Simultaneously Reconciled Quantile Forecasting of Hierarchically Related Time Series

Xing Han
UT Austin
aaronhan223@utexas.edu

Sambarta Dasgupta
IntuitAI
dasgupta.sambarta@gmail.com

Joydeep Ghosh
UT Austin
jghosh@utexas.edu

Contents

A Further Discussion on Related Works

As we mentioned in Section (1), state-of-the-art hierarchical forecasting algorithms (Ben Taieb and Koo, 2019; Hyndman et al., 2011, 2016; Wickramasuriya et al., 2015) involves computing the optimal P matrix to combine the base forecasts under different situations linearly. We now summarize these methods as follows.

A.1 Generalized Least Squares (GLS) Reconciliation

Denote $b_t \in \mathbb{R}^m$, $a_t \in \mathbb{R}^k$ as the observations at time t for the m and k series at the bottom and aggregation level(s), respectively. $S \in \{0,1\}^{n\times m}$ is the summing matrix. Each entry S_{ij} equals to 1 if the i^{th} aggregate series contains the j^{th} bottom-level series, where $i = 1, ..., k$ and $j = 1, ..., m$. Denote $\mathcal{I}_T = \{y_1, y_2, ..., y_T\}$ as the time series data observed up to time T; $\hat{b}_T(h)$ and $\hat{y}_T(h)$ as the $h-$step ahead forecast on the bottom-level and all levels based on $\mathcal{I}_T$.

Let $\hat{e}_T(h) = y_{T+h} - \hat{y}_T(h)$ be the $h-$step ahead conditional base forecast errors and $\beta_T(h) = E[\hat{b}_T(h) \mid \mathcal{I}_T]$ be the bottom-level mean forecasts. We then have $E[\hat{y}_T(h) \mid \mathcal{I}_T] = S\beta_T(h)$. Assume that $E[\hat{e}_T(h) \mid \mathcal{I}_T] = 0$, then a set of reconciled forecasts will be unbiased iff $SPS = S$, i.e.,
Assumption A1:

$$\mathbb{E}[\tilde{y}_T(h)|\mathcal{I}_T] = \mathbb{E}[\hat{y}_T(h)|\mathcal{I}_T] = S\beta_T(h) \tag{1}$$

The optimal combination approach proposed by Hyndman et al. (2011), is based on solving the above regression problem using the generalized least square method:

$$\hat{y}_T(h) = S\beta_T(h) + \varepsilon_h, \tag{2}$$

where ε_h is the independent coherency error with zero mean and $\text{Var}(\varepsilon_h) = \Sigma_h$. The GLS estimator of $\beta_T(h)$ is given by

$$\hat{\beta}_T(h) = (S'\Sigma_h'S)^{-1}S'\Sigma_h'\hat{y}_T(h), \tag{3}$$

which is an unbiased, minimum variance estimator. The optimal P is $(S'\Sigma_h'S)^{-1}S'\Sigma_h'$. The reconciled mean and variance can therefore be obtained accordingly.

A.2 Trace Minimization (MinT) Reconciliation

Defining the reconciliation error as $\tilde{e}_T(h) = y_{T+h} - \tilde{y}_T(h)$, the original problem can also be formulated as

$$\begin{aligned} &\min_{P\in\mathcal{P}} \mathbb{E}[\|\tilde{e}_T(h)\|_2^2 \mid \mathcal{I}_T] \\ &\text{subject to } \mathbb{E}[\tilde{y}_T(h) \mid \mathcal{I}_T] = \mathbb{E}[\tilde{y}_T(h) \mid \mathcal{I}_T] \end{aligned} \tag{4}$$

If the assumption **A1** still holds, then minimizing Eq.(4) reduces to

$$\min_{P\in\mathcal{P}} \text{Tr}(\text{Var}[\tilde{e}_T(h) \mid \mathcal{I}_T]) \quad \text{subject to } \mathbf{A1}, \tag{5}$$

where Tr(.) denotes the trace of a matrix. In Wickramasuriya et al. (2019), the proposed optimal solution of P obtained by solving this problem is given by

$$P = (S'W_h^{-1}S)^{-1}S'W_h^{-1}, \tag{6}$$

where $W_h = \mathbb{E}[\hat{e}_T(h)\hat{e}_T'(h) \mid \mathcal{I}_T]$ is the variance-covariance matrix of the $h-$step-ahead base forecast errors, which is different from the coherence errors Σ_h in GLS reconciliation method given in Eq.(3). There are various covariance estimators for W_h considered in Wickramasuriya et al. (2019), the most effective one is the shrinkage estimator with diagonal target, and can be computed by

$$\hat{W}_h = (1-\alpha)\hat{W}_s + \alpha\hat{W}_d, \quad \hat{W}_s = \frac{1}{T}\sum_{t=1}^{T}\hat{e}_t(1)\hat{e}_t(1)', \tag{7}$$

where $\hat{W}_d = \text{diag}(\hat{W}_s)$ and $\alpha \in (0, 1]$.

A.3 Empirical Risk Minimization (ERM) Reconciliation

Most recently, Ben Taieb and Koo (2019) proposed a new method to relax the unbiasedness assumption **A1**. Specifically, the objective function in (4) can be decomposed as

$$\mathbb{E}[\|y_{T+h} - \tilde{y}_T(h)\|_2^2 \mid \mathcal{I}_T] \tag{8}$$
$$= \|SP(\mathbb{E}[\hat{y}_T(h)|\mathcal{I}_T] - \mathbb{E}[y_{T+h}|\mathcal{I}_T]) + (S - SPS)\mathbb{E}[b_{T+h}|\mathcal{I}_T]\|_2^2 \tag{9}$$
$$+ \mathrm{Tr}(\mathrm{Var}[y_{T+h} - \tilde{y}_T(h)|\mathcal{I}_T], \tag{10}$$

where (9) and (10) are the bias and variance terms of the revised forecasts $\tilde{y}_T(h)$. The assumption **A1** in MinT method renders (9) to 0. Obviously, directly minimize the objective in (8) provides a more general form of reconciliation represented by following empirical risk minimization (ERM) problem:

$$\min_{P \in \mathcal{P}} \frac{1}{(T - T_1 - h + 1)n} \sum_{t=T_1}^{T-h} \|y_{t+h} - SP\hat{y}_t(h)\|_2^2, \tag{11}$$

where $T_1 < T$ is the number of observations used for model fitting. Empirically, this method demonstrates better performance than MinT according to Ben Taieb and Koo (2019), particularly when the forecasting models are mis-specified.

B Non-Additive Property of Quantile Loss

Here we prove the non-additive property of quantile loss as mentioned in Section (2.2).

THEOREM 1. ***(Non-additive Property)*** *Assume two independent random variables $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$, and define $Y = X_1 + X_2$. Then $Q_Y(\tau) \neq Q_{X_1}(\tau) + Q_{X_2}(\tau)$.*

PROOF. The τ^{th} quantile of X_1 is given by:

$$Q_{X_1}(\tau) = F_{X_1}^{-1}(\tau) = inf\{x : F_{X_1}(x) \geq \tau\}, \tag{12}$$

where $F_{X_1}(x)$ is $\frac{1}{2}\left[1 + erf\left(\frac{x-\mu_1}{\sigma_1\sqrt{2}}\right)\right]$, and $erf(x) = \frac{1}{\sqrt{\pi}}\int_{-x}^{x} e^{-t^2} dt$. Therefore, we can further get:

$$Q_{X_1}(\tau) = \mu_1 + \sigma_1\Phi^{-1}(\tau) = \mu_1 + \sigma_1\sqrt{2}\ erf^{-1}(2\tau - 1)$$
$$Q_{X_2}(\tau) = \mu_2 + \sigma_2\Phi^{-1}(\tau) = \mu_2 + \sigma_2\sqrt{2}\ erf^{-1}(2\tau - 1)$$

According to the additive property of Gaussian distribution, we have $Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$, and

$$Q_Y(\tau) = \mu_1 + \mu_2 + \sqrt{\sigma_1^2 + \sigma_2^2}\ \Phi^{-1}(\tau) = \mu_1 + \mu_2 + \sqrt{\sigma_1^2 + \sigma_2^2}\ \sqrt{2}\ erf^{-1}(2\tau - 1). \tag{13}$$

Therefore, even if we have `i.i.d.`normal distribution with $Y = X_1 + X_2$, it still doesn't imply $Q_Y(\tau) = Q_{X_1}(\tau) + Q_{X_2}(\tau)$. The only case that the addition property hold true in any quantile is when $X_1 = C \times X_2$, where C is arbitrary constant. Obviously, this is not applicable. □

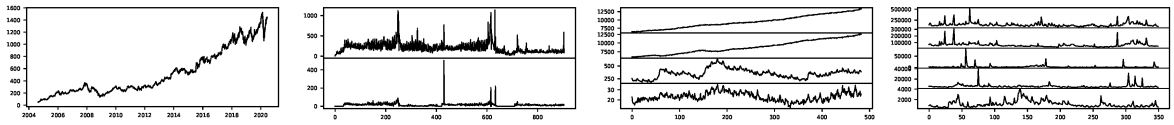
In fact, under Gaussian assumption, we have the following additive property holds for any τ:

$$(Q_Y^\tau - \mu_Y)^2 = (Q_{X_1}^\tau - \mu_{X_1})^2 + (Q_{X_2}^\tau - \mu_{X_2})^2. \tag{14}$$

Since by Eq.(13), the left hand side of Eq.(14) is $2(\sigma_1^2 + \sigma_2^2)\ \left[erf^{-1}(2\tau - 1)\right]^2$, and the right hand side of Eq.(14) is $2\sigma_1^2\left[erf^{-1}(2\tau - 1)\right]^2 + 2\sigma_2^2\left[erf^{-1}(2\tau - 1)\right]^2$. Therefore, the additive property holds for any τ assume the RVs follow Gaussian distribution.

Table 1: Details of four hierarchical time-series datasets. Note that hierarchical levels mean the number of aggregation levels from bottom to top in the hierarchical structure used in the experiments.

Dataset	Total number of time series	Total length of time series	Hierarchical Levels
FTSE	73	2512	4
M5	42840	1969	4
Wiki	145000	550	5
Labour	755	500	4



(a) FTSE (b) M5 (c) Labour (d) Wikipedia

Figure 1: Visualization of hierarchical time series data. (a) Bottom level time series of FTSE (the open stock price of Google); (b) bottom and top level of unit sales record; (c) Australian Labour Force data at all aggregation levels; (d) Wikipedia page views data at all aggregation levels.

C KKT Conditions

An alternative way of solving the optimization problem defined in Section (2.1) Eq.(1) is to obtain the KKT conditions (Boyd and Vandenberghe, 2004). For notational simplicity, we express the constrained loss for i^{th} vertex and m^{th} data point as $L_c(i,m)$. As the optimization problem is unconstrained, the KKT conditions will lead to:

$$\frac{\partial}{\partial[\theta_i,\Theta_i]}L_c(i,m) = [\ \frac{\partial}{\partial\theta_i}L_c(i,m)\ ,\ \frac{\partial}{\partial\Theta_i}L_c(i,m)\ ] = 0.$$

which will further imply that

$$\lambda_i \left[\frac{\partial}{\partial g_i}L(g_i(X_m^i,\theta_i),\ Y_m^i)\right]^\top \left[g_i(X_m^i,\theta_i) - \sum_{e_{i,k}\in E} e_{i,k}\ .\ g_k(X_m^k,\theta_k)\right] + \frac{\partial}{\partial g_i}L(g_i(X_m^i,\theta_i),\ Y_m^i)\ .\ \frac{\partial g_i}{\partial\theta_i} = 0,$$

and

$$\big(e_{i,j}\ .\ g_j(X_m^j,\theta_j)\big)^T \left(g_i(X_m^i,\theta_i) - \sum_{e_{i,k}\in E} e_{i,k}\ .\ g_k(X_m^k,\theta_k)\right) = 0, \quad \forall j | e_{i,j}\in E.$$

However, we found that SHARQ performs better and more efficiently than the KKT approach during our empirical evaluation. Solving the KKT conditions requires matrix inversion in most situations. Besides, SHARQ is more flexible in incorporating various forecasting models and performs probabilistic forecasts.

D Dataset Details

We first describe the details (dataset generation, processing, etc.) of each dataset used in the experiment. A summary of each dataset is shown in Table 1. Visualizations for some raw time series can be found in Figure 1.

D.1 FTSE Stock Market Data

The FTSE Global Classification System is a universally accepted classification scheme based on a market's division into Economic Groups, Industrial Sectors, and Industrial Sub-sectors. This system has been used to classify company data for over 30,000 companies from 59 countries. The FTSE 100 (Doherty et al., 2005) is the top 100 capitalized blue-chip companies in the UK and is recognized as the measure of UK stock market performance (Russell, 2017). Base on the FTSE classification system, we formulate a 4-level hierarchical structure (Economic

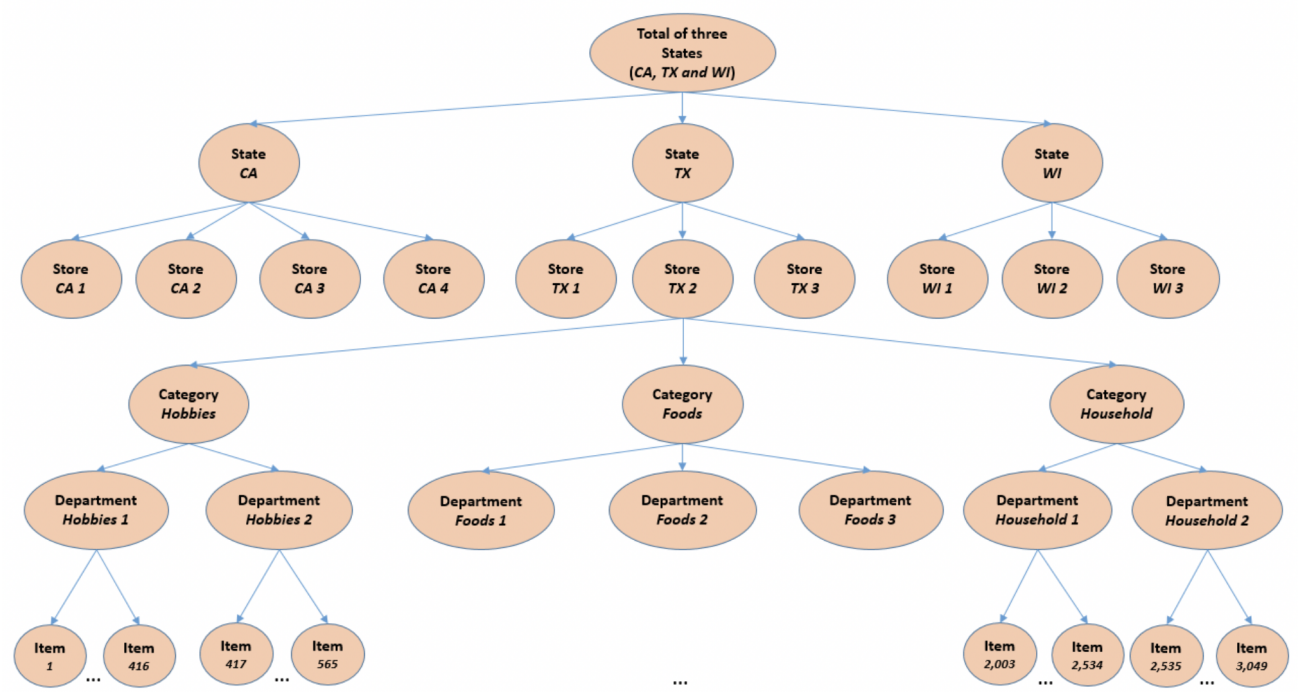


Figure 2: Hierarchical structure of the M5 dataset.

Groups, Industrial Sectors, Industrial Sub-sectors, and companies) of 73 companies in Doherty et al. (2005). Our task is to model the stock market time series for each company. The stock market data of each company is available from the Yahoo Finance package[1]. Since the stock market time series starting time of each company is not the same, we use a common time window ranging from January 4, 2010, to May 1, 2020.

D.2 M5 Competition Data

The M5 dataset[2] involves the unit sales of various products ranging from January 2011 to June 2016 in Walmart. It involves the unit sales of 3,049 products, classified into 3 product categories (Hobbies, Foods, and Household) and 7 product departments, where the categories mentioned above are disaggregated. The products are sold across ten stores in three states (CA, TX, and WI). An overview of how the M5 series are organized is shown in Figure 2. Here, we formulate a 4-level hierarchy, starting from the bottom-level individual item to unit sales of all products aggregated for each store.

D.3 Wikipedia Webpage Views

This dataset[3] contains the number of daily views of 145k various Wikipedia articles ranging from July 2015 to Dec. 2016. We follow the data processing approach used in Ben Taieb and Koo (2019) to sample 150 bottom-level series from the 145k series and aggregate to obtain the upper-level series. The aggregation features include the type of agent, type of access, and country codes. We then obtain a 5-level hierarchical structure with 150 bottom series.

[1] https://pypi.org/project/yfinance/
[2] https://mofc.unic.ac.cy/wp-content/uploads/2020/03/M5-Competitors-Guide-Final-10-March-2020.docx
[3] https://www.kaggle.com/c/web-traffic-time-series-forecasting

D.4 Australian Labour Force

This dataset[1] contains monthly employment information ranging from Feb. 1978 to Aug. 2019 with 500 records for each series. The original dataset provides a detailed hierarchical classification of labor force data, while we choose three aggregation features to formulate a 4-level symmetric structure. Specifically, the 32 bottom level series are hierarchically aggregated using labor force location, gender, and employment status.

E Additional Experiments

In this section, we demonstrate our additional experiment results, including the full results on FTSE and Wiki as well as additional simulation experiments under unbiasedness and Gaussian assumptions. Reconciliation error is also measured for each method. We start by discussing our evaluation metrics.

E.1 Evaluation Metrics

We denote $\hat{Y}_T(h)$ and $Y_T(h)$ as the $h-$step ahead forecast at time T and its ground truth, respectively. To construct confidence intervals, we use the $95^{th}, 50^{th}$, and 5^{th} quantiles as upper, median and lower forecasts.

E.1.1 Mean Absolute Percentage Error (MAPE)

The MAPE is commonly used to evaluate forecasting performance. It is defined by

$$\text{MAPE} = \frac{100}{H}\sum_{h=1}^{H}\frac{|Y_T(h) - \hat{Y}_T(h)|}{|Y_T(h)|}. \tag{15}$$

E.1.2 Likelihood Ratio

We compute the likelihood ratio between the quantile prediction intervals versus the trivial predictors, which gives the specified quantile of training samples as forecasts. Specifically, define N ($N = 3$ in our case) as the number of quantile predictors. Then the likelihood ratio at $h-$step forecast is:

$$\alpha = \frac{\sum_{i=1}^{N}\rho_{\tau_i}(Y_T(h) - Q_{Y_T(h)}(\tau_i))}{\sum_{i=1}^{N}\rho_{\tau_i}(Y_T(h) - Q_{\mathcal{I}_T}(\tau_i))}. \tag{16}$$

Ideally, we should have $\alpha < 1$ if our estimator performs better than the trivial estimator.

E.1.3 Continuous Ranked Probability Score (CRPS)

CRPS measures the compatibility of a cumulative distribution function F with an observation x as:

$$\text{CRPS}(F, x) = \int_{\mathbb{R}}(F(z) - \mathbb{I}\{x \le z\})^2\, dz \tag{17}$$

where $\mathbb{I}\{x \le z\}$ is the indicator function which is one if $x \le z$ and zero otherwise. Therefore, CRPS attains its minimum when the predictive distribution F and the data are equal. We used this library[2] to compute CRPS.

E.1.4 Reconciliation Error

We compute the reconciliation error of forecasts generated by each method on each dataset to measure the forecasting coherency. More specifically, assume a total of m vertices in the hierarchy at time T, the reconciliation error for the mean forecast is defined as

$$\frac{1}{H}\sum_{h=1}^{H}\sum_{i=1}^{m}\|\hat{Y}_T^i(h) - \sum_{e_{i,k}\in E}\hat{Y}_T^k(h)\|_1. \tag{18}$$

[1] https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/6202.0Dec%202019?OpenDocument
[2] https://github.com/TheClimateCorporation/properscoring

Table 2: MAPE for small and large simulation dataset. The likelihood ratios are given in parentheses.

MAPE	Simulation Small			Simulation Large			
	Top level	Level 1	Level 2	Top level	Level 1	Level 2	Level 3
Base	1.29 (.69)	1.50 (.77)	2.41 (.91)	2.08 (.43)	2.20 (.61)	1.41 (.75)	0.72 (.85)
BU	2.14 (.73)	1.76 (.79)	2.41 (.91)	4.19 (.46)	3.48 (.64)	1.48 (.76)	0.72 (.85)
MinT-sam	0.54 (.66)	1.48 (.77)	2.24 (.89)	1.48 (.42)	2.55 (.65)	1.38 (.74)	0.63 (.83)
MinT-shr	0.45 (.65)	1.47 (.77)	**2.23** (.89)	**1.28** (.39)	2.31 (.63)	**1.35** (.74)	**0.59** (.81)
MinT-ols	**0.20** (.64)	1.72 (.78)	2.41 (.91)	1.69 (.41)	2.15 (.60)	1.41 (.75)	0.71 (.85)
ERM	1.23 (.69)	1.73 (.78)	2.55 (.93)	2.78 (.44)	2.86 (.69)	1.50 (.76)	0.75 (.86)
SHARQ	1.54 (.41)	**1.42** (.45)	2.41 (.73)	2.16 (.23)	**2.13** (.49)	1.44 (.67)	0.72 (.82)

Table 3: MAPE results on FTSE dataset, lower values are better. Level 1 is the top aggregation level, and 4 is the bottom level.

Algorithm	RNN				Autoregressive				LST-Skip				N-Beats			
Reconciliation	Level				Level				Level				Level			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
BU	6.11	8.48	9.41	9.54	10.01	12.15	11.77	12.43	7.48	8.96	9.29	9.49	6.63	8.04	8.23	8.41
Base	4.82	6.27	8.55	9.54	**8.65**	10.46	10.88	12.43	6.02	7.79	8.76	9.49	5.86	**7.56**	**8.01**	8.41
MinT-sam	4.68	8.53	8.77	10.13	9.72	11.25	11.57	12.26	6.47	8.24	8.93	10.62	5.94	7.89	8.35	8.86
MinT-shr	**4.43**	8.46	8.59	9.75	9.23	10.91	11.02	**12.13**	6.12	8.11	8.81	10.57	5.67	7.74	8.22	**8.54**
MinT-ols	4.71	8.92	8.74	10.31	9.96	11.01	11.25	12.32	6.31	8.56	8.74	10.88	5.87	8.12	8.41	9.84
ERM	5.74	9.52	9.54	12.41	9.92	10.61	12.03	13.23	8.12	9.38	9.76	13.01	6.19	8.89	9.26	10.22
SHARQ	4.51	**8.28**	**8.08**	**9.54**	9.13	**9.35**	**10.61**	12.43	**5.01**	**7.14**	**8.52**	**9.49**	**5.44**	7.83	7.93	8.41

E.2 Simulation under Unbiased Assumption

We follow the data simulation mechanism developed in Wickramasuriya et al. (2019); Ben Taieb and Koo (2019), which satisfies the ideal unbiased base forecast and Gaussian error assumptions. The bottom level series were first generated through an ARIMA(p, d, q) process, where the coefficients are uniformly sampled from a predefined parameter space. The contemporaneous error covariance matrix is designed to introduce a positive error correlation among sibling series, while moderately positive correlation among others. We simulate a small and a large hierarchy with 4 and 160 bottom series, respectively. The bottom series are then aggregated to obtain the whole hierarchical time series in groups of two and four. For each series in the hierarchy, we generate 500 observations, and the final $h = 8, 16$ observations are used for evaluation for both the large and small hierarchies. We run the above simulation 100 times and report the average results. Table 2 shows the average MAPE by fitting an ARIMA model followed by reconciliation on two simulation datasets. We can see that the MinT methods generally perform the best, particularly for MinT methods with shrinkage estimators. This confirms the statements from Ben Taieb and Koo (2019); Hyndman et al. (2011) that under ideal unbiasedness assumption if the forecasting models are well specified, the MinT methods will provide the optimal solution. Simultaneously, the results of SHARQ are also satisfactory. In fact, it outperforms MinT methods at some levels.

E.3 Additional Results

Table 3 and 4 show the MAPE results of FTSE and Wiki dataset. Moreover, table 5 is the average likelihood ratio of each reconciliation method across four algorithms. The reported results are average across three random runs. We can see that SHARQ performs better overall in providing accurate probabilistic forecasts. Table 6 compares the average training and inference time across all forecasting models. Overall, the training time of SHARQ and base forecast are roughly the same, but the inference time of SHARQ is ignorable relative to MinT, and ERM approaches. Since both these methods require matrix inversion to compute the weight matrix. Even if ERM could calculate the weight matrix on a separate validation set before inference, additional matrix computations are required to obtain the results.

Table 4: MAPE results on Wiki dataset, lower values are better. Level 1 is the top aggregation level, and 5 is the bottom level.

Algorithm	RNN					Autoregressive					LST-Skip					N-Beats				
Reconciliation	Level					Level					Level					Level				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
BU	11.71	12.36	14.47	16.45	16.74	15.67	15.99	16.67	18.99	20.32	11.44	11.88	13.31	14.76	15.77	11.92	12.57	14.45	15.22	16.21
Base	11.12	11.52	14.06	16.11	16.74	**15.04**	15.23	16.02	17.83	20.32	11.21	11.24	12.88	14.35	15.77	11.84	12.02	14.17	15.16	16.21
MinT-sam	11.65	12.02	14.19	16.23	17.66	15.22	15.65	16.33	18.12	19.87	11.38	11.46	13.13	14.57	16.22	11.96	12.26	14.29	15.25	16.45
MinT-shr	11.32	11.86	13.87	**16.07**	17.54	15.17	15.12	15.98	17.69	**19.54**	11.24	11.15	12.91	**14.32**	16.14	11.75	12.19	14.03	15.02	16.39
MinT-ols	11.48	12.11	14.52	16.34	17.59	15.37	15.74	16.23	18.01	20.21	11.42	11.52	13.05	14.78	16.59	11.88	12.39	14.21	15.16	16.45
ERM	12.08	13.62	15.96	18.11	18.97	15.29	15.85	16.12	**17.58**	21.56	12.08	12.85	14.56	15.96	17.42	12.14	12.83	15.49	16.17	17.41
SHARQ	**10.84**	**11.07**	**13.54**	16.08	**16.74**	15.07	**15.05**	**15.87**	17.79	20.32	**11.07**	**11.09**	**12.65**	14.41	**15.77**	**11.64**	**11.67**	**13.81**	**15.02**	**16.21**

Table 5: Average likelihood ratio across forecasting horizons and models.

Likelihood Ratio	Labour	M5	FTSE	Wiki
BU	0.36	0.48	0.50	0.66
Base	0.36	0.48	0.51	0.66
MinT-sam	0.36	0.47	0.50	0.66
MinT-shr	0.35	0.49	0.51	0.68
MinT-ols	0.34	0.48	0.51	0.66
ERM	0.35	0.48	0.51	0.67
SHARQ	**0.07**	**0.25**	**0.32**	**0.65**

Table 6: Training and inference time (in second) comparison for each data set.

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

Table 7: Average forecasting coherency on each dataset across 4 forecasting models. Bottom-level $\lambda = 3.0$, higher-level λs are decreased gradually.

Reconciliation	Dataset			
	FTSE	Labour	M5	Wiki
Base	28.01	5.59	12.56	20.71
BU	0	0	0	0
MINTsam	4.21E-15	4.60E-12	0	5.46E-10
MINTshr	2.50E-15	4.19E-12	0	6.40E-11
MINTols	6.22E-15	6.10E-12	0	1.08E-10
ERM	6.48E-12	2.27E-08	5.86E-12	2.40E-07
SHARQ	1.59	0.53	0.22	2.63

E.4 Forecasting Coherency

Table 7 compares the forecasting coherency of each reconciliation method. We use the metric defined in (18) to compute the forecasting reconciliation error generated by previous experiments. As expected, the MinT and ERM approach give almost perfect coherent forecasts, as these methods can directly compute the close form of weight matrix P to optimally combine the original forecasts. Even though MinT and ERM can give perfectly coherent

forecasts, the accuracy can sometimes be worse than the base method, which coincides with Proposition 2 (Hard Constraint). Although SHARQ could not give the most coherent results, there is still a significant improvement compared to incoherent base forecasts. Note that this can be further improved by increasing the penalty of the regularization term.

F Hyper-parameter Configurations

Table 8: Common Hyper-parameters for all experiments.

	Train/Valid/Test	Epoch	Learning Rate	Batch Size	Window Size	Horizon
Quantile Simulation	0.6/0.2/0.2	300	1.00E-03	64	128	1
Unbiased Simulation	0.6/0.2/0.2	100	1.00E-03	128	10	1-8
Real-world Data	0.6/0.2/0.2	1000	0.1	128	168	1-8

We present hyper-parameters for all the experiments mentioned above. Table 8 lists the common hyper-parameters used on each experiment. Model-specific hyper-parameters are as follows.

Quantile Simulation Experiment We simulate 500 samples for both step function and sinusoidal function; the data is trained on a vanilla RNN model with hidden dimension 5, layer dimension 2, and *tanh* nonlinearity. We used 10 ensembles of estimators for bagging, and each model is trained using random 64 samples.

LSTNet The number of CNN hidden units: 100; the number of RNN hidden units: 100; kernel size of the CNN layers: 6; window size of the highway component: 24; gradient clipping: 10; dropout: 0.2; skip connection: 24. Note that to enable LSTNet to produce multi-quantile forecast, we add the final layer of each quantile estimator after the fully connected layer of the original model. The same linear bypass then adds the obtained quantile estimators to produce the final results.

N-Beats We use the same parameter setting as shown in the GitHub repository[1]. PyTorch is used to train the model.

References

Souhaib Ben Taieb and Bonsoo Koo. Regularized regression for hierarchical forecasting without unbiasedness conditions. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1337–1347, 2019.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Kevin AJ Doherty, Rod G Adams, Neil Davey, and Wanida Pensuwon. Hierarchical topological clustering learns stock market sectors. In *2005 ICSC Congress on Computational Intelligence Methods and Applications*, pages 6–pp. IEEE, 2005.

Rob J Hyndman, Roman A Ahmed, George Athanasopoulos, and Han Lin Shang. Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, 55(9):2579–2589, 2011.

Rob J Hyndman, Alan J Lee, and Earo Wang. Fast computation of reconciled forecasts for hierarchical and grouped time series. *Computational statistics & data analysis*, 97:16–32, 2016.

FTSE Russell. Ftse uk index series. *Retrieved February*, 5:2017, 2017.

Shanika L Wickramasuriya, George Athanasopoulos, Rob J Hyndman, et al. Forecasting hierarchical and grouped time series through trace minimization. *Department of Econometrics and Business Statistics, Monash University*, 2015.

Shanika L Wickramasuriya, George Athanasopoulos, and Rob J Hyndman. Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association*, 114(526):804–819, 2019.

[1] https://github.com/philipperemy/n-beats