Supplement for "COPULA-BASED ANALYSIS FOR COUNT TIME SERIES"

Mingze Zhang[1], Huixia Judy Wang[2] and James Livsey[3]

[1] *JPMorgan Chase & Co.*

[2] *Department of Statistics, The George Washington University*

[3] *U.S. Census Bureau*

Supplementary Material

S1 Basic statistics of the proposed model

In our paper, we employ copulas to capture serial dependence based on the assumed Markov property, which is in the same spirit as considered in Chen and Fan (2006), Chen et al. (2009), and Tang et al. (2019). Given this and our assumed regression setup, the autocorrelation function (ACF) depends on the covariates; even the unconditional (not dependent on X_t) value is an open problem in the literature. However, we can note some potentially useful insights.

Firstly, the mean of the series $\{Y_t\}$ is the same as that of the marginal distribution. For instance, for the simulation design considered in Section 3, $E(Y_t|X_t) = \exp\{\beta_1 + \beta_2 X_t\}$. Secondly, for certain copulas, one can apply Hermite expansion-type arguments to calculate the variance and ACF of $\{Y_t\}$, as demonstrated in Jia et al. (2021). For example, consider the Gaussian copula model in our setup

$$\Phi_2\left(\phi^{-1}(G(y_t|x)), \phi^{-1}(G(y_{t-1})|x)\right).$$

The observations conditioned on the covariates can be thought of as a transformed Gaussian process $\{Z_t\}$:

$$Z_t = \Phi^{-1}(G(y_t|x)) := L(Z_t).$$

The calculations in Jia et al. (2021) can now be extended to the bivariate case as in Section 5.7 of Pipiras and Taqqu (2017).

Finally, although it is possible to employ the Hermite expansion for approximating the variance and ACF, this calculation demands more in-depth study, and this will be an interesting topic for future research. On the other hand, note that $\{Z_t\}$ is a monotone increasing transformation of $\{Y_t|x\}$ and the former is a Gaussian process. It is known that rank correlation such as Kendal's τ correlation is invariant to monotone transformation. Therefore, the Kendal's τ autocorrelation for $\{Y_t|x\}$ is identical to that for $\{Z_t\}$. Thus in practice, we can use Kendal's τ ACF estimation based on

$\hat{Z}_t = \Phi^{-1}(\hat{G}(y_t|x))$ to perform model diagnosis.

We conducted a small-scale simulation to validate the previously mentioned conclusions. The data was generated following the same design as in Case 1, with $\alpha = 0.454$, corresponding to Kendall's $\tau = 0.3$. For each value of X, we generated a series $\{Y_t, t = 1, \ldots, n = 500\}$. In Figure S.1 below, we present: (a) the curves for sample mean and population mean against X; (b) the curves for sample variance and marginal variance against X; (c) the sample Kendall's τ for lag 1 correlation obtained for each X (depicted as circles) and the true Kendall's τ (which remains a constant invariant of X in our assumed model setup and is represented by the solid horizontal line); (d) the sample Kendall's τ autocorrelation function (ACF) for one example series $\{Y_t, t = 1, \ldots, n = 500\}$.

The results show that the sample conditional mean closely aligns with the population mean. The sample conditional Kendall's τ values fluctuate around the true value of 0.3, which remains constant invariant of X under our model framework. As expected, the sample variance is dependent on X in a manner similar to how the mean varies with X. Figure S.1 (d) demonstrates that Kendall's τ ACF exhibits a similar pattern to what we would expect from the ACF of a continuous AR(1) process, supporting our earlier recommendation to use Kendall's τ as a diagnostic tool for model assessment.

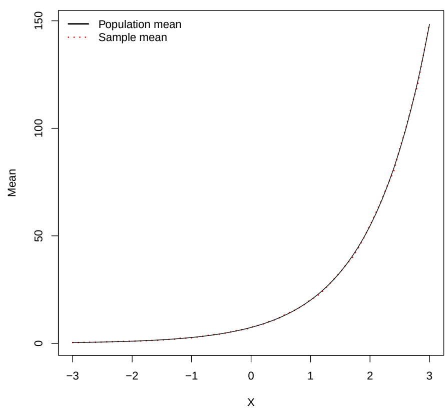
S2 Likelihood when $p \geq 1$

For general p-th order Markov processes, the PMF involves 2^{p+1} terms of copula functions. Similar to the case when $p = 1$, if the $(p+1)$-dimensional copula C_{p+1} is selected to be copula families without closed expressions, such as elliptical copulas, including Gaussian copula or t copula, then the joint distribution of $\boldsymbol{Y}_t$ can be obtained as

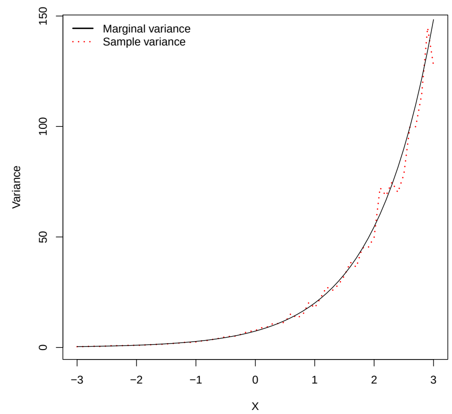
$$
\begin{aligned}
&P(Y_t = y_t, \ldots, Y_{t-p} = y_{t-p} | \boldsymbol{X}) \\
&= \int_{\Phi_1^{-1}\left(F_{t-p}^2\right)}^{\Phi_1^{-1}\left(F_{t-p}^1\right)} \cdots \int_{\Phi_1^{-1}\left(F_t^2\right)}^{\Phi_1^{-1}\left(F_t^1\right)} \phi_{p+1}\left(\psi_1, \ldots, \psi_{p+1} | \boldsymbol{X}; \boldsymbol{\alpha}\right) d\psi_1 \cdots d\psi_{p+1}. \quad \text{(S2.1)}
\end{aligned}
$$

Here, $\phi_{p+1}(\cdot; \boldsymbol{X}, \boldsymbol{\alpha})$ denotes the probability density function (PDF) of a $(p+1)$-dimensional elliptical distribution with location 0 and correlation given by the parameters $\boldsymbol{\alpha}$, Φ_1 denotes the CDF of the univariate margin of the same elliptical distribution, $F_t^1 := \mathrm{P}\left(Y_t \leq y_t | \boldsymbol{X}; \boldsymbol{\beta}\right)$ and $F_t^2 := \mathrm{P}\left(Y_t \leq y_t - 1 | \boldsymbol{X}; \boldsymbol{\beta}\right)$. Otherwise, if the $(p+1)$-dimensional copula has a closed expression for the CDF, we can then use the finite difference form to calculate the conditional joint distribution, that is,

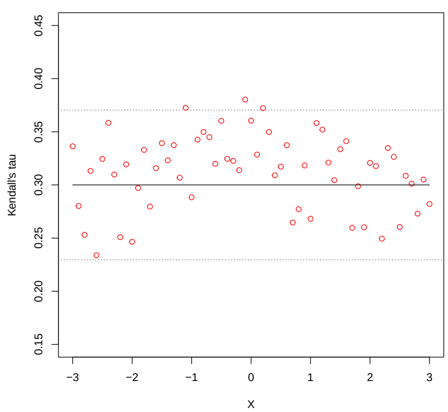
$$
P(Y_t = y_t, \ldots, Y_{t-p} = y_{t-p} | \boldsymbol{X})
$$



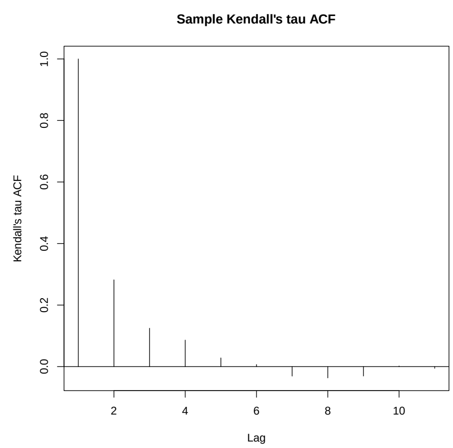
[]



[]



[]



[]

Figure S.1: Simulation for Case 1: (a) the sample and population conditional mean against X; (b) the sample and marginal conditional variance against X; (c) the sample conditional Kendall's τ for lag 1 correlation against X (circles) and the true Kendall's τ (solid horizontal line); and (d) the sample Kendall's τ ACF for one example series.

$$
= \sum_{j_0=1}^{2} \cdots \sum_{j_p=1}^{2} (-1)^{j_0+\cdots+j_p} C_{p+1}\left(F_t^{j_0}, \ldots, F_{t-p}^{j_p} | \boldsymbol{X}; \boldsymbol{\alpha}\right). \tag{S2.2}
$$

Then the conditional PMF of Y_t given $Y_{t-1}, \ldots, Y_{t-p}$ is given by

$$
\begin{aligned}
& P(Y_t = y_t | Y_{t-1} = y_{t-1}, \ldots, Y_{t-p} = y_{t-p}, \boldsymbol{X}) \\
=& \frac{P(Y_t = y_t, \ldots, Y_{t-p} = y_{t-p} | \boldsymbol{X})}{P(Y_{t-1} = y_{t-1}, \ldots, Y_{t-p} = y_{t-p} | \boldsymbol{X})} \\
=& \begin{cases}
\dfrac{\int_{\Phi_1^{-1}\left(F_{t-p}^2\right)}^{\Phi_1^{-1}\left(F_{t-p}^1\right)} \cdots \int_{\Phi_1^{-1}\left(F_t^2\right)}^{\Phi_1^{-1}\left(F_t^1\right)} \phi_{p+1}(\psi_1, \ldots, \psi_{p+1} | \boldsymbol{X}; \boldsymbol{\alpha}) d\psi_1 \cdots d\psi_{p+1}}{\int_{\Phi_1^{-1}\left(F_{t-p}^2\right)}^{\Phi_1^{-1}\left(F_{t-p}^1\right)} \cdots \int_{\Phi_1^{-1}\left(F_{t-1}^2\right)}^{\Phi_1^{-1}\left(F_{t-1}^1\right)} \phi_p(\psi_1, \ldots, \psi_p | \boldsymbol{X}; \boldsymbol{\alpha}) d\psi_1 \cdots d\psi_p} & C_{p+1}\text{: without closed expression} \\
\dfrac{\sum_{j_0=1}^{2} \cdots \sum_{j_p=1}^{2} (-1)^{j_0+\cdots+j_p} C_{p+1}\left(F_t^{j_0}, \ldots, F_{t-p}^{j_p} | \boldsymbol{X}; \boldsymbol{\alpha}\right)}{\sum_{j_1=1}^{2} \cdots \sum_{j_p=1}^{2} (-1)^{j_1+\cdots+j_p} C_p\left(F_{t-1}^{j_1}, \ldots, F_{t-p}^{j_p} | \boldsymbol{X}; \boldsymbol{\alpha}\right)} & C_{p+1}\text{: with closed expression}
\end{cases}.
\end{aligned} \tag{S2.3}
$$

where C_p is the p-dimensional marginal distribution of C_{p+1}.

Thus, the likelihood function of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ can be derived as:

$$
\begin{aligned}
& L(\boldsymbol{\beta}, \boldsymbol{\alpha}) \\
=& P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n | \boldsymbol{X}; \boldsymbol{\beta}, \boldsymbol{\alpha}) \\
=& P(Y_1 = y_1 | \boldsymbol{X}) P(Y_2 = y_2 | Y_1 = y_1, \boldsymbol{X}) \ldots P(Y_n = y_n | Y_{n-1} = y_{n-1}, \ldots, Y_{n-p} = y_{n-p} | \boldsymbol{X})
\end{aligned}
$$

$$
= \begin{cases} \dfrac{\prod_{t=p+1}^{n} \int_{\Phi_1^{-1}\left(F_{t-p}^2\right)}^{\Phi_1^{-1}\left(F_{t-p}^1\right)} \cdots \int_{\Phi_1^{-1}\left(F_t^2\right)}^{\Phi_1^{-1}\left(F_t^1\right)} \phi_{p+1}(\psi_1,\ldots,\psi_{p+1}|\boldsymbol{X};\boldsymbol{\alpha})d\psi_1\cdots d\psi_{p+1}}{\prod_{t=p+1}^{n-1} \int_{\Phi_1^{-1}\left(F_{t-p}^2\right)}^{\Phi_1^{-1}\left(F_{t-p}^1\right)} \cdots \int_{\Phi_1^{-1}\left(F_{t-1}^2\right)}^{\Phi_1^{-1}\left(F_{t-1}^1\right)} \phi_p(\psi_1,\ldots,\psi_p|\boldsymbol{X};\boldsymbol{\alpha})d\psi_1\cdots d\psi_p} & \\ & C_p\text{: without closed expression} \\ \dfrac{\prod_{t=p+1}^{n}\left\{\sum_{j_0=1}^{2}\cdots\sum_{j_p=1}^{2}(-1)^{j_0+\cdots+j_p}C_{p+1}\left(F_t^{j_0},\ldots,F_{t-p}^{j_p}|\boldsymbol{X};\boldsymbol{\alpha}\right)\right\}}{\prod_{t=p+1}^{n-1}\left\{\sum_{j_1=1}^{2}\cdots\sum_{j_p=1}^{2}(-1)^{j_1+\cdots+j_p}C_p\left(F_{t-1}^{j_1},\ldots,F_{t-p}^{j_p}|\boldsymbol{X};\boldsymbol{\alpha}\right)\right\}} & \\ & C_p\text{: with closed expression} \end{cases} . \tag{S2.4}
$$

S3 Proof of Proposition 1

Proof. Without loss of generality, we consider $p = 1$ and $d = 1$ with X being a continuous random variable. We need to show that $C(\cdot)$ in Assumption A1(c) is uniquely determined. To avoid confusion, we denote the marginal CDFs of $Y_t|x$ and $Y_{t-1}|x$ as $F(\cdot|x)$ and $G(\cdot|x)$. Let $C_{\mathcal{H}}(\cdot)$ be the set of copulas for which

$$
\mathcal{H}(y_t, y_{t-1}|x) = C\{F(y_t|x), G(y_{t-1}|x)\}, \quad y_1, y_2 \in \mathbb{R} \tag{S3.5}
$$

holds. Thus, for $\forall\ C(\cdot), D(\cdot) \in C_{\mathcal{H}}$, and $\forall\ u, v \in (0,1) \times (0,1)$, we have

$$
\begin{aligned}
&|C(u,v) - D(u,v)| \\
\leq &|C(u,v) - C\{F(i|x), G(j|x)\}| + |D(u,v) - D\{F(i|x), G(j|x)\}|.
\end{aligned}
$$

Note that for a given $X = x$, the conditional CDFs $F(y|x)$ and $G(y|x)$ cannot take any value in $(0, 1)$. However, since X is a continuous random variable, $F(y|x)$ and $G(y|x)$ can take any value in $(0, 1)$ across x (as we vary the values of x). Therefore, for $\forall\ u, v \in (0, 1)$, $\exists\ x, i, j$, where i and j depend on x, such that $F(i|x) = u$ and $G(j|x) = v$, and consequently

$$
\begin{aligned}
&|C(u,v) - D(u,v)| \\
\leq &|C(u,v) - C\{F(i|x), G(j|x)\}| + |D(u,v) - D\{F(i|x), G(j|x)\}| \\
= &0.
\end{aligned}
$$

$\square$

S4 Proof of Theorem 1

We first introduce one lemma. The statement was mentioned in Longla and Peligrad (2012) but without a formal proof. Below, we provide a proof for this lemma to ensure completeness.

Lemma 1. *The mixing coefficients of a Markov chain generated by a given copula C and marginal distribution uniform on $[0, 1]$, are larger than or equal to those of a Markov chain generated by the same copula and another marginal distribution G that is not necessarily continuous.*

Proof. According to the definition of the $\beta-$mixing coefficients, we have the $\beta-$mixing coefficients between two $\sigma-$fields $\mathcal{A}$, $\mathcal{B}$ is

$$\beta(\mathcal{A},\mathcal{B}) = \frac{1}{2}\sup_{\{A_i\},\{B_j\}}\sum_{i=1}^{n}\sum_{j=1}^{m}\left|\Pr\left(A_i\cap B_j\right)-\Pr\left(A_i\right)\Pr\left(B_j\right)\right|,$$

where the supremum is taken over all positive integers n and m, and all finite partitions $\{A_i\}$, $\{B_j\}$ with $A_i\in\mathcal{A}$ and $B_j\in\mathcal{B}$.

Let $\{U_n\}_{n\in\mathbb{Z}}$ be the Markov chain generated by a given copula C and marginal distribution uniform on $[0,1]$, that is, from the model in Assumption A1(c) with joint distribution $P(U_t\leq u_t,\ldots,U_{t-p}\leq u_{t-p}) = C(u_t,\ldots,u_{t-p})$. Let $\mathcal{P}_1=\sigma\{U_k, k\leq 0\}$, and $\mathcal{F}_{1n}=\sigma\{U_k, k\geq n\}$. Let G be a cumulative distribution function, which is not necessarily continuous. Define $Y_n = G^{-1}(u_n)\doteq\inf\{y, G(y)\geq u_n\}$, then $\{Y_n\}_{n\in\mathbb{Z}}$ is a Markov chain generated by the same copula C and the marginal distribution G. Furthermore, define $\mathcal{P}_2=\sigma\{Y_k, k\leq 0\}$ and $\mathcal{F}_{2n}=\sigma\{Y_k, k\geq n\}$. Therefore, we can express the $\beta-$mixing coefficients of $\{U_n\}_{n\in\mathbb{Z}}$ and $\{Y_n\}_{n\in\mathbb{Z}}$ as

$$\beta_{1n}=\beta(\mathcal{P}_1,\mathcal{F}_{1n}) = \frac{1}{2}\sup_{\{A_{1i}\},\{B_{1j}\}}\sum_{i=1}^{n}\sum_{j=1}^{m}\left|\Pr\left(A_{1i}\cap B_{1j}\right)-\Pr\left(A_{1i}\right)\Pr\left(B_{1j}\right)\right|,$$

where the supremum is taken over all positive integers n and m, and all finite partitions $\{A_{1i}\}$, $\{B_{1j}\}$ with $A_{1i}\in\mathcal{P}_1$ and $B_{1j}\in\mathcal{F}_{1n}$, and

$$\beta_{2n}=\beta(\mathcal{P}_2,\mathcal{F}_{2n}) = \frac{1}{2}\sup_{\{A_{2i}\},\{B_{2j}\}}\sum_{i=1}^{n}\sum_{j=1}^{m}\left|\Pr\left(A_{2i}\cap B_{2j}\right)-\Pr\left(A_{2i}\right)\Pr\left(B_{2j}\right)\right|,$$

where the supremum is taken over all positive integers n and m, and all finite partitions $\{A_{2i}\}$, $\{B_{2j}\}$ with $A_{2i} \in \mathcal{P}_2$ and $B_{2j} \in \mathcal{F}_{2n}$.

Note that $\mathcal{P}_1 \supset \mathcal{P}_2$ and $\mathcal{F}_{1n} \supset \mathcal{F}_{2n}$, so we get $\beta_{1n} \geq \beta_{2n}$, and Lemma 1 is proven.

$\square$

Proof of Theorem 1: Theorem 1 can be proven by applying Theorem 6.4 in White (1996) using the standard maximum likelihood theory. To do so, below we will verify that the required conditions in Theorem 6.4 are met.

Firstly, condition 2.1 in White (1996) is met under the model assumption A1. Assumptions A2(a) and A2(b) ensure that condition 2.3 is satisfied. Assumption A3 ensures conditions 3.1(a) and 3.1(b). Assumption A4 can be easily verified by using the previously introduced Lemma 1. By combining A4 with A1(b), A1(c), A2(a), A3, and A8(a), and Corollary 1 from Pötscher and Prucha (1989), we can demonstrate that the strong uniform law of large numbers (ULLN) holds for $\log\{f_t(Y_t; \boldsymbol{\theta_0})\}_{t=1}^{n}$. This, in turn, verifies condition 3.1(c) in White (1996). In addition, Assumption A2(c) confirms condition 3.2′, while A5 ensures condition 3.6 in White (1996). Assumption A6(a) ensures condition 3.7(a), and A6(b) along with A6(c) fulfill conditions 3.8(a) and 3.8(b), respectively. Condition 3.8(c) can be validated in a manner analogous to 3.1(c), utilizing assumption A8(b). Assumption

A7 validates condition 3.9. Applying Theorem 2.8.1 from Lehmann (2004) in conjunction with assumptions A1(a), A1(b), and A8(c), we can show that the central limit theorem applies to the array $\{n^{-1/2}\nabla \log\{f_t(Y_t;\boldsymbol{\theta}_0)\}\}_{t=1}^{n}$, with covariance matrix B_{n0}. Consequently, condition 6.1 in White (1996) is met.

Bibliography

Chen, X. and Fan, Y. (2006). Estimation of copula-based semiparametric time series models. *Journal of Econometrics*, 130(2):307–335.

Chen, X., Wu, W. B., Yi, Y., et al. (2009). Efficient estimation of copula-based semiparametric Markov models. *The Annals of Statistics*, 37(6B):4214–4253.

Jia, Y., Kechagias, S., Livsey, J., Lund, R., and Pipiras, V. (2021). Latent Gaussian count time series. *Journal of the American Statistical Association*, pages 1–28.

Lehmann, E. L. (2004). *Elements of large-sample theory.* Springer Science & Business Media.

Longla, M. and Peligrad, M. (2012). Some aspects of modeling dependence in copula-based markov chains. *Journal of Multivariate Analysis*, 111:234–240.

Pipiras, V. and Taqqu, M. S. (2017). *Long-range dependence and self-similarity*, volume 45. Cambridge university press.

Pötscher, B. M. and Prucha, I. R. (1989). A uniform law of large numbers for dependent and heterogeneous data processes. *Econometrica: Journal of the Econometric Society*, pages 675–683.

Tang, Y., Wang, H. J., Sun, Y., and Hering, A. S. (2019). Copula-based semiparametric models for spatiotemporal data. *Biometrics*, 75(4):1156–1167.

White, H. (1996). *Estimation, Inference and Specification Analysis*. Cambridge University Press.