Highly-scalable distributed modelling and forecasting with dependent data

Feng Li
@f3ngli

School of Statistics and Mathematics
Central University of Finance and Economics

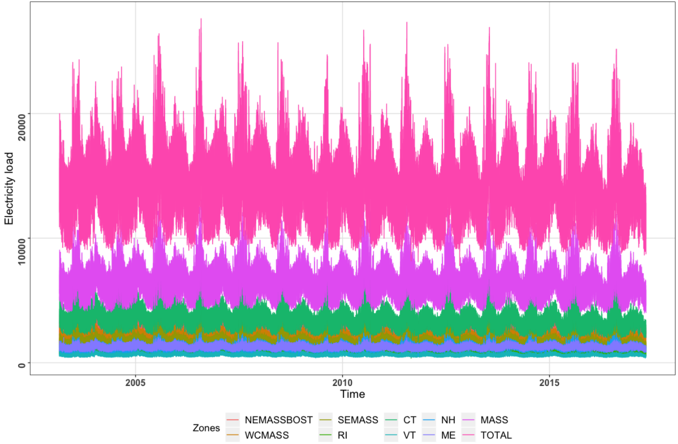
Outline

Distributed forecasting with ultra-long time series

↪ Motivation

- Ultra-long time series are increasingly accumulated in many cases,
  - hourly electricity demands
  - daily maximum temperatures
  - streaming data generated in real-time
- Forecasting ultra-long time series is challenging.
  - Time series are distributed stored with distributed file systems.
  - Training Bayesian models is time-consuming, and model updating is a demand for streaming time series.
  - Ultra-long time series require ultra-long forecasts.
  - Assuming the DGP remains invariant (e.g. stationary assumptions), over ultra-long time interval is unrealistic and requires efficient treatments.
- Few, *if not no*, Bayesian time models have even been applied to the industrial standard distributed computing systems.
  - Forecasters could not take advantage of probabilistic forecasting,
  - making it difficult for inventory planning in business.

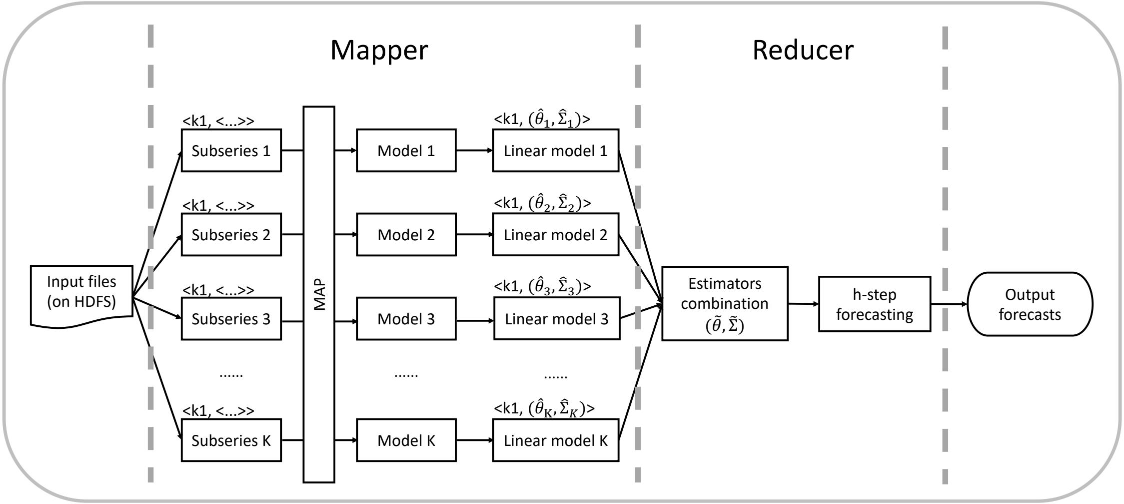
Distributed forecasting with ultra-long time series
↪ Electricity load data (Hong et al. 2019)



- The electricity load data set consists of 10 time series of hourly data, ranging from 1 March 2003 to 30 April 2017, spanning $124,171$ time points.
- The forecasting horizon is at least one month ahead to allow for earlier management plans.

Distributed forecasting with ultra-long time series

↬ The forecasting framework on a distributed system



Least-square approximation for a distributed system

- Let $\mathcal{L}(\theta; Z)$ be a plausible twice-differentiable loss function. Define the global loss function as $\mathcal{L}(\theta) = N^{-1} \sum_{i=1}^{N} \mathcal{L}(\theta; Z_i)$, whose global minimizer is $\widetilde{\theta} = \arg\min \mathcal{L}(\theta)$ and the true value is θ_0.
- Decomposing and approximating the global loss function using Taylor's expansion techniques as follows:

$$\mathcal{L}(\theta) \approx N^{-1} \sum_{k=1}^{K} \sum_{i \in \mathcal{S}_k} (\theta - \widetilde{\theta}_k)^\top \ddot{\mathcal{L}}(\widetilde{\theta}_k; Z_i)(\theta - \widetilde{\theta}_k) + C$$

- The quadratic form should be a good local approximation of the global loss function (Wang & Leng 2007, JASA). This leads to the following least squares objective function for a distributed model (DLSA) (Zhu, Li & Wang 2021, JCGS),

$$\widetilde{\mathcal{L}}(\theta) \stackrel{\text{def}}{=} \sum_{k} (\theta - \widetilde{\theta}_k)^\top \alpha_k \widetilde{\Sigma}_k^{-1} (\theta - \widetilde{\theta}_k)$$

The distributed Bayesian VARs I

- We consider a general VAR model written as

$$\boldsymbol{Y} = \boldsymbol{Y}_{-1}\boldsymbol{A}_1 + \boldsymbol{Y}_{-2}\boldsymbol{A}_2 + \ldots + \boldsymbol{Y}_{-p}\boldsymbol{A}_p + \boldsymbol{Z}\boldsymbol{C} + \boldsymbol{U} = \boldsymbol{X}\boldsymbol{\Gamma} + \boldsymbol{U}$$

  where $\boldsymbol{Y}$ is a $T \times n$ matrix, $\boldsymbol{X} = (\boldsymbol{Y}_{-1}, \boldsymbol{Y}_{-2}, \ldots, \boldsymbol{Y}_{-p}, \boldsymbol{Z})$is a $T \times m$ matrix to represent the lagged terms of $\boldsymbol{Y}$ and other explanatory variables, $\boldsymbol{\Gamma} = \left(\boldsymbol{A}_1', \boldsymbol{A}_2', \ldots, \boldsymbol{A}_p', \boldsymbol{C}'\right)'$ is the $m \times n$ coefficients matrix, and $\boldsymbol{U} = (\boldsymbol{u_1}, \boldsymbol{u_2}, \ldots, \boldsymbol{u_T})'$ while $\boldsymbol{u_t} \sim N(\boldsymbol{0}, \boldsymbol{\Psi})$ is the normally distributed errors.
- To keep things simple, we take an uniform prior distribution for $\boldsymbol{\Gamma}$ and a Jeffreys' prior for $\boldsymbol{\Psi}$,

$$p(\boldsymbol{\Gamma}, \boldsymbol{\Psi}) \propto |\boldsymbol{\Psi}|^{-(n+1)/2}. \tag{1}$$

- Then, we know that $\boldsymbol{\Gamma}$ follows a multivariate normal distribution conditional on $\boldsymbol{\Psi}$, and $\boldsymbol{\Psi}$ follows an inverse Wishart distribution.

The distributed Bayesian VARs II

- The marginal posterior for $\mathbf{\Gamma}$ is a matricvariate t-distribution

$$\mathbf{\Gamma}|\boldsymbol{Y}_T \sim MVT_{mn}(\widetilde{\gamma}, \boldsymbol{X}'\boldsymbol{X}, \boldsymbol{S}, T-m)$$

  for $\gamma = \operatorname{vec}\mathbf{\Gamma}$ and $\widetilde{\gamma} = \operatorname{vec}\widetilde{\mathbf{\Gamma}} = \operatorname{vec}\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}\right)$.

- Thus,

$$E(\mathbf{\Gamma}) = \widetilde{\gamma}, \qquad V(\operatorname{vec}\mathbf{\Gamma}) = \frac{1}{T-m-n-1}\boldsymbol{S}\otimes(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

- In this way, we could directly work out the posterior expectation and covariance matrix of coefficients for subdatas, and then the master could calculate global estimator $\widetilde{\mathbf{\Gamma}}$ using the DLSA (Zhu, Li & Wang 2021, JCGS).

$$\operatorname{vec}\widetilde{\mathbf{\Gamma}} = \left[\sum_k \alpha_k V_k^{-1}(\operatorname{vec}\mathbf{\Gamma})\right]^{-1}\left[\sum_k \alpha_k V_k^{-1}(\operatorname{vec}\mathbf{\Gamma})\,\widetilde{\gamma}_k\right]$$

- Model updating for streaming data is simple with Bayesian scheme
  — *Yesterday's posterior is today's prior*.
- With the DLSA method, if $\widetilde{V}_K^{-1}$ represents $\sum_k \alpha_K \widetilde{\Sigma}_K^{-1}$ and the initial value is set as $\widetilde{V}_1^{-1} = \alpha_1 \widetilde{\Sigma}_1^{-1}$ and $\widetilde{\theta}_1 = \widetilde{\theta}_1$, then we can get

$$\begin{aligned}
\widetilde{V}_{K+1}^{-1} &= \widetilde{V}_K^{-1} + \alpha_{k+1}\widetilde{\Sigma}_{k+1}^{-1} \\
\widetilde{\theta}_{K+1} &= \widetilde{V}_{K+1}\left(\widetilde{V}_K^{-1}\widetilde{\theta}_K + \alpha_{k+1}\widetilde{\Sigma}_{k+1}^{-1}\widetilde{\theta}_{k+1}\right)
\end{aligned}$$

- Therefore, we can constantly update the expected estimate $\widetilde{\theta}_{K+1}$ and $\widetilde{V}_{K+1}^{-1}$ based on the initial value and the parameters $\alpha_{k+1}\widetilde{\Sigma}_{k+1}^{-1}$ and $\widetilde{\theta}_{k+1}$ calculated based on the new data.

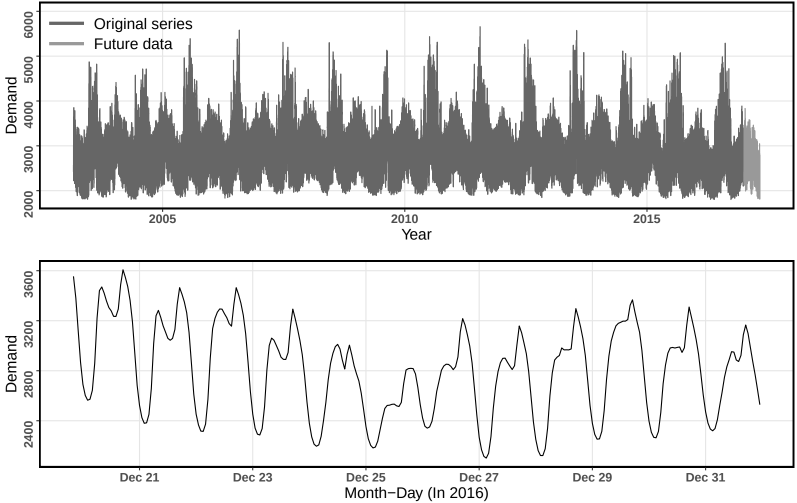
Application to the electricity demand data

Variable name	Description
Date	date in MM/DD/YYYY format
Hour	hour of the observation, in hour ending and 24-hour convention
DA_DEMD	day-ahead demand consists of fixed and price sensitive demand bids plus decrement bids & increment offers
DEMAND	non-PTF demand = [non-dispatchable + unmetered + station service] as determined by metering
DA_LMP	day-ahead locational marginal price
DA_EC	day-ahead energy component
RT_LMP	real-time locational marginal price
RT_EC	real-time energy component
DryBulb	dry bulb temperature in °F
DewPnt	dew point temperature in °F
SYSLoad	System load = [generation - pumping load + net interchange] as determined by metering
RegCP	Regulation Clearing Price

* Demand & load are in MWh. All prices are in $/MWh.

Distributed forecasting with ultra-long time series

↪ Electricity load data

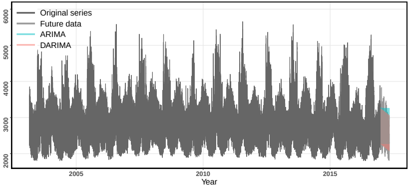


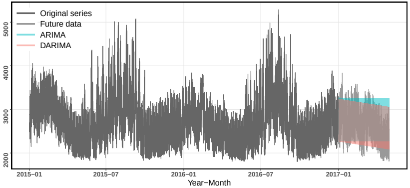
Distributed forecasting with ultra-long time series

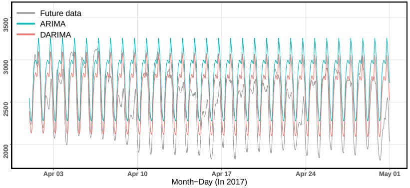
↪ Need for speed!

Max orders	Method	MASE	MSIS	Execution time (mins)
(5, 2, 5)	ARIMA	1.430	19.733	4.596
	DARIMA	**1.297**	**15.078**	**1.219**
(5, 2, 7)	ARIMA	1.410	18.695	14.189
	DARIMA	**1.297**	**15.078**	**1.211**
(6, 2, 7)	ARIMA	1.410	18.695	15.081
	DARIMA	**1.298**	**15.108**	**1.326**
(6, 3, 7)	ARIMA	1.413	15.444	21.072
	DARIMA	**1.324**	**12.590**	**1.709**
(6, 3, 10)	ARIMA	1.413	15.654	76.272
	DARIMA	**1.324**	**12.590**	**1.769**
(7, 3, 10)	ARIMA	1.413	15.654	83.077
	DARIMA	**1.327**	**12.561**	**1.829**
(7, 4, 10)	ARIMA	1.409	13.667	111.292
	DARIMA	**1.338**	**12.079**	**2.267**
(8, 4, 10)	ARIMA	1.409	13.667	117.875
	DARIMA	**1.335**	**12.076**	**2.224**

Distributed forecasting with ultra-long time series







Applications

↛ Forecasting Results of GEFCom2017 Data with VARs

	Distributed	FALSE	TRUE		
	Nodes	1	5	10	50
h=24	RMSE	0.129	0.124	**0.123**	0.125
	MASE	0.419	0.406	**0.400**	0.413
	LPS	32.223	32.910	**33.152**	32.372
h=48	RMSE	0.119	0.116	**0.116**	0.118
	MASE	0.371	0.366	**0.363**	0.374
	LPS	61.367	62.127	**62.284**	60.912
h=168	RMSE	0.127	0.125	**0.125**	0.125
	MASE	0.402	0.398	**0.393**	0.400
	LPS	187.877	189.688	**190.311**	189.660
h=720	RMSE	0.144	0.144	0.144	**0.143**
	MASE	0.451	0.453	0.449	**0.451**
	LPS	580.820	582.794	583.121	**587.239**

Discussions

- Distributed forecasting not only speeds up the computation but also improves forecasting performance (Wang et al. 2020, arXiv).
- Distributed systems like Apache Spark are the *de facto* standard in the data science industry, but it is costly to run time-consuming programs.
- Scalable Bayesian forecasting models empower rapid business planning.
- Theoretical details are available in
  - DLSA — Zhu, Li & Wang (2021, JCGS),
  - DQR — Pan, Ren, Guo, Li, Guodong & Wang (2021, JBES),
  - DVAR — Ma, Li, Karlsson & Kang (2021, soon on arXiv),
  - DARIMA — Wang, Kang, Hyndman & Li (2020, arXiv).
- Try our software `https://github.com/feng-li/dstats/` implemented for the Spark distributed computing platform.

Thank you!

https://kllab.org
feng.li@cufe.edu.cn

References

Hong, T., Xie, J. & Black, J. (2019), 'Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting', *International Journal of Forecasting* **35**(4), 1389–1399.

Ma, X., Li, F., Karlsson, S. & Kang, Y. (2021), 'Distributed forecasting with large Bayesian VAR models', *Working Paper* .

Pan, R., Ren, T., Guo, B., Li, F., Guodong & Wang, H. (2021), 'A note on distributed quantile regression by pilot sampling and one-step updating', *Journal of Business and Economic Statistics* **Accepted**.

Wang, H. & Leng, C. (2007), 'Unified lasso estimation by least squares approximation', *Journal of the American Statistical Association* **102**(479), 1039–1048.

Wang, X., Kang, Y., Hyndman, R. J. & Li, F. (2020), 'Distributed ARIMA models for ultra-long time series', *arXiv preprint arXiv:2007.09577* .

Zhu, X., Li, F. & Wang, H. (2021), 'Least squares approximation for a distributed system', *Journal of Computational and Graphical Statistics* **Accepted**.