行政院國家科學委員會專題研究計畫 成果報告

具一般發生原因分配之競爭型風險模型
研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 97-2118-M-032-006-
執 行 期 間 ：97 年 08 月 01 日至 98 年 07 月 31 日
執 行 單 位 ：淡江大學數學系

計 畫 主 持 人：溫啟仲
共 同 主 持 人：張憶壽
計畫參與人員：碩士班研究生-兼任助理人員：鄭欣怡
碩士班研究生-兼任助理人員：王國龍

報 告 附 件 ：出席國際會議研究心得報告及發表論文

公 開 資 訊 ：本計畫可公開查詢

中 華 民 國 98 年 10 月 12 日

行政院國家科學委員會補助專題研究計畫 ■ 成 果 報 告 □期中進度報告

Competing Risk Models with General Marginal Distribution of Causes

(具一般發生原因分配之競爭型風險模型)

計畫類別：■ 個別型計畫　□ 整合型計畫
計畫編號：NSC　97－2118－M －032 －006 －
執行期間：　97 年 08 月 01 日至 98 年 07 月 31 日

計畫主持人：溫 啟 仲
共同主持人：張 憶 壽
計畫參與人員： 鄭欣怡、王國龍

成果報告類型(依經費核定清單規定繳交)：■精簡報告　□完整報告

本成果報告包括以下應繳交之附件：
□赴國外出差或研習心得報告一份
□赴大陸地區出差或研習心得報告一份
■出席國際學術會議心得報告及發表之論文各一份
□國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、列管計畫及下列情形者外，得立即公開查詢
□涉及專利或其他智慧財產權，□一年□二年後可公開查詢

執行單位：

中　華　民　國　98　年　10　月　12　日

計畫名稱：

Competing Risk Models with General Marginal Distribution of Causes
(具一般發生原因分配之競爭型風險模型)

中文摘要：

對於競爭型風險數據，此研究提出一個以伯氏多項式建構發生原因分配和以正比風險模型建構條件存活函數之半母數混合模型。我們將研究此競爭型風險模型之無母數最大概然估計。我們建立此模型的漸近剖析概然函數理論；並提供一有效計算無母數最大概然估計的演算法。模擬試驗說明無母數最大概然估計的良好數值表現，以及這個以伯氏多項式建構發生原因分配的模型比傳統以羅氏函數建構發生原因分配的模型更具廣泛應用性。

英文摘要：

This study proposes a semiparametric mixture model for competing risks data, in which the failure time has proportional hazard rate conditional on cause type and the marginal distribution of cause conditional on covariates is described by Bernstein polynomials. We establish an asymptotic profile likelihood theory for this model and provide efficient algorithms for the computation of the nonparametric maximum likelihood estimate (NPMLE). The simulation studies indicate that the NPMLE perform excellently and that the Bernstein polynomial based model is more flexible than the popular logistic function based model.

關鍵詞

Bernstein polynomials; competing risks model; distribution of the failure type; 伯氏多項式; 競爭型風險模型; 發生原因分配

報告內容

一 前言、研究目的、文獻探討

Survival analysis deal with data measured from a specific time of origin until a specific endpoint. Competing risks models deal with survival data in which the endpoint consists of several distinct events of interests; these events are called causes or types of failure. One important and popular approach to competing risks data starts with characterizing the joint distribution of the failure time and type in terms of cause- specific hazard functions; see Prentice *et al*. (1987). In order to assess the direct effect of a covariate on the cumulative risk of a particular type, Larson and Dinse (1985) proposed a mixture model which incorporates covariates into a multinomial logistic model for the marginal distribution of failure type and parametric specifications of the conditional distribution of time to failure, given failure type. Fine (1999) proposes to analyze competing risks data by transformation models, in which the cumulative incidence function is decomposed into two parts: conditional survival distributions, given failure type, and the marginal distribution of the failure types.

Let $T_i \geq 0$, $C_i \geq 0$, $Z_i \in \Re^d$, and W_i be the time to event, the censoring time, the covariate, and the failure type of the *i*-th individual. Here $W_i \in \{1,2\}$ for simplicity. Larson and Dinse (1985) and Fine (1999) assume the marginal probability of the failure from cause 1, for example, satisfies

$$P(W_i = 1 \mid Z_i = z) = g(\alpha_1 + \alpha_2^T z) \qquad (1)$$

for some $\alpha_1 \in \Re^1$ and $\alpha_2 \in \Re^d$, and a known, positive and increasing function g. One popular choice of this function g is the logistic function. Much as this approach is convenient, popular and useful, it introduces certain constraints. Suppose the dimension *d* is 1 and Z_i is age, then (1) postulates that age-specific failure rate is a monotone function of age; it might seem desirable to have a model that does not impose monotonicity, if monotonicity is not suggested by substantive knowledge. Suppose the dimension *d* is 2, Z_{i1} is age and Z_{i2} is gender, then (1) postulates that either age-specific failure rate for female is always greater than that for male for any age group, or always smaller than that for male for any age group; it does not allow, for example, for younger people, female has larger failure rate than male, while, for older people, female has smaller failure rate than male. The purpose of this paper is to indicate that Bernstein polynomials provide a useful tool to model the marginal distribution of the causes conditional on covariates, without the aforementioned drawbacks.

A function *f* defined by $f(z)=\sum_{j=0}^{J} a_j C_j^J z^j (1-z)^{J-j}$ is called a Bernstein polynomial. Our approach capitalizes on the fact that many geometric properties of a Bernstein polynomial on [0,1] can be read off from its coefficients, and a continuous function satisfying certain shape restrictions can often be approximated by Bernstein polynomials with coefficients satisfying likewise conditions. One simple, yet useful such example is as follows. If $\{a_0,\cdots a_J\}\subset[0,1]$, then $f(z)\subset[0,1]$ for every *z* in [0,1]; a continuous function on [0,1] with values in [0,1] can be approximated by Bernstein polynomials with coefficients contained in [0,1], whose proof can be obtained by the same arguments in Chang *et al*. (2005) or Chang *et al*. (2007a). In fact, this example motivates the following model (2).

Suppose that $Z_i=(Z_{i1},Z_{i2})$ consists of a (0,1)-valued continuous covariate Z_{i1} and a categorical covariate Z_{i2} with *K* categories. We assume in this study that for some $\alpha_c=(\alpha_{01},\alpha_{11},\cdots,\alpha_{J1},\cdots,\alpha_{0K},\alpha_{1K},\cdots,\alpha_{JK})^T\in(0,1)^{(J+1)K}$,

$$P(W_i=1\mid Z_i)=\sum_{j=0}^{J}\Big(\prod_{k=0}^{K}\alpha_{jk}^{[Z_{i2}=k]}\Big)C_j^J Z_{i1}^j(1-Z_{i1})^{J-j}, \qquad (2)$$

which will be denoted by $\alpha(Z_i)$. Here $[Z_{i2}=k]$ is a shorthand for the indicator $1_{[Z_{i2}=k]}$; thus it is either 0 or 1 and it is 1 if and only if $Z_{i2}=k$. Expression (2) postulates that conditional on $Z_{i2}=k$; the marginal distribution of failure type 1 is a Bernstein polynomial. The flexibility of model (2) lies in the fact that every continuous function can be approximated by Bernstein polynomials.

We note that although both model (1) and (2) are in the category of the so called direct approach to modelling the cumulative incidence function in the sense of Jeong and Fine (2006, 2007), our approach based on (2) is different from Jeong and Fine (2006, 2007), in which the right hand side of (1) is expressed in terms of Gompertz distribution. We also note that the approach based on (1) to competing risks data is an important and standard component in the cure models; see, for example, Farewell (1997), Kuk (1992), Taylor (1995), Betensky and Schoenfeld (2001), Peng (2003), Lu and Ying (2004) and Fang et al. (2005), among others. Our approach based on (2) might also be useful in examining these cure models.

二 研究方法

To simplify the presentation of models, we assume $d\le 2$. We note that *K* = 1 in expression (2) when *d* = 1. We assume that the marginal probability of the failure from

cause 1 is given by (2), and assume that, for *j=1, 2* and *z* in $\mathbb{R}^d$, the conditional hazard of T_i at *t* given $W_i=j$ and $Z_i=z$ is

$$\lambda_j(t)\exp(\beta_j^T z),$$

where $\lambda_j(\cdot)$ is a non-negative deterministic baseline hazard function and $\beta_j \in \mathbb{R}^d$ is the regression coefficient of covariate *z*. *=1.*

With right censoring, the observed data consist of $\{(X_i,\Delta_i,Z_i)\,|\,i=1,\cdots,n\}$. Here $X_i=\min\{T_i,C_i\}$, $[\Delta_i=1]=[T_i\le C_i, W_i=1]$, $[\Delta_i=2]=[T_i\le C_i, W_i=2]$, and $[\Delta_i=3]=[T_i>C_i]$. Let $\Lambda_j(t)=\int_0^t \lambda_j(u)du$ for *j=1, 2* and $\theta=(\alpha_c^T,\beta_1^T,\beta_2^T,\Lambda_1,\Lambda_2)$. We will estimate θ based on observable data $\{(X_i,\Delta_i,Z_i)\,|\,i=1,\cdots,n\}$, assuming that $\{(T_i,C_i,Z_i,W_i)\,|\,i=1,\cdots,n\}$ are independent and identically distributed.

We now present the likelihood for the observable data. Assume that (T_i,W_i) and C_i are conditionally independent given Z_i and that the distribution of (C_i,Z_i) has nothing to do with parameter θ. Then the likelihood function is

$$\begin{aligned}L_n(\theta)=\prod_{i=1}^n &\left\{\alpha(Z_i)\Lambda_1\{X_i\}e^{\beta_1^T Z_i}\exp(-\Lambda_1(X_i)e^{\beta_1^T Z_i})\right\}^{[\Delta_i=1]}\\ &\times\left\{\alpha(Z_i)\Lambda_2\{X_i\}e^{\beta_2^T Z_i}\exp(-\Lambda_2(X_i)e^{\beta_2^T Z_i})\right\}^{[\Delta_i=2]}\\ &\times\left\{\alpha(Z_i)\exp(-\Lambda_1(X_i)e^{\beta_1^T Z_i})+(1-\alpha(Z_i))\exp(-\Lambda_2(X_i)e^{\beta_2^T Z_i})\right\}^{[\Delta_i=3]},\end{aligned}$$

where $\Lambda_j\{t\}=\Lambda_j(t)-\Lambda_j(t-)$, the jump of size of Λ_j at time *t*. The maximizer of this likelihood function is referred to as the nonparametric maximum likelihood estimator (NPMLE) of the parameter θ.

We establish the existence of NPMLE, derive the score functions, and to develop an efficient iterative algorithm for computing the NPMLE. The details are omitted here. In addition, we establish the consistency of the NPMLE, and follow Murphy & van der Vaart (2000) to get a profile likelihood theory of this model. The main results (Theorem 1~5) are given in the next section. Furthermore, we conduct simulation studies to evaluate the performance of proposed method and indicate Bernstein polynomial based model (2) is more flexible than the popular logistic function based model (1).

三 結果與討論

（1）Asymptotic properties

We establish the following asymptotic properties of the NPMLE. The proofs for them are skipped in this report. Denote the true parameter by $\zeta_0 = (\alpha_{c0}, \beta_{10}, \beta_{20})$ and $\theta_0 = (\zeta_0, \Lambda_{10}, \Lambda_{20})$. Let the profile likelihood for ζ be denoted by $pL_n(\zeta)$, which is equal to $\sup_{\Lambda_1, \Lambda_2} L_n(\theta)$. Then we have

Theorem 1 $\| \hat{\zeta}_n - \zeta_0 \|$, $\| \hat{\Lambda}_{1n} - \Lambda_{10} \|_\infty$, and $\| \hat{\Lambda}_{2n} - \Lambda_{20} \|_\infty$ converge to 0 almost surely, as n tends to infinity. Here $\| \cdot \|$ is the Euclidean norm and $\| \cdot \|_\infty$ is the uniform norm.

Theorem 2 $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges weakly to a tight Gaussian process with mean zero and some covariance process.

Theorem 3 $\sqrt{n}(\hat{\zeta}_n - \zeta_0)$ is asymptotically normal distributed with mean zero and some covariance Σ^{-1}. Here Σ^{-1} is the efficient variance of estimating ζ.

Theorem 4 For all sequences $v_n \xrightarrow{P} v \in R^{(J+1)K+2d}$ and $\rho_n \xrightarrow{P} 0$ such that $(\sqrt{n}\rho_n)^{-1} = O_P(1)$, we have $-2\dfrac{\log pL_n(\hat{\zeta}_n + \rho_n v_n) - \log pL_n(\hat{\zeta}_n)}{n\rho_n^2} \xrightarrow{P} v^T \Sigma v$

Theorem 5 Under the null hypothesis $\zeta = \zeta_0$, the likelihood ratio statistic $2\log \dfrac{pL_n(\hat{\zeta}_n)}{pL_n(\zeta_0)}$ is asymptotically chi-squared distributed with ((*J+1)K+2d*) degrees of freedom.

（2）Simulation studies

There are two studies in this section. In the first one, the marginal distribution of failure

type is a Bernstein polynomial; in the second one, the marginal distribution of failure type is not. While the first one is meant to provide information as to the performance of the method under the model assumptions, the second serves to indicate the robustness of our method.

In the first study, we set $d=1;\ J=2;\ \alpha_{c0}\equiv(\alpha_{010},\alpha_{110},\alpha_{210})^T=(0.3,0.8,0.4);$ $\beta_{10}=1,\beta_{20}=-1,\ \lambda_{10}(t)=1/4$ and $\lambda_{20}(t)=1/5$; the censoring variable C_i is exponential with parameter 25; the distribution of the covariate Z_{i1} is uniform(0,1).There are 500 replicates in this study and each replicate is a random sample with sample size 150. Based on the data from these 500 replicates, about 45% of them fail from cause 1, 38% of them fail from cause 2, and 17% of them censored.

The number of iterations in using the algorithm is set at 300, and the starting values are set as $\alpha_c=(0.5,0.5,0.5)^T,\beta_1=\beta_2=0,$ and $\Lambda_1(t)=\Lambda_2(t)=t.$

Table 1 summarizes the results of this simulation study. The second column of Table 1 lists the true values of the parameters. The third, fourth and fifth columns report respectively the sample mean, sample standard deviation (SD) and sample mean-squared error (MSE) of the 500 estimates. The sixth column reports the average of the 500 standard deviations computed by profile likelihood theory (SD^{prof}); the final column gives the 95% coverage probability (CP) based on the normal approximation (Theorem 3). It is clear from Table 1 that the numerical performance of our method is excellent.

The only difference between the model assumptions in the first study and those in the second study is that in the second study, the true probability of failure

$$P(W_i=1\mid Z_i=z)=0.2+\sin(2.5z)/2,$$

which is not a Bernstein polynomial. The data is analyzed by model (2) based on Bernstein polynomial of order 2 and by model (1) with logistic function

$$g(t)=\frac{e^t}{1+e^t}, \tag{3}$$

which was developed in Chang et al: [3]. The results of the second study are reported in Figure 1 and Table 2; entries in Table 2 bear the same meaning as those in Table 1. The mean integrated square error of $\alpha(\cdot)$ based on Bernstein polynomial model (2) and logistic model (3) are 0.0047 and 0.0132 respectively. These results seem to suggest that while both models provide quite good estimates of the relative risk coefficients in the conditional hazard rates, Bernstein polynomial model (2) with order $J=2$ provides much better estimate of the marginal probability of failure type. We note that Bernstein polynomial

model (2) with order $J = 1$ does not provide reasonable estimate (data not shown).

（3） Disscussions

We have introduced a new class of marginal distributions of competing cause using Bernstein polynomials. With suitably chosen Bernstein polynomial order, our approach is more flexible than the classical logistic model in describing the marginal distribution of failure type. We have presented an efficient algorithm and profile likelihood theory for the NPMLE.

Our simulation studies indicate that the numerical performance of the NPMLE is excellent. It might be of some interest to explore the possibility of replacing logistic function by Bernstein polynomial in other regression models. One notable and closely related class of examples would be the cure models mentioned in the introduction. Another example is the mixture model for competing risks data in which the conditional survival time, given the failure type, is a proportional odds model (Fine 1999).

Our simulation studies also suggest that when the model assumption on the marginal distribution of failure type is not correct, it is still likely that the NPMLE gives reasonable estimate of the marginal distribution of failure type. While higher order Bernstein polynomials provide better approximation to the true marginal distribution of failure type, which is not known to the scientist, they involve more parameters. Further investigation in this regard is needed.

The main idea of our approach is to capitalize on the fact that continuous functions valued in (0,1) can be approximated by Bernstein polynomials with coefficients in (0,1). Bernstein polynomials were used by Chang et al (2005) and Chang et al (2007b) to study shape restricted inference. In particular, they made use of the facts that if $a_0 \le a_1 \le \cdots \le a_J$ then *f* is increasing on [0,1] and a continuous increasing function on [0,1] can be approximated by Bernstein polynomials with non-decreasing coefficients. We note that similar statements can be made for convexity and unimodality of a continuous function. These remarks suggest that we may incorporate substantive knowledge like monotonicity, convexity or unimodality into (2) by putting suitable restrictions on the coefficients of the Bernstein polynomials. We will take up this study in the future.

參考文獻

1. R.A. Betensky and D.A. Schoenfeld, Nonparametric estimation in a cure model with random cure times. Biometrics 57 (2001), pp. 282-286.
2. I.S. Chang, L.C. Chien, C.A. Hsiung, C.C.Wen, and Y.J.Wu, Shape restricted regression with random Bernstein polynomials. In R. Liu, W. Strawderman and C.H. Zhang (eds), Complex Dataset and Inverse Problems. IMS Lecture Notes - Monograph Series 54 (2007), pp. 187-202.
3. I.S. Chang, C.A. Hsiung, C.C. Wen, Y.J. Wu, and C.C. Yang, Nonparametric maximum likelihood estimator in a semiparametric mixture model for competing risks data. Scandinavian Journal of Statistics 34 (2007), pp. 870-895.
4. I.S. Chang, C.A. Hsiung, Y.J. Wu, and C.C. Yang, Bayesian Bayesian survival analysis using Bernstein polynomials. Scandinavian Journal of Statistics 32 (2005), pp. 447-466.
5. H.B. Fang, G. Li, and J. Sun, Maximum likelihood estimation in a semiparametric logistic/proportional-hazards mixture model. Scandinavian Journal of Statistics 32 (2005), pp. 59-75.
6. V.T. Farewell, A model for a binary variable with time-censored observations. Biometrika 64 (1977), pp. 43-46.
7. J.P. Fine, Analyzing competing risks data with transformation models. Journal of the Royal Statistical Society, Series B 61 (1999), pp. 817-830.
8. J. Jeong and J.P. Fine, Direct parametric inference for the cumulative incidence function. Journal of the Royal Statistical Society, Series C 55 (2006), pp. 187-200.
9. J. Jeong and J.P. Fine, |||, Parametric regression on cumulative incidence function. Biostatistics 8 (2007), pp. 184-196.
10. A.Y.C. Kuk, A semiparametric mixture model for the analysis of competing risks data. Australian Journal of Statistics 34 (1992), pp. 169-180.
11. M.G. Larson and G.E.Dinse, A mixture model for the regression analysis of competing risks data. Journal of the Royal Statistical Society, Series C 34 (1985), pp. 201-211.
12. W. Lu and Z. Ying, On semiparametric transformation cure models. Biometrika 91 (2004), pp. 331-343.
13. Y. Peng, Fitting semiparametric cure models. Computational statistics and data analysis 41 (2003), pp. 481-490.
14. R.L.Prentice, J.D. Kalb°eisch, A.V. Peterson, N. Flournoy, V.T. Farewell, and N.E. Breslow, The analysis of failure times in the presence of competing risks. Biometrics 34 (1978), pp. 541-554.

15. J.M.G. Taylor, Semi-parametric estimation in failure time mixture models. Biometrics 51 (1995), pp.899-907.

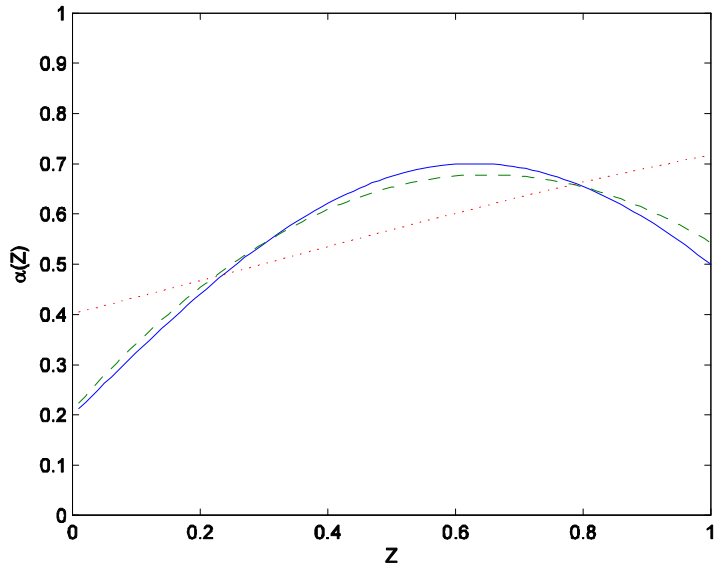
Table 1. Simulation study when the model assumptions are satisfied.

Parameter	True value	mean	SD	MSE	SD^{prof}	CP
α_{01}	0.3000	0.3126	0.1227	0.0152	0.1302	0.9460
α_{11}	0.8000	0.7922	0.1745	0.0305	0.1975	0.9680
α_{21}	0.4000	0.4075	0.1308	0.0172	0.1271	0.9360
β_1	1.0000	1.0093	0.5630	0.3171	0.5330	0.9360
β_2	-1.0000	-1.0393	0.5476	0.3014	0.4916	0.9400

Table 2. Simulation study when the Bernstein polynomial model (2) is not satisfied. Numbers in brackets are estimates using logistic model (3), others are using Bernstein polynomial model (2).

Parameter	True value	mean	SD	MSE	SD^{prof}	CP
β_1	1.0000	1.0131	0.5441	0.2962	0.5201	0.9300
		[1.0101]	[0.5667]	[0.3213]	[0.5239]	[0.9200]
β_2	-1.0000	-1.0209	0.5658	0.3206	0.5200	0.9280
		[-1.0403]	[0.5710]	[0.3277]	[0.5219]	[0.9260]

Figure 1: Simulation study when the Bernstein polynomial model (2) is not satisfied. The solid line gives true probability of failure from cause 1, $P(W_i = 1 | Z_i = z) = 0.2 + \sin(2.5z)/2$, which is not a Bernstein polynomial; the dash line gives the estimate using Bernstein polynomial model (2); the dot line gives the estimate using logistic model (3).



計畫成果自評

原計畫預期完成之工作項目(詳列如下)全部達成

- ☐ Establish the consistency and the asymptotic normality of the NPMLE.
- ☐ Provide asymptotic theories for standard error estimate and profile likelihood ratio inference.
- ☐ Develop algorithms for computing NPMLE and its asymptotic variance based on the integral characterization of the score functions.
- ☐ Conduct simulation studies to evaluate the performance of our method.
- ☐ Conduct simulation studies to indicate the flexibility of our model.

研究成果之學術或應用價值

For analyzing competing risks data, we introduce a new class of marginal distributions of competing cause using Bernstein polynomials. With suitably chosen Bernstein polynomial order, our approach is more flexible than the classical logistic model in describing the marginal distribution of failure type.

是否適合在學術期刊發表

Yes, we will submit our study to a appropriate journal.

出席國際學術會議心得報告

計畫編號	96WFD0400226
計畫名稱	具一般發生原因分配之競爭型風險模型
出國人員姓名 服務機關及職稱	溫啟仲 淡江大學數學系
會議時間地點	28 June ~01 July 2009, Seoul, Korea,
會議名稱	1st Institute of Mathematical Statistics-Asia Pacific Rim Meeting (APRM 2009)
發表論文題目	Competing risks models with general marginal distribution of causes

一、參加會議經過

在會議進行的幾天，參與了好幾個 talk sessions。其中有趣的，重要的或具啟發性的包含 Dr. Xiaoli Meng 發表之 Self-consistency: a general recipe for semi- parametric and non-parametric estimation; Dr. Thomas Lee 發表之 Further applications of the self-consistency principle for missing data problems; Dr. Jianguo Sun 發表之 Statistical analysis of informatively censored failure time data; Dr. Joe Wellner 發表之 On and off semiparametric models; Dr. Michael Kosorok 發表之 Optima; semiparametric inference under parameter constraints 等等是與個人目前研究相關的。整個會議議程內容豐富，討論也非常熱烈。此外，個人也參與了首天傍晚約兩小時的 poster session。緊接是接待會，配合以晚餐 buffet 方式輕鬆地舉行，這讓的方式讓與會者有寬裕的時間能夠認識彼此及討論交流。

二、與會心得

此次會議期間，個人也發表了近來的研究成果。對於競爭型風險數據，我們提出一個以伯氏多項式建構發生原因分配和以正比風險模型建構條件存活函數之半母數混合模型，並研究此模型之無母數最大概然估計。我們指出以伯氏多項式建構發生原因分配的模型比傳統以羅氏函數建構發生原因分配的模型更具廣泛應用性。此研究除了吸引了包括大會主席 Dr. Runze Li 等不少與會者的發問及對個人研究的肯定，也得到了他們很多寶貴的評論和建議。例如:應在模型中考慮多變量共變異(multivariate covariate)因子；並且將之應用於一個實際數據分析來說明此統計方法等等。此外，這次統計會議裡，認識了許多不同統計領域的專家學者，與他們交換個別的研究心得，收穫良多。此次與會是一次豐富且難得的經驗。