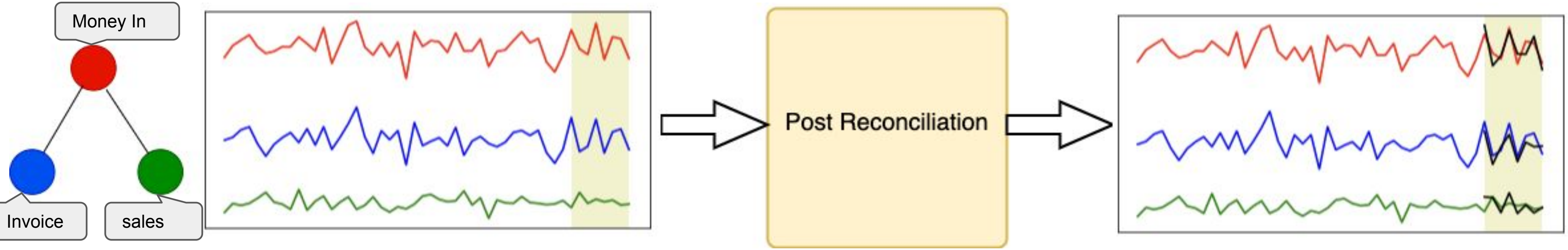
Simultaneously Reconciled Quantile Forecasting of Hierarchically Related Time Series (SHARQ)

Han, X., Dasgupta, S., Ghosh, J.
Simultaneously Reconciled Quantile Forecasting of Hierarchically Related Time Series
International Conference on Artificial Intelligence and Statistics (AISTATS), 2021

One-Slide Summary

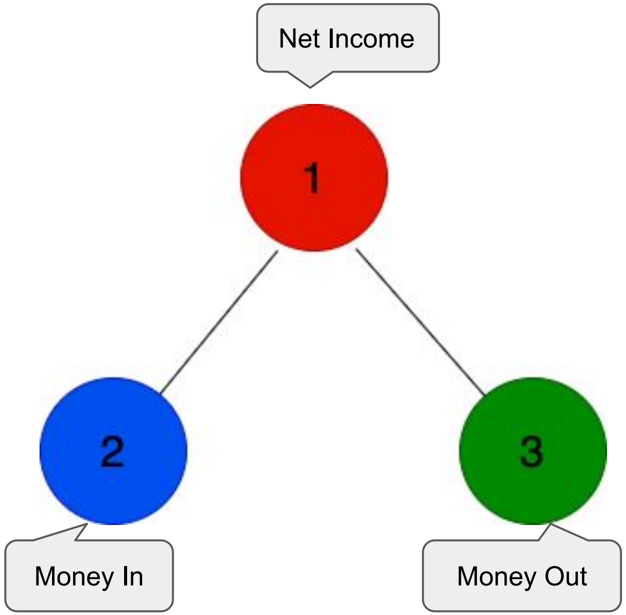
- Motivation:
  - Provide coherent point and probabilistic multi-step forecasts for hierarchical time-series
- Approach - Simultaneous Reconciled Quantile Forecasting:
  - Loss function
    - Captures hierarchical relationships among time-series from adjacent aggregation levels
    - Learns multiple probabilistic forecasts at the same time
  - Probabilistic forecasts simultaneously reconciled during model training
- Empirical Values:
  - Produced accurate and coherent point forecasts
  - Provided prediction intervals at any specified level

Existing Works: Post Reconciliation



- Post Reconciliation can be posed as OLS
- The reconciled Forecasts are computed as: $\tilde{\boldsymbol{y}}_T(h) = \boldsymbol{S}(\boldsymbol{S}'\boldsymbol{\Sigma}_h^{\dagger}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{\Sigma}_h^{\dagger}\hat{\boldsymbol{y}}_T(h),$
- Challenges:
  - Only works on assumptions of unbiased base forecasts and Gaussian noise
  - Matrix inversion is not stable and time consuming
  - Forecasts are “forced” to sum up (a hard constraint)
  - Cannot provide prediction intervals

SHARQ: Reconciling Point Forecast



Estimate Multiple Quantiles

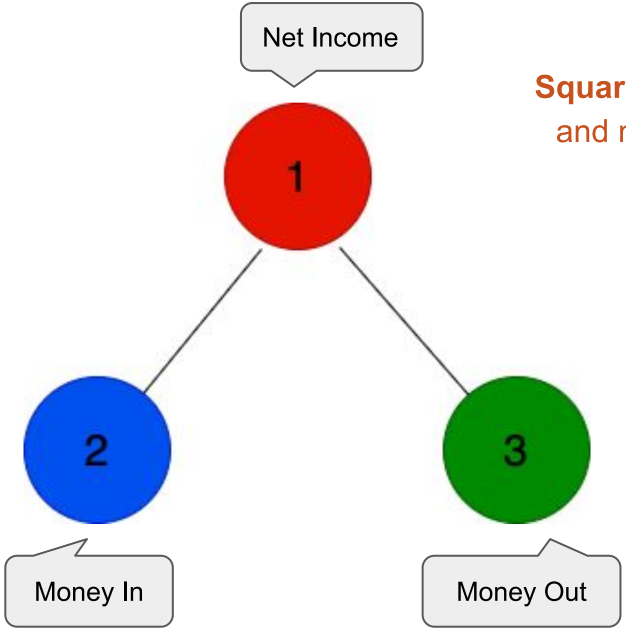
Reconciliation

$$L(\hat{y}_1, y_1) = \boxed{\sum_{\tau=\tau_0}^{\tau_q} \rho_\tau(\hat{y}_1, y_1)} + \lambda \boxed{(\hat{y}_1^{50} - \sum_{i \in \{2,3\}} \hat{y}_i^{50})^2}$$

$$L(\hat{y}_i, y_i) = \sum_{\tau=\tau_0}^{\tau_q} \rho_\tau(\hat{y}_i^\tau, y_i) \quad i = 2, 3$$

Where $\hat{y}_{t+1}^\tau \in \underset{q}{\operatorname{argmin}}\ \rho_\tau(q, y_{t+1}) = \begin{cases} (y_{t+1} - q)\,\tau & y_{t+1} \geq q \\ (y_{t+1} - q)\,(\tau - 1) & y_{t+1} < q \end{cases}$

SHARQ: Reconciling Quantile Forecasts



$$L(\hat{y}_1,\, \hat{y}_2,\, \hat{y}_3) = \sum_{\tau=\tau_0}^{\tau_q} \left[(\hat{y}_1^{\tau} - \hat{y}_1^{50})^2 - (\hat{y}_2^{\tau} - \hat{y}_2^{50})^2 - (\hat{y}_3^{\tau} - \hat{y}_3^{50})^2 \right]^2$$

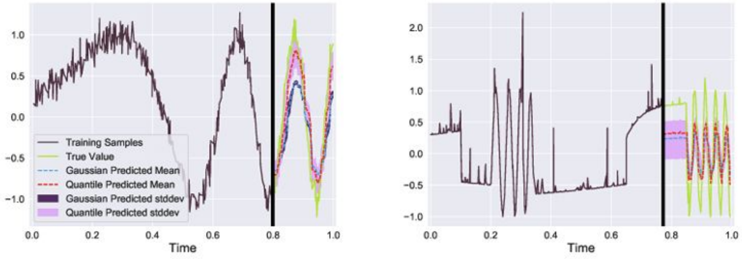
Squared distance between each quantile and median follows additive property in Gaussian distribution.

$$L(\hat{y}_1, y_1) = \sum_{\tau=\tau_0}^{\tau_q} \rho_{\tau}(\hat{y}_1, y_1) + \lambda\,(\hat{y}_1^{50} - \sum_{i\in\{2,3\}} \hat{y}_i^{50})^2$$

$$L(\hat{y}_i, y_i) = \sum_{\tau=\tau_0}^{\tau_q} \rho_{\tau}(\hat{y}_i^{\tau}, y_i) \quad i = 2, 3$$

1. Han, X., Dasgupta, S., Ghosh, J.. (2021). Simultaneously Reconciled Quantile Forecasting of Hierarchically Related Time Series. Proceedings of The 24th International Conference on Artificial Intelligence and Statistics

Results & Discussions



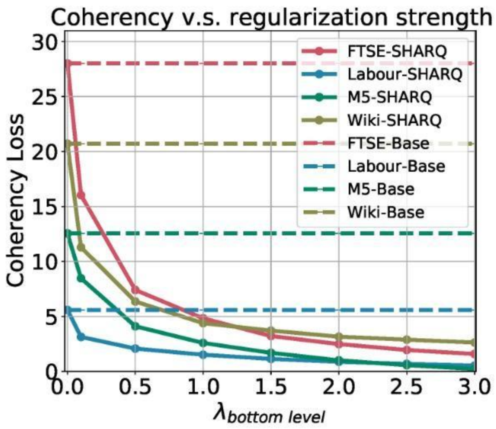
(a) Log-normal error (b) Gamma error

Algorithm	RNN			
Reconciliation	Level			
	1	2	3	4
BU	15.23	15.88	19.41	**17.96**
Base	12.89	14.26	16.96	**17.96**
MinT-sam	14.98	15.94	17.79	19.23
MinT-shr	14.46	15.43	16.94	18.75
MinT-ols	15.01	15.96	18.75	19.21
ERM	14.73	16.62	19.51	20.13
SHARQ	**12.55**	**13.21**	**16.01**	**17.96**

Australian Labour data: 40 vertices, 4 levels;
Performance measured by MAPE;
Level1 is the top aggregation level

- Quantile forecast is robust to multiple types of noise (log normal, gamma)
- Performs better than bagging method in terms of prediction intervals
- Performed better than other baseline methods, particularly at higher aggregation levels

Results & Discussions



Coherency loss drops dramatically after incorporating the regularization

Time (s)	FTSE		Labour		M5		Wikipedia	
	training	inference	training	inference	training	inference	training	inference
Base	115.96	0.01	68.35	0.00	181.58	0.00	205.47	0.01
BU	65.83	0.03	57.06	0.00	105.45	0.00	142.53	0.01
MinT-sam	106.55	1,784.77	72.24	430.42	172.11	1,461.81	208.26	1,106.70
MinT-shr	104.35	1,148.49	60.83	317.02	175.83	1,039.53	198.16	788.31
MinT-ols	103.23	1,129.45	64.14	310.13	163.24	977.88	196.88	702.02
ERM	547.66	0.05	497.88	0.01	551.60	0.01	1,299.30	0.04
SHARQ	121.84	0.01	99.96	0.00	201.40	0.00	241.97	0.01

Comparing the average training and inference time across forecasting models

- SHARQ provides a learnable trade-off between coherency and accuracy

- SHARQ reduces inference time from post-training reconciliation methods

Dynamic Combination of Heterogeneous Models for Hierarchical Time Series (DYCHEM)

Hu, J, Han, X., Ghosh, J.
Dynamic Combination of Heterogeneous Models for Hierarchical Time Series
International Conference on Data Mining (ICDM), 2022

One-Slide Summary

- Motivations:
  - Borrow strength from heterogeneous forecasting models
  - Allow any time series time algorithm to be used as a base forecasting method
  - Eliminate quantile crossing

- Dynamic Combination of Heterogeneous Models:
  - Improves point forecasts by utilizing multiple forecasting models via a mixture-of-expert framework
  - Allows mixing of different time series forecasting algorithms
  - Quantile forecasts are coherent and without quantile crossing

- Empirical Values:
  - Significantly improved point & probabilistic forecasting performance

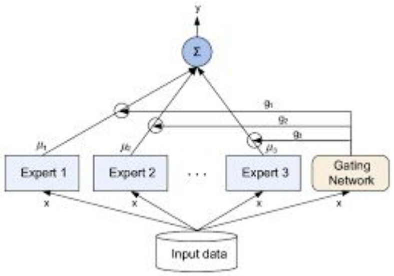
Motivations

- We wish to borrow the strength of heterogeneous forecasting models (DLM, Prophet, Auto-ARIMA, Deep learning methods), given their expertise on different types of data

- SHARQ requires to modify the objective function of forecasting models, which is not always available for existing state-of-the-art forecasting libraries

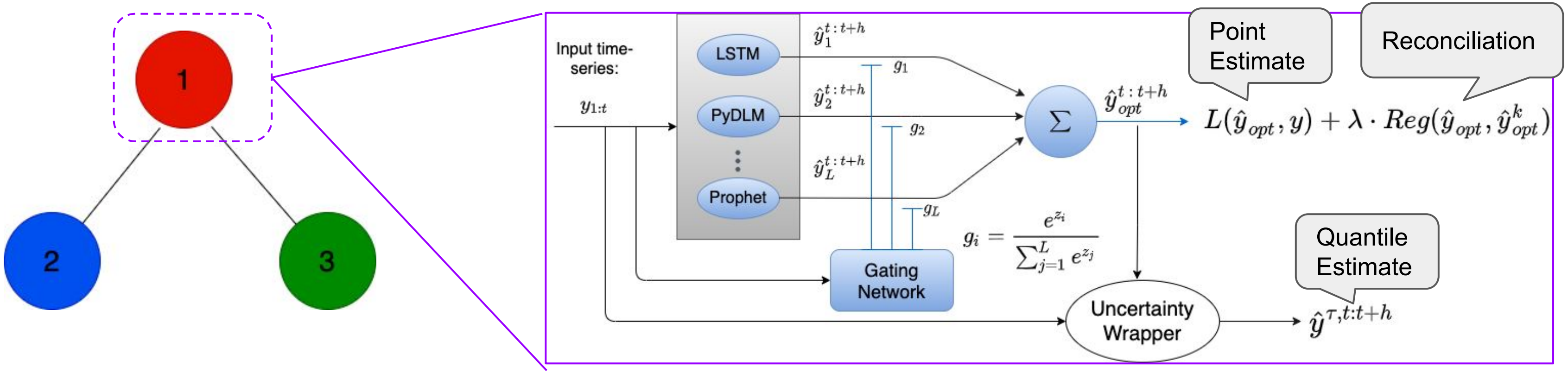
- We wish to build a framework where the forecasting models can be independent and user-specified

- We wish to keep simultaneous quantile estimations, without quantile crossing



Mixture-of-Experts: is an ensemble learning method that utilize gating network to combine outputs from each expert.

MECATS[1] - Mixture of Experts



Improved SHARQ via:

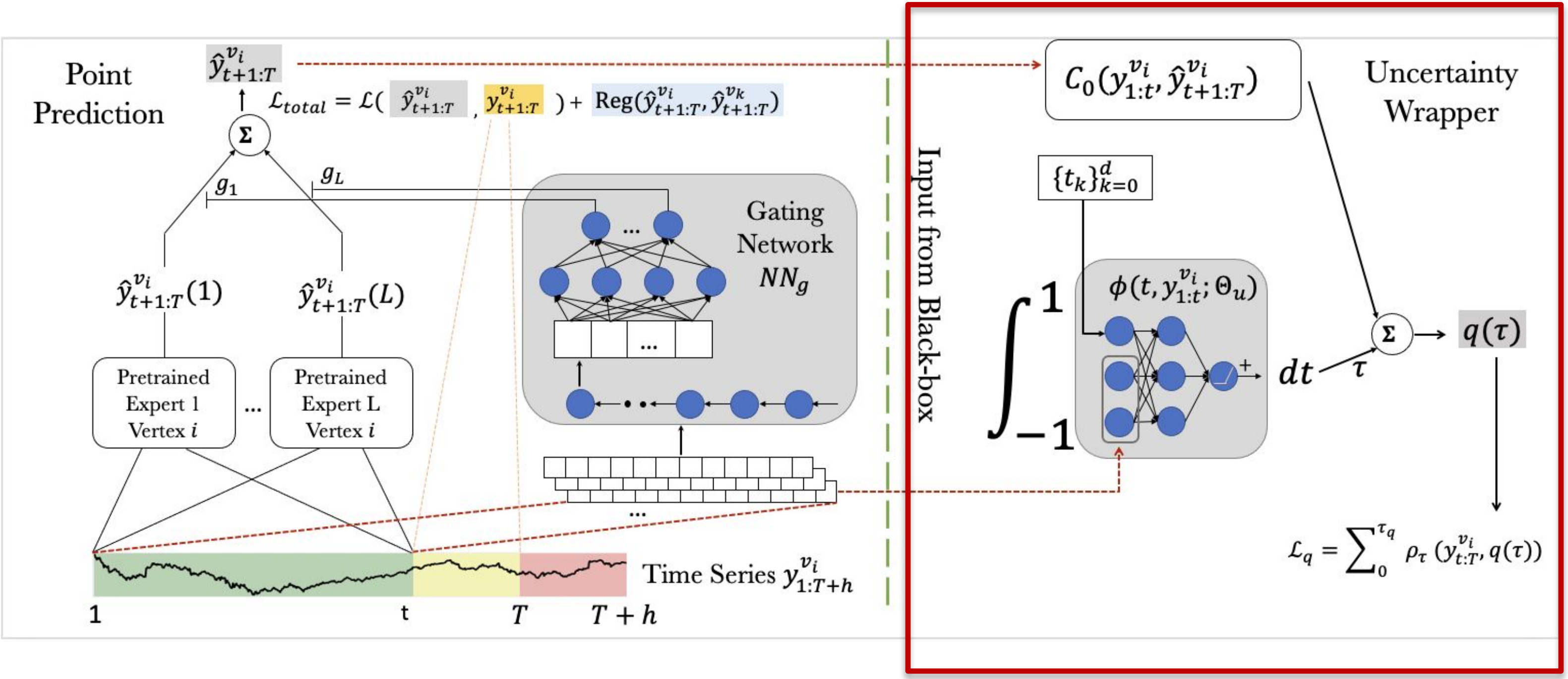
- Combine heterogeneous models at each vertex.
- Each expert is independently configurable and replaceable; overall is a black-box model.
- How about quantile estimations? Simply combining distributions won't work.

1. MECATS stands for Mixture-of-Experts for Coherent and Aggregated Temporal Sequences

Overall Structure

Requires 3 inputs:
- Point prediction
- Time series training data
- Quantile levels

Outputs:
- Quantile estimations at specified levels



- The heterogeneous experts bring robustness to the prediction.

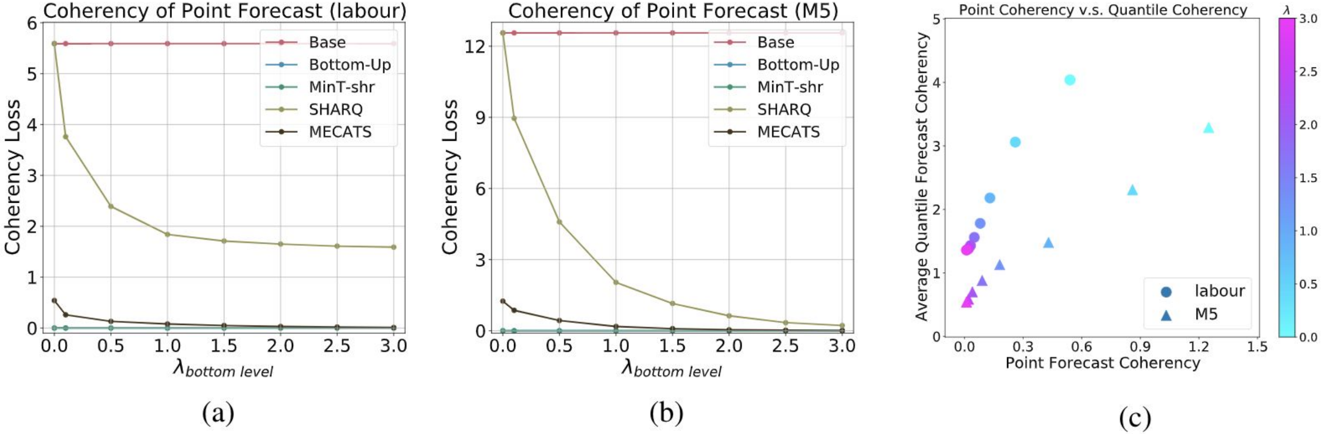
DYCHEM: Comparison with Baselines

Data \ Method	Level	DYCHEM-LOO	SHARQ	HIER-E2E	Average	DYCHEM
Labour	1	43.33 ±0.42 (.054)	52.07 ±0.45 (.085)	45.12 ±0.23 (.085)	49.34 ±0.65 (.075)	**38.84 ±0.04 (.045)**
	2	53.68 ±0.68 (.104)	58.69 ±0.41 (.120)	55.61 ±0.74 (.107)	60.87 ±0.33 (.119)	**48.64 ±0.78 (.092)**
	3	57.16 ±0.25 (.135)	64.02 ±0.09 (**.132**)	60.03 ±0.26 (.134)	69.29 ±0.42 (.138)	**49.17 ±0.36** (.144)
	4	65.05 ±0.18 (**.153**)	72.13 ±0.34 (.167)	71.38 ±0.15 (.154)	75.56 ±0.94 (.156)	**61.22 ±0.14** (.163)
M5	1	49.29 ±0.34 (.071)	56.31 ±0.17 (.054)	51.69 ±0.05 (.070)	59.61 ±0.38 (.104)	**42.61 ±0.14 (.046)**
	2	54.36 ±0.28 (.127)	62.16 ±0.27 (**.079**)	54.72 ±0.63 (.116)	60.48 ±0.58 (.133)	**49.75 ±0.22** (.084)
	3	55.18 ±0.22 (.142)	65.37 ±0.63 (.134)	65.02 ±0.24 (.142)	68.29 ±0.25 (.143)	**53.61 ±0.42 (.101)**
	4	59.04 ±0.36 (.164)	72.86 ±0.27 (.189)	72.04 ±0.36 (.164)	70.29 ±0.34 (.168)	**57.89 ±0.47 (.109)**
AEDemand	1	61.35 ±0.76 (.132)	68.19 ±0.29 (**.113**)	64.45 ±0.48 (.213)	67.32 ±0.29 (.164)	**59.89 ±0.32** (.145)
	2	58.12 ±0.46 (.152)	66.57 ±0.24 (.199)	63.72 ±0.36(.131)	63.58 ±0.72 (.129)	**55.72 ±0.73 (.122)**
	3	66.38 ±0.78 (.124)	68.25 ±0.47 (.131)	68.01 ±0.22 (.126)	70.44 ±0.09 (.124)	**62.55 ±0.14 (.111)**
	4	76.58 ±0.63 (.136)	87.35 ±0.69 (.225)	82.47 ±0.28 (.192)	73.22 ±0.37 (.135)	**71.45 ±0.43 (.125)**
Wiki	1	65.98 ±0.22 (.121)	70.36 ±0.24 (.147)	69.67 ±0.58 (**.067**)	66.42 ±0.16 (.128)	**63.27 ±0.73** (.117)
	2	68.54 ±0.47 (.157)	73.06 ±0.42 (.159)	68.24 ±0.33 (**.108**)	72.01 ±0.52 (.157)	**65.14 ±0.46** (.143)
	3	72.42 ±0.36 (.149)	76.15 ±0.34 (**.135**)	74.62 ±0.19 (.155)	74.37 ±0.83 (.147)	**69.48 ±0.33** (.156)
	4	77.12 ±0.23 (.268)	78.42 ±0.34 (.201)	79.63 ±0.41 (.291)	81.38 ±0.65 (.278)	**75.69 ±0.76 (.189)**
	5	84.77 ±0.49 (.241)	85.12 ±0.62 (.345)	79.65 ±0.24 (.326)	84.68 ±0.42 (.221)	**76.88 ±0.72 (.213)**

- Forecasting performance measured by averaged MASE and CRPS (within bracket). All experiments are repeated 5 times.
- Baselines: simple averaging (Average); SHARQ; leave-one-out of one expert; HIER-E2E [1]

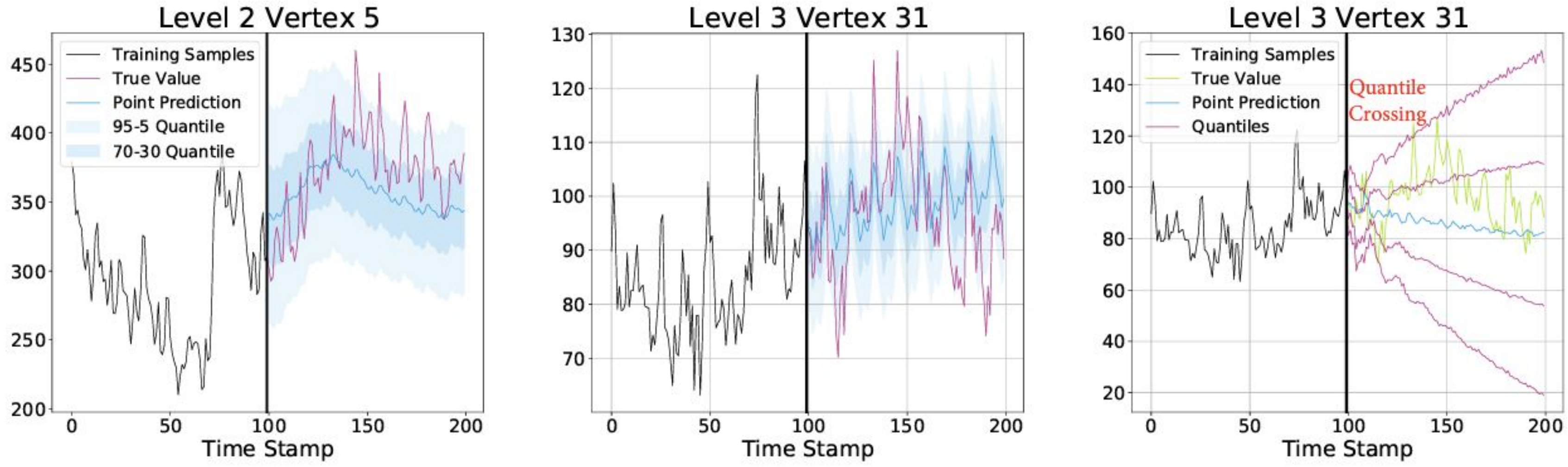
[1] End-to-End Learning of Coherent Probabilistic Forecasts for Hierarchical Time Series

DYCHEM: Coherency Analysis



(a), (b) Coherent loss of point forecast w.r.t. regularization strength λ on two datasets. (c) Relationship between point forecast coherency and average of quantile coherency

Generating Non-Crossing Probabilistic Forecasts



(a) (b) (c)

Fig. 4. (a), (b) Quantile forecasting results generated by DYCHEM at vertex 5 and 31 of the Australian Labour data, where $\tau_s = [0.05, 0.3, 0.5, 0.7, 0.95]$. (c) SHARQ at same τ_s, results showing mild quantile crossing.

Efficient Forecasting of Large-Scale Hierarchical Time Series via Multilevel Clustering

Hu, J., Han, X., Ren, T., Ghosh, J., Ho, N.
Efficient Forecasting of Large-Scale Hierarchical Time Series via Multilevel Clustering
International Conference on Time Series and Forecasting, 2023

One-Slide Summary

- Motivations:
  - Accelerate large-scale hierarchical time-series (HTS) forecasting
  - Provide <u>analytical insights</u> to large-scale user records with multi-level structure

- Hierarchical Time Series Clustering:
  - Enables multi-level time-series clustering with <u>arbitrary lengths</u>
  - Leveraged <u>local aggregation information</u> to improve time series clustering

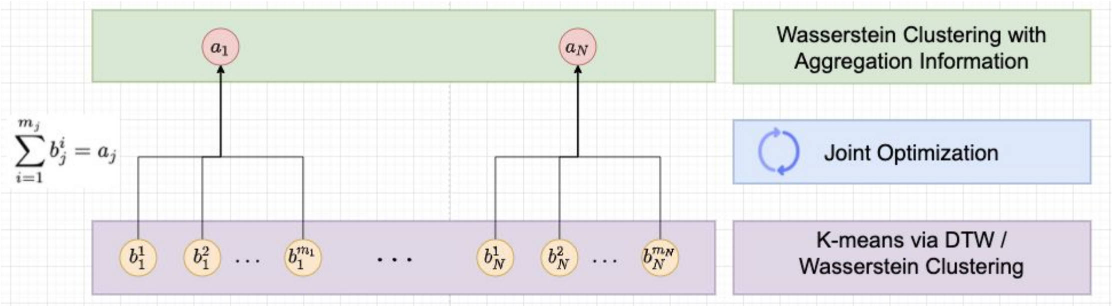
- Empirical Values:
  - Provided superior clustering results for time-series with multilevel structures
  - Improved forecasting efficiency for large-scale hierarchical time-series

Existing Approaches

- Mainly focused on time series data without hierarchical structure
  - Distance based approach
    - Define an appropriate distance measurement and cluster based on the distance
    - Extract temporal information capturing features and cluster in the embedding space
  - Model based approach
    - Specify the generative model type and estimate the parameters using MLE
    - Suitable if the data property is known
  - Distance based approach has no model assumptions and is computationally more efficient than Model based approach

High Level View of The Approach

- Our approach to a 2-level hierarchical time-series data:
  - Cluster bottom level data via DTW based K-means or Wasserstein Clustering
  - Assign a probability measure to each aggregated-level time series according to the cluster assignment of the child node.
    - Leverage local information in hierarchical time series clustering
  - Cluster top level data via Wasserstein Clustering incorporating the aggregation information
  - Using Wasserstein Clustering, both levels can be simultaneously clustered through joint optimization



Accelerating Hierarchical Forecasting via Clustering

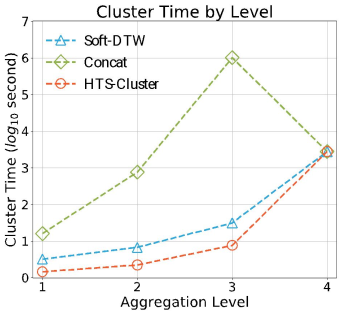
- Forecasts for individual hierarchical time series can “borrow strength” from the forecasts of nearest cluster means at each level.

- Utilizing fuzzy clustering, forecast of each time series can be represented as the weighted combination of forecasts of corresponding cluster means.

- One can quickly finetune results of each time-series to tailor the individual specifics.

Simulation & E-Commerce Clustering Results

	Method\Metric	Time (s)	Global NMI	Global AMI	Global ARI	Local NMI	Local AMI	Local ARI
Simulation	DTCR	132	0.325 ± 0.012	0.257 ± 0.023	0.21 ± 0.011	0.392 ± 0.014	0.313 ± 0.006	0.284 ± 0.009
	Soft-DTW	67	0.412 ± 0.009	0.326 ± 0.019	0.277 ± 0.008	0.411 ± 0.022	0.342 ± 0.009	0.304 ± 0.014
	Concat	186	0.436 ± 0.015	0.342 ± 0.014	**0.314** ± 0.016	0.411 ± 0.022	0.342 ± 0.009	0.304 ± 0.014
	HTS-Cluster	**37**	**0.455** ± 0.018	**0.354** ± 0.015	0.302 ± 0.013	**0.424** ± 0.018	**0.366** ± 0.013	**0.321** ± 0.018
Real-world	DTCR	72	0.065 ± 0.002	0.015 ± 0.001	0.008 ± 0.002	0.105 ± 0.011	0.059 ± 0.002	0.054 ± 0.003
	Soft-DTW	49	0.119 ± 0.005	0.043 ± 0.003	0.027 ± 0.003	0.126 ± 0.008	**0.082** ± 0.006	0.061 ± 0.005
	Concat	174	**0.135** ± 0.004	0.073 ± 0.007	**0.045** ± 0.006	0.126 ± 0.008	**0.082** ± 0.006	0.061 ± 0.005
	HTS-Cluster	**34**	0.134 ± 0.005	**0.075** ± 0.005	0.041 ± 0.004	**0.128** ± 0.014	0.064 ± 0.005	**0.065** ± 0.002



- Baseline method
  - DTCR: Deep Temporal Clustering Representation
    - Regular multivariate time series clustering without considering hierarchical structure.
- Competing methods
  - Soft-DTW: Soft-DTW divergence-based K-means on all levels.
  - Concat: simply concatenate local time-series for global time series.
  - HTS-Cluster: Our method, jointly optimizing two levels.

Accelerating Hierarchical Forecasting

Method \ Level	1	2	3	4	Total Time
Without cluster	62.39	76.26	78.25	84.14	1
DTCR	82.35	96.09	104.85	104.33	0.39
Soft-DTW	78.61	93.04	93.12	96.76	0.27
Concat	74.24	84.65	83.73	96.76	0.57
HTS-Cluster	72.99	80.07	85.29	96.76	0.16

Forecasting massive HTS with the help of clustering, results are measured by mean absolute scaled error (MASE) and relative computing time.

- Jointly model similar HTS via clustering is more efficient than building independent models without sacrificing too much accuracy.