PORTFOLIO

지영빈

OUTLINE

A study on multiple imputation for multivariate count time series data

- Introduction
  - ✓ Motivation
  - ✓ Proposal
- Proposed method
  - ✓ Multivariate multiple imputation for multivariate count time series data
  - ✓ Algorithm
- Simulation

MOTIVATION

- Situation
  - ✓ In factories and hospitals, various items are consumed everyday.
  - ✓ Consumption of these items are daily or weekly observed and managed in terms of countable quantities.
  - ✓ Unexpected situations in real life may result in missing information in data.
    ex. system error, human error.
  - ✓ For efficient inventory management, the purpose is to replace the missing values in the existing data rather than collecting additional data.

➔ Imputation for multivariate count time series data

PROPOSAL

- Problem Statement
  - Not continuous But count. Replace missing values with count values.
    - ✓ The properties of the count data may be lost.
    - ✓ If the missing values are replaced with continuous, and the results may be difficult to interpret.
  - Multivariate count time series data
    - ✓ Consider multivariate & time lag dependence.

➔ Multivariate multiple imputation for multivariate count time series data based on a poisson regression model considering time lags.

ALGORITHM

1. Consider incomplete data of multivariate count time series data $X_{j,t} = \left(X_j^{obs}, X_{j,t}^{*}\right)$
2. Choose maximum iterations K of the chain in multiple imputation algorithm
3. Generate M datasets $X_{j,t(m)} = \left(X_j^{obs}, X_{j,t(m)}^{*}\right), m = 1, \dots, M$ to replace missing values $X_{j,t}^{*}$ from incomplete data $X_{j,t} = \left(X_j^{obs}, X_{j,t}^{*}\right)$
4. For $m = 1, \dots, M$ & $j = 1, \dots, p$,

   Sample and assign the initial values $X_{j,t(m)}^{*(0)}$ of the chain process on missing values $X_{j,t(m)}^{*}$ based on the observed X_j^{obs}
5. Repeat for $m = 1, \dots, M$,
6. Set iteration $k \leftarrow 1$ of the chain
7. While $k \leq K$ do
8. Repeat for $j = 1, \dots, p$,
9. Estimate $\mu_{j,t(m)}^{*(k)}$ using the following model:

ALGORITHM

9. Estimate $\mu_{j,t(m)}^{(k)}$ using the following model:

since $X_{j,t(m)}^{(k-1)} \sim Poisson(\mu_{j,t(m)}^{(k-1)})$ from $X_{j,t(m)}^{(k-1)} = \left(X_j^{obs}, X_{j,t(m)}^{*(k-1)}\right)$

$$\log\left(\mu_{j,t(m)}^{(k)}\right) = Z^T\boldsymbol{\beta}, \quad Z^T = (X_{(-j),t(m)}^T, X_{\cdot,(t-1)(m)}^T, \dots, X_{\cdot,(t-q)(m)}^T)$$

$$X_{(-j),t(m)}^T = (X_{1,t(m)}^{(k)}, \dots, X_{(j-1),t(m)}^{(k)}, X_{(j+1),t(m)}^{(k-1)}, \dots, X_{p,t(m)}^{(k-1)})$$

$$X_{\cdot,(t-q)(m)}^T = (X_{1,(t-q)(m)}^{(k)}, \dots, X_{(j-1),(t-q)(m)}^{(k)}, X_{j,(t-q)(m)}^{(k-1)}, \dots, X_{p,(t-q)(m)}^{(k-1)})$$

10. Generate poisson random number using estimated $\widehat{\mu_{j,t(m)}^{(k)}}$ from 9,

replace $X_{j,t(m)}^{*(k)}$, and obtain $X_{j,t(m)}^{(k)} = \left(X_j^{obs}, X_{j,t(m)}^{*(k)}\right)$

11. End repeat
12. Set $k \leftarrow k+1$
13. End while
14. End repeat
15. Finally, obtain M datasets $X_{j,t(m)}^{(k)} = \left(X_j^{obs}, X_{j,t(m)}^{*(k)}\right)$ with missing values replaced

SETTING

- Multivariate count time series data from Vector Autoregressive Model(VAR)
  - ✓ Time series variables: $j = 1, \ldots, p,\ t = 1, \ldots, T,\ q = 1, \ldots, t$

    $$\ln(\boldsymbol{\mu_t}) = \boldsymbol{c} + \Theta_1 \ln(\boldsymbol{\mu_{t-1}}) + \Theta_2 \ln(\boldsymbol{\mu_{t-2}}) + \boldsymbol{\epsilon_t}$$

    $$X_{j,t} \sim Poisson(\mu_{j,t})$$

    ➔ $p = 5$ & $p = 10$ & $T = 400$ & $q = 1,2$
  - ✓ Missing percent: 10% & 20%
  - ✓ Imputed model: Poisson & Poisson lasso
  - ✓ Imputed dataset: $M = 100$
  - ✓ Chain: $K = 19$

Variable (p)	Missing percent	Imputed model
5	10%	Poisson
		Poisson lasso
	20%	Poisson
		Poisson lasso
10	10%	Poisson
		Poisson lasso
	20%	Poisson
		Poisson lasso

SETTING

- To check the distribution of the replaced data,

$$mean\left(X_{j,t(m)}^{*(19)}\right) = \overline{X_{j,t(m)}^{*(19)}} = \frac{1}{N_j}\sum X_{j,t(m)}^{*(19)}$$

$$sd\left(X_{j,t(m)}^{*(19)}\right) = \sqrt{\frac{1}{N_j}\sum(X_{j,t(m)}^{*(19)} - \overline{X_{j,t(m)}^{*(19)}})^2}\ ,\ j = 1, \ldots, p,\ \ m = 1, \ldots, 100$$ where,

- To compare the differences between the real and imputed data

➔ Calculate the differences between distributions using the Kullback–Leibler divergence

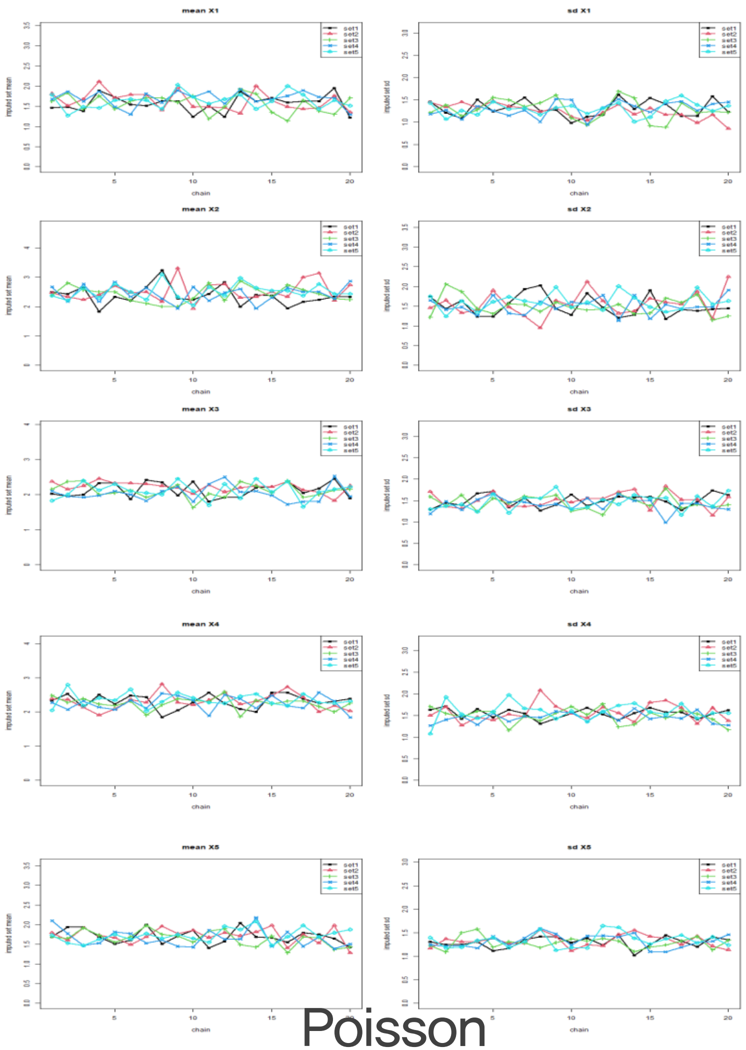
Kullback–Leibler divergence : $D_{KL}(P||Q) = H(P,Q) - H(P) = \int p(x)\log\frac{p(x)}{q(x)}dx$

Where $H(P,Q)$: cross entropy of P and Q, $\mathrm{H}(P)$: entropy of P

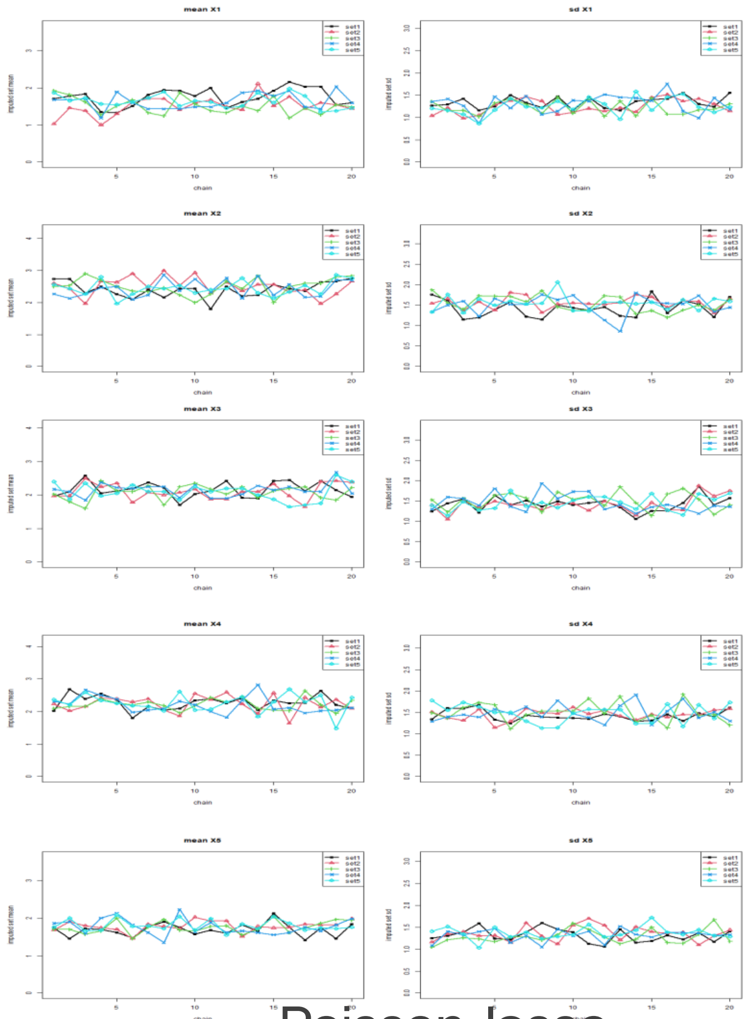
Variable: 5 & missing percent: 10%

Variable (p)	Missing percent	Imputed model	Variable (j)	imputed data			
				KL of mean dist		KL of sd dist	
				mean	sd	mean	sd
5	10%	Poisson	1	0.0098	0.0026	0.0114	0.0029
			2	0.0104	0.0019	0.0144	0.0027
			3	0.0102	0.0026	0.0147	0.0028
			4	0.0108	0.0022	0.0148	0.0032
			5	0.0109	0.0019	0.0151	0.0021
		Poisson lasso	1	0.0097	0.0026	0.0142	0.0031
			2	0.0104	0.0022	0.0150	0.0028
			3	0.0101	0.0029	0.0141	0.0033
			4	0.0103	0.0032	0.0148	0.0030
			5	0.0104	0.0021	0.0145	0.0021

Variable: 5 & missing percent: 10%

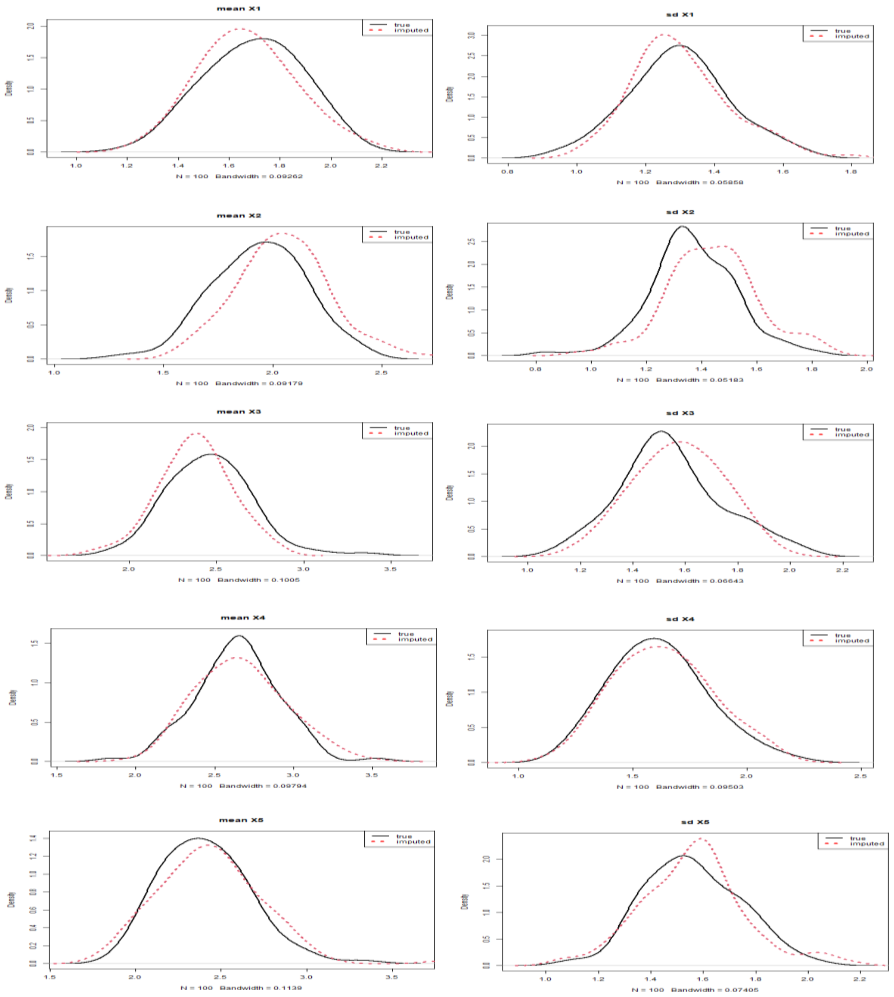


Poisson

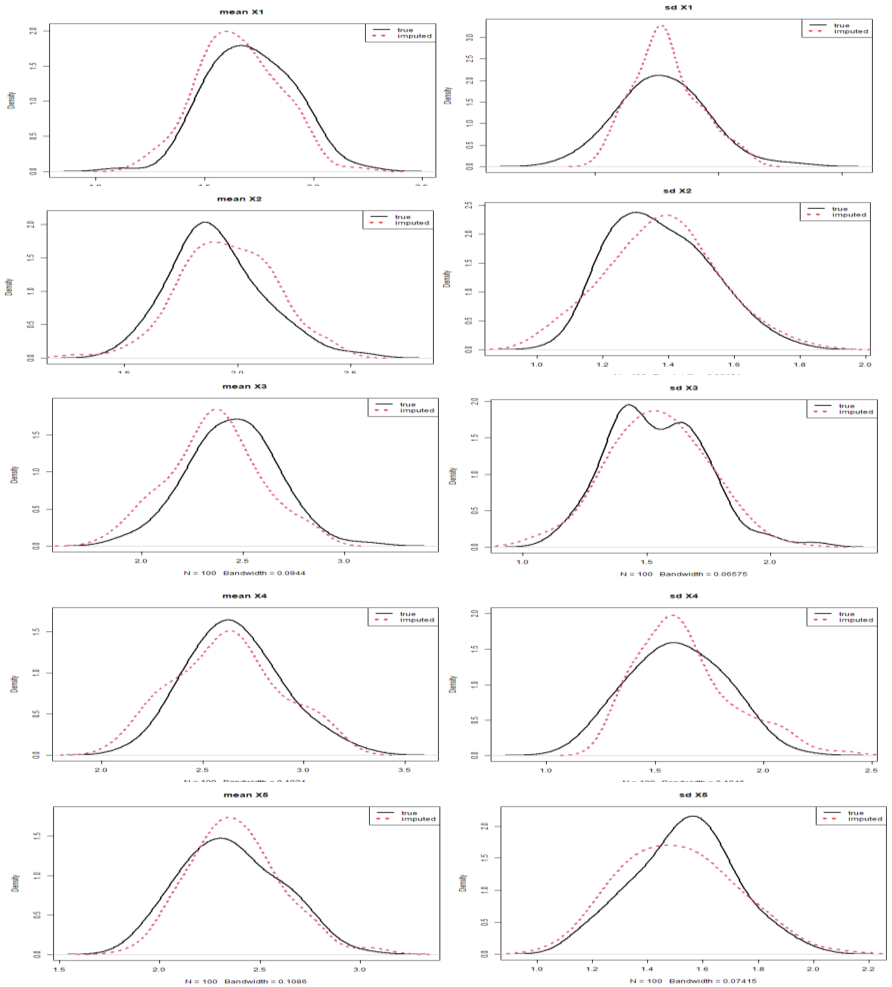


Poisson lasso

Variable: 5 & missing percent: 10%



Poisson



Poisson lasso

THANK YOU

감사합니다

APPROACHES TO HANDLING MISSING DATA

- Multiple imputation: to combine estimated imputation values that were generated by repeatedly using a single imputation to fill in any missing values.
  - ✓ goal: using the knowledge from the available data, to provide estimates that are identical to those generated using the whole data set without bias.

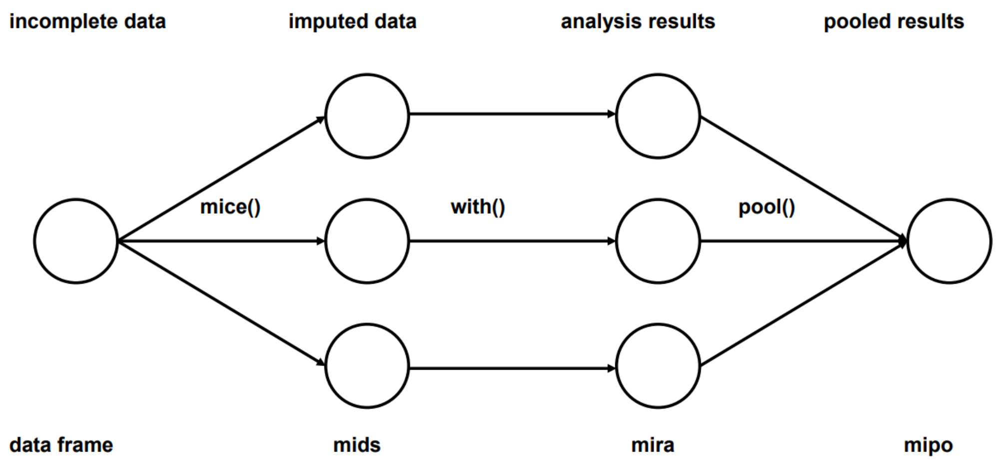


Fig. 1. Main steps used in multiple imputation
(Van Buuren와 Groothuis-Oudshoorn, 2011)

APPROACHES TO HANDLING MISSING DATA

- Step 1: Depending on the algorithm used to generate M datasets, replace each missing value with a different value.
- Step 2: Analyze each replaced dataset using selected statistical methods and estimate the parameters of interest.
- Step 3: Integrate the M association scales of each imputed dataset into an overall estimate and a variance-covariance matrix using Rubin's approach (1987)

✓ $\hat{\theta}_m$: estimates of a univariate or multivariate quantity of interest obtained from the mth imputed dataset m

✓ W_m : estimated variance of $\hat{\theta}_m$

✓ $\hat{\theta} = \frac{1}{M}\sum_{m=1}^{M} \widehat{\theta_m}$ $\qquad var(\hat{\theta}) = W + \left(1 + \frac{1}{M}\right) B$

Where $W = \frac{1}{M}\sum_{m=1}^{M} W_m$: within-imputation variance, $B = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\theta}_m - \hat{\theta})^2$: between-imputation variance.

MICE (Multivariate Imputation by Chained Equations)

(Van Buuren와 Groothuis–Oudshoorn, 2011)

- MICE: To generate a replacement value based on each model for each variable with missing data.

$$
\begin{aligned}
\theta_1^{(0)} &\sim P\left(\theta_1 \middle| X_1^{obs}\right) \\
X_1^{*(0)} &\sim P\left(X_1 \middle| \theta_1^{(0)}\right) \\
X_1^{(0)} &= (X_1^{obs}, X_1^{*(0)}) \\
&\vdots \\
\theta_p^{(0)} &\sim P\left(\theta_p \middle| X_1^{(0)}, \dots, X_{p-1}^{(0)}, X_p^{obs}\right) \\
X_p^{*(0)} &\sim P(X_p | \theta_p^{(0)}) \\
X_p^{(0)} &= (X_p^{obs}, X_p^{*(0)})
\end{aligned}
$$

$$
\begin{aligned}
\theta_1^{*(k)} &\sim P(\theta_1 | X_1^{obs}, X_2^{(k-1)}, \dots, X_p^{(k-1)}) \\
X_1^{*(k)} &\sim P(X_1 | X_1^{obs}, X_2^{(k-1)}, \dots, X_p^{(k-1)}, \theta_1^{*(k)}) \\
\theta_2^{*(k)} &\sim P(\theta_2 | X_2^{obs}, X_1^{(k)}, \dots, X_p^{(k-1)}) \\
X_2^{*(k)} &\sim P(X_2 | X_2^{obs}, X_1^{(k)}, \dots, X_p^{(k-1)}, \theta_2^{*(k)}) \\
&\vdots \\
\theta_p^{*(k)} &\sim P(\theta_p | X_p^{obs}, X_1^{(k)}, \dots, X_{p-1}^{(k)}) \\
X_p^{*(k)} &\sim P(X_p | X_p^{obs}, X_1^{(k)}, \dots, X_{p-1}^{(k)}, \theta_p^{*(k)})
\end{aligned}
$$

Where $X_j^{(k)} = (X_j^{obs}, X_j^{*(k)})$: jth imputed variable at iteration k.

X_j^* : missing values of jth imputed variable, X_j^{obs}: observed values of jth imputed variable.

POISSON REGRESSION MODEL

- Poisson regression model

$$y_i \sim P(\mu_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, y_i = 0,1,2,\ldots$$

$$g(\mu) = X'\boldsymbol{\beta} \quad \rightarrow \quad \mu = X'\boldsymbol{\beta} \text{ or } \ln \mu = X'\boldsymbol{\beta}$$

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ln \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} = \sum_{i=1}^{n}(-\mu_i + y_i \ln \mu_i - \ln y_i!) \text{ , where } \mu_i = \exp(X_i'\boldsymbol{\beta})$$

- Poisson lasso regression model (*Multicollinearity)

$$l_{lasso}(\boldsymbol{\beta}) = \sum_{i=1}^{n}(-\mu_i + y_i \ln \mu_i - \ln y_i!) + \lambda\|\boldsymbol{\beta}\|_1 \text{ , where } \lambda\|\boldsymbol{\beta}\|_1 = \lambda \sum_{j=1}^{p}|\beta_j|$$