ใบรับรองวิทยานิพนธ์

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

วิทยาศาสตรมหาบัณฑิต (สถิติ)

ปริญญา

สถิติ | สถิติ

สาขา | **ภาควิชา**

เรื่อง การประมาณค่าข้อมูลสูญหายในการวัดซ้ำด้วยวิธีมาร์คอฟเชนมอนติคาร์โลและวิธีคอปูลาส์

The Estimation of Missing Data in Repeated Measurements Using Markov Chain Monte Carlo and Copulas Methods

นามผู้วิจัย นางสาวดวงภรณ์ โปทาวี

ได้พิจารณาเห็นชอบโดย

อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก .....

(ผู้ช่วยศาสตราจารย์ลี่ลี อิงศรีสว่าง, Ph.D.)

อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม .....

(รองศาสตราจารย์จิรัชย์ สุขะเกตุ, วท.ม.)

หัวหน้าภาควิชา .....

(อาจารย์อำไพ ทองธีรภาพ, Ph.D.)

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์รับรองแล้ว

.....

(รองศาสตราจารย์กัญจนา ธีระกุล, D.Agr.)

คณบดีบัณฑิตวิทยาลัย

วันที่ ..... **เดือน** ..... **พ.ศ.** .....

วิทยานิพนธ์

เรื่อง

การประมาณค่าข้อมูลสูญหายในข้อมูลวัดซ้ำด้วยวิธีมาร์คอฟเชนมอนติคาร์โลและวิธีคอปูลาส์

The Estimation of Missing Data in Repeated Measurements Using Markov Chain Monte Carlo and Copulas Methods

โดย

นางสาวดวงภรณ์ โปทาวี

เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์
เพื่อความสมบูรณ์แห่งปริญญาวิทยาศาสตรมหาบัณฑิต (สถิติ)
พ.ศ. 2552

ดวงภรณ์ โปทาวี 2552: การประมาณค่าข้อมูลสูญหายในการวัดซ้ำด้วยวิธีมาร์คอฟเชน มอนติคาร์โลและวิธีคอปูลาส์ ปริญญาวิทยาศาสตรมหาบัณฑิต (สถิติ) สาขาสถิติ ภาควิชาสถิติ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก: ผู้ช่วยศาสตราจารย์ลี่ลี อิงศรีสว่าง, Ph.D. 99 หน้า

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาวิธีการประมาณค่าสูญหายในข้อมูลที่มีการวัดซ้ำด้วยวิธี Markov Chain Monte Carlo (MCMC) และวิธี Copulas โดยเปรียบเทียบกับวิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย ใช้ข้อมูลที่ได้จากการจำลองสถานการณ์ที่มีการวัดซ้ำ 3 ครั้ง ด้วยเทคนิคมอนติคาร์โล ภายใต้เงื่อนไขที่กำหนดคือ 1) ข้อมูลมีการแจกแจงแบบปกติหลายตัวแปร มีโครงสร้างเมตริกซ์ความสัมพันธ์ 2 แบบ คือ เมตริกซ์ความสัมพันธ์แบบ Compound Symmetry (CS) หรือ แบบ Autoregressive (AR) 2) กำหนดระดับความสัมพันธ์ระหว่างค่าวัดซ้ำมี 3 ระดับ คือ ระดับต่ำ($\rho = 0.3$) ระดับปานกลาง ($\rho = 0.5$) และ ระดับสูง ($\rho = 0.7$) 3) ขนาดตัวอย่างที่ศึกษา 3 ขนาดคือ 30, 70 และ 100 และ 4) กำหนดให้ตำแหน่งการวัดซ้ำครั้งสุดท้ายมีข้อมูลสูญหายเกิดขึ้นแบบสุ่ม โดยมีระดับการสูญหายของข้อมูล 5% , 10% , 20% และ 30% ตามลำดับ ทำให้มีสถานการณ์ที่เป็นไปได้ทั้งหมด 72 สถานการณ์ สำหรับการประมาณค่าสูญหายด้วยวิธี MCMC กำหนดจำนวนการแทนค่าข้อมูลสูญหายแต่ละค่า 5 ครั้ง ทำการจำลองแต่ละสถานการณ์ 1,000 ครั้ง ด้วยโปรแกรม SAS และวัดประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 3 วิธี ด้วยค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) วิธีการประมาณค่าสูญหายที่ให้ค่า MSE ต่ำกว่า ถือว่าเป็นวิธีที่ดีกว่า

ผลการจำลองทุกสถานการณ์ของการประมาณค่าข้อมูลสูญหาย พบว่า วิธี Copulas มีประสิทธิภาพดีกว่า วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย และ วิธี MCMC โดยให้ค่า MSE ต่ำสุด สำหรับวิธี MCMC มีประสิทธิภาพดีกว่าวิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย เฉพาะในกรณีโครงสร้างเมตริกซ์ความสัมพันธ์แบบ Autoregressive ส่วนผลที่ได้จากการนำวิธีประมาณค่าสูญหายไปประยุกต์กับข้อมูลจริงจำนวน 2 ชุด คือ 1) ข้อมูลรอบเอวของผู้เข้าร่วมโครงการ “นพรัตน์ไร้พุง มุ่งสุขภาพดี” จากการวัดรอบเอว 4 ครั้ง และ 2)ข้อมูลปริมาณน้ำฝนรายเดือนของเดือน มิถุนายน – สิงหาคม 2550 ของกรมอุตุนิยมวิทยาภาคเหนือ พบว่า วิธี Copulas มีประสิทธิภาพดีกว่า วิธี MCMC และ วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย ซึ่งสอดคล้องกับผลที่ได้จาก ข้อมูลจำลองสถานการณ์

_____ _____ ___ / ___ / ___

ลายมือชื่อนิสิต ลายมือชื่ออาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

Duangporn Potawee 2009: The Estimation of Missing Data in Repeated Measurements Using Markov Chain Monte Carlo and Copulas Methods. Master of Science (Statistics), Major Field: Statistics, Department of Statistics. Thesis Advisor: Assistant Professor Lily Ingsisawang, Ph.D. 99 pages.

The objectives of this research were to study two data imputation methods: Markov Chain Monte Carlo (MCMC) and Copulas, and compare these two methods with the simple form of mean imputation in repeated measures data. The datasets used in this study were simulated by the Monte Carlo technique. Each unit was measured on three occasions under the following conditions: 1) data had multivariate normal distribution with two types of correlation structures: Compound Symmetry (CS) and Autoregressive (AR), 2) the level of inter-correlation among repeated observations are determined as low level (0.3), middle level (0.5) and high level (0.7), 3) the sample sizes used are: 30, 70, and 100 units, and 4) the last measurement was assigned to be missing at random (MAR) by rates of missing data of 5%, 10%, 20% and 30%. These gave rise to a total of 72 possible situations. For the MCMC method, the number of imputations was set to be five. Each defined situation was repeated 1,000 times by SAS program, and the performance of each imputation method was evaluated using mean squared error (MSE). The lower MSE was used to indicate the more effective imputation method.

The results from all situations in the simulation study showed that the Copulas method was the most effective method, yielding the lowest value of MSE. The MCMC method was more effective than the mean imputation when the repeated data had correlation structure as AR. In addition, we applied the above three imputation methods with two datasets in practice: 1) the "Nopparat without Fat, but Healthy" project where the data on waist circumference of participants were repeat measured four times during November 21, 2008 – March 3, 2009; and 2) the data on monthly rainfall in the northern part of Thailand where the Department of Meteorology reported the volume of rainfall by month and by rain station in the north region during June - August 2009. The results also indicated that the Copulas was the most effective method, which was consistent with the simulation data.

_____ _____ ___ / ___ / ___

Student's signature Thesis Advisor's signature

กิตติกรรมประกาศ

ข้าพเจ้าขอขอบพระคุณ ผศ.ดร.ลี่ลี อิงศรีสว่างอาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก ที่ได้ช่วยเหลือในการวางแผนงานวิจัยในวิทยานิพนธ์ฉบับนี้ ตลอดจนการให้คำปรึกษา แนะนำ และตรวจแก้ไขข้อบกพร่องต่าง ๆ ขอกราบขอบพระคุณ รศ.จิรัชย์ สุขะเกตุ อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม ที่กรุณาให้คำแนะนำและช่วยเหลือในการทำวิทยานิพนธ์ให้สำเร็จลุล่วงไปด้วยดี ขอขอบพระคุณ รศ. เปรมใจ ตรีสรานุวัฒนา ประธานกรรมการสอบ และ รศ. ดร.นิภา โรจน์รุ่งวศินกุล กรรมการสอบ และท่านอาจารย์ทุกท่านที่ประสิทธิ์ประสาทความรู้ให้แก่ข้าพเจ้าตลอดมา

ขอขอบพระคุณ พ่อ แม่ พี่เอ พี่อุ๊ ป้าสร พี่นันต์ ที่เป็นกำลังใจในขณะศึกษาและทำวิทยานิพนธ์

ขอขอบคุณ พี่ ๆ เพื่อน ๆ ทุกคนที่ให้ความช่วยเหลือและเป็นกำลังใจขณะทำวิทยานิพนธ์

สุดท้ายนี้ คุณค่าและประโยชน์อันจะพึงมีจากวิทยานิพนธ์ฉบับนี้ ผู้วิจัยขอมอบแด่ครอบครัว บูรพาจารย์ และผู้มีพระคุณทุกท่าน

ดวงภรณ์ โปทาวี

เมษายน 2552

สารบัญ

สารบัญตาราง

สารบัญตาราง (ต่อ)

สารบัญตาราง (ต่อ)

สารบัญภาพ

สารบัญภาพ (ต่อ)

สารบัญภาพ (ต่อ)

สารบัญภาพ (ต่อ)

คำอธิบายสัญลักษณ์และคำย่อ

Mean	=	การแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย
MCMC	=	วิธีมาร์คอฟเชนมอนติคาร์โล
Copulas	=	วิธีคอปูลาส์
Imputation	=	การแทนค่าข้อมูลสูญหาย
MCAR	=	การสูญหายแบบสุ่มอย่างสมบูรณ์
MAR	=	การสูญหายแบบสุ่ม
NMAR	=	การสูญหายแบบไม่สุ่ม
n	=	ขนาดตัวอย่าง
p	=	จำนวนตัวแปร
ρ	=	ความสัมพันธ์ระหว่างค่าวัดซ้ำ
σ	=	ค่าเบี่ยงเบนมาตรฐานของข้อมูล
Σ	=	เมตริกซ์ความแปรปรวน-ความแปรปรวนร่วม

การประมาณค่าข้อมูลสูญหายในการวัดซ้ำด้วยวิธีมาร์คอฟเชนมอนติคาร์โล และวิธีคอปูลาส์

The Estimation of Missing Data in Repeated Measurements Using Markov Chain Monte Carlo and Copulas Methods

คำนำ

ในการบริหารงานด้านต่าง ๆ เช่น ด้านการแพทย์ การศึกษา สังคมศาสตร์ ข้อมูลมีบทบาทสำคัญต่อการนำไปวิเคราะห์ประมวลผลเพื่อให้ผลลัพธ์ที่ถูกต้องเป็นประโยชน์ต่อการนำไปใช้ในการวางแผนตัดสินใจต่าง ๆ ปัจจุบันได้มีการนำสถิติมาช่วยในการตัดสินใจมากขึ้น แต่ปัญหาที่พบส่วนใหญ่คือข้อมูลไม่สมบูรณ์ (Incomplete data) อาจเกิดจากผู้ให้ข้อมูลตอบคำถามไม่ครบหรือบางครั้งไม่ตอบคำถามที่มีการสำรวจซ้ำ ข้อมูลที่กลุ่มตัวอย่างไม่ตอบหรือมีการลงข้อมูลไม่ครบถ้วนจะเรียกว่าเป็น “ข้อมูลสูญหาย (Missing data)” สำหรับการสำรวจซ้ำของข้อมูลมักพบในงานด้านการแพทย์ การวิจัยเชิงทดลอง ซึ่งเป็นการศึกษาที่ต้องใช้เงินและเวลาในการดำเนินงาน ซึ่งส่วนใหญ่ในการสำรวจซ้ำจะมีข้อมูลบางค่าสูญหายไป จึงทำให้ต้องสูญเสียทั้งเงินและเวลา ทั้งยังมีผลกระทบต่อการดำเนินงานและผลลัพธ์ในการวิเคราะห์ข้อมูล ดังนั้นก่อนนำข้อมูลไปวิเคราะห์ควรมีการตรวจสอบข้อมูลในเบื้องต้นว่ามีข้อมูลสูญหายหรือไม่

สาเหตุของข้อมูลสูญหายนั้นในทางปฏิบัติอาจจะไม่ทราบว่าการสูญหายของข้อมูลนั้นมีสาเหตุมาจากอะไร แต่สำหรับการจำลองการสูญหายของข้อมูลต้องกำหนดสาเหตุและรูปแบบของการสูญหาย ซึ่งแบ่งออกเป็น 3 ประเภท คือ 1) การสูญหายแบบสุ่มสมบูรณ์ (Missing Complete At Random: MCAR) เกิดขึ้นเมื่อความน่าจะเป็นของข้อมูลสูญหายไม่มีความสัมพันธ์กับค่าของข้อมูลตัวอื่น ๆ 2) การสูญหายแบบสุ่ม (Missing At Random: MAR) เกิดขึ้นเมื่อความน่าจะเป็นของข้อมูลสูญหายอาจจะขึ้นอยู่กับตัวแปรอื่นหรือสามารถทำนายจากตัวแปรอื่นได้ และ 3) การสูญหายแบบไม่สุ่ม (Not Missing At Random: NMAR) เมื่อสาเหตุของการสูญหายจะเกี่ยวข้องกับตัวแปรที่มีข้อมูลสูญหาย แต่ไม่มีความสัมพันธ์กับค่าของตัวแปรอื่น ๆ (Little and Rubin,1987) โดยส่วนใหญ่การจำลองการสูญหายของข้อมูลใช้การสูญหายแบบสุ่ม เพราะวิธีการทางสถิติที่พัฒนาขึ้น มักดำเนินการภายใต้ข้อสมมติของข้อมูลมีการสูญหายแบบสุ่มเป็นหลัก

การแก้ปัญหาการสูญหายของข้อมูลนั้น มักจะแก้ปัญหาโดย การตัดข้อมูลที่สูญหายทิ้งและนำข้อมูลที่สมบูรณ์เท่านั้นมาวิเคราะห์ จึงทำให้มีปริมาณข้อมูลลดน้อยลง มีผลต่ออำนาจการทดสอบทางสถิติและการประมาณค่าพารามิเตอร์มีความเอนเอียง (Roth, 1994) นอกจากนี้บางครั้งข้อมูลที่ทำการศึกษามีปริมาณน้อยอยู่แล้ว ผู้วิเคราะห์ไม่สามารถที่จะตัดข้อมูลทิ้งได้อีก ดังนั้นข้อมูลที่สูญหายไปจึงมีความสำคัญเพราะทำให้รายละเอียดบางอย่างของข้อมูลสูญหายไปด้วย ทำให้มีผลกระทบต่อผลลัพธ์และการสรุปผลของการวิเคราะห์ข้อมูล วิธีแก้ปัญหาข้อมูลที่สูญหายที่นิยมใช้กันเช่น การใช้ค่าเฉลี่ย (Mean) หรือใช้ค่าฐานนิยม (Mode) ของข้อมูล แทนค่าตำแหน่งข้อมูลที่สูญหายของตัวแปรต่าง ๆ แต่วิธีการเหล่านี้ทำให้ค่าประมาณข้อมูลที่สูญหายเบี่ยงเบนไปจากความเป็นจริง ด้วยเหตุนี้ Little และ Rubin (1987) จึงได้คิดค้นวิธีการจัดการข้อมูลสูญหาย เช่น การแทนค่าข้อมูลสูญหายด้วยวิธีการถดถอย วิธี Expectation Maximization (EM) หรือ Multiple Imputations (MI) เพื่อให้การแทนค่าข้อมูลที่สูญหายมีความสมบูรณ์และใกล้เคียงกับความเป็นจริงก่อนที่จะนำไปวิเคราะห์

เมื่อมีข้อมูลสูญหายเกิดขึ้นในการวัดซ้ำทำให้มีผลกระทบต่อดำเนินงานและผลลัพธ์การวิเคราะห์ข้อมูล การศึกษานี้จึงสนใจวิธีการประมาณค่าข้อมูลสูญหายสำหรับข้อมูลที่มีการวัดซ้ำโดยมีข้อตกลงว่าข้อมูลที่ใช้ในการศึกษาต้องเป็นข้อมูลที่มีการแจกแจงปกติ ในที่นี้สนใจ 2 วิธี คือ 1) วิธีการแทนค่าข้อมูลสูญหายด้วยวิธี MCMC ซึ่งเป็นวิธีแทนค่าข้อมูลสูญหายแต่ละค่าหลายครั้งโดยทดลองแทนค่าข้อมูลที่สูญหายด้วยค่าประมาณที่แตกต่างกันในแต่ละครั้ง และ 2) วิธีการแทนค่าข้อมูลสูญหายด้วย วิธี Copulas เป็นวิธีที่ยังไม่แพร่หลายนัก โดย Skar (1959) คิดค้นขึ้น และ Kaarik (2006) นำมาประยุกต์ใช้กับข้อมูลสูญหาย ในการศึกษาครั้งนี้ยังสนใจเปรียบเทียบประสิทธิภาพของการประมาณค่าข้อมูลสูญหายทั้งสองวิธีกับวิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ยซึ่งเป็นวิธีดั้งเดิม

วัตถุประสงค์

1. เพื่อศึกษาวิธีประมาณค่าข้อมูลที่สูญหาย วิธี MCMC และวิธี Copulas เมื่อค่าข้อมูลสูญหายเป็นการสูญหายแบบสุ่ม (MAR) กรณีข้อมูลวัดซ้ำ

2. เปรียบเทียบประสิทธิภาพของการประมาณค่าข้อมูลที่สูญหาย 3 วิธีคือ วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย วิธี MCMC และ วิธี Copulas

ขอบเขตการวิจัย

การศึกษานี้ได้กำหนดขอบเขตของงานวิจัยไว้ดังนี้

1. ข้อมูลที่ใช้ในการศึกษาเป็นข้อมูลที่มีการวัดซ้ำ 3 ครั้ง โดยให้ Y_1 เป็นตัวแปรแทนข้อมูลจากการวัดซ้ำครั้งที่ 1 ของหน่วยศึกษาใด ๆ Y_2 เป็นตัวแปรแทนข้อมูลจากการวัดซ้ำครั้งที่ 2 และ Y_3 เป็นตัวแปรแทนข้อมูลจากการวัดซ้ำครั้งที่ 3 ดังนั้น Y_1, Y_2 และ Y_3 เป็นตัวแปรที่มีความสัมพันธ์กัน ถูกสร้างขึ้นจากการแจกแจงปกติหลายตัวแปรด้วยเวกเตอร์ของค่าเฉลี่ยเท่ากับ 0 เมตริกซ์ความแปรปรวน-ความแปรปรวนร่วม (variance-covariance matrix)

$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}$ และเมตริกซ์โครงสร้างความสัมพันธ์ 2 แบบคือ

1.1. เมตริกซ์โครงสร้างความสัมพันธ์แบบ Compound Symmetry

$$R_{CS} = \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

1.2. เมตริกซ์โครงสร้างความสัมพันธ์แบบ Autoregressive

$$R_{AR} = \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}$$

เมื่อ σ_{ii} เป็นความแปรปรวนของตัวแปร Y_i มีค่าเท่ากับ 1

σ_{ij} เป็นความแปรปรวนร่วมระหว่างตัวแปร Y_i และ Y_j

โดยที่ $\sigma_{ij} = \rho_{ij} \sqrt{\sigma_{ii}} \sqrt{\sigma_{jj}}$; i, j = 1, 2, 3

2. กำหนดระดับความสัมพันธ์ระหว่างค่าวัดซ้ำ (ρ) มี 3 ระดับ คือ คือ ความสัมพันธ์ระดับต่ำ ($\rho = 0.3$) ความสัมพันธ์ระดับปานกลาง ($\rho = 0.5$) และ ความสัมพันธ์ระดับสูง ($\rho = 0.7$)

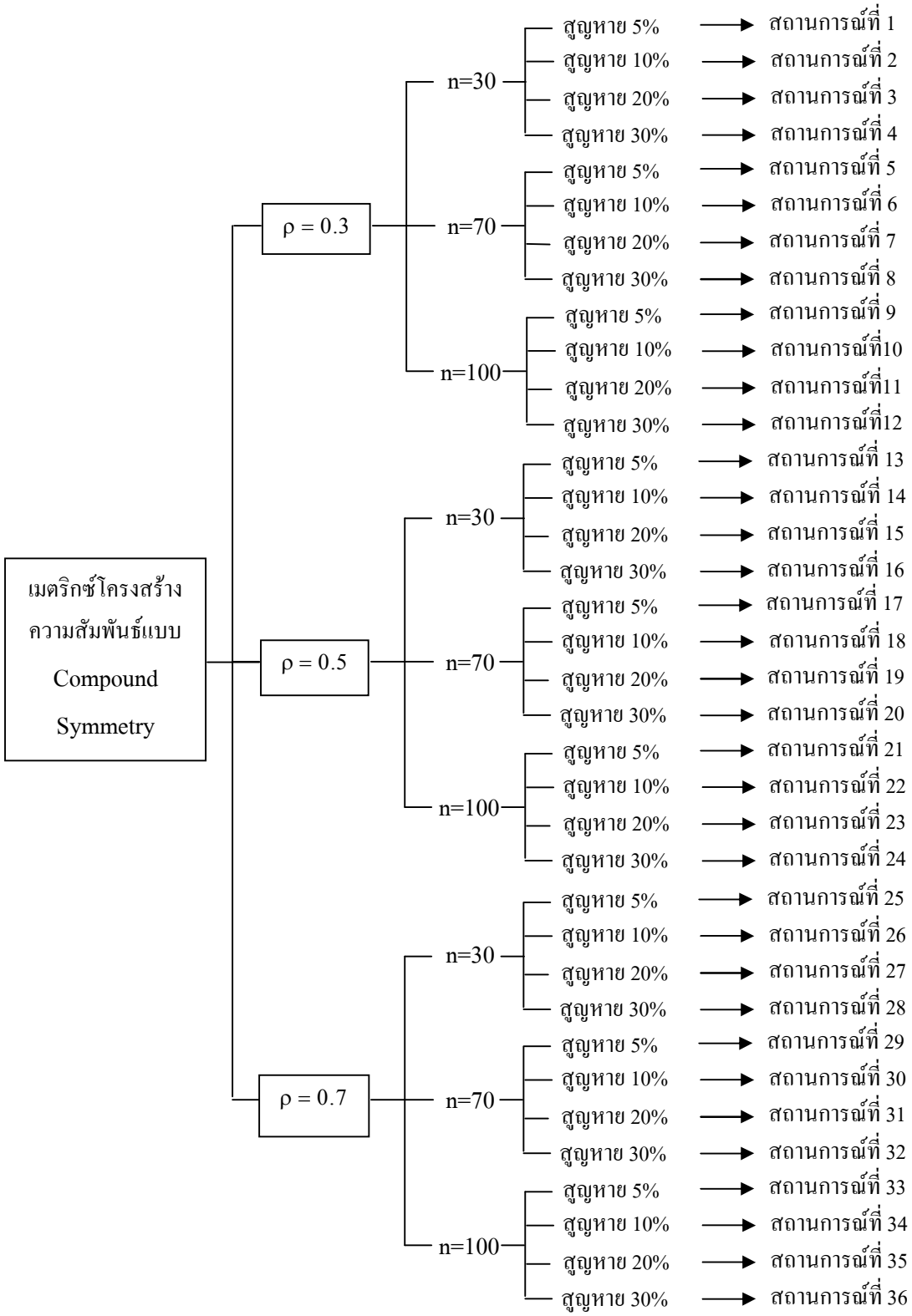
3. กำหนดขนาดตัวอย่างเท่ากับ 30, 70 และ 100

4. กำหนดตัวแปร Y_3 มีข้อมูลสูญหายเกิดขึ้นและมีการสูญหายแบบสุ่ม ที่ระดับการสูญหาย 5%, 10% , 20% และ 30% ของข้อมูลที่วัดซ้ำครั้งสุดท้าย

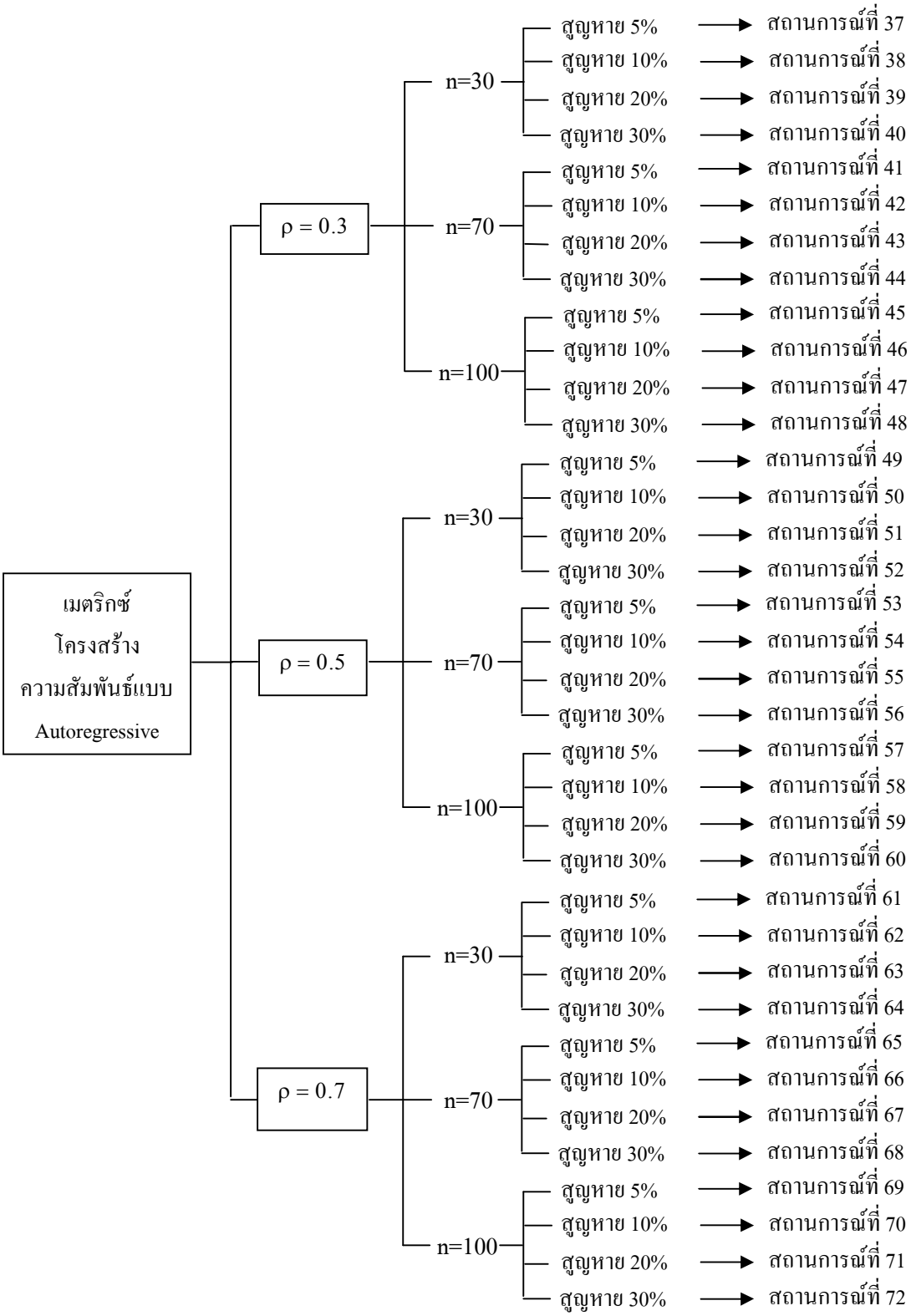
5. สำหรับการประมาณค่าสูญหายด้วยวิธี MCMC กำหนดจำนวนของการแทนค่าข้อมูลสูญหายแต่ละค่าจำนวน 5 ครั้ง

6. ข้อมูลที่ใช้ในการศึกษาเป็นข้อมูลที่ได้จากการจำลองสถานการณ์โดยใช้เทคนิควิธีมอนติคาร์โล (Monte Carlo Method) ทำการจำลองซ้ำ 1,000 ครั้งในแต่ละสถานการณ์ด้วยโปรแกรม SAS (Statistical Analysis System) และมีจำนวนสถานการณ์ทั้งหมด 72 สถานการณ์ โดยแบ่งเป็นสถานการณ์ที่มีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Compound Symmetry จำนวน 36 สถาณการณ์ และเมตริกซ์โครงสร้างความสัมพันธ์แบบ Autoregressive จำนวน 36 สถานการณ์ดังภาพที่ 1-2

7. ประยุกต์วิธี MCMC และ Copulas กับข้อมูล 2 ชุด คือ 1) ข้อมูลรอบเอวของผู้เข้าร่วมโครงการ “นพรัตน์ไร้พุง มุ่งสุขภาพดี” ของโรงพยาบาลนพรัตน์ราชธานี จากการวัดรอบเอว 4 ครั้ง โดยกำหนดให้ข้อมูลรอบเอวจากการวัดครั้งที่ 4 มีข้อมูลสูญหายเกิดขึ้น และ 2) ข้อมูลปริมาณน้ำฝนรายเดือนของเดือน มิถุนายน – สิงหาคม 2550 ของสถานีกรมอุตุนิยมวิทยาภาคเหนือ โดยกำหนดให้ข้อมูลเดือนสิงหาคมมีข้อมูลสูญหายเกิดขึ้น เมื่อการสูญหายที่เกิดขึ้นเป็นแบบสุ่ม ที่ระดับการสูญหาย 5%, 10%, 20% และ 30% ของข้อมูลสำหรับการวัดครั้งสุดท้าย ตามลำดับ



ภาพที่ 1 การจำลองสถานการณ์ที่มีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Compound Symmetry จำนวน 36 สถานการณ์



ภาพที่ 2 การจำลองสถานการณ์ที่มีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Autoregressive จำนวน 36 สถานการณ์

ประโยชน์ที่คาดว่าจะได้รับ

1. ได้เรียนรู้และสามารถประมาณค่าสูญหายของข้อมูลด้วยวิธี MCMC และวิธี Copulas

2. ได้แนวทางการเลือกวิธีการประมาณค่าสูญหายของข้อมูลที่เหมาะสม เมื่อข้อมูลมีการวัดซ้ำและมีค่าข้อมูลสูญหายในการวัดซ้ำครั้งสุดท้าย โดยที่มีระดับการสูญหายของข้อมูลแตกต่างกัน

การตรวจเอกสาร

การตรวจเอกสารแบ่งออกเป็น 2 ส่วน คือ ส่วนแรกกล่าวถึงวิธีการทางสถิติที่ใช้ในการวิจัย และส่วนที่สอง กล่าวถึงผลงานวิจัยที่เกี่ยวข้อง

วิธีการทางสถิติ

1. ประเภทของข้อมูลสูญหาย (Type of Missing Data)

ประสิทธิภาพของขั้นตอนวิธีหรือเทคนิคที่นำมาใช้ในการประมาณค่าสูญหายของข้อมูลนั้นจะดีหรือไม่ส่วนหนึ่งขึ้นอยู่กับรูปแบบการสูญหายของข้อมูล และหากทราบสาเหตุที่ทำให้เกิดข้อมูลสูญหาย ก็จะสามารถเติมเต็มหรือเดาข้อมูลส่วนนั้นได้ไม่ยาก แต่ในการทำงานจริงไม่ค่อยทราบว่าการสูญหายของข้อมูลนั้นมีสาเหตุมาจากอะไร และสูญหายในลักษณะใด ดังนั้นในการทดลองต่าง ๆ จึงมักกำหนดรูปแบบการสูญหายของค่าข้อมูลออกเป็น 3 ประเภท (Little and Rubin,1987) คือ

1.1. การสูญหายแบบสุ่มอย่างสมบูรณ์ (Missing Complete At Random: MCAR) การสูญหายของข้อมูลด้วยวิธีการสุ่มอย่างสมบูรณ์เกิดขึ้นเมื่อความน่าจะเป็นของการสูญหายของข้อมูลไม่มีความสัมพันธ์กับค่าของข้อมูลตัวอื่น ๆ ไม่ว่าจะเป็นข้อมูลที่ทราบค่า หรือข้อมูลที่เกิดการสูญหายด้วยกันก็ตาม นั่นคือความน่าจะเป็นที่จะเกิดการสูญหายของค่าข้อมูลในทุก ๆ ตำแหน่งมีค่าเท่ากัน Chantala และ Suchindran (nd) ได้ยกตัวอย่างการสูญหายแบบ MCAR ไว้เช่น ข้อมูลรายได้ของพนักงานเป็น MCAR ถ้าพนักงานที่ไม่รายงานรายได้ของตนเอง จะถูกสมมติว่ามีค่าเฉลี่ยรายได้เหมือนกับพนักงานคนอื่น ๆ ที่รายงานรายได้

สำหรับข้อมูลสูญหายประเภทนี้จัดเป็นข้อมูลที่ก่อให้เกิดปัญหาน้อยที่สุด เพราะว่าข้อมูลสูญหายไม่มีความเกี่ยวข้องต่อผลลัพธ์ของข้อมูล เพราะฉะนั้นสามารถเลือกทำการวิเคราะห์ข้อมูลในส่วนที่สมบูรณ์ได้ (ปิยะภรณ์ และ สุคนธ์, 2551)

1.2. การสูญหายแบบสุ่ม (Missing At Random: MAR) พิจารณาตัวแปร Y1, Y2 และ Y3 ให้ Y1 และ Y2 มีข้อมูลสมบูรณ์ และ Y3 มีข้อมูลสูญหาย ข้อมูลสูญหายใน Y3 เป็น MAR ถ้าความน่าจะเป็นของ Y3 อาจจะขึ้นอยู่กับกับ Y1 และ Y2 หรือสามารถทำนายจากตัวแปร Y1 และ Y2 ได้ แต่ ข้อมูลสูญหายใน Y3 จะไม่ขึ้นอยู่กับค่าสูญหายของตัวมันเอง เช่น รายได้ของพนักงานเป็น MAR ถ้าความน่าจะเป็นของข้อมูลรายได้ที่สูญหาย ขึ้นอยู่กับสถานภาพสมรส เช่น โสด แต่งงาน หรือหย่าร้าง แต่ความน่าจะเป็นของข้อมูลรายได้ที่สูญหายไม่ขึ้นอยู่กับข้อมูลรายได้ของตัวมันเอง

1.3. การสูญหายแบบไม่สุ่ม (Not Missing At Random: NMAR) สาเหตุของการสูญหายไม่สามารถบอกได้ และ สาเหตุของการสูญหายจะเกี่ยวข้องกับตัวแปรที่มีข้อมูลสูญหาย จะเรียกว่าเป็น Nonignorable ข้อมูลสูญหายสำหรับตัวแปร Y เป็น Nonignorable ถ้าความน่าจะเป็นของค่าสูญหายไปของ Y ไม่มีความสัมพันธ์กับค่าของตัวแปรอื่น ๆ แต่จะมีความสัมพันธ์กับค่าของตัวมันเอง เช่น ข้อมูลรายได้จะถือว่าเป็น NMAR ถ้าครอบครัวที่มีรายได้สูงส่วนใหญ่จะไม่ชอบรายงานรายได้ของตนเอง จึงทำให้ข้อมูลสูญหาย หรือ คนที่ดื่มสุรามาก ๆ จะหลีกเลี่ยงการตรวจแอลกอฮอล์มากกว่าคนที่ดื่มน้อย ทำให้เกิดการสูญหายของข้อมูลและหาสาเหตุได้ยาก

2. รูปแบบข้อมูลสูญหาย (Patterns of Missing Data)

มีรูปแบบดังนี้ (Chantala and Suchindran , nd)

2.1. ข้อมูลสูญหายหนึ่งตัวแปร(Univariate nonresponse)คือ ตัวแปร 1 ตัว มีข้อมูลสูญหาย

Case	Y1	Y2	Y3
A	4	7	8
B	7	6	
C	5	8	
D	6	6	8

2.2. ข้อมูลขาดหายมากกว่าหนึ่งตัวแปร (Multivariate two patterns) คือ มีข้อมูลสูญหายมากกว่าหนึ่งตัวแปรในหน่วยตัวอย่างเดียวกัน

Case	Y1	Y2	Y3
A	4	7	8
B	7	5	6
C	5		
D	6		

2.3. ข้อมูลสูญหายเป็นไปในทิศทางเดียวกัน (Monotone) คือ อันดับของตัวแปรหรืออันดับของค่าสังเกตในตัวแปรมีความสำคัญ นิยามคือ ให้เซตของตัวแปรคือ $Y_1, Y_2, ..., Y_p$ ถ้า Y_i มีค่าสูญหาย แล้ว $Y_{i+1}, Y_{i+2}, ..., Y_p$ จะที่มีค่าสูญหายด้วย

Case	Y1	Y2	Y3	Y4
A	4	7	4	8
B	7	5	6	
C	5	6		
D	6	7	8	5

2.4. ข้อมูลสูญหายแบบไม่เป็นระบบ (Arbitrary) โดยข้อมูลสูญหายสามารถเกิดขึ้นตรงจุดไหนก็ได้และอันดับของตัวแปรไม่มีความสำคัญ

Case	Y1	Y2	Y3	Y4
A	4	7	4	8
B		5	6	
C	5	6		7
D	6		5	8

3. เทคนิคที่ใช้ในการประมาณค่าที่สูญหายของข้อมูล (Method for Treating Missing Data)

จากการศึกษาพบว่าเทคนิคที่ใช้ในการประมาณค่าสูญหายของข้อมูลมีหลายวิธี Little และ Rubin (1987) ได้แบ่งวิธีการจัดการข้อมูลสูญหายไว้ดังนี้

3.1. การตัดข้อมูลทิ้ง (Ignoring and discarding data)

วิธีนี้เป็นวิธีที่ง่ายและสะดวกที่สุด โดยการตัดข้อมูลที่สูญหายออกไปแล้ววิเคราะห์ข้อมูลที่สมบูรณ์เท่านั้น ซึ่งนิยมใช้ในงานด้านสถิติ ตัวอย่างการตัดข้อมูลทิ้งมี 2 แบบ คือ

3.1.1. Listwise deletion เป็นการตัดค่าสังเกตหรือแถวที่มีข้อมูลสูญหายออกไปและจะนำค่าสังเกตที่สมบูรณ์เท่านั้นไปวิเคราะห์ วิธีนี้จะใช้ได้ก็ต่อเมื่อข้อมูลมีการสูญหายแบบสุ่มอย่างสมบูรณ์ (MCAR) เท่านั้น เนื่องจากการสูญหายในลักษณะอื่น ๆ นั้นจะมีความน่าจะเป็นของการสูญหายของข้อมูลขึ้นอยู่กับค่าของข้อมูลอื่น หรือค่าของข้อมูลที่เกิดการสูญหายเองด้วย ข้อเสียของวิธีการนี้คือ เนื่องจากมีการตัดข้อมูลที่มีการสูญหายออกทำให้ขนาดของกลุ่มตัวอย่างลดน้อยลง ดังนั้นอำนาจการทดสอบทางสถิติจึงลดลงด้วย

3.2.2. Pairwise deletion วิธีนี้จะไม่ตัดแถวที่ขาดความสมบูรณ์ทิ้งแต่จะนำแถวเหล่านั้นมาใช้ในการประมวลผลด้วย โดยจะพิจารณาทุก ๆ แถวที่มีค่าข้อมูลในตัวแปรที่กำลังสนใจ ถึงแม้ว่าตัวแปรอื่นจะไม่สมบูรณ์ก็ตาม วิธีการนี้มีข้อดีในส่วนของการใช้งานข้อมูลได้เต็มที่และมีประสิทธิภาพแต่มีขั้นตอนที่ซับซ้อนกว่า Listwise เล็กน้อยและเสียเวลามากกว่า จึงได้รับความนิยมน้อย แต่วิธีนี้จะทำให้สูญเสียอำนาจการทดสอบน้อยกว่า Listwise

3.2. การแทนค่าข้อมูลสูญหายด้วยค่าประมาณที่ได้จากวิธีการต่าง ๆ (Imputation) ดังต่อไปนี้

3.2.1. การแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ยหรือค่ามัธยฐาน เป็นวิธีที่ง่าย และนิยมใช้มากพอสมควรในการจัดการข้อมูลสูญหาย ในกรณีที่ข้อมูลเป็นข้อมูลตัวเลขและมีการแจกแจงแบบปกติจะแทนค่าข้อมูลที่สูญหายด้วยค่าเฉลี่ยของตัวแปรจากค่าสังเกตที่มีข้อมูลสมบูรณ์ แต่ถ้าข้อมูลมีการแจกแจงแบบเบ้ควรแทนค่าข้อมูลสูญหายด้วยค่ามัธยฐาน และสำหรับข้อมูลที่เป็นสัญลักษณ์หรือข้อมูลกลุ่มมักแทนค่าข้อมูลสูญหายด้วยค่าฐานนิยม

3.2.2. การแทนค่าข้อมูลสูญหายแบบ Hot Deck คือ เป็นการแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ยของกลุ่ม ซึ่งแบ่งเป็น 2 ขั้นตอนคือ 1) การแบ่งข้อมูลออกเป็นกลุ่มที่มีลักษณะใกล้เคียงกัน และ 2) แล้วหาค่าเฉลี่ยหรือค่ามัธยฐานของกลุ่มที่มีข้อมูลสมบูรณ์มาแทนที่ในตำแหน่งที่สูญหายไปของแถวข้อมูลที่ไม่สมบูรณ์ ซึ่งข้อมูลแต่ละแถวจะเป็นสมาชิกในกลุ่มใดกลุ่มหนึ่งที่ถูกแบ่งมาจากขั้นตอนแรก

3.2.3. การแทนค่าข้อมูลสูญหายด้วยวิธีการถดถอย (Regression imputation) วิธีนี้ใช้การวิเคราะห์การถดถอยเพื่อสร้างสมการทำนายข้อมูลสูญหายจากข้อมูลสมบูรณ์ที่มีอยู่ โดยกำหนดให้ตัวแปรอิสระ (x) มีข้อมูลสมบูรณ์และมีความสัมพันธ์กับ ตัวแปรตาม (y) ซึ่งเป็นตัวแปรที่มีข้อมูลสูญหาย แต่ถ้าตัวแปรอิสระ กับตัวแปรตาม ไม่มีความสัมพันธ์กันอาจจะทำให้ตัวแบบการประมาณค่าที่ได้ไม่ถูกต้องเท่าที่ควร เนื่องจากการสร้างตัวแบบการประมาณค่าในวิธีการนี้จะต้องอาศัยความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระเป็นหลัก

3.3. วิธีที่อาศัยตัวแบบ (Model-based procedures) วิธีนี้ใช้หลักการเกี่ยวกับ Likelihood การประมาณค่าพารามิเตอร์ ที่เรียกว่า Maximum likelihood โดยมีการทำซ้ำ เช่น วิธี EM (Expectation maximization) หรือ MI (Multiple imputations) เป็นต้น

เนื่องจากการประมาณค่าสูญหายด้วยวิธี MCMC และ วิธี Copulas มีข้อตกลงเบื้องต้นว่าข้อมูลต้องมีการแจกแจงปกติหลายตัวแปร (Schafer, 2005; Eaarik, 2006) ซึ่งรายละเอียดของการแจกแจงปกติหลายตัวแปรมีดังต่อไปนี้

4. การแจกแจงหลายตัวแปร

4.1. การแจกแจงปกติหลายตัวแปร (Multivariate Normal Distribution) Morrison (2005) ได้กำหนดนิยามไว้ดังนี้

นิยาม ให้ $\underline{Y}$ เป็นเวกเตอร์สุ่มที่มีฟังก์ชันความหนาแน่นน่าจะเป็น (Probability density function)

$$f(\underline{Y}) = \frac{1}{(2\pi)^{p/2}\,|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\underline{Y}-\mu)'\Sigma^{-1}(\underline{Y}-\underline{\mu})\right] \qquad (1)$$

นั่นคือ $\underline{Y}$ มีการแจกแจงแบบปกติหลายตัวแปร (Multivariate Normal Distribution) ด้วยค่าเฉลี่ย $\underline{\mu}$ และ ค่า เมตริกซ์ความแปรปรวน- ความแปรปรวนร่วม (variance-covariance matrix) Σ ดังนั้น $\underline{Y} \sim N_p(\underline{\mu}, \Sigma)$

$$\text{เมื่อ} \quad \underline{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{bmatrix} \quad , \quad \underline{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} \quad \text{และ} \quad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

เมตริกซ์โครงสร้างความสัมพันธ์ (Correlation matrix)

ถ้าให้ ρ_{ij} เป็นค่าโครงสร้างความสัมพันธ์ระหว่าง Y_i กับ Y_j จะได้ว่า

$$\rho_{ij} = \rho_{ji} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}} \tag{2}$$

ดังนั้น $$\sigma_{ij} = \rho_{ij}\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}$$

ถ้าเขียน ρ_{ij} ทั้งหมดอยู่ในรูปของเมตริกซ์สมมาตร (Symmetric matrix) โดยค่าที่เส้นทแยงมุม (Diagonal element) เท่ากับ $\rho_{ii} = \frac{\sigma_{ii}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{ii}}} = \frac{\sigma_{ii}}{\sigma_{ii}} = 1$ จะได้ เมตริกซ์โครงสร้างความสัมพันธ์ของประชากรที่สมมาตร คือเมตริกซ์ R ตามรูปแบบข้างล่างนี้ (Morrison, 2005)

$$R = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}$$

4.2. การแจกแจงแบบมีเงื่อนไขของตัวแปรปกติหลายตัวแปร (Morrison, 2005)

เมื่อแบ่ง partition ของ $\underline{Y}$ ที่เป็น nonsingular มีประชากรแจกแจงปกติขนาด (p+q) แบ่งเป็นสองส่วนย่อยคือ $\underline{Y} = \begin{bmatrix} \underline{Y}_1 \\ \cdots \\ \underline{Y}_2 \end{bmatrix} \sim N_{p+q}(\underline{\mu}, \Sigma)$; โดยที่ $\underline{Y}_1$ เป็นเวกเตอร์ มี p ค่า และ $\underline{Y}_2$ เป็นเวกเตอร์ มี q ค่าที่เหลือ ด้วยค่าเฉลี่ยและเมตริกซ์ความแปรปรวนร่วมที่ถูกแบ่ง partition คือ $\underline{\mu} = \begin{bmatrix} \underline{\mu}_1 \\ \cdots \\ \underline{\mu}_2 \end{bmatrix}$, $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma'_{12} & \Sigma_{22} \end{bmatrix}$ เมตริกซ์ย่อย Σ_{11} , Σ_{12} และ Σ_{22} มีขนาด $p \times p$, $p \times q$ และ $q \times q$ ตามลำดับ ดังนั้นฟังก์ชัน การแจกแจงแบบมีเงื่อนไขของ $\underline{Y_1}$ ขึ้นอยู่กับ $\underline{Y_2} = \underline{y_2}$ ที่ถูกกำหนดค่า คือ

$$g(\underline{y}_1 \mid \underline{y}_2) = \frac{f(\underline{y}_1, \underline{y}_2)}{h(\underline{y}_2)} \tag{3}$$

ได้ค่า

$$h(\underline{y}_2) = \frac{1}{(2\pi)^{q/2} \mid \Sigma_{22} \mid^{1/2}} \exp\left[-\frac{1}{2} (\underline{y}_2 - \underline{\mu}_2)' \Sigma_{22}^{-1} (\underline{y}_2 - \underline{\mu}_2) \right] \tag{4}$$

สำหรับการคำนวณ $g(y_1 \mid y_2)$ ต้องอาศัยฟังก์ชันความหนาแน่นร่วม $f(y_1, y_2)$ ที่ถูกแสดงในเทอมของเมตริกซ์ย่อยของ Σ ในรูปของอินเวิร์สของเมตริกซ์แสดงได้ดังนี้

$$\Sigma^{-1} = \begin{bmatrix} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma'_{12})^{-1} & -(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma'_{12})^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\ -\Sigma_{12}\Sigma_{22}^{-1}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma'_{12})^{-1} & \Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma'_{12}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma'_{12})^{-1}\Sigma_{12}\Sigma_{22}^{-1} \end{bmatrix}$$

และ determinant ของเมตริกซ์ Σ คือ

$$\mid \Sigma \mid = \mid \Sigma_{22} \mid \cdot \mid \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma'_{12} \mid \tag{5}$$

ความหนาแน่นร่วมสามารถเขียนได้ดังนี้

$$
\begin{aligned}
f(\underline{y}_1, \underline{y}_2) = & \frac{1}{(2\pi)^{(p+q)} \mid \Sigma_{22} \mid^{1/2} \times \mid \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}' \mid^{1/2}} \\
& \times \exp\Big\{-\frac{1}{2}[(\underline{y}_1 - \underline{\mu}_1)'(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}')^{-1}(\underline{y}_1 - \underline{\mu}_1) \\
& -(\underline{y}_2 - \underline{\mu}_2)'\Sigma_{22}^{-1}\Sigma_{12}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}')^{-1}(\underline{y}_1 - \underline{\mu}_1) \\
& -(\underline{y}_1 - \underline{\mu}_1)'(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}')^{-1}\Sigma_{12}\Sigma_{22}^{-1}(\underline{y}_2 - \underline{\mu}_2) \\
& +(\underline{y}_2 - \underline{\mu}_2)'\Sigma_{22}^{-1}\Sigma_{12}'(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}')^{-1}\Sigma_{12}\Sigma_{22}^{-1}(\underline{y}_2 - \underline{\mu}_2) \\
& +(\underline{y}_2 - \underline{\mu}_2)'\Sigma_{22}^{-1}(\underline{y}_2 - \underline{\mu}_2)]\Big\}
\end{aligned}
\tag{6}
$$

นำไปหารด้วย $h(\underline{y}_2)$ ได้ฟังก์ชันแบบมีเงื่อนไขของการแจกแจงปกติหลายตัวแปร คือ

$$
g(\underline{y}_1 \mid \underline{y}_2) = \frac{1}{(2\pi)^{p/2} \mid \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}' \mid^{1/2}} \times \exp\Big\{ \frac{1}{2}\Big[\underline{y}_1 - \underline{\mu}_1 - \Sigma_{12}\Sigma_{22}^{-1}(\underline{y}_2 - \underline{\mu}_2)\Big]'
$$

$$
\times\Big(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}'\Big)^{-1}\Big[\underline{y}_1 - \underline{\mu}_1 - \Sigma_{12}\Sigma_{22}^{-1}(\underline{y}_2 - \underline{\mu}_2)\Big]\Big\} \tag{7}
$$

ด้วยเวกเตอร์ค่าเฉลี่ย $\underline{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\underline{y}_2 - \underline{\mu}_2)$

และเมตริกซ์ความแปรปรวนร่วม คือ $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}'$

4.3. การแจกแจงแบบ Wishart (Wishart distribution)

Daniel (2002) ได้กำหนดนิยามการแจกแจงแบบ Wishart และ การแจกแจงส่วนกลับ Wishartไว้ดังนี้

ให้ $Y=\{y_{ij}\}$ เป็นเมตริกซ์ขนาด $m \times p$ เมื่อ $i = 1, 2, \ldots, m$ และ $j= 1, 2, \ldots, p$ ซึ่งแต่ละแถวของ Y เป็นอิสระกัน และมีการแจกแจงแบบปกติหลายตัวแปรด้วยค่าเฉลี่ยเท่ากับ 0 และความแปรปรวนร่วม Λ นั่นคือ

$$
y_i = (y_{i1}, y_{i2}, \ldots, y_{ip})' \sim \text{iid } N_p(0, \Lambda) \tag{8}
$$

แล้วผลรวมของผลคูณ Sum of Squares Cross Products เมตริกซ์ $A = Y'Y = \sum_{i=1}^{n} y_i y_i'$ ขนาด $p \times p$ มีการแจกแจงแบบ Wishart ของอันดับ p ด้วยองศาอิสระ m ดังนั้นเขียนได้ว่า $A \sim W_p(m, \Lambda)$ มีฟังก์ชันความหนาแน่นน่าจะเป็น คือ

$$f(A \mid \Lambda, m) = \frac{|A|^{(m-p-1)/2} \exp\left[-\frac{1}{2}\mathrm{tr}(\Lambda^{-1}A)\right]}{2^{mp/2}\pi^{p(p-1)/4}|\Lambda|^{m/2}\prod_{i=1}^{p}\Gamma[\frac{1}{2}(m+1-i)]} \tag{9}$$

$$\propto |A|^{(m-p-1)/2} \exp\left[-\frac{1}{2}\mathrm{tr}(\Lambda^{-1}A)\right] \qquad , |A| > 0$$

โดยที่ $A = \{A_{ij}\}$ และ $\Lambda = \{\sigma_{ij}\}$ สามารถหาค่าคาดหวัง (expected value) ของเมตริกซ์ A ได้จาก

$$E(A \mid \Lambda, m) = m\Lambda$$

ดังนั้น สำหรับค่าใด ๆ ของเมตริกซ์ $E(A_{ij}) = m\sigma_{ij}$

คุณสมบัติของการแจกแจงแบบ Wishart (Morrison, 2005) มีดังนี้

(1) เวกเตอร์ของค่าเฉลี่ยตัวอย่าง (sample mean vector) $\bar{y}$ และเมตริกซ์ A ถูกคำนวณจากตัวอย่างที่สุ่มมาจากประชากรที่มีการแจกแจงปกติหลายตัวแปรและเป็นอิสระกัน

(2) ถ้า $A_1, \ldots, A_k$ ถูกแจกแจงเป็นอย่างอิสระและเป็นเมตริกซ์ Wishart ที่มีความแปรปรวนร่วม Λ และองศาอิสระ $m_1, \ldots, m_k$ ตามลำดับ ผลรวมของ A_1 ถึง A_k หรือ $\sum_{i=1}^{k} A_i$ จะมีการแจกแจงแบบ Wishart ที่มีความแปรปรวนร่วม Λ และองศาอิสระ $m = \sum_{i=1}^{k} m_i$

ถ้า p=1 และ $\Sigma = 1$ การแจกแจงแบบ Wishart คือลักษณะทั่วไปของการแจกแจง Chi-squared ด้วยองศาอิสระ n

4.4. การแจกแจงส่วนกลับ Wishart (Inverted Wishart Distribution)

ถ้า $A \sim W_p(m, \Lambda)$ แล้ว $B = A^{-1}$ โดยมี $|\Lambda| > 0$ และ $m \geq p$ จะมีการแจกแจงส่วนกลับ Wishart คือ $B \sim W_p^{-1}(m, \Lambda)$ เมื่อ m คือองศาอิสระ โดยมีฟังก์ชันความหนาแน่นน่าจะเป็นดังนี้

$$p(B \mid \Lambda, m) = \frac{|B|^{-(m+p+1)/2} \exp\left[-\frac{1}{2}\mathrm{tr}(\Lambda^{-1}B^{-1})\right]}{2^{mp/2}\pi^{p(p-1)/4}|\Lambda|^{m/2}\prod_{i=1}^{p}\Gamma[1/2(m+1-i)]}$$

$$\propto |B|^{-(m+p+1)/2} \exp\left[-\frac{1}{2}\mathrm{tr}(\Lambda^{-1}B^{-1})\right] \quad , \ B > 0 \qquad (10)$$

สำหรับ $m < p$ เมตริกซ์ A เป็นซิงกูลาร์ (singular) จะหา A^{-1} ไม่ได้ ดังนั้นฟังก์ชันความหนาแน่นของ B ใช้ได้สำหรับตัวเลือกใด ๆ ของ $m \geq p$ และ $\Sigma > 0$ (Schafer, 1997)

และค่าคาดหวัง (expected value) ของ B คือ

$$E(B \mid \Lambda, m) = \frac{\Lambda^{-1}}{(m-p-1)} \quad ; \ m \geq p+2$$ และ p คือจำนวนตัวแปร y

5. วิธีการ Markov chain Monte Carlo (MCMC)

กระบวนการการแทนค่าข้อมูลสูญหายแต่ละค่าหลายครั้ง (Multiple Imputation) มีกลไกที่นิยมใช้ในการประมาณค่าข้อมูลที่สูญหายเพื่อให้ได้ข้อมูลที่สมบูรณ์ เช่น วิธีการถดถอย (Regression method) วิธีการไม่ใช้พารามิเตอร์ (Nonparametric method) วิธี MCMC หรือ วิธี Propensity score การเลือกใช้แต่ละวิธีจะขึ้นอยู่กับรูปแบบการสูญหายของข้อมูล สำหรับข้อมูลที่มีการแจกแจงแบบปกติหลายตัวแปรและมีรูปแบบการสูญหายไปในทิศทางเดียวกัน (Monotone) ควรเลือกใช้วิธีการถดถอยและวิธี Propensity Score ส่วนข้อมูลที่มีการแจกแจงแบบปกติหลายตัวแปรแต่มีรูปแบบการสูญหายแบบไม่เป็นระบบ (Arbitrary) ควรใช้วิธี MCMC จึงจะเหมาะสม (Schafer, 1997)

สำหรับวิธี MCMC มีข้อตกลงว่า

1. ข้อมูลที่ใช้ต้องมีการแจกแจงแบบปกติหลายตัวแปร

2. รูปแบบการสูญหายของข้อมูลแบบเป็นไปในทิศทางเดียวกันและแบบไม่เป็นระบบ

3. สาเหตุการสูญหายของข้อมูลเป็นแบบ MCAR และ MAR

วิธี MCMC จะสร้าง Markov chain ด้วยการจำลอง Monte Carlo ซ้ำหลาย ๆ ครั้ง จนกระทั่งได้ข้อมูลที่มีขนาดใหญ่ (Asymptotic) และมีการแจกแจงคงที่ (Stationary distribution)

5.1. MCMC สำหรับ การสูญหายของข้อมูล

ตัวประมาณค่าของ $\underline{\mu}$ และ Σ ที่หาได้จากวิธีการ EM ถูกนำมาใช้เป็นค่าพารามิเตอร์เริ่มต้นของการแจกแจงที่ใช้ในกระบวนการ MCMC ซึ่งมีขั้นตอนดังนี้

วิธี Expectation-maximization (EM)

วิธีการนี้เป็นการหาค่าประมาณ ของค่าเฉลี่ยและความแปรปรวน โดยอาศัยหลักของกระบวนการวนซ้ำ (Iterative procedure) ระหว่าง 2 ขั้นตอนดังนี้

สมมติให้ Y คือเมตริกซ์ขนาด n x p ที่มีขนาดตัวอย่างเท่ากับ n ด้วยค่าเฉลี่ย ด้วยเวกเตอร์ค่าเฉลี่ย μ และเมตริกซ์ความแปรปรวนร่วม Σ

Step 1 :M-Step

เริ่มต้นด้วยค่าประมาณของ μ และ Σ จากตัวอย่างหาได้จาก $\overline{Y}$ และ S ตามลำดับ และข้อมูลต้องไม่มีค่าสูญหาย ถ้าแต่ละแถวของชุดข้อมูลมีค่าสูญหายให้เริ่มต้นค่า $\mu = 0$ และ $\Sigma = I$

Step 2: E-Step

หาค่า $y_{i(miss)}$ ได้จากการคำนวณค่า $E(y_{i(miss)} \mid y_{i(obs)}, \hat{\mu}, \hat{\Sigma})$ แล้วคำนวณค่า $COV(y_{i(miss)} \mid y_{i(obs)}, \hat{\mu}, \hat{\Sigma})$, i= 1,2,…,N โดย $\hat{\mu}, \hat{\Sigma}$ คือค่าประมาณจาก M-Step

ทำซ้ำขั้นตอน M-Step และ E-Step จนกว่าค่า $(\hat{\mu}_{k+1}, \hat{\Sigma}_{k+1})$ จะเหมือนกันกับค่า $(\hat{\mu}_k, \hat{\Sigma}_k)$ (Schafer, 1997)

วิธี MCMC มี 2 ขั้นตอนมีดังนี้

5.1.1. The imputation: I-step

เมื่อกำหนดค่า ของ $\underline{\mu}$ และ Σ ที่ได้จากวิธี EM จากนั้นในขั้นตอน I-step จำลองค่าสูญหายของแต่ละค่าสังเกต Y_i อย่างเป็นอิสระกัน คือถ้าให้ $Y_{i(miss)}$ แทนตัวแปรของค่าสังเกตที่ i ที่เป็นค่าสูญหาย และให้ $Y_{i(obs)}$ เป็นตัวแปรของค่าสังเกตที่ i ที่เป็นค่าสังเกตจริงของ Y_i แล้วทำการสุ่ม $Y_{i(miss)}$ จากการแจกแจงแบบมีเงื่อนไขของ $Y_{i(miss)}$ ที่ถูกกำหนดด้วย $Y_{i(obs)}$ โดยมีรายละเอียดดังนี้

สำหรับค่าสังเกตที่ i , i= 1, 2, …, n แบ่ง $\underline{Y} = \begin{bmatrix} \underline{Y}_1 \\ \cdots \\ \underline{Y}_2 \end{bmatrix}$ เป็นสองส่วนย่อย

โดยที่ $\underline{Y}_1$ คือเวกเตอร์ของตัวแปร Y_{obs} และ $\underline{Y}_2$ คือเวกเตอร์ของตัวแปร Y_{miss} เวกเตอร์ค่าเฉลี่ย $\underline{\mu} = \left[\underline{\mu}'_1 , \underline{\mu}'_2 \right]'$ ถูกแบ่งเป็นสองส่วนย่อยที่สอดคล้องกับตัวแปร Y_{obs} และ Y_{miss} โดยที่ $\underline{\mu}_1$ คือเวกเตอร์ค่าเฉลี่ยสำหรับ Y_{obs} และ $\underline{\mu}_2$ คือเวกเตอร์ค่าเฉลี่ยสำหรับ Y_{miss} และให้ เมตริกซ์ความแปรปรวนร่วม $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ ถูกแบ่งเป็นเมตริกซ์ความแปรปรวนร่วมสำหรับตัวแปรเหล่านี้ คือ Σ_{11} คือ เมตริกซ์ความแปรปรวนสำหรับ Y_{obs}, Σ_{22} คือ เมตริกซ์ความแปรปรวนสำหรับสำหรับ Y_{miss} และ Σ_{12} คือ เมตริกซ์ความแปรปรวนร่วมระหว่าง Y_{obs} และ Y_{miss}

ดังนั้นจาก Morrison (2005) ข้อ 4.2 หน้า 14 การแจกแจงแบบมีเงื่อนไขของ Y_{miss} ที่ถูกกำหนดด้วย $Y_{obs} = \underline{y}_1$ เป็นการแจกแจงแบบ ปกติหลายตัวแปร โดยมี เวกเตอร์ค่าเฉลี่ยคือ

$$\underline{\mu}_{2.1} = \underline{\mu}_2 + \Sigma'_{12}\Sigma_{11}^{-1}(\underline{y}_1 - \underline{\mu}_1)$$

เมตริกซ์ความแปรปรวน- ความแปรปรวนร่วม คือ

$$\Sigma_{22.1} = \Sigma_{22} + \Sigma'_{12}\Sigma_{11}^{-1}\Sigma_{12}$$

การประมาณค่าแบบเบย์ของเวกเตอร์ค่าเฉลี่ย และ เมตริกซ์ความแปรปรวนร่วม

ในการอนุมานแบบเบย์เกี่ยวกับพารามิเตอร์ที่ไม่ทราบค่าต้องอาศัยฟังก์ชันการแจกแจงความน่าจะเป็นภายหลัง ซึ่งการแจกแจงภายหลังนี้ถูกคำนวณโดยใช้ทฤษฎีของเบย์ ในกรณีข้อมูลสมบูรณ์$Y=(Y_{miss}, Y_{obs})$ การแจกแจงภายหลังจะแปรผันตามการแจกแจงก่อนหน้า (Prior distribution) กับฟังก์ชัน likelihood (Schafer, 1997) ตามสมการต่อไปนี้

$$P(\underline{\theta} \mid Y) \propto \pi(\underline{\theta}) L(\underline{\theta} \mid Y) \tag{11}$$

เมื่อ $P(\underline{\theta} \mid Y)$ คือ การแจกแจงภายหลังของ θ ที่ถูกกำหนดด้วย Y

$\pi(\underline{\theta})$ คือ การแจกแจงก่อนหน้า (Prior distribution) ของ θ

$L(\theta \mid Y)$ คือ ฟังก์ชัน likelihood

$\underline{\theta}=(\underline{\mu}, \Sigma)$ คือพารามิเตอร์ที่ไม่ทราบค่า

เมื่อไม่ทราบการแจกแจงก่อนหน้าของ θ วิธีการของเบย์จะใช้ Conjugate prior คือใช้การแจกแจงเดียวกับ y สำหรับรูปแบบการแจกแจงปกติหลายตัวแปร คือ ใช้ family ของส่วนกลับ Wishart ที่เป็นแบบปกติ ให้แต่ละ y_i มีการแจกแจงแบบปกติหลายตัวแปร ด้วยค่าเฉลี่ยμ และความแปรปรวนร่วม Σ ดังนั้น จากนิยามการแจกแจงปกติหลายตัวแปรและการแจกแจงส่วนกลับ Wishart จะได้การแจกแจงก่อนหน้าของส่วนกลับ Wishart (Prior inverted Wishart distribution) สำหรับ Σ และ การแจกแจงก่อนหน้าปกติหลายตัวแปร (Prior multivariate normal distribution) สำหรับ μ (Schafer, 1997) คือ

$$\Sigma \sim w^{-1}(m, \Lambda) \tag{12}$$

$$\mu \mid \Sigma \sim N(\mu_0, \frac{1}{\tau}\Sigma) \qquad \text{เมื่อ } \tau > 0 \text{ เป็นค่าคงที่} \tag{13}$$

ตามที่อธิบายในหัวข้อ 4.2 – 4.4 หน้า 9-11 สำหรับพารามิเตอร์ที่กำหนดค่า $m \geq p$ และ $\Lambda > 0$ ฟังก์ชันการแจกแจงก่อนหน้าของ θ คือ

$$\pi(\theta) \propto |B|^{-(m+p+1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\Lambda^{-1}B^{-1})\right\} \times \exp\left\{-\frac{\tau}{2}(\mu-\mu_0)'\Sigma^{-1}(\mu-\mu_0)\right\} \tag{14}$$

และฟังก์ชัน likelihood ของข้อมูลสมบูรณ์คือ

$$L(\theta \mid Y) \propto |\Sigma|^{-n/2} \exp\left\{-\frac{n}{2} tr\Sigma^{-1}S\right\} \times \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)'\Sigma^{-1}(y_i - \mu)\right\} \quad (15)$$

เมื่อ S คือ เมตริกซ์ความแปรร่วมของตัวอย่าง (sample covariance matrix)

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})' \quad (16)$$

จากสมการ (11) นำ $\pi(\theta)$ คูณกับ $L(\theta \mid Y)$ ได้การแจกแจงภายหลังของค่าเฉลี่ย μ และความแปรปรวนร่วม Σ สำหรับข้อมูลสมบูรณ์ (Schafer, 1997) คือ

$$\Sigma \mid Y \sim w^{-1}\left(m + n, [(n-1)S + \Lambda^{-1} + \frac{n\tau}{n+\tau}(\bar{y} - \mu_0)(\bar{y} - \mu_0)']^{-1}\right) \quad (17)$$

$$\mu \mid (\Sigma, Y) \sim N\left(\frac{1}{n+\tau}(n\bar{y} - \tau\mu_0), \frac{1}{n+\tau}\Sigma\right) \quad (18)$$

5.1.2. The posterior : P-Step

เมื่อได้ค่า $Y_{i(miss)}$ ที่สร้างจาก I-Step ทำให้ได้ชุดข้อมูลสมบูรณ์ ขั้นตอน P-Step จำลองเวกเตอร์ของค่าเฉลี่ยประชากร μ และ เมตริกซ์ความแปรปรวนร่วม Σ จากการแจกแจงภายหลัง (Posterior Distribution)ของการแจกแจงปกติหลายตัวแปรและการแจกแจงส่วนกลับ Wishart ตามลำดับ แล้วค่าประมาณ μ และ Σ ตัวใหม่จะถูกนำไปใช้ในขั้นตอน I-step ต่อไป กรณีที่ไม่ทราบการแจกแจงข้อมูลก่อนหน้าของพารามิเตอร์ θ Schafer (1997) แนะนำให้ใช้ Noninformative prior แต่ถ้าทราบการแจกแจงข้อมูลก่อนหน้าของพารามิเตอร์ θ ก็จะเป็นประโยชน์ในการประมาณค่า θ ที่ดียิ่งขึ้น ซึ่งในการศึกษาครั้งนี้ศึกษากรณีที่ไม่ทราบข้อมูลก่อนหน้าเกี่ยวกับพารามิเตอร์ซึ่งมีรายละเอียดดังนี้

เมื่อไม่มีข้อมูลก่อนหน้าเกี่ยวกับพารามิเตอร์ θ ใช้ทฤษฎีของเบย์ด้วยการแจกแจงก่อนหน้า

$$\pi(\theta) \propto |\Sigma|^{-\frac{(p+1)}{2}} \quad (19)$$

ซึ่งเป็นรูปแบบจำกัดของการแจกแจงปกติและส่วนกลับ Wishart ในสมการ (12) – (13) ขณะที่ $\tau \to 0$, $m \to -1$ และ $\Lambda^{-1} \to 0$ ดังนั้นจากสมการ (17) – (18) ได้การแจกแจงภายหลังของ μ และ Σ สำหรับข้อมูลสมบูรณ์ กรณีไม่มีข้อมูลก่อนหน้าที่เกี่ยวข้องกับพารามิเตอร์ μ และ Σ (Schafer, 1997) คือ

$$\Sigma^{(t+1)} \mid Y \sim W^{-1}\left(n-1, (n-1)S\right) \tag{20}$$

$$\mu^{(t+1)} \mid (\Sigma^{(t+1)}, Y) \sim N\left(\bar{y}, \frac{1}{n}\Sigma^{(t+1)}\right) \tag{21}$$

ขั้นตอน I-step และ P-Step ถูกวนซ้ำ จนกว่าจะได้ผลลัพธ์ที่น่าเชื่อถือสำหรับชุดข้อมูลที่แทนค่าแล้วหลาย ๆ ชุด นั่นคือ ด้วยค่าพารามิเตอร์ที่ใช้อยู่ $\theta^t = (\mu^t, \Sigma^t)$ จากการวนซ้ำครั้งที่ t ขั้นตอน I-step จะสุ่มค่า Y_{miss}^{t+1} จาก $P(Y_{miss} \mid Y_{obs}, \theta^t)$ และ P-Step สุ่ม θ^{t+1} จากข้อมูลสมบูรณ์ภายหลังด้วย $P(\theta \mid Y_{obs}, Y_{miss}^{t+1})$ การวนซ้ำนี้เป็นการสร้าง Markov Chain

$$(Y_{miss}^1, \theta^1), (Y_{miss}^2, \theta^2), \ldots$$

และวนซ้ำจนกว่าจะลู่เข้าสู่การแจกแจงคงที่ด้วยค่าเฉลี่ยและความแปรปรวนคงที่ (SAS Institute Inc., 1999)

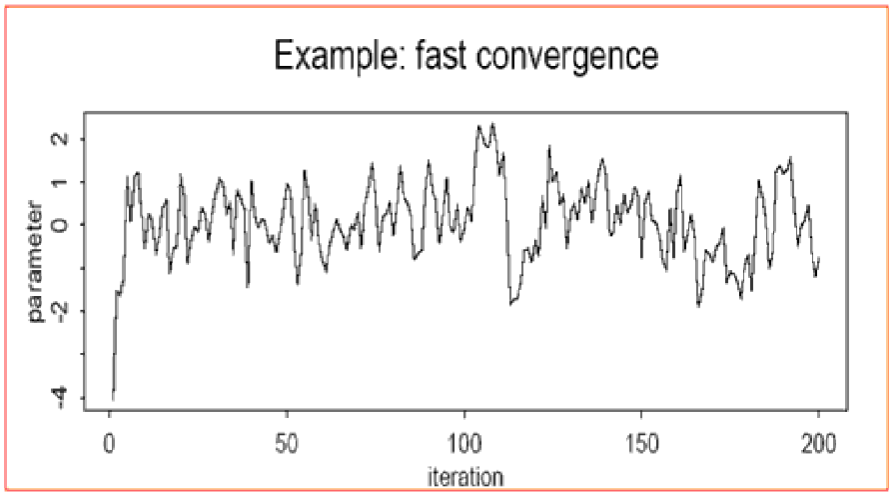
5.2. ตรวจสอบการลู่เข้าในขั้นตอนคำนวณแบบวนซ้ำของ MCMC (Check the Convergence of the MCMC Process)

วิธี MCMC มีข้อสมมติฐานที่ว่าตัวแปรต้องมาจากการแจกแจงแบบปกติหลายตัวแปร (Multivariate normal distribution) บางครั้งที่ตัวแปรเกือบจะมีการแจกแจงแบบปกติ เราสามารถแปลงค่า (transform) ตัวแปร เพื่อแปลงตัวแปรให้เป็นไปตามข้อตกลงของการแจกแจงแบบปกติหลายตัวแปร ซึ่งตัวแปรจะต้องถูกแปลงค่าก่อนเข้าสู่กระบวนการการแทนค่าข้อมูลสูญหาย (Imputation) และในวิธี MCMC ต้องทำวนซ้ำ I-Step และ P-Step เพื่อสร้าง Markov Chain ให้ลู่เข้าสู่การแจกแจงคงที่คือมีค่าเฉลี่ยและความแปรปรวนของการแจกแจงมีค่าคงที่ ดังนั้นควรตรวจสอบการลู่เข้าใน MCMC เพื่อตรวจสอบการแจกแจงคงที่หรือไม่ ก่อนจะนำไปสู่กระบวนการการแทน

ค่าข้อมูลสูญหายควรให้ข้อมูลเป็นไปตามข้อตกลงที่กำหนดไว้ ถ้าการลู่เข้าไม่สำเร็จ ให้เพิ่มจำนวนการวนซ้ำ (SAS Institute Inc., 1999) โดยใช้หลักการตรวจสอบดังนี้

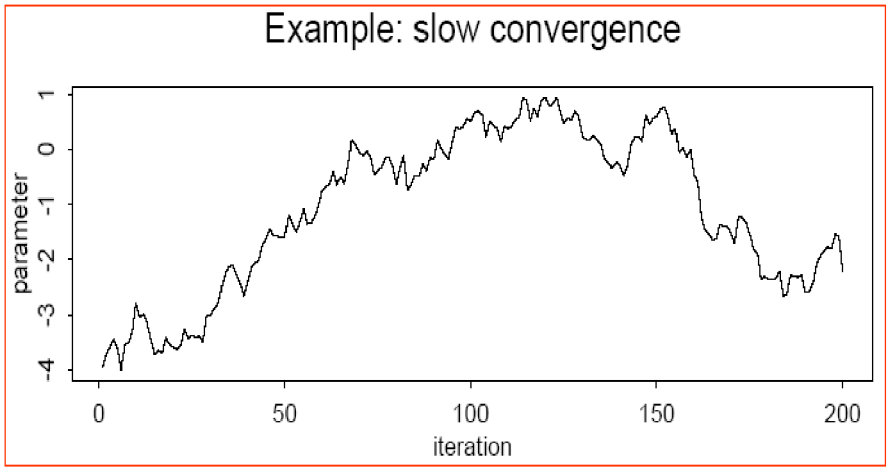
5.2.1. พล๊อตอนุกรมเวลา (Time-Series Plot)

การพล๊อตอนุกรมเวลาสำหรับพารามิเตอร์ θ คือการพล๊อตระหว่างฟังก์ชันของค่าประมาณพารามิเตอร์ θ กับจำนวนการวนซ้ำที่ i เพื่อตรวจดูคุณสมบัติการลู่เข้าของวิธีการประมาณค่าสำหรับ θ ถ้าพล๊อตอนุกรมเวลาแสดงถึงแนวโน้มแสดงว่าการวนซ้ำ (iteration)ที่ต่อเนื่องมีความสัมพันธ์กันสูงและอนุกรมของการวนซ้ำ (iteration)ไม่มีผลทำให้ลู่เข้าสู่การแจกแจงคงที่ นั่นคือความต้องการการวนซ้ำยังไม่เพียงพอ (SAS Institute Inc., 1999)



ภาพที่ 3 ตัวอย่างอนุกรมเวลาคงที่

จากภาพที่ 3 แสดงตัวอย่างการพล๊อตอนุกรมเวลาที่คงที่ไม่มีแนวโน้ม เนื่องจากแผนภาพกระจายของค่าประมาณพารามิเตอร์ θ ในแต่ละการวนซ้ำ มีการเคลื่อนไหวรอบ 0 แสดงว่าอนุกรมของการวนซ้ำ (iteration) มีผลทำให้ลู่เข้าสู่การแจกแจงคงที่ ดังนั้น การวนซ้ำ I-Step และ P-Step เพื่อสร้าง Markov Chain เพียงพอแล้ว (SAS Institute Inc., 1999)



ภาพที่ 4 ตัวอย่างอนุกรมเวลาที่มีแนวโน้ม

จากภาพที่ 4 แสดงตัวอย่างการพล๊อตอนุกรมเวลาที่มีแนวโน้ม เนื่องจากแผนภาพกระจายของค่าประมาณพารามิเตอร์ θ ในแต่ละการวนซ้ำ มีการเคลื่อนไหวให้เห็นถึงแนวโน้ม แสดงว่าอนุกรมของการรวนซ้ำ (iteration) ไม่มีผลทำให้ลู่เข้าสู่การแจกแจงคงที่ ดังนั้น การวนซ้ำ I-Step และ P-Step เพื่อสร้าง Markov Chain ยังไม่เพียงพอ ควรเพิ่มการวนซ้ำ

5.2.2. Autocorrelation Function Plot (ACF)

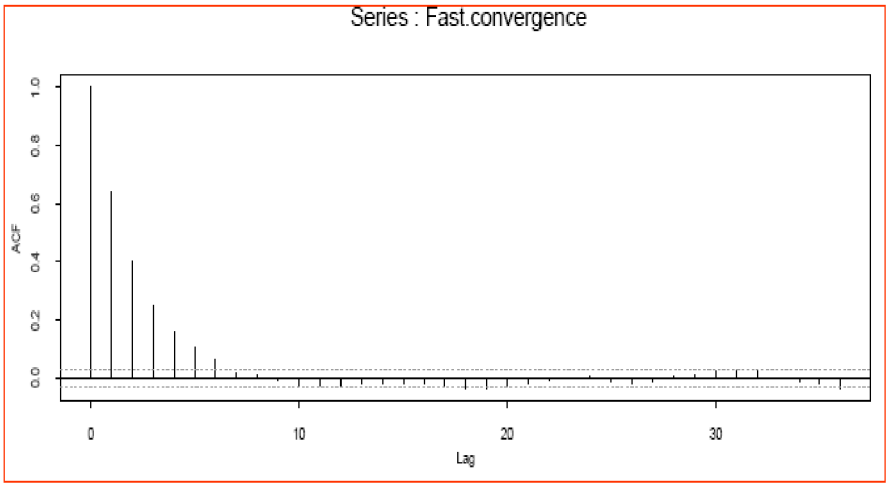
ACF ถูกนำมาใช้เพื่อตรวจดูความสัมพันธ์ของตัวประมาณพารามิเตอร์ θ สำหรับอนุกรมคงที่ ในข้อมูลอนุกรมเวลา สำหรับข้อมูลอนุกรมเวลา θ_t เมื่อ t = 1,2,…,m ค่า ACF สำหรับ lag ที่ k คือ

$$\rho_k = \frac{\text{cov}(\theta_t, \theta_{t+k})}{\text{Var}(\theta_t)} \tag{22}$$

สูตรการคำนวณ ค่า ACF ระหว่าง lag ที่ k สำหรับข้อมูลของตัวอย่าง คือ

$$r_k = \frac{\sum_{t=1}^{m-k}(\theta_t - \bar{\theta})(\theta_{t+k} - \bar{\theta})}{\sum_{t=1}^{m}(\theta_t - \bar{\theta})^2} \tag{23}$$

สามารถแสดงการพล๊อต ACF สำหรับฟังก์ชันเชิงเส้นเลวที่สุด, ค่าเฉลี่ยตัวแปร, ความแปรปรวนและความแปรปรวนร่วมของตัวแปร



ภาพที่ 5 ตัวอย่าง ACF เมื่อการวนซ้ำ I-step และ P-step เพียงพอแล้ว

จากภาพที่ 5 เป็นกราฟของค่า autocorrelation ตาม lag ต่าง ๆ พร้อมทั้งแสดงขอบเขตบนและขอบเขตล่าง ในการทดสอบ $\rho_k = 0$ ถ้าไม่มี ACF ที่เลยออกนอกช่วงวิกฤติ (เส้นประ) หมายความว่าตัวประมาณค่าพารามิเตอร์ต่อเนื่อง θ ไม่มีความสัมพันธ์กันที่ lag ต่าง ๆ นั่นคือความต้องการวนซ้ำ I-Step และ P-Step เพื่อสร้าง Markov Chain เพียงพอแล้ว

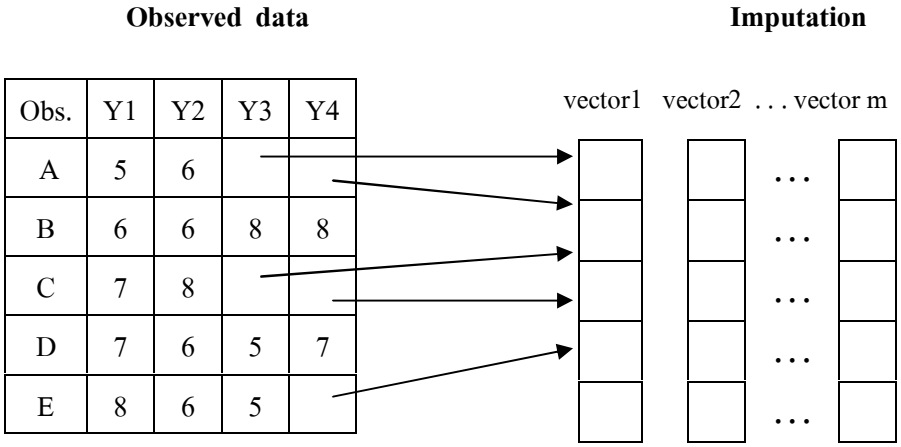
ภาพที่ 6 ตัวอย่าง ACF เมื่อการวนซ้ำ I-step และ P-step ยังไม่เพียงพอ

ภาพที่ 6 แสดง AFC ที่เลยออกนอกช่วงวิกฤติ (เส้นประ) หมายความว่าตัวประมาณค่าพารามิเตอร์ θ มีความสัมพันธ์กันที่ lag ต่าง ๆ นั่นคือความต้องการวนซ้ำ I-Step และ P-Step เพื่อสร้าง Markov Chain ยังไม่เพียงพอ ควรเพิ่มการวนซ้ำ (SAS Institute Inc., 1999)

5.3. การแทนค่าข้อมูลที่สูญหายแต่ละค่าหลายครั้ง (Multiple Imputations -MI)

เมื่อประมาณค่าสูญหายด้วย วิธี MCMC แล้วตรวจสอบว่าการวนซ้ำขั้นตอน I-step และ P-step เพียงพอแล้ว เมื่อค่าเฉลี่ยและความแปรปรวนของการแจกแจงมีค่าคงที่ ขั้นตอนต่อไปจึงเข้าสู่กระบวนการแทนค่าข้อมูลสูญหาย วิธีการแทนค่าข้อมูลที่สูญหายแต่ละค่าหลายครั้งเป็นการแทนค่าสูญหายแต่ละค่า โดยสร้างค่าประมาณข้อมูลสูญหายแต่ละค่าให้อยู่ในรูปแบบของเวกเตอร์ และใช้ค่าที่ประมาณได้จากเวกเตอร์ M ที่มีค่าตั้งแต่สองค่าขึ้นไป ซึ่ง $2 \le m < n$ นำไปแทนข้อมูลที่สูญหายแต่ละค่า เมื่อ เวกเตอร์ M คือ เวกเตอร์ของค่าประมาณที่จะนำมาแทนค่าสูญหายแต่ละค่าซึ่งนำมาจัดเรียงเป็นลำดับแล้ว ดังภาพที่ 3 เป็นตัวอย่างแบบแผนการใส่ค่าหลายครั้งแทนข้อมูลที่สูญหายแต่ละค่า เพื่อให้ได้ข้อมูลที่สมบูรณ์สมบูรณ์ m ชุด ซึ่งค่าสูญหายแต่ละค่าจะถูกแทนด้วยจุดที่ชี้ไปยังเวกเตอร์ m (Imputation) ซึ่งค่า m ค่าจะจัดเรียงเป็นลำดับ โดยที่ค่าลำดับแรกของเวกเตอร์เมื่อนำมาแทนข้อมูลสูญหายก็จะให้ข้อมูลที่สมบูรณ์ชุดที่หนึ่ง เมื่อนำค่าลำดับที่สองของเวกเตอร์ไปใส่แทนข้อมูลที่สูญหายก็จะให้ข้อมูลที่สมบูรณ์ชุดที่สอง เป็นเช่นนี้ไปเรื่อย ๆ จนถึงค่าลำดับที่ m เมื่อนำไปใส่แทนข้อมูลที่สูญหายก็จะให้ข้อมูลที่สมบูรณ์ชุดที่ m ดังนั้นจะได้ชุดข้อมูลสมบูรณ์ $Y= (Y_{obs}, Y_{miss}^{(i)}); i=1,2,...,m$ (Little and Rubin, 1987)

Rubin (1978) เป็นคนแรกที่เสนอแนวคิดเกี่ยวกับการแทนค่าข้อมูลที่สูญหายแต่ละค่าหลายครั้ง และได้ถูกนำเผยแพร่ในปี 1987 โดยมีหลักการของ การแทนค่าข้อมูลที่สูญหายแต่ละค่าหลายครั้ง ดังนี้ คือแทนค่าข้อมูลสูญหายแต่ละค่าจากค่าข้อมูลที่เป็นไปได้ใน m เวกเตอร์จะได้ทั้งหมด m ค่า จะเป็นการทำซ้ำ m ครั้งจากการแจกแจงที่คาดการณ์ได้ภายหลังของข้อมูลสูญหาย (Posterior predictive distribution) ซึ่งการทำซ้ำแต่ละครั้งจะสร้างพารามิเตอร์ $\theta = (\mu, \Sigma)$ และค่าสูญหาย



เมื่อช่องว่างของObserved data หมายถึงตำแหน่งที่มีข้อมูลสูญหาย

ภาพที่ 7 ตัวอย่างแบบแผนของการใส่ค่าแทนข้อมูลสูญหายแค่ละค่าหลายครั้ง เมื่อ m คือจำนวนครั้งของการแทนค่าสูญหาย

5.4 การรวมค่าอนุมานของการแทนค่าหลายครั้ง

จากการแทนค่าสูญหายแต่ละค่าทำให้ได้ค่าสูญหายที่ประมาณได้จากวิธี MCMC ของแต่ละค่าสังเกตที่มีค่าสูญหาย จำนวน m ค่า ดังนั้นรวมค่าสูญหายที่ประมาณได้ให้เป็นหนึ่งค่าดังนี้

$$\hat{y}_{i(miss)} = \frac{\sum_{j=1}^{m} y_{i(miss)}^{j}}{m} \quad , \quad j= 1, 2, \ldots, m \qquad (24)$$

6. การประมาณค่าสูญหายของข้อมูลโดยแทนค่าข้อมูลสูญหายใช้วิธีการCopulas

คำว่า copula มาจากภาษาละติน หมายถึงการเชื่อมเข้าด้วยกัน เนื่องจากการประมาณค่าสูญหายจะใช้การแจกแจงแบบมีเงื่อนไข โดยต้องทราบการแจกแจงร่วมของการวัดซ้ำก่อนหน้าค่าสังเกตที่เป็นข้อมูลสูญหาย สมมติให้ค่าสังเกตที่ได้จากการวัดครั้งที่ k เป็นค่าของข้อมูลสูญหาย ดังนั้น copulas จะถูกใช้เพื่อสร้างการแจกแจงร่วม (Joint distribution) จากการแจกแจงมาร์จินอล (Marginal distribution) ที่ได้จากการวัดซ้ำครั้งที่ 1, 2, …, k-1 เมื่อทราบการแจกแจงร่วมก็จะสามารถหาการแจกแจงแบบมีเงื่อนไขของการวัดวัดซ้ำข้อมูลครั้งที่ k ที่ขึ้นอยู่กับการวัดซ้ำข้อมูลครั้งที่ 1, 2, …, k-1 (Kaarik, 2006)

6.1. สัญลักษณ์และนิยามขั้นพื้นฐาน

Copula คือ ฟังก์ชันที่ยอมให้การแจกแจงร่วมของตัวแปรหลายตัวเป็นฟังก์ชันมาร์จินอลของตัวแปรเดียวที่สามารถระบุโครงสร้างความสัมพันธ์การขึ้นต่อกันได้ (Dependence structure) นั่นคือ copula จะเชื่อมโยงฟังก์ชันการแจกแจงมาร์จินอลของตัวแปรเดียวทั้งหลายเข้าเป็นฟังก์ชันการแจกแจงหลายตัวแปรร่วมกัน

เมื่อ copula เป็นฟังก์ชันการแจกแจงความน่าจะเป็นหลายตัวแปรที่มี n มิติ ที่มาจากการแจกแจงมาร์จินอลของตัวแปรเดียวที่เป็น Uniform บนช่วง $[0,1]$ ดังนั้นจะได้ copulas = $[0,1]^n$

Copulas มีหลายชนิดที่แตกต่างกัน ที่ง่ายที่สุดคือ independence copula หรือเรียกเป็น Copula ผลคูณ (Product copula) นั่นคือถ้าตัวแปรสุ่มที่เป็นอิสระต่อกันแล้วฟังก์ชัน copula จะเชื่อมโยงการแจกแจงมาร์จินอลของตัวแปรเล่านี้เป็น copula ผลคูณซึ่งสามารถใช้สร้างค่าทดแทนข้อมูลสูญหายที่การสูญหายเป็นแบบ MCAR ได้

ข้อมูลที่นำมาทดลองมีลักษณะแบบ Data Matrix เขียนแทนด้วย $X = \{x_{ij}\}$ เมื่อให้ i เป็นจำนวนหน่วยศึกษา; i = 1, 2, 3,…, n และ j เป็นครั้งที่ของการวัดซ้ำ; j = 1, 2, 3,…,p ถ้าละ subscript i และให้ $X_1, X_2, ..., X_k, ..., X_p$ เป็นตัวแปรสุ่มของการวัดครั้งที่ j โดยกำหนดให้ X_k เป็นเวกเตอร์ที่มีข้อมูลสูญหาย ดังนั้นเราจะพิจารณาลำดับของการวัดซ้ำข้อมูลถึงครั้งที่k คือ $X_1, X_2, ..., X_k$ และให้ H เป็นเวกเตอร์ที่มีข้อมูลในอดีตสมบูรณ์ นั่นคือ $H = X_1, X_2, ..., X_{k-1}$

สมมติให้ X_j มีการแจกแจงมาร์จินอล F_j , j = 1,2, … , k ที่มีการแจกแจงแบบเดียวกัน แต่โดยปกติเรามักจะไม่ทราบการแจกแจงร่วมของเวกเตอร์ $X=(X_1, X_2, ..., X_k)$ ดังนั้นจึงสร้างการแจกแจงร่วมโดยใช้ copula ซึ่งมีพื้นฐานทางทฤษฎีมาจากการสร้างตัวแบบของหลายตัวแปร (Multivariate modeling) ที่เสนอโดย Sklar (1959) โดยการแสดงให้ฟังก์ชันการแจกแจงร่วม $X_1, X_2, ..., X_k$ ที่มี k มิติ สามารถที่จะแตกออกเป็นการแจกแจงมาร์จินอลได้จำนวน k ฟังก์ชันมาร์จินอล ถ้าทราบการแจกแจงร่วมของเวกเตอร์ $X=(X_1, X_2, ..., X_k)$ ก็จะสามารถประมาณค่าข้อมูลสูญหายจากการแจกแจงแบบมีเงื่อนไขของ X_k ที่ขึ้นกับ H ได้ (Kaarik, 2006, 2007)

6.1.1. สร้างฟังก์ชันการแจกแจงร่วม (joint distribution function) ด้วย copula

ทฤษฎีของ Skar จะทำให้เห็นภาพของ Copula ที่แสดงบทบาทความสัมพันธ์ระหว่างฟังก์ชันการแจกแจงปกติหลายตัวแปร (Multivariate distribution function) กับการแจกแจงมาร์จินอลของมัน โดยสมมติให้ F เป็นฟังก์ชันการแจกแจงบน $\Re^k$ ซึ่งเป็นเซตของจำนวนจริงที่มี

k มิติ กับฟังก์ชันการแจกแจงมาร์จินอลหนึ่งมิติ $F_1(x_1), F_2(x_2), \ldots, F_k(x_k)$ เป็นฟังก์ชันการแจกแจงร่วม แล้วจะมี copula C เกิดขึ้นคือ

$$F(x_1, x_2, \ldots, x_k) = C(F_1(x_1), F_2(x_2), \ldots, F_k(x_k)) \tag{25}$$

สมมติให้ $C(u_1, u_2, \ldots, u_k)$ คือฟังก์ชันการแจกแจงร่วมด้วยมาร์จินอลที่เป็นยูนิฟอร์ม ถ้าการแจกแจงมาร์จินอล $F_1(x_1), F_2(x_2), \ldots, F_k(x_k)$ เป็นการแจกแจงแบบต่อเนื่อง แล้ว copula C จะมีค่าเดียว (unique) สำหรับทุก ๆ F โดย $u_j = F_j(x_j)$ แล้ว $x_j = F_j^{-1}(u_j)$ ดังนั้นจะได้

$$C(u_1, u_2, \ldots, u_k) = F(F_1^{-1}(u_1), F_1^{-1}(u_2), \ldots, F_1^{-1}(u_k)) \tag{26}$$

โดยที่ $F_1^{-1}, F_2^{-1}, \ldots, F_k^{-1}$ เป็นฟังก์ชันควอไทล์ของมาร์จินอลที่กำหนด $F_1(x_1), F_2(x_2), \ldots, F_k(x_k)$ และ $u_1, u_2, \ldots, u_k$ คือ ตัวแปรยูนิฟอร์มที่มีค่าอยู่ในช่วง [0,1] (Kaarik, 2006, 2007)

6.1.2. ความหนาแน่นร่วมและความหนาแน่นแบบมีเงื่อนไขของการวัดซ้ำข้อมูล

ถ้า C และ $F_1, F_2, \ldots, F_k$ สามารถหาอนุพันธ์ได้ ความหนาแน่นร่วม $f(x_1, x_2, \ldots, x_k)$ จะตรงกับการแจกแจงร่วม $F(x_1, x_2, \ldots, x_k)$ สามารถเขียนเป็นผลคูณของความหนาแน่นมาร์จินอลและความหนาแน่น copula ที่สามารถหาอนุพันธ์ของสมการที่ 25 ได้ดังนี้

$$f(x_1, x_2, \ldots, x_k) = f_1(x_1) \cdot f_2(x_2) \cdot \ldots \cdot f_k(x_k) \cdot c(F_1, F_2, \ldots, F_k) \tag{27}$$

โดยที่ $f_j(x_j)$ คือความหนาแน่นที่สอดคล้องกับ F_j (j = 1, 2, …,k) และความหนาแน่น copula c ถูกนิยามเป็นอนุพันธ์ของ copula หรือเรียกว่า ฟังก์ชันที่ไม่อิสระ (Dependence function) นั่นคือ

$$c(F_1, F_2, \ldots, F_k) = \frac{\partial^k C(F_1, F_2, \ldots, F_k)}{(\partial F_1\, \partial F_2 \ldots \partial F_k)} \tag{28}$$

จากอนุพันธ์ของ copula สามารถหาความหนาแน่นร่วมในสมการ (27) ได้โดยพิจารณาจากตัวอย่างต่อไปนี้

ในกรณีตัวอย่างของความหนาแน่น copula c สำหรับสองตัวแปร จะพิจารณาตัวแปรสุ่ม X_1 และ X_2 ที่มีการแจกแจงมาร์จินอล $F_1(x_1)$ และ $F_2(x_2)$ ตามลำดับ ให้ฟังก์ชันการแจกแจงร่วม F ถูกกำหนดโดย copula C ได้ $C(F_1(x_1),F_2(x_2)) = F(x_1,x_2)$ เมื่อให้ $u_1 = F_1(x_1)$ และ $u_2 = F_2(x_2)$ จะได้ $x_1 = F_1^{-1}(u_1)$ และ $x_1 = F_2^{-1}(u_2)$ ดังนั้นได้ copula c ที่ต่อเนื่อง (Kaarik 2007) ได้ดังนี้

$$c(u_1,u_2) = \frac{\partial^2 C}{\partial u_1 \partial u_2} = f((F_1^{-1}(u_1),F_2^{-1}(u_2))|J|$$

เมื่อ $$|J| = \begin{vmatrix} \frac{\partial x_1}{\partial u_1} & \frac{\partial x_1}{\partial u_2} \\ \frac{\partial x_2}{\partial u_1} & \frac{\partial x_2}{\partial u_2} \end{vmatrix}$$

ซึ่ง $$\frac{\partial x_i}{\partial u_i} = \left(\frac{\partial u_i}{\partial x_i}\right)^{-1} = \left(\frac{\partial F_i(x_i)}{\partial x_i}\right)^{-1} = f_i^{-1}(x_i)$$ และ $$\frac{\partial x_i}{\partial u_j} = \frac{\partial x_j}{\partial u_i} = 0$$

$, i \neq j,\ i,j = 1,2$

ดังนั้นได้ $$c(u_1,u_2) = f(F_1^{-1}(u_1),F_2^{-1}(u_2)) \cdot f_1^{-1}(F_1^{-1}(u_1)) \cdot f_2^{-1}(F_2^{-1}(u_2))$$

$$c(F_1(x_1),F_2(x_2)) = \frac{f(x_1,x_2)}{f_1(x_1) \cdot f_2(x_2)} \qquad (29)$$

ได้ความหนาแน่นร่วมคือ $f(x_1,x_2) = f_1(x_1) \cdot f_2(x_2) \cdot c(F_1(x_1),F_2(x_2))$

ในทำนองเดียวกัน ถ้าพิจารณาตัวแปรสุ่ม $X_1,X_2,...,X_k$ จะได้ความหนาแน่นร่วมตามสมการ (27)

เมื่อทราบความหนาแน่นร่วมจากสมการ (27) และนิยามความน่าจะเป็นแบบมีเงื่อนไข สามารถหาความหนาแน่นแบบมีเงื่อนไขได้ดังนี้

$$f(x_k \mid x_1,x_2,...,x_{k-1}) = \frac{f(x_1,x_2,...,x_{k-1},x_k)}{f(x_1,x_2,...,x_{k-1})}$$

$$= \frac{f_1(x_1)\cdot f_2(x_2)\cdot \ldots \cdot f_k(x_{k-1})\cdot f_k(x_k)\cdot c(F_1,F_2,\ldots,F_k)}{f_1(x_1)\cdot f_2(x_2)\cdot \ldots \cdot f_k(x_{k-1})\cdot c(F_1,F_2,\ldots,F_{k-1})}$$

$$= f_k(x_k)\frac{c(F_1,F_2,\ldots,F_k)}{c(F_1,F_2,\ldots,F_{k-1})} \qquad (30)$$

เมื่อ $c(F_1,F_2,\ldots,F_k)$ และ $c(F_1,F_2,\ldots,F_{k-1})$ คือ ความหนาแน่น copula

6.1.3. Gaussian copula

Copula ที่สำคัญมากอันหนึ่งคือ Normal copula หรือ Gaussian copula (Clemen and Reilly, 1999; Reilly, 1999; Song, 2000; Lindsay and Lindsay, 2002; Lambert and Vandenhende, 2002) โดยที่ Gaussian copula จัดอยู่ในกลุ่มของ implicit copula ที่ได้มาจากการแจกแจงปกติ implicit copula ไม่มีรูปแบบปิด (close form) ที่หาคำตอบได้ง่ายแต่สามารถนำไปประยุกต์ได้กับฟังก์ชันการแจกแจงปกติ จากนิยาม Gaussian copula ของ k ตัวแปร ถ้ามี k มาร์จินอลที่เป็นการแจกแจง Gaussian โดยมุมมองของcopula หลาย ๆ ตัว จะได้ว่าการแจกแจงปกติหลายตัวแปรจะมี การแจกแจงมาร์จินอลที่เป็นปกติและขึ้นต่อกัน

Gaussian copula จะแสดงการขึ้นต่อกันระหว่างการแจกแจงมาร์จินอลของตัวแปรเดียวผ่านเมตริกซ์โครงสร้างความสัมพันธ์ R ที่พิจารณาความสัมพันธ์ระหว่างตัวแปรเป็นคู่ที่มีทั้งหมด k(k-1)/2 คู่ได้

Kaarik (2007) ได้กำหนดนิยามพื้นฐานของ Gaussian copula ไว้ดังนี้

นิยาม ให้ R เป็นเมตริกซ์สมมาตรและมีค่าเป็นบวกเสมอ (Symmetric and positive difinite matrix) โดยมีค่าที่เส้นทแยงมุม (diagonal element) เท่ากับ 1 และ Φ_k เป็นฟังก์ชันการแจกแจงปกติมาตรฐานของ k ตัวแปร ที่มีเมตริกซ์โครงสร้างความสัมพันธ์ R แล้ว Gaussian copula หลายตัวแปร (multivariate Gaussian copula) ถูกนิยามดังนี้

$$C_k(u_1,u_2,\ldots,u_k;R) = \Phi_k(\Phi_1^{-1}(u_1),\Phi_1^{-1}(u_2),\ldots,\Phi_1^{-1}(u_k)) \qquad (31)$$

เมื่อ C_k คือ Gaussian copulaหลายตัวแปร

6.1.4. การสร้างตัวแบบความหนาแน่นของการวัดซ้ำข้อมูลโดย Gaussian copula

สมมติให้ X_j มีฟังก์ชันการแจกแจง F_j , j = 1,2, … , k และสามารถทำให้เป็นฟังก์ชันที่มีการแจกแจงปกติได้โดยวิธีการแปลง คือ

$$Y_j = \Phi_1^{-1}[F_j(X_j)] \ , \ j = 1, 2, \ldots, k \tag{32}$$

เมื่อ Φ_1^{-1} คือ อินเวิร์สของฟังก์ชันการแจกแจงปกติมาตรฐานตัวแปรเดียว

ดังนั้นใช้ ฟังก์ชัน copula ที่แจกแจงปกติ หรือ Gaussian copula k ตัวแปร จะได้ฟังก์ชันของการแจกแจงร่วมหลายตัวแปร (Joint multivariate distribution function) ดังนี้

$$F(y_1, y_2, \ldots, y_k; R) = C_k(u_1, u_2, \ldots, u_k; R) = \Phi_k[\Phi_1^{-1}(u_1), \Phi_1^{-1}(u_2), \ldots, \Phi_1^{-1}(u_k); R]$$

โดยที่ $u_j \in (0,1)$, $j = 1,2,\ldots,k$ และ Φ_k คือ ฟังก์ชันการแจกแจงปกติมาตรฐาน k ตัวแปร ที่มีเมตริกซ์โครงสร้างความสัมพันธ์ R

เพื่อให้ได้ฟังก์ชันความหนาแน่นร่วมที่มีการแจกแจงปกติเราจะต้องหาความหาความหนาแน่นของ Gaussian copula c_k ที่ได้มาจากการการหาอนุพันธ์ ของ Gaussian copula C_k ที่สอดคล้องกับสมการที่ (27) ดังนี้

$$\phi_k(y_1, \ldots, y_k; R) = \phi_1(y_1) \cdot \phi_1(y_2) \cdot \ldots \cdot \phi_1(y_k) \cdot c_k[\Phi_1(y_1), \ldots, \Phi_1(y_k); R^*] \tag{33}$$

เมื่อ Φ_1 คือฟังก์ชันการแจกแจงปกติมาตรฐานของตัวแปรเดียว

ϕ_1 คือฟังก์ชันความหนาแน่น (Density function)

c_k คือ ฟังก์ชันความหนาแน่น Gaussian copula

R^* คือ เมตริกซ์ของการวัดซ้ำที่ขึ้นต่อกันหรือมีความสัมพันธ์กัน

จากสมการ (33) ได้ความหนาแน่น Gaussian copula คือ

$$c_k[\Phi_1(y_1),\Phi_1(y_2),\ldots,\Phi_1(y_k);R^*] = \frac{\phi_k(y_1,y_2,\ldots,y_k;R)}{\phi_1(y_1)\cdot\phi_1(y_2)\cdot\ldots\cdot\phi_1(y_k)} \qquad (34)$$

ความหนาแน่น copula มีข้อมูลการขึ้นต่อกันระหว่างมาร์จินอลที่เรียกเป็นฟังก์ชันความหนาแน่น และแทนการคำนวณเชิงพีชคณิตทำให้ได้ผลลัพธ์ ความหนาแน่นของ Gaussian copula (Clemen and Reilly , 1999; Song , 2000) คือ

$$\begin{aligned} c_k\left[\Phi_1(y_1),\ldots,\Phi_1(y_k);R^*\right] &= |R|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}Y^tR^{-1}Y+\frac{1}{2}Y^tY\right\} \\ &= |R|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}Y^t(R^{-1}-I)Y\right\} \end{aligned} \qquad (35)$$

$Y = (Y_1, Y_2, \ldots, Y_k)$ และ I คือ เมตริกซ์ identity ขนาด $k \times k$

ผลลัพธ์ที่ได้จากสมการ 35 ไปแทนสมการ 27 (หน้า 29)

$$\begin{aligned} f(x_1,x_2,\ldots,x_k) &= f(x_1)\times\ldots\times f_k(x_k)\times |R|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}Y^t(R^{-1}-I)Y\right\} \\ &= f(x_1)\times\ldots\times f_k(x_k)\times |R|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}Q_k^t(R^{-1}-I)Q_k\right\} \end{aligned} \qquad (36)$$

เมื่อ $Q_k = (\Phi_1^{-1}[F_1(x_1)], \ldots, \Phi_1^{-1}[F_k(x_k)])$

6.1.5. การแจกแจงแบบมีเงื่อนไขและการประมาณค่าสูญหายจากการแจกแจงแบบมีเงื่อนไข

เพื่อแสดงการประมาณค่าสำหรับการแทนค่าข้อมูลที่สูญหาย จะหาความหนาแน่นแบบมีเงื่อนไขสำหรับ X_k ที่ขึ้นกับ $H = (X_1, X_2, \ldots, X_{k-1})$ โดยใช้การคำนวณ copula ที่มีการแจกแจงปกติหลายตัวแปร และกำหนดให้เมตริกโครงสร้างความสัมพันธ์ R= $\{r_{ij}\}$, i , j = 1,2, … ,k ถูกแบ่งเป็นส่วน ๆ ได้ดังนี้

$$R = \begin{pmatrix} R_{k-1} & r \\ r^{t} & 1 \end{pmatrix}$$

เมื่อ R_{k-1} คือเมตริกซ์โครงสร้างความสัมพันธ์ของ $H=(X_1, X_2, ..., X_{k-1})$ และ $r=(r_{1k}, ..., r_{(k-1)k})^{t}$ คือเวกเตอร์ของความสัมพันธ์ระหว่าง H และ X_k

ดังนั้นจาก การแจกแจงแบบมีเงื่อนไขในสมการ (30), ความหนาแน่น copula ในสมการ (34) และ partition ของเมตริกซ์โครงสร้างความสัมพันธ์ R ได้ความหนาแน่นแบบมีเงื่อนไข (Clemen and Reilly, 1999) คือ

$$f(x_k \mid x_1, ..., x_{k-1}; R^{*}) = f_k(x_k) \times \frac{c_k[\Phi_1(y_1), \Phi_1(y_2), ..., \Phi_1(y_k); R^{*}]}{c_{k-1}[\Phi_1(y_1), \Phi_1(y_2), ..., \Phi_1(y_{k-1}); R_{k-1}{}^{*}]}$$

$$= f_k(x_k) \times \frac{\phi_k(y_1, ..., y_k]; R)}{\phi_1(y_1) \cdot \phi_2(y_2) \cdot ... \cdot \phi_k(y_k)} \times \frac{\phi_1(y_1) \cdot \phi_2(y_2) \cdot ... \cdot \phi_{k-1}(y_{k-1})}{\phi_k(y_1, ..., y_{k-1}]; R_{k-1})}$$

$$= f_k(x_k) \times \frac{\phi_k(y_1, ..., y_k]; R)}{\phi_k(y_k) \cdot \phi_k(y_1, ..., y_{k-1}]; R_{k-1})} \qquad (37)$$

เมื่อ $Y_j = \Phi_1^{-1}[F_j(X_j)]$ ดังนั้น

$$f(x_k \mid x_1, x_2 .. x_k; R^{*}) = f_k(x_k) \times \frac{\phi_k(\Phi^{-1}[F_1(x_1)], ..., \Phi^{-1}[F_k(x_k)]; R)}{\phi_1(\Phi^{-1}[F_k(x_k)] \times \phi_{k-1}(\Phi^{-1}[F_1(x_1)], ..., \Phi^{-1}[F_{k-1}(x_{k-1})]; R_{k-1})} \qquad (38)$$

นำสมการ (38) ให้อยู่ในรูปของการแจกแจงปกติ และลดรูป

$$f(x_k \mid H; R) = f_k(x_k) \times \exp\left\{ -\frac{1}{2} \left[\frac{(\Phi^{-1}[F_k(x_k)] - r^{t} R_{k-1}^{-1} y_{k-1}^{*})^2}{(1 - r^{t} R_{k-1}^{-1} r)} - (\Phi^{-1}[F_k(x_k)])^2 \right] \right\} (1 - r^{t} R_{k-1}^{-1} r)^{-1/2} \qquad (39)$$

และได้การแจกแจงแบบมีเงื่อนไขของ y_k (kaarik, 2006) คือ

$$f(y_k \mid H, R^*) = \phi(y_k) \times \exp\left\{-\frac{1}{2}\left[\frac{y_k - r^t R_{k-1}^{-1}(y_{k-1}^*)^2}{(1 - r^t R_{k-1}^{-1} r)} - y_k^2\right]\right\} (1 - r^t R_{k-1}^{-1} r)^{-\frac{1}{2}} \tag{40}$$

เมื่อ $y_{k-1}^* = (y_1, \ldots, y_{k-1})^t$

ประมาณค่า y_k จากการแจกแจงแบบมีเงื่อนไขด้วยวิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood)

เมื่อ $y_1, \ldots, y_k$ เป็นตัวแปรสุ่มที่มีฟังก์ชันความหนาแน่น $f(y_k \mid H, R^*)$ ฟังก์ชันภาวะน่าจะเป็นของตัวแปรสุ่มแทนด้วยสัญลักษณ์ $L(y_k)$ (kaarik, 2007) จะได้

$$L(y_k) = \phi(y_k) \times \exp\left\{-\frac{1}{2}\left[\frac{y_k - r^t R_{k-1}^{-1} y_{k-1}^*)^2}{(1 - r^t R_{k-1}^{-1} r)} - y_k^2\right]\right\} (1 - r^t R_{k-1}^{-1} r)^{-\frac{1}{2}} \tag{41}$$

$$\ln L(y_k) = \ln [\phi_1(y_k)] - \frac{1}{2}\ln(1 - r^t R_{k-1}^{-1} r) - \frac{1}{2}\left[\frac{y_k - r^t R_{k-1}^{-1}(y_{k-1}^*)^2}{(1 - r^t R_{k-1}^{-1} r)} - y_k^2\right] \tag{42}$$

$$\frac{\partial \ln L(y_k)}{\partial y_k} = \frac{-y_k + r^t R_{k-1}^{-1} y_{k-1}^*}{(1 - r^t R_{k-1}^{-1} r)} = 0 \tag{43}$$

ดังนั้นได้ตัวประมาณคือ

$$\hat{y}_k = r^t R_{k-1}^{-1} y_{k-1}^* \tag{44}$$

เมื่อ $\hat{y}_k$ คือ ตัวประมาณค่าภาวะน่าจะเป็นสูงสุด (MLE) ของ y_k โดยที่ y_k คือตำแหน่งที่มีข้อมูลสูญหาย

6.2. การแทนค่าข้อมูลที่สูญหาย (Imputation)

แบ่งตามเมตริกซ์โครงสร้างความสัมพันธ์ออกเป็น 2 กรณีคือ

6.2.1. กรณีของ Compound Symmetry (CS)

เมื่อความสัมพันธ์ระหว่างตัวแปรคือ $r_{ij} = \rho$, $i, j = 1,2,...,k$; $i \neq j$ โดยเวกเตอร์สหสัมพันธ์ระหว่าง $y_1,...,y_{k-1}$ กับ y_k คือ $r = (\rho,\rho,...,\rho)^t$ และเมตริกซ์โครงสร้างความสัมพันธ์ สำหรับ $y_1,...,y_{k-1}$ ขนาด $(k-1)\times(k-1)$ คือ

$$R_{k-1} = \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{bmatrix} \tag{45}$$

หา inverse เมตริกซ์โครงสร้างความสัมพันธ์ R_{k-1}^{-1} ได้จาก

$$R_{k-1} = \begin{bmatrix} a & b & \dots & b \\ b & a & \cdots & b \\ \vdots & \vdots & \ddots & \vdots \\ b & b & \dots & a \end{bmatrix} \tag{46}$$

เมื่อ $a = 1+\dfrac{(k-2)\rho^2}{1-(k-2)\rho^2+(k-3)\rho}$ และ $b = 1+\dfrac{\rho}{1-(k-2)\rho^2+(k-3)\rho}$

นำค่าเหล่านี้แทนในสมการที่ (44) ได้ค่าประมาณของข้อมูลสูญหายในกรณี Compound Symmetry คือ

$$\hat{y}_k^{CS} = \frac{\rho}{1+(k-2)\rho}\sum_{j=1}^{k-1} y_j \tag{47}$$

เมื่อ $\hat{y}_k^{CS}$ คือ ตัวประมาณค่าของตำแหน่งที่มีข้อมูลสูญหาย ในกรณี Compound Symmetry

6.2.2. กรณีของ Autoregressive (AR)

เมื่อความสัมพันธ์ระหว่างตัวแปรคือ $r_{ij} = \rho^{|i-j|}$, $i, j = 1,2,...,k$; $i \neq j$ โดยเวกเตอร์โครงสร้างความสัมพันธ์ระหว่าง $y_1,...,y_{k-1}$ กับ y_k คือ $r = (\rho^{k-1},\rho^{k-2},...,\rho)^t$ และเมตริกซ์โครงสร้างความสัมพันธ์ สำหรับ $y_1,...,y_{k-1}$ ขนาด $(k-1)\times(k-1)$ คือ

$$R_{k-1} = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{k-2} \\ \rho & 1 & \rho & \cdots & \rho^{k-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{k-2} & \rho^{k-3} & \rho^{k-4} & \cdots & 1 \end{bmatrix} \tag{48}$$

หา inverse เมตริกซ์โครงสร้างความสัมพันธ์ R_{k-1}^{-1} ได้โดยให้ A เป็นเมตริกซ์โครงสร้างความสัมพันธ์ที่ไม่คงที่ของลำดับที่ k-1และ B เป็น three – diagonal matrix ของลำดับที่ k-1 ดังนั้น inverse เมตริกซ์ $A^{-1} = cB$ เมื่อ $c = \frac{1}{\rho^2 - 1}$ และ ค่าของเมตริกซ์ B คือ

1. $b_{ij} = 0$, ถ้า $|i - j| > 1$;
2. $b_{11} = b_{pp} = -1$ และ $b_{ii} = -(1-\rho)^2$, $i = 2,3,...,n-1$
3. $b_{ij} = \rho$ ถ้า $|i - j| = 1$

ดังนั้นจะได้ inverse เมตริกซ์โครงสร้างความสัมพันธ์ คือ

$$R_{k-1}^{-1} = \frac{1}{\rho^2 - 1}\begin{bmatrix} -1 & \rho & 0 & \cdots & 0 & 0 \\ \rho & -(1-\rho^2) & \rho & \cdots & 0 & 0 \\ 0 & \rho & -(1-\rho^2) & \cdots & 0 & 0 \\ 0 & 0 & \rho & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -(1-\rho^2) & \rho \\ 0 & 0 & 0 & \cdots & \rho & -1 \end{bmatrix} \tag{49}$$

นำค่า R_{k-1}^{-1} และค่า r^t แทนค่า ใน (44) ได้

$$\hat{y}_k = (\rho^{k-1},\rho^{k-2},...,\rho) \cdot R_{k-1}^{-1} \cdot (y_1,...,y_{k-1})^t$$

$$\hat{y}_k = \rho \cdot y_{k-1}$$

นำแนวโน้มทั่วไปของข้อมูลมาช่วยอธิบาย Kaarik (2006) ได้ปรับปรุงสูตรสำหรับการแทนค่าข้อมูลสูญหาย ในกรณีของ Autoregressive dependencies คือ

$$\hat{y}_k^{AR} = \rho \frac{S_k}{S_{k-1}} (y_{k-1} - \overline{Y}_{k-1}) + \overline{Y}_k \qquad (50)$$

เมื่อ $\hat{y}_k$ คือ ตัวประมาณค่าของตำแหน่งที่มีข้อมูลสูญหาย, $\overline{Y}_{k-1}$, $\overline{Y}_k$ คือค่าเฉลี่ยที่จุด k-1 และ k และ S_{k-1}, S_k คือค่าเบี่ยงเบนมาตรฐาน ที่จุด k-1 และ k

ค่าสหสัมพันธ์ (ρ) ที่ใช้ในวิธี Copulas จะใช้ค่าสหสัมพันธ์ของ Spearman ซึ่งหาได้ดังนี้

$$\rho_{ij} = \frac{6}{\pi} \arcsin\left(\frac{r_{ij}}{2}\right)$$

เมื่อ r_{ij} คือ ค่าสหสัมพันธ์ของ Pearson

7. จุดอ่อนและจุดแข็งของวิธี MCMC และ วิธี Copulas

การประมาณค่าข้อมูลที่สูญหายด้วย วิธี MCMC และวิธี Copulas มีทั้งจุดแข็งและจุดอ่อนดังต่อไปนี้

ตารางที่ 1 จุดแข็งและจุดอ่อนของวิธี MCMC และ วิธี Copulas

วิธีการ	จุดแข็ง	จุดอ่อน
MCMC	1. มีคำสั่งที่ใช้ในโปรแกรม SAS สำหรับการประมาณค่าและแทนค่าข้อมูลสูญหายทำให้ง่ายต่อการวิเคราะห์ข้อมูลและไม่ยุ่งยากเมื่อมีข้อมูลสูญหายจำนวนมาก	1. ใช้โปรแกรมเฉพาะ 2. ต้องตรวจสอบการลู่เข้าสู่การแจกแจงคงที่ในการะบวนการ MCMC

ตารางที่ 1 (ต่อ)

วิธีการ	จุดแข็ง	จุดอ่อน
MCMC	2. การสูญหายของข้อมูลเกิดขึ้นตำแหน่งไหนก็ได้	
Copulas	1. ง่ายในการคำนวณ 2. กรณีมีขนาดตัวอย่างน้อยกว่าจำนวนครั้งของการวัดซ้ำสามารถใช้วิธีนี้ได้	1. การวัดซ้ำก่อนหน้าต้องมีข้อมูลสมบูรณ์ 2. ยุ่งยากสำหรับการเทนค่าข้อมูลสูญหายเมื่อมีข้อมูลสูญหายจำนวนมาก 3. ต้องตรวจสอบรูปแบบความสัมพันธ์ก่อนว่าเป็นแบบ Compound Symmetry หรือ Autoregressive

8. เกณฑ์ที่ใช้เปรียบเทียบ

งานวิจัยนี้ได้ทำการวัดประสิทธิภาพของผลการทดลองโดยใช้ค่า MSE เพื่อให้เห็นความแตกต่างของประสิทธิภาพระหว่างวิธีต่าง ๆ ซึ่งสามารถหาค่า MSE ได้ดังนี้

ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Squares Error: MSE) เป็นการวัดประสิทธิภาพของการประมาณค่าสูญหายของข้อมูลที่คำนวณได้จากวิธีการต่าง ๆ เปรียบกับค่าของข้อมูลจริงที่ทราบอยู่แล้ว วิธีการใดให้ค่าประมาณของ MSE ต่ำกว่าจะเป็นวิธีการประมาณค่าสูญหายที่มีประสิทธิภาพมากกว่า เมื่อจำลองข้อมูลจำนวน 1,000 ครั้ง ได้ค่า MSE คือ

$$\text{MSE} = \frac{1}{1{,}000}\sum_{j=1}^{1{,}000}\left(\frac{\sum_{i=1}^{n}(Y_{ij}-\hat{Y}_{ij})^2}{n}\right) \qquad (51)$$

n คือ จำนวนข้อมูลที่มีการสูญหาย

$\hat{Y}_{ij}$ คือ ค่าประมาณของข้อมูลสูญหาย

Y_{ij} คือ ค่าจริงของข้อมูล

j คือจำนวนรอบของการทำซ้ำ $j=1,2,\ldots,1{,}000$

งานวิจัยที่เกี่ยวข้อง

บวรวรรณ (2543) ได้ประยุกต์ใช้วิธีการใส่ค่าหลายค่าแทนข้อมูลที่สูญหายแต่ละค่าในการวิเคราะห์ข้อมูลอุบัติเหตุผู้ขับขี่จักรยานยนต์ที่บาดเจ็บเข้ารับการรักษาที่แผนกฉุกเฉิน โรงพยาบาลราชวิถี จำนวน 2,668 ราย ตั้งแต่วันที่ 1 มกราคม – 31 ธันวาคม 2538 โดยใช้วิธีการใส่ค่าหลายค่าแทนข้อมูลที่สูญหายแต่ละค่าแบบสุ่มอย่างง่าย (Multiple Hot – Deck Imputation) ซึ่งเป็นวิธีการใส่ค่าแทนข้อมูลที่สูญหายอย่างง่าย โดยการสุ่มเลือกค่าที่จะนำไปใส่แทนค่าที่สูญหายซึ่งได้จากค่าที่มีคำตอบจากหน่วยตัวอย่างซึ่งจับคู่กันกับตัวแปรที่มีค่าสังเกต โดยจะทำการสุ่มค่าที่จะนำไปแทนข้อมูลที่สูญหายของหน่วยตัวอย่างนั้นหลายครั้ง จำนวนค่าที่ใส่แทนค่าสูญหายแต่ละค่า m = 3 และ 5 กับการใส่ค่าเพียงค่าเดียว m = 1 โดยให้ข้อมูลมีการสูญหายแบบเชิงสุ่ม (MAR) ผลการศึกษาพบว่า ค่าความคลาดเคลื่อนของสัมประสิทธิ์การถดถอยแบบลอจิสติคเมื่อใช้ m =3 และ 5 มีค่ามากกว่าการใส่ค่าแทนข้อมูลที่สูญหายเพียงค่าเดียว m = 1

เชาว์ (2547) ได้ศึกษาการพัฒนาวิธีการจัดการข้อมูลสูญหายแบบอีพีเอสเอสอี และตรวจสอบความแม่นยำ และอำนาจการทดสอบที่ได้จากวิธีการจัดการข้อมูลสูญหายแบบอีพีเอสเอสอีกับแบบอีเอ็มและแบบลิสท์ไวส์ ตามวิธีการสุ่มตัวอย่างแบบแบ่งชั้น แบบกลุ่ม และแบบหลายขั้นตอน ที่ความสัมพันธ์ระหว่างตัวแปรระดับต่ำ (r = .30) ปานกลาง (r = .50) และสูง (r = .70) และจำนวนข้อมูลสูญหาย 5% 10% 20% และ 30% และศึกษาปฏิสัมพันธ์ระหว่างวิธีการสุ่มตัวอย่าง วิธีการจัดการข้อมูลสูญหาย จำนวนข้อมูลสูญหาย และความสัมพันธ์ระหว่างตัวแปร ที่มีต่อความแม่นยำของค่าเฉลี่ยเลขคณิต ความแปรปรวน และสัมประสิทธิ์สหสัมพันธ์ ข้อมูลที่ใช้ศึกษามีลักษณะการแจกแจงแบบปกติสองตัวแปร และใช้เทคนิคมอนติคาร์โลซิมูเลชั่น จำลองการทดลอง ผลการศึกษาพบว่า วิธีการจัดการข้อมูลสูญหายโดยการแทนค่าแบบอีพีเอสเอสอีได้ค่าความแม่นยำของค่าเฉลี่ยเลขคณิต ไม่แตกต่างจากวิธีอีเอ็ม อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 วิธีการจัดการข้อมูลสูญหายโดยการตัดออกแบบลิสท์ไวส์ ได้ค่าความแม่นยำของความแปรปรวนแตกต่างจากวิธีอื่น ๆ อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05 และวิธีการจัดการข้อมูลสูญหายแบบลิสท์ไวส์ ได้ค่าความแม่นยำของสัมประสิทธิ์สหสัมพันธ์แตกต่างจากวิธีการจัดการข้อมูลสูญหายแบบอื่น ๆ อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05

Yuan (2000) ได้ศึกษาและพัฒนา วิธีการจัดการข้อมูลสูญหาย โดยใช้ Multiple Imputation โดยได้ยกตัวอย่างข้อมูล Fitness โดยกำหนดให้สาเหตุการสูญหายของข้อมูลเป็นแบบสุ่ม (MAR) และ เลือกใช้วิธี Regression และ วิธี Propensity score method กับรูปแบบการสูญหายแบบคล้ายกัน

(monotone) และ ใช้วิธี Markov chain Mote Carlo (MCMC) กับรูปแบบการสูญหายแบบไม่เป็น (arbitrary) โดยยกตัวอย่างการวิเคราะห์และแสดงผลลัพธ์การวิเคราะห์ด้วยโปรแกรม SAS

Huang (2004) ได้ศึกษาเปรียบเทียบวิธีการจัดการข้อมูลสูญหายในข้อมูลแบบวัดซ้ำในกลุ่มตัวอย่างขนาดเล็ก โดยมีวัตถุประสงค์ เพื่อ พัฒนาวิธีการใส่ค่าแทนข้อมูลที่สูญหายแต่ละค่า (Multiple Imputation Procedure) ของข้อมูลแบบวัดซ้ำในกลุ่มตัวอย่างขนาดเล็ก กระทำภายใต้ข้อตกเบื้องต้นข้อมูลมีการแจกแจงปกติหลายตัวแปร (multivariate normal distribution) ทำการศึกษาโดยใช้สถานการณ์จำลอง (simulation) และเปรียบเทียบวิธีการจัดการข้อมูลที่สูญหายระหว่างวิธีไม่ใส่ค่าแทนข้อมูลที่สูญหายโดยใช้วิธีภาวะน่าจะเป็นสูงสุด (maximum likelihood method) กับวิธีใส่ค่าแทนข้อมูลที่สูญหายด้วยวิธีเอ็มไอ (ใช้จำนวนค่าที่ใส่แทนข้อมูลที่สูญหาย m = 5) เพื่อทดสอบสมมติฐานของอิทธิพลของทรีเมนต์ และอิทธิพลของความคลาดเคลื่อนของทรีเมนต์ ผลการศึกษาปรากฏว่า วิธีการใส่ค่าแทนข้อมูลที่สูญหายด้วยวิธีเอ็มไอ จะให้ผลการทดสอบไม่แตกต่างจากวิธีไม่ใส่ค่าแทนข้อมูลที่สูญหายโดยวิธีภาวะน่าจะเป็นสูงสุด

Kaarik (2006) ได้ทำการศึกษาการประมาณค่าสูญหายของข้อมูลกรณีข้อมูลมีการวัดซ้ำ โดยใช้วิธี Copulas ซึ่งข้อมูลที่ใช้ในการทดลองเป็นข้อมูลที่มีขนาดเล็ก และให้รูปแบบการสูญหายของข้อมูลเป็นแบบสุ่มอย่างสมบูรณ์ (MCAR) แบบสุ่ม (MAR) และแบบไม่สุ่ม (MAR) การศึกษาครั้งนี้ได้ศึกษาในกรณีของ compound symmetry และ autoregressive dependencies นอกจากนี้ได้เปรียบเทียบ compound symmetry กับวิธี Last Observation Carried Forward (LOCF) และ autoregressive dependencies กับวิธีแทนค่าสูญหายโดยใช้การถดถอยเชิงเส้น โดยให้ ค่า $\rho = 0.5$, $\rho = 0.7$ และ n= 10 , 20 และให้ข้อมูลสูญหายที่ตำแหน่ง 3 , 6, 12 ผลการทดลอง พบว่า compound symmetry และ autoregressive มีประสิทธิภาพดีกว่าในกรณีที่รูปแบบการสูญหายของข้อมูลเป็นแบบสุ่มอย่างสมบูรณ์ (MCAR) และ แบบสุ่ม (MAR)

อุปกรณ์และวิธีการ

อุปกรณ์

1.เครื่องไมโครคอมพิวเตอร์ของภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

2. โปรแกรม SAS เวอร์ชัน 9.1

วิธีการ

1.การเตรียมข้อมูล

ข้อมูลที่นำมาใช้ในการศึกษาเป็นข้อมูลที่ได้จากการจำลองโดยวิธีมอนติคาร์โล กระทำซ้ำ 1,000 ครั้ง ในแต่ละสถานการณ์ จำนวนทั้งหมด 72 สถานการณ์ แบ่งเป็นสถานการณ์ที่มีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Compound Symmetry 36 สถาการณ์ และเมตริกซ์ความแบบ Autoregressive 36 สถานการณ์ดังภาพที่ 1-2 โดยจำลองข้อมูลวัดซ้ำที่มีการแจกแจงปกติหลายตัวแปร ด้วยจำนวนการวัดซ้ำ 3 ครั้ง มีเมตริซ์ความสัมพันธ์สัมพันธ์ 2 แบบคือ 1) แบบ Compound Symmetry และ 2) แบบ Autoregressive พร้อมทั้งนำวิธีการประมาณค่าสูญหายมาประยุกต์ใช้กับข้อมูลจริง 2 ชุด คือ 1) ข้อมูลรอบเอวของผู้เข้าร่วมโครงการ “นพรัตน์ไร้พุง มุ่งสุขภาพดี” จากการวัดรอบเอว 4 ครั้ง และ 2) ข้อมูลปริมาณน้ำฝนรายเดือนของเดือน มิถุนายน – สิงหาคม 2550 ของสถานีกรมอุตุนิยมวิทยาภาคเหนือ

2.การจำลองการสูญหายของข้อมูล

หลังจากผ่านขั้นตอนการเตรียมข้อมูลเรียบร้อยแล้ว สำหรับข้อมูลที่ได้จากการจำลองโดยวิธีมอนติคาร์โล ได้กำหนดให้การวัดซ้ำครั้งสุดท้าย มีข้อมูลสูญหายเกิดขึ้นและมีการสูญหายแบบสุ่ม ที่ระดับการสูญหายของข้อมูล 5%, 10%, 20% และ 30% ตามลำดับ ด้วยขนาดตัวอย่างเท่ากับ 30, 70 และ 100 ส่วนข้อมูลรอบเอวของผู้เข้าร่วมโครงการ “นพรัตน์ไร้พุง มุ่งสุขภาพดี” กำหนดให้การวัดรอบเอวครั้งที่ 4 มีข้อมูลสูญหาย และ ข้อมูลปริมาณน้ำฝนรายเดือน กำหนดให้

เดือนสิงหาคม มีข้อมูลสูญหาย ที่ระดับการสูญหาย 5%, 10%, 20% และ 30% ของข้อมูลที่วัดซ้ำครั้งสุดท้าย โดยการคำนวณหาจำนวนข้อมูลที่สูญหาย หาได้จาก

$$\text{จำนวนข้อมูลที่สูญหาย} = \frac{\text{ข้อมูลวัดซ้ำครั้งสุดท้าย} \times \text{เปอร์เซ็นต์ของการสูญหาย}}{100}$$

เช่น ข้อมูลที่วัดซ้ำครั้งสุดท้าย มีขนาดตัวอย่างเท่ากับ 30 เมื่อกำหนดระดับการสูญหายของข้อมูลคือ 10% ดังนั้นจำนวนข้อมูลสูญหายสำหรับกรณีนี้เท่ากับ 3 ตำแน่ง

3.ทดสอบการแจกแจงปกติ

หลังจากจำลองการสูญหายของข้อมูลที่ระดับต่าง ๆ แล้ว นำข้อมูลแต่ละชุดมาทดสอบการแจกแจงปกติ เพื่อให้เป็นไปตามข้อตกลงเบื้องต้นของวิธี MCMC และ Copulas ก่อนจะทำการประมาณค่าสูญหาย โดยใช้ Kolmogorov – Smirnov ซึ่งมีขั้นตอนการวิเคราะห์ดังนี้

3.1. <u>กำหนดสมมติฐาน</u> สมมติฐานหลัก (H_0) : ข้อมูลมีการแจกแจงปกติ
สมมติฐานรอง (H_1) : ข้อมูลไม่มีการแจกแจงปกติ

3.2. <u>สถิติทดสอบ</u> กำหนดระดับนัยสำคัญ 0.05

$$D = |F(x) - S(x)|$$

เมื่อ F(x) คือ ความน่าจะเป็นสะสมภายใต้ H_0
S(x) คือ ความน่าจะเป็นสะสมของตัวอย่าง

3.3. <u>เขตปฏิเสธหรือเขตวิกฤต</u> ถ้า $D \geq D_t$ ปฏิเสธ H_0 (โดยที่ D_t คือค่าวิกฤตจากตาราง Komogorov – Smirnov)

4. การประมาณค่าสูญหายโดยวิธี แทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย

วิธีการคือประมาณค่าสูญหายของข้อมูลโดยใช้ค่าเฉลี่ยของข้อมูลที่ไม่สูญหายแทนค่าข้อมูลสูญหาย สำหรับค่าสังเกตที่ i ซึ่งหาได้ดังนี้ (Schafer, 2005)

$$y_{i(miss)} = \frac{\sum_{j=1}^{p} r_{ij} y_{ij}}{\sum_{j=1}^{p} r_{ij}} \qquad \text{เมื่อ } i= 1,2,\ldots n;\ j= 1,2,\ldots,p$$

เมื่อ $y_{i(miss)}$ คือข้อมูลสูญหายของค่าสังเกตที่ i

y_{ij} คือข้อมูลค่าสังเกตที่ i ของการวัดซ้ำครั้งที่ j

$$r_{ij} = \begin{cases} 1 & \text{ถ้า } y_{ij} \text{ เป็นค่าสังเกตที่ไม่สูญหาย} \\ 0 & \text{ถ้า } y_{ij} \text{ เป็นค่าสังเกตที่สูญหาย} \end{cases}$$

5. การประมาณค่าสูญหายโดยวิธี MCMC

5.1. กำหนดจำนวนครั้งของการแทนค่าข้อมูลสูญหาย (m) เท่ากับ 5 เนื่องจากสัดส่วนการสูญหายของข้อมูลที่จำลองขึ้นสำหรับการศึกษานี้มีสัดส่วนการสูญหายน้อยกว่า 30% ควรใช้ m=3 หรือ 5 (Huang and Carriere, 2006)

5.2. สร้างชุดข้อมูลที่ได้จาการแทนค่าข้อมูลสูญหาย โดยแทนค่าข้อมูลสูญหาย ด้วยวิธี MCMC โดยการวนซ้ำ 2 ขั้นตอนจำนวน 100 รอบ ดังนี้

1) ขั้นตอน Imputation เพื่อสร้างค่าสูญหาย

2) ขั้นตอน Posterior เพื่อประมาณค่าเฉลี่ยและความแปรปรวน

สำหรับการสร้างชุดข้อมูลที่ได้จากการแทนค่าข้อมูลที่สูญหายด้วยวิธี MCMC ในโปรแกรม SAS ใช้ PROC MI

5.3. ก่อนเข้าสู่ขั้นตอนการแทนค่าข้อมูลสูญหายต้องตรวจสอบconvergence ใน MCMC โดยพิจารณาจาก การพล๊อตอนุกรมเวลา (Time-series Plot) และ ACF ถ้าพบว่าค่าเฉลี่ยและความแปรปรวนไม่คงที่ให้เพิ่มจำนวนการวนซ้ำของ ขั้นตอน Imputation และ ขั้นตอน Posterior ในกระบวนการ MCMC

5.4. เมื่อแทนค่าข้อมูลที่สูญหายแต่ละค่าจำนวน 5 ครั้งที่แตกต่างกัน ทำให้ได้ข้อมูลสมบูรณ์ 5 ชุด และได้ค่าประมาณของข้อมูลสูญหายแต่ละตำแหน่งจำนวน 5 ค่า

5.5. รวมการอนุมานจากการแทนค่าสูญหายหลายครั้งให้เป็นค่าเดียวโดยได้ค่าประมาณข้อมูลสูญหายจากการคำนวณในสมการ (24)

6. ประมาณค่าข้อมูลสูญหายโดยใช้วิธี Copulas

เมื่อจำลองการสูญหายของข้อมูลและตรวจสอบการแจกแจงปกติแล้วทำตามขั้นตอนต่อไปนี้

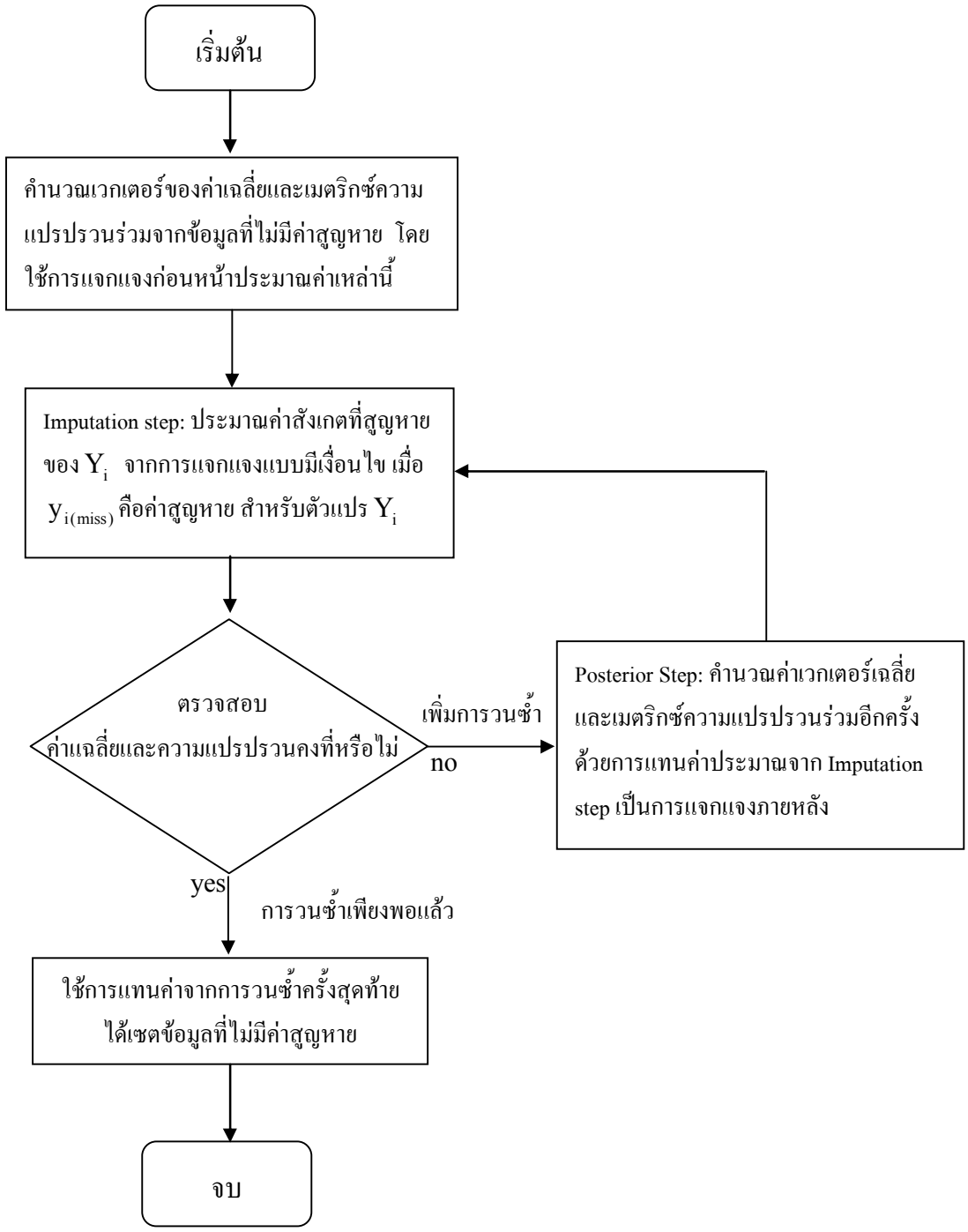
6.1. ประมาณค่าสูญหายในกรณีของ Compound Symmetry สำหรับเมตริกซ์โครงสร้างความสัมพันธ์แบบ Compound Symmetry จากสมการ (47)

6.2. ประมาณค่าสูญหายในกรณีของ Autoregressive dependencies สำหรับเมตริกซ์โครงสร้างความสัมพันธ์แบบAutoregressive จากสมการ (50)

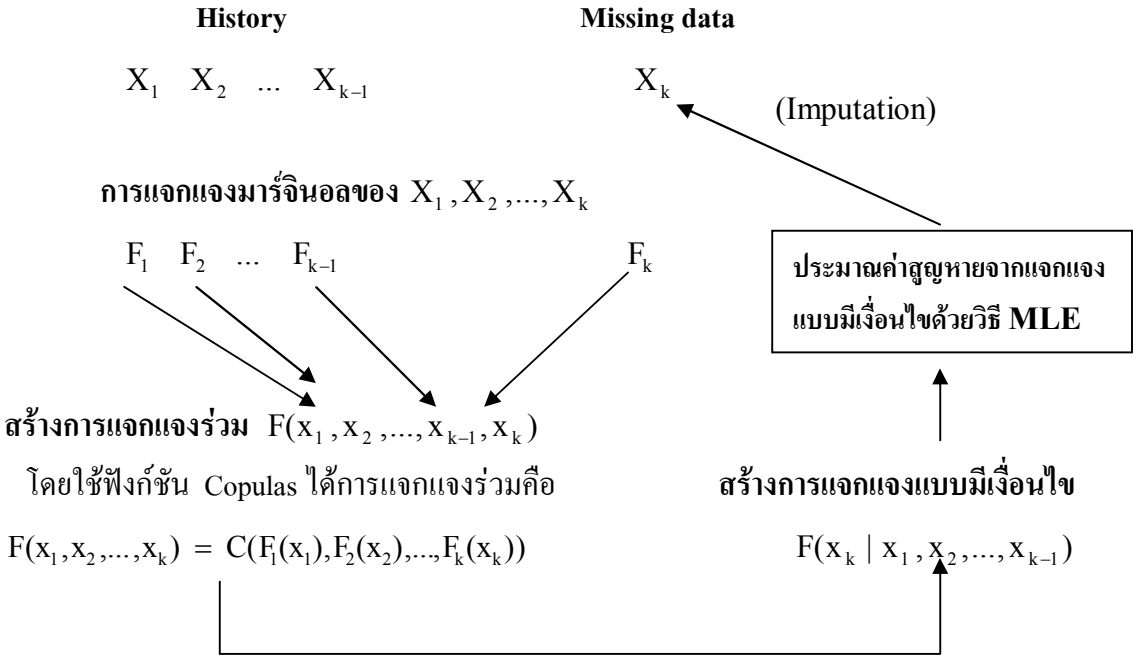
6.3. นำค่าประมาณข้อมูลสูญหายที่หาได้จากการคำนวณในข้อ 6.1 และ 6.2 แทนที่ในตำแหน่งที่มีข้อมูลสูญหาย

7. เปรียบเทียบประสิทธิภาพในการประมาณค่าสูญหายทั้งสามวิธี

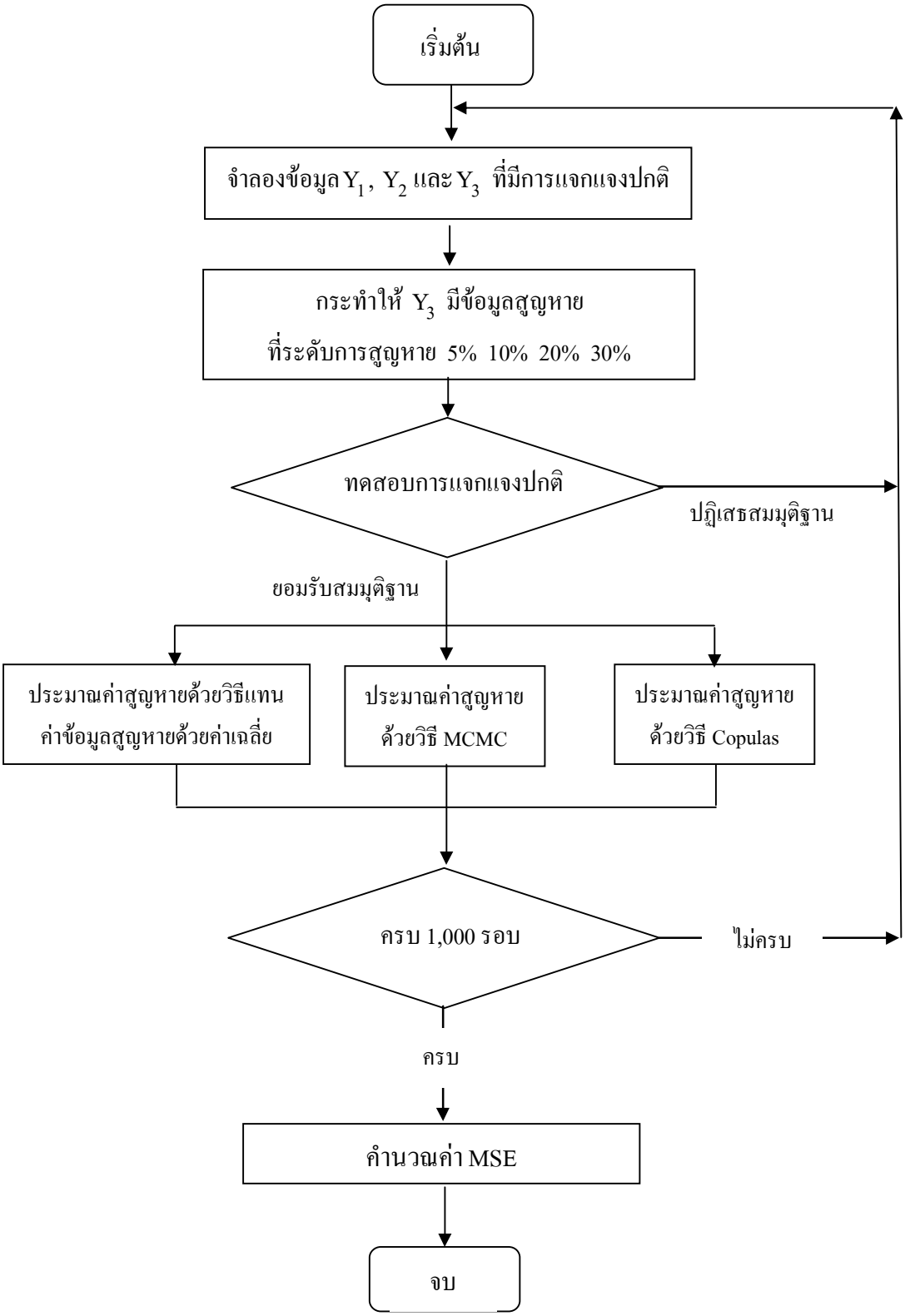
การเปรียบเทียบประสิทธิภาพโดยใช้การวัดค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (MSE) วิธีการประมาณค่าสูญหายที่ให้ค่า ค่าเฉลี่ยของค่า MSE ต่ำกว่า แสดงว่าวิธีการประมาณค่าสูญหายนั้นมีประสิทธิภาพมากกว่า โดยคำนวณจากสมการ (51)



ภาพที่ 8 แผนผังขั้นตอนการดำเนินงานของวิธี MCMC



ภาพที่ 9 แผนผังขั้นตอนการดำเนินงานของวิธี Copulas



ภาพที่ 10 ผังงานของขั้นตอนการดำเนินงานสำหรับข้อมูลที่ได้จากการจำลองสถานการณ์

ผลและวิจารณ์

ผล

การศึกษาวิธีประมาณค่าสูญหายของข้อมูล 3 วิธี คือ วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย วิธีMCMC และวิธี Copulas ได้ทำการศึกษากับข้อมูลที่ได้จากการจำลองสถานการณ์โดยเทคนิคมอนติคาร์โล ทำการจำลองซ้ำ 1,000 ครั้งในแต่ละสถานการณ์ จำนวน 72 สถานการณ์ และได้นำวิธีการประมาณค่าสูญหายทั้ง 3 วิธีมาประยุกต์ใช้กับข้อมูลจริง โดยผลการเปรียบเทียบประสิทธิภาพของทั้ง 3 วิธี แบ่งเป็น 2 กรณี คือ กรณีข้อมูลจากจำลองสถานการณ์โดยเทคนิคมอนติคาร์โล และ กรณีข้อมูลจริงที่นำมาประยุกต์ใช้กับวิธีการประมาณค่าสูญหายของข้อมูลทั้ง3 วิธี

1.ข้อมูลจากจำลองสถานการณ์โดยเทคนิคมอนติคาร์โล

ข้อมูลที่ได้จากจำลองสถานการณ์โดยเทคนิคมอนติคาร์โล กำหนดให้ตัวแปร Y_3 มีข้อมูลสูญหายที่ระดับการสูญหาย 5%, 10%, 20% และ 30% ของข้อมูล Y_3 เมื่อนำชุดข้อมูลที่มีการสูญหายในระดับต่าง ๆ มาทดสอบการแจกแจงปกติโดยใช้การทดสอบ Kolmogorov-Smirnov พบว่าเมื่อข้อมูลที่มีการสูญหายแบบสุ่ม จากการจำลอง 1,000 ครั้ง ในแต่ละสถานการณ์ ข้อมูลมีการแจกแจงปกติที่ระดับนัยสำคัญ 0.01 และเนื่องจากเมตริกซ์โครงสร้างความสัมพันธ์ของข้อมูลที่จำลองขึ้นมี 2 แบบ คือ 1) แบบ Compound Symmetry และ 2) แบบ Autoregressive ดังนั้นผลการเปรียบเทียบประสิทธิภาพของวิธีประมาณค่าข้อมูลสูญหายของทั้ง 3 วิธีโดยใช้ค่า MSE เป็นเกณฑ์ในการเปรียบเทียบแบ่งเป็น 2 กรณีคือ

1. กรณีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Compound Symmetry
2. กรณีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Autoregressive

1. 1. กรณีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Compound Symmetry

ผลการเปรียบเทียบประสิทธิภาพของวิธีประมาณค่าข้อมูลสูญหายของทั้ง 3 วิธีกรณีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Compound Symmetry แบ่งเป็น 3 ส่วนดังนี้

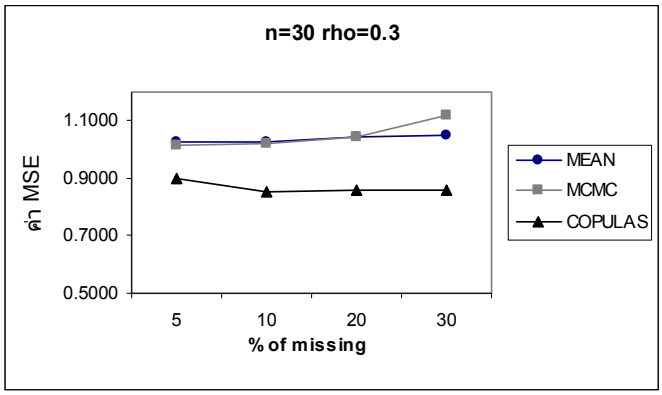
1.1.1. เมื่อค่าวัดซ้ำมีความสัมพันธ์ระดับต่ำ

เมื่อพิจารณาค่า MSE ของแต่ละวิธี ในการจำลองสถานการณ์ โดยกระทำซ้ำ 1,000 ครั้ง จากตารางที่ 2 และภาพที่ 11 ถ้าวิธีการประมาณค่าสูญหายวิธีใดให้ค่า MSE ต่ำสุด แสดงว่าวิธีนั้นมีประสิทธิภาพดีกว่า สำหรับกรณีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Compound Symmetry และความสัมพันธ์ระหว่างตัวแปร (ρ) เท่ากับ 0.3 สำหรับขนาดตัวอย่างเท่ากับ 30, 70 และ 100 ทุกระดับการสูญหายของข้อมูล พบว่าวิธี Copulas ให้ค่า MSE ต่ำที่สุด ดังนั้นการประมาณค่าสูญหายด้วยวิธี Copulas มีประสิทธิภาพดีกว่า วิธี MCMC และ วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย

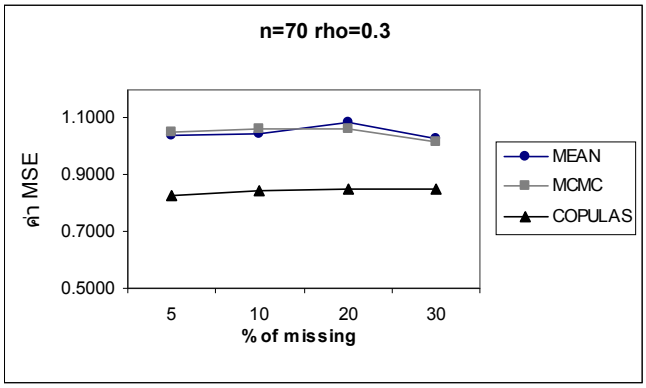
ตารางที่ 2 ค่า MSE ของวิธีประมาณค่าข้อมูลสูญหาย วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย วิธี MCMC และวิธี Copulas กรณีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Compound Symmetry เมื่อความสัมพันธ์ระหว่างค่าวัดซ้ำเท่ากับ 0.3 และขนาดตัวอย่างเท่ากับ 30, 70 และ 100 จำแนกตามเปอร์เซ็นต์การสูญหายของข้อมูลที่วัดซ้ำครั้งสุดท้าย และขนาดตัวอย่าง

ขนาดตัวอย่าง	% ของข้อมูลสูญหาย	ค่า MSE		
		Mean	MCMC	Copulas
30	5	1.0250	1.0171	0.8985*
	10	1.0240	1.0229	0.8536*
	20	1.0441	1.0440	0.8595*
	30	1.0500	1.1163	0.8561*
70	5	1.0399	1.0536	0.8298*
	10	1.0449	1.0620	0.8417*
	20	1.0872	1.0643	0.8484*
	30	1.0295	1.0137	0.8518*
100	5	1.0399	1.0079	0.8636*
	10	1.0210	1.0620	0.8648*
	20	1.0584	1.1036	0.8668*
	30	1.0558	1.0751	0.8638*

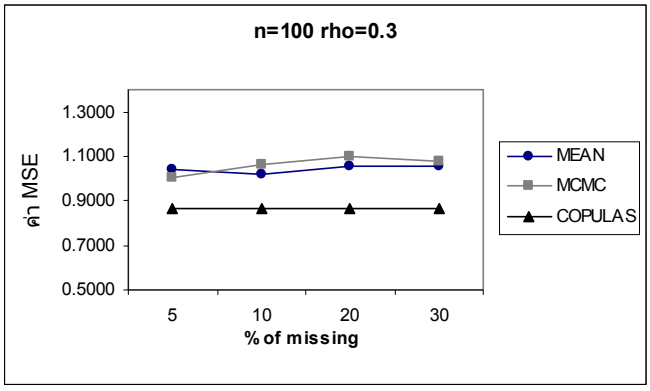
หมายเหตุ * ค่า MSE ที่มีค่าต่ำสุด



(ก) ขนาดตัวอย่างเท่ากับ 30



(ข) ขนาดตัวอย่างเท่ากับ 70



(ค) ขนาดตัวอย่างเท่ากับ 100

ภาพที่ 11 ค่า MSE ของวิธีประมาณค่าข้อมูลสูญหาย วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย วิธี MCMC และวิธี Copulas กรณีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Compound Symmetry เมื่อความสัมพันธ์ระหว่างค่าวัดซ้ำเท่ากับ 0.3 สำหรับขนาดตัวอย่างเท่ากับ 30, 70 และ 100 ที่ระดับการสูญหาย 5%, 10%, 20% และ 30% ของข้อมูลที่วัดซ้ำครั้งสุดท้าย

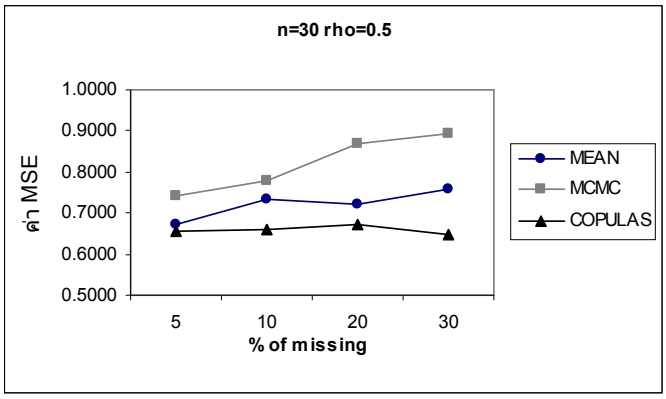
1.1.2. เมื่อค่าวัดซ้ำมีความสัมพันธ์ระดับปานกลาง

เมื่อพิจารณาค่า MSE ของแต่ละวิธี ในการจำลองสถานการณ์ โดยกระทำซ้ำ 1,000 ครั้ง จากตารางที่ 3 และภาพที่ 12 ถ้าวิธีการประมาณค่าสูญหายวิธีใดให้ค่า MSE ต่ำสุด แสดงว่าวิธีนั้นมีประสิทธิภาพดีกว่า สำหรับกรณีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Compound Symmetry และความสัมพันธ์ระหว่างค่าวัดซ้ำ (ρ) เท่ากับ 0.5 สำหรับขนาดตัวอย่างเท่ากับ 30, 70 และ 100 ทุกระดับการสูญหายของข้อมูล พบว่าวิธี Copulas ให้ค่า MSE ต่ำที่สุด ดังนั้นการประมาณค่าสูญหายด้วยวิธี Copulas มีประสิทธิภาพดีกว่า วิธี MCMC และ วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย

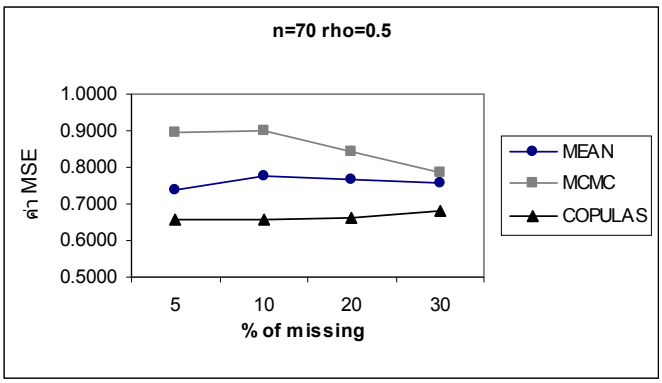
ตารางที่ 3 ค่า MSE ของวิธีประมาณค่าข้อมูลสูญหาย วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย วิธี MCMC และวิธี Copulas กรณีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Compound Symmetry เมื่อความสัมพันธ์ระหว่างค่าวัดซ้ำเท่ากับ 0.5 และขนาดตัวอย่างเท่ากับ 30, 70 และ 100 จำแนกตามเปอร์เซ็นต์การสูญหายของข้อมูลที่วัดซ้ำครั้งสุดท้าย และขนาดตัวอย่าง

ขนาดตัวอย่าง	% ของข้อมูลสูญหาย	ค่า MSE		
		Mean	MCMC	Copulas
30	5	0.6712	0.7434	0.6539*
	10	0.7347	0.7782	0.6610*
	20	0.7221	0.8705	0.6724*
	30	0.7591	0.8942	0.6493*
70	5	0.7386	0.8930	0.6587*
	10	0.7783	0.8977	0.6583*
	20	0.7646	0.8427	0.6612*
	30	0.7581	0.7869	0.6830*
100	5	0.8515	0.8882	0.6796*
	10	0.8223	0.8270	0.6760*
	20	0.7874	0.8703	0.6722*
	30	0.7770	0.8383	0.6717*

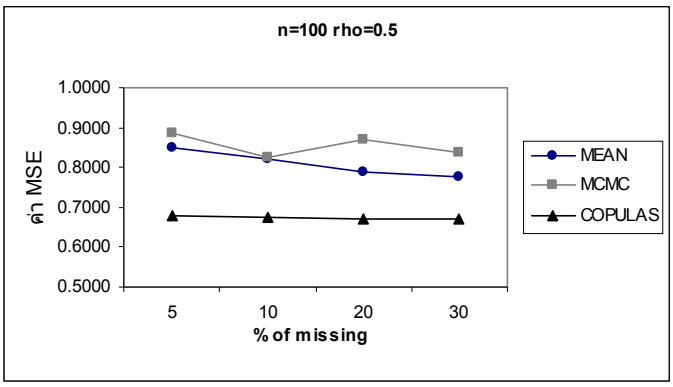
หมายเหตุ * ค่า MSE ที่มีค่าต่ำสุด



(ก) ขนาดตัวอย่างเท่ากับ 30



(ข) ขนาดตัวอย่างเท่ากับ 70



(ค) ขนาดตัวอย่างเท่ากับ 100

ภาพที่ 12 ค่า MSE ของวิธีประมาณค่าข้อมูลสูญหาย วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย วิธี MCMC และวิธี Copulas กรณีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Compound Symmetry เมื่อความสัมพันธ์ระหว่างค่าวัดซ้ำเท่ากับ 0.5 สำหรับขนาดตัวอย่างเท่ากับ 30, 70 และ 100 ที่ระดับการสูญหาย 5%, 10%, 20% และ 30% ของข้อมูลที่วัดซ้ำครั้งสุดท้าย

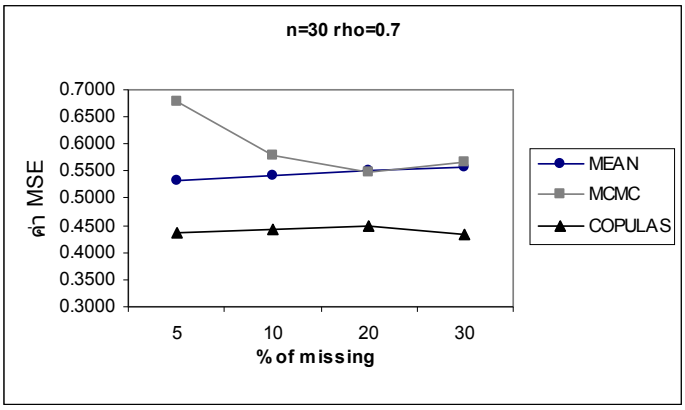
1.1.3. เมื่อค่าวัดซ้ำมีความสัมพันธ์ระดับสูง

เมื่อพิจารณาค่า MSE ของแต่ละวิธี ในการจำลองสถานการณ์ โดยกระทำซ้ำ 1,000 ครั้ง จากตารางที่ 4 และภาพที่ 13 สำหรับกรณีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Compound Symmetry และความสัมพันธ์ระหว่างค่าวัดซ้ำ (ρ) เท่ากับ 0.7 สำหรับขนาดตัวอย่าง เท่ากับ 30, 70 และ 100 ทุกระดับการสูญหายของข้อมูล พบว่าวิธี Copulas ให้ค่า MSE ต่ำที่สุด ดังนั้นการประมาณค่าสูญหายด้วยวิธี Copulas มีประสิทธิภาพดีกว่า วิธี MCMC และ วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย

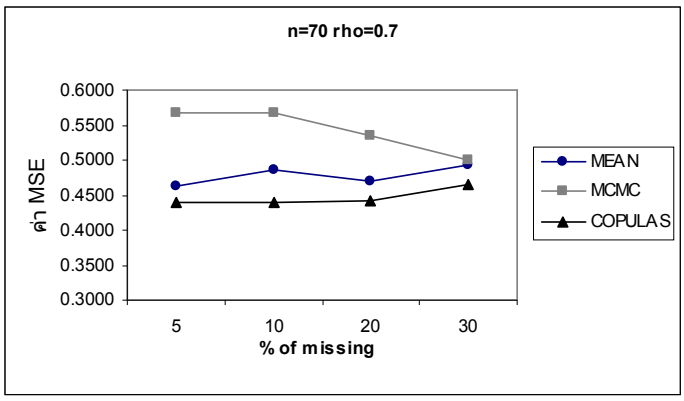
ตารางที่ 4 ค่า MSE ของวิธีประมาณค่าข้อมูลสูญหาย วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย วิธี MCMC และวิธี Copulas กรณีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Compound Symmetry เมื่อความสัมพันธ์ระหว่างค่าวัดซ้ำเท่ากับ 0.7 และขนาดตัวอย่างเท่ากับ 30, 70 และ 100 จำแนกตามเปอร์เซ็นต์การสูญหายของข้อมูลที่วัดซ้ำครั้งสุดท้าย และขนาดตัวอย่าง

ขนาดตัวอย่าง	% ของข้อมูลสูญหาย	ค่า MSE		
		Mean	MCMC	Copulas
30	5	0.5321	0.6773	0.4351*
	10	0.5431	0.5790	0.4436*
	20	0.5512	0.5494	0.4492*
	30	0.5578	0.5673	0.4343*
70	5	0.4631	0.5678	0.4399*
	10	0.4849	0.5679	0.4403*
	20	0.4690	0.5355	0.4418*
	30	0.4920	0.5008	0.4642*
100	5	0.5703	0.5663	0.4516*
	10	0.5349	0.5273	0.4494*
	20	0.4923	0.6480	0.4473*
	30	0.4806	0.5326	0.4473*

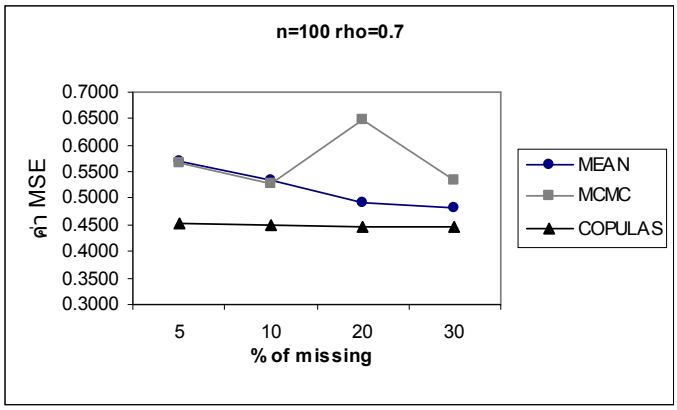
หมายเหตุ * ค่า MSE ที่มีค่าต่ำสุด



(ก) ขนาดตัวอย่างเท่ากับ 30



(ข) ขนาดตัวอย่างเท่ากับ 70



(ค) ขนาดตัวอย่างเท่ากับ 100

ภาพที่ 13 ค่า MSE ของวิธีประมาณค่าข้อมูลสูญหาย วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย วิธี MCMC และวิธี Copulas กรณีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Compound Symmetry เมื่อความสัมพันธ์ระหว่างค่าวัดซ้ำเท่ากับ 0.7 สำหรับขนาดตัวอย่างเท่ากับ 30, 70 และ 100 ที่ระดับการสูญหาย 5%, 10%, 20% และ 30% ของข้อมูลที่วัดซ้ำครั้งสุดท้าย

1. 2. กรณีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Autoregressive

ผลการเปรียบเทียบประสิทธิภาพของวิธีประมาณค่าข้อมูลสูญหายของวิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย วิธี MCMC และวิธี Copulas กรณีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Autoregressiveแบ่งเป็น 2 ส่วนดังนี้

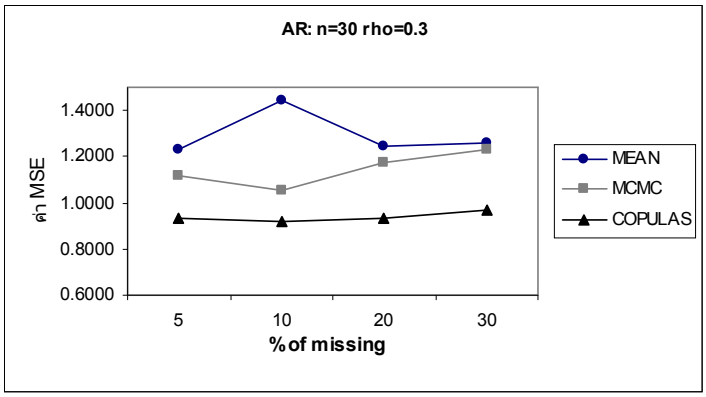
1.2.1. เมื่อค่าวัดซ้ำมีความสัมพันธ์ระดับต่ำ

เมื่อพิจารณาค่า MSE ของแต่ละวิธี ในการจำลองสถานการณ์ โดยกระทำซ้ำ 1,000 ครั้ง จากตารางที่ 5 และภาพที่ 14 สำหรับกรณีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Autoregressive และความสัมพันธ์ระหว่างค่าวัดซ้ำ (ρ) เท่ากับ 0.3 สำหรับขนาดตัวอย่างเท่ากับ 30, 70 และ 100 ทุกระดับการสูญหายของข้อมูล พบว่าวิธี Copulas ให้ค่า MSE ต่ำที่สุด ดังนั้นการประมาณค่าสูญหายด้วยวิธี Copulas มีประสิทธิภาพดีกว่า วิธี MCMC และ วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย

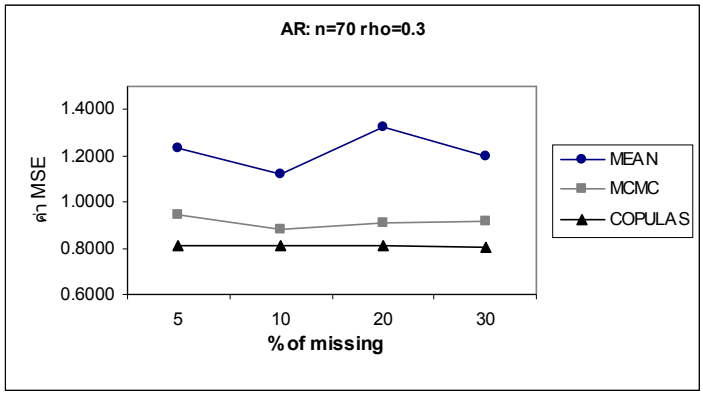
ตารางที่ 5 ค่า MSE ของวิธีประมาณค่าข้อมูลสูญหายวิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย วิธี MCMC และวิธี Copulas กรณีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Autoregressive เมื่อความสัมพันธ์ระหว่างค่าวัดซ้ำเท่ากับ 0.3 และขนาดตัวอย่างเท่ากับ 30, 70 และ 100 จำแนกตามเปอร์เซ็นต์การสูญหายของข้อมูลที่วัดซ้ำครั้งสุดท้าย และขนาดตัวอย่าง

ขนาดตัวอย่าง	% ของข้อมูลสูญหาย	ค่า MSE		
		Mean	MCMC	Copulas
30	5	1.2283	1.1165	0.9350*
	10	1.4438	1.0514	0.9220*
	20	1.2449	1.1752	0.9342*
	30	1.2571	1.2290	0.9667*
70	5	1.2321	0.9474	0.8125*
	10	1.1231	0.8816	0.8108*
	20	1.3218	0.9089	0.8091*
	30	1.1973	0.9196	0.8065*
100	5	1.2443	1.1152	0.9250*
	10	1.2579	1.1126	0.9314*
	20	1.2680	1.2100	0.9350*
	30	1.2633	1.1391	0.9342*

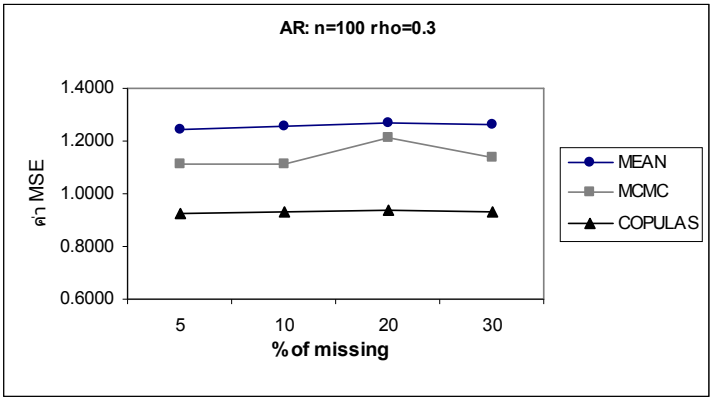
หมายเหตุ * ค่า MSE ที่มีค่าต่ำสุด



(ก) ขนาดตัวอย่างเท่ากับ 30



(ข) ขนาดตัวอย่างเท่ากับ 70



(ค) ขนาดตัวอย่างเท่ากับ 100

ภาพที่ 14 ค่า MSE ของวิธีประมาณค่าข้อมูลสูญหาย วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย วิธี MCMC และวิธี Copulas กรณีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Autoregressive เมื่อความสัมพันธ์ระหว่างค่าวัดซ้ำเท่ากับ 0.3 สำหรับขนาดตัวอย่างเท่ากับ 30, 70 และ 100 ที่ระดับการสูญหาย 5%, 10%, 20% และ 30% ของข้อมูลที่วัดซ้ำครั้งสุดท้าย

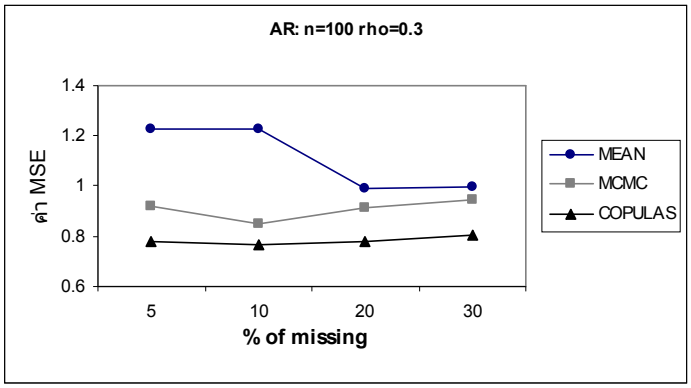
1.2.2. เมื่อค่าวัดซ้ำมีความสัมพันธ์ระดับปานกลาง

เมื่อพิจารณาค่า MSE ของแต่ละวิธี ในการจำลองสถานการณ์ โดยกระทำซ้ำ 1,000 ครั้ง จากตารางที่ 6 และภาพที่ 15 สำหรับกรณีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Autoregressive และความสัมพันธ์ระหว่างค่าวัดซ้ำ (ρ) เท่ากับ 0.5 สำหรับขนาดตัวอย่างเท่ากับ 30, 70 และ 100 ทุกระดับการสูญหายของข้อมูล พบว่าวิธี Copulas ให้ค่า MSE ต่ำที่สุด ดังนั้นการประมาณค่าสูญหายด้วยวิธี Copulas มีประสิทธิภาพดีกว่า วิธี MCMC และ วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย

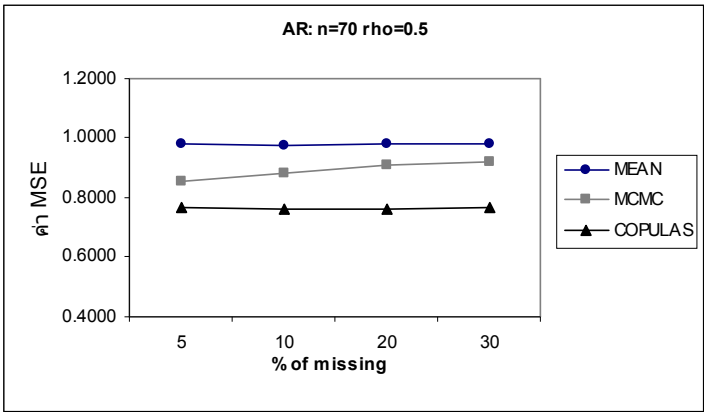
ตารางที่ 6 ค่า MSE ของวิธีประมาณค่าข้อมูลสูญหาย วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย วิธี MCMC และวิธี Copulas กรณีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Autoregressive เมื่อความสัมพันธ์ระหว่างค่าวัดซ้ำเท่ากับ 0.5 และขนาดตัวอย่างเท่ากับ 30, 70 และ 100 จำแนกตามเปอร์เซ็นต์การสูญหายของข้อมูลที่วัดซ้ำครั้งสุดท้าย และขนาดตัวอย่าง

ขนาดตัวอย่าง	% ของข้อมูลสูญหาย	ค่า MSE		
		Mean	MCMC	Copulas
30	5	1.2283	0.9196	0.7777*
	10	1.2251	0.8479	0.7643*
	20	0.9884	0.9110	0.7776*
	30	0.9974	0.9448	0.8049*
70	5	0.9805	0.8544	0.7651*
	10	0.9762	0.8816	0.7594*
	20	0.9790	0.9089	0.7638*
	30	0.9812	0.9196	0.7645*
100	5	0.9982	0.9187	0.7631*
	10	0.9976	0.9176	0.7685*
	20	0.9821	0.8942	0.7718*
	30	0.9781	0.8946	0.7715*

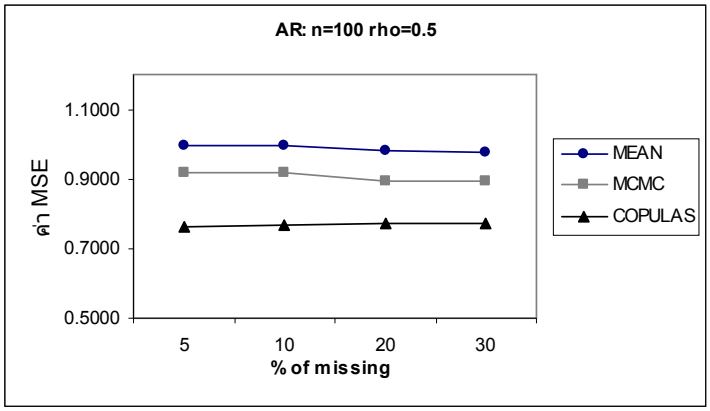
หมายเหตุ * ค่า MSE ที่มีค่าต่ำสุด



(ก) ขนาดตัวอย่างเท่ากับ 30



(ข) ขนาดตัวอย่างเท่ากับ 70



(ค) ขนาดตัวอย่างเท่ากับ 100

ภาพที่ 15 ค่า MSE ของวิธีประมาณค่าข้อมูลสูญหาย วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย วิธี MCMC และวิธี Copulas กรณีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Autoregressive เมื่อความสัมพันธ์ระหว่างค่าวัดซ้ำเท่ากับ 0.5 สำหรับขนาดตัวอย่างเท่ากับ 30, 70 และ 100 ที่ระดับการสูญหาย 5%, 10%, 20% และ 30% ของข้อมูลที่วัดซ้ำครั้งสุดท้าย

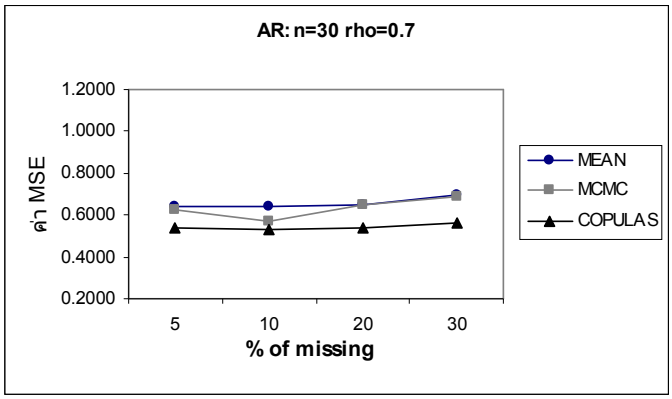
1.2.3. เมื่อค่าวัดซ้ำมีความสัมพันธ์ระดับสูง

เมื่อพิจารณาค่า MSE ของแต่ละวิธี ในการจำลองสถานการณ์ โดยกระทำซ้ำ 1,000 ครั้ง จากตารางที่ 7 และภาพที่ 16 สำหรับกรณีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Autoregressive และความสัมพันธ์ระหว่างค่าวัดซ้ำ (ρ) เท่ากับ 0.7 สำหรับขนาดตัวอย่างเท่ากับ 30, 70 และ 100 ทุกระดับการสูญหายของข้อมูล พบว่าวิธี Copulas ให้ค่า MSE ต่ำที่สุด ดังนั้นการประมาณค่าสูญหายด้วยวิธี Copulas มีประสิทธิภาพดีกว่า วิธี MCMC และ วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย

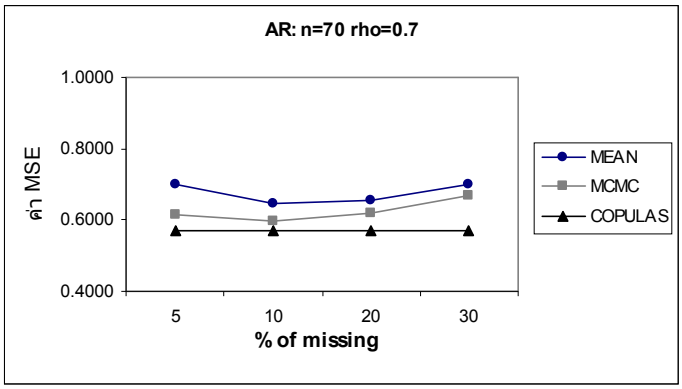
ตารางที่ 7 ค่า MSE ของวิธีประมาณค่าข้อมูลสูญหาย วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย วิธี MCMC และวิธี Copulas กรณีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Autoregressive เมื่อความสัมพันธ์ระหว่างค่าวัดซ้ำเท่ากับ 0.7 และขนาดตัวอย่างเท่ากับ 30, 70 และ 100 จำแนกตามเปอร์เซ็นต์การสูญหายของข้อมูลที่วัดซ้ำครั้งสุดท้าย และขนาดตัวอย่าง

ขนาดตัวอย่าง	% ของข้อมูลสูญหาย	ค่า MSE		
		Mean	MCMC	Copulas
30	5	0.6394	0.6253	0.5367*
	10	0.6417	0.5736	0.5268*
	20	0.6525	0.6451	0.5393*
	30	0.6972	0.6860	0.5600*
70	5	0.6983	0.6149	0.5698*
	10	0.6471	0.5960	0.5682*
	20	0.6571	0.6175	0.5685*
	30	0.6982	0.6671	0.5687*
100	5	0.6831	0.6243	0.5655*
	10	0.6580	0.6242	0.5682*
	20	0.6972	0.6757	0.5712*
	30	0.7123	0.6369	0.5722*

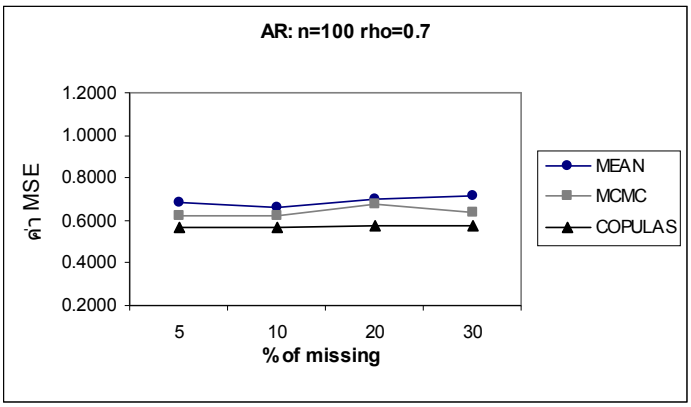
หมายเหตุ * ค่า MSE ที่มีค่าต่ำสุด



(ก) ขนาดตัวอย่างเท่ากับ 30



(ข) ขนาดตัวอย่างเท่ากับ 70



(ค) ขนาดตัวอย่างเท่ากับ 100

ภาพที่ 16 ค่า MSE ของวิธีประมาณค่าข้อมูลสูญหาย วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย วิธี MCMC และวิธี Copulas กรณีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Autoregressive เมื่อความสัมพันธ์ระหว่างค่าวัดซ้ำเท่ากับ 0.7 สำหรับขนาดตัวอย่างเท่ากับ 30, 70 และ 100 ที่ระดับการสูญหาย 5%, 10%, 20% และ 30% ของข้อมูลที่วัดซ้ำครั้งสุดท้าย

จากผลการศึกษาวิธีการประมาณค่าข้อมูลสูญหาย วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย วิธี MCMC และวิธี Copulas แสดงวิธีที่มีประสิทธิภาพดีที่สุดสำหรับแต่ละสถาการณ์ดังตารางที่ 8 ของข้อมูลที่ได้จากการจำลองสถานการณ์

ตารางที่ 8 วิธีที่มีประสิทธิภาพที่สุดในแต่ละสถานการณ์ของการประมาณค่าสูญหาย เมื่อความสัมพันธ์ระหว่างค่าวัดซ้ำเท่ากับ 0.3 , 0.5 และ 0.7

Correlation structure	% ของข้อมูลสูญหาย	ขนาดตัวอย่าง		
		30	70	100
Compound Symmetry	5	Copulas	Copulas	Copulas
	10	Copulas	Copulas	Copulas
	20	Copulas	Copulas	Copulas
	30	Copulas	Copulas	Copulas
Autoregressive	5	Copulas	Copulas	Copulas
	10	Copulas	Copulas	Copulas
	20	Copulas	Copulas	Copulas
	30	Copulas	Copulas	Copulas

2. ข้อมูลจริงที่นำมาประยุกต์ใช้กับวิธีการประมาณค่าสูญหายของข้อมูล

ข้อมูลจริงที่ใช้ในการศึกษาครั้งนี้มีข้อมูล 2 ชุด คือ 1) ข้อมูลรอบเอวของผู้เข้าร่วมโครงการ “นพรัตน์ไร้พุง มุ่งสุขภาพดี” ของโรงพยาบาลนพรัตนราชธานี และ 2) ข้อมูลปริมาณน้ำฝนรายเดือนของเดือน มิถุนายน – สิงหาคม 2550 ของสถานีอุตุนิยมวิทยาภาคเหนือจำนวน 28 สถานี ซึ่งได้ผลการศึกษาดังนี้

2.1. ข้อมูลรอบเอวของผู้เข้าร่วมโครงการ “นพรัตน์ไร้พุง มุ่งสุขภาพดี”

โครงการ “นพรัตน์ไร้พุง มุ่งสุขภาพดี” เป็นโครงการที่โรงพยาบาลนพรัตน์ราชธานีจัดทำขึ้น โดยมีวัตถุประสงค์ให้เจ้าหน้าที่ในโรงพยาบาลมีน้ำหนักและรอบพุงอยู่ในเกณฑ์มาตรฐาน จากการปฏิบัติตนอย่างถูกต้อง เพื่อลดอัตราการเกิดโรคเรื้อรัง และเพิ่มคุณภาพชีวิต มีผู้เข้าร่วมโครงการเป็นเจ้าหน้าที่ในโรงพยาบาลจำนวน 105 คน สำหรับผู้ที่อ้วนลงพุง หมายถึง รอบเอวผู้ชายเกิน 90 เซนติเมตรและผู้หญิงเกิน 80 เซนติเมตร

การดำเนินงานของโครงการมีการอบรมแนะนำ 2 เรื่องหลัก คือ โภชนาการ และ การออกกำลังกาย และมีการวัด น้ำหนัก ส่วนสูง รอบเอว ของผู้เข้าร่วมโครง จำนวน 4 ครั้งดังนี้

ครั้งที่ 1 วัดวันที่ 21 พฤศจิกายน 2551
ครั้งที่ 2 วัดวันที่ 22 ธันวาคม 2551
ครั้งที่ 3 วัดวันที่ 24 มกราคม 2552
ครั้งที่ 4 วัดวันที่ 3 มีนาคม 2552

สำหรับข้อมูลที่นำมาประยุกต์ใช้กับวิธีการประมาณค่าสูญหายของข้อมูล 3 วิธี คือ 1) ค่าเฉลี่ย 2) MCMC และ 3) Copulas เป็นข้อมูลรอบเอว (หน่วยเป็นเซนติเมตร) ของผู้ที่มาวัดตามกำหนดครบ 4 ครั้ง จำนวน44 คน ซึ่งเป็นข้อมูลที่สมบูรณ์ไม่มีข้อมูลสูญหายดังแสดงในตารางที่ 10

เนื่องจากวิธี Copulas ต้องทราบโครงสร้างของเมตริกซ์โครงสร้างความสัมพันธ์ของข้อมูลรอบเอวของผู้เข้าร่วมโครงการ “นพรัตน์ไร้พุง มุ่งสุขภาพดี” จากการวัด 4 ครั้ง จำนวน 44 คน ว่ามีเมตริกซ์โครงสร้างความสัมพันธ์แบบไหน ดังนั้นจึงหาเมตริกซ์โครงสร้างความสัมพันธ์ของข้อมูลรอบเอวได้ดังนี้

$$R_{CS} = \begin{bmatrix} 1 & 0.9026 & 0.8328 & 0.8434 \\ 0.9026 & 1 & 0.9207 & 0.8492 \\ 0.8328 & 0.9207 & 1 & 0.8815 \\ 0.8434 & 0.8492 & 0.8815 & 1 \end{bmatrix}$$

เมื่อพิจารณาเมตริกซ์โครงสร้างความสัมพันธ์พบว่าเมตริกซ์โครงสร้างความสัมพันธ์ R มีโครงสร้างแบบ Compound Symmetry ดังนั้นในวิธี Copulas จึงประมาณข้อมูลสูญหายได้จากสมการ (47) และ ได้กำหนดให้ $\rho_{34} = \rho$ เป็นค่าสหสัมพันธ์ของ Spearman ซึ่งมีความสัมพันธ์ระหว่างค่าวัดซ้ำระดับสูง

ผลการประมาณค่าสูญหายของข้อมูลสูญหายด้วยวิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย วิธี MCMC และวิธี Copulas สำหรับข้อมูลรอบเอวของผู้เข้าร่วมโครงการ “นพรัตน์ไร้พุง มุ่งสุขภาพดี” จากการวัดรอบเอว 4 ครั้ง โดยกำหนดให้ข้อมูลรอบเอวจากการวัดครั้งที่ 4 มีข้อมูลสูญหายเกิดขึ้นแบบสุ่ม ที่ระดับการสูญหายของข้อมูล 5% , 10%, 20% และ 30% ตามลำดับ และใช้ค่า MSE เป็นเกณฑ์ในการเปรียบเทียบประสิทธิภาพของวิธีประมาณค่าสูญหายทั้ง 3 วิธี โดยทำการวิเคราะห์ออกเป็น 3 ส่วน คือ

1. ทดสอบการแจกแจงปกติของข้อมูลรอบเอวเมื่อมีข้อมูลสูญหายเกิดขึ้น
2. ตรวจสอบการลู่เข้าในขั้นตอนคำนวณแบบวนซ้ำของ MCMC
3. ผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายของข้อมูล 3 วิธี คือ วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย MCMC และ Copulas

2.1.1. ทดสอบการแจกแจงปกติ

สมมติให้ข้อมูลรอบเอวที่ได้จากการวัดวัดครั้งที่ 4 มีข้อมูลสูญหายเกิดขึ้นแบบสุ่ม ที่ระดับการสูญหาย 5% , 10%, 20% และ 30% ตามลำดับ แล้วตรวจสอบการแจกแจงปกติของข้อมูลรอบเอว โดยใช้การทดสอบ Kolmogorov – Smirnov พบว่าข้อมูลมีการแจกแจงแบบปกติที่ระดับนัยสำคัญ 0.05 รายละเอียดแสดงดังตารางที่ 9

ตารางที่ 9 ค่า P-value ของการทดสอบ Kolmogorov – Smirnov จำแนกตามเปอร์เซ็นต์การสูญหายของข้อมูลที่วัดซ้ำครั้งที่ 4 และมีการวัดรอบเอวครั้งที่ 1 ถึงครั้งที่ 4

% ของข้อมูลสูญหาย	ค่า P-value โดยใช้ Kolmogorov – Smirnov test			
	วัดครั้งที่ 1	วัดครั้งที่ 2	วัดครั้งที่ 3	วัดครั้งที่ 4
5%	>0.1500	0.0743	>0.1500	0.0831
10%	>0.1500	0.0743	>0.1500	0.0774
20%	>0.1500	0.0743	>0.1500	>0.1500
30%	>0.1500	0.0743	>0.1500	>0.1500

ตารางที่ 10 ข้อมูลรอบเอวของผู้เข้าร่วมโครงการ “นพรัตน์ไร้พุง มุ่งสุขภาพดี” ของโรงพยาบาลนพรัตนราชธานี จากการวัดรอบเอว 4 ครั้ง

ลำดับที่	รอบเอว (cm) ของการวัด 4 ครั้ง			
	ครั้งที่ 1	ครั้งที่ 2	ครั้งที่ 3	ครั้งที่ 4
1	87	88	85	75
2	78	77	72	69
3	112	108	109	106
4	88	84	86	85
5	95	96	92	89
6	100	102	101	96
7	84	79	84	83
8	84	82	87	86
9	80	78	77	76
10	83	79	77	76
11	92.5	90	90	89
12	99	97	95	100
13	83	78	79	80
14	109	108	102	101
15	83	77	76	76

ตารางที่ 10 (ต่อ)

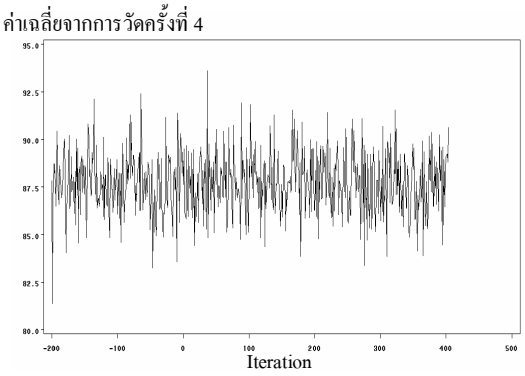
ลำดับที่	รอบเอว (cm) ของการวัด 4 ครั้ง			
	ครั้งที่ 1	ครั้งที่ 2	ครั้งที่ 3	ครั้งที่ 4
16	99	90	87	78
17	98	90	87	87
18	101	95	95	96
19	83	81	75	69
20	89	88	82	79
21	98	101	95	99
22	87	81	81	82
23	99	98	95	89
24	101	100	94	86
25	95	85	86	88
26	104	103	101	100
27	94	96	96	86
28	88	82	85	86
29	91	90	88	84
30	95	88	87	87
31	91	75	72	73
32	90	88	85	85
33	92	91	91	92
34	109	112	106	111
35	95	91	88	88
72	120	107	101	111
37	96	98	100	93
38	94	90	92	97
39	89	84	93	87
40	91	92	92	85

ตารางที่ 10 (ต่อ)

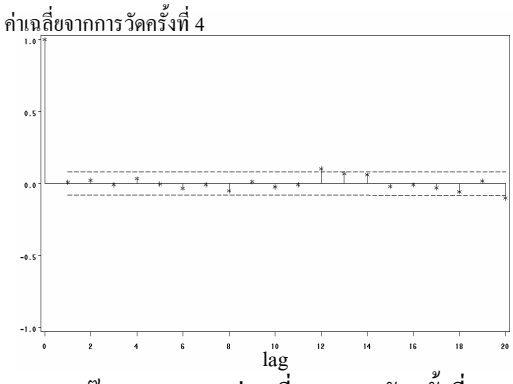
ลำดับที่	รอบเอว (cm) ของการวัด 4 ครั้ง			
	ครั้งที่ 1	ครั้งที่ 2	ครั้งที่ 3	ครั้งที่ 4
41	87	86	79	84
42	97	90	88	89
43	89	87	91	93
44	92	89	95	82

2.1.2. ตรวจสอบการลู่เข้าในขั้นตอนคำนวณแบบวนซ้ำของ MCMC

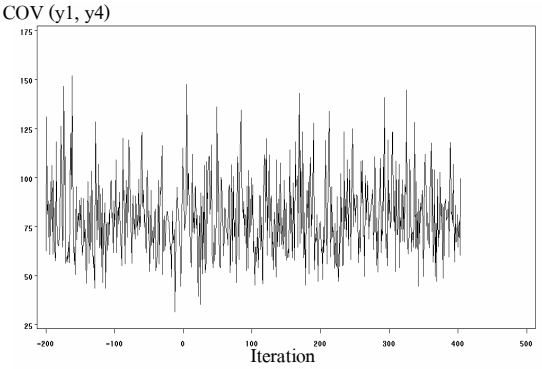
สำหรับการประมาณค่าสูญหายด้วยวิธี MCMC เมื่อประมาณค่าสูญหายได้แล้ว ต้องตรวจสอบว่าค่าเฉลี่ยและความแปรปรวนของค่าประมาณมีค่าคงที่หรือไม่ก่อนที่จะนำค่าที่ประมาณได้ไปแทนค่าสูญหายในแต่ละครั้ง การตรวจสอบทำโดยการพล๊อตอนุกรมเวลาและ ACF เมื่อทดลองการสูญหายของข้อมูลรอบเอวจากการวัดครั้งที่ 4 เป็นแบบสุ่มที่ระดับการสูญหายของข้อมูล 5% , 10%, 20% , และ 30% ตามลำดับ จากภาพที่ 17-20 พบว่าสำหรับทุกระดับการสูญหาย การพล๊อตอนุกรมเวลาและ ACF จากการวนซ้ำขั้นตอน I-step และ P-step 100 รอบ ได้ค่าเฉลี่ยและความแปรปรวนร่วมของค่าประมาณมีค่าคงที่ ดังนั้นสามารถนำค่าประมาณที่ได้จากวิธี MCMC แทนค่าสูญหายแต่ละครั้งได้



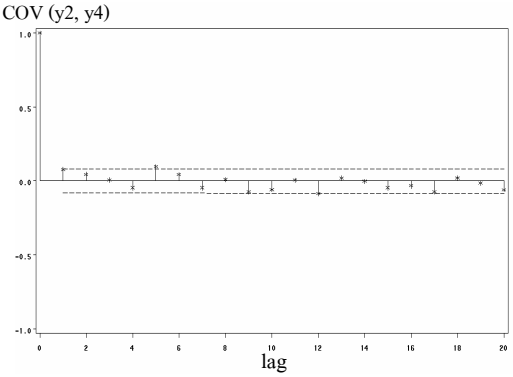
(ก) พล๊อตอนุกรมเวลาสำหรับการวนซ้ำของค่าเฉลี่ยจากการวัดครั้งที่4



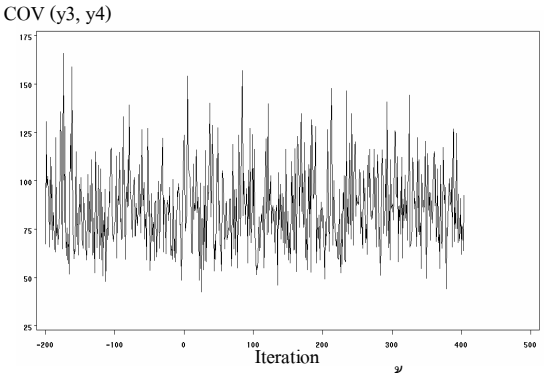
(ข) พล๊อต ACF ของค่าเฉลี่ยจากการวัดครั้งที่4



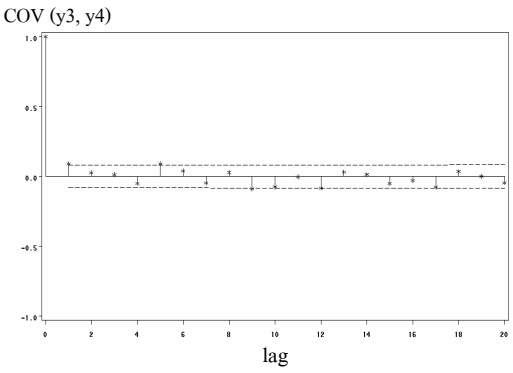
(ค) พล๊อตอนุกรมเวลาสำหรับการวนซ้ำของความแปรปรวนร่วมจากการวัดครั้งที่ 1 กับครั้งที่4



(ง) พล๊อต ACF ของความแปรปรวนร่วมจากการวัดครั้งที่ 2 กับครั้งที่4

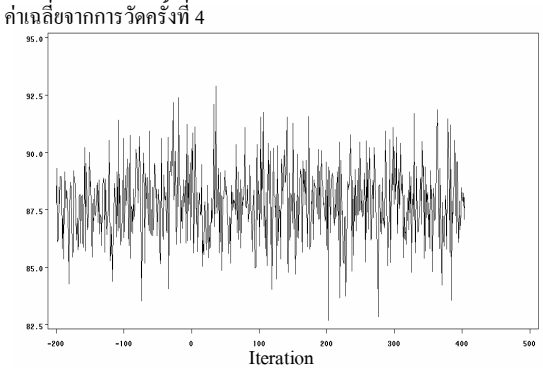


(จ) พล๊อตอนุกรมเวลาสำหรับการวนซ้ำของความแปรปรวนร่วมจากการวัดครั้งที่ 3 กับครั้งที่4

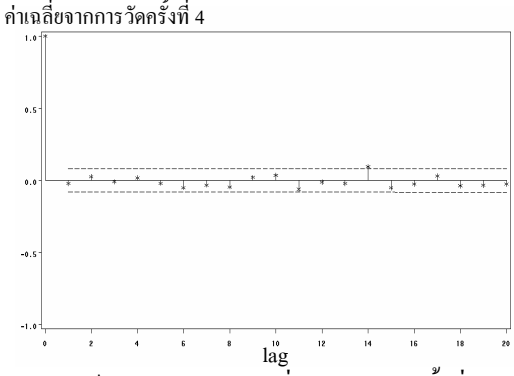


(ฉ) พล๊อต ACF ของความแปรปรวนร่วมจากการวัดครั้งที่ 3 กับครั้งที่4

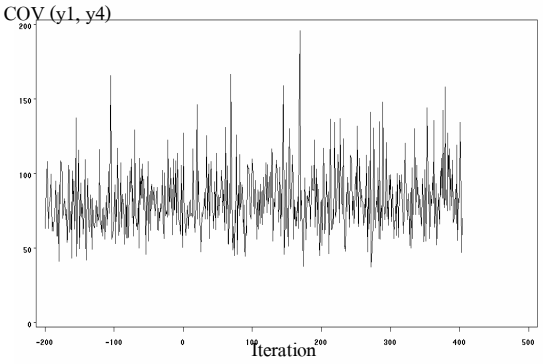
ภาพที่ 17 พล๊อตอนุกรมเวลาสำหรับการวนซ้ำและ ACF ของค่าเฉลี่ยและความแปรปรวนร่วมของข้อมูลรอบเอวของผู้เข้าร่วมโครงการ “นพรัตน์ไร้พุง มุ่งสุขภาพดี” ที่ได้จากการประมาณค่าสูญหายด้วยวิธี MCMC จากการวนซ้ำ I-step และ P-step จำนวน 100 รอบ ที่ระดับการสูญหาย 5% ของข้อมูลที่วัดครั้งที่ 4



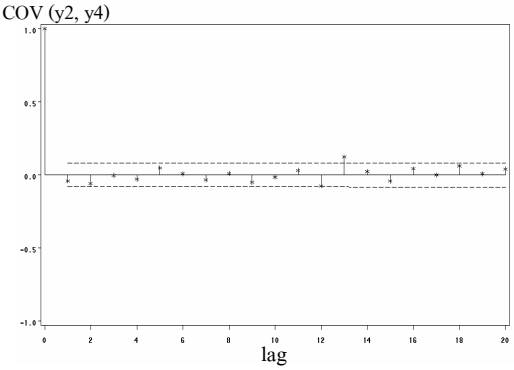
(ก) พล๊อตอนุกรมเวลาสำหรับการวนซ้ำของค่าเฉลี่ยจากการวัดครั้งที่4



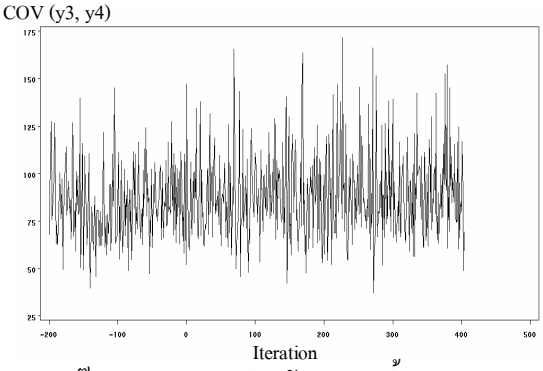
(ข) พล๊อต ACF ของค่าเฉลี่ยจากการวัดครั้งที่4



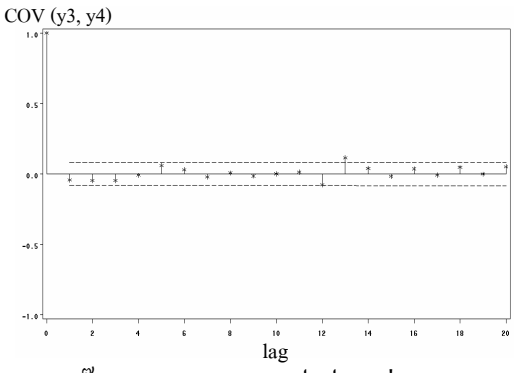
(ค) พล๊อตอนุกรมเวลาสำหรับการวนซ้ำของความแปรปรวนร่วมจากการวัดครั้งที่ 1 กับครั้งที่4



(ง) พล๊อต ACF ของความแปรปรวนร่วมจากการวัดครั้งที่ 2 กับครั้งที่4

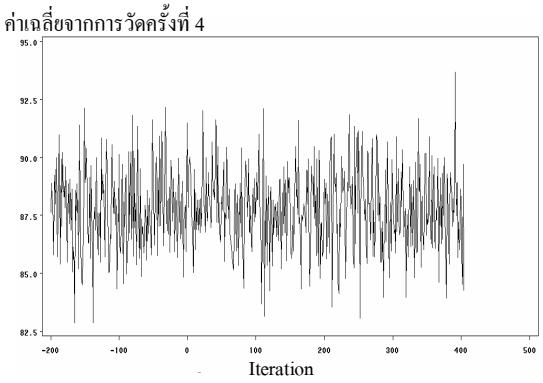


(จ) พล๊อตอนุกรมเวลาสำหรับการวนซ้ำของความแปรปรวนร่วมจากการวัดครั้งที่ 3 กับครั้งที่4

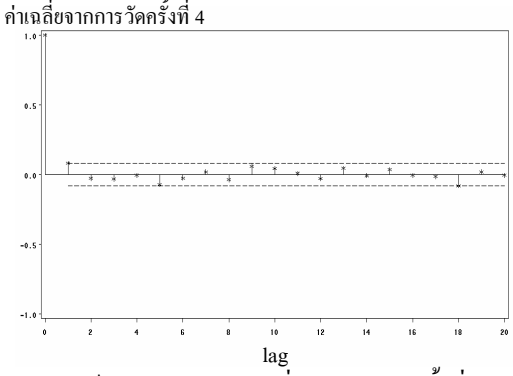


(ฉ) พล๊อต ACF ของความแปรปรวนร่วมจากการวัดครั้งที่ 3 กับครั้งที่4

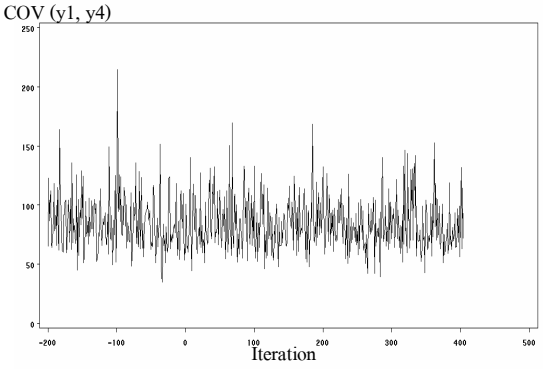
ภาพที่ 18 พล๊อตอนุกรมเวลาสำหรับการวนซ้ำและ ACF ของค่าเฉลี่ยและความแปรปรวนร่วมของข้อมูลรอบเอวของผู้เข้าร่วมโครงการ “นพรัตน์ไร้พุง มุ่งสุขภาพดี” ที่ได้จากการประมาณค่าสูญหายด้วยวิธี MCMC จากการวนซ้ำ I-step และ P-step จำนวน 100 รอบ ที่ระดับการสูญหาย 10% ของข้อมูลที่วัดครั้งที่ 4



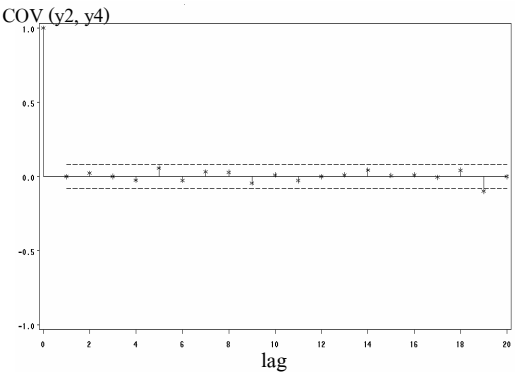
(ก) พล๊อตอนุกรมเวลาสำหรับการวนซ้ำของค่าเฉลี่ยจากการวัดครั้งที่4



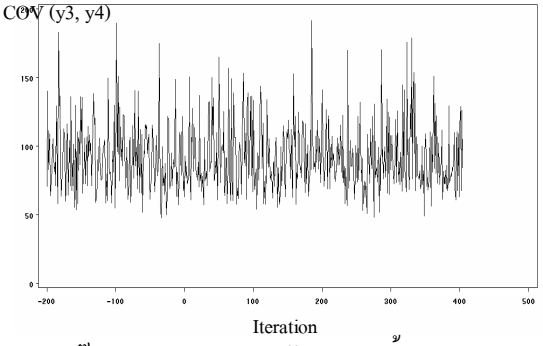
(ข) พล๊อต ACF ของค่าเฉลี่ยจากการวัดครั้งที่4



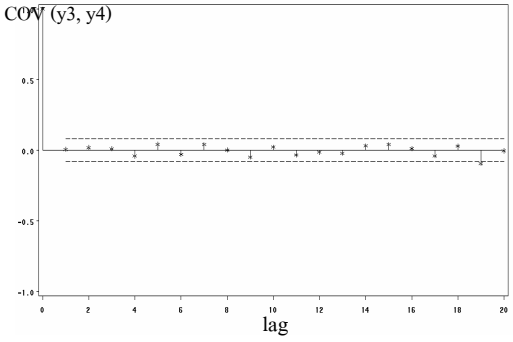
(ค) พล๊อตอนุกรมเวลาสำหรับการวนซ้ำของความแปรปรวนร่วมจากการวัดครั้งที่ 1 กับครั้งที่4



(ง) พล๊อต ACF ของความแปรปรวนร่วมจากการวัดครั้งที่ 2 กับครั้งที่4

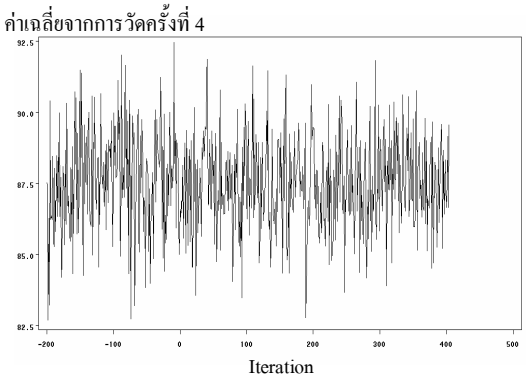


(จ) พล๊อตอนุกรมเวลาสำหรับการวนซ้ำของความแปรปรวนร่วมจากการวัดครั้งที่ 3 กับครั้งที่4

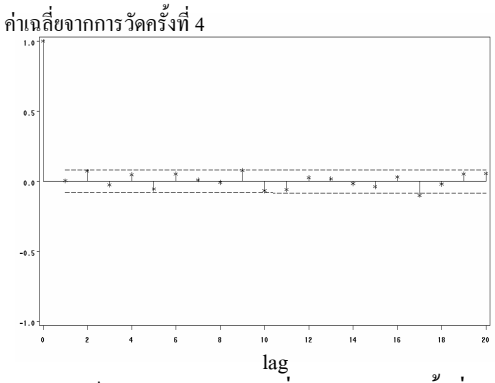


(ฉ) พล๊อต ACF ของความแปรปรวนร่วมจากการวัดครั้งที่ 3 กับครั้งที่4

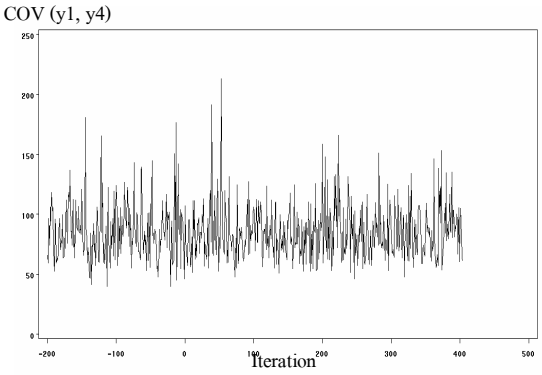
ภาพที่ 19 พล๊อตอนุกรมเวลาสำหรับการวนซ้ำและ ACF ของค่าเฉลี่ยและความแปรปรวนร่วมของข้อมูลรอบเอวของผู้เข้าร่วมโครงการ “นพรัตน์ไร้พุง มุ่งสุขภาพดี” ที่ได้จากการประมาณค่าสูญหายด้วยวิธี MCMC จากการวนซ้ำ I-step และ P-step จำนวน 100 รอบ ที่ระดับการสูญหาย 20% ของข้อมูลที่วัดครั้งที่ 4



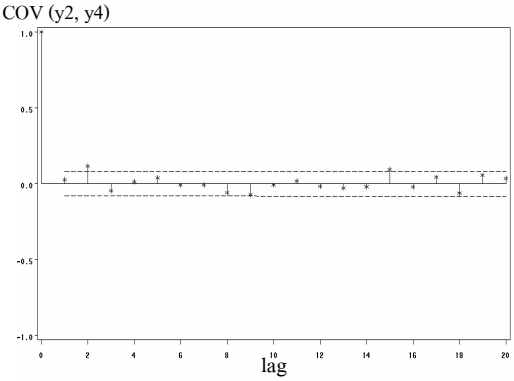
(ก) พล๊อตอนุกรมเวลาสำหรับการวนซ้ำของค่าเฉลี่ยจากการวัดครั้งที่4



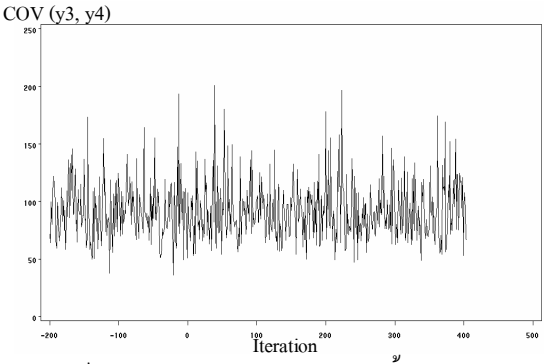
(ข) พล๊อต ACF ของค่าเฉลี่ยจากการวัดครั้งที่4



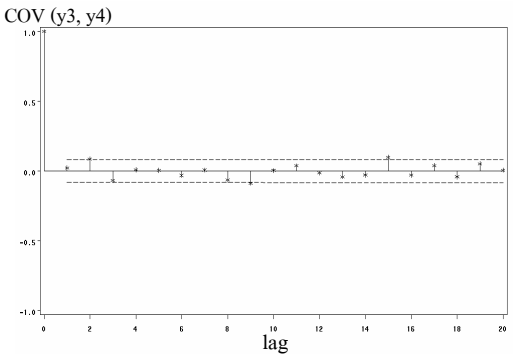
(ค) พล๊อตอนุกรมเวลาสำหรับการวนซ้ำของความแปรปรวนร่วมจากการวัดครั้งที่ 1 กับครั้งที่4



(ง) พล๊อต ACF ของความแปรปรวนร่วมจากการวัดครั้งที่ 2 กับครั้งที่4



(จ) พล๊อตอนุกรมเวลาสำหรับการวนซ้ำของความแปรปรวนร่วมจากการวัดครั้งที่ 1 กับครั้งที่4



(ฉ) พล๊อต ACF ของความแปรปรวนร่วมจากการวัดครั้งที่ 3 กับครั้งที่4

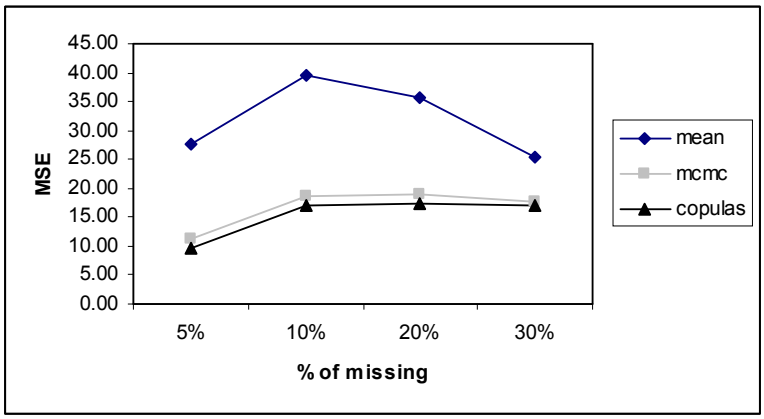
ภาพที่ 20 พล๊อตอนุกรมเวลาสำหรับการวนซ้ำและACF ของค่าเฉลี่ยและความแปรปรวนร่วมของข้อมูลรอบเอวของผู้เข้าร่วมโครงการ “นพรัตน์ไร้พุง มุ่งสุขภาพดี” ที่ได้จากการประมาณค่าสูญหายด้วยวิธี MCMC จากการวนซ้ำ I-step และ P-step จำนวน 100 รอบ ที่ระดับการสูญหาย 30% ของข้อมูลที่วัดครั้งที่ 4

2.1.3. เปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายของข้อมูล 3 วิธี คือ วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย วิธี MCMC และวิธี Copulas

เมื่อข้อมูลรอบเอวจากการวัดครั้ง 4 มีการสูญหายของข้อมูลเกิดขึ้นแบบสุ่มที่ระดับการสูญหายของข้อมูล5% , 10%, 20% และ 30% ตามลำดับ จากการประมาณค่าสูญหายข้อมูลรอบเอวของผู้เข้าร่วมโครงการ “นพรัตน์ไร้พุง มุ่งสุขภาพดี” จากการวัดรอบเอว 4 ครั้ง ด้วยวิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย วิธี MCMC และ วิธี Copulas พบว่า วิธี Copulas มีประสิทธิภาพดีกว่าวิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย และ วิธี MCMC โดยพิจารณาจากค่า MSE ดังแสดงในตารางที่ 11 และภาพที่ 21 จะเห็นว่า MSE ของวิธี Copulas มีค่าต่ำสุดในทุกระดับการสูญหาย

ตารางที่ 11 ค่า MSE ของวิธีประมาณค่าข้อมูลสูญหาย วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย วิธี MCMC และวิธี Copulas สำหรับข้อมูลรอบเอวของผู้เข้าร่วมโครงการ “นพรัตน์ไร้พุง มุ่งสุขภาพดี” จากการวัดรอบเอว 4 ครั้ง จำแนกตามเปอร์เซ็นต์การสูญหายของข้อมูลที่มีการวัดซ้ำครั้งที่ 4

% ของข้อมูลสูญหาย	ค่า MSE		
	mean	MCMC	Copulas
5%	27.78	11.11	9.80
10%	39.62	18.66	17.06
20%	35.65	18.96	17.44
30%	25.52	17.68	16.89



ภาพที่ 21 ค่า MSE ของวิธีประมาณค่าข้อมูลรอบเอวของผู้เข้าร่วมโครงการ “นพรัตน์ไร้พุง มุ่งสุขภาพดี”

2.2. ข้อมูลปริมาณน้ำฝนรายเดือนของเดือน มิถุนายน – สิงหาคม 2550

ข้อมูลที่นำมาศึกษาชุดที่สอง เป็นข้อมูลปริมาณน้ำฝนรายเดือนระหว่างเดือน มิถุนายน – สิงหาคม 2550 ของสถานีอุตุนิยมวิทยาภาคเหนือจำนวน 28 สถานี ดังแสดงในตารางที่ 13 ซึ่งเป็นข้อมูลที่สมบูรณ์ ทดลองให้ข้อมูลเดือนสิงหาคมมีข้อมูลสูญหายเกิดขึ้นแบบสุ่ม ที่ระดับการสูญหายของข้อมูล 5% , 10%, 20% และ 30% ตามลำดับ

เมื่อพิจารณาเมตริกซ์โครงสร้างความสัมพันธ์ R ของข้อมูลปริมาณน้ำฝนรายเดือนจากข้างล่างนี้ พบว่าข้อมูลมีรูปแบบโครงสร้างความสัมพันธ์แบบ Autoregresive ดังในวิธี Copulas จึงประมาณข้อมูลสูญหายจากสมการ (50) และให้ $\rho_{23} = \rho$ ซึ่งมีความสัมพันธ์ระหว่างค่าวัดซ้ำระดับปานกลาง

$$R_{AR} = \begin{bmatrix} 1 & 0.3566 & 0.189 \\ 0.3566 & 1 & 0.452 \\ 0.189 & 0.452 & 1 \end{bmatrix}$$

ผลการประมาณค่าข้อมูลสูญหายของวิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย วิธี MCMC และวิธี Copulas โดยใช้ค่า MSE เป็นเกณฑ์ในการเปรียบเทียบประสิทธิภาพของทั้งสามวิธี โดยทำการวิเคราะห์ออกเป็น 3 ส่วน คือ

1. ทดสอบการแจกแจงปกติของข้อมูลปริมาณน้ำฝนเมื่อมีข้อมูลสูญหายเกิดขึ้น
2. ตรวจสอบการลู่เข้าในขั้นตอนคำนวณแบบวนซ้ำของ MCMC
3. ผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายของข้อมูล 3 วิธี คือ วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย MCMC และ Copulas

เพื่อให้ง่ายต่อการสรุปผลการศึกษาจึงเขียนแทนข้อมูลปริมาณน้ำฝนเดือนมิถุนายน – สิงหาคม ด้วยตัวแปรต่อไปนี้

1. Y_1 : ข้อมูลปริมาณน้ำฝนเดือนมิถุนายน
2. Y_2 : ข้อมูลปริมาณน้ำฝนเดือนกรกฎาคม
3. Y_3 : ข้อมูลปริมาณน้ำฝนเดือนสิงหาคม

2.2.1. ทดสอบการแจกแจงปกติ

เมื่อกำหนดให้ข้อมูลปริมาณน้ำฝนของเดือนสิงหาคมมีข้อมูลสูญหายเกิดขึ้นแบบสุ่ม ที่ระดับการสูญหาย 5% , 10%, 20% และ 30% ตามลำดับ แล้วตรวจสอบการแจกแจงปกติ โดยใช้ การทดสอบ Kolmogorov – Smirnov พบว่าข้อมูลมีการแจกแจงปกติ ที่ระดับนัยสำคัญ 0.05 รายละเอียดแสดงดังตารางที่ 12

ตารางที่ 12 ค่า P-value ของ การทดสอบ Kolmogorov – Smirnov จำแนกตามเปอร์เซ็นต์การสูญหายของข้อมูลเดือนสิงหาคมและเดือน มิถุนายน – สิงหาคม

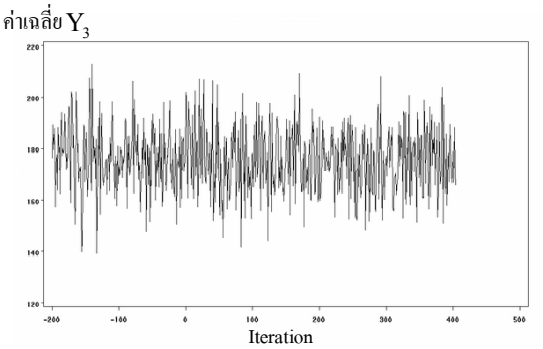
% ของข้อมูลสูญหาย	ค่า P-value โดยใช้การทดสอบ Kolmogorov – Smirnov		
	Y_1	Y_2	Y_3
5%	0.0811	0.0961	>0.1500
10%	0.0811	0.0961	>0.1500
20%	0.0811	0.0961	>0.1500
30%	0.0811	0.0961	>0.1500

ตารางที่ 13 ข้อมูลปริมาณน้ำฝนรายเดือนของเดือน เมษายน – มิถุนายน 2550 ของสถานีกรมอุตุนิยมวิทยาภาคเหนือ 28 สถานี ข้อมูลจากกรมอุตุนิยมวิทยา

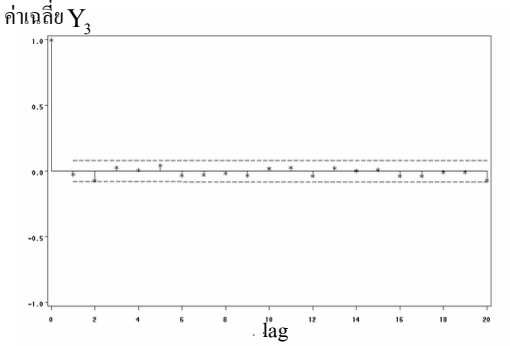
ลำดับที่	สถานี	เดือน		
		มิถุนายน	กรกฎาคม	สิงหาคม
1	Mae Hong Son	49	236.1	280.4
2	Mae Sariang	31.9	275.1	186.4
3	Chiang Rai	213.8	342	280.8
4	Chiang Rai Agromet	131.5	246.4	189.3
5	Phayao	195.4	215.1	128.4
6	Mae Jo	82.9	275.2	112.8
7	Chiang Mai	56	393.5	130.1
8	Lampang	85.2	332.7	80.4
9	Lampang Agromet	31.5	298.7	151.4
10	Lamphun	27	289.7	139.9
11	Phrae	78.9	240.5	122
12	Nan	114.9	208	131.2
13	Nan Agromet	64.9	171.5	216.9
14	Tha Wang Pha	96.9	214	160.1
15	Thung Chang	153.1	175	157
16	Uttaradit	25	363.9	244.8
17	Sukhothai	119.1	322.7	205.5
18	Si Samrong Agromet	99.5	364	129.5
19	Tak	22.3	422	115.9
20	Mae Sot	0.7	352.6	185.6
21	Bhumibol Dam	24.2	475.9	115.4
22	Doi Muser Agromet Stn.	18.6	277.1	118.7
23	Umphang	61.7	221.4	143.1
24	Phitsanulok	118.9	343	117.7
25	Phetchabun	109.4	233.1	89.4
26	Lom Sak	46.3	204	109.2
27	Wichian Buri	213.8	187.2	50.3
28	Kamphaeng Phet	85.6	300.4	223.5

2.2.2. ตรวจสอบการลู่เข้าในขั้นตอนคำนวณแบบวนซ้ำของ MCMC

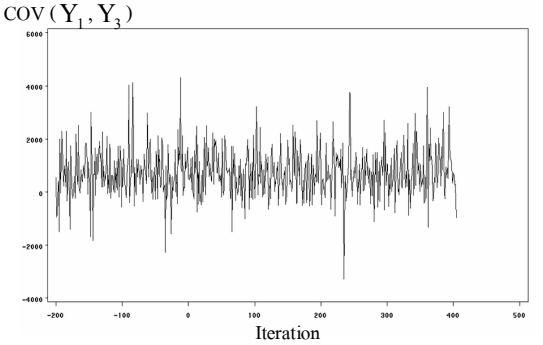
สำหรับการประมาณค่าสูญหายด้วยวิธี MCMC เมื่อประมาณค่าสูญหายได้แล้ว ต้องตรวจสอบว่าค่าเฉลี่ยและความแปรปรวนของค่าประมาณมีค่าคงที่หรือไม่ก่อนที่จะนำค่าที่ประมาณได้ไปแทนค่าสูญหายในแต่ละครั้ง การตรวจสอบทำโดยการพล๊อตอนุกรมเวลาและACF เมื่อ Y_3 มีข้อมูลสูญหายเกิดขึ้นแบบสุ่มที่ระดับการสูญหาย 5% , 10%, 20% , และ 30% ตามลำดับ จาก ภาพที่ 22-25 พบว่าสำหรับทุกระดับการสูญหาย การพล๊อตอนุกรมเวลาและACF จากการวนซ้ำขั้นตอน I-step และ P-step 100 รอบ ได้ค่าเฉลี่ยและความแปรปรวนร่วมของค่าประมาณมีค่าคงที่ ดังนั้นสามารถนำค่าประมาณที่ได้จากวิธี MCMC แทนค่าสูญหายแต่ละครั้งได้



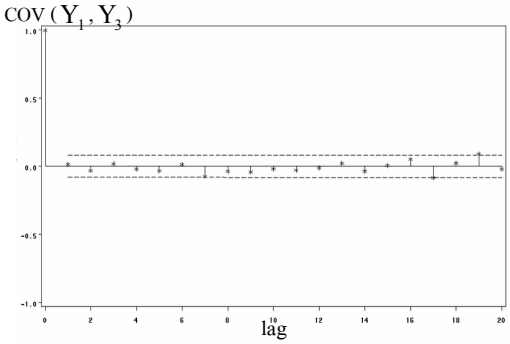
(ก) พล๊อตอนุกรมเวลาสำหรับการวนซ้ำของค่าเฉลี่ยเดือนสิงหาคม



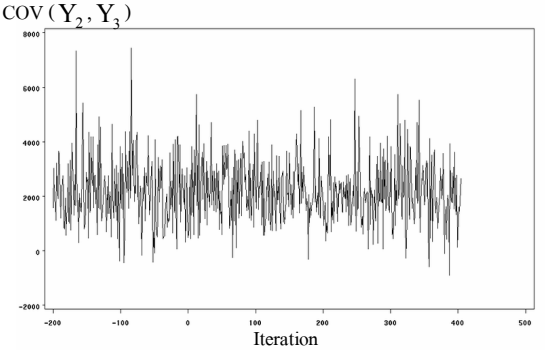
(ข) พล๊อต ACF ของค่าเฉลี่ยเดือนสิงหาคม



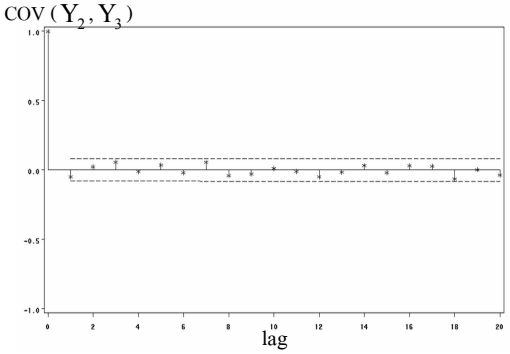
(ค) พล๊อตอนุกรมเวลาสำหรับการวนซ้ำของความแปรปรวนร่วมเดือนมิถุนายนกับเดือนสิงหาคม



(ง) พล๊อต ACF ของความแปรปรวนร่วมเดือนมิถุนายนกับเดือนสิงหาคม

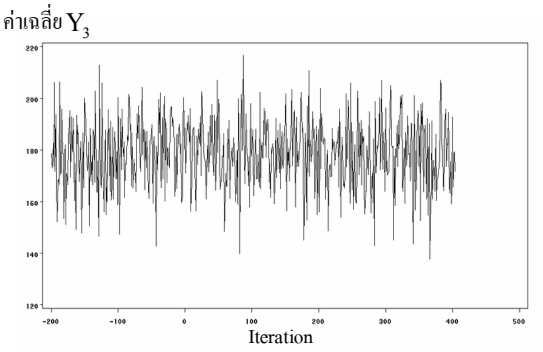


(จ) พล๊อตอนุกรมเวลาสำหรับการวนซ้ำของความแปรปรวนร่วมเดือนกรกฎาคมกับเดือนสิงหาคม

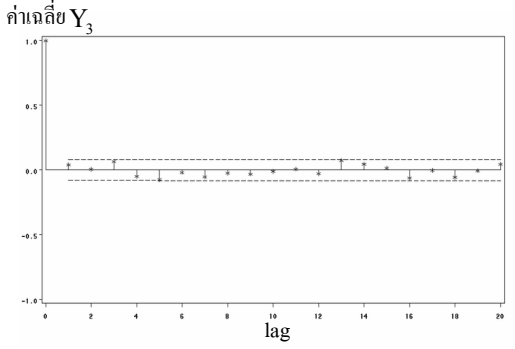


(ฉ) พล๊อต ACF ของความแปรปรวนร่วมเดือนกรกฎาคมกับเดือนสิงหาคม

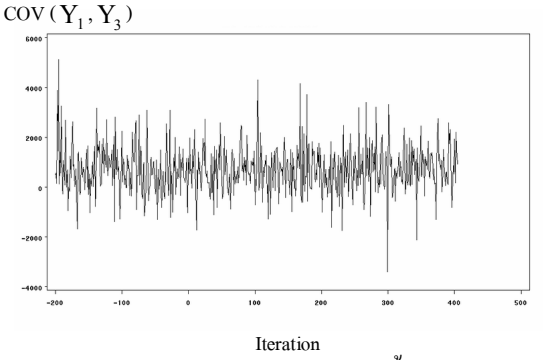
ภาพที่ 22 พล๊อตอนุกรมเวลาสำหรับการวนซ้ำและACF ของค่าเฉลี่ยและความแปรปรวนร่วมของข้อมูลปริมาณน้ำฝนรายเดือนของเดือน มิถุนายน – สิงหาคม 2550ที่ได้จากการประมาณค่าสูญหายด้วยวิธี MCMC จากการวนซ้ำ I-step และ P-step จำนวน 100 รอบ ที่ระดับการสูญหาย 5% ของข้อมูลเดือนสิงหาคม



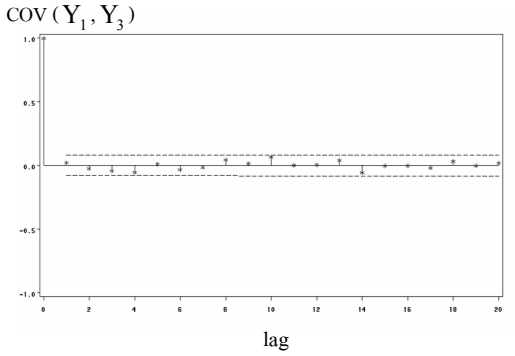
(ก) พล๊อตอนุกรมเวลาสำหรับการวนซ้ำของค่าเฉลี่ยเดือนสิงหาคม



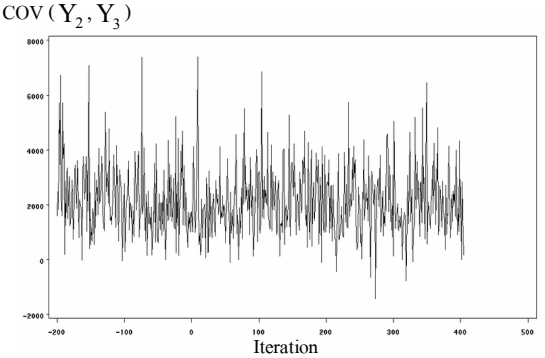
(ข) พล๊อต ACF ของค่าเฉลี่ยเดือนสิงหาคม



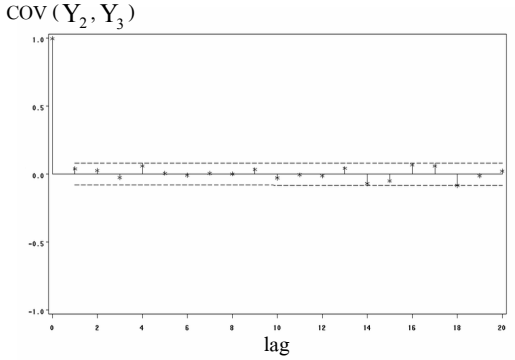
(ค) พล๊อตอนุกรมเวลาสำหรับการวนซ้ำของความแปรปรวนร่วมเดือนมิถุนายนกับเดือนสิงหาคม



(ง) พล๊อต ACF ของความแปรปรวนร่วมเดือนมิถุนายนกับเดือนสิงหาคม

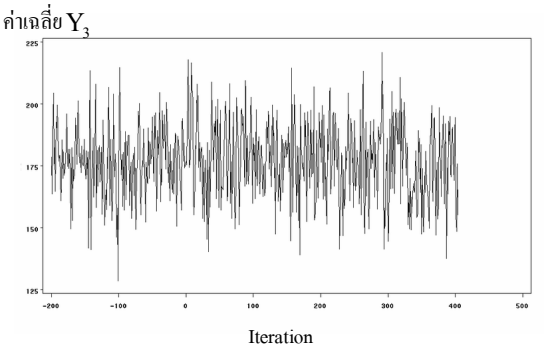


(จ) พล๊อตอนุกรมเวลาสำหรับการวนซ้ำของความแปรปรวนร่วมเดือนกรกฎาคมกับเดือนสิงหาคม

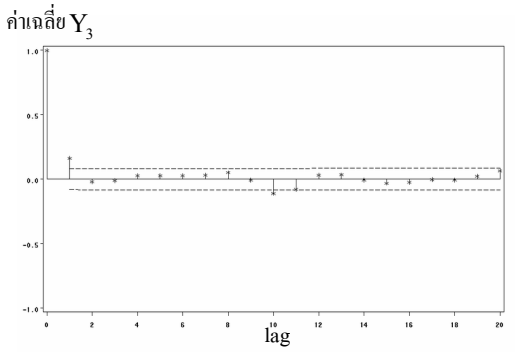


(ฉ) พล๊อต ACF ของความแปรปรวนร่วมเดือนกรกฎาคมกับเดือนสิงหาคม

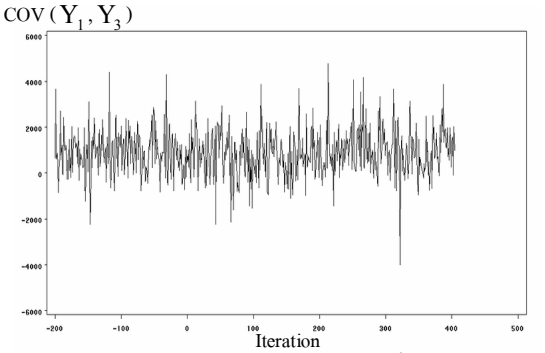
ภาพที่ 23 พล๊อตอนุกรมเวลาสำหรับการวนซ้ำและACF ของค่าเฉลี่ยและความแปรปรวนร่วมของข้อมูลปริมาณน้ำฝนรายเดือนของเดือน มิถุนายน – สิงหาคม 2550 ที่ได้จากการประมาณค่าสูญหายด้วยวิธี MCMC จากการวนซ้ำ I-step และ P-step จำนวน 100 รอบ ที่ระดับการสูญหาย 10% ของข้อมูลเดือนสิงหาคม



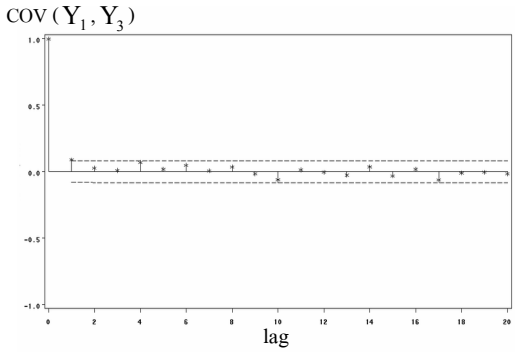
(ก) พล๊อตอนุกรมเวลาสำหรับการวนซ้ำของค่าเฉลี่ยเดือนสิงหาคม



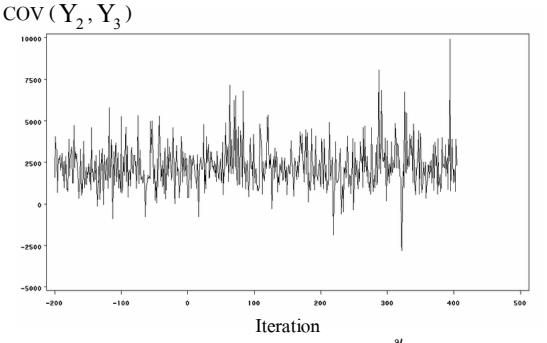
(ข) พล๊อต ACF ของค่าเฉลี่ยเดือนสิงหาคม



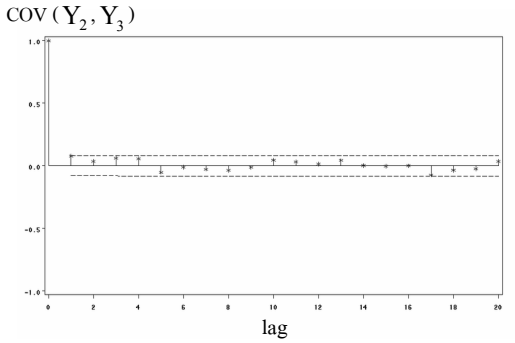
(ค) พล๊อตอนุกรมเวลาสำหรับการวนซ้ำของความแปรปรวนร่วมเดือนมิถุนายนกับเดือนสิงหาคม



(ง) พล๊อต ACF ของความแปรปรวนร่วมเดือนมิถุนายนกับเดือนสิงหาคม

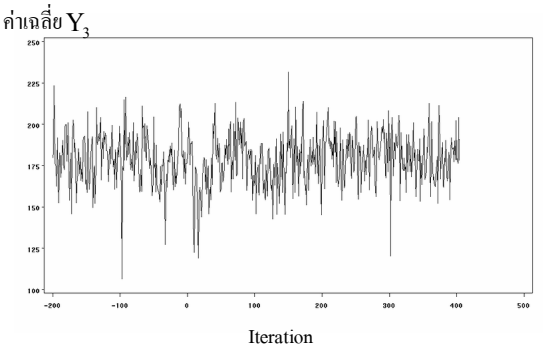


(จ) พล๊อตอนุกรมเวลาสำหรับการวนซ้ำของความแปรปรวนร่วมเดือนกรกฎาคมกับเดือนสิงหาคม

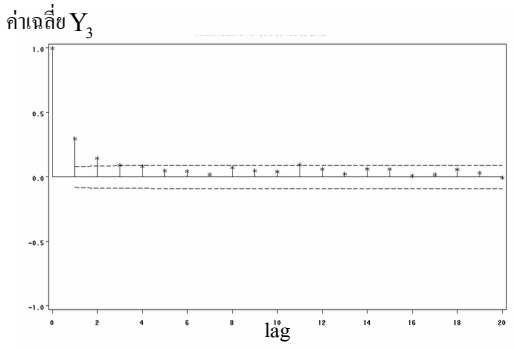


(ฉ) พล๊อต ACF ของความแปรปรวนร่วมเดือนกรกฎาคมกับเดือนสิงหาคม

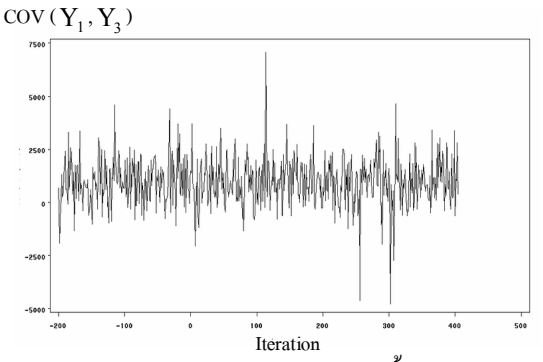
ภาพที่ 24 พล๊อตอนุกรมเวลาสำหรับการวนซ้ำและACF ของค่าเฉลี่ยและความแปรปรวนร่วมของข้อมูลปริมาณน้ำฝนรายเดือนของเดือน มิถุนายน – สิงหาคม 2550 ที่ได้จากการประมาณค่าสูญหายด้วยวิธี MCMC จากการวนซ้ำ I-step และ P-step จำนวน 100 รอบ ที่ระดับการสูญหาย 20% ของข้อมูลเดือนสิงหาคม



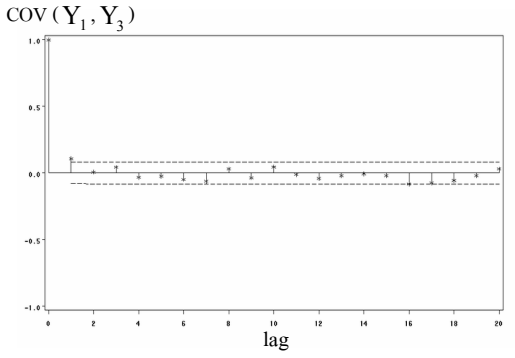
(ก) พล๊อตอนุกรมเวลาสำหรับการวนซ้ำของค่าเฉลี่ยเดือนสิงหาคม



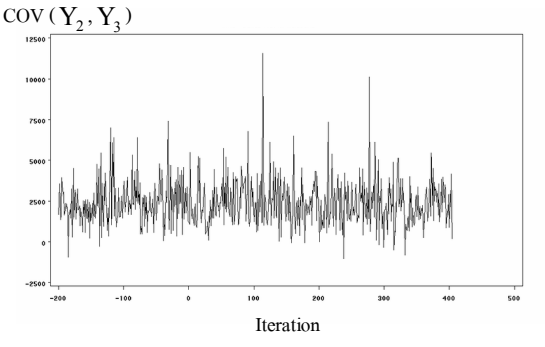
(ข) พล๊อต ACF ของค่าเฉลี่ยเดือนสิงหาคม



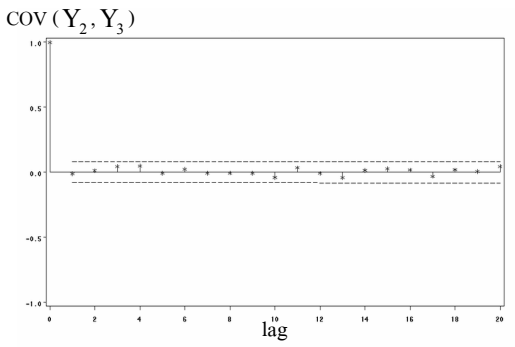
(ค) พล๊อตอนุกรมเวลาสำหรับการวนซ้ำของความแปรปรวนร่วมเดือนมิถุนายนกับเดือนสิงหาคม



(ง) พล๊อต ACF ของความแปรปรวนร่วมเดือนมิถุนายนกับเดือนสิงหาคม



(จ) พล๊อตอนุกรมเวลาสำหรับการวนซ้ำของความแปรปรวนร่วมเดือนกรกฎาคมกับเดือนสิงหาคม



(ฉ) พล๊อต ACF ของความแปรปรวนร่วมเดือนกรกฎาคมกับเดือนสิงหาคม

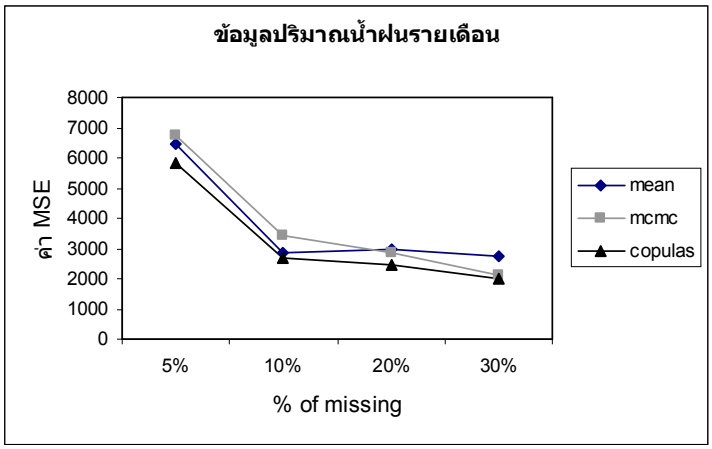
ภาพที่ 25 พล๊อตอนุกรมเวลาสำหรับการวนซ้ำและACF ของค่าเฉลี่ยและความแปรปรวนร่วมของข้อมูลปริมาณน้ำฝนรายเดือนของเดือน มิถุนายน – สิงหาคม 2550 ที่ได้จากการประมาณค่าสูญหายด้วยวิธี MCMC จากการวนซ้ำ I-step และ P-step จำนวน 100 รอบ ที่ระดับการสูญหาย 30% ของข้อมูลเดือนสิงหาคม

3. เปรียบเทียบวิธีการประมาณค่าสูญหายของข้อมูล 3 วิธี คือ วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย วิธี MCMC และวิธี Copulas

เมื่อข้อมูลปริมาณน้ำฝนเดือนสิงหาคมมีการสูญหายของข้อมูลเกิดขึ้นแบบสุ่มที่ระดับการสูญหาย 5% , 10%, 20% และ 30% ตามลำดับ จากการประมาณค่าสูญหายของข้อมูลปริมาณน้ำฝนรายเดือนของเดือนมิถุนายน – สิงหาคม 2550 โดยวิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย วิธี MCMC และ วิธี Copulas จากตาราง 14 และภาพ 26 เมื่อพิจารณาค่า MSE ที่น้อยที่สุด พบว่า วิธี Copulas มีประสิทธิภาพดีกว่าวิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย และ วิธี MCMC

ตารางที่ 14 ค่า MSE ของวิธีประมาณค่าข้อมูลสูญหาย วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย วิธี MCMC และวิธี Copulas สำหรับข้อมูลปริมาณน้ำฝนรายเดือนของเดือน มิถุนายน – สิงหาคม 2550 สถานีกรมอุตุนิยมวิทยาภาคเหนือ จำแนกตามเปอร์เซ็นต์การสูญหายของข้อมูลเดือนสิงหาคม

% ของข้อมูลสูญหาย	ค่า MSE		
	mean	MCMC	Copulas
5%	6451.4	6751.297	5856.309
10%	2830.72	3428.021	2671.946
20%	2987.15	2856.64	2439.963
30%	2739.919	2106.738	2027.227



ภาพที่ 26 ค่า MSE ของวิธีประมาณค่าข้อมูลสูญหาย วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย วิธี MCMC และวิธี Copulas สำหรับข้อมูลปริมาณน้ำฝนรายเดือนของเดือน มิถุนายน – สิงหาคม 2550 ของสถานีกรมอุตุนิยมวิทยาภาคเหนือ

วิจารณ์

ผลการศึกษาข้อมูลวัดซ้ำที่ได้จากการจำลองด้วยเทคนิคมอนติคาร์โล กรณีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Compound Symmetry และแบบ Autoregressive เมื่อขนาดตัวอย่างเท่ากับ 30, 70 และ 100 ความสัมพันธ์ระหว่างค่าวัดซ้ำเท่ากับ 0.3, 0.5 และ 0.7 สำหรับทุกระดับการสูญหายของข้อมูล พบว่าวิธีประมาณค่าสูญหายด้วยวิธี Copulas มีประสิทธิภาพดีกว่า วิธีแทนด้วยค่าเฉลี่ยและวิธี MCMC ซึ่งประสิทธิภาพของการสูญหายขึ้นอยู่กับสิ่งเหล่านี้ คือ

1. เปอร์เซ็นต์การสูญหายของข้อมูล พบว่าในขนาดตัวอย่างเดียวกันเมื่อเปอร์เซ็นต์การสูญหายของข้อมูลเพิ่มขึ้น วิธีการประมาณค่าสูญหายของ MCMC และ วิธี Copulas ให้ค่า MSE และ MAD แตกต่างกันมากขึ้น นั่นคือ ที่ระดับการสูญหาย 20% และ 30% ทั้งสองวิธีจะมีประสิทธิภาพแตกต่างกันมากขึ้น ยกเว้นกรณีข้อมูลวัดซ้ำที่มีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Compound Symmetry สำหรับขนาดตัวอย่างเท่ากับ 70 พบว่าเมื่อระดับการสูญหายของข้อมูลเพิ่มขึ้นวิธี MCMC และวิธี Copulas ให้ค่า MSE ใกล้เคียงกันมากขึ้น นั่นคือเมื่อระดับการสูญหายเพิ่มขึ้น ทั้งสองวิธีจะมีประสิทธิภาพไกล้เคียงกันมากขึ้น

2. ระดับความสัมพันธ์ของค่าวัดซ้ำ พบว่าในขนาดตัวอย่างและระดับการสูญหายของข้อมูลเดียวกัน เมื่อระดับความสัมพันธ์ระหว่างค่าวัดซ้ำเพิ่มขึ้น วิธีแทนด้วยค่าเฉลี่ย, วิธีMCMC และวิธี Copulas ให้ค่า MSE ลดลง นั่นคือทั้ง 3 วิธีมีประสิทธิภาพการประมาณค่าสูญหายเพิ่มขึ้น เมื่อระดับความสัมพันธ์ระหว่างค่าวัดซ้ำเพิ่มขึ้น

3. เมื่อระดับความสัมพันธ์ระหว่างค่าวัดซ้ำเพิ่มขึ้นวิธี MCMC และ วิธี Copulas มีค่า MSE ใกล้เคียงกันมากขึ้น นั้นคือ เมื่อระดับความสัมพันธ์ระหว่างค่าวัดซ้ำเพิ่มขึ้น ทั้ง 3 วิธีจะมีประสิทธิภาพใกล้เคียงกันมากขึ้น

สรุปและข้อเสนอแนะ

สรุป

การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของการประมาณค่าข้อมูลที่สูญหาย 3 วิธี คือ วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย วิธีMCMC และวิธี Copulas พร้อมทั้งได้นำข้อมูลจริง จำนวน 2 ชุด มาประยุกต์ใช้กับวิธีการประมาณค่าสูญหายทั้ง 3 วิธี โดยผลการศึกษาสรุปได้ดังนี้

1. ข้อมูลที่ได้จากการจำลอง

ผลการเปรียบเทียบประสิทธิภาพของการประมาณค่าข้อมูลที่สูญหายวิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย วิธี MCMC และวิธี Copulas ได้แบ่งผลสรุปออกเป็น 2 กรณีดังนี้

1.1. กรณีข้อมูลวัดซ้ำที่มีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Compound Symmetry เมื่อขนาดตัวอย่างเท่ากับ 30, 70 และ 100 ความสัมพันธ์ระหว่างค่าวัดซ้ำเท่ากับ 0.3, 0.5 และ 0.7 ที่ระดับการสูญหาย 5%, 10%, 20%, 30% ของข้อมูลที่วัดซ้ำครั้งสุดท้าย จากผลการศึกษาตารางที่ 2-4 และ ภาพ 11-13 โดยใช้ค่า MSE และ เป็นเกณฑ์ในการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 3 วิธี พบว่าในทุกสถานการณ์ การประมาณค่าสูญหายด้วยวิธี Copulas มีประสิทธิภาพดีกว่า วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย และ วิธี MCMC

และเมื่อพิจารณาวิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย และ วิธี MCMC พบว่า วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ยมีประสิทธิภาพดีกว่าวิธี MCMC แม้ในบางกรณี ที่ความสัมพันธ์ระหว่างค่าวัดซ้ำมีระดับสูง ด้วยขนาดตัวอย่าง 30 ที่ระดับการสูญหาย 20% และขนาดตัวอย่าง 100 ที่ระดับการสูญหาย 5% และที่ความสัมพันธ์ระหว่างค่าวัดซ้ำมีระดับต่ำด้วยขนาดตัวอย่าง 70 ที่ระดับการสูญหาย 20% และ 30% และขนาดตัวอย่าง 100 ที่ระดับการสูญหาย 5% พบว่าวิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ยและ วิธี MCMC มีค่า MSE ใกล้เคียงกันมาก ซึ่งเนื่องจากวิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ยเป็นวิธีที่ง่ายและไม่ยุ่งยาก ดังนั้นในกรณีโครงสร้างความสัมพันธ์แบบ Compound Symmetry วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย จึงเป็นวิธีที่เหมาะสม กว่าวิธี MCMC แต่มีประสิทธิภาพน้อยกว่าวิธี Copulas

1.2. กรณีข้อมูลวัดซ้ำที่มีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Autoregressive เมื่อขนาดตัวอย่างเท่ากับ 30, 70 และ 100 ความสัมพันธ์ระหว่างค่าวัดซ้ำเท่ากับ 0.5 และ 0.7 ที่ระดับการสูญหาย 5%, 10%, 20%, 30% ของข้อมูลที่วัดซ้ำครั้งสุดท้าย จากผลการศึกษาตารางที่ 5-7 และ ภาพ 14-16 พบว่าในทุกสถานการณ์ การประมาณค่าสูญหายด้วยวิธี Copulas มีประสิทธิภาพดีกว่า ดีกว่าวิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย และ วิธี MCMC เช่นเดียวกับกรณีเมตริกซ์โครงสร้างความสัมพันธ์แบบ Compound Symmetry สำหรับการเปรียบเทียบ วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย และ วิธี MCMC พบว่า ในกรณีนี้ วิธี MCMC มีประสิทธิภาพดีกว่า วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย

นอกจากนี้สำหรับข้อมูลที่ได้จากการจำลองที่มีรูปแบบเมตริกซ์โครงสร้างความสัมพันธ์ทั้ง 2 รูปแบบ พบว่าเมื่อความสัมพันธ์ระหว่างค่าวัดซ้ำเพิ่มขึ้นวิธีการประมาณค่าสูญหายของข้อมูลทั้ง 3 วิธี จะมีประสิทธิภาพใกล้เคียงกันมากขึ้น และจะประมาณค่าสูญหายของข้อมูลได้ดีขึ้น

2. ข้อมูลจริงที่นำมาประยุกต์ใช้กับวิธีประมาณค่าสูญหาย

จากการนำข้อมูลจริงมาประยุกต์ใช้กับวิธีประมาณค่าสูญหาย3 วิธีคือ 1) วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย 2) วิธี MCMC และ 3) วิธี Copulas โดยข้อมูลที่ใช้ในการศึกษาคือ ข้อมูลรอบเอวของผู้เข้าร่วมโครงการ “นพรัตน์ไร้พุง มุ่งสุขภาพดี” จากการวัดรอบเอว 4 ครั้ง และข้อมูลปริมาณน้ำฝนรายเดือนของเดือน มิถุนายน – สิงหาคม 2550 ของสถานีกรมอุตุนิยมวิทยาภาคเหนือ สรุปผลได้ว่า วิธี copulas มีประสิทธิภาพดีกว่า วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย และวิธี MCMC

ดังนั้นจากการศึกษาวิธีประมาณค่าสูญหายของทั้งสามวิธีโดยใช้ข้อมูลที่ได้จากการจำลองด้วยเทคนิคมอนติคาร์โลและข้อมูลจริงคือข้อมูลรอบเอวของผู้เข้าร่วมโครงการ “นพรัตน์ไร้พุง มุ่งสุขภาพดี” จากการวัดรอบเอว 4 ครั้ง และข้อมูลปริมาณน้ำฝนรายเดือนของเดือน มิถุนายน – สิงหาคม 2550 ของสถานีกรมอุตุนิยมวิทยาภาคเหนือ พบว่าผลสรุปการวัดประสิทธิภาพของวิธีการประมาณค่าข้อมูลสูญหายของข้อมูลทั้งสองแบบมีผลการศึกษาไปในทิศทางเดียวกัน

ข้อเสนอแนะ

1. ในการศึกษาครั้งต่อไปสามารถเพิ่มจำนวนการวัดซ้ำ โดยศึกษาในกรณีที่ข้อมูลวัดซ้ำมีจำนวนวัดซ้ำมากกว่า 3 ครั้ง เช่น 6 หรือ 12 ครั้ง

2. ศึกษาในกรณีที่ขนาดตัวอย่างที่ใหญ่เกิน 100

3. ศึกษากับข้อมูลที่มีรูปแบบการสูญหายแบบแบบไม่เป็นระบบ (Arbitrary)

4. ศึกษาในกรณีที่การวัดซ้ำแต่ละครั้งมีขนาดตัวอย่างไม่เท่ากัน

5. เปรียบเทียบวิธีการประมาณค่าสูญหายกับวิธีการอื่น ๆ นอกเหนือจากที่ใช้ในการศึกษาในครั้งนี้ เช่น การประมาณค่าสูญหายโดยวิธีการถดถอย (Regression imputation) วิธีประมาณค่าโดยใช้ค่าใกล้สุด (Nearest neighbor imputation)

เอกสารและสิ่งอ้างอิง

ชะไมพร ธรรมวัฒน์ไพศาล. 2547. **วิธีการประมาณค่าที่ขาดหายไปในการวิเคราะห์การถดถอย**. วิทยานิพนธ์ปริญญาโท, จุฬาลงกรณ์มหาวิทยาลัย.

เชาว์ อินใย. 2547. **การพัฒนาวิธีการจัดการข้อมูลสูญหายแบบอีพีเอสเอสอีและการตรวจสอบความแม่นยำและอำนาจการทดสอบเปรียบเทียบกับวิธีอีเอ็มและลิสท์ไวท์ เทคนิคมอนติคาร์โล.** วิทยานิพนธ์ปริญญาเอก, มหาวิทยาลัยนเรศวร.

ชัยชนะ บุญสุวรรณ. 2542. **การประมาณค่าสังเกตที่สูญหายด้วยสมการแม็กซิมั่มไลลิฮูดและสมการประมาณค่าทั่วไป**. วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยมหิดล.

บวรวรรณ ดิเรกโรค. 2543. **การประยุกต์ใช้วิธีการใส่ค่าหลายค่าแทนข้อมูลที่สูญหายแต่ละค่าในการวิเคราะห์ข้อมูลอุบัติเหตุผู้ขับขี่จักรยานยนต์**. วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยมหิดล.

ปิยะภรณ์ ประสิทธิ์วัฒนเสรี และ สุคนธ์ ประสิทธิ์วัฒนเสรี. 2551. **ข้อมูลสูญหายและแนวทางการจัดการ**. Data Management & Biostatistics Journal Vol.4 No.3 : 52-61.

วิรัชช พาณิชวงศ์. 2545**. การวิเคราะห์การถดถอย.** สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพฯ.

สายชล สินสมบูรณ์. 2544. **สถิติคณิตศาสตร์ 1**. สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง, กรุงเทพฯ.

Anderson, T.W. 1984. **An Introduction to Multivariate Statistical Analysis**, Second Edition. : John Wiley & Sons, Inc, New York.

Allison, Pual D. **Multiple Imputation for Missing Data: A Cautionary Tale**. University of Pennsylvania, Pennsylvania.

Barnard, J. and D.B. Rubin. 1999. **Small-Sample Degrees of freedom with Multiple Imputation**. Biometrika, 86:948-955.

Clemen, R.t. and T. Reilly. 1999**. Correlations and copulas for decision and risk analysis**. Management Science, 45:208-224.

Chantala ,Kim and C. Suchindran, nd. **Multiple Imputation for Missing Data**. Center for Population Studies, University of North Carolina, Chapel Hill. <hpp://www.cpc.unc.edu/services/computer/presentation/mi presentation 2.pdf>

Huang, R. and K.C. Carriere. 2006. Comparison of methods for incomplete repeated measure data analysis in small sample. **Journal of Statistical Planning and Inference**, 136:235-247.

Kaarik, Ene. 2006. **Imputation algorithm Using Copulas**. Metodoloski ki zvezki, 3, 1: 109-120.

Kaarik, Ene. 2007. **Modelling Dropouts by Conditional Distribution, a Copula-Based Approach.** Institute of Mathematical Statistical, University of Tartu.

Little, RJA. and DB. Rubin. 1987. **Statistical Analysis With Missing Data**. John Wiley and Sons, Inc., New York.

Mardia, K.V. 1970. **Measures of Multivariate Skewness and Kurtosis with applications**. Biometrika, 36: 519-530.

Morrison, Donald F. 2005. **Multivariate Statistical Methods**. The Wharton School University of Pennsylvania.

Nelsen, R.B. 1999. **An Introduction to Copula**. Lectures Notes in Statistic, 139, New York: Springer Verlag

Nathaneil Schenker and Jeremy M.G. Taylor. 1996. **Partially parametric techniques for multiple imputation**. Computational Statistics & Data Analysis,22:425-446.

Roth, P.L. 1994. **Missing data: A conceptual review for applied psychologists**. Personnel Psychology, Vol. 47, p. 537–559

Rubin, D.B. 1996. **Multiple Imputation After 18+ Years**. Journal of the American Statistical Association, 91: 473-489.

SAS Institute Inc. 1999. **SAS OnlineDoc®: Version 8**. Copyright (c) 1999 SAS Institute Inc., Cary, NC, USA.

Schafer, J.L. 1997. **Analysis of Incomplete Multivariate Data.** Chapman and Hall, Inc., London.

Schafer, Joe. 2005. **Missing DataIn Longitudinal Studies: A Review.** Department of Statistics and the Methodology Center, University of Pennsylvania.

Song, P.Xas.K. 2000. Multivariate dispersion models generated from Gaussian Copula. **Scandinavian Journal of Statistics**, 27:305-320.

Yuan, Yang C. 2000. **Multiple Imputation for Missing Data: Concepts and New Development**. SAS Institute, Inc., Rockville, MD.

ภาคผนวก

ส่วนหนึ่งของโปรแกรมที่ใช้ในการวิจัยทั้งหมด เขียนด้วยโปรแกรม SAS (Statistical Analysis System) ซึ่งมีรายละเอียดดังต่อไปนี้

1. โปรแกรมสำหรับหา factor pattern

```
DATA A (TYPE=CORR); _TYPE_='CORR';
        INPUT  X1-X3;
CARDS;
1.00    .       .
 .50    1.00    .
 .50    .50     1.00
;
PROC FACTOR N=3;
RUN;
```

2. โปรแกรมสำหรับจำลองข้อมูลที่มีการวัดซ้ำ

```
proc iml;
F={ 0.81650   0.57735   0.00000,
  0.81650   -0.28868   0.50000,
  0.81650   -0.28868  -0.50000};
data=rannor(J(30,3,12346));
data=data`;
x=F*data;
x=x`;
X1=x[,1];
X2=x[,2];
X3=x[,3];
x=X1||X2||X3;
Create MVN From X;
Append From X;
quit;
```

3. โปรแกรมสำหรับสุ่มหาตำแหน่งสูญหายของข้อมูล

```
data MVN;
input col1-col3 ;
datalines;
49      236.1   280.4
31.9    275.1   186.4
213.8   342     280.8
131.5   246.4   189.3
195.4   215.1   128.4
82.9    275.2   112.8
56      393.5   130.1
85.2    332.7   80.4
31.5    298.7   151.4
27      289.7   139.9
78.9    240.5   122
114.9   208     131.2
64.9    171.5   216.9
96.9    214     160.1
153.1   175     157
25      363.9   244.8
119.1   322.7   205.5
99.5    364     129.5
22.3    422     115.9
;
data MAR(keep=x1-x3)  miss(keep=m1-m3);
set MVN;
array x[3]          x1-x3;
array m[3]          m1-m3;
```

```
array c[3]          col1-col3;
do i=3 to 3;
if ranuni(6342) < 0.07 then do;
                    x[i]= .;
                    m[i]=c[i];
          end;
          else do;
                    m[i]= .;
                    x[i]=c[i];
          end;
end;
run;
data y(keep=x1-x3);
set MVN;
set MAR;
if x1=. then do;
  x1= col1;
end;
if x2=. then do;
  x2= col2;
end;
run;
```

4. โปรแกรมสำหรับตรวจสอบการแจกแจงปกติของข้อมูลหลังมีการสูญหายของข้อมูล

```
proc univariate data=y&i  normal noprint;
var  x1 x2 x3;
 output  out=test&i probn=PROBNx1 probn=PROBNx2  probn=PROBNx3;
run;
%end;
proc iml;
sum_a=0;
%do i=1 %to &nrun;
use test&i;
read all var _num_ into b;
if b >= 0.01 then
   do;
   a=0;
   end;
else if b<0.01 then
   do;
     a=1;
   end;
sum_a=sum_a+a;
%end;
print  sum_a;
quit;
%mend;
%runit
```

ประวัติการศึกษา และการทำงาน

ชื่อ –นามสกุล	นางสาวดวงภรณ์ โปทาวี
วัน เดือน ปี ที่เกิด	วันที่ 1 พฤศจิกายน 2525
สถานที่เกิด	น่าน
ประวัติการศึกษา	วท.บ. (คณิตศาสตร์) มหาวิทยาลัยราชภัฏอุตรดิตถ์
ตำแหน่งหน้าที่การงานปัจจุบัน	นักสถิติ
สถานที่ทำงานปัจจุบัน	โรงพยาบาลนพรัตนราชธานี
ผลงานดีเด่นและรางวัลทางวิชาการ	
ทุนการศึกษาที่ได้รับ	