Comment on "Model Uncertainty and Missing Data: An Objective Bayesian Perspective"

Yichen Ji[*], Monica J. Alexander[†] and Radu V. Craiu[‡]

Introduction

We congratulate the authors (henceforth, GDCCQF) on their contribution to the literature on model selection with missing data. Our discussion was inspired by the requirement to estimate the joint distribution of the covariates, $f(X_1,\dots,X_k|\nu)$, which is at the core of their method. The problem of estimating a joint distribution is central to modern statistics and its positive resolution has been known to impact other influential methods such as the knockoff method [1], the study of imputation efficiency in regression models [11], or sample surveys [3].

Estimation of the joint distribution of covariates

Although GDCCQF use, for simplicity, a multivariate Gaussian to generate the covariates and also to model their joint distribution, $f(X_1,\dots,X_k|\nu)$, real data can depart from such an ideal setup in several ways. First, some covariates may have marginal distributions that are not Gaussian or even continuous, thus falsifying the multivariate Gaussian assumption. Second, the type of dependence captured by the Gaussian distribution may differ substantially from the dependence patterns exhibited by f. For example, it is well known that the tail dependence coefficients are zero for multivariate Gaussian distributions, but not so for other multivariate laws [see 8, and references therein]. A general mathematical framework for such comparisons is provided by the copula function which links the marginal and joint distributions of a multivariate vector [10, 5]. Furthermore, copulas have been increasingly used in the development of statistical methods for multivariate-dependent data [e.g., 4, 2, 6, 12, 9]. These developments are accompanied by software packages [e.g., 7] or programs that make it easier to implement copula-based techniques.

A small numerical study

A copula formulation allows us to study empirically how the performance of GDCCQF's variable selection procedure is impacted when the marginals and the dependence structure of f are misspecified. The analysis produced by GDCCQF's programs assumes that f is a multivariate Gaussian distribution which is equivalent to a Gaussian copula model in which the marginals are all Gaussian. We considered data under three simulation scenarios that vary the copula and the marginals as follows:

arXiv: 2010.00000

[*] Department of Statistical Sciences, University of Toronto, yc.ji@mail.utoronto.ca

[†] Department of Sociology and Department of Statistical Sciences, University of Toronto, monica.alexander@utoronto.ca

[‡] Department of Statistical Sciences, University of Toronto, radu.craiu@utoronto.ca

CG We used a multivariate Clayton copula in which the Kendall tau between each pair of variables is 0.8, and all marginals are standard Gaussian. This copula choice introduces a strong lower tail dependence, unlike the posited model, which assumes that there is no tail dependence.

CL Same as **CG** but with marginal densities that are generalized Gaussian

$$g(x) \propto \exp(-|x|^8), \ \forall x \in \mathbb{R}$$

and thus have lighter tails than the posited model.

CH Same as **CL** but with marginal densities that are mixtures of a standard Gaussian (weight is 0.2) and an Exponential with parameter 2,

$$g(x) = 0.2\phi(x) + 1.6\exp(-2x)\mathbf{1}_{\{x\geq 0\}}(x), \ \forall x \in \mathbb{R},$$

where ϕ is the density of a standard normal, and $\mathbf{1}_{\{x\geq 0\}}(x)$ is equal to one if $x \geq 0$ and is zero otherwise. This yields marginals with a heavier right tail than in the posited model.

For all scenarios, each generated data set contained $n = 300$ observations, the number of covariates under consideration was $k = 10$, with active covariates X_1, X_2, X_6, X_7 having the corresponding regression coefficients, $\beta_1 = 1$, $\beta_2 = 2$, $\beta_6 = 1$ and $\beta_7 = 2$, and the missing probability was set to $p = 0.1$ under a MCAR scheme. We followed GDCCQF to produce $nMC = 500$ imputations. Each scenario was independently replicated $R = 500$ times. Table 1 presents the false negative (FN) and false positive (FP) rates for the three simulation scenarios when the data are analyzed assuming a multivariate Gaussian distribution. We note that the impact on the selection of active covariates seems to vary according to the size of their effect and the missing patterns produced in each replicate. Given that the error rates are not very high, we would need more replicates in order to see similar FP rates for all non-active covariates.

Conclusion

The simulations suggest that misspecification of the dependence can alter the performance of the method. Performance degradation is amplified when marginals are also misspecified. More work is required to understand if the difference in tail dependence between the Clayton copula and the Gaussian copula is responsible for these errors, or other copulas more similar to the Gaussian, such as a t copula, can also wreak havoc. Model misspecification is a well-known issue in Bayesian analysis, but in this case it can be realistically addressed by considering copula models to fit f. Our contribution to this discussion is not meant to be a criticism of GDCCQF's method, but is rather aimed at stimulating the development of flexible estimation methods of multivariate distributions when part of the data are missing. We thank GDCCQF for an inspiring article that opens several directions for further study.

Table 1: Error rates for each variable in the three scenarios. For the active covariates shown in bold, the error rates represent the fraction of false negatives, while for the remaining inactive covariates the error rates represent the fraction of false positives.

Scenario	Covariate									
	$\mathbf{X_1}$	$\mathbf{X_2}$	X_3	X_4	X_5	$\mathbf{X_6}$	$\mathbf{X_7}$	X_8	X_9	X_{10}
CG	0	0	0.216	0	0.108	0.002	0	0.002	0.002	0
CL	0.120	0	0.11	0	0.11	0.548	0.01	0	0	0
CH	0.108	0	0.318	0.002	0.106	0.008	0	0.014	0.002	0.002

References

[1] Candes, E., Fan, Y., Janson, L., and Lv, J. (2018). "Panning for gold:'model-X'knockoffs for high dimensional controlled variable selection." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3): 551–577. 1

[2] Dissmann, J., Brechmann, E. C., Czado, C., and Kurowicka, D. (2013). "Selecting and estimating regular vine copulae and application to financial returns." *Computational Statistics & Data Analysis*, 59: 52–69. 1

[3] Gelman, A., King, G., and Liu, C. (1998). "Not asked and not answered: Multiple imputation for multiple surveys." *Journal of the American Statistical Association*, 93(443): 846–857. 1

[4] Genest, C., Favre, A.-C., Beliveau, J., and Jacques, C. (2007). "Metaelliptical copulas and their use in frequency analysis of multivariate hydrological data." *Water Resources Research*, 43(9). 1

[5] Genest, C. and Rivest, L.-P. (1993). "Statistical inference procedures for bivariate Archimedean copulas." *Journal of the American Statistical Association*, 88: 1034–1043. 1

[6] Hasler, C., Craiu, R. V., and Rivest, L.-P. (2018). "Vine Copulas for Imputation of Monotone Non-response." *International Statistical Review*, 86(3): 488–511. 1

[7] Hofert, M., Kojadinovic, I., Maechler, M., and Yan, J. (2020). *copula: Multivariate Dependence with Copulas*. R package version 1.0-1. 1

[8] Hua, L. and Joe, H. (2011). "Tail order and intermediate tail dependence of multivariate copulas." *Journal of Multivariate Analysis*, 102(10): 1454–1471. 1

[9] Pan, R., Nieto-Barajas, L. E., and Craiu, R. V. (2025). "Bayesian Nonparametric Mixtures of Archimedean Copulas." *Journal of Agricultural, Biological and Environmental Statistics*, 1–25. 1

[10] Sklar, A. (1959). "Fonctions de répartition à *n* dimensions et leurs marges." *Publications de l'Institut de Statistique de l'Université de Paris*, 8: 229–231. 1

[11] White, I. R. and Carlin, J. B. (2010). "Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values." *Statistics in medicine*, 29(28): 2920–2931. 1

[12] Zimmerman, R., Craiu, R. V., and Leos-Barajas, V. (2024). "Copula modeling of serially correlated multivariate data with hidden structures." *Journal of the American Statistical Association*, 119(548): 2598–2609. 1