cally demonstrate the proposed method's effectiveness.
4. Through extensive testing, we show the newly proposed likelihood objective achieves SOTA performance on multiple real-world datasets.

2 Problem Statement

At a high level, the general problem is defined in a multivariate time series setting with the goal of modeling the joint distribution of unobserved values at particular time points conditioned on other observed values. Specifically, let $\mathbf{X}$ be a set of s univariate time series where each $\mathbf{X_i}$ is a vector containing d_i observations such that $\mathbf{X_i} \stackrel{\text{def}}{=} [X_{i1}, \ldots, X_{i,d_i}]$. For each realisation $\mathbf{x_i} \stackrel{\text{def}}{=} [x_{i1}, \ldots, x_{i,d_i}]$, there exists:

1. A Boolean value m_{ij} where $m_{ij} = 1$ if x_{ij} is observed or else $m_{ij} = 0$.
2. An associated set containing p time varying covariates or additionally available information (such as weather indicators) at each time step represented as $\mathbf{C_i} \stackrel{\text{def}}{=} [c_{i,1}, \ldots, c_{i,d_i}] \in \mathbb{R}^{p \times d_i}$.
3. $\mathbf{t_i} \stackrel{\text{def}}{=} [t_{i,1}, \ldots, t_{i,d_i}] \in \mathbb{R}^{d_i}$ being a vector of time stamps such that $t_{ij} < t_{i,j+1}$.

Then, if $\mathbf{X_i^{(m)}} = [X_{i,1}^{(m)}, \ldots, X_{i,s_m}^{(m)} \mid m_{i,j} = 0]$ and $\mathbf{X_i^{(o)}} = [X_{i,1}^{(o)}, \ldots, X_{i,s_o}^{(o)} \mid m_{i,j} = 1]$ represent the missing and observed components of $\mathbf{X_i}$, the problem of modelling the joint distribution can be written to be conditioned on all known information as the following:

$$P(\{\mathbf{X_i^{(m)}}\}_{i=1}^{s} | \{\mathbf{X_i^o}, \mathbf{C_i}, \mathbf{t_i}\}_{i=1}^{s}) \tag{1}$$

Note the formulation of time series problems is achieved by masking appropriately. For example setting the last t elements of each m_i to 0 conducts a t-step ahead probabilistic forecasting task.

Copulas

Now let n be the total number of random variables modelled within the joint distribution. We first redefine $\mathbf{X} = [X_1, \ldots, X_n]$ as a vector of n random variables. Then, following the work in TACTiS [4, 1] and Sklar's theorem [19], the joint multivariate CDF of any random vector can be modelled using a copula $C : [0, 1]^n \to [0, 1]$ and each random variable's marginal univariate CDF $F_i(x_i) \stackrel{\text{def}}{=} P(X_i \leq x_i) = u_i$. This allows the CDF of $\mathbf{X}$ to be expressed as:

$$P(X_1 \leq x_1, \ldots, X_n \leq x_n) = C(F_1(x_1), F_2(x_2), \ldots, F_n(x_n)) \tag{2}$$

Note that C represents the n dimensional joint CDF on the unit cube with uniform distributed marginals on [0, 1]. Specifically,

$$C(u_1, u_2, \ldots, u_n) \stackrel{\text{def}}{=} P(U_1 \leq u_1, U_2 \leq u_2, \ldots, U_n \leq u_n) \tag{3}$$

where $U_i \sim U_{[0,1]}$. We then construct the copula-based density function $g_\phi(x_1, \ldots, x_n)$ with parameters ϕ conditioned on observed variables:

$$g_\phi(x_1, \ldots, x_n) \stackrel{\text{def}}{=} c_{\phi_c}(F_{\phi_1}(x_1), \ldots, F_{\phi_n}(x_n)) \cdot \prod_{i=1}^{n} f_{\phi_i}(x_i) \tag{4}$$

f_i represents the marginal PDF of X_i while ϕ_i and ϕ_c represent the marginal and copula distributional parameters respectively.

Subsequently, we substitute the generic random vector $\mathbf{X}$ for $\mathbf{X_i^{(m)}}$ and obtain the final joint distribution of **unobserved** values conditioned on other observed values:

$$g_\phi(x_1^{(m)}, \ldots, x_{n_m}^{(m)}) \stackrel{\text{def}}{=} c_{\phi_c}(F_{\phi_1}(x_1^{(m)}), \ldots, F_{\phi_{s_m}}(x_{s_m}^{(m)})) \cdot \prod_{i=1}^{s_m} f_{\phi_i}(x_i^{(m)}) \tag{5}$$

Ultimately, this results in a likelihood optimization problem defined as the following

$$\underset{\Theta}{\operatorname{argmin}} \underset{x \sim X}{E} - \log g_\phi(x_1^m, \ldots, x_{s_m}^{(m)}) \tag{6}$$

This paper specifically aims to propose a new method that improves on this likelihood objective in two ways. Firstly, compared to existing methods we redefine a new non-parametric copula density function c_{ϕ_c} that has continuous gradients during optimization. Secondly, by incorporating a regularization term into Equation(6), we hope to introduce steeper gradients in regions that were previously flat within the original likelihood objective when being stuck in local minima. These adjustments aims to reduce instabilities and accelerate the convergence process when limited input data, model capacity, and training time prevents forbid the theoretical guarantee from being achieved.

3 Related Work

Traditional Methods

The origins of time series trace all the way back to the early 1920s when [26, 27, 22] adopted stochastic processes and autoregressive concepts for time series analysis. It relied on the idea of stationarity because only then could the correlation structure be consistent over time. Models proposed during that period include the classic AutoRegressive Integrated Moving Average (ARIMA) perfected by [2], SARIMA [11] that additionally accounted for seasonality, and the Holts-Winters seasonal method [24] that utilised exponential smoothing. Multivariate variants of these models included the Vector Autoregressive Model (VAR) [18] that started accounting for endogenously interrelated feature interaction. The Generalized Autoregressive Conditional Heteroskedasticity (GARCH) [5] was also popular through its introduction of co-integration where stationarity was obtained by linearly combining non-stationary processes.

While relatively simplistic, the robustness, explainability, and low computation costs of these methods allow them to still see use today. However, with modern data being increasingly sophisticated with complex temporal dependencies, the effectiveness of these methods begin the fall short.

Machine Learning Methods

To improve on traditional methods, machine learning models were introduced that focused on self-optimization from within a large feature space. The time-delay neural network (TDNN) [20] was one such model using feed forward neural networks with added time delays to the inputs within each layer. The recurrent neural network (RNN) [16] class of models were also introduced that allowed the output of the model at a given time step to be used within the input for the next time step. This 'memory' mechanism theoretically enables RNNs to capture longer-term dependencies. However it suffered from vanishing gradients caused by the repeated product of derivatives less than one resulting in gradients of early periods being exponentially small. Long-short-term memory (LSTM) [8]