Likelihood-based inference in temporal hierarchies

Jan Kloppenborg Møller, Hjorleifur G. Bergsteinson, Peter Nystrup,
and Henrik Madsen

DTU-Compute - Technical University of Denmark
jkmo@dtu.dk

IFF Workshop on Forecast Reconciliation
September 8. 2023

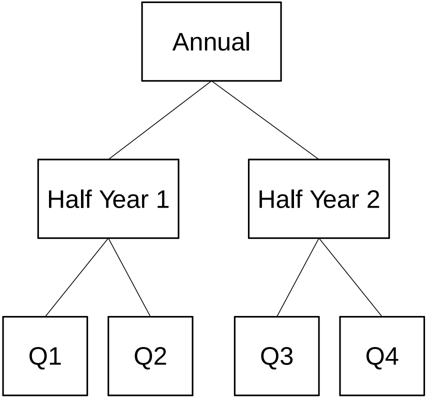
Technical University of Denmark DTU

Outline

Outline

Motivation

- Models on different aggregation levels
- Models on each level may not agree
- Reconciliation ensure consistent forecasts
- Reconciliation often improve forecast accuracy on all levels



Reconciliation

The reconciled forecast is calculated by

$$\tilde{\boldsymbol{y}} = (\boldsymbol{S}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{S})^{-1} \boldsymbol{S}^T \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{y}}$$

where:

- $\tilde{\boldsymbol{y}}$: reconciled forecast
- $\boldsymbol{S}$: the summation matrix
- $\boldsymbol{\Sigma}$: a variance-covariance matrix
- $\hat{\boldsymbol{y}}$: the base forecast

Regression setting

$$\hat{\boldsymbol{y}} = \boldsymbol{S}\tilde{\boldsymbol{y}} + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$$

$$\boldsymbol{S} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Choosing the correct variance

Different options for $\boldsymbol{\Sigma}$:

- $\boldsymbol{\Sigma} = \boldsymbol{I}$
- Variance scaling (proportional to volume of level)
- Use observed variance-covariance of base forecast error
- Ignore cross level correlation
- Use shrinkage on the observed variance-covariance (usually preferred). I.e.

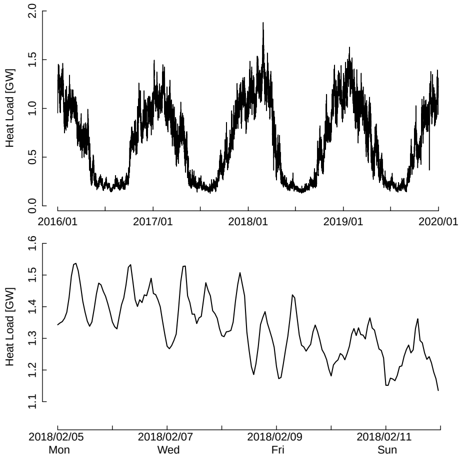
$$\hat{\boldsymbol{\Sigma}}_s = \lambda \hat{\boldsymbol{\Sigma}} + (1-\lambda)\mathrm{diag}\hat{\boldsymbol{\Sigma}}$$

Outline

Technical University of Denmark

Data

- District heating from an area of greater Copenhagen
- Clear annual and diurnal variation
- State of the art commercial hourly forecast (1-24 hours ahead)
- 2016 used for initialization

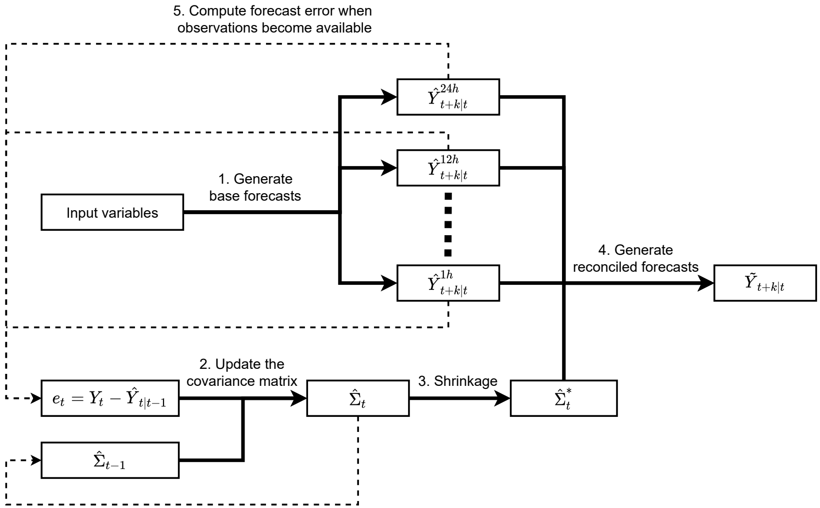


From Bergsteinsson et al. (2021).

A case study

- 1 hour level used as evaluation
- All other levels modeled in the study
- 2, 3, 4, 6, 8, 12, and 24 hours forecast:
  - Recursive Least Square
  - Forecast of ambient temperature
  - Diurnal variation
  - Auto-regressive parts
- $\boldsymbol{\Sigma}_t$ (60 by 60) estimated using the full variance covariance with shrinkage, and recursive updating.

Work flow of modeling



From Bergsteinsson et al. (2021).

Some results

	2017–2019			
	Base RMSE	Expanding Window	Rolling Window	Exponential Smoothing
Daily	0.5960	-23.75	-22.49	**-23.93**
Twelve-hourly	0.3516	-24.08	-22.83	**-24.2**
Eight-hourly	0.3538	-43.51	-42.72	**-43.69**
Six-hourly	0.2876	-44.64	-43.75	**-44.76**
Four-hourly	0.1765	-36.06	-35.19	**-36.37**
Three-hourly	0.1334	-33.03	-32.05	**-33.26**
Two-hourly	0.0884	-30.09	-29.07	**-30.36**
Hourly	0.0383	-14.75	-13.46	**-15.07**

Adapted from Bergsteinsson et al. (2021).

Outline

Motivation / Aim

The starting point

$$\tilde{\boldsymbol{y}} = (\boldsymbol{S}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{S})^{-1} \boldsymbol{S}^T \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{y}}$$

Comments and aim

- Observation appear only through the variance-covariance matrix $\boldsymbol{\Sigma}$
- The estimation of $\boldsymbol{\Sigma}$ include a large number of parameters
- Formulate a parameterized model for obtaining $\boldsymbol{\Sigma}$
- Reduce dimension of the the parameter space by well-known likelihood techniques

Example

Example: Assume that half year levels generated by

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \epsilon_t,$$

AR(1) models half-year levels and annual levels, i.e. the models

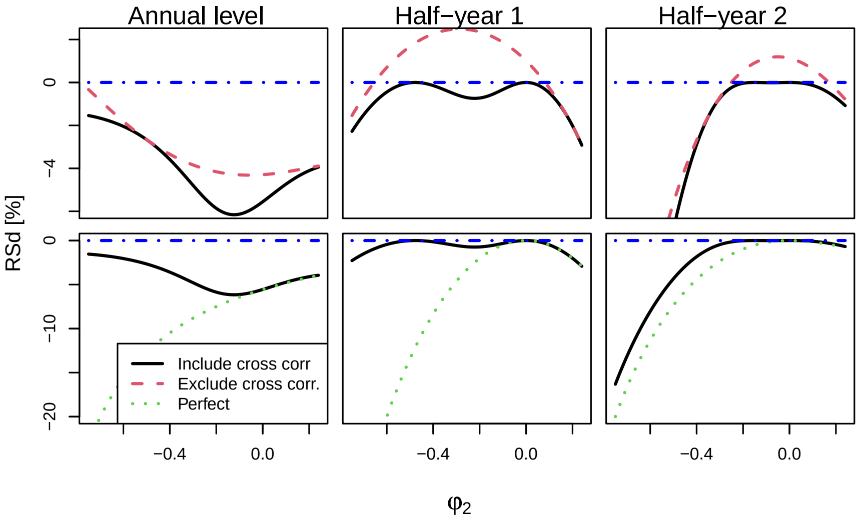
$$\begin{aligned} y_t^{\mathsf{A}} &= \phi_1^{\mathsf{A}} y_{t-1}^{\mathsf{A}} + \epsilon_t^{\mathsf{A}} \\ y_t^{\mathsf{H}} &= \phi_1^{\mathsf{H}} y_{t-1}^{\mathsf{H}} + \epsilon_t^{\mathsf{H}}. \end{aligned}$$

Full setup

$$\boldsymbol{y}_{2t+2|2t} = \begin{bmatrix} y^{\mathsf{A}}_{2t+2|2t} \\ y^{\mathsf{H}}_{2t+1|2t} \\ y^{\mathsf{H}}_{2t+2|2t} \end{bmatrix}; \quad \hat{\boldsymbol{y}}_{2t+2|2t} = \begin{bmatrix} \hat{y}^{\mathsf{A}}_{2t+2|2t} \\ \hat{y}^{\mathsf{H}}_{2t+1|2t} \\ \hat{y}^{\mathsf{H}}_{2t+2|2t} \end{bmatrix}; \quad \boldsymbol{S} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

and e.g. $\boldsymbol{\Sigma} = \mathrm{Var}[\boldsymbol{y}_{2t+2} - \hat{\boldsymbol{y}}_{2t+2}]$ can be calculated explicitly.

Choosing the correct variance



$\phi_1 = 0.75$

(from Møller et al. (2023))

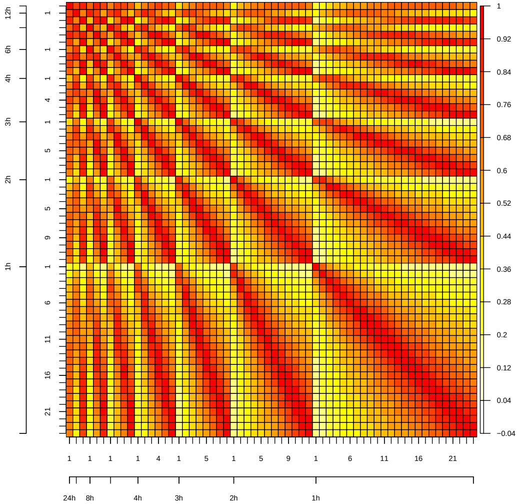
Modeling of variance-covariance matrix

$\boldsymbol{\Sigma}$ (some challenges):

- High-dimensional
- High correlations

Some suggestions

- Parametric models for the correlation
- Use statistical methods for reduction



Graphics: Møller et al. (2023)

A parametric model

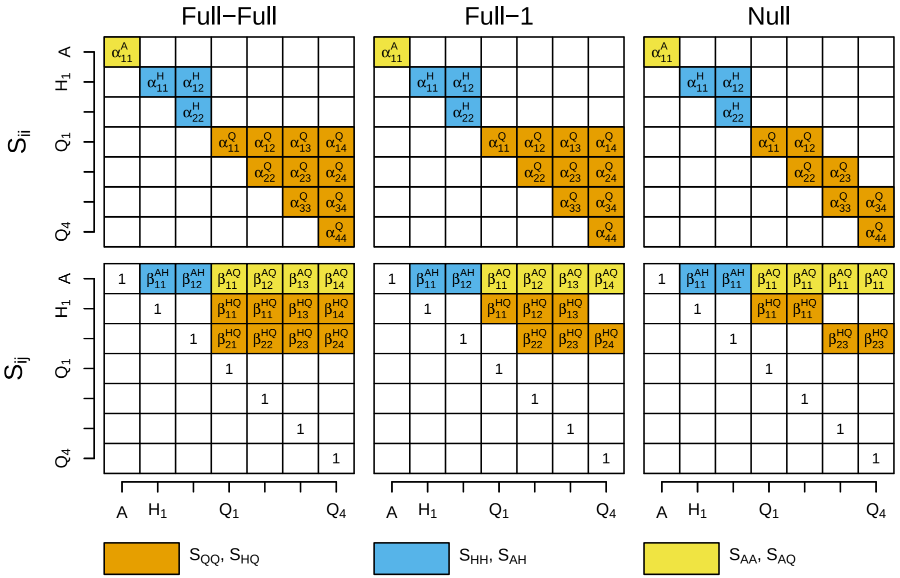
Starting from the bottom level define (Møller et al., 2023)

$$
\begin{aligned}
\boldsymbol{\epsilon}_1 &= \boldsymbol{u}_1 && ; \quad \boldsymbol{u}_1 \sim N(\mathbf{0}, \boldsymbol{\Sigma}_1),\\
\boldsymbol{\epsilon}_2 &= \boldsymbol{S}_{21}\boldsymbol{u}_1 + \boldsymbol{u}_2 && ; \quad \boldsymbol{u}_2 \sim N(\mathbf{0}, \boldsymbol{\Sigma}_2),\\
\boldsymbol{\epsilon}_3 &= \boldsymbol{S}_{31}\boldsymbol{u}_1 + \boldsymbol{S}_{32}\boldsymbol{u}_2 + \boldsymbol{u}_3 && ; \quad \boldsymbol{u}_3 \sim N(\mathbf{0}, \boldsymbol{\Sigma}_3),\\
&\vdots\\
\boldsymbol{\epsilon}_K &= \sum_{j=1}^{K-1} \boldsymbol{S}_{Kj}\boldsymbol{u}_j + \boldsymbol{u}_K && ; \quad \boldsymbol{u}_K \sim N(\mathbf{0}, \boldsymbol{\Sigma}_K),
\end{aligned}
$$

where $Cov[\boldsymbol{u}_i, \boldsymbol{u}_j] = \mathbf{0}$, $(i \neq j)$. And

$$\boldsymbol{\Sigma}_i^{-1} = \boldsymbol{S}_{ii}\boldsymbol{S}_{ii}^T,$$

with $\boldsymbol{S}_{ii}$ is an upper triangular matrix.



Full−Full
Full−1
Null
S_{ii}
S_{ij}
A
H_1
Q_1
Q_4
S_{QQ}, S_{HQ}
S_{HH}, S_{AH}
S_{AA}, S_{AQ}

Estimation

Likelihood

$$l(\boldsymbol{\beta}, \boldsymbol{\alpha}; \boldsymbol{V}) \propto -\frac{T-1}{2} Tr \boldsymbol{\Sigma}^{-1} \boldsymbol{V} + \frac{T-1}{2} \log |\boldsymbol{\Sigma}^{-1}|,$$

where $\boldsymbol{V}$ is the observed variance-covariance matrix, $\boldsymbol{\Sigma}^{-1}$ is parameterized through the model and estimation is done

- Sequentially starting from the bottom level
- Using:
  - a robust EM-like algorithm (solving normal equations) and
  - the Newton method (using the Hessian of the likelihood)

diagonal elements of $\boldsymbol{S}_{ii}$ is estimated in the log-domain.

Shrinkage

A simple modification of the likelihood by introducing weights in the following way:

$$l_s(\boldsymbol{\Sigma}; \boldsymbol{V}, \boldsymbol{w}) = l(\boldsymbol{\Sigma}; w_1\boldsymbol{V} + w_2 \text{blockdiag} \boldsymbol{V} + w_3 \text{diag} \boldsymbol{V}),$$

where $\sum_i w_i = 1$.

Statistical tests

As the method is based on likelihood estimation we have access to

- Wald test for individual parameters or sets of parameters
- Likelihood ratio test for individual parameters or specific hypothesis

In the work we explore

- Wald test for testing if parameters should be zero or equal and confirm using LRT
- Effect of different initial structures

Case study

- Electricity load in Sweden 2016-2020
- 2016-2019 used for estimating mean value structure
- Linear model including annual and diurnal variation
- Double seasonal AR models of the residuals (daily and and weekly)
- Temporal reconciliation of residuals

Some results

	SE		SE1		SE2		SE3		SE4	
	df	RRMSE	df	RRMSE	df	RRMSE	df	RRMSE	df	RRMSE
Obs-test	1830	-12.4	1830	-8.5	1830	-9.8	1830	-15.9	1830	-15.3
Obs-train	1830	0.1	1830	4.1	1830	2.1	1830	-2.3	1830	-5.2
Full-Full$_1$	402	-4.8	306	-1.7	413	**-4.1**	413	-5.6	464	-5.5
Full-Full$_2$	402	**-5.3**	306	0.8	413	-3.1	413	-5.8	464	-6.6
Full-2$_1$	219	-3.8	180	**-3.2**	237	-4.0	240	-3.9	264	-4.0
Full-2$_2$	219	-3.8	180	-2.2	237	-3.8	240	-4.9	264	-3.7
Null-Null$_1$	83	-4.3	77	-2.7	81	**-4.1**	85	-4.8	85	-4.6
Null-Null$_2$	83	-3.7	77	-0.7	81	-3.1	85	-4.0	85	-4.1
Shrink	0.015	**-5.3**	0.016	-1.3	0.035	-4.0	0.015	**-7.4**	0.016	**-7.0**
Auto-cov.	69	-2.2	49	-2.9	49	-3.5	71	-1.6	71	-2.2
Model-AR1	60	-2.4	60	-3.0	60	-3.2	60	-1.9	60	-2.2
Diag	60	-3.1	60	-2.4	60	-2.6	60	-2.9	60	-2.6

Outline

Conclusion

- Parameterized model for the variance-covarince matrix
- Statistical tests and huge reduction in number of parameters
- Similar performance as shrinkage
- Likelihood based inference including algorithms for estimation and testing
- Shrinkage needed

Open

- Error propagation from $\boldsymbol{\Sigma}$ to the weight matrix?
- Structured models for $\boldsymbol{S}_{ii}$ and $\boldsymbol{S}_{ij}$

References

[1] Bergsteinsson H.G., Møller J.K., Nystrup P., Pálsson Ó.P., Guericke D., and Madsen H. (2021). Heat load forecasting using adaptive temporal hierarchies. *Applied Energy*, **292**.

[2] Møller J.K., Nystrup P., and Madsen H. (2023). Likelihood-based inference in temporal hierarchies. *International Journal of Forecasting, in press*.

[3] Møller JK, Nystrup P, and Madsen H. (2022) Supplementary material for "Likelihood-based inference in temporal hierarchies. http://people.compute.dtu.dk/jkmo/

[4] Nystrup, P., Lindström, E., Møller, J. K., and Madsen, H. (2021). Dimensionality reduction in forecasting with temporal hierarchies. *International Journal of Forecasting, 37(3)*

[5] Nystrup, P., Lindström, E., Pinson, P., and Madsen, H. (2020). Temporal hierarchies with autocorrelation for load forecasting. *European Journal of Operational Research, 280, 876-888.*

Thank You!

Questions?