정규 코퓰라 모형

김성민

서울대학교
통계학과, 베이즈통계 연구실

2023. 11. 29

목차

목차

코퓰라 정의

균등 주변분포를 가지고 있는 $[0, 1]^m$에서 정의된 분포함수 C를
코퓰라라고 한다.

정리

Sklar 1959 [2]

F 가 $\mathbb{R}^m$에서의 분포함수이고 일차원 주변분포가 각각 $F_1, \ldots, F_m$ 이라고 하자. 이 때, 다음을 만족하는 코퓰라 C가 존재한다.

$$F(x_1, \ldots, x_m) = C(F_1(x_1), \ldots, F_m(x_m)) \tag{1}$$

만약 F가 연속이라면 (1)를 만족하는 코퓰라 C 가 유일하게 존재하고 다음을 만족한다.

$$C(u_1, \ldots, u_m) = F(F_1^{-1}(u_1), \ldots, F_m^{-1}(u_m)).$$

역으로 만약 C가 $[0, 1]^m$에서 정의된 코퓰라고, $F_1, \ldots, F_m$ 를 $\mathbb{R}$에서의 분포함수라고 하면 (1)에서 정의된 F는 주변분포 $F_1, \ldots, F_m$를 가지는 $\mathbb{R}^m$ 에서의 분포함수가 된다.

$N(\mu,\Sigma)$ 의 분포함수를 $F_{\mu,\Sigma}$라고 하면 $N(\mu_i,\sigma_i^2)$ 의 분포함수 F_i는 주변분포가 된다. F가 연속이기 때문에 앞의 정리에 의해 다음이 성립한다.

$$
\begin{aligned}
C(u_1,\ldots,u_m) &= F_{\mu,\Sigma}(F_1^{-1}(u_1),\ldots,F_m^{-1}(u_m)) \\
&= F_{\mu,\Sigma}(\sigma_1\Phi^{-1}(u_1)+\mu_1,\ldots,\sigma_m\Phi^{-1}(u_m)+\mu_m) \\
&= F_{0,R}(\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_m)), \quad (R\text{는 상관행렬})
\end{aligned}
$$

코퓰라 예시

- 앞에서 찾은 코퓰라를 통해서 주변분포가 Φ 인 $\mathbb{R}^m$에서의 분포함수 F를 찾아보자.
- 앞의 정리에 의해서 다음과 같이 분포함수 F를 표현할 수 있다.

$$\begin{aligned} F(x_1, \ldots, x_m) &= C(F_1(x_1), \ldots, F_m(x_m)) \\ &= C(\Phi(x_1), \ldots, \Phi(x_m)) \\ &= F_{0,R}(x_1, \ldots, x_m) \end{aligned}$$

- 따라서 $F_{0,R}$ 가 코퓰라 C에 의해 정의된 분포함수임을 알 수 있다.

목차

Motivation

- 다변량에서의 의존성을 확인.
- 많은 경우 다음과 같은 가정하에서 이루어진다.

$$(X_1, \ldots, X_p) \sim N(0, \Sigma)$$

- 만약 변수가 정규성을 띄지 않는다면 적절한 변환이 필요하다.
- X_i가 F_i를 따르면 다음을 알 수 있다.

$$\Phi^{-1}(F_i(X_i)) \sim N(0, 1).$$

- 다음을 가정하자. (R은 상관행렬)

$$(\Phi^{-1}(F_1(X_1)), \ldots, \Phi^{-1}(F_p(X_p))) \sim N(0, R).$$

$$(\Phi^{-1}(F_1(X_1)), \dots, \Phi^{-1}(F_p(X_p))) \sim N(0, R).$$

- Rescaled CDF (Hoff [1])

$$\tilde{F}_i(x) = \frac{n}{n+1}\hat{F}_i, \quad (\hat{F} : ECDF)$$

- Truncated ECDF (Liu [3])

$$\tilde{F}_i(x) = \begin{cases} \delta_n & \text{if } \hat{F}_i(x) < \delta_n \\ \hat{F}_i(x) & \text{if } 1-\delta_n \le \hat{F}_i(x) < \delta_n \\ 1-\delta_n & \text{if } \hat{F}_i(x) > 1-\delta_n \end{cases}$$

Nonparanormal distribution

확률변수 $X = (X_1, \ldots, X_p)$기 연속이고 단조함수인 함수 $\{f_d : d = 1, \ldots, p\}$가 존재하여 다음을 만족하면 nonparanormal distribution 이라고 한다.

$$Y = f(X) \sim N_p(\mu, \Sigma).$$

여기서 $f(X) = (f_1(X_1), \ldots, f_p(X_p))$이다.

베이즈 추정

다음과 같은 변환 함수 f를 생각하자.(Mulgrave2020[4])

$$f(X) = \sum_{i=1}^{J} \theta_j B_j(X) \sim N_p(\mu, \Sigma), \quad X \in \mathbb{R}^{n \times p}, \ \theta_j \in \mathbb{R}^p$$

$$\theta \sim N_J(\xi, \sigma^2 I)$$

- 식별성을 위해서는 $\mu = 0, \ \Sigma_{ii} = 1$ 로 설정해야 한다.
- 사전분포 설정이 어려우므로 μ, Σ를 제한하지 않고 f_d의 척도(scale)와 위치(location)를 제한한다.

식별성

- 켤레 정규 사전분포를 얻기 위해 다음과 같은 선형 제약을 고려하자.

$$0 = f_d(1/2) = \sum_{j=1}^{J} \theta_{dj} B_j(1/2),$$

$$1 = f_d(3/4) - f_d(1/4) = \sum_{i=1}^{J} \theta_{dj} [B_j(3/4) - B_j(1/4)].$$

- 선형 제약을 다음과 같이 표현할 수 있다. $A\theta = c$

$$A = \begin{bmatrix} B_1(1/2) & B_2(1/2) & \cdots & B_J(1/2) \\ B_1(3/4) - B_1(1/4) & B_2(3/4) - B_2(1/4) & \cdots & B_J(3/4) - B_J(1/4) \end{bmatrix}$$

$$c = (0, 1)^T$$

식별성

- 주어진 제약 하에서 다음과 같은 사전분포를 얻을 수 있다.

$$\theta | \{A\theta = c\} \sim N(\xi_1, \Gamma)$$
$$\xi_1 = \xi + A^T (AA^T)^{-1} (c - A\xi)$$
$$\Gamma = \sigma^2 [I - A^T (AA^T)^{-1} A].$$

- Γ가 비정칙행렬이라는 문제가 있다. 이를 해결하기 위해서 $\theta_{d,1}, \ldots, \theta_{d,J-2}$ 만 샘플링.

$$\bar{\theta} | \{A\theta = c\} \sim N_{J-2}(\bar{\xi}_1, \bar{\Gamma})$$

- $\theta_{d,J-1}, \theta_{d,J}$ 는 $A\theta = c$ 로 구하기.

단조성

$$f_d(x) = \sum_{j=1}^{J} \theta_{dj} B_j(x)$$

- 앞에서의 θ 샘플은 f_d의 단조성을 보장하지 않는다.
- 이를 위해서는 다음과 같은 조건이 필요하다.

$$\theta_{d1} < \theta_{d2} < \cdots < \theta_{dJ}.$$

단조성

- $A\theta = c$ 조건하에서 단조 제약은 다음과 같이 표현할 수 있다.

$$\bar{F}\bar{\theta} + \bar{g} > 0, \quad \bar{F} : (J-1) \times (J-2), \quad \bar{g} : (J-2)$$

$$\bar{F} = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \cdots & -1 & 1 \\ a_1 & a_2 & a_3 & \cdots & a_{J-3} & a_{J-2} - 1 \\ b_1 - a_1 & b_2 - a_2 & b_3 - a_3 & \cdots & b_{J-3} - a_{J-3} & b_{J-2} - a_{J-2} \end{bmatrix}$$

$$\bar{g} = (0, 0, 0, , a_0, b_0 - a_0)^T$$

사전분포

- 식별성 제약조건

$$A\theta = c$$

- 단조성 제약조건

$$\mathcal{T} := \{\bar{\theta} : \bar{F}\bar{\theta} + \bar{g} > 0\}$$

- 다음과 같이 사전분포가 표현된다.

$$\bar{\theta} | \{A\theta = c\} \sim TN_{J-2}(\bar{\xi}_1, \bar{\Gamma}, \mathcal{T})$$

목차

순위가능도 기반의 코풀라 모형

이산형 변수가 포함될 경우에 정규 코풀라 모형이 잘 작동하지 않는다.

- $y_2 \sim Ber(1/2)$
- $\tilde{z}_{ij} = \Phi^{-1}[\tilde{F}_j(y_{ij})]$.
- $\tilde{z}_{i,2}$는 0.5의 확률로 $\Phi^{-1}(\frac{n}{2(n+1)})$ 와 $\Phi^{-1}(\frac{n}{(n+1)})$ 값을 가짐.

정규 코풀라 모형은 각 변수에 대해서 변환 함수를 추정해야 한다.

$\Longrightarrow$ Hoff [1] 에서 (확장된) 순위가능도를 이용하여 추정하는 방법 제안.

순위가능도 기반의 코퓰라 모형

$$(Y_{i1}, \ldots, Y_{ip}) \overset{iid}{\sim} N(0, R), \quad i = 1, \ldots, n$$
$$Y_{ij} = \Phi^{-1}(F_j(X_{ij})), \quad j = 1, \ldots, p.$$

- F_j는 증가함수이므로 $X_{i_1,j} < X_{i_2,j}$ 이면 $Y_{i_1,j} < Y_{i_2,j}$ 가 성립한다.
- 일반적으로 표현하면 $X = (X_1, \ldots, X_n)^T$ 가 주어졌을 때 Y는 다음과 같은 집합에 포함될 수 밖에 없다.

$$D := \{Y \in \mathbb{R}^{n \times p} : \max\{y_{k,j} : x_{k,j} < x_{i,j}\} < y_{i,j} < \min\{y_{k,j} : x_{k,j} < x_{i,j}\}\}$$

순위가능도 기반의 코풀라 모형

다음과 같이 가능도를 계산할 수 있다.

$$\mathbb{P}(Y \in D | R, F_1, \ldots, F_p) = \int_D p(Y|R)dY = \mathbb{P}(Y \in D|R).$$

- 위 가능도는 $F_1, \ldots, F_p$와 무관한 R에만 의존하는 함수이다.
- 빈도론 : $\mathbb{P}(Y \in D|R)$ 를 최대화하는 R 찾기.
- 베이즈 : 사후분포를 다음과 같이 구할 수 있따.

$$\mathbb{P}(R|Y \in D) \propto p(C) \cdot \mathbb{P}(Y \in D|R).$$

베이즈 추정

$$\mathbb{P}(R|Y \in D) \propto p(C) \cdot \mathbb{P}(Y \in D|R).$$

- Mulgrave([5])가 식별 불가능한 정밀도 행렬 (Ω)를 이용하는 것을 제안.

$$\Psi = R^{-1} = A\Omega A, \quad A = diag(\sqrt{\Sigma_{ii}})$$

$$\mathbb{P}(\Psi|Y \in D) \propto p(\Psi) \cdot \mathbb{P}(Y \in D|\Psi).$$

- 깁스 샘플링을 이용.
  1. Ψ 하에서 Y 샘플링
  2. Y 하에서 Ω 샘플링
  3. Ω 를 Ψ로 변환.

목차

코퓰러 모형([6])

$$y_j \sim F_j(\cdot; \theta_j)$$

- 이 때, 정규 코퓰라 함수는 다음과 같다. (Γ : 상관행렬)

$$C(u_1, \ldots, u_m) = F_{0,\Gamma}(\phi^{-1}(u_1), \ldots, \phi^{-1}(u_m))$$

- 정규 코퓰라로 생성되는 정규 분포함수는 다음과 같다.

$$F(y) = C(F_1(y_1), \ldots, F_m(y_m)).$$

코퓰러 모형

- 만약에 주변분포가 모두 미분가능하다면 다음과 같은 밀도함수를 얻을 수 있다.

$$
\begin{aligned}
f(y) &= \frac{\partial}{\partial y} C(F_1(y_1), \ldots, F_m(y_m)) \\
&= c(u) \cdot \prod_{i=1}^{m} f_j(y_j)
\end{aligned}
$$

$$
u = (F_1(y_1), \ldots, F_m(y_m)), \quad c(u) = \frac{\partial}{\partial u} C(u)
$$

- 코퓰라 밀도함수 $c(u)$는 다음과 같다.

$$
c(u) = |\Gamma|^{-1/2} \exp\left(-\frac{1}{2} x^T (\Gamma^{-1} - I_m) x\right), \quad x = (\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_m))
$$

코퓰라 모형

- 코퓰라로 생성되는 밀도함수는 다음과 같다.

$$f(y|\Theta,\Gamma) = |\Gamma|^{-n/2}\left(\prod_{i=1}^{n}\exp\left\{-\frac{1}{2}x_i^T(\Gamma^{-1}-I_m)x_i\right\}\prod_{i=1}^{m}f_j(y_{ij};\theta_j)\right)$$

베이즈 추정

1. $f(\theta_j|\{\Theta\backslash\theta_j\},\Gamma,y)$ 에서 샘플링

$$
\begin{aligned}
& f(\theta_j|\{\Theta\backslash\theta_j\},\Gamma,y) \\
& \propto f(y|\Theta,\Gamma)\pi(\theta_j) \\
& \propto |\Gamma|^{-n/2}\left(\prod_{i=1}^{n}\exp\left\{-\frac{1}{2}x_i^T(\Gamma^{-1}-I_m)x_i\right\}\prod_{i=1}^{m}f_j(y_{ij};\theta_j)\right)\pi(\theta_j)
\end{aligned}
$$

2. $f(\Gamma|\Theta,y)$ 에서 샘플링
   1. $\Gamma = diag(\Sigma)^{-1/2}\Sigma diag(\Sigma)^{-1/2}$ 이고 $\Sigma^{-1} = LL^T$ 임을 이용하여 $f(l_{jk}|\{L\backslash l_{jk}\},\Theta,y)$ 에서 샘플링. (단, $l_{kk} = 1$로 고정)

$$
f(l_{jk}|\{L\backslash l_{jk}\},\Theta,y) \propto |\Gamma|^{-n/2}\left(\prod_{i=1}^{n}\exp\left\{-\frac{1}{2}x_i^T(\Gamma^{-1}-I_m)x_i\right\}\right)\pi(l_{kj}).
$$

   2. 샘플링한 R에서 Γ로 변환

References I

[1] HOFF, P. D. Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics* (2007), 265–283.

[2] KLAASSEN, C. A., AND WELLNER, J. A. Efficient estimation in the bivariate normal copula model: normal margins are least favourable. *Bernoulli* (1997), 55–77.

[3] LIU, H., LAFFERTY, J., AND WASSERMAN, L. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research 10*, 10 (2009).

[4] MULGRAVE, J. J., AND GHOSAL, S. Bayesian inference in nonparanormal graphical models. *Bayesian Analysis 15*, 2 (2020).

[5] MULGRAVE, J. J., AND GHOSAL, S. Bayesian analysis of nonparanormal graphical models using rank-likelihood. *Journal of Statistical Planning and Inference 222* (2023), 195–208.

References II

[6] Smith, M. S. Bayesian approaches to copula modelling. *Damien, PP Dellaportas, NG Polson, DA Stephens,(2013), Bayesian Theory and Applications, OUP* (2011), 336–358.