การคัดเลือกตัวแปรและการประมาณค่าสัมประสิทธิ์แบบเบย์เชิงประจักษ์สำหรับตัวแบบ Cox's proportional hazard ที่ข้อมูลมีมิติสูง

EMPIRICAL BAYES VARIABLE SELECTION AND ESTIMATION FOR THE COX'S PROPORTIONAL HAZARD MODEL WITH HIGH DIMENSIONAL DATA

อรณิชา ห่อนบุญเหิม[1*]และ วิฐรา พึ่งพาพงศ์[2]

Onnicha Honboonherm and Vitara Pungpapong

[1]นักศึกษาปริญญาโท สาขาวิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย

[2]อาจารย์ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย

Abstract

Cox's proportional hazard model is one of the most popular models in survival analysis. To analyses high dimensional data, Bayesian variable selection methods combined with Markov chain Monte Carlo (MCMC) algorithms have been explored for such tasks due to easy implementation. However, it has been known that they are slow to converge to stationary distribution especially for high dimensional case so. We therefore employ iterated conditional modes/medians (ICM/M) algorithm which is empirically faster and easy to implement. The objective of this dissertation is to study the performance of the proposed empirical Bayes variable selection method using the ICM/M algorithm. The effects from the ratio of sample size to the number of independent variables and the percentage of censored data are discussed here.

Keywords: *empirical Bayes, Cox's proportional hazard model, high dimensional data, sparse variable*

บทคัดย่อ

ตัวแบบ Cox's proportional hazard ถือเป็นตัวแบบหนึ่งที่ได้รับความนิยมมากในการวิเคราะห์การอยู่รอด การวิเคราะห์ข้อมูลที่มีมิติสูงทำได้หลายวิธีทั้งวิธี Penalized และวิธีแบบ Bayesian ซึ่งพบว่าเป็นวิธีที่ใช้งานได้สะดวก วิธีการคัดเลือกตัวแปรแบบ Bayesian มีเทคนิคที่ช่วยในการทำงานคือวิธีมาคอฟ เชน มอนติคาร์โล (MCMC) แต่เป็นที่ทราบกันดีว่าวิธีดังกล่าวใช้เวลานานในการรอให้ข้อมูลมีค่าความผันแปรเฉลี่ยมีลักษณะคงที่ตลอดช่วงเวลา (converge) โดยเฉพาะในกรณีที่จำนวนพารามิเตอร์มีขนาดใหญ่ เราจึงเลือกใช้เทคนิคIterated conditional modes/medians (ICM/M) ซึ่งเป็นเทคนิคที่สามารถคำนวณได้ง่ายและรวดเร็วกว่าวิธีมาคอฟเชน มอนติคาร์โล (MCMC) ในการศึกษาครั้งนี้ต้องการทราบผลกระทบของการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์ด้วยวิธีการของเบย์เชิงประจักษ์ โดยใชเทคนิคการ ICM/M เมื่อพิจารณาอัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ และร้อยละของข้อมูลเซ็นเซอร์ที่ระดับต่างๆต่อตัวแบบ Cox's proportional hazard ที่ข้อมูลมีมิติสูง

คำสำคัญ: *เบย์เชิงประจักษ์, ตัวแบบ Cox's proportional hazard, ข้อมูลมิติสูง, ตัวแปรที่ค่าสัมประสิทธิ์ส่วนใหญ่เป็นศูนย์*

บทนำ

เนื่องจากในปัจจุบันการวิเคราะห์การถดถอยเชิงเส้น(Regression Analysis) เข้าไปมีบทบาทในวงการต่างๆ มากมาย ในการวิเคราะห์การอยู่รอด(Survival Analysis) ถือเป็นการวิเคราะห์การถดถอยเชิงเส้นประเภทหนึ่ง ตัวแบบ Cox's proportional hazard เป็นตัวแบบเชิงเส้นที่ใช้ในการวิเคราะห์การอยู่รอดที่มีลักษณะแบบกึ่งพารามิเตอร์ (semi-parameter) อีกทั้งตัวแบบ Cox's proportional hazard ไม่จำเป็นต้องระบุถึงฟังก์ชัน hazard baseline ก็สามารถที่จะคำนวณหาค่าอัตราความเสี่ยง (hazard ratio) หรือค่าconstant rate over time ได้ ดังนั้นตัวแบบ Cox's proportional hazard จึงเป็นตัวแบบที่ได้รับความนิยมสูง

โดยทั่วไปการประมาณค่าสัมประสิทธิ์การถดถอยในตัวแบบCox's proportional hazard สามารถทำได้โดยวิธีการประมาณค่าภาวะน่าจะเป็นสูงสุด (Maximum likelihood estimation (MLE)) ซึ่งจำเป็นที่จะต้องมีขนาดตัวอย่างอย่างน้อยเท่ากับจำนวนตัวแปรอิสระจึงจะสามารถหาตัวประมาณ MLE ได้ ในการศึกษานี้ เราสนใจการประมาณค่าสัมประสิทธิ์กรณีที่ตัวแปรอิสระมีขนาดใหญ่กว่าขนาดตัวอย่าง รวมไปถึงการคัดเลือกตัวแปรอิสระที่เหมาะสมเข้ามาในตัวแบบ

Johnstone and Silverman (2004) ได้เสนอวิธี Empirical Bayes thresholding เพื่อใช้ในการสร้างThreshold แบบสุ่มสำหรับข้อมูลอิสระที่มีการแจกแจงแบบปกติโดยการให้prior สำหรับค่าเฉลี่ยของข้อมูลแต่ละตัวในรูปของการแจกแจงแบบผสมระหว่างส่วนที่ค่าพารามิเตอร์เป็นศูนย์และส่วนที่ค่าพารามิเตอร์ไม่เท่ากับศูนย์จาก prior ดังกล่าว ทำให้ได้การแจกแจง posterior แบบผสมระหว่างส่วนที่ค่าพารามิเตอร์เป็นศูนย์และส่วนที่ค่าพารามิเตอร์ไม่เท่ากับศูนย์เช่นเดียวกัน ดังนั้น หากเราเลือกใช้ตัวประมาณอย่างเหมาะสม เช่น ค่ามัธยฐาน (posterior median) จะทำให้ค่าพารามิเตอร์บางส่วนมีค่าเป็นศูนย์

จากแนวความคิดของJohnstone and Silverman ผู้ศึกษาจะนำมาต่อยอดในการเลือกตัวแปรอิสระเข้าสู่ตัวแบบ รวมไปถึงขั้นตอนการประมาณค่าสัมประสิทธิ์ (β) ในตัวแบบCox's proportional hazard โดยให้พารามิเตอร์ (β) อยู่ในรูปของการแจกแจงแบบผสม

สำหรับขั้นตอนวิธีที่ใช้เป็นเครื่องมือสำหรับวิธีการวิเคราะห์แบบ Bayesian กันอย่างแพร่หลายคือวิธีมาคอร์ฟเชน มัลติคาโร (MCMC) แต่เป็นที่ทราบกันดีว่าวิธีดังกล่าวใช้เวลานานในการรอให้ข้อมูลมีค่าความผันแปรเฉลี่ยมีลักษณะคงที่ตลอดช่วงเวลา (converge) โดยเฉพาะในกรณีที่จำนวนพารามิเตอร์มีขนาดใหญ่ โดยในที่นี้เราจะเลือกใช้เทคนิคIterated conditional modes/medians (ICM/M) เป็นเทคนิคที่ใช้การแจกแจงแบบมีเงื่อนไขซึ่งสามารถคำนวณได้ง่ายและรวดเร็วกว่าวิธีมาร์คอฟเชน มอนติคาร์โล (MCMC)

วัตถุประสงค์การวิจัย

เพื่อศึกษาการคัดเลือกตัวแปรอิสระและการประมาณค่าสัมประสิทธิ์สำหรับตัวแบบ Cox's proportional hazard ที่ข้อมูลมีมิติสูง

ทฤษฎีและกรอบแนวคิด

- Empirical Bayes thresholding method (Johnstone and Silverman, 2004)

ให้ $X=(X_1,X_2,...,X_n)$ คือค่าสังเกตุที่วัดได้ โดยที่

$X_i=\mu_i+\varepsilon_i \quad ,\varepsilon_i \sim N(0,1)$

โดยมีวัตถุประสงค์เพื่อประมาณค่า μ_i โดยสมมติว่าค่าพารามิเตอร์ส่วนใหญ่เท่ากับศูนย์ และจากปัญหาข้างต้นจะเห็นว่าค่าประมาณของ μ_i ที่ได้จากวิธี MLE คือ X_i แต่ตัวประมาณนี้เป็นตัวประมาณที่ไม่ค่อยดีนัก เนื่องจากพิจารณาข้อมูลเพียงข้อมูลเดียว นอกจากนี้ ตัวประมาณที่ได้ค่าส่วนใหญ่จะไม่เท่ากับศูนย์

ดังนั้น Johnstone and Silvermanจึงเสนอวิธีการแบบเบย์ โดยให้priorกับ μ_i ในรูปของการแจกแจงแบบผสม ดังนี้

$$\mu_i \sim (1-\omega)\delta_0(\mu_i)+\omega\gamma(\mu_i)$$

โดยที่ $\delta_0(.)$ คือ diract delta function จาก prior ข้างต้น จะเห็นได้ว่า $\mu_i=0$ ที่ความน่าจะเป็น $(1-\omega)$ และ $\mu_i \neq 0$ และมีความหนาแน่น $\gamma(.)$ ที่ความน่าจะเป็น ω Johnstone and Silverman ได้เสนอรูปแบบการแจกแจงสำหรับ $\gamma(.)$ ไว้ 2 รูปแบบคือการแจกแจง Laplace

$$\gamma(\mu\,|\,\alpha)=\frac{1}{2}\alpha\exp\{-\alpha|\mu|\},a>0$$

และการแจกแจง quasi-Cauchy

$$\gamma(\mu)=(2\pi)^{-\frac{1}{2}}\left\{1-|\mu|\frac{1-\Phi(|\mu|)}{\phi(\mu)}\right\}$$

จาก prior ดังกล่าว จะได้ว่าการแจกแจง posterior จะอยู่ในรูปของการแจกแจงผสมระหว่าง ส่วนที่ค่าพารามิเตอร์เป็นศูนย์และส่วนที่ค่าพารามิเตอร์ไม่เท่ากับศูนย์เช่นเดียวกันกับ prior Johnstone and Silverman เสนอการประมาณค่า μ_i ด้วยค่ามัธยฐาน (Posterior median) ซึ่งจะได้ค่า μ_i ส่วนใหญ่เป็นศูนย์ นอกจากนี้ ยังได้กล่าวถึงทฤษฎีความสัมพันธ์ระหว่างค่ามัธยฐาน (Posterior median $(\hat{\mu})$) และเขตกั้น (Threshold (τ)) กล่าวคือ สำหรับค่า ω ที่คงที่ $\exists\,\tau(\omega)>0$ ที่ทำให้ $\hat{\mu}(x,\omega)=0$ ก็ต่อเมื่อ $|x_i| \leq \tau(\omega)$

สำหรับในส่วนการประมาณค่า ω ด้วยวิธีเบย์เชิงประจักษ์สามารถทำได้โดยการหาค่า ω ที่ทำให้ความน่าจะเป็นส่วนริม(Marginal likelihood) สูงสุด

- Iterated Conditional Modes/Medians (ICM/M)

เป็นวิธีการสำหรับช่วยในการคำนวณให้รวดเร็วในการประมาณค่าพารามิเตอร์ด้วยวิธีการ Empirical Bayes ซึ่งมีลักษณะแนวคิดเช่นเดียวกับวิธี Iterated conditional modes (Besag,1986) ที่กล่าวถึงวิธี ICM ว่าเป็นวิธีการที่เหมาะสมในการประมาณค่าพารามิเตอร์แบบทั่วไปและแบบที่มีลักษณะมิติสูง (High-dimensional) ได้ดีกว่าการประมาณค่าพารามิเตอร์โดยการใช้สมการถดถอย ดังนั้นการกล่าวถึงวิธีการแบบ ICM/M จึงมีความเหมาะสมในการประมาณค่าพารามิเตอร์แบบที่ข้อมูลมีมิติสูง (High-dimensional)

Iterated conditional mediansใช้สำหรับเป็นวิธีในกระบวนการคัดเลือกตัวแปรอิสระ จาก Johnstone and Silverman (2004) กล่าวว่า เมื่อสร้างprior แบบผสมขึ้นมาแล้ว ค่าposterior medians จะนำไปสู่กฎของ Thresholdingซึ่งเป็นตัวคัดกลองตัวแปรได้ดี

จากคุณสมบัติของIterated conditional modes และ Iterated conditional medians ทำให้ ICM/M เป็นวิธีที่ช่วยให้การคัดเลือกตัวแปรอิสระจากการสร้าง priorที่ซับซ้อนเนื่องจากตัวแปรอสระที่พิจารณามีขนาดใหญ่ทำได้ง่ายและรวดเร็ว

วิธีการศึกษา

ตัวแปรที่ศึกษา

(t_i, x_i, ζ_i)

โดย t_i คือ Right-censored time

x_i คือ ตัวแปรอิสระที่มีมิติ p

ζ_i คือ ตัวแปรบ่งชี้ (Indicator variable) โดยที่ $\zeta_i = \begin{cases} 1; t_i \le c_i \\ 0; t_i > c_i \end{cases}$ เมื่อ c_i คือ เวลาคงที่ใดๆ

ขั้นตอนการดำเนินงาน

1.กำหนดและจำลองข้อมูล

1.1.กำหนดค่าเริ่มต้นดังนี้

- สร้างข้อมูลที่มีขนาด n และจำนวนตัวแปรอิสระ p ตัว และร้อยละของข้อมูลสูญหาย

โดยในการทดลองครั้งนี้จะพิจารณาที่ค่า n=100, p=300, 500และ1,000 ที่ร้อยละของข้อมูลสูญหายที่ 10% 50% และ 70% ดังนั้นข้อมูลที่ได้จะแบ่งเป็น 9 กรณี คือ

- กรณีที่ 1: n=100, p=300 ร้อยละของข้อมูลสูญหายคือ 10%
- กรณีที่ 2: n=100, p=300 ร้อยละของข้อมูลสูญหายคือ 50%
- กรณีที่ 3: n=100, p=300 ร้อยละของข้อมูลสูญหายคือ 70%
- กรณีที่ 4: n=100, p=500 ร้อยละของข้อมูลสูญหายคือ 10%
- กรณีที่ 5: n=100, p=500 ร้อยละของข้อมูลสูญหายคือ 50%
- กรณีที่ 6: n=100, p=500 ร้อยละของข้อมูลสูญหายคือ 70%
- กรณีที่ 7: n=100, p=1,000 ร้อยละของข้อมูลสูญหายคือ 10%
- กรณีที่ 8: n=100, p=1,000 ร้อยละของข้อมูลสูญหายคือ 50%
- กรณีที่ 9: n=100, p=1,000 ร้อยละของข้อมูลสูญหายคือ 70%

1.2.จำลองข้อมูลที่มีการแจกแจงแบบ Weibull

$$t = \left(-\frac{\log(U)}{\lambda \exp(\beta^T x)}\right)^{1/\nu}$$ โดยให้Scale parameter $\nu > 0$ และShape parameter $\lambda > 0$

3.นำข้อมูลที่จำลองขึ้นมาศึกษาการคัดเลือกตัวแปรแบบเบย์เชิงประจักษ์สำหรับตัวแบบ Cox's proportional Hazard โดยกำหนดให้

-Likelihood และ priorคือ

Likelihood: $$L=\prod_{i=1}^{n}\left[\left\{\lambda_0\left(t_i\right)e^{x_i^T\beta}\right\}^{\zeta_i}\exp\left\{-H_0\left(t_i\right)e^{x_i^T\beta}\right\}\right]$$

เมื่อ $H_0\left(t\right)=\sum_{j:y_j\le t_j}\Delta h_0\left(y_j\right)$ โดยที่ $y_1<y_2<\cdots<y_D$ เป็นค่า t_i ที่แตกต่างกัน

และ $\Delta\hat{h}_0\left(y_j\right)=\dfrac{d_j}{\sum_{i:t_j\ge y_j}e^{x_i^T\beta}},d_j=\sum_{i:t_i=y_i}\varsigma_i$

Prior: $$\beta_i\sim\left(1-\omega\right)\delta_0\left(\beta_i\right)+\omega\gamma\left(\beta_i\right)$$

เมื่อ $$\gamma\left(\beta\mid\alpha\right)=\frac{1}{2}\alpha\exp\left\{-\alpha\left|\beta\right|\right\},\alpha>0$$

ในการศึกษาครั้งนี้เราใช้ค่า $\alpha=0.5$ (เสนอแนะโดยJohnstone and Silverman (2004, 2005))

แนวทางการวิเคราะห์ข้อมูลและสถิติที่ใช้ในการวิเคราะห์

1. False negative rate
2. False positive rate

ผลการวิจัย

ตารางแสดงผลการศึกษาค่าความผิดพลาดเชิงบวก(ตาราง 1) และความผิดพลาดเชิงลบ (ตาราง 2) โดยที่ค่า standard deviation แสดงไว้ในวงเล็บ

ตาราง 1 แสดงค่า False positive rate

n:p	c	Lasso	EBVS-Cox	
			True beta	Lasso beta
100:300	10%	0.6603(0.0895)	0.0009(0.0046)	0.0020(0.0281)
	50%	0.6002(0.1084)	0.0019(0.0094)	0.0327(0.0307)
	70%	0.6216(0.0763)	0.0032(0.0122)	0.0519(0.0527)
100:500	10%	0.6451(0.0881)	0.0029(0.0114)	0.0699(0.1011)
	50%	0.6651(0.1014)	0.0033(0.0137)	0.0746(0.1165)
	70%	0.6406(0.1062)	0.0091(0.0227)	0.0896(0.1529)
	10%	0.6556(0.0598)	0.0033(0.0182)	0.1584(0.1816)

100:1,000	50%	0.6486(0.2125)	0.0024(0.0122)	0.3890(0.4291)
	70%	0.7780(0.1584)	0.0033(0.0182)	0.4433(0.4306)

หมายเหตุ: n คือ ขนาดตัวอย่าง, p คือ จำนวนตัวแปรอิสระ, c คือ ร้อยละของข้อมูลเซ็นเซอร์

EBVS-Cox คือ การคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์ด้วยวิธีการแบบเบย์เชิงประจักษ์สำหรับตัวแบบCox proportional hazard

ตาราง 2 แสดงค่า False negative rate

n:p	c	Lasso	EBVS-Cox	
			True beta	Lasso beta
100:300	10%	0.0148 (0.0204)	0(0)	0.0187(0.0129)
	50%	0.0353 (0.0123)	0(0)	0.0283(0.0154)
	70%	0.0351(0.0140)	0(0)	0.0304(0.0237)
100:500	10%	0.0114(0.0341)	0(0)	0.0195(0.0036)
	50%	0.0293(0.0303)	0(0)	0.0285(0.0064)
	70%	0.0364 (0.0639)	0(0)	0.0643(0.0074)
100:1,000	10%	0.0103(0.0106)	0(0)	0.0134(0.0034)
	50%	0.0179(0.0167)	0(0)	0.0196(0.0037)
	70%	0.0220(0.0313)	0(0)	0.0221(0.0054)

หมายเหตุ: n คือ ขนาดตัวอย่าง, p คือ จำนวนตัวแปรอิสระ, c คือ ร้อยละของข้อมูลเซ็นเซอร์

EBVS-Cox คือ การคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์ด้วยวิธีการแบบเบย์เชิงประจักษ์สำหรับตัวแบบCox proportional hazard

การอภิปรายผล

เมื่อพิจารณาอัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรอิสระพบว่าอัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรอิสระที่ระดับสูง (100/300) จะให้ค่าความผิดพลาดเชิงบวกต่ำกว่าอัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรอิสระที่ระดับกลาง (100/500) และระดับต่ำ (100/1,000) ตามลำดับ ร้อยละของข้อมูลเซ็นเซอร์ ส่งผลต่อการคัดเลือกตัวแปรและการประมาณค่าสัมประสิทธิ์แบบเบย์เชิงประจักษ์ โดยที่ร้อยละของข้อมูลเซ็นเซอร์ต่ำ (10%) จะให้ค่าความผิดพลาดเชิงบวกต่ำกว่ากรณีร้อยละของข้อมูลเซ็นเซอร์กลาง (50%) และร้อยละของข้อมูลเซ็นเซอร์สูง (70%) ตามลำดับ เมื่อเปรียบเทียบค่าความผิดพลาดเชิงบวกและเชิงลบระหว่างวิธีแบบเบย์เชิงประจักษ์และวิธี Lasso สำหรับตัวแบบCox proportional hazard พบว่า แม้วิธี Lasso จะให้ค่าความผิดพลาดเชิงลบน้อยกว่าวิธีวิธีแบบเบย์เชิงประจักษ์ แต่ค่าที่ได้ก็แตกต่างกันไม่มาก อีกทั้งค่าความผิดพลาดเชิงบวกของวิธี Lasso มีค่าสูงกว่าวิธีแบบเบย์เชิงประจักษ์ค่อนข้างมาก ดังนั้นเมื่อมองจากภาพรวมพบว่าวิธีการแบบเบย์เชิงประจักษ์จึงถือเป็นวิธีการที่ดีกว่า

อีกปัจจัยที่ส่งผลต่อความผิดพลาดเชิงบวกและเชิงลบก็คือการกำหนดค่าสัมประสิทธิ์ถดถอยเริ่มต้น ก่อนจะเข้าสู่ขั้นตอนการหาค่าสัมประสิทธ์ด้วยวิธีแบบเบย์เชิงประจักษ์ จากการทดลองเมื่อเรากำหนดค่าสัมประสิทธิ์

ถดถอยเริ่มต้นเป็นค่าสัมประสิทธิ์ที่แท้จริง ปัจจัยของอัตราส่วนระหว่างขนาดตัวอย่างต่อจำนวนพารามิเตอร์และร้อยละของข้อมูลเซ็นเซอร์จะส่งผลต่อความผิดพลาดเชิงบวกเพียงเล็กน้อย และไม่ส่งผลต่อความผิดพลาดเชิงลบเลย

เอกสารอ้างอิง

Andrew, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distributions, Journal of the Royal StatisticalSociety. Series B 36(1), 99-102.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association96,348-1360.

Fan, J. and Li, R. (2001).Variable selection for Cox's proportional hazards model and frailty model, The Annals of Statistics 30(1), 74-99.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks:empirical Bayes estimates of possibly sparse sequence. The Annals of Statistics 32, 1594-1649.

Lee, K.H., Chakraborty, S., and Sun, J. (2011). Bayesian variable selection in semiparametric proportional hazards model for high dimensional survival data. The International Journal of Biostatistics 7(1),

Mitchell, T. J. and Beauchamp, J. J. (1988).Bayesian variable selection in linear regression (with Discussion).Journal of the American Statistical Association, 83:1023-1036.

Park, T. and Casella, G. (2008).The Bayesian lasso.Journal of the American Statistical Association 103, 681-686.

Tibshirani, R. (1996).Regression shrinkage and selection via the lasso.Journal of the Royal Statistical Society. Series B 58(1), 267-288.

Tibshirani, R. (1997). The Lasso method for variable selection in the Cox model. Statistic in medicine,Vol.16, 385-395.

Zhang, H. H. and Lu, W. (2007).Adaptive lasso for Cox's proportional hazards model. Biometrika 94(3), 1-13.

Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American Statistical Association101, 1418-1429.