Milestone Meeting 1

Kai Reffert

February 26, 2025

Overview

1. Probabilistic Methods
2. IMS
3. DMS
4. LTSF models
5. Metrics
6. Implementation

Probabilistic Methods

- Distributional Forecasting ($\rightarrow$ first application to PatchTST ready)
- Quantile Regression ($\rightarrow$ first application to PatchTST ready)
  - Implicit Quantile Regression
- State Space Models
- Normalizing/Variational/Rectifying Flows
- Generative Methods?
  - Diffusion, VAE, GAN
- Conformal predictions?
- Bayesian NNs?

Implicit Quantile Regression

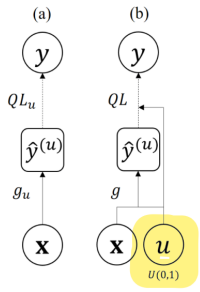


Figure: Solid arrow indicates forward computation with possibly multiple layers, and dashed arrow is the loss function linking prediction and truth. (a) Quantile Regression; (b) Generative Quantile Model.

(a) Quantile regression produces forecasts for pre-defined and fixed number of quantiles.

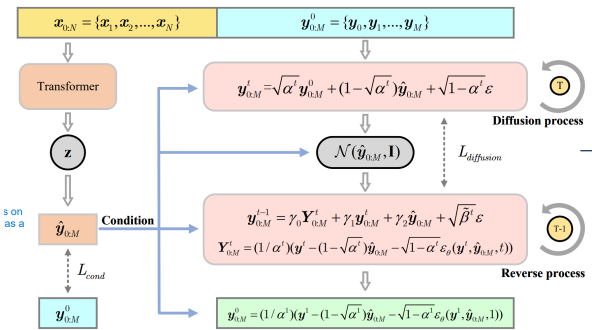
(b) Implicit quantile regression produces forecasts for a given (random) quantile level

- enables (generative) modelling of the whole distribution
- mitigates need to define quantiles a priori
- allows sampling of arbitrary quantile levels

Generative models - Diffusion

Model + Paper	Backbone	NAR/AR	Denoising
LDT [Feng et al., 2024]	Transformer	NAR	latent state
TSDiff [Kollovieh et al., 2023]	SSSD	NAR	target
TMDM [Li et al., 2023]	Transformer	NAR	point forecasts of target
TimeDiff [Shen and Kwok, 2023]	CNN	AR	target
mr-diff [Shen et al., 2023]	seasonal-trend decomposition AE	NAR	mulit-resolution target
Diffusion-TS [Yuan and Qiao, 2023]	Transformer	NAR	decomposed target
D^3VAE [Li et al., 2022]	BVAE	NAR	input and target
TimeGrad [Rasul et al., 2021a]	RNN	AR	target of next time step
CSDI [Tashiro et al., 2021]	-	NAR	single steps (comes from imputation)

TMDM



- use transformer prediction of conditional mean as prior
- then diffusion model estimates uncertainty

→ plug-and-play framework to incorporate arbitrary point forecasters

- additionally define $\mathcal{N}(\hat{y}_{0:M}, I)$ (instead of $\mathcal{N}(0, I)$) as the endpoint of the diffusion process → diffusion process is an interporlation between true data and transformer representation
- their Github probably not too easy to apply

IMS

Problem of IMS: Forecast of $p(\mathbf{x}_{L+1:L+H}|\mathbf{x}_{1:L})$ encounters error accumulation, since predictions are autocorrelated.

- [Sun et al., 2021] Adjusting for Autocorrelated Errors in Neural Networks for Time Series, NeurIPS
- [Pasini et al., 2024] Continuous Autoregressive Models with Noise Augmentation Avoid Error Accumulation, NeurIPS
- DMS on top of IMS
- other regularization or attention techniques

[Sun et al., 2021] Adjusting for Autocorrelated Errors in Neural Networks for Time Series, NeurIPS

- method to account for autocorrelation in NNs, based on econometrics
  $\rightarrow$ accounts for linear, first-order autocorrelation in the errors, i.e. $e_t = \rho e_{t-1} + \epsilon_t$
- reduces to modelling
  $X_t - \rho X_{t-1} = f(X_{t-1}, \ldots, X_{t-W}; \theta) - \rho f(X_{t-2}, \ldots, X_{t-W-1}; \theta) + \epsilon_t$
- Naive method: 1. fix $\hat{\rho}$ and train model parameters θ. 2. use errors e_t to update $\hat{\rho}$ by linearly regressing e_t on e_{t-1}, i.e.,

$$\hat{\rho} = \frac{\sum_{t=2}^{T} e_t e_{t-1}}{\sum_{t=1}^{T-1} e_t^2}$$

- But the naive procedure doesn't work well empirically

[Sun et al., 2021] Adjusting for Autocorrelated Errors in Neural Networks for Time Series, NeurIPS

Modifications

1. optimize $\hat{\rho}$ with θ jointly using stochastic gradient descent (SGD)
2. target $X_t - \hat{\rho}X_{t1}$ is related to the difference of two model outputs, which complicates the optimization. Use following approximation:

$$f(X_{t-1}, \ldots, X_{t-W}; \theta) - \hat{\rho} f(X_{t-2}, \ldots, X_{t-W-1}; \theta) \simeq f(X_{t-1}, \ldots, X_{t-W-1}; \theta, \hat{\rho})$$

   Now the minimization of MSE becomes $X_t - \hat{\rho}X_{t-1} = f(X_{t-1}, \ldots, X_{t-W-1}; \theta, \hat{\rho})$
3. However, in the equation above, input and target series are not in the same form. Input: X_t and Target: $X_t - \hat{\rho}X_{t-1}$, therefore we modify the equation into:

$$X_t - \hat{\rho}X_{t-1} = f(X_{t-1} - \hat{\rho}X_{t-2}, \ldots, X_{t-W} - \hat{\rho}X_{t-W-1}; \theta)$$

[Sun et al., 2021] Adjusting for Autocorrelated Errors in Neural Networks for Time Series, NeurIPS

Personal reflections

1. only needs slight modifications, i.e. input and target ($X_{t-1} \rightarrow X_{t-1} - \hat{\rho} X_{t-2}$), while introducing only a single new parameter $\hat{\rho}$
2. could be extended to **higher order autocorrelation errors** or possibly dynamically discover/learn which orders are most important
3. could be extended to **nonlinear modelling** by replacing $X_t - \rho X_{t-1}$ with $X_t - f(X_{t-1}; \theta')$
4. Problem: not directly applicable to (most) probabilistic NNs
   - Quantile regression $\rightarrow$ easy to apply
   - others: errors need to be expressible as:

$$e_t = \rho_1 e_{t-1} + \ldots + \rho_p e_{t-p} + \epsilon_t, \quad |\rho_i| < 1 \forall i$$

[Pasini et al., 2024] Continuous Autoregressive Models with Noise Augmentation Avoid Error Accumulation, NeurIPS

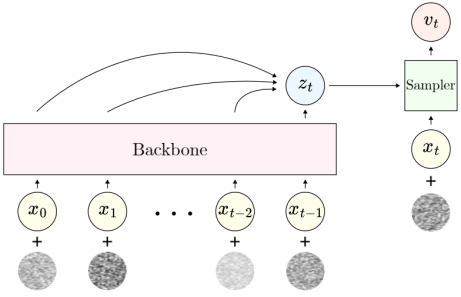


Figure 1: Training process of CAM. The causal Backbone receives as input a sequence of continuous embeddings with noise augmentation. It outputs z_t, which is used by the Sampler as conditioning to denoise a noise-corrupted version of x_t.

- continuous embeddings $x_0, x_1, \ldots, x_{t-1}$ = latent states of autoregressive VAE
- backbone receives corrupted continuous embeddings and produces z_t, then Sampler uses z_t to denoise x_t
- want to make the model more robust against error accumulation by injecting random noise during training to simulate erroneous predictions
- also add a small amount of artificial noise to the generated embeddings, further increasing resilience to accumulated errors

DMS

Problem of DMS: Forecast of $p(\mathbf{x}_{L+1:L+H}|\mathbf{x}_{1:L})$ without inherent temporal dependence between predictions, i.e. they are probabilistic time step predictors, but not probabilistic time series predictors.

- Post-Hoc approaches on top of DMS model
  - Gaussian Process, Copulas, Markov Processes, Kalman Filters
- latent variable models that share latent variable across time steps $\rightarrow$ dependence between predictions
- hybrid approaches, e.g. DMS and IMS hybrid
- also forecast the difference to next time step, to include dependence between predictions

Copula

A copula function $C : [0,1]^n \rightarrow [0,1]$ transforms an n-dimensional function on the interval [0, 1] into a unit-dimensional one.
In general, suppose $G_i(u_i) \in [0,1]$ is a univariate, uniform distribution with $u_i = u_1, \ldots, u_n$ and $i \in N$ (i is an element of set N). We, therefore, define a copula function as follows:

$$C[G_1(u_1), \ldots, G_n(u_n)] = F_n[F_1^{-1}(G_1(u_1)), \ldots, F_n^{-1}(G_n(u_n)); \rho_F]$$

where:

- $G_i(u_i)$ are the marginal distributions that have no well-known properties.
- F_n is the joint cumulative distribution function.
- F_i^{-1} is the inverse of F_i.
- ρ_F is the correlation structure of F_n.

Put in words, the above equation reads: Given the marginal distributions $G_1(u_1)$ to $G_n(u_n)$, there exists a copula function that allows the mapping of the marginal distributions $G_1(u_1)$ to $G_n(u_n)$ via F^{-1} and the joining of the (abscise values) $F_i^{-1}(G_i(u_i))$ to a single, n-variate function $F_n[F_1^{-1}(G_1(u_1)), \ldots, F_n^{-1}(G_n(u_n))]$ that has a correlation structure of ρ_F.

Copula

- copulas often used to model cross-series information [Salinas et al., 2019].
- sometimes additionally used to model across time as well [Ashok et al., 2023, Drouin et al., 2022].

$\rightarrow$ Use (Gaussian) Copula to model the dependence between the marginals of each time step

$$p(y_{1:t}|y_1, \ldots, y_t) = \mathcal{N}([f_1(z_1), f_2(z_2), \ldots, f_N(z_N)]^T|\mu(h), \Sigma(h))$$

Cons

- Possibly many parameters of covariance matrix ($\rightarrow$ need matrix approximation techniques)
- need to decide what copula to use

LTSF

Which models to try out probabilistically?

Model + Paper	Local/global	Univ./Multiv.	CD/CI	NAR/AR	Covariates
MTST [Zhang et al., 2024]	Global	Univ.	CI	NAR	No
PatchMixer [Gong et al., 2024]	Global	Univ.	CI	NAR	No
TimeMachine [Ahamed and Cheng, 2024]	Global	Univ./Multiv.	CI/CD	NAR	No
TiDE [Das et al., 2024]	Global	Univ.	CI	NAR	Yes
ModerTCN [Donghao and Xue, 2023]	Global	Multiv.	CD	NAR	No
PatchTST [Nie et al., 2023]	Global	Univ.	CI	NAR	No
TSMixer [Chen et al., 2023]	Global	Multiv.	CD	NAR	Yes
TimesNet [Wu et al., 2023]	Global	Multiv. (?)	CD (?)	NAR	No
SDCNet [Wang et al., 2023]	Global	Multiv.	CD	NAR	No

Metrics

- standard proper scoring rules for probabilistic forecasts
  - CRPS, CRPS-Sum
- empirical coverage
- PICP and QICE from [Li et al., 2023]
- scores for probabilistic time series forecasts
  - compare distributions of differences between lag time steps of true and predicted values (maybe via KL-divergence?) $\rightarrow$ but what is the difference of marginal distributions and how does it relate to the difference of discrete values
  - measure autocorrelation of the forecasts/errors (maybe only possible for realizations, i.e. quantiles), see methods to evaluate autocorrelation of OLS methods in [Wang and Akabay, 1994]

Implementation

- Repositories & Data
- Aktueller Stand & nächste Pläne

Repositories & Data - ProbTS

Models

Model	Orig. Eval. Horizon	Estimation	Decoding
Linear	-	Point	Auto / Non-auto
GRU	-	Point	AR / NAR
Transformer	-	Point	AR / NAR
Autoformer	Long	Point	NAR
N-HiTS	Long	Point	NAR
NLinear	Long	Point	NAR
DLinear	Long	Point	NAR
TSMixer	Long	Point	NAR
TimesNet	Short/ Long	Point	NAR
PatchTST	Long	Point	NAR
iTransformer	Long	Point	NAR
ElasTST	Long	Point	NAR
GRU NVP	Short	Probab.	AR
GRU MAF	Short	Probab.	AR
Trans MAF	Short	Probab.	AR
TimeGrad	Short	Probab.	AR
CSDI	Short	Probab.	NAR
TSDiff	Short	Probab.	NAR
+ Foundation models			

Data

Dataset	#Var	Time Steps	Description
STSF			
Exchange	8	6,071	exchange rates of 8 countries
Solar	137	7,009	Solar power production records
Electricity	370	5,833	Electricity consumption
Traffic	963	4,001	Road occupancy rates
Wikipedia	2,000	792	Page views of Wikipedia pages
LTSF			
ETTh	7	17,420	ET temperature per hour
ETTm	7	69,680	ET temperature every 15 min
Electricity	321	26,304	Electricity consumption (Kwh)
Weather	21	52,696	Local climatological data
Traffic	862	17,544	Road occupancy rates
Exchange	8	7,588	exchange rates of 8 countries
ILI	7	966	Ratio of influenza-like illness
Caiso	10	74,472	Electricity load California
Nordpool	18	70,128	Energy production Europe
Turkey Power	18	26,304	Electrical demand Turkey
Istanbul Traffic	3	14,244	Istanbul traffic

Repositories & Data - BasicTS [Shao et al., 2025]

Models (all LTSF)

Baseline	Venue	Baseline	Venue
SOFTS	NeurIPS'24	CrossGNN	NeurIPS'23
CATS	NeurIPS'24	DLinear, NLinear	AAAI'23
Sumba	NeurIPS'24	Crossformer	ICLR'23
GLAFF	NeurIPS'24	DSformer	CIKM'23
CycleNet	NeurIPS'24	SegRNN	arXiv
Koopa	NeurIPS'24	MTS-Mixers	arXiv
Fredformer	KDD'24	LightTS	arXiv
UMixer	AAAI'24	TiDE	TMLR'23
TimeMixer	ICLR'24	Triformer	IJCAI'22
Time-LLM	ICLR'24	NSformer	NeurIPS'22
SparseTSF	ICML'24	FiLM	NeurIPS'22
iTransformer	ICLR'24	FEDformer	ICML'22
ETSformer	NeurIPS'24	Pyraformer	ICLR'22
Autoformer	NeurIPS'21	HI	CIKM'21
Informer	AAAI'21	TimesNet	ICLR'23
+ Spatio temporal			

Data

Dataset	#Var	Time Steps	Description
LTSF			
ETTh1	7	14,400	
ETTh2	7	14,400	
ETTm1	7	57,600	
ETTm2	7	57,600	
Electricity	321	26,304	
Exchange	8	7,588	
ILI	7	966	Ratio of influenza-like illness
Traffic	862	17,544	Road occupancy rates
Weather	21	52,696	Local climatological data
BeijingAirQuality	7	36000	
+ Spatial temporal data			

Primarily focus on LTSF and STSF of spatio temproal models.

Repositories & Data - pytorch-ts

Models (built on top of gluonts)

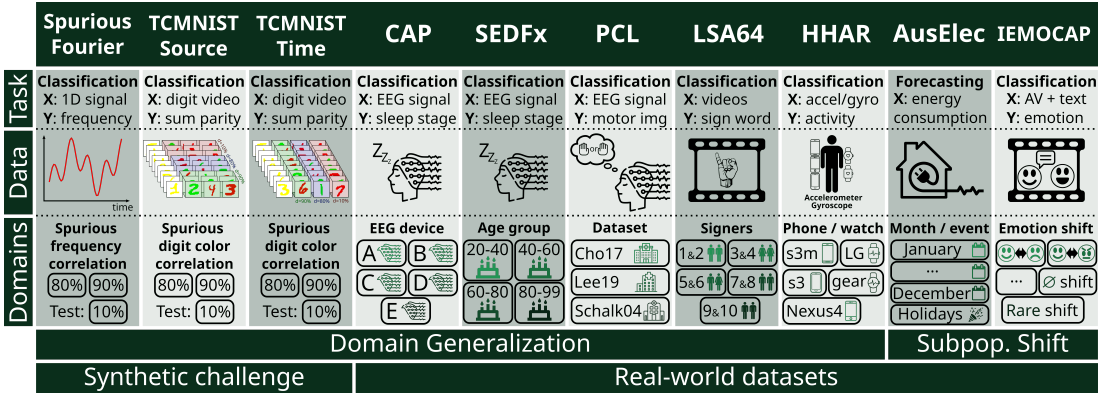
- Causal DeepAR
- DeepAR
- DeepVAR
- LSTNET
- N-Beats
- FNN
- Tempflow
- TFT
- Transformer
- IQN [Gouttes et al., 2021]
- Normalizing Flows[Rasul et al., 2021b]
- TimeGrad [Rasul et al., 2021a]

More

- gluonts
  - mainly method from 2017 to 2021
  - many probabilistic methods
- darts ('sklearn for time series')
  - includes many classical statistical models
  - some probabilistic methods
  - conformal prediction

Repositories & Data - WOODS [Gagnon-Audet et al., 2023]

Benchmarks for Out-of-Distribution Generalization in Time Series Tasks.



Aktueller Stand

- Notebook - Quantile and Distributional regression

Next Steps

- Sanitize and optimize implemented methods, and try them out for more LTSF methods and more datasets
- implicit quantile prediction head
- look for AR models and try to apply error accumulation counter measures
- metrics for probabilistic time **SERIES** forecasts
- Hyperparameter optimization
- start with Probabilistic time series forecasts
- TMDM

References I

Ahamed, M. A. and Cheng, Q. (2024).
TimeMachine: A Time Series is Worth 4 Mambas for Long-term Forecasting.
arXiv:2403.09898.

Ashok, A., Marcotte, , Zantedeschi, V., Chapados, N., and Drouin, A. (2023).
TACTiS-2: Better, Faster, Simpler Attentional Copulas for Multivariate Time Series.

Chen, S.-A., Li, C.-L., Yoder, N., Arik, S. O., and Pfister, T. (2023).
TSMixer: An All-MLP Architecture for Time Series Forecasting.
arXiv:2303.06053.

Das, A., Kong, W., Leach, A., Mathur, S., Sen, R., and Yu, R. (2024).
Long-term Forecasting with TiDE: Time-series Dense Encoder.
arXiv:2304.08424.

Donghao, L. and Xue, W. (2023).
ModernTCN: A Modern Pure Convolution Structure for General Time Series Analysis.

Drouin, A., Marcotte, , and Chapados, N. (2022).
TACTiS: Transformer-Attentional Copulas for Time Series.
arXiv:2202.03528 [cs].

Feng, S., Miao, C., Zhang, Z., and Zhao, P. (2024).
Latent Diffusion Transformer for Probabilistic Time Series Forecasting.
Proceedings of the AAAI Conference on Artificial Intelligence, 38(11):11979–11987.
Number: 11.

References II


Gagnon-Audet, J.-C., Ahuja, K., Bayazi, M. J. D., Mousavi, P., Dumas, G., and Rish, I. (2023).
WOODS: Benchmarks for Out-of-Distribution Generalization in Time Series.
Transactions on Machine Learning Research.


Gong, Z., Tang, Y., and Liang, J. (2024).
PatchMixer: A Patch-Mixing Architecture for Long-Term Time Series Forecasting.
arXiv:2310.00655.


Gouttes, A., Rasul, K., Koren, M., Stephan, J., and Naghibi, T. (2021).
Probabilistic Time Series Forecasting with Implicit Quantile Networks.
arXiv:2107.03743 [cs].


Kollovieh, M., Ansari, A. F., Bohlke-Schneider, M., Zschiegner, J., Wang, H., and Wang, Y. B. (2023).
Predict, Refine, Synthesize: Self-Guiding Diffusion Models for Probabilistic Time Series Forecasting.
Advances in Neural Information Processing Systems, 36:28341–28364.


Li, Y., Chen, W., Hu, X., Chen, B., Sun, B., and Zhou, M. (2023).
Transformer-Modulated Diffusion Models for Probabilistic Multivariate Time Series Forecasting.


Li, Y., Lu, X., Wang, Y., and Dou, D. (2022).
Generative Time Series Forecasting with Diffusion, Denoise, and Disentanglement.
Advances in Neural Information Processing Systems, 35:23009–23022.


Nie, Y., Nguyen, N. H., Sinthong, P., and Kalagnanam, J. (2023).
A Time Series is Worth 64 Words: Long-term Forecasting with Transformers.
arXiv:2211.14730.

References III


Pasini, M., Nistal, J., Lattner, S., and Fazekas, G. (2024).
Continuous Autoregressive Models with Noise Augmentation Avoid Error Accumulation.
arXiv:2411.18447 [cs].


Rasul, K., Seward, C., Schuster, I., and Vollgraf, R. (2021a).
Autoregressive Denoising Diffusion Models for Multivariate Probabilistic Time Series Forecasting.
In *Proceedings of the 38th International Conference on Machine Learning*, pages 8857–8868. PMLR.
ISSN: 2640-3498.


Rasul, K., Sheikh, A.-S., Schuster, I., Bergmann, U., and Vollgraf, R. (2021b).
Multivariate Probabilistic Time Series Forecasting via Conditioned Normalizing Flows.
arXiv:2002.06103 [cs].


Salinas, D., Bohlke-Schneider, M., Callot, L., Medico, R., and Gasthaus, J. (2019).
High-dimensional multivariate forecasting with low-rank Gaussian Copula Processes.
In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.


Shao, Z., Wang, F., Xu, Y., Wei, W., Yu, C., Zhang, Z., Yao, D., Sun, T., Jin, G., Cao, X., Cong, G., Jensen, C. S., and Cheng, X. (2025).
Exploring Progress in Multivariate Time Series Forecasting: Comprehensive Benchmarking and Heterogeneity Analysis.
IEEE Transactions on Knowledge and Data Engineering, 37(1):291–305.
Conference Name: IEEE Transactions on Knowledge and Data Engineering.


Shen, L., Chen, W., and Kwok, J. (2023).
Multi-Resolution Diffusion Models for Time Series Forecasting.

References IV

Shen, L. and Kwok, J. (2023).
Non-autoregressive Conditional Diffusion Models for Time Series Prediction.
In *Proceedings of the 40th International Conference on Machine Learning*, pages 31016–31029. PMLR.
ISSN: 2640-3498.

Sun, F.-K., Lang, C., and Boning, D. (2021).
Adjusting for Autocorrelated Errors in Neural Networks for Time Series.
In *Advances in Neural Information Processing Systems*, volume 34, pages 29806–29819. Curran Associates, Inc.

Tashiro, Y., Song, J., Song, Y., and Ermon, S. (2021).
CSDI: Conditional Score-based Diffusion Models for Probabilistic Time Series Imputation.
In *Advances in Neural Information Processing Systems*, volume 34, pages 24804–24816. Curran Associates, Inc.

Wang, G. C. S. and Akabay, C. K. (1994).
Autocorrelation: Problems and solutions in regression modeli.
The Journal of Business Forecasting Methods & Systems, 13(4):18.
Num Pages: 9 Place: Flushing, United States Publisher: Journal of Business Forecasting.

Wang, X., Liu, H., Du, J., Dong, X., and Yang, Z. (2023).
A long-term multivariate time series forecasting network combining series decomposition and convolutional neural networks.
Applied Soft Computing, 139:110214.

Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., and Long, M. (2023).
TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis.
arXiv:2210.02186 [cs].

References V

Yuan, X. and Qiao, Y. (2023).
Diffusion-TS: Interpretable Diffusion for General Time Series Generation.

Zhang, Y., Ma, L., Pal, S., Zhang, Y., and Coates, M. (2024).
Multi-resolution Time-Series Transformer for Long-term Forecasting.
In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pages 4222–4230. PMLR.
ISSN: 2640-3498.