การเปรียบเทียบประสิทธิภาพของการสุ่มตัวอย่างด้วยวิธีฮิตแอนด์รันและการสุ่มตัวอย่างด้วยวิธีกิบส์สำหรับการวิเคราะห์ปัจจัยเชิงเบส์

นางสาวไรยา พลเสน

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาสถิติศาสตรมหาบัณฑิต
สาขาวิชาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2553

A COMPARISON ON THE EFFICIENCY OF HIT-AND-RUN SAMPLER AND GIBBS SAMPLER FOR BAYESIAN FACTOR ANALYSIS

Miss Raiya Polsen

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Statistics

Department of Statistics
Faculty of Commerce and Accountancy
Chulalongkorn University

Academic Year 2010

หัวข้อวิทยานิพนธ์	การเปรียบเทียบประสิทธิภาพของการสุ่มตัวอย่างด้วยวิธีฮิตแอนด์รันและการสุ่มตัวอย่างด้วยวิธีกิบส์สำหรับการวิเคราะห์ปัจจัยเชิงเบส์
โดย	นางสาวไรยา พลเสน
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร. เสกสรร เกียรติสุไพบูลย์

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะพาณิชยศาสตร์และการบัญชี

(รองศาสตราจารย์ ดร. อรรณพ ตันละมัย)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(รองศาสตราจารย์ ดร. ธีระพร วีระถาวร)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(ผู้ช่วยศาสตราจารย์ ดร. เสกสรร เกียรติสุไพบูลย์)

..... กรรมการ

(รองศาสตราจารย์ ดร. สุพล ดุรงค์วัฒนา)

..... กรรมการภายนอกมหาวิทยาลัย

(อาจารย์ ดร. อัครินทร์ ไพบูลย์พานิช)

ไรยา พลเสน : การเปรียบเทียบประสิทธิภาพของการสุ่มตัวอย่างด้วยวิธีฮิตแอนด์รันและการสุ่มตัวอย่างด้วยวิธีกิบส์สำหรับการวิเคราะห์ปัจจัยเชิงเบส์. (A COMPARISON ON THE EFFICIENCY OF HIT-AND-RUN SAMPLER AND GIBBS SAMPLER FOR BAYESIAN FACTOR ANALYSIS) อ. ที่ปรึกษาวิทยานิพนธ์หลัก: ผศ. ดร. เสกสรร เกียรติสุไพบูลย์, 66 หน้า.

การศึกษานี้มีวัตถุประสงค์เพื่อเปรียบเทียบการสุ่มตัวอย่างด้วยวิธีฮิตแอนด์รันและการสุ่มตัวอย่างด้วยวิธีกิบส์สำหรับทำการวิเคราะห์ปัจจัยเมื่อมีจำนวนปัจจัย 1 ปัจจัย เพื่อใช้ในการจัดอันดับหน่วยทดลองจากตัวชี้วัดมากกว่า 1 ตัวซึ่งทั้ง 2 วิธีจัดอยู่ในกลุ่มวิธีลูกโซ่มาร์คอฟมอนติคาร์โล (MCMC) การสุ่มตัวอย่างทั้ง 2 วิธีจะถูกทดสอบในการจัดอันดับประเทศตามความเสี่ยงทางการเมือง-เศรษฐกิจด้วยข้อมูลตัวชี้วัดจาก MCMCpack R Package โดย Martin และ Quinn 2004 โดยใช้คะแนนของปัจจัยจากตัวแบบการวิเคราะห์ปัจจัยมาจัดอันดับ ในแต่ละวิธี MCMC จะประมาณคะแนนเฉลี่ยภายใต้การแจกแจงความน่าจะเป็นภายหลังและค่าคลาดเคลื่อนมาตรฐาน (SE) แล้วใช้ค่าคลาดเคลื่อนมาตรฐานเป็นเกณฑ์ในการเปรียบเทียบประสิทธิภาพความเร็วในการลู่เข้าของตัวประมาณจากการสุ่มตัวอย่างทั้ง 2 วิธี

จากการทดสอบประสิทธิภาพด้วยข้อมูลความเสี่ยงทางการเมือง-เศรษฐกิจจาก MCMCpack R Package พบว่าการสุ่มตัวอย่างด้วยวิธีฮิตแอนด์รันมีประสิทธิภาพด้อยกว่าวิธีกิบส์ โดยส่วนใหญ่ของค่าคลาดเคลื่อนมาตรฐานของคะแนนเฉลี่ยของแต่ละประเทศที่คำนวณได้จากการสุ่มตัวอย่างด้วยวิธีฮิตแอนด์รันมีค่ามากกว่าค่าคลาดเคลื่อนมาตรฐานที่คำนวณได้จากวิธีกิบส์ สังเกตได้ว่าการวิเคราะห์ปัจจัยเชิงเบส์ในกรณีที่มีพารามิเตอร์จำนวนมาก และมีขอบเขตของพารามิเตอร์ที่ซับซ้อน การเคลื่อนที่ในแต่ละรอบของวิธีฮิตแอนด์รันพารามิเตอร์มีมิติสูงจะเคลื่อนที่ได้อย่างจำกัด จึงทำให้การสุ่มตัวอย่างด้วยวิธีฮิตแอนด์รันมีประสิทธิภาพในเชิงการลู่เข้าของตัวประมาณด้อยกว่าการสุ่มตัวอย่างด้วยวิธีกิบส์

ภาควิชา สถิติ ลายมือชื่อนิสิต ไรยา พลเสน

สาขาวิชา สถิติ ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก

ปีการศึกษา 2553

5182900626 : MAJOR STATISTICS

KEYWORDS : BAYESIAN FACTOR ANALYSIS / MARKOV CHAIN MONTE CARLO / HIT-AND-RUN METHOD

RAIYA POLSEN: A COMPARISON ON THE EFFICIENCY OF HIT-AND-RUN SAMPLER AND GIBBS SAMPLER FOR BAYESIAN FACTOR ANALYSIS.

THESIS ADVISOR: ASST.PROF.SEKSAN KIATSUPAIBUL, Ph.D., 66 pp.

The purpose of this study is to compare the efficiency of Hit-and-run sampler and Gibb sampler, the two Markov chain Monte Carlo samplers, for single factor analysis with an application to ranking units by their multiple indicators. The two samplers are tested against each other on a cross-national political-economic risk ranking according to the indicator data provided in MCMCpack R package by Martin and Quinn 2004, where the factor scores of the factor analysis model are used as the ranking measure. Each of the MCMC samplers is employed to estimates the expectations of the factor scores under its posterior distribution, and the standard errors of the factor score estimates are adopted as the measurement of the efficiency of the samplers.

From the experiments with the test data set, Hit-and-run sampler is inferior to Gibbs sampler. The majority of the standard errors of the factor score estimates obtained from Hit-and-run sampler are higher than those obtained from Gibb sampler. From observations, Hit-and-run has difficulties in its transition due to the fact that the Bayesian factor analysis model imposes a large number of constraints in each step of Hit-and-run in high dimensional parameter space.

Department : Statistics Student's Signature

Field of Study : Statistics Advisor's Signature

Academic Year : 2010

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยความช่วยเหลือ คำแนะนำและข้อเสนอแนะ รวมถึงการเอาใจใส่อย่างใกล้ชิดจากผู้ช่วยศาสตราจารย์ ดร.เสกสรร เกียรติสุไพบูลย์ ผู้เขียนขอน้อมกราบขอบพระคุณต่อท่านอาจารย์เป็นอย่างสูง ตลอดทั้งอาจารย์ทุกท่านที่ได้สอนผู้เขียนมา ณ โอกาสนี้ด้วย

ผู้วิจัยขอกราบขอบพระคุณ รองศาสตราจารย์ ดร. ธีระพร วีระถาวร ประธานกรรมการ รองศาสตราจารย์ ดร. สุพล ดุรงค์วัฒนา และอาจารย์ ดร. อัครินทร์ ไพบูลย์พานิช กรรมการสอบวิทยานิพนธ์ ที่กรุณาตรวจสอบและแก้ไขให้วิทยานิพนธ์ฉบับนี้มีความสมบูรณ์มากยิ่งขึ้น รวมทั้งขอกราบขอบพระคุณคณาจารย์ทุกท่านที่ได้ประสิทธิประสาทวิชาความรู้ให้แก่ผู้วิจัย

ขอขอบคุณพี่สุมณฑา ที่มีส่วนสำคัญที่ทำให้เกิดวิทยานิพนธ์ฉบับนี้ ขอขอบคุณรุ่นพี่ และเพื่อน ๆ ทุกคนที่ให้ความร่วมมือและเป็นกำลังใจแก่ผู้เขียน

สุดท้ายนี้ ผู้วิจัยใคร่ขอกราบขอบพระคุณ บิดา มารดา และครอบครัวที่ได้ให้การสนับสนุนด้านการศึกษาและคอยเป็นกำลังให้ผู้วิจัยมาโดยตลอด

สารบัญ

บทที่ หน้า

สารบัญตาราง

สารบัญภาพ

บทที่ 1

บทนำ

ความเป็นมาและความสำคัญของปัญหา

ในการจัดอันดับ (Ranking) หรือ การประเมินผลการดำเนินงานเชิงสัมพัทธ์ มักจะทำการประเมินจากตัวชี้วัดมากกว่า 1 ตัว เช่น การจัดอันดับประเทศที่มีความเสี่ยงทางด้านการเมือง-เศรษฐกิจ โดยเรียงจากประเทศที่มีความเสี่ยงน้อยที่สุดไปถึงประเทศที่มีความเสี่ยงมากที่สุด สามารถประเมินความเสี่ยงจากดัชนีมากกว่า 1 ตัว เช่น ความเป็นอิสระของระบบตุลาการ, ความเสี่ยงจากการยึดทรัพย์สินในกิจการมาเป็นของรัฐ, การทุจริตในภาครัฐ, ส่วนต่างอัตราแลกเปลี่ยนเงินตราต่างประเทศในตลาดมืด, ผลิตภัณฑ์มวลรวมภายในประเทศ (GDP) เป็นต้น

วิธีการทางสถิติวิธีหนึ่งที่ช่วยในการจัดอันดับในกรณีที่มีตัวชี้วัดมากกว่า 1 ตัว ได้แก่ การวิเคราะห์ปัจจัย (Factor analysis) โดยระบุจำนวนปัจจัยเพียง 1 ปัจจัย และใช้คะแนนของปัจจัย (ϕ) ที่ประมาณได้เป็นดัชนีในการจัดอันดับ ซึ่งโดยทั่วไปแล้วตัวแปรหรือดัชนี ที่นำมาใช้จัดอันดับอาจมีทั้งตัวแปรแบบต่อเนื่องและตัวแปรเชิงอันดับ จึงใช้ตัวแบบการวิเคราะห์ปัจจัยสำหรับข้อมูลแบบผสมที่มีทั้งตัวแปรแบบต่อเนื่องและตัวแปรเชิงอันดับ ดังสมการถัดไป

$$x_i^* = \Lambda\phi_i + \varepsilon_i$$

โดยที่ค่าสังเกตของ x ได้มาจากค่าของตัวแปรแฝง (Latent variable) x^* ซึ่งเป็นค่าต่อเนื่อง แล้วทำการวิเคราะห์ปัจจัยเพื่อการจัดอันดับผ่านตัวแปรแฝง x^*

เนื่องจากวิเคราะห์ปัจจัยสำหรับข้อมูลแบบผสมที่มีทั้งตัวแปรแบบต่อเนื่องและตัวแปรเชิงอันดับต้องทำการวิเคราะห์ผ่านตัวแปรแฝง x^* ดังนั้น Quinn [1] จึงได้นำเสนอการใช้ตัวแบบการวิเคราะห์ปัจจัยเชิงเบส์ (Bayesian factor analysis) ซึ่งได้ปรับให้เหมาะกับการวิเคราะห์ข้อมูลแบบผสมที่มีทั้งตัวแปรแบบต่อเนื่องและตัวแปรเชิงอันดับ แต่ด้วยความซับซ้อนของพารามิเตอร์ Quinn [1] ยังนำเสนอการใช้วิธีการคำนวณที่วางอยู่บนพื้นฐานของการสุ่มตัวอย่างแบบกิบส์ และวิธีของ Cowles, M. K. [2] ซึ่งวิธีดังกล่าวจัดอยู่ในกลุ่มวิธีลูกโซ่มาร์คอฟมอนติคาร์โล (Markov chain Monte Carlo) หรือ MCMC แล้วใช้วิธีค่าเฉลี่ยกลุ่ม (Batch mean method) ที่จะช่วยแก้ปัญหาความไม่เป็นอิสระของพารามิเตอร์ที่จำลองได้ในแต่ละรอบของวิธี MCMC มาใช้ในการประมาณค่าพารามิเตอร์ และค่าคลาดเคลื่อนมาตรฐาน (SE) ของค่าประมาณ

อย่างไรก็ตามวิธีในกลุ่ม MCMC ที่มีประสิทธิภาพอีกวิธีหนึ่งคือ วิธีฮิตแอนด์รัน (Hit-and-run) ซึ่ง Lovasz, Vempala [3] ได้พิสูจน์ว่าในกรณีในการสร้างจุดตัวอย่างซึ่งลู่เข้าสู่การ

แจกแจงแบบสม่ำเสมอบน Log-concave วิธีฮิตแอนด์รันเป็นวิธีที่เร็วที่สุดที่เราทราบในปัจจุบัน และในงานวิจัยของ Chen M.-H, Schmeiser B.W [4] ได้ทำการเปรียบเทียบการสุ่มตัวอย่างทั้ง 2 วิธี พบว่าการสุ่มตัวอย่างแบบฮิตแอนด์รันมีประสิทธิภาพมากกว่าสำหรับในกรณีที่เป็นการแจกแจงแบบปกติของ 2 ตัวแปร จึงเป็นที่น่าสนใจว่าถ้าใช้ข้อมูลชุดเดียวกัน และตัวแบบเดียวกันกับ Quinn [1] แล้วหากลองเปลี่ยนจากการสุ่มตัวอย่างด้วยวิธีกิบส์มาเป็นการสุ่มตัวอย่างด้วยวิธีฮิจแอนด์รัน ผลที่ได้ยังคงสอดคล้องกับงานวิจัยในอดีตหรือไม่

ในงานวิจัยชิ้นนี้จึงต้องการเปรียบเทียบประสิทธิภาพของวิธีฮิตแอนด์รันกับวิธีกิบส์เมื่อพารามิเตอร์มีจำนวนมิติมากและมีขอบเขตที่ซับซ้อน เพื่อหาทางเลือกที่เหมาะสมในการคำนวณตัวแบบวิเคราะห์ปัจจัยเพื่อการจัดอันดับ โดยใช้ส่วนเบี่ยงเบนมาตรฐาน (SE) เป็นเกณฑ์ในการเปรียบเทียบ การเปรียบเทียบประสิทธิภาพเน้นที่ความเร็วในการลู่เข้าของตัวประมาณเบส์

วัตถุประสงค์ของการวิจัย

เพื่อเปรียบเทียบประสิทธิภาพของการสุ่มตัวอย่างด้วยวิธีกิบส์และการสุ่มตัวอย่างด้วยวิธีฮิตแอนด์รันสำหรับทำการวิเคราะห์ปัจจัยเชิงเบส์ เมื่อมีปัจจัย 1 ปัจจัย โดยมีวัตถุประสงค์เพื่อการจัดอันดับ

ขอบเขตของการวิจัย

การดำเนินงานวิจัยในครั้งนี้ มีขอบเขตของการวิจัย คือ

1. ปัจจัยในตัวแบบวิเคราะห์ปัจจัยจำนวน 1 ตัว
2. ประเภทตัวแปรหรือตัวชี้วัด รวมทั้ง ตัวแปรแบบต่อเนื่อง และ ตัวแปรเชิงอันดับ
3. ทดลองกับข้อมูลจากชุดข้อมูลความเสี่ยงทางการเมือง-เศรษฐกิจจาก MCMCpack 0.4-8 โดย Martin และ Quinn 2004 ใน R package ซึ่งเป็นข้อมูลที่เป็นมาตรฐานและได้รับการเผยแพร่ ซึ่งนักวิจัยสามารถในไปใช้ในการทวนสอบผลการศึกษาได้
4. เปรียบเทียบประสิทธิภาพของวิธี MCMC ด้วยวิธีค่าเฉลี่ยกลุ่ม (Batch Mean Method) วิธีที่มีประสิทธิภาพมากกว่าคือ วิธีที่ให้ความถูกต้องมากที่สุดในจำนวนการสุ่มเท่ากัน โดยจะพิจารณาจากค่าคลาดเคลื่อนมาตรฐาน (SE) ของตัวประมาณของพารามิเตอร์

คำจำกัดความที่ใช้ในการวิจัย

1. สัมประสิทธิ์ของปัจจัย (Factor loading) คือ ค่าสัมประสิทธิ์ของความสัมพันธ์ระหว่างตัวแปรและปัจจัย โดยตัวแปรที่มีความสัมพันธ์สูงที่สุดจะให้ความหมายในแง่ของการตีความหมายของปัจจัยมากที่สุด

2. คะแนนของปัจจัย (Factor Score) คือ คะแนนของแต่ละหน่วยสังเกต Observation บนปัจจัยใหม่แต่ละปัจจัย โดยคะแนนปัจจัยนี้ เป็นผลรวมเชิงเส้นของตัวแปรเดิมทั้งหมดที่เกี่ยวเนื่องกับปัจจัยใหม่แต่ละปัจจัย และใช้คะแนนเฉลี่ยเป็นดัชนีในการจัดอันดับ

3. ตัวแปรแฝง (Latent variable) เป็นตัวแปรที่ไม่ถูกสังเกตโดยตรง แต่จะถูกอนุมานผ่าน แบบจำลองทางคณิตศาสตร์จากตัวแปรอื่นๆ ที่สังเกตและวัดค่าได้ แบบจำลองทางคณิตศาสตร์ซึ่งมี วัตถุประสงค์เพื่ออธิบายตัวแปรสังเกตได้ในแง่ของตัวแปรแฝง เรียกว่า latent variable models ซึ่งนำไปใช้ในหลากหลายสาขา เช่น เศรษฐกิจ การเรียนรู้ ภาษาศาสตร์ จิตวิทยา และ สังคมวิทยา

ประโยชน์ที่คาดว่าจะได้รับ

1. เพื่อเป็นแนวทางในการศึกษาเปรียบเทียบหรือพัฒนาวิธีทางสถิติที่ใช้ในการจัดอันดับ

2. เพื่อเป็นแนวทางในการศึกษาเปรียบเทียบตัวแบบที่ใช้ในการจัดอันดับอื่นๆ ต่อไป

วิธีดำเนินการวิจัย

1. ทำความเข้าใจ และ ตรวจสอบความถูกต้องของขั้นตอนวิธีของ Quinn

2. เขียนโปรแกรมให้ดำเนินงานตามขั้นตอนวิธีของ Quinn โดยใช้โปรแกรม R

3. เขียนโปรแกรมให้ดำเนินงานตามขั้นตอนวิธีฮิตแอนด์รัน

4. ทดลองกับชุดข้อมูลความเสี่ยงทางการเมือง-เศรษฐกิจจาก MCMCpack 0.4-8 โดย Martin และ Quinn 2004 ใน R package

5. วิเคราะห์และเปรียบเทียบผล

ลำดับขั้นตอนในการเสนอผลการวิจัย

ในปี พ.ศ.2553 นี้ จุฬาลงกรณ์มหาวิทยาลัย ได้รับเกียรติจากคณะกรรมการเครือข่ายการวิจัยดำเนินงาน ให้เป็นเจ้าภาพจัดการประชุมวิชาการ OR-Net Conference 2010 ซึ่งคณะกรรมการจัดงานได้กำหนดให้จัดงานขึ้นในวันที่ 2-3 กันยายน 2553 ณ อาคารศศปาฐศาลา สถาบันบัณฑิตบริหารธุรกิจศศินทร์ จุฬาลงกรณ์มหาวิทยาลัย โดยมีลำดับขั้นตอนในการเสนอผลการวิจัย ดังต่อไปนี้

1. ส่งบทคัดย่อเพื่อให้กรรมการผู้ทรงคุณวุฒิพิจารณาว่าผ่านหรือไม่

2. เมื่อผ่านการพิจารณาบทคัดย่อแล้ว เตรียมบทความที่ประกอบด้วยบทคัดย่อ บทนำ วัตถุประสงค์ วิธีการวิจัยผลการวิจัย บทสรุป และเอกสารอ้างอิง ตามรูปแบบของบทความสำหรับการประชุมวิชาการวิจัยดำเนินการ OR-Net Conference 2010

3. ส่งบทความฉบับเต็มเข้าระบบเพื่อให้กรรมการผู้ทรงคุณวุฒิพิจารณาอีกครั้ง

4. แก้ไขบทความตามที่กรรมการผู้ทรงคุณวุฒิแนะนำเมื่อผลงานวิจัยผ่านการพิจารณาคัดเลือกผลงานให้เข้าร่วมการประชุมวิจัยดำเนินการ OR-Net Conference 2010

5. ส่งบทความฉบับสมบูรณ์ที่ทำการแก้ไขเรียบร้อยแล้ว

6. ผู้วิจัยนำเสนอผลงานวิจัยด้วยตนเองในวันที่ 3 กันยายน พ.ศ. 2553 เวลา 10.40 น.- 11.00 น. ณ อาคารศศปาฐศาลา สถาบันบัณฑิตบริหารธุรกิจศศินทร์ จุฬาลงกรณ์มหาวิทยาลัย

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

โดยทั่วไปแล้วเวลาเก็บข้อมูลเพื่อนำมาวิเคราะห์ ข้อมูลที่ได้มามักมีตัวแปรหลายประเภท ดังนั้นในการจัดอันดับโดยใช้การวิเคราะห์ปัจจัย ได้มีการปรับให้สามารถวิเคราะห์ข้อมูลแบบผสมที่มีทั้งตัวแปรแบบต่อเนื่องและตัวแปรเชิงอันดับ โดยทำการวิเคราะห์ปัจจัยผ่านตัวแปรแฝง x^* แต่ด้วยความซับซ้อนของพารามิเตอร์ทั้งหมดที่ต้องประมาณค่า รวมทั้งตัวแปรแฝงที่ต้องใช้ในการวิเคราะห์ จึงใช้การสุ่มตัวอย่างด้วยวิธีกิบส์และการสุ่มตัวอย่างวิธีฮิตแอนด์รันที่อยู่ในกลุ่มวิธีลูกโซ่มาร์คอฟมอนติคาร์โล (Markov chain Monte Carlo) หรือ MCMC ในการจำลองข้อมูล ซึ่งลูกโซ่มาร์คอฟจะลู่เข้าสู่การแจกแจงคงตัว (stationary distribution) ภายใต้เงื่อนไขบางประการ แล้วใช้วิธีค่าเฉลี่ยกลุ่ม (Batch mean method) สำหรับการประมาณค่าพารามิเตอร์ และค่าคลาดเคลื่อนมาตรฐาน (SE) ของค่าประมาณ เพื่อนำค่าคลาดเคลื่อนมาตรฐานจากวิธี MCMC ทั้ง 2 มาเปรียบเทียบประสิทธิภาพในการลู่เข้าของตัวประมาณ

1. ตัวแบบการวิเคราะห์ปัจจัยสำหรับข้อมูลแบบผสมที่มีทั้งตัวแปรแบบต่อเนื่องและตัวแปรเชิงอันดับ

ให้ X เป็นเมทริกซ์ของค่าสังเกตขนาด $N \times J$ ซึ่งแต่ละตัวแปรสามารถเป็นได้ทั้งตัวแปรแบบต่อเนื่องและตัวแปรเชิงอันดับที่มี $C_j > 1$กลุ่ม โดยให้ $i = 1,2,\ldots,N$ แทนดัชนีของหน่วยที่ให้ค่าสังเกต $j = 1,2,\ldots,J$ แทนดัชนีของตัวแปร

สมมติให้ ค่าของแต่ละสมาชิกใน X ได้มาจาก X^* ซึ่งเป็นเมทริกซ์ของตัวแปรแฝง (Latent variable) ขนาด $N \times J$ และจุดตัด γ

$$x_{ij} = \begin{cases} x_{ij}^* & \text{ถ้า ตัวแปรที่ } j \text{ เป็นตัวแปรแบบต่อเนื่อง} \\ c & \text{ถ้า } x_{ij}^* \in (\gamma_{j(c-1)}, \gamma_{jc}] \text{ และตัวแปรที่ } j \text{ เป็นตัวแปรเชิงอันดับ} \end{cases}$$

โดยที่ c มีค่าเป็นไปได้ตั้งแต่ $1,2,\ldots,C_j$ และให้ $\gamma_{j0} = -\infty, \gamma_{j1} = 0$ และ $\gamma_{jC_j} = \infty$

รูปแบบของการรวมตัวระหว่างตัวแปรสังเกตได้ X ถูกจำลองโดยตัวแบบการวิเคราะห์ปัจจัยผ่านตัวแปรแฝง X^*

$$x_i^* = \Lambda\phi_i + \varepsilon_i \quad , i = 1,\ldots,N$$

โดยที่ $\phi_i \overset{iid}{\sim} \mathrm{N}(0, \mathrm{I}) \quad , i = 1, \ldots, N$

และ $\varepsilon_i \overset{iid}{\sim} \mathrm{N}(0, \Psi) \quad , i = 1, \ldots, N$

ซึ่งสามารถเขียนให้อยู่ในรูปของ N สมการได้ดังนี้

$$X^*_{N \times J} = \Phi_{N \times 2} \Lambda'_{2 \times J} + \mathrm{E}_{N \times J}$$

เมื่อ X^* เป็นเมทริกซ์ของตัวแปรแฝงขนาด $N \times J$ Λ เป็นเมทริกซ์ขนาด $J \times 2$ ของสัมประสิทธิ์ของปัจจัย เมื่อกำหนดให้มี 1 ปัจจัย, Φ เป็นเมทริกซ์ขนาด $N \times 2$ โดยสมาชิกในหลักที่ 1 เป็น 1 ทั้งหมดและสมาชิกในหลักที่ 2 เป็นคะแนนของปัจจัย และ Ψ เป็นเมทริกซ์ทแยงมุมที่มีขนาด $J \times J$ โดยให้ Ψ_{jj} เท่ากับ 1 เมื่อตัวแปรที่ j เป็นตัวแปรอันดับ และ $\Psi_{jj} \overset{iid}{\sim} IG(a_o/2, b_o/2)$ เมื่อตัวแปรที่ j เป็นตัวแปรแบบต่อเนื่อง

ในการวิเคราะห์ปัจจัยเชิงเบส์สำหรับการแจกแจงก่อน (Prior distribution) ของพารามิเตอร์ที่เกี่ยวข้องเป็นดังต่อไปนี้

Λ จะต้องมีสมาชิกอย่างน้อยหนึ่งตัวในหลักที่ 2 ที่ถูกกำหนดให้มีเครื่องหมายเป็นบวกหรือลบเท่านั้น เพื่อกำจัดปัญหาการไม่แปรเปลี่ยนภายใต้การหมุน (Rotational invariance) โดยสมมติให้สมาชิกของ Λ ที่ถูกกำหนดเครื่องหมาย มีการแจกแจงปกติแบบตัดปลายที่ตัดความหนาแน่นบริเวณที่ต่ำกว่าหรือเหนือกว่าศูนย์ออก ขึ้นอยู่กับเครื่องหมายที่กำหนด และสมมติให้สมาชิกของ Λ ทุกตัว ทั้งที่ถูกกำหนดเครื่องหมายและไม่ได้ถูกกำหนดเครื่องหมายเป็นอิสระซึ่งกันและกัน มีการแจกแจงแบบปกติ มีค่าเฉลี่ยเท่ากับ l_{oj} และค่าความแม่นยำ (ซึ่งเป็นส่วนกลับของความแปรปรวน) เท่ากับ L_{oj} สมมติให้ γ มีการแจกแจงแบบสม่ำเสมอไม่ตรงแบบ (Improper Uniform) และ Λ, ϕ, Ψ และ γ เป็นอิสระซึ่งกันและกัน

ในการจัดอันดับโดยใช้การวิเคราะห์ปัจจัยเชิงเบส์ เราจะใช้ค่าคาดหวังภายใต้การแจกแจงความน่าจะเป็นภายหลังของคะแนนของปัจจัย (ϕ_i) มาใช้ในการจัดอันดับ ซึ่งการประมาณค่าคาดหวังดังกล่าวสามารถทำได้จากการจำลองโดยใช้เทคนิคลูกโซ่มาร์คอฟ (MCMC) โดยในงานวิจัยนี้จะใช้การสุ่มตัวอย่างแบบกิบส์ (Gibbs Sampler) และการสุ่มตัวอย่างแบบฮิตแอนด์รัน (Hit-and-run Sampler) แล้วนำผลจากการจำลองมาเฉลี่ยเพื่อมาประมาณค่าคาดหวัง แต่เนื่องจากการประมาณค่าย่อมมีความผิดพลาด จึงใช้วิธีค่าเฉลี่ยกลุ่ม (batch mean method)

ประมาณค่าคลาดเคลื่อนมาตรฐาน (standard error) ของตัวประมาณ และนำค่าดังกล่าวมาใช้ในการเปรียบเทียบประสิทธิภาพของการสุ่มตัวอย่างแบบกิบส์ กับการสุ่มตัวอย่างแบบฮิตแอนด์รัน

2. **ลูกโซ่มาร์คอฟ** (Markov Chain)

พิจารณากระบวนการเฟ้นสุ่มแบบไม่ต่อเนื่อง (discrete-time stochastic process) $\{X_n; n = 0,1,2,\ldots\}$ กระบวนการนี้เป็นลูกโซ่มาร์คอฟแบบไม่ต่อเนื่อง (discrete-time Markov chain) ถ้า $m, n \geq 0$ และสถานะ $i, j \in S, x_u$ และ $0 \leq u < m$ เป็นจำนวนเต็มที่ไม่เป็นลบ (nonnegative integers)

$$P(X_{m+n} = j \mid X_m = i, X_u = x_u, 0 \leq u < m) = P(X_{m+n} = j \mid X_m = i)$$

ลูกโซ่มาร์คอฟแบบไม่ต่อเนื่อง หรือลูกโซ่มาร์คอฟ เป็นกระบวนการเฟ้นสุ่มที่มีคุณสมบัติมาร์คอฟ (Markov property) คือ ความน่าจะเป็นแบบมีเงื่อนไขของอนาคต X_{m+n} เมื่อกำหนดปัจจุบัน X_m และอดีต X_u โดยที่ $0 \leq u < m$ จะขึ้นอยู่กับปัจจุบันเท่านั้น และเป็นอิสระกับอดีต X_u เรียก S ว่า สเปซของสถานะ (state space)

2.1 **ทฤษฎีลูกโซ่มาร์คอฟสำหรับสเปซทั่วไป**

สำหรับลูกโซ่มาร์คอฟบนสเปซ $(S, \Im)$ และ $\{X_n; n = 0,1,2,\ldots\}$ เป็นลูกโซ่มาร์คอฟ ถ้า

$$P(X_{n+1} \in A \mid X_n, X_{n-1}, \ldots, X_0) = P(X_{n+1} \in A \mid X_n)$$

$$= P(X_n, A) \qquad \text{สำหรับ } A \in \Im$$

ซึ่ง $P(X_n, A)$ มีคุณสมบัติดังนี้

1. สำหรับแต่ละ $x \in \Im$ ฟังก์ชัน $P(x, \cdot)$ เมเซอร์ความน่าจะเป็น (measure probability) บน $(S, \Im)$
2. สำหรับแต่ละ $A \in \Im$ ฟังก์ชัน $P(\cdot, A)$ สามารถวัดได้ (measurable)

<u>นิยาม</u> ทฤษฎีบทจำกัด

$$\mu P(A) = \int_S P(x, A)\mu(dx)$$

เมื่อกำหนดความน่าจะเป็นในการเปลี่ยนสถานะขั้นที่ 1 คือ

$$P^1(x,A) = P(x,A)$$

และความน่าจะเป็นในการเปลี่ยนสถานะขั้นที่ n คือ

$$P^n(x,A) = \int_S P^{n-1}(y,A)P(x,dy)$$

ดังนั้น

$$P(X_n \in A) = \mu P^n(A)$$

เมเซอร์ความน่าจะเป็น π บน $(S,\Im)$ เป็นการแจกแจงคงตัว (stationary distribution) ของ P ถ้า

$$\pi P = P$$

2.2 ทฤษฎีบทการลู่เข้าของลูกโซ่มาร์คอฟในกรณี total variation (Markov Chain Convergence in Total Variation)

ให้ P เป็น φ ที่ irreducible สำหรับบางค่า $\sigma-$ เมเซอร์ที่หาค่าได้ φ บน $(S,\Im)$ ถ้า P เป็น aperiodic แล้ว

$$\lim_{n\to\infty} P(X_n \in A \mid X_0 = x) = \pi(A) \text{ สำหรับทุก } A \in \Im$$

เมื่อ π เป็นการแจกแจงคงตัว (stationary distribution) ของ P ดังนั้น

$$\pi(\cdot) = \pi P(\cdot) = \int_S P(x,\cdot)\pi(dx)$$

1. The First Lyapunov Condition (FLC)

สำหรับเซตไม่ว่าง $B \subset S$ สเกลาร์ที่เป็นบวก (positive scalars) $a<1$ b และ δ เป็นจำนวนเต็ม $m \geq 1$ การแจกแจงความน่าจะเป็น φ บน S และฟังก์ชัน $V: S \to [1,\infty]$ จะได้ว่า

- $P(X_m \in \cdot \mid X_0 \in z) \geq \delta\varphi(\cdot)$ สำหรับทุก $z \in B$
- $E\big(V(X_1) \mid X_0 = z\big) \leq aV(z) + bI(z \in B)$ สำหรับทุก $z \in S$

หมายความว่า ถ้าลูกโซ่มาร์คอฟเป็นไปตามเงื่อนไข FLC และมีคุณสมบัติ aperiodic แล้วลูกโซ่มาร์คอฟนี้เป็น ergodic อย่างสม่ำเสมอ (Uniformly ergodic)

2. Markov Chain Strong Law of Large Numbers (MCSLLN)

ให้ $\{X_n; n = 0,1,2,\ldots\}$ เป็นลูกโซ่มาร์คอฟที่ ergodic อย่างสม่ำเสมอ (uniformly ergodic Markov chain) บนสเปซ S ที่มี π เป็นการแจกแจงคงตัว (stationary distribution) และฟังก์ชัน $h: S \to \Re$ ถ้า $\pi|h| = E_\pi\left[|h(X_0)|\right] = \int_S |h(x)|\pi(dx) < \infty$ แล้ว

$$\frac{1}{n}\sum_{i=0}^{n-1} h(X_i) \to \pi h = \int_S h(x)\pi(dx) \quad \text{เมื่อ } n \to \infty$$

3. Markov Chain Central Limit Theorem (MCCLT)

ให้ $\{X_n; n = 0,1,2,\ldots\}$ เป็นลูกโซ่มาร์คอฟที่ ergodic อย่างสม่ำเสมอ (uniformly ergodic Markov chain) บนสเปซ S และฟังก์ชัน $h: S \to \Re$ ด้วยค่า $h^2(z) \le cV(z)$ สำหรับบางค่า $c > 0$ และทุกค่า z จะได้ว่า

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=0}^{n-1} h(X_t) - \pi h\right) \Rightarrow \sigma\ N(0,1)$$

เมื่อ π เป็นการแจกแจงความน่าจะเป็นคงตัว (stationary probability distribution) ของ X และ $\sigma^2 = Var_\pi\left[h(X_0)\right] + 2\sum_{k=1}^{\infty} \mathrm{cov}_\pi\left[h(X_0), h(X_k)\right]$

4. Reversibility

ถ้าลูกโซ่มาร์คอฟใดมีคุณสมบัติ reversibility แล้ว จะสามารถออกแบบเมทริกซ์ของความน่าจะเป็นในการเปลี่ยนสถานะ (transition probability matrix) ได้ง่ายขึ้น

<u>ทฤษฎีบท</u> ถ้าความน่าจะเป็นในการเปลี่ยนสถานะ P มีคุณสมบัติ reversibility ตามความน่าจะเป็น π แล้ว π จะเป็นการแจกแจงคงตัวของ P

ให้ $A \subset S$ แล้ว

$$\begin{aligned}\pi P(A) &= \int_S P(x,A)\pi(dx) \\ &= \int_A P(x,S)\pi(dx) \\ &= \int_A 1\pi(dx) \\ &= \pi(A)\end{aligned}$$

2.3 **การวิเคราะห์สถานะเสถียรภาพของการจำลอง** (Steady State Analysis of a Simulation)

ในการจำลองกระบวนการเฟ้นสุ่มที่เป็นลูกโซ่มาร์คอฟ $\{X_n; n = 0,1,2,\ldots\}$ เมื่อ X_n เป็นเวกเตอร์ของตัวแปรสุ่ม ซึ่ง $X_n \in S$ และ S เป็นสเปซของสถานะ (state space) โดยเงื่อนไขทั่วไป การแจกแจงของ X_n จะลู่เข้าสู่การแจกแจงคงตัว π ซึ่งในการวิเคราะห์เชิงเบส์ก็คือการแจกแจงภายหลัง และเป็นอิสระกับจุดเริ่มต้น x_0 ดังนี้

$$p(X_n \in A \mid X_0 = x_0) \to \pi(A)$$

จาก MCSLLN จะได้ว่า ค่าเฉลี่ยในระยะยาว (long run average) จะสามารถประมาณค่า πh ดังนี้

$$\frac{1}{N}\sum_{i=0}^{N-1} h(X_i) \to \pi h \text{ สำหรับ } N \text{ ที่มีค่ามาก}$$

จาก MCCLT จะได้ว่า การแจกแจงของค่าเฉลี่ยในระยะยาว (distribution of long run average) จะประมาณ (approximate) ด้วยการแจกแจงแบบปกติ (normal distribution) ดังนี้

$$\frac{1}{N}\sum_{i=0}^{N-1} h(X_i) \sim N(\pi h, \sigma^2) \text{ สำหรับ } N \text{ ที่มีค่ามาก}$$

เมื่อ $\sigma^2 = Var_\pi[h(X_0)] + 2\sum_{k=1}^{\infty} \mathrm{cov}_\pi[h(X_0), h(X_k)]$

3. **เทคนิคลูกโซ่มาร์คอฟมอนติคาร์โล** (Markov Chain Monte Carlo)

ในทฤษฎีบทของเบส์กรณีที่การแจกแจงภายหลัง(Posterior distribution) ไม่ได้อยู่ในรูปแบบสามัญนั้น การคำนวณค่าคาดหวังของพารามิเตอร์จากฟังก์ชันความหนาแน่นน่าจะเป็นภายหลังเป็นปัญหาสำคัญของการสร้างแบบจำลองเบส์ เทคนิคลูกโซ่มาร์คอฟมอนติคาร์โล (Markov Chain Monte Carlo) หรือ MCMC จึงได้รับการเสนอเพื่อแก้ปัญหาดังกล่าว โดย MCMC เป็นเทคนิคที่ใช้ในการสุ่มตัวอย่างของตัวแปรสุ่มจากฟังก์ชันความหนาแน่นน่าจะเป็นภายหลัง

MCMC เป็นการสร้างลำดับของเวคเตอร์สุ่ม $\{\vec{\theta}_0, \vec{\theta}_1, \ldots, \vec{\theta}_n\}$ โดยที่การสุ่ม $\vec{\theta}_{j+1}$ จากการแจกแจง $p(\vec{\theta}_{j+1} | \vec{\theta}_j)$ เมื่อ $j \geq 0$ ซึ่งหมายถึง $\vec{\theta}_{j+1}$ ถูกสุ่มโดยขึ้นอยู่กับ $\vec{\theta}_j$ เพียงอย่างเดียวไม่ได้ขึ้นอยู่กับ $\{\vec{\theta}_0, \vec{\theta}_1, \ldots, \vec{\theta}_{j-1}\}$ นั่นคือ $p(\vec{\theta}_{j+1} | \vec{\theta}_j)$ เป็น Transition Kernel ซึ่งเป็นความ

น่าจะเป็นของการเปลี่ยนสถานะจากสถานะที่ j ไปยังสถานะที่ $j+1$ โดยเงื่อนไขทั่วไป เมื่อ j มีค่ามากขึ้น การแจกแจงของ $\vec{\theta}_j$ จะลู่เข้าสู่การแจกแจงคงตัว π (stationary distribution) ในกรณีการวิเคราะห์เชิงเบส์ π ก็คือการแจกแจงภายหลัง

ในการสร้าง MCMC นั้นจะมีขั้นตอนวิธี (Algorithm) หลายแบบ เช่น การสุ่มตัวอย่างแบบเมโทรโพลิส-เฮสติงส์ (Metropolis-Hastings Sampling), การสุ่มตัวอย่างแบบกิบส์ (Gibbs Sampling) และ การสุ่มตัวอย่างแบบฮิตแอนด์รัน ซึ่งมีรายละเอียดดังต่อไปนี้

3.1 การสุ่มตัวอย่างแบบเมโทรโปลิส-เฮสติงส์ (Metropolis-Hastings Sampling)

สมมติว่าต้องการสุ่มพารามิเตอร์ $\vec{\theta}$ จากการแจกแจง π ซึ่งก็คือการแจกแจงภายหลัง (posterior distribution) ของการวิเคราะห์เชิงเบส์ ในการกระบวนการ (Algorithm) ของเมโทรโปลิส-เฮสติงส์มีขั้นตอนดังต่อไปนี้

ขั้นที่ 1 กำหนดพารามิเตอร์เริ่มต้น $\vec{\theta}_0$ โดยที่ $\pi > 0$

ขั้นที่ 2 สุ่มพารามิเตอร์ $\vec{\theta}^{(can)}$ จาก Proposal Distribution $q\left(\vec{\theta}_j \middle| \vec{\theta}_{j+1}\right)$ ซึ่งเป็นความน่าจะเป็นในการเปลี่ยนค่าพารามิเตอร์จาก $\vec{\theta}_j$ ไปเป็น $\vec{\theta}_{j+1}$

ขั้นที่ 3 คำนวณหาค่า α ซึ่งเป็นอัตราส่วนของฟังก์ชันความหนาแน่น π ของพารามิเตอร์ที่สุ่มขึ้นมาใหม่ $\vec{\theta}^{(can)}$ เทียบกับฟังก์ชันความหนาแน่นของพารามิเตอร์ $\vec{\theta}_j$

$$\alpha = \min\left(\frac{\pi\left(\vec{\theta}^{(can)}\right) q\left(\vec{\theta}^{(can)} \middle| \vec{\theta}_j\right)}{\pi\left(\vec{\theta}_j\right) q\left(\vec{\theta}_j \middle| \vec{\theta}^{(can)}\right)}, 1 \right)$$

ขั้นที่ 4 สุ่ม u จากการแจกแจงยูนิฟอร์ม (Uniform Distribution) $U(0,1)$

ยอมรับ $\vec{\theta}^{(can)}$ ถ้า $u < \alpha$ และให้ $\vec{\theta}_{j+1} = \vec{\theta}^{(can)}$

หรือ ปฏิเสธ $\vec{\theta}^{(can)}$ ถ้า $u \geq \alpha$ และให้ $\vec{\theta}_{j+1} = \vec{\theta}_j$

ขั้นที่ 5 วนซ้ำในขั้นที่ 2 ใหม่

โดยเงื่อนไขทั่วไป การแจกแจงของ $\vec{\theta}_j$ จะลู่เข้าสู่การแจกแจงคงตัว π (stationary distribution) ซึ่งสามารถนำไปประมาณค่าคาดหวังภายใต้การแจกแจงภายหลังได้

3.2 **การสุ่มตัวอย่างแบบกิบส์** (Gibbs Sampling)

การสุ่มตัวอย่างแบบกิบส์เป็นกรณีพิเศษของ Single-component Metropolis Hasting Method กล่าวคือจะสุ่ม $\vec{\theta}^{(can)}$ จาก Full Condition Distribution แทนการสุ่มจาก Proposal Distribution นั่นคือ

$$q\left(\vec{\theta}_j^{(can)}\middle|\vec{\theta}_j,\vec{\theta}_{-j}\right)=\pi\left(\vec{\theta}_j^{(can)}\middle|\vec{\theta}_{-j}\right)$$

และ $$q\left(\vec{\theta}_j\middle|\vec{\theta}_j^{(can)},\vec{\theta}_{-j}\right)=\pi\left(\vec{\theta}_j\middle|\vec{\theta}_{-j}\right)$$

ซึ่งทำให้

$$\alpha=\min\left(\frac{\pi\left(\vec{\theta}_j^{(can)}\middle|\vec{\theta}_{-j}\right)\,q\left(\vec{\theta}_j\middle|\vec{\theta}_j^{(can)},\vec{\theta}_{-j}\right)}{\pi\left(\vec{\theta}_j|,\vec{\theta}_{-j}\right)\,q\left(\vec{\theta}_j^{(can)}\middle|\vec{\theta}_j,\vec{\theta}_{-j}\right)},\,1\right)$$

$$=\min\left(\frac{\pi\left(\vec{\theta}_j^{(can)}\middle|\vec{\theta}_{-j}\right)\,p\left(\vec{\theta}_j\middle|\vec{\theta}_{-j}\right)}{\pi\left(\vec{\theta}_j|,\vec{\theta}_{-j}\right)\,p\left(\vec{\theta}_j^{(can)}\middle|\vec{\theta}_{-j}\right)},\,1\right)=1$$

การสุ่ม $\vec{\theta}_j^{(can)}$ จาก Full Conditional Distribution จึงถูกยอมรับในทุกๆครั้ง

ดังนั้น การสุ่มพารามิเตอร์ในการสุ่มตัวอย่างแบบกิบส์จึงเป็นการสุ่มจาก Full Conditional Distribution เป็นดังนี้

$$\pi\left(\vec{\theta}_j\middle|\vec{\theta}_{-j}\right)=\frac{\pi\left(\vec{\theta}\right)}{\int\pi\left(\vec{\theta}\right)d\theta_j}\text{ โดยที่ }\vec{\theta}_{-j}=\left\{\theta_1,\theta_2,\ldots,\theta_{j-1},\theta_{j+1},\ldots,\theta_J\right\}$$

โดยเงื่อนไขทั่วไป การแจกแจงของ $\vec{\theta}_j$ จะลู่เข้าสู่การแจกแจงคงตัว π (stationary distribution) ซึ่งสามารถนำไปประมาณค่าคาดหวังภายใต้การแจกแจงภายหลังได้

3.3 **การสุ่มตัวอย่างแบบฮิตแอนด์รัน** (Hit-and-run sampler)

ให้ $S\subset\Re^n$ และให้การแจกแจงเป้าหมาย (Target distribution) มีฟังก์ชันความน่าจะเป็น π บน S

สร้าง $\vec{\theta}_n$ ด้วยกระบวนการดังนี้

ขั้นที่ 1 กำหนดพารามิเตอร์เริ่มต้น $\vec{\theta}_0\in S$

ขั้นที่ 2 ที่ $\vec{\theta}_n=\vec{\theta}\in S$ สุ่มเลือกทิศทาง $\vec{d}$ บนพื้นผิวทรงกลมรัศมี 1 หน่วย ลากเส้นตรง L ผ่าน $\vec{\theta}$ และตัดกับ S โดยให้ $L_n=S\cap\left\{l:l=\vec{\theta}+r\vec{d},r\in\Re\right\}$

ขั้นที่ 3 สุ่มเลือกจุด $\vec{\theta}_{n+1}$ บนเส้นตรง L_n ด้วยการแจกแจงแบบ π โดยมีเงื่อนไขบน L_n

ขั้นที่ 4 วนซ้ำในขั้นที่ 2 ใหม่

โดยเงื่อนไขทั่วไป การแจกแจงของ $\vec{\theta}_j$ จะลู่เข้าสู่การแจกแจงคงตัว π (stationary distribution) ซึ่งสามารถนำไปประมาณค่าคาดหวังภายใต้การแจกแจงภายหลังได้

จากการลู่เข้าของลูกโซ่มาร์คอฟจึงมีค่าความแปรปรวนร่วม (covariance) เพราะ $\vec{\theta}_j$ แต่ละตัวไม่เป็นอิสระกัน ดังนั้น งานวิจัยนี้จึงใช้วิธีค่าเฉลี่ยกลุ่ม (Batch Means method) สำหรับแก้ปัญหาค่าความแปรปรวนร่วม

4. วิธีค่าเฉลี่ยกลุ่ม (Batch means method)

เป็นวิธีการหนึ่งที่ใช้ประมาณค่าเฉลี่ยประชากร โดยใช้การคำนวณ Monte Carlo standard error (MCSE) สมมติว่าเราสนใจประมาณค่าเฉลี่ย $E_\pi\left(g\left(\vec{\theta}\right)\right)$ เมื่อ $\vec{\theta}$ มีการแจกแจง π ซึ่งกระบวนการของค่าเฉลี่ยกลุ่ม (Batch means) มีดังนี้

ขั้นที่ 1 จำลองลูกโซ่มาร์คอฟ $\left\{\vec{\theta}_n\right\}$ จำนวน $n = ab$ รอบ โดยที่ a และ b เป็นจำนวนเต็ม a เป็นจำนวน batch และแต่ละ batch มีขนาด b

ขั้นที่ 2 หาค่าเฉลี่ยของตัวอย่าง ใน batch ที่ k จากสูตร $\overline{Y}_k = \frac{\sum_{i=(k-1)b+1}^{kb} g\left(\vec{\theta}_i\right)}{b}$

ขั้นที่ 3 กำหนดให้ $\hat{\sigma}_g^2 = \frac{b}{a-1}\sum_{k=1}^{a}\left(\overline{Y}_k - \overline{g}_n\right)^2$ โดยที่ $\overline{g}_n = \frac{1}{a}\sum_{k=1}^{a}\overline{Y}_k$ จะได้ว่าค่าประมาณของ Monte Carlo standard error คือ $\frac{\hat{\sigma}_g}{\sqrt{n}}$

ขั้นที่ 4 คำนวณหาช่วงความเชื่อมั่นของ $E_\pi\left(g\left(\vec{\theta}\right)\right)$ ได้จากสูตร

$$\overline{g}_n \pm t_{\frac{\alpha}{2},(a-1)}\frac{\hat{\sigma}_g}{\sqrt{n}}$$

เอกสารและงานวิจัยที่เกี่ยวข้อง

Chen และ Schmeiser (1993) ได้ทำการเปรียบเทียบประสิทธิภาพการสุ่มตัวอย่างแบบฮิตแอนด์รัน และการสุ่มตัวอย่างแบบกิบส์ พบว่าการสุ่มตัวอย่างแบบฮิตแอนด์รันมีประสิทธิภาพมากว่าสำหรับในกรณีที่เป็นการแจกแจงแบบปกติของ 2 ตัวแปร

Quinn (2004) ได้นำเสนอการใช้ตัวแบบการวิเคราะห์ปัจจัยเชิงเบส์ (Bayesian factor analysis) ซึ่งได้ปรับให้เหมาะกับการวิเคราะห์ข้อมูลแบบผสมที่มีทั้งตัวแปรแบบต่อเนื่องและตัวแปรเชิงอันดับ และใช้วิธีการคำนวณที่วางอยู่บนพื้นฐานของการสุ่มตัวอย่างแบบกิบส์ และวิธีของ Cowles (1993) ซึ่งวิธีดังกล่าวจัดอยู่ในกลุ่มวิธีลูกโซ่มาร์คอฟ มอนติคาร์โล (Markov chain Monte Carlo) หรือ MCMC ในการประมาณค่าพารามิเตอร์

Lovasz และ Vempala (2006) ได้ศึกษาการสุ่มตัวอย่างแบบฮิตแอนด์รันบนบริเวณคอนเวกซ์ พบว่า การสุ่มตัวอย่างแบบฮิตแอนด์รันเป็นหนึ่งในวิธีที่เร็วที่สุด ในการสร้างจุดตัวอย่างซึ่งลู่เข้าสู่การแจกแจงแบบสม่ำเสมอบน Log-concave

จึงเป็นที่น่าสนใจว่า หากแทนคำนวณวิธีการสุ่มตัวอย่างแบบกิบส์ด้วยการสุ่มตัวอย่างแบบฮิตแอนด์รันในตัวแบบการวิเคราะห์ปัจจัยเชิงเบส์ที่ Quinn ได้นำเสนอ ผลที่ได้ยังคงสอดคล้องกับงานวิจัยของ Chen และ Schmeiser และงานวิจัยของ Lovasz และ Vempala หรือไม่

บทที่ 3

วิธีดำเนินการวิจัย

การศึกษาวิจัยในครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของการสุ่มตัวอย่างด้วยวิธีกิบส์และการสุ่มตัวอย่างด้วยวิธีฮิตแอนด์รันสำหรับทำการวิเคราะห์ปัจจัยเชิงเบส์เมื่อมีปัจจัย 1 ปัจจัย โดยมีวัตถุประสงค์เพื่อการจัดอันดับ ทำการทดลองโดยใช้ข้อมูลความเสี่ยงทางการเมือง-เศรษฐกิจจาก MCMCpack R Package โดย Martin และ Quinn 2004 เพื่อจัดอันดับประเทศ 62 ประเทศตามความเสี่ยงทางการเมือง-เศรษฐกิจ ซึ่งมีดัชนี 5 ตัว ตือ

1. ความเป็นอิสระของระบบตุลาการ (courts) เป็นตัวแปรเชิงอันดับที่มี 2 กลุ่ม
2. ส่วนต่างอัตราแลกเปลี่ยนเงินตราต่างประเทศในตลาดมืด (barb2) เป็นตัวแปรแบบต่อเนื่อง
3. ความเสี่ยงจากการยึดทรัพย์สินในกิจการมาเป็นของรัฐ (prsexp2) เป็นตัวแปรเชิงอันดับที่มี 6 กลุ่ม
4. การทุจริตในภาครัฐ (prscorr2) เป็นตัวแปรเชิงอันดับที่มี 6 กลุ่ม
5. ผลิตภัณฑ์มวลรวมภายในประเทศ (gdpw2) เป็นตัวแปรแบบต่อเนื่อง

ในการประมาณค่าพารามิเตอร์ ใช้ขั้นตอนวิธีของเทคนิคของลูกโซ่มาร์คอฟมอนติคาร์โล (Markov Chain Monte Carlo) หรือ MCMC ซึ่งคำนวณมาจากความน่าจะเป็นภายหลังซึ่งมีรูปแบบของความสัมพันธ์กับความน่าจะเป็นก่อนจากในบทที่ 2 ดังนี้

$$p(X^*, \gamma, \Lambda, \phi, \Psi \mid X) \propto p(X \mid X^*, \gamma)\, p(X^* \mid \Lambda, \phi, \Psi)\, p(\gamma)\, p(\Lambda),\, p(\Phi)\, p(\Psi)$$

โดย $p(\gamma)\, p(\Lambda),\, p(\Phi)\, p(\Psi)$ เป็นไปตามตัวแบบในบทที่ 2

เมื่อ X เป็นเมทริกซ์ของค่าสังเกต

X^* เป็นเมทริกซ์ของตัวแปรแฝง

γ เป็นเวกเตอร์ของจุดตัด

λ_1 เป็นเวกเตอร์หลักที่ 1 ของ Λ แสดงค่าเฉลี่ยของตัวแปรที่ j

λ_2 เป็นเวกเตอร์หลักที่ 2 ของ Λ แสดงน้ำหนักของปัจจัยของตัวแปรที่ j

ϕ_i เป็นเวกเตอร์หลักที่ 2 ของ Φ แสดงคะแนนของปัจจัยของหน่วยที่ i

ψ_{jj} เป็นความแปรปรวนของค่าคลาดเคลื่อนของตัวแปรที่ j

ในการสุ่มตัวอย่างแบบกิบส์จำเป็นจะต้องใช้ Full Conditional Distribution ซึ่งจากตัวแบบในบทที่ 2 Full Conditional Distribution ของพารามิเตอร์ที่เกี่ยวข้องเป็นดังนี้

สำหรับตัวแปรเชิงอันดับ Full conditional distribution ของ γ_{jc} มีการแจกแจงแบบสม่ำเสมอบนช่วงระหว่าง ค่าต่ำสุดเท่ากับ $\max(\max\{x_{ij}^* : x_{ij} = c\}, \gamma_{j(c-1)})$ และค่าสูงสุดเท่ากับ $\min(\min\{ x_{ij}^* : x_{ij} = c+1\}, \gamma_{j(c+1)})$ โดยให้ $\gamma_{j1} = 0$

กำหนดให้ λ_j เป็นเวกเตอร์แถวที่ j ของ Λ สำหรับ Full conditional distribution ของ x_{ij}^* คือ การแจกแจงแบบปกติตัดปลาย (Truncated Normal distribution) ที่มีค่าเฉลี่ยเท่ากับ $\lambda_j'\phi_i$ และมีความแปรปรวนเท่ากับ 1 โดยจะคำนวณความหนาแน่นเฉพาะในช่วง $(\gamma_{j(x_{ij}-1)}, \gamma_{j(x_{ij})}]$ เมื่อ j เป็นตัวแปรเชิงอันดับ และเป็นความหนาแน่น ณ จุด x_{ij} เมื่อ j เป็นตัวแปรแบบต่อเนื่อง

Full conditional distribution ของ λ_1 คือ การแจกแจงแบบปกติที่มีค่าเฉลี่ยเท่ากับ $\left(L_{0j} + \psi_{jj}^{-1}N\right)^{-1}\left(L_{0j}l_{0j} + \psi_{jj}^{-1}(x_j^* - \lambda_2'\phi_i)\right)$ และมีความแปรปรวนเท่ากับ $\left(L_{0j} + \psi_{jj}^{-1}N\right)^{-1}$ ในส่วนของ λ_2 จะต้องมีสมาชิกอย่างน้อย 1 ตัวที่ถูกกำหนดเครื่องหมาย ดังนั้น Full conditional distribution ของ λ_2 คือ การแจกแจงแบบปกติที่มีค่าเฉลี่ยเท่ากับ $\left(L_{0j} + \psi_{jj}^{-1}\phi_i'\phi_i\right)^{-1}\left(L_{0j}l_{0j} + \psi_{jj}^{-1}\phi_i'(x_j^* - \lambda_1)\right)$ และมีความแปรปรวนเท่ากับ $\left(L_{0j} + \psi_{jj}^{-1}\phi_i'\phi_i\right)^{-1}$ (การแจกแจงปกติแบบตัดปลายสำหรับตัวที่ถูกกำหนดเครื่องหมาย)

Full conditional distribution ของ ϕ_i คือ การแจกแจงแบบปกติที่มีค่าเฉลี่ยเท่ากับ $\left(\mathrm{I} + \lambda_2'\Psi^{-1}\lambda_2\right)^{-1}\left(\lambda_2'\Psi^{-1}\left(x_j^* - \lambda_1\right)\right)$ และมีความแปรปรวนเท่ากับ $\left(\mathrm{I} + \lambda_2'\Psi^{-1}\lambda_2\right)^{-1}$

สำหรับ ψ_{jj} เมื่อ j เป็นตัวแปรแบบต่อเนื่อง Full conditional distribution ของ ψ_{jj} คือ การแจกแจงแกมมาแบบผกผัน ที่มีพารามิเตอร์ $\left(a_{0j} + N\right)/2$ และ $\left(b_{0j} + \left(x_j^* - \Phi\lambda_j\right)'\left(x_j^* - \Phi\lambda_j\right)\right)$ เมื่อ j เป็นตัวแปรเชิงอันดับ ψ_{jj} มีค่าเท่ากับ 1

ใช้ขั้นตอนวิธีของเทคนิค MCMC ได้แก่ การสุ่มตัวอย่างแบบกิบส์ (Gibbs Sampler) และการสุ่มตัวอย่างแบบบฮิตแอนด์รัน (Hit-and-run Sampler) ในการจำลองแล้วใช้วิธีค่าเฉลี่ยกลุ่ม (Batch means method) ประมาณค่าคะแนนเฉลี่ยและค่าคลาดเคลื่อนมาตรฐาน (SE) ซึ่งจะใช้เป็นเกณฑ์ในการเปรียบเทียบประสิทธิภาพในการลู่เข้าของตัวประมาณ ในบทนี้จึงเป็นการนำเสนอวิธีดำเนินการวิจัยซึ่งผู้วิจัยได้ดำเนินการตามขั้นตอนต่อไปนี้

1. ดำเนินการด้วยวิธีการสุ่มตัวอย่างแบบกิบส์
2. ดำเนินการด้วยวิธีสุ่มตัวอย่างแบบฮิตแอนด์รัน
3. เปรียบเทียบประสิทธิภาพของการสุ่มตัวอย่างทั้ง 2 วิธี
4. สรุปผลการทดลอง

โดยขั้นตอนที่ 1-2 มีรายละเอียดของการดำเนินการดังนี้

ขั้นตอนในการดำเนินงานวิจัย

1.1 ขั้นตอนดำเนินการด้วยวิธีการสุ่มตัวอย่างแบบกิบส์

1) ทำการสร้างจุดเริ่มต้นของ γ_{jc} , x_{ij}^{*} , ϕ_i , λ_{j1} , λ_{j2} และ ψ_{jj}

2) เริ่มต้นให้จำนวนรอบ $n=1$

3) สุ่ม γ_{jc} สำหรับ j ที่เป็นตัวแปรเชิงอันดับจาก Full conditional distribution ที่มีการแจกแจงแบบสม่ำเสมอ

$$U\left(\max(\max\{x_{ij}^{*} : x_{ij}=c\}, \gamma_{j(c-1)}), \min(\min\{x_{ij}^{*} : x_{ij}=c+1\}, \gamma_{j(c+1)})\right)$$

4) สุ่ม x_{ij}^{*} สำหรับ j ที่เป็นตัวแปรเชิงอันดับจาก Full conditional distribution $N\left(\lambda_j'\phi_i, 1\right)$ โดยจะคำนวณความหนาแน่นเฉพาะในช่วง $(\gamma_{j(x_{ij}-1)}, \gamma_{j(x_{ij})}]$ โดยตัวแปร j ที่มี $C_j > 2$ จะสุ่ม x_{ij}^{*} ในแต่ละกลุ่มที่ $x_{ij}=c$

5) สุ่ม ϕ_i สำหรับทุก $i=1,\ldots,N$ จาก Full conditional distribution ที่มีการแจกแจงแบบปกติ $N\left(\left(\mathrm{I}+\lambda_2'\Psi^{-1}\lambda_2\right)^{-1}\left(\lambda_2'\Psi^{-1}\left(x_j^{*}-\lambda_1\right)\right), \left(\mathrm{I}+\lambda_2'\Psi^{-1}\lambda_2\right)^{-1}\right)$

6) สุ่ม λ_{j1} สำหรับ j ที่เป็นตัวแปรเชิงอันดับจากFull conditional distribution ที่มีการแจกแจงแบบปกติ $N\left(\left(L_{0j}+\psi_{jj}^{-1}N\right)^{-1}\left(L_{0j}l_{0j}+\psi_{jj}^{-1}(x_j^{*}-\lambda_2'\phi_i)\right), \left(L_{0j}+\psi_{jj}^{-1}N\right)^{-1}\right)$

7) สุ่ม λ_{j2} สำหรับทุก $j=1,\ldots,J$ โดยกำหนดให้ λ_{12} (λ_2 ของตัวแปร courts) มีเครื่องหมายเป็นลบ จาก Full conditional distribution ที่มีการแจกแจงแบบปกติ $N\left(\left(L_{0j}+\psi_{jj}^{-1}\phi_i'\phi_i\right)^{-1}\left(L_{0j}l_{0j}+\psi_{jj}^{-1}\phi_i'(x_j^{*}-\lambda_1)\right), \left(L_{0j}+\psi_{jj}^{-1}\phi_i'\phi_i\right)^{-1}\right)$ (การแจกแจงปกติแบบตัดปลาย สำหรับตัวที่ถูกกำหนดเครื่องหมาย)

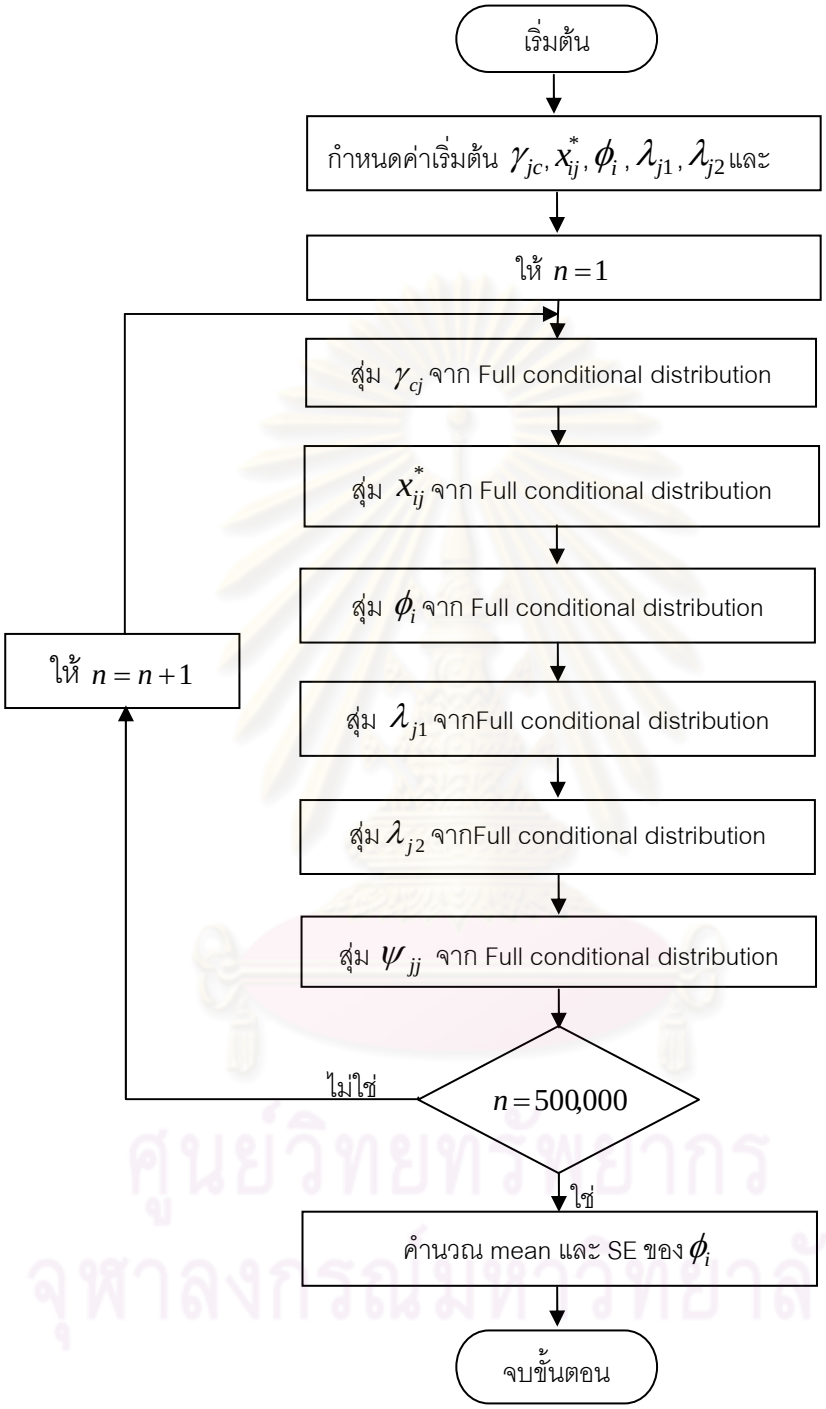
8) สุ่ม ψ_{jj} สำหรับ j ที่เป็นตัวแปรแบบต่อเนื่อง จาก Full conditional distribution ที่มีการแจกแจงแกมมาแบบผกผัน $IG\left(\left(a_{0j}+N\right)/2\ ,\left(b_{0j}+\left(x_j^*-\Phi\lambda_j\right)'\left(x_j^*-\Phi\lambda_j\right)\right)\right)$

9) ให้ $n=n+1$ ดำเนินการซ้ำจนกระทั่ง $n=500{,}000$

10) คำนวณคะแนนเฉลี่ย (mean) และค่าคลาดเคลื่อนมาตรฐาน (SE) ของ ϕ_i ด้วยวิธีค่าเฉลี่ยกลุ่ม โดยให้จำนวน batch เท่ากับ 20

ขั้นตอนดังกล่าวสามารถเขียนอยู่ในรูปที่ 3.1

รูปที่ 3.1 แสดงขั้นตอนการสุ่มตัวอย่างแบบกิบส์



3.2 ขั้นตอนดำเนินการด้วยวิธีการสุ่มตัวอย่างแบบฮิตแอนด์รัน

ในการเปรียบเทียบการสุ่มตัวอย่างแบบกิบส์กับการสุ่มตัวอย่างแบบฮิตแอนด์รันนั้น จะสุ่ม x_{ij}^* ด้วยการสุ่มตัวอย่างแบบฮิตแอนด์รัน เนื่องจาก x_{ij}^* มีการแจกแจงแบบปกติในหลายมิติที่มีขอบเขตจำกัดซึ่งสามารถหาการแจกแจงบนเส้นตรงในทิศทางของฮิตแอนด์รันได้ สำหรับพารามิเตอร์อื่นๆ การหาการแจกแจงบนเส้นตรงในทิศทางของฮิตแอนด์รันนั้นทำได้ยาก จึงยังคงใช้การสุ่มตัวอย่างด้วยวิธีกิบส์

1) ทำการสร้างจุดเริ่มต้นของ γ_{jc} , x_{ij}^* , ϕ_i , λ_{j1} , λ_{j2} และ ψ_{jj}

2) เริ่มต้นให้จำนวนรอบ $n=1$

3) สุ่ม γ_{jc} สำหรับ j ที่เป็นตัวแปรเชิงอันดับจาก Full conditional distribution ที่มีการแจกแจงแบบสม่ำเสมอ
$U\left(\max(\max\{x_{ij}^* : x_{ij}=c\}, \gamma_{j(c-1)}), \min(\min\{x_{ij}^* : x_{ij}=c+1\}, \gamma_{j(c+1)})\right)$

4) สุ่ม x_{ij}^* ด้วยการสุ่มตัวอย่างแบบฮิตแอนด์รัน สำหรับ j ที่เป็นตัวแปรเชิงอันดับที่มี $C_j > 2$ โดยจะสุ่ม x_{ij}^* ในแต่ละกลุ่มที่ $x_{ij}=c, c=1,2,\ldots,C_j$ พร้อมกัน

5) สุ่ม ϕ_i สำหรับทุก $i=1,\ldots,N$ จาก Full conditional distribution ที่มีการแจกแจงแบบปกติ $N\left(\left(\mathrm{I}+\lambda_2'\Psi^{-1}\lambda_2\right)^{-1}\left(\lambda_2'\Psi^{-1}\left(x_j^*-\lambda_1\right)\right), \left(\mathrm{I}+\lambda_2'\Psi^{-1}\lambda_2\right)^{-1}\right)$

6) สุ่ม λ_{j1} สำหรับ j ที่เป็นตัวแปรเชิงอันดับจากFull conditional distribution ที่มีการแจกแจงแบบปกติ $N\left(\left(L_{0j}+\psi_{jj}^{-1}N\right)^{-1}\left(L_{0j}l_{0j}+\psi_{jj}^{-1}(x_j^*-\lambda_2'\phi_i)\right), \left(L_{0j}+\psi_{jj}^{-1}N\right)^{-1}\right)$

7) สุ่ม λ_{j2} สำหรับทุก $j=1,\ldots,J$ โดยกำหนดให้ λ_{12} (λ_2 ของตัวแปร courts) มีเครื่องหมายเป็นลบ จาก Full conditional distribution ที่มีการแจกแจงแบบปกติ $N\left(\left(L_{0j}+\psi_{jj}^{-1}\phi_i'\phi_i\right)^{-1}\left(L_{0j}l_{0j}+\psi_{jj}^{-1}\phi_i'(x_j^*-\lambda_1)\right), \left(L_{0j}+\psi_{jj}^{-1}\phi_i'\phi_i\right)^{-1}\right)$ (การแจกแจงปกติแบบตัดปลาย สำหรับตัวที่ถูกกำหนดเครื่องหมาย)

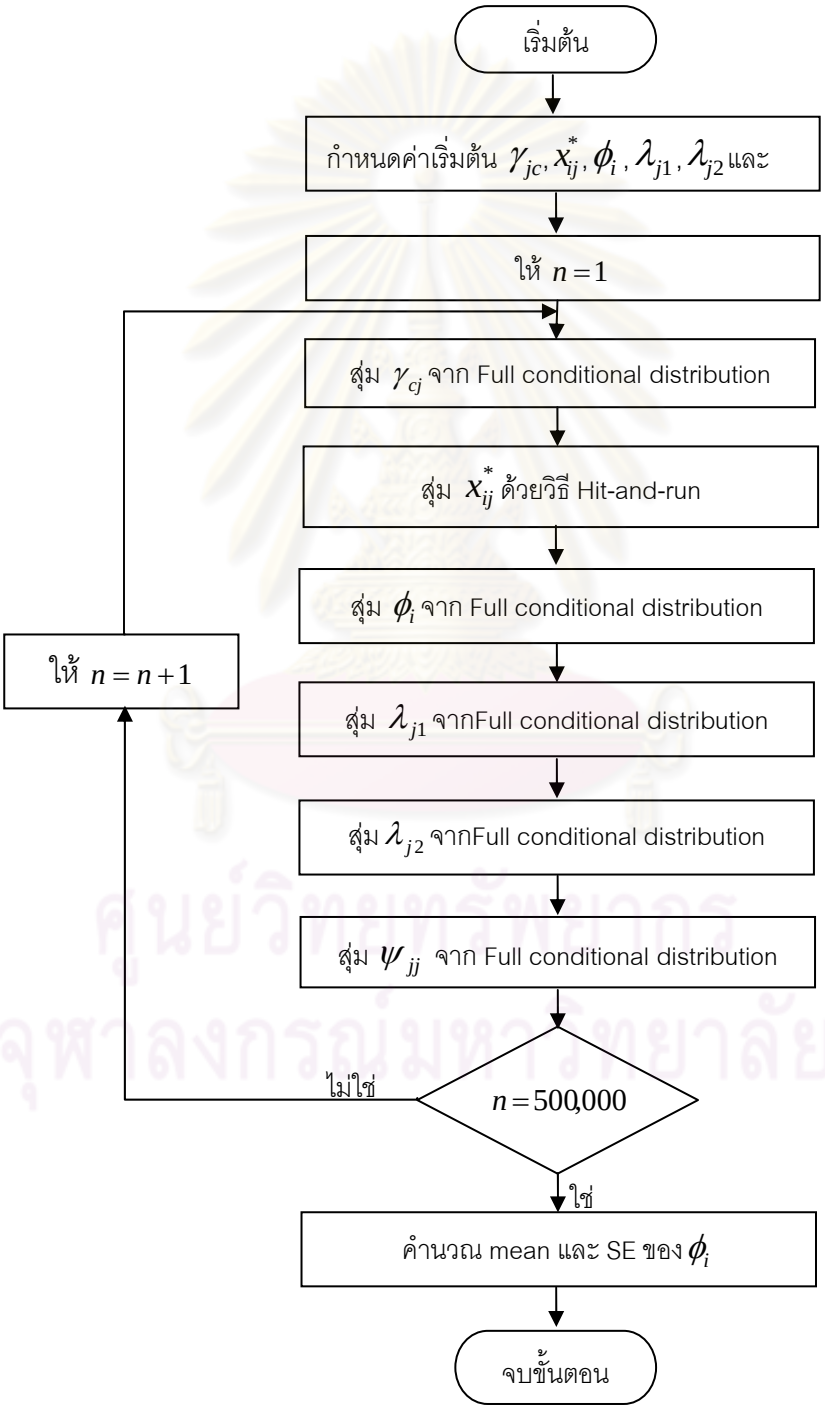
8) สุ่ม ψ_{jj} สำหรับ j ที่เป็นตัวแปรแบบต่อเนื่อง จาก Full conditional distribution ที่มีการแจกแจงแกมมาแบบผกผัน $IG\left(\left(a_{0j}+N\right)/2, \left(b_{0j}+\left(x_j^*-\Phi\lambda_j\right)'\left(x_j^*-\Phi\lambda_j\right)\right)\right)$

9) ให้ $n=n+1$ ดำเนินการซ้ำจนกระทั่ง $n=500{,}000$

10) คำนวณคะแนนเฉลี่ย (mean) และค่าคลาดเคลื่อนมาตรฐาน (SE) ของ ϕ_i ด้วยวิธีค่าเฉลี่ยกลุ่ม โดยให้จำนวน batch เท่ากับ 20

ขั้นตอนดังกล่าวสามารถเขียนอยู่ในรูปที่ 3.2

รูปที่ 3.2 แสดงขั้นตอนการสุ่มตัวอย่างแบบฮิตแอนด์รัน



ในงานวิจัยนี้ใช้โปรแกรม R 2.9.0 ในการประมวลผล ในการจัดอันดับความเสี่ยงทางการเมือง-เศรษฐกิจจะใช้คะแนนเฉลี่ย (Mean) ในการจัดอันดับ การเปรียบเทียบประสิทธิภาพระหว่างการดำเนินการด้วยการสุ่มตัวอย่างแบบฮิตแอนด์รันและการสุ่มตัวอย่างแบบกิบส์ จะใช้ค่าคลาดเคลื่อนมาตรฐานของคะแนนเฉลี่ย (ϕ_i) เป็นเกณฑ์ในการเปรียบเทียบประสิทธิภาพในการลู่เข้าของตัวประมาณ วิธีใดที่มีค่าคลาดเคลื่อนมาตรฐานน้อยกว่า ถือว่าเป็นวิธีที่มีประสิทธิภาพมากกว่า

บทที่ 4

ผลการวิเคราะห์ข้อมูล

ผลจากการทำการทดลองโดยใช้ข้อมูลความเสี่ยงทางการเมือง-เศรษฐกิจจาก MCMCpack R Package โดย Martin และ Quinn 2004 ซึ่งมีดัชนี 5 ตัว ได้แก่ ความเป็นอิสระของระบบตุลาการ (courts), ส่วนต่างอัตราแลกเปลี่ยนเงินตราต่างประเทศในตลาดมืด (barb2), ความเสี่ยงจากการยึดทรัพย์สินในกิจการมาเป็นของรัฐ (prsexp2), การทุจริตในภาครัฐ (prscorr2) และผลิตภัณฑ์มวลรวมภายในประเทศ (gdpw2) เป็นเพื่อจัดอันดับประเทศ 62 ประเทศตามความเสี่ยงทางการเมือง-เศรษฐกิจ มีดังนี้

ในการศึกษาครั้งนี้มีพารามิเตอร์ที่สนใจ คือ ค่าปัจจัย(ϕ, Phi) หรือ *คะแนน* ของทั้ง 62 ประเทศ ซึ่งใช้คะแนนเฉลี่ย (mean) ในการจัดอันดับความเสี่ยงทางการเมือง-เศรษฐกิจ ดังนั้นเปรียบเทียบประสิทธิภาพระหว่างการดำเนินการด้วยการสุ่มตัวอย่างแบบฮิตแอนด์รันและการสุ่มตัวอย่างแบบกิบส์ จะใช้ค่าคลาดเคลื่อนมาตรฐานของคะแนนเฉลี่ย เป็นเกณฑ์ในการเปรียบเทียบ โดยค่าคลาดเคลื่อนมาตรฐาน (SE) ของคะแนนเฉลี่ยของแต่ละประเทศจากการสุ่มตัวอย่างทั้ง 2 วิธีแสดงในตารางที่ 4.1

ตารางที่ 4.1 แสดงส่วนเบี่ยงเบนมาตรฐาน (SE) ของคะแนนเฉลี่ยจากการสุ่มตัวอย่างแบบฮิตแอนด์รันและการสุ่มตัวอย่างแบบกิบส์

Country	$SE_{Hit\text{-}and\text{-}run}$	SE_{Gibbs}	Country	$SE_{Hit\text{-}and\text{-}run}$	SE_{Gibbs}
Argentina	0.0270*	0.0228	Indonesia	0.0221	0.0173*
Australia	0.0084	0.0057*	Iran	0.0330	0.0235*
Austria	0.0081	0.0046*	Ireland	0.0086	0.0066*
Bangladesh	0.0063	0.0059*	Israel	0.0084*	0.0125
Belgium	0.0072	0.0045*	Italy	0.0031*	0.0032
Bolivia	0.0123	0.0059*	Japan	0.0090	0.0064*
Botswana	0.0021	0.0015*	Kenya	0.0070*	0.0083
Brazil	0.0025	0.0013*	Korea, South	0.0197	0.0149*
Burma	0.0145	0.0066*	Malawi	0.0083*	0.0119
Cameroon	0.0119	0.0086*	Malaysia	0.0028	0.0017*
Canada	0.0052	0.0043*	Mexico	0.0095	0.0084*
Chile	0.0028	0.0025*	Morocco	0.0112	0.0080*
Colombia	0.0035	0.0024*	New Zealand	0.0060*	0.0047

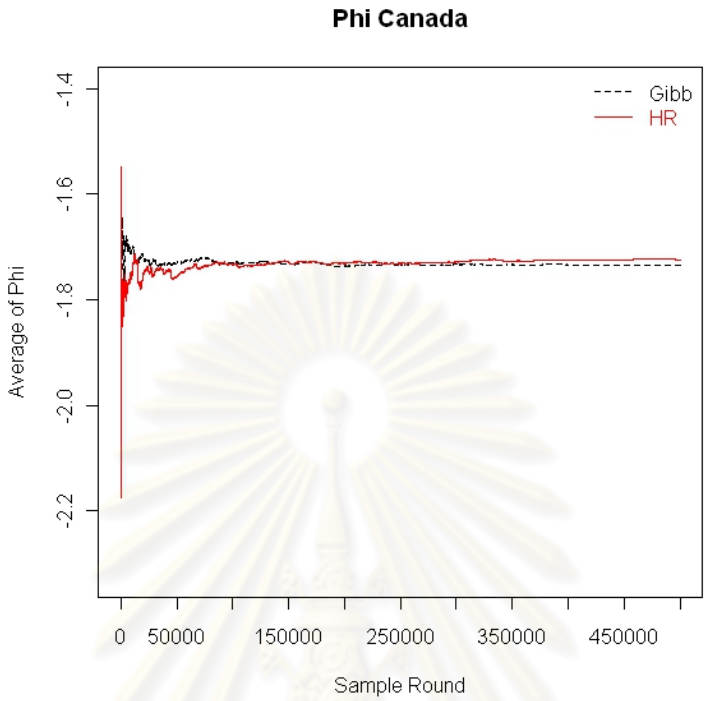
ตารางที่ 4.1 (ต่อ) แสดงส่วนเบี่ยงเบนมาตรฐาน (SE) ของคะแนนเฉลี่ยจากการสุ่มตัวอย่างแบบฮิตแอนด์รันและการสุ่มตัวอย่างแบบกิบส์

Country	$SE_{Hit\text{-}and\text{-}run}$	SE_{Gibbs}	Country	$SE_{Hit\text{-}and\text{-}run}$	SE_{Gibbs}
Congo-Kinshasa	0.0079	0.0071*	Nigeria	0.0062*	0.0063
Costa Rica	0.0164	0.0129*	Norway	0.0050	0.0032*
Cote d'Ivoire	0.0091	0.0084*	Papua New Guinea	0.0087	0.0084*
Denmark	0.0068	0.0052*	Paraguay	0.0229	0.0163*
Dominican Republic	0.0084	0.0078*	Philippines	0.0071	0.0049*
Ecuador	0.0031*	0.0035	Poland	0.0079	0.0065*
Finland	0.0064	0.0058*	Portugal	0.0025	0.0017*
Gambia, The	0.0126	0.0058*	Sierra Leone	0.0131	0.0067*
Ghana	0.0049	0.0018*	Singapore	0.0062*	0.0066
Greece	0.0085*	0.0111	South Africa	0.0140*	0.0178
Hungary	0.0025	0.0016*	Spain	0.0147	0.0123*
India	0.0118	0.0051*	Sri Lanka	0.0075*	0.0082
Sweden	0.0187	0.0154*	Turkey	0.0036*	0.0037
Switzerland	0.0068	0.0050*	United Kingdom	0.0083	0.0043*
Syria	0.0083	0.0057*	Uruguay	0.0090	0.0080*
Thailand	0.0034	0.0021*	Venezuela	0.0032	0.0024*
Togo	0.0226	0.0156*	Zambia	0.0132	0.0074*
Tunisia	0.0088	0.0084*	Zimbabwe	0.0055	0.0037*

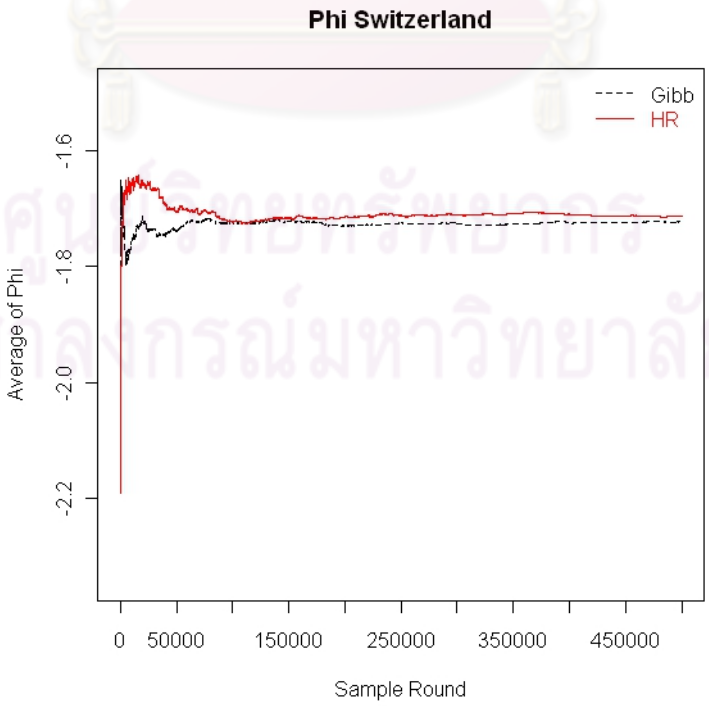
* แสดงว่าส่วนเบี่ยงเบนมาตรฐานจากการสุ่มด้วยวิธีดังกล่าวมีค่าน้อยกว่าการสุ่มตัวอย่างอีกวิธีหนึ่ง

ในการประมาณค่าคะแนนของประเทศทั้งหมด 62 ประเทศ มีอยู่ 51 ประเทศ ที่มี $SE_{Hit\text{-}and\text{-}run} > SE_{Gibbs}$ ซึ่งแสดงว่าการสุ่มตัวอย่างด้วยวิธีกิบส์มีประสิทธิภาพมากกว่าวิธีฮิตแอนด์รัน สามารถสังเกตการลู่เข้าของค่าประมาณได้จากกราฟคะแนนเฉลี่ยสะสม ในรูปที่ 4.1 ถึง 4.6

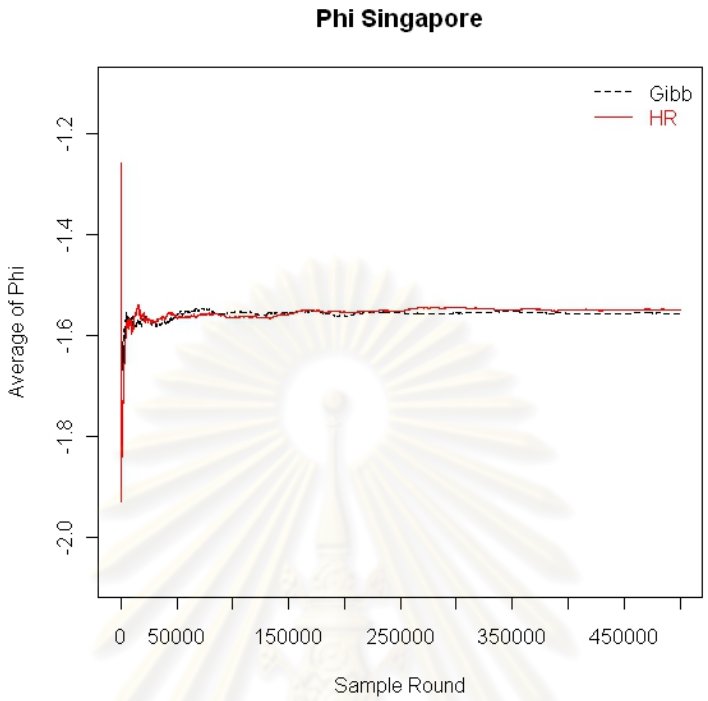
รูปที่ 4.1 คะแนนเฉลี่ยสะสมของประเทศแคนาดา



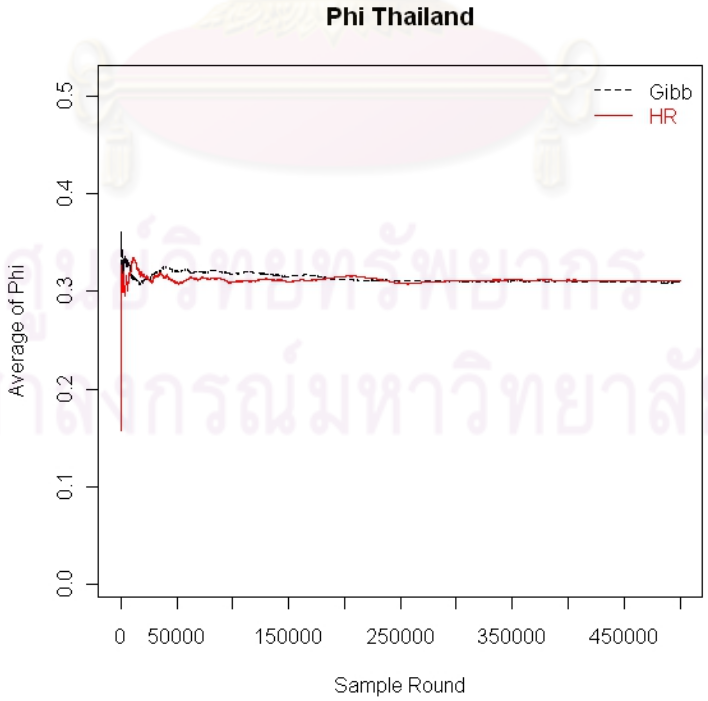
รูปที่ 4.2 คะแนนเฉลี่ยสะสมของประเทศสวิสเซอร์แลนด์



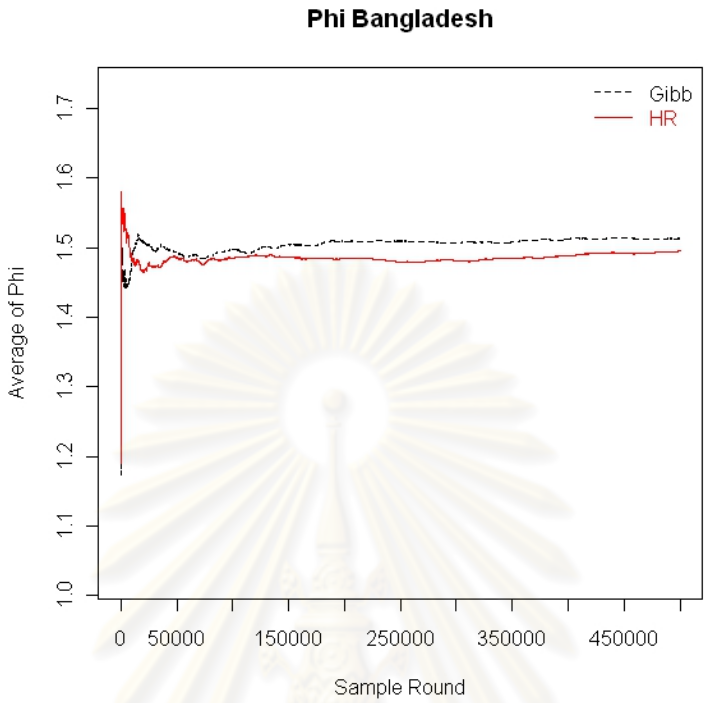
รูปที่ 4.3 คะแนนเฉลี่ยสะสมของประเทศสิงคโปร์



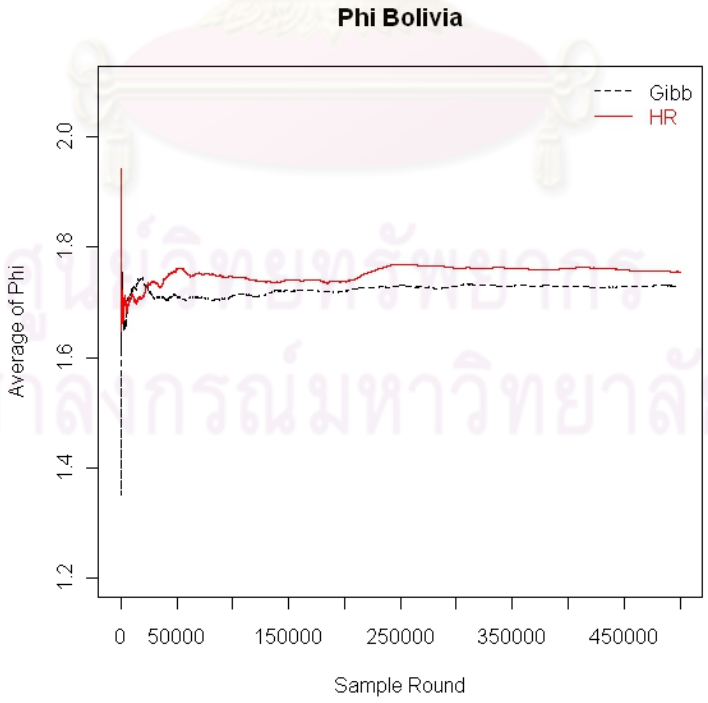
รูปที่ 4.4 คะแนนเฉลี่ยสะสมของประเทศไทย



รูปที่ 4.5 คะแนนเฉลี่ยสะสมของประเทศบังคลาเทศ



รูปที่ 4.6 คะแนนเฉลี่ยสะสมของประเทศโบลิเวีย



จากรูปที่ 4.1 ถึง 4.6 เส้นประสีดำแสดงคะแนนเฉลี่ยสะสมจากการสุ่มด้วยวิธีกิบส์ ส่วนเส้นทึบสีแดงแสดงคะแนนเฉลี่ยสะสมจากการสุ่มด้วยวิธีฮิตแอนด์รัน สังเกตได้ว่าคะแนนเฉลี่ยสะสมจากวิธีกิบส์มีการกระจายน้อยและลู่เข้าเร็วกว่าวิธีฮิตแอนด์รัน แสดงว่าในการประยุกต์ใช้กับการจัดอันดับด้วยการวิเคราะห์ปัจจัยเชิงเบส์นั้น การสุ่มตัวอย่างด้วยวิธีกิบส์มีประสิทธิภาพในการลู่เข้าของคะแนนเฉลี่ยมากกว่าวิธีฮิตแอนด์รัน

สำหรับพารามิเตอร์อื่นที่ไม่ได้กล่าวถึงในบทที่นี้ ได้แสดงผลไว้ในภาคผนวก

บทที่ 5
สรุปผลการวิจัย และข้อเสนอแนะ

การศึกษาวิจัยในครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของการสุ่มตัวอย่างด้วยวิธีกิบส์และการสุ่มตัวอย่างด้วยวิธีฮิตแอนด์รันสำหรับทำการวิเคราะห์ปัจจัยเชิงเบส์เมื่อมีปัจจัย 1 ปัจจัย โดยมีวัตถุประสงค์เพื่อการจัดอันดับ ใช้โปรแกรม R ทำการจำลองโดยใช้การสุ่มตัวอย่างด้วยวิธีกิบและการสุ่มตัวอย่างด้วยวิธีฮิตแอนด์รัน แล้วใช้วิธีค่าเฉลี่ยกลุ่ม (Batch mean method) คำนวณค่าคลาดเคลื่อนมาตรฐาน (SE) ที่ใช้เป็นเกณฑ์ในการเปรียบเทียบประสิทธิภาพการลู่เข้าของคะแนนเฉลี่ย

สรุปผลการวิจัย

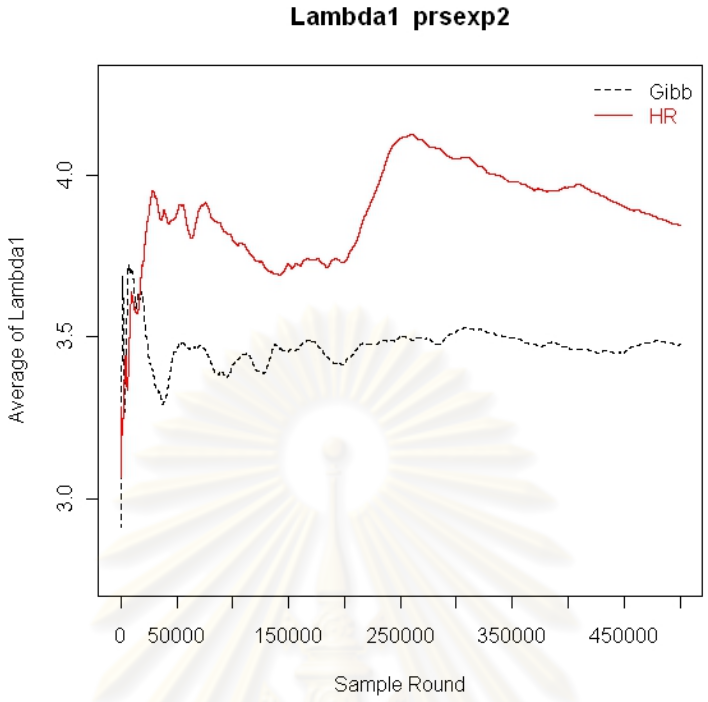
จากการทดสอบประสิทธิภาพด้วยข้อมูลความเสี่ยงทางการเมือง-เศรษฐกิจจาก MCMCPack 0.4-8 เมื่อเปรียบเทียบประสิทธิภาพของ MCMC ทั้ง 2 วิธี ด้วยค่าคลาดเคลื่อนมาตรฐาน (SE) ของคะแนนเฉลี่ยทั้ง 62 ตัว พบว่ามีค่าประมาณ 51 ตัว ที่ $SE_{\text{Hit-and-run}} > SE_{\text{Gibbs}}$ และอีก 11 ตัว ที่ $SE_{\text{Hit-and-run}} < SE_{\text{Gibbs}}$ ดังนั้นหากใช้เกณฑ์การตัดสินด้วยค่าคลาดเคลื่อนมาตรฐาน (SE) จึงสรุปว่า วิธีฮิตแอนด์รันมีประสิทธิภาพด้อยกว่าวิธีกิบส์

อภิปรายผลการวิจัย

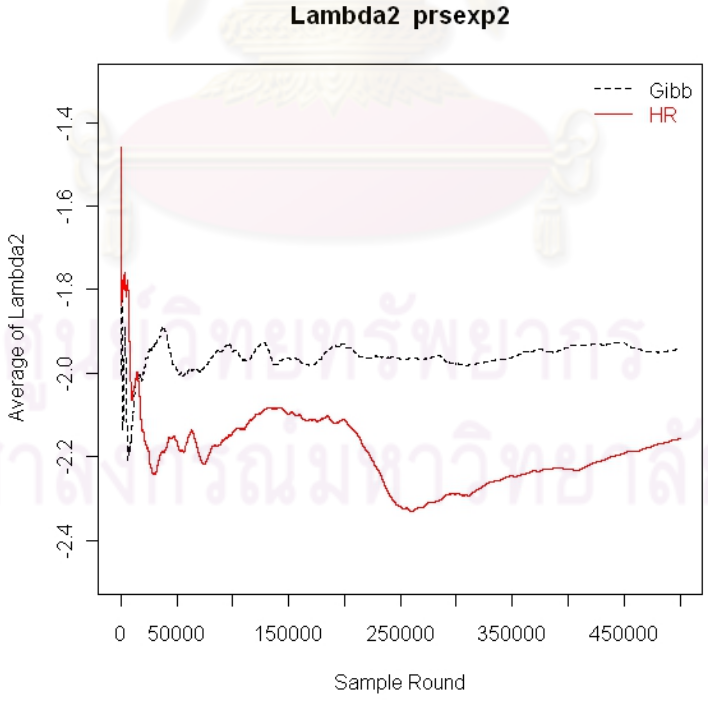
ผลการทดลองที่ได้แตกต่างจากงานวิจัยในอดีต เช่น งานวิจัยของ Chen และ Schmeiser [4] ที่พบว่าวิธีฮิตแอนด์รันเป็นวิธีที่มีประสิทธิภาพมากกว่าวิธีกิบส์สำหรับในกรณีที่มิติเท่ากับ 2 และไม่มีขอบเขตของพารามิเตอร์ ผลที่แตกต่างอาจเป็นเพราะปัญหาการวิเคราะห์ปัจจัยเพื่อการจัดอันดับมีจำนวนมิติมากกว่า 2 และพารามิเตอร์มีขอบเขตที่ซับซ้อน จึงทำให้วิธีฮิตแอนด์รันด้อยประสิทธิภาพลง

ถึงแม้ว่าพารามิเตอร์ที่เราสนใจ คือ ϕ จะลู่เข้าแล้ว แต่สาเหตุที่วิธีฮิตแอนด์รันมีประสิทธิภาพด้วยกว่า อาจเกิดจากเงื่อนไขของตัวแปรเชิงอันดับที่ต้องอาศัยจุดตัด γ ในการแบ่งกลุ่มตัวแปรแฝง x^* ซึ่งในแต่ละรอบของวิธีฮิตแอนด์รัน x^* ซึ่งถูกจำกัดขอบเขตด้วยจุดตัด γ เคลื่อนที่ในทิศทางที่สุ่มได้อย่างจำกัด จึงส่งผลให้พารามิเตอร์ตัวอื่น ๆที่ขึ้นอยู่กับ x^* ทั้ง γ, λ_1 และ λ_2 ลู่เข้าช้า สังเกตได้จากกราฟค่าเฉลี่ยสะสมของ γ, λ_1 และ λ_2 ดังรูปที่ 5.1 ถึง 5.3

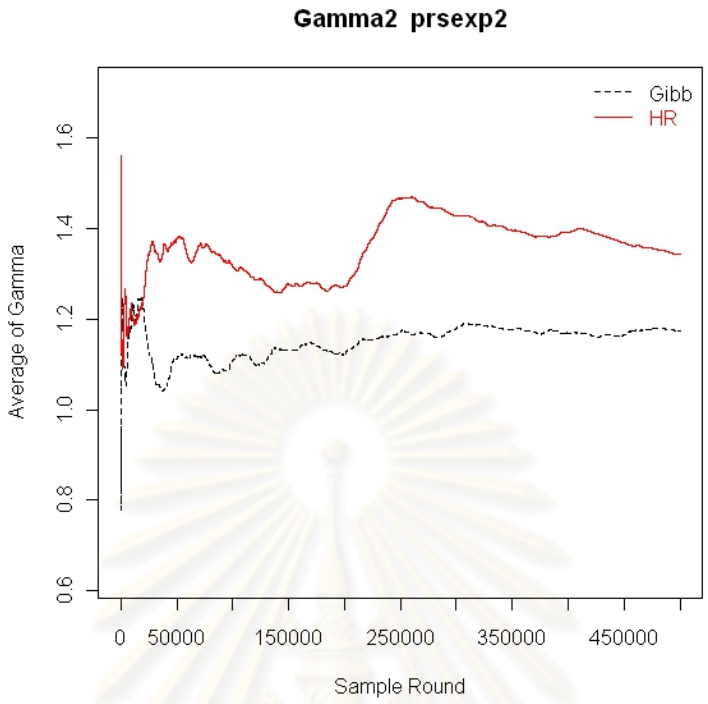
รูปที่ 5.1 แสดงค่าเฉลี่ยสะสมของ Lambda1 prsexp2



รูปที่ 5.2 แสดงค่าเฉลี่ยสะสมของ Lambda2 prsexp2



รูปที่ 5.2 แสดงค่าเฉลี่ยสะสมของ Gamma2 prsexp2



จากผลการศึกษาข้างต้นทำให้เห็นว่า ในกรณีการประมาณค่าพารามิเตอร์เชิงเบส์ที่มีพารามิเตอร์จำนวนมาก และมีขอบเขตของพารามิเตอร์ที่ซับซ้อน การสุ่มตัวอย่างด้วยวิธีฮิตแอนด์รันมีประสิทธิภาพด้อยกว่าการสุ่มตัวอย่างด้วยวิธีกิบส์

ข้อเสนอแนะ

1. เนื่องจากในงานวิจัยครั้งนี้ ใช้การสุ่มตัวอย่างแบบฮิตแอนด์รันสำหรับตัวแปรแฝง X^* เท่านั้น ดังนั้นในงานวิจัยต่อๆไปอาจปรับปรุงให้ใช้การสุ่มตัวอย่างแบบฮิตแอนด์รันสำหรับพารามิเตอร์ทั้งหมด เพื่อให้การเปรียบเทียบประสิทธิภาพชัดเจนมากยิ่งขึ้น

2. ในการสุ่มตัวอย่างแบบฮิตแอนด์รัน อาจใช้การสุ่มเลือกทิศทางแบบอื่นๆที่สอดคล้องกับของเขตของพารามิเตอร์ที่ค้องการประมาณค่า แทนการใช้การสุ่มเลือกทิศทางบนผิวทรงกลมรัศมี 1 หน่วย เพื่อเพิ่มประสิทธิภาพของการสุ่มตัวอย่างแบบฮิตแอนด์รัน

รายการอ้างอิง

[1] Quinn, Kevin M.. 2004. Bayesian Factor Analysis for Mixed Ordinal and Continuous Response. Political Analysis 12(4) : 338-353.

[2] Cowles, M. K.. 1996. Accelerating Monte Carlo Markov Chain Convergence for Cumulative Link Generalized Linear Models. Statistics and Computing 6 : 101-111.

[3] Lovasz and Vempala. 2006. Fast Algorithm for Log-concave Functions: Sampling, Rounding, Integration and Optimization. Proceedings of the 47th IEEE Symposium on Foundations of Computer Science (FOCS'06) : 57-68.

[4] Chen M.-H and Schmeiser B.W.. 1993. Performance of the Gibbs, Hit-and-Run, and Metropolis Samplers. Journal of Computational and Graphical Statistics 12, 3 : 251-272.

[5] Geman S. and Geman D.. 1984. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6 : 721-741.

[6] Smith, R. L. 1984. Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. Operations Research 6 : 1296-1308.

[7] Bélisle, C.J.P., Romeijn, H. E., and Smith, R. L.. 1993. Hit-and-Run algorithm for generating multivariate distributions. Mathematics of Operations Research 18 : 255-266.

[8] Romeijn H.E. and Smith R.L. 1994. Simulated annealing for constrained global optimization. Journal of Global Optimization 5 : 101-126.

[9] Flegal J.M., Haran M. and Jones G.L. 2008. Markov Chain Monte Carlo: Can We Trust the Third Significant Figure?. Statistical Science 23, 2 : 250-260.

ภาคผนวก

ภาคผนวก ก

โปรแกรมสำหรับงานวิจัย

โปรแกรม R สำหรับดำเนินงานตามขั้นตอนในการวิจัย

```
seednum<-101;
signvec<-c(-1,0,0,0,0);
roundnum<-500000;
numbatch<-20;
recname<-pedata[,1];
varname<-names(peframe);

#ดำเนินการด้วยการสุ่มตัวอย่างแบบกิบส์
set.seed(seednum);
Gibb2time<-system.time(Gibb2<-fmcmcG(peframe,signvec,roundnum,100));
outGibb2<-batchMeans(Gibb2$Record,numbatch);
Gibb2mean<-meanplot(Gibb2$Record,10);

#ดำเนินการด้วยการสุ่มตัวอย่างแบบฮิตแอนด์รัน
set.seed(seednum);
GHRtime<-system.time(GHR<-fmcmcGHR(peframe,signvec,roundnum,100));
outGHR<-batchMeans(GHR$Record,numbatch);
GHRmean<-meanplot(GHR$Record,10);
```

โปรแกรมสำหรับฟังก์ชัน fmcmcG

```
fmcmcG<-function(dframe,signvec,roundnum,batch){
levelvec<-sapply(sapply(dframe, levels),length);
varcode<-ifelse(levelvec==0,1,ifelse(levelvec==2,3,2));
numrec<-nrow(dframe);
numvar<-ncol(dframe);
recname<-row.names(dframe);
varname<-names(dframe);
burnin<-10000;

#Standardize continuous data
for(j in 1:numvar){
ifelse(varcode[j]==1,dframe[,j]<-(dframe[,j]-
        mean(dframe[,j]))/sd(dframe[,j]),dframe[,j]<-dframe[,j]);
}

#Prior Parameters
lparam<-matrix(0,numvar,2);
Lparam<-matrix(0.25,numvar,2);
aparam<-rep(0.001/2, length=numvar);
bparam<-rep(0.001/2, length=numvar);
t2param<-rep(0.01, length=numvar);

#Initialize gamma ordmax
gamma<-c();
for(j in 1:numvar){
if(varcode[j]!=2){
subgamma<-list(0);
}
else{
```

```
subgamma<-list(c(0,qnorm((seq(2:(levelvec[j]-1))/(2*levelvec[j])+0.5))));
}
gamma<-c(gamma,subgamma);
}

#Initialize xstar
xstar<-matrix(0,numrec,numvar);
for(j in 1:numvar){
if(varcode[j]==1) xstar[,j]<-dframe[,j];
if(varcode[j]==3) xstar[,j]<-as.numeric(dframe[,j])-1.5;
if(varcode[j]==2){
for(i in 1:numrec){
if(dframe[i,j]==1) xstar[i,j]<-0-0.5;
if((dframe[i,j]!=1)&(dframe[i,j]!=levelvec[j]))xstar[i,j]<-
        (0.5*(gamma[[j]][dframe[i,j]]+gamma[[j]][(as.numeric(dframe[i,j])-1)]));
if(dframe[i,j]==levelvec[j]) xstar[i,j]<-(gamma[[j]][(levelvec[j]-1)]+0.5);
}
}
}

#Initialize z
z<-rnorm(numrec);

#Initialize psi
psi<-rep(0,length=numvar);
psi<-ifelse(varcode==1,aparam*bparam,1);

#Initialize lambda constrain lambda2(courts) to be "-"
lambda<-matrix(0,numvar,2);
```

```
lambda[,1]<-ifelse(varcode==1,0,rnorm(numvar,lparam[,1],1 lambdatemp<-
        qnorm(runif(numvar),0,1);
lambda[,2]<-ifelse(signvec==-1,-
        abs(lambdatemp),ifelse(signvec==1,abs(lambdatemp),lambdatemp));

#Initialize output
lambda.col<-c();
gamma.col<-c();
## Get output column names
for(k in 1:ncol(lambda)){
lambda.col<-c(lambda.col,paste(varname,".l",k,sep=""));
}
for(j in 1:numvar){
if(levelvec[j]>2){
gamma.col<-c(gamma.col,paste(varname[j],".g",1:(levelvec[j]-1),sep=""));
}
}
outmat.col<-c(recname,lambda.col,gamma.col);
outmat<-
        matrix(0,nrow=roundnum,ncol=length(outmat.col),dimnames=list(c(1:roundnum),o
        utmat.col));

###Gibb Sampler
for(n in 1:(roundnum+burnin)){
outvec<-c();
gammavec<-c();
lambdavec<-c();

#Sample gamma
for(j in  1:numvar){
```

```
if(varcode[j]==2){
for(k in 2:(levelvec[j]-1)){
a<-max(max(xstar[,j][dframe[,j]==k]),gamma[[j]][k-1]);
if(k==(levelvec[j]-1)){
b<-min(xstar[,j][dframe[,j]==(k+1)])
}
else{
b<-min(min(xstar[,j][dframe[,j]==(k+1)]),gamma[[j]][k+1]);
}
gamma[[j]][k]<-runif(1,a,b);
}
gammavec<-c(gammavec,gamma[[j]]);
}#End if varcode[j]==2
}#End Sample Gamma[[j]]

#sample xstar for ordinal and binary variables
for(j in 1:numvar){
if(varcode[j]==2){
for(k in 1:levelvec[j]){
x<-xstar[dframe[,j]==k,j];
kk<-floor(runif(1,1,length(x)+1));
if(k==levelvec[j]){
rightnorm<-1;
}
else{
rightnorm<-pnorm(gamma[[j]][k],lambda[j,1]+lambda[j,2]*z[dframe[,j]==k][kk],1);
}
if(k==1){
leftnorm<-0;
}
```

```
else {
leftnorm<-pnorm(gamma[[j]][k-1],lambda[j,1]+lambda[j,2]*z[dframe[,j]==k][kk],1);
}
xstar[dframe[,j]==k,j][kk]<-qnorm(runif(1)*(rightnorm-
        leftnorm)+leftnorm,lambda[j,1]+lambda[j,2]*z[dframe[,j]==k][kk],1);
}
}
else{
if(varcode[j]==3){
for(i in 1:numrec){
A<-pnorm(0,lambda[j,1]+lambda[j,2]*z[i],1,lower.tail=TRUE);
if(dframe[i,j]==0){
xstar[i,j]<-qnorm(runif(1)*A,lambda[j,1]+lambda[j,2]*z[i],1);
}
else{
xstar[i,j]<-qnorm(runif(1)*(1-A)+A,lambda[j,1]+lambda[j,2]*z[i],1);
}
}
}#End sample xstar for binary variables
}#End if varcode[j]==3
}

#sample z
sigma2<-1/(1+t(lambda[,2])%*%diag(1/psi)%*%lambda[,2]);
sigma<-sqrt(sigma2);
mu<-sigma2%*%(t(lambda[,2])%*%diag(1/psi)%*%(t(xstar)-lambda[,1]));
z<-rnorm(numrec,mu,sigma*rep(1,length=numrec));

#sample lambda
for(j in 1:numvar){
```

```
#Sample lambda1
if(varcode[j]==1){
lambda[j,1]<-0;
}
else{
sigmaa2<-1/(Lparam[j,1]+numrec/psi[j]);
sigmaa<-sqrt(sigmaa2);
mua<-(Lparam[j,1]*lparam[j,1]+sum(xstar[,j]-z*lambda[j,2])/psi[j])*sigmaa2;
lambda[j,1]<-rnorm(1,mua,sigmaa);
}
#Sample lambda2
sigmab2<-1/(Lparam[j,2]+(z%*%z)/psi[j]);
sigmab<-sqrt(sigmab2);
mub<-(Lparam[j,2]*lparam[j,2]+(z%*%(xstar[,j]-lambda[j,1]))/psi[j])*sigmab2;
ifelse(j==1&templambda2>=0,lambda[j,2],templambda2)
if(signvec[j]==-1){
lambda[j,2]<-qnorm(pnorm(0,mub,sigmab,lower.tail=TRUE)*runif(1),mub,sigmab)
}
else{
if(signvec[j]==1){
lambda[j,2]<-qnorm(pnorm(0,mub,sigmab,lower.tail=FALSE)*runif(1)+
        pnorm(0,mub,sigmab,lower.tail=TRUE),mub,sigmab)
}
else{
lambda[j,2]<-qnorm(runif(1),mub,sigmab)
}
}
}#End sample lambda
lambdavec<-c(lambda[,1],lambda[,2]);
```

```
#Sample psi
for(j in 1:numvar){
shape<-(2*aparam[j]+numrec)/2;
tempvec<-xstar[,j]-lambda[j,1]-lambda[j,2]*z;
scale<-2*bparam[j]+tempvec%*%tempvec;
psi[j]<-ifelse(varcode[j]==1,1/rgamma(1,shape,scale),1);
}

#Colect output
outvec<-c(z,lambdavec,gammavec);
if(n>burnin){
outmat[(n-burnin),]<-outvec;
}
}#End Gibb

outmeanmat<-matrix(0,nrow=(roundnum/batch),ncol=length(outmat.col),
        dimnames=list(c(1:(roundnum/batch)),outmat.col));
for(n in 1:(roundnum/batch)){
outmeanmat[n,]<-outmat[(n*batch),];
}
outmean<-colMeans(outmeanmat);
Means<-outmean[(numrec+1):(numrec+numvar)];
Loadings<-outmean[(numrec+numvar+1):(numrec+(2*numvar))];
Scores<-outmean[1:numrec];
record<-outmat;
lambdaout<-as.data.frame(rbind(Means,Loadings),row.names=c("Means","Loadings"))
names(lambdaout)<-varname;
output<-list(Record=record,Lambda=lambdaout,Scores=Scores);
output
}#End function
```

โปรแกรมสำหรับฟังก์ชัน fmcmcGHR

```
fmcmcGHR<-function(dframe,signvec,roundnum,batch){
levelvec<-sapply(sapply(dframe, levels),length);
varcode<-ifelse(levelvec==0,1,ifelse(levelvec==2,3,2));
numrec<-nrow(dframe);
numvar<-ncol(dframe);
recname<-row.names(dframe);
varname<-names(dframe);
burnin<-10000;

#Standardize continuous data
for(j in 1:numvar){
ifelse(varcode[j]==1,dframe[,j]<-(dframe[,j]-
        mean(dframe[,j]))/sd(dframe[,j]),dframe[,j]<-dframe[,j]);
}

#Prior Parameters
lparam<-matrix(0,numvar,2);
Lparam<-matrix(0.25,numvar,2);
aparam<-rep(0.001/2, length=numvar);
bparam<-rep(0.001/2, length=numvar);
t2param<-rep(0.01, length=numvar);

#Initialize gamma ordmax
gamma<-c();
for(j in 1:numvar){
if(varcode[j]!=2){
subgamma<-list(0);
}
else{
```

```
subgamma<-list(c(0,qnorm((seq(2:(levelvec[j]-1))/(2*levelvec[j])+0.5))));
}
gamma<-c(gamma,subgamma);
}

#Initialize xstar
xstar<-matrix(0,numrec,numvar);
for(j in 1:numvar){
if(varcode[j]==1) xstar[,j]<-dframe[,j];
if(varcode[j]==3) xstar[,j]<-as.numeric(dframe[,j])-1.5;
if(varcode[j]==2){
for(i in 1:numrec){
if(dframe[i,j]==1) xstar[i,j]<-0-0.5;
if((dframe[i,j]!=1)&(dframe[i,j]!=levelvec[j]))xstar[i,j]<-
        (0.5*(gamma[[j]][dframe[i,j]]+gamma[[j]][(as.numeric(dframe[i,j])-1)]));
if(dframe[i,j]==levelvec[j]) xstar[i,j]<-(gamma[[j]][(levelvec[j]-1)]+0.5);
}
}
}

#Initialize z
z<-rnorm(numrec);

#Initialize psi
psi<-rep(0,length=numvar);
psi<-ifelse(varcode==1,aparam*bparam,1);

#Initialize lambda constrain lambda2(courts) to be "-"
lambda<-matrix(0,numvar,2);
```

```
lambda[,1]<-ifelse(varcode==1,0,rnorm(numvar,lparam[,1],1 lambdatemp<-
        qnorm(runif(numvar),0,1);
lambda[,2]<-ifelse(signvec==-1,-
        abs(lambdatemp),ifelse(signvec==1,abs(lambdatemp),lambdatemp));

#Initialize output
lambda.col<-c();
gamma.col<-c();
## Get output column names
for(k in 1:ncol(lambda)){
lambda.col<-c(lambda.col,paste(varname,".l",k,sep=""));
}
for(j in 1:numvar){
if(levelvec[j]>2){
gamma.col<-c(gamma.col,paste(varname[j],".g",1:(levelvec[j]-1),sep=""));
}
}
outmat.col<-c(recname,lambda.col,gamma.col);
outmat<-
        matrix(0,nrow=roundnum,ncol=length(outmat.col),dimnames=list(c(1:roundnum),o
        utmat.col));

###Gibb Sampler
for(n in 1:(roundnum+burnin)){
outvec<-c();
gammavec<-c();
lambdavec<-c();

#Sample gamma
for(j in  1:numvar){
```

```
if(varcode[j]==2){
for(k in 2:(levelvec[j]-1)){
a<-max(max(xstar[,j][dframe[,j]==k]),gamma[[j]][k-1]);
if(k==(levelvec[j]-1)){
b<-min(xstar[,j][dframe[,j]==(k+1)])
}
else{
b<-min(min(xstar[,j][dframe[,j]==(k+1)]),gamma[[j]][k+1]);
}
gamma[[j]][k]<-runif(1,a,b);
}
gammavec<-c(gammavec,gamma[[j]]);
}#End if varcode[j]==2
}#End Sample Gamma[[j]]

#sample xstar for ordinal and binary variables
for(j in 1:numvar){
if(varcode[j]==2){
for(k in 1:levelvec[j]){
x<-xstar[dframe[,j]==k,j];
d<-rnorm(length(x),0,1);
d<-d/sqrt(sum(d^2));
if(k == 1){
gleft<-c();
}
else{
gleft<-rep(gamma[[j]][k-1],length=length(x));
}
if(k==levelvec[j]){
gright<-c();
```

```
}
else{
gright<-rep(gamma[[j]][k],length=length(x));
}
rationvec<-c((gleft-x)/d,(gright-x)/d);
rmin<-max(rationvec[rationvec<0],-Inf);
rmax<-min(rationvec[rationvec>0],Inf);
zx<-z[dframe[,j]==k];
mux<-c(lambda[j,1]+lambda[j,2]*zx);
sigmar2<-1/(d%*%d);
sigmar<-sqrt(sigmar2);
mur<-((mux-x)%*%d)*sigmar2;
A<-pnorm(rmin,mur,sigmar);
B<-pnorm(rmax,mur,sigmar);
r<-qnorm(((B-A)*runif(1))+A,mur,sigmar);
xnew<-x+r*d;
xstar[dframe[,j]==k,j]<-xnew;
}
}
else{
if(varcode[j]==3){
for(i in 1:numrec){
A<-pnorm(0,lambda[j,1]+lambda[j,2]*z[i],1,lower.tail=TRUE);
if(dframe[i,j]==0){
xstar[i,j]<-qnorm(runif(1)*A,lambda[j,1]+lambda[j,2]*z[i],1);
}
else{
xstar[i,j]<-qnorm(runif(1)*(1-A)+A,lambda[j,1]+lambda[j,2]*z[i],1);
}
}
```

```
}#End sample xstar for binary variables
}#End if varcode[j]==3
                    }
#sample z
sigma2<-1/(1+t(lambda[,2])%*%diag(1/psi)%*%lambda[,2]);
sigma<-sqrt(sigma2);
mu<-sigma2%*%(t(lambda[,2])%*%diag(1/psi)%*%(t(xstar)-lambda[,1]));
z<-rnorm(numrec,mu,sigma*rep(1,length=numrec));

#sample lambda
for(j in 1:numvar){
#Sample lambda1
if(varcode[j]==1){
lambda[j,1]<-0;
}
else{
sigmaa2<-1/(Lparam[j,1]+numrec/psi[j]);
sigmaa<-sqrt(sigmaa2);
mua<-(Lparam[j,1]*lparam[j,1]+sum(xstar[,j]-z*lambda[j,2])/psi[j])*sigmaa2;
lambda[j,1]<-rnorm(1,mua,sigmaa);
}
#Sample lambda2
sigmab2<-1/(Lparam[j,2]+(z%*%z)/psi[j]);
sigmab<-sqrt(sigmab2);
mub<-(Lparam[j,2]*lparam[j,2]+(z%*%(xstar[,j]-lambda[j,1]))/psi[j])*sigmab2;
ifelse(j==1&templambda2>=0,lambda[j,2],templambda2)
if(signvec[j]==-1){
lambda[j,2]<-qnorm(pnorm(0,mub,sigmab,lower.tail=TRUE)*runif(1),mub,sigmab)
}
else{
```

```
if(signvec[j]==1){
lambda[j,2]<-qnorm(pnorm(0,mub,sigmab,lower.tail=FALSE)*runif(1)+
        pnorm(0,mub,sigmab,lower.tail=TRUE),mub,sigmab)
}
else{
lambda[j,2]<-qnorm(runif(1),mub,sigmab)
}
}
}#End sample lambda
lambdavec<-c(lambda[,1],lambda[,2]);

#Sample psi
for(j in 1:numvar){
shape<-(2*aparam[j]+numrec)/2;
tempvec<-xstar[,j]-lambda[j,1]-lambda[j,2]*z;
scale<-2*bparam[j]+tempvec%*%tempvec;
psi[j]<-ifelse(varcode[j]==1,1/rgamma(1,shape,scale),1);
}

#Colect output
outvec<-c(z,lambdavec,gammavec);
if(n>burnin){
outmat[(n-burnin),]<-outvec;
}
}#End Gibb

outmeanmat<-matrix(0,nrow=(roundnum/batch),ncol=length(outmat.col),
        dimnames=list(c(1:(roundnum/batch)),outmat.col));
for(n in 1:(roundnum/batch)){
outmeanmat[n,]<-outmat[(n*batch),];
```

```
}
outmean<-colMeans(outmeanmat);
Means<-outmean[(numrec+1):(numrec+numvar)];
Loadings<-outmean[(numrec+numvar+1):(numrec+(2*numvar))];
Scores<-outmean[1:numrec];
record<-outmat;
lambdaout<-as.data.frame(rbind(Means,Loadings),row.names=c("Means","Loadings"))
names(lambdaout)<-varname;
output<-list(Record=record,Lambda=lambdaout,Scores=Scores);
output
}#End function
```

โปรแกรมสำหรับฟังก์ชัน batchMeans

```
batchMeans<-function(x,numbatch){
x<-as.data.frame(x)
roundnum<-dim(x)[1];
batchmat<-{};
size<-roundnum/numbatch;
batchmat<-matrix(0,nrow=numbatch,ncol=length(names(x)),
        dimnames=list(c(1:numbatch),names(x)));

for(i in 1:numbatch){
batch<-colMeans(x[(((i-1)*size)+1):((i*size)),])
batchmat[i,]<-batch;
}
batchmean<-colMeans(batchmat);
batchsd<-apply(batchmat,2,sd);
batchSE<-batchsd/sqrt(numbatch);
data.frame(mean=batchmean,SE=batchSE,sd=batchsd);
}#End function
```

โปรแกรมสำหรับฟังก์ชัน meanplot

```
meanplot<-function(data,range){
numround<-dim(data)[1];
numrec<-dim(data)[2];
meanmat<-matrix(0,nrow=(numround/range),ncol=numrec,
        dimnames=list(c(1:(numround/range)),names(data)));
for(i in 1:(numround/range)){
meanmat[i,]<-colMeans(data[1:(i*range),]);
}
meanmat;
}#End function
```

ภาคผนวก ข

ตารางที่ 1 แสดงคะแนนเฉลี่ย (mean) และอันดับ (Rank) ของแต่ละประเทศจากการสุ่มตัวอย่างแบบฮิตแอนด์รันและการสุ่มตัวอย่างแบบกิบส์

Country	Hit-and-run		Gibbs	
	mean	Rank	mean	Rank
Argentina	0.5157	41	0.4271	39
Australia	-1.1885	9	-1.1647	10
Austria	-1.1077	13	-1.0864	14
Bangladesh	1.4951	60	1.5134	61
Belgium	-1.1100	12	-1.0881	13
Bolivia	1.7553	62	1.7296	62
Botswana	-0.3854	22	-0.3809	22
Brazil	-0.3992	21	-0.3946	21
Burma	0.8431	52	0.8821	52
Cameroon	0.6357	48	0.6746	49
Canada	-1.7241	1	-1.7345	1
Chile	-0.0147	27	-0.0058	28
Colombia	0.3809	37	0.3797	37
Congo-Kinshasa	1.4952	61	1.5132	60
Costa Rica	-0.4431	19	-0.5004	17
Cote d'Ivoire	-0.2236	24	-0.1878	24
Denmark	-1.6779	7	-1.6840	7
Dominican Republic	0.6063	45	0.5729	45
Ecuador	-0.0103	29	-0.0049	29
Finland	-1.6821	5	-1.6903	6
Gambia, The	0.2392	33	0.2717	34
Ghana	0.9643	55	0.9691	53
Greece	0.1407	31	0.1052	31
Hungary	-0.4101	20	-0.4041	20
India	0.2049	32	0.2338	32
Indonesia	0.9540	53	1.0282	54
Iran	1.1417	58	1.0316	55
Ireland	-1.1576	10	-1.1345	11
Israel	-0.2405	23	-0.2768	23

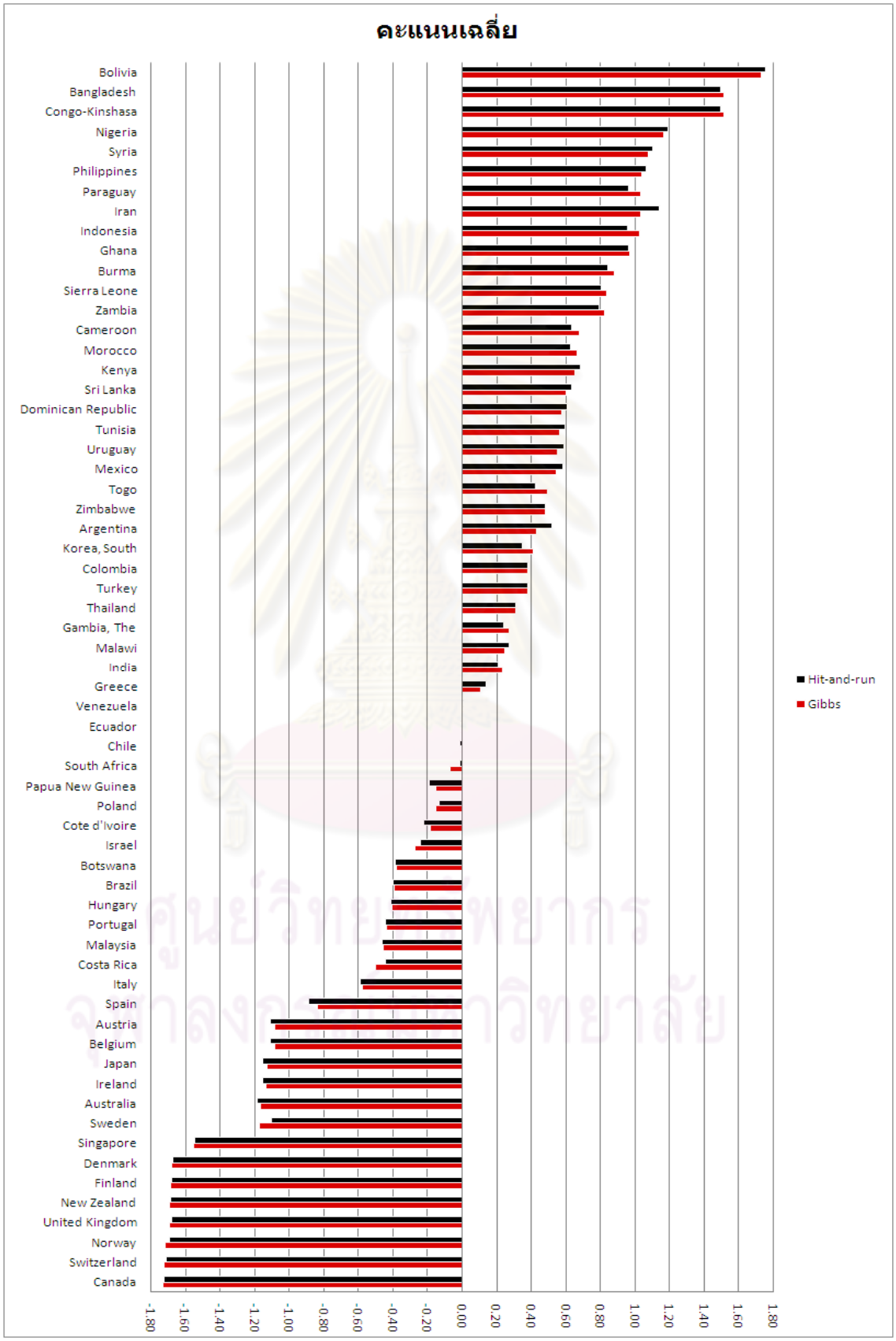
ตารางที่ 1 (ต่อ) แสดงคะแนนเฉลี่ย (mean) และอันดับ (Rank) ของแต่ละประเทศจากการสุ่มตัวอย่างแบบฮิตแอนด์รันและการสุ่มตัวอย่างแบบกิบส์

Country	Hit-and-run		Gibbs	
	mean	Rank	mean	Rank
Italy	-0.5892	16	-0.5768	16
Japan	-1.1573	11	-1.1291	12
Kenya	0.6810	49	0.6495	47
Korea, South	0.3485	36	0.4135	38
Malawi	0.2741	34	0.2435	33
Malaysia	-0.4647	17	-0.4552	18
Mexico	0.5802	42	0.5471	42
Morocco	0.6277	46	0.6614	48
New Zealand	-1.6861	4	-1.6932	5
Nigeria	1.1889	59	1.1651	59
Norway	-1.6925	3	-1.7188	3
Papua New Guinea	-0.1883	25	-0.1509	26
Paraguay	0.9623	54	1.0345	56
Philippines	1.0618	56	1.0375	57
Poland	-0.1330	26	-0.1559	25
Portugal	-0.4470	18	-0.4393	19
Sierra Leone	0.8013	51	0.8330	51
Singapore	-1.5509	8	-1.5574	8
South Africa	-0.0126	28	-0.0733	27
Spain	-0.8881	15	-0.8369	15
Sri Lanka	0.6342	47	0.6029	46
Sweden	-1.1074	14	-1.1719	9
Switzerland	-1.7130	2	-1.7225	2
Syria	1.1005	57	1.0746	58
Thailand	0.3108	35	0.3092	35
Togo	0.4237	39	0.4916	41
Tunisia	0.5926	44	0.5634	44
Turkey	0.3820	38	0.3792	36
United Kingdom	-1.6817	6	-1.6935	4

ตารางที่ 1 (ต่อ) แสดงคะแนนเฉลี่ย (mean) และอันดับ (Rank) ของแต่ละประเทศจากการสุ่มตัวอย่างแบบฮิตแอนด์รันและการสุ่มตัวอย่างแบบกิบส์

Country	Hit-and-run		Gibbs	
	mean	Rank	mean	Rank
Uruguay	0.5860	43	0.5526	43
Venezuela	-0.0041	30	0.0023	30
Zambia	0.7922	50	0.8232	50
Zimbabwe	0.4833	40	0.4801	40

รูปที่ 1 แสดงคะแนนเฉลี่ยเรียงตามความเสี่ยงทางการเมือง-เศรษฐกิจจากน้อยไปหามาก

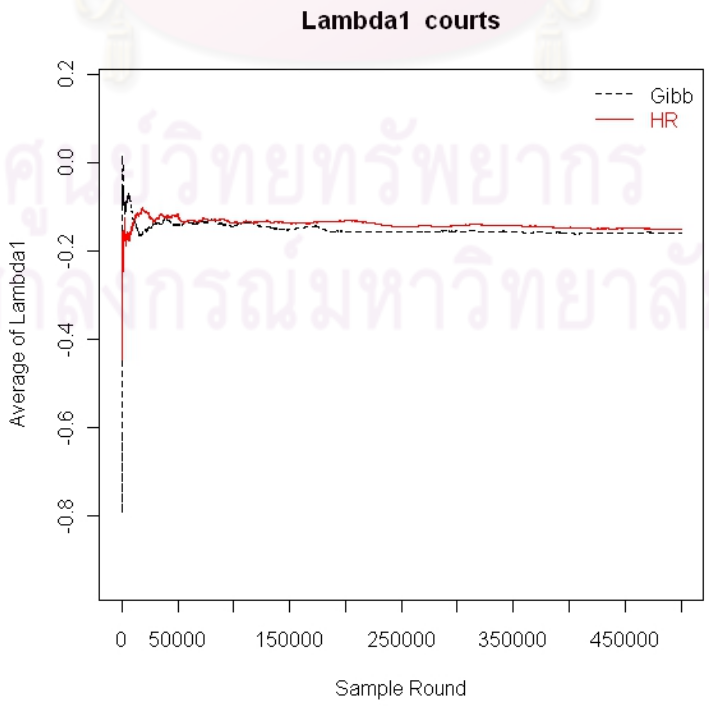


ตารางที่ 2 แสดงค่าประมาณ (mean) และส่วนเบี่ยงเบนมาตรฐาน (SE) ของสัมประสิทธิ์ปัจจัย (factor loading)

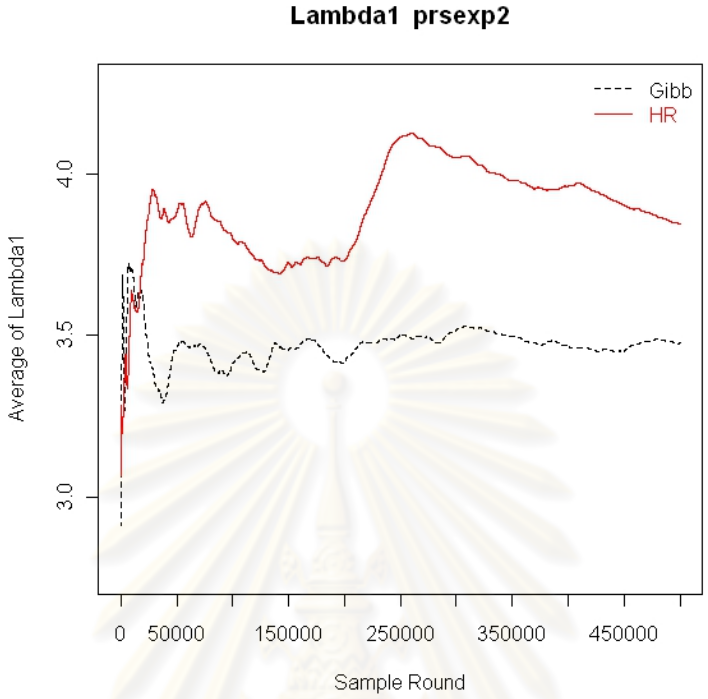
	Hit-and-run		Gibbs	
	mean	SE	mean	SE
Lambda1 courts	-0.1514	0.0067	-0.1596	0.0059*
Lambda1 barb2	0	0	0	0
Lambda1 prsexp2	3.8446	0.1467	3.4764	0.0616*
Lambda1 prscorr2	3.0004	0.0674*	3.2646	0.0866
Lambda1 gdpw2	0	0	0	0
Lambda2 courts	-3.1814	0.0245*	-3.0705	0.0324
Lambda2 barb2	0.7037	0.0009*	0.6997	0.0015
Lambda2 prsexp2	-2.1563	0.0838	-1.9433	0.0374*
Lambda2 prscorr2	-2.1559	0.0534*	-2.3948	0.0718
Lambda2 gdpw2	-0.6682	0.0036	-0.6774	0.0019*

* แสดงว่าส่วนเบี่ยงเบนมาตรฐานจากการสุ่มด้วยวิธีดังกล่าวมีค่าน้อยกว่าการสุ่มตัวอย่างอีกวิธีหนึ่ง

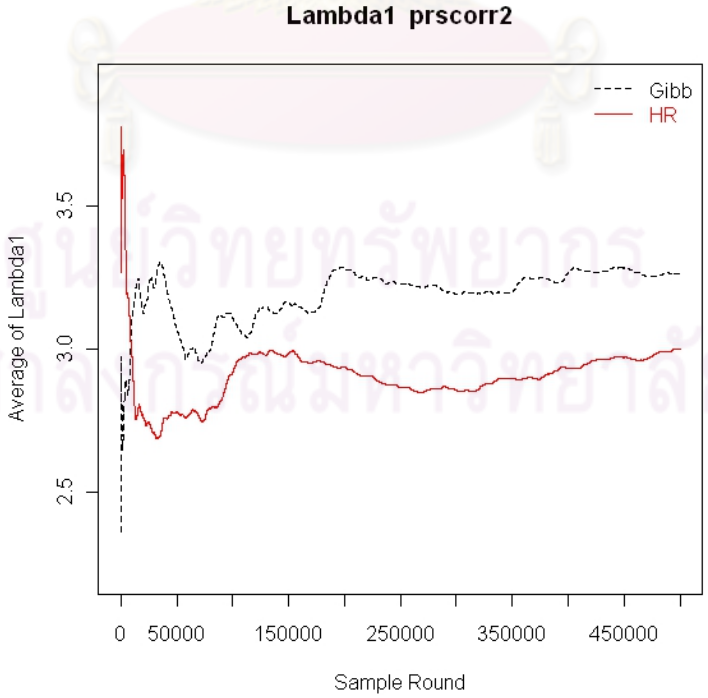
รูปที่ 2 แสดงค่าเฉลี่ยสะสมของ Lambda1 courts



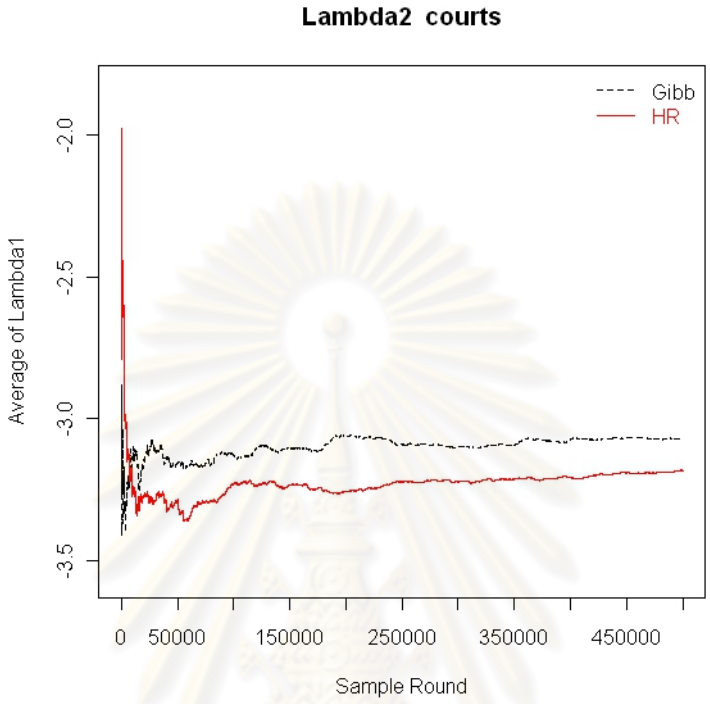
รูปที่ 3 แสดงค่าเฉลี่ยสะสมของ Lambda1 prsexp2



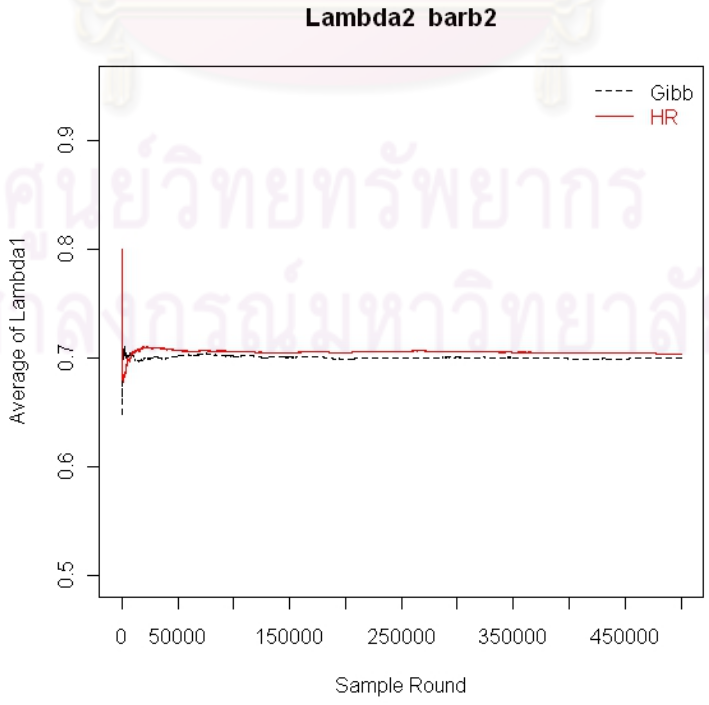
รูปที่ 4 แสดงค่าเฉลี่ยสะสมของ Lambda1 prscorr2



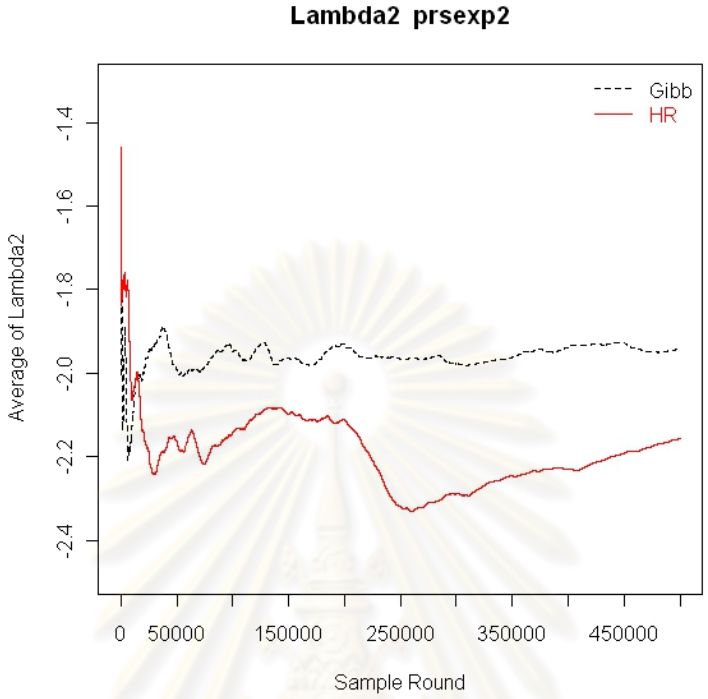
รูปที่ 5 แสดงค่าเฉลี่ยสะสมของ Lambda2 courts



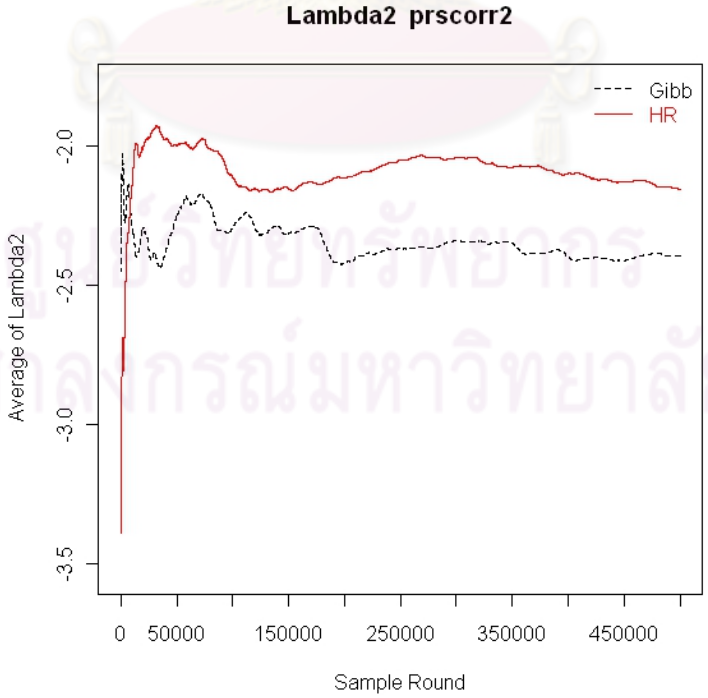
รูปที่ 6 แสดงค่าเฉลี่ยสะสมของ Lambda2 barb2



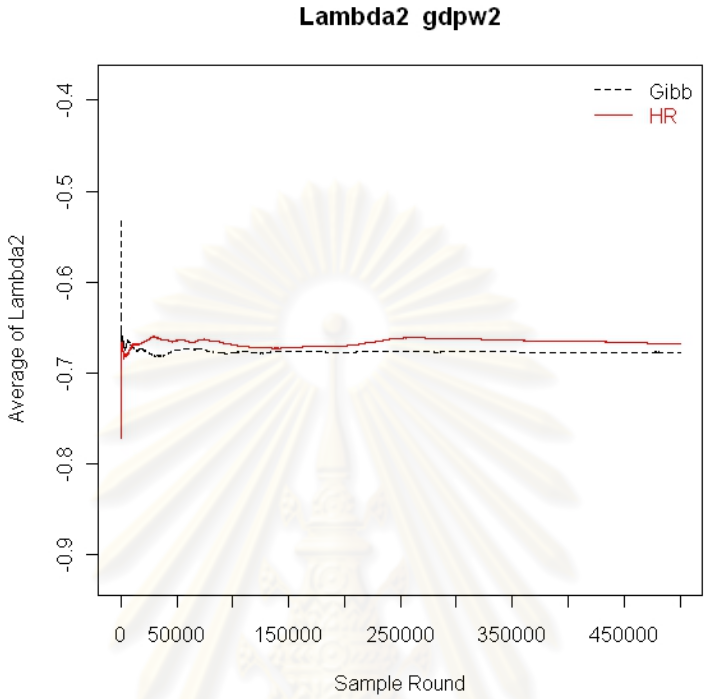
รูปที่ 7 แสดงค่าเฉลี่ยสะสมของ Lambda2 prsexp2



รูปที่ 8 แสดงค่าเฉลี่ยสะสมของ Lambda2 prscorr2



รูปที่ 9 แสดงค่าเฉลี่ยสะสมของ Lambda2 gdpw2

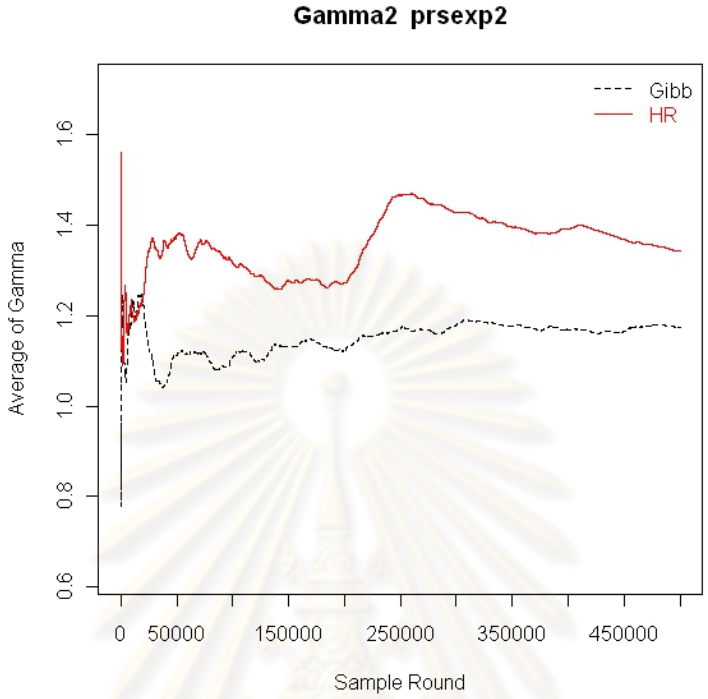


ตารางที่ 3 แสดงค่าประมาณ (mean) และส่วนเบี่ยงเบนมาตรฐาน (SE) ของ Gamma

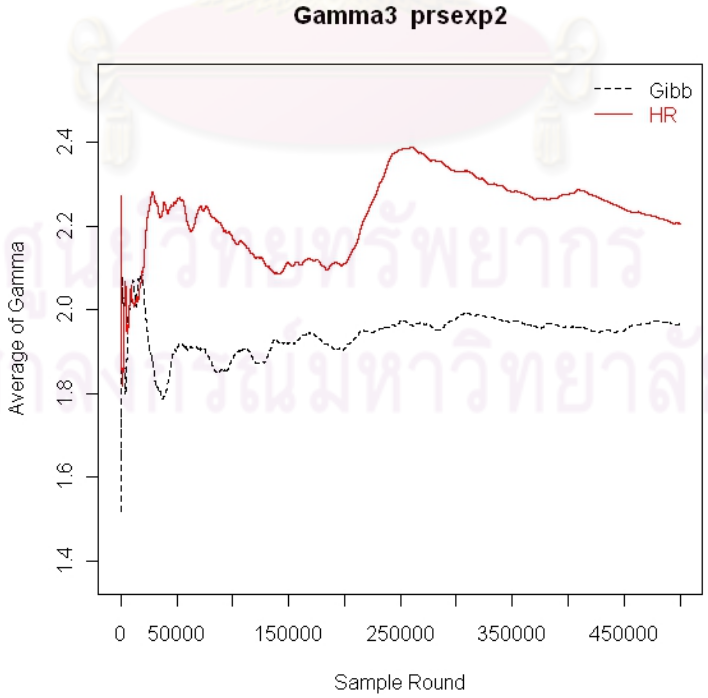
	Hit-and-run		Gibbs	
	mean	SE	mean	SE
Gamma1 prsexp2	0	0	0	0
Gamma2 prsexp2	1.3440	0.0733	1.1747	0.0285*
Gamma3 prsexp2	2.2064	0.1031	1.9662	0.0406*
Gamma4 prsexp2	3.9556	0.1540	3.5670	0.0641*
Gamma5 prsexp2	5.7865	0.2150	5.2483	0.0895*
Gamma1 prscorr2	0	0	0	0
Gamma2 prscorr2	1.4300	0.0353*	1.5483	0.0447
Gamma3 prscorr2	3.1409	0.0725*	3.4157	0.0962
Gamma4 prscorr2	4.4364	0.0965*	4.8289	0.1306
Gamma5 prscorr2	5.8301	0.1264*	6.3776	0.1742

* แสดงว่าส่วนเบี่ยงเบนมาตรฐานจากการสุ่มด้วยวิธีดังกล่าวมีค่าน้อยกว่าการสุ่มตัวอย่างอีกวิธีหนึ่ง

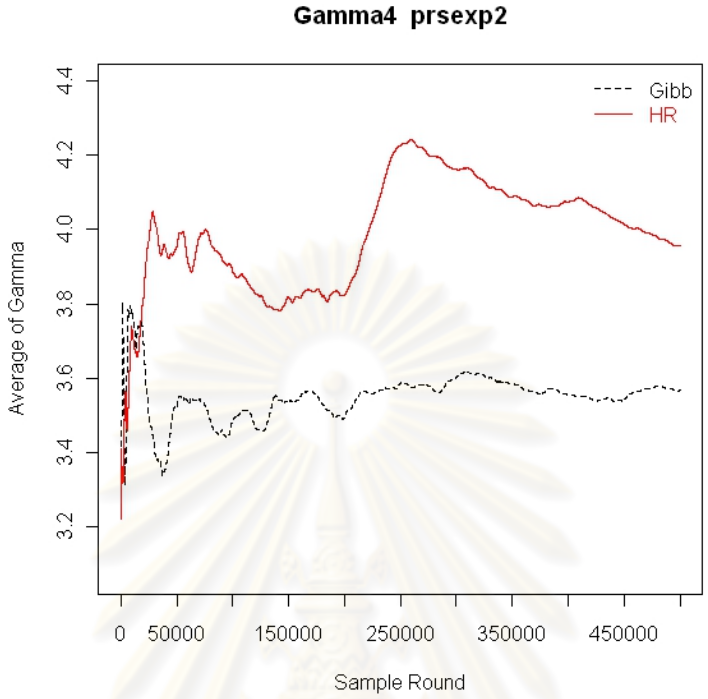
รูปที่ 10 แสดงค่าเฉลี่ยสะสมของ Gamma2 prsexp2



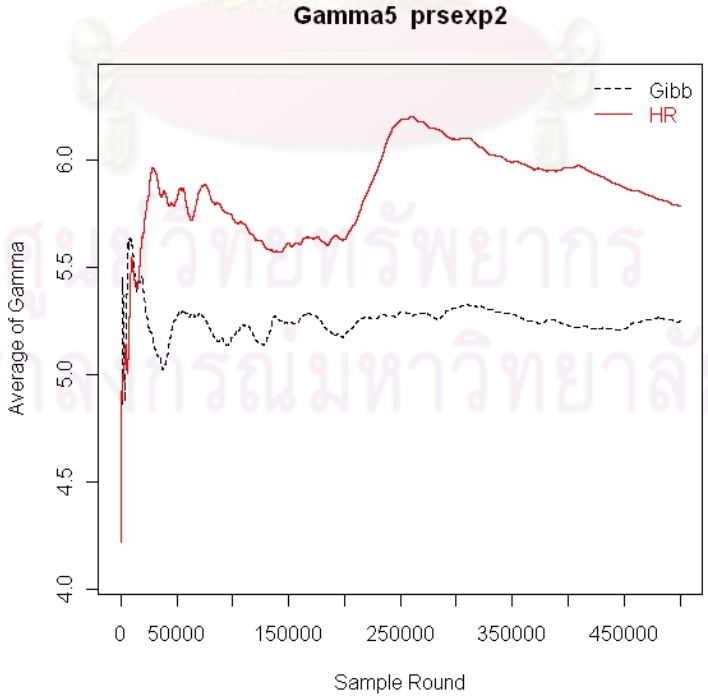
รูปที่ 11 แสดงค่าเฉลี่ยสะสมของ Gamma3 prsexp2



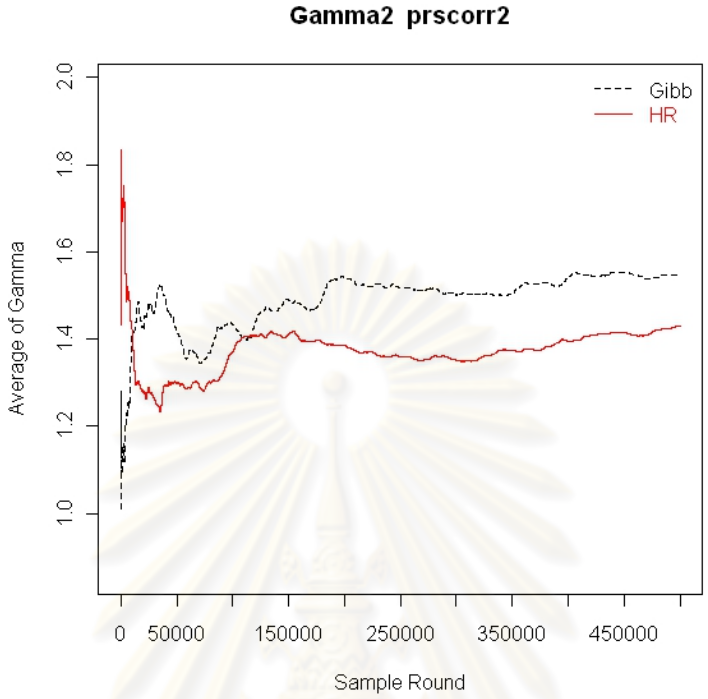
รูปที่ 12 แสดงค่าเฉลี่ยสะสมของ Gamma4 prsexp2



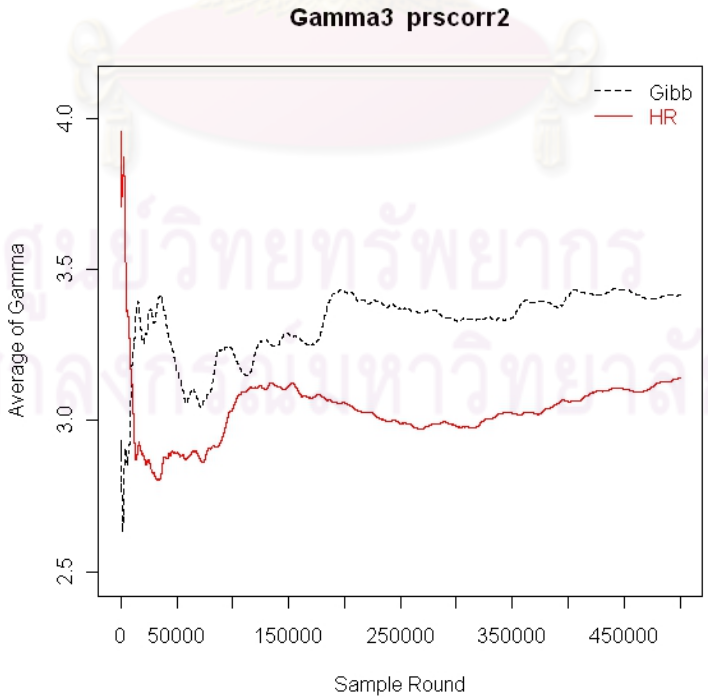
รูปที่ 13 แสดงค่าเฉลี่ยสะสมของ Gamma5 prsexp2



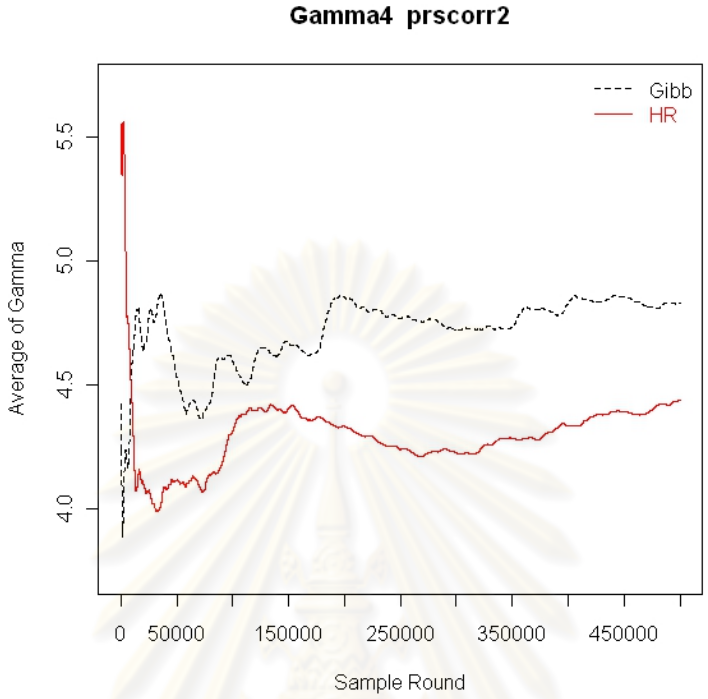
รูปที่ 14 แสดงค่าเฉลี่ยสะสมของ Gamma2 prscorr2



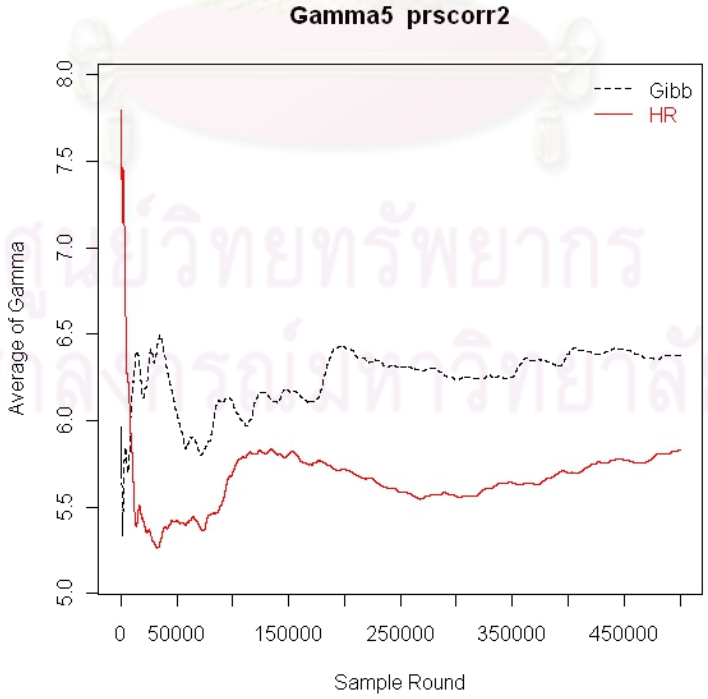
รูปที่ 15 แสดงค่าเฉลี่ยสะสมของ Gamma3 prscorr2



รูปที่ 16 แสดงค่าเฉลี่ยสะสมของ Gamma4 prscorr2



รูปที่ 17 แสดงค่าเฉลี่ยสะสมของ Gamma5 prscorr2



ประวัติผู้เขียนวิทยานิพนธ์

นางสาวไรยา พลเสน เกิดเมื่อวันพุธที่ 16 พฤษภาคม พ.ศ. 2527 สำเร็จการศึกษาระดับปริญญาบัณฑิต หลักสูตรสถิติศาสตรบัณฑิต (สต.บ.) ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2548 และเข้าศึกษาต่อในหลักสูตรสถิติศาสตรมหาบัณฑิต (สต.ม.) สาขาสถิติ ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย ในปี พ.ศ. 2551