Milestone Meeting 2

Kai Reffert

May 27, 2025

Overview

1. Research Questions (RQ)/Contributions
2. Evaluation metrics
3. Aktueller Stand
4. Thesis structure
5. Next steps

Contributions

1. Probabilistic Evaluation of LTSF models → methods that result in good marginals
2. Multivariate/Correlation Evaluation → methods that result in good **series** forecasts
3. Probabilistic framework for arbitrary time series model-backbones
4. Literature study

RQ 1 - Probabilistic Evaluation of LTSF models

GOAL: General (univariate) comparison & characterization of (all) implemented methods to turn LTSF models probabilistic. + strengths and weaknesses

- compare their performances across multiple datasets, prediction horizons, backbone LTSF models
- focus more on the simpler methods (distributional forecasting, quantile, implicit quantile, (maybe) Kernel density, ...) and not on more complicated ones, e.g. generative methods (diffusion, AE, VAE, GAN)
- **Evaluation**
  1. Compare with suitable/computable metrics (CRPS, NLL, ES, VS)
  2. use interval metrics to compare quantile forecasts with empirical quantiles of samples of distributional methods
  3. qualitatively

How do different methods for turning Long-Term Series Forecasting (LTSF) models into probabilistic models compare in terms of their ability to produce accurate marginal distributions and forecast uncertainty across various datasets, prediction horizons, and backbone LTSF models?

RQ 2 - Multivariate/Correlation Evaluation

GOAL: General comparison & characterization of methods, which turn marginals into multivariate probabilistic time **series** forecasts. + strengths and weaknesses

- compare their performances across multiple datasets, prediction horizons, backbone LTSF models
- Distributional methods: GP parameters, directly learning multivariate distributions, copulas, ...
- Quantile methods: maybe IMS?, but it doesn't model the correlation structure explicitly
- **Evaluation**
  1. Compare with suitable/computable metrics (NLL, ES, VS) (+ autocorrelation metrics for quantile forecasts?)
  2. qualitatively probably difficult → investigate samples + artificial data set

How do different methods for transforming marginal probabilistic forecasts into multivariate time series forecasts compare in terms of their ability to capture multivariate dependencies, produce accurate joint distributions, and handle forecast uncertainty across various datasets, prediction horizons, and backbone Long-Term Series Forecasting (LTSF) models?

Evaluation metrics - Introduction

Probabilistic forecasts aim for sharpness s.t. calibration, both are summarised by proper scoring rules $S(F, x)$, which assign a numerical score based on prediction F and realization x.

Definition: Assuming P is the true distribution, a scoring rule $S(F, x)$ is proper if:

$$s(P, P) \leq s(F, P) \quad \forall F, P \in \mathcal{F} \tag{1}$$

→ functions that are minimal in expectation for the ground truth distribution

Examples:

- Negative Log Likelihood:

$$\mathrm{NLL}(F, x) = -\log p(x) \tag{2}$$

- is a strictly proper local score, i.e. only depends on x and not distance of distribution to x
- requires density, which is sometimes a stronger limitation (→ approximation of the density may be used)

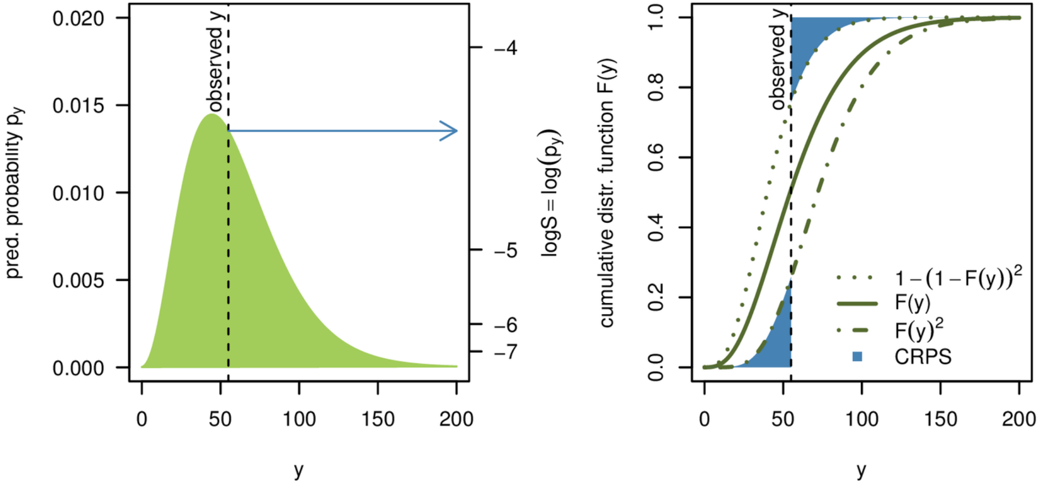
Evaluation metrics - Univariate Scoring Rules

- NLL → may result in $+\infty$ loss if very low probability assigned to $x \Rightarrow$ lacks robustness
- Continuous Ranked Probability Score (CRPS):

$$\mathrm{CRPS}(F, x) = \int_{-\infty}^{\infty} \{F(y) - I(y \geq x)\}^2 dy \tag{3}$$

- strictly proper (if distribution with finite first moment) more robust alternative to NLL
- defined via CDF F → many possible approximations, e.g. via empirical CDF $\hat{F} = \sum_{i=1}^{n} I(y \geq x_i)$
- alternative formulations: quantile score $\mathrm{CRPS}(F^{-1}, x) = \int_0^1 2 \cdot QS_\alpha(F^{-1}(\alpha), x)\ d\alpha$
- closed kernel score representation $\mathrm{CRPS}(F, x) = \mathbb{E}_F|X - x| - \frac{1}{2}\mathbb{E}_F|X - X'|$

Evaluation metrics - Univariate Scoring Rules



Evaluation metrics - Multivariate Scoring Rules

Evaluate multivariate forecast, i.e. with dependency structure.
Extension of CRPS to multivariate score - **Energy Score (ES):**

$$\mathrm{ES}(F, x) = \mathbb{E}_{X\sim F}||X - x||_p^\beta - \frac{1}{2}\mathbb{E}_{X,X'\sim F}||X - X'||_p^\beta \tag{4}$$

However, some studies question the ability of the ES to differentiate between correlation errors.
Use pairwise differences via Variogram Score (VS):

$$\mathrm{VS}_p(F, x) = \sum_i^H \sum_j^H w_{i,j}\Big(|x_i - x_j|^p - \mathbb{E}_{X\sim F}(|X_i - X_j|^p)\Big)^2 \tag{5}$$

Problems

- Quadratic complexity in the number of samples

→ only evaluate once the model is trained?

- *scoringrules* [*git*] implements most scores with exact and approximation options

Evaluation Metrics - Interval and Qunatile metrics

- Needed to compare direct quantile predictions with estimated quantile predictions from distributional methods
- quantile score: $\mathrm{QS}_\alpha(q,x) = \big(\alpha - I(q > x)\big)\big(x - q\big)$
- interval score: $IS_\alpha(F,x) = \frac{QS_{\alpha/2}(q_{\alpha/2},x)+QS_{1-\alpha/2}(q_{1-\alpha/2},x)}{\alpha} = (u-l) + \frac{2}{\alpha}\cdot(l-x)\cdot I(x<l) + \frac{2}{\alpha}\cdot(x-u)\cdot I(x>u)$
- weighted IS: $WIS_{\alpha_{0:K}}(l,u;x) = \frac{1}{K+0.5}\cdot\Big(w_0\cdot|x-m| + \sum_{k=1}^{K}\{w_k\cdot IS_\alpha(l,u;x)\}\Big)$

Problems

- Univariate: Would need autocorrelation based scores to assess correlation property

→ But technically quantile and interval forecasts don't model dependency structure

Evaluation metrics - Regions of reliability - Marcotte (ICML'23)

proper scoring rules = functions that are minimal **in expectation** for the ground truth distribution
→ However, this is not enough in the non-asymptotic case:

- use power analysis to test whether the rule (NLL, CRPS, ES, VS, DS) can discriminate an incorrect forecast distribution from the ground-truth distribution, given samples from both.
- results in the "region of reliability" of a scoring rule, i.e., the set of practical conditions (ground-truth samples n, number of variables d, forecast samples m) where it can identify forecasting errors reliably
- tune ϵ (=difference between distributions), such that NLL has statistical power of 0.8

Evaluation metrics - Marcotte (ICML'23)

Results:

- None of the rules served as a good surrogate to the NLL across all tests
- most rules fail in distinguishing correlation errors!
- VS and CRPS/ES behave complementary → higher reliability when the other score has lower reliability
- score approximations (CRPS, ES) do not affect reliability
- most recent studies have small forecast sample size $m \in [100, 1000]$ and ground truth samples $n \in [7, 57]$ (measured in number of evaluation windows) compared to data dimensionality $d \in [240, 60.000]$ (number of variables) → rules generally unreliable

⇒ Suggestions: increase n and m; look at the regions provided by them; use other diagnostic tools, i.e. inspecting calbration (Q-Q plots) and sharpness (interval width) and plotting correlations

Would it make sense to use simple example time series to specifically comapre the different metrics?
or even compute metrics by hand

Aktueller Stand

Probabilistic Heads

- quantile and implicit quantile heads ✓
- (univariate) distributional heads (Gaussian, student-T, negative-binomial, ...) ✓
- TODO: arbitrary marginals, e.g. via Normalizing Flows or Kernel Density Estimation on residuals
- multivariate Gaussian head ✓ → more multivariate distributions TODO
- GP (✓), not yet in pipeline, is end-to-end possible? → TODO: Copulas

LTSF Backbones

- PatchTST, iTransformer, DeepAR ✓
- For IMS models (like DeepAR):
  - How to construct multivariate distribution from marginals in an end-to-end way?
  - maybe via Copulas or the Covariance Matrix of a GP, but is that trainable end-to-end?
- TODO: maybe more backbones → see available models in Benchmark

BasicTS - Benchmark

- BasicTS+(IEEE Xplore) *Github Link*
- 1.1k Stars, last update 3 months ago
- LTSF and spatio-temporal forecasting focus
- 10 common LTSF datasets
- 32 LTSF models released from ’21 to ’24 (mainly from A* conferences)

Thesis structure

1. Abstract
2. Introduction
3. Literature Review
    - Related work on LTSF
    - Related work on probabilistic forecasting
    - Related work on scoring rules?
4. Body chapter
    - Probabilistic methods/heads: Distributional forecasting, (implicit) quantile forecasting, Flows
    - Multivariate methods: Gaussian Processes, Copulas → Background, fundamental definition, related work
    - Scoring rules?
5. Experiments
6. Conclusion

Next steps

1. Finalize Evaluation pipeline: Interval metrics, QQ-plots, Reliability diagrams, sharpness
2. Multivariate methods (Copulas, more multivariate distributions)
3. Gaussian Process into end-to-end pipeline → does it make sense to do this as another head?
4. goodness of fit test
5. try Kernel Density estimation
6. implement quantile metrics
7. HPO pipeline
8. more models?

References I