The Effectiveness of Discretization in Forecasting: An Empirical Study on Neural Time Series Models

6th Workshop on Mining and Learning from Time Series @ KDD 2020

Stephan Rabanser[1*] stephan.rabanser@mail.utoronto.ca
Tim Januschowski[2] tjnsch@amazon.com
Valentin Flunkert[2] flunkert@amazon.com
David Salinas[3] david.salinas@naverlabs.com
Jan Gasthaus[2] gasthaus@amazon.com

[1]University of Toronto
Vector Institute
*Work done at AWS AI Labs

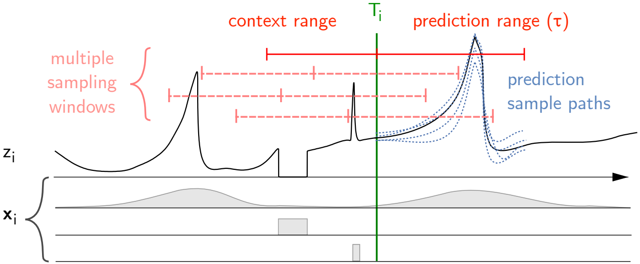
[2]Amazon
AWS AI Labs

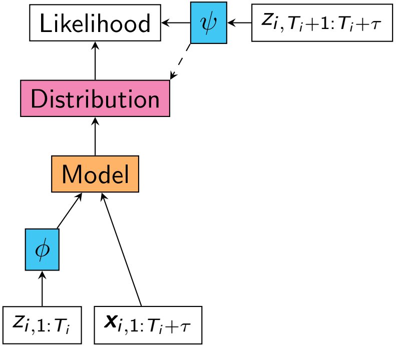
[3]NAVER LABS
Europe

August 24, 2020

Motivation & Setup

- Recent advancements in global forecasting: model architectures and probabilistic outputs.
- We investigate effects of (discrete) I/O representations.

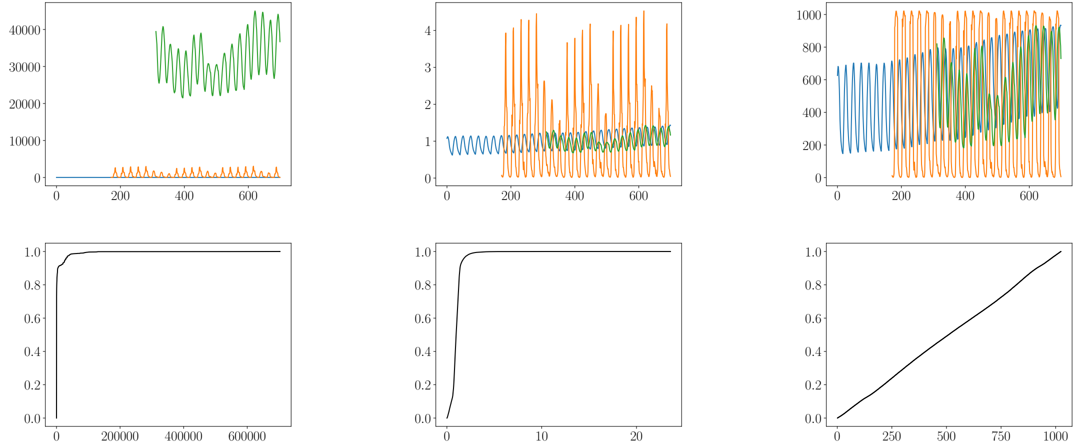




- ϕ: input transformation.
- ψ: output transformation, influences output distribution.

Scaling Problem: A Motivating Example (`m4_hourly`)

Original time series | Time series after scaling | Time series after q-transform



Continuous Transforms

Addressing the scaling problem in global forecasting is of utmost importance!

Scaling

Apply an affine transformation to each time series:

- General form: $z'_{i,t} = (z_{i,t} - b_i)/a_i$.
- Classic mean scaling (ms):
  - $a_i = \frac{1}{T_i} \sum_{t=1}^{T_i} |z_{i,t}|$
  - $b_i = 0$
- Lots of possible variations ...

Probability Integral Transform (pit)

Maps a RV Z through its CDF:

- $Y = F_Z(Z)$ with Y being uniform.
- Data preprocessing: make the empirical marginal of each time series approximately uniform [3].
- $z'_{i,t} = \hat{F}_i(z_{i,t})$ with $\hat{F}_i$ being the ECDF for time series $z_{i,1:T_i}$.

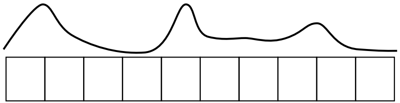
Discretizing Transforms

- Binning function $b : \mathbb{R} \to \{1, 2, \ldots, B\}$ mapping a real input to a discrete output.
- Each $b \in \{1, \ldots, B\}$ is tied to a bucket $S_b = [l_{b-1}, l_b)$: $b(z) = b$ iff $z \in S_b$.

Equally-Spaced Binning

Construct buckets to be equal in width:

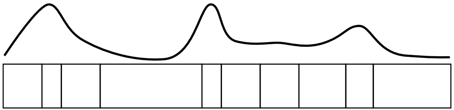
- Only optimal for uniform data.



Quantile Binning (discrete `pit`)

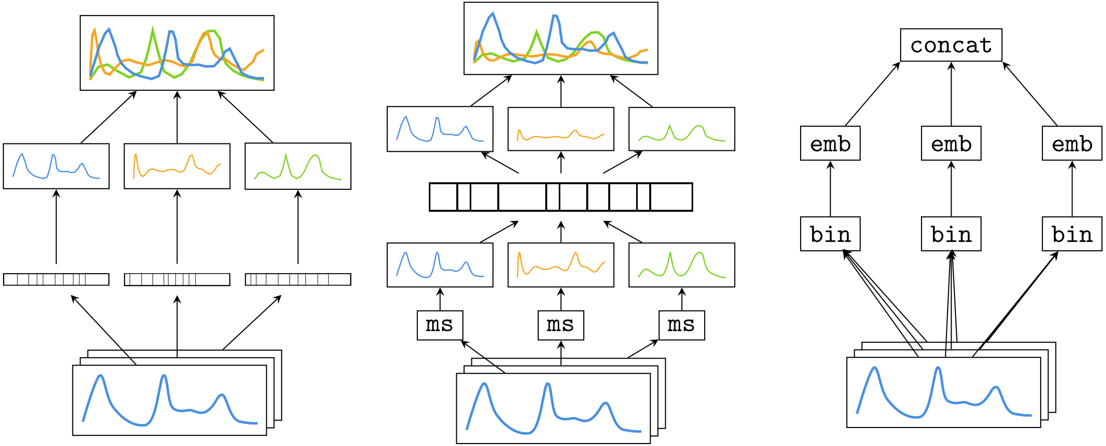
Construct buckets to be equal in mass:

- Adapts bins to fit the data distr.



Our Binning Strategies: Local Absolute & Global Relative Binning

Local Absolute Binning (`lab`) | Global Relative Binning (`grb`) | Hybrid Binning (`hyb`)

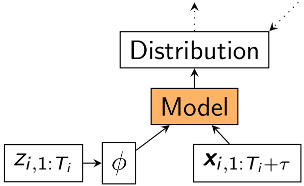


Models & Output Distributions

Models

We consider three different models which we combine with the aforementioned I/O transformations:

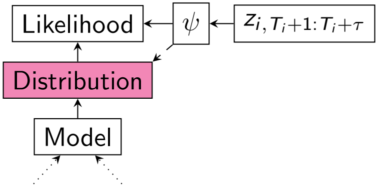
- Simple Feed Forward: SFF
- Autoregressive CNN: WaveNet [2]
- Autoregressive RNN: DeepAR [4]



Output Distributions

We compare three different approaches for modeling the output distribution $p(z_t|h_t)$:

- Student-t distribution (`st`);
- Piecewise-linear spline quantile function approach of [1] (`plqs`);
- Categorical distribution (`cat`);



Experimental Results

- Varying I/O representations with models on `m4`, `electricity`, `traffic`, `wiki`.

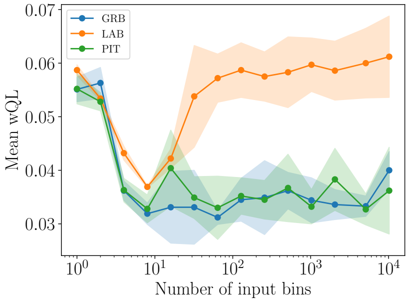
Output Scaling vs Binning

- Output representation has large perf. impact. Loss differences (max/min/avg):
  - WaveNet: 3.6x / 1.2x / 1.7x
  - DeepAR: 7.6x / 1.4x / 2.9x
  - SFF: 1.8x / 1.0x / 1.2x
- WaveNet profits a lot from binning (8/9), WaveNet with `grb` performs best (7/9).
- DeepAR shows degradation in perf. with binning over `ms` (avg 2.6x higher loss).
- Mixed results for SFF (no clear winner).

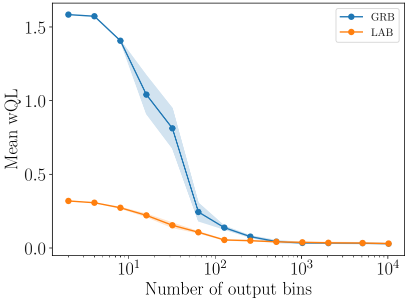
Input Scaling vs Binning

- Input representation has a smaller perf. impact. Loss differences (max/min/avg):
  - WaveNet: 3.0x / 1.4x / 1.9x
  - DeepAR: 5.7x / 1.0x / 1.9x
  - SFF: 1.8x / 1.0x / 1.2x
- There is no one clear dominant representation outperforming others.
- Multi-scale hybrid binning often does well (6/9), `lab` performs badly (9/9).
- `grb` and `pit` mostly on par (avg 1.4x).

Binning Resolution Effects (m4_hourly)



Performance effects of varying *input* binning resolutions w.r.t a fixed 1024-bin q-grb *output* binning.



Performance effects of varying *output* binning resolutions w.r.t a fixed 1024-bin q-grb *input* binning.

Summary

Picking a good I/O representation is equally important as selecting a good model!

Extended Paper: `https://arxiv.org/abs/2005.10111`

GluonTS: Probabilistic Time Series Modeling Library (Python):
`https://github.com/awslabs/gluon-ts`

References

J. Gasthaus, K. Benidis, Y. Wang, S. S. Rangapuram, D. Salinas, V. Flunkert, and T. Januschowski.
Probabilistic Forecasting with Spline Quantile Function RNNs.
In *The 22nd International Conference on Artificial Intelligence and Statistics*, 2019.

A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu.
Wavenet: A Generative Model for Raw Audio.
arXiv preprint arXiv:1609.03499, 2016.

D. Salinas, M. Bohlke-Schneider, L. Callot, R. Medico, and J. Gasthaus.
High-dimensional multivariate forecasting with low-rank gaussian copula processes.
In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 6824–6834. Curran Associates, Inc., 2019.

D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski.
DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks.
International Journal of Forecasting, 2019.