

Infinite forecast combinations based on Dirichlet process



Feng Li

feng.li@cufe.edu.cn

<http://feng.li/>

**School of Statistics and Mathematics
Central University of Finance and Economics**



Yinuo Ren
(PhD Student in CAS)



Feng Li*
(CUFE)



Yanfei Kang
(BUAA)



Jue Wang
(CAS)

- arXiv: <http://arxiv.org/abs/2311.12379>.
- Feng Li and Yanfei's research are supported by the National Natural Science Foundation of China.



*Forecast combination is **GREAT!***

- See our survey over the past 50 years (Wang, Hyndman, Li & Kang, 2023, IJF)



- In many machine learning applications, the performance depends on the number of candidate models.
- For example, N-BEATS (Oreshkin et al., 2020) are based on multi-scale aggregation with the median of bagging selected models. The model pool for ensembling consists of N-BEATS models fitted on one or several of the sMAPE, MASE and MAPE metrics and a selection of different window lengths e.g., $2 \times h$, $3 \times h$, ... , $7 \times h$, respectively, where h is the forecast horizon.
- In principle, it is desirable to have infinitely many candidate models to combine. In practice, it is not feasible.
- How to have a model pool with infinite models at a low cost, and then combine them?
- ... it is tricky.



- Not all best performed models will ensemble even better forecasts. The diversity of candidate models is of crucial importance (Kang, Cao, Petropoulos & Li, 2022, EJOR).
- With existing modeling architectures, we wish to generate a model pool with affordable costs.



Building up an infinite model pool

↪ Connecting the learning rate with model diversity

- The choice of learning rate plays a pivotal role as it is a critical factor in gradient descent during the optimization of the loss function.
- Many tools are developed to find an optimal learning rate to reach a local optimum.
- **But we do not! We think it differently!**
- Our approach in this study involves selecting a neural network as the single model and fixing the learning rate at a certain constant value.
- **The learning rate varying in the real line, e.g., $[0, 1]$, indicates infinite possibility of model.**



Then how do we collect (sample) from the infinite model pool?

↪ Dirichlet process

- Bayesian nonparametric models (Hjort et al., 2010) are a class of statistical models that allow for a flexible approach when modeling complex data patterns.
- Without the need for predefined parameters, it can capture intricate data patterns but might need more data for good performance.
- In this paper, we utilize an infinite mixture representation with the Dirichlet process.

$$G \sim \text{DP}(\alpha, H).$$

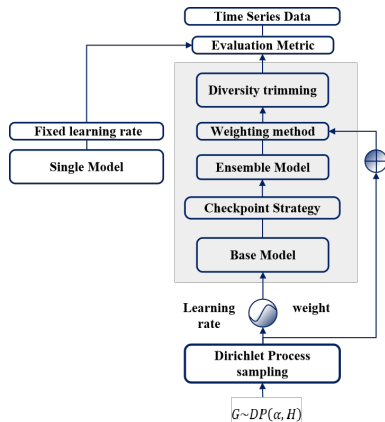
- The DP allows us to generate probability distributions with infinite dimensions. As a positive scale parameter, α determines the dispersion of the base distribution H . When $\alpha = 0$, the sample taken is degenerated as one value; $\alpha \rightarrow \infty$, it can be equated to the base distribution H . Thus, each sampling of its samples is a distribution, hence it is also referred to as “**a distribution of distributions**”.



- The samples drawn from the Dirichlet process are determined by the scaling parameter α and the base distribution H .
- Thus, the learning rate and the combination weight of the base learner m_i are derived from Dirichlet process sampling and correspond to β_i and π_i parameters in the stick-breaking process ($i = 1, 2, \dots, p$).

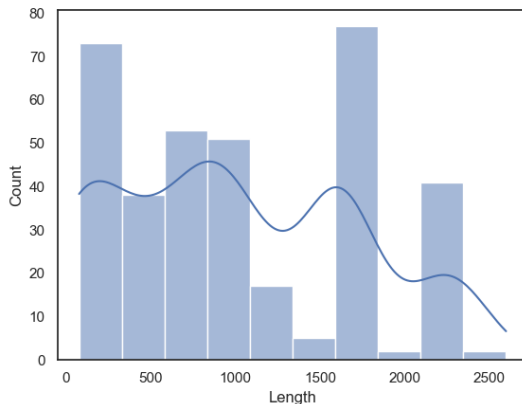
Combination strategies

- Upon the conclusion of the training phase, the straightforward procedure entails loading the previously stored weight parameter files to reconstruct the neural network architectures of the base models in the prediction stage.
- Consequently, for each data sequence, an ensemble of p prediction results is generated, and the ultimate prediction for that sequence is calculated by computing the weighted average value.

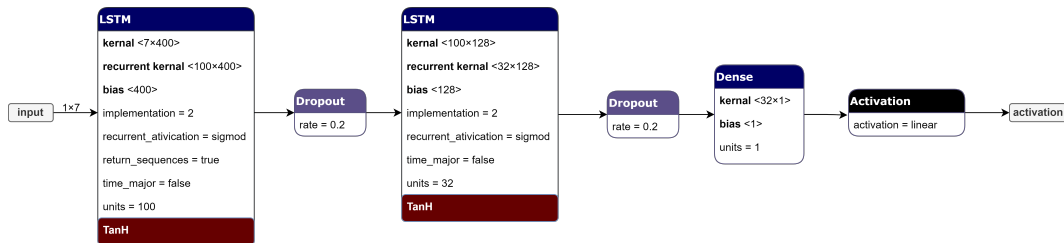


M4 Competition dataset

- The majority of series fall within the range of approximately 2000 weeks. The longest sequence spans up to 50 years, while the shortest sequence has a length of 276 weeks.



- Any machine learning model could be used and we use LSTM for illustration purposes.
- The schematic framework of the base model, where a lag of 7 steps is configured. It incorporates two LSTM modules and a Dropout layer, followed by a Dense layer and two activation functions to minimize the differences in weights and biases. To maintain consistency across other variables, a single model S is designated as the experimental group, the structure of which mirrors that of base models, except for a fixed learning rate set at 0.001.

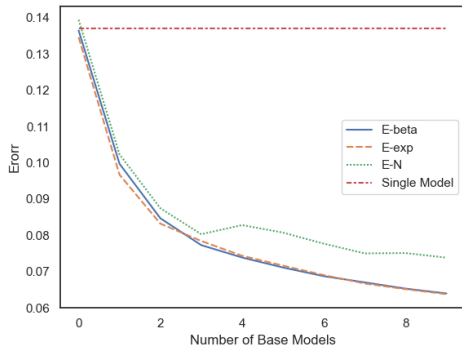
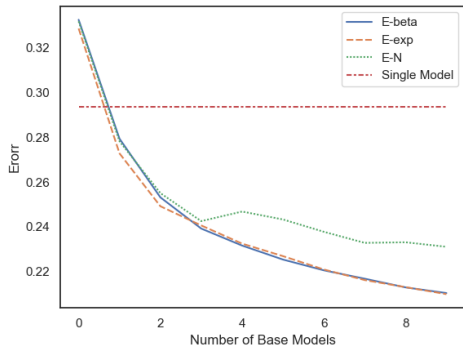




- **Input:** Time series data Y (Length T), Number of base models p , Hyperparameters (αH), Neural network M , Number of iterations I , Lag steps b .
 - **Output:** Base model m_i , Ensemble model E .
- ① Obtain a $(T - b) \cdot b$ training input matrix and a $(T - b)$ dependent variable vector after normalizing Y .
 - ② Derive a set of learning rates l and combination weight vectors w with the length of p .
 - ③ Insert the descending learning rate list l into the optimizer of M and at every l iteration save **Checkpoint** files using the decay algorithm, thereby completing the training of all base learners $m_i (i = 1, 2, \dots, p)$.
 - ④ Incorporate diversity trimming and weighting strategies to construct an ensemble prediction model and calculate prediction errors. The overall objective function is $\text{argminMetric}(E, l, w, p)$.

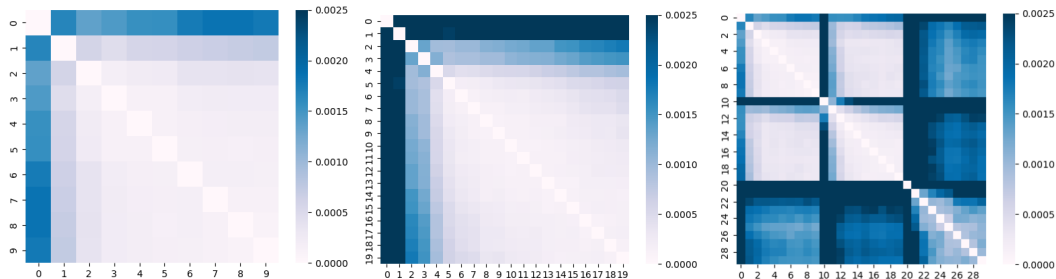
Model size and forecasting accuracy

- The Dirichlet process exhibits infinite possibilities, suggesting that the sample size p can be extended to infinity in a single draw.
- However, to translate the infinite concept into a finite context for practical empirical analysis, this paper initiates experiments exploring how the number of models impacts the ensemble effect.

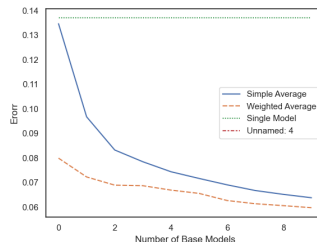
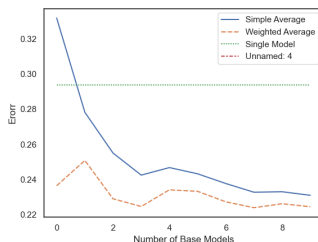
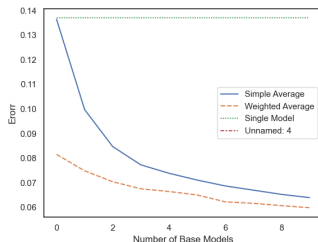
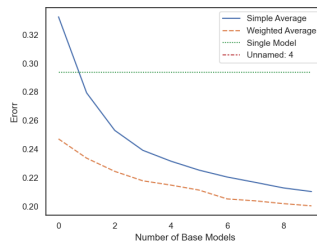
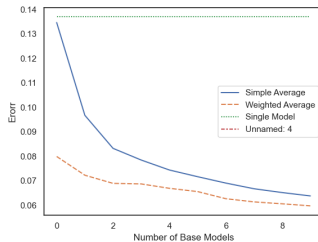
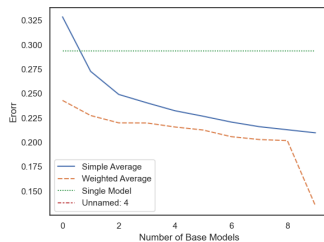


Model diversity on forecasting accuracy

- shows the diversity correlation matrices, and it can be seen that as the number of models increases, the diversity among the base learners steadily intensifies.
- This observation aligns with the earlier discussed trend of decreasing prediction error. Specifically, at a model number of 10, the lack of diversity renders E unable to enhance prediction accuracy.
- However, when p reaches 20, there is a significant order of magnitude increase in diversity, resulting in a remarkable reduction in prediction error.




Combination weights and prediction accuracy





- Multiple model classes.
- Different model pooling strategies.



-  Wang, X., Hyndman, R. J., Li, F. & Kang, Y. (2023). “Forecast Combinations: An over 50-Year Review”. *International Journal of Forecasting* 39.(4), pp. 1518–1547.
-  Kang, Y., Cao, W., Petropoulos, F. & Li, F. (2022). “Forecast with Forecasts: Diversity Matters”. *European Journal of Operational Research* 301.(1), pp. 180–190.
-  Oreshkin, B. N., Carпов, D., Chapados, N. & Bengio, Y. (2020). “N-BEATS: Neural Basis Expansion Analysis for Interpretable Time Series Forecasting”. In: *Proceedings of The 2020 International Conference on Learning Representations*.
-  Hjort, N. L., Holmes, C., Müller, P. & Walker, S. G., eds. (2010). *Bayesian Nonparametrics*. Illustrated edition. Cambridge: Cambridge University Press. 308 pp. ISBN: 978-0-521-51346-3.



Thank you!

Web: <https://feng.li>

Lab: <https://kllab.org>