.

[13] Sanchez-Marin F J, Padilla-Medina J A. Alternative performance index to analyze receiver operating characteristic data under the exponential assumption [J]. Journal of Electronic Imaging, 2006, 15:1-9.

[14] Wei L J, Lin D Y, Weissfeld L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions [J]. Journal of the American Statistical Association, 1989, 84:1065-1073.

[15] Tang L, Emerson S S, Zhou X H. Nonparametric and semiparametric group sequential methods for comparing accuracy of diagnostic tests [J]. Biometrics, 2008, 64:1137-1145.

[16] Ghosh D. Incorporating monotonicity into the evaluation of a biomarker[J]. Biostatistics, 2007, 2:402-413.

[17] Wolfinger R. The GLIMMIX SAS Macro [M]. Cary, NC: SAS Institute, 1996.

[18] Zheng Y, Heagerty P J. Semiparametric estimation of time-dependent ROC curves for longitudinal marker data [J]. Biostatistics, 2004, 5:615-632.

[19] Heagerty P J, Zheng Y. Survival model predictive accuracy and ROC curves [J]. Biometrics, 2005, 61: 92-105.

[20] Liu D, Zhou X H. Semiparametric estimation of the covariate-specific ROC curve in presence of ignorable verification bias [J]. Biometrics, 2011, 67:906-916.

[21] Cai T, Pepe M S, Zheng Y, et al. The sensitivity and specificity of markers for event times[J]. Biostatistics, 2006, 7:182-197.

[22] Heagerty P J, Lumley T, Pepe M S. Time-dependent ROC curves for censored survival data and a diagnostic marker[J]. Biometrics, 2000, 56:337-344.

[23] Pepe M, Longton G, Janes H. Estimation and comparison of receiver operating characteristic curves [J]. Stata Journal, 2009, 9:1.

[24] Huang Y, Pepe M S. Assessing risk prediction models in case-control studies using semiparametric and nonparametric methods[J]. Statistics in Medicine, 2010, 29:1391-1410.

[25] Wan S, Zhang B. Smooth semiparametric receiver operating characteristic curves for continuous diagnostic tests[J]. Statistics in Medicine, 2007, 26:2565-2586.

[26] Wan S, Zhang B. Semiparametric ROC surfaces for continuous diagnostic tests based on two test measurements[J]. Statistics in Medicine, 2009, 28:2370-2383.

[27] Zou K H, Tempany C M, Fielding J R, et al. Original smooth receiver operating characteristic curve estimation from continuous data: Statistical methods for analyzing the predictive value of spiral CT of ureteral stones[J]. Academic Radiology, 1998, 5:680-687.

[28] Branscum A J, Johnson W O, Hanson T E, et al. Bayesian semiparametric ROC curve estimation and disease diagnosis[J]. Statistics in Medicine, 2008, 27: 2474-2496.

[29] Huang Y, Pepe M S. A parametric ROC model based approach for evaluating the predictiveness of continuous markers in case-control studies[J]. Biometrics, 2009, 65:1133-1144.

[30] Gu J, Ghosal S, Roy A. Bayesian bootstrap estimation of ROC curve [J]. Statistics in Medicine, 2008, 27: 5407-5420.

[31] Hajian-Tilaki K O, Hanley J A. Comparison of three methods for estimating the standard error of the area under the curve in ROC analysis of quantitative data [J]. Academic Radiology, 2002, 9:1278-1285.

[32] Zou K H. Comparison of correlated receiver operating characteristic curves derived from repeated diagnostic test data [J]. Academic Radiology, 2001, 8:225-233.

[33] Zeng D, Lin D. Efficient estimation for the accelerated failure time model [J]. Journal of the American Statistical Association, 2007, 102:1387-1396.

[34] Kalbfleisch J D, Prentice R L. The Statistical Analysis of Failure Time Data [M]. 2nd ed. Hoboken: Wiley, 2002.

[35] Wikimedia Foundation Inc. Wikipedia, the free encyclopedia [EB/OL]. [2013-03-15]. http://en.wikipedia.org/wiki/Proportional_hazards_models.

[36] Gorfine M, Zucker D M, Hsu L. Case-control survival analysis with a general semiparametric shared frailty model - A pseudo full likelihood approach [J]. Annals of Statistics, 2009, 37:1489-1517.

[37] Murphy S A, Rossini A J, van der Vaart A W. Maximal likelihood estimation in the proportional odds model [J]. Journal of the American Statistical Association, 1997, 92:968-976.

[38] Wikimedia Foundation Inc. Wikipedia, the free encyclopedia [EB/OL]. [2013-03-15]. http://en.wikipedia.org/wiki/Accelerated_failure_time_model.

[39] Lambert P, Collett D, Kimber A, et al. Parametric accelerated failure time models with random effects and an application to kidney transplant survival [J]. Statistics in Medicine, 2004, 23:3177-3192.

[40] Keiding N, Andersen P K, Klein J P. The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates [J]. Statistics in Medicine, 1997, 16:215-224.

[41] Finkelstein D M, Wolfe R A. A semiparametric model for regression analysis of interval-censored failure time data [J]. Biometrics, 1985, 41: 933-945.

[42] Attali P, Prod'Homme S, Pelletier G, et al. Prognostic factors in patients with hepatocellular carcinoma. Attempts for the selection of patients with prolonged survival [J]. Cancer, 1987, 59: 2108-2111.

[43] Dinse G E. Estimating tumor incidence rates in animal carcinogenicity experiments [J]. Biometrics, 1988, 44: 405-415.

[44] Tsiatis A A. Estimating regression parameters using linear rank tests for censored data[J]. Annals of Statistics, 1990, 18: 354-372.

[45] Ritov Y. Estimation in a linear regression model with censored data[J]. Annals of Statistics , 1990, 18: 303-328.

[46] Jin Z, Lin D Y, Wei L J, et al. Rank-based inference for the accelerated failure time model [J]. Biometrika, 2003, 90: 341-353.

[47] Jin Z, Lin D Y, Ying Z. On least-squares regression with censored data [J]. Biometrika, 2006, 93: 147-161.

[48] Zou Y, Zhang J, Qin G. Semiparametric accelerated failure time partial linear model and its application to breast cancer [J]. Computational Statistics & Data Analysis, 2011, 55: 1479-1487.

[49] Zhang J, Peng Y. Accelerated hazards mixture cure model[J]. Lifetime Data Analysis, 2009, 15: 455-467.

[50] Ghosh P, Huang L, Yu B, et al. Semiparametric Bayesian approaches to joinpoint regression for population-based cancer survival data[J]. Computational Statistics & Data Analysis, 2009, 53: 4073-4040.

[51] Bickel P J, Klaassen C A J, Ritov Y, et al. Efficient and Adaptive Estimation for Semiparametric Models [M]. Baltimore: Johns Hopkins University Press, 1993.

[52] Zeng D, Lin D Y. Maximum likelihood estimation in semiparametric regression models with censored data (with discussion) [J]. Journal of the Royal Statistical Society Series B-Statistical Methodology, 2007, 69: 507-564.

[53] Zeng D, Lin D Y. A general asymptotic theory for maximum likelihood estimation in semiparametric regression models with censored data [J]. Statistica Sinica, 2010, 20: 871-910.

[54] Scheike T H, Martinussen T, Silver J D. Estimating haplotype effects for survival data [J]. Biometrics, 2010, 66: 705-715.

[55] Zhao L, Hanson T E. Spatially dependent polya tree modeling for survival data [J]. Biometrics, 2011, 67: 391-403.

[56] Verotta D, Beal S L, Sheiner L B. Semiparametric approach to pharmacokinetic-pharmacodynamic data[J]. American Journal of Physiology, 1989, 256 (Pt 2): R1005-R1010.

[57] Papper E M. The pharmacokinetics of inhalation anaesthetics: Clinical applications [J]. British Journal of Anaesthesia, 1964, 36: 124-128.

[58] Butler R A. Pharmacokinetics of halothane and ether [J]. British Journal of Anaesthesia, 1964, 36: 193-199.

[59] Sheiner L B, Stanski D R, Vozeh S, et al. Simultaneous modeling of pharmacokinetics and pharmacodynamics: Application to d-tubocurarine [J]. Clinical Pharmacology and Therapeutics, 1979, 25: 358-371.

[60] Shafer S L, Varvel J R, Gronert G A. A comparison of parametric with semiparametric analysis of the concentration versus effect relationship of metocurine in dogs and pigs [J]. Journal of Pharmacokinetics and Biopharmaceutics, 1989, 17: 291-304.

[61] Chiang S T, Ermer J C, Osman M, et al. Pharmacokinetic-pharmacodynamic relationships of bromfenac in mice and humans [J]. Pharmacotherapy, 1996, 16: 1179-1187.

[62] Shi J, Lasser T, Koziol T, et al. Kinetics and dynamics of sematilide [J]. Therapeutics Drug Monitoring, 1995, 17: 437-444.

[63] Mould D R, DeFeo T M, Reele S, et al. Simultaneous modeling of the pharmacokinetics and pharmacodynamics of midazolam and diazepam [J]. Clinical Pharmacology and Therapeutics, 1995, 58: 35-43.

[64] Verotta D, Sheiner L B. Simultaneous modeling of pharmacokinetics and pharmacodynamics: An improved algorithm[J]. CABIOS, 1987, 3: 345-349.

[65] Stanski D R, Maitre P O. Population pharmacokinetics and pharmacodynamics of thiopental: The effect of age revisited [J]. Anesthesiology, 1990, 72: 412-422.

[66] Wang Y, Eskridge K M, Zhang S. Semiparametric mixed-effects analysis of PK/PD models using differential equations [J]. Journal of Pharmacokinetics and Pharmacodynamics, 2008, 35: 443-463.

[67] Pitsiu M, Sathyan G, Gupta S, et al. A semiparametric deconvolution model to establish *in vivo - in vitro* correlation applied to OROS oxybutynin [J]. Journal of Pharmaceutical Sciences, 2001, 90: 702-712.

[68] Park K, Verotta D, Blaschke T F, et al. A semiparametric method for describing noisy population pharmacokinetic data [J]. Journal of Pharmacokinetics and Biopharmaceutics, 1997, 25: 615-642.

[69] Veng-Pedersen P, Wilhelm J A, Zakszewski T B, et al. Duration of opioid antagonism by nalmefene and naloxone in the dog: An integrated pharmacokinetic/pharmacodynamic comparison [J]. Journal of Pharmaceutical Sciences, 1995, 84: 1101-1106.

[70] Verotta D, Sheiner L B. Semiparametric analysis of non-steady-state pharmacodynamic data [J]. Journal of Pharmacokinetics and Biopharmaceutics, 1991, 19: 691-

712.

[71] Yuan Y, Yin G. Dose-response curve estimation: A semiparametric mixture approach[J]. Biometrics, 2011, 67:1543-1554.

[72] Lötsch J, Skarke C, Schmidt H, et al. The transfer half-life of morphine-6-glucuronide from plasma to effect site assessed by pupil size measurement in healthy volunteers [J]. Anesthesiology, 2001, 95:1329-1338.

[73] Tonne C, Yanosky J, Gryparis A, et al. Traffic particles and occurrence of acute myocardial infarction: A case-control analysis [J]. Occupational and Environmental Medicine, 2009, 66:797-804.

[74] Reich B J, Fuentes M, Dunson D B. Bayesian spatial quantile regression [J]. Journal of American Statistical Association, 2011, 106:6-20.

[75] Xu W, Zhou H. Mixed effect regression analysis for a cluster-based two-stage outcome-auxiliary-dependent sampling design with a continuous outcome [J]. Biostatistics, 2012, 13:650-664.

[76] Wang X, Zhou H. Design and inference for cancer biomarker study with an outcome and auxiliary-dependent subsampling [J]. Biometrics, 2010, 66:502-511.

[77] Zhou H, Song R, Wu Y, et al. Statistical inference for a two-stage outcome-dependent sampling design with a continuous outcome [J]. Biometrics, 2011, 67:194-202.

[78] Bhadra D, Daniels M J, Kim S, et al. A Bayesian semiparametric approach for incorporating longitudinal information on exposure history for inference in case-control studies [J]. Biometrics, 2012, 68:361-370.

[79] Wang M Z, Zheng S, He S L, et al. The association between diurnal temperature range and emergency room admissions for cardiovascular, respiratory, digestive and genitourinary disease among the elderly: A time series study [J]. Science of Total Environment, 2013, 456-457:370-375.

[80] Ge W Z, Xu F, Zhao Z H, et al. Association between diurnal temperature range and respiratory tract infections [J]. Biomedical and Environmental Sciences, 2013, 26:222-225.

[81] Tao Y, An X, Sun Z, et al. Association between dust weather and number of admissions for patients with respiratory diseases in spring in Lanzhou [J]. Science of Total Environment, 2012, 423:8-11.

[82] Erbas B, Akram M, Dharmage S C, et al. The role of seasonal grass pollen on childhood asthma emergency department presentations [J]. Clinical and Experimental Allergy, 2012, 42:799-805.

[83] Collier B A, Groce J E, Morrison M L, et al. Predicting patch occupancy in fragmented landscapes at the range-wide scale for an endangered species: An example of an American warbler [J]. Diversity and Distribution, 2012, 18:158-167.

[84] Bonner S J, Schwarz C J. Smoothing population size estimates for time-stratified mark-recapture experiments using Bayesian P-splines [J]. Biometrics, 2011, 67:1498-1507.

[85] Nkurunziza H, Gebhardt A, Pilz J. Geo-additive modelling of malaria in Burundi [J]. Malaria Journal, 2011, 11:234-241.

[86] Likhvar V, Honda Y, Ono M. Relation between temperature and suicide mortality in Japan in the presence of other confounding factors using time-series analysis with a semiparametric approach [J]. Environmental Health and Preventive Medicine, 2011, 16:36-43.

[87] Breitner S, Stölzel M, Cyrys J, et al. Short-term mortality rates during a decade of improved air quality in Erfurt, Germany [J]. Environmental Health Perspectives, 2009, 117:448-454.

[88] Otero U B, Chor D, Carvalho M S, et al. Association between socioeconomic position in earlier and later life and age at natural menopause: Estudo Pró-Saúde, Brazil [J]. Women's Health (London England), 2011, 7:719-727.

[89] He L, Huang G H, Lu H W. Health-risk-based groundwater remediation system optimization through clusterwise linear regression [J]. Environmental Science and Technology, 2008, 42:9237-9243.

[90] Wahlin K, Grimvall A. Roadmap for assessing regional trends in groundwater quality [J]. Environmental Monitoring and Assessment, 2010, 165:217-231.

[91] Wagner S E, Burch J B, Bottai M, et al. Groundwater uranium and cancer incidence in South Carolina [J]. Cancer Causes and Control, 2011, 22:41-50.

[92] Pollice A, Jona Lasinio G. Spatiotemporal analysis of the PM_{10} concentration over the Taranto area [J]. Environmental Monitoring and Assessment, 2010, 162:177-190.

[93] Fink D, Hochachka W M, Zuckerberg B, et al. Spatiotemporal exploratory models for broad-scale survey data [J]. Ecological Applications, 2010, 20:2131-2147.

[94] Yue Y R, Loh J M. Bayesian semiparametric intensity estimation for inhomogeneous spatial point processes [J]. Biometrics, 2011, 67:937-946.

[95] Warren J, Fuentes M, Herring A, et al. Bayesian spatialtemporal model for cardiac congenital anomalies and ambient air pollution risk assessment [J]. Environmetrics, 2012, 23:673-684.

[96] Muttray A, Gosepath J, Brieger J, et al. Acute effects of an exposure to 100ppm 1-methoxypropanol-2 on the upper airways of human subjects [J]. Toxicological Letters, 2013, 220:187-192.

[97] Touchette E, Dionne G, Forget-Dubois N, et al. Genetic and environmental influences on daytime and nighttime sleep duration in early childhood [J]. Pediatrics, 2013,

131:e1874-e1880.

[98] Li Y, Graubard B I. Pseudo semiparametric maximum likelihood estimation exploiting gene environment independence for population-based case-control studies with complex samples [J]. Biostatistics, 2012, 13: 711-723.

[99] Fardo D W, Liu J, Demeo D L, et al. Gene-environment interaction testing in family-based association studies with phenotypically ascertained samples: A causal inference approach [J]. Biostatistics, 2012, 13: 468-481.

[100] Cornelis M C, Tchetgen E J, Liang L, et al. Gene-environment interactions in genome-wide association studies: A comparative study of tests applied to empirical studies of type 2 diabetes [J]. American Journal of Epidemiology, 2012, 175: 191-202.

[101] Maity A, Carroll R J, Mammen E, et al. Testing in semiparametric models with interaction, with applications to gene-environment interactions [J]. Journal of the Royal Statistical Society, Series B Statistical Methodology, 2009, 71: 75-96.

[102] Chen J, Lin D, Hochner H. Semiparametric maximum likelihood methods for analyzing genetic and environmental effects with case-control mother-child pair data [J]. Biometrics, 2012, 68: 869-877.

[103] Diao G, Lin D Y. Semiparametric methods for genomewide linkage analysis of human gene expression data [J]. BMC Proceedings, 2007, 1 (Suppl 1): S83.

[104] Yuan A, He W. Semiparametric clustering method for microarray data analysis [J]. Journal of Bioinformatics and Computational Biology, 2008, 6: 261-282.

[105] Zou F, Huang H, Ibrahim J G. A semiparametric Bayesian approach for estimating the gene expression distribution [J]. Journal of Biopharmaceutical Statistics, 2010, 20: 267-280.

[106] Lin D Y, Zeng D. Correcting for population stratification in genomewide association studies [J]. Journal of American Statistical Association, 2011, 106: 997-1008.

[107] Kim I, Pang H, Zhao H. Bayesian semiparametric regression models for evaluating pathway effects on continuous and binary clinical outcomes [J]. Statistics in Medicine, 2012, 31: 1633-1651.

[108] Storey J D. The optimal discovery procedure: A new approach to simultaneous significance testing [EB/OL]. UW Biostatistics Working Paper Series, 2005, Working Paper 259[2013-03-15]. http://biostats.bepress.com/uwbiostat/paper259.

[109] Storey J D. The optimal discovery procedure: A new approach to simultaneous significance testing [J]. Journal of the Royal Statistical Society, Series B Statistical Methodology, 2007, 69: 347-368.

[110] Storey J D, Dai J Y, Leek J T. The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments [J]. Biostatistics, 2007, 8: 414-432.

[111] Neyman J, Pearson E S. On the problem of the most efficient tests of statistical hypotheses [J]. Philosophical Transactions of the Royal Society, 1933, 231: 289-337.

[112] Noma H, Matsui S. An empirical Bayes optimal discovery procedure based on simeparametric hierarchical mixture models [J]. Computational and Mathematical Methods in Medicine, 2013: 568480-568488. doi: 10.1155/293/568480.

[113] Ajaz S, Khaliq S, Hashmi A, et al. Polymorphisms in the methylene tetrahydrofolate reductase gene and their unique combinations are associated with an increased susceptibility to the renal cancers [J]. Genetic Testing and Molecular Biomarkers, 2012, 16: 346-352.

[114] Sun Y, Gilbert P B. Estimation of stratified mark-specific proportional hazards models with missing marks [J]. Scandinavian Journal of Statistics Theory and Applications, 2012, 39: 34-52.

[115] Frank R G, Lave J R. A comparison of hospital responses to reimbursement policies for Medicaid psychiatric patients [J]. Rand Journal of Economics, 1989, 20: 588-600.

[116] Smith L R, Milano C A, Molter B S, et al. Preoperative determinants of postoperative costs associated with coronary artery bypass graft surgery [J]. Circulation, 1994, 90 (Pt 2): 11124-11128.

[117] Herwartz H, Theilen B. The determinants of healthcare expenditure: New results from semiparametric estimation [J]. Health Economics, 2010, 19: 964-978.

[118] Pan W, Zeng D. Estimating mean cost using auxiliary covariates [J]. Biometrics, 2011, 67: 996-1006.

[119] Boyd-Ball A J, Dishion T J, Myers M W, et al. Predicting American Indian adolescent substance use trajectories following inpatient treatment [J]. Journal of Ethnicity in Substance Abuse, 2011, 10: 181-201.

[120] Hung M C, Lu H M, Chen L, et al. Life expectancies and incidence rates of patients under prolonged mechanical ventilation: A population-based study during 1998 to 2007 in Taiwan [J]. Critical Care, 2011, 15: R107-R105.

[121] Dawson R, Lavori P W. Efficient design and inference for multistage randomized trials of individualized treatment policies [J]. Biostatistics, 2012, 13: 142-152.

[122] Chi F W, Campbell C I, Sterling S, et al. Twelve-step attendance trajectories over 7 years among adolescents entering substance use treatment in an integrated health plan [J]. Addiction, 2012, 107: 933-942.

[123] Qin J, Ning J, Liu H, et al. Maximum likelihood esti-

mations and EM algorithms with length-biased data [J]. Journal of American Statistical Association, 2011,106:1434-1449.

[124] Liu X, Wang L, Liang H. Estimation and variable selection for semiparametric additive partial linear models (SS-09-140) [J]. Statistica Sinica, 2011, 21:1225-1248.

[125] Johnson L M, Strawderman R L. A smoothing expectation and substitution algorithm for the semiparametric accelerated failure time frailty model [J]. Statistics in Medicine, 2012, 31:2335-2358.

[126] Zhang M, Schaubel D E. Double-robust semiparametric estimator for differences in restricted mean lifetimes in observational studies [J]. Biometrics, 2012, 68:999-1009.

[127] Zhu B, Taylor J M, Song P X. Semiparametric stochastic modeling of the rate function in longitudinal studies [J]. Journal of American Statistical Association, 2011, 106:1485-1495.

[128] Song X Y, Xia Y M, Lee S Y. Bayesian semiparametric analysis of structural equation models with mixed continuous and unordered categorical variables [J]. Statistics in Medicine, 2009, 28:2253-2276.

[129] Guo R, Zhu H, Chow S M, et al. Bayesian lasso for semiparametric structural equation models[J]. Biometrics, 2012, 68:567-577.

[130] Chen Z, Shi N Z, Gao W, et al. Efficient semiparametric mean-association estimation for longitudinal binary responses [J]. Statistics in Medicine, 2012, 31:1323-1341.

[131] Kim M O, Yang Y. Semiparametric approach to a random effects quantile regression model [J]. Journal of American Statistical Association, 2011, 106: 1405-1417.

[132] Patil P N, Bagkavos D. Semiparametric smoothing of discrete failure time data [J]. Biometrical Journal, 2012, 54:5-19.

[133] Kai B, Li R, Zou H. New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models [J]. Annals of Statistics, 2011, 39:305-332.

[134] Chow S M, Tang N, Yuan Y, et al. Bayesian estimation of semiparametric nonlinear dynamic factor analysis models using the Dirichlet process prior [J]. British Journal of Mathematical and Statistical Psychology, 2011, 64 (Pt 1):69-106.

[135] Yilmaz Y E, Lawless J F. Likelihood ratio procedures and tests of fit in parametric and semiparametric copula models with censored data [J]. Lifetime Data Analysis, 2011, 17:386-408.

[136] van der Laan M J, Gruber S. Collaborative double robust targeted maximum likelihood estimation [J]. International Journal of Biostatistics, 2010, 6: Article 17.

[137] Wen S, Kedem B. A semiparametric cluster detection method—a comprehensive power comparison with Kulldorff's method[J]. International journal of Health Geographics, 2009, 8:73-89.

[138] Apanasovich T V, Carroll R J, Maity A. SIMEX and standard error estimation in semiparametric measurement error models [J]. Electronic Journal of Statistics, 2009, 3:318-348.

[139] Zhao Y, Lawless J F, McLeish D L. Likelihood methods for regression models with expensive variables missing by design [J]. Biometrical Journal, 2009, 51: 123-136.

[140] Tunes-da-Silva G, Pedroso-de-Lima A C, Sen P K. A semi-Markov multistate model for estimation of the mean quality-adjusted survival for non-progressive processes [J]. Lifetime Data Analysis, 2009, 15:216-240.

[141] Hirsch K, Wienke A. Software for semiparametric shared gamma and log-normal frailty models: An overview [J]. Computer Methods and Programs in Biomedicine, 2012, 107:582-597.

[142] The Institute for Statistics and Mathematics of WU (Wirtschaftsuniversität Wien). The Comprehensive R Archive Network[EB/OL]. [2013-03-15]. http://cran. r-project. org/.

[143] Scrucca L, Santucci A, Aversa F. Regression modeling of competing risk using R: An in depth guide for clinicians[J]. Bone Marrow Transplantation, 2010, 45: 1388-1395.

[144] Stare J, Perme M P, Henderson R. A measure of explained variation for event history data [J]. Biometrics, 2011, 67:750-759.

(责任编辑:尹　闯)