Error Encoding Feedback with Cross-variable Correlations for Multivariate Time Series Forecasting

Qiang Liu[1], Fu-Lee Wang[2], Xi-Zhao Wang[1]

[1]Big Data Institute, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, 518060, China
[2] School of Science and Technology, Hong Kong Metropolitan University, Hong Kong, 999077, China
Corresponding E-mail: xizhaowang@ieee.org

Abstract:

In multivariate time series forecasting, prediction errors can provide key information for both training and reasoning of the model. Accurately modeling the correlation structure of errors is critical for reliable uncertainty quantification. Existing modeling approaches often assume independence across different dimensions of errors for simplicity. However, this assumption ignores the cross-variable correlations in the multivariate time series, thereby seriously limiting the expressive capacity of uncertainty modeling. In this paper, prediction errors are treated as a random vector, and the impact of the correlation among these components of the vector on prediction performance is examined in formulation and experimentally. Furthermore, residual errors are projected into a low-dimensional latent space for back propagation to the decoder to avoid trivial solutions in the modeling process. The experimental results show that explicitly considering and modeling the cross-variable correlation of errors can effectively improve the stochastic model's performance, and mapping errors to a low-dimensional latent space avoids the problem of trivial solutions problem to some extent and enhances model stability.

Keywords:

Error encoding feedback; Latent variable; Cross-variable correlations; Time series forecasting

1. Introduction

Error modeling is crucial in statistical time series models. When a model can utilize all the correlations between consecutive time steps determined by previous values, the error is temporally uncorrelated. However, real-world data often violate this assumption, with errors exhibiting significant cross-correlation [1, 2]. Prediction errors can not only reflect model performance but also provide crucial guidance for model training, uncertainty estimation, and inference processes [3, 4]. Accurate modeling of errors directly determines the reliability of uncertainty quantification and the credibility of forecasting results, which is of critical importance in risk-aware scenarios such as industrial monitoring, financial analysis, and intelligent transportation.

In the field of error modeling for time series, the classic approaches generally assume that the error sequence follows a correlated time process, such as the ARIMA model [5]. However, in the deep learning framework, in order to simplify computational complexity, most deep learning probabilistic prediction methods generally adopt a simplified assumption, i.e., the prediction errors in different channels are independent, and even ignore the temporal correlation of the errors themselves [6].

To address this shortcoming, some existing studies have attempted to model the temporal correlation of errors. Sun et al. [7] considered the temporal autocorrelation of errors, jointly learned the error autocorrelation coefficients with model parameters, and effectively improved the prediction performance of the model; Zheng et al. [8] further adopted a matrix-variate autoregressive (AR) model to characterize the errors. In addition, many researchers also have focused on the temporal correlation of errors in the process of error modeling and proposed corresponding improved methods [e.g., 9, 10, 11].

In multivariate time series forecasting, the forecast error often includes both temporal correlations and the correlations among variables. Most existing methods focus only on the assumption of temporal dependence of errors, while typically ignoring their spatial correlation(i.e.,cross-variable correlation). That is, they assume that errors on different channels are independent, or that errors are completely independent in temporal and spatial, and use a diagonal covariance matrix for approximate modeling the errors. In real-world multivariate time series, channel variables naturally exhibit dependencies, and ignoring statistical dependencies between variables may limit the model's ability to express and characterize uncertainty, thereby reducing the accuracy of probabilistic predictions.

Based on the above-mentioned observations, this paper addresses the problem that most of existing methods generally ignore the cross-variable correlation of errors and are difficult to explore the cross-variable error dependence. By reasonably setting the forecasting horizon $H = 1$ (i.e., one-step prediction) in the experiment, this setting can effectively eliminate the temporal dependence and specifically explore the impact of cross-variable correlation of errors on multivariate time series prediction, thus clarifies the role of cross-variable correlation in error modeling. Meanwhile, for the trivial solution problem in error modeling [12], this paper also conducts discussions and experimental explorations. The main contributions of this paper can be summarized as follows:

1. We focus on exploring the specific impact of the cross-variable correlation of errors on the forecasting performance of multivariate time series. The result shows that leveraging cross-variable correlation of the model can improve the predictive performance.

2. We further experimentally analyze the trivial solution problem during the error modeling and demonstrate the necessity of mapping errors to low-dimensional variables.

3. By designing targeted experiments and constructing specialized experimental datasets, our proposed methods and core viewpoints have been fully validated.

The rest of this paper is organized as follows. Section 2 introduces the related work, section 3 is the methodology, section 4 provides an experimental validation, and section 5 gives the conclusions.

2 Related Work

Error modeling has always been a core research topic in the field of statistical time series analysis, and error correlation in time series has been widely studied in statistics [13].

In univariate time series, traditional methods usually assume that the prediction errors follow a stochastic process with temporal dependence, a typical example being the ARIMA model [5]. The Error Correction Model (ECM) [14] improves prediction accuracy by modeling the first-order autocorrelation of errors between predicted results and observed values. DeepAR [6] models errors by assuming that errors are independent over time during probabilistic prediction.

For multivariate time series prediction, most researches still assumes that error variables are independent in the temporal dimension. Choi et al. [9] modeled error variables as a dynamic mixture model of a zero-mean Gaussian distribution. Kim et al. [15] designed a learning module to calibrate the model prediction error, which takes the error and predicted value as input to predict future residual. Jia et al. [16] suggested that there is no reason to assume independence in the errors. Zheng et al. [17] assumed that error terms follow a matrix AR process with seasonal lags, further enhancing the ability to model the temporal correlation of errors. In addition, Zheng et al. [2] considered the cross-variable correlation pattern of errors, but their research focus is still on the modeling of temporal correlation, without exploration of cross-variable correlation of errors.

Although all the above methods claim to achieve effective modeling of errors, there are obvious limitations. Specifically, in statistics, only the errors of well-trained models satisfy the white noise assumption; however, current methods generally fail to achieve this. Henaff et al. [12] proposed the error encoding network (EEN), and their method employed a two-stage training procedure. First, the model was trained and its prediction errors were computed. Then, to avoid trivial solutions, these errors were encoded into low-dimensional latent variables, and the model parameters were further optimized with the aid of random variables. However, this method does not explicitly consider the statistical characteristics of errors, making it difficult to give full play to the guiding role of error correlation in model optimization.

In summary, there are still significant deficiencies in existing research on the correlation of time series prediction errors. This paper focuses on the correlation modeling of statistical errors, clearly characterizes the error distribution characteristics, explores the impact of cross-variable correlation of errors on model prediction performance, and confirms the impact of low-dimensional embedded variables on error coding backpropagation.

3 Methodology

In this section, we give the formulation of the time series forecasting problem, and then propose our method for modeling errors.

3.1 Problem formulation

Time series forecasting uses probabilistic and statistical methods to process a sequence of random vectors, and its essence lies in analyzing the observations of these random vectors (i.e., real-valued vector sequences). This type of modeling and analysis task based on observational data has been widely applied in many fields such as finance, medicine, and bio-informatics [e.g., 18, 19] after the data are transferred into the sequences.

The objective of the forecasting problem is to build an effective model which can accurately predict the values of the $(n+1)$th step (or more) based on the n historical observations of the series. Given a multivariate time series $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\} \in R^{D\times N}$ representing the measurements of D variables (or channels). We express the the forecasting model $\mathcal{F}_\theta$ as

$$\mathcal{Y}_t = \mathcal{F}_\theta(\mathcal{X}_t) + \boldsymbol{\varepsilon}_t \tag{1}$$

where $\mathcal{X}_t = \boldsymbol{X}_{t-L+1:t} = \{\boldsymbol{x}_{t-L+1}, \ldots, \boldsymbol{x}_t\} \in R^{D\times L}$ and $\mathcal{Y}_t = \boldsymbol{Y}_{t+1:t+H} = \{\boldsymbol{x}_{t+1}, \ldots, \boldsymbol{x}_{t+H}\} \in R^{D\times H}$ are the t-th training sample and its corresponding target, respectively. L represents the look-back window length, and H is the forecasting horizon length. The training samples $(\mathcal{X}_t, \mathcal{Y}_t), t = 1, 2, \cdots, n$ are constructed using a sliding window mechanism, and $\mathcal{X}_t$ and $\mathcal{Y}_t$ are temporally adjacent. $\widehat{\mathcal{Y}}_t = \mathcal{F}_\theta(\mathcal{X}_i)$ represents the prediction output. θ denotes the learnable parameters. $\boldsymbol{\varepsilon}_t = \widehat{\mathcal{Y}}_t - \mathcal{Y}_t$ is the residual prediction error, which is usually considered as a set of realized values of the random vector $\boldsymbol{r} \in R^{D\times H}$. It is well acknowledged that $\boldsymbol{\varepsilon}_t$ is usually considered as a non-interpretable part of output with respect to observed inputs which represents a type of irreducible noise with respect to prediction tasks. This noise is regarded as a source of uncertainty.

In multivariate time series forecasting, variables typically exhibit both temporal correlation across time steps (i.e., lag-covariance $\mathrm{Cov}(r_{i,h-\triangle}, r_{i,h}) \neq 0, 1 \leq h \leq H, 1 \leq i \leq D, \triangle = 1, 2, 3, \cdots H-1$) and spatial correlations correlation (i.e., cross-covariance $\mathrm{Cov}(r_{i,h}, r_{j,h}) \neq 0, 1 \leq j \leq D, i \neq j$ and cross-lag covariance $\mathrm{Cov}(r_{i,h-\triangle}, r_{j,h}) \neq 0$). For multi-step prediction tasks $H > 1$, accurately and perfectly accurate modeling the correlation of errors is highly complex. Generally, researchers often made some assumptions, such as $\mathrm{Cov}(r_{i,h-\triangle}, r_{i,h}) = 0$ [5, 6, 14], or only consider the impact of errors when $\mathrm{Cov}(r_{i,h-\triangle}, r_{i,h}) \neq 0$ [2, 16, 17].

In this paper, we primarily analyze the impact of cross-covariance $\mathrm{Cov}(r_{i,h}, r_{j,h}) \neq 0$ on error modeling. We set the prediction horizon to $H = 1$ (i.e., single-step prediction) to focus on the statistical cross-variable correlation of errors. Then, let $\boldsymbol{r}$ denote a D-dimensional random vector of the prediction errors with its realization $\boldsymbol{\varepsilon}_t \in R^D$ and suppose

$$\boldsymbol{r} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \tag{2}$$

a normal distribution with mean being zero vector $\mathbf{0} \in R^D$ and covariance matrix $\boldsymbol{\Sigma} \in R^{D\times D}$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathrm{Var}(r_1) & \mathrm{Cov}(r_1, r_2) & \cdots & \mathrm{Cov}(r_1, r_D) \\ \mathrm{Cov}(r_2, r_1) & \mathrm{Var}(r_2) & \cdots & \mathrm{Cov}(r_2, r_D) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(r_D, r_1) & \mathrm{Cov}(r_D, r_2) & \cdots & \mathrm{Var}(r_D) \end{bmatrix} \tag{3}$$

where $\mathrm{Cov}(r_i, r_j), (1 \leq i, j \leq D)$ represents the covariance of variables r_i and r_j. r_i and r_j are the i-th and j-th component of $\boldsymbol{r}$, respectively. The error distribution Eq. (2) can be used to retrain or calibrate $\mathcal{F}_\theta$ and improve prediction performance.

When $H = 1$, the temporal dependence of errors is naturally eliminated, thus excluding the interference of temporal correlations and allowing for a more accurate study of the impact of cross-variable error correlations on model's uncertainty and prediction performance.

3.2 Error Encoding Network

We introduce a distribution-based Error Encoding Network (dEEN) into $\mathcal{F}_\theta$ to model the error variable $\boldsymbol{r}$. The model architecture used in this paper is shown in Figure. 1.

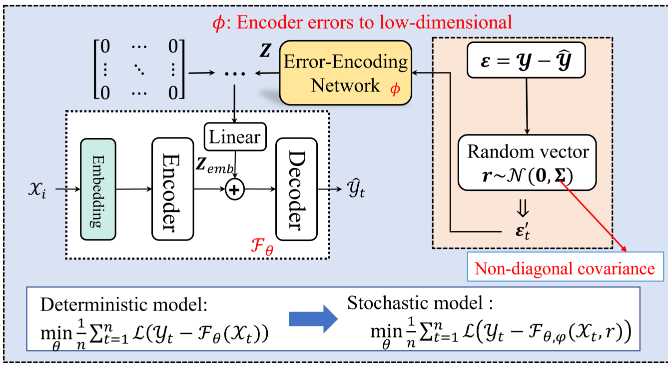

FIGURE 1. The Model Architecture

In our study, the process of modeling the error variable $\boldsymbol{r}$ can be classified into 3 steps. The 1st step is to choose an appropriate model $\mathcal{F}_\theta$ with $\mathcal{X}_t$ and $\mathcal{Y}_t$ as input and output, respectively. In this paper, we select the Transformer [20] as the basic deterministic model $\mathcal{F}_\theta$ since the Transformer has been well known as one of the most powerful tools for handling sequence learning. The deterministic model $\mathcal{F}_\theta$ is trained using $(\mathcal{X}_t, \mathcal{Y}_t)$, primarily through the following optimization objectives.

$$\min_\theta \frac{1}{n} \sum_{t=1}^{n} \mathcal{L}\left(\mathcal{Y}_t, \mathcal{F}_\theta(\mathcal{X}_t)\right) \tag{4}$$

where $\mathcal{L}$ represents the loss function, n is the number of training samples. In this step, the model is deterministic ($Z_{emb} = 0$), and its main purpose is to minimize the training loss and obtain the desired errors.

The 2nd step is to estimate the parameters of random vector r defined in Eq.(3). The residual prediction error is computed as:

$$\boldsymbol{\varepsilon}_t = \mathcal{Y}_t - \widehat{\mathcal{Y}}_t, \qquad t = 1, \ldots, n. \tag{5}$$

Collectively, the residual

$$\boldsymbol{\varepsilon} = [\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \cdots, \boldsymbol{\varepsilon}_n] \in R^{D\times n} \tag{6}$$

The parameters of $\boldsymbol{r}$ are estimated as:

$$\mu_i = \frac{1}{n}\sum_{t=1}^{n} \varepsilon_t^{(i)}, , i = 1, 2, \cdots, D \tag{7}$$

$$\mathrm{Cov}(r_i, r_j) = \frac{1}{n-1}\sum_{t=1}^{n}\left(\varepsilon_t^{(i)} - \hat{\mu}_i\right)(\varepsilon_t^{(j)} - \hat{\mu}_j)^\top \tag{8}$$

where μ_i and $\mathrm{Cov}(r_i, r_j)$ are the element of the vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Then, we can obtain a input $\boldsymbol{\varepsilon}_t'$ of dEEN, and $\boldsymbol{\varepsilon}_t'$ is a realization from the random vector $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

The 3rd step is to adjust the model parameters obtained on Step 1 together with the random vector r estimated in step 2. We enlarge the training set $(\mathcal{X}_t, \mathcal{Y}_t)$ into $((\mathcal{X}_t, \boldsymbol{r}), \mathcal{Y}_t)$ and denote the stochastic model by $\mathcal{F}_{\theta,\phi}(\cdot)$ where the newly added parameter vector ϕ is from the network of encoding residual $\boldsymbol{r}$, and $Z_{emb} = \boldsymbol{Z}$. Then, we try to tune the parameters θ to solve the following stochastic optimization problem

$$\min_{\theta} \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}\left(\mathcal{Y}_t, \mathcal{F}_{\theta,\phi}(\mathcal{X}_t, \boldsymbol{r})\right) \tag{9}$$

where r is the random vector defined in Eq.(3) and the other symbols have the same meaning as Eq.(4).

The Eq.(9) introduces two additional terms, i.e., the random vector of errors $\boldsymbol{r}$ and parameters ϕ of the dEEN, and the uncertainty caused by the error is incorporated into the optimization problem. Introducing the uncertainty information from $\boldsymbol{r}$ can enable the model to calibrate its outputs based on the correct pattern while simultaneously adjusting the decoder to propagate uncertainty information.

Here, we list two of our remarks about the model.

(1) For simplicity, when modeling error covariance in most research, it is common to assume $\mathrm{Cov}(r_i, r_j) = 0$ which is a popular approach when $H > 1$; and under this assumption, the covariance matrix takes the diagonal form $\boldsymbol{\Sigma} = \mathrm{diag}(\mathrm{Var}(r_1), \mathrm{Var}(r_2), \cdots, \mathrm{Var}(r_D))$. This paper explores the impact of two different forms (the diagonal covariance matrix and non-diagonal covariance matrix) on model performance. We hypothesize that using the non-diagonal covariance matrix to model the error might better improve model's performance. It may be because most cases of practice exhibit corrections among both time steps and variables, and the diagonal variance matrix is a special case of a general positive definite matrix.

(2). It is important to note that Henaff et.al [12] pointed out that $\boldsymbol{Z}$ needs to be designed as a low-dimensional latent variable to prevent the network from learning trivial solutions. The trivial mapping of dEEN means that the network can easily recover the target output by canceling the encoder representation without learning any meaningful structure of the prediction error. The trivial solution may lead to the failure of error modeling, and the model performance can not be improved.

4 Experiments

In this section, we conduct some experiments to verify our conclusions. We generate synthetic multivariate time series datasets and compare the effects of two types of covariance structures on error modeling. In addition, we discuss the problem of trivial solutions in the model.

4.1 Datasets Generation

We construct a function to generate a synthetic dataset, which is shown as:

$$\boldsymbol{y}_t = \boldsymbol{s}_t + \boldsymbol{\eta}_t, \tag{10}$$

where $\boldsymbol{s}_t \in R^D$ denotes the deterministic component, D is the number of channels, and $\boldsymbol{\eta}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_\eta)$ is a Gaussian white-noise vector. For simplicity, $\boldsymbol{\Sigma}_\eta$ is a diagonal matrix, and the same variance σ^2 is assigned to each diagonal entry.

$$\boldsymbol{s}_t = \boldsymbol{C}\left[\sin\left(2\pi\frac{t}{p} + 0.3(d-1)\right) + 0.5\cos\left(4\pi\frac{t}{p} + (d-1)\right)\right]_{d=1}^{D} \tag{11}$$

where $\boldsymbol{C} \in R^{D\times D}$ denotes a coupling matrix that introduces weak cross-variable dependencies into $\tilde{\boldsymbol{s}}_t$; $p = 24$ represents the hourly periodic patterns. The numbers $0.3(d-1)$ and $(d-1)$ ensure that there are distinct dynamic characteristics for different channels.

We generated three synthetic datasets with different noise variance assumptions, as follows:

$$\sigma^2 \in \{0.1, 0.3, 0.5\} \tag{12}$$

For each variance, a D-dimensional time series of length N is simulated, a larger σ^2 indicates that more noise is introduced into the data. Based on our previous results, the dEEN typically performs well when the noise variance $\sigma^2 < 0.5$. These synthetic datasets provide a hypothetical environment for our experimental exploration.

4.2 Experimental Details

This section details the experimental setup, including dataset splitting, comparison models, and training configuration. Specifically, we generated 5000 patterns ($N = 5000$) for the experiment. The dataset is divided into three parts: training, validation, and testing sets in chronological order with a split ratio of $7 : 1 : 2$. The input sequence length was $L = 96$. To investigate the trivial solution problem, we first set the dimension of the latent variable Z to 4. Subsequently, in the comparative experiments, we adjust the dimension to be consistent with the dimension of the Decoder input ($Z = 96$), so as to analyze the impact of different latent space dimensions on the trivial solution problem. For model training, we use the Adam optimizer, the initial learning rate is 10^{-4}, and the batch size is 32. The patience for early stopping to prevent overfitting is 10 . All experiments are performed on an NVIDIA V100 32GB GPU. The mean squared error (MSE) and mean absolute error (MAE) are adopted as basic metrics to assess the model's prediction performance.

4.3 Experimental Results

This section analyzes the proposed method experimentally from two perspectives: the impact of cross-variable correlation of errors and the exploration of trivial solutions.

After training the stochastic model, by fixing the model parameters and performing multiple independent inferences on some input sample, the stochastic model can conduct multiple prediction results, and we report the average MSE/MAE over 100 times independent testing. The results are shown in Table 1. Here, $Z = 96$ indicates that the dimension of the latent space Z is the same as the dimension of the Decoder's input, and $Z = 4$ indicates a low-dimensional latent variable space; "independent" assumes that the variables of erros are independent, i.e., the covariance matrix $\mathrm{Cov}(r_i, r_j) = 0$ (the covariance matrix is a diagonal matrix). "Correlated" refers to considering the cross-variable correlation, i.e., the covariance matrix $Cov(r_i, r_j) \neq 0$ (the covariance matrix is non-diagonal form).

First, we discuss the problem of trivial solutions regarding the model. As shown in Table 1, when $Z = 96$, we can see that the model's prediction performance is worse than the scenario of $Z = 4$. When $Z = 96$, the MSE is 0.532 and MAE is 0.583 in $\sigma^2 = 0.5$, while the MSE drops to 0.530 in the scenario of $Z = 4$. This phenomenon confirms that experimentally, to some extent, when the latent space dimension is the same as the Decoder's input dimension, the model may learn trivial solution which leads to a decrease in prediction performance. While constraining the latent space dimension to a low dimension can effectively avoid the trivial solution problem, allowing the model to focus more on learning the statistical characteristics of errors, thereby improving prediction performance. Furthermore, at different noise levels ($\sigma^2 = 0.1$, $\sigma^2 = 0.3$, $\sigma^2 = 0.5$), the prediction performance of the $Z = 4$ is better than that of the $Z = 96$, indicating that reasonably constraining the latent space dimension can effectively avoid the trivial solutions and improve model stability, which is perfectly consistent with the conclusion proposed by Henaff et al. ([12]) that "low-dimensional latent spaces can avoid trivial solutions".

TABLE 1. Experimental results on the Transformer under different error modeling configurations.

Synthetic Datasets	Deterministic model		Stochastic model					
			$Z = 96$		$Z = 4$			
			Independent		Independent		Correlated	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
$\sigma^2 = 0.1$	0.158	0.317	0.155	0.318	0.153	0.312	**0.152**	**0.311**
$\sigma^2 = 0.3$	0.372	0.487	0.364	0.484	**0.362**	0.482	**0.362**	**0.481**
$\sigma^2 = 0.5$	0.531	0.583	0.532	0.583	0.530	0.582	**0.529**	**0.580**

Then, we explore the impact of cross-variable correlation of errors on model's performance. It can be seen that, under the same latent space dimension ($Z = 4$), considering cross-variable correlation can significantly improve the model's performance. The overall trend shows stable optimization under different noise levels, which illustrates the importance of cross-variable correlation of errors. Ignoring this correlation (using a diagonal covariance matrix) may lead to the loss of statistical dependence information between variables, resulting in an inaccurate representation of the error's distribution. Explicitly considering the cross-variable correlation of errors (using the non-diagonal covariance matrix) can more realistically capture the statistical characteristics of errors, thereby further improving the model's predictive performance.

In summary, our experiments partially demonstrate that mapping Z to the dimension of the decoder input may lead to a decrease in model performance, and explicitly modeling the cross-variable correlation of errors can improve model performance. Furthermore, as the noise level σ^2 increases, both MSE and MAE show an upward trend, which is consistent with experimental expectations—the greater the noise, the more difficult the prediction becomes, and the greater the error. Besides, the method proposed in this paper is effective at the noise level given in Eq.(12), proving the stability of the method.

5. Conclusions

This paper focuses on error modeling in multivariate time series forecasting, specifically exploring the impact of cross-variable correlations of errors and trivial solutions on error modeling. Experimental results show that mapping errors to a low-dimensional space can effectively avoid the trivial solution problem and further enhance the model's stability. Explicitly considering the cross-variable correlation among the errors can optimize the model's representation of the error distribution and further improve prediction performance. Furthermore, the performance of the stochastic model outperforms the deterministic model, and the proposed method demonstrates stability under different noise levels, validating the effectiveness and practicality, and providing a reference for subsequent research on multivariate time series error modeling.

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China under Grants 62376161 and U24A20322.

References

[1] Gneiting T, Katzfuss M. Probabilistic forecasting[J]. Annual Review of Statistics and Its Application, 2014, 1(1): 125-151.

[2] Zheng V Z, Sun L. Multivariate probabilistic time series forecasting with correlated errors[J]. Advances in Neural Information Processing Systems, 2024, 37: 54288-54329.

[3] Hyndman R J, Athanasopoulos G. Forecasting: principles and practice[M]. OTexts, 2018.

[4] Prado R, West M. Time series: modeling, computation, and inference[M]. Chapman and Hall/CRC, 2010.

[5] Box G E P, Jenkins G M. Some recent advances in forecasting and control[J]. Journal of the Royal Statistical Society. Series C (Applied Statistics), 1968, 17(2): 91-109.

[6] Salinas D, Flunkert V, Gasthaus J, et al. DeepAR: Probabilistic forecasting with autoregressive recurrent networks[J]. International journal of forecasting, 2020, 36(3): 1181-1191.

[7] Sun F K, Lang C, Boning D. Adjusting for autocorrelated errors in neural networks for time series[J]. Advances in Neural Information Processing Systems, 2021, 34: 29806-29819.

[8] Zheng V Z, Choi S, Sun L. Probabilistic traffic forecasting with dynamic regression[J]. Transportation Science, 2025, 59(4): 689-707.

[9] Choi S, Saunier N, Zheng V Z, et al. Scalable dynamic mixture model with full covariance for probabilistic traffic forecasting[J]. Transportation Science, 2025, 59(4): 708-720.

[10] Saha A, Basu S, Datta A. Random forests for spatially dependent data[J]. Journal of the American Statistical Association, 2023, 118(541): 665-683.

[11] Zheng Z, Choi S, Sun L. Better batch for deep probabilistic time series forecasting[C]//International Conference on Artificial Intelligence and Statistics. PMLR, 2024: 91-99.

[12] Henaff M, Zhao J, LeCun Y. Prediction under uncertainty with error-encoding networks[J]. arXiv preprint arXiv:1711.04994, 2017.

[13] Hamilton J D. Time series analysis[M]. Princeton university press, 2020.

[14] Granger C W J, Weiss A A. Time series analysis of error-correction models[M]//Studies in econometrics, time series, and multivariate statistics. Academic Press, 1983: 255-278.

[15] Kim D, Cho Y, Kim D, et al. Residual correction in real-time traffic forecasting[C]//Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022: 962-971.

[16] Jia J, Benson A R. Residual correlation in graph neural network regression[C]//Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. 2020: 588-598.

[17] Zheng V Z, Choi S, Sun L. Enhancing deep traffic forecasting models with dynamic regression[J]. 2023.

[18] Ma L, Wang D D, Huang Y, et al. Decoding the EGFR mutation-induced drug resistance in lung cancer treatment by local surface geometric properties[J]. Computers in biology and medicine, 2015, 63: 293-300.

[19] Rahman N, Wang D D, Ng S H X, et al. Processing of electronic medical records for health services research in an academic medical center: methods and validation[J]. JMIR Medical Informatics, 2018, 6(4): e10933.

[20] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.