Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Likelihood-based inference in temporal hierarchies

Jan Kloppenborg Møller [*], Peter Nystrup, Henrik Madsen

Department of Applied Mathematics and Computer Science, Technical University of Denmark, Denmark

ARTICLE INFO

Keywords:
Maximum likelihood estimation
Dimensionality reduction
Forecast reconciliation
Hypothesis testing
Variance–covariance shrinkage
Statistical modelling
Load forecasting

ABSTRACT

We consider the importance of correctly specifying the variance–covariance matrix to allow information to be shared between aggregation levels when reconciling forecasts in a temporal hierarchy. We propose a novel framework for parametric modelling of the variance–covariance matrix, along with an iterative algorithm for maximum likelihood estimation. The covariance between aggregation levels can be modelled by aggregating the lower-level errors and disaggregating information from the higher levels. Using the likelihood approach, statistical inference can be applied to identify a parsimonious parametric structure for the variance–covariance matrix. We test and discuss different structures for how forecast errors are connected across aggregation levels and present a framework for simplifying these structures using Wald and likelihood-ratio tests. We evaluate the proposed method in a simulation study and through an application to day-ahead electricity load forecasting and find that it performs well compared to optimal shrinkage estimation.

1. Introduction

Real-life decision problems often involve forecasts for multiple levels of a hierarchy. During the last decade, the reconciliation of hierarchical forecasts has received a lot of attention from the forecasting community—from practitioners and academics alike. Reconciliation ensures unified predictions that support aligned decisions across all levels of a hierarchy, whether cross-sectional, temporal, or cross-temporal. Similar to forecast combination in general (Clemen, 1989; Timmermann, 2006), the process of reconciling forecasts from multiple aggregation levels is often beneficial, leading to improvements in forecast accuracy and/or a reduction in forecast error variance (Hollyman et al., 2021).

In addition to the practical motivation and beneficial properties, forecast reconciliation has become popular because it is easy to apply. It is accomplished by linearly combining all forecasts for a hierarchy to a set of adjusted bottom-level forecasts, which make use of all available information, and then aggregating these to reconciled forecasts for the entire hierarchy (Pritularga et al., 2021). This is intuitively better than simply aggregating forecasts from the bottom or disaggregating from the top level of a hierarchy, as has traditionally been done to ensure coherency (Athanasopoulos et al., 2009; Gross & Sohl, 1990). It is a model-independent and flexible approach that does not require all forecasts to come from a specific model or the same model. In fact, the forecasts do not even have to come from a model but could be entirely judgemental.

Temporal reconciliation has been successful in numerous energy applications where data are characterised by seasonal patterns, such as electricity load (Nystrup et al., 2020), heat load (Bergsteinsson et al., 2021), wind power (Jeon et al., 2019), and solar power forecasting (Yang et al., 2017). A lot of effort has been and is still being devoted to understanding when and why forecast reconciliation works and how it can be further improved (Di Fonzo & Girolimetto, 2022; Hollyman et al., 2021; Nystrup et al., 2021; Panagiotelis et al., 2021; Pritularga et al., 2021).

In what follows, we consider the importance of correctly specifying the variance–covariance matrix to allow

[*] Corresponding author.
E-mail address: jkmo@dtu.dk (J.K. Møller).

https://doi.org/10.1016/j.ijforecast.2022.12.005

for information to be shared between aggregation levels when reconciling forecasts in a temporal hierarchy. To illustrate this, we first consider an example where the actual variance–covariance matrix is known. In real-life applications it has to be estimated from the data available, which is difficult given its often high dimension and unknown structure. To this end, our contribution is to propose a novel framework for parametric modelling of the variance–covariance matrix, along with an iterative algorithm for maximum likelihood estimation. The covariance between aggregation levels is modelled by aggregating the lower-level errors and disaggregating information from the higher levels. Using the likelihood approach, we show how statistical inference can be applied to identify a parsimonious parametric structure for the variance–covariance matrix. We test and discuss different structures for how forecast errors are connected across aggregation levels and present a framework for simplifying these structures using Wald and likelihood-ratio tests. We evaluate the proposed method in a simulation study and through an application to day-ahead electricity load forecasting.

The outline of this article is as follows. We discuss related work in Section 2. In Section 3, we consider a simple example that illustrates the importance of correctly specifying the variance–covariance matrix. The likelihood approach is described in Section 4, including the model formulation and estimation setup. The framework for model reduction and shrinkage is introduced in Section 5. Results from the application to load forecasting are shown in Section 6. Section 7 presents a simulation study based on the load data, focusing on situations where the number of parameters in the variance–covariance matrix is high compared to the number of observations available for estimation. Finally, Section 8 concludes.

2. Related work

The reconciliation of hierarchical forecasts was first formulated as a linear regression problem by Hyndman et al. (2011), who used ordinary least squares (OLS) regression to reconcile forecasts for a cross-sectional hierarchy. Hyndman et al. (2016) proposed to use weighted least squares (WLS) regression instead, in order to take into account differences in the variances of the base forecast errors. Subsequently, Athanasopoulos et al. (2017) showed that the same approach can be used to produce coherent forecasts for temporal hierarchies, as temporally aggregated time series can be represented as hierarchical time series.

Wickramasuriya et al. (2019) found that the inclusion of covariance information through generalised least-squares (GLS) regression was beneficial for forecast accuracy in a cross-sectional hierarchy when combined with a linear shrinkage estimator. Similarly, Nystrup et al. (2020) showed that accuracy can be significantly improved across all aggregation levels in a temporal hierarchy by accounting for auto- and cross-covariances when reconciling forecasts.

The reconciliation weights are uncertain and must be estimated. This was illustrated by the recursive shrinkage approach employed by Bergsteinsson et al. (2021) to estimate time-varying weights capturing seasonality in the forecast errors. Uncertainties propagate from the variance–covariance matrix estimation to the reconciliation weights. Although the reconciled forecasts will be coherent as long as the reconciliation weights meet the coherency constraints, they might not be optimal. Pritularga et al. (2021) argued that the effect of uncertainties in forecast reconciliation has been overlooked. The uncertainty of reconciled forecasts comes from the incoherent base forecasts and propagates to the estimated reconciliation weights, which increases the uncertainty of the reconciled forecasts.

In general, the variance–covariance matrix for the coherency errors is unknown and unidentifiable. Wickramasuriya et al. (2019) provided theoretical justification for using the variance–covariance matrix for the base forecast errors as a proxy. A lot of work has gone into proposing and justifying different projections of base forecasts onto the coherent subspace, spanning from data-independent, structurally motivated approximations (Athanasopoulos et al., 2017) to approaches based on estimates of the full empirical variance–covariance matrix (Nystrup et al., 2020). Eckert et al. (2021) proposed a Bayesian approach to identify and shrink coherency errors that enables the inclusion of prior information.

As discussed by Pritularga et al. (2021), the accuracy improvement from forecast reconciliation depends on the quality of the estimated projection. At the same time, more complex approximations increase the variance of the reconciled forecasts. Nystrup et al. (2021) showed that due to the redundancy inherent in temporal hierarchies, the dimension of the estimation problem can be significantly reduced without lowering the accuracy of the reconciled forecasts. Unlike previous approaches to estimating or approximating the variance–covariance matrix for forecast reconciliation, we propose a statistical modelling and inference approach based on likelihood principles that can be applied to identify a parsimonious parametric structure.

3. Motivating example

A fundamental motivation for forecast reconciliation is that information from different aggregation levels is beneficial for all levels of a hierarchy. Even if we were only interested in a single level, it may be beneficial to create forecasts for other levels as well and reconcile them. One obvious reason for this is that in real-life applications the data-generating process is unknown, and models are prone to misspecification. The data-generating process is often a complex non-linear function with time-varying coefficients and combinations of (possibly higher-order and seasonal) auto-regressive and moving-average terms that is approximated by a lower-order linear model.

In order to illustrate the importance of correctly defining the full variance–covariance matrix, we consider a motivating example. Suppose that the data-generating, half-yearly process is a second-order auto-regressive, AR(2), process:

$$y_{t+1}^{\mathrm{H}} = \phi_1 y_t^{\mathrm{H}} + \phi_2 y_{t-1}^{\mathrm{H}} + \epsilon_{t+1}^{\mathrm{H}}; \quad \epsilon_{t+1}^{\mathrm{H}} \sim N(0, \sigma^2). \tag{1}$$

In order to mimic the misspecification that happens in practice, we adopt AR(1) models for the half-yearly level and the aggregate, annual level. The process is observed at the bottom, half-yearly level at $t = \{1, 2, \ldots\}$ and the aggregate, annual process is observed at times $2t = \{2, 4, \ldots\}$. For each observation of the annual process, forecasts are made two steps ahead at the half-yearly level and one step ahead at the annual level.

The data-generating process for the top, annual level is

$$y^{\mathrm{A}}_{2t+2} = y^{\mathrm{H}}_{2t+1} + y^{\mathrm{H}}_{2t+2}, \tag{2}$$

which can also be written as an ARMA(2,1) process (Amemiya & Wu, 1972). We use the formulation above to emphasise the hierarchical structure of the problem. Our models for the two levels are

$$y^{\mathrm{A}}_{2t+2} = \tilde{\phi}_{\mathrm{A}} y^{\mathrm{A}}_{2t} + \epsilon^{\mathrm{A}}_{2t+2}, \tag{3}$$

$$y^{\mathrm{H}}_{t+1} = \tilde{\phi}_{\mathrm{H}} y^{\mathrm{H}}_{t} + \epsilon^{\mathrm{H}}_{t+1}. \tag{4}$$

Although both models are misspecified (excluding the special case where ϕ_2 is zero, in which case only the annual and not the half-yearly model is misspecified), it seems reasonable that the annual level contains information about the half-yearly level and vice versa. The estimated coefficients in the models are the auto-correlation at lag one for each process.

We can write the process for the observations as

$$\boldsymbol{y}_{2t+2} = \begin{bmatrix} y^{\mathrm{A}}_{2t+2} \\ y^{\mathrm{H}}_{2t+1} \\ y^{\mathrm{H}}_{2t+2} \end{bmatrix} = \boldsymbol{\Phi} \begin{bmatrix} y^{\mathrm{H}}_{2t-1} \\ y^{\mathrm{H}}_{2t} \end{bmatrix} + \boldsymbol{\Phi}_{\epsilon} \begin{bmatrix} \epsilon^{\mathrm{H}}_{2t+1} \\ \epsilon^{\mathrm{H}}_{2t+2} \end{bmatrix}, \tag{5}$$

where the matrices $\boldsymbol{\Phi}$ and $\boldsymbol{\Phi}_{\epsilon}$ can be derived directly from the process definition. The details of the derivations and the exact forms of the matrices are shown in Appendix A.

The model forecast can be written in a similar way as follows:

$$\hat{\boldsymbol{y}}_{2t+2|2t} = \boldsymbol{\Gamma} \begin{bmatrix} y^{\mathrm{H}}_{2t-1} \\ y^{\mathrm{H}}_{2t} \end{bmatrix}, \tag{6}$$

where the elements of $\boldsymbol{\Gamma}$ are determined by the auto-correlations of the half-yearly and annual processes (see (A.5)). Here the random variable $\hat{\boldsymbol{y}}_{2t+2|2t} = E[\boldsymbol{y}_{2t+2}|\boldsymbol{y}_{2t}]$ is the conditional expectation of the forecast.

With the formulation above we can find the predictive variance–covariance matrices

$$\boldsymbol{\Sigma} = V[\boldsymbol{y}_{2t+2} - \hat{\boldsymbol{y}}_{2t+2|2t}], \tag{7}$$

$$\tilde{\boldsymbol{\Sigma}} = V[\boldsymbol{y}_{2t+2} - \boldsymbol{S}\boldsymbol{P}\hat{\boldsymbol{y}}_{2t+2|2t}], \tag{8}$$

where $\boldsymbol{S}$ is the summation matrix (see (12) for an example), $\boldsymbol{P}$ is given by

$$\boldsymbol{P} = (\boldsymbol{S}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{S})^{-1} \boldsymbol{S}^T \boldsymbol{\Sigma}^{-1}, \tag{9}$$

and $\boldsymbol{P}\hat{\boldsymbol{y}}_{2t+2|2t}$ is the reconciled forecast (Wickramasuriya et al., 2019).

Note that since we know the data-generating process, the variance–covariance matrices for the errors of the base and reconciled forecasts can be calculated from the process parameters, as shown in Appendix A. Hence, they do not have to be estimated in this example. The error introduced by estimating the parameters—i.e. by replacing $\boldsymbol{\Sigma}$ with $\hat{\boldsymbol{\Sigma}}$—is considered in the supplementary material for this article (Møller et al., 2022).

3.1. Accuracy measurement

To compare the accuracy in the simulation and case studies, we use the relative root mean square error:

$$RRMSE = \left(\frac{RMSE_{\mathrm{rec}} - RMSE_{\mathrm{base}}}{RMSE_{\mathrm{base}}} \right) \cdot 100\% \tag{10}$$

where subscripts rec and base refer to the reconciled and base forecasts, respectively. The RRMSE was recommended by Hyndman and Koehler (2006) and has been used frequently in studies on forecast reconciliation (Athanasopoulos et al., 2017; Nystrup et al., 2021, 2020). The RRMSE can be calculated for each aggregation level and each forecast horizon within each level.

The RMSE is basically an estimate of the standard deviation of the forecast error at a specific level (and horizon). Therefore, it is natural to use what we will refer to as the relative standard deviation as a measure of the improvement compared to the base forecast:

$$RSd_i = \left(\frac{\tilde{\sigma}_i - \sigma_i}{\sigma_i} \right) \cdot 100\%, \tag{11}$$

where σ_i and $\tilde{\sigma}_i$ are the standard deviations of the base and reconciled forecast errors, to compare the effect of reconciliation for different choices of ϕ_1 and ϕ_2.

Fig. 1 compares the improvements in terms of RSd_i as a function of ϕ_2 for two choices of $\boldsymbol{\Sigma}$ when $\phi_1 = 0.75$ and $\sigma^2 = 1$: the full (and correct) $\boldsymbol{\Sigma}$ and one constructed by ignoring cross-correlation. The figure also shows the improvement we could obtain if we knew the data-generating process.

For this specific choice of ϕ_1 (and for a range of ϕ_2), the variance of the forecast error at the annual level always improves. The largest improvements occur when $\boldsymbol{\Sigma}$ is correctly specified. At the half-yearly level, the variance actually increases when the correlation structure is not correctly specified. It is particularly interesting to note that when ϕ_2 is zero and the half-yearly model is equal to the data-generating process, the variance at the half-yearly level increases compared to the base forecast. For a range of ϕ_2 around zero, the variance of the reconciled forecast is very close to that of the perfect forecast using the data-generating process.

For most choices of ϕ_2, the residuals from the AR(1) model at the annual level contain additional information that can be used to improve forecasts at the half-yearly level through reconciliation. This illustrates the benefit of reconciliation. Even in this simple example, in order to unfold the full potential of reconciliation, the variance–covariance matrix for the combined vector of both forecasts is needed. In a real application, where the data-generating process is unknown, we would need to estimate $\boldsymbol{\Sigma}$, which would introduce an additional estimation error (see Møller et al., 2022).

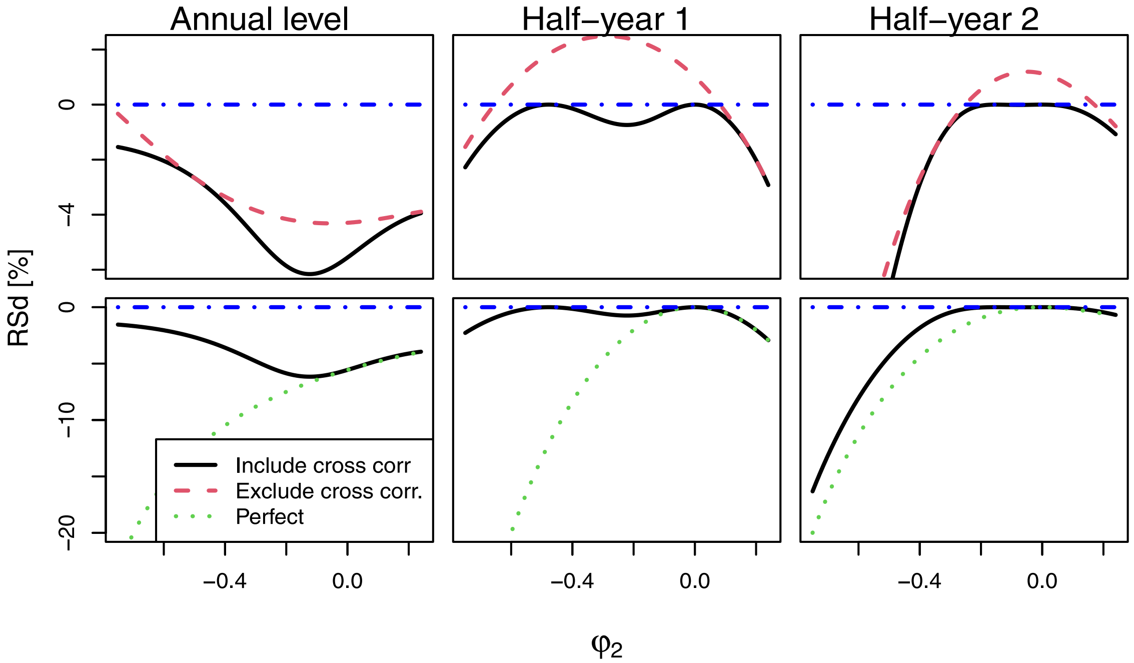

Fig. 1. RSd_i (negative means better than the base forecast) for two different constructions of the variance–covariance matrix (top row) and the perfect model (bottom row).

4. Model formulation and likelihood estimation

The main contribution of this article is to develop a framework for modelling and estimating $\boldsymbol{\Sigma}$, the variance–covariance matrix of the base forecast errors. Its often high dimension and unknown structure makes the estimation problem difficult and calls for tools for testing and simplifying its structure. To this end we develop an algorithm for likelihood estimation and model selection through statistical hypothesis testing. We formulate a statistical model using the framework of general linear models to arrive at a likelihood function. In order to find the maximum likelihood estimates in an efficient way, we derive the first and second derivatives of the log-likelihood with respect to the model parameters. The statistical tests in Section 5 follow directly from usual and well-known approximate results from likelihood theory. We begin by introducing our notation.

4.1. Notation

We use the following notation for the main variables, for a specific forecast. That is, we will have an index on some of them when we discuss the modelling:

- $\hat{\boldsymbol{y}} \in \mathbb{R}^n$ is the collection of *base* forecasts for all aggregation levels,
- $\tilde{\boldsymbol{y}} \in \mathbb{R}^n$ is the collection of *reconciled* forecasts for all aggregation levels,
- $\boldsymbol{y} \in \mathbb{R}^n$ is the observations for all aggregation levels,
- $\boldsymbol{S} \in \mathbb{R}^{n\times m}$ is the summation matrix,
- $\boldsymbol{e} \in \mathbb{R}^n$ is the base forecast error,
- $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$ is the variance–covariance matrix of $\boldsymbol{e}$, and
- $\hat{\boldsymbol{\Sigma}} \in \mathbb{R}^{n\times n}$ is an estimate of the variance–covariance matrix of $\boldsymbol{e}$.

The base forecast $\hat{\boldsymbol{y}}$ is treated as a known input to the system, which should be reconciled to give a *coherent* forecast; the reconciled forecast $\tilde{\boldsymbol{y}}$ is a linear function of $\hat{\boldsymbol{y}}$; $\boldsymbol{y}$ is the collection of the actual observations; and the base forecast error is $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$. Finally, $\boldsymbol{S}$ is the summation matrix that ensures aggregate coherency between the forecasts, which implies that

- $\tilde{\boldsymbol{y}} = \boldsymbol{S}\tilde{\boldsymbol{y}}_B$; $\tilde{\boldsymbol{y}}_B \in \mathbb{R}^m$,
- $\boldsymbol{y} = \boldsymbol{S}\boldsymbol{y}_B$; $\boldsymbol{y}_B \in \mathbb{R}^m$,

where subscript B refers to the forecasts or observations at the bottom level. There are $1, \dots, K$ aggregation levels, with $\hat{\boldsymbol{y}}_1 = \hat{\boldsymbol{y}}_B \in \mathbb{R}^m$ being the base forecast at the bottom level. The dimensions of the other base forecasts are defined by the aggregations.

As a small example, consider forecasts with quarterly, half-yearly, and annual resolution and a forecast horizon equal to one year. In this case, $\boldsymbol{y} = [y_1^{\mathrm{A}}, y_1^{\mathrm{H}}, y_2^{\mathrm{H}}, y_1^{\mathrm{Q}}, y_2^{\mathrm{Q}}, y_3^{\mathrm{Q}}, y_4^{\mathrm{Q}}]^T$, and similarly for $\hat{\boldsymbol{y}}$ and $\tilde{\boldsymbol{y}}$. The base forecast for the bottom level is $\hat{\boldsymbol{y}}_B = [\hat{y}_1^{\mathrm{Q}}, \hat{y}_2^{\mathrm{Q}}, \hat{y}_3^{\mathrm{Q}}, \hat{y}_4^{\mathrm{Q}}]^T$, and the summation matrix is given by

$$\boldsymbol{S} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{12}$$

Though there are multiple ways to define a summation matrix, in all cases its columns span the same coherent subspace, which is unique (Panagiotelis et al., 2021).

4.2. General linear model

Forecast reconciliation can be formulated as a general linear model:

$$\hat{\boldsymbol{y}} = \boldsymbol{S}\boldsymbol{b} + \boldsymbol{\epsilon}, \tag{13}$$

where $\boldsymbol{b}$ can be interpreted as $\boldsymbol{b} = E[\boldsymbol{y}_B]$. Assuming that $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_\epsilon)$ and $\boldsymbol{\Sigma}_\epsilon \in \mathbb{R}^{n\times n}$ is known, the reconciled forecasts, which are considered parameters in a linear regression model, can be estimated by

$$\tilde{\boldsymbol{y}}_B = \hat{\boldsymbol{b}} = (\boldsymbol{S}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}^T\boldsymbol{\Sigma}_\epsilon^{-1}\hat{\boldsymbol{y}}. \tag{14}$$

This leaves the problem of estimating the (possibly high-dimensional) variance–covariance matrix $\boldsymbol{\Sigma}_\epsilon$. The parameters in $\boldsymbol{\Sigma}_\epsilon$ cannot be estimated by maximum likelihood estimation directly, because the rank of the estimated $\boldsymbol{\Sigma}_\epsilon$ is lower than its dimension. It is well known in regression analysis that this is a consequence of treating the reconciled forecasts as parameters (see, e.g., Madsen & Thyregod, 2011). This is also addressed in the supplementary material for this article (Møller et al., 2022). Therefore, an estimate of the variance–covariance of the observed base forecast error is often used instead. Wickramasuriya et al. (2019) provided theoretical justification for this approach.

The variance–covariance of the base forecast error $\boldsymbol{e}_t$, where $t = \{1, \ldots, T\}$, may be observed directly from the data; but this involves estimating $\frac{n(n+1)}{2}$ parameters, where n is the dimension of $\boldsymbol{e}_t$. When perceived as a general linear model, the reconciled forecasts are parameters and $\boldsymbol{\epsilon}_t$ is the residual error. The variance of $\boldsymbol{e}_t$, $\boldsymbol{\Sigma}$, is assumed to be a good proxy for $V[\boldsymbol{\epsilon}_t] = \boldsymbol{\Sigma}_\epsilon$. Consequently, in this model we set $\boldsymbol{\Sigma}_\epsilon = \boldsymbol{\Sigma}$ and, in practice, we use some estimate $\hat{\boldsymbol{\Sigma}}$.

In the small example above, $\boldsymbol{\Sigma}$ has $7\cdot 8/2 = 28$ parameters. In the main application we consider in this article, the number of parameters is $60 \cdot 61/2 = 1830$. A number of approaches have been proposed in order to deal with this high-dimensional problem, ranging from diagonal approximations (Athanasopoulos et al., 2017) to shrinkage estimates of the full observed variance–covariance matrix (Nystrup et al., 2020). Here, we investigate a setup where the variance–covariance matrix is modelled using a hypothesis on the inverse variance–covariance matrices and summation matrices estimated using the maximum likelihood approach.

4.3. Model formulation

Next, we present the framework for parametric modelling of the variance–covariance matrix, which is the main contribution of this article. More technical issues related to parameter estimation are presented in Sections 4.4 and 4.5. The idea is that the covariance between aggregation levels can be modelled by aggregating the lower-level errors and disaggregating information from the higher levels.

Covariances between aggregation levels arise because data are shared between levels through the summation matrix. We construct the full variance–covariance matrix based on this data sharing. Starting from the bottom level, we define (omitting the index t)

$$\boldsymbol{\epsilon}_1 = \boldsymbol{u}_1 \quad ; \quad \boldsymbol{u}_1 \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_1), \tag{15}$$

$$\boldsymbol{\epsilon}_2 = \boldsymbol{S}_{21}\boldsymbol{u}_1 + \boldsymbol{u}_2 \quad ; \quad \boldsymbol{u}_2 \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_2), \tag{16}$$

$$\boldsymbol{\epsilon}_3 = \boldsymbol{S}_{31}\boldsymbol{u}_1 + \boldsymbol{S}_{32}\boldsymbol{u}_2 + \boldsymbol{u}_3 \quad ; \quad \boldsymbol{u}_3 \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_3), \tag{17}$$

$$\vdots$$

$$\boldsymbol{\epsilon}_K = \sum_{j=1}^{K-1}\boldsymbol{S}_{Kj}\boldsymbol{u}_j + \boldsymbol{u}_K \quad ; \quad \boldsymbol{u}_K \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_K), \tag{18}$$

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

The variance–covariance matrices $\boldsymbol{\Sigma}_i$ can be thought of as the within-level variance–covariance matrices, or as some parameterised estimate of the variance–covariance matrices. The dimension of $\boldsymbol{u}_i$ is equal to the dimension of the forecast at aggregation level i, with $\boldsymbol{u}_1 \in \mathbb{R}^m$. The matrices $\boldsymbol{S}_{ij}$ define the covariance between levels. $\boldsymbol{S}_{ij}$ is not equal to the summation matrix $\boldsymbol{S}$. They are related, however, and particular hypotheses about how they are related can be formulated directly through the structure of $\boldsymbol{S}_{ij}$. Although it seems reasonable that they are defined by the sharing of observations between aggregation levels, it is not a requirement.

Continuing the small quarterly-to-annual example from above, we have $\boldsymbol{\Sigma}_{\mathrm{Q}} \in \mathbb{R}^{4\times 4}$, $\boldsymbol{\Sigma}_{\mathrm{H}} \in \mathbb{R}^{2\times 2}$, and $\boldsymbol{\Sigma}_{\mathrm{A}} \in \mathbb{R}$. Further, $\boldsymbol{S}_{\mathrm{HQ}} \in \mathbb{R}^{2\times 4}$, $\boldsymbol{S}_{\mathrm{AQ}} \in \mathbb{R}^{1\times 4}$, and $\boldsymbol{S}_{\mathrm{AH}} \in \mathbb{R}^{1\times 2}$. Assuming $\boldsymbol{S}_{ij}$ is defined by the sharing of observations—i.e. that the estimate for the first half year is related to the first two quarters, the second half year to the last two quarters, and so on—we would expect that the parametric form is

$$\boldsymbol{S}_{\mathrm{HQ}} = \begin{bmatrix} \beta_{11}^{\mathrm{HQ}} & \beta_{12}^{\mathrm{HQ}} & 0 & 0 \\ 0 & 0 & \beta_{23}^{\mathrm{HQ}} & \beta_{24}^{\mathrm{HQ}} \end{bmatrix}, \tag{19}$$

$$\boldsymbol{S}_{\mathrm{AQ}} = \begin{bmatrix} \beta_{11}^{\mathrm{AQ}} & \beta_{12}^{\mathrm{AQ}} & \beta_{13}^{\mathrm{AQ}} & \beta_{14}^{\mathrm{AQ}} \end{bmatrix}, \tag{20}$$

$$\boldsymbol{S}_{\mathrm{AH}} = \begin{bmatrix} \beta_{11}^{\mathrm{AH}} & \beta_{12}^{\mathrm{AH}} \end{bmatrix}, \tag{21}$$

where β_{ij}^{kl} are parameters to be estimated from the data.

Of course, the variance–covariance matrices must also be estimated. While it is not reasonable to assume that the variance–covariance matrices are well approximated by sparse matrices, it is often reasonable to assume that their inverse is. This is, for example, the case for low-order AR processes, which are often good approximations of high-order AR processes. Therefore, we define $\boldsymbol{S}_{ii}$ such that

$$\boldsymbol{\Sigma}_i^{-1} = \boldsymbol{S}_{ii}\boldsymbol{S}_{ii}^T, \tag{22}$$

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

$$\boldsymbol{S}_{ii} = \begin{bmatrix} \alpha_{11}^i & \alpha_{12}^i & \cdots & \cdots & \alpha_{1n_1}^i \\ 0 & \alpha_{22}^i & \alpha_{23}^i & \cdots & \alpha_{2n_1}^i \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & & & \alpha_{n_i,n_i}^i \end{bmatrix}. \tag{23}$$

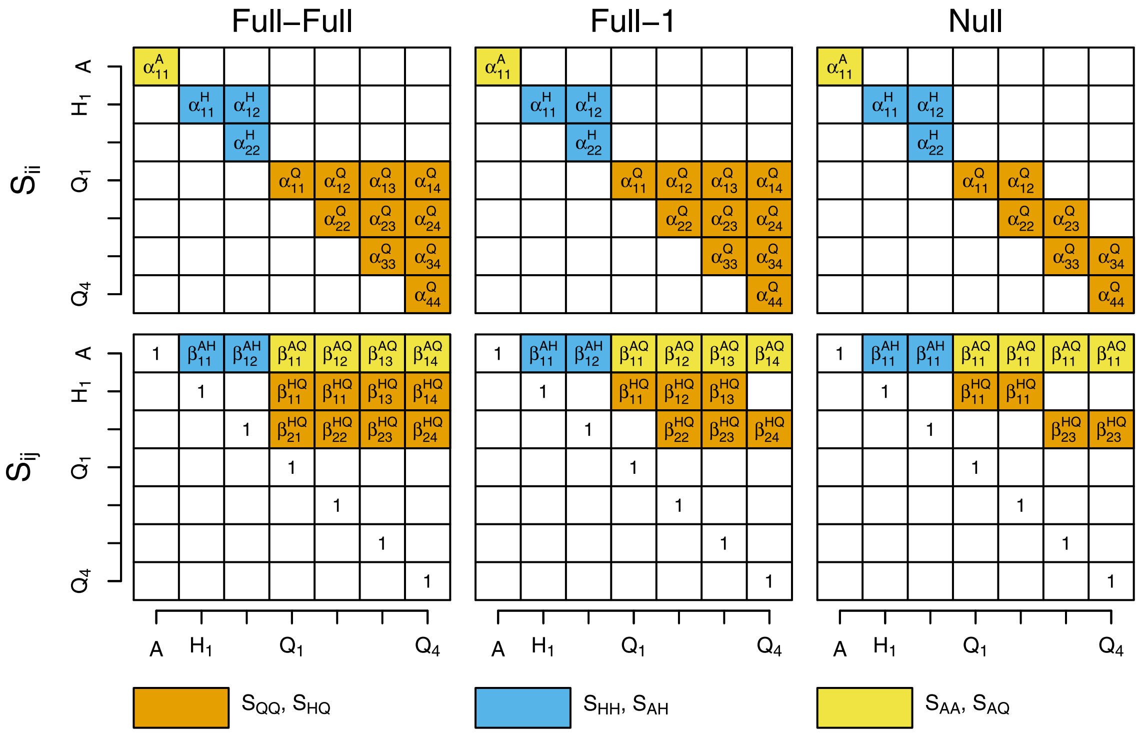
Full–Full
Full–1
Null
S_{ii}
S_{ij}
A
H_1
Q_1
Q_4
Full–Full S_{ii}: α^A_{11}; α^H_{11} α^H_{12}; α^H_{22}; α^Q_{11} α^Q_{12} α^Q_{13} α^Q_{14}; α^Q_{22} α^Q_{23} α^Q_{24}; α^Q_{33} α^Q_{34}; α^Q_{44}
Full–1 S_{ii}: α^A_{11}; α^H_{11} α^H_{12}; α^H_{22}; α^Q_{11} α^Q_{12} α^Q_{13} α^Q_{14}; α^Q_{22} α^Q_{23} α^Q_{24}; α^Q_{33} α^Q_{34}; α^Q_{44}
Null S_{ii}: α^A_{11}; α^H_{11} α^H_{12}; α^H_{22}; α^Q_{11} α^Q_{12}; α^Q_{22} α^Q_{23}; α^Q_{33} α^Q_{34}; α^Q_{44}
Full–Full S_{ij}: 1 β^{AH}_{11} β^{AH}_{12} β^{AQ}_{11} β^{AQ}_{12} β^{AQ}_{13} β^{AQ}_{14}; 1 β^{HQ}_{11} β^{HQ}_{11} β^{HQ}_{13} β^{HQ}_{14}; 1 β^{HQ}_{21} β^{HQ}_{22} β^{HQ}_{23} β^{HQ}_{24}; 1; 1; 1; 1
Full–1 S_{ij}: 1 β^{AH}_{11} β^{AH}_{12} β^{AQ}_{11} β^{AQ}_{12} β^{AQ}_{13} β^{AQ}_{14}; 1 β^{HQ}_{11} β^{HQ}_{12} β^{HQ}_{13}; 1 β^{HQ}_{22} β^{HQ}_{23} β^{HQ}_{24}; 1; 1; 1; 1
Null S_{ij}: 1 β^{AH}_{11} β^{AH}_{11} β^{AQ}_{11} β^{AQ}_{11} β^{AQ}_{11} β^{AQ}_{11}; 1 β^{HQ}_{11} β^{HQ}_{11}; 1 β^{HQ}_{23} β^{HQ}_{23}; 1; 1; 1; 1
A H_1 Q_1 Q_4
S_{QQ}, S_{HQ}
S_{HH}, S_{AH}
S_{AA}, S_{AQ}

Fig. 2. Illustration of the parameters in some of the tested models for a quarterly-to-annual hierarchy.

This is a Cholesky decomposition of $\boldsymbol{\Sigma}_i^{-1}$. If $\boldsymbol{\Sigma}_i^{-1}$ is sparse, then so is $\boldsymbol{S}_{ii}$. If, for some reason, it is reasonable to assume that $\boldsymbol{\Sigma}_i$ is sparse but its inverse is not, then the derivation done in this section should be redone, and the Wishart distribution should be replaced by the inverse Wishart distribution in Section 4.4. If neither $\boldsymbol{\Sigma}_i$ nor $\boldsymbol{\Sigma}_i^{-1}$ can reasonably be assumed to be sparse, then one has to estimate all values of α^i_{kl}.

In Fig. 2, the setup is illustrated for three different structures for the small quarterly-to-annual example. These structures are chosen to explain the models considered in Section 6. The first column (Full-Full) is a different representation of the full variance–covariance matrix with the same number of parameters. In the second column (Full-1), the variance–covariance of $\boldsymbol{u}_i$ is free, but the matrix $\boldsymbol{S}_{\mathrm{HQ}}$ is restricted (one extra parameter per row, compared to (19)). This implies that the number of parameters is reduced by two in this small example. Finally, in the third column (NULL), the variance–covariance of $\boldsymbol{u}_i$ is restricted to an AR(1) process, and only elements of $\boldsymbol{S}_{ij}$ directly related through data sharing are different from zero. In other words, $\boldsymbol{S}_{ij}$ is captured by just one parameter.

The three structures in Fig. 2 have 28, 26, and 15 parameters, respectively. In this small example, the reduction in the number of parameters is not large. For more realistic examples the reduction will be much larger. For example, in the case we consider in Sections 6 and 7, the number of parameters will be 1830, 930, and 140.

It is expected that $\boldsymbol{S}_{ii}$ is sparse in many cases. For example, for the stationary AR(1) process only the diagonal and one-lag off-diagonal elements of $\boldsymbol{\Sigma}_i^{-1}$ are different from zero. Hence, in that case $\alpha^i_{kl} = 0$ for $l - k > 1$. Furthermore, in the case of a stationary AR(1) process, we would have $\alpha^i_{kk} = \alpha^i_{ll}$ and $\alpha^i_{k,k+1} = \alpha^i_{k+1,k+2}$, which drastically reduces the number of parameters. In the continued investigation we will suppress the superscript of α and β.

We can write the model for the variance–covariance as

$$\begin{bmatrix}\boldsymbol{\epsilon}_K\\ \vdots\\ \boldsymbol{\epsilon}_1\end{bmatrix} = \begin{bmatrix}\boldsymbol{I} & \boldsymbol{S}_{K,K-1} & \cdots & \cdots & \boldsymbol{S}_{K,1}\\ \boldsymbol{0} & \boldsymbol{I} & \cdots & & \boldsymbol{S}_{K-1,1}\\ \vdots & & \ddots & & \vdots\\ \boldsymbol{0} & \cdots & & & \boldsymbol{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{u}_K\\ \vdots\\ \boldsymbol{u}_1\end{bmatrix}, \tag{24}$$

where the variance–covariance matrix for $\boldsymbol{u} = [\boldsymbol{u}_K^T, \dots, \boldsymbol{u}_1^T]^T$ is block diagonal.

As short-hand notation we will write

$$\boldsymbol{\epsilon} = \boldsymbol{S}_u\boldsymbol{u}; \quad \boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}^u), \tag{25}$$

and

$$V[\boldsymbol{\epsilon}] = \boldsymbol{S}_u\boldsymbol{\Sigma}^u\boldsymbol{S}_u^T. \tag{26}$$

The final model for the residual of the forecast errors is

$$\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{S}_u\boldsymbol{\Sigma}^u\boldsymbol{S}_u^T), \tag{27}$$

where $\boldsymbol{\Sigma}^u$ is a function of $\boldsymbol{\alpha}$, and $\boldsymbol{S}_u$ is a function of $\boldsymbol{\beta}$.

It should be emphasised that (15)–(18) define a representation of the observed variance–covariance matrix. In particular, if all values of $\boldsymbol{S}_{ii}$ and $\boldsymbol{S}_{ij}$ are properly adjusted (corresponding to the Full-Full model in Fig. 2), then the model is just another representation of the observed variance–covariance matrix. This is often referred to as a saturated model. We use the representation for maximum likelihood estimation of the parameters, defining $\boldsymbol{S}_{ii}$ and $\boldsymbol{S}_{ij}$.

4.4. Likelihood estimation

In order to formulate an estimation problem, we need to define a loss function or a distance between the parameterised version of the variance–covariance matrix and the observed one. As we assumed a specific distribution of all errors, it is natural to use a likelihood approach. In a univariate setting (i.e. $y_i \sim N(\mu, \sigma^2)$), the empirical variance is related to the χ^2-distribution by $\frac{(T-1)S^2}{\sigma^2} \sim \chi^2(T-1)$. In a multivariate setting (under the multivariate normal assumption), the observed variance–covariance matrix is related to the Wishart distribution by $(T-1)\boldsymbol{V} \sim W(\boldsymbol{\Sigma}, T-1)$, where we write the observed variance–covariance matrix as

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{V}_{KK} & \cdots & \boldsymbol{V}_{K1} \\ \vdots & \ddots & \vdots \\ \boldsymbol{V}_{K1}^T & \cdots & \boldsymbol{V}_{11} \end{bmatrix}. \tag{28}$$

With the assumptions outlined above, the likelihood is defined by the Wishart distribution, which implies that the log-likelihood is given by

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\alpha}; \boldsymbol{V}) &\propto -\frac{1}{2} Tr(\boldsymbol{\Sigma}^{-1}(T-1)\boldsymbol{V}) - \frac{T-1}{2} \log|\boldsymbol{\Sigma}| \\ &= -\frac{T-1}{2} Tr\, \boldsymbol{\Sigma}^{-1}\boldsymbol{V} + \frac{T-1}{2} \log|\boldsymbol{\Sigma}^{-1}|. \end{aligned} \tag{29}$$

The maximum likelihood estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are given by

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\alpha}} \end{bmatrix} = \arg\max_{\boldsymbol{\beta}, \boldsymbol{\alpha}} l(\boldsymbol{\beta}, \boldsymbol{\alpha}; \boldsymbol{V}), \tag{30}$$

and the maximum likelihood estimate of $\boldsymbol{\Sigma}$ is

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}). \tag{31}$$

The estimation is done for one aggregation level at a time, starting from the bottom level, assuming that the lower levels are correctly estimated. The rationale is that the variance–covariance matrix should be correct for the lower levels before we continue with higher levels in the hierarchy. This approximation is similar to the composite conditional likelihood presented in Varin et al. (2011).

This implies that it suffices to consider the setup

$$\boldsymbol{S}_u = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{S}_{21} \\ \boldsymbol{0} & \boldsymbol{T}_{11} \end{bmatrix} \tag{32}$$

and

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{S}_{22}\boldsymbol{S}_{22}^T & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_1^{-1} \end{bmatrix}. \tag{33}$$

Here, $\boldsymbol{T}_{11}$ is $\boldsymbol{S}_u$ for the level preceding the current level, which is fixed in the estimation. If we again consider our small example, we would start by estimating $\boldsymbol{S}_{11} \in \mathbb{R}^{4\times4}$ using $[y_1^Q, y_2^Q, y_3^Q, y_4^Q]^T$. Then, with $\boldsymbol{S}_{11}$ (and hence $\boldsymbol{\Sigma}_1^{-1}$) fixed, at $\boldsymbol{S}_{11}$, we would estimate $\boldsymbol{S}_{22}$ and $\boldsymbol{S}_{12}$ and set $\boldsymbol{\Sigma}_1^{-1} = \boldsymbol{S}_{11}\boldsymbol{S}_{11}^T$.

The next step would be to rename

$$\boldsymbol{T}_{11} := \begin{bmatrix} \boldsymbol{I} & \boldsymbol{S}_{21} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \tag{34}$$

$$\boldsymbol{\Sigma}_1 := \boldsymbol{T}_{11} \begin{bmatrix} \boldsymbol{\Sigma}_2 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_1 \end{bmatrix} \boldsymbol{T}_{11}^T \tag{35}$$

and estimate $\boldsymbol{S}_{33}$ (which is named $\boldsymbol{S}_{22}$) and the summation

$$\boldsymbol{S}_{21} := \begin{bmatrix} \boldsymbol{S}_{32} & \boldsymbol{S}_{31} \end{bmatrix}. \tag{36}$$

In order to set up the estimation, we examine the two terms in the log-likelihood (29). We begin with the term inside the trace:

$$\begin{aligned} \boldsymbol{\Sigma}^{-1}\boldsymbol{V} &= \boldsymbol{S}_u^{-T}\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_u^{-1}\boldsymbol{V} \\ &= \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ -\boldsymbol{T}_{11}^{-T}\boldsymbol{S}_{21}^T & \boldsymbol{T}_{11}^{-T} \end{bmatrix} \begin{bmatrix} \boldsymbol{S}_{22}\boldsymbol{S}_{22}^T & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_1^{-1} \end{bmatrix} \boldsymbol{S}_u^{-1}\boldsymbol{V} \\ &= \begin{bmatrix} \boldsymbol{S}_{22}\boldsymbol{S}_{22}^T & \boldsymbol{0} \\ -\boldsymbol{T}_{11}^{-T}\boldsymbol{S}_{21}^T\boldsymbol{S}_{22}\boldsymbol{S}_{22}^T & \boldsymbol{T}_{11}^{-T}\boldsymbol{\Sigma}_1^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{S}_{21}\boldsymbol{T}_{11}^{-1} \\ \boldsymbol{0} & \boldsymbol{T}_{11}^{-1} \end{bmatrix} \boldsymbol{V} \\ &= \begin{bmatrix} \boldsymbol{S}_{22}\boldsymbol{S}_{22}^T & -\boldsymbol{S}_{22}\boldsymbol{S}_{22}^T\boldsymbol{S}_{21}\boldsymbol{T}_{11}^{-1} \\ -\boldsymbol{T}_{11}^{-T}\boldsymbol{S}_{21}^T\boldsymbol{S}_{22}\boldsymbol{S}_{22}^T & \boldsymbol{T}_{11}^{-T}\boldsymbol{S}_{21}^T\boldsymbol{S}_{22}\boldsymbol{S}_{22}^T\boldsymbol{S}_{21}\boldsymbol{T}_{11}^{-1} + \boldsymbol{T}_{11}^{-T}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{T}_{11}^{-1} \end{bmatrix} \\ &\quad \times \begin{bmatrix} \boldsymbol{V}_{22} & \boldsymbol{V}_{21} \\ \boldsymbol{V}_{21}^T & \boldsymbol{V}_{11} \end{bmatrix}. \end{aligned} \tag{37}$$

Since we are interested in the trace, we only need the block-diagonal elements:

$$(\boldsymbol{\Sigma}^{-1}\boldsymbol{V})_{11} = \boldsymbol{S}_{22}\boldsymbol{S}_{22}^T\boldsymbol{V}_{22} - \boldsymbol{S}_{22}\boldsymbol{S}_{22}^T\boldsymbol{S}_{21}\boldsymbol{T}_{11}^{-1}\boldsymbol{V}_{21}^T, \tag{38}$$

$$\begin{aligned} (\boldsymbol{\Sigma}^{-1}\boldsymbol{V})_{22} &= -\boldsymbol{T}_{11}^{-T}\boldsymbol{S}_{21}^T\boldsymbol{S}_{22}\boldsymbol{S}_{22}^T\boldsymbol{V}_{21} \\ &\quad + \boldsymbol{T}_{11}^{-T}\boldsymbol{S}_{21}^T\boldsymbol{S}_{22}\boldsymbol{S}_{22}^T\boldsymbol{S}_{21}\boldsymbol{T}_{11}^{-1}\boldsymbol{V}_{11} \\ &\quad + \boldsymbol{T}_{11}^{-T}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{T}_{11}^{-1}\boldsymbol{V}_{11}. \end{aligned} \tag{39}$$

Omitting terms that do not contain $\boldsymbol{S}_{ij}$, we can write the first part of the log-likelihood as

$$\begin{aligned} Tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{V}) &\propto Tr(\boldsymbol{S}_{22}\boldsymbol{S}_{22}^T\boldsymbol{V}_{22}) - Tr(\boldsymbol{S}_{22}\boldsymbol{S}_{22}^T\boldsymbol{S}_{21}\boldsymbol{T}_{11}^{-1}\boldsymbol{V}_{21}^T) - \\ &\quad Tr(\boldsymbol{T}_{11}^{-T}\boldsymbol{S}_{21}^T\boldsymbol{S}_{22}\boldsymbol{S}_{22}^T\boldsymbol{V}_{21}) \\ &\quad + Tr(\boldsymbol{T}_{11}^{-T}\boldsymbol{S}_{21}^T\boldsymbol{S}_{22}\boldsymbol{S}_{22}^T\boldsymbol{S}_{21}\boldsymbol{T}_{11}^{-1}\boldsymbol{V}_{11}) \\ &= Tr(\boldsymbol{S}_{22}\boldsymbol{S}_{22}^T\boldsymbol{V}_{22}) - 2Tr(\boldsymbol{S}_{22}\boldsymbol{S}_{22}^T\boldsymbol{S}_{21}\boldsymbol{T}_{11}^{-1}\boldsymbol{V}_{21}^T) \\ &\quad + Tr(\boldsymbol{T}_{11}^{-T}\boldsymbol{S}_{21}^T\boldsymbol{S}_{22}\boldsymbol{S}_{22}^T\boldsymbol{S}_{21}\boldsymbol{T}_{11}^{-1}\boldsymbol{V}_{11}). \end{aligned} \tag{40}$$

For the second term in the log-likelihood, we have

$$\begin{aligned} \log|\boldsymbol{\Sigma}^{-1}| &= \log|\boldsymbol{S}_u^{-T}\boldsymbol{\Sigma}_u^{-1}\boldsymbol{S}_u^{-1}| \\ &= \log|\boldsymbol{S}_u^{-T}| + \log|\boldsymbol{\Sigma}_u^{-1}| + \log|\boldsymbol{S}_u^{-1}| \\ &= \log|\boldsymbol{\Sigma}_2^{-1}| + \log|\boldsymbol{\Sigma}_1^{-1}| + 2\log|\boldsymbol{S}_u^{-1}| \\ &= \log|\boldsymbol{S}_{22}\boldsymbol{S}_{22}^T| + \log|\boldsymbol{\Sigma}_1^{-1}| + 2\log|\boldsymbol{S}_u^{-1}| \\ &= 2\log|\boldsymbol{S}_{22}| + \log|\boldsymbol{\Sigma}_1^{-1}| + 2\log|\boldsymbol{S}_u^{-1}|. \end{aligned} \tag{41}$$

Both $\boldsymbol{S}_{22}$ and $\boldsymbol{S}_u^{-1}$ are upper triangular matrices. The diagonal elements of $\boldsymbol{S}_u^{-1}$ are all equal to one and, hence, $\log|\boldsymbol{S}_u^{-1}| = 0$. The diagonal elements of $\boldsymbol{S}_{22}$ are α_{ii}^2, which means that

$$\log|\boldsymbol{\Sigma}^{-1}| \propto \sum_i \log\alpha_{ii}^2. \tag{42}$$

We use the following notation:

- $\boldsymbol{I}_{ij}^{22}$ is a matrix with the same dimensions as $\boldsymbol{S}_{22}$ and elements corresponding to α_{ij} equal to one, and
- $\boldsymbol{I}_{ij}^{21}$ is a matrix with the same dimensions as $\boldsymbol{S}_{21}$ and elements corresponding to β_{ij} equal to one.

In order to find the estimation equations, we calculate the derivative of the log-likelihood function:

$$\begin{aligned}\frac{\partial l}{\partial \beta_{ij}} &= (T-1)Tr(\boldsymbol{S}_{22}\boldsymbol{S}_{22}^T\boldsymbol{I}_{ij}^{21}\boldsymbol{T}_{11}^{-1}\boldsymbol{V}_{21}^T)\\ &\quad -\frac{1}{2}(T-1)Tr(\boldsymbol{T}_{11}^{-T}(\boldsymbol{I}_{ij}^{21})^T\boldsymbol{S}_{22}\boldsymbol{S}_{22}^T\boldsymbol{S}_{21}\boldsymbol{T}_{11}^{-1}\boldsymbol{V}_{11})\\ &\quad -\frac{1}{2}(T-1)Tr(\boldsymbol{T}_{11}^{-T}\boldsymbol{S}_{21}^T\boldsymbol{S}_{22}\boldsymbol{S}_{22}^T\boldsymbol{I}_{ij}^{21}\boldsymbol{T}_{11}^{-1}\boldsymbol{V}_{11})\\ &= (T-1)Tr(\boldsymbol{S}_{22}\boldsymbol{S}_{22}^T\boldsymbol{I}_{ij}^{21}\boldsymbol{T}_{11}^{-1}\boldsymbol{V}_{21}^T)\\ &\quad -(T-1)Tr(\boldsymbol{T}_{11}^{-T}(\boldsymbol{I}_{ij}^{21})^T\boldsymbol{S}_{22}\boldsymbol{S}_{22}^T\boldsymbol{S}_{21}\boldsymbol{T}_{11}^{-1}\boldsymbol{V}_{11})\\ &= (T-1)Tr(\boldsymbol{S}_{22}\boldsymbol{S}_{22}^T\boldsymbol{I}_{ij}^{21}\boldsymbol{T}_{11}^{-1}\boldsymbol{V}_{21}^T)\\ &\quad -(T-1)\sum_{kl}\beta_{kl}Tr(\boldsymbol{T}_{11}^{-T}(\boldsymbol{I}_{ij}^{21})^T\boldsymbol{S}_{22}\boldsymbol{S}_{22}^T\boldsymbol{I}_{kl}^{21}\boldsymbol{T}_{11}^{-1}\boldsymbol{V}_{11}).\end{aligned} \tag{43}$$

The estimation equation ($\frac{\partial l}{\partial \boldsymbol{\beta}} = 0$) for $\boldsymbol{\beta}$ given $\boldsymbol{\alpha}$ can be formulated as

$$\boldsymbol{X}_{\beta}(\boldsymbol{\alpha})\boldsymbol{\beta} = \boldsymbol{Y}_{\beta}(\boldsymbol{\alpha}) \tag{44}$$

or

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}) = \boldsymbol{X}_{\beta}^{-1}(\boldsymbol{\alpha})\boldsymbol{Y}_{\beta}(\boldsymbol{\alpha}). \tag{45}$$

The estimation equation for $\boldsymbol{\alpha}$ is split into two parts: the diagonal elements ($\boldsymbol{\alpha}^{ii}$), and the off-diagonal elements ($\boldsymbol{\alpha}^{ij}$). We start by rewriting the trace

$$\begin{aligned}Tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{V}) &\propto Tr(\boldsymbol{S}_{22}\boldsymbol{S}_{22}^T\boldsymbol{V}_{22}) - 2Tr(\boldsymbol{S}_{22}\boldsymbol{S}_{22}^T\boldsymbol{S}_{21}\boldsymbol{T}_{11}^{-1}\boldsymbol{V}_{21}^T)\\ &\quad + Tr(\boldsymbol{T}_{11}^{-T}\boldsymbol{S}_{21}^T\boldsymbol{S}_{22}\boldsymbol{S}_{22}^T\boldsymbol{S}_{21}\boldsymbol{T}_{11}^{-1}\boldsymbol{V}_{11})\\ &= Tr(\boldsymbol{S}_{22}\boldsymbol{S}_{22}^T(\boldsymbol{V}_{22} - 2\boldsymbol{S}_{22}\boldsymbol{S}_{22}^T\boldsymbol{S}_{21}\boldsymbol{T}_{11}^{-1}\boldsymbol{V}_{21}^T))\\ &\quad + Tr(\boldsymbol{S}_{22}\boldsymbol{S}_{22}^T\boldsymbol{S}_{21}\boldsymbol{T}_{11}^{-1}\boldsymbol{V}_{11}\boldsymbol{T}_{11}^{-T}\boldsymbol{S}_{21}^T)\\ &= Tr(\boldsymbol{S}_{22}\boldsymbol{S}_{22}^T(\boldsymbol{V}_{22} - 2\boldsymbol{S}_{12}\boldsymbol{T}_{11}^{-1}\boldsymbol{V}_{21}^T\\ &\quad + \boldsymbol{S}_{21}\boldsymbol{T}_{11}^{-1}\boldsymbol{V}_{11}\boldsymbol{T}_{11}^{-T}\boldsymbol{S}_{21}^T))\\ &= Tr(\boldsymbol{S}_{22}\boldsymbol{S}_{22}^T\boldsymbol{F}(\boldsymbol{\beta})).\end{aligned} \tag{46}$$

The derivative of the log-likelihood with respect to α_{ij} ($i \neq j$) is

$$\begin{aligned}\frac{\partial l}{\partial \alpha_{ij}} &= -\frac{1}{2}(T-1)\left[Tr(\boldsymbol{I}_{ij}^{22}\boldsymbol{S}_{22}^T\boldsymbol{F}(\boldsymbol{\beta})) + Tr(\boldsymbol{S}_{22}(\boldsymbol{I}_{ij}^{22})^T\boldsymbol{F}(\boldsymbol{\beta}))\right]\\ &= -\frac{1}{2}(T-1)\left[Tr(\boldsymbol{I}_{ij}^{22}\boldsymbol{S}_{22}^T\boldsymbol{F}(\boldsymbol{\beta})) + Tr((\boldsymbol{I}_{ij}^{22})\boldsymbol{S}_{22}^T\boldsymbol{F}^T(\boldsymbol{\beta}))\right]\\ &= -\frac{1}{2}(T-1)Tr(\boldsymbol{I}_{ij}^{22}\boldsymbol{S}_{22}^T(\boldsymbol{F}(\boldsymbol{\beta}) + \boldsymbol{F}^T(\boldsymbol{\beta})))\\ &= -\frac{1}{2}(T-1)\sum_{k\neq l}\alpha_{kl}Tr(\boldsymbol{I}_{ij}^{22}(\boldsymbol{I}_{kl}^{22})^T(\boldsymbol{F}(\boldsymbol{\beta}) + \boldsymbol{F}^T(\boldsymbol{\beta})))\\ &\quad -\frac{1}{2}(T-1)\sum_{k}\alpha_{kk}Tr(\boldsymbol{I}_{ij}^{22}(\boldsymbol{I}_{kk}^{22})^T(\boldsymbol{F}(\boldsymbol{\beta}) + \boldsymbol{F}^T(\boldsymbol{\beta}))).\end{aligned} \tag{47}$$

The estimation equation ($\frac{\partial l}{\partial \boldsymbol{\alpha}^{ij}} = 0$) for $\boldsymbol{\alpha}^{ij}$ given $\boldsymbol{\alpha}^{ii}$ and $\boldsymbol{\beta}$ can be formulated as

$$\boldsymbol{X}_{\alpha^{ij}}(\boldsymbol{\beta})\boldsymbol{\alpha}^{ij} = \boldsymbol{Y}_{\alpha^{ij}}(\boldsymbol{\alpha}^{ii}, \boldsymbol{\beta}) \tag{48}$$

or

$$\hat{\boldsymbol{\alpha}}^{ij} = \boldsymbol{X}_{\alpha^{ij}}^{-1}(\boldsymbol{\beta})\boldsymbol{Y}_{\alpha^{ij}}(\boldsymbol{\alpha}^{ii}, \boldsymbol{\beta}). \tag{49}$$

The derivative of the log-likelihood with respect to α_{ii} is

$$\begin{aligned}\frac{\partial l}{\partial \alpha_{ii}} &= \frac{1}{2}(T-1)\left[-\sum_{k\neq l}\alpha_{kl}Tr(\boldsymbol{I}_{ii}^{22}(\boldsymbol{I}_{kl}^{22})^T(\boldsymbol{F}(\boldsymbol{\beta}) + \boldsymbol{F}^T(\boldsymbol{\beta})))\right.\\ &\quad \left. -\sum_{k}\alpha_{kk}Tr(\boldsymbol{I}_{ii}^{22}(\boldsymbol{I}_{kk}^{22})^T(\boldsymbol{F}(\boldsymbol{\beta}) + \boldsymbol{F}^T(\boldsymbol{\beta}))) + \frac{2}{\alpha_{ii}}Tr(\boldsymbol{I}_{ii}^{22})\right].\end{aligned} \tag{50}$$

Note that $\boldsymbol{I}_{ii}^{22}(\boldsymbol{I}_{kk}^{22})^T = \boldsymbol{0}$ for $i \neq k$. Therefore, we get the estimation equation

$$\begin{aligned}0 &= \alpha_{ii}\sum_{k\neq l}\alpha_{kl}Tr(\boldsymbol{I}_{ii}^{22}(\boldsymbol{I}_{kl}^{22})^T(\boldsymbol{F}(\boldsymbol{\beta}) + \boldsymbol{F}^T(\boldsymbol{\beta})))\\ &\quad + \alpha_{ii}^2 Tr(\boldsymbol{I}_{ii}^{22}(\boldsymbol{I}_{ii}^{22})^T(\boldsymbol{F}(\boldsymbol{\beta}) + \boldsymbol{F}^T(\boldsymbol{\beta}))) - 2Tr(\boldsymbol{I}_{ii}^{22})\\ &= a(\boldsymbol{\beta})\alpha_{ii}^2 + b(\boldsymbol{\alpha}^{ij}, \boldsymbol{\beta})\alpha_{ij} - c\end{aligned} \tag{51}$$

with solutions

$$\alpha_{ii}^{*} = \frac{-b(\boldsymbol{\alpha}^{ij}, \boldsymbol{\beta}) \pm \sqrt{b^2(\boldsymbol{\alpha}^{ij}, \boldsymbol{\beta}) + 8a(\boldsymbol{\beta})c}}{2a(\boldsymbol{\beta})}, \tag{52}$$

where c is clearly larger than zero, and $a(\boldsymbol{\beta}) > 0$. To see this, consider random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ with $V[\boldsymbol{x}] = \boldsymbol{V}_{22}$, $V[\boldsymbol{y}] = \boldsymbol{V}_{11}$, and $Cov[\boldsymbol{x}, \boldsymbol{y}] = \boldsymbol{V}_{12}$. Then $V[\boldsymbol{x} - \boldsymbol{S}_{12}\boldsymbol{T}_{11}^{-1}\boldsymbol{y}] = \boldsymbol{F} + \boldsymbol{F}^T$ and $a(\boldsymbol{\beta}) = Tr\left(\boldsymbol{I}_{ii}^{22}(\boldsymbol{I}_{ii}^{22})^T\left(\boldsymbol{F}(\boldsymbol{\beta}) + \boldsymbol{F}^T(\boldsymbol{\beta})\right)\right) = (\boldsymbol{F} + \boldsymbol{F}^T)_{ii} > 0$, meaning that α_{ii}^{*} is not complex. Thus,

$$\hat{\alpha}_{ii} = \frac{-b(\boldsymbol{\alpha}^{ij}, \boldsymbol{\beta}) + \sqrt{b^2(\boldsymbol{\alpha}^{ij}, \boldsymbol{\beta}) + 8a(\boldsymbol{\beta})c}}{2a(\boldsymbol{\beta})} > 0, \tag{53}$$

which is the solution we choose here.

Eqs. (45), (49), and (53) define a robust, iterative algorithm (similar to an expectation–maximisation algorithm) for this problem. The solution can be found by iterating between these three equations. The algorithm is robust because every iteration increases the value of the likelihood function. Therefore, it will eventually converge, although convergence might be to a local optimum (the upper bound for the likelihood is when $\boldsymbol{S}_u\boldsymbol{\Sigma}^u\boldsymbol{S}_u^T = \boldsymbol{V}$, cf. (27)).

4.5. Newton 's method and the Hessian

It is well known that algorithms like the one described in the previous subsection can be fairly slow. An alternative is to use Newton's method. To that end, the Hessian of the parameters is needed. The Hessian also serves as the basis for the Wald test that is used for model reduction and may be used for model building through the score test.

When using Newton's method, it is reasonable to require that $\alpha_{ii} > 0$. A simple way to do this is by estimating

$\tau_{ii} = \log(\alpha_{ii})$. We first note that

$$\frac{\partial l}{\partial \tau_{ii}} = \frac{\partial l}{\partial \alpha_{ii}} \frac{\partial \alpha_{ii}}{\partial \tau_{ii}} \\ = -\frac{1}{2}(T-1)e^{\tau_{ii}} Tr(\boldsymbol{I}_{ii}^{22}\boldsymbol{S}_{22}^{T}(\boldsymbol{F}(\boldsymbol{\beta}) + \boldsymbol{F}^{T}(\boldsymbol{\beta}))) + 2Tr(\boldsymbol{I}_{ii}^{22}). \tag{54}$$

The second derivative can be calculated by

$$\frac{\partial^2 l}{\partial \tau_{ii}\tau_{kk}} = -\frac{1}{2}(T-1)e^{\tau_{ii}+\tau_{kk}} Tr(\boldsymbol{I}_{ii}^{22}(\boldsymbol{I}_{kk}^{22})^{T}(\boldsymbol{F}(\boldsymbol{\beta}) + \boldsymbol{F}^{T}(\boldsymbol{\beta}))). \tag{55}$$

Hence,

$$\frac{\partial^2 l}{\partial \tau_{ii}^2} = -\frac{1}{2}(T-1)e^{2\tau_{ii}} Tr(\boldsymbol{I}_{ii}^{22}(\boldsymbol{F}(\boldsymbol{\beta}) + \boldsymbol{F}^{T}(\boldsymbol{\beta}))), \tag{56}$$

$$\frac{\partial^2 l}{\partial \tau_{ii}\tau_{kk}} = 0; \quad i \neq k, \tag{57}$$

and

$$\frac{\partial^2 l}{\partial \tau_{ii}\alpha_{kl}} = -\frac{1}{2}(T-1)e^{\tau_{ii}} Tr(\boldsymbol{I}_{ii}^{22}(\boldsymbol{I}_{ij}^{22})^{T}(\boldsymbol{F}(\boldsymbol{\beta})+\boldsymbol{F}^{T}(\boldsymbol{\beta}))); \quad k \neq l. \tag{58}$$

For α_{ij}, we get

$$\frac{\partial^2 l}{\partial \alpha_{ij}\alpha_{kl}} = \frac{1}{2}(T-1)Tr\left((\boldsymbol{I}_{ij}^{22})^{T}\boldsymbol{I}_{kl}^{22}\right). \tag{59}$$

Now note that

$$\frac{\partial l}{\partial \beta_{ij}} = (T-1)Tr(\boldsymbol{S}_{22}\boldsymbol{S}_{22}^{T}(\boldsymbol{I}_{ij}^{21}\boldsymbol{T}_{11}^{-1}\boldsymbol{V}_{21}^{T} - \boldsymbol{S}_{21}\boldsymbol{T}_{11}^{-1}\boldsymbol{V}_{11}\boldsymbol{T}_{11}^{-T}(\boldsymbol{I}_{ij}^{21})^{T})) \tag{60}$$

and, therefore,

$$\frac{\partial^2 l}{\partial \alpha_{kl}\partial \beta_{ij}} = (T-1)Tr((\boldsymbol{I}_{kl}^{22}\boldsymbol{S}_{22}^{T} + \boldsymbol{S}_{22}(\boldsymbol{I}_{kl}^{22})^{T})(\boldsymbol{I}_{ij}^{21}\boldsymbol{T}_{11}^{-1}\boldsymbol{V}_{21}^{T} \\ - \boldsymbol{S}_{21}\boldsymbol{T}_{11}^{-1}\boldsymbol{V}_{11}\boldsymbol{T}_{11}^{-T}(\boldsymbol{I}_{ij}^{21})^{T})), \tag{61}$$

$$\frac{\partial^2 l}{\partial \tau_{kk}\partial \beta_{ij}} = (T-1)e^{\tau_{kk}}Tr((\boldsymbol{I}_{kk}^{22}\boldsymbol{S}_{22}^{T} + \boldsymbol{S}_{22}(\boldsymbol{I}_{kk}^{22})^{T})(\boldsymbol{I}_{ij}^{21}\boldsymbol{T}_{11}^{-1}\boldsymbol{V}_{21}^{T} \\ - \boldsymbol{S}_{21}\boldsymbol{T}_{11}^{-1}\boldsymbol{V}_{11}\boldsymbol{T}_{11}^{-T}(\boldsymbol{I}_{ij}^{21})^{T})). \tag{62}$$

Finally, we have that

$$\frac{\partial^2 l}{\partial \beta_{ij}\partial \beta_{kl}} = -(T-1)Tr(\boldsymbol{S}_{22}\boldsymbol{S}_{22}^{T}\boldsymbol{I}_{kl}^{21}\boldsymbol{T}_{11}^{-1}\boldsymbol{V}_{11}\boldsymbol{T}_{11}^{-T}(\boldsymbol{I}_{ij}^{21})^{T}). \tag{63}$$

Newton's method is much faster (needs fewer iterations) than the robust, iterative algorithm defined by (45), (49), and (53), but it requires good initial values in order to converge. In the final algorithm, we use the robust algorithm to initialise before using Newton's method. In cases where Newton's method diverges even after initialisation, we go back to the slower but robust algorithm defined in Section 4.4.

5. Model reduction and shrinkage

Using the presented method, the full or saturated model has the same number of parameters as the empirical variance–covariance matrix. It is often of interest to reduce models using statistical inference. The likelihood framework allows us to explore different hypotheses about the correlation structure. We can, for example, use Wald or likelihood-ratio tests to remove insignificant parameters by setting them equal to zero, or test the hypothesis that certain parameters are equal. We can also use specific hypotheses when setting up the model, for example by assuming that the correlation matrix within each aggregation level has an AR(1) structure, or by assuming some degree of sparsity in the between-level summation matrix.

The Wald test can be calculated directly based on the results from Newton's method. The likelihood-ratio test, on the other hand, is time-consuming because the likelihood of the reduced model needs to be recalculated using the iterative method described above.

Consequently, we suggest using the Wald test to select candidates for model reduction and confirming or rejecting the reduction using a likelihood-ratio test. Let $\boldsymbol{\theta}$ denote the full set of parameters and $\boldsymbol{\theta}_0$ a subset of $\boldsymbol{\theta}$. We test two different types of hypotheses:

$$H_{0,a}: \quad \boldsymbol{\theta}_0 = \boldsymbol{0} \tag{64}$$

$$H_{0,b}: \quad \theta_i - \theta_j = 0. \tag{65}$$

The hypothesis $H_{0,a}$ is only relevant for parameters in $\boldsymbol{\alpha}^{ij}$ and $\boldsymbol{\beta}$. The hypothesis $H_{0,b}$ is only tested within each group, i.e. $\tau_{ii} = \tau_{jj}, \alpha_{ij} = \alpha_{kl}$ $(i \neq j, k \neq l$, and $(i,j) \neq (k,l))$, and $\beta_{ij} = \beta_{kl}$ (for $(i,j) \neq (k,l)$).

The variance–covariance matrix for all parameters is approximated by $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = -\boldsymbol{H}^{-1}$. The test statistic for hypothesis $H_{0,a}$ is

$$T_a = \hat{\boldsymbol{\theta}}_0^{T}\boldsymbol{\Sigma}_{\boldsymbol{\theta}_0}^{-1}\hat{\boldsymbol{\theta}}_0, \tag{66}$$

which is compared to a χ^2-distribution with $p_0 = \dim(\boldsymbol{\theta}_0)$ degrees of freedom. Here, $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_0}^{-1}$ is the part of $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ that corresponds to $\boldsymbol{\theta}_0$.

For hypothesis $H_{0,b}$, we need the variance of the difference

$$\sigma_{ij}^2 = \begin{bmatrix} 1 & -1 \end{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\theta}_{ij}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \tag{67}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_{ij}}$ is the part of $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ that corresponds to θ_i and θ_j. The test statistic for this hypothesis is

$$T_{0,b} = \frac{(\hat{\theta}_i - \hat{\theta}_j)^2}{\sigma_{ij}^2}, \tag{68}$$

which is compared to a χ^2-distribution with one degree of freedom.

Using statistical inference to remove insignificant parameters one by one leads to a multiple testing problem, as it involves testing of a large number of hypotheses. A simple solution is to require a stricter significance level for each individual comparison (e.g. using the Bonferroni

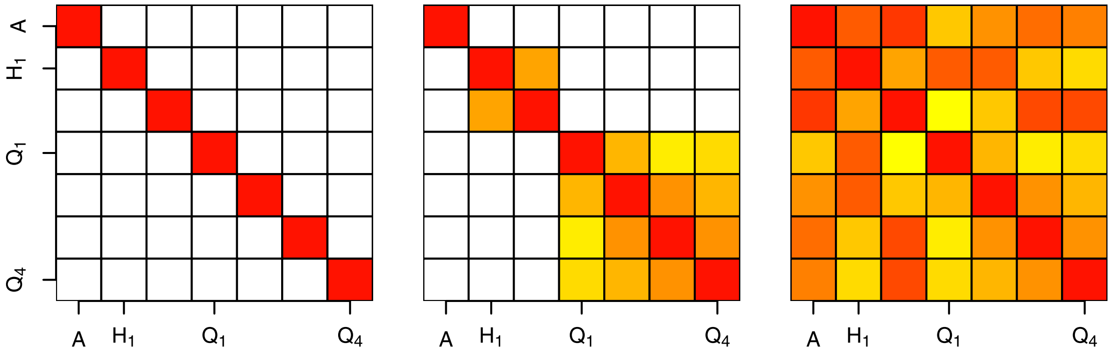

Fig. 3. Illustration of shrinkage targets for a quarterly-to-annual hierarchy.

correction), to compensate for the large number of statistical tests. By choosing a stricter significance level, the number of parameters can be further reduced.

5.1. Summary of algorithm

The entire workflow of the algorithm is summarised below:

1. Bottom level.
   (a) Estimation.
      i. Initialise a variance–covariance structure for the bottom level (i.e. an index set of non-zero elements of $\boldsymbol{S}_{11}$).
      ii. Iterate between (49) and (53) a fixed number of times (we do 10 iterations).
      iii. Use the result from 1(a)ii as the initial value for Newton's method. If Newton's method diverges, go back to 1(a)ii.
      iv. Report the parameter estimates, likelihood, and Hessian.
   (b) Model reduction.
      i. Choose in which order to test parameters for removal (we rank according to the smallest marginal Wald test statistic).
      ii. Calculate the test statistics (66) for the increasing set of parameters according to 1(b)i until significant (we use significance level 5%) and confirm by likelihood-ratio test using Step 1(a).
      iii. Calculate the test statistics for all relevant pairwise comparisons and individual parameters that can be set to zero, reduce the model, and confirm by likelihood-ratio test using 1(a).
      iv. Report the parameters and structure of $\boldsymbol{S}_{ii}$ and $\boldsymbol{S}_{ij}$.
2. Iterate through higher aggregation levels with the lower levels fixed in the same way as the bottom level, except that the estimation of $\boldsymbol{\beta}$ is included (i.e. (45) in 1(a)ii and in Newton's method).
3. Report the final result.

The result from the optimisation described above is used as input to the shrinkage approach described next.

5.2. Shrinkage

It has been shown before that shrinkage works well and is often needed in order to get more robust estimates of the variance–covariance matrix in both cross-sectional (Wickramasuriya et al., 2019) and temporal hierarchies (Nystrup et al., 2021, 2020). Shrinkage is effectively a dimensionality reduction technique. Usually this is done by shrinking the correlation matrix towards the identity matrix. The framework presented here allows us to shrink towards more general structures, similar to the idea of Nystrup et al. (2020) and Pritularga et al. (2021).

We propose to shrink towards block-diagonal matrices and diagonal matrices, as illustrated in Fig. 3. Shrinking towards the block diagonal corresponds to ignoring the correlation between forecasts from different aggregation levels, while shrinkage towards the diagonal corresponds to ignoring auto- and cross-correlation. These considerations are implemented directly through a simple modification of the likelihood by introducing weights in the following way:

$$l_s(\boldsymbol{\Sigma}; \boldsymbol{V}, \boldsymbol{w}) = l(\boldsymbol{\Sigma}; w_1\boldsymbol{V}+w_2\text{blockdiag}\boldsymbol{V}+w_3\text{diag}\boldsymbol{V}), \quad (69)$$

where $\sum_i w_i = 1$.

When $w_1 = 1$, the full variance–covariance matrix is considered; when $w_2 = 1$, the full auto-correlation structure within each aggregation level is considered, but cross-correlations between forecasts from different levels are ignored; and finally, when $w_3 = 1$, only the marginal variances are considered.

6. Reconciliation of load forecasts

In the case study, we consider hourly load data from the four price areas in Sweden shown in Fig. 4 plus the total for all of Sweden, i.e. five time series in total. We consider the period from 2016 to 2020.[1]

[1] The data were downloaded from https://www.nordpoolgroup.com/historical-market-data/.

Table 1
Cross-validation table for all the tested areas. RRMSE and RSd are shown in percent. Subscripts 1 and 2 refer to $(w_1, w_2) = (0.1, 0.01)$ and $(w_1, w_2) = (0.01, 0.002)$, respectively. The shrinkage parameter is given rather than the degrees of freedom for shrinkage.

	SE			SE1			SE2			SE3			SE4		
	df	RRMSE	Rsd	df	RRMSE	Rsd	df	RRMSE	Rsd	df	RRMSE	Rsd	df	RRMSE	Rsd
Obs-test	1830	−12.4	−24.6	1830	−8.5	−16.5	1830	−9.8	−18.8	1830	−15.9	−27.3	1830	−15.3	−27.3
Obs-train	1830	0.1	−2.6	1830	4.1	1.6	1830	2.1	5.8	1830	−2.3	−4.7	1830	−5.2	−15.6
Full-Full$_1$	402	−4.8	−6.9	306	−1.7	−5.4	413	**−4.1**	**−9.7**	413	−5.6	−8.6	464	−5.5	−6.3
Full-Full$_2$	402	**−5.3**	−8.8	306	0.8	−4.1	413	−3.1	−8.7	413	−5.8	−8.4	464	−6.6	−9.2
Full-2$_1$	219	−3.8	−6.5	180	**−3.2**	**−6.5**	237	−4.0	−9.1	240	−3.9	−6.2	264	−4.0	−4.5
Full-2$_2$	219	−3.8	−8.7	180	−2.2	−5.3	237	−3.8	−8.8	240	−4.9	−8.2	264	−3.7	−6.0
Null-Null$_1$	83	−4.3	−6.8	77	−2.7	−4.6	81	**−4.1**	−9.2	85	−4.8	−7.6	85	−4.6	−3.1
Null-Null$_2$	83	−3.7	−8.8	77	−0.7	−1.4	81	−3.1	−8.3	85	−4.0	−7.5	85	−4.1	−5.4
Shrink	0.015	**−5.3**	**−9.8**	0.016	−1.3	−4.4	0.035	−4.0	−8.8	0.015	**−7.4**	**−12.8**	0.016	**−7.0**	**−10.5**
Auto-covariance	69	−2.2	−0.9	49	−2.9	−4.7	49	−3.5	−6.6	71	−1.6	−1.7	71	−2.2	0.8
Model-AR1	60	−2.4	−1.5	60	−3.0	−4.8	60	−3.2	−6.2	60	−1.9	−2.2	60	−2.2	1.2
Diag	60	−3.1	−4.8	60	−2.4	−3.8	60	−2.6	−5.3	60	−2.9	−4.7	60	−2.6	−3.2

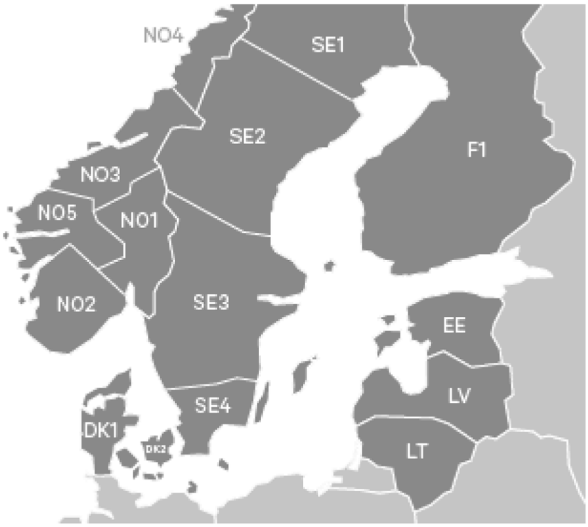

Fig. 4. Map showing the Nord Pool region, including the four price areas in Sweden.

6.1. Forecast models

The years 2016 to 2019 are used for estimating a mean-value structure for each of the five areas. These models are given by

$$\hat{y}_t = \beta_0 + \sum_{i=1}^{4} \beta_i^s \sin\left(\frac{2i\pi d_t}{366}\right) + \beta_i^c \cos\left(\frac{2i\pi d_t}{366}\right) + \gamma(h_t), \tag{70}$$

where d_t is the day of the year, and h_t is the hour of the day. The data used for modelling are the residuals from this linear model.

Due to non-stationarity, the original data have very high auto-correlation at all lags, as seen from the top panels in Fig. 5. In particular, the 24-h-lagged auto-correlation decays slowly. The residuals from the mean-value model (70) shown in the bottom panels of Fig. 5 have a faster decay in the auto-correlation function, though both 24-h and 168-h (one-week) seasonality are clearly visible.

Double-seasonal auto-regressive models of order (3,3,3) in the seasons (1,24,168) hours (except the 24-h model which has order (3,3) in (24,168) hours) are then fitted to the year 2019, for each level of the hierarchy. The residuals shown in Fig. 6 do not display clear seasonal patterns. Even though most of the systematic behaviour in the data seems to be captured by the model, it is also clear that the data contain some very large residuals in both directions. For a better understanding of the assumptions underlying the model, we assessed the proposed method in a simulation study based on real data, as described in Section 7.

6.2. Results

We use the year 2019 for estimating the variance–covariance matrix for the base forecast errors and 2020 to evaluate the method out of sample. The performance in terms of the RRMSE of the methodology is presented in Table 1 for two different choices of (w_1, w_2). In addition to RRMSE, the table includes the measure Rsd, which is calculated using observed standard error rather than RMSE. This is a measure of the risk of the method/model.

We test the following three different initial models for the variance–covariance matrix (see Fig. 2 for an illustration), each for the two different choices of w_i:

Full-Full: Reduction from a full model similar to the observed variance–covariance matrix but reduced by testing if parameters should be zero or if two parameters could be the same using likelihood-ratio and Wald tests.

Full-2: Reduction from a model with the full structure for the block diagonal, but only two entries away from the data-sharing entries for the cross-correlation. For example, in the second column of Fig. 2 this would imply that β_{14}^{HQ} and β_{21}^{HQ} would be included.

Null: Reduction from a model with only one off-diagonal element in S_{ii} and only one parameter for each cross-correlation.

Additionally, we consider the following benchmark models:

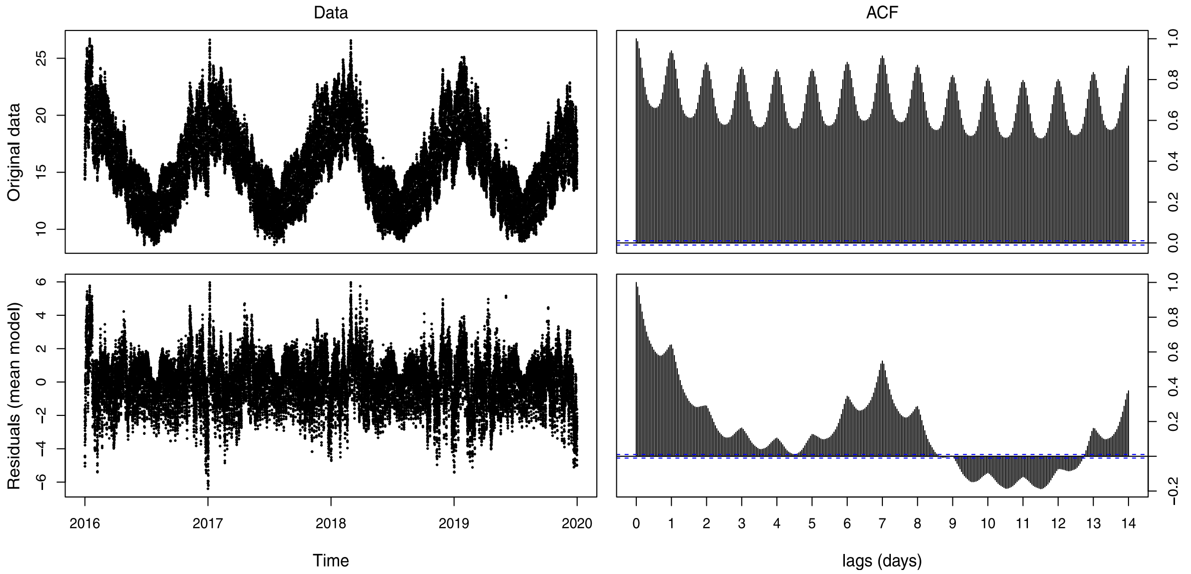

Fig. 5. Left panel: The original consumption data for all of Sweden and the residuals from the mean-value model. Right panel: Auto-correlation function for the data.

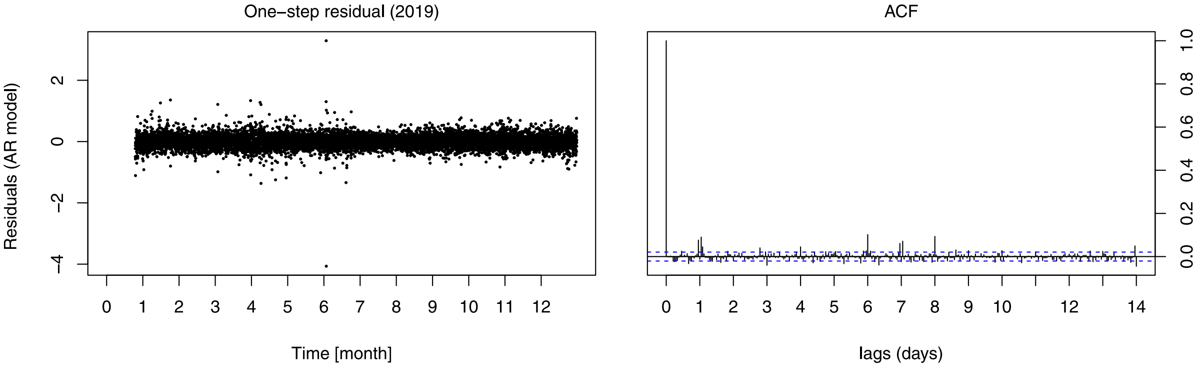

Fig. 6. One-hour-ahead forecast errors and auto-correlation function for the errors.

Obs-test: Using the variance–covariance matrix calculated on the test set. This is not a real benchmark but it provides an upper bound for the improvements.

Obs-train: Using the observed variance–covariance matrix from the training set, equivalent to the cross-covariance estimator considered by Nystrup et al. (2021, 2020).

Shrink: Using the observed variance–covariance matrix from the training set with optimal shrinkage (Ledoit & Wolf, 2003; Schäfer & Strimmer, 2005). Many consider this to be the state-of-the-art method in both cross-sectional (Wickramasuriya et al., 2019) and temporal (Nystrup et al., 2021) forecast reconciliation.

Auto-covariance: Estimating the block-diagonal matrix for each level using the presented methodology while ignoring all cross-covariances, as proposed by Nystrup et al. (2020).

Model-AR1: Using the lag-one auto-correlation obtained from the double seasonal AR model to calculate the forecast correlation matrix. This is similar to the idea behind Markov scaling proposed by Nystrup et al. (2020).

Diag: Only a diagonal variance matrix is used, thus ignoring all auto- and cross-covariances, equivalent to hierarchy variance scaling proposed by Athanasopoulos et al. (2017).

As seen from the top row of Table 1, the potential improvements in terms of RRMSE range from 8.5% to 15.9%. The actual improvements are always less than half of the potential. Using the variance–covariance matrix from the training period does not give good results, as the RRMSE is positive in three out of the five cases. This shows that some shrinkage/reduction is needed in order to get good results. The simple benchmark models, which ignore cross-correlation, consistently yield improvements in the range from 1.6% to 3.5%.

Shrinkage and Diag match the diagonal elements of the observed variance–covariance matrix. For the other models, the entire likelihood is matched with the available parameters, which can cause a slight shift of the diagonal elements. For the Auto-covariance approach, the presented algorithm is applied to each level while cross-covariances are ignored. In some cases, this leads to fewer degrees of freedom compared to some of the other simple benchmark models with similar performance. The number of parameters for the initial Auto-covariance model is 455, so the number of parameters is reduced by a factor between six and nine.

Shrinkage gives good results in most areas. Table 1 reports the shrinkage parameter in each area, rather than showing the degrees of freedom for shrinkage (which is 1830). The estimated shrinkage parameter is around 0.015 in all areas, with the exception of SE2, where it is higher. In this area the likelihood-based method outperforms shrinkage.

The likelihood-based approach generally leads to improvements in accuracy. Which of the models performs best differs between the areas. For example, Full-Full with low shrinkage has good performance in SE, while it performs poorly in SE1, where Full-2 with high shrinkage performs well. In SE2, Full-Full and Null with high shrinkage both perform well. In SE3 and SE4, the high-dimensional models have the best performance.

In all cases there is a significant parameter reduction. The initial models (before model reduction) have 1830 (Full-Full), 930 (Full-2), and 140 (Null) parameters, respectively. This means that the number of parameters is reduced by a factor of 3.9–6.0, 3.5–5.2, and 1.6–1.8, respectively, compared to the initial models. The Null model outperforms the simple benchmark models in most cases, and in many cases it performs similarly to Shrinkage, but with a much lower number of parameters.

We also see that the performance varies across the different areas. SE1 seems to be the most difficult, as none of the models are able to improve the RRMSE by more than 3.2%. This is also the area where Obs-train has the worst performance, indicating that the variance–covariance matrix changes significantly between 2019 and 2020. This favours the simple benchmark models and models with high shrinkage. At the other end of the spectrum, Obs-train performs quite well in SE4, where shrinkage and the likelihood-based models with low shrinkage have a better performance.

In general, across all models and areas, the risk measure (Rsd) and the performance measure (RRMSE) follow each other (i.e. good performance in RRMSE implies good performance in Rsd). To summarise, the likelihood-based method performs well in most areas with much fewer parameters than optimal shrinkage.

7. Simulation study

To evaluate the method, we consider a complex triple-seasonal AR model with seasonalities of 24, 168, and $168 \cdot 52$ and corresponding orders 18, 6, 12, and 1. At each level of the hierarchy, a double-seasonal AR model with seasonalities corresponding to 24 h and 168 h and order (3,3,3) is estimated based on 100 years of data. This yields 100 estimated AR processes. Fig. 7 shows the average correlation matrix for the 100 simulations, which clearly has a lot of structure.

Summary of simulation setup:

- Simulate 100 "years" of data (data-generating process $(18, 0, 0)\times(6, 0, 0)_{24}\times(12, 0, 0)_{168}\times(1, 0, 0)_{52\cdot 168}$.
- Estimate an AR process for each year (model $(3, 0, 0) \times (3, 0, 0)_{24h} \times (3, 0, 0)_{168}$, except for the daily level, which has one seasonal component with weekly season).
- Calculate 24-h forecasts in sample and find the variance–covariance for day-ahead forecasts.
- Calculate reconciled forecasts out of sample for the following year using different versions of the variance–covariance matrix.
- Compare the accuracy to the bottom-level base forecast.

In order to explore the effect of choosing different weights, these are optimised for the simple structure. The structure is such that

$$\alpha^{ii}_{kl} = 0; \quad l - k > 1, \tag{71}$$

$$\beta^{ij}_{kl} = \beta^{ij}; \quad \text{for shared observations between levels}, \tag{72}$$

$$\beta^{ij}_{kl} = 0; \quad \text{if observations are not shared between levels.} \tag{73}$$

The weights w_1 and w_2 are optimised for different lengths T of the training period for the variance–covariance matrix. Only the training set for the covariance is changed, while the training period for the AR model is kept at one year. It should be noted that the RRMSE for the variance–covariance is not a true out-of-sample evaluation, as the weights are optimised over all 100 realisations.

The average RRMSEs for optimal shrinkage and the likelihood approach are compared in Fig. 8. For large T, optimal shrinkage performs better than the likelihood approach, whereas the likelihood approach outperforms optimal shrinkage for T below 100. As T decreases, the average RRMSE for optimal shrinkage is above zero, meaning that it is worse than the base forecast at the one-hour level.

The results in Fig. 8 indicate that, when focusing on the average performance across all the forecasts, shrinkage is superior for large T. The risk that an individual forecast is worse can be evaluated by considering the variation of RRMSE or high quantiles. As seen in Fig. 9, the variation in RRMSE is much higher for the shrinkage method. The difference in performance is, at least in part, due to some very high RRMSEs. When $T = 182$, shrinkage has a better average RRMSE, but the worst case is worse than the RRMSE for the likelihood-based method.

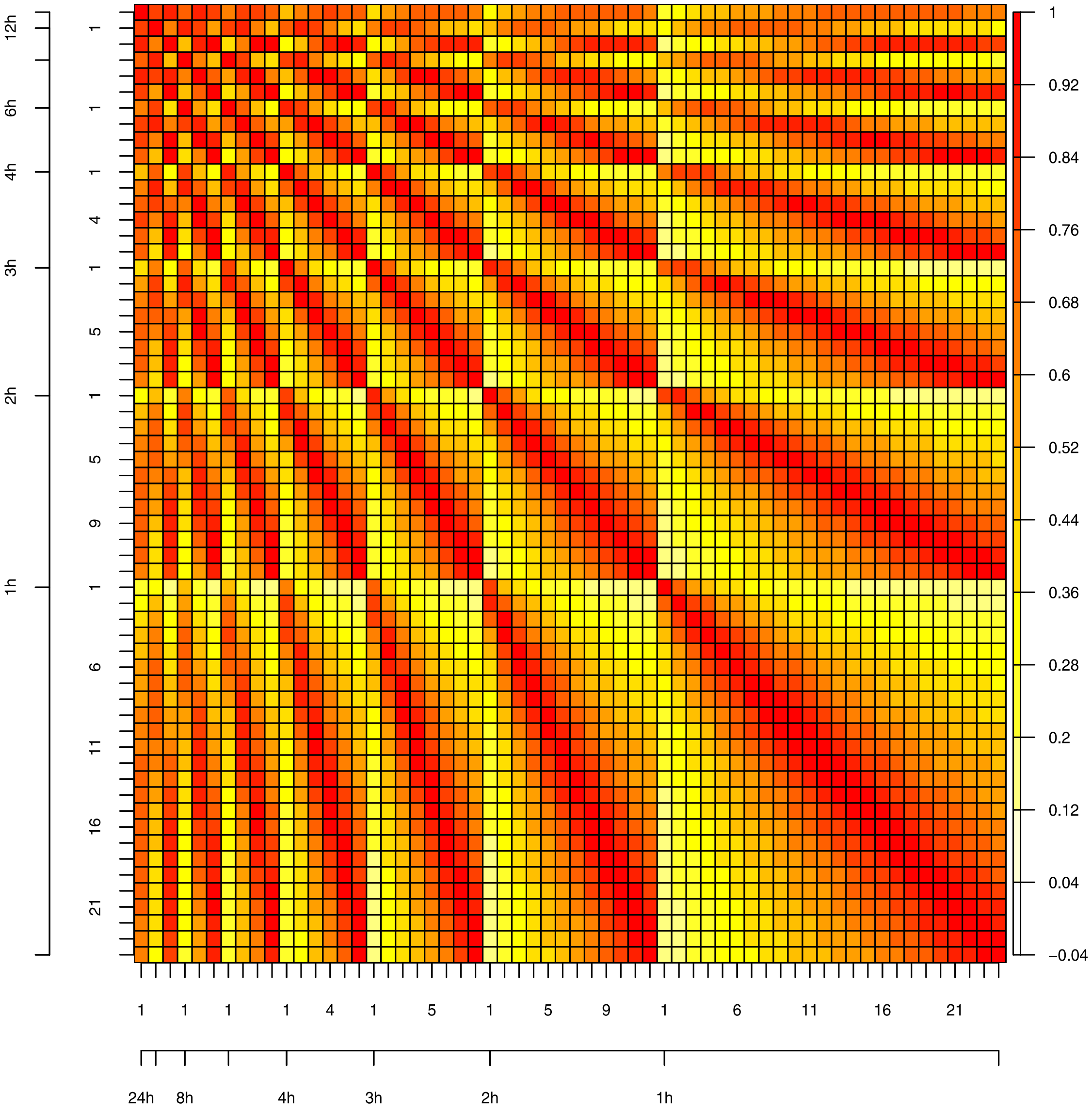

Fig. 7. Average correlation matrix for the 100 simulations.

The box-plots in Fig. 9 indicate that the lower number of parameters makes the likelihood-based method more robust than shrinkage estimation, as it reduces the risk of deteriorating forecast accuracy.

8. Conclusion

We proposed a novel framework for modelling and estimating the variance–covariance matrix used when reconciling forecasts in a temporal hierarchy. We derived a two-stage estimation procedure based on the formulated model and its likelihood function. Formulating a model with a likelihood function allowed us to apply statistical inference to identify a parsimonious parametric structure for the variance–covariance matrix. Furthermore, the likelihood-based approach offered a simple way of shrinking the variance–covariance matrix towards different targets, such as diagonal and block-diagonal matrices.

In the case study on load forecasting, the proposed method performed similarly to optimal shrinkage while requiring significantly fewer parameters. Optimal shrinkage did outperform the likelihood-based approach overall in terms of accuracy on the rather large data set. The best likelihood-based models outperformed optimal shrinkage in the most difficult cases, which illustrated the robustness of the proposed framework.

The simulation study highlighted the difficulties in estimating the variance–covariance matrix when it is of high dimension compared to the number of observations available for estimation. By simplifying its structure using

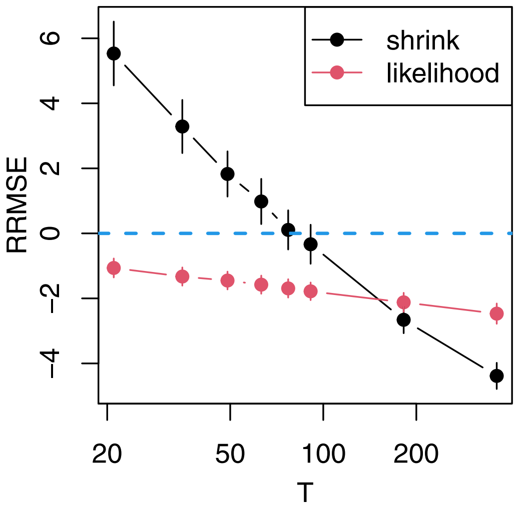

Fig. 8. RRMSE for the shrinkage- and likelihood-based methods. For the likelihood-based method, the weights have been optimised. The vertical lines indicate standard errors.

the proposed framework for model reduction based on hypothesis testing, the estimation could be made quite robust at the cost of missing some structure when more data were available for estimation. By lowering the number of parameters, the risk of deteriorating forecast accuracy through reconciliation could be reduced.

In future work we plan to explore the important choice of shrinkage parameters in greater detail. On a practical note, we expect a more efficient implementation to be able to reduce the computation time for the estimation procedure. On the modelling side, more research is needed on the choice of suitable variance–covariance structures given the data available. We believe that the formalism presented in this article contributes to the understanding of forecast reconciliation, and that it is an important step towards a model-based approach for describing noise propagation through aggregation in hierarchical forecasting.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank an Associate Editor and two anonymous reviewers for detailed feedback. The authors received funding from the following projects: ebalance-plus (H2020 no. 864283), Sustainable plus energy neighbourhoods (syn.ikia) (H2020 No. 869918), ARV (H2020 No. 101036723), Top-Up (Innovation Fund Denmark (IFD) 9045-00017B), SEM4Cities (IFD 0143-0004), Flexible Energy Denmark (IFD 8090-00069B), and IEA Wind Task 51 (EUDP nr. 134-22015). Those contributions are gratefully acknowledged.

Appendix A. Derivations for motivating example

In this appendix, we go through the detailed calculations for the example in Section 3. For the data-generating process in (2), the two-step-ahead observations of the process are given by

$$y^{\mathrm{H}}_{2t+1} = \phi_1 y^{\mathrm{H}}_{2t} + \phi_2 y^{\mathrm{H}}_{2t-1} + \epsilon^{\mathrm{H}}_{2t+1}, \tag{A.1}$$

$$\begin{aligned} y^{\mathrm{H}}_{2t+2} &= \phi_1 y^{\mathrm{H}}_{2t+1} + \phi_2 y^{\mathrm{H}}_{2t} + \epsilon^{\mathrm{H}}_{2t+2} \\ &= (\phi_1^2 + \phi_2) y^{\mathrm{H}}_{2t} + \phi_1\phi_2 y^{\mathrm{H}}_{2t-1} + \phi_1 \epsilon^{\mathrm{H}}_{2t+1} + \epsilon^{\mathrm{H}}_{2t+2}. \end{aligned} \tag{A.2}$$

The data-generating process for the annual level is

$$\begin{aligned} y^{\mathrm{A}}_{2t+2} &= y^{\mathrm{H}}_{2t+1} + y^{\mathrm{H}}_{2t+2} \\ &= (\phi_1 + \phi_1^2 + \phi_2) y^{\mathrm{H}}_{2t} + (\phi_1\phi_2 + \phi_2) y^{\mathrm{H}}_{2t-1} \\ &\quad + (1 + \phi_1)\epsilon^{\mathrm{H}}_{2t+1} + \epsilon^{\mathrm{H}}_{2t+2}. \end{aligned} \tag{A.3}$$

It follows that the data-generating process for the full three-dimensional vector can be written as

$$\boldsymbol{y}_{2t+2} = \begin{bmatrix} y^{\mathrm{A}}_{2t+2} \\ y^{\mathrm{H}}_{2t+1} \\ y^{\mathrm{H}}_{2t+2} \end{bmatrix}$$

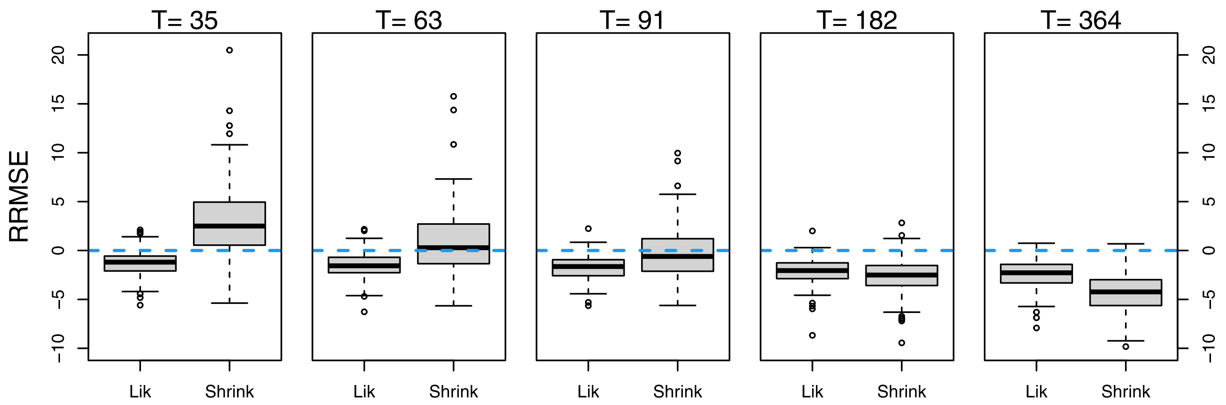

Fig. 9. Box-plots of RRMSEs for different numbers of observations used to estimate the variance–covariance matrix (using the optimal weights).

$$
\begin{aligned}
&= \begin{bmatrix} \phi_1\phi_2+\phi_2 & \phi_1+\phi_1^2+\phi_2 \\ \phi_2 & \phi_1 \\ \phi_1\phi_2 & \phi_1^2+\phi_2 \end{bmatrix} \begin{bmatrix} y^{\mathrm{H}}_{2t-1} \\ y^{\mathrm{H}}_{2t} \end{bmatrix} \\
&\quad + \begin{bmatrix} \phi_1+1 & 1 \\ 1 & 0 \\ \phi_1 & 1 \end{bmatrix} \begin{bmatrix} \epsilon^{\mathrm{H}}_{2t+1} \\ \epsilon^{\mathrm{H}}_{2t+2} \end{bmatrix} \\
&= \boldsymbol{\Phi} \boldsymbol{y}^{\mathrm{H}}_{2t} + \boldsymbol{\Phi}_{\epsilon} \boldsymbol{\epsilon}^{\mathrm{H}}_{2t+2}, \qquad \text{(A.4)}
\end{aligned}
$$

where $\boldsymbol{y}^{\mathrm{H}}_{2t} = [y^{\mathrm{H}}_{2t-1}, y^{\mathrm{H}}_{2t}]^T$ and $\boldsymbol{\epsilon}^{\mathrm{H}}_{2t+2} = [\epsilon^{\mathrm{H}}_{2t+1}, \epsilon^{\mathrm{H}}_{2t+2}]^T$.

For modelling purposes we assume AR(1) models at both levels, as outlined in (3) and (4). Given $\tilde{\phi}_{\mathrm{A}}$ and $\tilde{\phi}_{\mathrm{H}}$, we can write the forecast in matrix–vector notation as

$$
\hat{\boldsymbol{y}}_{2t+2|2t} = \begin{bmatrix} \hat{y}^{\mathrm{A}}_{2t+2|2t} \\ \hat{y}^{\mathrm{H}}_{2t+1|2t} \\ \hat{y}^{\mathrm{H}}_{2t+2|2t} \end{bmatrix} = \begin{bmatrix} \tilde{\phi}_{\mathrm{A}} & \tilde{\phi}_{\mathrm{A}} \\ 0 & \tilde{\phi}_{\mathrm{H}} \\ 0 & \tilde{\phi}^2_{\mathrm{H}} \end{bmatrix} \begin{bmatrix} y^{\mathrm{H}}_{2t-1} \\ y^{\mathrm{H}}_{2t} \end{bmatrix} = \boldsymbol{\Gamma} \boldsymbol{y}^{\mathrm{H}}_{2t}. \qquad \text{(A.5)}
$$

We can find the variance of the base forecast error as follows:

$$
\begin{aligned}
\boldsymbol{\Sigma} &= V[\boldsymbol{y}_{2t+2} - \hat{\boldsymbol{y}}_{2t+2|2t}] \\
&= V[\boldsymbol{\Phi}\boldsymbol{y}^{\mathrm{H}}_{2t} + \boldsymbol{\Phi}_{\epsilon}\boldsymbol{\epsilon}^{\mathrm{H}}_{2t+2} - \boldsymbol{\Gamma}\boldsymbol{y}^{\mathrm{H}}_{2t}] \\
&= V[(\boldsymbol{\Phi} - \boldsymbol{\Gamma})\boldsymbol{y}^{\mathrm{H}}_{2t} + \boldsymbol{\Phi}_{\epsilon}\boldsymbol{\epsilon}^{\mathrm{H}}_{2t+2}] \\
&= (\boldsymbol{\Phi} - \boldsymbol{\Gamma})V[\boldsymbol{y}^{\mathrm{H}}_{2t}](\boldsymbol{\Phi} - \boldsymbol{\Gamma})^T + \sigma^2 \boldsymbol{\Phi}_{\epsilon}\boldsymbol{\Phi}^T_{\epsilon}, \qquad \text{(A.6)}
\end{aligned}
$$

with

$$
V[\boldsymbol{y}^{\mathrm{H}}_{2t}] = \begin{bmatrix} \gamma_0 & \gamma_1 \\ \gamma_1 & \gamma_0 \end{bmatrix}, \qquad \text{(A.7)}
$$

where γ_i is the auto-covariance function for the data-generating process at lag i.

The variance of the reconciled forecast error is

$$
\begin{aligned}
\tilde{\boldsymbol{\Sigma}} &= V[\boldsymbol{y}_{2t+2} - \boldsymbol{SP}\hat{\boldsymbol{y}}_{2t+2|2t}] \\
&= V[\boldsymbol{\Phi}\boldsymbol{y}^{\mathrm{H}}_{2t} + \boldsymbol{\Phi}_{\epsilon}\boldsymbol{\epsilon}^{\mathrm{H}}_{2t+2} - \boldsymbol{SP}\boldsymbol{\Gamma}\boldsymbol{y}^{\mathrm{H}}_{2t}] \\
&= V[(\boldsymbol{\Phi} - \boldsymbol{SP}\boldsymbol{\Gamma})\boldsymbol{y}^{\mathrm{H}}_{2t} + \boldsymbol{\Phi}_{\epsilon}\boldsymbol{\epsilon}^{\mathrm{H}}_{2t+2}] \\
&= (\boldsymbol{\Phi} - \boldsymbol{SP}\boldsymbol{\Gamma})V[\boldsymbol{y}^{\mathrm{H}}_{2t}](\boldsymbol{\Phi} - \boldsymbol{SP}\boldsymbol{\Gamma})^T + \sigma^2 \boldsymbol{\Phi}_{\epsilon}\boldsymbol{\Phi}^T_{\epsilon}. \qquad \text{(A.8)}
\end{aligned}
$$

Here $\boldsymbol{SP}$ is the usual projection matrix for forecast reconciliation, and $\boldsymbol{P}$ is defined by (9).

In Appendix B, we show that (A.8) is equivalent to

$$
\tilde{\boldsymbol{\Sigma}} = \boldsymbol{SP\Sigma P}^T\boldsymbol{S}^T, \qquad \text{(A.9)}
$$

as derived by Wickramasuriya et al. (2019) (Lemma 1). An advantage of the formulation in (A.8) is that we can clearly see that the lower limit of the reconciled forecast error variance is $\sigma^2\boldsymbol{\Phi}_{\epsilon}\boldsymbol{\Phi}^T_{\epsilon}$, which is obtained when $\boldsymbol{SP\Gamma} = \boldsymbol{\Phi}$, i.e. when $E[\boldsymbol{y}_{2t+2}|\boldsymbol{y}_{2t}] = \boldsymbol{SP}\hat{\boldsymbol{y}}_{2t+2|2t}$.

In order to find the coefficients $\tilde{\phi}_{\mathrm{A}}$ and $\tilde{\phi}_{\mathrm{H}}$, we need the auto-covariance function up to lag three. Using the Yule–Walker equations, these are given by

$$
\begin{aligned}
\rho_1 &= \frac{\phi_1}{1-\phi_2}, \qquad \text{(A.10)} \\
\rho_2 &= \frac{\phi_1^2 + \phi_2(1-\phi_2)}{1-\phi_2}, \qquad \text{(A.11)} \\
\rho_3 &= \phi_1\rho_2 + \phi_2\rho_1. \qquad \text{(A.12)}
\end{aligned}
$$

The auto-variance is $\gamma_i = \gamma_0\rho_i$, with

$$
\gamma_0 = \frac{1}{1 - \phi_1\rho_2 - \phi_2\rho_1}. \qquad \text{(A.13)}
$$

The best estimate of $\tilde{\phi}_{\mathrm{H}}$ is ρ_1.

The auto-covariance at lag zero and one for the annual process is

$$
\begin{aligned}
\gamma^{\mathrm{A}}_0 &= Cov[y^{\mathrm{A}}_{2t}, y^{\mathrm{A}}_{2t}] \\
&= Cov[y^{\mathrm{H}}_{2t} + y^{\mathrm{H}}_{2t-1}, y^{\mathrm{H}}_{2t} + y^{\mathrm{H}}_{2t-1}] \\
&= 2\gamma_0 + 2\gamma_1, \qquad \text{(A.14)}
\end{aligned}
$$

and

$$
\begin{aligned}
\gamma^{\mathrm{A}}_1 &= Cov[y^{\mathrm{A}}_{2t}, y^{\mathrm{A}}_{2(t-1)}] \\
&= Cov[y^{\mathrm{H}}_{2t} + y^{\mathrm{H}}_{2t-1}, y^{\mathrm{H}}_{2t-2} + y^{\mathrm{H}}_{2t-3}] \\
&= \gamma_1 + 2\gamma_2 + \gamma_3. \qquad \text{(A.15)}
\end{aligned}
$$

The best estimate of $\tilde{\phi}_{\mathrm{A}}$ is the lag-one auto-correlation for the annual process

$$
\rho^{\mathrm{A}}_1 = \frac{\gamma_1 + 2\gamma_2 + \gamma_3}{2(\gamma_0 + \gamma_1)}. \qquad \text{(A.16)}
$$

This concludes the needed derivations, as all values needed for the example can be calculated with the above.

Appendix B. Proof that $\boldsymbol{SP\Sigma P}^T\boldsymbol{S}^T = \tilde{\boldsymbol{\Sigma}}$

The purpose of this section is to show that

$$
\boldsymbol{SP\Sigma P}^T\boldsymbol{S}^T = \tilde{\boldsymbol{\Sigma}}, \qquad \text{(B.1)}
$$

where $\boldsymbol{\Sigma}$ is defined by (A.6) and $\tilde{\boldsymbol{\Sigma}}$ is defined by (A.8). Hence there is no contradiction between the result given here and Lemma 1 of Wickramasuriya et al. (2019).

Using (A.6) we can write

$$
\begin{aligned}
\boldsymbol{SP\Sigma P}^T\boldsymbol{S}^T &= \boldsymbol{SP}(\boldsymbol{\Phi} - \boldsymbol{\Gamma})V[\boldsymbol{y}^{\mathrm{H}}_{2t}](\boldsymbol{\Phi} - \boldsymbol{\Gamma})^T\boldsymbol{P}^T\boldsymbol{S}^T \\
&\quad + \sigma^2\boldsymbol{SP}\boldsymbol{\Phi}_{\epsilon}\boldsymbol{\Phi}^T_{\epsilon}\boldsymbol{P}^T\boldsymbol{S}^T \\
&= \boldsymbol{SP\Phi}V[\boldsymbol{y}^{\mathrm{H}}_{2t}]\boldsymbol{\Phi}^T\boldsymbol{P}^T\boldsymbol{S}^T + \boldsymbol{SP\Gamma}V[\boldsymbol{y}^{\mathrm{H}}_{2t}]\boldsymbol{\Gamma}^T\boldsymbol{P}^T\boldsymbol{S}^T \\
&\quad - \boldsymbol{SP\Phi}V[\boldsymbol{y}^{\mathrm{H}}_{2t}]\boldsymbol{\Gamma}^T\boldsymbol{P}^T\boldsymbol{S}^T - \\
&\quad \boldsymbol{SP\Gamma}V[\boldsymbol{y}^{\mathrm{H}}_{2t}]\boldsymbol{\Phi}^T\boldsymbol{P}^T\boldsymbol{S}^T + \sigma^2\boldsymbol{SP}\boldsymbol{\Phi}_{\epsilon}\boldsymbol{\Phi}^T_{\epsilon}\boldsymbol{P}^T\boldsymbol{S}^T. \qquad \text{(B.2)}
\end{aligned}
$$

Next we write the equation $\boldsymbol{SPS} = \boldsymbol{S}$ in detail as

$$
\begin{aligned}
\boldsymbol{SPS} &= \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \\
&= \begin{bmatrix} P_{11} + P_{21} + P_{12} + P_{22} & P_{11} + P_{21} + P_{13} + P_{23} \\ P_{11} + P_{12} & P_{11} + P_{13} \\ P_{21} + P_{22} & P_{21} + P_{23} \end{bmatrix} \\
&= \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \text{(B.3)}
\end{aligned}
$$

and

$$
\boldsymbol{SP} = \begin{bmatrix} P_{11} + P_{21} & P_{11} + P_{22} & P_{13} + P_{23} \\ P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \end{bmatrix}. \qquad \text{(B.4)}
$$

Using (A.4) we can write the elements of $\boldsymbol{SP\Phi}$ as

$$\begin{aligned}(\boldsymbol{SP\Phi})_{11} &= \phi_1\phi_2(P_{11}+P_{21}+P_{13}+P_{23}) \\ &\quad + \phi_2(P_{11}+P_{21}+P_{12}+P_{22}) \\ &= \phi_1\phi_2+\phi_2 \end{aligned} \tag{B.5}$$

$$\begin{aligned}(\boldsymbol{SP\Phi})_{12} &= \phi_1(P_{11}+P_{21}+P_{12}+P_{22}) \\ &\quad + (\phi_2+\phi_1^2)(P_{11}+P_{21}+P_{13}+P_{23}) \\ &= \phi_1+\phi_2+\phi_1^2 \end{aligned} \tag{B.6}$$

$$\begin{aligned}(\boldsymbol{SP\Phi})_{21} &= \phi_1\phi_2(P_{11}+P_{13})+\phi_2(P_{21}+P_{22}) \\ &= \phi_2 \end{aligned} \tag{B.7}$$

$$\begin{aligned}(\boldsymbol{SP\Phi})_{22} &= \phi_1(P_{11}+P_{12})+(\phi_2+\phi_1^2)(P_{11}+P_{13}) \\ &= \phi_1 \end{aligned} \tag{B.8}$$

$$\begin{aligned}(\boldsymbol{SP\Phi})_{31} &= \phi_1\phi_2(P_{21}+P_{23})+\phi_2(P_{21}+P_{22}) \\ &= \phi_1\phi_2 \end{aligned} \tag{B.9}$$

$$\begin{aligned}(\boldsymbol{SP\Phi})_{32} &= \phi_1(P_{21}+P_{22})+(\phi_1^2+\phi_2)(P_{21}+P_{23}) \\ &= \phi_1^2+\phi_2 \end{aligned} \tag{B.10}$$

and hence,

$$\boldsymbol{SP\Phi} = \boldsymbol{\Phi}. \tag{B.11}$$

Additionally, using (A.4) we can write

$$\boldsymbol{\Phi}_\epsilon = \boldsymbol{S} + \phi_1\begin{bmatrix}1 & 0\\ 0 & 0\\ 1 & 0\end{bmatrix} \tag{B.12}$$

and

$$\boldsymbol{SP}\begin{bmatrix}1 & 0\\ 0 & 0\\ 1 & 0\end{bmatrix} = \begin{bmatrix}P_{11}+P_{21}+P_{13}+P_{23} & 0\\ P_{11}+P_{13} & 0\\ P_{21}+P_{23} & 0\end{bmatrix} = \begin{bmatrix}1 & 0\\ 0 & 0\\ 1 & 0\end{bmatrix} \tag{B.13}$$

and hence,

$$\begin{aligned}\boldsymbol{SP\Phi}_\epsilon &= \boldsymbol{SPS} + \phi_1\boldsymbol{SP}\begin{bmatrix}1 & 0\\ 0 & 0\\ 1 & 0\end{bmatrix} \\ &= \boldsymbol{S} + \phi_1\begin{bmatrix}1 & 0\\ 0 & 0\\ 1 & 0\end{bmatrix} \\ &= \boldsymbol{\Phi}_\epsilon. \end{aligned} \tag{B.14}$$

Therefore,

$$\begin{aligned}\boldsymbol{SP\Sigma P}^T\boldsymbol{S}^T &= \boldsymbol{SP\Phi}V[\boldsymbol{y}_{2t}^{\mathrm{H}}]\boldsymbol{\Phi}^T\boldsymbol{P}^T\boldsymbol{S}^T + \boldsymbol{SP\Gamma}V[\boldsymbol{y}_{2t}^{\mathrm{H}}]\boldsymbol{\Gamma}^T\boldsymbol{P}^T\boldsymbol{S}^T \\ &\quad - \boldsymbol{SP\Phi}V[\boldsymbol{y}_{2t}^{\mathrm{H}}]\boldsymbol{\Gamma}^T\boldsymbol{P}^T\boldsymbol{S}^T - \\ &\quad \boldsymbol{SP\Gamma}V[\boldsymbol{y}_{2t}^{\mathrm{H}}]\boldsymbol{\Phi}^T\boldsymbol{P}^T\boldsymbol{S}^T + \sigma^2\boldsymbol{SP\Phi}_\epsilon\boldsymbol{\Phi}_\epsilon^T\boldsymbol{P}^T\boldsymbol{S}^T \\ &= \boldsymbol{\Phi}V[\boldsymbol{y}_{2t}^{\mathrm{H}}]\boldsymbol{\Phi}^T + \boldsymbol{SP\Gamma}V[\boldsymbol{y}_{2t}^{\mathrm{H}}]\boldsymbol{\Gamma}^T\boldsymbol{P}^T\boldsymbol{S}^T \\ &\quad - \boldsymbol{\Phi}V[\boldsymbol{y}_{2t}^{\mathrm{H}}]\boldsymbol{\Gamma}^T\boldsymbol{P}^T\boldsymbol{S}^T - \\ &\quad \boldsymbol{SP\Gamma}V[\boldsymbol{y}_{2t}^{\mathrm{H}}]\boldsymbol{\Phi}^T + \sigma^2\boldsymbol{\Phi}_\epsilon\boldsymbol{\Phi}_\epsilon^T \\ &= (\boldsymbol{\Phi}-\boldsymbol{SP\Gamma})V[\boldsymbol{y}_{2t}^{\mathrm{H}}](\boldsymbol{\Phi}-\boldsymbol{SP\Gamma})^T + \sigma^2\boldsymbol{\Phi}_\epsilon\boldsymbol{\Phi}_\epsilon^T \\ &= \tilde{\boldsymbol{\Sigma}}. \end{aligned} \tag{B.15}$$

This concludes the proof.

References

Amemiya, T., & Wu, R. Y. (1972). The effect of aggregation on prediction in the autoregressive model. *Journal of the American Statistical Association*, *67*(339), 628–632.

Athanasopoulos, G., Ahmed, R. A., & Hyndman, R. J. (2009). Hierarchical forecasts for Australian domestic tourism. *International Journal of Forecasting*, *25*(1), 146–166.

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., & Petropoulos, F. (2017). Forecasting with temporal hierarchies. *European Journal of Operational Research*, *262*(1), 60–74.

Bergsteinsson, H. G., Møller, J. K., Nystrup, P., Pálsson, Ó. P., Guericke, D., & Madsen, H. (2021). Heat load forecasting using adaptive temporal hierarchies. *Applied Energy*, *292*, Article 116872.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583.

Di Fonzo, T., & Girolimetto, D. (2022). Forecast combination-based forecast reconciliation: Insights and extensions. *International Journal of Forecasting*, http://dx.doi.org/10.1016/j.ijforecast.2022.07.001.

Eckert, F., Hyndman, R. J., & Panagiotelis, A. (2021). Forecasting Swiss exports using Bayesian forecast reconciliation. *European Journal of Operational Research*, *291*(2), 693–710.

Gross, C. W., & Sohl, J. E. (1990). Disaggregation methods to expedite product line forecasting. *Journal of Forecasting*, *9*(3), 233–254.

Hollyman, R., Petropoulos, F., & Tipping, M. E. (2021). Understanding forecast reconciliation. *European Journal of Operational Research*, *294*(1), 149–160.

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., & Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, *55*(9), 2579–2589.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688.

Hyndman, R. J., Lee, A. J., & Wang, E. (2016). Fast computation of reconciled forecasts for hierarchical and grouped time series. *Computational Statistics & Data Analysis*, *97*, 16–32.

Jeon, J., Panagiotelis, A., & Petropoulos, F. (2019). Probabilistic forecast reconciliation with applications to wind power and electric load. *European Journal of Operational Research*, *279*(2), 364–379.

Ledoit, O., & Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, *10*(5), 603–621.

Madsen, H., & Thyregod, P. (2011). *Texts in statistical science, Introduction to general and generalized linear models*. CRC Press.

Møller, J. K., Nystrup, P., & Madsen, H. (2022). Supplementaty material for "Likelihood-based inference in temporal hierarchies". http://people.compute.dtu.dk/jkmo/.

Nystrup, P., Lindström, E., Møller, J. K., & Madsen, H. (2021). Dimensionality reduction in forecasting with temporal hierarchies. *International Journal of Forecasting*, *37*(3), 1127–1146.

Nystrup, P., Lindström, E., Pinson, P., & Madsen, H. (2020). Temporal hierarchies with autocorrelation for load forecasting. *European Journal of Operational Research*, *280*(3), 876–888.

Panagiotelis, A., Athanasopoulos, G., Gamakumara, P., & Hyndman, R. J. (2021). Forecast reconciliation: A geometric view with new insights on bias correction. *International Journal of Forecasting*, *37*(1), 343–359.

Pritularga, K. F., Svetunkov, I., & Kourentzes, N. (2021). Stochastic coherency in forecast reconciliation. *International Journal of Production Economics*, *240*, Article 108221.

Schäfer, J., & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, *4*(1), 32.

Timmermann, A. (2006). Forecast combinations. In G. Elliott, C. W. J. Granger, & A. Timmermann (Eds.), *Handbook of economic forecasting, Vol. 1* (pp. 135–196). Elsevier: Amsterdam.

Varin, C., Reid, N., & Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinicia*, *21*(1), 5–42.

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

Yang, D., Quan, H., Disfani, V. R., & Rodríguez-Gallegos, C. D. (2017). Reconciling solar forecasts: Temporal hierarchy. *Solar Energy*, *158*, 332–346.