Transformer Multivariate Forecasting: Less is More?

Jingjing Xu[1], Caesar Wu[1, 3], Yuan-Fang Li[2], Pascal Bouvry[1, 3]

[1] University of Luxembourg/FSTM
[2] Monash University
[3] Interdisciplinary Centre for Security, Reliability and Trust (SnT)
jingjing.xu@uni.lu, caesar.wu@uni.lu, yuanfang.li@monash.edu, pascal.bouvry@uni.lu

Abstract

In the domain of multivariate forecasting, transformer models stand out as powerful apparatus, displaying exceptional capabilities in handling messy datasets from real-world contexts. However, the inherent complexity of these datasets, characterized by numerous variables and lengthy temporal sequences, poses challenges, including increased noise and extended model runtime. This paper focuses on reducing redundant information to elevate forecasting accuracy while optimizing runtime efficiency. We propose a novel transformer forecasting framework enhanced by Principal Component Analysis (PCA) to tackle this challenge. The framework is evaluated by five state-of-the-art (SOTA) models and four diverse real-world datasets. Our experimental results demonstrate the framework's ability to minimize prediction errors across all models and datasets while significantly reducing runtime. From the model perspective, one of the PCA-enhanced models: PCA+Crossformer, reduces mean square errors (MSE) by 33.3% and decreases runtime by 49.2% on average. From the dataset perspective, the framework delivers 14.3% MSE and 76.6% runtime reduction on Electricity datasets, as well as 4.8% MSE and 86.9% runtime reduction on Traffic datasets. This study aims to advance various SOTA models and enhance transformer-based time series forecasting for intricate data. Code is available at: https://github.com/jingjing-unilu/PCA_Transformer

Introduction

Sequence modeling proves highly effective in capturing patterns within sequential data types like languages, time series, and biological data. Various forms of recurrent neural networks (RNNs) (Schuster and Paliwal 1997) play a vital role in this modeling. Unlike traditional fully connected neural networks (FCNs), which struggle to share features across different data points or locations, standard RNNs overcome this limitation by incorporating input from the previous time step, facilitating the connection of features across various locations within the sequence.

However, standard RNNs face barriers in capturing long-term dependencies due to issues like vanishing and exploding gradients. Proposed solutions, including the integration of Rectified Linear Unit (ReLU) activation functions, weight initialization with identity matrices, and the use of gates
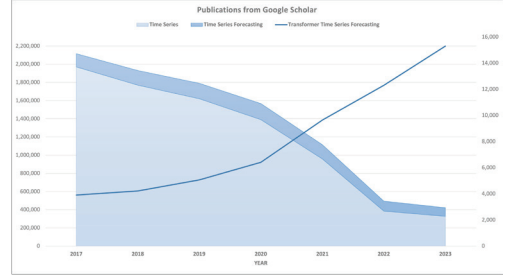
Figure 1: Trend of transformer in time series forecasting topic from 2017.

within RNNs, aim to mitigate these challenges. Other approaches, such as Gated Recurrent Units (GRU) (Cho et al. 2014) and Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997), intend to leverage gate cells to retain long-term dependencies. Nonetheless, challenges persist, particularly in unidirectional processing during training. In response to this limitation, bidirectional RNN structures have been introduced, enabling the processing of both forward and reverse-directional information within a sequence. Despite advancements, these models still face challenges attributed to a lack of parallelism. The Transformer architecture (Vaswani et al. 2017) offers a solution by concurrently processing sequence data through an encoder-decoder architecture and employing attention mechanisms on each word. This approach enhances both efficiency and performance. Transformers prove particularly adept at handling long-term sequence data, such as lengthy sentences in Natural Language Processing (NLP) (Gillioz et al. 2020), as well as image and video data in the Computer Vision (CV) (Han et al. 2022) domain. This characteristic of the transformer architecture in long-time series forecasting can also be applied in other domains.

The landscape of time series forecasting models spans from classic auto-regressive-moving-average (ARMA) models (Whittle 1951) to the era of deep learning. The Transformer, initially crafted for sequential data, emerges as an appealing solution for studying its application in time series forecasting tasks. Figure 1 shows the trend of Google Scholar's publications on Transformer techniques

since 2017. This paper specifically aims for long-term multivariate forecasting, seeking to predict the future based on multiple variables using transformer models.

While multivariate datasets provide richer information and insightful patterns for constructing prediction models, they are inherently complex and pose challenges such as sensitivity to errors and increased computational demands. Excessive runtimes during training and testing phases contribute to significant energy consumption and a large carbon footprint. Hence, the necessity arises for performing multivariate analysis, dimensionality reduction, and feature extraction to facilitate model implementation. Many researchers have made significant contributions by offering different approaches on dimensionality reduction, including Principal Components Analysis (PCA) (Pearson 1901; Hotelling 1933), Linear Discriminant Analysis (LDA), and t-Distributed Stochastic Neighbor Embedding (t-SNE) (Anowar, Sadaoui, and Selim 2021). However, the existing approaches to dimensionality reduction in datasets for transformer models remain insufficient. This paper addresses this gap through a series of intensive experiments. This study introduces a novel framework for multivariate analysis, emphasizing dimension reduction in the context of transformer-based multivariate time series forecasting from a dataset perspective.

Contributions

The primary contributions of this paper are as follows:

1. We present a benchmark test on transformer-based multivariate long-term forecasting with PCA dimension reduction. The comprehensive benchmark test is a particular scheme to evaluate the performance of multivariate long-term forecasting.

2. The work demonstrates that the proposed framework can significantly enhance long-term forecasting performance across five SOTA representative models and four real-world datasets.

3. The comprehensive study also provides detailed and interpretable insights for model transparency, which adds the value of the interpretability and explainability of multivariate analysis underpinned by PCA techniques.

The rest of this paper is organized as follows: Section 2 is the related work. Section 3 is the experiments part, which defines datasets, experimental formulation, experimental environment and configuration, as well as the framework construction. Section 4 shows the experimental results with analysis. Section 5 presents the conclusion and future works.

Related Work
Transformer in Time Series Forecasting

Transformer models display excellent performance due to its ability to capture long-range dependencies in sequential tokens. Consequently, the transformer model emerges as a good option for modelling time series problems. Several transformer-based forecasting models have been developed to address specific challenges in time series forecasting. Transformer-based forecasting models can be organized into different categories based on network modification criteria (Wen et al. 2022; Liu et al. 2023). This research adopts the classification method articulated by (Liu et al. 2023). Table 1 highlights four types of transformer models based on modified components and architecture.

Modi. Archi. \ Modi. Compo.	Yes	No
Yes	Crossformer	iTransformer
No	Autoformer	PatchTST Non-stationary Transformer

Table 1: Four categories of transformer-based forecasting models. Modi. Archi.: Modified Architecture; Modi. Compo.: Modified Components.

Informer (Zhou et al. 2021) and Autoformer (Wu et al. 2021) focuses on adapting components for temporal long sequences. The Crossformer model (Zhang and Yan 2022) presents its capability to preserve time and dimension information while effectively capturing cross-time and cross-dimension dependencies, enhancing its performance in multivariate time series forecasting. The Non-stationary Transformer (Liu et al. 2022)), on the other hand, unifies input, converts output, and solves the over-stationary problem to forecast results. PatchTST (Nie et al. 2022) utilizes patching and channel-independent architecture. It facilitates the model to capture local semantic information and longer lookback windows. iTransformer (Liu et al. 2023) inverts the structure of the transformer model without modify components to enhance the performance of forecasting. It provides an alternative architecture for time series forecasting. Additionally, many recent transformer-based models such as Scaleformer (Shabani et al. 2022), TDformer (Zhang et al. 2022), GCformer (Zhao et al. 2023), PDTrans (Tong, Xie, and Zhang 2023) etc, offering diverse approaches to time series forecasting. In addition, (WU et al. 2024a,b) introduces credit default swap (CDS) dataset for testing transformer models. A comprehensive survey by (Wen et al. 2022) explores transformer models for time series forecasting, which provides some valuable insights. Besides transformer models, researchers also adopt other techniques to address time series forecasting, such as graph representation learning (Jin et al. 2022) and Large Language Models (LLMs) (Jin et al. 2023).

PCA

From a data dimensionality reduction perspective, PCA has a distinct capacity to reduce dimensions in an easily interpretable way while preserving the essential information contained in the data. The roots of the PCA technique can be traced back to (Pearson 1901) and (Hotelling 1933). PCA has gained widespread acceptance, particularly in the era of high-dimensional big data, such as image, text, various stock market data (XIE 2019) and hospital patient data (Kutcher, Ferguson, and Cohen 2013) etc, PCA has widespread adoption. The methodology has evolved into several versions, such as Sparse PCA (Hotelling 1933; Merola 2015), Nonlinear PCA (Hastie and Stuetzle 1989),
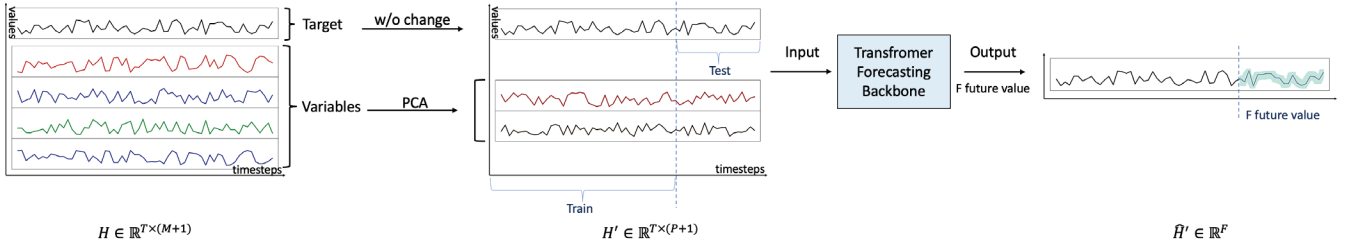
Figure 2: Structure overview of the PCA-enhanced transformer forecasting framework.
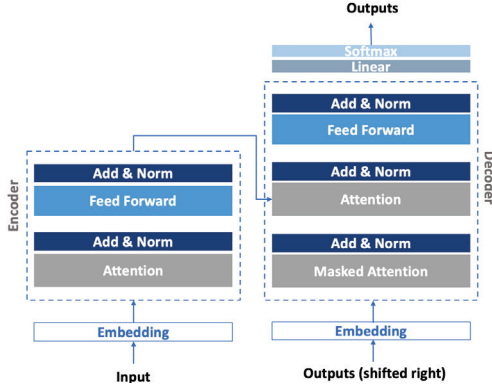

Figure 3: Simplified vanilla transformer architecture with three main components (Attention, Add&Norm, FeedForward).

Kernel PCA (Schölkopf, Smola, and Müller 1998) and Robust PCA (Kriegel et al. 2008). It has become a fundamental tool for unraveling and interpreting complex datasets across varied domains.

PCA+Transformer in Time Series Forecasting

Several studies have explored the integration of PCA with transformer in time series forecasting, and they focus on different perspectives. For instance, the work by (An et al. 2022) centers on combining PCA with Informer model for fault detection and prediction in nuclear power valves. However, this work suffers from shortage of experimental implementation across various transformer models and datasets. Alongside this development, PCA has gained popularity for its role in model interpretation. In recent studies (Madane et al. 2022) and (Li et al. 2022), PCA has been applied to assess model performance, specifically in examining similarities between synthetic and real data. Researchers (Zhou et al. 2023) present a distinctive viewpoint by connecting self-attention with PCA to illustrate the functioning of transformer models. Data analysts (Jin, Hou, and Chen 2022) adopt t-distributed stochastic neighbour Embedding (t-SNE) method (an unsupervised nonlinear dimensionality reduction technique, similar to the PCA) to explore the operational patterns of the model. Other investigators (Pandey et al. 2023) have introduced kernel PCA models for mul-

tivariate time series forecasting but have not applied them to various transformer-based forecasting models. The limitations of dimensionality reduction techniques remain when applied to datasets intended for transformer-based models. To address this gap, we propose the PCA-enhanced transformer-based time series forecasting model.

Experiments

Dataset

To evaluate all transformer models, we adopt four long-term time series datasets widely accepted as standard data sources. They are ETTh1, Weather, Traffic, and Electricity. The targeted or predicted variables for these datasets are oil temperature, wet bulb, 862th sensor, and 321th client, respectively. Furthermore, these four datasets contain different ranges of variables, varying from 7 to 862 with large time steps. Having dataset with large variables and long time steps can ensure stable results with less over-fitting. Table 2 highlights a summary of information about four datasets. A detailed description of variables for these datasets can be found in (Zhou et al. 2021; Wu et al. 2021).

Datasets	ETTh1	Weather	Electricity	Traffic
Variables	7	21	321	862
Timesteps	17420	52696	26304	17544

Table 2: Summary of information about four datasets (targets are counted into variables).

Experimental Formulation

We aim to solve the following problem: given past observations of time series $H = (x^1, x^2, ..., x^M, y) \in \mathbb{R}^{T \times (M+1)}$ where the M represents M variables, the y denotes the *target* and the T is the length of time series. The time series of the m-th variable is denoted as $x^m = (x^m_1, x^m_2, ..., x^m_T)^\top \in \mathbb{R}^T$. The time series of the *target* can be denoted as $y = (y_1, y_2, ..., y_T)^\top \in \mathbb{R}^T$. We aim to forecast F future values of the *target* $(y_{T+1}, y_{T+2}, ..., y_{T+F})^\top$.

Experimental Environment and Configuration

All experiments are implemented on a platform with a single Nvidia TU02 GPU. We test all transformer-based models (refer to Table 1), currently considered to be SOTA for

Dataset	Variables w/o Target	PCA Components	PCA+PatchTST (MSE | MAE)		PCA+Crossformer (MSE | MAE)		PCA+Autoformer (MSE | MAE)		PCA+N.S. Trans. (MSE | MAE)		PCA+iTransformer (MSE | MAE)		PCA+Transformer (MSE | MAE)	
ETTh1	6	2	**0.05561**	0.17868	0.19371	0.38307	**0.08919**	**0.22736**	**0.07144**	**0.20484**	0.05736	0.18378	0.85307	0.88175
		4	0.05671	0.18052	**0.16091**	**0.33508**	0.10550	0.25530	0.08048	0.21217	**0.05637**	**0.18241**	1.17205	1.01673
		w/o PCA	0.05575	**0.17838**	0.36994	0.55491	0.11395	0.25875	0.07814	0.21079	0.05662	0.18278	**0.72492**	**0.79393**
Weather	20	2	0.0013408	0.0267667	0.0066745	0.0669198	0.0390251	0.0985756	**0.0012783**	**0.0262055**	**0.0012391**	**0.0258336**	0.0074935	0.0697642
		5	0.0013426	0.0266463	0.0039031	0.0495619	0.0086512	0.0701804	0.0013453	0.0270548	0.0012655	0.0262108	0.0031227	0.0433567
		10	0.0013277	0.0265256	0.0044438	0.0535568	0.0114851	0.0851401	0.0015487	0.0288031	0.0012676	0.0260793	0.0051427	0.0554222
		15	0.0013124	0.0266787	**0.0019383**	**0.0324117**	**0.0072318**	**0.0679583**	0.0018783	0.0313342	0.0013160	0.0265606	0.0040705	0.0517972
		w/o PCA	**0.0013116**	**0.0263536**	0.0045753	0.0548066	0.0079332	0.0704268	0.0016366	0.0294900	0.0013511	0.0271509	**0.0025455**	**0.0372824**
Electricity	320	2	**0.30114**	**0.39454**	0.24117	0.35286	0.41615	0.49454	0.32548	0.42185	0.35119	0.43913	**0.31722**	**0.42333**
		20	0.30803	0.40253	0.31382	0.39153	0.39585	0.47261	0.35360	0.44082	0.31114	0.41341	0.44774	0.50390
		40	0.31029	0.40227	0.32395	0.39478	0.38023	0.46607	0.38404	0.44949	0.34803	0.43267	0.45457	0.49402
		80	0.30792	0.40005	**0.23528**	**0.34291**	0.42425	0.48486	0.31263	0.41109	0.28967	0.39682	0.36052	0.44804
		160	0.30951	0.39863	0.30753	0.38397	0.38894	0.47346	0.32938	0.42110	0.31660	0.41921	0.48044	0.51270
		240	0.31747	0.40354	0.26574	0.36454	0.42832	0.49689	0.34589	0.43446	**0.25611**	**0.36751**	0.37116	0.45557
		w/o PCA	0.30771	0.40156	0.27444	0.36760	**0.37565**	**0.46435**	**0.30231**	**0.40977**	0.54722	0.54924	0.39571	0.46279
Traffic	861	1	0.17485	0.25452	**0.14789**	**0.22577**	0.31848	0.41374	0.18998	0.29156	0.32595	0.40816	0.28672	0.35876
		2	0.17714	0.25728	0.16042	0.24255	0.29514	0.39557	**0.17894**	**0.28501**	1.86223	1.17306	**0.28144**	**0.35329**
		25	**0.16754**	0.24795	0.25004	0.31726	0.27335	0.36999	0.22374	0.33921	1.31162	0.93039	0.29579	0.37345
		50	0.16907	0.24932	0.20874	0.28413	**0.24967**	**0.34481**	0.20833	0.31271	1.86220	1.17306	0.41403	0.43881
		105	0.17121	0.25329	0.23546	0.29953	0.30357	0.40072	0.19733	0.30020	1.86222	1.17307	0.33498	0.40060
		215	0.17100	0.24890	0.16648	0.23886	0.37778	0.44868	0.23479	0.33047	1.86224	1.17306	0.35731	0.41112
		430	0.17507	0.25062	0.17663	0.25027	0.29786	0.40188	0.27937	0.38314	1.86222	1.17307	0.38011	0.41905
		645	0.17123	0.24950	0.21702	0.28981	0.31110	0.39311	0.37402	0.45356	1.25570	0.92052	0.33137	0.39332
		w/o PCA	0.16831	**0.24585**	0.15540	0.22792	0.30356	0.40767	0.21742	0.32678	**0.25060**	**0.35071**	0.29690	0.35520

Table 3: Accuracy results of transformer forecasting models (w/o PCA) and PCA-enhanced transformer forecasting models. Note: 1) w/o PCA: without PCA; 2) PCA+N.S. Trans.: PCA + Non-stationary Transformer.

Dataset	Variables w/o Target	PCA Components	PCA+ PatchTST	PCA+ Crossformer	PCA+ Autoformer	PCA+N.S. Transformer	PCA+ iTransformer	PCA+ Transformer
ETTh1	6	2	71	241	310	118	87	219
		4	64	**246**	311	141	**84**	126
		w/o PCA	74	252	310	177	88	**126**
Weather	20	2	231	1915	686	**415**	**653**	375
		5	241	1264	690	414	538	377
		10	250	1399	701	416	603	379
		15	283	**1224**	**689**	418	440	378
		w/o PCA	**311**	1768	685	417	625	**378**
Electricity	320	2	**221**	1430	975	876	200	**283**
		20	246	883	723	1198	209	241
		40	352	904	604	1043	207	241
		80	579	**925**	733	887	205	245
		160	1038	1620	748	1394	216	250
		240	1510	2847	622	1418	**380**	303
		w/o PCA	1953	3960	**759**	**1768**	287	303
Traffic	861	1	667	**2590**	320	146	346	135
		2	735	2672	322	**149**	378	**136**
		25	**716**	2395	321	152	162	138
		50	726	2670	**330**	153	329	139
		105	743	2760	330	155	358	141
		215	1285	6122	332	159	409	145
		430	2398	10330	348	167	683	154
		645	3554	10171	353	177	755	162
		w/o PCA	4626	19749	354	173	**625**	154

Table 4: Runtime results of transformer forecasting models (w/o PCA) and PCA-enhanced transformers forecasting model (unit: second).

time series forecasting. For each experiment, all models are configured with pre-selected default hyperparameters specific to each dataset for their best performance. The prediction length is set to 96 (based on the randomly selection) for all experiments as the long-term forecasting setting. For the models' evaluation metrics, we use mean squared error (MSE) and mean absolute error (MAE). The code[1] in this paper is constructed based on the Time Series Library by (Wu et al. 2023).

Framework Construction

Figure 2 represents the basic idea of the novel framework with various transformer-based forecasting models[2] serving as the backbone. The initial step of the experiments starts with loading the original dataset. The next step is to employ PCA and reduce a complete set of variables from M to P. It is the process of dimensionality reduction. The *target*

---

[1] https://github.com/jingjing-unilu/PCA_Transformer
[2] They are all based on the vanilla transformer model.

variable is deliberately excluded from these steps to prevent information or data leakage. The following step is to partition the dataset resulting from PCA into training, testing and validation (depends on the model). The post-processed sub-dataset is then fed into transformer-based forecasting models for training and validation. The final step is using forecasting model to predict F future values for the *target* variable.

- PCA Process. In the PCA process, we utilize the randomized singular value decomposition (SVD) based PCA (Pedregosa et al. 2011; Halko, Martinsson, and Tropp 2011; Szlam, Kluger, and Tygert 2014) due to our large dataset. The input data of the PCA process is $H_{T \times M} = (x^1, x^2, ..., x^M) \in \mathbb{R}^{T \times M}$ with $x^m \in \mathbb{R}^T$. The output data is $H'_{T \times P} = (c^1, c^2, ..., c^P) \in \mathbb{R}^{T \times P}$. P is principal components numbers of the PCA. The calculation steps are in the Algorithm 1.

Algorithm 1: SVD-based PCA

1: **procedure** PCA($H_{T \times M} = (x^1, x^2, ..., x^M) \in \mathbb{R}^{T \times M}$ with $x^m \in \mathbb{R}^T$)
2:     Center the dataset: $\tilde{H}_{T \times M} = (\tilde{x^1}, \tilde{x^2}, ..., \tilde{x^M})$, where $\tilde{x^m} = (x^m - \frac{1}{T}\sum x^m)$
3:     Calculate the covariance matrix: $C\tilde{H}_{T \times M} = cov(\tilde{H}_{T \times M})$
4:     Apply SVD: $S\tilde{H}_{T \times M} = svd(C\tilde{H}_{T \times M})$
5:     Get top P principal components: $H'_{T \times P} = S\tilde{H}_{T \times P} = (c^1, c^2, ..., c^P) \in \mathbb{R}^{T \times P}$.
6:     **return** $H'_{T \times P}$
7: **end procedure**

- Transformer-based Time Series Forecasting Models. Figure 3 depicts a vanilla transformer process with characteristic transformer components and architecture. Following the PCA process, we focus on a transformer-based time series forecasting process. The transformer process consists of three main components (attention, add & norm, and feedforward) with two primary functions (encoder and de-

Dataset	Variables w/o Target	PCA+PatchTST (MSE | Time)		PCA+Crossformer (MSE | Time)		PCA+Autoformer (MSE | Time)		PCA+N.S. Trans. (MSE | Time)		PCA+iTransformer (MSE | Time)		PCA+Transformer (MSE | Time)	
ETTh1	6	0.25%	4.05%	56.50%	2.38%	21.73%	0.00%	8.58%	33.33%	0.43%	4.55%	0.00%	0.00%
Weather	20	0.00%	0.00%	57.64%	30.77%	8.84%	-0.58%	21.89%	0.48%	8.29%	-4.48%	0.00%	0.00%
Electricity	320	2.14%	88.68%	14.27%	76.64%	0.00%	0.00%	0.00%	0.00%	53.20%	-32.40%	19.83%	6.60%
Traffic	861	0.46%	84.52%	4.83%	86.9%	17.75%	6.78%	17.70%	13.87%	0.00%	0.00%	5.21%	11.69%
Average Reduction		0.71%	44.32%	33.31%	49.17%	12.08%	1.55%	12.04%	11.92%	15.48%	-8.08%	6.26%	4.57%

Table 5: MSE and runtime reductions of PCA-enhanced transformer forecasting models. This table is the summary of the accuracy result Table 3 and corresponding runtime result Table 4. The baselines are corresponding original transformer models.

coder). According to the previous study (Liu et al. 2023), we can classify four transformer models (Crossformer, iTransformer, Autoformer, and PathTST) into modified components and architecture categories. Notice that non-stationary transformer belong to the same category of PatchTST, as well as our PCA-enhanced transformer framework.

Experimental Results with Analysis

Model Perspective

Table 5 summarises experiment results with PCA-enhanced and non-PCA-enhanced methods across all four datasets. Each entry in Table 5 represents the best-performing model from Table 3 and its runtime Table 4 (lowest MSE with its runtime, bold with blue cell color). Table 5 shows that the proposed PCA-enhanced method exhibits significant reduction MSE and running time for all models except iTransformer. The results highlight that we can improve accuracy and efficiency by reducing input data dimensions. The entire process emphasizes the principle of "less is more". Based on Table 5, we can summarize the experiment results as follows:

1. Compared with non-PCA models, the **PCA+PathTST** model can reduce the average by 0.71% on MSE and 44.32% on runtime. In other words, PCA-enhanced model increases the model's accuracy and time efficiency. For datasets of Electricity and Traffic, the **PCA+PatchTST** model can increase time efficiency 88.68% and 84.52%, respectively, while MSE can be reduced marginally by 2.14% and 0.46%. Similarly, the **PCA+Transformer** model improves performance when applying Electricity and Traffic datasets.

2. Regarding the **PCA+Non-stationary Transformer** model, we achieve a reduction of 12.04% on the MSE and 11.92% on the runtime. Likewise, **PCA+Autoformer** demonstrates similar results except for the weather dataset.

3. Across all PCA-enhanced models, the **PCA+Crossformer** model has the best performance regarding the MSE reduction (33.31%) and the runtime decreasing (49.17%) compared to its non-PCA model. Table 5 shows that all PCA-enhanced models reduce the MSE (between 0.71% and 33.31%) and the runtime (between 1.55% and 49.17%) except the PCA+iTransformer model.

4. Notice that the runtime of the **PCA+ iTransformer** model goes in the opposite direction, increased by 8.08%

on average due to the model's uniqueness of inverted design. Nevertheless, MSE is reduced by up to 53.20% in the best case.
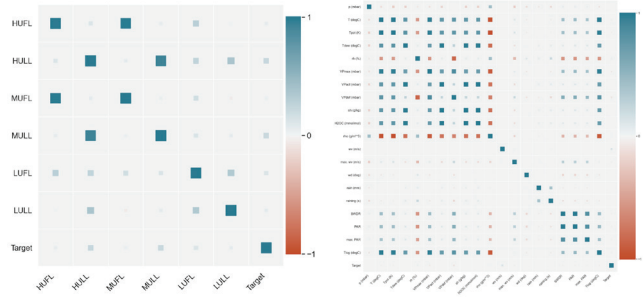

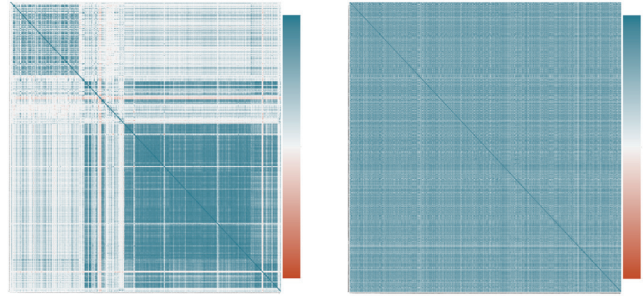Figure 4: Left: PCC of ETTh1 dataset. Right: PCC of Weather dataset.

Figure 5: Left: PCC of Electricity dataset. Right: PCC of Traffic dataset.

Dataset Perspective

The essence of PCA is a linear, unsupervised transformation algorithm. It simplifies dimension reduction by identifying maximum variance in the data and incorporating new features (Ghojogh et al. 2019). This PCA characteristic leads us to employ the Pearson Correlation Coefficient (PCC) for evaluating linear correlations within datasets. Nevertheless, it is crucial to select correlation measurement approaches and customize them for particular dimension-reduction techniques.

The basic logic of selecting the number of PCA components is based on the consideration of variable correlation. If all variables are independent, we should include all of

Dataset	Variables w/o Target	PCA Components	Information Kept (PCA)	Dataset Ratio	Best Result (MSE | Time)		Best Reduction (MSE | Time)		Model
ETTh1	6	2	70.1%	33.3%	0.05561	71	0.2%	4.1%	PCA+Patch.
Weather	20	2	63.1%	10.0%	0.00124	653	8.3%	-4.5%	PCA+iTrans.
Electricity	320	80	94.7%	25.0%	0.23528	925	14.3%	76.6%	PCA+Crossf.
Traffic	861	1	57.6%	0.1%	0.14789	2590	4.8%	86.9%	PCA+Crossf.

Table 6: Information Kept Ratio by PCA with the Best Performance Model (Lowest MSE on each Dataset). The baselines are corresponding original transformer models.

Dataset	Variables w/o Target	PCA Components	Information Kept (PCA)	Dataset Ratio
ETTh1	6	2	70.1%	33.3%
		4	99.6%	66.7%
		w/o PCA	-	100.0%
Weather	20	2	63.1%	10.0%
		5	82.4%	25.0%
		10	98.9%	50.0%
		15	100.0%	75.0%
		w/o PCA	-	100.0%
Electricity	320	2	67.2%	0.6%
		20	87.4%	6.3%
		40	91.2%	12.5%
		80	94.7%	25.0%
		160	97.7%	50.0%
		240	99.2%	75.0%
		w/o PCA	-	100.0%
Traffic	861	1	57.6%	0.1%
		2	71.2%	0.2%
		25	83.5%	2.9%
		50	87.0%	5.8%
		105	91.2%	12.2%
		215	95.3%	25.0%
		430	98.5%	49.9%
		645	99.6%	74.9%
		w/o PCA	-	100.0%

Table 7: Information Kept Ratio by PCA.

them. Otherwise, we only select a few or even one. The correlation patterns in Figures 4 and 5 differ among the four datasets examined in this paper. The ETTh1 dataset with six variables shows two correlated variables, while the Weather dataset exhibits about 50% of correlated variables in the right of the Figure 4. Table 7 shows that if we employ a two-component PCA algorithm, the ETTh1 dataset preserves 70.1% of the information, the Weather dataset retains 63.1%, the Electricity dataset maintains 67.2%, and the Traffic dataset keeps 71.2%. Nevertheless, we still selected the one-component-PCA method for the Traffic dataset experiment because this dataset has massive correlated variables (See the right of the Figure 5). In order to achieve the best MSE performance, we select a different number of PCA components for different datasets. Table 6 summarises information on PCA components selection for the experiments. PCA-enhanced models consistently outperform their non-PCA counterparts across all datasets, highlighting the effectiveness of PCA in enhancing transformer-based forecasting models through dimensionality reduction on the input data.

Table 6 also demonstrates that the PCA-enhanced model is particularly good for a dataset with a medium and large number of correlated variables. The more correlated variables are, the better PCA enhancement is. For example, the Electricity dataset illustrates this point, which can retain 94.7% of the information after PCA enhancement, achieving MSE reduction by 14.3% and runtime reduction by 76.6%. Also, the Traffic dataset's variables are highly correlated. The significant number of correlated variables means that all evaluation metrics benefit from PCA process: information preservation rate with one-component PCA is 57.6%, while MSE is reduced by 4.8% and runtime is decreased by 86.9%.

Conclusion and Future Works

We present a novel forecasting framework that leverages Principal Component Analysis (PCA) to enhance the performance of transformer time series forecasting models. This study conducted experiments across five SOTA transformer models with four diverse real-world datasets. The subsequent analyses provide insights from both model and data perspectives. The results underscore the significant improvements achieved by the PCA-enhanced transformer forecasting framework across all contexts. This framework illuminates the substantial potential of dimension reduction algorithms in terms of both the effectiveness and efficiency of transformer-based forecasting models. We argue less is more. Nonetheless, this study primarily focuses on compressing datasets from a variable perspective.

PCA-enhanced solution is a preliminary examination for dimension reduction in transformer models. Future research should extend its boundary to the temporal compression of datasets. Furthermore, it is may useful to consider other dimension reduction methods, such as linear discriminant analysis (LDA), Independent Component Analysis (ICA), GrandPrix, Zero-Inflation Modulation Analysis(ZIFA), t-distributed stochastic neighbour edging (t-SNE) etc. We hope that this study will serve as a stepping stone to stimulate future work on transformer-based time series forecasting.

Acknowledgments

References

An, Z.; Cheng, L.; Guo, Y.; Ren, M.; Feng, W.; Sun, B.; Ling, J.; Chen, H.; Chen, W.; Luo, Y.; et al. 2022. A novel principal component analysis-informer model for fault prediction of nuclear valves. *Machines*, 10(4): 240.

Anowar, F.; Sadaoui, S.; and Selim, B. 2021. Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne). *Computer Science Review*, 40: 100378.

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Ghojogh, B.; Samad, M. N.; Mashhadi, S. A.; Kapoor, T.; Ali, W.; Karray, F.; and Crowley, M. 2019. Feature selection and feature extraction in pattern analysis: A literature review. *arXiv preprint arXiv:1905.02845*.

Gillioz, A.; Casas, J.; Mugellini, E.; and Abou Khaled, O. 2020. Overview of the Transformer-based Models for NLP Tasks. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, 179–183. IEEE.

Halko, N.; Martinsson, P.-G.; and Tropp, J. A. 2011. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2): 217–288.

Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. 2022. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1): 87–110.

Hastie, T.; and Stuetzle, W. 1989. Principal curves. *Journal of the American Statistical Association*, 84(406): 502–516.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6): 417.

Jin, M.; Wang, S.; Ma, L.; Chu, Z.; Zhang, J. Y.; Shi, X.; Chen, P.-Y.; Liang, Y.; Li, Y.-F.; Pan, S.; et al. 2023. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*.

Jin, M.; Zheng, Y.; Li, Y.-F.; Chen, S.; Yang, B.; and Pan, S. 2022. Multivariate time series forecasting with dynamic graph neural odes. *IEEE Transactions on Knowledge and Data Engineering*.

Jin, Y.; Hou, L.; and Chen, Y. 2022. A time series transformer based method for the rotating machinery fault diagnosis. *Neurocomputing*, 494: 379–395.

Kriegel, H.-P.; Kröger, P.; Schubert, E.; and Zimek, A. 2008. A general framework for increasing the robustness of PCA-based correlation clustering algorithms. In *Scientific and Statistical Database Management: 20th International Conference, SSDBM 2008, Hong Kong, China, July 9-11, 2008 Proceedings 20*, 418–435. Springer.

Kutcher, M. E.; Ferguson, A. R.; and Cohen, M. J. 2013. A principal component analysis of coagulation after trauma. *The journal of trauma and acute care surgery*, 74(5): 1223.

Li, X.; Metsis, V.; Wang, H.; and Ngu, A. H. H. 2022. Ttsgan: A transformer-based time-series generative adversarial network. In *International Conference on Artificial Intelligence in Medicine*, 133–143. Springer.

Liu, Y.; Hu, T.; Zhang, H.; Wu, H.; Wang, S.; Ma, L.; and Long, M. 2023. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. *arXiv preprint arXiv:2310.06625*.

Liu, Y.; Wu, H.; Wang, J.; and Long, M. 2022. Non-stationary transformers: Exploring the stationarity in time series forecasting. *Advances in Neural Information Processing Systems*, 35: 9881–9893.

Madane, A.; Dilmi, M.-d.; Forest, F.; Azzag, H.; Lebbah, M.; and Lacaille, J. 2022. Transformer-based conditional generative adversarial network for multivariate time series generation. *arXiv preprint arXiv:2210.02089*.

Merola, G. M. 2015. Least squares sparse principal component analysis: A backward elimination approach to attain large loadings. *Australian & New Zealand Journal of Statistics*, 57(3): 391–429.

Nie, Y.; Nguyen, N. H.; Sinthong, P.; and Kalagnanam, J. 2022. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*.

Pandey, A.; De Meulemeester, H.; De Moor, B.; and Suykens, J. A. 2023. Multi-view Kernel PCA for Time series Forecasting. *arXiv preprint arXiv:2301.09811*.

Pearson, K. 1901. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11): 559–572.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Schölkopf, B.; Smola, A.; and Müller, K.-R. 1998. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5): 1299–1319.

Schuster, M.; and Paliwal, K. K. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11): 2673–2681.

Shabani, A.; Abdi, A.; Meng, L.; and Sylvain, T. 2022. Scaleformer: iterative multi-scale refining transformers for time series forecasting. *arXiv preprint arXiv:2206.04038*.

Szlam, A.; Kluger, Y.; and Tygert, M. 2014. An implementation of a randomized algorithm for principal component analysis. *arXiv preprint arXiv:1412.3510*.

Tong, J.; Xie, L.; and Zhang, K. 2023. Probabilistic decomposition transformer for time series forecasting. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, 478–486. SIAM.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wen, Q.; Zhou, T.; Zhang, C.; Chen, W.; Ma, Z.; Yan, J.; and Sun, L. 2022. Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125*.

Whittle, P. 1951. Hypothesis testing in time series analysis. *(No Title)*.

WU, C.; Li, Y.-F.; LI, J.; XU, J.; and BOUVRY, P. 2024a. Trustworthy AI: Deciding What to Decide. In *12th Computing Conference 2024*.

WU, C.; XU, J.; BOUVRY, P.; and Li, J. 2024b. Strategic Predictions and Explanations By Machine Learning. In *ICOIN 2024, The 38th International Conference on Information Networking*.

Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; and Long, M. 2023. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In *International Conference on Learning Representations*.

Wu, H.; Xu, J.; Wang, J.; and Long, M. 2021. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems*, 34: 22419–22430.

XIE, X. 2019. Principal component analysis. *Wiley interdisciplinary reviews*.

Zhang, X.; Jin, X.; Gopalswamy, K.; Gupta, G.; Park, Y.; Shi, X.; Wang, H.; Maddix, D. C.; and Wang, Y. 2022. First de-trend then attend: Rethinking attention for time-series forecasting. *arXiv preprint arXiv:2212.08151*.

Zhang, Y.; and Yan, J. 2022. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The Eleventh International Conference on Learning Representations*.

Zhao, Y.; Ma, Z.; Zhou, T.; Sun, L.; Ye, M.; and Qian, Y. 2023. GCformer: An Efficient Framework for Accurate and Scalable Long-Term Multivariate Time Series Forecasting. *arXiv preprint arXiv:2306.08325*.

Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; and Zhang, W. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, 11106–11115.

Zhou, T.; Niu, P.; Wang, X.; Sun, L.; and Jin, R. 2023. One Fits All: Power General Time Series Analysis by Pretrained LM. *arXiv preprint arXiv:2302.11939*.