경시적자료분석 기말 보고서

이름 : 정희철

논문 : Bayesian Joint Modelling of Longitudinal and Time-to-Event Data

저자 : Maha Alsefri

선택한 논문의 제목은 "Bayesian Joint Modelling of Longitudinal and Time-to-Event Data"이다. 해당 논문은 바이오 분야에서 많이 쓰이는 Joint Modelling 에 대해 다양한 소개를 해주는데, 이 때 Bayesian 으로 정한 이유는 이미 Frequentist 방법론들은 많은 연구가 진행되었고, 논문이 나와있기 때문이라고 하였다. Joint Modelling 이란, 경시적 자료 Longitudinal Data 와 생존 자료 Time-to-Event Data 의 관계성을 이용하여 더욱 정확한 설명력을 가지는 것을 목표로 하는 분석 방법이다. 이 때, 생존 자료 Time-to-Event Data 린 어떠한 사건이 일어났는가 와 해당 사건이 언제 일어났는지에 대한 정보가 담긴 자료로, 생존분석이 중용되는 바이오 분야에서 활발히 사용되고 있다.

저자는 관련 논문들을 모두 읽은 뒤 논문 별로 Response Variable 이 일변량 또는 다변량인지, 그리고 데이터 유형이 연속형인지, Count 인지, 어떤 모델링 기법을 사용했는지, 어떤 Error Distribution 을 가정했는지, 어떤 Association Structure 을 사용했는지 등등 매우 상세부분으로 나눠 각 방법 별 비율을 제시하였다.

저자가 논문을 모으고 분류한 방법은 아래 그림 1 을 참고하면서 설명하겠다. Medline, Scopus, Web of Science 에서 "joint model AND Bayesian", "joint models AND Bayesian", "joint modelling AND Bayesian", "longitudinal AND Bayesian", "survival AND Bayesian" 키워드들을 사용하여 나오는 논문 각 179 개, 412 개, 206 개의

논문들을 찾을 수 있었고, 이 중 중복된 논문들을 제거하여 495 개를 수집하였다. 이 중, 하나씩 읽어보면서 나열된 기준들을 만족하지 못하면 폐기하여 최종적으로 75 개의 methodological 논문들이 남았고, 이 논문들을 사용하여 앞서 기술한대로 세부사항들로 분류하였다.

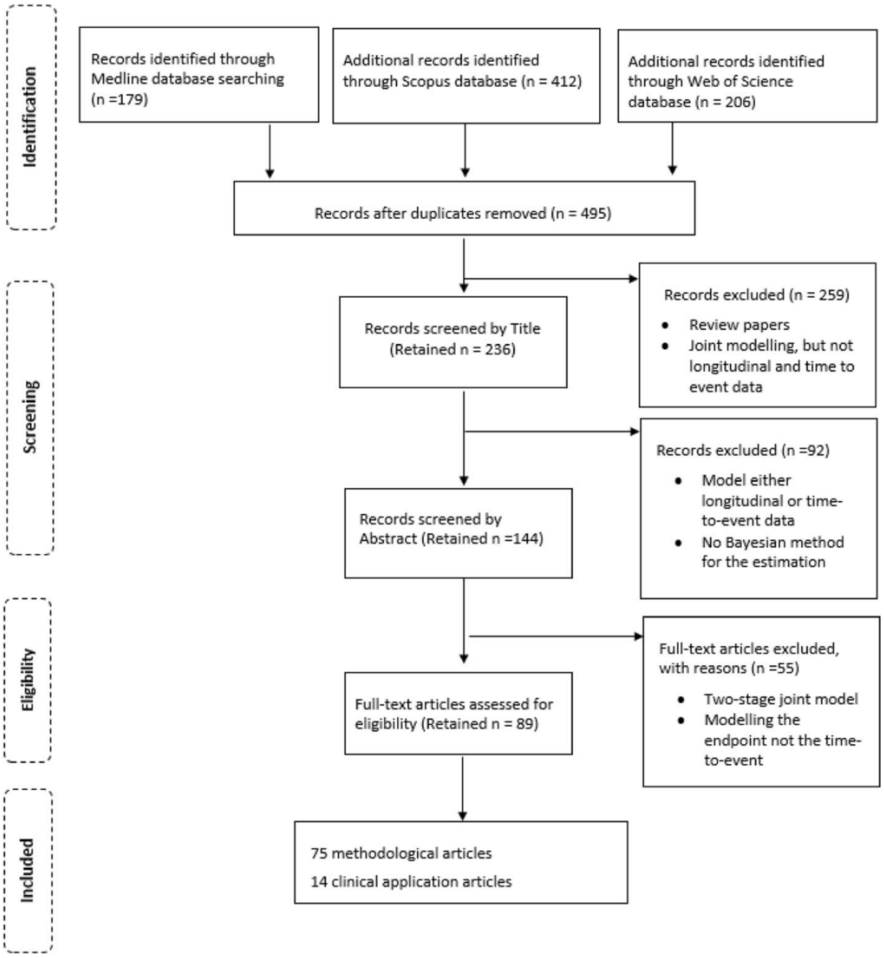


그림 1 : 논문 수집 및 분류 과정

각 세부사항마다 상당히 많은 기법 또는 가정이나 이론들이 소개되어 있었다. 논문 내용을 충실히 반영하게 된다면 단순히 표 1 의 내용을 그대로 문장으로 나열하고 설명하는 것 밖에 안되기 때문에, 세부사항 별 비율들은 간단히 표로 제공하고, 이 중

가장 많이 사용된 것과 흥미로웠던 것을 언급하고, 특히 흥미로웠던 것들에 대해서 이론적 설명을 덧붙이겠다.

	Number of articles (%)	Reference
Type of outcome		
Continuous	39(95.1%)	[13, 15–19, 22–25, 27, 33, 35, 37–40, 43, 44, 46–48, 51, 52, 55, 56, 59–62, 65, 66, 75–81]
Count	2(4.9%)	[31, 53]
Model		
GLM, NLME, SNLME, Semiparametric random-effects model[a]	5(12.2%)	[51, 52, 77, 78, 81]
LME	13(31.7%)	[15–19, 23, 33, 35, 46, 48, 62, 75, 76]
Partially LME	4(9.8%)	[22, 37, 56, 61]
Mixed effect model, Mixed effect model with IOU stochastic process, Mixed-effect varying coefficient Tobit model, Bent-cable mixed-effects model[a]	4(9.8%)	[25, 39, 44, 47]
Mixed-effects varying-coefficient model	3(7.3%)	[27, 38, 80]
LQMM, Quantile-based mixed model, QR- NLME, QR-NLMET[a]	4(9.8%)	[24, 59, 60, 79]
Hurdle two-part model and Longitudinal Tobit model[a]	2(4.8%)	[43, 66]
Random change point model, Multiple-change point model, Longitudinal model for the immune response [a]	4(9.8%)	[13, 40, 55, 65]
ZAB, Two zero-inflated count models[a]	2(4.8%)	[31, 53]
Random effect distribution		
Normal	17(47.4%)	[13, 15, 19, 25, 27, 33, 38, 46, 47, 51, 55, 60, 75–77, 79–81]
Multivariate normal	10(26.3%)	[16–18, 24, 39, 43, 44, 59, 62, 66]
Finite mixture of normal distributions, N/I[a]	3(7.9%)	[48], [23, 35]
Dirichlet process prior	2(5.3%)	[40, 52]
Spline	5(13.1%)	[22, 37, 56, 61, 78]
Error distribution		
Normal	18(48.6%)	[13, 16, 17, 19, 33, 39, 40, 46–48, 52, 55, 62, 65, 66, 75, 76, 81]
N/I, SN[a]	3(8.1%)	[23, 35], [51]
t-distribution	1(2.8%)	[18]
ST	6(16.2%)	[15, 22, 37, 38, 56, 61]
Multivariate ST	6(16.2%)	[25, 27, 44, 77, 78, 80]
ALD	3(8.1%)	[24, 59, 79]

표 1 : 세부사항 별 사용방법론 비율

MODELLING

Bayesian Joint Modelling 를 다룬 논문에서 가장 많이 사용된 모델은 Linear Mixed Model 이고, 흥미롭게 읽었던 모델은 Latent Variable Model 이다 (각 비율은 표 2 참고).

Model	
GLM, Partially LME[a]	2(9.1%)
Multivariate GLM	4(18.2%)
Multivariate mixed effect models	5(22.7%)
ZAB, Proportional-odds cumulative logit model[a]	2(9.1%)
GLM and CR mixed-effects model, Mixed-effect model and CR mixed-effects model, LME and continuous latent variable model, LME and a mixed-effects beta regression model, ZOIB[a]	5(22.7%)
MLIRT	2(9.1%)
MLLTM, MLTLM[a]	2(9.1%)

표 2 : 모형 별 논문출현 비율

여기서 Latent Variable 이란 관측되지 않았으나 Y 변수에 영향을 줄 것이라 예상되는 변수를 뜻하며, Latent Variable Modelling 의 주요 목적은 3 가지가 있다. 첫 째는 기존 변수로 설명되지 않는 분산을 잡아내는 거이고, 둘 째는 예측 변수에 직접적으로 영향을 주는 변수를 만드는 것, 그리고 마지막은 예측 변수를 직접적으로 설명하는 변수를 만드는 것이다. Latent Variable Modelling 의 대표적 예시로 Factor Analysis 가 있다.

RANDOM EFFECT DISTRIBUTION

랜덤효과 분포 가정에 사용된 분포들의 비율은 표 3 에 나와 있으며, 가장 많이 사용된 분포는 정규분포이고, 흥미로웠던 가정은 Dirichlet Process Prior 이다.

Random effect distribution	
Normal	12 (54.5%)
Multivariate normal	7(31.8%)
Dirichlet process prior	3(13.7%)

표 3 : 랜덤효과 분포 가정 비율

Dirichlet Process 는 단순하게 말하자면 분포들의 분포로, Dirichlet Process 를 이용해 sampling 을 진행한다는 것은 분포를 sampling 한다는 의미로 받아들이면 된다.

그림 2 를 참고하며 설명하겠다. Sampling 에서 사용하는 것은 Base Distribution G_0이 아니라, G_0과 같은 support 를 갖는 G 를 이용해서 sampling 을 진행하는 것이다. 이 때, Dirichlet Process 는 sample 인 X_n에 대한 분포가 아닌 G 에 대한 분포이기 때문에 $X_n \sim DP(\alpha, G_0)$이 아니라, 수식 1 과 같이 표기해야 함에 유의해야 한다.

아래와 같은 process G가 있다고 하자.

$$G \sim DP(\alpha, G_0)$$

(G_0는 sampling하게 될 Base Distribution이라고 하며,
$\alpha(>0)$는 Scaling Parameter라고 한다)
이 때, G는 Base Distribution G_0와 같은 Support를
가지는 Random Probability Measure이다.

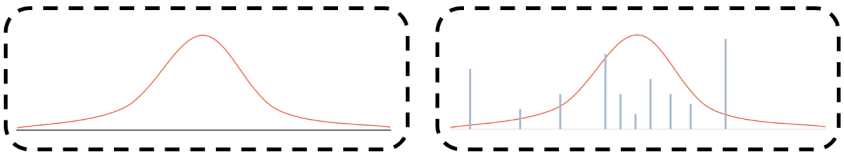


그림 2

$$\begin{aligned} X_n|G &\overset{iid}{\sim} G \quad \text{for } n = \{1, \dots, N\} \\ G &\sim DP(\alpha, G_0) \end{aligned}$$

수식 1

이를 수식 2 처럼 베이즈 정리를 이용해 정리하고, $X_1, \dots, X_n$에 대하여 marginalization 을 시켜주면, $X_n|X_1, \dots, X_{n-1}$이 어떤 형태를 취하는지 알 수 있다

(결과는 수삭 3 에 나와 있다). 그리고 이를 K 개의 sample 로 확장하면 수식 4 처럼 나오는 것을 확인할 수 있다.

$$P(X_1, \ldots, X_N) = \int P(G) \prod_{n=1}^{N} P(X_n|G) dG$$

수식 2

$$X_n|X_1, \ldots, X_{n-1} = \begin{cases} X_i & \text{with prob. } \frac{1}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with prob. } \frac{\alpha}{n-1+\alpha} \end{cases}$$

수식 3

Let there be K unique values for the variables: $X_k^\star$ for $k \in \{1, \ldots, K\}$. Then, we can rewrite it as

$$X_n|X_1, \ldots, X_{n-1} = \begin{cases} X_k^\star & \text{with prob. } \frac{\text{num}_{n-1}(X_k^\star)}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with prob. } \frac{\alpha}{n-1+\alpha} \end{cases}$$

Then,

$$\begin{aligned} P(X_1, \ldots, X_N) &= P(X_1)P(X_2|X_1)\ldots P(X_N|X_1, \ldots, X_{N-1}) \\ &= \frac{\alpha^K \prod_{k=1}^{K}(n_k - 1)!}{\alpha(1+\alpha)\cdots(N-1+\alpha)} \prod_{k=1}^{K} G_0(X_k^\star), \end{aligned}$$

where n_k is the number of observations having value $X_k^\star$.

수식 4

ERROR DISTRIBUTION

오차의 분포에 대한 가정으로 가장 많이 사용된 것은 당연하게도 정규 분포이다. 48.6%로 압도적으로 높았으며, 이 중 흥미로웠던 것은 8.1%의 Skew Normal Distribution 이었다. Skew Normal Distribution 은 이상치가 많을수록 Robust 한 결과를

출력하기 때문에 사용하고, 자세한 설명은 그림 3 에 나와 있다. 이를 통해 알 수 있는 것은 정규분포는 Skew Normal Distribution 의 한 부분이라는 것이다. (α = 0)

- 정규분포에서 파생된 분포 (shape parameter = 0)

$$f(x|\alpha) = 2\phi(x|\xi,\omega)\Phi(\alpha x)$$

→ $\xi : location, \omega : scale, \alpha = shape$

$$pdf : \frac{1}{(\omega\pi)^e}\int_{-\infty}^{\alpha(\frac{x-\xi}{\omega})} \exp\{-\frac{t^2}{2}\}\,dt$$

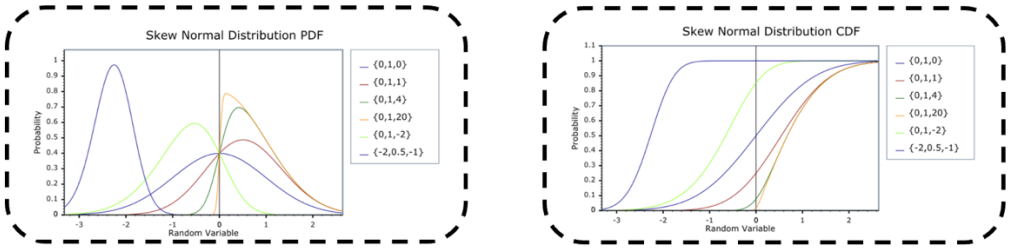


그림 3

BAYESIAN SAMPLING ALGORITHM

샘플링 기법으로 가장 많이 사용된건 38.8%로 MCMC 이고(표 4), 이 중 눈에 띄었던 것은 1.4%의 점유율을 가졌던 Hamiltonian Monte Carlo 였다. 이름에서 알 수 있듯이 MCMC 의 변형으로, transition 생성방법과 적분을 근사하는 방식이 매우 다르다. Hamiltonian Monte Carlo 는 대략 3 단계로 구성되어 있다. 3 단계를 설명하기 전에 알아야 할 개념인 Auxiliary Momentum Variable 과 Hamiltonian 에 대해서는 설명 1 과 설명 2 로 대체하겠다.

1 단계는 θ는 current value = $\theta^{(i)}$, ρ는 $\rho^{(i)} \sim MultiNormal(0,\Sigma)$로 initial value 를 정하고, Hamiltonian 을 각각 ρ와 θ에 대해 미분한다. 2 단계는 미분을 통해 얻은

수식들을 사용하여 $\rho^{(i+1)} = \rho^{(i)} - \frac{\epsilon}{2}\left(\frac{\partial V}{\partial \theta}\right)$, $\theta^{(i+1)} = \theta^{(i)} + \epsilon\Sigma\rho^{(i+1)}$로 업데이트 하고, 이러한 과정을 L 번 반복한 뒤, $\rho^{(L)} = \rho^*, \theta^{(L)} = \theta^*$로 정의한다. 그리고 마지막 3 단계에서는 $\exp\left(H\left(\rho^{(1)}, \theta^{(1)}\right) - H(\rho^*, \theta^*)\right)$을 계산하여 1 이상이면 ρ^*와 θ^*를 표본으로 받아들이고, 만약 1 보다 작으면 ρ^*와 θ^*를 initial value 로 설정하고 1, 2 단계를 반복한다.

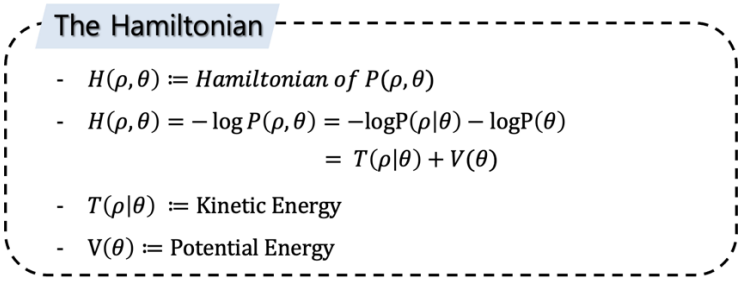
Sampling algorithm	Number of articles (%)
Markov Chain Monte Carlo (MCMC)	28(38.8%)
Gibbs sampler and Metropolis Hastings (MH)	24(33.3%)
Gibbs sampling	9(12.5%)
Gibbs sampling with adaptive rejection and MH	3(4.2%)
Block Gibbs sampling and MH	2(2.8%)
Bayesian Lasso	1(1.4%)
Newton-Raphson procedure and a derivative-based MCMC	1(1.4%)
No-U-Turn sampler	2(2.8%)
Hamiltonian Monte Carlo (HMC)	1(1.4%)
HMC and No-U-Turn sampler	1(1.4%)

표 4

Auxiliary Momentum Variable

- ρ : Auxiliary Momentum Variable
- $P(\rho, \theta) = P(\rho|\theta)P(\theta)$에서 sampling
- $\rho \sim MultiNormal(0, \Sigma),\ where\ \Sigma = I$ assumed mostly

설명 1



설명 2

CONCLUSION

사실 처음 논문을 읽었을 때 예상했던 것보다 지나치게 이론적 그리고 수리적 내용이 없어서 당황하였다. 해당 논문은 단순히 저자 자신이 찾은 논문들을 읽고, 이에 대한 statistical summary 만 제공했다고 해도 무방할 정도로 논문의 구조가 매우 단순했고, 사실 내용이랄게 없었다고 생각했다. 논문 그대로를 반영하여 발표를 하고, 이에 대한 보고서를 쓰기에는 무리라고 판단되어, 고민한 끝에 위와 같은 방식을 선택하게 되었다. 사실 아직 이론적으로 이해가 안가는 부분이 많고, 제대로 설명한 것인지 확신이 서지 않지만, 흥미로운 이론, 방법론, 분포에 대해 새로 알아냈고, 공부할 수 있었다는 사실에 만족한다.