TACTiS-2: Better, Faster, Simpler Attentional Copulas for Multivariate Time Series

Arjun Ashok[1,2,3] *Étienne Marcotte[1] *Valentina Zantedeschi[1] †Nicolas Chapados[1,2] †Alexandre Drouin[1,2]

[1]ServiceNow Research [2]Mila-Quebec AI Institute [3]Université de Montréal *Equal Contribution †Equal Contribution

Summary

Problem: Multivariate Probabilistic Time Series Prediction, i.e., estimating the joint distribution of high-dimensional multivariate time series

General-Purpose Models: We seek models that support

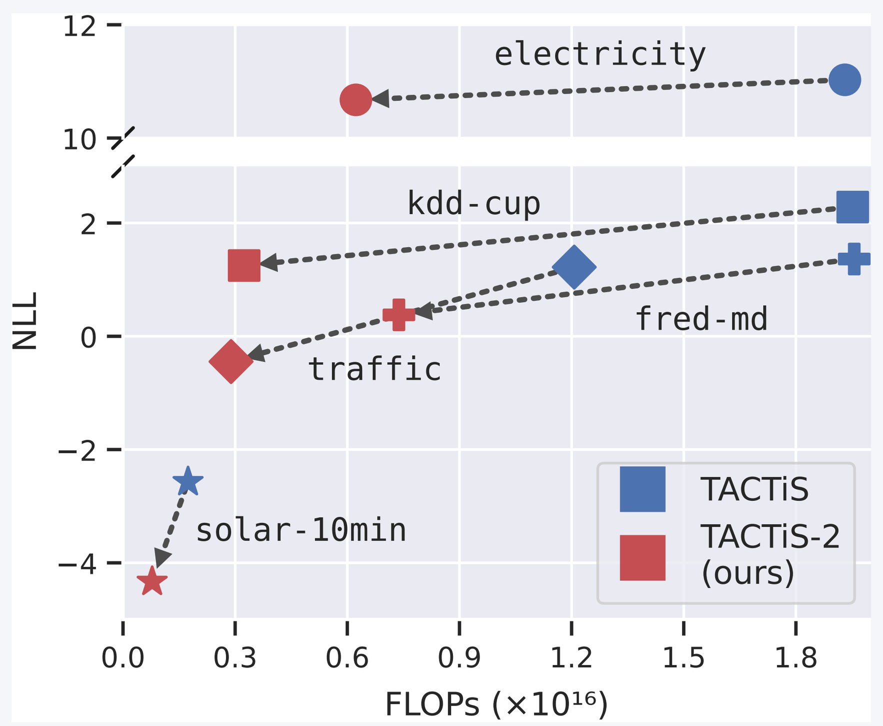
- ☑ Arbitrarily complex data distributions
- ☑ Heterogeneous/irregular sampling frequencies
- ☑ Hundreds of time series with missing data
- ☑ Deterministic covariates for conditioning (e.g., holidays)
- ☑ Tasks: forecasting, interpolation, and hybrids

Contributions:

- We show that Transformer Attentional Copulas for Time Series (TACTiS) [1], while flexible, are highly inefficient
- We propose a simpler and faster approach to learning valid attentional copulas and prove its correctness
- We show that this results in significantly better training dynamics and empirical results on real-world datasets

Main Takeaway

TACTiS-2 obtains better performance with lesser compute in real-world forecasting tasks, compared to TACTiS



Problem setting

- **Multivariate time series:** a collection of univariate time series $\mathbf{X} \stackrel{\text{def}}{=} \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$, where each $\mathbf{X}_i \stackrel{\text{def}}{=} [X_{i1}, \dots, X_{i,\ell_i}]$ is a random vector representing ℓ_i observations of some real-valued process in time
- **Additional data:** for any realization $\mathbf{x}_i \stackrel{\text{def}}{=} [x_{i1}, \dots, x_{i,\ell_i}]$ of $\mathbf{X}_i$, each x_{ij} is paired with:
  - a timestamp, $t_{ij} \in \mathbb{R}$ marking its measurement time
  - a vector of covariates $\mathbf{c}_{ij} \in \mathbb{R}^p$ that represents arbitrary additional information available
- **Learning Tasks:** defined with the help of a mask $m_{ij} \in \{0, 1\}$, which determines if any X_{ij} should be considered as observed ($m_{ij} = 1$) or to be inferred ($m_{ij} = 0$)
- **Goal:** estimating the joint distribution of missing values ($m_{ij} = 0$), given the observed ones ($m_{ij} = 1$), covariates, and timestamps:

$$P\Big(\underbrace{\mathbf{X}^{(m)}}_{\text{Missing values}} \;\Big|\; \underbrace{\mathbf{X}^{(o)},}_{\text{Observed values}} \underbrace{\mathbf{C}^{(m)}, \mathbf{C}^{(o)},}_{\text{Covariates}} \underbrace{\mathbf{T}^{(m)}, \mathbf{T}^{(o)}}_{\text{Timestamps}}\Big)$$

What is a copula?

Informally: a mathematical construct that expresses the coupling (dependency structure) of multiple random variables, irrespective of their marginal distributions (individual structure)

According to Sklar's theorem [3], the joint CDF of any random vector $[X_1, \dots, X_d]$ can be expressed as combination of:

- The marginal CDF of each random variable $F_i(x_i) \stackrel{\text{def}}{=} P(X_i \le x_i)$,
- The copula: a distribution on the unit cube with CDF $C : [0,1]^d \to [0,1]$ and $U_{[0,1]}$ marginals

$$P(X_1 \le x_1, \dots, X_d \le x_d) = C\Big(F_1(x_1), \dots, F_d(x_d)\Big)$$

Improved Learning of Non-Parametric Copulas

Copula-Based Density Estimators [1]: Joint density decomposed into marginals and copula:

$$g_{\boldsymbol{\phi}}\Big(x_1, \dots, x_d\Big) \stackrel{\text{def}}{=} c_{\phi_c}\Big(F_{\phi_1}\big(x_1\big), \dots, F_{\phi_d}\big(x_d\big)\Big) \times f_{\phi_1}\big(x_1\big) \times \cdots \times f_{\phi_d}\big(x_d\big) \tag{1}$$

$\boldsymbol{\phi} = \{\phi_1, \dots, \phi_d; \phi_c\}$ where $\{\phi_i\}_{i=1}^d$ are parameters of the marginal distributions, and ϕ_c are the parameters of the copula density c_{ϕ_c} estimated by minimizing negative log-likelihood:

$$\arg\min_{\boldsymbol{\phi}} \; - \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathbf{X}} \log g_{\boldsymbol{\phi}}(x_1, \dots, x_d) \tag{2}$$

Proposition 1: Problem (2) has infinitely many invalid solutions wherein c_{ϕ_c} is not the density function of a valid copula.

! The true marginals and copula can be entangled → Non-trivial to learn valid non-parametric copula-based density estimators.

Permutation-based Nonparametric Copulas (TACTiS): Nonparametric copulas learned using a permutation-based objective.

Considers an autoregressive factorization of c_{ϕ_c} according to an arbitrary permutation of the variables $\pi = [\pi_1, \dots, \pi_d]$:

$$c_{\boldsymbol{\phi}_c^\pi}(u_1, \dots, u_d) \stackrel{\text{def}}{=} c_{\phi_{c,1}^\pi}(u_{\pi_1}) \times c_{\phi_{c,2}^\pi}(u_{\pi_2} \mid u_{\pi_1}) \times \cdots \times c_{\phi_{c,d}^\pi}(u_{\pi_d} \mid u_{\pi_1}, \dots, u_{\pi_{d-1}}) \tag{3}$$

where $u_{\pi_k} = F_{\phi_{\pi_k}}\big(x_{\pi_k}\big)$. Optimizes a permutation-based objective over Π, where Π is the set of all $d!$ permutations:

$$\arg\min_{\phi_1, \dots, \phi_d, \phi_c^\pi} \; - \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathbf{X}} \mathop{\mathbb{E}}_{\pi \sim \Pi} \log c_{\phi_c^\pi}\Big(F_{\phi_1}\big(x_1\big), \dots, F_{\phi_d}\big(x_d\big)\Big) \times f_{\phi_1}\big(x_1\big) \times \cdots \times f_{\phi_d}\big(x_d\big) \tag{4}$$

✗ **Limitations:** The model needs the capacity to fit all d! permutations → results in slow convergence and sub-optimal solutions.

Two-Stage Nonparametric Copulas (ours): Learn marginal parameters (eq. (6)), then learn copula parameters (eq. (5)):

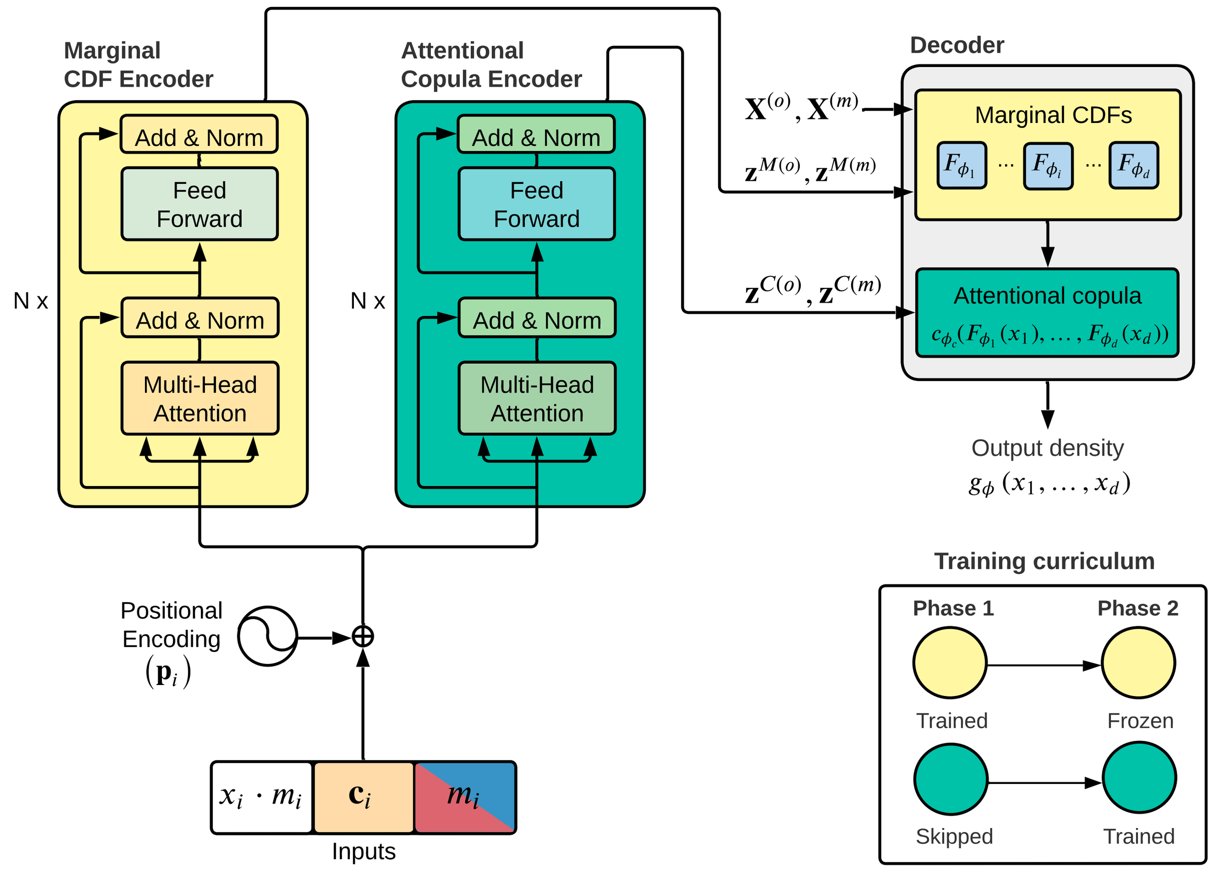
$$\arg\min_{\phi_c} \; - \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathbf{X}} \log c_{\phi_c}\Big(F_{\phi_1^\star}(x_1), \dots, F_{\phi_d^\star}(x_d)\Big) \tag{5}$$

$$s.t. \quad (\phi_1^\star, \dots, \phi_d^\star) \in \arg\min_{\phi_1, \dots, \phi_d} - \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathbf{X}} \log \prod_{i=1}^{d} f_{\phi_i}(x_i) \tag{6}$$

Proposition 2: Solving Problem (5) yields a solution to Problem (2) where c_{ϕ_c} is a valid copula. Proof builds on Sklar's theorem [3].

✓ **Advantages:** The model needs to fit just 1 permutation → simpler objective with faster convergence to better solutions.

Putting Theory into Practice



Architecture of TACTiS-2: The two encoders serve to parametrize the decoder in the two stages of the training curriculum (bottom right). Phase 1 solves Problem (6), while Phase 2 solves Problem (5).

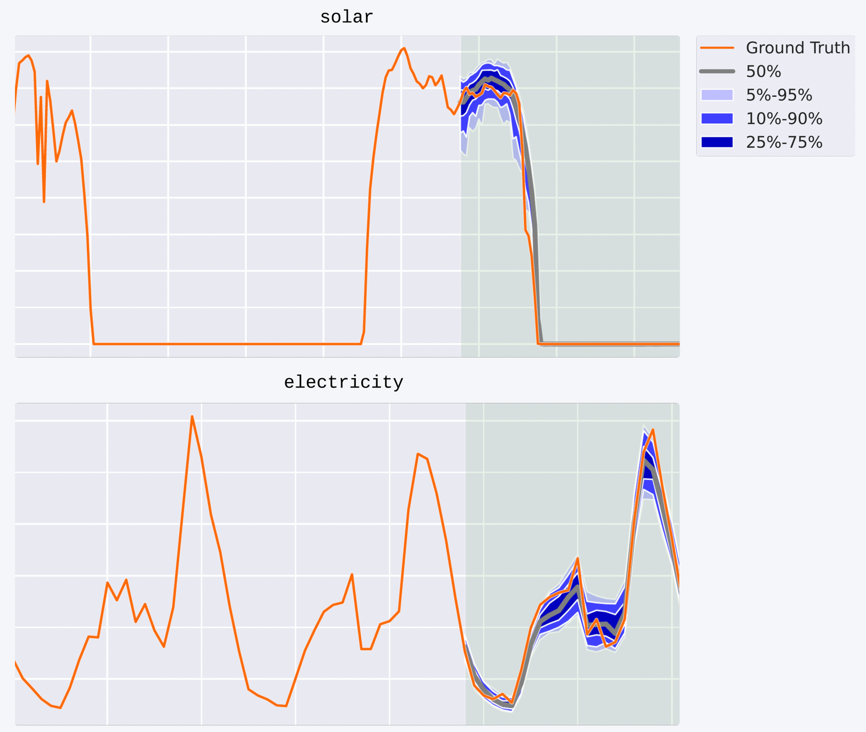
Results

Mean CRPS-Sum values for the forecasting experiments (± standard errors). Lower is better. Best results in bold.

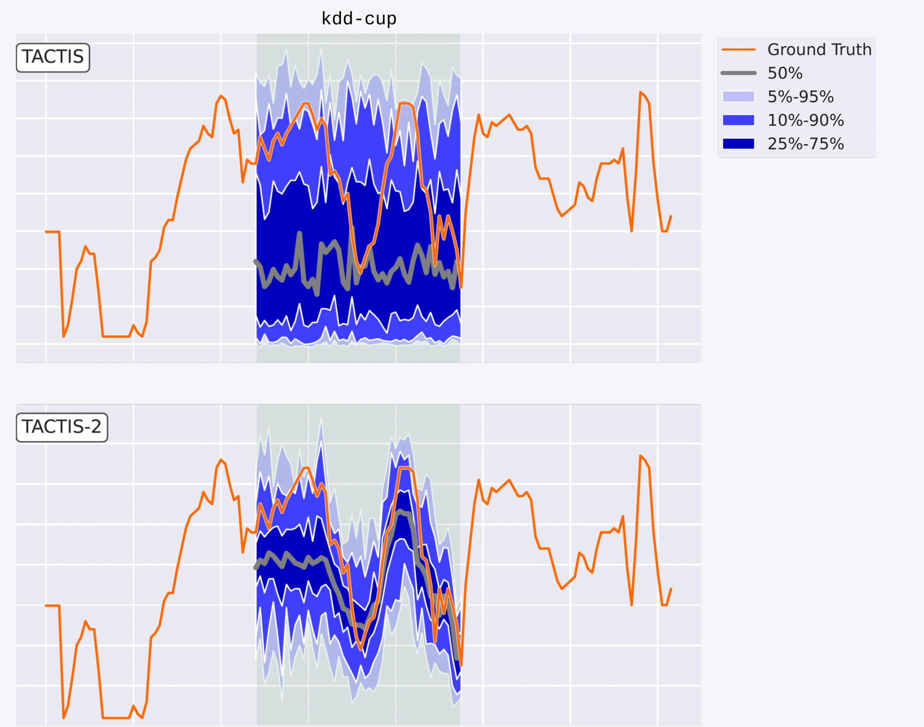
Model	electricity	fred-md	kdd-cup	solar-10min	traffic	Avg. Rank
Auto-ARIMA	0.077 ± 0.016	0.043 ± 0.005	0.625 ± 0.066	0.994 ± 0.216	0.222 ± 0.005	6.2 ± 0.3
ETS	0.059 ± 0.011	0.037 ± 0.010	0.408 ± 0.030	0.678 ± 0.097	0.353 ± 0.011	6.0 ± 0.2
TempFlow	0.075 ± 0.024	0.095 ± 0.004	0.250 ± 0.010	0.507 ± 0.034	0.242 ± 0.020	5.4 ± 0.2
SPD	0.062 ± 0.016	0.048 ± 0.011	0.319 ± 0.013	0.568 ± 0.061	0.228 ± 0.013	5.2 ± 0.3
TimeGrad	0.067 ± 0.028	0.094 ± 0.030	0.326 ± 0.024	0.540 ± 0.044	0.126 ± 0.019	5.0 ± 0.2
GPVar	0.035 ± 0.011	0.067 ± 0.008	0.290 ± 0.005	0.254 ± 0.028	0.145 ± 0.010	3.8 ± 0.2
TACTiS	0.021 ± 0.005	0.042 ± 0.009	0.237 ± 0.013	0.311 ± 0.061	**0.071 ± 0.008**	2.4 ± 0.2
TACTiS-2	**0.020 ± 0.005**	**0.035 ± 0.005**	**0.234 ± 0.011**	**0.240 ± 0.027**	0.078 ± 0.008	**1.9 ± 0.2**

Mean NLL values for forecasting experiments and training FLOP counts (± standard errors). Lower is better. Best results in bold. .

Model		electricity	fred-md	kdd-cup	solar-10min	traffic
TACTiS	NLL	11.028 ± 3.616	1.364 ± 0.253	2.281 ± 0.770	−2.572 ± 0.093	1.249 ± 0.080
	FLOPs (×10¹⁶)	1.931 ± 0.182	1.956 ± 0.192	1.952 ± 0.208	0.174 ± 0.018	1.207 ± 0.517
TACTiS-2	NLL	**10.674 ± 2.867**	**0.378 ± 0.076**	**1.055 ± 0.713**	**−4.333 ± 0.181**	**−0.358 ± 0.077**
	FLOPs (×10¹⁶)	**0.623 ± 0.018**	**0.738 ± 0.022**	**0.324 ± 0.014**	**0.078 ± 0.005**	**0.289 ± 0.061**



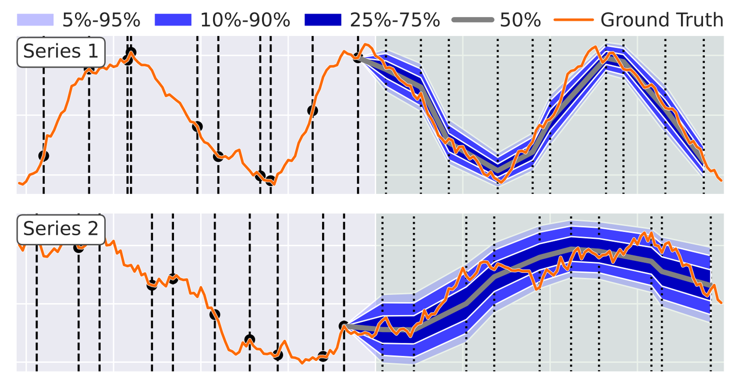
Example forecasts of TACTiS-2



Example interpolations by TACTiS (top) and TACTiS-2 (bottom)

Model flexibility

TACTiS-2 maintains the flexibility of TACTiS



Example forecasts of TACTiS-2 on irregular and unevenly sampled data

References

[1] Alexandre Drouin, Étienne Marcotte, and Nicolas Chapados. TACTiS: Transformer-attentional copulas for time series. In *ICML*, 2022.

[2] David Salinas, Michael Bohlke-Schneider, Laurent Callot, Roberto Medico, and Jan Gasthaus. High-dimensional multivariate forecasting with low-rank Gaussian copula processes. *NeurIPS*, 2019.

[3] Abe Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publ. de l'Institut Statistique de l'Université de Paris*, 8:229–231, 1959.